Math 235: Mathematical Problem Solving

Darij Grinberg*

December 25, 2021 (unfinished!)

Contents

*Drexel University, Korman Center, 15 S 33rd Street, Philadelphia PA, 19104, USA

Preface

This is the text accompanying my Math 235 (Mathematical Problem Solving) class at Drexel University in Fall 2020. The website of this class can be found at

`http://www.cip.ifi.lmu.de/~grinberg/t/20f` .

This document is a work in progress. It might become a textbook one day, but for now is a construction zone.

Please report any errors you find to `darijgrinberg@gmail.com` **.**

What is this?

This course is an introduction to the art of mathematical problem solving. This means solving the type of problems that usually are posed in competitions (like the Putnam Competition or the International Mathematical Olympiad) or in journals (like the American Mathematical Monthly or Crux Mathematicorum): self-contained mathematical questions that do not require any specialized knowledge (beyond the basic undergraduate curriculum) to understand, but often a significant amount of ingenuity, creativity and effort to solve, although the resulting solutions again are readable without specialized knowledge.

There is no panacea for solving such problems; the hardest ones constitute serious challenges for even the most well-trained solvers. Yet there is a variety of ideas, techniques and tools that can help attacking these problems; while they offer no

guarantee of success, they are sufficiently useful that they nowadays get systematically taught to olympiad participants in many countries and universities. Most of these techniques are also helpful in mathematical research – even though the typical research question poses a rather different kind of challenge than a competition problem.[1]

Some of these techniques are very basic. Mathematical induction, for example, is fundamental to all of mathematics. The Pigeonhole Principle says (roughly speaking) that if you put 6 or more pigeons into 5 pigeonholes, then at least two pigeons will be roommates. The Extremal Principle says that (at least) one of the pigeons has the longest beak. These principles are neither surprising nor deep; but they can often be applied in unexpected and surprising ways. We will study them mainly by exploring their applications, as there is little to say about these principles themselves.

Another part of the problem solving toolkit is a selection of "little" theories ("little" in the sense of not going deep; typically, the main results and their proofs fit on 20 pages at most). The most well-known such theory is elementary number theory, due to a large extent to Euclid (or to whoever discovered the results in his *Elements*) and expanded somewhat by Fermat, Euler and Gauss. The lore of polynomials (commonly over the rational or complex numbers) can be regarded as another such theory, with results like Viete's theorems. The art of elementary inequalities (like Cauchy–Schwarz, Hölder, rearrangement), too, can be counted as a theory.

The tools and theories we will see are thus a cross-section of elementary mathematics. This text will omit most of analysis (which will only appear in some exercises; see [GelAnd17] for a lot more) and all of elementary geometry (which is its own subject, worthy of its own books; see [Honsbe12] for a selection of neat results and [Hadama09] for a systematic introduction). We will also not use abstract algebra beyond a few basic ideas that can be explained from scratch. As to linear algebra, we will occasionally make use of it, but we will not introduce it, as there are enough good texts for that.

This is not the first introductory text on mathematical problem solving, and has no ambitions to be the last. The classics in this genre are Polya's [Polya81] and Engel's [Engel98][2]; many other sources have been published since, including

[1] For starters, a researcher attacking a research question does not know whether it has an answer and, if it has one, how difficult it will be to find; meanwhile, questions posed in competitions have been "pre-solved" by their proposers and/or the selection committees, and usually have a somewhat predictable level of difficulty. Also, one of the frequent challenges in research is asking the right question; but competitions preempt this challenge (although it tends to resurface in the solving process, as one looks for useful auxiliary results).

However, competition problems can (and often do) have a lot in common with research questions; in particular, several have crystallized out of the research of their proposers. For the solver, mathematical competitions can provide inspiration and training for proper research; when designed well, they are research "writ small".

[2] The even earlier book [Polya73] by Polya is, in some ways, a harbinger of this kind of book, although in itself it is too short (just a few dozen solved problems), too philosophical (replete

[GelAnd17], [Zeitz17], [ZawHit09] and [Chen20, OTIS Excerpts][3]. There is also a plenitude of problem books – i.e., collections of (solved) problems either taken from a specific contest (or journal) over the years (such as [DJMP11] or [Strasz65], or the "Hungarian Problem Books" published by the MAA), or chosen from specific subjects (the classics in the latter genre are [DynUsp06], [ShChYa62], [YagYag64] and [YagYag67]). The present text intends to differ from these both in the choice of subjects and in the choice of examples and problems (based on my personal taste); that said, there is likely to be more overlap than difference, since both the most important techniques and their most prominent examples are generally agreed upon.

Prerequisites

I expect you, the reader, to be familiar with mathematical proofs. In particular, you should be able to

- apply (and verify) arguments by mathematical induction, by contradiction, by case distinction and by similar fundamental tactics,
- write up your proofs at a reasonable level of mathematical rigor, and read similarly rigorous proofs written by others (as frequently seen in textbooks written for mathematics majors),
- understand and use basic mathematical language such as the summation sign[4] or set builder notation[5].

If not, there is a whole industry of texts from which these skills can be learned (such as [LeLeMe16, Chapters 1–5], [Day16], [Hammac15], [Newste20, Part I and

with didactical comments and Socratic dialogues) and too basic (typical strategies suggested are "do you know a related problem?" and "introduce suitable notations"). Even in the much more content-rich and concrete [Polya81], Polya spends quite a lot of space attempting to construct a general theory of mathematical problem solving, with input from philosophy and psychology (two chapters are called "Rules of Discovery?" and "On Learning, Teaching, and Learning Teaching", although they are more hands-on than the names may suggest).

Engel's main innovation in [Engel98] was to scrap the philosophy and fill as much of his book as possible with actual problems and solutions. The problem solving strategies he presents are mathematical ones, and they are taught by doing (worked examples, then problems). The occasional personal story or heuristical remark gives some variety (but the mathematics itself is already pretty varied, as one and the same strategy often has applications in very different fields).

In the notes you are now reading, I intend to stick to Engel's paradigm (but aim at a higher level of detail).

[3] also, various books in foreign languages: e.g., [Grinbe08] and [Carl17] (in German) and [Jarvin20] (in Finnish) and [DLPS16] (in Dutch)

[4] for example, $\sum_{k=a}^{b} \frac{1}{k} = \frac{1}{a} + \frac{1}{a+1} + \cdots + \frac{1}{b}$

[5] for example, $\{i \in \mathbb{Z} \mid i \text{ is even}\}$ for the set of all even integers

Appendices A–B] and [Swanso20, Chapters 1–3], just to mention a few freely available ones[6]).

Acknowledgments

I thank Natalia Grinberg, Sam Hopkins, Pasha Pylyavskyy, Tom Roby and David Speyer for helpful conversations on the problem solving literature.

[6] I recommend the MIT text [LeLeMe16] in particular for its lively writing and its enjoyable exercises. It is long and goes far beyond the basics; it has significant overlap with what I will be doing in this class.

Newstead's [Newste20] (work in progress) is also far more than an introduction to proofs. It covers a lot of the "no man's land" between such an introduction and advanced courses.

Swanson's [Swanso20] is focussed on setting up the prerequisites for a rigorous treatment of analysis.

1. Introduction

1.1. Homework set #0: Diagnostic

This is a special problem set: Its main purpose is to give you an idea of what is to come in this course, and to give me an idea of your level of familiarity with certain things (including proof writing). Do not expect to solve all these problems.

This problem set **(and only this one!)** will be graded **on completion** – each problem will be graded out of 2; any serious work on a problem (e.g., a partial solution, a particular case solved, or even an attempt that shows you have thought about it) will get 2 points. (All future problem sets will be graded **on correctness**, with a correct and well-written solution worth 10 points.)

Tips:

- Ways to make partial progress can include
  - trying out small values of n (in problems that depend on an integer n – even if it is not explicitly mentioned[7]),
  - replacing some numbers by smaller numbers (to make the problem easier to check on examples),
  - ruling some cases out, or
  - rewriting some of the expressions involved in the problem.
- You don't need to be too detailed in your solutions (nor will I always be in mine). If you solve (say) Exercise 1.1.2 by induction on n, I trust you that you can do the base case without you going into the details. Likewise, you don't need to show any intermediate steps when claiming algebraic identities like $\left(a^2+b^2+c^2\right)-(bc+ca+ab)=\frac{1}{2}\left((b-c)^2+(c-a)^2+(a-b)^2\right)$; it is clear enough that the reader can just expand both sides and compare. (Thus, your solution to Exercise 1.1.7 can consist of a single equation.) Details are needed when a reader (who has been following the course, but doesn't already know the solution of the specific problem) might realistically stumble or get stuck.
- I **will** proofread your solutions, so this is a good chance to get feedback on your writing before we get to homework worth more serious points.

Exercise 1.1.1. Let n be an even positive integer. Find a $q \in \{1,2,\ldots,2n\}$ such that

$$\frac{1!\cdot 2!\cdot\cdots\cdot(2n)!}{q!} \text{ is a perfect square.}$$

[7] E.g., a problem about a polynomial always has an "implicit n", namely the degree of the polynomial involved. Thus one can start by considering polynomials of degrees $1,2,3$.

Exercise 1.1.2. Define a sequence $(t_0, t_1, t_2, \ldots)$ of positive rational numbers recursively by setting

$$t_0 = 1, \qquad t_1 = 1, \qquad t_2 = 1, \qquad \text{and}$$
$$t_n = \frac{1 + t_{n-1}t_{n-2}}{t_{n-3}} \qquad \text{for each } n \geq 3.$$

(For example, $t_3 = \dfrac{1 + t_2 t_1}{t_0} = \dfrac{1 + 1 \cdot 1}{1} = 2$ and $t_4 = \dfrac{1 + t_3 t_2}{t_1} = \dfrac{1 + 2 \cdot 1}{1} = 3$.)

(a) Prove that $t_{n+2} = 4t_n - t_{n-2}$ for each $n \geq 2$.

(b) Prove that t_n is a positive integer for each integer $n \geq 0$.

The next exercise uses the *floor function*, which is defined as follows:

Definition 1.1.1. For every real x, we let $\lfloor x \rfloor$ denote the largest integer that is $\leq x$. This integer $\lfloor x \rfloor$ is called the *floor* (or the *integer part*) of x.

For example, $\lfloor 3 \rfloor = \lfloor \pi \rfloor = 3$ and $\lfloor -\pi \rfloor = -4$. It is easy to see[8] that for every $x \in \mathbb{R}$, we have

$$\lfloor x \rfloor \leq x < \lfloor x \rfloor + 1. \tag{1}$$

(Note that the chain of inequalities (1) determines $\lfloor x \rfloor$ uniquely, and thus can be used as an alternative definition of $\lfloor x \rfloor$. This is how $\lfloor x \rfloor$ is defined in [Grinbe16, Definition 1.1.1], for example.) Older sources use the notation $[x]$ for $\lfloor x \rfloor$; this notation, however, can mean many other things.

Exercise 1.1.3. Let $x \in \mathbb{R}$ and let n be a positive integer. Prove that

$$\sum_{k=0}^{n-1} \left\lfloor x + \frac{k}{n} \right\rfloor = \lfloor nx \rfloor .$$

Exercise 1.1.4. Let a, b, c, n be positive integers such that $a \mid b^n$ and $b \mid c^n$ and $c \mid a^n$. Prove that $abc \mid (a + b + c)^{n^2+n+1}$.

Exercise 1.1.5. A mountain ridge has the form of a (finite) line segment, bordered on each end by a cliff. Several (finitely many) lemmings are walking along the ridge, with equal speeds (but not necessarily in the same direction). Whenever two lemmings meet, they "bounce off" one another, keeping their respective speeds but reversing their directions. Whenever a lemming arrives at an endpoint of the ridge, it falls off the cliff. Prove that eventually, all lemmings will fall off the cliff.

[8] and well-known enough to be used without proof on any contest

[**Example:** Here is a possible lemming configuration:

$$\overrightarrow{1} \quad \overleftarrow{2} \quad \overleftarrow{3} \qquad \overrightarrow{4} \qquad \overleftarrow{5}$$

(with $1, 2, 3, 4, 5$ being the lemmings, and the arrows signifying their walking directions). The first two lemmings to meet here will be 1 and 2, after which they both change their directions:

$$\overleftarrow{1}\,\overrightarrow{2} \quad \overleftarrow{3} \qquad \overrightarrow{4} \quad \overleftarrow{5}$$

Now lemming 1 is on its way to the cliff, which it will reach without interference from other lemmings.]

Exercise 1.1.6. You have 20 white socks and 20 black socks hanging on a clothesline, in some order.

(a) Prove that you can pick 10 consecutive socks from the clothesline such that 5 of them are black and the other 5 white. (You can call such a pick *color-balanced*.)

(b) Prove the same if the number 20 is replaced by 13 (so you have 13 white and 13 black socks).

[**Example:** If the number 20 is replaced by 7, then the claim does not hold. For example, the configuration $B^3W^7B^4$ (standing for "3 black socks, followed by 7 white socks, followed by 4 white socks") has no color-balanced pick of 10 consecutive socks.]

Exercise 1.1.7. Factor the polynomial

$$bc\,(b-c)\,(b+c) + ca\,(c-a)\,(c+a) + ab\,(a-b)\,(a+b)$$

into a product of four linear (i.e., degree-1) polynomials.

Exercise 1.1.8. **(a)** Let a, b, c be three nonnegative reals. Prove that

$$|ca-ab| + |ab-bc| + |bc-ca| \leq \left|b^2-c^2\right| + \left|c^2-a^2\right| + \left|a^2-b^2\right|.$$

(b) Is this still true if the word "nonnegative" is omitted?

Exercise 1.1.9. Let $n \geq 1$. Let $a_1, a_2, \ldots, a_n$ be any n integers. Prove that there exist some $p, q \in \{1, 2, \ldots, n\}$ with $p \leq q$ and $n \mid a_p + a_{p+1} + \cdots + a_q$.

Exercise 1.1.10. Briefly review the problems above: Which ones did you like? Which ones did you not like? Why? How long did they take you? Which parts did you get stuck on? Did you learn anything from solving (or trying to solve) them? If you knew the solution already (nothing wrong with that!), where did you learn it? (No need to rate every exercise; just say some words about some 4 of them.)

2. Induction I

Mathematical induction is essentially the main strategy for proving theorems in mathematics: Any statement that involves a natural number (even implicitly, such as by talking of a finite set) potentially lends itself to be proved by induction. This does not mean induction is always the best way to go, or even a viable one, but it is a method you should certainly not be **surprised** to see working. Only recently, two long books [AndCri17] and [Gunder10] (with 400 and 850 pages, respectively) entirely devoted to applications of induction have appeared[9]. We will not go into this depth or detail; but we will show examples for the most frequent ways in which induction can be applied.

2.1. Standard induction

We assume you are familiar with standard mathematical induction. Just as a reminder, it is a way to prove that a certain statement $\mathcal{A}(n)$ holds for every integer $n \geq g$, where g is a fixed integer. Stated as a theorem itself, here is what it says:

Theorem 2.1.1. Let $g \in \mathbb{Z}$. For each integer $n \geq g$, let $\mathcal{A}(n)$ be a logical statement. Assume the following:

- *Assumption 1:* The statement $\mathcal{A}(g)$ holds.
- *Assumption 2:* If m is an integer such that $m \geq g$ and such that $\mathcal{A}(m)$ holds, then $\mathcal{A}(m+1)$ also holds.

Then, $\mathcal{A}(n)$ holds for all integers $n \geq g$.

Theorem 2.1.1 is one of the facts known as the *principle of mathematical induction*. Some consider it to be an axiom; others use essentially equivalent axioms that it can easily be derived from.[10] We will not dwell on its logical status, but rather explore its use.

Theorem 2.1.1 provides a way to prove any result that can be stated in the form "some statement $\mathcal{A}(n)$ holds for each integer $n \geq g$" (commonly, g is taken to be 0 or 1, but any integer is fine). All one needs to do is to prove Assumption 1 and Assumption 2. The proof of Assumption 1 is commonly called the *induction base* (or *base case*), while the proof of Assumption 2 is commonly called the *induction step*. In the induction step, the assumption that $\mathcal{A}(m)$ holds is called the *induction hypothesis* (or *induction assumption*). The whole proof is called an "*induction on n*" (or "*induction over n*"). Here is a simple example:

[9] This is not to say that the method has not been appreciated before; it is just that the idea of gathering mathematical results by their method of proof (as opposed to their objects of concern) is fairly new.

[10] You may have seen it stated for $g = 0$ only (or for $g = 1$ only). But the general case of Theorem 2.1.1 easily follows from these particular cases (see, e.g., [Grinbe15, proof of Theorem 2.53]).

Exercise 2.1.1. Let b be a real number. Prove that

$$(b-1)\left(b^0+b^1+\cdots+b^{n-1}\right)=b^n-1 \tag{2}$$

for each integer $n \geq 0$. [11]

Solution to Exercise 2.1.1. We will prove (2) by induction on n (that is, we shall apply Theorem 2.1.1 to $g=0$ and to $\mathcal{A}(n)$ being the statement (2)).

Induction base: We have $(b-1)\underbrace{\left(b^0+b^1+\cdots+b^{0-1}\right)}_{=(\text{empty sum})=0}=0=b^0-1$ (since $b^0=1$). In other words, (2) holds for $n=0$. This completes the induction base.

Induction step: Let $m \geq 0$ be an integer. Assume (as the induction hypothesis) that (2) holds for $n=m$. We must prove that (2) holds for $n=m+1$. In other words, we must prove that $(b-1)\left(b^0+b^1+\cdots+b^{(m+1)-1}\right)=b^{m+1}-1$.

Our induction hypothesis says that (2) holds for $n=m$. In other words,

$$(b-1)\left(b^0+b^1+\cdots+b^{m-1}\right)=b^m-1. \tag{3}$$

Now, $(m+1)-1=m$, so that

$$\begin{aligned}
(b-1)\left(b^0+b^1+\cdots+b^{(m+1)-1}\right) &= (b-1)\underbrace{\left(b^0+b^1+\cdots+b^m\right)}_{=\left(b^0+b^1+\cdots+b^{m-1}\right)+b^m} \\
&= (b-1)\left(\left(b^0+b^1+\cdots+b^{m-1}\right)+b^m\right) \\
&= \underbrace{(b-1)\left(b^0+b^1+\cdots+b^{m-1}\right)}_{\substack{=b^m-1 \\ (\text{by } (3))}}+\underbrace{(b-1)\,b^m}_{=bb^m-b^m} \\
&= b^m-1+bb^m-b^m=\underbrace{bb^m}_{=b^{m+1}}-1=b^{m+1}-1.
\end{aligned}$$

But this is precisely what we need to prove. Thus, the induction step is complete. Hence, (2) is proved. Thus, Exercise 2.1.1 is solved. □

We note that if b is a real number distinct from 1, then the claim of (2) entails

$$b^0+b^1+\cdots+b^{n-1}=\frac{b^n-1}{b-1}=\frac{1-b^n}{1-b} \tag{4}$$

for each integer $n \geq 0$. This is the classical formula for the sum of a (finite) geometric progression. There are other proofs of this formula (in particular, we shall

[11] Note that if $n=0$, then the sum $b^0+b^1+\cdots+b^{n-1}$ is an empty sum and thus equals 0 by definition.

prove something slightly more general in a later chapter), but the induction proof that we gave above has the advantage of being entirely straightforward.

A particularly simple particular case of Exercise 2.1.1 is the following: Each integer $n \geq 0$ satisfies

$$2^0 + 2^1 + \cdots + 2^{n-1} = 2^n - 1. \tag{5}$$

Indeed, this follows by applying Exercise 2.1.1 to $b = 2$ (since $2 - 1 = 1$).

Here is another basic example for a proof by induction (which appears, e.g., in [Grinbe15, Proposition 2.164], in [Gunder10, Exercise 74] and in [AndCri17, Problem 2.2]):

Exercise 2.1.2. Prove that every integer $n \geq 0$ satisfies

$$\frac{1}{1} - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots + \frac{1}{2n-1} - \frac{1}{2n} = \frac{1}{n+1} + \frac{1}{n+2} + \cdots + \frac{1}{2n}. \tag{6}$$

Solution to Exercise 2.1.2. Let us denote the left hand side of (6) by $L(n)$, and let us denote the right hand side by $R(n)$. That is, we set

$$L(n) = \frac{1}{1} - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots + \frac{1}{2n-1} - \frac{1}{2n} \qquad \text{and}$$
$$R(n) = \frac{1}{n+1} + \frac{1}{n+2} + \cdots + \frac{1}{2n}$$

for each $n \geq 0$. Thus, we need to prove that $L(n) = R(n)$ for each $n \geq 0$. We shall do this by induction on n (that is, we shall apply Theorem 2.1.1 to $g = 0$ and $\mathcal{A}(n) = (\text{“}L(n) = R(n)\text{”})$).

Induction base: The equality $L(0) = R(0)$ holds, since both $L(0)$ and $R(0)$ are empty sums (and thus 0). In other words, $L(n) = R(n)$ holds for $n = 0$. This completes the induction base.

Induction step: Let $m \geq 0$ be an integer. Assume (as the induction hypothesis) that $L(m) = R(m)$. We must prove that $L(m+1) = R(m+1)$.

The definition of $L(n)$ shows that $L(m+1)$ is the same sum as $L(m)$ except with the two extra addends $\dfrac{1}{2(m+1)-1}$ and $-\dfrac{1}{2(m+1)}$ added to it. Thus,

$$L(m+1) = L(m) + \frac{1}{2(m+1)-1} - \frac{1}{2(m+1)}. \tag{7}$$

The definition of $R(n)$ shows that $R(m+1)$ is the same sum as $R(m)$ except without the addend $\dfrac{1}{m+1}$ but with the two new addends $\dfrac{1}{2m+1}$ and $\dfrac{1}{2m+2}$ added to it. Thus,

$$R(m+1) = R(m) - \frac{1}{m+1} + \frac{1}{2m+1} + \frac{1}{2m+2}. \tag{8}$$

Now it is clear what to do: We need to show that the left hand sides of the two equalities (7) and (8) are equal, but we know that the $L(m)$ and $R(m)$ terms on

their right hand sides are equal (since $L(m) = R(m)$ by assumption). So we need to show that the remaining terms on the right hand sides are also equal – i.e., that we have

$$\frac{1}{2(m+1)-1} - \frac{1}{2(m+1)} = -\frac{1}{m+1} + \frac{1}{2m+1} + \frac{1}{2m+2}.$$

But this can be checked by a straightforward computation (which is simplified by the facts that $2(m+1)-1 = 2m+1$ and $2m+2 = 2(m+1)$). Thus, the right hand sides of (7) and (8) are equal (since $L(m) = R(m)$), and therefore so are the left hand sides. In other words, $L(m+1) = R(m+1)$. This completes the induction step. Exercise 2.1.2 is thus solved. □

Equalities between finite sums (such as in Exercise 2.1.2) provide lots of exercise in using induction; for example, you can use induction to prove the equalities

$$1+2+\cdots+n = \frac{n(n+1)}{2}; \tag{9}$$

$$1-2+3-4\pm\cdots+(-1)^n(n+1) = \begin{cases} n/2+1, & \text{if } n \text{ is even;} \\ -(n+1)/2, & \text{if } n \text{ is odd;} \end{cases}$$

$$1^2+2^2+\cdots+n^2 = \frac{n(n+1)(2n+1)}{6};$$

$$1^3+2^3+\cdots+n^3 = \left(\frac{n(n+1)}{2}\right)^2;$$

$$1^4+2^4+\cdots+n^4 = \frac{n(2n+1)(n+1)\left(3n+3n^2-1\right)}{30}$$

for each integer $n \geq 0$ (see [Grinbe15, §2.4] and [Grinbe15, Exercise 2.9], respectively, for detailed proofs of the first two equalities).

We shall now come to a more combinatorial application. Recall that a *bit* is defined to be an element of the 2-element set $\{0,1\}$.

Exercise 2.1.3. Fix a positive integer n. An *n-bitstring* shall mean an n-tuple $(a_1, a_2, \ldots, a_n) \in \{0,1\}^n$ of bits. Prove that there exists a list $(b_1, b_2, \ldots, b_{2^n})$ containing all n-bitstrings (each exactly once) such that for every $i \in \{1, 2, \ldots, 2^n\}$, the two n-bitstrings b_i and b_{i-1} differ in exactly one entry. Here, we understand b_0 to mean b_{2^n}.

[**Example:** For $n = 3$, one such list is $(b_1, b_2, \ldots, b_8)$, where

$$\begin{aligned} b_1 &= (0,0,0), & b_2 &= (0,0,1), & b_3 &= (0,1,1), & b_4 &= (0,1,0), \\ b_5 &= (1,1,0), & b_6 &= (1,1,1), & b_7 &= (1,0,1), & b_8 &= (1,0,0). \end{aligned}$$

It is far from the only such list.]

Before we solve this exercise, a few comments on its significance are in order. A list $(b_1, b_2, \ldots, b_{2^n})$ satisfying the requirement of Exercise 2.1.3 is known as a *(circular) Gray code*. Quoting the (extensive) Wikipedia article on these:

> "Gray codes are used in linear and rotary position encoders (absolute encoders and quadrature encoders) in preference to weighted binary encoding. This avoids the possibility that, when multiple bits change in the binary representation of a position, a misread will result from some of the bits changing before others."

That is, the property that b_i and b_{i-1} differ in exactly one entry ensures that b_{i-1} can be turned into b_i by only changing a single bit, which removes the need for changing several bits simultaneously to prevent unintended intermediate states.

Another way to think of Exercise 2.1.3 is as follows: Imagine a combination lock with n dials, where each dial has exactly 2 disks labeled 0 and 1. Then, a Gray code provides a way to cycle through all 2^n possible combinations in a way that getting from each combination to the other (and from the last back to the first) only requires rotating a single dial. Variants of Gray codes exist for dials with more than 2 disks, and for various other similar situations; a comprehensive discussion of various such contraptions can be found in [TAoCP4A, §7.2.1.1].[12]

Solution to Exercise 2.1.3 (sketched). We define an *n-Gray code* to be a list $(b_1, b_2, \ldots, b_{2^n})$ containing all n-bitstrings (each exactly once) such that for every $i \in \{1, 2, \ldots, 2^n\}$, the two n-bitstrings b_i and b_{i-1} differ in exactly one entry, where b_0 means b_{2^n}. Thus, the exercise wants us to prove that there exists an n-Gray code.

We shall prove this by induction on n, so we forget that we fixed n. Note that we have to start our induction at $n = 1$ (that is, we must apply Theorem 2.1.1 to $g = 1$), since the exercise is stated for all positive integers n (rather than for all integers $n \geq 0$, as the two previous exercises were).

Induction base: There exists a 1-Gray code – namely, the list $((0), (1))$. Thus, Exercise 2.1.3 is proved[13] for $n = 1$. This completes the induction base.

Induction step: Let $m \geq 1$ be an integer. Assume (as the induction hypothesis) that there exists an m-Gray code. We must show that there exists an $(m+1)$-Gray code.

We introduce a notation: For any m-bitstring $a = (a_1, a_2, \ldots, a_m)$, we let $0a$ denote the $(m+1)$-bitstring $(0, a_1, a_2, \ldots, a_m)$, and we let $1a$ denote the $(m+1)$-bitstring $(1, a_1, a_2, \ldots, a_m)$. Thus, $0a$ and $1a$ are the two possible $(m+1)$-bitstrings that can be obtained from a by inserting a new bit at the front. For example, if $a = (1, 1, 0)$ (so $m = 3$), then $0a = (0, 1, 1, 0)$ and $1a = (1, 1, 1, 0)$.

We have assumed that there exists an m-Gray code; let $(b_1, b_2, \ldots, b_{2^m-1}, b_{2^m})$ be this m-Gray code. Then,

$$(0b_1, 0b_2, \ldots, 0b_{2^m-1}, 0b_{2^m}, 1b_{2^m}, 1b_{2^m-1}, \ldots, 1b_2, 1b_1)$$

[12] We note that Exercise 2.1.3 is also a popular contest problem. For instance, Problem A3 on the Putnam Mathematical Competition 1968 is a thinly-veiled restatement of Exercise 2.1.3 (more precisely, of a weaker version of this exercise, which requires b_i and b_{i-1} differ in exactly one entry only for $i \in \{1, 2, \ldots, 2^n - 1\}$ rather than for all $i \in \{1, 2, \ldots, 2^n\}$).

[13] Here, we are being slightly sloppy with our language: When we say "Exercise 2.1.3", we mean "the claim of Exercise 2.1.3". Thus, "Exercise 2.1.3 is proved" means "the claim of Exercise 2.1.3 is proved". We shall use this sloppy language occasionally, as it saves us space.

(note that the subscripts first go up from 1 to 2^m, then go back down from 2^m to 1) is an $(m+1)$-Gray code (check this!). Thus, there exists an $(m+1)$-Gray code. This completes the induction step. Thus, Exercise 2.1.3 is solved by induction. $\square$

The verb "*induct*" is often used to mean "perform an induction". Thus, for instance, in the above proof, we could have said "We induct on n" instead of "We shall prove this by induction on n". A proof by induction is also known as an *inductive proof*.[14]

As mentioned above, when one inducts on n to prove a statement, the n needs not explicitly appear in the statement; it suffices that the statement can be restated in a way that contains an n. As an example, let us show the following (seemingly evident) fact:

Proposition 2.1.2. Let S be a nonempty finite set of integers. Then, S has a maximum.

Recall that a "*maximum*" (also known as a "*largest element*") of a set S of integers[15] is defined to be an element $s \in S$ such that we have $s \geq t$ for all $t \in S$. It is easy to see that any set of integers has at most one maximum (because if s_1 and s_2 are two maxima of the same set, then $s_1 \geq s_2$ and $s_2 \geq s_1$, whence $s_1 = s_2$); but some sets of integers have none. (For example, the sets $\varnothing$ and $\mathbb{Z}$ have none.) Proposition 2.1.2 says that for a nonempty finite set, there always is a maximum (and thus, according to what we just said, a unique maximum).

Proof of Proposition 2.1.2 (sketched). (See [Grinbe15, first proof of Theorem 2.35] for details[16].) Forget that we fixed S. We thus must prove the following:

> *Claim 1:* Every nonempty finite set S of integers has a maximum.

We can rewrite this claim as follows:

> *Claim 2:* Let n be a positive integer. Then, every n-element set of integers has a maximum.

Claim 2 is equivalent to Claim 1, because a set S is a nonempty and finite if and only if its size $|S|$ is a positive integer. But Claim 2 has the advantage of depending on a positive integer n, which gives us an opportunity to use induction. And that's what we will do now:

[*Proof of Claim 2:* We induct on n:

[14] This is not to be confused with the epistemological concept of "inductive reasoning", which is not a proof method.

[15] You can replace "integers" by "rational numbers" or "real numbers" here, and nothing will change.

[16] This said, [Grinbe15, first proof of Theorem 2.35] is organized somewhat differently (in particular, most of the argument is relegated into a lemma), and uses the nonnegative integer $|S| - 1$ instead of the positive integer $|S|$ to do the induction.

Induction base: Every 1-element set of integers has a maximum (namely, its unique element). In other words, Claim 2 holds for $n = 1$. This completes the induction base.

Induction step: Let m be a positive integer. Assume (as the induction hypothesis) that Claim 2 holds for $n = m$. We must prove that Claim 2 holds for $n = m + 1$. In other words, we must prove that every $(m + 1)$-element set of integers has a maximum.

So let S be an $(m + 1)$-element set of integers. Then, $|S| = m + 1 \geq 1 > 0$, so that S is nonempty. Hence, there exists some $s \in S$. Consider such an s. The set $S \setminus \{s\}$ then must be an m-element set (since S is an $(m + 1)$-element set), and thus has a maximum (because our induction hypothesis says that every m-element set of integers has a maximum). Let t be this maximum.

Now, we can see that S has a maximum: Indeed, if $t \geq s$, then t is a maximum of S, whereas otherwise, s is a maximum of S. (Check this!)

Forget that we fixed S. We thus have proved that every $(m + 1)$-element set S of integers has a maximum. In other words, Claim 2 holds for $n = m + 1$. This completes the induction step. Claim 2 is thus proved.]

Since Claim 2 is equivalent to Claim 1, we thus conclude that Claim 1 also holds, i.e., Proposition 2.1.2 is proved. □

In advanced mathematics, most inductions are of the kind we just showed (although usually more complicated) – i.e., the n that is being inducted upon does not explicitly appear in the claim that is being proved, but rather is a "derived quantity" (like the size $|S|$ of the set S in the above proof). Commonly, this is done quickly and tacitly – that is, instead of restating the claim in terms of n as we did above, one simply says that one is doing an induction on the derived quantity (i.e., in the above example, an induction on $|S|$). See [Grinbe15, §2.5.3] for the exact convention that is being used here; let me just show how the above proof could be rewritten using this convention:

Proof of Proposition 2.1.2 (short version). (See [Grinbe15, second proof of Theorem 2.35] for details.) Forget that we fixed S. Notice that $|S|$ is a positive integer whenever S is a nonempty finite set. Hence, we induct on $|S|$:

Induction base: If S is a nonempty finite set of integers satisfying $|S| = 1$, then S has a maximum (namely, its unique element). In other words, Proposition 2.1.2 holds for $|S| = 1$. This completes the induction base.

Induction step: Let m be a positive integer. Assume (as the induction hypothesis) that Proposition 2.1.2 holds for $|S| = m$. We must prove that Proposition 2.1.2 holds for $|S| = m + 1$. In other words, we must prove that every nonempty finite set S of integers satisfying $|S| = m + 1$ has a maximum.

So let S be a nonempty finite set of integers satisfying $|S| = m + 1$. Then, $|S| = m + 1 \geq 1 > 0$, so that S is nonempty. Hence, there exists some $s \in S$. Consider such an s. The set $S \setminus \{s\}$ then satisfies $|S \setminus \{s\}| = |S| - 1 = m$ (since $|S| = m + 1$), and thus is nonempty (since m is a positive integer); hence, this set $S \setminus \{s\}$ has a

maximum (because our induction hypothesis says that every nonempty finite set of integers that has size m has a maximum). Let t be this maximum.

Now, we can see that S has a maximum: Indeed, if $t \geq s$, then t is a maximum of S, whereas otherwise, s is a maximum of S. (Check this!)

Forget that we fixed S. We thus have proved that every nonempty finite set S of integers satisfying $|S| = m + 1$ has a maximum. In other words, Proposition 2.1.2 holds for $|S| = m + 1$. This completes the induction step. Proposition 2.1.2 is thus proved. □

We shall see more complicated examples of induction on a derived quantity soon.

2.2. Fibonacci numbers I

The Fibonacci sequence, with its recursive definition and multiple properties, is a veritable induction playground. But it is also an object of serious research; a whole book about it has been written [Vorobi02], and the *Fibonacci Association* publishes the *Fibonacci Quarterly* since 1963 and organizes the two-yearly *Fibonacci Conference*. Arguably these are only partly concerned with the Fibonacci sequence, but its role in them is substantial if not leading.

Let us recall the definition of this venerable sequence and show a few of its properties; we'll see more of it at times later on.

Definition 2.2.1. The *Fibonacci sequence* is the sequence $(f_0, f_1, f_2, \ldots)$ of integers which is defined recursively by

$$f_0 = 0, \qquad f_1 = 1, \qquad \text{and} \qquad f_n = f_{n-1} + f_{n-2} \text{ for all } n \geq 2.$$

The entries of this sequence are called the *Fibonacci numbers*.

The first Fibonacci numbers are

$$\begin{gathered} f_0 = 0, \quad f_1 = 1, \quad f_2 = 1, \quad f_3 = 2, \quad f_4 = 3, \quad f_5 = 5, \\ f_6 = 8, \quad f_7 = 13, \quad f_8 = 21, \quad f_9 = 34, \quad f_{10} = 55, \\ f_{11} = 89, \quad f_{12} = 144, \quad f_{13} = 233. \end{gathered}$$

Some authors[17] prefer to start the sequence at f_1 rather than f_0; of course, the recursive definition then needs to be modified to require $f_2 = 1$ instead of $f_0 = 0$.

The first property of Fibonacci numbers that we prove is the following:

Exercise 2.2.1. Let $(f_0, f_1, f_2, \ldots)$ be the Fibonacci sequence. Prove that each integer $n \geq 0$ satisfies

$$f_1 + f_2 + \cdots + f_n = f_{n+2} - 1. \tag{10}$$

[17] such as Vorobiev in his book [Vorobi02]

Solution to Exercise 2.2.1. We prove (10) by induction on n:

Induction base: We have $f_1 + f_2 + \cdots + f_0 = (\text{empty sum}) = 0 = f_2 - 1$ (since $f_2 = 1$). In other words, (10) holds for $n = 0$. This completes the induction base.[18]

Induction step: Let $m \geq 0$. Assume (as the induction hypothesis) that (10) holds for $n = m$. We must show that (10) holds for $n = m + 1$.

Our assumption says that (10) holds for $n = m$; in other words, $f_1 + f_2 + \cdots + f_m = f_{m+2} - 1$.

Now, our goal is to prove that (10) holds for $n = m + 1$; in other words, our goal is to prove that $f_1 + f_2 + \cdots + f_{m+1} = f_{m+3} - 1$ (since this is what (10) says for $n = m + 1$). But this follows by comparing

$$f_1 + f_2 + \cdots + f_{m+1} = \underbrace{(f_1 + f_2 + \cdots + f_m)}_{=f_{m+2}-1} + f_{m+1} = f_{m+2} - 1 + f_{m+1}$$

with

$$\underbrace{f_{m+3}}_{\substack{=f_{m+2}+f_{m+1} \\ \text{(by the recursive definition} \\ \text{of the Fibonacci sequence)}}} - 1 = f_{m+2} + f_{m+1} - 1 = f_{m+2} - 1 + f_{m+1}.$$

So we have shown that (10) holds for $n = m + 1$. This completes the induction step, and thus proves (10). Hence, Exercise 2.2.1 is solved. □

Let us briefly discuss two notational aspects of induction. So far, we have always been using the letter m for the integer that is introduced in the induction step, because it was called m in the statement of Theorem 2.1.1. Clearly, we can use any other letter (for example, k) instead, as long as that letter does not already have a different meaning. It is also perfectly fine to use $k - 1$ for it – i.e., instead of assuming that the claim holds for $n = m$ and proving that it holds for $n = m + 1$, we can just as well assume that the claim holds for $n = k - 1$ and prove that it holds for $n = k$. (Of course, if we do this, then we need to assume $k > g$ instead of $k \geq g$.) This just boils down to substituting $k - 1$ for m in Theorem 2.1.1. Here is how Theorem 2.1.1 looks like after this substitution:

Theorem 2.2.2. Let $g \in \mathbb{Z}$. For each integer $n \geq g$, let $\mathcal{A}(n)$ be a logical statement. Assume the following:

- *Assumption 1:* The statement $\mathcal{A}(g)$ holds.
- *Assumption 2:* If k is an integer such that $k > g$ and such that $\mathcal{A}(k-1)$ holds, then $\mathcal{A}(k)$ also holds.

Then, $\mathcal{A}(n)$ holds for all integers $n \geq g$.

[18] We will no longer write this sentence at the end of an induction base, since the "*Induction step:*" that follows it should make it clear enough that the induction base ends here.

As an example for an induction proof written in this way, let us prove another property of the Fibonacci sequence (the so-called *Cassini identity*):

Exercise 2.2.2. Let $(f_0, f_1, f_2, \ldots)$ be the Fibonacci sequence. Prove that

$$f_{n+1}f_{n-1} - f_n^2 = (-1)^n$$

for any positive integer n.

Solution to Exercise 2.2.2. We induct on n (using Theorem 2.2.2):

Induction base: We have $f_0 = 0$, $f_1 = 1$ and $f_2 = 1$. Thus, $f_2 f_0 - f_1^2 = 1 \cdot 0 - 1^2 = -1 = (-1)^1$. In other words, Exercise 2.2.2 holds for $n = 1$.

Induction step: Let k be an integer such that $k > 1$. Assume (as the induction hypothesis) that Exercise 2.2.2 holds for $n = k - 1$. We must show that Exercise 2.2.2 holds for $n = k$. In other words, we must show that $f_{k+1}f_{k-1} - f_k^2 = (-1)^k$.

Our induction hypothesis says that Exercise 2.2.2 holds for $n = k - 1$. In other words, it says that $f_k f_{k-2} - f_{k-1}^2 = (-1)^{k-1}$. But the recursive definition of the Fibonacci sequence yields $f_{k+1} = f_k + f_{k-1}$ (since $k > 1$). Hence,

$$\begin{aligned}
\underbrace{f_{k+1}}_{=f_k+f_{k-1}} f_{k-1} &= (f_k + f_{k-1}) f_{k-1} = f_k f_{k-1} + \underbrace{f_{k-1}^2}_{\substack{=f_k f_{k-2} - (-1)^{k-1} \\ (\text{since } f_k f_{k-2} - f_{k-1}^2 = (-1)^{k-1})}} \\
&= f_k f_{k-1} + f_k f_{k-2} - (-1)^{k-1} = f_k \underbrace{(f_{k-1} + f_{k-2})}_{\substack{=f_k \\ (\text{since the recursive definition} \\ \text{of the Fibonacci sequence} \\ \text{yields } f_k = f_{k-1} + f_{k-2})}} - \underbrace{(-1)^{k-1}}_{=-(-1)^k} \\
&= f_k f_k - \left(-(-1)^k\right) = f_k^2 + (-1)^k,
\end{aligned}$$

so that $f_{k+1}f_{k-1} - f_k^2 = (-1)^k$. This is exactly what we wanted to prove. This completes the induction step, and with it the solution to Exercise 2.2.2. □

Some writers shorten their induction proofs even further by reusing the letter n itself (instead of m or k as we did above) in the induction step. That is, instead of fixing a $k > g$ and assuming that the claim holds for $n = k - 1$ and proving that it holds for $n = k$, they fix an $n > g$ and assume that the claim holds "for $n - 1$ instead of n" (i.e., the claim holds if n is replaced by $n - 1$ in it) and prove that it holds for n as well. This saves a letter and a bit of writing, at the cost of being potentially more confusing and slippery (as it requires you to check that you have properly replaced n by $n - 1$ in the induction hypothesis); thus I do not recommend it. But you should be aware that it is a commonly used "figure of speech".[19]

The next exercise (a particular case of [Grinbe15, Theorem 2.26 **(a)**]) illustrates a point about choosing the right claim when doing induction.

[19]For example, here is how the above solution to Exercise 2.2.2 could be rewritten using this con-

Exercise 2.2.3. Let $(f_0, f_1, f_2, \ldots)$ be the Fibonacci sequence. Prove that

$$f_{n+m+1} = f_n f_m + f_{n+1} f_{m+1} \tag{11}$$

for any nonnegative integers n and m.

We have two nonnegative integer variables here (n and m), so we can try to do induction on either. As to the other, it appears reasonable to just fix it beforehand. So we try the following:

First attempt at solving Exercise 2.2.3. Let us fix m and induct on n (using Theorem 2.1.1).

Induction base: The equality (11) holds for $n = 0$. Indeed, for $n = 0$, this equality boils down to $f_{0+m+1} = f_0 f_m + f_1 f_{m+1}$, which follows easily from $f_0 = 0$ and $f_1 = 1$.

Induction step: Let k be a nonnegative integer.[20] Assume (as the induction hypothesis) that (11) holds for $n = k$. We must show that (11) holds for $n = k + 1$. In other words, we must show that $f_{(k+1)+m+1} = f_{k+1} f_m + f_{(k+1)+1} f_{m+1}$. In other words, we must show that $f_{k+m+2} = f_{k+1} f_m + f_{k+2} f_{m+1}$.

Our induction hypothesis says that (11) holds for $n = k$, that is, we have $f_{k+m+1} = f_k f_m + f_{k+1} f_{m+1}$.

The recursive definition of the Fibonacci sequence yields

$$f_{k+m+2} = f_{k+m+1} + f_{k+m}.$$

vention:

We induct on n (using Theorem 2.2.2):

Induction base: We have $f_0 = 0$, $f_1 = 1$ and $f_2 = 1$. Thus, $f_2 f_0 - f_1^2 = 1 \cdot 0 - 1^2 = -1 = (-1)^1$. In other words, Exercise 2.2.2 holds for $n = 1$.

Induction step: Let n be an integer such that $n > 1$. Assume (as the induction hypothesis) that Exercise 2.2.2 holds for $n - 1$ instead of n. We must show that Exercise 2.2.2 holds for n. In other words, we must show that $f_{n+1} f_{n-1} - f_n^2 = (-1)^n$.

Our induction hypothesis says that Exercise 2.2.2 holds for $n - 1$ instead of n. In other words, it says that $f_n f_{n-2} - f_{n-1}^2 = (-1)^{n-1}$. But the recursive definition of the Fibonacci sequence yields $f_{n+1} = f_n + f_{n-1}$ (since $n > 1$). Hence,

$$\begin{aligned} \underbrace{f_{n+1}}_{=f_n+f_{n-1}} f_{n-1} &= (f_n + f_{n-1}) f_{n-1} = f_n f_{n-1} + \underbrace{f_{n-1}^2}_{\substack{=f_n f_{n-2}-(-1)^{n-1} \\ \text{(since } f_n f_{n-2}-f_{n-1}^2=(-1)^{n-1})}} = f_n f_{n-1} + f_n f_{n-2} - (-1)^{n-1} \\ &= f_n \underbrace{(f_{n-1} + f_{n-2})}_{\substack{=f_n \\ \text{(since the recursive definition} \\ \text{of the Fibonacci sequence} \\ \text{yields } f_n=f_{n-1}+f_{n-2})}} - \underbrace{(-1)^{n-1}}_{=-(-1)^n} = f_n f_n - \left(-(-1)^n\right) = f_n^2 + (-1)^n, \end{aligned}$$

so that $f_{n+1} f_{n-1} - f_n^2 = (-1)^n$. This is exactly what we wanted to prove. This completes the induction step, and with it the solution to Exercise 2.2.2.

[20] This k will play the role of the m in Theorem 2.1.1, since we cannot use the letter m for it here (because m is already a fixed number).

Now, our induction hypothesis lets us decompose the f_{k+m+1} on the right hand side as $f_k f_m + f_{k+1} f_{m+1}$; but how do we decompose the f_{k+m}? The induction hypothesis does not help us here, and we are stuck. $\square$

This suggests that we need a better induction hypothesis. The easiest way to get one is to **not** fix m before starting the induction. The result is that we have to carry a "for all nonnegative integers m" through our induction; but this extra weight turns out to be useful:

Second attempt at solving Exercise 2.2.3. Let us induct on n.

Induction base: The equality (11) holds for $n = 0$ and every nonnegative integer m. Indeed, for $n = 0$, this equality boils down to $f_{0+m+1} = f_0 f_m + f_1 f_{m+1}$, which follows easily from $f_0 = 0$ and $f_1 = 1$.

Induction step: Let k be a nonnegative integer. Assume (as the induction hypothesis) that (11) holds for $n = k$ and every nonnegative integer m. We must show that (11) holds for $n = k + 1$ and every nonnegative integer m. In other words, we must show that $f_{(k+1)+m+1} = f_{k+1} f_m + f_{(k+1)+1} f_{m+1}$ for every nonnegative integer m. In other words, we must show that $f_{k+m+2} = f_{k+1} f_m + f_{k+2} f_{m+1}$ for every nonnegative integer m.

Our induction hypothesis says that (11) holds for $n = k$ and every nonnegative integer m. In other words, we have

$$f_{k+m+1} = f_k f_m + f_{k+1} f_{m+1} \tag{12}$$

for every nonnegative integer m.

Now, let m be a nonnegative integer. The recursive definition of the Fibonacci sequence yields

$$f_{k+m+2} = f_{k+m+1} + f_{k+m}. \tag{13}$$

Once again, we can use our induction hypothesis (12) to rewrite the f_{k+m+1} on the right hand side as $f_k f_m + f_{k+1} f_{m+1}$. But this time, we can also apply (12) to $m - 1$ instead of m (since the m in (12) was an arbitrary nonnegative integer, not related to the m we have currently fixed), and thus obtain $f_{k+(m-1)+1} = f_k f_{m-1} + f_{k+1} f_{(m-1)+1}$. To be fully honest, we can only do this when $m - 1$ is a nonnegative integer, which means that we cannot do this when $m = 0$; but the $m = 0$ case is easy and left to the reader. So we WLOG assume that $m \neq 0$; therefore, $m \geq 1$, and thus $m - 1$ is a nonnegative integer. Hence, (12) (applied to $m - 1$ instead of m) yields $f_{k+(m-1)+1} = f_k f_{m-1} + f_{k+1} f_{(m-1)+1}$. Since $(m - 1) + 1 = m$, this rewrites as

$$f_{k+m} = f_k f_{m-1} + f_{k+1} f_m. \tag{14}$$

Now, (13) becomes

$$\begin{aligned}
f_{k+m+2} &= \underbrace{f_{k+m+1}}_{\substack{=f_kf_m+f_{k+1}f_{m+1} \\ \text{(by (12))}}} + \underbrace{f_{k+m}}_{\substack{=f_kf_{m-1}+f_{k+1}f_m \\ \text{(by (14))}}} \\
&= f_kf_m + f_{k+1}f_{m+1} + f_kf_{m-1} + f_{k+1}f_m \\
&= f_{k+1}f_m + f_k \underbrace{(f_m + f_{m-1})}_{\substack{=f_{m+1} \\ \text{(since the recursive definition} \\ \text{of the Fibonacci sequence} \\ \text{yields } f_{m+1}=f_m+f_{m-1})}} + f_{k+1}f_{m+1} \\
&= f_{k+1}f_m + f_kf_{m+1} + f_{k+1}f_{m+1} \\
&= f_{k+1}f_m + \underbrace{(f_k + f_{k+1})}_{\substack{=f_{k+2} \\ \text{(since the recursive definition} \\ \text{of the Fibonacci sequence} \\ \text{yields } f_{k+2}=f_{k+1}+f_k=f_k+f_{k+1})}} f_{m+1} = f_{k+1}f_m + f_{k+2}f_{m+1}.
\end{aligned}$$

But this is exactly what we wanted to show. Thus, the induction step is complete, and Exercise 2.2.3 is solved. □

The solution we just gave is not the simplest possible; there is a different one, which avoids having to special-case the $m = 0$ case (as we did above). The trick is to rewrite f_{k+m+2} not as $f_{k+m+1} + f_{k+m}$, but rather as follows:

$$f_{k+m+2} = f_{k+(m+1)+1} = f_kf_{m+1} + f_{k+1}f_{m+2}$$

(by (12), applied to $m + 1$ instead of m). It is a nice exercise for the reader to finish this argument (see [Grinbe15, proof of Theorem 2.26 **(a)**] for the answer).

Another version of induction is tailored to proving theorems about integers n in a finite interval[21] rather than about all integers $n \geq g$. Instead of relying on Theorem 2.1.1, it relies on the following theorem ([Grinbe15, Theorem 2.74]):

Theorem 2.2.3. Let $g \in \mathbb{Z}$ and $h \in \mathbb{Z}$. For each $n \in \{g, g+1, \ldots, h\}$, let $\mathcal{A}(n)$ be a logical statement.

Assume the following:

Assumption 1: If $g \leq h$, then the statement $\mathcal{A}(g)$ holds.[22]

Assumption 2: If $m \in \{g, g+1, \ldots, h-1\}$ is such that $\mathcal{A}(m)$ holds, then $\mathcal{A}(m+1)$ also holds.

Then, $\mathcal{A}(n)$ holds for each $n \in \{g, g+1, \ldots, h\}$.

[21] Here, the word "interval" means a set of the form

$$\{g, g+1, \ldots, h\} = \{\text{all integers between } g \text{ and } h \text{ (inclusive)}\}$$

for two fixed integers g and h. Thus, unlike in analysis, our intervals here consist entirely of integers.

Theorem 2.2.3 can easily be derived from Theorem 2.1.1, by applying the latter to the statement "if $n \leq h$, then $\mathcal{A}(n)$" instead of $\mathcal{A}(n)$. (See [Grinbe15, proof of Theorem 2.74] for the details of this derivation.) It is clear how to perform induction using Theorem 2.2.3: It differs from standard induction only in that we assume $g \leq m \leq h-1$ (instead of only assuming $m \geq g$) in the induction step.

Here is an example of an induction proof using Theorem 2.2.3:

Proposition 2.2.4. Let g and h be integers such that $g \leq h$. Let $b_g, b_{g+1}, \ldots, b_h$ be any $h-g+1$ nonzero integers. Assume that $b_g \geq 0$. Assume further that

$$|b_{i+1} - b_i| \leq 1 \qquad \text{for every } i \in \{g, g+1, \ldots, h-1\}. \tag{15}$$

Then, $b_n > 0$ for each $n \in \{g, g+1, \ldots, h\}$.

Proposition 2.2.4 is often called the "*discrete intermediate value theorem*" or the "*discrete continuity principle*". Its intuitive meaning is that if a finite list of nonzero integers starts with a nonnegative integer, and every further entry of this list differs from its preceding entry by at most 1 (you can think of this as a discrete version of continuity), then all entries of this list must be positive.[23] Intuitively, this is obvious: It just says that it isn't possible to go from a nonnegative integer to a negative integer by steps of 1 without ever stepping at 0. The following proof is just a rigorous restatement of this intuitive argument:

Proof of Proposition 2.2.4 (sketched). (See [Grinbe15, proof of Proposition 2.75] for details.) We induct on n using Theorem 2.2.3 (that is, we apply Theorem 2.2.3 to the statement $\mathcal{A}(n) = (\text{"}b_n > 0\text{"})$):

Induction base: We must prove that if $g \leq h$, then the statement $b_g > 0$ holds. So let us assume that $g \leq h$. (Actually, we have already assumed so in the statement of the proposition.) Recall that $b_g, b_{g+1}, \ldots, b_h$ are nonzero; thus, in particular, $b_g \neq 0$. But we also have $b_g \geq 0$. Combining these, we find $b_g > 0$. This completes the induction base.

Induction step: Let $m \in \{g, g+1, \ldots, h-1\}$ be such that $b_m > 0$ holds. We must show that $b_{m+1} > 0$ also holds.

Applying (15) to $i = m$, we find $|b_{m+1} - b_m| \leq 1$. But it is well-known (and easy to see) that every integer x satisfies $-x \leq |x|$. Applying this to $x = b_{m+1} - b_m$, we obtain $-(b_{m+1} - b_m) \leq |b_{m+1} - b_m| \leq 1$. In other words, $1 \geq -(b_{m+1} - b_m) = b_m - b_{m+1}$. In other words, $1 + b_{m+1} \geq b_m$. Hence, $1 + b_{m+1} \geq b_m > 0$, so that $1 + b_{m+1} \geq 1$ (since $1 + b_{m+1}$ is an integer). In other words, $b_{m+1} \geq 0$. However, $b_{m+1} \neq 0$ (since $b_g, b_{g+1}, \ldots, b_h$ are nonzero). Combining this with $b_{m+1} \geq 0$, we obtain $b_{m+1} > 0$. This completes the induction step. Hence, Proposition 2.2.4 is proved. □

[22] The "If $g \leq h$" in this assumption is a bit silly: If we don't have $g \leq h$, then the claim we are proving is vacuous to begin with (because the set $\{g, g+1, \ldots, h\}$ is empty in this case). But sometimes it is good to have vacuous cases covered, too.

[23] An example of such a list is $(2, 3, 3, 2, 3, 4, 4, 3, 2, 3, 2, 3, 2, 1)$.

In the future, we will no longer say "We induct on n using Theorem 2.2.3", but instead just say "We induct on n". Indeed, it is clear that we must be using Theorem 2.2.3 if we are inducting on a variable that is bound to an interval (like the n in Proposition 2.2.4).

2.3. Strong induction

The induction principle (in the form of Theorem 2.2.2) can be briefly summarized as follows: If we want to prove that a statement $\mathcal{A}(n)$ holds for all integers $n \geq g$, then it suffices to prove it under the assumption that $\mathcal{A}(n-1)$ holds (as long as one has proved that $\mathcal{A}(g)$ holds). In our above examples, this idea was enough to make the proof straightforward or at least fairly easy. Sometimes, however, this assumption is not enough. Here is one example, known as *Binet's formula*:

Theorem 2.3.1. Let $\varphi = \dfrac{1+\sqrt{5}}{2}$ and $\psi = \dfrac{1-\sqrt{5}}{2}$ be the two solutions of the quadratic equation $X^2 - X - 1 = 0$. Let $(f_0, f_1, f_2, \ldots)$ be the Fibonacci sequence. Then,

$$f_n = \frac{1}{\sqrt{5}}\varphi^n - \frac{1}{\sqrt{5}}\psi^n \tag{16}$$

for every nonnegative integer n.

Theorem 2.3.1 is an explicit formula for Fibonacci numbers, and what a surprising one – would you have expected the irrational numbers φ and ψ to appear in a formula for an integer sequence? Note that the number $\varphi = \dfrac{1+\sqrt{5}}{2} \approx 1.618$ in Theorem 2.3.1 is known as the *golden ratio* (although it shares that distinction with its reciprocal $\dfrac{1}{\varphi} \approx 0.618$), and the number ψ is what is known as a *conjugate* of φ (although the notion is only properly understood in abstract algebra[24]). The golden ratio is famous for appearing in various places in mathematics, and this is one of them.

How can we prove Theorem 2.3.1? It appears reasonable to induct on n (using Theorem 2.2.2); the base case ($n = 0$) is easy, and the induction step would have us assume that (16) holds for $n = k-1$ and try to prove that (16) holds for $n = k$. Unfortunately, here we hit a wall: While the induction hypothesis lets us compute f_{k-1}, it would not let us compute f_{k-2}, but we would need both f_{k-1} and f_{k-2} in order to simplify $f_k = f_{k-1} + f_{k-2}$.

[24] In a nutshell: If a (real or complex) number ζ is a root of an irreducible polynomial P with rational coefficients, then all roots of P are called the *conjugates* of ζ. For example, the conjugates of the number $\sqrt{2}$ are $\sqrt{2}$ and $-\sqrt{2}$, since they are the roots of the irreducible polynomial $x^2 - 2$ whose root $\sqrt{2}$ is. One of the ideas of Galois theory is that conjugates of a number "belong together with it"; while we will not formalize this idea in this class, we will see its manifestations several times.

Thus we are in want of a stronger induction hypothesis: one that claims (16) not only for $n = k - 1$, but also for $n = k - 2$. Better yet, why not claim that (16) holds for all $n < k$?

This is what *strong induction* does: It strengthens the induction hypothesis from "the claim holds for $n = k - 1$" to "the claim holds for all $n < k$". The underlying principle is the following:[25]

Theorem 2.3.2. Let g be an integer. For each integer $n \geq g$, let $\mathcal{A}(n)$ be a logical statement.

Assume the following:

- *Assumption 1:* If k is an integer such that $k \geq g$ and such that

$$(\mathcal{A}(n) \text{ holds for every integer } n \geq g \text{ satisfying } n < k), \tag{17}$$

  then $\mathcal{A}(k)$ holds.

Then, $\mathcal{A}(n)$ holds for each integer $n \geq g$.

This looks a bit tortuous, so let us unravel what it means to prove something using Theorem 2.3.2. This will be called a proof by *strong induction*.

Let g be an integer (which is typically taken to be 0 or 1, as in standard induction). Say that we want to prove that some statement $\mathcal{A}(n)$ holds for every integer $n \geq g$. Doing this via standard induction (i.e., using Theorem 2.2.2) would require us to prove (in the induction step) that $\mathcal{A}(k)$ follows from $\mathcal{A}(k-1)$ (whenever $k > g$ is an integer). (It would also require an induction base, which we disregard for now.) On the other hand, doing this via strong induction (i.e., using Theorem 2.3.2) would instead require us to prove that $\mathcal{A}(k)$ follows from (17) – that is, from $\mathcal{A}(g) \wedge \mathcal{A}(g+1) \wedge \cdots \wedge \mathcal{A}(k-1)$. Obviously, when $k > g$, then $\mathcal{A}(g) \wedge \mathcal{A}(g+1) \wedge \cdots \wedge \mathcal{A}(k-1)$ is a stronger statement than $\mathcal{A}(k-1)$; thus, when we are applying strong induction, we have a stronger induction hypothesis than when applying standard induction.

There is one more difference between standard and strong induction: A strong induction needs no induction base. Indeed, in Theorem 2.2.2 there are two assumptions, but in Theorem 2.3.2 there is only one. This might appear strange, because how comes we can afford omitting the induction base? However, this is not as fishy as it looks like; it turns out that Assumption 1 in Theorem 2.3.2 already contains a "base case" in it. To be more precise, I claim that if Assumption 1 in Theorem 2.3.2 is satisfied, then $\mathcal{A}(g)$ must hold. Indeed, Assumption 1 in Theorem 2.3.2 (applied to $k = g$) says that if

$$(\mathcal{A}(n) \text{ holds for every integer } n \geq g \text{ satisfying } n < g), \tag{18}$$

then $\mathcal{A}(g)$ holds. But (18) is a vacuous statement (since there is no integer $n \geq g$

[25] Theorem 2.3.2 is [Grinbe15, Theorem 2.60], with m renamed as k.

satisfying $n < g$), and thus is vacuously true[26]; therefore, the previous sentence entails that $\mathcal{A}(g)$ holds. This argument might appear like a sleight of hand, but it is logically sound.

All of this theoretical chatter is probably less useful than an actual example of a proof by strong induction; so let us have one now:

Proof of Theorem 2.3.1. Let us try to prove Theorem 2.3.1 by strong induction on n. This means that we apply Theorem 2.3.2 to $g = 0$ and to the statement

$$\mathcal{A}(n) = \left(\text{“} f_n = \frac{1}{\sqrt{5}}\varphi^n - \frac{1}{\sqrt{5}}\psi^n \text{”} \right).$$

Induction step:[27] Let k be an integer such that $k \geq 0$. Assume that

$$(\mathcal{A}(n) \text{ holds for every integer } n \geq 0 \text{ satisfying } n < k). \tag{19}$$

We must show that $\mathcal{A}(k)$ holds. In other words, we must prove that

$$f_k = \frac{1}{\sqrt{5}}\varphi^k - \frac{1}{\sqrt{5}}\psi^k. \tag{20}$$

Of course, we would want to apply the recursive definition of the Fibonacci sequence to obtain the equality $f_k = f_{k-1} + f_{k-2}$. This is slightly complicated by the fact that this equality holds only for $k \geq 2$ (since we have not defined f_{-1} and f_{-2}); thus we need to handle the case $k < 2$ separately. Fortunately, this is straightforward (and fairly quick): If $k < 2$, then $k = 0$ or $k = 1$, and in each of these two cases we can verify (20) by hand (noticing that the definitions of φ and ψ easily imply $\varphi - \psi = \sqrt{5}$). Thus, we WLOG assume that $k \geq 2$. Hence, we do have $f_k = f_{k-1} + f_{k-2}$. Note also that (because of $k \geq 2$) we have $k - 2 \geq 0$ and $k - 1 \geq 0$.

Our assumption (19) plays the role of an induction hypothesis. In particular, we can apply it to $n = k - 1$ (since $k - 1 \geq 0$ is an integer satisfying $k - 1 < k$). We thus obtain that $\mathcal{A}(k-1)$ holds, i.e., that we have

$$f_{k-1} = \frac{1}{\sqrt{5}}\varphi^{k-1} - \frac{1}{\sqrt{5}}\psi^{k-1}. \tag{21}$$

Likewise, we can apply (19) to $n = k - 2$ (since $k - 2 \geq 0$ is an integer satisfying $k - 2 < k$). We thus conclude that $\mathcal{A}(k-2)$ holds, i.e., that we have

$$f_{k-2} = \frac{1}{\sqrt{5}}\varphi^{k-2} - \frac{1}{\sqrt{5}}\psi^{k-2}. \tag{22}$$

[26] For the meaning of the words “vacuously true”, see the Wikipedia page on “vacuous truth” (or most introductions to mathematical proof).

[27] In a strong induction, “induction step” means the part where we check that Assumption 1 in Theorem 2.3.2 holds. As we just discussed, a strong induction needs no induction base, so the induction step is the only real part of it. Thus, many writers don’t even bother to say “*Induction step:*”.

(Note that we have used $k-1 \geq 0$ and $k-2 \geq 0$ here; this is another reason why we needed to handle the case $k < 2$ separately.)

But φ is a solution of the quadratic equation $X^2 - X - 1 = 0$; thus, $\varphi^2 - \varphi - 1 = 0$, so that $\varphi + 1 = \varphi^2$. Likewise, $\psi + 1 = \psi^2$. Now,

$$\begin{aligned}
f_k &= \underbrace{f_{k-1}}_{\substack{=\frac{1}{\sqrt{5}}\varphi^{k-1} - \frac{1}{\sqrt{5}}\psi^{k-1} \\ \text{(by (21))}}} + \underbrace{f_{k-2}}_{\substack{=\frac{1}{\sqrt{5}}\varphi^{k-2} - \frac{1}{\sqrt{5}}\psi^{k-2} \\ \text{(by (22))}}} \\
&= \frac{1}{\sqrt{5}}\varphi^{k-1} - \frac{1}{\sqrt{5}}\psi^{k-1} + \frac{1}{\sqrt{5}}\varphi^{k-2} - \frac{1}{\sqrt{5}}\psi^{k-2} \\
&= \frac{1}{\sqrt{5}}\underbrace{\left(\varphi^{k-1} + \varphi^{k-2}\right)}_{=\varphi^{k-2}(\varphi+1)} - \frac{1}{\sqrt{5}}\underbrace{\left(\psi^{k-1} + \psi^{k-2}\right)}_{=\psi^{k-2}(\psi+1)} \\
&= \frac{1}{\sqrt{5}}\varphi^{k-2}\underbrace{(\varphi+1)}_{=\varphi^2} - \frac{1}{\sqrt{5}}\psi^{k-2}\underbrace{(\psi+1)}_{=\psi^2} \\
&= \frac{1}{\sqrt{5}}\underbrace{\varphi^{k-2}\varphi^2}_{=\varphi^k} - \frac{1}{\sqrt{5}}\underbrace{\psi^{k-2}\psi^2}_{=\psi^k} = \frac{1}{\sqrt{5}}\varphi^k - \frac{1}{\sqrt{5}}\psi^k.
\end{aligned}$$

In other words, (20) holds. In other words, $\mathcal{A}(k)$ holds. Thus, the induction step is complete, and we are done proving Theorem 2.3.1. $\square$

A few general remarks are in order. First of all, as we already saw in the proof of Theorem 2.3.1, the statement (17) that is being assumed in the induction step of a strong induction plays the role of an induction hypothesis; thus it is commonly referred to as the induction hypothesis. Second, while (technically speaking) our strong induction had no induction base, we nevertheless had to treat some "small-k" cases by hand (in our case, the case $k < 2$, which split into $k = 0$ and $k = 1$). Thus, we had a "de-facto" induction base in our induction step, even though we didn't **call** it an induction base. This is fairly common for proofs by strong induction.

Strong induction is a standard method for proving properties of Fibonacci numbers. Indeed, the latter are defined by a recursion ($f_k = f_{k-1} + f_{k-2}$) that refers back not just to the previous entry f_{k-1} but to the entry f_{k-2} as well; thus, when proving a claim about f_k, it is useful to have not just the corresponding claim about f_{k-1}, but also the corresponding claim about f_{k-2} at one's disposal. With its stronger induction hypothesis, strong induction is uniquely suited to providing both claims in its induction step.

Our proof of Theorem 2.3.1 was fairly straightforward, but it does nothing to demystify it. In particular, if you didn't know this formula, how could you come up with it in the first place? We will see an answer to this question in Subsection

4.9.4 (specifically, in the second proof of Theorem 4.9.11).[28]

Let me note in passing that Theorem 2.3.1 is useful as a tool in proving identities for Fibonacci numbers. In particular, Exercise 2.2.3 boils down to a straightforward computation once both sides are rewritten using Theorem 2.3.1.

Proofs by strong induction rely on Theorem 2.3.2. This theorem itself can be proved by standard induction (so it is not a new fundamental principle, but merely a "better interface" to standard induction); see [Grinbe15, proof of Theorem 2.60] for this proof. See also [Grinbe15, §2.8] for more examples of proofs by strong induction.

The next result we are going to prove using strong induction is yet another property of Fibonacci numbers, namely a combinatorial interpretation of them. This relies on the following notion:

Definition 2.3.3. A set S of integers is said to be *lacunar* if it contains no two consecutive integers (i.e., there is no $s \in S$ such that $s+1 \in S$).

For example, the sets $\{1,4\}$ and $\{3,5\}$ and $\{3,7,9\}$ are lacunar, while the sets $\{1,2,5\}$ and $\{3,4\}$ are not. (The empty set is lacunar, and so is any 1-element set of integers.)

We can now state the result we want to prove:

Theorem 2.3.4. Let $(f_0, f_1, f_2, \ldots)$ be the Fibonacci sequence. For each nonnegative integer n, let $[n]$ denote the n-element set $\{1, 2, \ldots, n\}$.

Let $n \geq 0$ be an integer. Then, the number of all lacunar subsets of $[n]$ is f_{n+2}.

Example 2.3.5. Let us verify Theorem 2.3.4 for small values of n:

- The only lacunar subset of $[0]$ is $\varnothing$ (since the set $[0]$ itself is empty). Thus, the number of all lacunar subsets of $[0]$ is 1, which is precisely what Theorem 2.3.4 says for $n = 0$ (since $f_{0+2} = f_2 = 1$).
- The only lacunar subsets of $[1]$ are $\varnothing$ and $\{1\}$ (since $[1] = \{1\}$). Thus, the number of all lacunar subsets of $[1]$ is 2, which is precisely what Theorem 2.3.4 says for $n = 1$ (since $f_{1+2} = f_3 = 2$).
- The only lacunar subsets of $[2]$ are $\varnothing$, $\{1\}$ and $\{2\}$ (since $[2] = \{1,2\}$). Thus, the number of all lacunar subsets of $[2]$ is 3, which is precisely what Theorem 2.3.4 says for $n = 2$ (since $f_{2+2} = f_4 = 3$).
- The only lacunar subsets of $[3]$ are $\varnothing$, $\{1\}$, $\{2\}$, $\{3\}$ and $\{1,3\}$ (since $[3] = \{1,2,3\}$). Thus, the number of all lacunar subsets of $[3]$ is 5, which is precisely what Theorem 2.3.4 says for $n = 3$ (since $f_{3+2} = f_5 = 5$).

[28] A rather elementary answer is found in [Grinbe15, Remark 4.3], but note that linear algebra (specifically, the notion of eigenvalues and diagonalization) provides some more motivation.

Proof of Theorem 2.3.4 (sketched). Apply strong induction on n. (That is, apply Theorem 2.3.2 with $g = 0$.)

Induction step:[29] Let $k \in \mathbb{N}$. Assume (as the induction hypothesis) that Theorem 2.3.4 holds for each integer $n \geq 0$ satisfying $n < k$. (This is precisely the assumption (17) in Theorem 2.3.2.) We must show that Theorem 2.3.4 holds for $n = k$. In other words, we must prove that the number of all lacunar subsets of $[k]$ is f_{k+2}.

Here is a sketch of how this is done (see [19fco, first proof of Proposition 1.4.9] for details): If $k < 2$, then this has already been done in Example 2.3.5; thus, we WLOG assume that $k \geq 2$. (We will see later why this assumption is needed.[30]) We shall call a subset of $[k]$

- *red* if it contains k;
- *green* if it does not contain k.

Thus, each subset of $[k]$ is either red or green (but not both). Hence, in order to count the lacunar subsets of $[k]$, we can count the red ones and the green ones separately, and then add the results.

Our induction hypothesis says that Theorem 2.3.4 holds for each integer $n \geq 0$ satisfying $n < k$. In other words, for each integer $n \geq 0$ satisfying $n < k$, we have

$$(\text{the number of all lacunar subsets of } [n]) = f_{n+2}. \tag{23}$$

Counting the green lacunar subsets of $[k]$ is easy: They are plainly the lacunar subsets of $[k-1]$. Thus,

$$\begin{aligned} &(\text{the number of all green lacunar subsets of } [k]) \\ &= (\text{the number of all lacunar subsets of } [k-1]) \\ &= f_{(k-1)+2} \qquad (\text{by (23), applied to } n = k-1) \\ &= f_{k+1}. \end{aligned} \tag{24}$$

Here, in applying (23) to $n = k-1$, we have tacitly used the facts that $k - 1 \geq 0$ (since $k \geq 2 \geq 1$) and that $k - 1 < k$ (obviously).

Now let us count the red lacunar subsets of $[k]$. These subsets contain k, and thus do not contain $k-1$ (since they are lacunar, but k and $k-1$ are consecutive integers). Hence, if we remove k from them, then we end up with lacunar subsets of $[k-2]$. Thus, to each red lacunar subset S of $[k]$ corresponds a lacunar subset $S \setminus \{k\}$ of $[k-2]$. This is a 1-to-1 correspondence[31], because conversely, if T is a

[29] As we said, no induction base is needed in a strong induction.

[30] Of course, if you were to solve a problem, you wouldn't make this WLOG assumption here; you would make it at the point you need it. I just found it better to make it here in order to uncrowd the argument later on.

[31] Here is an example of this correspondence: If $k = 5$, then the red lacunar subsets

$$\{5\}, \ \{1,5\}, \ \{2,5\}, \ \{3,5\}, \ \{1,3,5\} \qquad \text{of } [5]$$

lacunar subset of $[k-2]$, then $T \cup \{k\}$ is a red lacunar subset of $[k]$ (check this!). This 1-to-1 correspondence entails that

$$\begin{aligned}
&(\text{the number of all red lacunar subsets of } [k]) \\
&= (\text{the number of all lacunar subsets of } [k-2]) \\
&= f_{(k-2)+2} \qquad (\text{by (23), applied to } n = k-2) \\
&= f_k
\end{aligned} \tag{25}$$

(where we now have used $k-2 \geq 0$ and $k-2 < k$).

Now, recall that each subset of $[k]$ is either red or green (but not both). Hence,

$$\begin{aligned}
&(\text{the number of all lacunar subsets of } [k]) \\
&= (\text{the number of all red lacunar subsets of } [k]) \\
&\qquad + (\text{the number of all green lacunar subsets of } [k]) \qquad (26) \\
&= f_k + f_{k+1} \qquad (\text{by (25) and (24)}) \\
&= f_{k+1} + f_k = f_{k+2}
\end{aligned}$$

(since the recursive definition of the Fibonacci sequence yields $f_{k+2} = f_{k+1} + f_k$). In other words, Theorem 2.3.4 holds for $n = k$. This completes the induction step, and so Theorem 2.3.4 is proved. □

Let me stress a basic fact from set theory that we have tacitly used in the above proof:

Theorem 2.3.6. If P and Q are two disjoint finite sets, then the set $P \cup Q$ is finite as well, and satisfies

$$|P \cup Q| = |P| + |Q| . \tag{27}$$

Theorem 2.3.6 is known as the *sum rule for two sets*; it is the reason why the equality (26) holds. Indeed, since each lacunar subset of $[k]$ is either red or green, we have

$$\begin{aligned}
&\{\text{lacunar subsets of } [k]\} \\
&= \{\text{red lacunar subsets of } [k]\} \cup \{\text{green lacunar subsets of } [k]\}
\end{aligned}$$

and thus

$$\begin{aligned}
&|\{\text{lacunar subsets of } [k]\}| \\
&= |\{\text{red lacunar subsets of } [k]\} \cup \{\text{green lacunar subsets of } [k]\}| \\
&= |\{\text{red lacunar subsets of } [k]\}| + |\{\text{green lacunar subsets of } [k]\}|
\end{aligned}$$

correspond to the lacunar subsets

$$\varnothing, \ \{1\}, \ \{2\}, \ \{3\}, \ \{1,3\} \qquad \text{of } [3],$$

respectively.

(by (27), applied to

$$P = \{\text{red lacunar subsets of } [k]\} \qquad \text{and}$$
$$Q = \{\text{green lacunar subsets of } [k]\},$$

because the sets $\{\text{red lacunar subsets of } [k]\}$ and $\{\text{green lacunar subsets of } [k]\}$ are disjoint); but this is precisely the equality (26).

2.4. Recitation #1: More induction problems

Here are some exercises that use induction in non-obvious ways. The first one is about an infinite series:

Exercise 2.4.1. Prove that

$$\sum_{i=1}^{\infty} \frac{1}{i(i+1)} = 1.$$

Discussion of Exercise 2.4.1. There is no "n" here to induct over, so one might wonder whether induction really can help. But let's not give up. Induction can be very useful for computing **finite** sums; but an infinite sum is the limit of its partial sums, which are finite. So let us consider some partial sums.

For each integer $n \geq 0$, let $S(n)$ denote the sum $\sum_{i=1}^{n} \frac{1}{i(i+1)}$. This is a partial sum of the infinite sum $\sum_{i=1}^{\infty} \frac{1}{i(i+1)}$; thus,

$$\sum_{i=1}^{\infty} \frac{1}{i(i+1)} = \lim_{n\to\infty} S(n). \tag{28}$$

(This assumes that the limit $\lim_{n\to\infty} S(n)$ exists in the first place; we don't know this yet, but it will fall out of our argument in the end.)

To get a feeling for the $S(n)$, let us compute the first few values:

$$S(0) = 0, \qquad S(1) = \frac{1}{2}, \qquad S(2) = \frac{2}{3}, \qquad S(3) = \frac{3}{4}.$$

This suggests a conjecture: Namely, we suspect that

$$S(n) = \frac{n}{n+1} \qquad \text{for every integer } n \geq 0. \tag{29}$$

But this conjecture can easily be verified by induction:

[*Proof of (29):* Induct on n.

Base case: The definition of $S(0)$ yields $S(0) = \sum_{i=1}^{0} \frac{1}{i(i+1)} = (\text{empty sum}) = 0 = \frac{0}{1}$. In other words, (29) holds for $n = 0$.

Induction step: Let $m \geq 0$ be an integer. Assume (as the induction hypothesis) that (29) holds for $n = m$. We must prove that (29) holds for $n = m + 1$. In other words, we must prove that $S(m+1) = \dfrac{m+1}{m+2}$.

Our induction hypothesis says that (29) holds for $n = m$. In other words, it says that $S(m) = \dfrac{m}{m+1}$.

Both $S(m)$ and $S(m+1)$ are partial sums of $\sum\limits_{i=1}^{\infty} \dfrac{1}{i(i+1)}$; they differ only in that $S(m+1)$ has an extra addend $\dfrac{1}{(m+1)(m+2)}$ that $S(m)$ does not have. Thus,

$$S(m+1) = \underbrace{S(m)}_{=\frac{m}{m+1}} + \frac{1}{(m+1)(m+2)} = \frac{m}{m+1} + \frac{1}{(m+1)(m+2)} = \frac{m+1}{m+2}$$

(by straightforward computations). Thus, we have shown that $S(m+1) = \dfrac{m+1}{m+2}$. In other words, (29) holds for $n = m + 1$. This completes the induction step, and with it the proof of (29).]

Now, (28) becomes

$$\sum_{i=1}^{\infty} \frac{1}{i(i+1)} = \lim_{n\to\infty} \underbrace{S(n)}_{\substack{=\frac{n}{n+1} \\ \text{(by (29))}}} = \lim_{n\to\infty} \underbrace{\frac{n}{n+1}}_{=\frac{1}{1+\frac{1}{n}}} = \lim_{n\to\infty} \frac{1}{1+\frac{1}{n}} = \frac{1}{1+\lim\limits_{n\to\infty}\frac{1}{n}}$$

$$= \frac{1}{1+0} = 1.$$

This solves the exercise. □

The "induct, then take the limit" approach from the previous exercise has other uses as well. The following exercise is a more intricate example:

Exercise 2.4.2. Let $\varphi = \dfrac{1+\sqrt{5}}{2}$ (so that $\varphi \approx 1.618\ldots$). Prove that

$$\varphi = 1 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{\ddots}}}}. \tag{30}$$

Note: The infinite nested fraction on the right hand side of (30) is called an *(infinite) continued fraction*. It is rigorously defined as the limit of the sequence of

its finite initial segments – i.e., in our case, of the sequence

$$\left(1, 1+\frac{1}{1}, 1+\frac{1}{1+\frac{1}{1}}, 1+\frac{1}{1+\frac{1}{1+\frac{1}{1}}}, \ldots\right).$$

Discussion of Exercise 2.4.2. Let us first restate the claim without scary-looking nested fractions. Define a sequence $(x_1, x_2, x_3, \ldots)$ of rational numbers recursively by

$$x_1 = 1, \qquad \text{and } x_n = 1 + \frac{1}{x_{n-1}} \qquad \text{for all } n \geq 2.$$

Thus,

$$x_1 = 1, \qquad x_2 = 1+\frac{1}{1}, \qquad x_3 = 1+\frac{1}{1+\frac{1}{1}}, \qquad x_4 = 1+\frac{1}{1+\frac{1}{1+\frac{1}{1}}}, \qquad \ldots.$$

Thus, $(x_1, x_2, x_3, \ldots)$ is the sequence of finite initial segments of the infinite continued fraction on the right hand side of (30). The latter fraction is thus defined as $\lim\limits_{n\to\infty} x_n$. Hence, our claim rewrites as $\varphi = \lim\limits_{n\to\infty} x_n$. So we need to prove that $\varphi = \lim\limits_{n\to\infty} x_n$.

There are many ways to go about this; let us outline two:

First approach. Computing the first few entries of the sequence $(x_1, x_2, x_3, \ldots)$ shows that

$$x_1 = 1, \qquad x_2 = 2, \qquad x_3 = \frac{3}{2}, \qquad x_4 = \frac{5}{3}, \qquad x_5 = \frac{8}{5}, \qquad x_6 = \frac{13}{8}.$$

Both the numerators and the denominators in these fractions belong to the Fibonacci sequence $(f_0, f_1, f_2, \ldots)$ introduced in Definition 2.2.1. A quick comparison suggests the conjecture that

$$x_n = \frac{f_{n+1}}{f_n} \qquad \text{for each } n \geq 1. \tag{31}$$

This is indeed true; indeed, (31) can easily be shown by induction on n (exercise!).

Better yet, even if you forgot about the Fibonacci sequence, you could easily discover the formula (31) as follows: The values of $x_1, x_2, \ldots, x_6$ we computed above do not look very random; they are fractions[32], with the numerator of each x_i reappearing as the denominator

[32]Here we rewrite $x_1 = 1$ as $\frac{1}{1}$, and $x_2 = 2$ as $\frac{2}{1}$.

of x_{i+1}. (This is not surprising, given the recursive definition of $(x_1, x_2, x_3, \ldots)$.) This suggests that the numbers x_n have the form $x_n = \dfrac{a_{n+1}}{a_n}$ for some sequence $(a_1, a_2, a_3, \ldots)$ of integers. And indeed, we can try to define such a sequence $(a_1, a_2, a_3, \ldots)$ by setting $a_i = x_1 x_2 \cdots x_{i-1}$ for each $i \geq 1$ (which, in particular, entails that $a_1 = x_1 x_2 \cdots x_0 =$ (empty product) $= 1$, because empty products are defined to be 1). Now, the recursive equation $x_n = 1 + \dfrac{1}{x_{n-1}}$ can be restated in terms of the sequence $(a_1, a_2, a_3, \ldots)$ as follows:

$$\begin{aligned}
&\left(x_n = 1 + \frac{1}{x_{n-1}}\right) \\
\iff &\left(\frac{a_{n+1}}{a_n} = 1 + \frac{1}{\left(\dfrac{a_n}{a_{n-1}}\right)}\right) \qquad \left(\text{since } x_n = \frac{a_{n+1}}{a_n} \text{ and } x_{n-1} = \frac{a_n}{a_{n-1}}\right) \\
\iff &\left(\frac{a_{n+1}}{a_n} = 1 + \frac{a_{n-1}}{a_n}\right) \\
\iff &\left(a_{n+1} = a_n + a_{n-1}\right) \qquad \left(\text{here, we multiplied both sides by } a_n\right).
\end{aligned}$$

Thus, the sequence $(a_1, a_2, a_3, \ldots)$ must satisfy the recursive equation $a_{n+1} = a_n + a_{n-1}$ for each $n \geq 2$. But this is the same recursive equation that the Fibonacci sequence $(f_0, f_1, f_2, \ldots)$ satisfies! Moreover, the first two entries a_1 and a_2 of the sequence $(a_1, a_2, a_3, \ldots)$ are the second and third entries f_1 and f_2 of the Fibonacci sequence $(f_0, f_1, f_2, \ldots)$ (since $a_1 = 1 = f_1$ and $a_2 = x_1 = 1 = f_2$). Thus, the sequence $(a_1, a_2, a_3, \ldots)$ is precisely the "shifted" Fibonacci sequence $(f_1, f_2, f_3, \ldots)$ ("shifted" in the sense that the first entry $f_0 = 1$ is discarded). Now, $x_n = \dfrac{a_{n+1}}{a_n}$ becomes $x_n = \dfrac{f_{n+1}}{f_n}$, so we have arrived at (31) without relying on our integer sequence identification skills.

What can we do with (31)? We still need to prove that $\varphi = \lim\limits_{n\to\infty} x_n$. While (31) gives a formula for x_n in terms of Fibonacci numbers, we still need to find $\lim\limits_{n\to\infty} x_n$ somehow.

Binet's formula (Theorem 2.3.1) comes to our help now. Set $\psi = \dfrac{1-\sqrt{5}}{2}$. Note that our φ and ψ here are precisely the φ and ψ from Theorem 2.3.1. Thus, for every nonnegative integer n, we have

$$\begin{aligned}
f_n &= \frac{1}{\sqrt{5}}\varphi^n - \frac{1}{\sqrt{5}}\psi^n \qquad \text{(by Theorem 2.3.1)} \\
&= \frac{1}{\sqrt{5}}\left(\varphi^n - \psi^n\right) = \frac{1}{\sqrt{5}}\varphi^n\left(1 - \frac{\psi^n}{\varphi^n}\right) = \frac{1}{\sqrt{5}}\varphi^n\left(1 - \left(\frac{\psi}{\varphi}\right)^n\right) \\
&= \frac{1}{\sqrt{5}}\varphi^n\left(1 - \rho^n\right), \qquad (32)
\end{aligned}$$

where we have set $\rho := \dfrac{\psi}{\varphi}$. Note that this ρ is explicitly given by $\rho = \dfrac{1-\sqrt{5}}{1+\sqrt{5}} =$

$\dfrac{\sqrt{5}-3}{2} \approx -0.381\,97\ldots$.

Now, for each $n \geq 1$, we have

$$\begin{aligned} x_n &= \frac{f_{n+1}}{f_n} \qquad \text{(by (31))} \\ &= \underbrace{f_{n+1}}_{\substack{=\frac{1}{\sqrt{5}}\varphi^{n+1}\left(1-\rho^{n+1}\right) \\ \text{(by (32))}}} \Big/ \underbrace{f_n}_{\substack{=\frac{1}{\sqrt{5}}\varphi^{n}\left(1-\rho^{n}\right) \\ \text{(by (32))}}} \\ &= \left(\frac{1}{\sqrt{5}}\varphi^{n+1}\left(1-\rho^{n+1}\right)\right) \Big/ \left(\frac{1}{\sqrt{5}}\varphi^{n}\left(1-\rho^{n}\right)\right) = \varphi \cdot \frac{1-\rho^{n+1}}{1-\rho^{n}}. \end{aligned}$$

Hence,

$$\lim_{n\to\infty} x_n = \lim_{n\to\infty}\left(\varphi \cdot \frac{1-\rho^{n+1}}{1-\rho^{n}}\right) = \varphi \cdot \lim_{n\to\infty} \frac{1-\rho^{n+1}}{1-\rho^{n}}. \tag{33}$$

The limit on the right hand side is now easy to find: Since $|\rho| < 1$, we have $\rho^n \to 0$ as $n \to \infty$, and therefore also $\rho^{n+1} \to 0$ as $n \to \infty$. Hence, $\dfrac{1-\rho^{n+1}}{1-\rho^{n}} \to \dfrac{1-0}{1-0} = 1$ as $n \to \infty$. In other words, $\lim\limits_{n\to\infty} \dfrac{1-\rho^{n+1}}{1-\rho^{n}} = 1$. Hence, (33) becomes

$$\lim_{n\to\infty} x_n = \varphi \cdot \underbrace{\lim_{n\to\infty} \frac{1-\rho^{n+1}}{1-\rho^{n}}}_{=1} = \varphi.$$

This proves $\varphi = \lim\limits_{n\to\infty} x_n$, and thus solves the exercise.

Second approach. The first approach showed above relied on an exact computation of the first few entries of the sequence $(x_1, x_2, x_3, \ldots)$. In contrast, for our second approach (which is mostly taken from [GelAnd17, §3.1.4]), let us compute them approximately:

$$x_1 = 1, \qquad x_2 = 2, \qquad x_3 = 1.5, \qquad x_4 \approx 1.667, \qquad x_5 = 1.6, \qquad x_6 \approx 1.625.$$

This creates the impression that the sequence $(x_1, x_2, x_3, \ldots)$ converges to $\varphi \approx 1.618$ alternatingly from the bottom and the top: the odd-numbered entries $x_1, x_3, x_5, \ldots$ appear to be smaller than φ, while the even-numbered entries $x_2, x_4, x_6, \ldots$ appear to be larger than φ. Better yet, it seems that

$$x_2 < x_4 < x_6 < \cdots < \varphi < \cdots < x_5 < x_3 < x_1. \tag{34}$$

While this alone would not suffice to solve the exercise, it seems like a good step forward, so let us try to prove this. We need to show the following two statements:

- If $i \geq 1$ satisfies $x_i > \varphi$, then $x_{i+1} < \varphi$ and $x_{i+2} < x_i$.

- If $i \geq 1$ satisfies $x_i < \varphi$, then $x_{i+1} > \varphi$ and $x_{i+2} > x_i$.

Both of these statements follow by some manipulation of inequalities.[33] Thus, (34) is proved.

Now, (34) shows that the sequence $(x_1, x_3, x_5, \ldots)$ is decreasing and bounded from below; hence, the monotone convergence theorem shows that this sequence $(x_1, x_3, x_5, \ldots)$ has a limit $a := \lim\limits_{n\to\infty} x_{2n+1}$. Likewise, the sequence $(x_2, x_4, x_6, \ldots)$ has a limit $b := \lim\limits_{n\to\infty} x_{2n}$ (since (34) shows that this sequence is increasing and bounded from above). We shall now show that $a = b = \varphi$.

Indeed, we have

$$a = \lim_{n\to\infty} \underbrace{x_{2n+1}}_{=1+\frac{1}{x_{2n}}} = \lim_{n\to\infty}\left(1+\frac{1}{x_{2n}}\right) = 1+\frac{1}{\lim\limits_{n\to\infty} x_{2n}} = 1+\frac{1}{b} \qquad \left(\text{since } \lim_{n\to\infty} x_{2n} = b\right)$$

and

$$b = \lim_{n\to\infty} x_{2n} = \lim_{n\to\infty} \underbrace{\underbrace{x_{2(n+1)}}_{=x_{2n+2}}}_{=1+\frac{1}{x_{2n+1}}} \qquad \text{(since we can substitute } n+1 \text{ for } n \text{ in a limit)}$$

$$= \lim_{n\to\infty}\left(1+\frac{1}{x_{2n+1}}\right) = 1+\frac{1}{\lim\limits_{n\to\infty} x_{2n+1}} = 1+\frac{1}{a} \qquad \left(\text{since } \lim_{n\to\infty} x_{2n+1} = a\right).$$

How can we solve these two equalities for a and b ? The simplest way is perhaps the following: Multiplying the equation $a = 1+\frac{1}{b}$ by b, we find $ab = b+1$; on

[33] For example, let us prove the first statement. So let $i \geq 1$ be an integer satisfying $x_i > \varphi$. We must prove that $x_{i+1} < \varphi$ and $x_{i+2} < x_i$. The recursive definition of x_{i+1} yields $x_{i+1} = 1+\frac{1}{x_i} < 1+\frac{1}{\varphi}$ (since $x_i > \varphi$). But a straightforward computation shows that $1+\frac{1}{\varphi} = \varphi$. Hence, $x_{i+1} < 1+\frac{1}{\varphi} = \varphi$. Furthermore, the recursive definition of x_{i+2} yields

$$\begin{aligned} x_{i+2} &= 1+\frac{1}{x_{i+1}} = 1+\frac{1}{1+\frac{1}{x_i}} \qquad \left(\text{since } x_{i+1} = 1+\frac{1}{x_i}\right) \\ &= 1+\frac{x_i}{1+x_i} < x_i, \end{aligned}$$

where the last inequality is a consequence of the fact that $x_i > \varphi$ and the (straightforward) fact that $1+\frac{x}{1+x} < x$ for every real $x > \varphi$. Thus, both $x_{i+1} < \varphi$ and $x_{i+2} < x_i$ are proved.

The second statement is proved similarly, except that this time $1+\frac{x_i}{1+x_i} > x_i$ follows from $x_i < \varphi$ and $x_i \geq 0$. We leave the details to the reader.

the other hand, multiplying the equation $b = 1 + \dfrac{1}{a}$ by a, we find $ab = a + 1$. Comparing $ab = a + 1$ with $ab = b + 1$, we find $a + 1 = b + 1$, thus $a = b$. However, $b = \lim\limits_{n\to\infty} \underbrace{x_{2n}}_{\substack{<\varphi \\ \text{(by (34))}}} \leq \lim\limits_{n\to\infty} \varphi = \varphi$ and $a = \lim\limits_{n\to\infty} \underbrace{x_{2n+1}}_{\substack{>\varphi \\ \text{(by (34))}}} \geq \lim\limits_{n\to\infty} \varphi = \varphi$, so that $\varphi \leq a$. Combined with $a = b \leq \varphi$, this yields $a = \varphi$. Similarly, $b = \varphi$.

Now, we know that both subsequences $(x_1, x_3, x_5, \ldots)$ and $(x_2, x_4, x_6, \ldots)$ of the sequence $(x_1, x_2, x_3, \ldots)$ converge to φ (since $\lim\limits_{n\to\infty} x_{2n+1} = a = \varphi$ and $\lim\limits_{n\to\infty} x_{2n} = b = \varphi$). By basic properties of limits, this implies that the whole sequence $(x_1, x_2, x_3, \ldots)$ converges to φ. In other words, $\varphi = \lim\limits_{n\to\infty} x_n$. Again, the exercise is solved. $\square$

Here is another exercise that does not look like an induction problem, yet is one:

Exercise 2.4.3. We say that a number is *funny* if it can be written in the form

$$\pm 1^2 \pm 2^2 \pm 3^2 \pm \cdots \pm m^2$$

for some nonnegative integer m and some choice of $\pm$ signs. For example, 4 is funny because $4 = -1^2 - 2^2 + 3^2$, whereas 5 is funny because $5 = +1^2 + 2^2$. Also, 0 is funny since $0 = (\text{empty sum})$ (this corresponds to choosing $m = 0$).

Prove that every integer is funny.

Discussion of Exercise 2.4.3. I will not give a full solution, but here is a sequence of hints that should suffice:

1. First, let's try to solve the "little brother" of the problem, which is obtained by replacing the squares by 1-st powers: We say that a number is *giggly* if it can be written in the form

   $$\pm 1 \pm 2 \pm 3 \pm \cdots \pm m$$

   for some nonnegative integer m and some choice of $\pm$ signs. Prove that every integer is giggly.

2. It suffices to prove that every positive integer is giggly (resp. funny), since flipping all the $\pm$ signs will flip the sign of the number.

3. If some number n is giggly, then so is $n + 1$, because $n = \pm 1 \pm 2 \pm 3 \pm \cdots \pm m$ entails $n + 1 = \pm 1 \pm 2 \pm 3 \pm \cdots \pm m - (m + 1) + (m + 2)$.

4. So much for the "little brother". What about the original problem? We need an analogue of the formula $-(m + 1) + (m + 2) = 1$ that we just used.

5. The most obvious thing to try is $-(m + 1)^2 + (m + 2)^2$. Unfortunately, this is not 1 but $2m + 3$, but this is already a good step forward, since it is linear (not quadratic) in m.

6. Now, let us subtract $-(m+1)^2 + (m+2)^2 = 2m+3$ from $-(m+3)^2 + (m+4)^2 = 2(m+2)+3$, since this should get rid of the linear term and leave a constant behind. We obtain

$$\begin{aligned}
&(m+1)^2 - (m+2)^2 - (m+3)^2 + (m+4)^2 \\
&= -\underbrace{\left((m+1)^2 - (m+2)^2\right)}_{=2m+3} + \underbrace{\left(-(m+3)^2 + (m+4)^2\right)}_{=2(m+2)+3=2m+7} \\
&= -(2m+3) + (2m+7) = 4. \qquad (35)
\end{aligned}$$

7. Thus, if some number n is funny, then so is $n+4$.

8. Hence, by strong induction, it suffices to show that every $n \in \{0,1,2,3\}$ is funny.

9. Do it! $\square$

This ends our first excursion into the uses of induction. We will see more uses, and even more variants of induction, in the following chapters.

3. Number Theory I: Divisibility and congruence

We shall now discuss some of the most classical results in mathematics: the properties of divisibility and modular congruence of integers. Most of them were known to Euclid, although the concept of congruence was first defined by Gauss. These results are doubly important: first, these facts are basic and commonly used in contest mathematics; second, they make good examples for induction proofs. Proofs will be mostly skipped or sketched; detailed proofs can be found in every text on elementary number theory (or in [19s, Chapter 2], from which I am copypasting some of the statements). We are not trying to be comprehensive or detailed; there are better sources for this (e.g., [Burton11], [NiZuMo91] or [UspHea39], or, for a shorter introduction, [Dudley12]).

3.1. Quotients and remainders

First, a few definitions. I begin by weighing in on one of the most controversial issues of our time:

Definition 3.1.1. The symbol $\mathbb{N}$ will denote the set $\{0, 1, 2, \ldots\}$ of all nonnegative integers.

(The faultline between the "$\mathbb{N} = \{0, 1, 2, \ldots\}$ people" and the "$\mathbb{N} = \{1, 2, 3, \ldots\}$ people" is rather close to that between algebraists/combinatorialists and number theorists; as you can see, I am one of the former.)

Definition 3.1.2. Let a and b be two integers. We say that $a \mid b$ (or "*a divides b*" or "b is *divisible by a*" or "b is a *multiple* of a" or "a is a *divisor* of b") if there exists an integer c such that $b = ac$.

We furthermore say that $a \nmid b$ if a does not divide b.

This definition, too, is a bit controversial: It implies that $0 \mid 0$. Some authors don't like this (arguing that $a \mid b$ should mean that b/a is uniquely determined, which $0/0$ is not); while I understand their thinking, I believe that forbidding 0 from dividing itself would create more headaches than it would prevent. The words "divides" and "divisor", too, can mean different things depending on whom you ask; in particular, Knuth (e.g. in [GrKnPa94]) likes to define them differently from how I do[34].

Note that $1 \mid b$ holds for any $b \in \mathbb{Z}$; but $0 \mid b$ holds only for $b = 0$. Note also that $a \mid -a$ for each $a \in \mathbb{Z}$.

[34] He defines "a divides b" to mean "$b > 0$, and there exists an integer c such that $b = ac$". This is to be distinguished from "b is a multiple of a", which he (like me) defines to mean only "there exists an integer c such that $b = ac$". Thus, quoting [GrKnPa94, §4.1]: "Every integer is a multiple of -1, but no integer is divisible by -1 (strictly speaking)". Again, I understand the reasons for this (e.g., this way, the divisors of 6 are $1, 2, 3, 6$ rather than the unnecessarily duplicated $-6, -3, -2, -1, 1, 2, 3, 6$), but I am not convinced that it is worth the headache.

Here are some basic properties of divisibility (all of which are easy to prove[35]):

Proposition 3.1.3. Let a and b be two integers.

(a) We have $a \mid b$ if and only if $|a| \mid |b|$. (Here, "$|a| \mid |b|$" means "$|a|$ divides $|b|$".)

(b) If $a \mid b$ and $b \neq 0$, then $|a| \leq |b|$.

(c) If $a \mid b$ and $b \mid a$, then $|a| = |b|$.

(d) Assume that $a \neq 0$. Then, $a \mid b$ if and only if $\dfrac{b}{a} \in \mathbb{Z}$.

Proposition 3.1.4. **(a)** We have $a \mid a$ for every $a \in \mathbb{Z}$. (This is called the *reflexivity of divisibility*.)

(b) If $a, b, c \in \mathbb{Z}$ satisfy $a \mid b$ and $b \mid c$, then $a \mid c$. (This is called the *transitivity of divisibility*.)

(c) If $a_1, a_2, b_1, b_2 \in \mathbb{Z}$ satisfy $a_1 \mid b_1$ and $a_2 \mid b_2$, then $a_1 a_2 \mid b_1 b_2$.

(d) If $a, b \in \mathbb{Z}$ satisfy $a \mid b$, then $a^k \mid b^k$ for any nonnegative integer k.

Proposition 3.1.5. Let a, b, c be three integers such that $c \neq 0$. Then, $a \mid b$ holds if and only if $ac \mid bc$.

Proposition 3.1.6. Let $n \in \mathbb{Z}$. Let $a, b \in \mathbb{N}$ be such that $a \leq b$. Then, $n^a \mid n^b$.

Proposition 3.1.7. Let $a, b, c \in \mathbb{Z}$ be such that $a \mid b$ and $a \mid c$. Then, $a \mid b + c$ and $a \mid b - c$.

We shall use these facts many times (and often without saying). They provide basic rules for manipulating divisibilities[36]. For example, Proposition 3.1.4 **(c)** allows multiplying two divisibilities, while Proposition 3.1.5 allows multiplying both sides of a divisibility by a nonzero integer c or, conversely, cancelling the factor c from both sides. Proposition 3.1.3 **(a)** ensures that both sides of a divisibility can be replaced by their absolute values, which is why it often suffices to consider nonnegative integers in divisibility arguments.

The next fact ([19s, Theorem 2.6.1]) is significantly more important:

Theorem 3.1.8. Let n be a positive integer. Let $u \in \mathbb{Z}$. Then, there exists a unique pair $(q, r) \in \mathbb{Z} \times \{0, 1, \ldots, n-1\}$ such that $u = qn + r$.

Theorem 3.1.8 essentially says that any integer u can be uniquely divided with remainder by any positive integer n. The entries q and r of the pair (q, r) in Theorem 3.1.8 are called the *quotient* and the *remainder* of this division.

[35] For what it's worth: Proposition 3.1.3 is a combination of [19s, Proposition 2.2.3 and Exercise 2.2.2]. Proposition 3.1.4 is a combination of [19s, Proposition 2.2.4 and Exercise 2.2.6]. Proposition 3.1.5 is [19s, Exercise 2.2.3]. Proposition 3.1.6 is [19s, Exercise 2.2.4]. Finally, Proposition 3.1.7 follows easily from [19s, Proposition 2.3.4 **(d)**] or can be straightforwardly proved using the definition of divisibility.

[36] A *divisibility* means a statement of the form "$a \mid b$".

We shall not prove Theorem 3.1.8 in detail, but the reader is advised to recall (or construct) at least one proof – as it is a nice exercise on induction. One detailed proof (using strong induction) is given in [19s, §2.6]; another appears in [Grinbe15, proof of Theorem 2.153]. Let me say a few words about the second proof, as it illustrates a little variation on standard induction: *two-sided induction*. This means using the following *two-sided induction principle* ([Grinbe15, Theorem 2.149]):

Theorem 3.1.9. Let $g \in \mathbb{Z}$. For each integer n, let $\mathcal{A}(n)$ be a logical statement. Assume the following:

- *Assumption 1:* The statement $\mathcal{A}(g)$ holds.
- *Assumption 2:* If m is an integer such that $m \geq g$ and such that $\mathcal{A}(m)$ holds, then $\mathcal{A}(m+1)$ also holds.
- *Assumption 3:* If m is an integer such that $m \leq g$ and such that $\mathcal{A}(m)$ holds, then $\mathcal{A}(m-1)$ also holds.

Then, $\mathcal{A}(n)$ holds for all integers n.

While standard induction (Theorem 2.1.1) and strong induction (Theorem 2.3.2) are tailored for proving theorems about nonnegative (or positive) integers, two-sided induction (Theorem 3.1.9) is best suited for proving theorems about all integers. Thus, a two-sided induction has an induction base and **two** induction steps, one of which goes "upwards" (from $\mathcal{A}(m)$ to $\mathcal{A}(m+1)$) while the other goes "downwards" (from $\mathcal{A}(m)$ to $\mathcal{A}(m-1)$). Now, Theorem 3.1.8 can be proved by two-sided induction on u (for fixed n). More precisely, we can use two-sided induction on u to prove the existence of a pair $(q, r) \in \mathbb{Z} \times \{0, 1, \ldots, n-1\}$ such that $u = qn + r$; the uniqueness of this pair is easiest to prove directly. (See [Grinbe15, proof of Theorem 2.153] for details.)

3.2. Modular arithmetic I: Congruences

Judged by their definition, *congruences* – or, more precisely, *modular congruences* – are just reformulated divisibilities. But the reformulation is worth it, as it exposes their most useful qualities. In this section, we will define congruences and state their most basic properties; later we shall return to take a deeper look at them.

First, the definition:

Definition 3.2.1. Let $n, a, b \in \mathbb{Z}$. We say that *a is congruent to b modulo n* if and only if $n \mid a - b$. We shall use the notation "$a \equiv b \bmod n$" for "a is congruent to b modulo n".

We furthermore shall use the notation "$a \not\equiv b \bmod n$" for "a is not congruent to b modulo n".

Example 3.2.2. **(a)** Is $3 \equiv 7 \bmod 2$? Yes, since $2 \mid 3 - 7 = -4$.

(b) Is $3 \equiv 6 \bmod 2$? No, since $2 \nmid 3 - 6 = -3$. So we have $3 \not\equiv 6 \bmod 2$.

Now, let a and b be two integers.

(c) We have $a \equiv b \bmod 0$ if and only if $a = b$. (Indeed, $a \equiv b \bmod 0$ is defined to mean $0 \mid a - b$, but the latter divisibility happens only when $a - b = 0$, which is tantamount to saying $a = b$.)

(d) We have $a \equiv b \bmod 1$ always, since $1 \mid a - b$ always holds (remember: 1 divides everything).

Statements of the form "$a \equiv b \bmod n$" are called *congruences* (just as statements of the form "$a = b$" are called equalities). The number n is called the *modulus* of the congruence $a \equiv b \bmod n$. The following properties of congruences are proved in [19s, §2.3] (but the proofs make good exercises!):

Proposition 3.2.3. Let $n \in \mathbb{Z}$ and $a \in \mathbb{Z}$. Then, $a \equiv 0 \bmod n$ if and only if $n \mid a$. (Thus, in particular, $n \equiv 0 \bmod n$ always holds.)

Proposition 3.2.4. Let $a, b, n \in \mathbb{Z}$. Then, $a \equiv b \bmod n$ if and only if there exists some $d \in \mathbb{Z}$ such that $b = a + nd$.

Proposition 3.2.5. Let $a, b, c, n \in \mathbb{Z}$. Then, $a - b \equiv c \bmod n$ if and only if $a \equiv b + c \bmod n$.

Proposition 3.2.6. Let $n \in \mathbb{Z}$.

(a) We have $a \equiv a \bmod n$ for every $a \in \mathbb{Z}$. (This is called the *reflexivity of congruence*.)

(b) If $a, b, c \in \mathbb{Z}$ satisfy $a \equiv b \bmod n$ and $b \equiv c \bmod n$, then $a \equiv c \bmod n$. (This is called the *transitivity of congruence*.)

(c) If $a, b \in \mathbb{Z}$ satisfy $a \equiv b \bmod n$, then $b \equiv a \bmod n$. (This is called the *symmetry of congruence*.)

(d) If $a_1, a_2, b_1, b_2 \in \mathbb{Z}$ satisfy $a_1 \equiv b_1 \bmod n$ and $a_2 \equiv b_2 \bmod n$, then

$$a_1 + a_2 \equiv b_1 + b_2 \bmod n; \tag{36}$$

$$a_1 - a_2 \equiv b_1 - b_2 \bmod n; \tag{37}$$

$$a_1 a_2 \equiv b_1 b_2 \bmod n. \tag{38}$$

(e) Let $m \in \mathbb{Z}$ be such that $m \mid n$. If $a, b \in \mathbb{Z}$ satisfy $a \equiv b \bmod n$, then $a \equiv b \bmod m$.

Proposition 3.2.7. Let $n, a, b \in \mathbb{Z}$ be such that $a \equiv b \bmod n$. Then, $a^k \equiv b^k \bmod n$ for each $k \in \mathbb{N}$.

Proposition 3.2.8. Let n be an integer. Let S be a finite set. For each $s \in S$, let a_s and b_s be two integers. Assume that

$$a_s \equiv b_s \bmod n \qquad \text{for each } s \in S. \tag{39}$$

Then,

$$\sum_{s \in S} a_s \equiv \sum_{s \in S} b_s \bmod n \tag{40}$$

and

$$\prod_{s \in S} a_s \equiv \prod_{s \in S} b_s \bmod n. \tag{41}$$

Proposition 3.2.6 **(d)** shows that any two congruences can be added, subtracted and multiplied like equalities (as long as their moduli are equal; e.g., we cannot add the congruences $1 \equiv 3 \bmod 2$ and $1 \equiv 4 \bmod 3$, because their moduli 2 and 3 are distinct)[37]. Proposition 3.2.8 extends this to any (finite) number of congruences (instead of just two). Proposition 3.2.7 says that a congruence can be taken to the k-th power for every $k \in \mathbb{N}$ (again like an equality).[38] There is yet another way in which congruences behave like equalities: Namely, they can be chained together like equalities. To state this precisely, we need a definition:

Definition 3.2.9. Let $n \in \mathbb{Z}$. If $a_1, a_2, \ldots, a_k$ are k integers, then the statement "$a_1 \equiv a_2 \equiv \cdots \equiv a_k \bmod n$" shall mean that

$$a_i \equiv a_{i+1} \bmod n \text{ holds for each } i \in \{1, 2, \ldots, k-1\}.$$

(In other words, it shall mean that $a_1 \equiv a_2 \bmod n$ and $a_2 \equiv a_3 \bmod n$ and $a_3 \equiv a_4 \bmod n$ and $\cdots$ and $a_{k-1} \equiv a_k \bmod n$. This is vacuously true when $k \leq 1$. If $k = 2$, then it simply means that $a_1 \equiv a_2 \bmod n$.)

Such a statement will be called a *chain of congruences modulo n*.

Proposition 3.2.10. Let $n \in \mathbb{Z}$. Let $a_1, a_2, \ldots, a_k$ be k integers such that $a_1 \equiv a_2 \equiv \cdots \equiv a_k \bmod n$. Let u and v be two elements of $\{1, 2, \ldots, k\}$. Then, $a_u \equiv a_v \bmod n$.

Proposition 3.2.10 follows easily from Proposition 3.2.6 by induction. (See [Grinbe15, proof of Proposition 2.16] for the details of this proof.)

Proposition 3.2.10 allows chaining congruences (with equal moduli) together. Thus, for example, we can quickly see that $7^{15} \equiv 7 \bmod 8$ via the following compu-

[37] That said, congruences cannot be **divided**. That is, if we have $a_1 \equiv b_1 \bmod n$ and $a_2 \equiv b_2 \bmod n$, then we **cannot** conclude that $a_1/a_2 \equiv b_1/b_2 \bmod n$, even if we assume something like $a_2 \not\equiv 0 \bmod n$ and $b_2 \not\equiv 0 \bmod n$. (Most of the time, the congruence $a_1/a_2 \equiv b_1/b_2 \bmod n$ will be meaningless, since a_1/a_2 and b_1/b_2 are usually not integers. But even if they are integers, the congruence need not hold.)

[38] Note, however, that $a \equiv b \bmod n$ does **not** imply $k^a \equiv k^b \bmod n$.

tation: Since $7 \equiv -1 \bmod 8$, we have

$$\begin{aligned} 7^{15} &\equiv \underbrace{(-1)^{15}}_{=-1} \qquad \text{(by Proposition 3.2.7, applied to } n=8,\ a=7,\ b=-1 \text{ and } k=15) \\ &\equiv -1 \equiv 7 \bmod 8. \end{aligned}$$

Note that the second "$\equiv$" sign in this chain is actually an equality sign, because every equality $a = b$ is automatically a congruence $a \equiv b \bmod n$ (by Proposition 3.2.6 **(a)**).

Proposition 3.1.4 **(b)** shows that divisibilities can also be chained together:

Definition 3.2.11. If $a_1, a_2, \ldots, a_k$ are k integers, then the statement "$a_1 \mid a_2 \mid \cdots \mid a_k$" shall mean that

$$a_i \mid a_{i+1} \text{ holds for each } i \in \{1, 2, \ldots, k-1\}.$$

(In other words, it shall mean that $a_1 \mid a_2$ and $a_2 \mid a_3$ and $a_3 \mid a_4$ and $\cdots$ and $a_{k-1} \mid a_k$. Again, this is vacuously true when $k \leq 1$.)

Such a statement will be called a *chain of divisibilities*.

Thus, the analogue of Proposition 3.2.10 for divisibilities is the following:

Proposition 3.2.12. Let $a_1, a_2, \ldots, a_k$ be k integers such that $a_1 \mid a_2 \mid \cdots \mid a_k$. Let u and v be two elements of $\{1, 2, \ldots, k\}$ such that $u \leq v$. Then, $a_u \mid a_v$.

For example, if four integers a, b, c, d satisfy $a \mid b \mid c \mid d$ (by which we mean $a \mid b$ and $b \mid c$ and $c \mid d$), then $a \mid d$. Note that we required $u \leq v$ in Proposition 3.2.12, because chains of divisibilities cannot be reversed (e.g., we cannot derive $d \mid a$ from $a \mid b \mid c \mid d$). Proposition 3.2.12 is easily proved by induction.

Here is a sample exercise to illustrate manipulation of congruences:

Exercise 3.2.1. Let $n \in \mathbb{N}$. Show that $7 \mid 3^{2n+1} + 2^{n+2}$.

Solution to Exercise 3.2.1. We have $3^2 = 9 \equiv 2 \bmod 7$. Thus, Proposition 3.2.7 (applied to 7, 3^2, 2 and n instead of n, a, b and k) yields $\left(3^2\right)^n \equiv 2^n \bmod 7$. Multiplying this congruence[39] by the obvious congruence $3 \equiv 3 \bmod 7$, we obtain $\left(3^2\right)^n \cdot 3 \equiv 2^n \cdot 3 \bmod 7$. Thus, $3^{2n+1} = \left(3^2\right)^n \cdot 3 \equiv 2^n \cdot 3 \bmod 7$. On the other hand, $2^{n+2} = 2^n \cdot 2^2 = 2^n \cdot 4$, so that $2^{n+2} \equiv 2^n \cdot 4 \bmod 7$ (since any equality is a congruence).

Now, adding the two congruences[40] $3^{2n+1} \equiv 2^n \cdot 3 \bmod 7$ and $2^{n+2} \equiv 2^n \cdot 4 \bmod 7$, we obtain

$$3^{2n+1} + 2^{n+2} \equiv 2^n \cdot 3 + 2^n \cdot 4 = 2^n \cdot \underbrace{(3+4)}_{=7} = 2^n \cdot 7 \equiv 0 \bmod 7$$

[39] i.e., applying (38)

[40] i.e., applying (36)

(since $2^n \cdot 7$ is clearly divisible by 7). In other words, $7 \mid 3^{2n+1} + 2^{n+2}$. This solves Exercise 3.2.1. $\square$

Congruences can not only be added, subtracted, multiplied and chained together, but they can also be substituted into one another, in the following sense:[41]

Example 3.2.13. Let n be an integer. Assume you have two integers a and b that are congruent modulo n (that is, they satisfy $a \equiv b \bmod n$). Then, if you have any polynomial expression P involving a, then you can substitute b for a in this expression, and obtain a new expression Q that satisfies $P \equiv Q \bmod n$. For example,

$$\begin{aligned} a + 2 &\equiv b + 2 \bmod n; \\ a^5 &\equiv b^5 \bmod n; \\ (a+2)(a+9) - a &\equiv (b+2)(b+9) - b \bmod n; \\ a^2 + c^2 + ac &\equiv b^2 + c^2 + bc \bmod n \qquad \text{for any integer } c; \\ \sum_{i=0}^{k} a^i &\equiv \sum_{i=0}^{k} b^i \bmod n \qquad \text{for every } k \in \mathbb{N}. \end{aligned}$$

This is called the *substitution principle for congruences*. I will not formalize this principle (see [19s, §2.5] for a more detailed treatment, which too stops short of properly formalizing it), but I will make four comments:

First, you don't have to replace **every** a by b when substituting; you can choose **some** a's to replace by b's while leaving the remaining a's unchanged. Thus, for example, $a \equiv b \bmod n$ yields $a^7 + a^4 + a \equiv b^7 + a^4 + b \bmod n$ and $a^7 + a^4 + a \equiv b^7 + a^4 + a \bmod n$ and various other such congruences.

Second, it is worth stressing that the only powers that can appear in a polynomial expression are powers with constant exponents. In particular, we cannot have "2^a" in a polynomial expression. This is important because we could **not** substitute b for a in 2^a; it is not usually true that $a \equiv b \bmod n$ implies $2^a \equiv 2^b \bmod n$.

Third, when using the substitution principle for congruences, I will use underbraces to point out what is being replaced by what. Thus, for example, when substituting b's for the three a's in $(a+2)(a+9) - a$, I will write

$$\left(\underbrace{a}_{\equiv b \bmod n} + 2\right)\left(\underbrace{a}_{\equiv b \bmod n} + 9\right) - \underbrace{a}_{\equiv b \bmod n} \equiv (b+2)(b+9) - b \bmod n.$$

(This is the same convention that I am using when substituting equal things in an equality.)

[41] Here, a "*polynomial expression*" means an expression that contains only variables, integers, the symbols "$+$", "$-$" and "$\cdot$" and powers with constant exponents (i.e., things like "a^3" or "a^9", but never "2^a" or "a^b") such that all exponents are nonnegative integers (so "a^{-1}" is not allowed either).

Fourth, instead of proving the substitution principle for congruences, let me explain how any specific application of this principle (i.e., any substitution in a congruence) can be justified by references to Proposition 3.2.6, Proposition 3.2.7 and Proposition 3.2.8. For example, let's say we want to justify that $a \equiv b \bmod n$ entails $(a+2)(a+9) - a \equiv (b+2)(b+9) - b \bmod n$. To do so, we first add the two congruences $a \equiv a \bmod n$ and $2 \equiv 2 \bmod n$ to obtain $a + 2 \equiv b + 2 \bmod n$. Likewise, we can get $a + 9 \equiv b + 9 \bmod n$. Multiplying the latter two congruences, we find $(a+2)(a+9) \equiv (b+2)(b+9) \bmod n$. Subtracting the congruence $a \equiv b \bmod n$ from the latter congruence, we obtain $(a+2)(a+9) - a \equiv (b+2)(b+9) - b \bmod n$, as desired. Similar reasoning can be used to rigorously prove any instance of the substitution principle for congruences. Thus, we don't really need the substitution principle; we can always circumvent it (by adding, subtracting and multiplying congruences, and by taking congruences to powers). However, the substitution principle saves time and mental effort.

The following exercise is a slightly more intricate example of the use of congruences:

Exercise 3.2.2. Let $(f_0, f_1, f_2, \ldots)$ be the Fibonacci sequence. Prove the following: If $a, b \in \mathbb{N}$ satisfy $a \mid b$, then $f_a \mid f_b$.

For example, this exercise entails that $f_7 \mid f_{21}$ (since $7 \mid 21$).

Solution to Exercise 3.2.2. Nothing forces us to use congruence here (after all, this exercise only talks about divisibility), and it is indeed quite easy to avoid it; but I believe the solution is easier to find using congruences.

Fix $a \in \mathbb{N}$. We must prove that for each $b \in \mathbb{N}$, the following statement holds:

$$\text{if } a \mid b, \text{ then } f_a \mid f_b. \tag{42}$$

We shall prove (42) by strong induction on b:

Induction step: Let $k \in \mathbb{N}$. Assume (as the induction hypothesis) that (42) holds for all $b < k$. We must prove that (42) holds for $b = k$. In other words, we must prove that

$$\text{if } a \mid k, \text{ then } f_a \mid f_k. \tag{43}$$

So let us assume that $a \mid k$. We must then prove that $f_a \mid f_k$.

If $k = 0$, then this is clearly true (since we have $f_k = f_0 = 0$ in this case, and since $f_a \mid 0$ is true[42]). Thus, for the rest of this proof, we WLOG assume that $k \neq 0$. Hence, $k \geq 1$ (since $k \in \mathbb{N}$).

[42] because any integer divides 0

It is fairly easy to see (from $a \mid k$ and $k \geq 1$ and $a \in \mathbb{N}$) that $a \geq 1$ and $k - a \in \mathbb{N}$ and $k - a < k$ and $a \mid k - a$ [43].

Our induction hypothesis says that (42) holds for all $b < k$. In other words, for each $b \in \mathbb{N}$ satisfying $b < k$, the statement (42) holds. We can apply this to $b = k - a$ (since $k - a \in \mathbb{N}$ and $k - a < k$), and thus conclude that the statement (42) holds for $b = k - a$. In other words, if $a \mid k - a$, then $f_a \mid f_{k-a}$. Thus, $f_a \mid f_{k-a}$ (since we know that $a \mid k - a$). In other words, $f_{k-a} \equiv 0 \bmod f_a$. Note also that $f_a \equiv 0 \bmod f_a$ (since every integer n satisfies $n \equiv 0 \bmod n$).

But $a - 1 \in \mathbb{N}$ (since $a \geq 1$). Hence, Exercise 2.2.3 (applied to $n = a - 1$ and $m = k - a$) yields

$$f_{(a-1)+(k-a)+1} = f_{a-1} f_{k-a} + f_{(a-1)+1} f_{(k-a)+1}.$$

In view of $(a - 1) + (k - a) + 1 = k$ and $(a - 1) + 1 = a$, this rewrites as

$$f_k = f_{a-1} f_{k-a} + f_a f_{(k-a)+1}.$$

Hence,

$$\begin{aligned}
f_k &= f_{a-1} \underbrace{f_{k-a}}_{\equiv 0 \bmod f_a} + f_a f_{(k-a)+1} \\
&\equiv f_{a-1} \cdot 0 + \underbrace{f_a}_{\equiv 0 \bmod f_a} f_{(k-a)+1} \\
&\qquad \left(\begin{array}{c} \text{here, we have used the substitution principle} \\ \text{(from Example 3.2.13) to replace } f_{k-a} \text{ by } 0 \end{array} \right) \\
&\equiv f_{a-1} \cdot 0 + 0 \cdot f_{(k-a)+1} \\
&\qquad \left(\begin{array}{c} \text{here, we have used the substitution principle} \\ \text{(from Example 3.2.13) to replace } f_a \text{ by } 0 \end{array} \right) \\
&= 0 \bmod f_a, \qquad (44)
\end{aligned}$$

and thus $f_a \mid f_k$. This is precisely what we wanted to show. Thus, we have proved that (42) holds for $b = k$. This completes the induction step. Thus, Exercise 3.2.2 is solved.

Two remarks are in order:

[43] *Proof.* We have $a \mid k$ (by assumption); in other words, there exists an integer c such that $k = ac$. Consider this c. If we had $a = 0$, then we would have $k = \underbrace{a}_{=0} c = 0$, which would contradict $k \neq 0$. Hence, we have $a \neq 0$. Thus, $a \geq 1$ (since $a \in \mathbb{N}$). Also, if we had $c \leq 0$, then we would have $k = a \underbrace{c}_{\leq 0} \leq 0$ (since $a \geq 0$), which would contradict $k \geq 1 > 0$. Thus, we cannot have $c \leq 0$; hence, we have $c > 0$. (This is all pretty obvious if you think about it; I'm just arguing that the quotient of two positive integers must be positive.)

From $c > 0$, we obtain $c \geq 1$ (since c is an integer), thus $a \underbrace{c}_{\geq 1} \geq a$ (since $a \geq 0$). Hence, $k = ac \geq a$, so that $k - a \in \mathbb{N}$. Moreover, $a \mid k - a$ (since $\underbrace{k}_{=ac} - a = ac - a = a(c - 1)$ is obviously a multiple of a). Finally, $k - a < k$ (since $a \geq 1 > 0$).

- We could have avoided using the substitution principle in the computation that led to (44). In order to do so, we should have argued as follows: Multiplying the congruence $f_{a-1} \equiv f_{a-1} \bmod f_a$ (which is obvious) with the congruence $f_{k-a} \equiv 0 \bmod f_a$, we get $f_{a-1} f_{k-a} \equiv f_{a-1} \cdot 0 \bmod f_a$. Adding the obvious congruence $f_a f_{(k-a)+1} \equiv f_a f_{(k-a)+1} \bmod f_a$ to this congruence, we obtain $f_{a-1} f_{k-a} + f_a f_{(k-a)+1} \equiv f_{a-1} \cdot 0 + f_a f_{(k-a)+1} \bmod f_a$. This justifies the first "$\equiv$" sign in (44). A similar argument justifies the second "$\equiv$" sign.

  In the future, we will, however, be even terser and not only use the substitution principle, but also do both substitutions at once:

$$f_{a-1} \underbrace{f_{k-a}}_{\equiv 0 \bmod f_a} + \underbrace{f_a}_{\equiv 0 \bmod f_a} f_{(k-a)+1} \equiv f_{a-1} \cdot 0 + 0 \cdot f_{(k-a)+1} \bmod f_a.$$

- Instead of using a strong induction on b, we could have used a (standard) induction on b/a, after first ruling out the case $a = 0$. This is essentially how Exercise 3.2.2 (actually a slight generalization thereof) is solved in [Grinbe15, proof of Theorem 2.26 **(b)**].

□

3.3. Congruences vs. remainders

Let us now come back to division with remainder. As we mentioned above, the entries of the pair (q, r) in Theorem 3.1.8 have names; let us also give them notations[44]:

Definition 3.3.1. Let n be a positive integer. Let $u \in \mathbb{Z}$. Theorem 3.1.8 shows that there exists a unique pair $(q, r) \in \mathbb{Z} \times \{0, 1, \ldots, n-1\}$ such that $u = qn + r$. Consider this pair.

(a) We denote the integer q by $u//n$, and refer to it as the *quotient of the division of u by n* (or as the *quotient obtained when u is divided by n*).

(b) We denote the integer r by $u\%n$, and refer to it as the *remainder of the division of u by n* (or as the *remainder obtained when u is divided by n*).

Here are some basic properties of these integers:

Proposition 3.3.2. Let n be a positive integer. Let $u \in \mathbb{Z}$.

(a) Then, $u\%n \in \{0, 1, \ldots, n-1\}$ and $u\%n \equiv u \bmod n$.

(b) We have $n \mid u$ if and only if $u\%n = 0$.

(c) If $c \in \{0, 1, \ldots, n-1\}$ is such that $c \equiv u \bmod n$, then $c = u\%n$.

(d) We have $u = (u//n)\, n + (u\%n)$.

[44] Warning: The notations $u//n$ and $u\%n$ are not standard across the literature; I have taken them from the Python programming language. Various authors write $u \bmod n$ for what I call $u\%n$, but the notation $u \bmod n$ typically means a different things (which we will meet in a later chapter), so I prefer not to overload it thus.

Proposition 3.3.3. Let n be a positive integer. Let $u, v \in \mathbb{Z}$.
(a) We have $u\%n + v\%n - (u+v)\%n \in \{0, n\}$.
(b) We have $(u+v)//n - u//n - v//n \in \{0, 1\}$.

Proposition 3.3.4. Let n be a positive integer. Let u and v be integers. Then, $u \equiv v \bmod n$ if and only if $u\%n = v\%n$.

Proposition 3.3.5. Let n be a positive integer. Let $u \in \mathbb{Z}$. Then, $u//n = \left\lfloor \frac{u}{n} \right\rfloor$. (See Section 1.1 for the definition of the floor $\left\lfloor \frac{u}{n} \right\rfloor$ of $\frac{u}{n}$.)

(The first three of these four propositions are proved in [19s, §2.6]; all four are easy to prove.)

Note that Proposition 3.3.4 can be viewed as another definition of congruence modulo n (at least when n is a positive integer).

Recall that an integer is said to be *even* if it is divisible by 2, and *odd* if it is not. The following exercise illustrates how remainders interact with congruence:

Exercise 3.3.1. Prove that the sum of any two odd integers is even.

Solution to Exercise 3.3.1. Let a and b be two odd integers. We must prove that $a+b$ is even.

The integer a is odd, i.e., not divisible by 2 (by the definition of "odd"). In other words, we don't have $2 \mid a$.

Proposition 3.3.2 **(b)** (applied to $n = 2$ and $u = a$) yields that we have $2 \mid a$ if and only if $a\%2 = 0$. Thus, we don't have $a\%2 = 0$ (since we don't have $2 \mid a$). But Proposition 3.3.2 **(a)** (applied to $n = 2$ and $u = a$) yields that $a\%2 \in \{0, 1\}$ and $a\%2 \equiv a \bmod 2$. From $a\%2 \in \{0, 1\}$, we obtain $a\%2 = 1$ (since we don't have $a\%2 = 0$). But from $a\%2 \equiv a \bmod 2$, we obtain $a \equiv a\%2 \bmod 2$ (by Proposition 3.2.6 **(c)**). This rewrites as $a \equiv 1 \bmod 2$ (since $a\%2 = 1$). Similarly, $b \equiv 1 \bmod 2$. Adding these two congruences[45], we obtain $a + b \equiv 1 + 1 \bmod 2$. But $1 + 1 = 2 \equiv 0 \bmod 2$ (since $2 \mid 2 - 0$). Hence, $a + b \equiv 1 + 1 \equiv 0 \bmod 2$. In other words, $2 \mid (a+b) - 0 = a + b$. In other words, $a + b$ is divisible by 2. In other words, $a + b$ is even. This solves Exercise 3.3.1. □

The following simple exercises (see [19s, §2.7] for solutions) collect various basic properties of even and odd integers:

Exercise 3.3.2. Let u be an integer.
(a) Prove that u is even if and only if $u\%2 = 0$.
(b) Prove that u is odd if and only if $u\%2 = 1$.
(c) Prove that u is even if and only if $u \equiv 0 \bmod 2$.
(d) Prove that u is odd if and only if $u \equiv 1 \bmod 2$.

[45] i.e., applying (36)

(e) Prove that u is odd if and only if $u+1$ is even.
(f) Prove that exactly one of the two numbers u and $u+1$ is even.
(g) Prove that $u(u+1) \equiv 0 \bmod 2$.
(h) Prove that $u^2 \equiv -u \equiv u \bmod 2$.
(i) Let v be a further integer. Prove that $u \equiv v \bmod 2$ holds if and only if u and v are either both odd or both even.

Exercise 3.3.3. **(a)** Prove that each even integer u satisfies $u^2 \equiv 0 \bmod 4$.
(b) Prove that each odd integer u satisfies $u^2 \equiv 1 \bmod 4$.
(c) Prove that no two integers x and y satisfy $x^2 + y^2 \equiv 3 \bmod 4$.
(d) Prove that if x and y are two integers satisfying $x^2 + y^2 \equiv 2 \bmod 4$, then x and y are both odd.

As an application of these simple facts, let us prove something a little bit less trivial:

Exercise 3.3.4. Let n be an odd integer. Prove that $8 \mid n^2 - 1$.

Solution to Exercise 3.3.4. Exercise 3.3.2 **(d)** (applied to $u = n$) shows that n is odd if and only if $n \equiv 1 \bmod 2$. Hence, $n \equiv 1 \bmod 2$ (since n is odd). In other words, $2 \mid n - 1$. In other words, there exists an integer c such that $n - 1 = 2c$. Consider this c. From $n - 1 = 2c$, we obtain $n = 2c + 1$, hence $n^2 = (2c+1)^2 = 4c^2 + 4c + 1$ and therefore $n^2 - 1 = 4c^2 + 4c = 4c(c+1)$. But Exercise 3.3.2 **(g)** (applied to $u = c$) yields $c(c+1) \equiv 0 \bmod 2$; in other words, $2 \mid c(c+1)$. Hence, there exists an integer d such that $c(c+1) = 2d$. Consider this d. Now,

$$n^2 - 1 = 4\underbrace{c(c+1)}_{=2d} = 4 \cdot 2d = 8d.$$

Thus, $8 \mid n^2 - 1$ (since d is an integer). This solves Exercise 3.3.4. $\square$

Here is another, very similar exercise to illustrate how remainders can be used to prove congruences almost mechanically:

Exercise 3.3.5. Let n be an integer such that $3 \nmid n$. Prove that $3 \mid n^2 - 1$.

Solution to Exercise 3.3.5. Proposition 3.3.2 **(a)** (applied to 3 and n instead of n and u) yields that $n\%3 \in \{0, 1, 2\}$ and $n\%3 \equiv n \bmod 3$. Symmetry of congruence yields $n \equiv n\%3 \bmod 3$ (since $n\%3 \equiv n \bmod 3$).

Proposition 3.3.2 **(b)** (applied to 3 and n instead of n and u) yields that we have $3 \mid n$ if and only if $n\%3 = 0$. Since we don't have $3 \mid n$ (because we assumed $3 \nmid n$), we thus conclude that we don't have $n\%3 = 0$.

Thus we know that $n\%3 \in \{0, 1, 2\}$, but we don't have $n\%3 = 0$. Hence, $n\%3 \in \{0, 1, 2\} \setminus \{0\} = \{1, 2\}$. Hence, we are in one of the following two cases:

Case 1: We have $n\%3 = 1$.

Case 2: We have $n\%3 = 2$.

Let us first consider Case 1. In this case, we have $n\%3 = 1$. Recall that $n \equiv n\%3 = 1 \bmod 3$. We can square both sides of this congruence (by applying Proposition 3.2.7 to 3, n, 1 and 2 instead of n, a, b and k), and thus obtain $n^2 \equiv 1^2 = 1 \bmod 3$. In other words, $3 \mid n^2 - 1$. Hence, Exercise 3.3.5 is solved in Case 1.

Let us now consider Case 2. In this case, we have $n\%3 = 2$. Recall that $n \equiv n\%3 = 2 \bmod 3$. We can square both sides of this congruence (by applying Proposition 3.2.7 to 3, n, 2 and 2 instead of n, a, b and k), and thus obtain $n^2 \equiv 2^2 = 4 \equiv 1 \bmod 3$ (where the last "$\equiv$" sign is a consequence of $3 \mid 4 - 1$). In other words, $3 \mid n^2 - 1$. Hence, Exercise 3.3.5 is solved in Case 2.

We have now solved Exercise 3.3.5 in both possible cases, so we are done. □

The solution to Exercise 3.3.5 we just gave is an example of the "try all possible remainders" technique for proving divisibilities and congruences. It should be clear that we could also have used it to get a more-or-less mechanical solution to Exercise 3.3.4: Since $n\%8 \in \{0, 1, \ldots, 7\}$, we would just have to check that $8 \mid n^2 - 1$ holds for all possible values $0, 1, \ldots, 7$ of $n\%8$. (Out of these 8 values, only $1, 3, 5, 7$ are possible, because n is assumed to be odd in Exercise 3.3.4.) In a similar fashion, we can prove that every integer n satisfies the divisibilities

$$6 \mid n^3 - n; \qquad 12 \mid n^4 - n^2; \qquad 10 \mid n^5 - n; \qquad 24 \mid n^5 - n^3;$$
$$6 \mid n(n+1)(n+2); \qquad 24 \mid n(n+1)(n+2)(n+3)$$

and many others.

As another example for the use of congruence arguments, let us find out when Fibonacci numbers are even and when they are odd. A look at the first values suggests that every third Fibonacci number (starting with f_0) is even, while the remaining ones are odd. Equipped with the notion of a congruence (and parts **(c)** and **(d)** of Exercise 3.3.2), we can restate this as follows:

Exercise 3.3.6. Let $(f_0, f_1, f_2, \ldots)$ be the Fibonacci sequence. Then,

$$f_n \equiv \begin{cases} 0, & \text{if } 3 \mid n; \\ 1, & \text{if } 3 \nmid n \end{cases} \bmod 2 \tag{45}$$

for every nonnegative integer n.

Solution to Exercise 3.3.6 (sketched). We shall prove (45) by strong induction on n:

Induction step: Let $m \in \mathbb{N}$. Assume (as the induction hypothesis) that (45) holds for all $n < m$. We must prove that (45) for $n = m$. In other words, we must prove that

$$f_m \equiv \begin{cases} 0, & \text{if } 3 \mid m; \\ 1, & \text{if } 3 \nmid m \end{cases} \bmod 2. \tag{46}$$

If $m < 2$, then we can see this directly from $f_0 = 0$ and $f_1 = 1$. Thus, we WLOG assume that $m \geq 2$. Hence, $m - 2$ and $m - 1$ are nonnegative integers. Since these

two nonnegative integers $m-2$ and $m-1$ are $< m$, we can thus apply (45) to $n = m-2$ and to $n = m-1$ (by our induction hypothesis). We thus obtain

$$f_{m-1} \equiv \begin{cases} 0, & \text{if } 3 \mid m-1; \\ 1, & \text{if } 3 \nmid m-1 \end{cases} \mod 2$$

and

$$f_{m-2} \equiv \begin{cases} 0, & \text{if } 3 \mid m-2; \\ 1, & \text{if } 3 \nmid m-2 \end{cases} \mod 2.$$

Adding these two congruences, we obtain

$$f_{m-1} + f_{m-2} \equiv \begin{cases} 0, & \text{if } 3 \mid m-1; \\ 1, & \text{if } 3 \nmid m-1 \end{cases} + \begin{cases} 0, & \text{if } 3 \mid m-2; \\ 1, & \text{if } 3 \nmid m-2 \end{cases} \mod 2.$$

This rewrites as

$$f_m \equiv \begin{cases} 0, & \text{if } 3 \mid m-1; \\ 1, & \text{if } 3 \nmid m-1 \end{cases} + \begin{cases} 0, & \text{if } 3 \mid m-2; \\ 1, & \text{if } 3 \nmid m-2 \end{cases} \mod 2 \tag{47}$$

(since the recursive definition of the Fibonacci sequence yields $f_m = f_{m-1} + f_{m-2}$). Our goal is now to deduce (46) from this congruence. In order to do so, it suffices to show that

$$\begin{aligned} &\begin{cases} 0, & \text{if } 3 \mid m-1; \\ 1, & \text{if } 3 \nmid m-1 \end{cases} + \begin{cases} 0, & \text{if } 3 \mid m-2; \\ 1, & \text{if } 3 \nmid m-2 \end{cases} \\ &\equiv \begin{cases} 0, & \text{if } 3 \mid m; \\ 1, & \text{if } 3 \nmid m \end{cases} \mod 2 \end{aligned} \tag{48}$$

(because then, combining (47) with (48) will immediately yield (46) by the transitivity of congruence).

The proof of (48) is a straightforward case distinction. Indeed, Proposition 3.3.2 **(a)** (applied to $n = 3$ and $u = m$) yields that $m\%3 \in \{0, 1, 2\}$ and $m\%3 \equiv m \mod 3$. Symmetry of congruence yields $m \equiv m\%3 \mod 3$ (since $m\%3 \equiv m \mod 3$). Since $m\%3 \in \{0, 1, 2\}$, we are in one of the following three cases:

Case 1: We have $m\%3 = 0$.

Case 2: We have $m\%3 = 1$.

Case 3: We have $m\%3 = 2$.

Let me work through Case 1 in detail, leaving the other two cases to the reader (the arguments are closely similar). In Case 1, we have $m\%3 = 0$. Thus, $m \equiv m\%3 = 0 \mod 3$. According to Proposition 3.2.3 (applied to $n = 3$ and $a = m$), we have $m \equiv 0 \mod 3$ if and only if $3 \mid m$. Thus, $3 \mid m$ (since $m \equiv 0 \mod 3$). Furthermore, subtracting the congruence $1 \equiv 1 \mod 3$ from the congruence $m \equiv 0 \mod 3$, we obtain $m - 1 \equiv 0 - 1 = -1 \mod 3$. Thus, we do not have $m - 1 \equiv 0 \mod 3$ (because

if we had $m-1 \equiv 0 \bmod 3$, then we would have $0 \equiv m-1 \equiv -1 \bmod 3$, or, equivalently, $3 \mid 0-(-1) = 1$, which is absurd). According to Proposition 3.2.3 (applied to $n=3$ and $a=m-1$), we have $m-1 \equiv 0 \bmod 3$ if and only if $3 \mid m-1$. Hence, we do not have $3 \mid m-1$ (since we do not have $m-1 \equiv 0 \bmod 3$). In other words, we have $3 \nmid m-1$. A similar computation yields $3 \nmid m-2$ (since $3 \mid 2$ is as absurd as $3 \mid 1$). Now, we know that $3 \mid m$ and $3 \nmid m-1$ and $3 \nmid m-2$. In light of these facts, the congruence (48) (which we need to prove) rewrites as $1+1 \equiv 0 \bmod 2$. In other words, it is equivalent to $2 \mid (1+1)-0$, which is obvious. Thus, we have proved (48) in Case 1.

The proof of (48) in Case 2 is similar, except that this time we have $m \equiv 1 \bmod 3$ (instead of $m \equiv 0 \bmod 3$) and thus $3 \mid m-1$ and $3 \nmid m$ and $3 \nmid m-2$.

The proof of (48) in Case 3 is similar, too. □

3.4. Greatest common divisors

3.4.1. The definitions

One of the main workhorses of number theory is the notion of a common divisor, defined exactly as one would expect:

Definition 3.4.1. Let $b_1, b_2, \ldots, b_k$ be integers. Then, the *common divisors* of $b_1, b_2, \ldots, b_k$ are defined to be the integers a that satisfy

$$(a \mid b_i \text{ for all } i \in \{1, 2, \ldots, k\}) \tag{49}$$

(in other words, that divide all of the integers $b_1, b_2, \ldots, b_k$).

For example, the common divisors of 6 and 8 are $-2, -1, 1, 2$. We refer to [19s, §2.9] for a thorough treatment of common divisors; here we shall only sketch the main steps (particularly to the extent they illustrate induction). The notion of a *greatest common divisor* is crucial:[46]

Definition 3.4.2. Let $b_1, b_2, \ldots, b_k$ be finitely many integers. The *greatest common divisor* (or, for short, the *gcd*) of $b_1, b_2, \ldots, b_k$ is the nonnegative integer $\gcd(b_1, b_2, \ldots, b_k)$ defined as follows:

- If $b_1, b_2, \ldots, b_k$ are not all 0, then it is defined as the name suggests: It is the largest of all common divisors of $b_1, b_2, \ldots, b_k$.
- If $b_1, b_2, \ldots, b_k$ are all 0, then it is defined to be 0.

[46] It is easy to see that any nonzero integer b has only finitely many divisors; indeed, each divisor of b is an integer between $-|b|$ and $|b|$. Hence, if $b_1, b_2, \ldots, b_k$ are finitely many integers that are not all 0, then there are only finitely many common divisors of $b_1, b_2, \ldots, b_k$. Moreover, there is at least one common divisor of $b_1, b_2, \ldots, b_k$ (since 1 is always such a divisor). Hence, the set of all common divisors of $b_1, b_2, \ldots, b_k$ is nonempty and finite (when $b_1, b_2, \ldots, b_k$ are not all zero); therefore, this set has a largest element. This is used implicitly in Definition 3.4.2.

(Note that the second case should make some eyes roll; if $b_1, b_2, \ldots, b_k$ are all 0, then **every** integer is a common divisor of $b_1, b_2, \ldots, b_k$; it is thus strange to designate 0 the greatest common divisor. But this makes the more sense the more you learn about greatest common divisors. For now, treat it as an annoying special case.)

For example, $\gcd(4,6) = 2$ and $\gcd(3,5) = 1$ and $\gcd(6,10,15) = 1$. Definition 3.4.2 easily entails the following (see [19s, Definition 2.9.6] for the details):

Proposition 3.4.3. Let $b_1, b_2, \ldots, b_k$ be finitely many integers.
(a) The number $\gcd(b_1, b_2, \ldots, b_k)$ is a nonnegative integer.
(b) If $b_1, b_2, \ldots, b_k$ are not all 0, then $\gcd(b_1, b_2, \ldots, b_k)$ is a positive integer.

3.4.2. Basic properties

The following properties of greatest common divisors ([19s, Proposition 2.9.7]) are easy to check using the definition:

Proposition 3.4.4. **(a)** We have $\gcd(a,0) = \gcd(a) = |a|$ for all $a \in \mathbb{Z}$.
(b) We have $\gcd(a,b) = \gcd(b,a)$ for all $a, b \in \mathbb{Z}$.
(c) We have $\gcd(a, ua+b) = \gcd(a,b)$ for all $a, b, u \in \mathbb{Z}$.
(d) If $a, b, c \in \mathbb{Z}$ satisfy $b \equiv c \bmod a$, then $\gcd(a,b) = \gcd(a,c)$.
(e) If $a, b \in \mathbb{Z}$ are such that a is positive, then $\gcd(a,b) = \gcd(a, b\%a)$.
(f) We have $\gcd(a,b) \mid a$ and $\gcd(a,b) \mid b$ for all $a, b \in \mathbb{Z}$.
(g) We have $\gcd(-a,b) = \gcd(a,b)$ for all $a, b \in \mathbb{Z}$.
(h) We have $\gcd(a,-b) = \gcd(a,b)$ for all $a, b \in \mathbb{Z}$.
(i) If $a, b \in \mathbb{Z}$ satisfy $a \mid b$, then $\gcd(a,b) = |a|$.
(j) The greatest common divisor of the empty list of integers is $\gcd() = 0$.

Note that we are focusing on gcds of two or fewer numbers for now; we will eventually come back to the general case.

3.4.3. Bezout's theorem

The most important fact about greatest common divisors is the following fact ([19s, Theorem 2.9.12]), known as *Bezout's theorem* (or *Bezout's identity*, despite being an existence statement rather than a literal identity):

Theorem 3.4.5 (Bezout's theorem)**.** Let a and b be two integers. Then, there exist integers $x \in \mathbb{Z}$ and $y \in \mathbb{Z}$ such that

$$\gcd(a,b) = xa + yb.$$

I shall prove this theorem here not just because it is crucial for the development of number theory, but also because its proof is an instructive example of strong induction on a derived quantity (namely, $a+b$). First, an example:

Example 3.4.6. Set $a = 6$ and $b = 10$. Then, $\gcd(a, b) = \gcd(6, 10) = 2$. Theorem 3.4.5 says that there exist integers $x \in \mathbb{Z}$ and $y \in \mathbb{Z}$ such that $\gcd(a, b) = xa + yb$, that is, $2 = x \cdot 6 + y \cdot 10$. And indeed, it is not hard to find such x and y: For example, we can take $x = 2$ and $y = -1$. (Alternatively, we can take $x = 7$ and $y = -4$. There are infinitely many valid choices.)

Proof of Theorem 3.4.5 (sketched). Forget that we fixed a and b.

If a and b are two integers, then $\mathbb{Z}a + \mathbb{Z}b$ shall denote the set

$$\{xa + yb \mid x \in \mathbb{Z} \text{ and } y \in \mathbb{Z}\}.$$

This is a subset of $\mathbb{Z}$. Thus, Theorem 3.4.5 is saying that for any two integers a and b, we have

$$\gcd(a, b) \in \mathbb{Z}a + \mathbb{Z}b. \tag{50}$$

We aim to prove this by strong induction on $a + b$, but first we need to ensure that $a + b$ belongs to $\mathbb{N}$ (or at least to $\mathbb{Z}_{\geq g}$ for some $g \in \mathbb{Z}$); this does not come for free, since $a + b$ can be arbitrarily small when a and b range over $\mathbb{Z}$. We will ensure $a + b \in \mathbb{N}$ by restricting ourselves to the case $a, b \in \mathbb{N}$ and then deducing the general case from this case.

So, at first, let us prove (50) for $a, b \in \mathbb{N}$ only. We shall prove this by strong induction on $a + b$; that is, we shall prove the following claim:

> *Claim 1:* Let $n \in \mathbb{N}$. Then, any $a, b \in \mathbb{N}$ satisfying $a + b = n$ satisfy $\gcd(a, b) \in \mathbb{Z}a + \mathbb{Z}b$.

[*Proof of Claim 1:* Apply strong induction on n:

Induction step: Let $k \in \mathbb{N}$. Assume (as the induction hypothesis) that Claim 1 is true for all $n < k$. We must prove that Claim 1 holds for $n = k$.

So let $a, b \in \mathbb{N}$ be such that $a + b = k$. We shall show that $\gcd(a, b) \in \mathbb{Z}a + \mathbb{Z}b$.

Proposition 3.4.4 **(b)** yields $\gcd(a, b) = \gcd(b, a)$. Proposition 3.4.4 **(a)** yields $\gcd(a, 0) = \gcd(a) = |a| = a$ (since $a \in \mathbb{N}$). The same argument (applied to b instead of a) yields $\gcd(b, 0) = b$.

Note that a and b play symmetric roles in Claim 1 (since $\gcd(a, b) = \gcd(b, a)$ and $\mathbb{Z}a + \mathbb{Z}b = \mathbb{Z}b + \mathbb{Z}a$), and thus can be swapped at will. By swapping a and b if necessary, we can ensure that $a \leq b$. Hence, we WLOG assume that $a \leq b$. Thus, $b - a \in \mathbb{N}$.

If $a = 0$, then

$$\begin{aligned} \gcd(a, b) &= \gcd(b, a) = \gcd(b, 0) \qquad (\text{since } a = 0) \\ &= b = 0a + 1b \in \mathbb{Z}a + \mathbb{Z}b. \end{aligned}$$

Thus, we are done if $a = 0$. Hence, we WLOG assume that $a \neq 0$. Therefore, $a > 0$ (since $a \in \mathbb{N}$). Thus, $a + b > b$, so that $b < a + b = k$.

But our induction hypothesis says that Claim 1 is true for all $n < k$. Hence, we can apply Claim 1 to $b - a$ and b instead of b and n (since $b - a \in \mathbb{N}$ and $a + (b - a) = b$ and $b < k$). We thus obtain $\gcd(a, b - a) \in \mathbb{Z}a + \mathbb{Z}(b - a)$. However,

$$\begin{aligned} \gcd(a, b - a) &= \gcd(a, (-1)a + b) \qquad (\text{since } b - a = (-1)a + b) \\ &= \gcd(a, b) \qquad (\text{by Proposition 3.4.4 } \textbf{(c)}\text{, applied to } u = -1), \end{aligned}$$

so that $\gcd(a, b) = \gcd(a, b - a) \in \mathbb{Z}a + \mathbb{Z}(b - a)$.

But it is easy to see that $\mathbb{Z}a + \mathbb{Z}(b - a) \subseteq \mathbb{Z}a + \mathbb{Z}b$ [47]. (Actually, it is easy to see that $\mathbb{Z}a + \mathbb{Z}(b - a) = \mathbb{Z}a + \mathbb{Z}b$, but we will not need this.)

Hence, $\gcd(a, b) \in \mathbb{Z}a + \mathbb{Z}(b - a) \subseteq \mathbb{Z}a + \mathbb{Z}b$.

Now, forget that we fixed a, b. We thus have shown that any $a, b \in \mathbb{N}$ satisfying $a + b = k$ satisfy $\gcd(a, b) \in \mathbb{Z}a + \mathbb{Z}b$. In other words, Claim 1 holds for $n = k$. This completes the induction step. Thus, Claim 1 is proved.]

We still need to prove (50) for arbitrary integers a and b. Claim 1 handles the case when $a, b \in \mathbb{N}$ (that is, when a, b are nonnegative). The general case can be reduced to this case as follows:

- If a is negative, then we replace a by the positive integer $-a \in \mathbb{N}$. This does not change our claim because $\gcd(-a, b) = \gcd(a, b)$ and $\mathbb{Z}(-a) + \mathbb{Z}b = \mathbb{Z}a + \mathbb{Z}b$ (check this!).
- If b is negative, then we replace b by the positive integer $-b \in \mathbb{N}$. This does not change our claim because $\gcd(a, -b) = \gcd(a, b)$ and $\mathbb{Z}a + \mathbb{Z}(-b) = \mathbb{Z}a + \mathbb{Z}b$ (check this!).

Thus, in proving (50) for arbitrary integers a and b, we can WLOG assume that a and b belong to $\mathbb{N}$. But if they do, then Claim 1 (applied to $n = a + b$) yields that $\gcd(a, b) \in \mathbb{Z}a + \mathbb{Z}b$, and we are done. Theorem 3.4.5 is proved. $\square$

Our above proof of Theorem 3.4.5 essentially encodes (the most basic form of) the *extended Euclidean algorithm*, which computes $\gcd(a, b)$ (for $a, b \in \mathbb{N}$) and represents $\gcd(a, b)$ in the form $xa + yb$ (with $x, y \in \mathbb{Z}$) by repeatedly subtracting one of the numbers a and b from the other until one of the numbers becomes 0. While we have not explicitly shown any algorithm in our proof, it can be recovered by unraveling our (strong) induction; generally, inductive proofs encode recursive algorithms. We shall not usually dwell on the algorithmic content of our proofs.

[47] *Proof.* Let $z \in \mathbb{Z}a + \mathbb{Z}(b - a)$. Then, there exist $x \in \mathbb{Z}$ and $y \in \mathbb{Z}$ such that $z = xa + y(b - a)$ (by the definition of $\mathbb{Z}a + \mathbb{Z}(b - a)$). Consider these x and y. Then,

$$z = xa + y(b - a) = xa + yb - ya = (x - y)a + yb.$$

Hence, there exist $x' \in \mathbb{Z}$ and $y' \in \mathbb{Z}$ such that $z = x'a + y'b$ (namely, $x' = x - y$ and $y' = y$). In other words, $z \in \mathbb{Z}a + \mathbb{Z}b$.

Forget that we fixed z. We thus have shown that $z \in \mathbb{Z}a + \mathbb{Z}b$ for each $z \in \mathbb{Z}a + \mathbb{Z}(b - a)$. In other words, $\mathbb{Z}a + \mathbb{Z}(b - a) \subseteq \mathbb{Z}a + \mathbb{Z}b$.

3.4.4. The universal property

Theorem 3.4.5 is the workhorse of classical Euclidean number theory (i.e., the number theory done in Euclid's *Elements*); it quickly proves many important results. For example, watch how the following fact ([19s, Theorem 2.9.15 **(a)**]) – sometimes known as the *universal property of the greatest common divisor* – easily follows from it:

Theorem 3.4.7. Let $a, b \in \mathbb{Z}$ and $m \in \mathbb{Z}$. Then, we have the following logical equivalence:

$$(m \mid a \text{ and } m \mid b) \iff (m \mid \gcd(a,b)). \tag{51}$$

Proof of Theorem 3.4.7. In order to prove (51), we need to prove the "$\Longrightarrow$" and "$\Longleftarrow$" directions of the equivalence (51).

Proof of the "$\Longrightarrow$" direction:[48] We must show that the statement $(m \mid a \text{ and } m \mid b)$ implies the statement $(m \mid \gcd(a,b))$. So let us assume that $m \mid a$ and $m \mid b$. Theorem 3.4.5 shows that there exist integers $x \in \mathbb{Z}$ and $y \in \mathbb{Z}$ such that

$$\gcd(a,b) = xa + yb. \tag{52}$$

Consider these x and y. Now, (using the transitivity of divisibility) we have $m \mid a \mid xa$, so that $xa \equiv 0 \bmod m$. Also, $m \mid b \mid yb$, thus $yb \equiv 0 \bmod m$. Adding the congruences $xa \equiv 0 \bmod m$ and $yb \equiv 0 \bmod m$ together, we find $xa + yb \equiv 0 + 0 = 0 \bmod m$; in other words, $m \mid xa + yb$. In view of (52), this rewrites as $m \mid \gcd(a,b)$. This proves the "$\Longrightarrow$" direction of the equivalence (51).

Proof of the "$\Longleftarrow$" direction:[49] We must show that the statement $(m \mid \gcd(a,b))$ implies the statement $(m \mid a \text{ and } m \mid b)$. So let us assume that $m \mid \gcd(a,b)$. Then,

$$m \mid \gcd(a,b) \mid a \qquad \text{(by Proposition 3.4.4 \textbf{(f)})}$$

and

$$m \mid \gcd(a,b) \mid b \qquad \text{(by Proposition 3.4.4 \textbf{(f)})}.$$

Thus, we have $m \mid a$ and $m \mid b$. This proves the "$\Longleftarrow$" direction of the equivalence (51).

Now, both directions of the equivalence (51) are proved, so the equivalence holds. This proves Theorem 3.4.7. □

Theorem 3.4.7 can be restated as "an integer m divides two integers a and b if and only if it divides their gcd". This gives a useful way to show that something divides a gcd. A sample application of Theorem 3.4.7 (as well as a useful fact in itself) is the following theorem ([19s, Theorem 2.9.20]):

[48] Commonly, mathematicians just write "$\Longrightarrow$:" instead of "*Proof of the "$\Longrightarrow$" direction:*".
[49] Commonly, mathematicians just write "$\Longleftarrow$:" instead of "*Proof of the "$\Longleftarrow$" direction:*".

Theorem 3.4.8. Let $s, a, b \in \mathbb{Z}$. Then,

$$\gcd(sa, sb) = |s| \gcd(a, b).$$

Proof of Theorem 3.4.8 (sketched). What follows is perhaps not the simplest way to prove Theorem 3.4.8, but it serves as a wonderful illustration of Theorem 3.4.7, as it will use the latter theorem three times.

If we want to prove that two nonnegative integers x and y are equal, it suffices to show that they mutually divide each other (i.e., that they satisfy $x \mid y$ and $y \mid x$). In fact, once this is shown, Proposition 3.1.3 **(c)** (applied to x and y instead of a and b) will yield $|x| = |y|$, and this will rewrite as $x = y$ (since x and y are nonnegative). This might appear like a roundabout approach to proving equalities, but it turns out to be pretty useful when x and y are characterized through their divisibility properties (and this is the case for gcds, among other things).

We shall apply this approach to $x := \gcd(sa, sb)$ and $y := |s| \gcd(a, b)$. Our goal is to show that $x = y$; thus, we shall achieve it by showing that $x \mid y$ and $y \mid x$. (Indeed, Proposition 3.4.3 **(a)** shows that x and y are nonnegative.)

If $s = 0$, then both x and y are 0 (since $\gcd(0, 0) = 0$), and thus we are done. Hence, we WLOG assume that $s \neq 0$. Therefore, $|s| \neq 0$. This will come handy as we will divide by $|s|$ soon.

Let us first prove $y \mid x$. Indeed, Proposition 3.4.4 **(f)** yields $\gcd(a, b) \mid a$. We can multiply both sides of this divisibility by s (by applying Proposition 3.1.5 to $\gcd(a, b)$, a and s instead of a, b and c), and thus obtain $\gcd(a, b) \cdot s \mid as$. Note that the two integers $\gcd(a, b) \cdot s$ and y are equal up to sign, since

$$y = \underbrace{|s|}_{=\pm s} \gcd(a, b) = \pm s \gcd(a, b) = \pm \gcd(a, b) \cdot s.$$

Thus, they mutually divide each other; in particular, we have $y \mid \gcd(a, b) \cdot s \mid as = sa$. Similarly, $y \mid sb$. Thus, the integer y divides both sa and sb. Hence, Theorem 3.4.7 (applied to sa, sb and y instead of a, b and m) shows that it divides $\gcd(sa, sb)$. In other words, $y \mid \gcd(sa, sb)$. In other words, $y \mid x$ (since $x = \gcd(sa, sb)$).

Let us next prove $x \mid y$. The integer $|s|$ divides both sa and sb (since $|s| \mid s \mid sa$ and $|s| \mid s \mid sb$). Thus, Theorem 3.4.7 (applied to sa, sb and $|s|$ instead of a, b and m) shows that it divides $\gcd(sa, sb)$. In other words, it divides x (since $x = \gcd(sa, sb)$). Hence, $\dfrac{x}{|s|}$ is an integer (since $|s| \neq 0$). We shall now show that this integer $\dfrac{x}{|s|}$ divides $\gcd(a, b)$. According to Theorem 3.4.7, this will follow if we can show that it divides both a and b. So let us prove that it divides a and b. Indeed, $|s|$ equals either s or $-s$ (depending on the sign of s); thus, $\dfrac{x}{|s|} \cdot s$ equals either x or $-x$. In either case, we have $\dfrac{x}{|s|} \cdot s \mid x$. Now,

$$\frac{x}{|s|} \cdot s \mid x = \gcd(sa, sb) \mid sa = as.$$

We can "cancel" the factor s from this divisibility (by applying Proposition 3.1.5 to $\frac{x}{|s|}$, a and s instead of a, b and c), and thus obtain $\frac{x}{|s|} \mid a$. Similarly, $\frac{x}{|s|} \mid b$. Combining $\frac{x}{|s|} \mid a$ with $\frac{x}{|s|} \mid b$, we obtain $\frac{x}{|s|} \mid \gcd(a,b)$ (by Theorem 3.4.7, applied to $m = \frac{x}{|s|}$). This, in turn, is equivalent to $\frac{x}{|s|} \cdot s \mid \gcd(a,b) \cdot s$ (by Proposition 3.1.5, applied to $\frac{x}{|s|}$, $\gcd(a,b)$ and s instead of a, b and c). In view of $\gcd(a,b) \cdot s = s \gcd(a,b) = y$, this rewrites as $\frac{x}{|s|} \cdot s \mid y$. But as we recall, $\frac{x}{|s|} \cdot s$ equals either x or $-x$; in either case, we have $x \mid \frac{x}{|s|} \cdot s \mid y$. Hence, $x \mid y$ is proved.

We have now proved both $x \mid y$ and $y \mid x$. As we have seen above, this entails $x = y$ (since x and y are nonnegative). This proves Theorem 3.4.8. □

3.4.5. Using gcds

The next theorem ([19s, Theorem 2.9.17]) is crucial in a way that might not immediately meet one's eye:

Theorem 3.4.9. Let $a, b, c \in \mathbb{Z}$ satisfy $a \mid c$ and $b \mid c$. Then, $ab \mid \gcd(a,b) \cdot c$.

Before we prove this theorem, let us chat about it a bit. Here is an example first:

Example 3.4.10. Let $a = 6$ and $b = 10$ and $c = 30$. Then, $a = 6 \mid 30 = c$ and $b = 10 \mid 30 = c$. Thus, Theorem 3.4.9 yields $ab \mid \gcd(a,b) \cdot c$. In view of $a = 6$, $b = 10$, $c = 13$ and $\gcd(a,b) = \gcd(6,10) = 2$, we can rewrite this as $6 \cdot 10 \mid 2 \cdot 30$, which is not only true but actually an equality (we have $6 \cdot 10 = 2 \cdot 30$). Note that we do **not** have $ab \mid c$.

What makes Theorem 3.4.9 useful is that it lets us "swim upstream" in divisibility arguments, in the sense of deriving "stronger" divisibilities from "weaker" ones. (Here, we not-so-rigorously designate a divisibility $x \mid y$ as "weak" if the ratio y/x is large and "strong" if it is small. Thus, for example, the divisibility $4 \mid 24$ is much weaker than either of the two divisibilities $4 \mid 12$ and $12 \mid 24$, so that we are "swimming downstream" when we derive $4 \mid 24$ from $4 \mid 12 \mid 24$. We are also "swimming downstream" when we apply Proposition 3.1.4 **(c)**; indeed, $a_1 a_2 \mid b_1 b_2$ is "weaker" than $a_1 \mid b_1$ and $a_2 \mid b_2$ (or "equally strong" at best). In contrast, as we saw in Example 3.4.10, the divisibility $ab \mid \gcd(a,b) \cdot c$ that we gain from Theorem 3.4.9 can be much "stronger" than the two divisibilities $a \mid c$ and $b \mid c$ we have invested; thus, Theorem 3.4.9 is taking us "upstream". All this should not be taken literally, but it gives a useful intuition.)

Again, as the following proof shows, the active ingredient in Theorem 3.4.9 is Bezout's theorem:

First proof of Theorem 3.4.9. Theorem 3.4.5 yields that there exist integers $x \in \mathbb{Z}$ and $y \in \mathbb{Z}$ such that $\gcd(a,b) = xa + yb$. Consider these x and y.

There exists an integer u such that $c = au$ (since $a \mid c$). Consider this u.

There exists an integer v such that $c = bv$ (since $b \mid c$). Consider this v.

Now,

$$\underbrace{\gcd(a,b)}_{=xa+yb} \cdot c = (xa+yb)\,c = xa\underbrace{c}_{=bv} + yb\underbrace{c}_{=au} = xabv + ybau = ab(xv+yu).$$

Thus, there exists an integer d such that $\gcd(a,b) \cdot c = abd$ (namely, $d = xv + yu$). In other words, $ab \mid \gcd(a,b) \cdot c$. This proves Theorem 3.4.9. □

Second proof of Theorem 3.4.9. We have $a \mid a$ and $b \mid c$, thus $ab \mid ac$ (by Proposition 3.1.4 **(c)**, applied to $a_1 = a$, $a_2 = b$, $b_1 = a$ and $b_2 = c$). Also, we have $a \mid c$ and $b \mid b$, thus $ab \mid cb$ (by Proposition 3.1.4 **(c)**, applied to $a_1 = a$, $a_2 = c$, $b_1 = b$ and $b_2 = b$). However, we have the logical equivalence

$$(ab \mid ac \text{ and } ab \mid cb) \iff (ab \mid \gcd(ac, cb))$$

(by Theorem 3.4.7, applied to ac, cb and ab instead of a, b and m). Therefore, we have $ab \mid \gcd(ac, cb)$ (since we have $ab \mid ac$ and $ab \mid cb$). This rewrites as $ab \mid \gcd(ca, cb)$ (since $ac = ca$).

But Theorem 3.4.8 (applied to $s = c$) yields

$$\gcd(ca, cb) = \underbrace{|c|}_{=\pm c} \gcd(a,b) = \pm c \gcd(a,b) \mid c \gcd(a,b) = \gcd(a,b) \cdot c.$$

Hence,

$$ab \mid \gcd(ca, cb) \mid \gcd(a,b) \cdot c.$$

This proves Theorem 3.4.9 again. □

Here is yet another fact ([19s, Theorem 2.9.19]) that follows from Bezout's theorem:

Theorem 3.4.11. Let $a, b, c \in \mathbb{Z}$ satisfy $a \mid bc$. Then, $a \mid \gcd(a,b) \cdot c$.

Theorem 3.4.11, too, is a tool for "swimming upstream" (as the resulting divisibility $a \mid \gcd(a,b) \cdot c$ is usually "stronger" than $a \mid bc$).

First proof of Theorem 3.4.11. Theorem 3.4.5 yields that there exist integers $x \in \mathbb{Z}$ and $y \in \mathbb{Z}$ such that $\gcd(a,b) = xa + yb$. Consider these x and y.

We have $axc \equiv 0 \bmod a$ (since $a \mid axc$) and $ybc \equiv 0 \bmod a$ (since $a \mid bc \mid ybc$). Adding these two congruences together, we obtain $axc + ybc \equiv 0 + 0 = 0 \bmod a$. In view of $axc + ybc = \underbrace{(xa+yb)}_{=\gcd(a,b)} c = \gcd(a,b) \cdot c$, this rewrites as $\gcd(a,b) \cdot c \equiv 0 \bmod a$. In other words, $a \mid \gcd(a,b) \cdot c$. This proves Theorem 3.4.11. □

Second proof of Theorem 3.4.11. We have the logical equivalence

$$(a \mid ac \text{ and } a \mid bc) \iff (a \mid \gcd(ac, bc))$$

(by Theorem 3.4.7, applied to ac, bc and a instead of a, b and m). Therefore, we have $a \mid \gcd(ac, bc)$ (since we have $a \mid ac$ and $a \mid bc$). This rewrites as $a \mid \gcd(ca, cb)$ (since $ac = ca$ and $bc = cb$). But just as in the Second proof of Theorem 3.4.9, we can show that $\gcd(ca, cb) \mid \gcd(a, b) \cdot c$. Hence, $a \mid \gcd(ca, cb) \mid \gcd(a, b) \cdot c$. This proves Theorem 3.4.11 again. □

Here is yet another property of gcds:

Proposition 3.4.12. Let $a_1, a_2, b_1, b_2 \in \mathbb{Z}$ satisfy $a_1 \mid b_1$ and $a_2 \mid b_2$. Then,

$$\gcd(a_1, a_2) \mid \gcd(b_1, b_2).$$

Hint to proof of Proposition 3.4.12. Show that $\gcd(a_1, a_2)$ divides both b_1 and b_2; then argue by Theorem 3.4.7. (See [19s, Exercise 2.9.4] for details.) □

3.4.6. Gcds of multiple numbers

A few words are worth saying about gcds of more than two numbers. In many ways, they behave similarly to those of two numbers. For example, Theorem 3.4.5 can be generalized to multiple numbers:

Theorem 3.4.13. Let $b_1, b_2, \ldots, b_k$ be integers. Then, there exist integers $x_1, x_2, \ldots, x_k$ such that

$$\gcd(b_1, b_2, \ldots, b_k) = x_1 b_1 + x_2 b_2 + \cdots + x_k b_k.$$

Theorem 3.4.13 can be proved similarly to Theorem 3.4.5: If $b_1, b_2, \ldots, b_k \in \mathbb{N}$, then the claim follows by strong induction on $b_1 + b_2 + \cdots + b_k$ (where the induction step proceeds by subtracting the smallest nonzero number among $b_1, b_2, \ldots, b_k$ from the largest[50]); the general case can be reduced to this case by observing that replacing any b_i by $-b_i$ changes nothing. Another proof of Theorem 3.4.13 can be found in [19s, proof of Theorem 2.9.22].

A generalization of Theorem 3.4.7 to multiple numbers also exists:

Theorem 3.4.14. Let $k \in \mathbb{N}$, let $b_1, b_2, \ldots, b_k \in \mathbb{Z}$ and $m \in \mathbb{Z}$. Then, we have the following logical equivalence:

$$(m \mid b_i \text{ for all } i \in \{1, 2, \ldots, k\}) \iff (m \mid \gcd(b_1, b_2, \ldots, b_k)).$$

[50] If there is only one nonzero number among $b_1, b_2, \ldots, b_k$, then the claim is easily verified by hand.

Theorem 3.4.14 can easily be derived from Theorem 3.4.13 just as Theorem 3.4.7 was derived from Theorem 3.4.5. An alternative proof of Theorem 3.4.14 is found in [19s, proof of Theorem 2.9.21 **(a)**].

It is an instructive exercise to derive from Theorem 3.4.14 the following consequence:

Theorem 3.4.15. Let $b_1, b_2, \ldots, b_k$ be integers, and let $c_1, c_2, \ldots, c_\ell$ be integers. Then,

$$\gcd(b_1, b_2, \ldots, b_k, c_1, c_2, \ldots, c_\ell) = \gcd(\gcd(b_1, b_2, \ldots, b_k), \gcd(c_1, c_2, \ldots, c_\ell)).$$

See [19s, proof of Theorem 2.9.26] for a detailed proof of Theorem 3.4.15.

It is also not hard to extend Theorem 3.4.8 to multiple numbers ([19s, Exercise 2.9.7]):

Theorem 3.4.16. Let $s \in \mathbb{Z}$, and let $b_1, b_2, \ldots, b_k$ be integers. Then, $\gcd(sb_1, sb_2, \ldots, sb_k) = |s| \gcd(b_1, b_2, \ldots, b_k)$.

3.4.7. An exercise

Here is a sample exercise on greatest common divisors, before we go on to properly exploit them:

Exercise 3.4.1. Let u be an integer.

(a) Prove that $u^b - 1 \equiv u^a - 1 \bmod u^{b-a} - 1$ for any $a \in \mathbb{N}$ and $b \in \mathbb{N}$ satisfying $b \geq a$.

(b) Prove that $\gcd\left(u^a - 1, u^b - 1\right) = \left|u^{\gcd(a,b)} - 1\right|$ for all $a \in \mathbb{N}$ and $b \in \mathbb{N}$.

Solution to Exercise 3.4.1 (sketched). This is an outline; see [19s, solution to Exercise 2.9.3] for a detailed version.

(a) Let $a \in \mathbb{N}$ and $b \in \mathbb{N}$ be such that $b \geq a$. We have $b - a \in \mathbb{N}$ (since $b \geq a$). Hence, u^{b-a} is an integer. We have

$$\left(u^b - 1\right) - (u^a - 1) = u^b - u^a = \left(u^{b-a} - 1\right) u^a.$$

Thus, $u^{b-a} - 1 \mid \left(u^b - 1\right) - (u^a - 1)$ (since u^a is an integer). In other words, $u^b - 1 \equiv u^a - 1 \bmod u^{b-a} - 1$. This solves Exercise 3.4.1 **(a)**.

(b) The following argument will imitate our proof of Theorem 3.4.5 above (specifically the proof of Claim 1 in it).

We use strong induction on $a + b$:

Induction step: Let $k \in \mathbb{N}$. Assume (as the induction hypothesis) that Exercise 3.4.1 **(b)** is true for $a + b < k$. We must prove that Exercise 3.4.1 **(b)** is true for $a + b = k$.

So let $a, b \in \mathbb{N}$ be such that $a + b = k$. We must show that $\gcd\left(u^a - 1, u^b - 1\right) = \left|u^{\gcd(a,b)} - 1\right|$.

Note that a and b play symmetric roles in this claim[51], and thus can be swapped at will. By swapping a and b if necessary, we can ensure that $a \leq b$. Hence, we WLOG assume that $a \leq b$. Thus, $b - a \in \mathbb{N}$.

It is easy to see that our claim $\gcd\left(u^a - 1, u^b - 1\right) = \left|u^{\gcd(a,b)} - 1\right|$ holds if $a = 0$ [52]. Thus, we are done if $a = 0$. Hence, we WLOG assume that $a \neq 0$. Therefore, $a > 0$ (since $a \in \mathbb{N}$). Thus, $a + b > b$, so that $b < a + b = k$.

But our induction hypothesis says that Exercise 3.4.1 **(b)** is true for $a + b < k$. Hence, we can apply Exercise 3.4.1 **(b)** to $b - a$ instead of b (since $b - a \in \mathbb{N}$ and $a + (b - a) = b < k$). We thus obtain

$$\gcd\left(u^a - 1, u^{b-a} - 1\right) = \left|u^{\gcd(a,b-a)} - 1\right|. \tag{53}$$

But we have $\gcd(a, b - a) = \gcd(a, b)$ (this has already been proved during our proof of Theorem 3.4.5). Furthermore, Exercise 3.4.1 **(a)** (applied to $b - a$ instead of a) yields $u^b - 1 \equiv u^{b-a} - 1 \bmod u^{b-(b-a)} - 1$ (since $b - a \in \mathbb{N}$ and $b \geq b - a$). Since $b - (b - a) = a$, this rewrites as $u^b - 1 \equiv u^{b-a} - 1 \bmod u^a - 1$. Hence, Proposition 3.4.4 **(d)** (applied to $u^a - 1$, $u^b - 1$ and $u^{b-a} - 1$ instead of a, b and c) yields

$$\begin{aligned}
\gcd\left(u^a - 1, u^b - 1\right) &= \gcd\left(u^a - 1, u^{b-a} - 1\right) = \left|u^{\gcd(a,b-a)} - 1\right| \qquad \text{(by (53))} \\
&= \left|u^{\gcd(a,b)} - 1\right| \qquad (\text{since } \gcd(a, b - a) = \gcd(a, b)).
\end{aligned}$$

Now, forget that we fixed a, b. We thus have shown that any $a, b \in \mathbb{N}$ satisfying $a + b = k$ satisfy $\gcd\left(u^a - 1, u^b - 1\right) = \left|u^{\gcd(a,b)} - 1\right|$. In other words, Exercise 3.4.1 **(b)** is true for $a + b = k$. This completes the induction step. Thus, Exercise 3.4.1 **(b)** is solved.

[See also `https://math.stackexchange.com/questions/7473/` for various solutions of Exercise 3.4.1 **(b)**.] □

[51] because Proposition 3.4.4 **(b)** yields $\gcd(a, b) = \gcd(b, a)$ and $\gcd\left(u^a - 1, u^b - 1\right) = \gcd\left(u^b - 1, u^a - 1\right)$

[52] *Proof.* Assume that $a = 0$. Then, $\gcd(a, b) = b$ (this has already been proved during our above proof of Theorem 3.4.5) and thus $b = \gcd(a, b)$. Furthermore, from $a = 0$, we obtain $u^a - 1 = u^0 - 1 = 0$ (since $u^0 = 1$) and therefore

$$\begin{aligned}
\gcd\left(u^a - 1, u^b - 1\right) &= \gcd\left(0, u^b - 1\right) \\
&= \gcd\left(u^b - 1, 0\right) \qquad \text{(by Proposition 3.4.4 } \textbf{(b)}\text{)} \\
&= \left|u^b - 1\right| \qquad \text{(by Proposition 3.4.4 } \textbf{(a)}\text{)} \\
&= \left|u^{\gcd(a,b)} - 1\right| \qquad (\text{since } b = \gcd(a, b)),
\end{aligned}$$

qed.

3.5. Coprimality

3.5.1. Definition and basic properties

Perhaps ironically, gcds are at their most useful when they equal 1. This situation has a name:

Definition 3.5.1. Let a and b be two integers. We say that a is *coprime* to b if and only if $\gcd(a,b) = 1$.

Instead of "coprime", some authors say "relatively prime"[53].

Example 3.5.2. **(a)** The number 2 is coprime to 3, since $\gcd(2,3) = 1$. More generally, if a is any integer, then a is coprime to $a+1$. (Check this! Or see [19s, Example 2.10.2 **(c)**] for the proof.)

(b) The number 6 is not coprime to 15, since $\gcd(6,15) = 3 \neq 1$.

(c) Let a be an integer. Then, it is easy to see (see [19s, Example 2.10.2 **(d)**] for the proof) that

$$\gcd(a, a+2) = \gcd(a,2) = \begin{cases} 2, & \text{if } a \text{ is even;} \\ 1, & \text{if } a \text{ is odd.} \end{cases}$$

Hence, a is coprime to $a+2$ if and only if a is odd.

Following the book [GrKnPa94], we introduce a slightly quaint notation:

Definition 3.5.3. Let a and b be two integers. We write "$a \perp b$" to signify that a is coprime to b.

The relation "$\perp$" is symmetric:

Proposition 3.5.4. Let a and b be two integers. Then, $a \perp b$ if and only if $b \perp a$.

Proof of Proposition 3.5.4. Follows from Proposition 3.4.4 **(b)**. (See [19s, proof of Proposition 2.10.4] for details.) □

Note that coprimality is not transitive: i.e., if we have $a \perp b$ and $b \perp c$, then we don't usually have $a \perp c$. (A simple counterexample is $a = 2$, $b = 1$ and $c = 2$.)

Definition 3.5.5. Let a and b be two integers. Proposition 3.5.4 shows that a is coprime to b if and only if b is coprime to a. Hence, we shall sometimes use a more symmetric terminology for this situation: We shall say that "*a and b are coprime*" to mean that a is coprime to b (or, equivalently, that b is coprime to a).

[53] The book [NiZuMo91] even abbreviates "coprime" as "prime", which I find somewhat misleading (as the concept is only mildly related to the notion of a "prime" that we will discuss later on).

3.5.2. More properties and examples

The following is easy ([19s, Exercise 2.10.1]) and will be used without saying:

Exercise 3.5.1. Let $a \in \mathbb{Z}$. Prove the following:
(a) We have $1 \perp a$.
(b) We have $0 \perp a$ if and only if $|a| = 1$.

We can get more examples of coprime integers from the Fibonacci sequence:

Exercise 3.5.2. Let $(f_0, f_1, f_2, \ldots)$ be the Fibonacci sequence. Prove that $f_n \perp f_{n+1}$ for each $n \in \mathbb{N}$.

Solution to Exercise 3.5.2. We use induction on n:

Induction base: We have $\gcd(0,1) = 1$. In other words, $0 \perp 1$ (by the definition of "coprime"). In other words, $f_0 \perp f_1$ (since $f_0 = 0$ and $f_1 = 1$). In other words, Exercise 3.5.2 holds for $n = 0$.

Induction step: Let $m \in \mathbb{N}$. Assume (as the induction hypothesis) that Exercise 3.5.2 holds for $n = m$. We must prove that Exercise 3.5.2 holds for $n = m + 1$. In other words, we must prove that $f_{m+1} \perp f_{m+2}$.

Our induction hypothesis says that Exercise 3.5.2 holds for $n = m$. In other words, we have $f_m \perp f_{m+1}$. According to Proposition 3.5.4 (applied to $a = f_m$ and $b = f_{m+1}$), this entails $f_{m+1} \perp f_m$. In other words, $\gcd(f_{m+1}, f_m) = 1$ (by the definition of "coprime"). But the recursive definition of the Fibonacci sequence yields $f_{m+2} = f_{m+1} + f_m = 1f_{m+1} + f_m$. Hence,

$$\begin{aligned}\gcd(f_{m+1}, f_{m+2}) &= \gcd(f_{m+1}, 1f_{m+1} + f_m) = \gcd(f_{m+1}, f_m) \\ &\qquad \text{(by Proposition 3.4.4 \textbf{(c)}, applied to } a = f_{m+1},\ b = f_m \text{ and } u = 1) \\ &= 1.\end{aligned}$$

In other words, $f_{m+1} \perp f_{m+2}$ (by the definition of "coprime"). This is exactly what we needed to prove. Thus, the induction step is complete, and Exercise 3.5.2 is solved. □

3.5.3. Using coprimality

We can now state two important theorems about coprime numbers (and, as a reward for our previous troubles, their proofs will follow immediately from the properties of gcds). Both of them are tools for "swimming upstream" (in the sense explained in the previous section). The first one states that we can "cancel" a factor b from a divisibility $a \mid bc$ as long as this factor is coprime to a:

Theorem 3.5.6. Let $a, b, c \in \mathbb{Z}$ satisfy $a \mid bc$ and $a \perp b$. Then, $a \mid c$.

Proof of Theorem 3.5.6. We have $a \perp b$; in other words, a is coprime to b (by Definition 3.5.3). In other words, $\gcd(a, b) = 1$ (by the definition of "coprime"). Now, Theorem 3.4.11 yields $a \mid \underbrace{\gcd(a, b)}_{=1} \cdot c = c$. This proves Theorem 3.5.6. □

I like to think of Theorem 3.5.7 as a way of removing "unsolicited guests" from divisibilities. Indeed, it says that we can remove the factor b from $a \mid bc$ if we know that b is "unrelated" (i.e., coprime) to a. This is used all over number theory.

The next theorem lets us "combine" two divisibilities $a \mid c$ and $b \mid c$ to $ab \mid c$ as long as a and b are coprime:

Theorem 3.5.7. Let $a, b, c \in \mathbb{Z}$ satisfy $a \mid c$ and $b \mid c$ and $a \perp b$. Then, $ab \mid c$.

Proof of Theorem 3.5.7. We have $\gcd(a, b) = 1$ (as in the proof of Theorem 3.5.6). Now, Theorem 3.4.9 yields $ab \mid \underbrace{\gcd(a, b)}_{=1} \cdot c = c$. This proves Theorem 3.5.7. □

Theorem 3.5.7 is highly useful, and will become more so once we have learnt about prime factorization. For now, here is a quick sample application:

Exercise 3.5.3. Let n be an integer such that $2 \nmid n$ and $3 \nmid n$. Prove that $24 \mid n^2 - 1$.

Solution to Exercise 3.5.3. The integer n is odd (since $2 \nmid n$); thus, Exercise 3.3.4 yields $8 \mid n^2 - 1$. Also, $3 \nmid n$; thus Exercise 3.3.5 yields $3 \mid n^2 - 1$. But it is easy to see that $\gcd(8, 3) = 1$, so that $8 \perp 3$ (by the definition of "coprime"). Thus, Theorem 3.5.7 (applied to $a = 8$, $b = 3$ and $c = n^2 - 1$) yields $8 \cdot 3 \mid n^2 - 1$. In other words, $24 \mid n^2 - 1$ (since $8 \cdot 3 = 24$). This solves Exercise 3.5.3. □

Coprimality is "inherited" by divisors, in the sense that coprime integers have coprime divisors. To be more precise:

Proposition 3.5.8. Let $a_1, a_2, b_1, b_2 \in \mathbb{Z}$ satisfy $a_1 \mid b_1$ and $a_2 \mid b_2$ and $b_1 \perp b_2$. Then, $a_1 \perp a_2$.

Proof of Proposition 3.5.8. We have $b_1 \perp b_2$; in other words, $\gcd(b_1, b_2) = 1$ (by the definition of "coprime"). But Proposition 3.4.12 yields $\gcd(a_1, a_2) \mid \gcd(b_1, b_2) = 1$. Since $\gcd(a_1, a_2)$ is a nonnegative integer (by Proposition 3.4.3 **(a)**), this entails that $\gcd(a_1, a_2) = 1$ (since the only nonnegative integer that divides 1 is 1). In other words, $a_1 \perp a_2$ (by the definition of "coprime"). This proves Proposition 3.5.8. □

The next theorem (still part of the fallout of Bezout's theorem) is important, but we will not truly appreciate it until later:

Theorem 3.5.9. Let $a, n \in \mathbb{Z}$.
(a) There exists an $a' \in \mathbb{Z}$ such that $aa' \equiv \gcd(a, n) \bmod n$.
(b) If $a \perp n$, then there exists an $a' \in \mathbb{Z}$ such that $aa' \equiv 1 \bmod n$.
(c) If there exists an $a' \in \mathbb{Z}$ such that $aa' \equiv 1 \bmod n$, then $a \perp n$.

If $a, n \in \mathbb{Z}$, then an integer $a' \in \mathbb{Z}$ satisfying $aa' \equiv 1 \bmod n$ is called a *modular inverse* of a modulo n. The word "modular inverse" is chosen in analogy to the usual concept of an "inverse" in $\mathbb{Z}$ (which stands for an integer $a' \in \mathbb{Z}$ satisfying

$aa' = 1$; this exists if and only if a equals 1 or -1). Theorem 3.5.9 **(b)** shows that such a modular inverse always exists when $a \perp n$; Theorem 3.5.9 **(c)** is the converse of this statement (i.e., it says that if a modular inverse of a modulo n exists, then $a \perp n$).

Proof of Theorem 3.5.9 (sketched). (See [19s, proof of Theorem 2.10.8] for details.)

(a) Theorem 3.4.5 (applied to $b = n$) yields that there exist integers $x \in \mathbb{Z}$ and $y \in \mathbb{Z}$ such that $\gcd(a, n) = xa + yn$. Consider these x and y. We have[54] $\gcd(a, n) = xa + \underbrace{yn}_{\equiv 0 \bmod n} \equiv xa = ax \bmod n$, so that $ax \equiv \gcd(a, n) \bmod n$. Thus, there exists an $a' \in \mathbb{Z}$ such that $aa' \equiv \gcd(a, n) \bmod n$ (namely, $a' = x$). This proves Theorem 3.5.9 **(a)**.

(b) Assume that $a \perp n$. In other words, $\gcd(a, n) = 1$ (by the definition of "coprime"). Hence, the claim of Theorem 3.5.9 **(a)** rewrites immediately as the claim of Theorem 3.5.9 **(b)**.

(c) Assume that there exists an $a' \in \mathbb{Z}$ such that $aa' \equiv 1 \bmod n$. Consider this a'.

Let $g = \gcd(a, n)$. Then, g is a nonnegative integer[55], and Proposition 3.4.4 **(f)** (applied to $b = n$) yields $g \mid a$ and $g \mid n$.

Now, $g \mid a \mid aa'$, so that $aa' \equiv 0 \bmod g$. But also $g \mid n$. Hence, from $aa' \equiv 1 \bmod n$, we obtain $aa' \equiv 1 \bmod g$ (by Proposition 3.2.6 **(e)**, applied to g, aa' and 1 instead of m, a and b). Hence, $1 \equiv aa' \equiv 0 \bmod g$. Equivalently, $g \mid 1 - 0 = 1$. Hence, $g = 1$ (since the only nonnegative integer that divides 1 is 1). Thus, $\gcd(a, n) = g = 1$. In other words, $a \perp n$. This proves Theorem 3.5.9 **(c)**. □

3.5.4. Multiplying coprimalities

The next theorem provides a way to "multiply" coprimalities. Note how modular inverses are used in its proof:

Theorem 3.5.10. Let $a, b, c \in \mathbb{Z}$ such that $a \perp c$ and $b \perp c$. Then, $ab \perp c$.

Proof of Theorem 3.5.10. Theorem 3.5.9 **(b)** (applied to $n = c$) yields that there exists an $a' \in \mathbb{Z}$ such that $aa' \equiv 1 \bmod c$. Likewise, Theorem 3.5.9 **(b)** (applied to b and c instead of a and n) yields that there exists a $b' \in \mathbb{Z}$ such that $bb' \equiv 1 \bmod c$. Consider these a' and b'.

Multiplying the two congruences $aa' \equiv 1 \bmod c$ and $bb' \equiv 1 \bmod c$, we obtain $(aa')(bb') \equiv 1 \cdot 1 = 1 \bmod c$.

Now, define the integers $r = ab$ and $s = a'b'$. Then, $\underbrace{r}_{=ab}\underbrace{s}_{=a'b'} = (ab)(a'b') = (aa')(bb') \equiv 1 \bmod c$. Hence, there exists an $r' \in \mathbb{Z}$ such that $rr' \equiv 1 \bmod c$ (namely, $r' = s$). Thus, Theorem 3.5.9 **(c)** (applied to r and c instead of a and n) yields that $r \perp c$. In view of $r = ab$, this rewrites as $ab \perp c$. This proves Theorem 3.5.10. □

[54] Here, we are again using the substitution principle for congruences (from Example 3.2.13). Convince yourself that you know how to do without this principle.

[55] by Proposition 3.4.3 **(a)**

Theorem 3.5.10 can be generalized (in a straightforward way) to products of several numbers:

Exercise 3.5.4. Let $c \in \mathbb{Z}$. Let $a_1, a_2, \ldots, a_k$ be integers such that each $i \in \{1, 2, \ldots, k\}$ satisfies $a_i \perp c$. Prove that $a_1 a_2 \cdots a_k \perp c$.

Hint to Exercise 3.5.4. Induct on k. (See [19s, solution to Exercise 2.10.2] for the details.) □

A quick consequence of this exercise is that powers of coprime integers are still coprime:

Exercise 3.5.5. Let $a, b \in \mathbb{Z}$ be such that $a \perp b$. Let $n, m \in \mathbb{N}$. Prove that $a^n \perp b^m$.

Thus, for example, $2^{14} \perp 3^6$, because $2 \perp 3$.

Hint to Exercise 3.5.5. We have $a \perp b$. Thus, Exercise 3.5.4 (applied to $k = n$ and $a_i = a$) yields $a^n \perp b$. According to Proposition 3.5.4, this entails $b \perp a^n$. Hence, another application of Exercise 3.5.4 yields $b^m \perp a^n$. Thus, by Proposition 3.5.4, we get $a^n \perp b^m$. (See [19s, solution to Exercise 2.10.4] for the details.) □

We can generalize Theorem 3.5.7 to show that the product of several mutually coprime divisors of an integer c must again be a divisor of c:

Exercise 3.5.6. Let $c \in \mathbb{Z}$. Let $b_1, b_2, \ldots, b_k$ be integers that are mutually coprime (i.e., they satisfy $b_i \perp b_j$ for all $i \neq j$). Assume that $b_i \mid c$ for each $i \in \{1, 2, \ldots, k\}$. Prove that $b_1 b_2 \cdots b_k \mid c$.

Hint to Exercise 3.5.6. Induct on k using Exercise 3.5.4. (See [19s, solution to Exercise 2.10.3] for the details.) □

The above results have one important application to congruences. Recall that if a, b, c are integers satisfying $ab = ac$, then we can "cancel" a from the equality $ab = ac$ to obtain $b = c$ as long as a is nonzero. Something similar is true for congruences modulo n, but the condition "a is nonzero" has to be replaced by "a is coprime to n":

Lemma 3.5.11. Let a, b, c, n be integers such that $a \perp n$ and $ab \equiv ac \bmod n$. Then, $b \equiv c \bmod n$.

Lemma 3.5.11 says that we can cancel an integer a from a congruence $ab \equiv ac \bmod n$ as long as a is coprime to n. Let us give two proofs of this lemma, to illustrate the uses of some of the previous results:

First proof of Lemma 3.5.11. We have $ab \equiv ac \bmod n$. In other words, $n \mid ab - ac = a(b - c)$. But Proposition 3.5.4 (applied to n instead of b) shows that $a \perp n$ if and only if $n \perp a$. Thus, we have $n \perp a$ (since $a \perp n$).

Thus, we know that $n \mid a(b - c)$ and $n \perp a$. Hence, Theorem 3.5.6 (applied to n, a and $b - c$ instead of a, b and c) yields $n \mid b - c$. In other words, $b \equiv c \bmod n$. This proves Lemma 3.5.11. □

Second proof of Lemma 3.5.11. Theorem 3.5.9 **(b)** yields that there exists an $a' \in \mathbb{Z}$ such that $aa' \equiv 1 \bmod n$ (since $a \perp n$). Consider this a'. Now, multiplying both sides of the congruence $ab \equiv ac \bmod n$ by a', we obtain

$$a'ab \equiv a'ac \bmod n.$$

But we have $\underbrace{a'a}_{=aa' \equiv 1 \bmod n} c \equiv 1c = c \bmod n$ and similarly $a'ab \equiv b \bmod n$. Hence,

$$b \equiv a'ab \equiv a'ac \equiv c \bmod n.$$

This proves Lemma 3.5.11. □

3.5.5. Reduced fractions

We notice that any two integers can be "made coprime" by factoring out their gcd, unless they are both 0:

Proposition 3.5.12. Let a and b be two integers such that $(a, b) \neq (0, 0)$. Let $g = \gcd(a, b)$. Then, $g > 0$ and $\dfrac{a}{g} \perp \dfrac{b}{g}$.

Proof of Proposition 3.5.12. The integers a and b are not both 0 (since $(a, b) \neq (0, 0)$). Hence, $\gcd(a, b)$ is a positive integer (by Proposition 3.4.3 **(b)**). Thus, $\gcd(a, b) > 0$. In other words, $g > 0$ (since $g = \gcd(a, b)$).

Moreover, $g = \gcd(a, b)$, so that Proposition 3.4.4 **(f)** yields $g \mid a$ and $g \mid b$. Hence, $\dfrac{a}{g}$ and $\dfrac{b}{g}$ are integers. Thus, Theorem 3.4.8 (applied to g, $\dfrac{a}{g}$ and $\dfrac{b}{g}$) yields

$$\gcd\left(g \cdot \frac{a}{g}, g \cdot \frac{b}{g}\right) = \underbrace{|g|}_{\substack{=g \\ (\text{since } g>0)}} \gcd\left(\frac{a}{g}, \frac{b}{g}\right) = g \gcd\left(\frac{a}{g}, \frac{b}{g}\right).$$

Therefore,

$$\gcd\left(\frac{a}{g}, \frac{b}{g}\right) = \frac{1}{g} \gcd\left(\underbrace{g \cdot \frac{a}{g}}_{=a}, \underbrace{g \cdot \frac{b}{g}}_{=b}\right) = \frac{1}{g} \gcd(a, b) = 1$$

(since $g = \gcd(a, b)$). In other words, $\dfrac{a}{g} \perp \dfrac{b}{g}$ (by the definition of "coprime"). This proves Proposition 3.5.12. □

An easy consequence of Proposition 3.5.12 is the following fundamental fact ([19s, Exercise 2.10.14]):

Corollary 3.5.13. Let $r \in \mathbb{Q}$. Then, there exist two **coprime** integers a and b satisfying $r = a/b$.

Proof of Corollary 3.5.13. Write r as $r = x/y$ for some integers x and y (with $y \neq 0$). Apply Proposition 3.5.12 to $a = x$ and $b = y$. (Details are found in [19s, solution to Exercise 2.10.14].) □

Corollary 3.5.13 is commonly stated in the form "any rational number can be represented as a reduced fraction". Here, "reduced fraction" means a fraction in which the numerator and the denominator are coprime integers. (Such fractions are also known as "irreducible fractions" or "fractions in lowest terms".)

3.5.6. The rational root test

Exercise 3.5.7. Find three integers x, y and z such that $\gcd(x,y,z) = 1$, yet no two of x, y and z are coprime.

Solution to Exercise 3.5.7 (sketched). Computer experiments quickly reveal that $x = 6$, $y = 10$ and $z = 15$ is an answer. Indeed, $\gcd(6,10,15) = 1$, yet no two of 6, 10 and 15 are coprime (since $\gcd(6,10) = 2 \neq 1$ and $\gcd(6,15) = 3 \neq 1$ and $\gcd(10,15) = 5 \neq 1$).

This can be found without a computer. The trick is to find three integers a, b and c that are mutually coprime (i.e., any two of a, b and c are coprime) and greater than 1, and then set $x = bc$, $y = ca$ and $z = ab$. Then, these integers x, y and z satisfy

$$\begin{aligned}
\gcd(x,y) &= \gcd(bc,ca) = \gcd(ca,bc) = \gcd(ca,cb) \\
&= |c| \underbrace{\gcd(a,b)}_{\substack{=1 \\ \text{(since any two of } a,\, b \text{ and } c \\ \text{are coprime)}}} \qquad \text{(by Theorem 3.4.8)} \\
&= |c| = c \qquad (\text{since } c > 1 \geq 0) \\
&> 1
\end{aligned}$$

and similarly $\gcd(y,z) > 1$ and $\gcd(z,x) > 1$ (which shows that no two of x, y and z are coprime), but also

$$\begin{aligned}
\gcd(x,y,z) &= \gcd\left(\underbrace{\gcd(x,y)}_{=c}, \underbrace{\gcd(z)}_{\substack{=z \\ \text{(by Proposition 3.4.4 } \textbf{(a)})}}\right) \qquad \text{(by Theorem 3.4.15)} \\
&= \gcd\left(c, \underbrace{z}_{=ab}\right) = \gcd(c,ab) = 1
\end{aligned}$$

(since Theorem 3.5.10 yields $ab \perp c$). Thus, in order to solve the exercise, it remains to find three integers a, b and c that are mutually coprime and greater than 1. But this is easy – for example, one can take $a = 5$ and $b = 3$ and $c = 2$. Taking these yields $x = 6$, $y = 10$ and $z = 15$, which is our above example. $\square$

Another application of coprime integers is the following fact (known as the *rational root test*):

Theorem 3.5.14. Let $P(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_0 x^0$ be a polynomial in a single variable x with integer coefficients (i.e., all of $a_0, a_1, \ldots, a_n$ are integers). Let r be a rational root of $P(x)$ (that is, a rational number satisfying $P(r) = 0$). Write r in the form $r = p/q$ for some integers p and q satisfying $p \perp q$. Then, $p \mid a_0$ and $q \mid a_n$.

Remark 3.5.15. Theorem 3.5.14 provides an algorithm for finding all rational roots of a polynomial with integer coefficients. Indeed, let $P(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_0 x^0$ be such a polynomial. We can WLOG assume that $a_n \neq 0$ (otherwise, we throw away the $a_n x^n$ term and replace n by $n-1$) and that $a_0 \neq 0$ (otherwise, we record 0 as a root of our polynomial and divide $P(x)$ by x). Any rational root r of $P(x)$ can be written in the form $r = p/q$ for some integers p and q satisfying $p \perp q$ (by Corollary 3.5.13). Writing it in this form, we then conclude from Theorem 3.5.14 that $p \mid a_0$ and $q \mid a_n$. But $p \mid a_0$ yields only finitely many options for p (since a_0 is nonzero and thus has only finitely many divisors), and $q \mid a_n$ yields only finitely many options for q (since a_n is nonzero and thus has only finitely many divisors). Thus, we have only finitely many options for r (since $r = p/q$). By listing all these options and checking which of them actually satisfy $P(r) = 0$, we can identify all rational roots of $P(x)$. (Note that divisors of a positive integer can be negative; however, we can WLOG assume that $q > 0$ because otherwise we can replace p and q by $-p$ and $-q$.)

Proof of Theorem 3.5.14. We have $P(r) = 0$ (since r is a root of $P(x)$), thus

$$\begin{aligned}
0 &= P(r) = P(p/q) \qquad (\text{since } r = p/q) \\
&= a_n (p/q)^n + a_{n-1} (p/q)^{n-1} + \cdots + a_0 (p/q)^0 \\
&\qquad \left(\text{since } P(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_0 x^0\right) \\
&= a_n \cdot \frac{p^n}{q^n} + a_{n-1} \cdot \frac{p^{n-1}}{q^{n-1}} + \cdots + a_0 \cdot \frac{p^0}{q^0} \\
&= \frac{1}{q^n} \left(a_n p^n + a_{n-1} p^{n-1} q + \cdots + a_1 p q^{n-1} + a_0 q^n\right).
\end{aligned}$$

Multiplying both sides of this equality by q^n, we find

$$0 = a_n p^n + a_{n-1} p^{n-1} q + \cdots + a_1 p q^{n-1} + a_0 q^n. \tag{54}$$

Solving the equation (54) for a_0q^n, we find

$$\begin{aligned} a_0q^n &= -\left(a_np^n + a_{n-1}p^{n-1}q + \cdots + a_1pq^{n-1}\right) \\ &= -p\left(a_np^{n-1} + a_{n-1}p^{n-2}q + \cdots + a_1q^{n-1}\right). \end{aligned}$$

Thus, $-p \mid a_0q^n$ (since $a_np^{n-1} + a_{n-1}p^{n-2}q + \cdots + a_1q^{n-1}$ is an integer). Hence, $p \mid -p \mid a_0q^n = q^na_0$.

But $p \perp q$. Hence, Exercise 3.5.5 (applied to p, q, 1 and n instead of a, b, n and m) yields $p^1 \perp q^n$. In other words, $p \perp q^n$. Hence, Theorem 3.5.6 (applied to p, q^n and a_0 instead of a, b and c) yields $p \mid a_0$ (since $p \mid q^na_0$).

The proof of $q \mid a_n$ is similar: Solving the equation (54) for a_np^n, we find

$$\begin{aligned} a_np^n &= -\left(a_{n-1}p^{n-1}q + \cdots + a_1pq^{n-1} + a_0q^n\right) \\ &= -q\left(a_{n-1}p^{n-1} + \cdots + a_1pq^{n-2} + a_0q^{n-1}\right). \end{aligned}$$

Thus, $-q \mid a_np^n$ (since $a_{n-1}p^{n-1} + \cdots + a_1pq^{n-2} + a_0q^{n-1}$ is an integer). Hence, $q \mid -q \mid a_np^n = p^na_n$.

But $p \perp q$. Hence, Exercise 3.5.5 (applied to p, q, n and 1 instead of a, b, n and m) yields $p^n \perp q^1$. In other words, $p^n \perp q$. By Proposition 3.5.4, this yields $q \perp p^n$. Hence, Theorem 3.5.6 (applied to q, p^n and a_n instead of a, b and c) yields $q \mid a_n$ (since $q \mid p^na_n$).

Thus, Theorem 3.5.14 is proved. □

Corollary 3.5.16. Let $P(x) = a_nx^n + a_{n-1}x^{n-1} + \cdots + a_0x^0$ be a polynomial in a single variable x with integer coefficients (i.e., all of $a_0, a_1, \ldots, a_n$ are integers). Let r be an integer root of $P(x)$ (that is, an integer satisfying $P(r) = 0$). Then, $r \mid a_0$.

Proof of Corollary 3.5.16. We have $r = r/1$ and $r \perp 1$. Hence, Theorem 3.5.14 (applied to $p = r$ and $q = 1$) yields $r \mid a_0$ and $1 \mid a_n$. Corollary 3.5.16 is thus proved. □

Corollary 3.5.17. Let $P(x) = a_nx^n + a_{n-1}x^{n-1} + \cdots + a_0x^0$ be a polynomial in a single variable x with integer coefficients (i.e., all of $a_0, a_1, \ldots, a_n$ are integers) and with $a_n = \pm 1$. Let r be a rational root of $P(x)$ (that is, a rational number satisfying $P(r) = 0$). Then, r is an integer and satisfies $r \mid a_0$.

Proof of Corollary 3.5.16. Corollary 3.5.13 yields that there exist two **coprime** integers a and b satisfying $r = a/b$. Consider these a and b, and denote them by p and q. Thus, p and q are two coprime integers satisfying $r = p/q$. Hence, $p \perp q$ (since p and q are coprime). Thus, Theorem 3.5.14 yields $p \mid a_0$ and $q \mid a_n$. However, from $a_n = \pm 1$, we obtain $a_n \mid 1$, so that $q \mid a_n \mid 1$ and therefore $q = \pm 1$ (since the only divisors of 1 are 1 and -1). Thus, $r = p/\underbrace{q}_{=\pm 1} = p/(\pm 1) = \pm p \in \mathbb{Z}$ (since $p \in \mathbb{Z}$). In other words, r is an integer. Moreover, from $r = \pm p$, we obtain $r \mid p \mid a_0$. Thus, the proof of Corollary 3.5.17 is complete. □

We state another property of gcds and coprimality, which will come useful in an exercise later on:

Proposition 3.5.18. Let $a, b, c \in \mathbb{Z}$ satisfy $a \perp c$. Then, $\gcd(a, bc) = \gcd(a, b)$.

Proof of Proposition 3.5.18. Proposition 3.4.4 **(f)** (applied to bc instead of b) yields $\gcd(a, bc) \mid a$ and $\gcd(a, bc) \mid bc$. Proposition 3.5.8 (applied to $a_1 = \gcd(a, bc)$, $a_2 = c$, $b_1 = a$ and $b_2 = c$) yields $\gcd(a, bc) \perp c$ (since $\gcd(a, bc) \mid a$ and $c \mid c$ and $a \perp c$).

We have $\gcd(a, bc) \mid bc = cb$ and $\gcd(a, bc) \perp c$. Hence, Theorem 3.5.6 (applied to $\gcd(a, bc)$, c and b instead of a, b and c) yields $\gcd(a, bc) \mid b$.

But Theorem 3.4.7 (applied to $m = \gcd(a, bc)$) shows that we have the following logical equivalence:

$$(\gcd(a, bc) \mid a \text{ and } \gcd(a, bc) \mid b) \iff (\gcd(a, bc) \mid \gcd(a, b)).$$

Hence, we have $\gcd(a, bc) \mid \gcd(a, b)$ (since we have $\gcd(a, bc) \mid a$ and $\gcd(a, bc) \mid b$).

On the other hand, $a \mid a$ and $b \mid bc$. Hence, Proposition 3.4.12 (applied to $a_1 = a$, $a_2 = b$, $b_1 = a$ and $b_2 = bc$) yields

$$\gcd(a, b) \mid \gcd(a, bc).$$

Recall also that $\gcd(a, bc) \mid \gcd(a, b)$. Thus, Proposition 3.1.3 **(c)** (applied to $\gcd(a, bc)$ and $\gcd(a, b)$ instead of a and b) yields $|\gcd(a, bc)| = |\gcd(a, b)|$. But a gcd is always a nonnegative integer (by Proposition 3.4.3 **(a)**); thus, both $\gcd(a, bc)$ and $\gcd(a, b)$ are nonnegative integers, and therefore satisfy $|\gcd(a, bc)| = \gcd(a, bc)$ and $|\gcd(a, b)| = \gcd(a, b)$. Hence, $\gcd(a, bc) = |\gcd(a, bc)| = |\gcd(a, b)| = \gcd(a, b)$. This proves Proposition 3.5.18. □

3.6. Lowest common multiples

The notion of a *lowest common multiple* is, so to speak, an "upside-down" counterpart to the notion of a greatest common divisor. We shall restrict ourselves to defining this notion and stating its most important properties.

First, we define the notion of a common multiple:

Definition 3.6.1. Let $b_1, b_2, \ldots, b_k$ be integers. Then, the *common multiples* of $b_1, b_2, \ldots, b_k$ are defined to be the integers a that satisfy

$$(b_i \mid a \text{ for all } i \in \{1, 2, \ldots, k\}).$$

(In other words, a *common multiple* of $b_1, b_2, \ldots, b_k$ is an integer that is a multiple of each of $b_1, b_2, \ldots, b_k$.)

For example, the common multiples of the two integers $6, 9$ are $\ldots, -54, -36, -18, 0, 18, 36, 54, \ldots$, that is, all multiples of 18. Clearly, if $b_1, b_2, \ldots, b_k$ are k integers, then the product $b_1 b_2 \cdots b_k$ is a common multiple of $b_1, b_2, \ldots, b_k$ (and so is any multiple of this product, including the number 0).

Now, we can define the *lowest common multiple* of k integers $b_1, b_2, \ldots, b_k$. This concept is a bit of a misnomer, since the word "lowest" is not to be taken literally; to be more precise, it should be called "lowest positive common multiple" (and even this would be wrong when one of $b_1, b_2, \ldots, b_k$ is 0). So here is the rigorous definition of the lowest common multiple:

Definition 3.6.2. Let $b_1, b_2, \ldots, b_k$ be finitely many integers. The *lowest common multiple* (also known as the *least common multiple*, or, for short, the *lcm*) of $b_1, b_2, \ldots, b_k$ is the nonnegative integer $\operatorname{lcm}(b_1, b_2, \ldots, b_k)$ defined as follows:

- If $b_1, b_2, \ldots, b_k$ are all nonzero, then it is defined as the smallest **positive** common multiple of $b_1, b_2, \ldots, b_k$. (It is easy to see that this is well-defined[56].)
- If $b_1, b_2, \ldots, b_k$ are not all nonzero (i.e., at least one of $b_1, b_2, \ldots, b_k$ is zero), then it is defined to be 0.

For example, $\operatorname{lcm}(6, 9) = 18$ and $\operatorname{lcm}(3, 5) = 15$ and $\operatorname{lcm}(6, 10, 15) = 30$ and $\operatorname{lcm}(2, 4, 0) = 0$. Definition 3.6.2 immediately yields the following:

Proposition 3.6.3. Let $b_1, b_2, \ldots, b_k$ be finitely many integers.
(a) The number $\operatorname{lcm}(b_1, b_2, \ldots, b_k)$ is a nonnegative integer.
(b) If $b_1, b_2, \ldots, b_k$ are all nonzero, then $\operatorname{lcm}(b_1, b_2, \ldots, b_k)$ is a positive integer.

The following theorem ([19s, Theorem 2.11.6]) gives an alternative characterization for lcms of two integers:

Theorem 3.6.4. Let $a, b \in \mathbb{Z}$. Then, $\gcd(a, b) \cdot \operatorname{lcm}(a, b) = |ab|$.

Hints to the proof of Theorem 3.6.4. (See [19s, Theorem 2.11.6] for the details.) We WLOG assume that $a \neq 0$ and $b \neq 0$ (since otherwise, our claim immediately boils down to $0 = 0$). Then, set $c = \dfrac{ab}{\gcd(a, b)}$. It is now easy to see that c is a nonzero integer (since $\gcd(a, b) \mid a \mid ab$ and $ab \neq 0$) and is a common multiple of a and b

[56] *Proof.* Assume that $b_1, b_2, \ldots, b_k$ are nonzero. Then, the product $b_1 b_2 \cdots b_k$ is nonzero, so that its absolute value $|b_1 b_2 \cdots b_k|$ is a positive integer. Moreover, $|b_1 b_2 \cdots b_k|$ is a common multiple of $b_1, b_2, \ldots, b_k$ (indeed, each $i \in \{1, 2, \ldots, k\}$ satisfies $b_i \mid b_1 b_2 \cdots b_k \mid |b_1 b_2 \cdots b_k|$). Hence, there exists at least one positive common multiple of $b_1, b_2, \ldots, b_k$ (namely, $|b_1 b_2 \cdots b_k|$). Hence, the set of all positive common multiples of $b_1, b_2, \ldots, b_k$ is nonempty. Since this set is a set of nonnegative integers, we thus conclude (using Theorem 5.1.2) that it has a minimum. In other words, the smallest **positive** common multiple of $b_1, b_2, \ldots, b_k$ is well-defined.

(since $\gcd(a,b) \mid b$ and $\gcd(a,b) \mid a$). Thus, $|c|$ is a positive common multiple of a and b (indeed, $c \neq 0$, so that $|c|$ is positive).

On the other hand, the definition of c yields $\gcd(a,b) \cdot c = ab$. Now, if x is any positive common multiple of a and b, then Theorem 3.4.9 (applied to x instead of c) yields $ab \mid \gcd(a,b) \cdot x$, so that $\gcd(a,b) \cdot c = ab \mid \gcd(a,b) \cdot x$ and therefore $c \mid x$ (here, we have cancelled the nonzero factor $\gcd(a,b)$ from our divisibility), which entails $x \geq |c|$. In other words, any positive common multiple of a and b is $\geq |c|$. Thus, $|c|$ is the **smallest** positive common multiple of a and b (since we already know that $|c|$ is a positive common multiple of a and b). In other words, $|c| = \operatorname{lcm}(a,b)$. Hence,

$$\operatorname{lcm}(a,b) = |c| = \left| \frac{ab}{\gcd(a,b)} \right| = \frac{|ab|}{\gcd(a,b)}$$

(since $\gcd(a,b)$ is positive), and this yields the claim of Theorem 3.6.4. □

The lowest common multiple $\operatorname{lcm}(a,b)$ of two integers a and b has a universal property analogous to that of the greatest common divisor (Theorem 3.4.7):

Theorem 3.6.5. Let $a, b \in \mathbb{Z}$ and $m \in \mathbb{Z}$. Then, we have the following logical equivalence:

$$(a \mid m \text{ and } b \mid m) \iff (\operatorname{lcm}(a,b) \mid m). \tag{55}$$

Hints to the proof of Theorem 3.6.5. (See [19s, Theorem 2.11.7 **(a)**] for the details.) We WLOG assume that $a \neq 0$ and $b \neq 0$ (since otherwise, both sides of (55) are equivalent to the statement "$m = 0$" and therefore are equivalent to each other). Let $n = \operatorname{lcm}(a,b)$. Thus, n is a positive integer. Now, our goal is to prove the equivalence (55). The "$\Longleftarrow$" direction of this equivalence is easy (in fact, if $\operatorname{lcm}(a,b) \mid m$, then $a \mid \operatorname{lcm}(a,b) \mid m$ and $b \mid \operatorname{lcm}(a,b) \mid m$), so we only need to prove the "$\Longrightarrow$" direction. To achieve this, we assume that $a \mid m$ and $b \mid m$. Thus, m is a common multiple of a and b. Therefore, the remainder $m\%n$ must also be a common multiple of a and b (why?). However, this remainder $m\%n$ belongs to $\{0, 1, \ldots, n-1\}$, and thus is either 0 or a positive integer smaller than n. However, as a common multiple of a and b, it cannot be a positive integer smaller than n (because $n = \operatorname{lcm}(a,b)$ is the **smallest** positive common multiple of a and b). Hence, it must be 0. In other words, $m\%n = 0$. Equivalently, $n \mid m$, so that $\operatorname{lcm}(a,b) = n \mid m$. This proves the "$\Longrightarrow$" direction of the equivalence (55); thus, our proof of Theorem 3.6.5 is complete. □

Theorem 3.6.5 can be generalized to k integers:

Theorem 3.6.6. Let $k \in \mathbb{N}$, let $b_1, b_2, \ldots, b_k \in \mathbb{Z}$ and $m \in \mathbb{Z}$. Then, we have the following logical equivalence:

$$(b_i \mid m \text{ for all } i \in \{1, 2, \ldots, k\}) \iff (\operatorname{lcm}(b_1, b_2, \ldots, b_k) \mid m).$$

Hints to the proof of Theorem 3.6.6. This is a straightforward adaptation of our proof of Theorem 3.6.5. □

We end our first dive into number theory here, but we shall come back to it a few more times during this course. One last comment at this point: It is commonly believed that number theory quickly starts requiring advanced mathematics (analysis, geometry, abstract algebra) as one goes beyond the basics. This is not true; there are several books full of beautiful results treated with elementary means ([Stein09], [UspHea39], [NiZuMo91]; see also [AnDoMu17] for a text specifically targeted at olympiad problems). There also are miraculous applications accessible at the most basic level, such as the RSA cryptosystem; as these are nowadays treated in most courses on abstract algebra or cryptography, I shall not discuss them here.

3.7. Homework set #1: Induction and number theory

This is a **regular** problem set. Each problem will be graded out of 10 points based on correctness and readability. You don't need to motivate your proofs, but they should be readable to a reasonably competent reader (say, to your fellow students).

The topic of this homework set is induction and elementary number theory; this means that several (but not all!) of its problems are about these topics.

Please solve **at most 5** problems. (No points will be given for further solutions.)

Exercise 3.7.1. Let k be a positive integer. A set S of integers is said to be *k-lacunar* if every two distinct elements $u, v \in S$ satisfy $|u - v| \geq k$. (Thus, a 2-lacunar set is the same as a lacunar set as defined in Definition 2.3.3.)

Let $n \in \mathbb{N}$. Let S be a k-lacunar subset of $\{1, 2, \ldots, n\}$. Prove that

$$|S| \leq \frac{n + k - 1}{k}.$$

(Thus, in particular, if S is a lacunar subset of $\{1, 2, \ldots, n\}$, then $|S| \leq \dfrac{n+1}{2}$.)

Exercise 3.7.2. Let $(f_0, f_1, f_2, \ldots)$ be the Fibonacci sequence. Prove that any $n, m \in \mathbb{N}$ satisfy

$$\gcd(f_n, f_m) = f_{\gcd(n,m)}.$$

Exercise 3.7.3. Let $(F_0, F_1, F_2, \ldots)$ be the *Fermat sequence* – that is, the sequence of integers defined by

$$F_n = 2^{2^n} + 1 \qquad \text{for each } n \in \mathbb{N}.$$

(Keep in mind that nested powers are to be read top-to-bottom: That is, the expression "a^{b^c}" means $a^{(b^c)}$ rather than $\left(a^b\right)^c$.)

(a) Prove that

$$F_n = F_0 F_1 \cdots F_{n-1} + 2 \qquad \text{for every integer } n \geq 0.$$

(b) Prove that $\gcd(F_n, F_m) = 1$ for any two distinct nonnegative integers n and m.

Exercise 3.7.4. Prove that there exist **infinitely many odd** positive integers n for which

$$\frac{1! \cdot 2! \cdot \cdots \cdot (2n)!}{(n+1)!} \text{ is a perfect square.}$$

Exercise 3.7.5. Let $x \in \mathbb{R}$. Let n and m be positive integers. Prove that

$$\sum_{k=0}^{mn-1} \left\lfloor x + \frac{k}{n} \right\rfloor = m \left(\lfloor nx \rfloor + \frac{n(m-1)}{2} \right).$$

Exercise 3.7.6. Let $(f_0, f_1, f_2, \ldots)$ be the Fibonacci sequence. Prove that each nonnegative integer n satisfies

$$f_n \equiv f_{n\%5} \cdot 3^{n//5} \bmod 5.$$

(See Definition 3.3.1 for the notations used here.)

Exercise 3.7.7. Let a and b be two **positive** integers. Prove that there exist **positive** integers x and y such that

$$\gcd(a, b) = xa - yb.$$

Exercise 3.7.8. Prove that any positive integer a can be uniquely expressed in the form

$$a = 3^m + b_{m-1} 3^{m-1} + b_{m-2} 3^{m-2} + \cdots + b_0 3^0$$

where m is a nonnegative integer, and where $b_0, b_1, \ldots, b_{m-1} \in \{0, 1, -1\}$. (This is called the *balanced ternary representation*.)

Exercise 3.7.9. Let $p, q, m, n \in \mathbb{N}$ with $p \leq m$ and $q \leq n$. Consider an $m \times n$-table T of integers, with all entries distinct. In each column of T, we mark the p largest entries with a *cyan* marker. In each row of T, we mark the q largest entries with a *red* marker. Prove that at least pq entries of T are marked twice (i.e., with both colors).

[**Example:** Let $p = 2$ and $q = 2$ and $m = 3$ and $n = 3$ and

$$T = \begin{pmatrix} 1 & 2 & 9 \\ 4 & 3 & 8 \\ 5 & 6 & 7 \end{pmatrix}.$$

Then,

$$\begin{aligned}&\text{the cyan entries are } 4,5,3,6,8,9, \text{ while}\\&\text{the red entries are } 2,9,4,8,6,7.\end{aligned}$$

Thus, the entries $4,6,8,9$ are marked twice. This is exactly the pq entries claimed in the exercise. You can easily find situations in which there are more than pq doubly-marked entries.]

Exercise 3.7.10. A *bitstring* shall mean a finite sequence consisting of 0's and 1's. (This is what we called an "n-bitstring" in Exercise 2.1.3, except that the length is no longer fixed.) We shall write our bitstrings without commas and parentheses – i.e., we shall simply write $a_1a_2\cdots a_n$ for the bitstring $(a_1, a_2, \ldots, a_n)$.

Bitstrings can be transformed by *moves*. In each *move*, you pick two consecutive entries 01 in the bitstring (appearing in this order), and replace them by three consecutive entries 100 (in this order). In other words, in each, move you replace a bitstring of the form $\ldots 01 \ldots$ by $\ldots 100 \ldots$, where the two "$\ldots$" parts stay unchanged. For example, here is a move:

$$011\underline{01}101 \to 011\underline{100}101$$

(where we are using an underscore to mark the place where the move is happening). Here is a sequence of moves:

$$01\underline{01} \to \underline{01}100 \to 10\underline{01}00 \to 1\underline{01}0000 \to 11000000$$

(where we are putting an underscore under the position of the next move). Note that the last bitstring in this sequence has no two consecutive entries 01 any more, and thus no more moves can be applied to it.

(a) Prove that there are no infinite sequences of moves. That is, if you start with a bitstring a, then any sequence of moves that can be applied successively must have an end.

(b) A bitstring shall be called *immovable* if no move can be applied to it. Part **(a)** shows that, starting with any bitstring a, we can always reach an immovable bitstring by performing moves until no more moves are possible. Prove that this immovable bitstring is uniquely determined by a – that is, no matter how you perform the moves, the immovable bitstring that results at the end will be the same. Moreover, the number of moves needed to reach the immovable bitstring will be the same.

3.8. Recitation #2: Coprimality and more number theory

We shall now discuss a sample of exercises related to divisibility and coprimality.

The following exercise is essentially Problem 1 from the 1st International Mathematical Olympiad 1959:

Exercise 3.8.1. Let $n \in \mathbb{Z}$. Prove that $21n + 4 \perp 14n + 3$.

We shall give two solutions for this exercise:

First solution to Exercise 3.8.1. We shall apply Proposition 3.4.4 similarly to how it is employed in the Euclidean algorithm – using Proposition 3.4.4 **(b)** to interchange the two arguments in $\gcd(a, b)$, and using Proposition 3.4.4 **(c)** to subtract a multiple of the first argument from the second in $\gcd(a, b)$. We perform these operations in such a way that the two arguments gradually get simpler.

Here is how this looks like:

$$\begin{aligned}
&\gcd(21n + 4, 14n + 3) \\
&= \gcd\left(14n + 3, \underbrace{21n + 4}_{=1\cdot(14n+3)+(7n+1)}\right) \qquad \text{(by Proposition 3.4.4 \textbf{(b)})} \\
&= \gcd(14n + 3, 1 \cdot (14n + 3) + (7n + 1)) \\
&= \gcd(14n + 3, 7n + 1) \\
&\qquad \text{(by Proposition 3.4.4 \textbf{(c)}, applied to } a = 14n + 3,\ b = 7n + 1 \text{ and } u = 1) \\
&= \gcd\left(7n + 1, \underbrace{14n + 3}_{=2\cdot(7n+1)+1}\right) \qquad \text{(by Proposition 3.4.4 \textbf{(b)})} \\
&= \gcd(7n + 1, 2 \cdot (7n + 1) + 1) \\
&= \gcd(7n + 1, 1) \\
&\qquad \text{(by Proposition 3.4.4 \textbf{(c)}, applied to } a = 7n + 1,\ b = 1 \text{ and } u = 2) \\
&= \gcd(1, 7n + 1) \qquad \text{(by Proposition 3.4.4 \textbf{(b)})} \\
&= 1
\end{aligned}$$

(since Exercise 3.5.1 **(a)** yields $1 \perp 7n + 1$ and thus $\gcd(1, 7n + 1) = 1$). In other words, $21n + 4 \perp 14n + 3$. This solves Exercise 3.8.1. □

A second solution to Exercise 3.8.1 can be given using the following converse of Bezout's theorem:

Proposition 3.8.1. Let a, b, x and y be four integers. Then:
(a) We have $\gcd(a, b) \mid xa + yb$.
(b) If $xa + yb = 1$, then $a \perp b$.

Proof of Proposition 3.8.1. **(a)** Let $g = \gcd(a, b)$. Thus, $g = \gcd(a, b) \mid a$ (by Proposition 3.4.4 **(f)**), so that $a \equiv 0 \bmod g$. Similarly, $b \equiv 0 \bmod g$. Now, $x \underbrace{a}_{\equiv 0 \bmod g} + y \underbrace{b}_{\equiv 0 \bmod g} \equiv x \cdot 0 + y \cdot 0 = 0 \bmod g$. In other words, $g \mid xa + yb$. In other words, $\gcd(a, b) \mid xa + yb$ (since $g = \gcd(a, b)$). This proves Proposition 3.8.1 **(a)**.

(b) We know (from Proposition 3.4.3 **(a)**) that $\gcd(a,b)$ is a nonnegative integer. But Proposition 3.8.1 **(a)** yields $\gcd(a,b) \mid xa + yb = 1$. Thus, $\gcd(a,b)$ is a divisor of 1, and therefore a nonnegative divisor of 1 (since $\gcd(a,b)$ is nonnegative). But the only nonnegative divisor of 1 is 1 itself. Hence, we conclude that $\gcd(a,b) = 1$. In other words, $a \perp b$. This proves Proposition 3.8.1 **(b)**. $\square$

Second solution to Exercise 3.8.1. We have $(-2)\cdot(21n+4)+3\cdot(14n+3) = 1$. Hence, Proposition 3.8.1 **(b)** (applied to $a = 21n+4$, $x = -2$, $b = 14n+3$ and $y = 3$) yields $21n+4 \perp 14n+3$. This solves Exercise 3.8.1 again. $\square$

This second solution to Exercise 3.8.1 is slick, but how could we have found it? The answer turns out to be "pretty easily, once we made up our mind to apply Proposition 3.8.1 **(b)**". Indeed, in order to prove $21n+4 \perp 14n+3$ using Proposition 3.8.1 **(b)**, it is necessary to find two integers x and y that satisfy $x\cdot(21n+4)+y\cdot(14n+3) = 1$. In theory, these x and y could depend on n, but the first thing you should try are constants. So you are looking for two (constant) integers x and y that satisfy $x\cdot(21n+4)+y\cdot(14n+3) = 1$ for each $n \in \mathbb{Z}$. But this boils down to a system of infinitely many linear equations in x and y (one for each n); any two of them determine x and y uniquely. Solving the system thus yields $x = -2$ and $y = 3$, which is exactly the two numbers used in the solution above.

Here is another little exercise:

Exercise 3.8.2. You have a corridor with 1000 lamps, which are initially all off. Each lamp has a light switch controlling its state.

Every night, a ghost glides through the corridor (always in the same direction) and flips some of the switches:

On the 1st night, the ghost flips every switch.

On the 2nd night, the ghost flips switches $2,4,6,8,10,\ldots$.

On the 3rd night, the ghost flips switches $3,6,9,12,15,\ldots$.

etc.

(That is: For each $k \in \{1,2,\ldots,1000\}$, the ghost spends the k-th night flipping switches $k, 2k, 3k, \ldots$.)

Which lamps will be on after 1000 nights?

Discussion of Exercise 3.8.2. Let us first make the problem more manageable by focussing on a single lamp. So let us fix some $m \in \{1,2,\ldots,1000\}$ and find out whether lamp m is on after 1000 nights.

Indeed, switch m gets flipped on the k-th night if and only if m is a multiple of k. In other words, switch m gets flipped on the k-th night if and only if k is a divisor of m. Thus, the number of times that switch m gets flipped (during the entire 1000 nights) is precisely the number of positive divisors of m. Of course, lamp m will be on after 1000 nights if and only if this number of times is odd. Thus, asking which lamps will be on after 1000 nights is equivalent to asking which of the numbers $1,2,\ldots,1000$ have an odd number of positive divisors.

Experiments reveal that among the first 10 positive integers, only three have an odd number of positive divisors: namely, 1, 4 and 9. (For example, 9 has the 3 positive divisors 1, 3 and 9.) This suggests the following: □

Exercise 3.8.3. Let n be a positive integer. Prove that the number of positive divisors of n is even if and only if n is not a perfect square.

Solution to Exercise 3.8.3 (sketched). A detailed proof can be found in [19s, Proposition 2.14.7], so we shall keep to the main idea. We will use the word "*posdiv*" as shorthand for "positive divisor". If d is a posdiv of n, then $\frac{n}{d}$ is an integer and again a posdiv of n. We shall refer to $\frac{n}{d}$ as the *complement* of d. Note that being the complement is a symmetric relation: If d is a posdiv of n, and if e is the complement of d, then d is in turn the complement of e.

We can thus pair up each posdiv of n with its complement. This results in a "pairing" that covers all posdivs of n, except that it might fail to be a proper pairing: Namely, if a posdiv of n is its own complement, then this posdiv will be paired with itself.

Example 3.8.2. Let us see how this "pairing" looks like:

- If $n = 12$, then the posdivs of n are $1, 2, 3, 4, 6, 12$. Their complements are $12, 6, 4, 3, 2, 1$, respectively. Thus, our "pairing" pairs 1 with 12, pairs 2 with 6, and pairs 3 with 4.
- If $n = 16$, then the posdivs of n are $1, 2, 4, 8, 16$. Their complements are $16, 8, 4, 2, 1$, respectively. Thus, our "pairing" pairs 1 with 16, pairs 2 with 8, and pairs 4 with itself.

When does n have a posdiv that gets paired with itself? In other words, when is there a posdiv d of n that is its own complement? Clearly, this means that $d = \frac{n}{d}$, or, equivalently, $n = d^2$. If n is a perfect square, then there is exactly one posdiv d that satisfies $n = d^2$ (namely, $\sqrt{n}$); otherwise, there is no such d. Thus:

- If n is a perfect square, then exactly one posdiv of n gets paired with itself.
- If n is not a perfect square, then no posdiv of n gets paired with itself.

Hence:

- If n is a perfect square, then our "pairing" of the posdivs of n leaves exactly one posdiv of n paired with itself, while all others have proper partners. Therefore, in this case, the number of posdivs of n is odd.

- If n is not a perfect square, then our "pairing" of the posdivs of n is a proper pairing (i.e., no posdiv gets paired with itself). Therefore, in this case, the number of posdivs of n is even.

Combining these facts, we conclude that the number of posdivs of n is even if and only if n is not a perfect square. This solves Exercise 3.8.3. □

Exercise 3.8.3 quickly yields an answer to Exercise 3.8.2: Namely, the lamps that will be on after 1000 nights are precisely the m-th lamps where $m \in \{1, 2, \ldots, 1000\}$ is a perfect square – i.e., the 1^2-th, 2^2-th, 3^2-th etc. lamps.

Exercise 3.8.3 is due to John Wallis (1685, [Wallis85, Additional Treatises, Chapter III, §16]).

Next, we shall discuss some variations on Bezout's theorem. As a warmup exercise, here is an easy one:

Exercise 3.8.4. Let a and b be two coprime integers. Let n be an integer.

Prove that there exist integers x and y such that $n = xa + yb$.

Solution to Exercise 3.8.4. The numbers a and b are coprime; in other words, $\gcd(a, b) = 1$.

But Theorem 3.4.5 yields that there exist integers $x \in \mathbb{Z}$ and $y \in \mathbb{Z}$ such that $\gcd(a, b) = xa + yb$. Consider these x and y, and denote them by x' and y'. (We do not want to call them x and y, since they are not the x and y we are looking for.) Thus, x' and y' are integers satisfying $\gcd(a, b) = x'a + y'b$. Hence, $x'a + y'b = \gcd(a, b) = 1$. Multiplying both sides of this equality by n, we find $n(x'a + y'b) = n \cdot 1 = n$. Hence, $n = n(x'a + y'b) = nx'a + ny'b$. Thus, there exist integers x and y such that $n = xa + yb$ (namely, $x = nx'$ and $y = ny'$). This solves Exercise 3.8.4. □

Now, what happens if we require x and y to be **nonnegative** in Exercise 3.8.4? Then, Exercise 3.8.4 is no longer valid. For example, if $a = 3$ and $b = 5$, then there are no nonnegative integers x and y such that $4 = xa + yb$. This is, in a way, due to 4 being "too small"; one is thus tempted to ask whether requiring n to be large enough (for a and b positive) will force those nonnegative x and y to exist. If so, then how large is large enough?

The answer is given by a famous result of J. J. Sylvester (who found it in 1882 when studying invariant theory). We state it as an exercise:

Exercise 3.8.5. Let a and b be two coprime positive integers. Let n be an integer such that $n > ab - a - b$.

Prove that there exist **nonnegative** integers x and y such that $n = xa + yb$.

The "$n > ab - a - b$" bound in Exercise 3.8.5 is sharp: The claim of the exercise would no longer hold if we allowed $n = ab - a - b$. (See Exercise 4.5.2 **(b)** for why.)

Exercise 3.8.5 is often stated in terms of coins: If a and b are two coprime positive integers, and if you have unlimited supplies of a-cent coins and of b-cent coins, but no other coins, then Exercise 3.8.5 says that you can pay any amount that is larger

than $ab - a - b$ cents[57] without asking for change. See Wikipedia on the Sylvester Coin Problem, which also studies the generalization to more than two sorts of coins. (This generalization is significantly harder.)[58]

Discussion of Exercise 3.8.5. Exercise 3.8.4 yields that there exist integers x and y such that $n = xa + yb$. Consider these x and y, and denote them by x' and y'. (We do not want to call them x and y, since they are not the x and y we are looking for.) Thus, x' and y' are integers satisfying $n = x'a + y'b$.

If x' and y' are nonnegative, then we are already done (since we can just take $x = x'$ and $y = y'$). But x' and y' might not be nonnegative yet. The trick is now to modify x' and y' in such a way that they become nonnegative but the sum $x'a + y'b$ is unchanged. (If you want, you can view this as an instance of the "find a not-quite-answer and tweak it until it becomes a full answer" technique; the time-honored *regula falsi* is another instance of this technique.)

How can we modify x' and y' in a way that $x'a + y'b$ is unchanged? A simple way to do so is to add b to x' and subtract a from y' (because this causes $x'a + y'b$ to become $(x' + b)\,a + (y' - a)\,b$, which is still the same as $x'a + y'b$). More generally, we can pick any $d \in \mathbb{Z}$ and add db to x' and subtract da from y'. Obviously, if we pick d large enough, then $x' + db$ will become nonnegative, whereas a small enough d will make $y' - da$ nonnegative. The question is: Is there a (nonempty) "goldilocks zone" in which d is sufficiently large for $x' + db$ to be nonnegative and yet sufficiently small for $y' - da$ to be nonnegative?

We can try to describe this "goldilocks zone" explicitly. The number $x' + db$ is nonnegative if and only if $d \geq \dfrac{-x'}{b}$ (check this!); on the other hand, the number $y' - da$ is nonnegative if and only if $d \leq \dfrac{y'}{a}$. Hence, the "goldilocks zone" for d is the interval $\left[\dfrac{-x'}{b}, \dfrac{y'}{a}\right]$. Let us check whether this interval is nonempty. We have

$$\begin{aligned} \frac{y'}{a} - \frac{-x'}{b} &= \frac{x'a + y'b}{ab} > \frac{ab - a - b}{ab} \qquad \left(\text{since } x'a + y'b = n > ab - a - b\right) \\ &= 1 - \frac{1}{a} - \frac{1}{b}, \end{aligned}$$

which is ≥ 0 in all interesting cases (the only exception being when a or b is 1, but this case is easy to handle separately). Thus, in all interesting cases, we have $\dfrac{y'}{a} \geq \dfrac{-x'}{b}$.

Unfortunately, this is not the whole story. We are not looking for a **real** number d in the interval $\left[\dfrac{-x'}{b}, \dfrac{y'}{a}\right]$, but for an **integer** d in this interval. Having $\dfrac{y'}{a} \geq \dfrac{-x'}{b}$

[57] obviously an integer amount of cents

[58] See also `https://artofproblemsolving.com/community/c6h373` for further discussions on what amounts can be paid using two kinds of coins.

guarantees the former but not the latter. How can we guarantee the latter? What must the difference $\beta - \alpha$ of two real numbers α and β satisfy to ensure that the interval $[\alpha, \beta]$ contains an integer?

It is not hard to see that the answer to this question is "they must satisfy $\beta - \alpha \geq 1$". Unfortunately, this is not true for our "goldilocks zone"; the difference $\dfrac{y'}{a} - \dfrac{-x'}{b}$ is always < 1. So we are stuck.

Being stuck, let us revisit what we did, in the hopes of finding something improvable. Did we perhaps throw away some information too early, or make too weak an estimate? One thing that stands out is that we never really used the fact that x' and y' are integers.

This turns out to be the key. We said that the number $x' + db$ is nonnegative if and only if $d \geq \dfrac{-x'}{b}$. But we can say something slightly stronger: The number $x' + db$ is nonnegative if and only if $d > \dfrac{-x'-1}{b}$. Indeed, we have the following chain of equivalences:

$$\begin{aligned}
(x' + db \text{ is nonnegative}) &\iff (x' + db \geq 0) \\
&\iff (x' + db > -1) \qquad (\text{since } x' + db \text{ is an integer}) \\
&\iff (db > -x' - 1) \iff \left(d > \frac{-x'-1}{b}\right).
\end{aligned}$$

Notice how rewriting "$x' + db \geq 0$" as "$x' + db > -1$" has improved our inequality! Likewise, we can see that $y' - da$ is nonnegative if and only if $d < \dfrac{y'+1}{a}$. Thus, our "goldilocks zone" $\left[\dfrac{-x'}{b}, \dfrac{y'}{a}\right]$ has grown to a wider interval $\left(\dfrac{-x'-1}{b}, \dfrac{y'+1}{a}\right)$ (an open interval, not a pair of numbers). And as it turns out, this latter interval has size > 1, since

$$\begin{aligned}
\frac{y'+1}{a} - \frac{-x'-1}{b} &= \frac{x'a + y'b + a + b}{ab} = \frac{n + a + b}{ab} \qquad (\text{since } x'a + y'b = n) \\
&> 1 \qquad (\text{since } n + a + b > ab \text{ (because } n > ab - a - b)).
\end{aligned}$$

But an open interval of size > 1 must always contain an integer[59]. Hence, the open interval $\left(\dfrac{-x'-1}{b}, \dfrac{y'+1}{a}\right)$ contains an integer d. Picking such a d, we then conclude that both $x' + db$ and $y' - da$ are nonnegative, and thus there exist **nonnegative** integers x and y such that $n = xa + yb$ (namely, $x = x' + db$ and $y = y' - da$). Thus, we have found a solution to Exercise 3.8.5. $\square$

[59] *Proof.* Let (α, β) be an open interval of size > 1. Thus, α and β are real numbers such that $\beta - \alpha > 1$. We must prove that the interval (α, β) contains an integer.

Indeed, it is easy to see that $\lfloor \alpha \rfloor + 1$ is an integer contained in this interval (α, β). Check this!

It is instructive to see how short this solution becomes if we forget about the pains we took finding it, and only write down the parts that ended up necessary for the proof:

Solution to Exercise 3.8.5 (final copy). Exercise 3.8.4 yields that there exist integers x and y such that $n = xa + yb$. Consider these x and y, and denote them by x' and y'. Thus, x' and y' are integers satisfying $n = x'a + y'b$.

Define an integer d by $d = \left\lfloor \frac{-x'-1}{b} \right\rfloor + 1$. Now, we claim the following:

Claim 1: Both integers $x' + db$ and $y' - da$ are nonnegative.

[*Proof of Claim 1:* The chain of inequalities (1) (applied to $x = \frac{-x'-1}{b}$) shows that

$$\left\lfloor \frac{-x'-1}{b} \right\rfloor \leq \frac{-x'-1}{b} < \left\lfloor \frac{-x'-1}{b} \right\rfloor + 1. \tag{56}$$

Hence,

$$\frac{-x'-1}{b} < \left\lfloor \frac{-x'-1}{b} \right\rfloor + 1 = d \qquad \left(\text{since } d = \left\lfloor \frac{-x'-1}{b} \right\rfloor + 1\right).$$

We can multiply this inequality by b (since b is positive), and thus obtain $-x' - 1 < db$. Hence, $db > -x' - 1$, so that $x' + db > -1$. Since $x' + db$ is an integer, this entails that $x' + db \geq 0$. In other words, $x' + db$ is nonnegative.

It remains to show that $y' - da$ is nonnegative. But $n + a + b > ab$ (because $n > ab - a - b$). We can divide this inequality by ab (since $ab > 0$), and thus obtain $\frac{n+a+b}{ab} > 1$. Now,

$$\begin{aligned} \frac{y'+1}{a} - \frac{-x'-1}{b} &= \frac{x'a + y'b + a + b}{ab} = \frac{n+a+b}{ab} \qquad (\text{since } x'a + y'b = n) \\ &> 1, \end{aligned}$$

so that

$$\frac{y'+1}{a} > \underbrace{\frac{-x'-1}{b}}_{\substack{\geq \left\lfloor \frac{-x'-1}{b} \right\rfloor \\ \text{(by the first inequality in (56))}}} + 1 \geq \left\lfloor \frac{-x'-1}{b} \right\rfloor + 1 = d.$$

We can multiply this inequality by a (since a is positive), and find $y' + 1 > da$. In other words, $y' - da > -1$. Since $y' - da$ is an integer, this entails $y' - da \geq 0$. In other words, $y' - da$ is nonnegative. This completes our proof of Claim 1.]

Thus, the integers $x' + db$ and $y' - da$ are nonnegative. They furthermore satisfy

$$\left(x' + db\right) a + \left(y' - da\right) b = x'a + dba + y'b - dab = x'a + y'b = n,$$

so that $n = (x' + db)\, a + (y' - da)\, b$. Thus, there exist **nonnegative** integers x and y such that $n = xa + yb$ (namely, $x = x' + db$ and $y = y' - da$). This solves Exercise 3.8.5. $\square$

4. Sequences and sums

In this chapter, we shall explore sequences and finite sums and products. This is a vast subject and one that is popular on mathematical competitions; my choice of material is "informed" more by my own tastes than by relevance for current olympiads. In particular, I will mostly avoid any analysis (limits and infinite sums); good sources for this are [GelAnd17, §3.1] and much of [PolSze78]. Serious number-theoretical questions will have to wait for future chapters. In this chapter, I will begin with some rules and tricks for dealing with finite sums and products; then I will introduce the binomial coefficients and their most fundamental properties (nothing too combinatorial yet). After that, I will turn to some more specific kinds of sequences: ones that are integer (i.e., their entries are integers) although their definitions involve fractions; ones that are defined by recursions yet can be explicitly computed; periodic ones; and ones satisfying linear recurrences.

More diverse selections of problems on sequences can be found in [Engel98, Chapter 9] and [GelAnd17, §3.1]. It is worth keeping in mind that there is an *Online Encyclopedia of Integer Sequences* that knows hundreds of thousands of integer sequences and can be used to answer questions such as "what meaningful sequences begin with $1, 0, 1, 2, 9, 44, 265$, and what facts are known about them?" (check it out if you are curious). It is a useful tool for mathematical researchers, not just for puzzle solvers, as integer sequences appear frequently in mathematics, and knowing that two questions lead to one and the same integer sequence often foreshadows some deeper connection between the questions. There is also a *Journal of Integer Sequences*.

4.1. Finite sums

We begin with finite sums and products. While we have already used these concepts and the corresponding notations (the $\sum$ and $\prod$ symbols), let us nevertheless quickly recall some of their basic properties. We refer to [Grinbe15, §1.4] for a much more thorough (although probably not very exciting) survey of the properties of finite sums and products with detailed proofs, and to [GrKnPa94, Chapter 2] for a hands-on introduction to the art of manipulating them. Also, [AndTet18] is an entire book devoted to (mostly finite) sums and products.

This section will be a long yet (most likely) fast read, as much of it just will be spent stating rules and definitions that you have likely already encountered in your mathematical life history. Furthermore, almost all of these rules are just formalizing common sense about sums and products, so you won't be surprised by them even if you see them for the first time. The rules will be illustrated by exercises, and some of the latter might actually be surprising. Thus, skim this section, but do not skip the exercises!

4.1.1. The $\sum$ sign

The most well-known type of finite sums are those of the form $\sum\limits_{i=u}^{v} a_i$ (for example, $\sum\limits_{i=2}^{10} i^2$). But let us first define finite sums of a more general shape (from which we shall later obtain the $\sum\limits_{i=u}^{v} a_i$ kind as a special case):[60]

Definition 4.1.1. If S is a finite set, and if a_s is a number for each $s \in S$, then $\sum\limits_{s\in S} a_s$ denotes the sum of all of these numbers a_s. Formally, this sum is defined by recursion on $|S|$, as follows:

- If $|S| = 0$, then $\sum\limits_{s\in S} a_s$ is defined to be 0. (In this case, $\sum\limits_{s\in S} a_s$ is called an *empty sum*.)
- Let $n \in \mathbb{N}$. Assume that we have defined $\sum\limits_{s\in S} a_s$ for every finite set S with $|S| = n$ (and every choice of numbers a_s). Now, if S is a finite set with $|S| = n+1$ (and if a number a_s is chosen for each $s \in S$), then $\sum\limits_{s\in S} a_s$ is defined by picking any $t \in S$ and setting

$$\sum_{s\in S} a_s = a_t + \sum_{s\in S\setminus\{t\}} a_s. \tag{57}$$

It is not immediately clear why this definition is legitimate; in fact, the right hand side of (57) is defined using a choice of t, but theoretically one can imagine that different choices of t would lead to different results. Nevertheless, one can prove (see, e.g., [Grinbe15, Theorem 2.118] for this proof) that this definition is indeed legitimate (i.e., the right hand side of (57) does not depend on t). This is essentially saying that if we sum the numbers a_s for all $s \in S$ by starting with 0 and adding these numbers one by one, then the result does not depend on the order in which the numbers are being added. (Note that this is not true of multiplying matrices, or of putting on clothes, or of chemical reactions. So it's not an obvious claim!)

Expressions of the form $\sum\limits_{s\in S} a_s$ for finite sets S are called *finite sums*. Here are some examples:

[60] Throughout Section 4.1, "number" means an integer or a rational number or a real number or a complex number. More generally, the concept of a finite sum can be applied to any sort of object that has an addition operation with reasonable properties (such as commutativity and associativity); e.g., in linear algebra, one defines finite sums of vectors in a vector space, and in abstract algebra, one generalizes this further to finite sums of elements of an abelian group (or even monoid).

Example 4.1.2. **(a)** If $S = \{2,6,9\}$ and $a_s = 3^s$ for each $s \in S$, then $\sum\limits_{s\in S} a_s = a_2 + a_6 + a_9 = 3^2 + 3^6 + 3^9$.

(b) If $S = \{1,2,\ldots,n\}$ (for some $n \in \mathbb{N}$) and $a_s = s^2$ for every $s \in S$, then $\sum\limits_{s\in S} a_s = \sum\limits_{s\in S} s^2 = 1^2 + 2^2 + \cdots + n^2$.

(c) If $S = \varnothing$, then $\sum\limits_{s\in S} a_s = 0$ (since $|S| = 0$).

The sum $\sum\limits_{s\in S} a_s$ is usually pronounced "sum of the a_s over all $s \in S$" or "sum of the a_s with s ranging over S" or "sum of the a_s with s running through all elements of S". The letter "s" in the sum is called the "summation index", and its exact choice is immaterial (for example, you can rewrite $\sum\limits_{s\in S} a_s$ as $\sum\limits_{t\in S} a_t$ or as $\sum\limits_{\Phi\in S} a_\Phi$ or as $\sum\limits_{\spadesuit\in S} a_\spadesuit$), as long as it does not already have a different meaning outside of the sum. The sign $\sum$ itself is called "the summation sign" or "the $\sum$ sign". The numbers a_s are called the *addends* (or *summands* or *terms*) of the sum $\sum\limits_{s\in S} a_s$. More precisely, for any given $t \in S$, we can refer to the number a_t as the "addend corresponding to the index t" (or as the "addend for $s = t$", or as the "addend for t") of the sum $\sum\limits_{s\in S} a_s$.

The summation index in a finite sum does not always have to be a single letter. For instance, if S is a set of pairs, then we can write $\sum\limits_{(x,y)\in S} a_{(x,y)}$ (meaning the same as $\sum\limits_{s\in S} a_s$). Here is an example of this notation:

$$\sum_{(x,y)\in\{1,2,3\}^2} \frac{x}{y} = \frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \frac{2}{1} + \frac{2}{2} + \frac{2}{3} + \frac{3}{1} + \frac{3}{2} + \frac{3}{3}$$

(here, we are using the notation $\sum\limits_{(x,y)\in S} a_{(x,y)}$ with $S = \{1,2,3\}^2$ and $a_{(x,y)} = \dfrac{x}{y}$).

Warning 4.1.3. There is no agreement on the operator precedence of the $\sum$ sign versus the $+$ sign. By this I mean the following question: Does $\sum\limits_{s\in S} a_s + b$ (where b is some other number) mean $\sum\limits_{s\in S} (a_s + b)$ or $\left(\sum\limits_{s\in S} a_s\right) + b$? I will use the second interpretation (i.e., I will read it as $\left(\sum\limits_{s\in S} a_s\right) + b$), but I have seen both used in the literature (although I believe that the second is more popular).

However, the $\cdot$ sign definitely has higher precedence than the $\sum$ sign. That is, an expression of the form $\sum\limits_{s\in S} ba_sc$ is always understood to mean $\sum\limits_{s\in S} (ba_sc)$.

As I mentioned above, the most common finite sums are of a special kind:

Definition 4.1.4. Let u and v be integers. We agree to understand the set $\{u, u+1, \ldots, v\}$ to be empty if $u > v$.

Let a_s be a number for each $s \in \{u, u+1, \ldots, v\}$. Then, the finite sum $\sum\limits_{s \in \{u, u+1, \ldots, v\}} a_s$ will also be denoted by $\sum\limits_{s=u}^{v} a_s$ or by $a_u + a_{u+1} + \cdots + a_v$.

Thus, finite sums of the form $\sum\limits_{s=u}^{v} a_s$ satisfy

$$\sum_{s=u}^{v} a_s = 0 \qquad \text{for } v < u, \text{ and} \tag{58}$$

$$\sum_{s=u}^{v} a_s = a_u + \sum_{s=u+1}^{v} a_s = \sum_{s=u}^{v-1} a_s + a_v \qquad \text{for } v \geq u. \tag{59}$$

The equality (58) says that a sum that "ends before it begins" is 0. For example, $\sum\limits_{i=5}^{2} i = 0$. This is not completely uncontroversial[61], but it shall be so in this course.

In a sum of the form $\sum\limits_{s=u}^{v} a_s$, the integer u is called the *lower limit* (or *lower bound*) of the sum, whereas the integer v is called the *upper limit* (or *upper bound*) of the sum.[62] The sum is said to *start* (or *begin*) at u and *end* at v.

Another way to use $\sum$ signs is in describing sums that sum over all elements of a given set that satisfy a given statement. This is done as follows:

Definition 4.1.5. Let S be a finite set, and let $\mathcal{A}(s)$ be a logical statement defined for every $s \in S$. (For example, if S is a set of integers, then $\mathcal{A}(s)$ can be the statement "s is even".) For each $s \in S$ satisfying $\mathcal{A}(s)$, let a_s be a number. Then, the sum $\sum\limits_{\substack{s \in S; \\ \mathcal{A}(s)}} a_s$ is defined by

$$\sum_{\substack{s \in S; \\ \mathcal{A}(s)}} a_s = \sum_{s \in \{t \in S \mid \mathcal{A}(t)\}} a_s.$$

In other words, $\sum\limits_{\substack{s \in S; \\ \mathcal{A}(s)}} a_s$ is the sum of the a_s for all $s \in S$ which satisfy $\mathcal{A}(s)$.

[61] I have seen authors who would "sum backwards" when $v < u$, thus interpreting $\sum\limits_{i=5}^{2} i$ as $-4-3$. This has some advantages (because why let a notation go to waste?), but it would mean that $\sum\limits_{s=u}^{v} a_s$ cannot be rewritten as $\sum\limits_{s \in \{u, u+1, \ldots, v\}} a_s$ without checking the order relation between u and v, and I don't want this extra headache.

[62] This has nothing to do with the notions of "bounds" and "limits" in analysis.

Example 4.1.6. If $S = \{1,2,3,4,5\}$, then $\sum\limits_{\substack{s\in S;\\ s \text{ is even}}} a_s = a_2 + a_4$ and $\sum\limits_{\substack{s\in S;\\ s \text{ is odd}}} a_s = a_1 + a_3 + a_5$.

4.1.2. The simplest rules: factoring out and splitting addends

Having defined the notations, let us recall some rules for manipulating finite sums. We shall give no proofs (see [Grinbe15, §1.4] for them, or treat them as induction exercises), but we shall illustrate them with examples whenever reasonable.

We begin with four very simple rules. The first one ([Grinbe15, (6)]) says that if all addends in a sum are equal to some number a, then the sum equals the number of addends times a:

Theorem 4.1.7. Let S be a finite set. Let a be a number. Then,

$$\sum_{s\in S} a = |S| \cdot a. \tag{60}$$

This is trivial to prove by induction and needs no further explanation. Applying (60) to $a = 0$ yields $\sum\limits_{s\in S} 0 = |S| \cdot 0 = 0$.

The next rule ([Grinbe15, (9)]) is just a distributive law for finite sums; it says that a factor common to all addends of a sum can be factored out:

Theorem 4.1.8. Let S be a finite set. For every $s \in S$, let a_s be a number. Also, let λ be a number. Then,

$$\sum_{s\in S} \lambda a_s = \lambda \sum_{s\in S} a_s. \tag{61}$$

Example 4.1.9. Let u and v be two integers. Let a_s be a number for each $s \in \{u, u+1, \ldots, v\}$. Also, let λ be a number. Then, Theorem 4.1.8 (applied to $S = \{u, u+1, \ldots, v\}$) yields

$$\sum_{s=u}^{v} \lambda a_s = \lambda \sum_{s=u}^{v} a_s. \tag{62}$$

Equivalently,

$$\lambda a_u + \lambda a_{u+1} + \cdots + \lambda a_v = \lambda \left(a_u + a_{u+1} + \cdots + a_v\right). \tag{63}$$

For example, if $n \in \mathbb{N}$, then the sum of the first n even positive integers is

$$\begin{aligned}
2 + 4 + \cdots + 2n &= 2\cdot 1 + 2\cdot 2 + \cdots + 2\cdot n \\
&= 2\left(1 + 2 + \cdots + n\right) \qquad \left(\begin{array}{c}\text{by (63), applied to } \lambda = 2 \\ \text{and } u = 1 \text{ and } v = n \text{ and } a_s = s\end{array}\right) \\
&= 2\cdot \frac{n(n+1)}{2} \qquad \text{(by (9))} \\
&= n(n+1).
\end{aligned}$$

The next rule ([Grinbe15, (7)]) says that if each addend in a sum is itself a sum of two numbers, then we can split the sum up:

Theorem 4.1.10. Let S be a finite set. For every $s \in S$, let a_s and b_s be numbers. Then,

$$\sum_{s \in S} (a_s + b_s) = \sum_{s \in S} a_s + \sum_{s \in S} b_s. \tag{64}$$

Example 4.1.11. Let $(f_0, f_1, f_2, \ldots)$ be the Fibonacci sequence. Let $n \in \mathbb{N}$. Then,

$$\begin{aligned}
\sum_{s=1}^{n} (f_s + 1) &= \underbrace{\sum_{s=1}^{n} f_s}_{\substack{=f_1+f_2+\cdots+f_n \\ =f_{n+2}-1 \\ \text{(by Exercise 2.2.1)}}} + \underbrace{\sum_{s=1}^{n} 1}_{\substack{=n\cdot 1 \\ \text{(as a consequence of (60))}}} \\
&\qquad \text{(by (64), applied to } S = \{1, 2, \ldots, n\} \text{ and } a_s = f_s \text{ and } b_s = 1) \\
&= f_{n+2} - 1 + n \cdot 1 = f_{n+2} - 1 + n.
\end{aligned}$$

The equality (64) is often used forwards (i.e., in order to split a sum of the form $\sum_{s \in S} (a_s + b_s)$ into $\sum_{s \in S} a_s + \sum_{s \in S} b_s$) and backwards (i.e., in order to combine two sums $\sum_{s \in S} a_s + \sum_{s \in S} b_s$ into $\sum_{s \in S} (a_s + b_s)$). When applying it backwards, don't forget to check that the two sums being combined are ranging over the same set!

An easy consequence of Theorem 4.1.10 is the following:

Exercise 4.1.1. Let S be a finite set. For every $s \in S$, let a_s and b_s be numbers. Then,

$$\sum_{s \in S} (a_s - b_s) = \sum_{s \in S} a_s - \sum_{s \in S} b_s.$$

Solution to Exercise 4.1.1. Theorem 4.1.10 (applied to $a_s - b_s$ instead of a_s) yields

$$\sum_{s \in S} ((a_s - b_s) + b_s) = \sum_{s \in S} (a_s - b_s) + \sum_{s \in S} b_s.$$

Solving this equation for $\sum_{s \in S} (a_s - b_s)$, we find

$$\sum_{s \in S} (a_s - b_s) = \sum_{s \in S} \underbrace{((a_s - b_s) + b_s)}_{=a_s} - \sum_{s \in S} b_s = \sum_{s \in S} a_s - \sum_{s \in S} b_s.$$

This solves Exercise 4.1.1. □

4.1.3. Substitution of the summation index

The next rule ([Grinbe15, (12)]) is intuitively simple, yet highly useful:

Theorem 4.1.12. Let S and T be two finite sets. Let $f : S \to T$ be a **bijective** map[63]. Let a_t be a number for each $t \in T$. Then,

$$\sum_{t \in T} a_t = \sum_{s \in S} a_{f(s)}. \tag{65}$$

Roughly speaking, (65) holds because the sum $\sum_{s \in S} a_{f(s)}$ contains the same addends as the sum $\sum_{t \in T} a_t$. We say that the sum $\sum_{s \in S} a_{f(s)}$ is obtained from the sum $\sum_{t \in T} a_t$ by *substituting $f(s)$ for t*. (Conversely, $\sum_{t \in T} a_t$ is obtained from $\sum_{s \in S} a_{f(s)}$ by *substituting t for $f(s)$*.)

Example 4.1.13. For any $n \in \mathbb{N}$, we have

$$\sum_{t \in \{1,2,\ldots,n\}} t^3 = \sum_{s \in \{-n,-n+1,\ldots,-1\}} (-s)^3.$$

Indeed, this follows from (65), applied to $S = \{-n, -n+1, \ldots, -1\}$, $T = \{1, 2, \ldots, n\}$, $f(s) = -s$, and $a_t = t^3$.

When substituting the index in a sum, it is perfectly okay to re-use the same letter for the new index. Thus, (65) can be rewritten as

$$\sum_{s \in T} a_s = \sum_{s \in S} a_{f(s)}.$$

Theorem 4.1.12 has several well-known and oft-used consequences. The first one says that we can shift the index of a sum by any integer k:

Corollary 4.1.14. Let u, v and k be integers. Let a_t be a number for each $t \in \{u+k, u+k+1, \ldots, v+k\}$. Then,

$$\sum_{t=u+k}^{v+k} a_t = \sum_{s=u}^{v} a_{s+k}.$$

Proof of Corollary 4.1.14. Let f denote the map

$$\begin{aligned} \{u, u+1, \ldots, v\} &\to \{u+k, u+k+1, \ldots, v+k\}, \\ s &\mapsto s+k. \end{aligned}$$

[63] Recall: A map is said to be *bijective* if it is injective (i.e., one-to-one) and surjective (i.e., onto). Bijective maps are also known as *bijections* or *one-to-one correspondences*. See, e.g., the detailed Wikipedia page for more about these kinds of maps.

This map f is easily seen to be well-defined and bijective[64]. Thus, Theorem 4.1.12 (applied to $S = \{u, u+1, \ldots, v\}$ and $T = \{u+k, u+k+1, \ldots, v+k\}$) yields

$$\sum_{t \in \{u+k, u+k+1, \ldots, v+k\}} a_t = \sum_{s \in \{u, u+1, \ldots, v\}} \underbrace{a_{f(s)}}_{\substack{= a_{s+k} \\ \text{(since } f(s) = s+k)}} = \sum_{s \in \{u, u+1, \ldots, v\}} a_{s+k}.$$

In other words,

$$\sum_{t=u+k}^{v+k} a_t = \sum_{s=u}^{v} a_{s+k}.$$

This proves Corollary 4.1.14. □

The next corollary says that sums can be turned around: i.e., we have

$$a_u + a_{u+1} + \cdots + a_v = a_v + a_{v-1} + \cdots + a_u. \tag{66}$$

Restated using $\sum$ signs, and generalized by shifting the index, this takes the following form:

Corollary 4.1.15. Let u, v and k be integers. Let a_t be a number for each $t \in \{k-v, k-v+1, \ldots, k-u\}$. Then,

$$\sum_{t=k-v}^{k-u} a_t = \sum_{s=u}^{v} a_{k-s}. \tag{67}$$

Proof of Corollary 4.1.15. Let f denote the map

$$\begin{aligned} \{u, u+1, \ldots, v\} &\to \{k-v, k-v+1, \ldots, k-u\}, \\ s &\mapsto k-s. \end{aligned}$$

This map f is easily seen to be well-defined and bijective[65]. Thus, Theorem 4.1.12 (applied to $S = \{u, u+1, \ldots, v\}$ and $T = \{k-v, k-v+1, \ldots, k-u\}$) yields

$$\sum_{t \in \{k-v, k-v+1, \ldots, k-u\}} a_t = \sum_{s \in \{u, u+1, \ldots, v\}} \underbrace{a_{f(s)}}_{\substack{= a_{k-s} \\ \text{(since } f(s) = k-s)}} = \sum_{s \in \{u, u+1, \ldots, v\}} a_{k-s}.$$

In other words,

$$\sum_{t=k-v}^{k-u} a_t = \sum_{s=u}^{v} a_{k-s}.$$

This proves Corollary 4.1.15. □

Here is an example for the use of substitution in sums. Recall the formula (9), which is easily proved by induction. We shall now prove it in a more elegant way:

[64] Indeed, it has an inverse map, which sends each $t \in \{u+k, u+k+1, \ldots, v+k\}$ to $t-k$.
[65] Indeed, it has an inverse map, which sends each $t \in \{k-v, k-v+1, \ldots, k-u\}$ to $k-t$.

Exercise 4.1.2. Let $n \in \mathbb{N}$. Prove the equality (9) without using induction.

Solution to Exercise 4.1.2. First of all, we notice that

$$\sum_{i=0}^{n} i = 0 + \sum_{i=1}^{n} i = \sum_{i=1}^{n} i = 1 + 2 + \cdots + n. \tag{68}$$

Thus, we shall focus on computing $\sum\limits_{i=0}^{n} i$.

We can substitute $n - i$ for i in the sum $\sum\limits_{i=0}^{n} i$ (since the map $\{0, 1, \ldots, n\} \to \{0, 1, \ldots, n\}$, $i \mapsto n - i$ is a bijection). Thus, we obtain

$$\sum_{i=0}^{n} i = \sum_{i=0}^{n} (n - i).$$

(If you are too lazy to think about bijections, you can also derive this immediately by applying Corollary 4.1.15 to $u = 0$, $v = n$, $k = n$ and $a_t = t$. This is the main purpose of Corollary 4.1.15: avoid the use of bijections.)

Now comes Gauss's "doubling trick": Recall that $2q = q + q$ for every $q \in \mathbb{Q}$. Hence,

$$\begin{aligned}
2\sum_{i=0}^{n} i &= \sum_{i=0}^{n} i + \underbrace{\sum_{i=0}^{n} i}_{=\sum\limits_{i=0}^{n}(n-i)} = \sum_{i=0}^{n} i + \sum_{i=0}^{n} (n - i) \\
&= \sum_{i=0}^{n} \underbrace{(i + (n - i))}_{=n} \qquad \text{(here, we have used (64) backwards)} \\
&= \sum_{i=0}^{n} n = (n + 1)\, n \qquad \text{(by (60))} \\
&= n\,(n + 1), \qquad (69)
\end{aligned}$$

and therefore

$$\sum_{i=0}^{n} i = \frac{n\,(n + 1)}{2}. \tag{70}$$

In view of (68), this rewrites as $1 + 2 + \cdots + n = \dfrac{n\,(n + 1)}{2}$, which is precisely the equality (9). Hence, Exercise 4.1.2 is solved.

Let us remark how the computation (69) could be rewritten without the use of $\sum$

signs:

$$\begin{aligned}
&2\cdot(0+1+\cdots+n)\\
&=(0+1+\cdots+n)+(0+1+\cdots+n)\\
&=(0+1+\cdots+n)+(n+(n-1)+\cdots+0)\\
&\qquad\qquad\text{(here, we have turned the second sum around using (66))}\\
&=\underbrace{(0+n)}_{=n}+\underbrace{(1+(n-1))}_{=n}+\cdots+\underbrace{(n+0)}_{=n}\\
&\qquad\qquad\text{(here, we have used (64) backwards)}\\
&=\underbrace{n+n+\cdots+n}_{n+1\text{ times}}=(n+1)\,n=n\,(n+1)\,.
\end{aligned}$$

[66] This arguably looks simpler than (69), but it is clear that with more complicated sums it will become progressively harder to avoid the use of $\sum$ signs. □

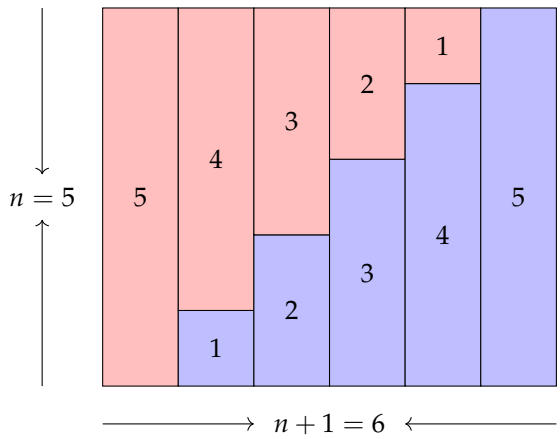
Gauss's "doubling trick" used in the above solution is not a one-trick pony. Here is a more advanced use:

Exercise 4.1.3. Let $n\in\mathbb{N}$. Let d be an odd positive integer. Prove that

$$1+2+\cdots+n\mid 1^d+2^d+\cdots+n^d.$$

Solution to Exercise 4.1.3 (sketched). This is solved in detail in [19s, Exercise 2.10.8]; thus, we only give the skeleton of the argument. In view of (9), the claim that we

[66] We can also visualize this computation as a "picture proof". For example, here is the picture for the case $n=5$:



.

The addends of the sum $0+1+\cdots+n$ are drawn as purple rectangles, whereas the addends of the sum $n+(n-1)+\cdots+0$ are drawn as red rectangles. (Rectangles of area 0, corresponding to 0 addends, are degenerate and thus invisible.) The rectangles are matched in such a way that the i-th column (counted from left, started with 0) has a purple rectangle of area i and a red rectangle of area $n-i$. This matching corresponds precisely to the way we matched our addends: $\underbrace{(0+n)}_{=n}+\underbrace{(1+(n-1))}_{=n}+\cdots+\underbrace{(n+0)}_{=n}$.

must prove rewrites as

$$\frac{n(n+1)}{2} \mid 1^d + 2^d + \cdots + n^d.$$

This is equivalent to

$$n(n+1) \mid 2\left(1^d + 2^d + \cdots + n^d\right) \tag{71}$$

(by Proposition 3.1.5). Hence, it suffices to prove (71).

In order to prove (71), it suffices to show that

$$n \mid 2\left(1^d + 2^d + \cdots + n^d\right) \qquad \text{and} \tag{72}$$

$$n+1 \mid 2\left(1^d + 2^d + \cdots + n^d\right). \tag{73}$$

Indeed, Example 3.5.2 **(a)** (applied to $a = n$) yields that n is coprime to $n+1$; in other words, $n \perp n+1$. Hence, if we can prove (72) and (73), then we will get (71) by applying Theorem 3.5.7.

We shall prove (73) first:

[*Proof of (73):* We have

$$\begin{aligned}
2\left(1^d + 2^d + \cdots + n^d\right) &= \left(1^d + 2^d + \cdots + n^d\right) + \left(1^d + 2^d + \cdots + n^d\right) \\
&= \left(1^d + 2^d + \cdots + n^d\right) + \left(n^d + (n-1)^d + \cdots + 1^d\right) \\
&= \sum_{k=1}^{n} k^d + \sum_{k=1}^{n} \underbrace{(n+1-k)^d}_{\substack{\equiv (-k)^d \bmod n+1 \\ \text{(by taking the congruence } n+1-k \equiv -k \bmod n+1 \\ \text{to the } d\text{-th power)}}} \\
&\equiv \sum_{k=1}^{n} k^d + \sum_{k=1}^{n} (-k)^d = \sum_{k=1}^{n} \underbrace{\left(k^d + (-k)^d\right)}_{\substack{=0 \\ \text{(since } d \text{ is odd, so that } (-k)^d = -k^d)}} \\
&= \sum_{k=1}^{n} 0 = 0 \bmod n+1.
\end{aligned}$$

In other words, $n+1 \mid 2\left(1^d + 2^d + \cdots + n^d\right)$. Thus, (73) is proven.]

[*Proof of (72):* If $n = 0$, then (72) boils down to $0 \mid 2 \cdot 0$, which is obvious. Thus, for the rest of this proof, we WLOG assume that $n \neq 0$. Hence, $n - 1 \in \mathbb{N}$. Therefore, we can apply (73) to $n-1$ instead of n (since we have already proven (73) for each $n \in \mathbb{N}$). We thus obtain

$$n \mid 2\left(1^d + 2^d + \cdots + (n-1)^d\right).$$

In other words, $2\left(1^d+2^d+\cdots+(n-1)^d\right) \equiv 0 \bmod n$. Also, $d>0$ (since d is odd), so that $n \mid n^d$ and thus $n^d \equiv 0 \bmod n$. Now,

$$\begin{aligned}2\left(1^d+2^d+\cdots+n^d\right) &= 2\left(1^d+2^d+\cdots+(n-1)^d+\underbrace{n^d}_{\equiv 0 \bmod n}\right)\\ &\equiv 2\left(1^d+2^d+\cdots+(n-1)^d\right) \equiv 0 \bmod n.\end{aligned}$$

That is, $n \mid 2\left(1^d+2^d+\cdots+n^d\right)$. This proves (72).]

We have now proven both (72) and (73). As we have explained, this yields (71), which in turn solves Exercise 4.1.3. □

4.1.4. The telescope principle

The next summation rule is the so-called *telescope principle* ([Grinbe15, (16)]), which has a number of simple yet elegant applications:

Theorem 4.1.16. Let u and v be two integers such that $u-1 \leq v$. Let a_s be a number for each $s \in \{u-1, u, \ldots, v\}$. Then,

$$\sum_{s=u}^{v}\left(a_s-a_{s-1}\right)=a_v-a_{u-1}. \tag{74}$$

Intuitively, the claim of Theorem 4.1.16 is obvious: The left hand side of (74) is

$$\begin{aligned}\sum_{s=u}^{v}\underbrace{\left(a_s-a_{s-1}\right)}_{=-a_{s-1}+a_s} &= \sum_{s=u}^{v}\left(-a_{s-1}+a_s\right)\\ &= \left(-a_{u-1}+a_u\right)+\left(-a_u+a_{u+1}\right)+\left(-a_{u+1}+a_{u+2}\right)+\cdots+\left(-a_{v-1}+a_v\right).\end{aligned}$$

If we expand the right hand side of this equality, then it "contracts like a telescope" (thus the name of Theorem 4.1.16): All addends except for the $-a_{u-1}$ and the a_v cancel each other, and you are left with $-a_{u-1}+a_v=a_v-a_{u-1}$, which is exactly the right hand side of (74). (Be careful with this argument, though: It does not work for $v=u-1$. But this case is trivial anyway.) This intuitive argument can be formalized. Alternatively, Theorem 4.1.16 can be proved by induction on v (the proof is utterly straightforward), or derived from Exercise 4.1.1.

Note that Theorem 4.1.16 can be regarded as a discrete version of the Second Part of the Fundamental Theorem of Calculus. In fact, the latter fact says that $\int_u^v F'(x)\,dx = F(v)-F(u)$ for any differentiable function F on a (real) interval $[u,v]$. But the $\sum$ sign is a discrete analogue of the $\int$ sign, whereas the consecutive differences a_s-a_{s-1} are discrete analogues of the values of the derivative F'.

For all its simplicity, Theorem 4.1.16 is a surprisingly helpful tool for simplifying sums. The next four exercises are examples of this:

Exercise 4.1.4. Let $n \in \mathbb{N}$. Simplify the sum $\sum\limits_{i=1}^{n} i \cdot i!$ (that is, rewrite it without using the $\sum$ sign).

Solution to Exercise 4.1.4. The trick is to realize that each $i \in \mathbb{N}$ satisfies

$$i \cdot i! = (i+1)! - i! \tag{75}$$

(because $(i+1)! = (i+1) \cdot i! = i \cdot i! + i!$). Thus,

$$\begin{aligned}
\sum_{i=1}^{n} \underbrace{i \cdot i!}_{\substack{=(i+1)!-i! \\ \text{(by (75))}}} &= \sum_{i=1}^{n} ((i+1)! - i!) = \sum_{s=2}^{n+1} \left(\left(\underbrace{(s-1)+1}_{=s} \right)! - (s-1)! \right) \\
&\qquad \text{(here, we have substituted } s \text{ for } i+1 \text{ in the sum)} \\
&= \sum_{s=2}^{n+1} (s! - (s-1)!) = (n+1)! - \underbrace{(2-1)!}_{=1!=1} \\
&\qquad \left(\begin{array}{c} \text{by Theorem 4.1.16, applied} \\ \text{to } u = 2,\ v = n+1 \text{ and } a_s = s! \end{array} \right) \\
&= (n+1)! - 1. \qquad (76)
\end{aligned}$$

This solves Exercise 4.1.4. □

Before we move on to the next example, let us observe what we have gained from the telescope principle. This equality (76), once it has been found, can easily be proved by induction on n (see [19s-hw0s, solution to Exercise 2 **(b)**] for such a proof); but the telescope principle has helped us find this equality in the first place (once we had the fancy to rewrite $i \cdot i!$ as $(i+1)! - i!$). Thus, the telescope principle turns the (often difficult) question of simplifying a sum into an (often simpler) question of rewriting its addends as differences.[67] Of course, it is not magic: Any proof using the telescope principle can easily be rewritten as an induction proof, because the telescope principle itself is easily proved by induction. The main advantage of the principle is its convenience.

In future proofs, we shall be less detailed than in our above solution to Exercise 4.1.4, and simply say "by the telescope principle" instead of specifying what we are applying Theorem 4.1.16 to.

The next exercise is about generalizing the formula (29):

Exercise 4.1.5. Let p be a positive integer, and let $n \in \mathbb{N}$. Simplify the sum

$$\sum_{i=1}^{n} \frac{1}{i\,(i+1)\,(i+2) \cdots (i+p)}.$$

[67] This is similar to computing an integral by finding an antiderivative of the function under the integral sign.

Solution to Exercise 4.1.5. We define

$$a_i := \frac{1}{(i+1)(i+2)\cdots(i+p)}$$

for each $i \in \mathbb{N}$. Then, it is easy to see that

$$\frac{1}{i(i+1)(i+2)\cdots(i+p)} = \frac{a_{i-1}}{p} - \frac{a_i}{p} \tag{77}$$

for any positive integer i.

Indeed, if i is a positive integer, then

$$\begin{aligned}
&\underbrace{\frac{1}{i(i+1)\cdots(i+p-1)}}_{\substack{= a_{i-1} \\ \text{(by the definition of } a_{i-1})}} - \underbrace{\frac{1}{(i+1)(i+2)\cdots(i+p)}}_{\substack{= a_i \\ \text{(by the definition of } a_i)}} \\
&= \underbrace{\frac{1}{i(i+1)\cdots(i+p-1)}}_{= \frac{i+p}{i(i+1)\cdots(i+p-1)\cdot(i+p)}} - \underbrace{\frac{1}{(i+1)(i+2)\cdots(i+p)}}_{= \frac{i}{i\cdot(i+1)(i+2)\cdots(i+p)}} \\
&= \frac{i+p}{i(i+1)\cdots(i+p-1)\cdot(i+p)} - \frac{i}{i\cdot(i+1)(i+2)\cdots(i+p)} \\
&= \frac{i+p}{i(i+1)(i+2)\cdots(i+p)} - \frac{i}{i(i+1)(i+2)\cdots(i+p)} \\
&= \frac{(i+p)-i}{i(i+1)(i+2)\cdots(i+p)} = \frac{p}{i(i+1)(i+2)\cdots(i+p)}
\end{aligned} \tag{78}$$

and thus

$$\begin{aligned}
\frac{a_{i-1}}{p} - \frac{a_i}{p} &= \frac{1}{p}(a_{i-1} - a_i) = \frac{1}{p} \cdot \frac{p}{i(i+1)(i+2)\cdots(i+p)} \qquad \text{(by (78))} \\
&= \frac{1}{i(i+1)(i+2)\cdots(i+p)};
\end{aligned}$$

this proves (77).

Now,

$$\begin{aligned}
&\sum_{i=1}^{n} \underbrace{\frac{1}{i\,(i+1)\,(i+2)\cdots(i+p)}}_{\substack{=\frac{a_{i-1}}{p}-\frac{a_i}{p} \\ \text{(by (77))}}} \\
&= \sum_{i=1}^{n} \underbrace{\left(\frac{a_{i-1}}{p} - \frac{a_i}{p}\right)}_{=\frac{-a_i}{p}-\frac{-a_{i-1}}{p}} = \sum_{i=1}^{n} \left(\frac{-a_i}{p} - \frac{-a_{i-1}}{p}\right) \\
&= \frac{-a_n}{p} - \frac{-a_{1-1}}{p} \qquad \text{(by the telescope principle)} \\
&= \frac{1}{p}\left(\underbrace{a_{1-1}}_{=a_0} - a_n\right) = \frac{1}{p}\left(\underbrace{a_0}_{\substack{=\frac{1}{1\cdot 2\cdot\cdots\cdot p} \\ \text{(by the definition of } a_0)}} - \underbrace{a_n}_{\substack{=\frac{1}{(n+1)\,(n+2)\cdots(n+p)} \\ \text{(by the definition of } a_n)}}\right) \\
&= \frac{1}{p}\left(\frac{1}{1\cdot 2\cdot\cdots\cdot p} - \frac{1}{(n+1)\,(n+2)\cdots(n+p)}\right). \qquad (79)
\end{aligned}$$

This solves the exercise. □

Note that Exercise 4.1.5 has no good answer for $p = 0$. Indeed, the sum $\sum_{i=1}^{n} \frac{1}{i}$ cannot be simplified. (It is known as the *n-th harmonic number*.)

Note also that, as an easy consequence of (79), we have

$$\sum_{i=1}^{\infty} \frac{1}{i\,(i+1)\,(i+2)\cdots(i+p)} = \frac{1}{p} \cdot \frac{1}{1\cdot 2\cdot\cdots\cdot p} \qquad (80)$$

for any positive integer p. This also "holds" for $p = 0$, in the sense that the infinite series $\sum_{i=1}^{\infty} \frac{1}{i}$ (known as the *harmonic series*) diverges (which is precisely what one would expect from seeing the $\frac{1}{p} = \frac{1}{0}$ term on the right hand side of (80)). This is a well-known fact, with several proofs on the Wikipedia.

To facilitate its future application, let us restate Theorem 4.1.16 in a form that will allow us to apply it without substituting indices:

Corollary 4.1.17. Let u and v be two integers such that $u-1 \leq v$. Let a_i be a number for each $i \in \{u, u+1, \ldots, v+1\}$. Then,

$$\sum_{i=u}^{v} (a_{i+1} - a_i) = a_{v+1} - a_u.$$

Proof of Corollary 4.1.17. We have $(u+1)-1 = u \leq v+1$ (since $u-1 \leq v$). Also,

$$\begin{aligned}
\sum_{i=u}^{v} (a_{i+1} - a_i) &= \sum_{s=u+1}^{v+1} \left(\underbrace{a_{(s-1)+1}}_{=a_s} - a_{s-1} \right) \\
&\qquad \left(\begin{array}{c} \text{here, we have substituted } s-1 \text{ for } i \text{ in the sum} \\ \text{(i.e., formally speaking, applied Corollary 4.1.14)} \end{array} \right) \\
&= \sum_{s=u+1}^{v+1} (a_s - a_{s-1}) = a_{v+1} - \underbrace{a_{(u+1)-1}}_{=a_u} \\
&\qquad \left(\begin{array}{c} \text{by Theorem 4.1.16,} \\ \text{applied to } u+1 \text{ and } v+1 \text{ instead of } u \text{ and } v \end{array} \right) \\
&= a_{v+1} - a_u.
\end{aligned}$$

This proves Corollary 4.1.17. □

Corollary 4.1.17 is just Theorem 4.1.16 restated with a shifted index; thus, we shall still refer to it as the "telescope principle".

Exercise 4.1.6. Let $n \in \mathbb{N}$. Simplify the sum $\sum_{i=1}^{n} \frac{1}{\sqrt{i} + \sqrt{i+1}}$.

Solution to Exercise 4.1.6. If x and y are two distinct positive reals, then

$$\left(\sqrt{x} - \sqrt{y}\right)\left(\sqrt{x} + \sqrt{y}\right) = \left(\sqrt{x}\right)^2 - \left(\sqrt{y}\right)^2 = x - y$$

and thus

$$\frac{1}{\sqrt{x} + \sqrt{y}} = \frac{\sqrt{x} - \sqrt{y}}{x - y}. \tag{81}$$

This formula might be well-known from high school (where it is used to rationalize denominators).

Now, for each positive real i, we have

$$\begin{aligned}
\frac{1}{\sqrt{i} + \sqrt{i+1}} &= \frac{\sqrt{i} - \sqrt{i+1}}{i - (i+1)} && \text{(by (81))} \\
&= \frac{\sqrt{i} - \sqrt{i+1}}{-1} && \text{(since } i - (i+1) = -1\text{)} \\
&= \sqrt{i+1} - \sqrt{i}.
\end{aligned}$$

Hence,

$$\sum_{i=1}^{n} \underbrace{\frac{1}{\sqrt{i}+\sqrt{i+1}}}_{=\sqrt{i+1}-\sqrt{i}} = \sum_{i=1}^{n} \left(\sqrt{i+1}-\sqrt{i}\right) = \sqrt{n+1}-\sqrt{1}$$

(by the telescope principle[68]). This simplifies further to $\sqrt{n+1}-1$. □

Finally, here is what might be the simplest application of the telescope principle; we state it mainly because of its usefulness:

Exercise 4.1.7. Let a and b be any numbers. Let $m \in \mathbb{N}$. Then,

$$(a-b)\sum_{i=0}^{m-1} a^i b^{m-1-i} = a^m - b^m. \tag{82}$$

Solution to Exercise 4.1.7. From (61), we obtain

$$\begin{aligned}
(a-b)\sum_{i=0}^{m-1} a^i b^{m-1-i} &= \sum_{i=0}^{m-1} \underbrace{(a-b)\,a^i b^{m-1-i}}_{=aa^ib^{m-1-i}-ba^ib^{m-1-i}} = \sum_{i=0}^{m-1}\left(\underbrace{aa^ib^{m-1-i}}_{=a^{i+1}b^{m-(i+1)}} - \underbrace{ba^ib^{m-1-i}}_{=a^ib^{m-i}}\right) \\
&= \sum_{i=0}^{m-1}\left(a^{i+1}b^{m-(i+1)} - a^ib^{m-i}\right) \\
&= \underbrace{a^{(m-1)+1}}_{=a^m}\underbrace{b^{m-((m-1)+1)}}_{=b^0=1} - \underbrace{a^0}_{=1}\underbrace{b^{m-0}}_{=b^m} \\
&\qquad \left(\begin{array}{c}\text{by the telescope principle (Corollary 4.1.17,} \\ \text{applied to } u=0,\ v=m-1 \text{ and } a_i = a^ib^{m-i})\end{array}\right) \\
&= a^m - b^m.
\end{aligned}$$

□

We notice that Exercise 4.1.7 generalizes Exercise 2.1.1 and also yields a rather explicit new proof of Proposition 3.2.7 (check it!).

4.1.5. Splitting a sum into two

Another general rule for sums is the following ([Grinbe15, (3)]):

Theorem 4.1.18. Let S be a finite set. Let X and Y be two subsets of S such that $X \cap Y = \varnothing$ and $X \cup Y = S$. (Equivalently, X and Y are two subsets of S such that

[68] specifically, by Corollary 4.1.17, applied to $u = 1$, $v = n$ and $a_i = \sqrt{i}$

each element of S lies in **exactly** one of X and Y.) Let a_s be a number for each $s \in S$. Then,

$$\sum_{s \in S} a_s = \sum_{s \in X} a_s + \sum_{s \in Y} a_s. \tag{83}$$

Here, as we explained, $\sum_{s \in X} a_s + \sum_{s \in Y} a_s$ stands for $\left(\sum_{s \in X} a_s \right) + \left(\sum_{s \in Y} a_s \right)$.

Behind Theorem 4.1.18 stands the intuitively obvious fact that a sum $\sum_{s \in S} a_s$ can be computed by first sorting its addends into two "heaps" $\sum_{s \in X} a_s$ and $\sum_{s \in Y} a_s$, then summing each heap separately, and finally adding the two heap sums together. For example, this says that

$$a_1 + a_2 + a_3 + a_4 + a_5 = (a_1 + a_4) + (a_2 + a_3 + a_5).$$

(This is the particular case of (83) for $S = \{1, 2, 3, 4, 5\}$, $X = \{1, 4\}$ and $Y = \{2, 3, 5\}$.)

Most commonly, Theorem 4.1.18 is used to split a sum of the form $\sum_{s=u}^{w} a_s$ into two parts:

Corollary 4.1.19. Let u, v and w be three integers such that $u - 1 \leq v \leq w$. Let a_s be a number for each $s \in \{u, u+1, \ldots, w\}$. Then,

$$\sum_{s=u}^{w} a_s = \sum_{s=u}^{v} a_s + \sum_{s=v+1}^{w} a_s. \tag{84}$$

Proof of Corollary 4.1.19. Recall the assumption $u - 1 \leq v \leq w$. Thus, the sets $\{u, u+1, \ldots, v\}$ and $\{v+1, v+2, \ldots, w\}$ are subsets of the set $\{u, u+1, \ldots, w\}$ and satisfy

$$\{u, u+1, \ldots, v\} \cap \{v+1, v+2, \ldots, w\} = \varnothing \qquad \text{and}$$
$$\{u, u+1, \ldots, v\} \cup \{v+1, v+2, \ldots, w\} = \{u, u+1, \ldots, w\}.$$

Thus, Theorem 4.1.18 (applied to $S = \{u, u+1, \ldots, w\}$ and $X = \{u, u+1, \ldots, v\}$ and $Y = \{v+1, v+2, \ldots, w\}$) yields

$$\sum_{s \in \{u, u+1, \ldots, w\}} a_s = \sum_{s \in \{u, u+1, \ldots, v\}} a_s + \sum_{s \in \{v+1, v+2, \ldots, w\}} a_s.$$

But this rewrites as

$$\sum_{s=u}^{w} a_s = \sum_{s=u}^{v} a_s + \sum_{s=v+1}^{w} a_s.$$

Hence, Corollary 4.1.19 is proved. □

Here are two examples of sums getting split:

Exercise 4.1.8. Let $n \in \mathbb{N}$.

(a) Compute $\sum\limits_{s=-n}^{n} |s|$.

(b) Compute $\sum\limits_{s=0}^{2n-1} (s\%2)$.

Solution to Exercise 4.1.8 (sketched). **(a)** Corollary 4.1.19 (applied to $u = -n$, $v = 0$, $w = n$ and $a_s = |s|$) yields

$$\sum_{s=-n}^{n} |s| = \sum_{s=-n}^{0} \underbrace{|s|}_{\substack{=-s \\ \text{(since } s\leq 0)}} + \sum_{s=1}^{n} \underbrace{|s|}_{\substack{=s \\ \text{(since } s\geq 1\geq 0)}} = \underbrace{\sum_{s=-n}^{0} (-s)}_{= \sum\limits_{s=-n}^{-1} (-s) + (-0)} + \sum_{s=1}^{n} s$$

$$= \underbrace{\sum_{s=-n}^{-1} (-s)}_{\substack{= \sum\limits_{t=1}^{n} t \\ \text{(here, we have substituted } t \text{ for } -s \\ \text{in the sum, since the} \\ \text{map } \{1,2,\ldots,n\} \to \{-n,-n+1,\ldots,-1\},\ s \mapsto -s \\ \text{is a bijection)}}} + \underbrace{(-0)}_{=0} + \sum_{s=1}^{n} s = \sum_{t=1}^{n} t + \sum_{s=1}^{n} s$$

$$= \sum_{s=1}^{n} s + \sum_{s=1}^{n} s \qquad \left(\begin{array}{c} \text{here, we have renamed the summation index } t \\ \text{as } s \text{ in the first sum} \end{array} \right)$$

$$= 2 \cdot \sum_{s=1}^{n} s = 2 \cdot (1 + 2 + \cdots + n) = n(n+1) \qquad (\text{by } (70)).$$

(b) Set $S = \{0, 1, \ldots, 2n-1\}$. Let X be the set of all even elements of S, and let Y be the set of all odd elements of S. Then, each element of S lies in **exactly** one of X and Y (since each element of S is either even or odd, but not both). In other words, $X \cap Y = \varnothing$ and $X \cup Y = S$. Therefore, we can apply Theorem 4.1.18 to $a_s = s\%2$. We obtain

$$\sum_{s\in S} (s\%2) = \sum_{s\in X} (s\%2) + \sum_{s\in Y} (s\%2). \tag{85}$$

Let us now simplify the right hand side. If $s \in X$, then s is even (by the definition of X) and thus satisfies $s\%2 = 0$ (by Exercise 3.3.2 **(a)**). Hence, $\sum\limits_{s\in X} \underbrace{(s\%2)}_{=0} = \sum\limits_{s\in X} 0 = 0$. On the other hand, if $s \in Y$, then s is odd (by the definition of Y) and thus satisfies $s\%2 = 1$ (by Exercise 3.3.2 **(b)**). Hence, $\sum\limits_{s\in Y} \underbrace{(s\%2)}_{=1} = \sum\limits_{s\in Y} 1 = |Y| \cdot 1$ (by (60)). Now, what is $|Y|$? We defined Y as the set of all odd elements of $S = \{0, 1, \ldots, 2n-1\}$; hence, the elements of Y are $1, 3, 5, \ldots, 2n-1$. These are precisely n elements[69].

[69] Strictly speaking, this needs to be proved. It can be easily proved by induction on n, or using Theorem 3.1.8.

Thus, $|Y| = n$. Hence, $\sum_{s \in Y} (s\%2) = |Y| \cdot 1 = |Y| = n$. Now, (85) becomes

$$\sum_{s \in S} (s\%2) = \underbrace{\sum_{s \in X} (s\%2)}_{=0} + \underbrace{\sum_{s \in Y} (s\%2)}_{=n} = 0 + n = n. \tag{86}$$

Since $S = \{0, 1, \ldots, 2n-1\}$, we can rewrite the $\sum_{s \in S}$ sign on the left hand side as $\sum_{s=0}^{2n-1}$. Thus, (86) rewrites as $\sum_{s=0}^{2n-1} (s\%2) = n$. □

The following is just a restatement of Theorem 4.1.18 using the notation from Definition 4.1.5:

Theorem 4.1.20. Let S be a finite set. For each $s \in S$, let a_s be a number, and let $\mathcal{A}(s)$ be a logical statement. Then,

$$\sum_{s \in S} a_s = \sum_{\substack{s \in S; \\ \mathcal{A}(s) \text{ is true}}} a_s + \sum_{\substack{s \in S; \\ \mathcal{A}(s) \text{ is false}}} a_s = \sum_{\substack{s \in S; \\ \mathcal{A}(s)}} a_s + \sum_{\substack{s \in S; \\ \text{not } \mathcal{A}(s)}} a_s.$$

(The right hand side is just a shorter way to rewrite the middle hand side; there is nothing profound happening here.)

4.1.6. Splitting a sum into several

Theorem 4.1.18 has a rather natural generalization ([Grinbe15, (26)]), in which a sum is split into n instead of 2 smaller sums:

Theorem 4.1.21. Let S be a finite set. Let $S_1, S_2, \ldots, S_n$ be finitely many subsets of S. Assume that these subsets $S_1, S_2, \ldots, S_n$ are pairwise disjoint (i.e., we have $S_i \cap S_j = \varnothing$ for any two distinct elements i and j of $\{1, 2, \ldots, n\}$) and their union is S. (Thus, every element of S lies in precisely one of the subsets $S_1, S_2, \ldots, S_n$.) Let a_s be a number for each $s \in S$. Then,

$$\sum_{s \in S} a_s = \sum_{w=1}^{n} \sum_{s \in S_w} a_s \tag{87}$$

$$= \sum_{s \in S_1} a_s + \sum_{s \in S_2} a_s + \cdots + \sum_{s \in S_n} a_s. \tag{88}$$

Note that the sum on the right hand side of (87) is a *double sum* – i.e., a sum of sums.

Theorem 4.1.18 is the particular case of Theorem 4.1.21 obtained if we set $n = 2$, $S_1 = X$ and $S_2 = Y$.

As an example for the use of Theorem 4.1.21, here is a pretty obvious exercise:

Exercise 4.1.9. Let I be a finite set of integers. For each pair $(i,j) \in I \times I$, let $a_{(i,j)}$ be a number. Prove that

$$\sum_{(i,j)\in I\times I} a_{(i,j)} = \sum_{i\in I} a_{(i,i)} + \sum_{\substack{(i,j)\in I\times I;\\ i<j}} \left(a_{(i,j)} + a_{(j,i)}\right).$$

Solution to Exercise 4.1.9. Let S be the set $I \times I$. Define three subsets S_1, S_2 and S_3 of S by

$$\begin{aligned} S_1 &= \{(i,j) \in I \times I \mid i < j\}, \\ S_2 &= \{(i,j) \in I \times I \mid i = j\}, \\ S_3 &= \{(i,j) \in I \times I \mid i > j\}. \end{aligned}$$

Then, every element of S lies in precisely one of the subsets S_1, S_2, S_3. In other words, the three subsets S_1, S_2, S_3 are pairwise disjoint and their union is S. Hence, (88) (applied to $n = 3$) yields

$$\sum_{s\in S} a_s = \sum_{s\in S_1} a_s + \sum_{s\in S_2} a_s + \sum_{s\in S_3} a_s.$$

Renaming the index s as (i,j) in each sum of this equality, we obtain

$$\sum_{(i,j)\in S} a_{(i,j)} = \sum_{(i,j)\in S_1} a_{(i,j)} + \sum_{(i,j)\in S_2} a_{(i,j)} + \sum_{(i,j)\in S_3} a_{(i,j)}. \tag{89}$$

We shall now take a closer look at the sums on the right hand side.

The set $S_2 = \{(i,j) \in I \times I \mid i = j\}$ consists of the pairs (i,i) for all $i \in I$. More precisely, there is a bijection

$$\begin{aligned} I &\to S_2, \\ i &\mapsto (i,i). \end{aligned}$$

Thus, we can substitute (i,i) for (i,j) in the sum $\sum_{(i,j)\in S_2} a_{(i,j)}$. This sum therefore rewrites as follows:

$$\sum_{(i,j)\in S_2} a_{(i,j)} = \sum_{i\in I} a_{(i,i)}. \tag{90}$$

Furthermore, there is a bijection between the two sets $S_1 = \{(i,j) \in I \times I \mid i < j\}$ and $S_3 = \{(i,j) \in I \times I \mid i > j\}$. Namely, the map

$$\begin{aligned} S_1 &\to S_3, \\ (i,j) &\mapsto (j,i) \end{aligned}$$

is a bijection (because if $(i,j) \in I \times I$ satisfies $i < j$, then $j > i$, and vice versa). Therefore, we can substitute (j,i) for (i,j) in the sum $\sum_{(i,j)\in S_3} a_{(i,j)}$. This sum therefore rewrites as follows:

$$\sum_{(i,j)\in S_3} a_{(i,j)} = \sum_{(i,j)\in S_1} a_{(j,i)}. \tag{91}$$

Finally, $I \times I = S$, so that

$$\begin{aligned}
\sum_{(i,j)\in I\times I} a_{(i,j)} &= \sum_{(i,j)\in S} a_{(i,j)} = \sum_{(i,j)\in S_1} a_{(i,j)} + \underbrace{\sum_{(i,j)\in S_2} a_{(i,j)}}_{\substack{=\sum_{i\in I} a_{(i,i)} \\ \text{(by (90))}}} + \underbrace{\sum_{(i,j)\in S_3} a_{(i,j)}}_{\substack{=\sum_{(i,j)\in S_1} a_{(j,i)} \\ \text{(by (91))}}} \qquad \text{(by (89))} \\
&= \sum_{(i,j)\in S_1} a_{(i,j)} + \sum_{i\in I} a_{(i,i)} + \sum_{(i,j)\in S_1} a_{(j,i)} = \sum_{i\in I} a_{(i,i)} + \underbrace{\sum_{(i,j)\in S_1} a_{(i,j)} + \sum_{(i,j)\in S_1} a_{(j,i)}}_{\substack{=\sum_{(i,j)\in S_1}\left(a_{(i,j)}+a_{(j,i)}\right) \\ \text{(by (64))}}} \\
&= \sum_{i\in I} a_{(i,i)} + \underbrace{\sum_{(i,j)\in S_1}\left(a_{(i,j)} + a_{(j,i)}\right)}_{\substack{=\sum_{\substack{(i,j)\in I\times I; \\ i<j}}\left(a_{(i,j)}+a_{(j,i)}\right) \\ \text{(since } S_1=\{(i,j)\in I\times I \mid i<j\}\text{)}}} = \sum_{i\in I} a_{(i,i)} + \sum_{\substack{(i,j)\in I\times I; \\ i<j}}\left(a_{(i,j)} + a_{(j,i)}\right).
\end{aligned}$$

This solves Exercise 4.1.9. □

The following theorem ([Grinbe15, (22)]) provides a more elaborate way of splitting a sum:

Theorem 4.1.22. Let S be a finite set. Let W be a finite set. Let $f : S \to W$ be a map. Let a_s be a number for each $s \in S$. Then,

$$\sum_{s\in S} a_s = \sum_{w\in W} \sum_{\substack{s\in S; \\ f(s)=w}} a_s. \tag{92}$$

The idea behind the formula (92) is the following: The left hand side is the sum of all a_s for $s \in S$. The right hand side is what you get if you first subdivide the collection of numbers a_s into heaps according to the value of $f(s)$ (with one heap for each $w \in W$), then sum each heap, and then sum the resulting heap sums. That the two sides are equal is thus quite obvious, although the rigorous proof would require some bookkeeping[70].

[70] See [Grinbe15, Theorem 2.127] for a proof by induction on $|S|$.

When we apply Theorem 4.1.22 to rewrite a sum $\sum_{s\in S} a_s$ as $\sum_{w\in W} \sum_{\substack{s\in S;\\ f(s)=w}} a_s$, we say that we are *splitting* the sum $\sum_{s\in S} a_s$ *according to the value of* $f(s)$.

Let us illustrate the use of Theorem 4.1.22 by an exercise that generalizes Exercise 4.1.8 **(b)**:

Exercise 4.1.10. Let n be a positive integer. Let $m \in \mathbb{N}$. Prove that

$$\sum_{k=0}^{mn-1} (k\%n) = \frac{mn(n-1)}{2}.$$

(It is easy to see that Exercise 4.1.8 **(b)** follows from Exercise 4.1.10, applied to 2 and n instead of n and m.)

Solution to Exercise 4.1.10 (sketched). We let S be the set $\{0, 1, \ldots, mn-1\}$. Let W be the set $\{0, 1, \ldots, n-1\}$. Define a map

$$\begin{aligned} f : S &\to W, \\ k &\mapsto k\%n. \end{aligned}$$

(This is well-defined, since each $k \in S$ satisfies $k\%n \in \{0, 1, \ldots, n-1\} = W$.) Thus, (92) (applied to $a_s = k\%n$) yields

$$\begin{aligned} \sum_{s\in S} (s\%n) &= \sum_{w\in W} \underbrace{\sum_{\substack{s\in S;\\ f(s)=w}} (s\%n)}_{\substack{= \sum_{\substack{s\in S;\\ s\%n=w}} (s\%n) \\ \text{(since } f(s)=s\%n \text{ for each } s\in S)}} = \sum_{w\in W} \sum_{\substack{s\in S;\\ s\%n=w}} \underbrace{(s\%n)}_{=w} \\ &= \sum_{w\in W} \sum_{\substack{s\in S;\\ s\%n=w}} w. \end{aligned} \tag{93}$$

Now, let us simplify the sum $\sum_{\substack{s\in S;\\ s\%n=w}} w$ for each $w \in W$. Indeed, fix $w \in W$. The summation sign $\sum_{\substack{s\in S;\\ s\%n=w}}$ is just shorthand for $\sum_{s\in\{k\in S \mid k\%n=w\}}$ (by Definition 4.1.5). Thus, we have

$$\sum_{\substack{s\in S;\\ s\%n=w}} w = \sum_{s\in\{k\in S \mid k\%n=w\}} w = |\{k \in S \mid k\%n = w\}| \cdot w \tag{94}$$

(by (60)). Now, we need to compute $|\{k \in S \mid k\%n = w\}|$, that is, find the number of all $k \in S$ satisfying $k\%n = w$.

Recall that $S = \{0, 1, \ldots, mn-1\}$ and $w \in W = \{0, 1, \ldots, n-1\}$. Thus, the elements $k \in S$ satisfying $k\%n = w$ are the m elements

$$w, \quad n+w, \quad 2n+w, \quad 3n+w, \quad \ldots, \quad (m-1)\,n+w.$$

Hence, there are exactly m such elements. In other words,

$$|\{k \in S \mid k\%n = w\}| = m.$$

In light of this, (94) becomes

$$\sum_{\substack{s \in S; \\ s\%n = w}} w = \underbrace{|\{k \in S \mid k\%n = w\}|}_{=m} \cdot w = mw. \tag{95}$$

Forget that we fixed w. We thus have proved (95) for each $w \in W$. Now, (93) becomes

$$\begin{aligned}
\sum_{s \in S} (s\%n) &= \sum_{w \in W} \underbrace{\sum_{\substack{s \in S; \\ s\%n = w}} w}_{\substack{=mw \\ \text{(by (95))}}} = \sum_{w \in W} mw = \sum_{w=0}^{n-1} mw \qquad (\text{since } W = \{0, 1, \ldots, n-1\}) \\
&= m \underbrace{\sum_{w=0}^{n-1} w}_{\substack{=0+1+2+\cdots+(n-1) \\ =1+2+\cdots+(n-1)}} = m \frac{(n-1)\,((n-1)+1)}{2} = m \frac{(n-1)\,n}{2} \\
&= \underbrace{\frac{(n-1)\,((n-1)+1)}{2}}_{\substack{\text{(by (70), applied} \\ \text{to } n-1 \text{ instead of } n)}} \\
&= \frac{mn\,(n-1)}{2}.
\end{aligned}$$

This solves Exercise 4.1.10. □

4.1.7. Fubini's principle and interchange of summation signs

Probably the most complicated rule for summation signs (although this is not saying a lot) is *Fubini's theorem for finite sums*. We first state it in a particular case ([Grinbe15, (28)]):

Theorem 4.1.23. Let $n \in \mathbb{N}$ and $m \in \mathbb{N}$. Let $a_{(x,y)}$ be a number for each $(x, y) \in \{1, 2, \ldots, n\} \times \{1, 2, \ldots, m\}$. Then,

$$\sum_{x=1}^{n} \sum_{y=1}^{m} a_{(x,y)} = \sum_{(x,y) \in \{1,2,\ldots,n\} \times \{1,2,\ldots,m\}} a_{(x,y)} = \sum_{y=1}^{m} \sum_{x=1}^{n} a_{(x,y)}. \tag{96}$$

Intuitively, Theorem 4.1.24 says that if you have a rectangular table filled with numbers, then you can sum the entries of the table by going along rows (and then summing the row tallies), or by going along columns (and then summing the column tallies), or plainly by summing all the entries in some unspecified way. Indeed, if our table is the rectangular table

$$\boxed{\begin{matrix} a_{(1,1)} & a_{(1,2)} & \cdots & a_{(1,m)} \\ a_{(2,1)} & a_{(2,2)} & \cdots & a_{(2,m)} \\ \vdots & \vdots & \ddots & \vdots \\ a_{(n,1)} & a_{(n,2)} & \cdots & a_{(n,m)} \end{matrix}},$$

then the sum obtained by going along rows is

$$\begin{aligned} & \left(a_{(1,1)} + a_{(1,2)} + \cdots + a_{(1,m)}\right) \\ & + \left(a_{(2,1)} + a_{(2,2)} + \cdots + a_{(2,m)}\right) \\ & + \cdots \\ & + \left(a_{(n,1)} + a_{(n,2)} + \cdots + a_{(n,m)}\right) \\ = & \sum_{x=1}^{n} \sum_{y=1}^{m} a_{(x,y)}, \end{aligned}$$

whereas the sum obtained by going along columns is

$$\begin{aligned} & \left(a_{(1,1)} + a_{(2,1)} + \cdots + a_{(n,1)}\right) \\ & + \left(a_{(1,2)} + a_{(2,2)} + \cdots + a_{(n,2)}\right) \\ & + \cdots \\ & + \left(a_{(1,m)} + a_{(2,m)} + \cdots + a_{(n,m)}\right) \\ = & \sum_{y=1}^{m} \sum_{x=1}^{n} a_{(x,y)}, \end{aligned}$$

and the sum obtained in an unspecified way is simply

$$\sum_{(x,y) \in \{1,2,\ldots,n\} \times \{1,2,\ldots,m\}} a_{(x,y)}.$$

Thus, (96) states precisely that these three sums are equal.

We can generalize Theorem 4.1.23 by replacing the sets $\{1, 2, \ldots, n\}$ and $\{1, 2, \ldots, m\}$ by any two finite sets X and Y:

Theorem 4.1.24. Let X and Y be two finite sets. Let $a_{(x,y)}$ be a number for each $(x,y) \in X \times Y$. Then,

$$\sum_{x\in X}\sum_{y\in Y} a_{(x,y)} = \sum_{(x,y)\in X\times Y} a_{(x,y)} = \sum_{y\in Y}\sum_{x\in X} a_{(x,y)}. \tag{97}$$

Theorem 4.1.24 is [Grinbe15, (27)] (and a proof can be found there). Of course, Theorem 4.1.24 can in turn be recovered from Theorem 4.1.23, because any two finite sets X and Y can be relabeled as $\{1,2,\ldots,n\}$ and $\{1,2,\ldots,m\}$ for appropriate n and m.

Theorem 4.1.24 is called *Fubini's theorem for finite sums*, due to its similarity to the (much deeper and subtler) Fubini's theorems in analysis (which allow interchanging infinite sums or integrals). Unlike the latter, it comes with no restrictions (like absolute convergence or measurability).

Theorem 4.1.24 can be used in multiple ways. When we use it to rewrite a sum $\sum\limits_{(x,y)\in X\times Y} a_{(x,y)}$ as $\sum\limits_{x\in X}\sum\limits_{y\in Y} a_{(x,y)}$, we say that we are *decomposing* the summation sign $\sum\limits_{(x,y)\in X\times Y}$. When we use it to rewrite a double sum $\sum\limits_{x\in X}\sum\limits_{y\in Y} a_{(x,y)}$ as $\sum\limits_{y\in Y}\sum\limits_{x\in X} a_{(x,y)}$, we say that we are *interchanging* the summation signs $\sum\limits_{x\in X}$ and $\sum\limits_{y\in Y}$. Note that two summation signs can only be interchanged in this way if the sets X and Y that they range over are independent of x and y; for example, we can interchange the two summation signs in the double sum $\sum\limits_{x=1}^{n}\sum\limits_{y=1}^{m}$, but we cannot interchange the two summation signs in the double sum $\sum\limits_{x=1}^{n}\sum\limits_{y=1}^{x}$. (But we shall see below how we can nevertheless transform the latter double sum.)

Here is a simple (yet quite useful) application of Theorem 4.1.23:

Exercise 4.1.11. Let $n \in \mathbb{N}$ and $m \in \mathbb{N}$. Let $u_1, u_2, \ldots, u_n$ and $v_1, v_2, \ldots, v_m$ be numbers. Prove that

$$\left(\sum_{i=1}^{n} u_i\right)\left(\sum_{j=1}^{m} v_j\right) = \sum_{(i,j)\in\{1,2,\ldots,n\}\times\{1,2,\ldots,m\}} u_i v_j.$$

Obviously, Exercise 4.1.11 is nothing else than a souped-up distributivity law, saying that a product of two finite sums can be expanded into a sum of all possible products of an addend in the former with an addend in the latter. Facts like this can be used without proof (and even without saying) on any exam, but let us give a formal proof to illustrate the use of Theorem 4.1.23:

Solution to Exercise 4.1.11. Applying (61) to $\lambda = \sum\limits_{i=1}^{n} u_i$ and $S = \{1,2,\ldots,m\}$ and

$a_s = v_s$, we find

$$\sum_{s=1}^{m} \left(\sum_{i=1}^{n} u_i \right) v_s = \left(\sum_{i=1}^{n} u_i \right) \left(\sum_{s=1}^{m} v_s \right).$$

Renaming the summation index s as j in this equality, we obtain

$$\sum_{j=1}^{m} \left(\sum_{i=1}^{n} u_i \right) v_j = \left(\sum_{i=1}^{n} u_i \right) \left(\sum_{j=1}^{m} v_j \right).$$

Hence,

$$\begin{aligned} \left(\sum_{i=1}^{n} u_i \right) \left(\sum_{j=1}^{m} v_j \right) &= \sum_{j=1}^{m} \underbrace{\left(\sum_{i=1}^{n} u_i \right) v_j}_{\substack{= v_j \left(\sum_{i=1}^{n} u_i \right) \\ = \sum_{i=1}^{n} v_j u_i \\ \text{(again by (61))}}} = \sum_{j=1}^{m} \sum_{i=1}^{n} \underbrace{v_j u_i}_{= u_i v_j} \\ &= \sum_{j=1}^{m} \sum_{i=1}^{n} u_i v_j. \end{aligned} \tag{98}$$

But Theorem 4.1.23 (applied to $a_{(x,y)} = u_x v_y$) yields

$$\sum_{x=1}^{n} \sum_{y=1}^{m} u_x v_y = \sum_{(x,y) \in \{1,2,\ldots,n\} \times \{1,2,\ldots,m\}} u_x v_y = \sum_{y=1}^{m} \sum_{x=1}^{n} u_x v_y.$$

Renaming the indices x and y as i and j in this chain of equalities, we obtain

$$\sum_{i=1}^{n} \sum_{j=1}^{m} u_i v_j = \sum_{(i,j) \in \{1,2,\ldots,n\} \times \{1,2,\ldots,m\}} u_i v_j = \sum_{j=1}^{m} \sum_{i=1}^{n} u_i v_j.$$

Hence,

$$\sum_{(i,j) \in \{1,2,\ldots,n\} \times \{1,2,\ldots,m\}} u_i v_j = \sum_{j=1}^{m} \sum_{i=1}^{n} u_i v_j = \left(\sum_{i=1}^{n} u_i \right) \left(\sum_{j=1}^{m} v_j \right)$$

(by (98)). This solves Exercise 4.1.11. □

A consequence of Exercise 4.1.11 is the "multinomial formula" for the square of a finite sum:

Exercise 4.1.12. Let $n \in \mathbb{N}$. Let $u_1, u_2, \ldots, u_n$ be numbers. Prove that

$$\left(\sum_{i=1}^{n} u_i \right)^2 = \sum_{i=1}^{n} u_i^2 + 2 \sum_{1 \le i < j \le n} u_i u_j. \tag{99}$$

Here, the summation sign "$\sum\limits_{1\leq i<j\leq n}$" is shorthand for $\sum\limits_{\substack{(i,j)\in\{1,2,\ldots,n\}^2;\\ i<j}}$. (For example, $\sum\limits_{1\leq i<j\leq 3} u_i u_j = u_1u_2 + u_1u_3 + u_2u_3$.)

Solution to Exercise 4.1.12. Exercise 4.1.11 (applied to $m = n$ and $v_j = u_j$) yields

$$\begin{aligned}
\left(\sum_{i=1}^n u_i\right)\left(\sum_{j=1}^n u_j\right) &= \sum_{(i,j)\in\{1,2,\ldots,n\}\times\{1,2,\ldots,n\}} u_i u_j \\
&= \sum_{i\in\{1,2,\ldots,n\}} \underbrace{u_i u_i}_{=u_i^2} + \sum_{\substack{(i,j)\in\{1,2,\ldots,n\}\times\{1,2,\ldots,n\};\\ i<j}} \underbrace{\left(u_i u_j + u_j u_i\right)}_{=2u_iu_j} \\
&\qquad \left(\begin{array}{c}\text{by Exercise 4.1.9, applied to } I = \{1,2,\ldots,n\} \\ \text{and } a_{(i,j)} = u_i u_j\end{array}\right) \\
&= \sum_{i\in\{1,2,\ldots,n\}} u_i^2 + \sum_{\substack{(i,j)\in\{1,2,\ldots,n\}\times\{1,2,\ldots,n\};\\ i<j}} 2u_iu_j \\
&= \sum_{i=1}^n u_i^2 + \underbrace{\sum_{1\leq i<j\leq n} 2u_iu_j}_{=2\sum\limits_{1\leq i<j\leq n} u_iu_j} \\
&\qquad \left(\begin{array}{c}\text{here, we have rewritten the summation} \\ \text{signs } \sum\limits_{i\in\{1,2,\ldots,n\}} \text{ and } \sum\limits_{\substack{(i,j)\in\{1,2,\ldots,n\}\times\{1,2,\ldots,n\};\\ i<j}} \\ \text{as } \sum\limits_{i=1}^n \text{ and } \sum\limits_{1\leq i<j\leq n}\end{array}\right) \\
&= \sum_{i=1}^n u_i^2 + 2\sum_{1\leq i<j\leq n} u_iu_j.
\end{aligned}$$

In view of

$$\begin{aligned}
\left(\sum_{i=1}^n u_i\right)\left(\sum_{j=1}^n u_j\right) &= \left(\sum_{i=1}^n u_i\right)\left(\sum_{i=1}^n u_i\right) \qquad \left(\begin{array}{c}\text{here, we have renamed the} \\ \text{index } j \text{ in the second sum as } i\end{array}\right) \\
&= \left(\sum_{i=1}^n u_i\right)^2,
\end{aligned}$$

this rewrites as

$$\left(\sum_{i=1}^n u_i\right)^2 = \sum_{i=1}^n u_i^2 + 2\sum_{1\leq i<j\leq n} u_iu_j.$$

This solves Exercise 4.1.12. □

Fubini's theorem can be generalized a little bit more ([Grinbe15, (34)]):

Theorem 4.1.25. Let X and Y be two finite sets. For every pair $(x,y) \in X \times Y$, let $\mathcal{A}(x,y)$ be a logical statement. For each $(x,y) \in X \times Y$ satisfying $\mathcal{A}(x,y)$, let $a_{(x,y)}$ be a number. Then,

$$\sum_{x \in X} \sum_{\substack{y \in Y; \\ \mathcal{A}(x,y)}} a_{(x,y)} = \sum_{\substack{(x,y) \in X \times Y; \\ \mathcal{A}(x,y)}} a_{(x,y)} = \sum_{y \in Y} \sum_{\substack{x \in X; \\ \mathcal{A}(x,y)}} a_{(x,y)}. \tag{100}$$

Theorem 4.1.25 differs from Theorem 4.1.24 in that the sums are restricted to only those $y \in Y$ or only those $(x,y) \in X \times Y$ or only those $x \in X$ that satisfy $\mathcal{A}(x,y)$. In other words, if we regard the three sides of (97) as three ways to sum the entries of a rectangular table, then the three sides of (100) are likewise understood as three ways to sum the entries of a gappy table (i.e., a rectangular table in which not every cell has an entry). This added generality is convenient, but it is not profound; indeed, Theorem 4.1.25 can easily be derived from Theorem 4.1.24 by defining $a_{(x,y)}$ to be 0 for any $(x,y) \in X \times Y$ that does **not** satisfy $\mathcal{A}(x,y)$.

Example 4.1.26. For any $n \in \mathbb{N}$ and $m \in \mathbb{N}$, we have

$$\sum_{x \in \{1,2,\ldots,n\}} \sum_{\substack{y \in \{1,2,\ldots,m\}; \\ x+y \text{ is even}}} xy = \sum_{\substack{(x,y) \in \{1,2,\ldots,n\} \times \{1,2,\ldots,m\}; \\ x+y \text{ is even}}} xy = \sum_{y \in \{1,2,\ldots,m\}} \sum_{\substack{x \in \{1,2,\ldots,n\}; \\ x+y \text{ is even}}} xy.$$

(This follows from (100), applied to $X = \{1,2,\ldots,n\}$, $Y = \{1,2,\ldots,m\}$ and $\mathcal{A}(x,y) =$ ("$x+y$ is even").)

Finally, the last two variants of Fubini's theorem we will give are analogues of Theorem 4.1.23 for triangular (instead of rectangular) tables. Let us state them and then give an application:

Theorem 4.1.27. Let $n \in \mathbb{N}$. Let $a_{(x,y)}$ be a number for each pair $(x,y) \in \{1,2,3,\ldots\}^2$ satisfying $x+y \leq n$. Then,

$$\sum_{x=1}^{n} \sum_{y=1}^{n-x} a_{(x,y)} = \sum_{\substack{(x,y) \in \{1,2,3,\ldots\}^2; \\ x+y \leq n}} a_{(x,y)} = \sum_{y=1}^{n} \sum_{x=1}^{n-y} a_{(x,y)}. \tag{101}$$

Example 4.1.28. If $n = 4$, then the formula (101) (rewritten without the use of $\sum$

signs) looks as follows:

$$\begin{aligned}
&\left(a_{(1,1)} + a_{(1,2)} + a_{(1,3)}\right) + \left(a_{(2,1)} + a_{(2,2)}\right) + a_{(3,1)} + (\text{empty sum}) \\
&= \left(\text{the sum of all entries of the table } \begin{array}{|ccc|} \hline a_{(1,1)} & a_{(1,2)} & a_{(1,3)} \\ a_{(2,1)} & a_{(2,2)} & \\ a_{(3,1)} & & \end{array}\right) \\
&= \left(a_{(1,1)} + a_{(2,1)} + a_{(3,1)}\right) + \left(a_{(1,2)} + a_{(2,2)}\right) + a_{(1,3)} + (\text{empty sum}).
\end{aligned}$$

Theorem 4.1.29. Let $n \in \mathbb{N}$. Let $a_{(x,y)}$ be a number for each $(x,y) \in \{1,2,\ldots,n\}^2$ satisfying $x \leq y$. Then,

$$\sum_{x=1}^{n} \sum_{y=x}^{n} a_{(x,y)} = \sum_{\substack{(x,y)\in\{1,2,\ldots,n\}^2; \\ x\leq y}} a_{(x,y)} = \sum_{y=1}^{n} \sum_{x=1}^{y} a_{(x,y)}. \tag{102}$$

Example 4.1.30. If $n = 3$, then the formula (102) (rewritten without the use of $\sum$ signs) looks as follows:

$$\begin{aligned}
&\left(a_{(1,1)} + a_{(1,2)} + a_{(1,3)}\right) + \left(a_{(2,2)} + a_{(2,3)}\right) + a_{(3,3)} \\
&= \left(\text{the sum of all entries of the table } \begin{array}{|ccc|} \hline a_{(1,1)} & a_{(1,2)} & a_{(1,3)} \\ & a_{(2,2)} & a_{(2,3)} \\ & & a_{(3,3)} \end{array}\right) \\
&= a_{(1,1)} + \left(a_{(1,2)} + a_{(2,2)}\right) + \left(a_{(1,3)} + a_{(2,3)} + a_{(3,3)}\right).
\end{aligned}$$

Theorem 4.1.27 and Theorem 4.1.29 are [Grinbe15, (31) and (33), respectively]. They are commonly used to interchange summation signs: i.e., to rewrite a sum of the form $\sum_{x=1}^{n} \sum_{y=1}^{n-x} a_{(x,y)}$ as $\sum_{y=1}^{n} \sum_{x=1}^{n-y} a_{(x,y)}$ or vice versa, or to rewrite a sum of the form $\sum_{x=1}^{n} \sum_{y=x}^{n} a_{(x,y)}$ as $\sum_{y=1}^{n} \sum_{x=1}^{y} a_{(x,y)}$. This kind of interchange is somewhat trickier than the straightforward interchange facilitated by Theorem 4.1.23; indeed, unlike the latter, it requires a change of the limits of summation. However, it is not hard to remember how the limits have to change: After all, the resulting double sum needs to include the same $a_{(x,y)}$ as the original one. Thus, for example, we rewrite

$\sum\limits_{x=1}^{n}\sum\limits_{y=1}^{n-x} a_{(x,y)}$ as $\sum\limits_{y=1}^{n}\sum\limits_{x=1}^{n-y} a_{(x,y)}$ because both of these double sums are really summing over all pairs (x,y) of positive integers that satisfy $x+y \leq n$.

An example of this interchange is given by the following exercise:

Exercise 4.1.13. Let x be a number distinct from 1. Simplify the sum $\sum\limits_{i=1}^{n} ix^i = 1x^1 + 2x^2 + \cdots + nx^n$.

Solution to Exercise 4.1.13. For each $m \in \mathbb{N}$, we have

$$(x-1)\left(x^0 + x^1 + \cdots + x^m\right) = x^{m+1} - 1$$

(by Exercise 2.1.1, applied to x and $m+1$ instead of b and n) and therefore

$$x^0 + x^1 + \cdots + x^m = \frac{x^{m+1}-1}{x-1} \tag{103}$$

(here, we have divided by $x-1$, which is allowed since $x \neq 1$).

Now, we proceed – informally – as follows:

$$\begin{aligned}
\sum_{i=1}^{n} i x^i &= 1x^1 + 2x^2 + 3x^3 + 4x^4 + \cdots + nx^n \\
&= \begin{array}{ccccccccc}
x^1 & + & x^2 & + & x^3 & + & \cdots & + & x^n \\
 & + & x^2 & + & x^3 & + & \cdots & + & x^n \\
 & & & + & x^3 & + & \cdots & + & x^n \\
 & & & & & \vdots & \vdots & \vdots & \vdots \\
 & & & & & & & + & x^n
\end{array} \\
&= \sum_{j=1}^{n} \underbrace{\left(x^j + x^{j+1} + \cdots + x^n\right)}_{=x^j\left(x^0+x^1+\cdots+x^{n-j}\right)} = \sum_{j=1}^{n} x^j \underbrace{\left(x^0 + x^1 + \cdots + x^{n-j}\right)}_{\substack{=\dfrac{x^{n-j+1}-1}{x-1} \\ \text{(by (103), applied to } m=n-j)}} \\
&= \sum_{j=1}^{n} x^j \frac{x^{n-j+1}-1}{x-1} = \frac{1}{x-1} \sum_{j=1}^{n} \underbrace{x^j \left(x^{n-j+1}-1\right)}_{=x^{n+1}-x^j} = \frac{1}{x-1} \underbrace{\sum_{j=1}^{n} \left(x^{n+1}-x^j\right)}_{=\sum\limits_{j=1}^{n} x^{n+1} - \sum\limits_{j=1}^{n} x^j} \\
&= \frac{1}{x-1} \left(\underbrace{\sum_{j=1}^{n} x^{n+1}}_{=nx^{n+1}} - \underbrace{\sum_{j=1}^{n} x^j}_{\substack{=x^1+x^2+\cdots+x^n \\ =x\left(x^0+x^1+\cdots+x^{n-1}\right)}} \right) \\
&= \frac{1}{x-1} \left(nx^{n+1} - x \underbrace{\left(x^0 + x^1 + \cdots + x^{n-1}\right)}_{\substack{=\dfrac{x^n-1}{x-1} \\ \text{(by (103), applied to } m=n-1)}} \right) \\
&= \frac{1}{x-1} \left(nx^{n+1} - x \cdot \frac{x^n-1}{x-1} \right).
\end{aligned}$$

Let us formalize this computation – specifically, its second and third equality

sign (since the rest is already fully rigorous). It becomes

$$\sum_{i=1}^{n} \underbrace{ix^i}_{\substack{=\sum_{j=1}^{i} x^i \\ \text{(by (60)} \\ \text{applied backwards)}}} = \sum_{i=1}^{n} \sum_{j=1}^{i} x^i = \sum_{j=1}^{n} \underbrace{\sum_{i=j}^{n} x^i}_{=x^j+x^{j+1}+\cdots+x^n}$$

$$\left(\begin{array}{c}\text{here, we have interchanged the two summation signs} \\ \text{using (102)}\end{array}\right)$$

$$= \sum_{j=1}^{n} \left(x^j + x^{j+1} + \cdots + x^n\right).$$

□

4.2. Finite products

Finite products are defined just like finite sums, but using multiplication instead of addition (and using 1 instead of 0). In other words, they are *multiplicative analogues* (i.e., analogues in which addition is replaced by multiplication) of finite sums. Let me just state the analogues of Definition 4.1.1, Definition 4.1.4 and Definition 4.1.5 for products:

Definition 4.2.1. If S is a finite set, and if a_s is a number for each $s \in S$, then $\prod_{s\in S} a_s$ denotes the product of all of these numbers a_s. Formally, this product is defined by recursion on $|S|$, as follows:

- If $|S| = 0$, then $\prod_{s\in S} a_s$ is defined to be 1. (In this case, $\prod_{s\in S} a_s$ is called an *empty product*.)
- Let $n \in \mathbb{N}$. Assume that we have defined $\prod_{s\in S} a_s$ for every finite set S with $|S| = n$ (and every choice of numbers a_s). Now, if S is a finite set with $|S| = n+1$ (and if a number a_s is chosen for each $s \in S$), then $\prod_{s\in S} a_s$ is defined by picking any $t \in S$ and setting

$$\prod_{s\in S} a_s = a_t \cdot \prod_{s\in S\setminus\{t\}} a_s. \tag{104}$$

Definition 4.2.2. Let u and v be integers. Let a_s be a number for each $s \in \{u, u+1, \ldots, v\}$. Then, the finite product $\prod_{s\in\{u,u+1,\ldots,v\}} a_s$ will also be denoted by $\prod_{s=u}^{v} a_s$ or by $a_u a_{u+1} \cdots a_v$ (or $a_u \cdot a_{u+1} \cdot \cdots \cdot a_v$).

Definition 4.2.3. Let S be a finite set, and let $\mathcal{A}(s)$ be a logical statement defined for every $s \in S$. For each $s \in S$ satisfying $\mathcal{A}(s)$, let a_s be a number. Then, the product $\prod\limits_{\substack{s \in S; \\ \mathcal{A}(s)}} a_s$ is defined by

$$\prod_{\substack{s \in S; \\ \mathcal{A}(s)}} a_s = \prod_{s \in \{t \in S \mid \mathcal{A}(t)\}} a_s.$$

In other words, $\prod\limits_{\substack{s \in S; \\ \mathcal{A}(s)}} a_s$ is the product of the a_s for all $s \in S$ which satisfy $\mathcal{A}(s)$.

Note that $a^k = \underbrace{aa \cdots a}_{k \text{ times}} = \prod\limits_{i=1}^{k} a$ for every number a and every $k \in \mathbb{N}$. Thus, in particular, $a^0 = (\text{empty product}) = 1$. This includes $0^0 = 1$.

All the notations we have introduced for finite sums exist just as well for finite products, mutatis mutandis[71]. For example, expressions of the form $\prod\limits_{s \in S} a_s$ are called *finite products*, and the terms a_s in such expressions are called their *factors* (or *terms*). There is no agreement on the operator precedence of the $\prod$ sign versus the $\cdot$ sign (i.e., expressions like $\prod\limits_{s \in S} a_s b$ are ambiguous and need to be avoided), but the $\prod$ sign definitely has higher precedence than the $+$ and $-$ signs (so an expression like $\prod\limits_{s \in S} a_s + b$ always means $\left(\prod\limits_{s \in S} a_s\right) + b$).

In all reasonable regards, finite products (of numbers) behave just like finite sums, as long as the obvious changes are made (addition replaced by multiplication, subtraction by division, 0 by 1, products by powers, etc.). Let me state (without comment) the multiplicative analogues of Theorem 4.1.7, Theorem 4.1.8, Theorem 4.1.10, Theorem 4.1.1, Theorem 4.1.12, Theorem 4.1.16, Theorem 4.1.18, Corollary 4.1.19, Theorem 4.1.21, Theorem 4.1.22 and Theorem 4.1.24, respectively:[72]

Theorem 4.2.4. Let S be a finite set. Let a be a number. Then,

$$\prod_{s \in S} a = a^{|S|}. \tag{105}$$

Theorem 4.2.5. Let S be a finite set. For every $s \in S$, let a_s be a number. Also, let $\lambda \in \mathbb{N}$. Then,

$$\prod_{s \in S} a_s^{\lambda} = \left(\prod_{s \in S} a_s\right)^{\lambda}. \tag{106}$$

[71] "Mutatis mutandis" means "once the things that need to be changed have been changed" (i.e., "once the necessary changes are made").

[72] The proofs for all these analogues are analogous to the proofs of the original theorems for sums.

This also holds for $\lambda \in \mathbb{Z}$ if all the numbers a_s are nonzero; and this also holds for $\lambda \in \mathbb{R}$ if all the numbers a_s are positive reals.

Theorem 4.2.6. Let S be a finite set. For every $s \in S$, let a_s and b_s be numbers. Then,

$$\prod_{s \in S} (a_s b_s) = \left(\prod_{s \in S} a_s \right) \left(\prod_{s \in S} b_s \right). \tag{107}$$

Exercise 4.2.1. Let S be a finite set. For every $s \in S$, let a_s and b_s be numbers such that $b_s \neq 0$. Then,

$$\prod_{s \in S} \frac{a_s}{b_s} = \left(\prod_{s \in S} a_s \right) / \left(\prod_{s \in S} b_s \right).$$

Theorem 4.2.7. Let S and T be two finite sets. Let $f : S \to T$ be a **bijective** map. Let a_t be a number for each $t \in T$. Then,

$$\prod_{t \in T} a_t = \prod_{s \in S} a_{f(s)}. \tag{108}$$

Theorem 4.2.8. Let u and v be two integers such that $u - 1 \leq v$. Let a_s be a number for each $s \in \{u-1, u, \ldots, v\}$. Assume that $a_{u-1}, a_u, \ldots, a_{v-1}$ are nonzero. Then,

$$\prod_{s=u}^{v} \frac{a_s}{a_{s-1}} = \frac{a_v}{a_{u-1}}. \tag{109}$$

Theorem 4.2.9. Let S be a finite set. Let X and Y be two subsets of S such that $X \cap Y = \varnothing$ and $X \cup Y = S$. (Equivalently, X and Y are two subsets of S such that each element of S lies in **exactly** one of X and Y.) Let a_s be a number for each $s \in S$. Then,

$$\prod_{s \in S} a_s = \left(\prod_{s \in X} a_s \right) \left(\prod_{s \in Y} a_s \right). \tag{110}$$

Corollary 4.2.10. Let u, v and w be three integers such that $u - 1 \leq v \leq w$. Let a_s be a number for each $s \in \{u, u+1, \ldots, w\}$. Then,

$$\prod_{s=u}^{w} a_s = \left(\prod_{s=u}^{v} a_s \right) \left(\prod_{s=v+1}^{w} a_s \right). \tag{111}$$

Theorem 4.2.11. Let S be a finite set. Let $S_1, S_2, \ldots, S_n$ be finitely many subsets of S. Assume that these subsets $S_1, S_2, \ldots, S_n$ are pairwise disjoint (i.e., we have $S_i \cap S_j = \varnothing$ for any two distinct elements i and j of $\{1, 2, \ldots, n\}$) and their union is S. (Thus, every element of S lies in precisely one of the subsets $S_1, S_2, \ldots, S_n$.) Let a_s be a number for each $s \in S$. Then,

$$\prod_{s \in S} a_s = \prod_{w=1}^{n} \prod_{s \in S_w} a_s \tag{112}$$

$$= \left(\prod_{s \in S_1} a_s \right) \left(\prod_{s \in S_2} a_s \right) \cdots \left(\prod_{s \in S_n} a_s \right). \tag{113}$$

Theorem 4.2.12. Let S be a finite set. Let W be a finite set. Let $f : S \to W$ be a map. Let a_s be a number for each $s \in S$. Then,

$$\prod_{s \in S} a_s = \prod_{w \in W} \prod_{\substack{s \in S; \\ f(s) = w}} a_s. \tag{114}$$

Theorem 4.2.13. Let X and Y be two finite sets. Let $a_{(x,y)}$ be a number for each $(x, y) \in X \times Y$. Then,

$$\prod_{x \in X} \prod_{y \in Y} a_{(x,y)} = \prod_{(x,y) \in X \times Y} a_{(x,y)} = \prod_{y \in Y} \prod_{x \in X} a_{(x,y)}. \tag{115}$$

Using common sense, the reader can easily figure out how all these theorems are named (for example, Theorem 4.2.8 is known as the *telescope principle for products*, or the *multiplicative telescope principle*), and also state multiplicative analogues for the remaining rules for finite sums (e.g., for Theorem 4.1.25 and for Theorem 4.1.27).

The most famous finite products are the factorials

$$n! = 1 \cdot 2 \cdot \cdots \cdot n = \prod_{i=1}^{n} i \qquad \text{for all } n \in \mathbb{N}.$$

(This includes $0! = (\text{empty product}) = 1$.)

While there are numerous good exercises on finite products (see [AndTet18] for a large collection), we shall here only show two. The first is a token application of the telescope principle for products:

Exercise 4.2.2. Let n be a positive integer. Simplify $\prod_{s=2}^{n} \left(1 - \frac{1}{s}\right)$.

Solution to Exercise 4.2.2. For each integer $s \geq 2$, we have

$$1 - \frac{1}{s} = \frac{s-1}{s} = \frac{1/s}{1/(s-1)}.$$

Hence,

$$\begin{aligned}
\prod_{s=2}^{n} \underbrace{\left(1 - \frac{1}{s}\right)}_{=\frac{1/s}{1/(s-1)}} &= \prod_{s=2}^{n} \frac{1/s}{1/(s-1)} = \frac{1/n}{1/(2-1)} \\
&\qquad \text{(by (109), applied to } u = 2,\ v = n \text{ and } a_s = 1/s) \\
&= \frac{1}{n}.
\end{aligned}$$

□

The next exercise is a simple formula that is useful in combinatorics (more on its use perhaps later on):

Exercise 4.2.3. Let $n \in \mathbb{N}$. Prove that

$$1 \cdot 3 \cdot 5 \cdot \cdots \cdot (2n-1) = \frac{(2n)!}{2^n n!}.$$

(The left hand side of this equality is understood to be the product of all odd integers from 1 to $2n-1$.)

Solution to Exercise 4.2.3 (sketched). See [Grinbe15, Exercise 3.2 **(a)**] or [17f-hw2s, Exercise 1 **(a)**] for a detailed solution. The main idea:

$$\begin{aligned}
(2n)! &= \prod_{i \in \{1,2,\ldots,2n\}} i = \underbrace{\left(\prod_{\substack{i \in \{1,2,\ldots,2n\}; \\ i \text{ is even}}} i\right)}_{\substack{=2\cdot4\cdot6\cdot\cdots\cdot(2n) \\ =2^n\cdot(1\cdot2\cdot3\cdot\cdots\cdot n) \\ \text{(by (107) and (105))}}} \cdot \underbrace{\left(\prod_{\substack{i \in \{1,2,\ldots,2n\}; \\ i \text{ is odd}}} i\right)}_{=1\cdot3\cdot5\cdot\cdots\cdot(2n-1)} \qquad \text{(by Theorem 4.2.9)} \\
&= 2^n \cdot \underbrace{(1 \cdot 2 \cdot 3 \cdot \cdots \cdot n)}_{=n!} \cdot (1 \cdot 3 \cdot 5 \cdot \cdots \cdot (2n-1)) = 2^n n! \cdot (1 \cdot 3 \cdot 5 \cdot \cdots \cdot (2n-1)).
\end{aligned}$$

□

Finally, let us state the natural generalization of Exercise 4.1.11 to products with several factors (each of which is a finite sum):

Theorem 4.2.14. For every $n \in \mathbb{N}$, let $[n]$ denote the set $\{1, 2, \ldots, n\}$.

Let $n \in \mathbb{N}$. For every $i \in [n]$, let $p_{i,1}, p_{i,2}, \ldots, p_{i,m_i}$ be finitely many numbers. Then,

$$\prod_{i=1}^{n} \sum_{k=1}^{m_i} p_{i,k} = \sum_{(k_1,k_2,\ldots,k_n) \in [m_1] \times [m_2] \times \cdots \times [m_n]} \prod_{i=1}^{n} p_{i,k_i}. \tag{116}$$

(**Pedantic remark:** If $n = 0$, then the Cartesian product $[m_1] \times [m_2] \times \cdots \times [m_n]$ has no factors; it is what is called an *empty Cartesian product*. It is understood to be a 1-element set, and its single element is the 0-tuple $()$ (also known as the empty list). Thus the equality (116) indeed holds for $n = 0$, as it just says $1 = 1$ in this case.)

The left hand side of (116) can be rewritten as

$$\left(p_{1,1} + p_{1,2} + \cdots + p_{1,m_1}\right) \left(p_{2,1} + p_{2,2} + \cdots + p_{2,m_2}\right) \cdots \left(p_{n,1} + p_{n,2} + \cdots + p_{n,m_n}\right),$$

and the right hand side is precisely what one obtains when expanding this product into a sum of $m_1 m_2 \cdots m_n$ monomials (i.e., products of single $p_{i,k}$'s). Thus, there is nothing surprising about Theorem 4.2.14. A formal proof of Theorem 4.2.14 can be found in [Grinbe15, Lemma 6.20].

4.3. Binomial coefficients

Our next topic are the *binomial coefficients*. In this section, we shall only briefly survey these coefficients from an algebraic perspective, but we will return to them later on as we get to enumerative combinatorics. Most of the claims in this section are easily proved by computation or induction (and also found in many easily accessible sources), so we will almost entirely omit their proofs. Much deeper treatments of binomial coefficients are found in the book [GrKnPa94] (whose entire Chapter 5 is devoted to binomial coefficients), in the book [Comtet74] (one of the classics on enumerative combinatorics and binomial identities), in the recent book [Spivey19], and in [Grinbe15, Chapter 3]. (Of these, I can recommend [GrKnPa94] as a first introduction.)[73] Another (somewhat silly but mathematically content-rich) introduction to binomial coefficients is a recent booklet by McCleary [Mcclea17].

We begin by defining binomial coefficients.

[73] See also the Wikipedia page for Pascal's triangle for an overview of many properties.

Definition 4.3.1. Let n and k be any two numbers. We define a number $\binom{n}{k}$ as follows:

- If $k \in \mathbb{N}$, then we set
$$\binom{n}{k} = \frac{n(n-1)(n-2)\cdots(n-k+1)}{k!}. \tag{117}$$
- If $k \notin \mathbb{N}$, then we set
$$\binom{n}{k} = 0. \tag{118}$$

We call $\binom{n}{k}$ a *binomial coefficient*, and we refer to it as "*n choose k*". (We will soon see what motivated this terminology.)

I believe this definition is the best at balancing generality and simplicity. There are two simpler definitions (one that defines $\binom{n}{k}$ as $\frac{n!}{k!(n-k)!}$, and another that defines $\binom{n}{k}$ as the number of k-element subsets of an n-element set) that work only in certain cases (e.g., when n is a nonnegative integer)[74], and thus are less suited as definitions; meanwhile, the only definitions more general than Definition 4.3.1 that I have seen require significant new concepts.

Warning 4.3.2. Definition 4.3.1 is pretty widespread: For example, it is followed in [GrKnPa94], [Comtet74] and [Grinbe15]. But it is not an unquestioned standard across the literature. All definitions I am aware of are equivalent in the "core region" of the binomial coefficients – that is, in the case when $n \in \mathbb{N}$ and $k \in \{0, 1, \ldots, n\}$. However, some definitions yield values differing from ours when $n < 0$. Many authors prefer to define $\binom{n}{k}$ only for $n \in \mathbb{N}$, or only for $k \in \mathbb{N}$, or even only in the most restrictive case (when $n \in \mathbb{N}$ and $k \in \{0, 1, \ldots, n\}$). Loehr, in [Loehr11], seems to avoid defining $\binom{n}{k}$ for negative n at all, despite this case being rather useful.

Some authors use notations like C_k^n or ${}^nC_k$ or ${}_nC_k$ for $\binom{n}{k}$.

Do not confuse the binomial coefficient $\binom{n}{k}$ with the column vector $\begin{pmatrix} n \\ k \end{pmatrix}$. (Note the differences in horizontal spacing.)

[74] We shall soon see that these definitions are equivalent to ours (in said cases).

Example 4.3.3. Let us see some consequences of Definition 4.3.1.

(a) For any number n, we have

$$\begin{aligned}\binom{n}{0} &= \frac{n(n-1)(n-2)\cdots(n-0+1)}{0!} \qquad \text{(by (117), applied to } k=0\text{)} \\ &= \frac{1}{1} \qquad \left(\begin{array}{c}\text{since } n(n-1)(n-2)\cdots(n-0+1) = (\text{empty product}) = 1 \\ \text{and } 0! = 1\end{array}\right) \\ &= 1. \end{aligned} \tag{119}$$

(b) For any number n, we have

$$\begin{aligned}\binom{n}{1} &= \frac{n(n-1)(n-2)\cdots(n-1+1)}{1!} \qquad \text{(by (117), applied to } k=1\text{)} \\ &= \frac{n}{1} \qquad \text{(since } n(n-1)(n-2)\cdots(n-1+1) = n \text{ and } 1! = 1\text{)} \\ &= n. \end{aligned} \tag{120}$$

(c) For any number n, we have

$$\begin{aligned}\binom{n}{2} &= \frac{n(n-1)(n-2)\cdots(n-2+1)}{2!} \qquad \text{(by (117), applied to } k=2\text{)} \\ &= \frac{n(n-1)}{2}. \end{aligned} \tag{121}$$

(d) For any number n, we have

$$\binom{n}{3} = \frac{n(n-1)(n-2)}{3!} = \frac{n(n-1)(n-2)}{6} \qquad \text{(likewise)}.$$

(e) The equality (117) (applied to $n=-1$ and $k=5$) yields

$$\begin{aligned}\binom{-1}{5} &= \frac{(-1)(-1-1)(-1-2)\cdots(-1-5+1)}{5!} = \frac{(-1)(-2)(-3)(-4)(-5)}{5!} \\ &= \frac{(-1)(-2)(-3)(-4)(-5)}{1\cdot 2\cdot 3\cdot 4\cdot 5} = -1.\end{aligned}$$

(f) More generally, for any $k \in \mathbb{N}$, we have

$$\begin{aligned}\binom{-1}{k} &= \frac{(-1)(-1-1)(-1-2)\cdots(-1-k+1)}{k!} \qquad \text{(by (117))} \\ &= \frac{(-1)(-2)\cdots(-k)}{k!} = \frac{(-1)(-2)\cdots(-k)}{1\cdot 2\cdot\cdots\cdot k} \\ &= (-1)^k. \end{aligned} \tag{122}$$

(g) The equality (121) (applied to $n=\sqrt{2}$) yields

$$\binom{\sqrt{2}}{2} = \frac{\sqrt{2}\left(\sqrt{2}-1\right)}{2}.$$

(h) The equality (118) (applied to $n = 2$ and $k = \sqrt{2}$) yields

$$\binom{2}{\sqrt{2}} = 0, \qquad \text{since } \sqrt{2} \notin \mathbb{N}.$$

Instead of further examples, let us see a table of the most important binomial coefficients $\binom{n}{k}$ – namely those with $n \in \mathbb{N}$ and $k \in \{0, 1, \ldots, n\}$. This table is called *Pascal's triangle*; here are its first 9 rows (i.e., the part that covers $n \in \{0, 1, \ldots, 8\}$):

											$k=0$ ↙						
$n = 0$	$\to$									1		$k=1$ ↙					
$n = 1$	$\to$								1		1		$k=2$ ↙				
$n = 2$	$\to$							1		2		1		$k=3$ ↙			
$n = 3$	$\to$						1		3		3		1		$k=4$ ↙		
$n = 4$	$\to$					1		4		6		4		1		$k=5$ ↙	
$n = 5$	$\to$				1		5		10		10		5		1		$k=6$ ↙
$n = 6$	$\to$			1		6		15		20		15		6		1		$k=7$ ↙
$n = 7$	$\to$		1		7		21		35		35		21		7		1	
$n = 8$	$\to$	1		8		28		56		70		56		28		8		1

Why did we restrict ourselves only to $k \in \{0, 1, \ldots, n\}$ when making this table? What about the binomial coefficients $\binom{n}{k}$ "left of Pascal's triangle" – i.e., those with $k < 0$? They are not shown because they are all 0 (by (118)). The binomial coefficients $\binom{n}{k}$ "right of Pascal's triangle" – i.e., those with $k > n$ – are also 0, as the following proposition shows:

Proposition 4.3.4. Let $n \in \mathbb{N}$ and $k \in \mathbb{R}$ be such that $k > n$. Then, $\binom{n}{k} = 0$.

Hints to the proof of Proposition 4.3.4. (See [19fco, Proposition 1.3.6].) For $k \notin \mathbb{N}$, this follows from (118). Thus, assume $k \in \mathbb{N}$. Now, recall the old joke problem "Simplify $(x - a)(x - b) \cdots (x - z)$". $\square$

Warning 4.3.5. Proposition 4.3.4 does not usually hold for $n \notin \mathbb{N}$.

A more interesting question is what we can say about the binomial coefficients $\binom{n}{k}$ "above Pascal's triangle" – i.e., those with $n < 0$. It turns out that they are just "mirror images" of the ones below (up to sign); namely, we have the following:

Proposition 4.3.6 (Upper negation formula)**.** Let $n \in \mathbb{R}$ and $k \in \mathbb{Z}$. Then,

$$\binom{-n}{k} = (-1)^k \binom{n+k-1}{k}. \tag{123}$$

Hints to the proof of Proposition 4.3.6. This is a simple computation using Definition 4.3.1. (See [19fco, Proposition 1.3.7] for details.) □

Proposition 4.3.4 and Proposition 4.3.6 are the reasons why almost all tables of binomial coefficients you will find in the literature show only (parts of) Pascal's triangle. (Of course, there are also the binomial coefficients $\binom{n}{k}$ with non-integer n, but it is not really clear how to tabulate them.)

Binomial coefficients tend to appear in almost all parts of mathematics (as coefficients or as standalone values). Being able to identify them when one encounters them is thus a useful skill. Myself, I know few numbers by heart, but the binomial coefficients in the first six rows of Pascal's triangle are among them (I never deliberately memorized them; I just see them too often to forget). If you need any further rows of Pascal's triangle and have no computer handy, the easiest way to construct them is using a surprising pattern in Pascal's triangle: Every entry (except for the 1 at the peak) is the sum of the two adjacent entries above it[75] (for example, $56 = 21 + 35$). This pattern holds not only for the entries of Pascal's triangle, but more generally for all binomial coefficients:

Theorem 4.3.7 (Recurrence of the binomial coefficients)**.** Let $n \in \mathbb{R}$ and $k \in \mathbb{R}$. Then,

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}.$$

Hints to the proof of Theorem 4.3.7. Since there were two cases in Definition 4.3.1, we must distinguish between three cases here: $k \in \{1, 2, 3, \ldots\}$, $k = 0$ and $k \notin \mathbb{N}$. The first case is the interesting one (the latter two are trivial). In this first case, rewrite $\binom{n-1}{k-1}$ and $\binom{n-1}{k}$ as fractions and add them by finding a common denominator. □

[75] If there is only one adjacent entry above it (as it happens for the 1's along the left and right sides of Pascal's triangle), then we treat the missing other entry as a 0 (after all, there should be a 0 at its position; we just didn't put it in the table).

Theorem 4.3.7 is known as *Pascal's rule*. It provides perhaps the quickest way to compute Pascal's triangle up to a given row; however, Definition 4.3.1 might still be a better way to compute a specific entry (as long as one remembers not to multiply the products out, since there will be a lot of cancellation visible in the fraction). There is an even more explicit formula:

Theorem 4.3.8 (Factorial formula for the binomial coefficients)**.** Let $n \in \mathbb{N}$ and $k \in \mathbb{N}$ be such that $k \leq n$. Then,

$$\binom{n}{k} = \frac{n!}{k! \cdot (n-k)!}.$$

Hints to the proof of Theorem 4.3.8. Notice that

$$n(n-1)(n-2)\cdots(n-k+1) = n!/(n-k)!.$$

(See [19fco, Theorem 1.3.9] for details.) □

Warning 4.3.9. Theorem 4.3.8 cannot be used to compute $\binom{-3}{5}$ or $\binom{1/3}{4}$ or $\binom{\pi}{\sqrt{2}}$; it only applies under the given assumptions ($n \in \mathbb{N}$ and $k \in \mathbb{N}$ and $k \leq n$). (This is the reason why, unlike various authors, we do **not define** $\binom{n}{k}$ through Theorem 4.3.8.)

Another immediately visible pattern exhibited by Pascal's triangle is its symmetry across a vertical axis:

Theorem 4.3.10 (Symmetry of the binomial coefficients)**.** Let $n \in \mathbb{N}$ and $k \in \mathbb{R}$. Then,

$$\binom{n}{k} = \binom{n}{n-k}.$$

Hints to the proof of Theorem 4.3.10. If $k \in \mathbb{N}$ and $k \leq n$, then rewrite both sides using Theorem 4.3.8. Otherwise, argue that both sides are 0. (See [19fco, Theorem 1.3.11] for details.) □

Warning 4.3.11. Theorem 4.3.10 would be false without the requirement $n \in \mathbb{N}$. For example, $n = -1$ and $k = 0$ provides a counterexample.

An easy consequence of Theorem 4.3.10 is that

$$\binom{n}{n} = 1 \qquad \text{for each } n \in \mathbb{N}. \tag{124}$$

Indeed, Theorem 4.3.10 (applied to $k = n$) yields

$$\binom{n}{n} = \binom{n}{n-n} = \binom{n}{0} = 1 \qquad \text{(by (119))}$$

for each $n \in \mathbb{N}$.

The following fact is one of the most important properties of binomial coefficients; it explains why they are often regarded as combinatorial in nature:

Theorem 4.3.12 (Combinatorial interpretation of the binomial coefficients)**.** Let $n \in \mathbb{N}$ and $k \in \mathbb{R}$. Let S be an n-element set. Then,

$$\binom{n}{k} = (\text{the number of } k\text{-element subsets of } S).$$

Example 4.3.13. (a) Let $n = 4$ and $k = 2$ and $S = \{1, 2, 3, 4\}$. Then, the 2-element subsets of S are $\{1, 2\}$, $\{1, 3\}$, $\{1, 4\}$, $\{2, 3\}$, $\{2, 4\}$ and $\{3, 4\}$. The number of these subsets is $6 = \binom{4}{2}$, which is exactly what Theorem 4.3.12 predicts.

(b) Now, let $k = 5$ instead (while n is still 4, and S is still $\{1, 2, 3, 4\}$). Then, there are no 5-element subsets of S, since S only has 4 elements. Thus, the number of these 5-element subsets is $0 = \binom{4}{5}$, which is exactly what Theorem 4.3.12 predicts.

Hints to the proof of Theorem 4.3.12. There is a fairly straightforward proof of Theorem 4.3.12 by induction on n. In the induction step, pick an element $s \in S$ and classify the k-element subsets of S as "red" or "green" according to whether they contain s or not (similarly to our above proof of Theorem 2.3.4). This proof can be found in [19fco, proof of Theorem 1.3.12] or [Grinbe15, solution to Exercise 3.4].

An alternative proof can be given using enumerative combinatorics (see [LeLeMe16, §15.5] or [19fco, §2.7]). □

Warning 4.3.14. Theorem 4.3.12 says nothing about $\binom{n}{k}$ when $n \notin \mathbb{N}$.

The following important fact is **not** obvious from Definition 4.3.1:

Theorem 4.3.15 (Integrality of the binomial coefficients)**.** Let $n \in \mathbb{Z}$ and $k \in \mathbb{Z}$. Then, $\binom{n}{k} \in \mathbb{Z}$.

Hints to the proof of Theorem 4.3.15. In the case $n \in \mathbb{N}$, this can either be shown by induction on n (using Theorem 4.3.7), or obtained immediately from Theorem 4.3.12. The general case $n \in \mathbb{Z}$ can be reduced to the case $n \in \mathbb{N}$ via Proposition 4.3.6 and (118). (See [19fco, proof of Theorem 1.3.16] for details.) □

The most famous property of binomial coefficients is the *binomial theorem*:

Theorem 4.3.16 (the binomial formula)**.** Let x and y be any numbers. Let $n \in \mathbb{N}$. Then,

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}.$$

This is why the $\binom{n}{k}$ are called the "binomial coefficients".

Hints to the proof of Theorem 4.3.16. Either induct on n (see [19fco, proof of Theorem 1.3.24] for details), or argue combinatorially using Theorem 4.2.14 (best done after familiarizing yourself with basic bijections). □

Let us state two famous consequences of Theorem 4.3.16. The first one says that the sum of all entries in the n-th row of Pascal's triangle (where rows are counted from 0) is 2^n:

Corollary 4.3.17. Let $n \in \mathbb{N}$. Then, $\sum_{k=0}^{n} \binom{n}{k} = 2^n$.

Hints to the proof of Corollary 4.3.17. Apply Theorem 4.3.16 to $x = 1$ and $y = 1$. (See [19fco, Corollary 1.3.27] for details.) □

The second consequence ([19fco, Proposition 1.3.28], [Spivey19, §3.4, Identity 12]) is a bit subtler:

Proposition 4.3.18. Let $n \in \mathbb{N}$. Then,

$$\sum_{k=0}^{n} (-1)^k \binom{n}{k} = [n = 0]. \tag{125}$$

Here, we are using the so-called *Iverson bracket notation* (which despite its trivial definition is surprisingly useful):

Definition 4.3.19. If $\mathcal{A}$ is any logical statement, then we define an integer $[\mathcal{A}] \in \{0, 1\}$ by

$$[\mathcal{A}] = \begin{cases} 1, & \text{if } \mathcal{A} \text{ is true;} \\ 0, & \text{if } \mathcal{A} \text{ is false.} \end{cases}$$

For example, $[2+2=4] = 1$ but $[2+2=5] = 0$.

If $\mathcal{A}$ is any logical statement, then $[\mathcal{A}]$ is known as the *truth value* of $\mathcal{A}$.

Hints to the proof of Proposition 4.3.18. Apply Theorem 4.3.16 to $x = -1$ and $y = 1$, and recall what power of 0 is nonzero. (See [19fco, Proposition 1.3.28] for details.) □

We finish our first incursion into binomial coefficients with a formula that expresses Fibonacci numbers as sums thereof:

Proposition 4.3.20. Let $(f_0, f_1, f_2, \ldots)$ be the Fibonacci sequence. Let $n \in \mathbb{N}$. Then, the Fibonacci number f_{n+1} is

$$f_{n+1} = \sum_{k=0}^{n} \binom{n-k}{k} = \binom{n-0}{0} + \binom{n-1}{1} + \binom{n-2}{2} + \cdots + \binom{n-n}{n}.$$

Note that roughly half the addends on the right hand side of Proposition 4.3.20 are 0 (indeed, Proposition 4.3.4 shows that $\binom{n-k}{k} = 0$ for any $k \in \{0, 1, \ldots, n\}$ satisfying $k > n/2$) and thus can be discarded; nevertheless it is easier to have the sum end at $k = n$ rather than figure out where exactly its nonzero addends stop.

Hints to the proof of Proposition 4.3.20. This is probably the hardest statement in this section, which is not saying much. There is a proof by strong induction on n (using Theorem 4.3.7); it is not very enlightening and somewhat laborious due to the fiddling involved in making all the sums have the same upper limit. (See [Vorobi02, §15] or [AndCri17, Problem 2.4] for a sketch of this proof, but notice that both of these sources are sloppy with the summation limits.) We will state a more general result (Proposition 4.9.18) later on, at which point we will give this proof in full detail; Proposition 4.3.20 is merely a particular case of this latter result.

Alternatively, there is a combinatorial proof, which proceeds as follows: WLOG assume that $n \geq 1$ (else, just check it by hand). Let $[n-1]$ denote the set $\{1, 2, \ldots, n-1\}$. Then, Theorem 2.3.4 yields that the number of all lacunar subsets of $[n-1]$ is f_{n+1}. However, for each $k \in \{0, 1, \ldots, n\}$, the number of all lacunar k-element subsets of $[n-1]$ is $\binom{n-k}{k}$, as can easily be proved by induction on n (or even by a nicer, combinatorial argument: see [19fco, Proposition 1.4.10]). Hence, the total number of all lacunar subsets of $[n-1]$ is $\sum_{k=0}^{n} \binom{n-k}{k}$. Comparing these two results yields the claim of Proposition 4.3.20. (See [19fco, §1.4.5, proof of Proposition 1.3.32] for this proof.) □

4.4. Recitation #3: Sums, products, binomial coefficients

Next comes an exercise in finite sums that is rather similar to Exercise 2.2.1:

Exercise 4.4.1. Let $(f_0, f_1, f_2, \ldots)$ be the Fibonacci sequence. Prove that each integer $n \geq 0$ satisfies

$$f_1 + f_3 + f_5 + \cdots + f_{2n-1} = f_{2n}.$$

Discussion of Exercise 4.4.1. This is easy to prove by induction on n, but let us try to prove it by the telescope principle.

Let $n \geq 0$ be an integer. For each integer $i \geq 1$, we have $f_{i+1} = f_i + f_{i-1}$ (by the recursive definition of the Fibonacci numbers) and therefore

$$f_i = f_{i+1} - f_{i-1}. \tag{126}$$

Thus, for each positive integer s, we have

$$\begin{aligned} f_{2s-1} &= f_{(2s-1)+1} - f_{(2s-1)-1} \qquad \text{(by (126), applied to } i = 2s-1) \\ &= f_{2s} - f_{2(s-1)} \end{aligned} \tag{127}$$

(since $(2s-1)+1 = 2s$ and $(2s-1)-1 = 2(s-1)$). Now,

$$\begin{aligned} f_1 + f_3 + f_5 + \cdots + f_{2n-1} &= \sum_{s=1}^{n} \underbrace{f_{2s-1}}_{\substack{=f_{2s}-f_{2(s-1)} \\ \text{(by (127))}}} = \sum_{s=1}^{n} \left(f_{2s} - f_{2(s-1)} \right) \\ &= f_{2n} - f_{2(1-1)} \\ &\qquad \text{(by (74), applied to } u = 1,\ v = n \text{ and } a_s = f_{2s}) \\ &= f_{2n} - 0 \qquad \left(\text{since } f_{2(1-1)} = f_0 = 0 \right) \\ &= f_{2n}. \end{aligned}$$

This solves Exercise 4.4.1.

Exercise 4.4.1 has other solutions. For example, we may solve it (for $n \geq 1$) by noticing that

$$\begin{aligned} &\underbrace{f_1}_{=1} + \underbrace{f_3}_{\substack{=f_2+f_1 \\ =f_1+f_2}} + \underbrace{f_5}_{\substack{=f_4+f_3 \\ =f_3+f_4}} + \cdots + \underbrace{f_{2n-1}}_{\substack{=f_{2n-2}+f_{2n-3} \\ =f_{2n-3}+f_{2n-2}}} \\ &= 1 + \underbrace{(f_1+f_2) + (f_3+f_4) + \cdots + (f_{2n-3}+f_{2n-2})}_{\substack{=f_1+f_2+\cdots+f_{2n-2} \\ =f_{(2n-2)+2}-1 \\ \text{(by Exercise 2.2.1, applied to } 2n-2 \text{ instead of } n)}} \\ &= 1 + f_{(2n-2)+2} - 1 = f_{(2n-2)+2} = f_{2n}. \end{aligned}$$

(The case $n = 0$ is easily checked manually.) □

The next exercise appears, e.g., in [AndTet18, Introduction]:

Exercise 4.4.2. "Simplify" $\prod_{k=0}^{n-1} \left(1 + a^{2^k} \right)$ for $a \neq 1$ and $n \in \mathbb{N}$. (The answer should not contain any $\prod$ signs.)

Discussion of Exercise 4.4.2. Let a be a number such that $a \neq 1$. Let $n \in \mathbb{N}$. For each $k \in \mathbb{N}$, we have

$$\begin{aligned} 1 - \underbrace{a^{2^{k+1}}}_{\substack{=a^{2^k \cdot 2} \\ (\text{since } 2^{k+1} = 2^k \cdot 2)}} &= 1 - \underbrace{a^{2^k \cdot 2}}_{=\left(a^{2^k}\right)^2} = 1 - \left(a^{2^k}\right)^2 \\ &= \left(1 - a^{2^k}\right)\left(1 + a^{2^k}\right) \end{aligned} \tag{128}$$

(by the classical formula $1 - b^2 = (1-b)(1+b)$, applied to $b = a^{2^k}$) and therefore

$$1 + a^{2^k} = \frac{1 - a^{2^{k+1}}}{1 - a^{2^k}}. \tag{129}$$

Now,

$$\begin{aligned} \prod_{k=0}^{n-1} \underbrace{\left(1 + a^{2^k}\right)}_{\substack{= \dfrac{1 - a^{2^{k+1}}}{1 - a^{2^k}} \\ (\text{by } (129))}} &= \prod_{k=0}^{n-1} \frac{1 - a^{2^{k+1}}}{1 - a^{2^k}} = \prod_{s=1}^{n} \frac{1 - a^{2^s}}{1 - a^{2^{s-1}}} \\ &\qquad (\text{here, we have substituted } s-1 \text{ for } k \text{ in the product}) \\ &= \frac{1 - a^{2^n}}{1 - a^{2^{1-1}}} \\ &\qquad \left(\text{by (109), applied to } u = 1,\ v = n \text{ and } a_s = a^{2^s}\right) \\ &= \frac{1 - a^{2^n}}{1 - a} \qquad \left(\text{since } a^{2^{1-1}} = a^{2^0} = a^1 = a\right). \end{aligned}$$

So we found our answer:

$$\prod_{k=0}^{n-1} \left(1 + a^{2^k}\right) = \frac{1 - a^{2^n}}{1 - a}. \tag{130}$$

But wait – did you spot the subtle error?

The error is the following: We divided the equality (128) by $1 - a^{2^k}$ to get (129), but $1 - a^{2^k}$ can be zero. Namely, if $a = -1$, then $1 - a^{2^k}$ will be zero for every $k > 0$; thus, in all these cases, we have divided by zero. Nevertheless, our final result (130) holds even in these cases – it is only the proof that went wrong.

Fortunately, now that we know the equality (130), it is utterly straightforward to prove it by induction on n (using (128) in the induction step).[76] Thus, while the

[76] Here is this proof in detail:

Proof of (130): Forget that we fixed n. We shall prove (130) by induction on n.

argument by which we originally arrived at (130) was flawed, we nevertheless were able to put it to good use, since it helped us find the formula (130), which we were then able to prove by a different (perfectly sound) argument.

This was a typical example of a phenomenon that is often observed in mathematics (particularly in modern research): Results are often discovered by means of reasoning that is not quite rigorous (to the point that it often sounds like quackery or cannot be communicated at all); then, one is forced to come up with different arguments to find valid proofs for the results. In our case, we did not have to look far (in essence, the induction proof was just a restatement of the not-quite-correct telescope argument avoiding unnecessary divisions, which helped it avoid division by zero); but this sort of "parallel construction" of proofs can be quite a long and difficult process. □

The next exercises are about binomial coefficients:

Exercise 4.4.3. Show that $\sum_{k=0}^{(n-1)/2} \binom{n}{k} = 2^{n-1}$ for any odd $n \in \mathbb{N}$.

Induction base: We have

$$\prod_{k=0}^{1-1} \left(1 + a^{2^k}\right) = \prod_{k=0}^{0} \left(1 + a^{2^k}\right) = (\text{empty product}) = 1 = \frac{1 - a^{2^0}}{1 - a}$$

(since $\frac{1 - a^{2^0}}{1 - a} = \frac{1 - a^1}{1 - a} = \frac{1 - a}{1 - a} = 1$). In other words, (130) holds for $n = 0$.

Induction step: Let $m \in \mathbb{N}$. Assume (as the induction hypothesis) that (130) holds for $n = m$. We must prove that (130) holds for $n = m + 1$.

Our induction hypothesis says that (130) holds for $n = m$; in other words, it says that $\prod_{k=0}^{m-1} \left(1 + a^{2^k}\right) = \frac{1 - a^{2^m}}{1 - a}$.

Now, (128) (applied to $k = m$) yields

$$1 - a^{2^{m+1}} = \left(1 - a^{2^m}\right)\left(1 + a^{2^m}\right). \tag{131}$$

But

$$\prod_{k=0}^{m} \left(1 + a^{2^k}\right) = \underbrace{\left(\prod_{k=0}^{m-1} \left(1 + a^{2^k}\right)\right)}_{= \frac{1 - a^{2^m}}{1 - a}} \cdot \left(1 + a^{2^m}\right) = \frac{1 - a^{2^m}}{1 - a} \cdot \left(1 + a^{2^m}\right)$$

$$= \frac{1}{1 - a} \cdot \underbrace{\left(1 - a^{2^m}\right)\left(1 + a^{2^m}\right)}_{\substack{= 1 - a^{2^{m+1}} \\ (\text{by } (131))}} = \frac{1}{1 - a} \cdot \left(1 - a^{2^{m+1}}\right) = \frac{1 - a^{2^{m+1}}}{1 - a}.$$

In other words, (130) holds for $n = m + 1$. This completes the induction step. Thus, (130) is proved.

Discussion of Exercise 4.4.3. Intuitively, the idea is clear from a look at Pascal's triangle: Pascal's triangle is symmetric across the vertical axis (by Theorem 4.3.10). Thus, if $n \in \mathbb{N}$ is odd, then the n-th row of Pascal's triangle (i.e., the row that begins with 1 and n) splits neatly into two equal halves (one half being to the left of the vertical axis, and the other half being to the right of it), and therefore the sum of the left half of this row is exactly $\frac{1}{2}$ of the sum of all entries in this row; but the latter sum is 2^n according to Corollary 4.3.17. Thus, the former sum is 2^{n-1}.

Here is a formal way to restate this argument: Let $n \in \mathbb{N}$ be odd. Thus, $n = 2u+1$ for some $u \in \mathbb{N}$. Consider this u. From $n = 2u+1$, we obtain $2u = n-1$ and $u = (n-1)/2$ and $u+1 = n-u$ and $2u+1 = n$. Now, Corollary 4.3.17 yields $\sum_{k=0}^{n} \binom{n}{k} = 2^n = 2^{2u+1}$ (since $n = 2u+1$), so that

$$\begin{aligned} 2^{2u+1} &= \sum_{k=0}^{n} \binom{n}{k} = \sum_{k=0}^{2u+1} \binom{n}{k} \qquad (\text{since } n = 2u+1) \\ &= \sum_{k=0}^{u} \binom{n}{k} + \sum_{k=u+1}^{2u+1} \binom{n}{k} \end{aligned} \tag{132}$$

(here, we have split the sum at $k = u$; that is, we have applied (84) to 0, u and $2u+1$ instead of u, v and w). But

$$\begin{aligned} \sum_{k=u+1}^{2u+1} \binom{n}{k} &= \sum_{k=n-u}^{n} \binom{n}{k} \qquad (\text{since } u+1 = n-u \text{ and } 2u+1 = n) \\ &= \sum_{k=0}^{u} \underbrace{\binom{n}{n-k}}_{\substack{= \binom{n}{k} \\ \text{(by Theorem 4.3.10)}}} \\ &\qquad \left(\begin{array}{c} \text{here, we have substituted } n-k \text{ for } k \text{ in the sum,} \\ \text{since the map } \{0,1,\ldots,u\} \to \{n-u, n-u+1, \ldots, n\} \\ \text{that sends each } k \text{ to } n-k \text{ is a bijection} \end{array} \right) \\ &= \sum_{k=0}^{u} \binom{n}{k}. \end{aligned}$$

Hence, (132) becomes

$$2^{2u+1} = \sum_{k=0}^{u} \binom{n}{k} + \underbrace{\sum_{k=u+1}^{2u+1} \binom{n}{k}}_{= \sum_{k=0}^{u} \binom{n}{k}} = \sum_{k=0}^{u} \binom{n}{k} + \sum_{k=0}^{u} \binom{n}{k} = 2 \cdot \sum_{k=0}^{u} \binom{n}{k}.$$

Dividing both sides of this equality by 2, we find

$$2^{2u} = \sum_{k=0}^{u} \binom{n}{k} = \sum_{k=0}^{(n-1)/2} \binom{n}{k} \qquad (\text{since } u = (n-1)/2).$$

Hence,

$$\sum_{k=0}^{(n-1)/2} \binom{n}{k} = 2^{2u} = 2^{n-1} \qquad (\text{since } 2u = n-1).$$

This solves Exercise 4.4.3. □

Exercise 4.4.4. Recall once again the Fibonacci sequence $(f_0, f_1, f_2, \ldots)$, which is defined recursively by $f_0 = 0$, $f_1 = 1$, and

$$f_n = f_{n-1} + f_{n-2} \qquad \text{for all } n \geq 2. \tag{133}$$

Now, let us define f_n for **negative** integers n as well, by "applying (133) backwards": This means that we set $f_{n-2} = f_n - f_{n-1}$ for all integers $n \leq 1$. This allows us to recursively compute $f_{-1}, f_{-2}, f_{-3}, \ldots$ (in this order). For example,

$$\begin{aligned} f_{-1} &= f_1 - f_0 = 1 - 0 = 1; \\ f_{-2} &= f_0 - f_{-1} = 0 - 1 = -1; \\ f_{-3} &= f_{-1} - f_{-2} = 1 - (-1) = 2, \end{aligned}$$

etc.

(a) Prove that $f_{-n} = (-1)^{n-1} f_n$ for each $n \in \mathbb{Z}$.
(b) Prove that $f_{n+m+1} = f_n f_m + f_{n+1} f_{m+1}$ for all $n \in \mathbb{Z}$ and $m \in \mathbb{Z}$.
(c) Prove that $7 f_n = f_{n-4} + f_{n+4}$ for all $n \in \mathbb{Z}$.
(d) Prove that if $a, b \in \mathbb{Z}$ satisfy $a \mid b$, then $f_a \mid f_b$.

Discussion of Exercise 4.4.4. Parts **(a)**, **(b)** and **(c)** of Exercise 4.4.4 appear (with detailed solutions) in [18f-mt1s, Exercise 4]; thus, we shall only sketch the arguments here.

Because of our definition of the f_n, the equation

$$f_n = f_{n-1} + f_{n-2} \tag{134}$$

holds for all $n \in \mathbb{Z}$.

(a) First, prove (by strong induction on n) that

$$f_{-n} = (-1)^{n-1} f_n \qquad \text{for each } n \in \mathbb{N}. \tag{135}$$

Then, conclude that the equality $f_{-n} = (-1)^{n-1} f_n$ holds for each $n \in \mathbb{Z}$ as well, because:

- If n is nonnegative, then it follows from (135).

- If n is negative, then (135) (applied to $-n$ instead of n) yields $f_n = (-1)^{-n-1} f_{-n}$; but this quickly yields $f_{-n} = (-1)^{n-1} f_n$.

Thus, Exercise 4.4.4 **(a)** is solved.

An alternative solution for Exercise 4.4.4 **(a)** can be given by strong induction on $|n|$. (Note that strong induction on n will not work, since n ranges over $\mathbb{Z}$.) We leave the details to the reader.

(b) First, we claim that

$$f_{n+m+1} = f_n f_m + f_{n+1} f_{m+1} \qquad \text{for all } n \in \mathbb{N} \text{ and } m \in \mathbb{Z}. \tag{136}$$

Indeed, (136) can be proved in the same way as we proved (11) for all nonnegative integers n and m (see our "Second attempt at solving Exercise 2.2.3" above) – i.e., by induction on n. (The proof even becomes a bit simpler, since we no longer have to treat the case $m = 0$ separately: It is no longer problematic that $m - 1$ may be negative.)

Now, recall that we need to prove that $f_{n+m+1} = f_n f_m + f_{n+1} f_{m+1}$ for all $n \in \mathbb{Z}$ and $m \in \mathbb{Z}$. So let us fix $n \in \mathbb{Z}$ and $m \in \mathbb{Z}$. We must prove that $f_{n+m+1} = f_n f_m + f_{n+1} f_{m+1}$. If $n \in \mathbb{N}$, then this follows from (136). Hence, we WLOG assume that $n \notin \mathbb{N}$. Hence, n is negative, so that $n \leq -1$ and therefore $-1 - \underbrace{n}_{\leq -1} \geq -1 - (-1) = 0$. In other words, $-1 - n \in \mathbb{N}$. Thus, (136) (applied to $-1 - n$ and $-1 - m$ instead of n and m) yields

$$\begin{aligned}
& f_{(-1-n)+(-1-m)+1} \\
&= \underbrace{f_{-1-n}}_{\substack{=f_{-(n+1)} \\ =(-1)^{(n+1)-1} f_{n+1} \\ \text{(by Exercise 4.4.4 (a),} \\ \text{applied to } n+1 \\ \text{instead of } n)}} \underbrace{f_{-1-m}}_{\substack{=f_{-(m+1)} \\ =(-1)^{(m+1)-1} f_{m+1} \\ \text{(by Exercise 4.4.4 (a),} \\ \text{applied to } m+1 \\ \text{instead of } n)}} + \underbrace{f_{(-1-n)+1}}_{\substack{=f_{-n} \\ =(-1)^{n-1} f_n \\ \text{(by Exercise 4.4.4 (a))}}} \underbrace{f_{(-1-m)+1}}_{\substack{=f_{-m} \\ =(-1)^{m-1} f_m \\ \text{(by Exercise 4.4.4 (a),} \\ \text{applied to } m \\ \text{instead of } n)}} \\
&= \underbrace{(-1)^{(n+1)-1}}_{=(-1)^n} f_{n+1} \cdot \underbrace{(-1)^{(m+1)-1}}_{=(-1)^m} f_{m+1} + \underbrace{(-1)^{n-1}}_{=-(-1)^n} f_n \cdot \underbrace{(-1)^{m-1}}_{=-(-1)^m} f_m \\
&= (-1)^n f_{n+1} \cdot (-1)^m f_{m+1} + \left(-(-1)^n\right) f_n \cdot \left(-(-1)^m\right) f_m \\
&= \underbrace{(-1)^n (-1)^m}_{=(-1)^{n+m}} \underbrace{(f_{n+1} f_{m+1} + f_n f_m)}_{=f_n f_m + f_{n+1} f_{m+1}} = (-1)^{n+m} \left(f_n f_m + f_{n+1} f_{m+1}\right).
\end{aligned}$$

Comparing this with

$$\begin{aligned}
& f_{(-1-n)+(-1-m)+1} \\
&= f_{-(n+m+1)} \qquad (\text{since } (-1-n) + (-1-m) + 1 = -(n+m+1)) \\
&= \underbrace{(-1)^{(n+m+1)-1}}_{=(-1)^{n+m}} f_{n+m+1} \qquad (\text{by Exercise 4.4.4 } \textbf{(a)}\text{, applied to } n+m+1 \text{ instead of } n) \\
&= (-1)^{n+m} f_{n+m+1},
\end{aligned}$$

we obtain

$$(-1)^{n+m} f_{n+m+1} = (-1)^{n+m} (f_n f_m + f_{n+1} f_{m+1}).$$

Dividing both sides of this equality by $(-1)^{n+m}$, we obtain $f_{n+m+1} = f_n f_m + f_{n+1} f_{m+1}$. This solves Exercise 4.4.4 **(b)**.

An alternative solution to Exercise 4.4.4 **(b)** can be found in [18f-mt1s, Exercise 4]; it relies on two-sided induction (Theorem 3.1.9).

(c) This is just an exercise in applying the recursive equation of the Fibonacci sequence over and over:[77]

$$\begin{aligned}
f_{n-4} + \underbrace{f_{n+4}}_{\overset{(134)}{=} f_{n+3}+f_{n+2}} &= f_{n-4} + \underbrace{f_{n+3}}_{\overset{(134)}{=} f_{n+2}+f_{n+1}} + \underbrace{f_{n+2}}_{\overset{(134)}{=} f_{n+1}+f_n} \\
&= f_{n-4} + f_{n+2} + f_{n+1} + f_{n+1} + f_n = f_{n-4} + \underbrace{f_{n+2}}_{\overset{(134)}{=} f_{n+1}+f_n} + 2f_{n+1} + f_n \\
&= f_{n-4} + f_{n+1} + f_n + 2f_{n+1} + f_n = f_{n-4} + 2f_n + 3 \underbrace{f_{n+1}}_{\overset{(134)}{=} f_n + f_{n-1}} \\
&= f_{n-4} + 2f_n + 3(f_n + f_{n-1}) = f_{n-4} + 5f_n + 3 \underbrace{f_{n-1}}_{\overset{(134)}{=} f_{n-2}+f_{n-3}} \\
&= f_{n-4} + 5f_n + 3(f_{n-2} + f_{n-3}) = \underbrace{f_{n-4} + f_{n-3}}_{\overset{(134)}{=} f_{n-2}} + 5f_n + 3f_{n-2} + 2f_{n-3} \\
&= f_{n-2} + 5f_n + 3f_{n-2} + 2f_{n-3} = 5f_n + 2f_{n-2} + 2\underbrace{(f_{n-2} + f_{n-3})}_{\overset{(134)}{=} f_{n-1}} \\
&= 5f_n + 2f_{n-2} + 2f_{n-1} = 5f_n + 2\underbrace{(f_{n-1} + f_{n-2})}_{\overset{(134)}{=} f_n} \\
&= 5f_n + 2f_n = 7f_n.
\end{aligned}$$

Alternatively, Exercise 4.4.4 **(c)** can be obtained by applying Exercise 4.4.4 **(b)** to $m = -5$ and again to $m = 3$, and then adding the resulting equalities together. (See [18f-mt1s, Exercise 4] for the details.)

(d) Let $a, b \in \mathbb{Z}$ satisfy $a \mid b$. We must prove that $f_a \mid f_b$.

The numbers $|a|$ and $|b|$ belong to $\mathbb{N}$ (since a and b belong to $\mathbb{Z}$). Furthermore, Proposition 3.1.3 **(a)** yields $|a| \mid |b|$ (since $a \mid b$). Hence, Exercise 3.2.2 (applied to $|a|$ and $|b|$ instead of a and b) yields $f_{|a|} \mid f_{|b|}$.

It remains to derive $f_a \mid f_b$ from this. But this is easy: Exercise 3.2.2 **(a)** (applied to $n = a$) yields $f_{-a} = (-1)^{a-1} f_a = \pm f_a$. Thus, $f_{|a|} = \pm f_a$ [78]. Likewise, $f_{|b|} = \pm f_b$;

[77] The symbol "$\overset{(134)}{=}$" designates an equality that follows from (134). For example, $f_5 \overset{(134)}{=} f_4 + f_3$.

[78] *Proof.* We are in one of the following two cases:

thus, $f_b = \pm f_{|b|}$. Hence, $f_{|b|} \mid f_b$. Also, from $f_{|a|} = \pm f_a$, we obtain $f_a \mid f_{|a|}$. Thus, $f_a \mid f_{|a|} \mid f_{|b|} \mid f_b$. This solves Exercise 4.4.4 **(d)**. $\square$

4.5. Homework set #2: More number theory and sums

This is a **regular** problem set. See Section 3.7 for details on grading.

This homework set covers the end of Chapter 3 and the above parts of Chapter 4. Some of the problems may also be unrelated.

Please solve **at most 5** problems. (No points will be given for further solutions.)

Exercise 4.5.1. Let $n \in \mathbb{N}$. Let $a_1, a_2, \ldots, a_n$ be n **odd** integers. Prove that

$$a_1 a_2 + a_2 a_3 + \cdots + a_{n-1} a_n + a_n a_1 \equiv n \bmod 4.$$

Exercise 4.5.2. Let a and b be two coprime positive integers.

(a) Prove that there do **not** exist any positive integers x and y satisfying $ab = xa + yb$.

(b) Prove that there do **not** exist any $x, y \in \mathbb{N}$ satisfying $ab - a - b = xa + yb$.

Exercise 4.5.3. Let n and m be two coprime positive integers. Let $u \in \mathbb{Z}$. Prove that

$$(u^n - 1)(u^m - 1) \mid (u - 1)(u^{nm} - 1).$$

Exercise 4.5.4. Let $n, m \in \mathbb{N}$ satisfy $n > 0$. Prove the following:

(a) We have $\dfrac{m}{n}\dbinom{n}{m} = \dbinom{n-1}{m-1}$.

(b) We have $\dfrac{\gcd(n,m)}{n}\dbinom{n}{m} \in \mathbb{Z}$.

Exercise 4.5.5. Let n, i and j be positive integers such that $i < n$ and $j < n$. Prove that $\gcd\left(\dbinom{n}{i}, \dbinom{n}{j}\right) > 1$.

Case 1: We have $a \geq 0$.

Case 2: We have $a < 0$.

Let us first consider Case 1. In this case, we have $a \geq 0$. Thus, $|a| = a$, so that $f_{|a|} = f_a = \pm f_a$. Hence, $f_{|a|} = \pm f_a$ is proved in Case 1.

Let us now consider Case 2. In this case, we have $a < 0$. Thus, $|a| = -a$, so that $f_{|a|} = f_{-a} = \pm f_a$. Hence, $f_{|a|} = \pm f_a$ is proved in Case 2.

We have now proved that $f_{|a|} = \pm f_a$ in each of the two Cases 1 and 2. Hence, $f_{|a|} = \pm f_a$ always holds, qed.

Exercise 4.5.6. Let n be a positive integer. We let $\phi(n)$ denote the number of all $i \in \{1, 2, \ldots, n\}$ satisfying $i \perp n$. (For example, $\phi(12) = 4$, because there are exactly 4 numbers $i \in \{1, 2, \ldots, 12\}$ satisfying $i \perp 12$: namely, 1, 5, 7 and 11.)

(a) Prove that $\phi(n)$ is even if $n > 2$.

(b) Prove that the sum of all $i \in \{1, 2, \ldots, n\}$ satisfying $i \perp n$ equals $\frac{1}{2} n\phi(n)$ if $n > 1$.

[**Remark:** The function $\phi : \{1, 2, 3, \ldots\} \to \mathbb{N}$ that sends each positive integer n to $\phi(n)$ is known as the *Euler totient function* (or the *phi-function*). Here is a table of its first few values:

n	1	2	3	4	5	6	7	8	9	10	11	12	13
$\phi(n)$	1	1	2	2	4	2	6	4	6	4	10	4	12

.

Can you spot any patterns?]

Exercise 4.5.7. Let $(f_0, f_1, f_2, \ldots)$ be the Fibonacci sequence. Find $\sum_{k=2}^{\infty} \frac{f_k}{f_{k-1} f_{k+1}}$.

Exercise 4.5.8. **(a)** Prove that

$$\sum_{i=0}^{n} \binom{i}{k} = \binom{n+1}{k+1} \qquad \text{for each } n \in \mathbb{N} \text{ and } k \in \mathbb{N}.$$

(b) Prove that[79]

$$\sum_{k=0}^{m} (-1)^k \binom{n}{k} = (-1)^m \binom{n-1}{m} \qquad \text{for each } n \in \mathbb{C} \text{ and } m \in \mathbb{N}.$$

Exercise 4.5.9. Let $(f_0, f_1, f_2, \ldots)$ be the Fibonacci sequence. Prove that

$$2^{n-1} \cdot f_n = \sum_{k=0}^{n} \binom{n}{2k+1} 5^k \qquad \text{for each } n \in \mathbb{N}.$$

Exercise 4.5.10. Let $(f_0, f_1, f_2, \ldots)$ be the Fibonacci sequence. For this exercise, we also set $f_{-1} = 1$.

For any $n \in \mathbb{N}$ and $k \in \mathbb{Z}$, define the rational number $\binom{n}{k}_F$ (a slight variation

[79] Readers unfamiliar with complex numbers can imagine that the symbol $\mathbb{C}$ is replaced by $\mathbb{R}$ here. The solution doesn't really depend on what kind of number n is.

on the corresponding binomial coefficient) by

$$\binom{n}{k}_F = \begin{cases} \dfrac{f_n f_{n-1} \cdots f_{n-k+1}}{f_k f_{k-1} \cdots f_1}, & \text{if } n \geq k \geq 0; \\ 0, & \text{otherwise.} \end{cases}$$

(a) Prove that $\binom{n}{k}_F = \binom{n}{n-k}_F$ for any $n \in \mathbb{N}$ and $k \in \mathbb{N}$.

(b) Let n be a positive integer, and let $k \in \mathbb{N}$ be such that $n \geq k$. Prove that

$$\binom{n}{k}_F = f_{k+1}\binom{n-1}{k}_F + f_{n-k-1}\binom{n-1}{k-1}_F.$$

(c) Prove that $\binom{n}{k}_F \in \mathbb{N}$ for any $n \in \mathbb{N}$ and $k \in \mathbb{N}$.

4.6. Guessing sequences

Our next topic are sequences. We begin with a sequence of puzzle-style exercises in which a sequence is defined recursively and an explicit formula is asked for. Of course, not every recursively defined sequence has an explicit formula, so these sorts of exercises often are somewhat artificial. Yet the methods of solving them can be instructive, and problems of this kind appear not just on mathematical competitions but also in real research. Thus, we shall discuss a few such exercises with solutions.

The first sequence we consider is a common generalization of arithmetic progressions (i.e., sequences $(x_0, x_1, x_2, \ldots)$ satisfying $x_n = x_{n-1} + d$ for all $n \geq 1$, where d is a given number) and geometric progressions (i.e., sequences $(x_0, x_1, x_2, \ldots)$ satisfying $x_n = qx_{n-1}$ for all $n \geq 1$, where q is a given number):[80]

Exercise 4.6.1. Let q and d be two numbers. Let $(x_0, x_1, x_2, \ldots)$ be a sequence of numbers that satisfies the recursive equation

$$x_n = qx_{n-1} + d \qquad \text{for each } n \geq 1. \tag{137}$$

Find an explicit formula for x_n in terms of x_0, q and d.

[80] As before, the word "number" means an integer, rational, real or complex number.

Discussion of Exercise 4.6.1. We compute the first few entries of our sequence:

$$\begin{aligned}
x_0 &= x_0; \\
x_1 &= qx_0 + d; \\
x_2 &= q \underbrace{x_1}_{=qx_0+d} + d = q(qx_0 + d) + d = q^2 x_0 + qd + d; \\
x_3 &= q^3 x_0 + q^2 d + qd + d; \\
x_4 &= q^4 x_0 + q^3 d + q^2 d + qd + d.
\end{aligned}$$

Thus, we are led to guessing the equation

$$x_n = q^n x_0 + \left(q^0 + q^1 + \cdots + q^{n-1}\right) d \tag{138}$$

for each $n \in \mathbb{N}$. Once this equation is found, proving it by induction on n is completely straightforward (hence left to the reader).

We can make (138) more explicit by getting rid of the "$\cdots$", but the answer will depend on the value of q: If $q = 1$, then (138) simplifies to $x_n = x_0 + nd$. On the other hand, if $q \neq 1$, then (2) (applied to $b = q$) yields $(q-1)\left(q^0 + q^1 + \cdots + q^{n-1}\right) = q^n - 1$, so that we have $q^0 + q^1 + \cdots + q^{n-1} = \dfrac{q^n - 1}{q - 1}$, and therefore the equality (138) rewrites as

$$x_n = q^n x_0 + \frac{q^n - 1}{q - 1} d. \tag{139}$$

□

Here is a different sequence puzzle, which has appeared in the finals of the Norwegian mathematical olympiad 1994-95 (problem 1a) and the British mathematical olympiad 1996:

Exercise 4.6.2. Let $(a_1, a_2, a_3, \ldots)$ be a sequence of numbers defined recursively by $a_1 = 1$ and

$$a_1 + a_2 + \cdots + a_n = n^2 \cdot a_n \qquad \text{for all } n \geq 2. \tag{140}$$

Find an explicit formula for a_n.

Discussion of Exercise 4.6.2. Define a new sequence $(b_0, b_1, b_2, \ldots)$ by setting

$$b_n = a_1 + a_2 + \cdots + a_n \qquad \text{for each } n \in \mathbb{N}. \tag{141}$$

Then, for each positive integer n, we have

$$\begin{aligned}
\underbrace{b_n}_{=a_1+a_2+\cdots+a_n} - \underbrace{b_{n-1}}_{=a_1+a_2+\cdots+a_{n-1}} &= (a_1 + a_2 + \cdots + a_n) + (a_1 + a_2 + \cdots + a_{n-1}) \\
&= a_n.
\end{aligned} \tag{142}$$

Note also that $b_1 = a_1 = 1$.

Now, for each integer $n \geq 2$, we have

$$\begin{aligned} b_n &= a_1 + a_2 + \cdots + a_n = n^2 \cdot \underbrace{a_n}_{\substack{=b_n - b_{n-1} \\ \text{(by (142))}}} \qquad \text{(by (140))} \\ &= n^2 \cdot (b_n - b_{n-1}) = n^2 b_n - n^2 b_{n-1}, \end{aligned}$$

hence

$$n^2 b_{n-1} = n^2 b_n - b_n = \left(n^2 - 1\right) b_n = (n-1)(n+1) b_n.$$

Solving this for b_n, we find

$$b_n = \frac{n^2 b_{n-1}}{(n-1)(n+1)} = \frac{n}{n-1} \cdot \frac{n}{n+1} \cdot b_{n-1}. \tag{143}$$

This is a recursive equation for the sequence $(b_0, b_1, b_2, \ldots)$ that is much easier to work with than the original recursion (140) for the sequence $(a_1, a_2, a_3, \ldots)$; in particular, it relies only on one preceding value b_{n-1} rather than the $n-1$ values $a_1, a_2, \ldots, a_{n-1}$. This is why we introduced the sequence $(b_0, b_1, b_2, \ldots)$. The moral of the story (so far – we haven't solved the exercise yet!) is that if the sums $a_1 + a_2 + \cdots + a_n$ appear in a recursive equation, it is worth introducing a new sequence $(b_0, b_1, b_2, \ldots)$ defined by (141); then, these sums can be rewritten as b_n, whereas single entries a_n of the original sequence $(a_1, a_2, a_3, \ldots)$ can be rewritten as $b_n - b_{n-1}$ (by (142)). This way, we trade finite sums for differences of two numbers; the latter are usually easier to deal with than the former.[81]

Using (143), we can now easily compute the entries of the sequence $(b_0, b_1, b_2, \ldots)$, starting with b_1:

$$\begin{aligned} b_1 &= 1; \\ b_2 &= \frac{2}{1} \cdot \frac{2}{3} \cdot \underbrace{b_1}_{=1} = \frac{2}{1} \cdot \frac{2}{3}; \\ b_3 &= \frac{3}{2} \cdot \frac{3}{4} \cdot \underbrace{b_2}_{=\frac{2}{1} \cdot \frac{2}{3}} = \frac{3}{2} \cdot \frac{3}{4} \cdot \frac{2}{1} \cdot \frac{2}{3}; \\ b_4 &= \frac{4}{3} \cdot \frac{4}{5} \cdot \underbrace{b_3}_{=\frac{3}{2} \cdot \frac{3}{4} \cdot \frac{2}{1} \cdot \frac{2}{3}} = \frac{4}{3} \cdot \frac{4}{5} \cdot \frac{3}{2} \cdot \frac{3}{4} \cdot \frac{2}{1} \cdot \frac{2}{3}; \\ &\ldots. \end{aligned}$$

[81] Note also that not only sums of the form $a_1 + a_2 + \cdots + a_n$, but more generally any sums of the form $a_{i+1} + a_{i+2} + \cdots + a_j$ can be rewritten in a simpler form using the $b_1, b_2, b_3, \ldots$. Indeed, we have $a_{i+1} + a_{i+2} + \cdots + a_j = b_j - b_i$ for all positive integers i and j satisfying $i \leq j$.

The general rule is that

$$b_m = \prod_{n=2}^{m} \left(\frac{n}{n-1} \cdot \frac{n}{n+1} \right) \qquad \text{for each integer } m \geq 1. \tag{144}$$

(Of course, this formula is easily proved by induction on m, using (144).)

Now, for each integer $m \geq 1$, we have

$$\begin{aligned} b_m &= \prod_{n=2}^{m} \underbrace{\left(\frac{n}{n-1} \cdot \frac{n}{n+1} \right)}_{= \frac{n}{n+1} / \frac{n-1}{n}} = \prod_{n=2}^{m} \left(\frac{n}{n+1} / \frac{n-1}{n} \right) = \prod_{s=2}^{m} \left(\frac{s}{s+1} / \frac{s-1}{s} \right) \\ &= \frac{m}{m+1} / \frac{1}{2} \qquad \left(\text{by (109), applied to } u = 2,\ v = m \text{ and } a_s = \frac{s}{s+1} \right) \\ &= \frac{2m}{m+1}. \end{aligned} \tag{145}$$

It is easy to see that this equality holds for $m = 0$ as well (indeed, both b_m and $\frac{2m}{m+1}$ are 0 when $m = 0$). Hence, (145) holds for each $m \in \mathbb{N}$.

Now, for each positive integer n, we have

$$\begin{aligned} a_n &= \underbrace{b_n}_{\substack{= \frac{2n}{n+1} \\ \text{(by (145))}}} - \underbrace{b_{n-1}}_{\substack{= \frac{2(n-1)}{n} \\ \text{(by (145))}}} \qquad \text{(by (142))} \\ &= \frac{2n}{n+1} - \frac{2(n-1)}{n} = \frac{2}{n(n+1)}. \end{aligned}$$

Thus, the exercise is solved.[82] □

The next sequence-guessing exercise is [Galvin20, §2.4, Exercise 2]:

Exercise 4.6.3. Define a sequence $(a_1, a_2, a_3, \ldots)$ of rational numbers recursively by

$$a_1 = \frac{5}{2} \qquad \text{and} \qquad a_n = a_{n-1}^2 - 2 \text{ for every } n \geq 2.$$

Find an explicit formula for a_n.

[82] Try checking that our explicit answer $a_n = \frac{2}{n(n+1)}$ does indeed satisfy the recursion (140)! (The formula you should obtain is precisely (29).)

Discussion of Exercise 4.6.3. Computing the first few entries of our sequence yields

$$\begin{aligned} a_1 &= \frac{5}{2} = 2.5; \\ a_2 &= \frac{17}{4} = 4.25; \\ a_3 &= \frac{257}{16} \approx 16.063; \\ a_4 &= \frac{66\,049}{256} \approx 256.0; \\ &\ldots. \end{aligned}$$

The first thing that will catch your eye is the omnipresence of powers of 2 here – both in the denominators (where they are not surprising due to the repeated squaring in the construction of the sequence) and as approximate values. The latter is a dead giveaway: It appears that a_n is very close to a power of 2, and the difference between a_n and said power of 2 converges to 0 (rather fast) as $n \to \infty$. It is easily observed that the relevant power of 2 is $2^{2^{n-1}}$. Thus, we suspect that $a_n \approx 2^{2^{n-1}}$.

To get an exact formula for a_n, we take a look at the differences $a_n - 2^{2^{n-1}}$ (as it is always a good idea to subtract the part we know to get a closer look at the part we don't):

$$\begin{aligned} a_1 - 2^{2^0} &= \frac{5}{2} - 2 = \frac{1}{2}; \\ a_2 - 2^{2^1} &= \frac{17}{4} - 4 = \frac{1}{4}; \\ a_3 - 2^{2^2} &= \frac{257}{16} - 16 = \frac{1}{16}; \\ &\ldots. \end{aligned}$$

These suggest that $a_n - 2^{2^{n-1}} = \dfrac{1}{2^{2^{n-1}}}$, and thus

$$a_n = 2^{2^{n-1}} + \frac{1}{2^{2^{n-1}}} \qquad \text{for each } n \geq 1. \tag{146}$$

Once this formula has been guessed, its proof is a straightforward induction on n (which we leave to the reader, who can find it in [Galvin20, §2.4, Exercise 2]).

Could we have found this formula without staring at numbers? Yes, although that would have required a more sophisticated trick: A contest mathematics connoisseur will recognize the "scent of $x + \dfrac{1}{x}$" in the recursive formula $a_n = a_{n-1}^2 - 2$. What does this mean? The crucial observation is that if a number y has the form $x + \dfrac{1}{x}$ for some number $x \neq 0$, then

$$y^2 - 2 = x^2 + \frac{1}{x^2} \tag{147}$$

(check this!). Knowing this, when you see an expression of the form $y^2 - 2$ anywhere, you may want to try rewriting the y as $x + \frac{1}{x}$, so that $y^2 - 2$ becomes $x^2 + \frac{1}{x^2}$. In the case of the present problem, we thus suspect that there is a "secret" sequence $(x_1, x_2, x_3, \ldots)$ of numbers with the property that

$$a_n = x_n + \frac{1}{x_n} \qquad \text{for every } n \geq 1. \tag{148}$$

If this sequence exists, then the recursion $a_n = a_{n-1}^2 - 2$ can be rewritten as $x_n + \frac{1}{x_n} = x_{n-1}^2 + \frac{1}{x_{n-1}^2}$.

Constructing such a "secret" sequence $(x_1, x_2, x_3, \ldots)$ is easy: We find x_1 by solving the equation $\frac{5}{2} = a_1 = x_1 + \frac{1}{x_1}$ for x_1 (this boils down to a quadratic equation, whose solutions are $\frac{1}{2}$ and 2); then we set $x_n = x_{n-1}^2$ for each $n \geq 2$ (this is the easiest way to satisfy the recursion $x_n + \frac{1}{x_n} = x_{n-1}^2 + \frac{1}{x_{n-1}^2}$; of course, we could have also set $x_n = \frac{1}{x_{n-1}^2}$, but why complicate things?). This leads to

$$\begin{aligned}
x_1 &= 2 \qquad \left(\text{we could have just as well chosen } \frac{1}{2}\right); \\
x_2 &= 2^2; \\
x_3 &= \left(2^2\right)^2 = 2^4; \\
x_4 &= \left(2^4\right)^2 = 2^8; \\
&\ldots;
\end{aligned}$$

the general formula is easily seen to be $x_n = 2^{2^{n-1}}$. Hence, (148) rewrites as $a_n = 2^{2^{n-1}} + \frac{1}{2^{2^{n-1}}}$. Thus, we have discovered (146) again.

The rational expression $x + \frac{1}{x}$ is a sufficiently common occurrence in mathematics that its properties are worth remembering (cf. the Wikipedia page on the Joukowsky transform). The equality (147) is a close relative of the formula

$$\cos(2\varphi) = 2\cos^2\varphi - 1$$

from trigonometry; indeed, using complex numbers, the latter formula can easily be derived from (147) (by applying (147) to $x = e^{i\varphi}$ and $y = 2\cos\varphi$). $\square$

The following sequence-guessing exercise was problem A3 on the Putnam contest 2004 (see [GelAnd17, §3.1.1, Example]):

Exercise 4.6.4. Define a sequence $(a_0, a_1, a_2, \ldots)$ of rational numbers recursively by

$$a_0 = a_1 = a_2 = 1$$

and

$$a_n a_{n+3} - a_{n+1} a_{n+2} = n! \qquad \text{for every } n \in \mathbb{N}. \tag{149}$$

Prove that a_n is an integer for all $n \in \mathbb{N}$.

(The original statement on the Putnam contest wrote $\det \begin{pmatrix} a_n & a_{n+1} \\ a_{n+2} & a_{n+3} \end{pmatrix}$ instead of $a_n a_{n+3} - a_{n+1} a_{n+2}$, but this clearly makes no difference.)

Discussion of Exercise 4.6.4. Let us compute the first few entries of our sequence. We are given that $a_0 = a_1 = a_2 = 1$. Now, in order to find a_3, we apply (149) to $n = 0$, obtaining $a_0 a_3 - a_1 a_2 = 0! = 1$. Solving this for a_3 (using $a_0 = a_1 = a_2 = 1$) yields $a_3 = 2$. Next, in order to find a_4, we apply (149) to $n = 1$, obtaining $a_1 a_4 - a_2 a_3 = 1! = 1$. Solving this for a_4 (using $a_1 = a_2 = 1$ and $a_3 = 2$) yields $a_4 = 3$. Going further in the same vein, we find $a_5 = 8$ and $a_6 = 15$ and $a_7 = 48$ and $a_8 = 105$.

Do we see any pattern? Perhaps not yet. But with these entries being integers, we can try to analyze them in a way only integers can be analyzed: by factoring them into primes.[83] The prime factorizations of the entries $a_3, a_4, \ldots, a_8$ are

$$a_3 = 2, \qquad a_4 = 3, \qquad a_5 = 2^3, \qquad a_6 = 3 \cdot 5, \qquad a_7 = 2^4 \cdot 3, \qquad a_8 = 3 \cdot 5 \cdot 7.$$

Now a pattern in the a_4, a_6, a_8 meets the eye: It appears that

$$a_{2m} = \underbrace{1 \cdot 3 \cdot 5 \cdot \cdots \cdot (2m-1)}_{\substack{\text{the product of the first } m \\ \text{odd positive integers}}} \qquad \text{for each } m \in \mathbb{N}. \tag{150}$$

The odd-indexed values a_{2m+1} are disguising somewhat better. Unlike the integers a_{2m}, which are all odd, these are even (at least as far as we have computed them). Moreover, they seem to be multiples of larger and larger powers of 2. More precisely, a_{2m+1} appears to be a multiple of 2^m; thus, we can try to look at $\dfrac{a_{2m+1}}{2^m}$ in order to pin down the other factor. We see that

$$\frac{a_3}{2^1} = 1, \qquad \frac{a_5}{2^2} = 2, \qquad \frac{a_7}{2^3} = 6, \qquad \frac{a_9}{2^4} = 24.$$

These look familiar – aren't these the factorials? Thus, we are led to conjecturing that

$$a_{2m+1} = 2^m \cdot m! = \underbrace{2 \cdot 4 \cdot 6 \cdot \cdots \cdot (2m)}_{\substack{\text{the product of the first } m \\ \text{even positive integers}}} \qquad \text{for each } m \in \mathbb{N} \tag{151}$$

[83] We won't properly discuss prime factorization until Section 9.2, but for now let us take it for granted.

(where the last equality sign is an easy consequence of the definition of $m!$ and (107)). Another way to guess this formula (151) would have been to observe experimentally that $a_3 \mid a_5 \mid a_7 \mid a_9$ and thus it is reasonable to look at the consecutive quotients $\frac{a_5}{a_3}, \frac{a_7}{a_5}, \frac{a_9}{a_7}$. (The same reasoning could have been used to come up with (150), if we hadn't thought of using prime factorization.)

At this point, the two formulas we have guessed – that is, (150) and (151) – cover all entries of our sequence $(a_0, a_1, a_2, \ldots)$; thus, we can try proving them by induction. Note that proving any one of these two formulas alone would be difficult: For example, if we tried to prove (150) by induction on m, then we could not get much mileage out of our recursion (149), since it would require us to know a_{2m-1} and a_{2m-3}, which would not be covered by (150). Likewise, if we tried to prove (151) by induction on m, then we would need to know a_{2m} and a_{2m-2}, which (151) could not help us compute. However, (150) and (151) complement each other so neatly that proving them **together** is a straightforward induction exercise. Let us do this in detail just to drive this point home:

Claim 1: For each $m \in \mathbb{N}$, the equalities (150) and (151) hold.

[*Proof of Claim 1:* We proceed by strong induction on m:

Induction step: Let $k \in \mathbb{N}$. Assume (as the induction hypothesis) that the equalities (150) and (151) hold for all $m < k$. We must prove that the equalities (150) and (151) hold for $m = k$. In other words, we must prove that

$$a_{2k} = 1 \cdot 3 \cdot 5 \cdot \cdots \cdot (2k-1) \tag{152}$$

and

$$a_{2k+1} = 2 \cdot 4 \cdot 6 \cdot \cdots \cdot (2k). \tag{153}$$

If $k \leq 1$, then this is straightforward. Thus, we WLOG assume that $k > 1$. Hence, $2k > 2$, so that $2k \geq 3$. Now, our induction hypothesis yields that the equalities (150) and (151) hold for $m = k-2$. In other words, we have

$$a_{2k-4} = 1 \cdot 3 \cdot 5 \cdot \cdots \cdot (2k-5)$$

and

$$a_{2k-3} = 2 \cdot 4 \cdot 6 \cdot \cdots \cdot (2k-4). \tag{154}$$

Also, our induction hypothesis yields that the equalities (150) and (151) hold for $m = k-1$. In other words, we have

$$a_{2k-2} = 1 \cdot 3 \cdot 5 \cdot \cdots \cdot (2k-3) \tag{155}$$

and

$$a_{2k-1} = 2 \cdot 4 \cdot 6 \cdot \cdots \cdot (2k-2). \tag{156}$$

Now, (149) (applied to $n = 2k-3$) yields

$$a_{2k-3}a_{2k} - a_{2k-2}a_{2k-1} = (2k-3)!,$$

so that

$$\begin{aligned}
a_{2k-3}a_{2k} &= (2k-3)! + \underbrace{a_{2k-2}}_{\substack{=1\cdot3\cdot5\cdot\cdots\cdot(2k-3)\\ \text{(by (155))}}} \underbrace{a_{2k-1}}_{\substack{=2\cdot4\cdot6\cdot\cdots\cdot(2k-2)\\ \text{(by (156))}}} \\
&= (2k-3)! + \underbrace{(1\cdot3\cdot5\cdot\cdots\cdot(2k-3))}_{\substack{=\prod\limits_{\substack{i\in\{1,2,\ldots,2k-2\};\\ i\text{ is odd}}} i}} \cdot \underbrace{(2\cdot4\cdot6\cdot\cdots\cdot(2k-2))}_{\substack{=\prod\limits_{\substack{i\in\{1,2,\ldots,2k-2\};\\ i\text{ is even}}} i}} \\
&= (2k-3)! + \underbrace{\left(\prod_{\substack{i\in\{1,2,\ldots,2k-2\};\\ i\text{ is odd}}} i\right) \cdot \left(\prod_{\substack{i\in\{1,2,\ldots,2k-2\};\\ i\text{ is even}}} i\right)}_{\substack{=\prod\limits_{i\in\{1,2,\ldots,2k-2\}} i\\ \text{(by (110))}}} = (2k-3)! + \underbrace{\prod_{i\in\{1,2,\ldots,2k-2\}} i}_{\substack{=1\cdot2\cdot\cdots\cdot(2k-2)\\ =(2k-2)!\\ =(2k-2)\cdot(2k-3)!}} \\
&= (2k-3)! + (2k-2)\cdot(2k-3)! \\
&= \underbrace{(1+(2k-2))}_{=2k-1}\cdot(2k-3)! = (2k-1)\cdot(2k-3)!.
\end{aligned}$$

Comparing this with

$$\begin{aligned}
&\underbrace{a_{2k-3}}_{\substack{=2\cdot4\cdot6\cdot\cdots\cdot(2k-4)\\ \text{(by (154))}}} \cdot \underbrace{(1\cdot3\cdot5\cdot\cdots\cdot(2k-1))}_{=(1\cdot3\cdot5\cdot\cdots\cdot(2k-3))\cdot(2k-1)} \\
&= \underbrace{(2\cdot4\cdot6\cdot\cdots\cdot(2k-4))}_{\substack{=\prod\limits_{\substack{i\in\{1,2,\ldots,2k-3\};\\ i\text{ is even}}} i}} \cdot \underbrace{(1\cdot3\cdot5\cdot\cdots\cdot(2k-3))}_{\substack{=\prod\limits_{\substack{i\in\{1,2,\ldots,2k-3\};\\ i\text{ is odd}}} i}} \cdot (2k-1) \\
&= \underbrace{\left(\prod_{\substack{i\in\{1,2,\ldots,2k-3\};\\ i\text{ is even}}} i\right) \cdot \left(\prod_{\substack{i\in\{1,2,\ldots,2k-3\};\\ i\text{ is odd}}} i\right)}_{\substack{=\prod\limits_{i\in\{1,2,\ldots,2k-3\}} i\\ \text{(by (110))}}} \cdot (2k-1) = \underbrace{\left(\prod_{i\in\{1,2,\ldots,2k-3\}} i\right)}_{\substack{=1\cdot2\cdot\cdots\cdot(2k-3)\\ =(2k-3)!}} \cdot (2k-1) \\
&= (2k-3)!\cdot(2k-1) = (2k-1)\cdot(2k-3)!,
\end{aligned}$$

we obtain

$$a_{2k-3}a_{2k} = a_{2k-3}\cdot(1\cdot3\cdot5\cdot\cdots\cdot(2k-1)).$$

We can cancel the factor a_{2k-3} from this equality (since $a_{2k-3} = 2\cdot4\cdot6\cdot\cdots\cdot(2k-4) \neq 0$), and thus find $a_{2k} = 1\cdot3\cdot5\cdot\cdots\cdot(2k-1)$. This proves (152).

Furthermore, (149) (applied to $n = 2k-2$) yields

$$a_{2k-2}a_{2k+1} - a_{2k-1}a_{2k} = (2k-2)!,$$

so that

$$\begin{aligned}
a_{2k-2}a_{2k+1} &= (2k-2)! + \underbrace{a_{2k-1}}_{\substack{=2\cdot4\cdot6\cdot\cdots\cdot(2k-2)\\ \text{(by (156))}}} \underbrace{a_{2k}}_{\substack{=1\cdot3\cdot5\cdot\cdots\cdot(2k-1)\\ \text{(by (152))}}} \\
&= (2k-2)! + \underbrace{(2\cdot4\cdot6\cdot\cdots\cdot(2k-2))}_{\substack{=\prod\limits_{\substack{i\in\{1,2,\ldots,2k-1\};\\ i\text{ is even}}} i}} \cdot \underbrace{(1\cdot3\cdot5\cdot\cdots\cdot(2k-1))}_{\substack{=\prod\limits_{\substack{i\in\{1,2,\ldots,2k-1\};\\ i\text{ is odd}}} i}} \\
&= (2k-2)! + \underbrace{\left(\prod_{\substack{i\in\{1,2,\ldots,2k-1\};\\ i\text{ is even}}} i\right) \cdot \left(\prod_{\substack{i\in\{1,2,\ldots,2k-1\};\\ i\text{ is odd}}} i\right)}_{\substack{=\prod\limits_{i\in\{1,2,\ldots,2k-1\}} i\\ \text{(by (110))}}} = (2k-2)! + \underbrace{\prod_{i\in\{1,2,\ldots,2k-1\}} i}_{\substack{=1\cdot2\cdot\cdots\cdot(2k-1)\\ =(2k-1)!\\ =(2k-1)\cdot(2k-2)!}} \\
&= (2k-2)! + (2k-1)\cdot(2k-2)! \\
&= \underbrace{(1+(2k-1))}_{=2k}\cdot(2k-2)! = 2k\cdot(2k-2)!.
\end{aligned}$$

Comparing this with

$$\begin{aligned}
&\underbrace{a_{2k-2}}_{\substack{=1\cdot3\cdot5\cdot\cdots\cdot(2k-3)\\ \text{(by (155))}}} \cdot \underbrace{(2\cdot4\cdot6\cdot\cdots\cdot(2k))}_{=(2\cdot4\cdot6\cdot\cdots\cdot(2k-2))\cdot(2k)} \\
&= \underbrace{(1\cdot3\cdot5\cdot\cdots\cdot(2k-3))}_{\substack{=\prod\limits_{\substack{i\in\{1,2,\ldots,2k-2\};\\ i\text{ is odd}}} i}} \cdot \underbrace{(2\cdot4\cdot6\cdot\cdots\cdot(2k-2))}_{\substack{=\prod\limits_{\substack{i\in\{1,2,\ldots,2k-2\};\\ i\text{ is even}}} i}} \cdot (2k) \\
&= \underbrace{\left(\prod_{\substack{i\in\{1,2,\ldots,2k-2\};\\ i\text{ is odd}}} i\right) \cdot \left(\prod_{\substack{i\in\{1,2,\ldots,2k-2\};\\ i\text{ is even}}} i\right)}_{\substack{=\prod\limits_{i\in\{1,2,\ldots,2k-2\}} i\\ \text{(by (110))}}} \cdot (2k) = \underbrace{\left(\prod_{i\in\{1,2,\ldots,2k-2\}} i\right)}_{\substack{=1\cdot2\cdot\cdots\cdot(2k-2)\\ =(2k-2)!}} \cdot (2k) \\
&= (2k-2)!\cdot(2k) = 2k\cdot(2k-2)!,
\end{aligned}$$

we obtain

$$a_{2k-2}a_{2k+1} = a_{2k-2}\cdot(2\cdot4\cdot6\cdot\cdots\cdot(2k)).$$

We can cancel the factor a_{2k-2} from this equality (since $a_{2k-2} = 1\cdot3\cdot5\cdot\cdots\cdot(2k-3) \neq 0$), and thus find $a_{2k+1} = 2\cdot4\cdot6\cdot\cdots\cdot(2k)$. This proves (153).

Thus, both (152) and (153) are proved. This completes the induction step. Thus, Claim 1 is proved.]

Obviously, Claim 1 entails that a_{2m} and a_{2m+1} are integers for each $m \in \mathbb{N}$. Hence, a_n is an integer for each $n \in \mathbb{N}$ (since each $n \in \mathbb{N}$ has the form $2m$ or $2m+1$ for some $m \in \mathbb{N}$). This solves Exercise 4.6.4.

[*Remark:* Using Exercise 4.2.3, we could have rewritten (150) as $a_{2m} = \dfrac{(2m)!}{2^m m!}$. This would have made (150) shorter and a bit easier to work with, but it would have obscured the fact that a_{2m} is an integer.

We could have also combined the two formulas (150) and (151) into a single formula

$$a_n = (n-1) \cdot (n-3) \cdot (n-5) \cdot \cdots ,$$

where the product on the right hand side is understood to end at 1 if n is even and at 2 if n is odd.] □

4.7. Periodicity

4.7.1. Periodic sequences

Next, we shall discuss a property that some sequences have: *periodicity*. Here is a way to define it:

Definition 4.7.1. Let $u = (u_0, u_1, u_2, \ldots)$ be an infinite sequence (of any kinds of objects – e.g., of numbers).

(a) A positive integer d is said to be a *period* of u if every $i \in \mathbb{N}$ satisfies $u_i = u_{i+d}$.

(b) The sequence u is said to be *periodic* if it has a period (i.e., if a period of u exists).

(c) Let d be a positive integer. The sequence u is said to be *d-periodic* if d is a period of u.

Example 4.7.2. Let $u = (u_0, u_1, u_2, \ldots)$ be an infinite sequence. Then, 1 is a period of u if every $i \in \mathbb{N}$ satisfies $u_i = u_{i+1}$ (by Definition 4.7.1 **(a)**). In other words, 1 is a period of u if $u_0 = u_1 = u_2 = \cdots$. In other words, 1 is a period of u if all entries of u are equal. Sequences whose all entries are equal are said to be *constant*; thus, the 1-periodic sequences are precisely the constant sequences.

Example 4.7.3. Let u be the sequence $\left((-1)^0, (-1)^1, (-1)^2, (-1)^3, \ldots\right) = (1, -1, 1, -1, 1, -1, \ldots)$.

Then, 2 is a period of u, since every $i \in \mathbb{N}$ satisfies $(-1)^i = (-1)^{i+2}$. Thus, the sequence u is 2-periodic and periodic.

On the other hand, 1 is not a period of u, because not every $i \in \mathbb{N}$ satisfies $(-1)^i = (-1)^{i+1}$. (Actually, no $i \in \mathbb{N}$ satisfies this.)

Example 4.7.4. Let u be the sequence $\left(0^0, 0^1, 0^2, 0^3, \ldots\right) = (1, 0, 0, 0, \ldots)$.

Then, u has no period. In fact, if d was a period of u, then every $i \in \mathbb{N}$ would satisfy $0^i = 0^{i+d}$; but this cannot hold for $i = 0$ (because in this case, $0^i = 0^0 = 1$ but $0^{i+d} = 0^{(\text{something positive})} = 0$). Thus, u is not periodic.

On the other hand, if we remove the first entry of u, then u becomes the sequence $(0,0,0,0,\ldots)$, which is periodic (and even 1-periodic). The sequence u is therefore "periodic except for a few early entries". Such sequences are called *eventually periodic*, but we will not have much use for this word.

Example 4.7.5. Complex numbers give a simple example of a 4-periodic sequence: If i denotes the complex number $\sqrt{-1}$, then the sequence $\left(i^0, i^1, i^2, i^3, \ldots\right) = (1, i, -1, -i, 1, i, -1, -i, 1, i, -1, -i, \ldots)$ is 4-periodic, since $i^4 = 1$.

Example 4.7.6. Let n be a positive integer. The sequence

$$\begin{aligned}&(0\%n, 1\%n, 2\%n, 3\%n, \ldots)\\ &= (0,1,2,\ldots,n-1,0,1,2,\ldots,n-1,0,1,2,\ldots,n-1,\ldots)\end{aligned}$$

is n-periodic. Indeed, every $i \in \mathbb{N}$ satisfies $i\%n = (i+n)\%n$.

Example 4.7.7. Let us define a sequence $(g_0, g_1, g_2, \ldots)$ of integers recursively by

$$g_0 = 0, \qquad g_1 = 1, \qquad \text{and } g_n = g_{n-1} - g_{n-2} \text{ for all } n \geq 2.$$

Note that this differs from the definition of the Fibonacci sequence only in a single sign. Here is a table of the first few entries of this new sequence:

n	0	1	2	3	4	5	6	7	8	9	10	11	12
g_n	0	1	1	0	-1	-1	0	1	1	0	-1	-1	0

.

What a difference a sign can make! We observe from the table that the sequence $(g_0, g_1, g_2, \ldots)$ is 6-periodic (i.e., its entries repeat every 6 terms: that is, $g_i = g_{i+6}$ for each $i \in \mathbb{N}$), and can be rewritten explicitly as

$$g_n = [3 \nmid n]\,(-1)^{n//3} \qquad \text{for each } n \in \mathbb{N}$$

using the Iverson bracket notation (Definition 4.3.19). All of this can be proved by a straightforward strong induction on n.

The following theorem gives basic properties of periods of sequences:

Theorem 4.7.8. Let $u = (u_0, u_1, u_2, \ldots)$ be an infinite sequence (of any kinds of objects – e.g., of numbers). Then:

(a) If a and b are two periods of u, then $a+b$ is a period of u.
(b) If a and b are two periods of u such that $a > b$, then $a-b$ is a period of u.
(c) If a is a period of u, then na is a period of u for every positive integer n.
(d) If a and b are two periods of u, then $\gcd(a,b)$ is a period of u.

(e) If a is a period of u, and if p and q are two nonnegative integers satisfying $p \equiv q \bmod a$, then $u_p = u_q$.

Proof of Theorem 4.7.8. **(a)** Let a and b be two periods of u. We must prove that $a+b$ is a period of u.

Recall that a is a period of u if and only if every $i \in \mathbb{N}$ satisfies $u_i = u_{i+a}$ (by Definition 4.7.1 **(a)**). Hence, every $i \in \mathbb{N}$ satisfies

$$u_i = u_{i+a} \tag{157}$$

(since a is a period of u). The same argument (applied to b instead of a) shows that every $i \in \mathbb{N}$ satisfies

$$u_i = u_{i+b}. \tag{158}$$

Now, every $i \in \mathbb{N}$ satisfies

$$\begin{aligned} u_i &= u_{i+a} && \text{(by (157))} \\ &= u_{(i+a)+b} && \text{(by (158), applied to } i+a \text{ instead of } i) \\ &= u_{i+(a+b)} && \text{(since } (i+a)+b = i+(a+b)). \end{aligned}$$

But $a+b$ is a period of u if and only if every $i \in \mathbb{N}$ satisfies $u_i = u_{i+(a+b)}$ (by Definition 4.7.1 **(a)**). Hence, $a+b$ is a period of u (since every $i \in \mathbb{N}$ satisfies $u_i = u_{i+(a+b)}$). This proves Theorem 4.7.8 **(a)**.

(b) Let a and b be two periods of u such that $a > b$. We must prove that $a-b$ is a period of u.

Every $i \in \mathbb{N}$ satisfies

$$u_i = u_{i+a} \tag{159}$$

and

$$u_i = u_{i+b}. \tag{160}$$

(Indeed, these two equalities are precisely the two equalities (157) and (158) that were already proved in the proof of Theorem 4.7.8 **(a)** above.)

Now, let $i \in \mathbb{N}$. Then, $i + \underbrace{a}_{>b} - b > i + b - b = i \geq 0$ (since $i \in \mathbb{N}$), so that $i + a - b \in \mathbb{N}$. Hence, applying (160) to $i+a-b$ instead of i, we obtain $u_{i+a-b} = u_{(i+a-b)+b} = u_{i+a}$. Comparing this with (159), we obtain $u_i = u_{i+a-b} = u_{i+(a-b)}$.

Forget that we fixed i. We thus have shown that every $i \in \mathbb{N}$ satisfies $u_i = u_{i+(a-b)}$.

Now, notice that $a - b > 0$ (since $a > b$). Hence, $a-b$ is a positive integer. Thus, $a-b$ is a period of u if and only if every $i \in \mathbb{N}$ satisfies $u_i = u_{i+(a-b)}$ (by Definition 4.7.1 **(a)**). Hence, $a-b$ is a period of u (since every $i \in \mathbb{N}$ satisfies $u_i = u_{i+(a-b)}$). This proves Theorem 4.7.8 **(b)**.

(c) This is a straightforward induction on n, using Theorem 4.7.8 **(a)** in the induction step.[84]

[84] Here are the details:

(d) Let a and b be two periods of u. Then, a and b are two positive integers (since a period must be a positive integer by definition). Hence, Exercise 3.7.7 shows that there exist **positive** integers x and y such that $\gcd(a,b) = xa - yb$. Consider these x and y. Theorem 4.7.8 **(c)** (applied to $n = x$) yields that xa is a period of u. Also, Theorem 4.7.8 **(c)** (applied to y and b instead of n and a) yields that yb is a period of u.

Note that $\gcd(a,b)$ is itself a positive integer (by Proposition 3.4.3 **(b)**). Hence, $\gcd(a,b) > 0$, so that $xa - yb = \gcd(a,b) > 0$ and thus $xa > yb$. Therefore, Theorem 4.7.8 **(b)** (applied to xa and yb instead of a and b) shows that $xa - yb$ is a period of u. In other words, $\gcd(a,b)$ is a period of u (since $\gcd(a,b) = xa - yb$). This proves Theorem 4.7.8 **(d)**.

(e) Let a be a period of u, and let p and q are two nonnegative integers satisfying $p \equiv q \bmod a$. We must prove that $u_p = u_q$.

From $p \equiv q \bmod a$, we obtain $q \equiv p \bmod a$ (by Proposition 3.2.6 **(c)**). Hence, p and q play symmetric roles in our situation (and, of course, also in the claim $u_p = u_q$ that we need to prove). Thus, we can WLOG assume that $p \geq q$ (since otherwise, we can simply swap p with q). Assume this.

We must prove that $u_p = u_q$. If $p = q$, then this is obvious. Hence, for the rest of this proof, we WLOG assume that $p \neq q$. Hence, $p > q$ (since $p \neq q$ and $p \geq q$).

We have $p \equiv q \bmod a$. In other words, $a \mid p - q$. In other words, there exists an integer c such that $p - q = ac$. Consider this c. We have $ac = p - q > 0$. Since a is positive (because a is a period of u), we can divide this inequality by a, and thus obtain $c > 0$. Hence, c is a positive integer. Thus, Theorem 4.7.8 **(c)** (applied to $n = c$) shows that ca is a period of u.

However, ca is a period of u if and only if every $i \in \mathbb{N}$ satisfies $u_i = u_{i+ca}$ (by Definition 4.7.1 **(a)**). Thus, every $i \in \mathbb{N}$ satisfies $u_i = u_{i+ca}$ (since ca is a period of u). Applying this to $i = q$, we find $u_q = u_{q+ca} = u_p$ (since $q + ca = p$ (because $ca = ac = p - q$)). In other words, $u_p = u_q$. Hence, Theorem 4.7.8 **(e)** is proved. □

As a consequence of Theorem 4.7.8, the periods of any given periodic sequence have a rather simple structure:

Let a be a period of u. We must show that

$$na \text{ is a period of } u \tag{161}$$

for every positive integer n.

We shall prove (161) by induction on n:

Induction base: We know that a is a period of u. In other words, $1a$ is a period of u (since $1a = a$). In other words, (161) holds for $n = 1$.

Induction step: Let m be a positive integer. Assume (as the induction hypothesis) that (161) holds for $n = m$. We must prove that (161) holds for $n = m + 1$.

We have assumed that (161) holds for $n = m$. In other words, ma is a period of u. Hence, Theorem 4.7.8 **(a)** (applied to $b = ma$) yields that $a + ma$ is a period of u. In other words, $(m+1)a$ is a period of u (since $a + ma = (m+1)a$). In other words, (161) holds for $n = m + 1$. This completes the induction step. Thus, (161) is proved. In other words, Theorem 4.7.8 **(c)** is proved.

Corollary 4.7.9. Let u be a periodic infinite sequence. Let m be the **smallest** period of u (that is, the smallest element of the set $\{\text{periods of } u\}$). Then,

$$\{\text{periods of } u\} = \{\text{positive multiples of } m\}.$$

Proof of Corollary 4.7.9. Write the sequence u as $u = (u_0, u_1, u_2, \ldots)$.

The number m is a period of u, and thus is a positive integer. Hence, $m > 0$, so that $m \neq 0$ and $|m| = m$.

We shall prove the following two claims separately:

Claim 1: We have $\{\text{periods of } u\} \subseteq \{\text{positive multiples of } m\}$.

Claim 2: We have $\{\text{positive multiples of } m\} \subseteq \{\text{periods of } u\}$.

[*Proof of Claim 1:* Let $a \in \{\text{periods of } u\}$. We will show that $a \in \{\text{positive multiples of } m\}$.

We have $a \in \{\text{periods of } u\}$. In other words, a is a period of u. Also, m is a period of u. Hence, Theorem 4.7.8 **(d)** (applied to $b = m$) shows that $\gcd(a, m)$ is a period of u. Furthermore, Proposition 3.4.4 **(f)** (applied to $b = m$) yields $\gcd(a, m) \mid a$ and $\gcd(a, m) \mid m$. From the latter divisibility, we easily obtain $\gcd(a, m) \leq m$ [85].

But m is the **smallest** period of u. Hence, every period b of u satisfies $b \geq m$. Applying this to $b = \gcd(a, m)$, we obtain $\gcd(a, m) \geq m$ (since $\gcd(a, m)$ is a period of u). Combining this with $\gcd(a, m) \leq m$, we find $\gcd(a, m) = m$. Hence, $m = \gcd(a, m) \mid a$. In other words, a is a multiple of m. Since a is positive (because a is a period of u), we thus conclude that a is a positive multiple of m. In other words, $a \in \{\text{positive multiples of } m\}$.

Forget that we fixed a. We thus have shown that $a \in \{\text{positive multiples of } m\}$ for each $a \in \{\text{periods of } u\}$. In other words, $\{\text{periods of } u\} \subseteq \{\text{positive multiples of } m\}$. This proves Claim 1.]

[*Proof of Claim 2:* Let $a \in \{\text{positive multiples of } m\}$. We will show that $a \in \{\text{periods of } u\}$.

We have $a \in \{\text{positive multiples of } m\}$. In other words, a is a positive multiple of m. Hence, a is a multiple of m; in other words, there exists an integer c such that $a = mc$. Consider this c.

We have $mc = a > 0$ (since a is positive). We can divide this inequality by m (since $m > 0$), and thus find $c > 0$. Hence, c is a positive integer. Thus, Theorem 4.7.8 **(c)** (applied to m and c instead of a and n) yields that cm is a period of u. In other words, $cm \in \{\text{periods of } u\}$. In view of $cm = mc = a$, this rewrites as $a \in \{\text{periods of } u\}$.

[85] *Proof.* We know that $\gcd(a, m)$ is nonnegative (by Proposition 3.4.3 **(a)**), so that $|\gcd(a, m)| = \gcd(a, m)$.

But Proposition 3.1.3 **(b)** (applied to $\gcd(a, m)$ and m instead of a and b) yields $|\gcd(a, m)| \leq |m|$ (since $\gcd(a, m) \mid m$ and $m \neq 0$). In view of $|\gcd(a, m)| = \gcd(a, m)$ and $|m| = m$, this rewrites as $\gcd(a, m) \leq m$.

Forget that we fixed a. We thus have shown that $a \in \{\text{periods of } u\}$ for each $a \in \{\text{positive multiples of } m\}$. In other words, $\{\text{positive multiples of } m\} \subseteq \{\text{periods of } u\}$. This proves Claim 2.]

Combining Claim 1 with Claim 2, we obtain $\{\text{periods of } u\} = \{\text{positive multiples of } m\}$. This proves Corollary 4.7.9. $\square$

Note that Theorem 4.7.8 and Corollary 4.7.9 rely substantially on the assumption that the sequence $(u_0, u_1, u_2, \ldots)$ be infinite. A theory of periodic finite sequences exists, but is less well-behaved. For example, the 6-tuple $(0, 1, 0, 0, 1, 0)$ is 3-periodic (in the obvious meaning of this word) and 5-periodic, but not $\gcd(3, 5)$-periodic (since its entries are not all equal); thus, the most natural analogue of Theorem 4.7.8 **(d)** for finite sequences does not hold. The same applies to Theorem 4.7.8 **(b)**; indeed, the 5-tuple $(0, 0, 1, 0, 0)$ is 4-periodic and 3-periodic but not $(4-3)$-periodic.

4.7.2. Periodic functions on $\mathbb{R}$ and on $\mathbb{R}_+$

Infinite sequences can be viewed as functions from $\mathbb{N}$: Namely, an infinite sequence $(u_0, u_1, u_2, \ldots)$ of real numbers can be identified with the function $\mathbb{N} \to \mathbb{R}$ that sends each $i \in \mathbb{N}$ to u_i. (Likewise for sequences of other objects.) Thus, we can view infinite sequences as analogues of functions from $\mathbb{R}$ or from $\mathbb{R}_+ = \{\text{positive reals}\}$ or from other sets. Under this analogy, the i-th entry u_i of a sequence $u = (u_0, u_1, u_2, \ldots)$ corresponds to the value $u(i)$ of a function u. The analogue of Definition 4.7.1 takes the following form:

Definition 4.7.10. Let $\mathbb{A}$ be either the set $\mathbb{R}$ or the set $\mathbb{R}_+ := \{\text{positive reals}\}$. Let S be any set. Let $u : \mathbb{A} \to S$ be a function.

(a) A positive real d is said to be a *period* of u if every $x \in \mathbb{A}$ satisfies $u(x) = u(x+d)$.

(b) The function u is said to be *periodic* if it has a period (i.e., if a period of u exists).

(c) Let d be a positive real. The function u is said to be *d-periodic* if d is a period of u.

Example 4.7.11. The most famous examples of periodic functions are the trigonometric functions $\sin : \mathbb{R} \to \mathbb{R}$ and $\cos : \mathbb{R} \to \mathbb{R}$. Both of these functions are 2π-periodic (i.e., have period 2π), since every $x \in \mathbb{R}$ satisfies $\sin x = \sin(x + 2\pi)$ and $\cos x = \cos(x + 2\pi)$.

Example 4.7.12. For each $x \in \mathbb{R}$, the real $x - \lfloor x \rfloor$ is known as the *fractional part* of x; it satisfies $0 \leq x - \lfloor x \rfloor < 1$. (Some authors denote it by $\{x\}$, though this notation clashes with the set builder notation.) Now, the function

$$\begin{aligned} &\mathbb{R} \to \mathbb{R}, \\ &x \mapsto x - \lfloor x \rfloor \end{aligned}$$

is 1-periodic. (Check this!)

Example 4.7.13. The *delta function* δ_0 is the function $\mathbb{R} \to \mathbb{R}$ that sends each $x \in \mathbb{R}$ to $[x = 0]$ (where we are using Definition 4.3.19). Thus, it sends 0 to 1, but sends all nonzero reals to 0. It is easy to see that this function δ_0 is not periodic.

The following is an analogue of Theorem 4.7.8 for functions (to the extent such an analogue is possible):

Theorem 4.7.14. Let $\mathbb{A}$ be either the set $\mathbb{R}$ or the set $\mathbb{R}_+ := \{\text{positive reals}\}$. Let S be any set. Let $u : \mathbb{A} \to S$ be a function.

(a) If a and b are two periods of u, then $a + b$ is a period of u.

(b) If a and b are two periods of u such that $a > b$, then $a - b$ is a period of u.

(c) If a is a period of u, then na is a period of u for every positive integer n.

(d) If a and b are two positive integers that are periods of u, then $\gcd(a, b)$ is a period of u.

(e) If a is a period of u, and if p and q are two elements of $\mathbb{A}$ such that $p - q$ is an integer multiple of a, then $u(p) = u(q)$.

Here, an "integer multiple of a" means a number of the form ac for some $c \in \mathbb{Z}$.

Proof of Theorem 4.7.14. Each part of Theorem 4.7.14 is an analogue of the corresponding part of Theorem 4.7.8 (because the statement "$p - q$ is an integer multiple of a" is an analogue of "$p \equiv q \bmod a$"). Its proof is analogous to the proof of Theorem 4.7.8 as well (check this!). □

A useful property of a-periodic functions is that they are uniquely determined by their values on any given half-open interval of length a. More precisely:

Proposition 4.7.15. Let $\mathbb{A}$ be either the set $\mathbb{R}$ or the set $\mathbb{R}_+ := \{\text{positive reals}\}$. Let S be any set. Let a be a positive real. Let $u : \mathbb{A} \to S$ and $v : \mathbb{A} \to S$ be two a-periodic functions. Let $b \in \mathbb{A}$. Assume that

$$u(x) = v(x) \qquad \text{for each } x \in [b, b + a). \tag{162}$$

Then, $u = v$.

Proof of Proposition 4.7.15. Let $y \in \mathbb{A}$. We shall show that $u(y) = v(y)$.

We shall find an $x \in [b, b + a)$ such that $y - x$ is an integer multiple of a. Namely, we set

$$x := y - a \left\lfloor \frac{y - b}{a} \right\rfloor.$$

Then, it is straightforward to see that $x \in [b, b + a)$ [86]. Thus, (162) shows that $u(x) = v(x)$.

[86] *Proof.* The chain of inequalities (1) (applied to $\frac{y-b}{a}$ instead of x) yields $\left\lfloor \frac{y-b}{a} \right\rfloor \leq \frac{y-b}{a} < \left\lfloor \frac{y-b}{a} \right\rfloor + 1$. We can multiply this chain of inequalities by a (since a is positive), and thus obtain

Furthermore, from $x = y - a\left\lfloor \frac{y-b}{a} \right\rfloor$, we obtain $y - x = a\left\lfloor \frac{y-b}{a} \right\rfloor$. Thus, $y - x$ is an integer multiple of a (since $\left\lfloor \frac{y-b}{a} \right\rfloor$ is an integer). But the function u is a-periodic; in other words, a is a period of u. Hence, Theorem 4.7.14 **(e)** (applied to $p = y$ and $q = x$) yields that $u(y) = u(x)$ (since $y - x$ is an integer multiple of a). In other words, $u(x) = u(y)$. The same argument (applied to v instead of u) yields $v(x) = v(y)$. Now, $u(y) = u(x) = v(x) = v(y)$.

Forget that we fixed y. We thus have shown that $u(y) = v(y)$ for each $y \in \mathbb{A}$. In other words, $u = v$. This proves Proposition 4.7.15. □

An analogue of Proposition 4.7.15 exists for periodic sequences.

Proposition 4.7.16. Let a be a positive integer. Let $u = (u_0, u_1, u_2, \ldots)$ and $v = (v_0, v_1, v_2, \ldots)$ be two a-periodic infinite sequences. Let $b \in \mathbb{N}$. Assume that

$$u_i = v_i \qquad \text{for each } i \in \{b, b+1, \ldots, b+a-1\}.$$

Then, $u = v$.

Proof of Proposition 4.7.16. Analogous to the proof of Proposition 4.7.15. □

We can use Proposition 4.7.15 to give a simple new solution to Exercise 1.1.3:

Second solution to Exercise 1.1.3 (sketched). Forget that we fixed x. Set $a = \frac{1}{n}$ and $b = 0$. Note that a is a positive real. Define a function $u : \mathbb{R} \to \mathbb{R}$ by setting

$$u(x) = \sum_{k=0}^{n-1} \left\lfloor x + \frac{k}{n} \right\rfloor - \lfloor nx \rfloor \qquad \text{for each } x \in \mathbb{R}.$$

$a\left\lfloor \frac{y-b}{a} \right\rfloor \leq y - b < a\left(\left\lfloor \frac{y-b}{a} \right\rfloor + 1\right)$. Now, combining

$$x = y - \underbrace{a\left\lfloor \frac{y-b}{a} \right\rfloor}_{\leq y-b} \geq y - (y-b) = b$$

with

$$\begin{aligned} x &= \underbrace{y}_{=y-b+b} - a\left\lfloor \frac{y-b}{a} \right\rfloor = \underbrace{y-b}_{<a\left(\left\lfloor \frac{y-b}{a} \right\rfloor + 1\right)} + b - a\left\lfloor \frac{y-b}{a} \right\rfloor \\ &< a\left(\left\lfloor \frac{y-b}{a} \right\rfloor + 1\right) + b - a\left\lfloor \frac{y-b}{a} \right\rfloor = b + a, \end{aligned}$$

we obtain $b \leq x < b + a$. In other words, $x \in [b, b+a)$.

We define another function $v : \mathbb{R} \to \mathbb{R}$ by setting

$$v(x) = 0 \qquad \text{for each } x \in \mathbb{R}.$$

We shall now show the following three claims:

Claim 1: The function u is a-periodic.

Claim 2: The function v is a-periodic.

Claim 3: We have $u(x) = v(x)$ for each $x \in [b, b+a)$.

Once these three claims are proved, we will be able to apply Proposition 4.7.15 and conclude that $u = v$; this will easily yield the claim of Exercise 1.1.3. Let us thus prove the three claims:

[*Proof of Claim 1:* We shall show that every $x \in \mathbb{R}$ satisfies $u(x) = u(x+a)$.

Indeed, let $x \in \mathbb{R}$. We observe that

$$\lfloor y + k \rfloor = \lfloor y \rfloor + k \qquad \text{for each } y \in \mathbb{R} \text{ and } k \in \mathbb{Z}. \tag{163}$$

(This is a well-known property of floors; it is easy to prove.[87]) Now, recall that $a = \dfrac{1}{n}$. Hence, $n(x+a) = n\left(x + \dfrac{1}{n}\right) = nx + 1$. Thus,

$$\lfloor n(x+a) \rfloor = \lfloor nx + 1 \rfloor = \lfloor nx \rfloor + 1 \tag{164}$$

(by (163), applied to $y = nx$ and $k = 1$). Furthermore, each $k \in \{0, 1, \ldots, n-1\}$ satisfies

$$x + \underbrace{a}_{=\frac{1}{n}} + \frac{k}{n} = x + \frac{1}{n} + \frac{k}{n} = x + \frac{k+1}{n}. \tag{165}$$

[87] It also appears (with y renamed as x) as Proposition A.2.3 in our first solution to Exercise 1.1.3.

Hence,

$$\sum_{k=0}^{n-1} \underbrace{\left\lfloor x+a+\frac{k}{n}\right\rfloor}_{\substack{=\left\lfloor x+\frac{k+1}{n}\right\rfloor \\ \text{(by (165))}}} = \sum_{k=0}^{n-1} \left\lfloor x+\frac{k+1}{n}\right\rfloor = \sum_{k=1}^{n} \left\lfloor x+\frac{k}{n}\right\rfloor$$

(here, we have substituted k for $k+1$ in the sum)

$$\begin{aligned}
&= \underbrace{\sum_{k=0}^{n} \left\lfloor x+\frac{k}{n}\right\rfloor}_{=\sum_{k=0}^{n-1}\left\lfloor x+\frac{k}{n}\right\rfloor+\left\lfloor x+\frac{n}{n}\right\rfloor} - \left\lfloor \underbrace{x+\frac{0}{n}}_{=x}\right\rfloor \\
&= \sum_{k=0}^{n-1} \left\lfloor x+\frac{k}{n}\right\rfloor + \left\lfloor x+\underbrace{\frac{n}{n}}_{=1}\right\rfloor - \lfloor x\rfloor \\
&= \sum_{k=0}^{n-1} \left\lfloor x+\frac{k}{n}\right\rfloor + \underbrace{\lfloor x+1\rfloor}_{\substack{=\lfloor x\rfloor+1 \\ \text{(by (163), applied to } y=x \text{ and } k=1)}} - \lfloor x\rfloor \\
&= \sum_{k=0}^{n-1} \left\lfloor x+\frac{k}{n}\right\rfloor + \lfloor x\rfloor + 1 - \lfloor x\rfloor \\
&= \sum_{k=0}^{n-1} \left\lfloor x+\frac{k}{n}\right\rfloor + 1. \qquad (166)
\end{aligned}$$

Now, the definition of u yields

$$\begin{aligned}
u(x+a) &= \underbrace{\sum_{k=0}^{n-1} \left\lfloor x+a+\frac{k}{n}\right\rfloor}_{\substack{=\sum_{k=0}^{n-1}\left\lfloor x+\frac{k}{n}\right\rfloor+1 \\ \text{(by (166))}}} - \underbrace{\lfloor n(x+a)\rfloor}_{\substack{=\lfloor nx\rfloor+1 \\ \text{(by (164))}}} \\
&= \left(\sum_{k=0}^{n-1} \left\lfloor x+\frac{k}{n}\right\rfloor + 1\right) - (\lfloor nx\rfloor + 1) = \sum_{k=0}^{n-1} \left\lfloor x+\frac{k}{n}\right\rfloor - \lfloor nx\rfloor = u(x)
\end{aligned}$$

(by the definition of u). In other words, $u(x) = u(x+a)$.

Now, forget that we fixed x. We thus have shown that every $x \in \mathbb{R}$ satisfies $u(x) = u(x+a)$.

However, a is a period of u if and only if every $x \in \mathbb{R}$ satisfies $u(x) = u(x+a)$ (by Definition 4.7.10 **(a)**). Thus, a is a period of u (since every $x \in \mathbb{R}$ satisfies

$u(x) = u(x+a)$). In other words, the function u is a-periodic (by Definition 4.7.10 **(c)**). This proves Claim 1.]

[*Proof of Claim 2:* Each $x \in \mathbb{R}$ satisfies $v(x) = 0$ (by the definition of v) and $v(x+a) = 0$ (for the same reason) and thus $v(x) = 0 = v(x+a)$.

However, a is a period of v if and only if every $x \in \mathbb{R}$ satisfies $v(x) = v(x+a)$ (by Definition 4.7.10 **(a)**). Thus, a is a period of v (since every $x \in \mathbb{R}$ satisfies $v(x) = v(x+a)$). In other words, the function v is a-periodic (by Definition 4.7.10 **(c)**). This proves Claim 2.]

[*Proof of Claim 3:* Let $x \in [b, b+a)$. We must show that $u(x) = v(x)$.

We observe the following: If y is a real such that $0 \leq y < 1$, then

$$\lfloor y \rfloor = 0. \tag{167}$$

(This follows almost immediately from the definition of $\lfloor y \rfloor$; indeed, $0 \leq y < 1$ entails that the largest integer that is $\leq y$ is 0.)

We have $x \in [b, b+a) = [0, 0+a)$ (since $b = 0$). In other words, $0 \leq x < 0 + a$. Hence, $x < 0 + a = a = \dfrac{1}{n}$, so that $nx < 1$. Also, $0 \leq nx$ (since $0 \leq x$). Thus, $0 \leq nx < 1$. Hence, (167) (applied to $y = nx$) yields

$$\lfloor nx \rfloor = 0. \tag{168}$$

Next, let $k \in \{0, 1, \ldots, n-1\}$. Then, $0 \leq k \leq n-1$. We have $\underbrace{x}_{<\frac{1}{n}} + \underbrace{\frac{k}{n}}_{\substack{\leq \frac{n-1}{n} \\ \text{(since } k \leq n-1)}} < \dfrac{1}{n} + \dfrac{n-1}{n} = 1$ and $\underbrace{x}_{\geq 0} + \underbrace{\dfrac{k}{n}}_{\substack{\geq 0 \\ \text{(since } k \geq 0)}} \geq 0$. Thus, $0 \leq x + \dfrac{k}{n} < 1$. Hence, (167) (applied to $y = x + \dfrac{k}{n}$) yields

$$\left\lfloor x + \frac{k}{n} \right\rfloor = 0. \tag{169}$$

Forget that we fixed k. We thus have proved (169) for each $k \in \{0, 1, \ldots, n-1\}$.

Now, the definition of u yields

$$u(x) = \sum_{k=0}^{n-1} \underbrace{\left\lfloor x + \frac{k}{n} \right\rfloor}_{\substack{=0 \\ \text{(by (169))}}} - \underbrace{\lfloor nx \rfloor}_{\substack{=0 \\ \text{(by (168))}}} = \underbrace{\sum_{k=0}^{n-1} 0}_{=0} - 0 = 0.$$

Comparing this with

$$v(x) = 0 \qquad \text{(by the definition of } v\text{)},$$

we find $u(x) = v(x)$. This proves Claim 3.]

Now, we can apply Proposition 4.7.15 to $\mathbb{A} = \mathbb{R}$ and $S = \mathbb{R}$ (because of Claim 1, Claim 2 and Claim 3). Hence, we obtain $u = v$. Now, for each $x \in \mathbb{R}$, we have

$$\begin{aligned}\sum_{k=0}^{n-1} \left\lfloor x + \frac{k}{n} \right\rfloor - \lfloor nx \rfloor &= \underbrace{u}_{=v}(x) && \text{(by the definition of } u\text{)} \\ &= v(x) = 0 && \text{(by the definition of } v\text{)},\end{aligned}$$

so that $\sum_{k=0}^{n-1} \left\lfloor x + \frac{k}{n} \right\rfloor = \lfloor nx \rfloor$. This solves Exercise 1.1.3 again. □

4.8. Homework set #3: Sequences and more sums

This is a **regular** problem set. See Section 3.7 for details on grading.

This homework set covers the above parts of Chapter 4. Some of the problems may be unrelated. Note that the problems are ordered by (approximate) topic, not by difficulty!

Please solve **at most 5** problems. (No points will be given for further solutions.)

Exercise 4.8.1. Let $(a_0, a_1, a_2, \ldots)$ be a sequence of integers defined recursively by

$$\begin{aligned} &a_0 = 0, \qquad a_1 = 1, \qquad \text{and} \\ &a_n = 1 + a_{n-1}a_{n-2} \qquad \text{for each integer } n \geq 2. \end{aligned}$$

Prove the following:

(a) For any $k \in \mathbb{N}$ and $n \in \mathbb{N}$, we have $a_{k+n} \equiv a_k \bmod a_n$.

(b) If $u, v \in \mathbb{N}$ satisfy $u \mid v$, then $a_u \mid a_v$.

(c) For any $n, m \in \mathbb{N}$, we have $\gcd(a_n, a_m) = a_{\gcd(n,m)}$.

Exercise 4.8.2. For any positive integer n, we let $d(n)$ denote the number of all positive divisors of n. (For example, $d(6) = 4$.)

Let n be a positive integer. Prove that

$$d(1) + d(2) + \cdots + d(n) = \left\lfloor \frac{n}{1} \right\rfloor + \left\lfloor \frac{n}{2} \right\rfloor + \cdots + \left\lfloor \frac{n}{n} \right\rfloor.$$

Exercise 4.8.3. Let $n \in \mathbb{N}$.

(a) Simplify $\sum_{k=0}^{n} k \binom{n}{k}$. ("Simplify" means "get rid of the $\sum$ sign".)

(b) Simplify $\sum_{k=0}^{n} \binom{2n}{k}$.

Exercise 4.8.4. Let $n \in \mathbb{N}$. Let $x_1, x_2, \ldots, x_n$ be any numbers, and let y be a further number. Let $[n]$ denote the set $\{1, 2, \ldots, n\}$.

(a) Prove that every $m \in \{0, 1, \ldots, n-1\}$ satisfies

$$\sum_{I \subseteq [n]} (-1)^{n-|I|} \left(y + \sum_{i \in I} x_i \right)^m = 0.$$

(b) Prove that

$$\sum_{I \subseteq [n]} (-1)^{n-|I|} \left(y + \sum_{i \in I} x_i \right)^n = n! x_1 x_2 \cdots x_n.$$

[**Remark:** For $n = 2$, the statement of Exercise 4.8.4 **(b)** says that

$$(y + x_1 + x_2)^2 - (y + x_1)^2 - (y + x_2)^2 + y^2 = 2! x_1 x_2.$$

Note that (35) is the particular case of this equality for $y = m + 1$, $x_1 = 1$ and $x_2 = 2$.]

Exercise 4.8.5. Let $n \in \mathbb{N}$ and $x \in \mathbb{R} \setminus \{1, -1\}$. For each $i \in \mathbb{N}$, set $y_i = 1 - x^i$. Prove that

$$\sum_{k=0}^{n-1} \frac{y_n y_{n-1} \cdots y_{n-k}}{y_{k+1}} = n.$$

Exercise 4.8.6. Define a sequence $(a_0, a_1, a_2, \ldots)$ of integers recursively by

$$\begin{aligned} &a_0 = 0, \qquad a_1 = 1, \qquad \text{and} \\ &a_n = n\left(a_{n-1} + (n-1)\, a_{n-2}\right) \qquad \text{for each integer } n \geq 2. \end{aligned}$$

Compute a_n explicitly (in terms of sequences we already know).

Exercise 4.8.7. Let a, b, u and v be four reals.

We define two sequences $(a_0, a_1, a_2, \ldots)$ and $(b_0, b_1, b_2, \ldots)$ of reals recursively by setting

$$a_0 = a \qquad \text{and} \qquad b_0 = b$$

and

$$a_n = u a_{n-1} + v b_{n-1} \qquad \text{and} \qquad b_n = u b_{n-1} + v a_{n-1}$$

for each $n \geq 1$.

Find explicit formulas for a_n and b_n.

Recall that a sequence $(a_0, a_1, a_2, \ldots)$ of reals is said to be *weakly increasing* if it satisfies $a_0 \leq a_1 \leq a_2 \leq \cdots$ (that is, $a_i \leq a_{i+1}$ for each $i \in \mathbb{N}$). (Some authors call

such sequences *nondecreasing*, but I find this terminology slightly counterintuitive; in particular, "nondecreasing" is not the same as "not decreasing".)

Exercise 4.8.8. Let $(a_0, a_1, a_2, \ldots)$ be the unique weakly increasing sequence of positive integers that contains each positive integer i exactly i times. Thus,

$$(a_0, a_1, a_2, \ldots) = (1, 2, 2, 3, 3, 3, 4, 4, 4, 4, 5, 5, 5, 5, 5, \ldots).$$

Prove that

$$a_n = \left\lfloor \frac{1}{2}\sqrt{8n+1} + \frac{1}{2} \right\rfloor \qquad \text{for each } n \in \mathbb{N}.$$

Exercise 4.8.9. Let a and b be two positive reals. Let $f : \mathbb{R} \to \mathbb{R}$ be an a-periodic function. Let $g : \mathbb{R} \to \mathbb{R}$ be a b-periodic function. Consider the function $f + g : \mathbb{R} \to \mathbb{R}$ (which sends each $x \in \mathbb{R}$ to $(f+g)(x)$).

(a) If $a/b \in \mathbb{Q}$, then prove that $f + g$ is again a periodic function.

(b) Show that $f + g$ need not be periodic if $a/b \notin \mathbb{Q}$. (Feel free to interpret this either as "Find an example where $f + g$ is not periodic for **some** pair of a and b satisfying $a/b \notin \mathbb{Q}$" or as "Find an example where $f + g$ is not periodic for **every** pair of a and b satisfying $a/b \notin \mathbb{Q}$".)

The following is a **bonus problem**: You can get points on it even if you solved 5 others. Be warned that it is **rather hard**.

Exercise 4.8.10. Let $n \in \mathbb{Z}$ and $k \in \mathbb{Z}$. Prove that

$$\gcd\left(\binom{n-1}{k-1}, \binom{n}{k+1}, \binom{n+1}{k}\right) = \gcd\left(\binom{n-1}{k}, \binom{n}{k-1}, \binom{n+1}{k+1}\right).$$

4.9. Linear recurrences

A wider and more interesting class of integers are the *linearly recurrent sequences* (or, more precisely, *sequences satisfying linear recurrences with constant coefficients*). This class contains the arithmetic progressions, the geometric progressions, the Fibonacci sequence and some others. These sequences have an interesting yet manageable theory; in particular, they can all be described by "explicit" formulas (similar to Binet's formula for Fibonacci numbers – i.e., Theorem 2.3.1) as long as one is ready to accept the appearance of irrational numbers in these formulas. We shall not prove these formulas in full generality here, but we shall fully analyze the most commonly used particular case – that of the *two-term recurrences* (also known as *recurrences of second degree*) – and say a few things about the general case. We note that the explicit formula is not the be-all and end-all of linearly recurrent sequences; more can be said and proved, often independently of the explicit formula.

(For example, we didn't use Binet's formula all that often in our study of Fibonacci numbers!)

Some resources on linearly recurrent sequences are [Markus83], [BeDeQu11], [Melian01] and [Ivanov08].

4.9.1. Two-term recurrences: definition and examples

We begin by studying sequences $(x_0, x_1, x_2, \ldots)$ that satisfy two-term recurrences – i.e., recurrent equations of the form $x_n = ax_{n-1} + bx_{n-2}$ for all $n \geq 2$, where a and b are two fixed numbers.[88] We begin by giving them a short name:[89]

Definition 4.9.1. Let a and b be two numbers. A sequence $(x_0, x_1, x_2, \ldots)$ of numbers will be called (a, b)-*recurrent*[90] if every $n \geq 2$ satisfies

$$x_n = ax_{n-1} + bx_{n-2}. \tag{170}$$

It is clear that an (a, b)-recurrent sequence $(x_0, x_1, x_2, \ldots)$ is uniquely determined by the four numbers a, b, x_0 and x_1, since the equality (170) can be used to compute all entries of the sequence using these four numbers. Thus, if two (a, b)-recurrent sequences $(x_0, x_1, x_2, \ldots)$ and $(y_0, y_1, y_2, \ldots)$ (for a given pair (a, b)) agree in their first two entries (that is, satisfy $x_0 = y_0$ and $x_1 = y_1$), then they are identical (that is, $x_n = y_n$ for each $n \in \mathbb{N}$).

Here are some examples[91] of (a, b)-recurrent sequences for various pairs (a, b):

Example 4.9.2. The Fibonacci sequence $(f_0, f_1, f_2, \ldots) = (0, 1, 1, 2, 3, 5, \ldots)$ from Definition 2.2.1 is $(1, 1)$-recurrent, since every $n \geq 2$ satisfies $f_n = f_{n-1} + f_{n-2} = 1f_{n-1} + 1f_{n-2}$.

Example 4.9.3. The *Lucas sequence* $(\ell_0, \ell_1, \ell_2, \ldots)$ is another famous $(1, 1)$-recurrent sequence of integers. It is defined recursively by $\ell_0 = 2$, $\ell_1 = 1$, and $\ell_n = \ell_{n-1} + \ell_{n-2}$ for all $n \geq 2$. Its first terms are $2, 1, 3, 4, 7, 11, 18$.

Example 4.9.4. A sequence $(x_0, x_1, x_2, \ldots)$ is $(2, -1)$-recurrent if and only if every $n \geq 2$ satisfies $x_n = 2x_{n-1} - x_{n-2}$. In other words, a sequence $(x_0, x_1, x_2, \ldots)$ is $(2, -1)$-recurrent if and only if every $n \geq 2$ satisfies $x_n - x_{n-1} = x_{n-1} - x_{n-2}$.

[88] The name "two-term recurrences" refers to the two terms ax_{n-1} and bx_{n-2} on the right hand side. Some people prefer to also count the x_n term on the left hand side, which makes these equations "three-term recurrences" instead.

[89] As before, "numbers" can mean integer, rational, real or complex numbers.

[90] The name "(a, b)-recurrent" is not standard; I have picked it merely for the sake of brevity. A more standard way to say "(a, b)-recurrent" would be "linearly recurrent with characteristic polynomial $x^2 - ax - b$"; this is related to the polynomial point of view that we might discuss later on.

[91] mostly taken from [Grinbe15, Chapter 4]

In other words, a sequence $(x_0, x_1, x_2, \ldots)$ is $(2,-1)$-recurrent if and only if $x_1 - x_0 = x_2 - x_1 = x_3 - x_2 = \cdots$. In other words, the $(2,-1)$-recurrent sequences are precisely the arithmetic progressions.

Example 4.9.5. Geometric progressions are also (a,b)-recurrent for appropriate a and b. Namely, any geometric progression $(u, uq, uq^2, uq^3, \ldots)$ is $(q,0)$-recurrent, since every $n \geq 2$ satisfies $uq^n = q \cdot uq^{n-1} + 0 \cdot uq^{n-2}$. However, not every $(q,0)$-recurrent sequence $(x_0, x_1, x_2, \ldots)$ is a geometric progression (since the condition $x_n = qx_{n-1} + 0x_{n-2}$ for all $n \geq 2$ says nothing about x_0, and thus x_0 can be arbitrary).

Example 4.9.6. A sequence $(x_0, x_1, x_2, \ldots)$ is $(0,1)$-recurrent if and only if every $n \geq 2$ satisfies $x_n = x_{n-2}$. In other words, a sequence $(x_0, x_1, x_2, \ldots)$ is $(0,1)$-recurrent if and only if it has the form $(u, v, u, v, u, v, \ldots)$ for two numbers u and v.

Example 4.9.7. A sequence $(x_0, x_1, x_2, \ldots)$ is $(1,0)$-recurrent if and only if every $n \geq 2$ satisfies $x_n = x_{n-1}$. In other words, a sequence $(x_0, x_1, x_2, \ldots)$ is $(1,0)$-recurrent if and only if it has the form $(u, v, v, v, v, \ldots)$ for two numbers u and v. Notice that u is not required to be equal to v, because we never claimed that $x_n = x_{n-1}$ holds for $n = 1$.

Example 4.9.8. A sequence $(x_0, x_1, x_2, \ldots)$ is $(1,-1)$-recurrent if and only if every $n \geq 2$ satisfies $x_n = x_{n-1} - x_{n-2}$. In Example 4.7.7, we have already seen an example of such a sequence, and observed that it is 6-periodic. But this holds more generally: Any $(1,-1)$-recurrent sequence $(x_0, x_1, x_2, \ldots)$ is 6-periodic (i.e., it satisfies $x_{n+6} = x_n$ for every $n \in \mathbb{N}$). This is because every $n \in \mathbb{N}$ satisfies

$$\begin{aligned} x_{n+6} &= \underbrace{x_{n+5}}_{=x_{n+4}-x_{n+3}} - x_{n+4} = (x_{n+4} - x_{n+3}) - x_{n+4} = -\underbrace{x_{n+3}}_{=x_{n+2}-x_{n+1}} \\ &= -\left(\underbrace{x_{n+2}}_{=x_{n+1}-x_n} - x_{n+1}\right) = -(x_{n+1} - x_n - x_{n+1}) = x_n. \end{aligned}$$

We can describe the $(1,-1)$-recurrent sequences explicitly: A sequence $(x_0, x_1, x_2, \ldots)$ is $(1,-1)$-recurrent if and only if it has the form $(u, v, v-u, -u, -v, u-v, \ldots)$ (where the "…" stands for "repeat the preceding 6 values over and over" here) for two numbers u and v.

Example 4.9.9. The sequence $(t_0, t_1, t_2, \ldots)$ from Exercise 1.1.2 itself is not (a,b)-recurrent for any pair (a,b); however, its two subsequences $(t_0, t_2, t_4, t_6, \ldots)$ and $(t_1, t_3, t_5, t_7, \ldots)$ are $(4,-1)$-recurrent, since Exercise 1.1.2 **(a)** shows that $t_{2n} = 4t_{2(n-1)} - t_{2(n-2)}$ and $t_{2n+1} = 4t_{2(n-1)+1} - t_{2(n-2)+1}$ for every $n \geq 2$.

Another family of examples of (a, b)-recurrent sequences comes from trigonometry:

Proposition 4.9.10. Let α be any angle. Then, the sequences

$$(\sin(0\alpha), \sin(1\alpha), \sin(2\alpha), \ldots) \qquad \text{and} \tag{171}$$
$$(\cos(0\alpha), \cos(1\alpha), \cos(2\alpha), \ldots) \tag{172}$$

are $(2\cos\alpha, -1)$-recurrent. More generally, if α and β are two angles, then the sequence

$$(\sin(\beta + 0\alpha), \sin(\beta + 1\alpha), \sin(\beta + 2\alpha), \ldots) \tag{173}$$

is $(2\cos\alpha, -1)$-recurrent.

Proof of Proposition 4.9.10 (sketched). Let α and β be two angles. It suffices to show that the sequence $(\sin(\beta + 0\alpha), \sin(\beta + 1\alpha), \sin(\beta + 2\alpha), \ldots)$ is $(2\cos\alpha, -1)$-recurrent (because the two sequences (171) and (172) are particular cases of this sequence for $\beta = 0$ and $\beta = \frac{\pi}{2}$, respectively). In other words, it suffices to show that

$$\sin(\beta + n\alpha) = 2\cos\alpha \sin(\beta + (n-1)\alpha) + (-1)\sin(\beta + (n-2)\alpha)$$

for every $n \geq 2$. So let us fix $n \geq 2$.

One of the well-known trigonometric identities states that

$$\sin x + \sin y = 2\sin\frac{x+y}{2}\cos\frac{x-y}{2}$$

for any two angles x and y. Applying this to $x = \beta + n\alpha$ and $y = \beta + (n-2)\alpha$, we obtain

$$\begin{aligned}
&\sin(\beta + n\alpha) + \sin(\beta + (n-2)\alpha) \\
&= 2\sin\underbrace{\frac{(\beta + n\alpha) + (\beta + (n-2)\alpha)}{2}}_{=\beta+(n-1)\alpha}\cos\underbrace{\frac{(\beta + n\alpha) - (\beta + (n-2)\alpha)}{2}}_{=\alpha} \\
&= 2\sin(\beta + (n-1)\alpha)\cos\alpha = 2\cos\alpha\sin(\beta + (n-1)\alpha).
\end{aligned}$$

Hence,

$$\begin{aligned}
\sin(\beta + n\alpha) &= 2\cos\alpha\sin(\beta + (n-1)\alpha) - \sin(\beta + (n-2)\alpha) \\
&= 2\cos\alpha\sin(\beta + (n-1)\alpha) + (-1)\sin(\beta + (n-2)\alpha).
\end{aligned}$$

This proves Proposition 4.9.10. □

Note that all three sequences in Proposition 4.9.10 are periodic when α is a rational multiple of 2π (that is, when $\alpha = 2\pi r$ for some $r \in \mathbb{Q}$). This provides an ample source of periodic recurrent sequences (of any possible period).

4.9.2. Two-term recurrences: Binet-like formulas

The following question now suggests itself:[92]

Exercise 4.9.1. Let a and b be two numbers. Let $(x_0, x_1, x_2, \ldots)$ be an (a, b)-recurrent sequence. Is there an explicit formula for x_n in terms of a, b, x_0 and x_1, similar to Binet's formula (Theorem 2.3.1) for Fibonacci numbers?

Discussion of Exercise 4.9.1. Let us first solve Exercise 4.9.1 in a "barbaric" way – by experimenting and guessing. We will later see a more cultured way (or two) of solving it.

We begin by taking a look at Binet's formula, in order to get an idea of what we should expect.

First, we observe that Binet's formula involves the two irrational numbers φ and ψ, even though the Fibonacci numbers are integers. Thus, we should expect irrational numbers (specifically, square roots) to appear in our formula, even if a, b, x_0 and x_1 are integers.

Binet's formula said $f_n = \frac{1}{\sqrt{5}}\varphi^n - \frac{1}{\sqrt{5}}\psi^n$. While it would be strange to expect the same specific numbers $\frac{1}{\sqrt{5}}$, φ and ψ to appear in the general case, we can hope that the overall form of the formula will be the same:

$$x_n = \gamma\lambda^n + \delta\mu^n \tag{174}$$

for four constants $\gamma, \delta, \lambda, \mu$. (Binet's formula is an example of this form, with $\gamma = \frac{1}{\sqrt{5}}$, $\delta = -\frac{1}{\sqrt{5}}$, $\lambda = \varphi$ and $\mu = \psi$).

So we **guess** that there is an explicit formula of the form (174) for four constants $\gamma, \delta, \lambda, \mu$. ("Constants" means "numbers not depending on n".) We can now try to find these four constants. (Of course, nothing guarantees us that they actually exist – indeed, they don't always do, as we will eventually see. But it can't hurt to try!)

Our sequence $(x_0, x_1, x_2, \ldots)$ is (a, b)-recurrent; i.e., it satisfies the equation (170) for each $n \geq 2$. Let us rewrite this equation with the help of (174). Indeed, fix an integer $n \geq 2$. Then, (174) shows that $x_n = \gamma\lambda^n + \delta\mu^n$ and $x_{n-1} = \gamma\lambda^{n-1} + \delta\mu^{n-1}$ and $x_{n-2} = \gamma\lambda^{n-2} + \delta\mu^{n-2}$. Thus, (170) rewrites as

$$\gamma\lambda^n + \delta\mu^n = a\left(\gamma\lambda^{n-1} + \delta\mu^{n-1}\right) + b\left(\gamma\lambda^{n-2} + \delta\mu^{n-2}\right). \tag{175}$$

[92] When we say "an (a, b)-recurrent sequence", we of course mean an (a, b)-recurrent sequence of numbers.

Therefore,

$$\begin{aligned}
0 &= \gamma\lambda^n + \delta\mu^n - \left(a\left(\gamma\lambda^{n-1} + \delta\mu^{n-1}\right) + b\left(\gamma\lambda^{n-2} + \delta\mu^{n-2}\right)\right) \\
&= \underbrace{\left(\gamma\lambda^n - a\gamma\lambda^{n-1} - b\gamma\lambda^{n-2}\right)}_{=\gamma(\lambda^2 - a\lambda - b)\lambda^{n-2}} + \underbrace{\left(\delta\mu^n - a\delta\mu^{n-1} - b\delta\mu^{n-2}\right)}_{=\delta(\mu^2 - a\mu - b)\mu^{n-2}} \\
&= \gamma\left(\lambda^2 - a\lambda - b\right)\lambda^{n-2} + \delta\left(\mu^2 - a\mu - b\right)\mu^{n-2}. \qquad (176)
\end{aligned}$$

Now, forget that we fixed n. We thus have shown that the equality (176) must hold for every integer $n \geq 2$. Since the only "moving parts" in this equality are the λ^{n-2} and the μ^{n-2} terms (in the sense that all the other terms are independent of n), we sense an easy way to ensure that it holds: Namely, it will hold whenever $\lambda^2 - a\lambda - b = 0$ and $\mu^2 - a\mu - b = 0$. Technically, this is not the only way to make (176) hold; for example, the same would hold if $\gamma = 0$ and $\delta = 0$. However, it is clear that $\gamma = 0$ and $\delta = 0$ can only work if the sequence $(x_0, x_1, x_2, \ldots)$ consists of zeros, while we are trying to study the general case. In the "typical" case, neither γ nor δ will be 0, since otherwise the sequence $(x_0, x_1, x_2, \ldots)$ would have to be a geometric progression (by (174)).

So we need to pick λ and μ in such a way that $\lambda^2 - a\lambda - b = 0$ and $\mu^2 - a\mu - b = 0$. In other words, λ and μ must be roots of the quadratic equation $X^2 - aX - b = 0$. As you know from high school(?), this quadratic equation has two complex roots (counted with multiplicity[93]). It seems a good idea to let λ and μ be the two roots of this equation.[94] Explicitly, this means that

$$\lambda = \frac{a + \sqrt{a^2 + 4b}}{2} \qquad \text{and} \qquad \mu = \frac{a - \sqrt{a^2 + 4b}}{2}. \qquad (177)$$

(We could of course swap the roles of λ and μ, but this would not change anything.) Note that λ and μ can fail to be real numbers, even if a and b are reals. (For example, this will happen if $a = b = -1$.)

By defining λ and μ through (177), we have ensured that (176) holds for each integer $n \geq 2$. Thus, (175) holds for each integer $n \geq 2$ as well (since (176) is equivalent to (175)). In other words, the sequence $\left(\gamma\lambda^0 + \delta\mu^0, \gamma\lambda^1 + \delta\mu^1, \gamma\lambda^2 + \delta\mu^2, \ldots\right)$ is (a, b)-recurrent. Hence, if the formula (174) holds for $n = 0$ and for $n = 1$, then it also holds for all $n \in \mathbb{N}$ (because two (a, b)-recurrent sequences that agree in their first two entries must be identical). It thus remains to pick γ and δ in such a way that the formula (174) holds for $n = 0$ and for $n = 1$. In other words, we must pick

[93] This means that sometimes, the equation has only one root, but we should count it as two equal roots. (This happens precisely when $a^2 + 4b = 0$.)

[94] Technically, we could of course let λ and μ be one and the same root. But this feels like wasting a variable (and is indeed so); in fact, setting $\lambda = \mu$ in (174) would force $(x_0, x_1, x_2, \ldots)$ to be a geometric progression, which is not representative of the "typical" case. So it stands to reason that we don't take λ and μ to be one and the same root (unless there is only one root).

γ and δ in such a way that

$$\begin{cases} x_0 = \gamma\lambda^0 + \delta\mu^0; \\ x_1 = \gamma\lambda^1 + \delta\mu^1. \end{cases} \tag{178}$$

This is a system of two linear equations in the two unknowns γ and δ (since x_0, x_1, λ, μ are known). Using Gaussian elimination or Cramer's rule (or any other method for solving systems of linear equations), we see that its solution is

$$\gamma = \frac{x_1 - \mu x_0}{\lambda - \mu} \qquad \text{and} \qquad \delta = \frac{\lambda x_0 - x_1}{\lambda - \mu} \tag{179}$$

when $\lambda \neq \mu$. Note that $\lambda \neq \mu$ holds if and only if $a^2 + 4b \neq 0$ (this follows easily from (177)). If $a^2 + 4b = 0$, then the system (178) has no solutions or infinitely many solutions; let us ignore this case for now.

Thus, whenever $a^2 + 4b \neq 0$, we have found four constants $\gamma, \delta, \lambda, \mu$ that ensure that (174) does indeed hold for each $n \in \mathbb{N}$: Namely, λ and μ are given by (177), whereas γ and δ are given by (179). It is easy to see that these four constants $\gamma, \delta, \lambda, \mu$ indeed work (i.e., that the formula (174) indeed holds for all $n \in \mathbb{N}$ when these four constants are being used)[95].

We have thus obtained a partial answer to Exercise 4.9.1: We have found a formula that works whenever $a^2 + 4b \neq 0$. We will soon turn to the other case; but let us say a few general words first. Our approach to finding a formula for x_n had the following structure: We **guessed** that there should be a formula of the form (174) for four constants $\gamma, \delta, \lambda, \mu$. We could not guess the exact values of $\gamma, \delta, \lambda, \mu$, however, so we treated them as unknowns and solved for them. That is, we guessed only the part of the answer that was easily guessable (viz., the structure of the formula), and then used that guess to figure out the rest through analytic reasoning. This kind of "incomplete guess" is called an *ansatz* (German word, plural *ansätze*) and is particularly popular in mathematical physics, but contest problems also tend to provide opportunities for ansätze to those who know to look out for them. Of course, as with any kind of guess, an answer found using an ansatz needs to be checked (i.e., proved); in our case, this was essentially trivial, because the reasoning we made to find $\gamma, \delta, \lambda, \mu$ was reversible. In other situations, this can be nontrivial or even difficult – but knowing the answer is still a good step forward.

Let us now complete the solution of Exercise 4.9.1 by studying the case $a^2 + 4b = 0$. Consider this case. In general, the system (178) may fail to have a solution, so our ansatz (174) will not work here. Instead, we can use the condition $a^2 + 4b = 0$ to rewrite b as $\dfrac{-a^2}{4}$ (thus eliminating one parameter from our problem), and compute

[95] Indeed, we already explained that, due to our choice of λ and μ, if the formula (174) holds for $n = 0$ and for $n = 1$, then it also holds for all $n \in \mathbb{N}$. But our choice of γ and δ ensures that the formula (174) does indeed hold for $n = 0$ and $n = 1$. Thus, this formula holds for all $n \in \mathbb{N}$, as desired.

the first few entries of our sequence to look for a pattern:

$$
\begin{aligned}
x_0 &= x_0; \\
x_1 &= x_1; \\
x_2 &= ax_1 + \underbrace{b}_{=\frac{-a^2}{4}} x_0 = ax_1 + \frac{-a^2}{4} x_0 = a\left(x_1 - \frac{a}{4}x_0\right); \\
x_3 &= a \underbrace{x_2}_{=a\left(x_1 - \frac{a}{4}x_0\right)} + \underbrace{b}_{=\frac{-a^2}{4}} x_1 = aa\left(x_1 - \frac{a}{4}x_0\right) + \frac{-a^2}{4}x_1 = a^2\left(\frac{3}{4}x_1 - \frac{a}{4}x_0\right); \\
x_4 &= a \underbrace{x_3}_{=a^2\left(\frac{3}{4}x_1 - \frac{a}{4}x_0\right)} + \underbrace{b}_{=\frac{-a^2}{4}} \underbrace{x_2}_{=a\left(x_1 - \frac{a}{4}x_0\right)} \\
&= aa^2\left(\frac{3}{4}x_1 - \frac{a}{4}x_0\right) + \frac{-a^2}{4}a\left(x_1 - \frac{a}{4}x_0\right) = a^3\left(\frac{1}{2}x_1 - \frac{3a}{16}x_0\right); \\
x_5 &= a^4\left(\frac{5}{16}x_1 - \frac{a}{8}x_0\right) \qquad \text{(we are now omitting the intermediate steps)}; \\
x_6 &= a^5\left(\frac{3}{16}x_1 - \frac{5a}{64}x_0\right).
\end{aligned}
$$

The answer may not immediately stare at us from the equations, but again we can make an **ansatz**: Namely, we expect to have

$$x_n = a^{n-1}\left(u_n x_1 - v_n a x_0\right) \qquad \text{for each } n > 0, \tag{180}$$

[96] where u_n and v_n are two rational numbers (dependent on n, but independent on a, x_0 and x_1). Our goal is now to find these u_n and v_n. Let us list the first few of them:

$$
\begin{aligned}
(u_1, u_2, u_3, u_4, u_5, u_6) &= \left(1, 1, \frac{3}{4}, \frac{1}{2}, \frac{5}{16}, \frac{3}{16}\right); \\
(v_1, v_2, v_3, v_4, v_5, v_6) &= \left(0, \frac{1}{4}, \frac{1}{4}, \frac{3}{16}, \frac{1}{8}, \frac{5}{64}\right).
\end{aligned}
$$

This suggests that both the u_i and the v_i are fractions whose denominators are powers of 2. The exact denominators appear to oscillate, but it is noticeable that the denominator of u_i is never larger than 2^{i-1}, while that of v_i is never larger than

[96] We are saying "for each $n > 0$" instead of "for each $n \in \mathbb{N}$" here, since a^{n-1} is undefined when $n = 0$ and $a = 0$. (In truth, we are being over-careful here; the $a = 0$ case is easy enough that we can just leave it aside.)

2^i. Thus, we may want to look at $2^{i-1}u_i$ and $2^i v_i$, which will be integers:

$$\left(2^0u_1, 2^1u_2, 2^2u_3, 2^3u_4, 2^4u_5, 2^5u_6\right) = (1,2,3,4,5,6)\,;$$
$$\left(2^1v_1, 2^2v_2, 2^3v_3, 2^4v_4, 2^5v_5, 2^6v_6\right) = (0,1,2,3,4,5)\,.$$

Voilà: now the pattern is impossible to miss.[97] We guess that $2^{i-1}u_i = i$ and $2^i v_i = i-1$. If this is true, then (180) becomes

$$x_n = a^{n-1}\left(\frac{n}{2^{n-1}}x_1 - \frac{n-1}{2^n}ax_0\right) \qquad \text{for each } n > 0. \tag{181}$$

In other words,

$$x_n = \frac{1}{2^n}\left(2na^{n-1}x_1 - (n-1)\,a^n x_0\right) \qquad \text{for each } n > 0. \tag{182}$$

This is, so far, a guess. But clearly, if a formula like (182) is true, then it can be proved by a straightforward strong induction on n. And this is indeed the case; (182) is true, and its induction proof can be found in [Grinbe15, solution to Exercise 4.1]. Note that (182) holds for $n = 0$ as well, if we agree to understand na^{n-1} as 0 for $n = 0$ (even if a^{n-1} may fail to be defined).

Thus, we have found explicit formulas for x_n in both cases $a^2 + 4b \neq 0$ and $a^2 + 4b = 0$. Exercise 4.9.1 is solved. □

Combining our results obtained in the discussion of Exercise 4.9.1 above, we obtain the following result (a "generalized Binet formula"):

Theorem 4.9.11. Let a and b be two numbers. Let $(x_0, x_1, x_2, \ldots)$ be an (a,b)-recurrent sequence. Then:

(a) If $a^2 + 4b \neq 0$, then every $n \in \mathbb{N}$ satisfies

$$x_n = \gamma\lambda^n + \delta\mu^n, \tag{183}$$

where we set

$$\lambda = \frac{a + \sqrt{a^2+4b}}{2} \qquad \text{and} \qquad \mu = \frac{a - \sqrt{a^2+4b}}{2} \tag{184}$$

and

$$\gamma = \frac{x_1 - \mu x_0}{\lambda - \mu} \qquad \text{and} \qquad \delta = \frac{\lambda x_0 - x_1}{\lambda - \mu}. \tag{185}$$

[97] If you found the idea of considering $2^{i-1}u_i$ and $2^i v_i$ somewhat far-fetched: there are other ways to guess the same result. For example, after computing u_7 and u_9, it becomes easy to notice that the numerator of u_i is i whenever i is odd. With this in mind, you will then be motivated to look at the sequence $\left(\frac{u_1}{1}, \frac{u_2}{2}, \frac{u_3}{3}, \frac{u_4}{4}, \ldots\right)$, which will quickly reveal itself to be the rather famous $(1,2,4,8,16,\ldots)$.

(b) If $a^2 + 4b = 0$, then every $n \in \mathbb{N}$ satisfies

$$x_n = \frac{1}{2^n}\left(2na^{n-1}x_1 - (n-1)\,a^n x_0\right)$$

(where we agree to understand na^{n-1} as 0 for $n = 0$).

It is easy to see that Theorem 2.3.1 is a particular case of Theorem 4.9.11 **(a)** (obtained by setting $a = 1$ and $b = 1$ and $x_i = f_i$). Likewise, we can compute explicit formulas for other examples of (a, b)-recurrent sequences:

Example 4.9.12. Consider the Lucas sequence $(\ell_0, \ell_1, \ell_2, \ldots)$ from Example 4.9.3. This sequence is $(1, 1)$-recurrent; thus, Theorem 4.9.11 **(a)** (applied to $a = 1$ and $b = 1$ and $x_i = \ell_i$) yields that every $n \in \mathbb{N}$ satisfies $\ell_n = \gamma\lambda^n + \delta\mu^n$, where (as you can see by straightforward computations) we have $\lambda = \dfrac{1+\sqrt{5}}{2}$ and $\mu = \dfrac{1-\sqrt{5}}{2}$ and $\gamma = 1$ and $\delta = 1$. That is, every $n \in \mathbb{N}$ satisfies $\ell_n = \varphi^n + \psi^n$, where φ and ψ are as in Theorem 2.3.1.

Example 4.9.13. As we know from Example 4.9.4, a $(2, -1)$-recurrent sequence is the same as an arithmetic progression. We can obtain an explicit formula for any entry x_n of such a sequence $(x_0, x_1, x_2, \ldots)$ using Theorem 4.9.11 **(b)** (indeed, this is a case where Theorem 4.9.11 **(b)** applies, since $2^2 + 4(-1) = 0$); but this formula readily simplifies to the well-known $x_n = x_0 + n(x_1 - x_0)$.

Example 4.9.14. Let $(x_0, x_1, x_2, \ldots) = (u, v, u, v, u, v, \ldots)$ for two numbers u and v. As we saw in Example 4.9.6, this is a $(0, 1)$-recurrent sequence. Thus, Theorem 4.9.11 **(a)** (applied to $a = 0$ and $b = 1$) yields that every $n \in \mathbb{N}$ satisfies $x_n = \gamma\lambda^n + \delta\mu^n$, where (again by computation) $\lambda = 1$ and $\mu = -1$ and $\gamma = \dfrac{u+v}{2}$ and $\delta = \dfrac{u-v}{2}$. That is, every $n \in \mathbb{N}$ satisfies

$$x_n = \frac{u+v}{2} \cdot 1^n + \frac{u-v}{2} \cdot (-1)^n = \frac{u+v}{2} + (-1)^n \cdot \frac{u-v}{2}.$$

This is easy to check by hand.

Example 4.9.15. Let $(x_0, x_1, x_2, \ldots) = (u, v, v, v, v, \ldots)$ for two numbers u and v. As we saw in Example 4.9.7, this is a $(1, 0)$-recurrent sequence. Thus, Theorem 4.9.11 **(a)** (applied to $a = 1$ and $b = 0$) yields that every $n \in \mathbb{N}$ satisfies $x_n = \gamma\lambda^n + \delta\mu^n$, where (again by computation) $\lambda = 1$ and $\mu = 0$ and $\gamma = v$ and $\delta = u - v$. That is, every $n \in \mathbb{N}$ satisfies

$$x_n = v1^n + (u-v)\,0^n = v + (u-v)\,0^n.$$

This is, of course, again true due to the way 0^n behaves when $n = 0$ and when $n > 0$.

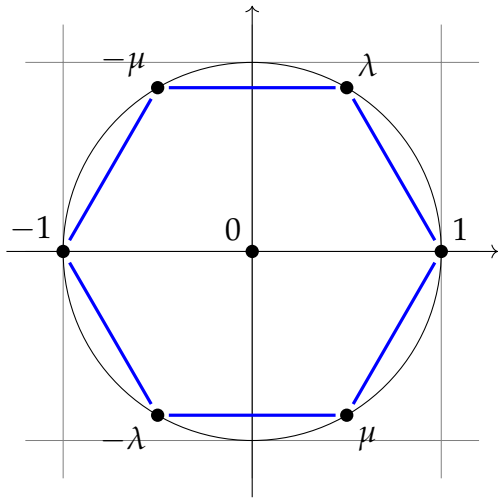
Example 4.9.16. Now, consider the $(1,-1)$-recurrent sequence $(x_0, x_1, x_2, \ldots)$ from Example 4.9.8. Here, Theorem 4.9.11 **(a)** (applied to $a = 1$ and $b = -1$) yields that every $n \in \mathbb{N}$ satisfies $x_n = \gamma\lambda^n + \delta\mu^n$, where (again by computation)

$$\lambda = \frac{1+\sqrt{-3}}{2} \qquad \text{and} \qquad \mu = \frac{1-\sqrt{-3}}{2}$$

and

$$\gamma = \frac{x_1 - \mu x_0}{\lambda - \mu} \qquad \text{and} \qquad \delta = \frac{\lambda x_0 - x_1}{\lambda - \mu}.$$

The numbers λ and μ here are complex – but they are more than that. If we plot them on the Argand diagram (see [19s, §4.1.8] for details on this), we realize that – together with 1, -1, $-\lambda$ and $-\mu$ – they form the vertices of a regular hexagon inscribed in the unit circle:



.

Thus, due to the way multiplication of complex numbers corresponds to composition of rotations, the numbers λ and μ (as well as the other vertices 1, -1, $-\lambda$ and $-\mu$) are 6-th roots of unity – i.e., they satisfy $\lambda^6 = 1$ and $\mu^6 = 1$. Hence, every $n \in \mathbb{N}$ satisfies $\lambda^{n+6} = \lambda^n$ and $\mu^{n+6} = \mu^n$ and therefore $\gamma\lambda^{n+6} + \delta\mu^{n+6} = \gamma\lambda^n + \delta\mu^n$. Since $x_{n+6} = \gamma\lambda^{n+6} + \delta\mu^{n+6}$ and $x_n = \gamma\lambda^n + \delta\mu^n$, we can rewrite this as $x_{n+6} = x_n$. This shows once again (but geometrically this time) that the sequence $(x_0, x_1, x_2, \ldots)$ is 6-periodic.

Example 4.9.17. Let α and β be two angles. Proposition 4.9.10 shows that the sequence

$$(\sin(\beta + 0\alpha), \sin(\beta + 1\alpha), \sin(\beta + 2\alpha), \ldots)$$

is $(2\cos\alpha, -1)$-recurrent. Obviously, we already have an explicit formula for its entries, but let us contrast it with what Theorem 4.9.11 **(a)** would yield. Namely, Theorem 4.9.11 **(a)** (applied to $a = 2\cos\alpha$ and $b = -1$ and $x_n = \sin(\beta + n\alpha)$)

yields that $\sin(\beta + n\alpha) = \gamma\lambda^n + \delta\mu^n$, where

$$\lambda = \frac{2\cos\alpha + \sqrt{(2\cos\alpha)^2 - 4}}{2} \qquad \text{and} \qquad \mu = \frac{2\cos\alpha - \sqrt{(2\cos\alpha)^2 - 4}}{2}$$

and

$$\gamma = \frac{\sin(\beta + \alpha) - \mu\sin\beta}{\lambda - \mu} \qquad \text{and} \qquad \delta = \frac{\lambda\sin(\beta + \alpha) - \sin\beta}{\lambda - \mu}.$$

(To be fully precise, Theorem 4.9.11 **(a)** only applies when $(2\cos\alpha)^2 - 4 \neq 0$, that is, when α is not a multiple of π; but the other case is easy.) We note that the square root $\sqrt{(2\cos\alpha)^2 - 4}$ is an imaginary number (since $|2\cos\alpha| \leq 2$ and thus $(2\cos\alpha)^2 - 4 \leq 0$) and can be rewritten as $2i\,|\sin\alpha|$, where $i = \sqrt{-1}$ (this follows easily from $(\sin\alpha)^2 + (\cos\alpha)^2 = 1$). We assume that $\alpha \in (0, \pi)$, so that we can rewrite $|\sin\alpha|$ as $\sin\alpha$ (the other case is no harder). Then, our above formulas for λ and μ become

$$\lambda = \cos\alpha + i\sin\alpha \qquad \text{and} \qquad \mu = \cos\alpha - i\sin\alpha.$$

Anyone familiar with Euler's formula will rewrite this as

$$\lambda = e^{i\alpha} \qquad \text{and} \qquad \mu = e^{-i\alpha}.$$

The formula $\sin(\beta + n\alpha) = \gamma\lambda^n + \delta\mu^n$ can now be straightforwardly (if laboriously) derived from de Moivre's formula.

4.9.3. Two-term recurrences: various properties

Various properties of the Fibonacci sequence can be generalized to arbitrary (a, b)-recurrent sequences. For example, here is how the Cassini identity (Exercise 2.2.2) can be generalized:

Exercise 4.9.2. Let a and b be two numbers. Let $(x_0, x_1, x_2, \ldots)$ be an (a, b)-recurrent sequence. Prove that

$$x_{n+1}x_{n-1} - x_n^2 = (-b)^{n-1}\left(x_2x_0 - x_1^2\right) \tag{186}$$

for any positive integer n.

Solution to Exercise 4.9.2. Let us prove (186) by induction on n:

Induction base: We have

$$\underbrace{x_{1+1}}_{=x_2}\underbrace{x_{1-1}}_{=x_0} - \underbrace{x_1^2}_{=x_1^2} = x_2x_0 - x_1^2 = (-b)^{1-1}\left(x_2x_0 - x_1^2\right)$$

(since $\underbrace{(-b)^{1-1}}_{=(-b)^0=1}\left(x_2x_0 - x_1^2\right) = x_2x_0 - x_1^2$). In other words, (186) holds for $n = 0$.

Induction step: Let m be a positive integer. Assume (as the induction hypothesis) that (186) holds for $n = m$. We must prove that (186) holds for $n = m + 1$.

The sequence $(x_0, x_1, x_2, \ldots)$ is (a, b)-recurrent. In other words, every $n \geq 2$ satisfies $x_n = ax_{n-1} + bx_{n-2}$ (by the definition of "(a, b)-recurrent"). Applying this equality to $n = m + 2$, we obtain

$$x_{m+2} = ax_{(m+2)-1} + bx_{(m+2)-2} = ax_{m+1} + bx_m.$$

Also, applying the same equality $x_n = ax_{n-1} + bx_{n-2}$ to $n = m + 1$, we obtain

$$x_{m+1} = ax_{(m+1)-1} + bx_{(m+1)-2} = ax_m + bx_{m-1}. \tag{187}$$

We have assumed that (186) holds for $n = m$. In other words, we have

$$x_{m+1}x_{m-1} - x_m^2 = (-b)^{m-1}\left(x_2x_0 - x_1^2\right). \tag{188}$$

Now,

$$\begin{aligned}
& x_{(m+1)+1}x_{(m+1)-1} - x_{m+1}^2 \\
&= \underbrace{x_{m+2}}_{=ax_{m+1}+bx_m} x_m - x_{m+1}^2 = \left(ax_{m+1} + bx_m\right)x_m - x_{m+1}^2 \\
&= ax_{m+1}x_m + bx_m^2 - x_{m+1}^2 = bx_m^2 - x_{m+1}\underbrace{\left(x_{m+1} - ax_m\right)}_{\substack{=bx_{m-1} \\ \text{(by (187))}}} \\
&= bx_m^2 - x_{m+1}\cdot bx_{m-1} = b\left(x_m^2 - x_{m+1}x_{m-1}\right) = (-b)\underbrace{\left(x_{m+1}x_{m-1} - x_m^2\right)}_{\substack{=(-b)^{m-1}\left(x_2x_0 - x_1^2\right) \\ \text{(by (188))}}} \\
&= \underbrace{(-b)(-b)^{m-1}}_{=(-b)^m=(-b)^{(m+1)-1}}\left(x_2x_0 - x_1^2\right) = (-b)^{(m+1)-1}\left(x_2x_0 - x_1^2\right).
\end{aligned}$$

In other words, (186) holds for $n = m + 1$. This completes the induction step. Thus, Exercise 4.9.2 is solved by induction. □

Alternatively, Exercise 4.9.2 can be solved in a straightforward way using Theorem 4.9.11, provided that one is willing to get one's hands dirty with the necessary computations.

Next, let us generalize the addition formula for Fibonacci numbers (Exercise 2.2.3) to (a, b)-recurrent sequences:

Exercise 4.9.3. Let a and b be two numbers. Let $(x_0, x_1, x_2, \ldots)$ and $(y_0, y_1, y_2, \ldots)$ be two (a, b)-recurrent sequences such that $x_0 = 0$ and $x_1 = 1$. Prove that

$$y_{n+m+1} = bx_n y_m + x_{n+1} y_{m+1}$$

for any nonnegative integers n and m.

Note that Exercise 2.2.3 is the particular case of Exercise 4.9.3 for $a = 1$ and $b = 1$ and $x_i = f_i$ and $y_i = f_i$. Indeed, the Fibonacci sequence satisfies $f_0 = 0$ and $f_1 = 1$, so Exercise 4.9.3 can be applied to it.

Once you start generalizing, it is hard to stop. Thus, rather than solving Exercise 4.9.3 directly, let us generalize it even further (dropping the requirements $x_0 = 0$ and $x_1 = 1$) and solve the generalization instead:

Exercise 4.9.4. Let a and b be two numbers. Let $(x_0, x_1, x_2, \ldots)$ and $(y_0, y_1, y_2, \ldots)$ be two (a, b)-recurrent sequences. Prove that

$$bx_0 y_{n+m} + x_1 y_{n+m+1} = bx_n y_m + x_{n+1} y_{m+1} \tag{189}$$

for any nonnegative integers n and m.

Solution to Exercise 4.9.4. Rather than try to adapt our above solution to Exercise 2.2.3 to our now-increased generality, we start from scratch and prove Exercise 4.9.4 by induction. (In this we follow [Grinbe15, Theorem 2.26], which is a particular case of Exercise 4.9.4.)

Our goal is to prove that (189) holds for all nonnegative integers n and m.

We proceed by induction on n (without fixing m):

Induction base: For any nonnegative integer m, we have

$$bx_0 \underbrace{y_{0+m}}_{=y_m} + \underbrace{x_1}_{=x_{0+1}} \underbrace{y_{0+m+1}}_{=y_{m+1}} = bx_0 y_m + x_{0+1} y_{m+1}.$$

In other words, (189) holds for $n = 0$ (and every nonnegative integer m).

Induction step: Let $k \in \mathbb{N}$. Assume (as the induction hypothesis) that (189) holds for $n = k$. We must now prove that (189) holds for $n = k + 1$.

We have assumed that (189) holds for $n = k$. In other words, we have

$$bx_0 y_{k+m} + x_1 y_{k+m+1} = bx_k y_m + x_{k+1} y_{m+1} \tag{190}$$

for every nonnegative integer m.

Let m be a nonnegative integer.

The sequence $(x_0, x_1, x_2, \ldots)$ is (a, b)-recurrent. In other words, every $n \geq 2$ satisfies $x_n = ax_{n-1} + bx_{n-2}$ (by the definition of "(a, b)-recurrent"). Applying this equality to $n = k + 2$, we obtain

$$x_{k+2} = ax_{(k+2)-1} + bx_{(k+2)-2} = ax_{k+1} + bx_k. \tag{191}$$

The sequence $(y_0, y_1, y_2, \ldots)$ is (a,b)-recurrent. In other words, every $n \geq 2$ satisfies $y_n = ay_{n-1} + by_{n-2}$ (by the definition of "(a,b)-recurrent"). Applying this equality to $n = m+2$, we obtain

$$y_{m+2} = ay_{(m+2)-1} + by_{(m+2)-2} = ay_{m+1} + by_m. \tag{192}$$

Now,

$$\begin{aligned}
bx_{k+1}y_m + \underbrace{x_{(k+1)+1}}_{\substack{=x_{k+2} \\ =ax_{k+1}+bx_k \\ \text{(by (191))}}} y_{m+1} &= bx_{k+1}y_m + (ax_{k+1} + bx_k)\, y_{m+1} \\
&= bx_{k+1}y_m + ax_{k+1}y_{m+1} + bx_k y_{m+1} \\
&= bx_k y_{m+1} + x_{k+1} \underbrace{(ay_{m+1} + by_m)}_{\substack{=y_{m+2} \\ \text{(by (192))}}} \\
&= bx_k y_{m+1} + x_{k+1} \underbrace{y_{m+2}}_{=y_{(m+1)+1}} = bx_k y_{m+1} + x_{k+1} y_{(m+1)+1}.
\end{aligned}$$

On the other hand, applying (190) to $m+1$ instead of m, we obtain

$$bx_0 y_{k+(m+1)} + x_1 y_{k+(m+1)+1} = bx_k y_{m+1} + x_{k+1} y_{(m+1)+1}.$$

Comparing these two equalities yields

$$bx_{k+1}y_m + x_{(k+1)+1}y_{m+1} = bx_0 y_{k+(m+1)} + x_1 y_{k+(m+1)+1} = bx_0 y_{(k+1)+m} + x_1 y_{(k+1)+m+1}$$

(since $k + (m+1) = (k+1) + m$). In other words,

$$bx_0 y_{(k+1)+m} + x_1 y_{(k+1)+m+1} = bx_{k+1}y_m + x_{(k+1)+1}y_{m+1}.$$

Forget that we fixed m. We thus have proved that

$$bx_0 y_{(k+1)+m} + x_1 y_{(k+1)+m+1} = bx_{k+1}y_m + x_{(k+1)+1}y_{m+1}$$

for every nonnegative integer m. In other words, (189) holds for $n = k+1$. This completes the induction step. Thus, (189) is proved, so that Exercise 4.9.4 is solved. □

Again, an alternative solution could have been given using Theorem 4.9.11.

As we said, Exercise 4.9.4 is a generalization of Exercise 4.9.3; thus, having solved the former, we can now quickly obtain the latter.

Solution to Exercise 4.9.3. Let n and m be two nonnegative integers. Then, Exercise 4.9.4 yields

$$bx_0 y_{n+m} + x_1 y_{n+m+1} = bx_n y_m + x_{n+1} y_{m+1}.$$

Comparing this with

$$b \underbrace{x_0}_{=0} y_{n+m} + \underbrace{x_1}_{=1} y_{n+m+1} = \underbrace{b \cdot 0 y_{n+m}}_{=0} + y_{n+m+1} = y_{n+m+1},$$

we obtain $y_{n+m+1} = bx_n y_m + x_{n+1} y_{m+1}$. This solves Exercise 4.9.3. □

Here is a generalization of Proposition 4.3.20 to (a,b)-recurrent sequences:

Proposition 4.9.18. Let a and b be two numbers such that $a \neq 0$. Let $(x_0, x_1, x_2, \ldots)$ be an (a,b)-recurrent sequence with $x_0 = 0$ and $x_1 = 1$. Let $n \in \{-1, 0, 1, \ldots\}$. Then,

$$x_{n+1} = \sum_{k=0}^{n} \binom{n-k}{k} a^{n-2k} b^k. \tag{193}$$

Note that the "$a \neq 0$" condition in Proposition 4.9.18 is needed only to ensure that the a^{n-2k} term does not become undefined when $n-2k$ is negative. In truth, the addends on the right hand side of (193) in which $n-2k$ is negative are irrelevant, since the binomial coefficients $\binom{n-k}{k}$ in them vanish anyway; thus, we could just as well change the upper limit of the sum from n to $\lfloor n/2 \rfloor$ and stop worrying about what happens if $a = 0$. However, letting the sum range from 0 to n makes for a slightly simpler formula and a simpler proof.

Proof of Proposition 4.9.18. We proceed by strong induction on n:

Induction step: Let $m \in \{-1, 0, 1, \ldots\}$. Assume (as the induction hypothesis) that Proposition 4.9.18 holds for $n < m$. We must prove that Proposition 4.9.18 holds for $n = m$. In other words, we must prove that

$$x_{m+1} = \sum_{k=0}^{m} \binom{m-k}{k} a^{m-2k} b^k. \tag{194}$$

If $m < 1$, then this is easy to do directly[98]. Thus, for the rest of this proof, we WLOG assume that $m \geq 1$. Hence, the numbers $m-1$ and $m-2$ both belong to

[98] *Proof.* Assume that $m < 1$. We must prove that (194) holds.

We have $m < 1$. Thus, either $m = -1$ or $m = 0$ (since $m \in \{-1, 0, 1, \ldots\}$). In other words, we are in one of the following two cases:

Case 1: We have $m = -1$.

Case 2: We have $m = 0$.

Let us first consider Case 1. In this case, we have $m = -1$. Hence, $m + 1 = 0$, so that

$$x_{m+1} = x_0 = 0.$$

Comparing this with

$$\begin{aligned} \sum_{k=0}^{m} \binom{m-k}{k} a^{m-2k} b^k &= \sum_{k=0}^{-1} \binom{m-k}{k} a^{m-2k} b^k \qquad (\text{since } m = -1) \\ &= (\text{empty sum}) = 0, \end{aligned}$$

we obtain $x_{m+1} = \sum_{k=0}^{m} \binom{m-k}{k} a^{m-2k} b^k$. Hence, we have shown that (194) holds in Case 1.

Let us now consider Case 2. In this case, we have $m = 0$. Hence, $m + 1 = 1$, so that

$$x_{m+1} = x_1 = 1.$$

$\{-1, 0, 1, \ldots\}$. Also, $m - 1 \geq 0$ (since $m \geq 1$). Furthermore, $m \geq 1 > 0$; therefore, Proposition 4.3.4 (applied to $n = 0$ and $k = m$) yields $\binom{0}{m} = 0$. (We'll need this later.)

We have assumed that Proposition 4.9.18 holds for $n < m$. Hence, Proposition 4.9.18 holds for $n = m - 1$ (since $m - 1 \in \{-1, 0, 1, \ldots\}$ and $m - 1 < m$). In other words, we have

$$x_{(m-1)+1} = \sum_{k=0}^{m-1} \binom{m-1-k}{k} a^{m-1-2k} b^k.$$

In view of $(m - 1) + 1 = m$, this rewrites as

$$x_m = \sum_{k=0}^{m-1} \binom{m-1-k}{k} a^{m-1-2k} b^k. \tag{195}$$

We have assumed that Proposition 4.9.18 holds for $n < m$. Hence, Proposition 4.9.18 holds for $n = m - 2$ (since $m - 2 \in \{-1, 0, 1, \ldots\}$ and $m - 2 < m$). In other words, we have

$$x_{(m-2)+1} = \sum_{k=0}^{m-2} \binom{m-2-k}{k} a^{m-2-2k} b^k.$$

In view of $(m - 2) + 1 = m - 1$, this rewrites as

$$x_{m-1} = \sum_{k=0}^{m-2} \binom{m-2-k}{k} a^{m-2-2k} b^k. \tag{196}$$

The sequence $(x_0, x_1, x_2, \ldots)$ is (a, b)-recurrent. In other words, every $n \geq 2$ satisfies $x_n = ax_{n-1} + bx_{n-2}$ (by the definition of "(a, b)-recurrent"). Applying this equality to $n = m + 1$, we obtain $x_{m+1} = ax_{(m+1)-1} + bx_{(m+1)-2} = ax_m + bx_{m-1}$.

Comparing this with

$$\begin{aligned}\sum_{k=0}^{m} \binom{m-k}{k} a^{m-2k} b^k &= \sum_{k=0}^{0} \binom{0-k}{k} a^{0-2k} b^k \qquad \text{(since } m = 0\text{)} \\ &= \underbrace{\binom{0-0}{0}}_{=1} \underbrace{a^{0-2\cdot 0}}_{=a^0=1} \underbrace{b^0}_{=1} = 1,\end{aligned}$$

we obtain $x_{m+1} = \sum_{k=0}^{m} \binom{m-k}{k} a^{m-2k} b^k$. Hence, we have shown that (194) holds in Case 2.

We have now shown that (194) holds in each of the two Cases 1 and 2. Thus, (194) always holds (under our assumption that $m < 1$). Qed.

Now, multiplying both sides of the equality (195) by a, we find

$$ax_m = a\sum_{k=0}^{m-1}\binom{m-1-k}{k}a^{m-1-2k}b^k = \sum_{k=0}^{m-1}\underbrace{\binom{m-1-k}{k}}_{\substack{=\binom{m-k-1}{k} \\ \text{(since } m-1-k=m-k-1)}}\underbrace{aa^{m-1-2k}}_{=a^{(m-1-2k)+1}=a^{m-2k}}b^k$$

$$= \sum_{k=0}^{m-1}\binom{m-k-1}{k}a^{m-2k}b^k. \tag{197}$$

Furthermore, multiplying both sides of the equality (196) by b, we find

$$\begin{aligned} bx_{m-1} &= b\sum_{k=0}^{m-2}\binom{m-2-k}{k}a^{m-2-2k}b^k = \sum_{k=0}^{m-2}\binom{m-2-k}{k}a^{m-2-2k}bb^k \\ &= \sum_{k=1}^{m-1}\underbrace{\binom{m-2-(k-1)}{k-1}}_{\substack{=\binom{m-k-1}{k-1} \\ \text{(since } m-2-(k-1)=m-k-1)}}\underbrace{a^{m-2-2(k-1)}}_{\substack{=a^{m-2k} \\ \text{(since } m-2-2(k-1)=m-2k)}}\underbrace{bb^{k-1}}_{=b^k} \\ &\qquad \text{(here, we have substituted } k-1 \text{ for } k \text{ in the sum)} \\ &= \sum_{k=1}^{m-1}\binom{m-k-1}{k-1}a^{m-2k}b^k. \end{aligned}$$

Comparing this with

$$\begin{aligned} &\sum_{k=0}^{m-1}\binom{m-k-1}{k-1}a^{m-2k}b^k \\ &= \underbrace{\binom{m-0-1}{0-1}}_{\substack{=0 \\ \text{(by (118), since } 0-1\notin\mathbb{N})}}a^{m-2\cdot 0}b^0 + \sum_{k=1}^{m-1}\binom{m-k-1}{k-1}a^{m-2k}b^k \qquad (\text{since } m-1\geq 0) \\ &= \underbrace{0a^{m-2\cdot 0}b^0}_{=0} + \sum_{k=1}^{m-1}\binom{m-k-1}{k-1}a^{m-2k}b^k = \sum_{k=1}^{m-1}\binom{m-k-1}{k-1}a^{m-2k}b^k, \end{aligned}$$

we obtain

$$bx_{m-1} = \sum_{k=0}^{m-1}\binom{m-k-1}{k-1}a^{m-2k}b^k. \tag{198}$$

Now,

$$\begin{aligned}
x_{m+1} &= ax_m + bx_{m-1} = bx_{m-1} + ax_m \\
&= \sum_{k=0}^{m-1} \binom{m-k-1}{k-1} a^{m-2k} b^k + \sum_{k=0}^{m-1} \binom{m-k-1}{k} a^{m-2k} b^k \\
&\qquad \text{(by adding the equalities (198) and (197))} \\
&= \sum_{k=0}^{m-1} \left(\binom{m-k-1}{k-1} a^{m-2k} b^k + \binom{m-k-1}{k} a^{m-2k} b^k \right) \\
&= \sum_{k=0}^{m-1} \left(\binom{m-k-1}{k-1} + \binom{m-k-1}{k} \right) a^{m-2k} b^k. \qquad (199)
\end{aligned}$$

Comparing this with

$$\begin{aligned}
\sum_{k=0}^{m} \binom{m-k}{k} a^{m-2k} b^k &= \sum_{k=0}^{m-1} \underbrace{\binom{m-k}{k}}_{\substack{=\binom{m-k-1}{k-1}+\binom{m-k-1}{k} \\ \text{(by Theorem 4.3.7, applied to } n=m-k)}} a^{m-2k} b^k + \underbrace{\binom{m-m}{m}}_{=\binom{0}{m}=0} a^{m-2m} b^m \\
&= \sum_{k=0}^{m-1} \left(\binom{m-k-1}{k-1} + \binom{m-k-1}{k} \right) a^{m-2k} b^k + \underbrace{0 a^{m-2m} b^m}_{=0} \\
&= \sum_{k=0}^{m-1} \left(\binom{m-k-1}{k-1} + \binom{m-k-1}{k} \right) a^{m-2k} b^k,
\end{aligned}$$

we obtain

$$x_{m+1} = \sum_{k=0}^{m} \binom{m-k}{k} a^{m-2k} b^k.$$

In other words, Proposition 4.9.18 holds for $n = m$. This completes the induction step. Thus, Proposition 4.9.18 is proved. □

Proposition 4.9.18 is, in a way, an explicit formula for any entry of an (a, b)-recurrent sequence that begins with 0 and 1. Unlike the formulas in Theorem 4.9.11, it involves no irrational numbers; however, it involves a finite sum and binomial coefficients. Its other disadvantage is that it only applies to a very specific (a, b)-recurrent sequence (namely, the one starting with 0 and 1) rather than to any arbitrary (a, b)-recurrent sequence. This disadvantage, however, can easily be amended: In fact, every (a, b)-recurrent sequence can be expressed through the very specific one that starts with 0 and 1. Namely, we have the following:

Proposition 4.9.19. Let a and b be two numbers. Let $(x_0, x_1, x_2, \ldots)$ be an (a, b)-recurrent sequence with $x_0 = 0$ and $x_1 = 1$. Let $(y_0, y_1, y_2, \ldots)$ be an arbitrary (a, b)-recurrent sequence. Then, $y_{n+1} = by_0 x_n + y_1 x_{n+1}$ for each $n \in \mathbb{N}$.

Proof of Proposition 4.9.19. Exercise 4.9.4 (applied to $m = 0$) yields

$$bx_0y_{n+0} + x_1y_{n+0+1} = bx_ny_0 + x_{n+1}\underbrace{y_{0+1}}_{=y_1} = bx_ny_0 + x_{n+1}y_1 = by_0x_n + y_1x_{n+1}.$$

In view of

$$b\underbrace{x_0}_{=0}y_{n+0} + \underbrace{x_1}_{=1}\underbrace{y_{n+0+1}}_{=y_{n+1}} = \underbrace{b0y_{n+0}}_{=0} + y_{n+1} = y_{n+1},$$

this rewrites as $y_{n+1} = by_0x_n + y_1x_{n+1}$. This proves Proposition 4.9.19. □

Proposition 4.3.20 is a particular case of Proposition 4.9.18:

Proof of Proposition 4.3.20. We have $n \in \mathbb{N} \subseteq \{-1, 0, 1, \ldots\}$ and $1 \neq 0$. The Fibonacci sequence $(f_0, f_1, f_2, \ldots)$ is $(1, 1)$-recurrent (as we have already seen) and satisfies $f_0 = 0$ and $f_1 = 1$. Hence, Proposition 4.9.18 (applied to $a = 1$, $b = 1$ and $x_i = f_i$) yields

$$\begin{aligned} f_{n+1} &= \sum_{k=0}^{n} \binom{n-k}{k} \underbrace{1^{n-2k}}_{=1} \underbrace{1^k}_{=1} \\ &= \sum_{k=0}^{n} \binom{n-k}{k} = \binom{n-0}{0} + \binom{n-1}{1} + \binom{n-2}{2} + \cdots + \binom{n-n}{n}. \end{aligned}$$

This proves Proposition 4.3.20. □

Another particular case of Proposition 4.9.18 is [Grinbe15, Exercise 4.4].

Yet another particular case of Proposition 4.9.18 is the following corollary, whose connection to (a, b)-recurrent sequences is not immediately visible:

Corollary 4.9.20. Let $n \in \{-1, 0, 1, \ldots\}$. Then,

$$n + 1 = \sum_{k=0}^{n} (-1)^k \binom{n-k}{k} 2^{n-2k}.$$

Proof of Corollary 4.9.20. We have $2 \neq 0$. The sequence $(0, 1, 2, 3, \ldots)$ is $(2, -1)$-recurrent[99] and starts with the entries $0 = 0$ and $1 = 1$. Hence, Proposition 4.9.18 (applied to $a = 2$, $b = -1$ and $x_i = i$) yields

$$n + 1 = \sum_{k=0}^{n} \underbrace{\binom{n-k}{k} 2^{n-2k} (-1)^k}_{=(-1)^k \binom{n-k}{k} 2^{n-2k}} = \sum_{k=0}^{n} (-1)^k \binom{n-k}{k} 2^{n-2k}.$$

This proves Corollary 4.9.20. □

[99] Indeed, this follows from the fact that this sequence is an arithmetic progression, but (as we saw in Example 4.9.4) every arithmetic progression is $(2, -1)$-recurrent. Alternatively, you can check it directly: Each integer $n \geq 2$ satisfies $n = 2(n-1) + (-1)(n-2)$; but this is saying precisely that the sequence $(0, 1, 2, 3, \ldots)$ is $(2, -1)$-recurrent (by the definition of "$(2, -1)$-recurrent").

4.9.4. Two-term recurrences: the matrix approach

For now, we have been approaching (a, b)-recurrent sequences in an "entry-by-entry" way. However, there are some more conceptual approaches. The simplest one is the *matrix approach*. In essence, it says that instead of considering single entries x_i of an (a, b)-recurrent sequence $(x_0, x_1, x_2, \ldots)$, it is better to consider pairs (x_i, x_{i+1}) of two consecutive entries. Better yet, we can encode these pairs (x_i, x_{i+1}) as column vectors $\begin{pmatrix} x_i \\ x_{i+1} \end{pmatrix}$. Here is what we gain when doing this:

Proposition 4.9.21. Let a and b be any two numbers. Let A be the 2×2-matrix $\begin{pmatrix} 0 & 1 \\ b & a \end{pmatrix}$.

Let $(x_0, x_1, x_2, \ldots)$ be an (a, b)-recurrent sequence of numbers. For each $i \in \mathbb{N}$, define a column vector v_i by $v_i = \begin{pmatrix} x_i \\ x_{i+1} \end{pmatrix}$. Then:

(a) We have $Av_i = v_{i+1}$ for each $i \in \mathbb{N}$.

(b) We have $A^n v_i = v_{i+n}$ for each $i \in \mathbb{N}$ and $n \in \mathbb{N}$.

Proposition 4.9.21 shows what is so nice about the vectors v_i as opposed to the single entries x_i: We cannot compute an entry x_{i+1} from knowing x_i alone (we need x_{i-1} as well), but we can compute a vector v_{i+1} from knowing v_i alone (by Proposition 4.9.21 **(a)**). Furthermore, Proposition 4.9.21 **(b)** (applied to $i = 0$) says that $A^n v_0 = v_n$ for each $n \in \mathbb{N}$; this lets us quickly compute v_n (and thus x_n) if we can quickly compute A^n. Fortunately, powers of matrices can be computed quickly (exponentiation by squaring does the trick). This yields one of the quickest ways to exactly compute any entry of an (a, b)-recurrent sequence.[100]

The proof of Proposition 4.9.21 is rather easy:

Proof of Proposition 4.9.21. **(a)** Let $i \in \mathbb{N}$. The definition of v_i yields $v_i = \begin{pmatrix} x_i \\ x_{i+1} \end{pmatrix}$. The definition of v_{i+1} yields $v_{i+1} = \begin{pmatrix} x_{i+1} \\ x_{i+2} \end{pmatrix}$.

The sequence $(x_0, x_1, x_2, \ldots)$ is (a, b)-recurrent. In other words, every $n \geq 2$ satisfies $x_n = ax_{n-1} + bx_{n-2}$ (by the definition of "(a, b)-recurrent"). Applying this equality to $n = i + 2$, we obtain $x_{i+2} = ax_{i+1} + bx_i = bx_i + ax_{i+1}$. Now, recall that $A = \begin{pmatrix} 0 & 1 \\ b & a \end{pmatrix}$ (by the definition of A). Multiplying this equality by $v_i = \begin{pmatrix} x_i \\ x_{i+1} \end{pmatrix}$, we find

$$Av_i = \begin{pmatrix} 0 & 1 \\ b & a \end{pmatrix} \begin{pmatrix} x_i \\ x_{i+1} \end{pmatrix} = \begin{pmatrix} 0x_i + 1x_{i+1} \\ bx_i + ax_{i+1} \end{pmatrix} = \begin{pmatrix} x_{i+1} \\ bx_i + ax_{i+1} \end{pmatrix}$$

[100] This phenomenon, where it is easier to work with a "bundle" of several entries of a sequence (or values of a function) instead of a single entry (or value), is omnipresent in mathematics and computer science.

(since $0x_i + 1x_{i+1} = x_{i+1}$). Comparing this with

$$v_{i+1} = \begin{pmatrix} x_{i+1} \\ x_{i+2} \end{pmatrix} = \begin{pmatrix} x_{i+1} \\ bx_i + ax_{i+1} \end{pmatrix} \qquad (\text{since } x_{i+2} = bx_i + ax_{i+1}),$$

we obtain $Av_i = v_{i+1}$. This proves Proposition 4.9.21 **(a)**.

(b) Fix $i \in \mathbb{N}$. We must show that

$$A^n v_i = v_{i+n} \tag{200}$$

for each $n \in \mathbb{N}$. We show this by induction on n:

Induction base: We use the notation I_2 for the 2×2 identity matrix. Then, $A^0 = I_2$. Thus, $\underbrace{A^0}_{=I_2} v_i = I_2 v_i = v_i = v_{i+0}$ (since $i = i + 0$). In other words, our claim (200) holds for $n = 0$.

Induction step: Let $m \in \mathbb{N}$. Assume (as the induction hypothesis) that (200) holds for $n = m$. We must prove that (200) holds for $n = m + 1$. In other words, we must prove that $A^{m+1} v_i = v_{i+m+1}$.

We have assumed that (200) holds for $n = m$. In other words, $A^m v_i = v_{i+m}$. Now,

$$\underbrace{A^{m+1}}_{=AA^m} v_i = A \underbrace{A^m v_i}_{=v_{i+m}} = Av_{i+m} = v_{i+m+1}$$

(by Proposition 4.9.21 **(a)**, applied to $i + m$ instead of i). In other words, (200) holds for $n = m + 1$. This completes the induction step. Thus, (200) is proved. In other words, Proposition 4.9.21 **(b)** is proved. □

As an illustration of how useful Proposition 4.9.21 is, let us use it to solve Exercise 4.9.2 again:

Second solution to Exercise 4.9.2 (sketched). If u and v are two column vectors of size 2, then $(u \mid v)$ shall denote the 2×2-matrix whose columns are u and v (from left to right). It is well-known (see, e.g., [Heffer20, Chapter Three, Section IV, Lemma 3.7]) that any 2×2-matrix B and any two column vectors u and v of size 2 satisfy

$$B(u \mid v) = (Bu \mid Bv). \tag{201}$$

For each $i \in \mathbb{N}$, let V_i be the 2×2-matrix $(v_i \mid v_{i+1})$. Thus, explicitly,

$$V_i = \begin{pmatrix} x_i & x_{i+1} \\ x_{i+1} & x_{i+2} \end{pmatrix} \tag{202}$$

(since the definitions of v_i and v_{i+1} yield $v_i = \begin{pmatrix} x_i \\ x_{i+1} \end{pmatrix}$ and $v_{i+1} = \begin{pmatrix} x_{i+1} \\ x_{i+2} \end{pmatrix}$). Therefore, the determinant of V_i is

$$\det(V_i) = \det \begin{pmatrix} x_i & x_{i+1} \\ x_{i+1} & x_{i+2} \end{pmatrix} = x_i x_{i+2} - x_{i+1}^2 \tag{203}$$

for each $i \in \mathbb{N}$.

It is well-known (see, e.g., [Grinbe15, Theorem 6.23]) that any $n \in \mathbb{N}$ and any two $n \times n$-matrices A and B satisfy

$$\det(AB) = \det A \cdot \det B. \tag{204}$$

(In the current solution, we shall only use this equality in the case $n = 2$; in this case it can be proved by straightforward computation.)

An easy consequence of (204) is the following formula (see, e.g., [Grinbe15, Corollary 6.25 **(b)**]): If $n \in \mathbb{N}$, and if B is an $n \times n$-matrix, then

$$\det\left(B^k\right) = (\det B)^k \tag{205}$$

for any $k \in \mathbb{N}$. (This follows by induction on k.)

Now, let n be a positive integer. Thus, $n - 1 \in \mathbb{N}$. Let A be the 2×2-matrix $\begin{pmatrix} 0 & 1 \\ b & a \end{pmatrix}$. Then, Proposition 4.9.21 **(b)** (applied to $n-1$ and 0 instead of n and i) yields $A^{n-1}v_0 = v_{0+n-1} = v_{n-1}$. Also, Proposition 4.9.21 **(b)** (applied to $n-1$ and 1 instead of n and i) yields $A^{n-1}v_1 = v_{1+n-1} = v_n$. But the definition of V_0 yields $V_0 = (v_0 \mid v_1)$. Hence,

$$\begin{aligned} A^{n-1} \underbrace{V_0}_{=(v_0 \mid v_1)} &= A^{n-1}(v_0 \mid v_1) = \left(A^{n-1}v_0 \mid A^{n-1}v_1\right) \\ &\qquad \left(\text{by (201), applied to } B = A^{n-1},\ u = v_0 \text{ and } v = v_1\right) \\ &= (v_{n-1} \mid v_n) \qquad \left(\text{since } A^{n-1}v_0 = v_{n-1} \text{ and } A^{n-1}v_1 = v_n\right) \\ &= V_{n-1} \qquad (\text{since } V_{n-1} \text{ was defined as } (v_{n-1} \mid v_n)). \end{aligned}$$

Hence,

$$\det\left(A^{n-1}V_0\right) = \det(V_{n-1}) = x_{n-1}x_{n+1} - x_n^2$$

(by (203), applied to $i = n-1$). Therefore,

$$\begin{aligned} & x_{n-1}x_{n+1} - x_n^2 \\ &= \det\left(A^{n-1}V_0\right) = \underbrace{\det\left(A^{n-1}\right)}_{\substack{=(\det A)^{n-1} \\ \text{(by (205), applied to } n=2,\ B=A \text{ and } k=n-1)}} \underbrace{\det(V_0)}_{\substack{=x_0x_2-x_1^2 \\ \text{(by (203), applied to } i=0)}} \\ &\qquad \left(\text{by (204), applied to 2, } A^{n-1} \text{ and } V_0 \text{ instead of } n,\ A \text{ and } B\right) \\ &= (\det A)^{n-1} \cdot \left(x_0x_2 - x_1^2\right). \end{aligned} \tag{206}$$

But $A = \begin{pmatrix} 0 & 1 \\ b & a \end{pmatrix}$ and thus $\det A = \det\begin{pmatrix} 0 & 1 \\ b & a \end{pmatrix} = 0a - 1b = -b$. Hence, (206) rewrites as $x_{n-1}x_{n+1} - x_n^2 = (-b)^{n-1} \cdot \left(x_0x_2 - x_1^2\right)$. In other words, $x_{n+1}x_{n-1} - x_n^2 = (-b)^{n-1}\left(x_2x_0 - x_1^2\right)$. This solves Exercise 4.9.2 again. $\square$

Another application of Proposition 4.9.21 is a second proof of Theorem 4.9.11 (recall that the first proof was sketched in our discussion of Exercise 4.9.1):

Second proof of Theorem 4.9.11 (sketched). Let A be the 2×2-matrix $\begin{pmatrix} 0 & 1 \\ b & a \end{pmatrix}$. Proposition 4.9.21 **(b)** (applied to $i = 0$) yields that each $n \in \mathbb{N}$ satisfies

$$A^n v_0 = v_{0+n} = v_n = \begin{pmatrix} x_n \\ x_{n+1} \end{pmatrix} \tag{207}$$

(by the definition of v_n). We thus obtain an explicit formula for v_n if we have an explicit formula for A^n.

How can we find an explicit formula for A^n ? In linear algebra, the powers of a matrix are easiest to compute by diagonalizing the matrix. Indeed, if we can diagonalize the matrix A as $A = TDT^{-1}$ (with T invertible and D diagonal), then

$$A^n = TD^nT^{-1} \qquad \text{for every } n \in \mathbb{N}. \tag{208}$$

(If you don't know this, you can easily prove it by induction on n.) For a diagonal matrix D, computing its powers D^n is a trivial task (it suffices to take each diagonal entry of D to the n-th power). Thus, in order to compute A^n, it will suffice to diagonalize A.

Not every square matrix can be diagonalized; however, over the complex numbers, "most" matrices can be[101]. Diagonalizing a matrix is particularly easy for a 2×2-matrix like A, since it only requires solving a quadratic (as opposed to a higher-degree) equation. Let me omit the details and merely state the result:

- If $a^2 + 4b \neq 0$, then A can be diagonalized as $A = TDT^{-1}$, where we define two complex numbers λ and μ as in (184) and set

  $$D = \operatorname{diag}(\lambda, \mu) = \begin{pmatrix} \lambda & 0 \\ 0 & \mu \end{pmatrix} \qquad \text{and} \qquad T = \begin{pmatrix} 1 & 1 \\ \lambda & \mu \end{pmatrix}.$$

  (The appearance of the numbers λ and μ should not be surprising: They are the two eigenvalues of A.)

- If $a^2 + 4b = 0$, then A cannot be diagonalized.

Let us first draw some conclusions in the case $a^2 + 4b \neq 0$, and then discuss what can be done in the case $a^2 + 4b = 0$.

Assume that $a^2 + 4b \neq 0$. Then, we have just found a way to diagonalize A. Namely, we have defined two complex numbers λ and μ as in (184), and we have set

$$D = \operatorname{diag}(\lambda, \mu) = \begin{pmatrix} \lambda & 0 \\ 0 & \mu \end{pmatrix} \qquad \text{and} \qquad T = \begin{pmatrix} 1 & 1 \\ \lambda & \mu \end{pmatrix}.$$

[101] The word "most" here can be interpreted in many ways, most of which are correct :) We will soon see what it means in our specific case.

Then, $A = TDT^{-1}$. Thus, for each $n \in \mathbb{N}$, we have

$$\begin{aligned}
A^n &= TD^nT^{-1} \qquad \text{(by (208))} \\
&= \begin{pmatrix} 1 & 1 \\ \lambda & \mu \end{pmatrix} \begin{pmatrix} \lambda & 0 \\ 0 & \mu \end{pmatrix}^n \begin{pmatrix} 1 & 1 \\ \lambda & \mu \end{pmatrix}^{-1} \qquad \left(\text{since } T = \begin{pmatrix} 1 & 1 \\ \lambda & \mu \end{pmatrix} \text{ and } D = \begin{pmatrix} \lambda & 0 \\ 0 & \mu \end{pmatrix}\right) \\
&= \begin{pmatrix} 1 & 1 \\ \lambda & \mu \end{pmatrix} \begin{pmatrix} \lambda^n & 0 \\ 0 & \mu^n \end{pmatrix} \begin{pmatrix} 1 & 1 \\ \lambda & \mu \end{pmatrix}^{-1} \qquad \left(\text{since } \begin{pmatrix} \lambda & 0 \\ 0 & \mu \end{pmatrix}^n = \begin{pmatrix} \lambda^n & 0 \\ 0 & \mu^n \end{pmatrix}\right) \\
&= \begin{pmatrix} \dfrac{\lambda\mu^n - \lambda^n\mu}{\lambda - \mu} & \dfrac{\lambda^n - \mu^n}{\lambda - \mu} \\ -\lambda\mu\dfrac{\lambda^n - \mu^n}{\lambda - \mu} & \dfrac{\lambda\lambda^n - \mu\mu^n}{\lambda - \mu} \end{pmatrix}
\end{aligned}$$

(by straightforward matrix computations). Hence,

$$\begin{aligned}
A^n v_0 &= \begin{pmatrix} \dfrac{\lambda\mu^n - \lambda^n\mu}{\lambda - \mu} & \dfrac{\lambda^n - \mu^n}{\lambda - \mu} \\ -\lambda\mu\dfrac{\lambda^n - \mu^n}{\lambda - \mu} & \dfrac{\lambda\lambda^n - \mu\mu^n}{\lambda - \mu} \end{pmatrix} \begin{pmatrix} x_0 \\ x_1 \end{pmatrix} \\
&= \begin{pmatrix} \dfrac{\lambda\mu^n - \lambda^n\mu}{\lambda - \mu} x_0 + \dfrac{\lambda^n - \mu^n}{\lambda - \mu} x_1 \\ * \end{pmatrix}. \qquad (209)
\end{aligned}$$

Here, the asterisk $*$ in the bottom row means "we don't care what the value here is" (we will see why in a moment). Comparing this with (207), we obtain

$$\begin{pmatrix} x_n \\ x_{n+1} \end{pmatrix} = \begin{pmatrix} \dfrac{\lambda\mu^n - \lambda^n\mu}{\lambda - \mu} x_0 + \dfrac{\lambda^n - \mu^n}{\lambda - \mu} x_1 \\ * \end{pmatrix}.$$

Comparing the top entries of the vectors on both sides of this equality, we find

$$x_n = \frac{\lambda\mu^n - \lambda^n\mu}{\lambda - \mu} x_0 + \frac{\lambda^n - \mu^n}{\lambda - \mu} x_1 = \frac{x_1 - \mu x_0}{\lambda - \mu}\lambda^n + \frac{\lambda x_0 - x_1}{\lambda - \mu}\mu^n = \gamma\lambda^n + \delta\mu^n,$$

where γ and δ are as in (185). So we have recovered the formula (183) and thus reproved Theorem 4.9.11 **(a)**. Note that this approach to finding an explicit formula for x_n was entirely self-motivated (assuming that you know about diagonalization of matrices); unlike our first approach above, it did not rely on us reverse-engineering the Binet formula (which came out of clear skies).

Let us now consider the case when $a^2 + 4b = 0$. In this case, the matrix A cannot be diagonalized. The closest thing to a diagonalization in such a case is the *Jordan normal form* (see, e.g., [Heffer20, Chapter Five, Section IV]). Again omitting the details, a computation shows that the Jordan normal form of A is

$$A = TJT^{-1},$$

where

$$T = \begin{pmatrix} 0 & 1 \\ 1 & \frac{a}{2} \end{pmatrix} \qquad \text{and} \qquad J = \begin{pmatrix} \frac{a}{2} & 0 \\ 1 & \frac{a}{2} \end{pmatrix}.$$

(The computation is best done by rewriting b as $\frac{-a^2}{4}$ right away, thus getting rid of the dependent parameter b.) In order to compute A^n, we thus need to know how to compute J^n. Fortunately, there is a formula for this:

$$\begin{pmatrix} d & 0 \\ u & d \end{pmatrix}^n = \begin{pmatrix} d^n & 0 \\ nd^{n-1}u & d^n \end{pmatrix} \tag{210}$$

for any numbers d and u and any $n \in \mathbb{N}$ (where we agree to interpret nd^{n-1} as 0 when $n = 0$). This formula (210) is a particular case of a general rule for taking powers of Jordan blocks, but it also can easily be proved directly[102]. Now, from $A = TJT^{-1}$, we obtain

$$A^n = TJ^nT^{-1} \qquad \text{(by (208), applied to } J \text{ instead of } D\text{)}$$

for each $n \in \mathbb{N}$. Recalling how T and J were defined, and using (210) to compute J^n, we thus get an explicit expression for A^n, which we can then turn into an explicit expression for x_n in the same way as we did in the case $a^2 + 4b \neq 0$ above. Thus we recover Theorem 4.9.11 **(b)**. □

4.9.5. Recitation #4: Two-term recurrences

Here are two more exercises on (a, b)-recurrent sequences:

Exercise 4.9.5. Define a sequence $(a_0, a_1, a_2, \ldots)$ of integers recursively by setting $a_0 = 2$ and $a_1 = 1$ and $a_n = a_{n-1} + 2a_{n-2}$ for each $n \geq 2$.

Find an explicit formula for a_n.

Discussion of Exercise 4.9.5. The most systematic approach to this is by using Theorem 4.9.11. Indeed, the sequence $(a_0, a_1, a_2, \ldots)$ is $(1, 2)$-recurrent (due to the equality $a_n = a_{n-1} + 2a_{n-2}$ that holds for each $n \geq 2$). Since $1^2 + 4 \cdot 2 = 9 \neq 0$, we can thus apply Theorem 4.9.11 **(a)** to $a = 1$ and $b = 2$ and $x_i = a_i$. We thus conclude that every $n \in \mathbb{N}$ satisfies

$$a_n = \gamma\lambda^n + \delta\mu^n, \tag{211}$$

where we set

$$\lambda = \frac{1 + \sqrt{1^2 + 4 \cdot 1}}{2} \qquad \text{and} \qquad \mu = \frac{1 - \sqrt{1^2 + 4 \cdot 1}}{2} \tag{212}$$

[102] With our induction experience, we can prove it in our sleep.

and

$$\gamma = \frac{a_1 - \mu a_0}{\lambda - \mu} \qquad \text{and} \qquad \delta = \frac{\lambda a_0 - a_1}{\lambda - \mu}. \tag{213}$$

The equalities (212) yield

$$\lambda = \frac{1 + \sqrt{1^2 + 4 \cdot 1}}{2} = \frac{1 + \sqrt{9}}{2} = \frac{1+3}{2} = 2 \qquad \text{and}$$
$$\mu = \frac{1 - \sqrt{1^2 + 4 \cdot 1}}{2} = \frac{1 - \sqrt{9}}{2} = \frac{1-3}{2} = -1.$$

Hence, (213) yields

$$\gamma = \frac{a_1 - \mu a_0}{\lambda - \mu} = \frac{1 - (-1) \cdot 2}{2 - (-1)} \qquad (\text{since } \lambda = 2 \text{ and } \mu = -1 \text{ and } a_0 = 2 \text{ and } a_1 = 1)$$
$$= 1$$

and (by a similar computation) $\delta = 1$. Thus, (211) shows that every $n \in \mathbb{N}$ satisfies

$$a_n = \underbrace{\gamma}_{=1} \lambda^n + \underbrace{\delta}_{=1} \mu^n = \lambda^n + \mu^n = 2^n + (-1)^n$$

(since $\lambda = 2$ and $\mu = -1$). This solves Exercise 4.9.5.

Here are two alternative ways to get the formula $a_n = 2^n + (-1)^n$:

- One can simply compute the first few entries of the sequence $(a_0, a_1, a_2, \ldots)$ and observe that they each differ by 1 from the respective powers of 2. The formula $a_n = 2^n + (-1)^n$ is then easily guessed and just as easily proved (by strong induction on n).

- Here is a more sophisticated approach: Rewrite the recursive relation $a_n = a_{n-1} + 2a_{n-2}$ as $a_n + a_{n-1} = 2(a_{n-1} + a_{n-2})$ (indeed, adding a_{n-1} to both sides of $a_n = a_{n-1} + 2a_{n-2}$ yields $a_n + a_{n-1} = 2a_{n-1} + 2a_{n-2} = 2(a_{n-1} + a_{n-2})$). We can rewrite the latter equation as $b_{n-1} = 2b_{n-2}$, where we define an auxiliary sequence $(b_0, b_1, b_2, \ldots)$ by $b_m := a_m + a_{m+1}$. So we have $b_{n-1} = 2b_{n-2}$ for each $n \geq 2$; equivalently, $b_m = 2b_{m-1}$ for each $m \geq 1$. In other words, each entry of the sequence $(b_0, b_1, b_2, \ldots)$ equals twice the preceding entry. In other words, the sequence $(b_0, b_1, b_2, \ldots)$ is a geometric progression with ratio 2. Thus, $b_m = 2^m b_0$ for each $m \in \mathbb{N}$. In view of $b_0 = \underbrace{a_0}_{=2} + \underbrace{a_1}_{=1} = 2 + 1 = 3$, this rewrites as $b_m = 2^m \cdot 3$ for each $m \in \mathbb{N}$.

  Now, how do we recover the a_n from the b_m ? Experiments with small entries suggest (and a simple argument using the telescope principle proves) that

  $$a_n = b_{n-1} - b_{n-2} + b_{n-3} - b_{n-4} \pm \cdots \pm b_0 \pm (-a_0).$$

  (The last two $\pm$ signs are $+$ signs if n is odd, and $-$ signs if n is even.) Plugging our formula $b_m = 2^m \cdot 3$ into this equality, and summing the resulting geometric sum, we eventually obtain $a_n = 2^n + (-1)^n$.

□

Exercise 4.9.6. Let $(f_0, f_1, f_2, \ldots)$ be the Fibonacci sequence. Let $\varphi = \dfrac{1+\sqrt{5}}{2}$.

(a) Prove that $f_n = \left\lfloor \dfrac{1}{\sqrt{5}}\varphi^n + \dfrac{1}{2} \right\rfloor$ for each $n \in \mathbb{N}$.

(b) Prove that $\varphi^n = f_{n-1} + f_n\varphi$ for each positive integer n.

Discussion of Exercise 4.9.6. **(a)** First, we observe that if u is a real number, then $\left\lfloor u + \dfrac{1}{2} \right\rfloor$ is the result of "rounding" u to the nearest integer (where a possible tie is resolved in favor of the larger integer[103]). Thus, we have the following:

Claim 1: Let $u \in \mathbb{R}$. Let $v \in \mathbb{Z}$ be such that $v - \dfrac{1}{2} \leq u < v + \dfrac{1}{2}$. Then, $\left\lfloor u + \dfrac{1}{2} \right\rfloor = v$.

[*Proof of Claim 1:* For the sake of completeness, let us give a formal proof of this. Let $x = u + \dfrac{1}{2}$. Adding $\dfrac{1}{2}$ to all sides of the inequality chain $v - \dfrac{1}{2} \leq u < v + \dfrac{1}{2}$, we obtain $v \leq u + \dfrac{1}{2} < v + \dfrac{1}{2} + \dfrac{1}{2}$. In other words, $v \leq x < v + 1$ (since $u + \dfrac{1}{2} = x$ and $v + \dfrac{1}{2} + \dfrac{1}{2} = v + 1$). Thus, the integer v is $\leq x$ (since $v \leq x$), but the next integer $v + 1$ is no longer $\leq x$ (since $x < v + 1$). Hence, the largest integer that is $\leq x$ is v. In other words, $\lfloor x \rfloor$ is v (since $\lfloor x \rfloor$ was defined as the largest integer that is $\leq x$). In other words, $\lfloor x \rfloor = v$. This rewrites as $\left\lfloor u + \dfrac{1}{2} \right\rfloor = v$ (since $x = u + \dfrac{1}{2}$). This proves Claim 1.]

Now, let us apply this. Set $\psi = \dfrac{1-\sqrt{5}}{2}$. Then, $\psi \approx -0.618$, so that $|\psi| \leq 1$.

Let $n \in \mathbb{N}$. Then, Theorem 2.3.1 yields $f_n = \dfrac{1}{\sqrt{5}}\varphi^n - \dfrac{1}{\sqrt{5}}\psi^n$, so that

$$\frac{1}{\sqrt{5}}\varphi^n = f_n + \frac{1}{\sqrt{5}}\psi^n. \tag{214}$$

However, $|\psi^n| = |\psi|^n \leq 1$ (since $|\psi| \leq 1$), so that $\left| \dfrac{1}{\sqrt{5}}\psi^n \right| = \dfrac{1}{\sqrt{5}} \underbrace{|\psi^n|}_{\leq 1} \leq \dfrac{1}{\sqrt{5}} < \dfrac{1}{2}$. In other words, the number $\dfrac{1}{\sqrt{5}}\psi^n$ lies in the open interval $\left(-\dfrac{1}{2}, \dfrac{1}{2}\right)$; in other words,

[103] For example, $\dfrac{1}{2} = 0.5$ is rounded to 1, not to 0.

$-\frac{1}{2} < \frac{1}{\sqrt{5}}\psi^n < \frac{1}{2}$. Hence, (214) entails

$$\frac{1}{\sqrt{5}}\varphi^n = f_n + \underbrace{\frac{1}{\sqrt{5}}\psi^n}_{<\frac{1}{2}} < f_n + \frac{1}{2}$$

and

$$\frac{1}{\sqrt{5}}\varphi^n = f_n + \underbrace{\frac{1}{\sqrt{5}}\psi^n}_{>-\frac{1}{2}} > f_n + \left(-\frac{1}{2}\right) = f_n - \frac{1}{2}.$$

Combining the latter two inequalities, we find $f_n - \frac{1}{2} < \frac{1}{\sqrt{5}}\varphi^n < f_n + \frac{1}{2}$. Hence, Claim 1 (applied to $u = \frac{1}{\sqrt{5}}\varphi^n$ and $v = f_n$) yields that $\left\lfloor \frac{1}{\sqrt{5}}\varphi^n + \frac{1}{2} \right\rfloor = f_n$ (since $f_n \in \mathbb{Z}$). This solves Exercise 4.9.6 **(a)**.

(b) We apply induction on n:

Induction base: We have $\varphi^1 = f_0 + f_1\varphi$ (since $\underbrace{f_0}_{=0} + \underbrace{f_1}_{=1}\varphi = 0 + 1\varphi = \varphi = \varphi^1$). In other words, Exercise 4.9.6 **(b)** holds for $n = 1$.

Induction step: Let m be a positive integer. Assume (as the induction hypothesis) that Exercise 4.9.6 **(b)** holds for $n = m$. We must prove that Exercise 4.9.6 **(b)** holds for $n = m + 1$. In other words, we must prove that $\varphi^{m+1} = f_m + f_{m+1}\varphi$.

We have $\varphi^2 = \varphi + 1$. (This is easily seen by calculation using the definition of φ.) Also, the recursive definition of the Fibonacci sequence yields $f_{m+1} = f_m + f_{m-1} = f_{m-1} + f_m$.

Our induction hypothesis says that Exercise 4.9.6 **(b)** holds for $n = m$. In other words, we have $\varphi^m = f_{m-1} + f_m\varphi$. Now,

$$\begin{aligned}\varphi^{m+1} &= \underbrace{\varphi^m}_{=f_{m-1}+f_m\varphi}\varphi = (f_{m-1} + f_m\varphi)\varphi = f_{m-1}\varphi + f_m\underbrace{\varphi^2}_{=\varphi+1} = f_{m-1}\varphi + f_m(\varphi+1) \\ &= f_{m-1}\varphi + f_m\varphi + f_m = \underbrace{(f_{m-1} + f_m)}_{=f_{m+1}}\varphi + f_m = f_{m+1}\varphi + f_m = f_m + f_{m+1}\varphi.\end{aligned}$$

In other words, Exercise 4.9.6 **(b)** holds for $n = m + 1$. This completes the induction step. Thus, Exercise 4.9.6 **(b)** is solved. □

4.9.6. Two-term recurrences: arithmetical properties

Generalizing is a muscle worth training, so let us generalize a few more properties of Fibonacci numbers to (a, b)-recurrent sequences. The next exercise generalizes Exercise 3.2.2 (note that we are using (u, v) instead of (a, b) in order to avoid assigning double duty to the letters a and b):

Exercise 4.9.7. Let u and v be two integers. Let $(x_0, x_1, x_2, \ldots)$ be an (u, v)-recurrent sequence of integers with $x_0 = 0$ and $x_1 = 1$.

Show that all $a, b \in \mathbb{N}$ satisfying $a \mid b$ satisfy $x_a \mid x_b$.

Solution to Exercise 4.9.7. We essentially repeat our above solution to Exercise 3.2.2, with the obvious changes made at every step where they are necessary.

Fix $a \in \mathbb{N}$. We must prove that for each $b \in \mathbb{N}$, the following statement holds:

$$\text{if } a \mid b, \text{ then } x_a \mid x_b. \tag{215}$$

We shall prove (215) by strong induction on b:

Induction step: Let $k \in \mathbb{N}$. Assume (as the induction hypothesis) that (215) holds for all $b < k$. We must prove that (215) holds for $b = k$. In other words, we must prove that

$$\text{if } a \mid k, \text{ then } x_a \mid x_k.$$

So let us assume that $a \mid k$. We must then prove that $x_a \mid x_k$.

If $k = 0$, then this is clearly true (since we have $x_k = x_0 = 0$ in this case, and since $x_a \mid 0$ is true[104]). Thus, for the rest of this proof, we WLOG assume that $k \neq 0$. Hence, $k \geq 1$ (since $k \in \mathbb{N}$).

It is fairly easy to see (from $a \mid k$ and $k \geq 1$ and $a \in \mathbb{N}$) that $a \geq 1$ and $k - a \in \mathbb{N}$ and $k - a < k$ and $a \mid k - a$ [105]. From $a \geq 1$, we obtain $a - 1 \in \mathbb{N}$.

Our induction hypothesis says that (215) holds for all $b < k$. In other words, for each $b \in \mathbb{N}$ satisfying $b < k$, the statement (215) holds. We can apply this to $b = k - a$ (since $k - a \in \mathbb{N}$ and $k - a < k$), and thus conclude that the statement (215) holds for $b = k - a$. In other words, if $a \mid k - a$, then $x_a \mid x_{k-a}$. Thus, $x_a \mid x_{k-a}$ (since we know that $a \mid k - a$). In other words, $x_{k-a} \equiv 0 \bmod x_a$. Note also that $x_a \equiv 0 \bmod x_a$ (since every integer n satisfies $n \equiv 0 \bmod n$).

But Exercise 4.9.3 (applied to u, v and x_i instead of a, b and y_i) shows that

$$x_{n+m+1} = v x_n x_m + x_{n+1} x_{m+1}$$

for any nonnegative integers n and m. We can apply this to $n = a - 1$ and $m = k - a$ (since $a - 1 \in \mathbb{N}$ and $k - a \in \mathbb{N}$); thus, we obtain

$$x_{(a-1)+(k-a)+1} = v x_{a-1} x_{k-a} + x_{(a-1)+1} x_{(k-a)+1}.$$

In view of $(a - 1) + (k - a) + 1 = k$ and $(a - 1) + 1 = a$, this rewrites as

$$x_k = v x_{a-1} x_{k-a} + x_a x_{(k-a)+1}.$$

Hence,

$$x_k = v x_{a-1} \underbrace{x_{k-a}}_{\equiv 0 \bmod x_a} + \underbrace{x_a}_{\equiv 0 \bmod x_a} x_{(k-a)+1} \equiv v x_{a-1} \cdot 0 + 0 \cdot x_{(k-a)+1} = 0 \bmod x_a,$$

[104] because any integer divides 0

[105] These statements can be proved in the same way as they were proved in the respective part of the solution to Exercise 3.2.2.

and thus $x_a \mid x_k$. This is precisely what we wanted to show. Thus, we have proved that (215) holds for $b = k$. This completes the induction step. Thus, Exercise 4.9.7 is solved. $\square$

Exercise 4.9.7 also appears in [Grinbe15, Theorem 2.26 **(c)**] (with different notations). Note that the $x_1 = 1$ condition is not actually necessary for the claim to hold (see Exercise 4.10.2 below); it was merely convenient for our proof. The $x_0 = 0$ condition, on the other hand, is "more or less necessary" (i.e., if $x_0 \neq 0$, then the claim of Exercise 4.9.7 holds only in a few exceptional cases[106]).

Next, let us generalize Exercise 3.5.2:

Exercise 4.9.8. Let u and v be two integers satisfying $u \perp v$. Let $(x_0, x_1, x_2, \ldots)$ be an (u, v)-recurrent sequence of integers with $x_1 = 1$.

Prove that $x_n \perp x_{n+1}$ for each $n \in \mathbb{N}$.

Discussion of Exercise 4.9.8. Let us try to imitate the solution to Exercise 3.5.2. Thus, we use induction on n.

Induction base: Exercise 3.5.1 **(a)** (applied to $a = x_0$) yields $1 \perp x_0$. According to Proposition 3.5.4, this entails $x_0 \perp 1$. In other words, $x_0 \perp x_1$ (since $x_1 = 1$). In other words, Exercise 4.9.8 holds for $n = 0$.

Induction step: Let $m \in \mathbb{N}$. Assume (as the induction hypothesis) that Exercise 4.9.8 holds for $n = m$. We must prove that Exercise 4.9.8 holds for $n = m + 1$. In other words, we must prove that $x_{m+1} \perp x_{m+2}$.

Our induction hypothesis says that Exercise 4.9.8 holds for $n = m$. In other words, we have $x_m \perp x_{m+1}$. According to Proposition 3.5.4 (applied to $a = x_m$ and $b = x_{m+1}$), this entails $x_{m+1} \perp x_m$. In other words, $\gcd(x_{m+1}, x_m) = 1$ (by the definition of "coprime"). But the sequence $(x_0, x_1, x_2, \ldots)$ is (u, v)-recurrent; thus, $x_{m+2} = ux_{m+1} + vx_m$. Hence,

$$\gcd(x_{m+1}, x_{m+2}) = \gcd(x_{m+1}, ux_{m+1} + vx_m) = \gcd(x_{m+1}, vx_m)$$
$$\text{(by Proposition 3.4.4 \textbf{(c)}, applied to } a = x_{m+1} \text{ and } b = vx_m).$$

It would be nice if we knew that $x_{m+1} \perp vx_m$, since that would yield $\gcd(x_{m+1}, vx_m) = 1$ and we could proceed. Unfortunately, we only know that $x_{m+1} \perp x_m$, which is weaker. What should we do?

Theorem 3.5.10 suggests a way we could try to prove $x_{m+1} \perp vx_m$. Namely, if we know that $v \perp x_{m+1}$ and $x_m \perp x_{m+1}$, then an application of Theorem 3.5.10 would yield $vx_m \perp x_{m+1}$, and therefore $x_{m+1} \perp vx_m$ (by Proposition 3.5.4). The two necessary ingredients for this argument to work are $v \perp x_{m+1}$ and $x_m \perp x_{m+1}$. We already know that $x_m \perp x_{m+1}$, but how do we get $v \perp x_{m+1}$?

The way out of this predicament turns out to be a highly important strategy that can be useful wherever induction is in play. Namely, we **insert** "$v \perp x_{m+1}$" into our induction hypothesis! This means that, instead of proving the claim "$x_n \perp x_{n+1}$",

[106] such as when $u + v = 1$ and $x_0 = x_1 = x_2 = x_3 = \cdots$

we prove the stronger claim "$x_n \perp x_{n+1}$ and $v \perp x_{n+1}$". This may look more difficult (after all, we now have to prove two statements rather than one), but in practice it may turn out to be more doable, as the extra strength of our claim makes the induction hypothesis stronger as well (after all, the induction hypothesis now also contains two statements rather than one). Thus, in the induction step, we have both more work and more to work with.

What happens if we insert $v \perp x_{n+1}$ into our claim? In the base case, we just have to show (additionally) that $v \perp x_1$, which is easy (since $x_1 = 1$). In the induction step, we have to show (additionally) that $v \perp x_{m+2}$, but we can use $v \perp x_{m+1}$ (which is now part of the induction hypothesis).

So how do we show that $v \perp x_{m+2}$ in the induction step? We compute

$$\begin{aligned}
\gcd(v, x_{m+2}) &= \gcd(v, x_m v + u x_{m+1}) \\
&\qquad (\text{since } x_{m+2} = u x_{m+1} + v x_m = x_m v + u x_{m+1}) \\
&= \gcd(v, u x_{m+1}) \\
&\qquad \left(\begin{array}{c} \text{by Proposition 3.4.4 } \textbf{(c)}\text{, applied to } v,\ u x_{m+1} \text{ and } x_m \\ \text{instead of } a,\ b \text{ and } u \end{array} \right) \\
&= 1,
\end{aligned}$$

where the last equality sign follows from noticing that $v \perp u x_{m+1}$ (which, in turn, follows from $v \perp u$ and $v \perp x_{m+1}$ using Theorem 3.5.10). So we get $v \perp x_{m+2}$, and we win: Our stronger claim "$x_n \perp x_{n+1}$ and $v \perp x_{n+1}$" has turned out to be easier to prove than the original claim "$x_n \perp x_{n+1}$". □

For the sake of clarity, let me distill the above discussion into a detailed solution to Exercise 4.9.8:

Solution to Exercise 4.9.8 (final copy). We claim that each $n \in \mathbb{N}$ satisfies

$$x_n \perp x_{n+1} \qquad \text{and} \qquad v \perp x_{n+1}. \tag{216}$$

[*Proof of (216):* We shall prove (216) by induction on n:

Induction base: Exercise 3.5.1 **(a)** (applied to $a = x_0$) yields $1 \perp x_0$. According to Proposition 3.5.4 (applied to $a = 1$ and $b = x_0$), this entails $x_0 \perp 1$. In other words, $x_0 \perp x_1$ (since $x_1 = 1$). The same argument (applied to v instead of x_0) yields $v \perp x_1$. We have now shown that $x_0 \perp x_1$ and $v \perp x_1$. In other words, (216) holds for $n = 0$.

Induction step: Let $m \in \mathbb{N}$. Assume (as the induction hypothesis) that (216) holds for $n = m$. We must prove that (216) holds for $n = m + 1$. In other words, we must prove that $x_{m+1} \perp x_{m+2}$ and $v \perp x_{m+2}$.

Our induction hypothesis says that (216) holds for $n = m$. In other words, we have $x_m \perp x_{m+1}$ and $v \perp x_{m+1}$. Proposition 3.5.4 (applied to $a = v$ and $b = x_{m+1}$) shows that $x_{m+1} \perp v$ (since $v \perp x_{m+1}$).

Theorem 3.5.10 (applied to $a = v$, $b = x_m$ and $c = x_{m+1}$) yields $v x_m \perp x_{m+1}$ (since $v \perp x_{m+1}$ and $x_m \perp x_{m+1}$). Hence, $x_{m+1} \perp v x_m$ (by Proposition 3.5.4, applied

to $a = vx_m$ and $b = x_{m+1}$). In other words, $\gcd(x_{m+1}, vx_m) = 1$ (by the definition of "coprime").

But the sequence $(x_0, x_1, x_2, \ldots)$ is (u, v)-recurrent; in other words, every $n \geq 2$ satisfies $x_n = ux_{n-1} + vx_{n-2}$ (by the definition of "(u, v)-recurrent"). Applying this to $n = m + 2$, we obtain $x_{m+2} = ux_{(m+2)-1} + vx_{(m+2)-2} = ux_{m+1} + vx_m$. Hence,

$$\begin{aligned}\gcd(x_{m+1}, x_{m+2}) &= \gcd(x_{m+1}, ux_{m+1} + vx_m) = \gcd(x_{m+1}, vx_m) \\ &\qquad \text{(by Proposition 3.4.4 \textbf{(c)}, applied to } a = x_{m+1} \text{ and } b = vx_m) \\ &= 1.\end{aligned}$$

In other words, $x_{m+1} \perp x_{m+2}$.

Furthermore, we have $u \perp v$ and $x_{m+1} \perp v$. Hence, Theorem 3.5.10 (applied to $a = u$, $b = x_{m+1}$ and $c = v$) yields $ux_{m+1} \perp v$. Hence, $v \perp ux_{m+1}$ (by Proposition 3.5.4, applied to $a = ux_{m+1}$ and $b = v$). In other words, $\gcd(v, ux_{m+1}) = 1$.

Now, $x_{m+2} = ux_{m+1} + vx_m = x_m v + ux_{m+1}$, so that

$$\begin{aligned}\gcd(v, x_{m+2}) &= \gcd(v, x_m v + ux_{m+1}) = \gcd(v, ux_{m+1}) \\ &\qquad \left(\begin{array}{c}\text{by Proposition 3.4.4 \textbf{(c)}, applied to } v,\ ux_{m+1} \text{ and } x_m \\ \text{instead of } a,\ b \text{ and } u\end{array}\right) \\ &= 1.\end{aligned}$$

In other words, $v \perp x_{m+2}$.

Now, we have shown that $x_{m+1} \perp x_{m+2}$ and $v \perp x_{m+2}$. In other words, (216) holds for $n = m + 1$. This completes the induction step; thus, (216) is proved.]

The claim of Exercise 4.9.8 is part of (216). Thus, Exercise 4.9.8 is solved (since (216) is proved). □

The technique we used to solve Exercise 4.9.8 above is known as *strengthening the induction hypothesis*. The underlying phenomenon is that induction proofs often become simpler (or even doable in the first place) when the claim being proved is strengthened (by adding additional statements to it, or in another way). For a more classical example, try proving the inequality

$$\frac{1}{1^2} + \frac{1}{2^2} + \cdots + \frac{1}{n^2} < 2 \qquad \text{for all } n \in \mathbb{N}. \tag{217}$$

It is unclear how to prove this by induction on n, since the space between the left and the right hand sides keeps shrinking as n grows. However, the **stronger** inequality

$$\frac{1}{1^2} + \frac{1}{2^2} + \cdots + \frac{1}{n^2} \leq 2 - \frac{1}{n} \qquad \text{for all } n \in \mathbb{N} \tag{218}$$

can easily be proved by induction on n. (This example has been taken from `https://mathoverflow.net/a/40688` , where other instances of this phenomenon can also be found.)

I occasionally think of induction proofs as long travels. If the distance is long enough, one has to take provisions along, even though they encumber the backpack. Extra statements in the claim (like the $v \perp x_{n+1}$ part of (216) in the solution above, or like the $\frac{2}{n}$ in (218)) are like such provisions. The art is to find the "right" provisions, whose usefulness on the journey makes up for their weight. Incidentally, when strengthening the induction hypothesis, we also gain a stronger result in the end; unlike most kinds of provisions, they don't get consumed during the journey.

An astute reader will have noticed that we have already done something similar in our solution to Exercise 4.6.4: We proved the two equalities (150) and (151) together, rather than proving each of them separately. Each of them assisted the other in the induction step. Thus, the best way to prove (150) by induction (actually the only way I know) is to combine it with the extra claim (151).

4.9.7. Two-term recurrences: odds and ends

Let us give a few more remarks about (a, b)-recurrent sequences before moving on.

The vector space of (a, b)-recurrent sequences. First, while we have used matrices to study (a, b)-recurrent sequences in Subsection 4.9.4, there is yet another way in which linear algebra shines light on (a, b)-recurrent sequences. Namely, for any two fixed numbers a and b, the (a, b)-recurrent sequences form a 2-dimensional vector subspace of the vector space of all sequences. More precisely:

Proposition 4.9.22. Let $\mathbb{K}$ be one of the fields $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$. Fix two numbers $a, b \in \mathbb{K}$. Let $\operatorname{Rec}_{a,b}$ denote the set of all (a, b)-recurrent sequences $(x_0, x_1, x_2, \ldots) \in \mathbb{K}^\infty$. Consider $\mathbb{K}^\infty$ (that is, the set of all infinite sequences of elements of $\mathbb{K}$) as an (infinite-dimensional) $\mathbb{K}$-vector space, with addition and scaling defined entrywise[107]. Then, $\operatorname{Rec}_{a,b}$ is a 2-dimensional vector subspace of $\mathbb{K}^\infty$.

Proof of Proposition 4.9.22 (sketched). Since Proposition 4.9.22 is nowhere near essential to what we are planning to do, we shall be on the terse side. First of all, we claim that $\operatorname{Rec}_{a,b}$ is a vector subspace of $\mathbb{K}^\infty$. In order to prove this, we need to show the following three facts:

[107] This means that addition is defined by

$$(x_0, x_1, x_2, \ldots) + (y_0, y_1, y_2, \ldots) = (x_0 + y_0, x_1 + y_1, x_2 + y_2, \ldots) \quad \text{for all } (x_0, x_1, x_2, \ldots), (y_0, y_1, y_2, \ldots) \in \mathbb{K}^\infty,$$

and that scaling is defined by

$$\lambda (x_0, x_1, x_2, \ldots) = (\lambda x_0, \lambda x_1, \lambda x_2, \ldots) \qquad \text{for all } (x_0, x_1, x_2, \ldots) \in \mathbb{K}^\infty.$$

(In other words, sequences are considered as row vectors of infinite size.) The zero vector of this vector space is the sequence $(0, 0, 0, \ldots)$.

1. We have $(0,0,0,\ldots) \in \operatorname{Rec}_{a,b}$.
2. If $(x_0,x_1,x_2,\ldots) \in \operatorname{Rec}_{a,b}$ and $(y_0,y_1,y_2,\ldots) \in \operatorname{Rec}_{a,b}$, then $(x_0,x_1,x_2,\ldots) + (y_0,y_1,y_2,\ldots) \in \operatorname{Rec}_{a,b}$.
3. If $(x_0,x_1,x_2,\ldots) \in \operatorname{Rec}_{a,b}$ and $\lambda \in \mathbb{K}$, then $\lambda\,(x_0,x_1,x_2,\ldots) \in \operatorname{Rec}_{a,b}$.

All three of these facts are straightforward to verify[108]. Thus, we have shown that $\operatorname{Rec}_{a,b}$ is a vector subspace of $\mathbb{K}^\infty$. It remains to prove that $\operatorname{Rec}_{a,b}$ is 2-dimensional. We shall achieve this by constructing a basis (s,t) of $\operatorname{Rec}_{a,b}$.

Indeed, we define a sequence $s = (s_0,s_1,s_2,\ldots) \in \mathbb{K}^\infty$ recursively by setting $s_0 = 0$, $s_1 = 1$ and $s_n = as_{n-1} + bs_{n-2}$ for each $n \geq 2$. Thus, the sequence s is (a,b)-recurrent.

Furthermore, we define a sequence $t = (t_0,t_1,t_2,\ldots) \in \mathbb{K}^\infty$ recursively by setting $t_0 = 1$, $t_1 = 0$ and $t_n = at_{n-1} + bt_{n-2}$ for each $n \geq 2$. Thus, the sequence t is (a,b)-recurrent.

The two sequences s and t are (a,b)-recurrent, thus belong to $\operatorname{Rec}_{a,b}$. Furthermore, any $\mathbb{K}$-linear combination $\lambda s + \mu t$ (with $\lambda,\mu \in \mathbb{K}$) starts with the two entries $\lambda \underbrace{s_0}_{=0} + \mu \underbrace{t_0}_{=1} = \lambda \cdot 0 + \mu \cdot 1 = \mu$ and $\lambda \underbrace{s_1}_{=1} + \mu \underbrace{t_1}_{=0} = \lambda \cdot 1 + \mu \cdot 0 = \lambda$, and thus cannot equal the zero sequence $(0,0,0,\ldots)$ unless both λ and μ are 0. In other words, the sequences s and t are $\mathbb{K}$-linearly independent.

Moreover, we claim that s and t span $\operatorname{Rec}_{a,b}$. Indeed, every sequence $x = (x_0,x_1,x_2,\ldots) \in \operatorname{Rec}_{a,b}$ can be written as a $\mathbb{K}$-linear combination of s and t as follows:

$$x = x_1 s + x_0 t. \tag{219}$$

[*Proof of (219):* Let $x = (x_0,x_1,x_2,\ldots) \in \operatorname{Rec}_{a,b}$. We must prove (219). Since both s and t belong to $\operatorname{Rec}_{a,b}$, the linear combination $x_1 s + x_0 t$ also belongs to $\operatorname{Rec}_{a,b}$ (since $\operatorname{Rec}_{a,b}$ is a vector subspace of $\mathbb{K}^\infty$). In other words, $x_1 s + x_0 t$ is an (a,b)-recurrent sequence.

[108] Just in case, let me show the proof of fact 2:

Let $(x_0,x_1,x_2,\ldots) \in \operatorname{Rec}_{a,b}$ and $(y_0,y_1,y_2,\ldots) \in \operatorname{Rec}_{a,b}$. We must show that $(x_0,x_1,x_2,\ldots) + (y_0,y_1,y_2,\ldots) \in \operatorname{Rec}_{a,b}$.

We have $(x_0,x_1,x_2,\ldots) \in \operatorname{Rec}_{a,b}$. In other words, the sequence $(x_0,x_1,x_2,\ldots)$ is (a,b)-recurrent (since $\operatorname{Rec}_{a,b}$ is the set of all (a,b)-recurrent sequences). In other words, every $n \geq 2$ satisfies $x_n = ax_{n-1} + bx_{n-2}$. Likewise, every $n \geq 2$ satisfies $y_n = ay_{n-1} + by_{n-2}$. Hence, every $n \geq 2$ satisfies

$$\begin{aligned} \underbrace{x_n}_{=ax_{n-1}+bx_{n-2}} + \underbrace{y_n}_{=ay_{n-1}+by_{n-2}} &= ax_{n-1} + bx_{n-2} + ay_{n-1} + by_{n-2} \\ &= a\left(x_{n-1} + y_{n-1}\right) + b\left(x_{n-2} + y_{n-2}\right). \end{aligned}$$

In other words, the sequence $(x_0+y_0, x_1+y_1, x_2+y_2, \ldots)$ is (a,b)-recurrent (by the definition of "(a,b)-recurrent"). In other words, the sequence $(x_0,x_1,x_2,\ldots) + (y_0,y_1,y_2,\ldots)$ is (a,b)-recurrent (since $(x_0,x_1,x_2,\ldots) + (y_0,y_1,y_2,\ldots)$ was defined to be $(x_0+y_0, x_1+y_1, x_2+y_2, \ldots)$). In other words, $(x_0,x_1,x_2,\ldots) + (y_0,y_1,y_2,\ldots) \in \operatorname{Rec}_{a,b}$ (since $\operatorname{Rec}_{a,b}$ is the set of all (a,b)-recurrent sequences). This proves fact 2. The proofs of facts 1 and 3 are similar.

Now, recall that an (a,b)-recurrent sequence is uniquely determined by its first two entries (since its recursive equation allows all other entries to be computed in terms of these first two). Hence, if two (a,b)-recurrent sequences agree in their first two entries, then these two sequences must be identical. Thus, in order to prove that $x = x_1 s + x_0 t$, it suffices to show that the sequences x and $x_1 s + x_0 t$ agree in their first two entries (since both sequences x and $x_1 s + x_0 t$ are (a,b)-recurrent). But this is easy: The 0-th entry of $x_1 s + x_0 t$ is $x_1 \underbrace{s_0}_{=0} + x_0 \underbrace{t_0}_{=1} = x_1 \cdot 0 + x_0 \cdot 1 = x_0$, which is precisely the 0-th entry of x. The 1-st entry of $x_1 s + x_0 t$ is $x_1 \underbrace{s_1}_{=1} + x_0 \underbrace{t_1}_{=0} = x_1 \cdot 1 + x_0 \cdot 0 = x_1$, which is precisely the 1-st entry of x. Thus, the two sequences x and $x_1 s + x_0 t$ agree in their first two entries, and hence are identical (as we just explained). In other words, $x = x_1 s + x_0 t$. This proves (219).]

Now we have shown that the two elements s and t of $\operatorname{Rec}_{a,b}$ span $\operatorname{Rec}_{a,b}$ and are $\mathbb{K}$-linearly independent. In other words, (s,t) is a basis of the $\mathbb{K}$-vector space $\operatorname{Rec}_{a,b}$. Hence, the $\mathbb{K}$-vector space $\operatorname{Rec}_{a,b}$ has a basis consisting of 2 vectors, and therefore is 2-dimensional. Proposition 4.9.22 is now proved. □

The Chebyshev polynomials. We have so far been studying (a,b)-recurrent sequences of numbers; but we could apply (most of) the same reasoning to (a,b)-recurrent sequences of polynomials or other objects that can be added and multiplied. (In particular, a and b can themselves be polynomials.) The most important example of such sequences is a sequence of polynomials known as the *Chebyshev polynomials of the second kind*:

Example 4.9.23. Consider polynomials in a single variable x with integer coefficients. Define a sequence $(T_0, T_1, T_2, \ldots)$ of such polynomials recursively by setting

$$\begin{aligned} &T_0(x) = 1, \qquad T_1(x) = x, \qquad \text{and} \\ &T_n(x) = 2xT_{n-1}(x) - T_{n-2}(x) \qquad \text{for each } n \geq 2. \end{aligned}$$

The polynomials $T_0, T_1, T_2, \ldots$ in this sequence are called *Chebyshev polynomials of the first kind*. They have several properties (useful in approximation theory as well as in number theory); the most well-known one is the fact that

$$\cos(n\alpha) = T_n(\cos\alpha) \qquad \text{for any angle } \alpha. \tag{220}$$

(Thus, the polynomial T_n is the answer to the rather natural question "how can we express $\cos(n\alpha)$ through $\cos\alpha$ without using arc-cosines?". It is easy to prove (220) by induction on n.)

The definition of the Chebyshev polynomials T_n shows that the sequence $(T_0, T_1, T_2, \ldots)$ is $(2x, -1)$-recurrent (where we have extended the concept of "(a,b)-recurrent" to sequences of polynomials in the obvious way). We can now apply most of what we know about (a,b)-recurrent sequences to this sequence

$(T_0, T_1, T_2, \ldots)$. For example, Exercise 4.9.4 (applied to $a = 2x$ and $b = -1$ and $x_i = T_i$ and $y_i = T_i$) yields that

$$(-1)\, T_0(x) \cdot T_{n+m}(x) + T_1(x) \cdot T_{n+m+1}(x) = (-1)\, T_n(x) \cdot T_m(x) + T_{n+1}(x) \cdot T_{m+1}(x)$$

for any nonnegative integers n and m. In view of $T_0(x) = 1$ and $T_1(x) = x$, this rewrites as

$$-T_{n+m}(x) + xT_{n+m+1}(x) = -T_n(x) \cdot T_m(x) + T_{n+1}(x) \cdot T_{m+1}(x).$$

More advanced properties of (a, b)-recurrent sequences lead to more advanced properties of the polynomials T_n.

The *Chebyshev polynomials of the second kind* form another $(2x, -1)$-recurrent sequence $(U_0, U_1, U_2, \ldots)$, which starts with $U_0(x) = 1$ and $U_1(x) = 2x$.

We refer to the Wikipedia page for an overview of the major properties of Chebyshev polynomials, to [ChaSed97, Chapter 22] for an introduction, and to the book [Rivlin90] for an in-depth treatment.

4.9.8. k-term recurrences

In the definition of an (a, b)-recurrent sequence, the equality (170) expresses an entry x_n of the sequence in terms of the preceding two entries x_{n-1} and x_{n-2}. Nothing speaks against generalizing this definition to more than two parameters and, correspondingly, more than two preceding entries:

Definition 4.9.24. Let $k \in \mathbb{N}$. Let $a_1, a_2, \ldots, a_k$ be any k numbers. A sequence $(x_0, x_1, x_2, \ldots)$ of numbers will be called $(a_1, a_2, \ldots, a_k)$*-recurrent* if every $n \geq k$ satisfies

$$x_n = a_1 x_{n-1} + a_2 x_{n-2} + \cdots + a_k x_{n-k}. \tag{221}$$

Clearly, Definition 4.9.1 is a particular case of Definition 4.9.24. Here are some other particular cases:

- Here is the case $k = 1$ of Definition 4.9.24: If a is a single number, then an (a)-recurrent sequence is the same as a geometric progression with ratio a.
- The *Padovan sequence* is the sequence $(p_0, p_1, p_2, \ldots)$ of integers defined recursively by setting

  $$\begin{aligned} &p_0 = 1, \qquad p_1 = 1, \qquad p_2 = 1, \qquad \text{and} \\ &p_n = p_{n-2} + p_{n-3} \qquad \text{for each } n \geq 3. \end{aligned}$$

  This sequence is obviously $(0, 1, 1)$-recurrent. The *Perrin sequence* is also $(0, 1, 1)$-recurrent but has starting values 3, 0 and 2 (in this order). For the properties of these two sequences, we refer to their Wikipedia articles.

- The sequence $\left(0^1, 1^2, 2^2, 3^2, \ldots\right)$ is $(3, -3, 1)$-recurrent, since a straightforward computation shows that

$$n^2 = 3\left(n-1\right)^2 - 3\left(n-2\right)^2 + \left(n-3\right)^2 \qquad \text{for each } n \geq 3.$$

  We will generalize this in an exercise below (Exercise 5.4.2 **(e)**).

- The sequence $(t_0, t_1, t_2, \ldots)$ in Exercise 1.1.2 is $(0, 4, 0, -1)$-recurrent. Indeed, this is precisely the claim of Exercise 1.1.2 **(a)**.

- If $(x_0, x_1, x_2, \ldots)$ is an (a, b)-recurrent sequence (for some numbers a and b), then $\left(x_0^2, x_1^2, x_2^2, \ldots\right)$ is an $\left(a^2 + b, b\left(a^2 + b\right), -b^3\right)$-recurrent sequence. This is not hard to check by direct computation[109]. Similar results hold for $(a_1, a_2, \ldots, a_k)$-recurrent sequences, but they get messier as k grows, and require linear algebra to properly state and prove.

- If $(x_0, x_1, x_2, \ldots)$ is an (a, b)-recurrent sequence (for some numbers a and b), and if $(y_0, y_1, y_2, \ldots)$ is a (c, d)-recurrent sequence (for some numbers c and d), then $(x_0 + y_0, x_1 + y_1, x_2 + y_2, \ldots)$ is an $(a + c, b + d - ac, -ad - bc, -bd)$-recurrent sequence, whereas $(x_0 y_0, x_1 y_1, x_2 y_2, \ldots)$ is an $\left(ac, a^2 d + c^2 b + 2bd, abcd, -b^2 d^2\right)$-recurrent sequence. Both of these claims can be checked by (laborious yet fairly straightforward) computation; however, lurking behind them are certain properties of polynomials and matrices that are best understood from a linear-algebraic viewpoint. (This would also allow us to generalize them to $(a_1, a_2, \ldots, a_k)$-recurrent sequences.)

Theorem 4.9.11 can be generalized to $(a_1, a_2, \ldots, a_k)$-recurrent sequences, although it is a judgment call whether the result counts as an "explicit formula":

[109] Indeed, you need to check that $x_n^2 = \left(a^2 + b\right) x_{n-1}^2 + \left(a^2 + b\right) b x_{n-2}^2 - b^3 x_{n-3}^2$ for each $n \geq 3$. But this can be done by expressing x_n and x_{n-1} through x_{n-2} and x_{n-3} (via $x_{n-1} = a x_{n-2} + b x_{n-3}$ and $x_n = a \underbrace{x_{n-1}}_{= a x_{n-2} + b x_{n-3}} + b x_{n-2} = a\left(a x_{n-2} + b x_{n-3}\right) + b x_{n-2}$) and expanding both sides.

Theorem 4.9.25. Let $k \in \mathbb{N}$. Let $a_1, a_2, \ldots, a_k$ be any k numbers. Let $p(X)$ be the polynomial $X^k - a_1 X^{k-1} - a_2 X^{k-2} - \cdots - a_k X^{k-k}$. Let $\lambda_1, \lambda_2, \ldots, \lambda_h$ be all the distinct complex roots of $p(X)$. For each $i \in \{1, 2, \ldots, h\}$, let m_i be the multiplicity of the root λ_i of $p(X)$.

Let $(x_0, x_1, x_2, \ldots)$ be an $(a_1, a_2, \ldots, a_k)$-recurrent sequence. Then, there exist constants $\gamma_{i,j}$ for all $i \in \{1, 2, \ldots, h\}$ and $j \in \{0, 1, \ldots, m_i - 1\}$ with the property that each $n \in \mathbb{N}$ satisfies

$$x_n = \sum_{i=1}^{h} \sum_{j=0}^{m_i - 1} \gamma_{i,j} n^j \lambda_i^n. \tag{222}$$

In other words, there exist polynomials $\delta_i(X)$ of degree $< m_i$ for all $i \in \{1, 2, \ldots, h\}$ with the property that each $n \in \mathbb{N}$ satisfies

$$x_n = \sum_{i=1}^{h} \delta_i(n) \lambda_i^n. \tag{223}$$

(Here, we agree that the zero polynomial has degree $< m_i$, whatever m_i is.)

Thus, in particular, if all roots of $p(X)$ are distinct (i.e., we have $h = k$, and the multiplicities m_i are all equal to 1), then each $n \in \mathbb{N}$ satisfies

$$x_n = \gamma_{1,0} \lambda_1^n + \gamma_{2,0} \lambda_2^n + \cdots + \gamma_{k,0} \lambda_k^n. \tag{224}$$

We shall not dwell on the proof of Theorem 4.9.25, nor on its uses. Proofs of Theorem 4.9.25 can be found in [Melian01, Theorem 1] and in [Ivanov08, Theorem 2]. The proof in [Melian01, Theorem 1] is essentially a straightforward generalization of our above second proof of Theorem 4.9.11 in Subsection 4.9.4; it relies on the $k \times k$-matrix

$$\begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ a_k & a_{k-1} & a_{k-2} & \cdots & a_1 \end{pmatrix} \tag{225}$$

(a generalization of the matrix A from Proposition 4.9.21) and its Jordan normal form[110]. The generalization of Proposition 4.9.21 to $(a_1, a_2, \ldots, a_k)$-recurrent sequences is easily stated and proved, and this generalization provides a way to study $(a_1, a_2, \ldots, a_k)$-recurrent sequences via matrix algebra. Unlike the particular case $k = 2$ (in which case explicit computations are a viable method), it is this matrix approach that bears the most fruits in the general case. For example, here is a neat folklore result whose only proof I know uses the matrix approach:

[110] To be more precise, the matrix used in [Melian01, Theorem 1] is obtained from the above matrix by a 180°-rotation; but this is just a matter of relabeling rows and columns.

Proposition 4.9.26. Let $d \in \mathbb{N}$ and $k \in \mathbb{N}$. Let $a_1, a_2, \ldots, a_k$ be any k numbers. Then, there exist k numbers $b_1, b_2, \ldots, b_k$ with the following property: If $(x_0, x_1, x_2, \ldots)$ is any $(a_1, a_2, \ldots, a_k)$-recurrent sequence, then the sequence $\left(x_{p+0d}, x_{p+1d}, x_{p+2d}, \ldots\right)$ is $(b_1, b_2, \ldots, b_k)$-recurrent for each $p \in \mathbb{N}$.

Note that these k numbers $b_1, b_2, \ldots, b_k$ depend neither on the sequence $(x_0, x_1, x_2, \ldots)$ nor on the integer p; they are determined by d, k and $a_1, a_2, \ldots, a_k$ alone. For example, for $k = 2$ and $d = 2$, this is saying that if $(x_0, x_1, x_2, \ldots)$ is any (a_1, a_2)-recurrent sequence, then both sequences $(x_0, x_2, x_4, \ldots)$ and $(x_1, x_3, x_5, \ldots)$ are (b_1, b_2)-recurrent. In the particular case $k = 2$, Proposition 4.9.26 can be proved by elementary means (see [Grinbe15, solution to Exercise 4.2 **(c)**]); but in the general case, the only proof I know uses the matrix (225) and the Cayley–Hamilton theorem. (See [Grinbe19b, Theorem 1] for this proof.)

4.10. Homework set #4: More sequences

This is a **regular** problem set. See Section 3.7 for details on grading.

This homework set covers the above parts of Chapter 4. Some of the problems may be unrelated.

Please solve **at most 5** problems. (No points will be given for further solutions.)

Exercise 4.10.1. Let u be a positive integer. Let $k \in \mathbb{N}$. Define a sequence $(a_0, a_1, a_2, \ldots)$ of integers by setting

$$a_n = \binom{n}{k} \% u \qquad \text{for each } n \in \mathbb{N}.$$

Show that this sequence $(a_0, a_1, a_2, \ldots)$ is $uk!$-periodic.

The next exercise generalizes Exercise 4.9.7:

Exercise 4.10.2. Let u and v be two integers. Let $(x_0, x_1, x_2, \ldots)$ be a (u, v)-recurrent sequence of integers with $x_0 = 0$.

Show that all $a, b \in \mathbb{N}$ satisfying $a \mid b$ satisfy $x_a \mid x_b$.

Exercise 4.10.3. Generalize Exercise 4.9.2 further, to a claim about two (a, b)-recurrent sequences $(x_0, x_1, x_2, \ldots)$ and $(y_0, y_1, y_2, \ldots)$.

[**Hint:** The left hand side will be $x_{n+1}y_{n-1} - x_n y_n$.]

Exercise 4.10.4. Let a be any number. Let $(x_0, x_1, x_2, \ldots)$ and $(y_0, y_1, y_2, \ldots)$ be two $(a, 1)$-recurrent sequences of numbers with $x_0 = 0$. (We don't require anything of y_0.) Let $n, m \in \mathbb{N}$ satisfy $n \geq m$. Prove that

$$x_{n-m}y_{n+m} = x_n y_n - (-1)^{n+m} x_m y_m.$$

The next exercise is about generalizing Exercise 3.7.2 and Exercise 3.4.1 **(b)**:

Exercise 4.10.5. Let u and v be two integers such that $u \perp v$. Let $(x_0, x_1, x_2, \ldots)$ be a (u, v)-recurrent sequence of integers with $x_0 = 0$ and $x_1 = 1$. Show that all $a, b \in \mathbb{N}$ satisfy $\gcd(x_a, x_b) = \left| x_{\gcd(a,b)} \right|$.

Exercise 4.10.6. Let $n \in \mathbb{N}$. Prove that there exists some $m \in \mathbb{N}$ such that $\left(\sqrt{2} - 1\right)^n = \sqrt{m+1} - \sqrt{m}$.

Exercise 4.10.7. A sequence $(a_0, a_1, a_2, \ldots)$ of numbers is defined recursively by $a_0 = -1$ and $a_1 = 0$ and $a_n = a_{n-1}^2 - n^2 a_{n-2} - 1$ for all $n \geq 2$. Find a_{100}.

Exercise 4.10.8. Let $n \in \mathbb{N}$. Prove that $\left\lfloor \left(1 + \sqrt{2}\right)^n \right\rfloor$ is even if and only if n is odd.

Exercise 4.10.9. Let m and k be two positive integers such that $m \mid k + 1$. Define a sequence $(a_0, a_1, a_2, \ldots)$ of positive integers recursively by

$$a_0 = 1, \qquad a_1 = 1, \qquad a_2 = m \qquad \text{and}$$
$$a_n = \frac{k + a_{n-1} a_{n-2}}{a_{n-3}} \qquad \text{for each } n \geq 3.$$

Prove that a_n is an integer for each $n \in \mathbb{N}$.

[**Hint:** Prove that each of the two subsequences $(a_0, a_2, a_4, a_6, \ldots)$ and $(a_1, a_3, a_5, a_7, \ldots)$ is (a, b)-recurrent for some integers a and b.]

Exercise 4.10.10. Let $(a, a+d, a+2d, a+3d, \ldots)$ be any (infinite) arithmetic progression with $d \neq 0$. Prove that this arithmetic progression contains an infinite geometric progression as a subsequence (i.e., there is an infinite strictly increasing sequence $(i_0, i_1, i_2, \ldots)$ of nonnegative integers such that $(a + i_0 d, a + i_1 d, a + i_2 d, \ldots)$ is a geometric progression) if and only if $\dfrac{a}{d} \in \mathbb{Q}$.

4.11. More integer sequences

Exercise 1.1.2 **(b)**, Exercise 4.6.4 and Exercise 4.10.9 are three instances of a common type of problem: Given a sequence of numbers defined recursively, to prove that all entries of the sequence are integers. (Usually this is not obvious from the definition, since the definition involves division or even taking roots.) This type of problem has recently become popular; famous examples are the *Somos sequences* and various variants thereof. See [Gale98, Chapters 1 and 4] and [Grinbe15, §2.9] for a few examples of such problems. There is also a theory that unifies some (but

not all) instances of this "unexpected integrality" phenomenon, but properly understanding this theory goes beyond these notes. (See [AlCuHu16] for a large part of this theory. For more, google "cluster algebras" and "Laurent phenomenon".)

In this short section, we shall see more examples of such problems.

4.11.1. Propp's $t_n t_{n-k} = 1 + t_{n-1} t_{n-2} \ldots t_{n-k+1}$ recurrence

Let us generalize the sequence $(t_0, t_1, t_2, \ldots)$ from Exercise 1.1.2:

Exercise 4.11.1. Fix a positive integer $k \geq 2$. Define a sequence $(t_0, t_1, t_2, \ldots)$ of positive rational numbers recursively by setting

$$t_n = 1 \qquad \text{for each } n < k \tag{226}$$

and

$$t_n = \frac{1 + t_{n-1} t_{n-2} \cdots t_{n-k+1}}{t_{n-k}} \qquad \text{for each } n \geq k. \tag{227}$$

(For example, $t_k = \dfrac{1 + t_{k-1} t_{k-2} \cdots t_1}{t_0} = \dfrac{1 + 1 \cdot 1 \cdot \cdots \cdot 1}{1} = 2$ and likewise $t_{k+1} = 3$.)

Prove that t_n is a positive integer for each integer $n \geq 0$.

Let us see what this means for small values of k:

Example 4.11.1. Set $k = 3$ in Exercise 4.11.1. Then, $t_0 = t_1 = t_2 = 1$ and

$$t_n = \frac{1 + t_{n-1} t_{n-2}}{t_{n-3}} \qquad \text{for each } n \geq 3$$

(since the right hand side of (227) becomes $\dfrac{1 + t_{n-1} t_{n-2}}{t_{n-3}}$). Hence, the sequence $(t_0, t_1, t_2, \ldots)$ is precisely the sequence $(t_0, t_1, t_2, \ldots)$ from Exercise 1.1.2. Thus, the claim of Exercise 4.11.1 in the case $k = 3$ is precisely the claim of Exercise 1.1.2 **(b)**.

Let us make a table of the first few entries of the sequence $(t_0, t_1, t_2, \ldots)$ (for $k = 3$):

n	0	1	2	3	4	5	6	7	8	9	10	11	12	13
t_n	1	1	1	2	3	7	11	26	41	97	153	362	571	1351

.

This sequence is sequence A005246 in the OEIS, and is $(0, 4, 0, -1)$-recurrent (as we know from Exercise 1.1.2 **(a)**).

Example 4.11.2. Set $k = 2$ in Exercise 4.11.1. Then, $t_0 = t_1 = 1$ and

$$t_n = \frac{1 + t_{n-1}}{t_{n-2}} \qquad \text{for each } n \geq 2. \tag{228}$$

Let us make a table of the first few entries of the sequence $(t_0, t_1, t_2, \ldots)$:

n	0	1	2	3	4	5	6	7	8	9	10	11	12
t_n	1	1	2	3	2	1	1	2	3	2	1	1	2

.

This reveals that the sequence $(t_0, t_1, t_2, \ldots)$ is 5-periodic (with values $1, 1, 2, 3, 2$ repeating over and over). This can be proved by a straightforward computation (just use (228) four times to show that $t_n = t_{n+5}$). Despite its trivial nature, this sequence appears in the OEIS (as sequence A076839).

Example 4.11.3. Set $k = 4$ in Exercise 4.11.1. Then, $t_0 = t_1 = t_2 = t_3 = 1$ and

$$t_n = \frac{1 + t_{n-1}t_{n-2}t_{n-3}}{t_{n-4}} \qquad \text{for each } n \geq 4.$$

Let us make a table of the first few entries of the sequence $(t_0, t_1, t_2, \ldots)$:

n	0	1	2	3	4	5	6	7	8	9	10	11
t_n	1	1	1	1	2	3	7	43	452	45351	125920291	60027819184831

.

The rapid growth of the t_n (after a slow start for $n \leq 7$) strikes the eye. Actually, the sequence grows at least doubly exponentially[111], whence it is not $(a_1, a_2, \ldots, a_m)$-recurrent for any positive integer m and any numbers $a_1, a_2, \ldots, a_m$ (because Theorem 4.9.25 shows that any $(a_1, a_2, \ldots, a_m)$-recurrent sequence grows at most exponentially). The sequence $(t_0, t_1, t_2, \ldots)$ is A051786 in the OEIS.

Rather than solving Exercise 4.11.1 directly, let us generalize it even further:

[111] To make this more precise: It is not hard to prove that $t_n \geq 2^{f_{n-3}}$ for each $n \geq 3$ (where $(f_0, f_1, f_2, \ldots)$ is the Fibonacci sequence). Since the Fibonacci sequence grows exponentially, we thus conclude that t_n grows at least doubly exponentially.

The inequality $t_n \geq 2^{f_{n-3}}$ can be proved by induction on n. Here is an outline: The first step is to show (by strong induction) that $t_n \geq t_{n-1}$ for each positive integer n. The next step is to show that $t_n \geq t_{n-1}t_{n-2}$ for each integer $n \geq 2$. Once this is done, $t_n \geq 2^{f_{n-3}}$ can be proved by strong induction (the induction step is arguing that $t_m \geq \underbrace{t_{m-1}}_{\geq 2^{f_{m-4}}} \underbrace{t_{m-2}}_{\geq 2^{f_{m-5}}} \geq 2^{f_{m-4}} 2^{f_{m-5}} = 2^{f_{m-4}+f_{m-5}} = 2^{f_{m-3}}$).

Exercise 4.11.2. Fix a positive integer $k \geq 2$. Fix $k-1$ positive integers $p_1, p_2, \ldots, p_{k-1}$. Define a sequence $(t_0, t_1, t_2, \ldots)$ of positive rational numbers recursively by setting

$$t_n = 1 \qquad \text{for each } n < k \tag{229}$$

and

$$t_n = \frac{1 + t_{n-1}^{p_1} t_{n-2}^{p_2} \cdots t_{n-k+1}^{p_{k-1}}}{t_{n-k}} \qquad \text{for each } n \geq k. \tag{230}$$

(For example, $t_k = \dfrac{1 + t_{k-1}^{p_1} t_{k-2}^{p_2} \cdots t_1^{p_{k-1}}}{t_0} = \dfrac{1 + 1^{p_1} 1^{p_2} \cdots 1^{p_{k-1}}}{1} = 2$ and likewise $t_{k+1} = 1 + 2^{p_1}$.)

Prove that t_n is a positive integer for each integer $n \geq 0$.

Example 4.11.4. Exercise 4.11.1 is the particular case of Exercise 4.11.2 when all the integers $p_1, p_2, \ldots, p_{k-1}$ equal 1. (Indeed, the equality (230) turns into (227) in this case.)

Example 4.11.5. Let us illustrate why it is important that the integers $p_1, p_2, \ldots, p_{k-1}$ in Exercise 4.11.2 are positive. Indeed, let us pick $k = 4$, $p_1 = 1$, $p_2 = 1$ and $p_3 = 0$ (so p_3 is not positive but merely nonnegative). Then, the sequence $(t_0, t_1, t_2, \ldots)$ satisfies $t_0 = t_1 = t_2 = t_3 = 1$ and

$$t_n = \frac{1 + t_{n-1} t_{n-2}}{t_{n-4}} \qquad \text{for each } n \geq 4.$$

Let us make a table of the first few entries of the sequence $(t_0, t_1, t_2, \ldots)$:

n	0	1	2	3	4	5	6	7	8	9
t_n	1	1	1	1	2	3	7	22	$\frac{155}{2}$	$\frac{1706}{3}$

.

We see that t_8 and t_9 are not positive integers. So the claim of Exercise 4.11.2 does not hold in this case.

Solution to Exercise 4.11.2. For each integer $n \geq k$, we have

$$t_n t_{n-k} = 1 + t_{n-1}^{p_1} t_{n-2}^{p_2} \cdots t_{n-k+1}^{p_{k-1}}. \tag{231}$$

(Indeed, this follows by multiplying both sides of the equality (230) by t_{n-k}.)

We must prove that

$$t_n \text{ is a positive integer} \tag{232}$$

for each integer $n \geq 0$.

We shall prove (232) by strong induction on n:

Induction step: Let $m \geq 0$ be an integer. Assume (as the induction hypothesis) that (232) holds for $n < m$. We must prove that (232) holds for $n = m$. In other

words, we must prove that t_m is a positive integer. It clearly suffices to show that t_m is an integer (since we already know that t_m is positive).

Hence, our goal is to prove that t_m is an integer. If $m < k$, then this is clearly true[112]. Thus, for the rest of this proof, we WLOG assume that we don't have $m < k$. Hence, $m \geq k$. Thus, (230) (applied to $n = m$) yields

$$t_m = \frac{1 + t_{m-1}^{p_1} t_{m-2}^{p_2} \cdots t_{m-k+1}^{p_{k-1}}}{t_{m-k}}. \tag{233}$$

We have assumed that (232) holds for $n < m$. In other words, t_n is a positive integer for each $n < m$. In other words, $t_0, t_1, \ldots, t_{m-1}$ are positive integers. In other words, the numbers $t_{m-1}, t_{m-2}, t_{m-3}, \ldots, t_0$ are positive integers. This will allow us to do modular arithmetic with these numbers (i.e., to state congruences between these numbers).

Recall that our goal is to prove that t_m is an integer. If $m - k < k$, then this can easily be done[113]. Hence, for the rest of this proof, we WLOG assume that we don't have $m - k < k$. Thus, $m - k \geq k$.

Let $r = m - k$. Then, $r = m - k \geq k \geq 0$, so that $r \in \mathbb{N}$. Hence, (231) (applied to $n = r$) yields

$$t_r t_{r-k} = 1 + t_{r-1}^{p_1} t_{r-2}^{p_2} \cdots t_{r-k+1}^{p_{k-1}} \tag{234}$$

(since $r \geq k$). Note that the numbers $t_r, t_{r-1}, t_{r-2}, \ldots, t_{r-k}$ are among the numbers $t_{m-1}, t_{m-2}, t_{m-3}, \ldots, t_0$ [114], and thus are integers[115]. Thus, all the numbers that occur on either side of (234) are integers. Therefore,

$$\begin{aligned} t_r \mid t_r t_{r-k} &= 1 + t_{r-1}^{p_1} t_{r-2}^{p_2} \cdots t_{r-k+1}^{p_{k-1}} \qquad \text{(by (234))} \\ &= t_{r-1}^{p_1} t_{r-2}^{p_2} \cdots t_{r-k+1}^{p_{k-1}} - (-1). \end{aligned}$$

In other words,

$$t_{r-1}^{p_1} t_{r-2}^{p_2} \cdots t_{r-k+1}^{p_{k-1}} \equiv -1 \bmod t_r.$$

In other words,

$$-1 \equiv t_{r-1}^{p_1} t_{r-2}^{p_2} \cdots t_{r-k+1}^{p_{k-1}} \bmod t_r. \tag{235}$$

[112] Indeed, if $m < k$, then $t_m = 1$ (by (229), applied to $n = m$), and thus t_m is an integer (since 1 is an integer).

[113] *Proof.* Assume that $m - k < k$. Note that $m - k \in \mathbb{N}$ (since $m \geq k$). Hence, (229) (applied to $n = m - k$) yields $t_{m-k} = 1$ (since $m - k < k$). Now, (233) becomes

$$\begin{aligned} t_m &= \frac{1 + t_{m-1}^{p_1} t_{m-2}^{p_2} \cdots t_{m-k+1}^{p_{k-1}}}{t_{m-k}} = \frac{1 + t_{m-1}^{p_1} t_{m-2}^{p_2} \cdots t_{m-k+1}^{p_{k-1}}}{1} \qquad (\text{since } t_{m-k} = 1) \\ &= 1 + t_{m-1}^{p_1} t_{m-2}^{p_2} \cdots t_{m-k+1}^{p_{k-1}}. \end{aligned}$$

The right hand side of this equality is an integer (since $t_{m-1}, t_{m-2}, t_{m-3}, \ldots, t_0$ are integers). Therefore, so is the left hand side. In other words, t_m is an integer. Hence, we have shown that t_m is an integer under the assumption that $m - k < k$. Qed.

[114] because $r = m - \underbrace{k}_{\geq 2 \geq 1} \leq m - 1$ and $\underbrace{r}_{\geq k} - k \geq k - k = 0$

[115] since the numbers $t_{m-1}, t_{m-2}, t_{m-3}, \ldots, t_0$ are integers

Also, the numbers $t_{m-1}, t_{m-2}, \ldots, t_{m-k+1}$ are among the numbers $t_{m-1}, t_{m-2}, t_{m-3}, \ldots, t_0$ [116], and thus are integers[117]. Hence,

$$\begin{aligned}
&- t_{m-1}^{p_1} t_{m-2}^{p_2} \cdots t_{m-k+1}^{p_{k-1}} \\
&= \underbrace{(-1)}_{\substack{\equiv t_{r-1}^{p_1} t_{r-2}^{p_2} \cdots t_{r-k+1}^{p_{k-1}} \bmod t_r \\ \text{(by (235))}}} \left(t_{m-1}^{p_1} t_{m-2}^{p_2} \cdots t_{m-k+1}^{p_{k-1}} \right) \\
&\equiv \underbrace{\left(t_{r-1}^{p_1} t_{r-2}^{p_2} \cdots t_{r-k+1}^{p_{k-1}} \right)}_{= \prod_{i=1}^{k-1} t_{r-i}^{p_i}} \underbrace{\left(t_{m-1}^{p_1} t_{m-2}^{p_2} \cdots t_{m-k+1}^{p_{k-1}} \right)}_{= \prod_{i=1}^{k-1} t_{m-i}^{p_i}} = \left(\prod_{i=1}^{k-1} t_{r-i}^{p_i} \right) \left(\prod_{i=1}^{k-1} t_{m-i}^{p_i} \right) \\
&= \prod_{i=1}^{k-1} \underbrace{\left(t_{r-i}^{p_i} t_{m-i}^{p_i} \right)}_{= (t_{r-i} t_{m-i})^{p_i}} = \prod_{i=1}^{k-1} \left(t_{r-i} t_{m-i} \right)^{p_i} \bmod t_r. \qquad (236)
\end{aligned}$$

On the other hand, we claim that

$$t_{r-i} t_{m-i} \equiv 1 \bmod t_r \qquad (237)$$

for each $i \in \{1, 2, \ldots, k-1\}$.

[*Proof of (237):* Let $i \in \{1, 2, \ldots, k-1\}$. Then, $1 \leq i \leq k-1$, so that $m - \underbrace{i}_{\leq k-1 \leq k} \geq m - k \geq k \geq 0$. We now know that $m - i \geq k$, so that we can apply (231) to $n = m - i$. We thus find

$$t_{m-i} t_{m-i-k} = 1 + t_{m-i-1}^{p_1} t_{m-i-2}^{p_2} \cdots t_{m-i-k+1}^{p_{k-1}}.$$

In other words,

$$t_{m-i-1}^{p_1} t_{m-i-2}^{p_2} \cdots t_{m-i-k+1}^{p_{k-1}} = t_{m-i} t_{m-i-k} - 1. \qquad (238)$$

The numbers $t_{m-i}, t_{m-i-1}, \ldots, t_{m-i-k}$ are among the numbers $t_{m-1}, t_{m-2}, t_{m-3}, \ldots, t_0$ [118], and thus are integers[119]. Thus, all the numbers that occur on either side of (238) are integers. Let us now set $j = k - i$. Then, $j = k - i \in \{1, 2, \ldots, k-1\}$ (since $i \in \{1, 2, \ldots, k-1\}$). Hence, $t_{m-i-j}^{p_j}$ is one of the factors in the product $t_{m-i-1}^{p_1} t_{m-i-2}^{p_2} \cdots t_{m-i-k+1}^{p_{k-1}}$. Therefore, $t_{m-i-j}^{p_j} \mid t_{m-i-1}^{p_1} t_{m-i-2}^{p_2} \cdots t_{m-i-k+1}^{p_{k-1}}$ (since the

[116] since $\underbrace{m-k}_{=r} + 1 = r + 1 \geq r \geq 0$

[117] since the numbers $t_{m-1}, t_{m-2}, t_{m-3}, \ldots, t_0$ are integers

[118] because $m - \underbrace{i}_{\geq 1} \leq m - 1$ and $\underbrace{m-i}_{\geq k} - k \geq k - k = 0$

[119] since the numbers $t_{m-1}, t_{m-2}, t_{m-3}, \ldots, t_0$ are integers

numbers $t_{m-i}, t_{m-i-1}, \ldots, t_{m-i-k}$ are integers). Moreover, p_j is a positive integer[120]; thus, $t_{m-i-j} \mid t_{m-i-j}^{p_j}$. Hence,

$$\begin{aligned} t_{m-i-j} \mid t_{m-i-j}^{p_j} \mid t_{m-i-1}^{p_1} t_{m-i-2}^{p_2} \cdots t_{m-i-k+1}^{p_{k-1}} &= t_{m-i} t_{m-i-k} - 1 \qquad \text{(by (238))} \\ &= t_{m-i} t_{r-i} - 1 \qquad \left(\text{since } m-i-k = \underbrace{m-k}_{=r} - i = r-i \right) \\ &= t_{r-i} t_{m-i} - 1. \end{aligned}$$

In other words, $t_{r-i} t_{m-i} \equiv 1 \bmod t_{m-i-j}$. In view of $m - i - \underbrace{j}_{=k-i} = m - i - (k-i) = m - k = r$, this rewrites as $t_{r-i} t_{m-i} \equiv 1 \bmod t_r$. This proves (237).]

Now, each $i \in \{1, 2, \ldots, k-1\}$ satisfies $t_{r-i} t_{m-i} \equiv 1 \bmod t_r$ (by (237)) and therefore

$$\begin{aligned} (t_{r-i} t_{m-i})^{p_i} &\equiv 1^{p_i} \qquad \left(\begin{array}{c} \text{by Proposition 3.2.7, applied to } t_{r-i} t_{m-i}, 1, t_r \text{ and } p_i \\ \text{instead of } a, b, n \text{ and } k \end{array} \right) \\ &= 1 \bmod t_r. \end{aligned} \tag{239}$$

Hence, (236) becomes[121]

$$-t_{m-1}^{p_1} t_{m-2}^{p_2} \cdots t_{m-k+1}^{p_{k-1}} \equiv \prod_{i=1}^{k-1} \underbrace{(t_{r-i} t_{m-i})^{p_i}}_{\substack{\equiv 1 \bmod t_r \\ \text{(by (239))}}} \equiv \prod_{i=1}^{k-1} 1 = 1 \bmod t_r.$$

In other words, $1 \equiv -t_{m-1}^{p_1} t_{m-2}^{p_2} \cdots t_{m-k+1}^{p_{k-1}} \bmod t_r$. In other words,

$$t_r \mid 1 - \left(-t_{m-1}^{p_1} t_{m-2}^{p_2} \cdots t_{m-k+1}^{p_{k-1}} \right) = 1 + t_{m-1}^{p_1} t_{m-2}^{p_2} \cdots t_{m-k+1}^{p_{k-1}}. \tag{240}$$

The number t_r is among the numbers $t_{m-1}, t_{m-2}, t_{m-3}, \ldots, t_0$ [122], and thus is a positive integer[123], hence is nonzero. Thus, from (240), we conclude that

$$\frac{1 + t_{m-1}^{p_1} t_{m-2}^{p_2} \cdots t_{m-k+1}^{p_{k-1}}}{t_r} \in \mathbb{Z}.$$

In view of $r = m - k$, this rewrites as

$$\frac{1 + t_{m-1}^{p_1} t_{m-2}^{p_2} \cdots t_{m-k+1}^{p_{k-1}}}{t_{m-k}} \in \mathbb{Z}.$$

In view of (233), this rewrites as $t_m \in \mathbb{Z}$. In other words, t_m is an integer. As we know, this is sufficient to complete the induction step. Thus, the induction step is complete, and (232) is proved. This solves Exercise 4.11.2. $\square$

[120] since $p_1, p_2, \ldots, p_{k-1}$ are positive integers

[121] In the following computation, we use (41) to multiply several congruences modulo t_r.

[122] because $r = m - \underbrace{k}_{\geq 2 \geq 1} \leq m - 1$ and $r \geq k \geq 0$

[123] since the numbers $t_{m-1}, t_{m-2}, t_{m-3}, \ldots, t_0$ are positive integers

Thus, we have solved Exercise 4.11.1 as well (since Exercise 4.11.1 is the particular case of Exercise 4.11.2 when all of $p_1, p_2, \ldots, p_{k-1}$ equal 1). This yields a new solution to Exercise 1.1.2 **(b)** (since the claim of Exercise 4.11.1 in the case $k = 3$ is precisely the claim of Exercise 1.1.2 **(b)**).

4.11.2. The Somos sequences

Here is another problem of the "unexpected integrality" type ([Negut05, Chapter 2, Exercise 25], [Wemyss14, Theorem 2.6]):

Exercise 4.11.3. Define a sequence $(a_0, a_1, a_2, \ldots)$ of positive rational numbers recursively by setting

$$a_n = 1 \qquad \text{for each } n < 5 \tag{241}$$

and

$$a_n = \frac{a_{n-1}a_{n-4} + a_{n-2}a_{n-3}}{a_{n-5}} \qquad \text{for each } n \geq 5. \tag{242}$$

(For example, $a_5 = \dfrac{a_4 a_1 + a_3 a_2}{a_0} = \dfrac{1 \cdot 1 + 1 \cdot 1}{1} = 2$ and likewise $a_6 = 3$.)

Prove that a_n is a positive integer for each integer $n \geq 0$.

The sequence $(a_0, a_1, a_2, \ldots)$ in Exercise 4.11.3 is known as the *Somos*-5 *sequence*; it is sequence A006721 in the OEIS. Here is a table of its first few entries:

n	0	1	2	3	4	5	6	7	8	9	10	11
a_n	1	1	1	1	1	2	3	5	11	37	83	274

.

The growth rate of this sequence is faster than exponential[124]; thus, the sequence is not $(b_1, b_2, \ldots, b_k)$-recurrent (for any $b_1, b_2, \ldots, b_k$).

Here is a similar sequence ([Malouf92, Theorem 1]):

Exercise 4.11.4. Define a sequence $(a_0, a_1, a_2, \ldots)$ of positive rational numbers recursively by setting

$$a_n = 1 \qquad \text{for each } n < 4$$

and

$$a_n = \frac{a_{n-1}a_{n-3} + a_{n-2}^2}{a_{n-4}} \qquad \text{for each } n \geq 4.$$

Prove that a_n is a positive integer for each integer $n \geq 0$.

[124] In fact, using Landau's Big-O notation, we have $a_n = O\left(q^{n^2}\right)$ for a constant $q \approx 1.0728$. See [Brown20] for a proof.

The sequence $(a_0, a_1, a_2, \ldots)$ in Exercise 4.11.4 is known as the *Somos-4 sequence*; it is sequence A006720 in the OEIS. Here is a table of its first few entries:

n	0	1	2	3	4	5	6	7	8	9	10	11
a_n	1	1	1	1	2	3	7	23	59	314	1529	8209

.

At this point, it should be clear (both from their names and from their definitions) that the Somos-4 and Somos-5 sequences are part of a sequence of sequences:

Definition 4.11.6. Let $k \geq 2$ be an integer. The *Somos-k sequence* is the sequence $(a_0, a_1, a_2, \ldots)$ of positive rational numbers defined recursively by setting

$$a_n = 1 \qquad \text{for each } n < k$$

and

$$\begin{aligned} a_n &= \frac{a_{n-1}a_{n-k+1} + a_{n-2}a_{n-k+2} + \cdots + a_{n-\lfloor k/2 \rfloor}a_{n-k+\lfloor k/2 \rfloor}}{a_{n-k}} \\ &= \frac{\sum_{i=1}^{\lfloor k/2 \rfloor} a_{n-i}a_{n-k+i}}{a_{n-k}} \qquad \text{for each } n \geq k. \end{aligned}$$

It is easy to see that the Somos-2 and Somos-3 sequences are just the constant sequence $(1, 1, 1, \ldots)$. We have seen the Somos-4 and Somos-5 sequences. Surprisingly, all entries of the Somos-k sequences for all $k \in \{2, 3, \ldots, 7\}$ are integers. However, the magic stops at $k = 7$: The 17-th entry of the Somos-8 sequence is not an integer.

There is much to say about the Somos-k sequences, as well as other sequences with similar definitions and properties (we will in fact see some of them later on). Good starting points are the Wikipedia page for "Somos sequence" and Jim Propp's "Somos Sequences Site". As mentioned above, [Gale98, Chapters 1 and 4] gives an introduction with a few of the simplest proofs. Similar sequences can be found in [FomZel02], [AlCuHu16] and [Russel16, Chapters 5–6]. Most of these sequences have connections to combinatorics (in particular, the Somos-k sequences for all $k \in \{4, 5, 6, 7\}$ are known to enumerate perfect matchings in certain graphs); also, the Somos-k sequences are related to elliptic curves (see, e.g., [Poorte04], or – for a particularly short connection – the solution to Fifth Day problem 1 in [Zagier96]). The Somos-4 and Somos-5 sequences have nontrivial divisibility properties as well ([Kamp15]).

We shall not dwell on these wide and fertile grounds; however, we shall state (without proof) two theorems that generalize the integrality of the Somos-k sequences for $k \in \{4, 5, 6, 7\}$. The first generalizes the integrality of the Somos-4 and Somos-5 sequences:

Theorem 4.11.7. Let $k \geq 2$ be an integer. Let i and j be two elements of $\{1, 2, \ldots, k-1\}$. Let w and z be two positive integers. Define a sequence $(a_0, a_1, a_2, \ldots)$ of positive rational numbers recursively by setting

$$a_n = 1 \qquad \text{for each } n < k$$

and

$$a_n = \frac{w a_{n-i} a_{n-k+i} + z a_{n-j} a_{n-k+j}}{a_{n-k}} \qquad \text{for each } n \geq k. \tag{243}$$

Then, a_n is a positive integer for each integer $n \geq 0$.

Exercise 4.11.4 (that is, the integrality of the Somos-4 sequence) is the particular case of Theorem 4.11.7 for $k = 4$, $i = 1$, $j = 2$, $w = 1$ and $z = 1$. Exercise 4.11.3 (that is, the integrality of the Somos-5 sequence) is the particular case of Theorem 4.11.7 for $k = 5$, $i = 1$, $j = 2$, $w = 1$ and $z = 1$.

Note that the only reason why we required w and z to be positive in Theorem 4.11.7 is to ensure that all of $a_0, a_1, a_2, \ldots$ are positive; this prevents the denominator a_{n-k} in (243) from becoming zero. If $a_{n-k} \neq 0$ can be guaranteed in some other way, then we can drop the requirement that w and z be positive (although, of course, the claim must be changed: a_n will not generally be a positive integer, but merely an integer). The "grown-up" version of Theorem 4.11.7 does not deal with a sequence of integers at all; instead, it considers w and z as indeterminates, so all of $a_0, a_1, a_2, \ldots$ are rational functions in these two indeterminates w and z (with integer coefficients). The claim is then that these rational functions $a_0, a_1, a_2, \ldots$ are actually polynomials with nonnegative integer coefficients. This has been proved combinatorially (meaning that these polynomials a_n have been described explicitly, as sums over matchings in certain graphs!) in [BMPrWe09, Theorem 11]. An alternative algebraic proof (using the algebra of Laurent polynomials) has been given in [FomZel02, Example 1.7].

The next theorem generalizes the integrality of the Somos-6 and Somos-7 sequences:

Theorem 4.11.8. Let i, j and ℓ be three distinct positive integers. Let $k = i + j + \ell$. Let u, v and w be three positive integers. Define a sequence $(a_0, a_1, a_2, \ldots)$ of positive rational numbers recursively by setting

$$a_n = 1 \qquad \text{for each } n < k$$

and

$$a_n = \frac{u a_{n-i} a_{n-k+i} + v a_{n-j} a_{n-k+j} + w a_{n-\ell} a_{n-k+\ell}}{a_{n-k}} \qquad \text{for each } n \geq k.$$

Then, a_n is a positive integer for each integer $n \geq 0$.

The integrality of the Somos-6 sequence is the particular case of Theorem 4.11.8

for $k = 6$, $i = 1$, $j = 2$, $\ell = 3$, $u = 1$, $v = 1$ and $w = 1$. The integrality of the Somos-7 sequence is the particular case of Theorem 4.11.8 for $k = 7$, $i = 1$, $j = 2$, $\ell = 4$, $u = 1$, $v = 1$ and $w = 1$.

Theorem 4.11.8 is known as the *Gale-Robinson sequence theorem*; it has been proved combinatorially in [CarSpe04, Theorem 6], and algebraically in [FomZel02, Example 1.3]. Again, u, v and w are best considered as indeterminates rather than fixed positive integers; the resulting rational functions a_n are polynomials having explicit combinatorial interpretations.

4.11.3. Odds and ends

Exercise 4.11.4 can be generalized in yet another direction: by putting exponents on the a_{n-i}'s (just as in Exercise 4.11.2). Indeed, we have the following ([Gale98, Chapter 1, (8)]):

Exercise 4.11.5. Fix three positive integers p, q and r. Define a sequence $(a_0, a_1, a_2, \ldots)$ of positive rational numbers recursively by setting

$$a_n = 1 \qquad \text{for each } n < 4$$

and

$$a_n = \frac{a_{n-1}^p a_{n-3}^q + a_{n-2}^r}{a_{n-4}} \qquad \text{for each } n \geq 4. \tag{244}$$

Prove that a_n is a positive integer for each integer $n \geq 0$.

However, this generalization cannot be merged with the w and z from Theorem 4.11.7. To wit, if we replace (244) by

$$a_n = \frac{3a_{n-1}a_{n-3} + a_{n-2}}{a_{n-4}} \qquad \text{for each } n \geq 4,$$

then $a_8 = 10337/2$ will fail to be an integer. On the other hand, the recurrence

$$a_n = \frac{a_{n-1}a_{n-3} + za_{n-2}}{a_{n-4}} \qquad \text{for each } n \geq 4$$

appears to produce integers a_n for each $z \in \mathbb{N}$ and each $n \in \mathbb{N}$. (Is there a proof?)

This area still offers many challenges and surprises. Let me mention a counterexample and a few open problems:

Example 4.11.9. Define a sequence $(a_0, a_1, a_2, \ldots)$ of positive rational numbers recursively by setting $a_0 = 1$ and

$$a_n = \frac{1 + a_0^2 + a_1^2 + \cdots + a_{n-1}^2}{n} \qquad \text{for each } n \geq 1. \tag{245}$$

This is known as *Göbel's sequence* (see also OEIS sequence A003504, which is the same sequence shifted by an entry). It may appear that all its entries are integers; but this is not so! The smallest n for which a_n is not an integer is 43; the corresponding $a_n = a_{43}$ has about a billion digits before the decimal point. (See [Zagier96, Fifth Day, problem 3] for a more detailed discussion of this sequence; note, however, that it is defined starting with $a_1 = 2$ rather than $a_0 = 1$ there.)

Note that the recurrence relation (245) can be replaced by

$$a_n = \frac{(n-1)\, a_{n-1} + a_{n-1}^2}{n} \qquad \text{for each } n \geq 1.$$

Question 4.11.10. (MathOverflow questions #248604 and #323963:) Define a sequence $(a_0, a_1, a_2, \ldots)$ of positive rational numbers recursively by setting

$$a_n = 1 \qquad \text{for each } n < 4$$

and

$$a_n = \frac{(a_{n-1}+1)\,(a_{n-2}+1)\,(a_{n-3}+1)}{a_{n-4}} \qquad \text{for each } n \geq 4.$$

This is sequence A276123 in the OEIS. Is a_n an integer for each $n \in \mathbb{N}$? It has been shown by mercio (math.stackexchange #1906097) that each a_n is a "2-integer" (i.e., it becomes an integer when multiplied with a sufficiently large power of 2), but it is not clear whether the power of 2 can be dispensed with.

More generally, we can fix an integer $k \geq 2$, and define a sequence $(a_0, a_1, a_2, \ldots)$ of positive rational numbers recursively by setting

$$a_n = 1 \qquad \text{for each } n < k$$

and

$$a_n = \frac{(a_{n-1}+1)\,(a_{n-2}+1)\cdots(a_{n-k+1}+1)}{a_{n-k}} \qquad \text{for each } n \geq k.$$

The above sequence is obtained for $k = 4$. For each $k \geq 2$, one can ask whether all the a_n are integers. The cases $k = 2$ and $k = 3$ have been resolved in the positive (i.e., it is known that each a_n is an integer when $k \in \{2, 3\}$). The case $k = 5$ is also open (MathOverflow question #323963). For each $k > 5$, the answer is negative (viz., a_{2k} is not an integer).

Here are three similar questions (all found by the same Michael Somos who the Somos-k sequences originate with[125]) that have been answered in the positive:

[125] See math.stackexchange question #3586309.

Exercise 4.11.6. Define a sequence $(x_0, x_1, x_2, \ldots)$ of positive rational numbers recursively by setting

$$x_n = 1 \qquad \text{for each } n < 6, \tag{246}$$

and

$$x_n = \frac{x_{n-3}\left(x_{n-1} + x_{n-5}\right)}{x_{n-6}} \qquad \text{for each } n \geq 6. \tag{247}$$

(a) Prove that $x_n + x_{n-4} + x_{n-8} = 6x_{n-3}x_{n-4}x_{n-5}$ for each $n \geq 8$.
(b) Prove that x_n an integer for each $n \in \mathbb{N}$.

Here is a table of the first few entries of this sequence:

n	0	1	2	3	4	5	6	7	8	9	10	11
x_n	1	1	1	1	1	1	2	3	4	10	33	140

.

Exercise 4.11.7. Define a sequence $(a_0, a_1, a_2, \ldots)$ of positive rational numbers recursively by setting

$$a_0 = 2, \qquad a_1 = 1, \qquad a_2 = 1, \qquad a_3 = 1, \tag{248}$$

and

$$a_n = \frac{\left(a_{n-1} + a_{n-2}\right)\left(a_{n-2} + a_{n-3}\right)}{a_{n-4}} \qquad \text{for each } n \geq 4. \tag{249}$$

This is sequence A248049 in the OEIS.
(a) Prove that a_n an integer for each $n \in \mathbb{N}$.
(b) Prove that $a_n = x_{n+2}x_{n+1}x_nx_{n-1}$ for each $n \geq 1$, where $(x_0, x_1, x_2, \ldots)$ is the sequence defined in Exercise 4.11.6.

Here is a table of the first few entries of this sequence:

n	0	1	2	3	4	5	6	7	8	9	10	11
a_n	2	1	1	1	2	6	24	240	3960	184800	33033000	26125799700

.

Exercise 4.11.8. Define a sequence $(b_0, b_1, b_2, \ldots)$ of positive rational numbers recursively by setting

$$b_0 = 1, \qquad b_1 = 1, \qquad b_2 = 1, \qquad b_3 = 1, \tag{250}$$

and

$$b_n = \frac{b_{n-2}\left(b_{n-1} + b_{n-3}\right)}{b_{n-4}} \qquad \text{for each } n \geq 4. \tag{251}$$

This is sequence A078918 in the OEIS.
(a) Prove that b_n an integer for each $n \in \mathbb{N}$.
(b) Prove that $b_n = x_{n+2}x_n$ for each $n \geq 0$, where $(x_0, x_1, x_2, \ldots)$ is the sequence defined in Exercise 4.11.6.
(c) Prove that $a_n = b_nb_{n-1}$ for each $n \geq 1$, where $(a_0, a_1, a_2, \ldots)$ is the sequence defined in Exercise 4.11.7.

Here is a table of the first few entries of this sequence:

n	0	1	2	3	4	5	6	7	8	9	10	11
b_n	1	1	1	1	2	3	8	30	132	1400	23595	1107260

.

Solutions to Exercises 4.11.6, 4.11.7 and Exercise 4.11.8 can be found in Section A.5.

5. The Extremal Principle

This chapter is devoted to one of the simplest tricks in mathematics: **When you have a bunch of objects, look at the smallest or the largest among them**. This is known as the *Extremal Principle*, since the word "extremal" includes both "smallest" and "largest". This principle is far from a universal strategy that can be applied mechanically; in particular, figuring out the exact meaning of "smallest" and "largest" is often a delicate and creative matter. For example, if you have a bunch of finite sets of integers, which one is the smallest? The one of smallest size?[126] The one whose smallest element is the smallest? The one whose sum of elements is the smallest? The answer depends on what you are trying to achieve. When applying the Extremal Principle, it is often necessary to cycle through several possible meanings of "smallest" (or "largest") before finding one that helps solve the problem.

5.1. Existence theorems

Before we see the Extremal Principle applied, let me recall some theorems that guarantee the existence of extremal (i.e., minimal or maximal) objects. These theorems are all fairly basic and will mostly be used without proof; nevertheless it is worth recalling them at least once.

Theorem 5.1.1. Let S be a nonempty finite set of real numbers. Then, S has a minimum and a maximum.

Proof of Theorem 5.1.1. Proposition 2.1.2 shows that S has a maximum[127]. The same argument (mutatis mutandis[128]) shows that S has a minimum. This proves Theorem 5.1.1. □

The word "nonempty" in Theorem 5.1.1 is clearly necessary: The empty set has neither a minimum nor a maximum. The word "finite", too, cannot simply be dropped (for example, the set $\left\{\frac{1}{1}, \frac{1}{2}, \frac{1}{3}, \ldots\right\}$ has no minimum); but there are many infinite sets that nevertheless have minima or maxima. For example, the following holds ([Grinbe15, Theorem 2.44]):

[126] We are saying "The one" here, but of course there can be several ones. For example, among the three sets $\{1,2\}$, $\{1,3\}$ and $\{0,1,2\}$, the first two have the smallest size.

[127] To be more precise, Proposition 2.1.2 shows this under the assumption that S is a set of **integers**, while we are here only assuming that S is a set of **real numbers**. However, the proof is the same in either case.

[128] Once again, the words "mutatis mutandis" mean "once the things that need to be changed have been changed". The things that need to be changed here are the following: The "$\geq$" sign needs to be flipped (i.e., replaced by a "$\leq$" sign); the word "maximum" needs to be replaced by "minimum".

Theorem 5.1.2. Let S be a nonempty set of nonnegative integers. Then, S has a minimum.

Of course, S will not usually have a maximum in this situation; for example, $\mathbb{N}$ itself has no maximum.

Note that we have used Theorem 5.1.2 in Definition 3.6.2 (when arguing that the smallest positive common multiple of k nonzero integers is well-defined). We have also tacitly used Theorem 5.1.2 in Corollary 4.7.9 (in the corollary itself, not in the proof), when defining m as the smallest period of u. Indeed, the existence of "the smallest period of u" follows from the fact that the set $\{\text{periods of } u\}$ has a smallest element (i.e., a minimum); but this is a particular case of Theorem 5.1.2.

Theorem 5.1.2 can be generalized: For a set of integers, the existence of a *lower bound* guarantees the existence of a minimum, while the existence of an *upper bound* guarantees the existence of a maximum. Let us recall what lower and upper bounds are:

Definition 5.1.3. Let S be a set of real numbers. Let a be a real number.

(a) We say that a is a *lower bound* for S if and only if every $s \in S$ satisfies $a \leq s$.

(b) We say that a is an *upper bound* for S if and only if every $s \in S$ satisfies $a \geq s$.

For example, 0 and any negative real number are lower bounds for the set $\left\{\frac{1}{1}, \frac{1}{2}, \frac{1}{3}, \ldots\right\}$, whereas 1 and any larger number are upper bounds for this set. No positive real is a lower bound for this set, because if a is a positive real, then there exists a positive integer n such that $\frac{1}{n} < a$ (for example, we can take $n = \left\lfloor \frac{1}{a} + 1 \right\rfloor$).

Now, here is the promised generalization of Theorem 5.1.2:

Theorem 5.1.4. Let S be a nonempty set of integers. Then:

(a) If S has a lower bound, then S has a minimum.

(b) If S has an upper bound, then S has a maximum.

Note that the word "integers" in Theorem 5.1.4 is important. Even with rational numbers, the theorem would fail; for example, as we have seen above, the set $\left\{\frac{1}{1}, \frac{1}{2}, \frac{1}{3}, \ldots\right\}$ has a lower bound but no minimum. (And likewise, the set $\left\{-\frac{1}{1}, -\frac{1}{2}, -\frac{1}{3}, \ldots\right\}$ has an upper bound but no maximum.)

For the sake of completeness, let us give a proof of Theorem 5.1.4:

Proof of Theorem 5.1.4. **(a)** Assume that S has a lower bound. Let a be this lower bound.

Thus, a is a lower bound of S. In other words, every $s \in S$ satisfies

$$a \leq s \tag{252}$$

(by the definition of a "lower bound").

There exists some $t \in S$ (since t is nonempty). Consider this t. Applying (252) to $s = t$, we obtain $a \leq t$.

Let H denote the set of all integers z satisfying $a \leq z \leq t$. Then, the set H is finite (since there are only finitely many integers between a and t). Note that t is an integer z satisfying $a \leq z \leq t$ (since t is an integer[129] and since $a \leq t \leq t$). In other words, $t \in H$ (since H is the set of all such integers z).

Define a set T of real numbers by $T = S \cap H$. Then, the set T is a subset of H (since $T = S \cap H \subseteq H$), and thus is finite (since the set H is finite). Moreover, $t \in T$ [130]. Hence, the set T is nonempty. Now, we know that T is a nonempty finite set of real numbers. Therefore, Theorem 5.1.1 (applied to T instead of S) yields that T has a minimum and a maximum. In particular, this shows that T has a minimum. Let m be this minimum. By the definition of a "minimum", this means that $m \in T$ and that we have

$$m \leq u \text{ for all } u \in T. \tag{253}$$

Applying (253) to $u = t$, we find $m \leq t$ (since $t \in T$).

Now, it is easy to see that $m \in S$ (since $m \in T = S \cap H \subseteq S$) and that we have $m \leq u$ for all $u \in S$ [131]. In other words, m is a minimum of S (by the definition of a "minimum"). Therefore, the set S has a minimum. This proves Theorem 5.1.4 **(a)**.

(b) The same argument that we just used to prove Theorem 5.1.4 **(a)** can be reused (mutatis mutandis) to prove Theorem 5.1.4 **(b)**. ("Mutatis mutandis" here means replacing "minimum" by "maximum" and "lower bound" by "upper bound", as well as flipping all inequality signs.) □

Theorem 5.1.2 is a particular case of Theorem 5.1.4 **(a)**, since any set of nonnegative integers has a lower bound (namely, 0).

We note in passing that some things can be said about infinite sets of real numbers, even if neither Theorem 5.1.1 nor Theorem 5.1.4 apply to them. Namely, we can define two "weaker" versions of minima and maxima:

Definition 5.1.5. Let S be a set of real numbers.

(a) An *infimum* (or *greatest lower bound*) of S means a maximum of the set of all lower bounds of S. That is, it means a lower bound a of S with the property that every lower bound b of S satisfies $a \geq b$. An infimum of S is unique if it exists, and is denoted by $\inf S$.

[129] because $t \in S$ (and since S is a set of integers)

[130] *Proof.* We have $t \in S$ and $t \in H$. Thus, $t \in S \cap H$. In other words, $t \in T$ (since $T = S \cap H$).

[131] *Proof.* Let $u \in S$. We must show that $m \leq u$. Indeed, assume the contrary. Thus, $m > u$, so that $u < m \leq t$. Also, $u \in S$ and thus $a \leq u$ (by (252), applied to $s = u$). Also, u is an integer (since $u \in S$, but S is a set of integers). Hence, u is an integer and satisfies $a \leq u \leq t$ (since $u < t$). In other words, u is an integer z satisfying $a \leq z \leq t$. In other words, $u \in H$ (since H is the set of all such integers z).

Combining $u \in S$ with $u \in H$, we obtain $u \in S \cap H = T$. Hence, (253) shows that $m \leq u$. But this contradicts $m > u$. This contradiction shows that our assumption was false. Hence, $m \leq u$ is proved.

(b) A *supremum* (or *least upper bound*) of S means a minimum of the set of all upper bounds of S. That is, it means an upper bound a of S with the property that every upper bound b of S satisfies $a \leq b$. A supremum of S is unique if it exists, and is denoted by $\sup S$.

If a set S of real numbers has a minimum, then this minimum is also an infimum of S; but the converse is not true. For example, the set $\left\{\frac{1}{1}, \frac{1}{2}, \frac{1}{3}, \ldots\right\}$ has infimum 0 but no minimum. Similar claims hold for maxima and suprema. See [Swanso20, Example 2.7.3] for various examples of infima and suprema. Now, the following weakening of Theorem 5.1.4 holds for sets of reals:

Theorem 5.1.6. Let S be a nonempty set of reals. Then:
(a) If S has a lower bound, then S has an infimum. This infimum is a minimum of S if and only if it belongs to S.
(b) If S has an upper bound, then S has a supremum. This supremum is a maximum of S if and only if it belongs to S.

Theorem 5.1.6 is one of the most fundamental results in analysis. We refer to [Swanso20, Theorem 3.8.5] for its proof (more precisely, for the proofs of the first sentences of parts **(a)** and **(b)**; but the second sentences are easy exercises).

An infimum is a weaker concept than a minimum, and often less useful. Thus, one sometimes wants stronger results that guarantee the existence of minima and maxima rather than merely infima and suprema. Topology can sometimes provide such results; in particular, the following holds:[132]

Theorem 5.1.7. Let S be a nonempty set of reals that is closed with respect to the topology on $\mathbb{R}$. Then:
(a) If S has a lower bound, then S has a minimum.
(b) If S has an upper bound, then S has a maximum.

The proof of Theorem 5.1.7 is implicit in [Swanso20, proof of Theorem 5.2.2].

5.2. Applications

5.2.1. Writing numbers as sums of powers of 2

We shall now see various uses of the Extremal Principle. We begin with a well-known result:

Theorem 5.2.1. Let $n \in \mathbb{N}$. Then, there is a unique finite subset T of $\mathbb{N}$ such that $n = \sum_{t \in T} 2^t$.

[132] We refer to [Swanso20, §3.13] for the definition of the topology on $\mathbb{R}$.

In words, Theorem 5.2.1 is saying that each $n \in \mathbb{N}$ can be written in a unique way as a sum of (finitely many) distinct powers of 2 (where "powers of 2" means numbers of the form 2^t with $t \in \mathbb{N}$). Here, of course, "in a unique way" means that the set of powers of 2 being added is unique (so we aren't counting $2^3 + 2^6$ and $2^6 + 2^3$ as two different ways). Here are some examples:

Example 5.2.2. If $n = 12$, then $n = 2^2 + 2^3 = \sum\limits_{t \in \{2,3\}} 2^t$.

If $n = 23$, then $n = 2^0 + 2^1 + 2^2 + 2^4 = \sum\limits_{t \in \{0,1,2,4\}} 2^t$.

If $n = 0$, then $n = (\text{empty sum}) = \sum\limits_{t \in \varnothing} 2^t$.

The reader will probably have noticed that Theorem 5.2.1 is just the existence and uniqueness of the base-2 representation of n in disguise. Indeed, if the base-2 representation of an $n \in \mathbb{N}$ is

$$n = b_k \cdot 2^k + b_{k-1} \cdot 2^{k-1} + \cdots + b_0 \cdot 2^0$$

(with $b_0, b_1, \ldots, b_k \in \{0, 1\}$), then

$$n = \sum_{t \in T} 2^t, \qquad \text{where } T = \{i \in \{0, 1, \ldots, k\} \mid b_i = 1\}.$$

(And conversely, if we have a finite subset T of $\mathbb{N}$ such that $n = \sum\limits_{t \in T} 2^t$, then we can easily obtain a base-2 representation of n from it.) Thus, we don't have a pressing need to prove Theorem 5.2.1. Nevertheless, let us do so, since the proof illustrates the Extremal Principle rather nicely.

We split Theorem 5.2.1 into two pieces: the existence and the uniqueness. While it is possible to prove them together, I prefer shorter proofs, even if some of the work will end up duplicated. The existence part of Theorem 5.2.1 is the following proposition:

Proposition 5.2.3. Let $n \in \mathbb{N}$. Then, there exists a finite subset T of $\mathbb{N}$ such that $n = \sum\limits_{t \in T} 2^t$.

The uniqueness part of Theorem 5.2.1 is the following proposition:

Proposition 5.2.4. Let $n \in \mathbb{N}$. Let T and T' be two finite subsets of $\mathbb{N}$ such that $n = \sum\limits_{t \in T} 2^t$ and $n = \sum\limits_{t \in T'} 2^t$. Then, $T = T'$.

Proof of Proposition 5.2.3. We shall prove Proposition 5.2.3 by strong induction on n:

Induction step: Let $m \in \mathbb{N}$. Assume (as the induction hypothesis) that Proposition 5.2.3 holds for $n < m$. We must prove that Proposition 5.2.3 holds for $n = m$.

In other words, we must prove that there exists a finite subset T of $\mathbb{N}$ such that $m = \sum_{t \in T} 2^t$.

If $m = 0$, then this is obvious[133]. Thus, for the rest of this proof, we WLOG assume that $m \neq 0$. Hence, $m \geq 1$.

It is easy to see that $m < 2^m$ [134]. Hence, there exists at least one $p \in \mathbb{N}$ such that $m < 2^p$ (for example, $p = m$ works). In other words, the set $\{p \in \mathbb{N} \mid m < 2^p\}$ is nonempty. Hence, Theorem 5.1.2 shows that this set has a minimum (since it is a set of nonnegative integers). In other words, there is a **smallest** $p \in \mathbb{N}$ satisfying $m < 2^p$. Let q be this smallest p.

Thus, q itself is a $p \in \mathbb{N}$ satisfying $m < 2^p$. In other words, $q \in \mathbb{N}$ and $m < 2^q$. Hence, $2^q > m \geq 1$, so that $q > 0$. Therefore, $q - 1 \in \mathbb{N}$. If we had $m < 2^{q-1}$, then $q - 1$ would therefore be a $p \in \mathbb{N}$ satisfying $m < 2^p$; but this would contradict the fact that q is the **smallest** such p. Hence, we cannot have $m < 2^{q-1}$. In other words, we have $m \geq 2^{q-1}$.

Hence, $m - 2^{q-1} \in \mathbb{N}$. Also, $m - 2^{q-1} < m$ (since $2^{q-1} > 0$). Thus, Proposition 5.2.3 holds for $n = m - 2^{q-1}$ (since we assumed that Proposition 5.2.3 holds for $n < m$). In other words, there exists a finite subset T of $\mathbb{N}$ such that $m - 2^{q-1} = \sum_{t \in T} 2^t$. Consider this T, and denote it by T_0. Hence, T_0 is a finite subset of $\mathbb{N}$ such that $m - 2^{q-1} = \sum_{t \in T_0} 2^t$.

Now, from $m - 2^{q-1} = \sum_{t \in T_0} 2^t$, we obtain

$$m = 2^{q-1} + \sum_{t \in T_0} 2^t. \tag{254}$$

It is not hard to see that $q - 1 \notin T_0$ [135]. Hence, $\sum_{t \in T_0 \cup \{q-1\}} 2^t = 2^{q-1} + \sum_{t \in T_0} 2^t$. Comparing this with (254), we find $m = \sum_{t \in T_0 \cup \{q-1\}} 2^t$. Thus, there exists a finite

[133] *Proof.* Assume that $m = 0$. Then, the empty set $\varnothing$ is a finite subset of $\mathbb{N}$ and satisfies $\sum_{t \in \varnothing} 2^t = (\text{empty sum}) = 0$ and thus $m = 0 = \sum_{t \in \varnothing} 2^t$. Hence, there exists a finite subset T of $\mathbb{N}$ such that $m = \sum_{t \in T} 2^t$ (namely, $T = \varnothing$). Thus, we have proved our claim under the assumption that $m = 0$.

[134] The quickest way to see this is to apply (5) to $n = m$ and conclude that $2^0 + 2^1 + \cdots + 2^{m-1} = 2^m - 1 < 2^m$, so that $2^m > \underbrace{2^0}_{\geq 1} + \underbrace{2^1}_{\geq 1} + \cdots + \underbrace{2^{m-1}}_{\geq 1} \geq \underbrace{1 + 1 + \cdots + 1}_{m \text{ times}} = m$, and thus $m < 2^m$.

[135] *Proof.* Assume the contrary. Thus, $q - 1 \in T_0$. Hence, 2^{q-1} is an addend of the sum $\sum_{t \in T_0} 2^t$. Since all addends of this sum are nonnegative, we thus conclude that $\sum_{t \in T_0} 2^t \geq 2^{q-1}$ (because a sum of nonnegative reals is always $\geq$ to each of its addends). Hence, $m - 2^{q-1} = \sum_{t \in T_0} 2^t \geq 2^{q-1}$, so that $m \geq 2^{q-1} + 2^{q-1} = 2 \cdot 2^{q-1} = 2^q$. But this contradicts $m < 2^q$. This contradiction shows that our assumption was false. Qed.

subset T of $\mathbb{N}$ such that $m = \sum_{t \in T} 2^t$ (namely, $T = T_0 \cup \{q-1\}$). This completes the induction step. Thus, Proposition 5.2.3 is proved by strong induction. $\square$

Proof of Proposition 5.2.4. We shall prove Proposition 5.2.4 by strong induction on n:

Induction step: Let $m \in \mathbb{N}$. Assume (as the induction hypothesis) that Proposition 5.2.4 holds for $n < m$. We must prove that Proposition 5.2.4 holds for $n = m$.

We have assumed that Proposition 5.2.4 holds for $n < m$. In other words, the following claim holds:

> *Claim 1:* Let $n \in \mathbb{N}$ satisfy $n < m$. Let T and T' be two finite subsets of $\mathbb{N}$ such that $n = \sum_{t \in T} 2^t$ and $n = \sum_{t \in T'} 2^t$. Then, $T = T'$.

Now, let us prove that Proposition 5.2.4 holds for $n = m$. Let T and T' be two finite subsets of $\mathbb{N}$ such that $m = \sum_{t \in T} 2^t$ and $m = \sum_{t \in T'} 2^t$. We shall show that $T = T'$.

If $m = 0$, then this is easy to prove[136]. Thus, for the rest of this proof, we WLOG assume that $m \neq 0$.

If we had $T = \varnothing$, then we would have $\sum_{t \in T} 2^t = \sum_{t \in \varnothing} 2^t = (\text{empty sum}) = 0$, which would contradict $\sum_{t \in T} 2^t = m \neq 0$. Hence, we cannot have $T = \varnothing$. Thus, the set T is nonempty. Hence, T is a nonempty finite set of integers. Therefore, Theorem 5.1.1 (applied to $S = T$) shows that T has a minimum and a maximum. Thus, in particular, T has a maximum. Similarly, T' has a maximum. Consider these two maxima $\max T$ and $\max(T')$.

We shall show that $\max T \leq \max(T')$. Indeed, assume the contrary. Thus, $\max T > \max(T')$. The number $\max T$ is an element of T (by the definition of a maximum); therefore, $2^{\max T}$ is an addend of the sum $\sum_{t \in T} 2^t$. Since all addends of this sum are nonnegative integers, we thus conclude that

$$\sum_{t \in T} 2^t \geq 2^{\max T} \tag{255}$$

(since a sum of nonnegative integers is always $\geq$ to any addend of this sum). On the other hand, each element s of the set T' satisfies $s \leq \max(T')$ (by the definition of $\max(T')$) and therefore $s \leq \max(T') < \max T$ (since $\max T > \max(T')$) and therefore $s \in \{0, 1, \ldots, \max T - 1\}$ (since $s \in T' \subseteq \mathbb{N}$). In other words,

[136] *Proof.* Assume that $m = 0$. Then, $0 = m = \sum_{t \in T} 2^t$. Hence, $\sum_{t \in T} 2^t = 0$. In other words, the sum $\sum_{t \in T} 2^t$ is 0. But all addends of the sum $\sum_{t \in T} 2^t$ are positive. Thus, if this sum was nonempty, then it would be positive (because a nonempty sum of positive reals is positive), which would contradict the fact that this sum is 0. Hence, this sum must be empty. In other words, $T = \varnothing$. The same argument (applied to T' instead of T) yields $T' = \varnothing$. Hence, $T = \varnothing = T'$. Thus, we have proved that $T = T'$ under the assumption that $m = 0$.

$T' \subseteq \{0, 1, \ldots, \max T - 1\}$. Hence, the sum $\sum\limits_{t \in T'} 2^t$ is a subsum[137] of the sum $\sum\limits_{t \in \{0,1,\ldots,\max T - 1\}} 2^t$. Since all addends of the latter sum are nonnegative, we thus conclude that $\sum\limits_{t \in T'} 2^t \leq \sum\limits_{t \in \{0,1,\ldots,\max T - 1\}} 2^t$ (since a subsum of a sum of nonnegative reals is always $\leq$ to the whole sum). Hence,

$$\begin{aligned} m = \sum_{t \in T'} 2^t &\leq \sum_{t \in \{0,1,\ldots,\max T - 1\}} 2^t = 2^0 + 2^1 + \cdots + 2^{\max T - 1} \\ &= 2^{\max T} - 1 \qquad \text{(by (5), applied to } n = \max T) \\ &< 2^{\max T}. \end{aligned}$$

This contradicts

$$m = \sum_{t \in T} 2^t \geq 2^{\max T} \qquad \text{(by (255))}.$$

This contradiction shows that our assumption was wrong. Hence, $\max T \leq \max (T')$ is proved. The same argument (with the roles of T and T' interchanged) yields $\max (T') \leq \max T$. Combining these two inequalities, we obtain $\max T = \max (T')$. Set $g = \max T$; thus, $g = \max T \in T$ and $g = \max T = \max (T') \in T'$.

Define an integer $n \in \mathbb{N}$ by $n = \sum\limits_{t \in T \setminus \{g\}} 2^t$. Then,

$$\begin{aligned} m = \sum_{t \in T} 2^t &= 2^g + \underbrace{\sum_{t \in T \setminus \{g\}} 2^t}_{=n} \qquad \text{(since } g \in T) \\ &= 2^g + n, \end{aligned}$$

so that $n = m - 2^g < m$ (since $2^g > 0$). Also, from $n = m - 2^g$, we obtain

$$2^g + n = m = \sum_{t \in T'} 2^t = 2^g + \sum_{t \in T' \setminus \{g\}} 2^t \qquad \left(\text{since } g \in T'\right).$$

Subtracting 2^g from both sides of this equality, we find $n = \sum\limits_{t \in T' \setminus \{g\}} 2^t$.

Now, we know that $n \in \mathbb{N}$ satisfies $n < m$. Furthermore, $T \setminus \{g\}$ and $T' \setminus \{g\}$ are two finite subsets of $\mathbb{N}$ satisfying $n = \sum\limits_{t \in T \setminus \{g\}} 2^t$ and $n = \sum\limits_{t \in T' \setminus \{g\}} 2^t$. Hence, Claim 1 (applied to $T \setminus \{g\}$ and $T' \setminus \{g\}$ instead of T and T') yields $T \setminus \{g\} = T' \setminus \{g\}$. But $g \in T$ and thus

$$T = \underbrace{(T \setminus \{g\})}_{=T' \setminus \{g\}} \cup \{g\} = \left(T' \setminus \{g\}\right) \cup \{g\} = T' \qquad \left(\text{since } g \in T'\right).$$

[137] A *subsum* of a sum $\sum\limits_{s \in S} a_s$ means a sum of the form $\sum\limits_{s \in S'} a_s$, where S' is a subset of S.

Forget that we fixed T and T'. We thus have shown that if T and T' are two finite subsets of $\mathbb{N}$ such that $m = \sum_{t \in T} 2^t$ and $m = \sum_{t \in T'} 2^t$, then $T = T'$. In other words, Proposition 5.2.4 holds for $n = m$. This completes the induction step; thus, Proposition 5.2.4 is proved. □

Proof of Theorem 5.2.1. Proposition 5.2.3 shows that there exists a finite subset T of $\mathbb{N}$ such that $n = \sum_{t \in T} 2^t$. Proposition 5.2.4 then shows that such a T is unique. Theorem 5.2.1 is thus proved. □

The reader will have an opportunity to make a similar argument in Exercise 5.4.6 below.

5.2.2. Students in a lecture

Our next application of the extremal principle is [Grinbe08, Exercise 3.11] (slightly generalized):

Exercise 5.2.1. Let n be a positive integer. A lecture is attended by n students. Each student enters the classroom once and leaves it once (and does not come back). We know that among any three (distinct) students, there are at least two that are together in the room at some moment. The lecturer wants to make an announcement that every student will hear. Prove that the lecturer can pick two moments at which to make the announcement so that each student will hear it. (We assume that the announcement takes no time – i.e., if a student leaves at the same moment that another student enters, the lecturer can make the announcement at this moment and both students will hear it.)

Before we solve this exercise, let us formalize it mathematically. We do so by labelling the n students $1, 2, \ldots, n$, and by encoding the times that student i is present in the classroom as a closed interval I_i on the real axis. We denote the two moments at which the announcement is made by a and b. Thus, Exercise 5.2.1 takes the following shape:

Exercise 5.2.2. Let n be a positive integer. Let $I_1, I_2, \ldots, I_n$ be n nonempty finite closed intervals on the real axis. Assume that for any three distinct elements $i, j, k \in \{1, 2, \ldots, n\}$, at least two of the three intervals I_i, I_j, I_k intersect[138]. Prove that there exist two reals a and b such that each of the intervals $I_1, I_2, \ldots, I_n$ contains at least one of a and b.

[138] Two sets U and V are said to *intersect* if and only if $U \cap V \neq \varnothing$. Note that two intervals $[a, b]$ and $[b, c]$ always intersect, even though the intersection is a singleton set.

Solution to Exercise 5.2.2. Write each interval I_m as $[a_m, b_m]$ for two reals a_m and b_m. Let

$$a = \max\{a_m \mid m \in \{1,2,\ldots,n\}\} \qquad \text{and}$$
$$b = \min\{b_m \mid m \in \{1,2,\ldots,n\}\}.$$

[139] (Thus, in the language of Exercise 5.2.1, the number a is the moment at which the last student to enter enters; likewise, b is the moment at which the first student to leave leaves.)

We shall show that a and b have the required property – i.e., that each of the intervals $I_1, I_2, \ldots, I_n$ contains at least one of a and b.

Indeed, assume the contrary. Thus, there is a $p \in \{1,2,\ldots,n\}$ such that the interval I_p contains neither a nor b. Consider this p.

The interval I_p contains neither a nor b. That is, we have $a \notin I_p$ and $b \notin I_p$. Recall that $I_p = [a_p, b_p]$ (since $I_m = [a_m, b_m]$ for each $m \in \{1,2,\ldots,n\}$). Thus, $a_p \leq b_p$ (since the interval I_p is nonempty).

But we have $a = \max\{a_m \mid m \in \{1,2,\ldots,n\}\}$ and therefore $a \geq a_m$ for each $m \in \{1,2,\ldots,n\}$ (by the definition of a maximum). Applying this to $m = p$, we obtain $a \geq a_p$. Likewise, from $b = \min\{b_m \mid m \in \{1,2,\ldots,n\}\}$, we obtain $b \leq b_p$.

We have $a \geq a_p$. If we had $a \leq b_p$, then we would thus have $a \in [a_p, b_p] = I_p$, which would contradict $a \notin I_p$. Hence, we cannot have $a \leq b_p$. In other words, we have $a > b_p$.

We have $b \leq b_p$. If we had $b \geq a_p$, then we would thus have $b \in [a_p, b_p] = I_p$, which would contradict $b \notin I_p$. Hence, we cannot have $b \geq a_p$. In other words, we have $b < a_p$.

However, we also have $a = \max\{a_m \mid m \in \{1,2,\ldots,n\}\} \in \{a_m \mid m \in \{1,2,\ldots,n\}\}$ (since the maximum of a set is always an element of that set). In other words, there exists some $m \in \{1,2,\ldots,n\}$ such that $a = a_m$. Consider this m, and denote it by u. Thus, $u \in \{1,2,\ldots,n\}$ such that $a = a_u$. Similarly, we can find a $v \in \{1,2,\ldots,n\}$ such that $b = b_v$.

Recall that $I_m = [a_m, b_m]$ for each $m \in \{1,2,\ldots,n\}$. Hence, $I_u = [a_u, b_u]$ and $I_v = [a_v, b_v]$.

Now, $b_u = b < a_p$. Hence, the intervals $[a_u, b_u]$ and $[a_p, b_p]$ do not intersect (since the interval $[a_u, b_u]$ ends at b_u, whereas the interval $[a_p, b_p]$ begins at a_p).

Also, $a > b_p$, so that $b_p < a = a_v$. Hence, the intervals $[a_p, b_p]$ and $[a_v, b_v]$ do not intersect (since the interval $[a_p, b_p]$ ends at b_p, whereas the interval $[a_v, b_v]$ begins at a_v).

Finally, $b_u < a_p \leq b_p < a_v$. Hence, the intervals $[a_u, b_u]$ and $[a_v, b_v]$ do not intersect (since the interval $[a_u, b_u]$ ends at b_u, whereas the interval $[a_v, b_v]$ begins at a_v).

Combining the results of the previous three paragraphs, we conclude that no two of the three intervals $[a_u, b_u], [a_p, b_p], [a_v, b_v]$ intersect. In other words, no two of the three intervals I_u, I_p, I_v intersect (since $I_u = [a_u, b_u]$ and $I_p = [a_p, b_p]$ and

[139] These maximum and minimum exist because of Theorem 5.1.1.

$I_v = [a_v, b_v]$). This shows that u, p and v are distinct (since otherwise, two of the three intervals I_u, I_p, I_v would be equal, and thus would intersect[140]).

But we assumed that for any three distinct elements $i, j, k \in \{1, 2, \ldots, n\}$, at least two of the three intervals I_i, I_j, I_k intersect. Applying this to $i = u$, $j = p$ and $k = v$, we conclude that at least two of the three intervals I_u, I_p, I_v intersect. But this contradicts the fact that no two of the three intervals I_u, I_p, I_v intersect. This contradiction shows that our assumption was false. Hence, we have shown that each of the intervals $I_1, I_2, \ldots, I_n$ contains at least one of a and b. This solves Exercise 5.2.2. □

Does this mean we have solved Exercise 5.2.1? It looks like it does, but there is a minor wrinkle. Exercise 5.2.1 does not just claim that there exist two moments such that each student is present in the classroom at least at one of them. It also claims that the lecturer can pick these two moments. This is a slightly stronger claim, as the lecturer might not know in advance when the students will enter or leave the classroom until they have done so (and by then, it might be too late to make the announcements). Thus, the lecturer needs to work with incomplete information (or at least Exercise 5.2.1 allows for such an interpretation).

To some extent, our solution to Exercise 5.2.2 above works even with this incomplete information. Indeed, the numbers a and b we defined can be computed "on the fly" if the lecturer knows how many students she has (i.e., the number n) and pays attention to their comings and goings. The number a is the moment at which the last student enters the classroom (so the head count is complete), and the number b is the moment at which the first student leaves the classroom. At both of these moments, the lecturer (if she keeps a head count and pays attention to students leaving) can immediately make the announcement.

But what if the lecturer does not know n? Without knowing how many students there are in total, how to tell when the last one has entered? The way Exercise 5.2.1 is stated, it is not clear whether this is a situation to be considered[141], but we can choose to do so anyway. Can the lecturer find a and b independently of the number of students? Yes – but this needs a different choice of a and b. See Exercise 5.4.7 below for this. See also Exercise 5.4.8 for a generalization of Exercise 5.2.2.

5.2.3. Matching n points to n points with no intersection

The next exercise comes from the realm of combinatorial geometry (see, e.g., [Engel98, Chapter 3, Example E4] or [Grinbe08, Aufgabe 3.8] for somewhat weaker versions of it):

Exercise 5.2.3. Let $n \in \mathbb{N}$. Let $F_1, F_2, \ldots, F_n$ be n distinct points in the plane. Let $W_1, W_2, \ldots, W_n$ be n further distinct points in the plane. Assume that no four

[140] since these intervals are nonempty

[141] The original wording in [Grinbe08, Exercise 3.11] uses $n = 100$, so this situation needs not be addressed.

of the $2n$ points $F_1, F_2, \ldots, F_n, W_1, W_2, \ldots, W_n$ are collinear (i.e., lie on a common line). Let $[n] = \{1, 2, \ldots, n\}$. Prove that there exists a bijection $\sigma : [n] \to [n]$ such that no two of the n segments[142] $F_1 W_{\sigma(1)}, F_2 W_{\sigma(2)}, \ldots, F_n W_{\sigma(n)}$ intersect.

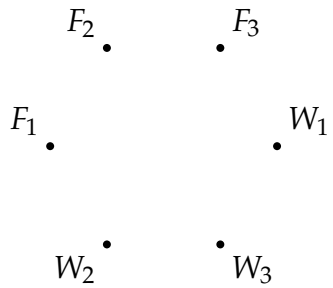
Exercise 5.2.3 is commonly worded as a road planning problem: The points $F_1, F_2, \ldots, F_n$ are the positions of n farms, while the points $W_1, W_2, \ldots, W_n$ are the positions of n wells. The exercise then claims that there is a way to connect each farm to a well by a straight-line road in such a way that no two roads intersect. (It is understood that different farms should be connected to different wells.)

We note that Exercise 5.2.3 would still hold if we replaced "plane" by "space" or even "d-dimensional space" for any $d \geq 1$; we have just stated it on the plane for reasons of familiarity. (Even the solution we will give below would still apply in d-dimensional space for any $d \geq 1$.)

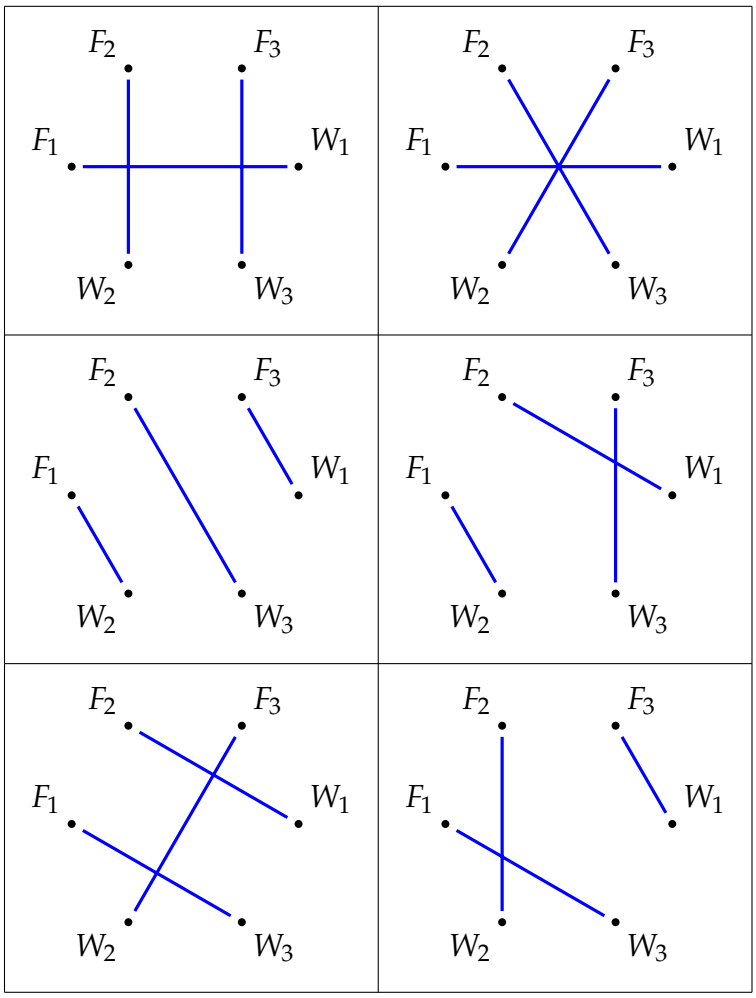
Before we solve Exercise 5.2.3, let us illustrate it on a small example.

[142] The word "segment" mens "line segment". Two sets U and V are said to *intersect* if and only if $U \cap V \neq \varnothing$. Thus, two line segments intersect even if they just have an endpoint in common (or an endpoint of one lies on the other).

Example 5.2.5. Let $n = 3$. Let $F_1, F_2, F_3, W_1, W_2, W_3$ be the following six points on the plane:



(Looks familiar?) There are six bijections $\sigma : [n] \to [n]$; let us draw the n segments $F_1W_{\sigma(1)}, F_2W_{\sigma(2)}, \ldots, F_nW_{\sigma(n)}$ for each of them:



The third bijection (i.e., the first one in the second row of the table) has the property required in the problem (viz., that no two of the n segments $F_1W_{\sigma(1)}, F_2W_{\sigma(2)}, \ldots, F_nW_{\sigma(n)}$ intersect).

Solution to Exercise 5.2.3 (sketched). For any bijection $\sigma : [n] \to [n]$, we define the *road length* $r(\sigma)$ by

$$r(\sigma) = \left|F_1 W_{\sigma(1)}\right| + \left|F_2 W_{\sigma(2)}\right| + \cdots + \left|F_n W_{\sigma(n)}\right| \in \mathbb{R}.$$

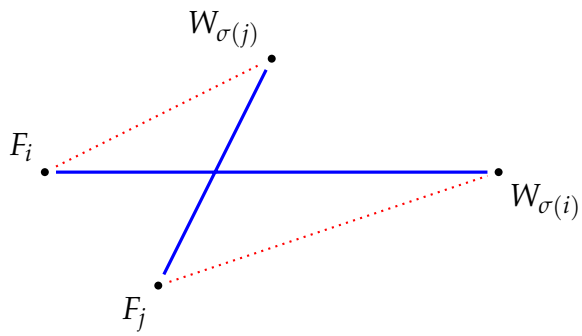
(Visually speaking, this is the total length of road that needs to be paved if we choose to connect the farms $F_1, F_2, \ldots, F_n$ to the wells $W_{\sigma(1)}, W_{\sigma(2)}, \ldots, W_{\sigma(n)}$, respectively.)

The set $\{\text{bijections } \sigma : [n] \to [n]\}$ is clearly nonempty and finite[143]. Thus, the set $\{r(\sigma) \mid \sigma : [n] \to [n] \text{ is a bijection}\}$ also is nonempty and finite. Hence, this set has a minimum element[144]. In other words, there exists a bijection $\sigma : [n] \to [n]$ for which $r(\sigma)$ is minimum (among all such bijections). Consider this σ. (If there are several σ that all give the same minimum value of $r(\sigma)$, then we just pick any of them.) Note that any bijection $\sigma' : [n] \to [n]$ satisfies

$$r(\sigma) \leq r(\sigma') \tag{256}$$

(since σ was chosen in such a way that $r(\sigma)$ is minimum). (Intuitively, the bijection σ is an assignment of wells to farms that minimizes the total road length.)

We shall now show that no two of the n segments $F_1 W_{\sigma(1)}, F_2 W_{\sigma(2)}, \ldots, F_n W_{\sigma(n)}$ intersect. Indeed, assume the contrary. Thus, there exist two distinct elements $i, j \in [n]$ such that the two segments $F_i W_{\sigma(i)}$ and $F_j W_{\sigma(j)}$ intersect. Consider these i, j. The segments $F_i W_{\sigma(i)}$ and $F_j W_{\sigma(j)}$ intersect. The situation is illustrated in the following picture (which also suggests the next step):



We have $i \neq j$ (since i, j are distinct) and thus $\sigma(i) \neq \sigma(j)$ (since σ is a bijection). Recall that no four of the $2n$ points $F_1, F_2, \ldots, F_n, W_1, W_2, \ldots, W_n$ are collinear. Hence, the four points $F_i, F_j, W_{\sigma(i)}, W_{\sigma(j)}$ are not collinear (since $i \neq j$ and $\sigma(i) \neq \sigma(j)$). Moreover, from $i \neq j$, we obtain $F_i \neq F_j$ (since the n points $F_1, F_2, \ldots, F_n$

[143] Indeed, anyone familiar with basic combinatorics (or abstract algebra) will recognize this set as the *n-th symmetric group* S_n, and will know that it has size $n!$. (See, e.g., Theorem 7.4.1 or [19fco, Theorem 1.7.2] for the proof of the latter fact.) But even if we don't know this, we can easily see that this set is nonempty and finite. Indeed, it is nonempty (since the identity map $\mathrm{id}_{[n]} : [n] \to [n]$ belongs to this set) and finite (since it is a subset of the finite set $\{\text{maps } [n] \to [n]\}$).

[144] by Theorem 5.1.1

are distinct). Also, from $\sigma(i) \neq \sigma(j)$, we obtain $W_{\sigma(i)} \neq W_{\sigma(j)}$ (since the n points $W_1, W_2, \ldots, W_n$ are distinct).

Now, we claim that

$$\left|F_i W_{\sigma(j)}\right| + \left|F_j W_{\sigma(i)}\right| < \left|F_i W_{\sigma(i)}\right| + \left|F_j W_{\sigma(j)}\right|. \tag{257}$$

This appears rather obvious from a look at the above picture, but a rigorous proof takes some work. We thus outsource this work into a lemma (Lemma 5.2.6 below). The inequality (257) follows from Lemma 5.2.6, applied to $X = F_i$, $Y = W_{\sigma(i)}$, $Z = F_j$ and $W = W_{\sigma(j)}$ (since $F_i \neq F_j$ and $W_{\sigma(i)} \neq W_{\sigma(j)}$, and since the points X, Z, Y, W are not collinear, and since the segments $F_i W_{\sigma(i)}$ and $F_j W_{\sigma(j)}$ intersect).

Now, what do we gain from the inequality (257)? Intuitively, it is telling us that if we replace the roads $F_i W_{\sigma(i)}$ and $F_j W_{\sigma(j)}$ by $F_i W_{\sigma(j)}$ and $F_j W_{\sigma(i)}$ (that is, if we reconnect the wells $W_{\sigma(i)}$ and $W_{\sigma(j)}$, which we originally connected to the farms F_j and F_i, to the farms F_j and F_i instead), then the total road length decreases. This observation is helpful, since it contradicts the fact that our original road connections were chosen to minimize the road length. So let us formalize this: We define a new map $\sigma' : [n] \to [n]$ by setting

$$\sigma'(k) = \begin{cases} \sigma(k), & \text{if } k \neq i \text{ and } k \neq j; \\ \sigma(j), & \text{if } k = i; \\ \sigma(i), & \text{if } k = j \end{cases} \qquad \text{for each } k \in [n].$$

That is, the map σ' is obtained from σ by swapping the values at i and j. It is clear that this map σ' is a bijection (since σ was a bijection, and we obtained σ' from σ by swapping two values). Consequently, (256) yields $r(\sigma) \leq r(\sigma')$. However, the definition of $r(\sigma)$ yields

$$\begin{aligned} r(\sigma) &= \left|F_1 W_{\sigma(1)}\right| + \left|F_2 W_{\sigma(2)}\right| + \cdots + \left|F_n W_{\sigma(n)}\right| = \sum_{k \in [n]} \left|F_k W_{\sigma(k)}\right| \\ &= \left|F_i W_{\sigma(i)}\right| + \left|F_j W_{\sigma(j)}\right| + \sum_{\substack{k \in [n]; \\ k \neq i \text{ and } k \neq j}} \left|F_k W_{\sigma(k)}\right| \end{aligned} \tag{258}$$

(here we have split off the addends for $k = i$ and for $k = j$ from the sum). The same

argument (applied to σ' instead of σ) yields

$$\begin{aligned} r(\sigma') &= \underbrace{\left|F_i W_{\sigma'(i)}\right|}_{\substack{=\left|F_i W_{\sigma(j)}\right| \\ \text{(since } \sigma'(i)=\sigma(j))}} + \underbrace{\left|F_j W_{\sigma'(j)}\right|}_{\substack{=\left|F_j W_{\sigma(i)}\right| \\ \text{(since } \sigma'(j)=\sigma(i))}} + \sum_{\substack{k\in[n]; \\ k\neq i \text{ and } k\neq j}} \underbrace{\left|F_k W_{\sigma'(k)}\right|}_{\substack{=\left|F_k W_{\sigma(k)}\right| \\ \text{(since } \sigma'(k)=\sigma(k) \\ \text{(because } k\neq i \text{ and } k\neq j))}} \\ &= \underbrace{\left|F_i W_{\sigma(j)}\right| + \left|F_j W_{\sigma(i)}\right|}_{\substack{<\left|F_i W_{\sigma(i)}\right|+\left|F_j W_{\sigma(j)}\right| \\ \text{(by (257))}}} + \sum_{\substack{k\in[n]; \\ k\neq i \text{ and } k\neq j}} \left|F_k W_{\sigma(k)}\right| \\ &< \left|F_i W_{\sigma(i)}\right| + \left|F_j W_{\sigma(j)}\right| + \sum_{\substack{k\in[n]; \\ k\neq i \text{ and } k\neq j}} \left|F_k W_{\sigma(k)}\right| = r(\sigma) \qquad \text{(by (258))}. \end{aligned}$$

This contradicts $r(\sigma) \leq r(\sigma')$. This contradiction shows that our assumption was wrong. Hence, no two of the n segments $F_1 W_{\sigma(1)}, F_2 W_{\sigma(2)}, \ldots, F_n W_{\sigma(n)}$ intersect. Thus, Exercise 5.2.3 is solved, except that we still owe a proof of the following lemma: □

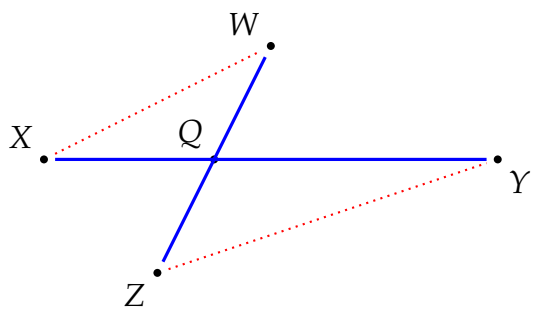
Lemma 5.2.6. Let X, Y, Z, W be four points in the plane. Assume that $X \neq Z$ and $Y \neq W$. Assume furthermore that

$$\text{the points } X, Z, Y, W \text{ are not collinear.} \tag{259}$$

Assume that the segments XY and ZW intersect. Then,

$$|XW| + |ZY| < |XY| + |ZW|.$$

Proof of Lemma 5.2.6. The segments XY and ZW intersect, i.e., have some point Q in common. Consider this point Q. We illustrate this situation with a picture, albeit we don't harbor any illusions about its generality:



.

Recall the triangle inequality, which says that any triangle ABC satisfies $|AC| < |AB| + |BC|$. This inequality holds even when the triangle ABC is degenerate (i.e.,

when the three points A, B, C are collinear), as long as we replace the "$<$" sign by a "$\leq$" sign. In other words, any three points A, B, C in the plane satisfy

$$|AC| \leq |AB| + |BC| . \tag{260}$$

Applying this to $A = X$, $B = Q$ and $C = W$, we obtain

$$|XW| \leq |XQ| + |QW| . \tag{261}$$

The same argument (applied to Z and Y instead of X and W) yields

$$|ZY| \leq |ZQ| + |QY| . \tag{262}$$

However, the point Q lies on the segment XY, and thus satisfies $|XQ| + |QY| = |XY|$. The same argument (applied to Z and W instead of X and Y) yields $|ZQ| + |QW| = |ZW|$.

Now, adding the two inequalities (261) and (262) together, we find

$$\begin{aligned} |XW| + |ZY| &\leq (|XQ| + |QW|) + (|ZQ| + |QY|) \\ &= \underbrace{|XQ| + |QY|}_{=|XY|} + \underbrace{|ZQ| + |QW|}_{=|ZW|} \\ &= |XY| + |ZW| . \end{aligned} \tag{263}$$

This is almost the inequality $|XW| + |ZY| < |XY| + |ZW|$ that we need to prove. But only "almost". In fact, the inequality in (263) has a "$\leq$" sign, while we want to prove the same inequality with a "$<$" sign. In other words, we want to prove that the inequality (263) cannot be an equality.

We assume the contrary. That is, we assume that the inequality (263) is an equality. Thus, we have

$$\begin{aligned} |XW| + |ZY| &= |XY| + |ZW| \\ &= (|XQ| + |QW|) + (|ZQ| + |QY|) \end{aligned} \tag{264}$$

(since we have previously shown that $(|XQ| + |QW|) + (|ZQ| + |QY|) = |XY| + |ZW|$).

We notice the following basic fact: If a, b, c, d are four real numbers satisfying $a \leq b$ and $c \leq d$ and $a + c = b + d$, then $a = b$ and $c = d$ [145]. We can apply this to $a = |XW|$ and $b = |XQ| + |QW|$ and $c = |ZY|$ and $d = |ZQ| + |QY|$ (because the three relations (261), (262) and (264) are saying precisely that these four numbers a, b, c, d satisfy $a \leq b$ and $c \leq d$ and $a + c = b + d$). As a result, we conclude that

$$|XW| = |XQ| + |QW| \tag{265}$$

[145] *Proof.* Let a, b, c, d be four real numbers satisfying $a \leq b$ and $c \leq d$ and $a + c = b + d$. Then, $b + d = \underbrace{a}_{\leq b} + c \leq b + c$ and $b + \underbrace{c}_{\leq d} \leq b + d$. Combining these two inequalities, we find $b + d = b + c$. Hence, $d = c$, so that $c = d$. Now, $a + c = b + \underbrace{d}_{=c} = b + c$, so that $a = b$. Hence, we have proved that $a = b$ and $c = d$, qed.

and

$$|ZY| = |ZQ| + |QY| . \tag{266}$$

Now, we recall the triangle inequality in its original form. It says that if three points A, B, C form a nondegenerate triangle (i.e., are not collinear), then $|AC| < |AB| + |BC|$. Hence, the contrapositive holds: If three points A, B, C satisfy $|AC| = |AB| + |BC|$, then they are collinear. A bit of thought reveals an even stronger conclusion: If three points A, B, C satisfy $|AC| = |AB| + |BC|$, then B lies on the segment AC. Applying this to $A = X$, $B = Q$ and $C = W$, we conclude that Q lies on the segment XW (since we have $|XW| = |XQ| + |QW|$). The same argument (applied to Z and Y instead of X and W) yields that Q lies on the segment ZY. Hence, in total, we know that the point Q lies on the four segments XY, XW, ZY, ZW.

It is now tempting to conclude right away that this forces all four points X, Z, Y, W are collinear. In truth, this is not totally obvious yet, since some of the segments XY, XW, ZY, ZW may be degenerate, or the point Q might be an endpoint of some of them. Thus, we proceed more cautiously.

We claim that we have $Q = X$ or $Q = Y$. Indeed, assume the contrary. Thus, $Q \neq X$ and $Q \neq Y$. Thus, the lines QX and QY are well-defined. These two lines QX and QY are furthermore equal (since Q lies on the segment XY), so we can call them the "line QXY". This line QXY contains the points X and Y, thus also the point W (since Q lies on the segment XW) and the point Z (since Q lies on the segment ZY); thus, it contains all four points X, Z, Y, W. Hence, these four points X, Z, Y, W are collinear. But this contradicts (259). This contradiction shows that our assumption was false.

Hence, we have shown that $Q = X$ or $Q = Y$. Similarly, we can see that $Q = Z$ or $Q = W$. Hence, we are in one of the following two cases:

Case 1: We have $Q = Z$.

Case 2: We have $Q = W$.

Let us first consider Case 1. In this case, we have $Q = Z$. Hence, $Q = Z \neq X$ (since $X \neq Z$). But recall that we have $Q = X$ or $Q = Y$. Hence, we must have $Q = Y$ (since $Q \neq X$). But Q lies on the segment XW. In other words, Z and Y lie on the segment XW (since $Q = Z$ and $Q = Y$). This shows that the four points X, Z, Y, W are collinear. But this contradicts (259). Thus, we have found a contradiction in Case 1.

Let us next consider Case 2. In this case, we have $Q = W$. Hence, $Q = W \neq Y$ (since $Y \neq W$). But recall that we have $Q = X$ or $Q = Y$. Hence, we must have $Q = X$ (since $Q \neq Y$). But Q lies on the segment ZY. In other words, W and X lie on the segment ZY (since $Q = W$ and $Q = X$). This shows that the four points X, Z, Y, W are collinear. But this contradicts (259). Thus, we have found a contradiction in Case 2.

We have now found a contradiction in each of the Cases 1 and 2. Hence, we always obtain a contradiction. Thus, our assumption was wrong. This means we have shown that the inequality (263) cannot be an equality. Hence, the "$\leq$" sign in this inequality can be replaced by a "$<$" sign. In other words, we have $|XW| + |ZY| < |XY| + |ZW|$. This proves Lemma 5.2.6. $\square$

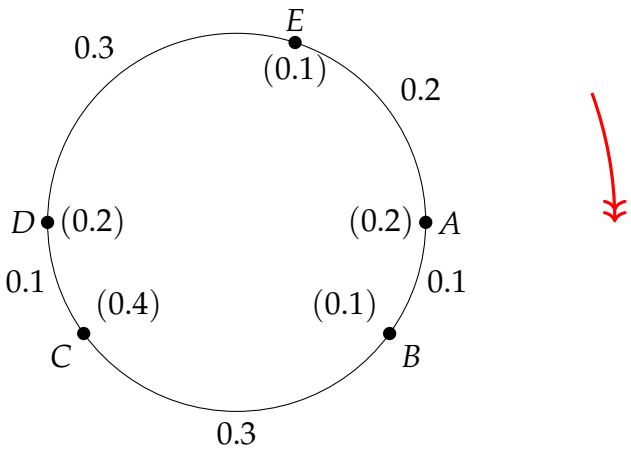
5.2.4. The round track puzzle

The next exercise is a classic (e.g., [Engel98, Exercise 3.15], [Engel98, Exercise 8.2], `https://math.stackexchange.com/questions/2338579/` just to mention three sources):

Exercise 5.2.4. Let n be a positive integer. Consider a circular track with n gas stations on it. Consider a car. Assume that, taken together, the n gas stations have just enough gas for the car to complete the entire track. Prove that at least one of these n gas stations has the property that if the car starts at this gas station with an initially empty gas tank, then it can traverse the entire track without ever running out of gas, provided that it refuels at every gas station it comes across (including the one at which it starts).

(It is understood that the car's tank is large enough to fit all the gas it can get. It is also understood that the track is one-way, so the car can only move in one direction. It is also understood that the use of gas is proportional to the length of road covered – i.e., independent of speed and acceleration.)

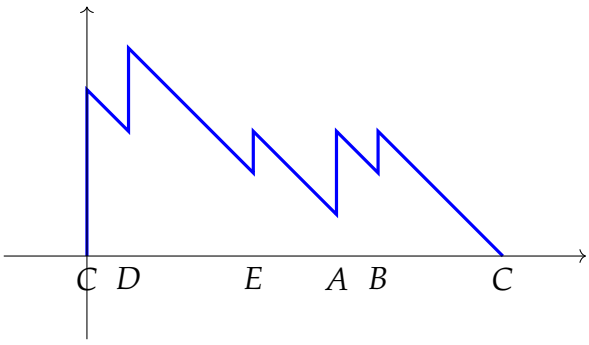
Example 5.2.7. The following picture shows an example for $n = 5$:



Here, the 5 gas stations are called A, B, C, D, E; the amount of gas available from each gas station is written in parentheses beside the station; and the length of road between two consecutive gas stations is written on the corresponding piece of track. The measures of gas and road length are normalized in such a way that the total length of the track is 1, and the total amount of gas in all stations is 1. We understand that the car can only go clockwise around the track.

In this example, we see that if a car with an empty tank were to start at A, then it would not be able to traverse the entire track. Indeed, it would first refuel at A; this would put 0.2 units of gas in its tank; it would then use 0.1 units of gas to get to B; then it would refuel at B; then it would have $0.2 - 0.1 + 0.1 = 0.2$ units of gas in its tank; but it would then run out of gas between B and C.

On the other hand, if a car with an empty tank were to start at C, then it would be able to traverse the entire track; here is a plot of the gas in its tank over the entire journey:



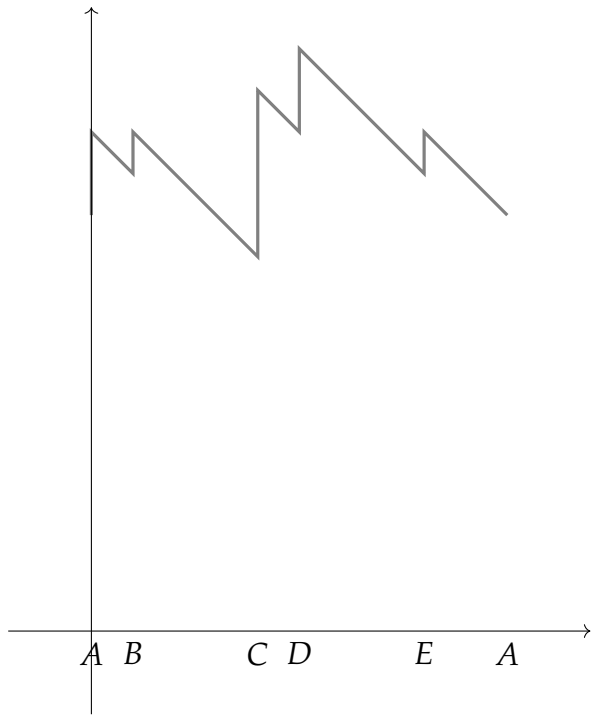
(with the horizontal axis standing for road length covered, while the vertical axis stands for the gas in the car's tank). (The labels C, D, E, A, B, C at the bottom mark the points at which the car is at a gas station.) Note that this is a "sawtooth function" plot, in which the vertical segments correspond to refuelling and the downward slopes correspond to gas being burned along the road.

Solution to Exercise 5.2.4 (sketched). Imagine a new car that starts somewhere on the track with enough gas to go through the entire track. We call this the *ghost car*. Now, let the ghost car go around the track[146], refueling at each gas station (even though it does not actually need this extra gas).

For example, in Example 5.2.7, if the ghost car starts at gas station A, then here

[146] We assume that the ghost car uses gas at the same rate as the original car.

is the plot of the gas in its tank:

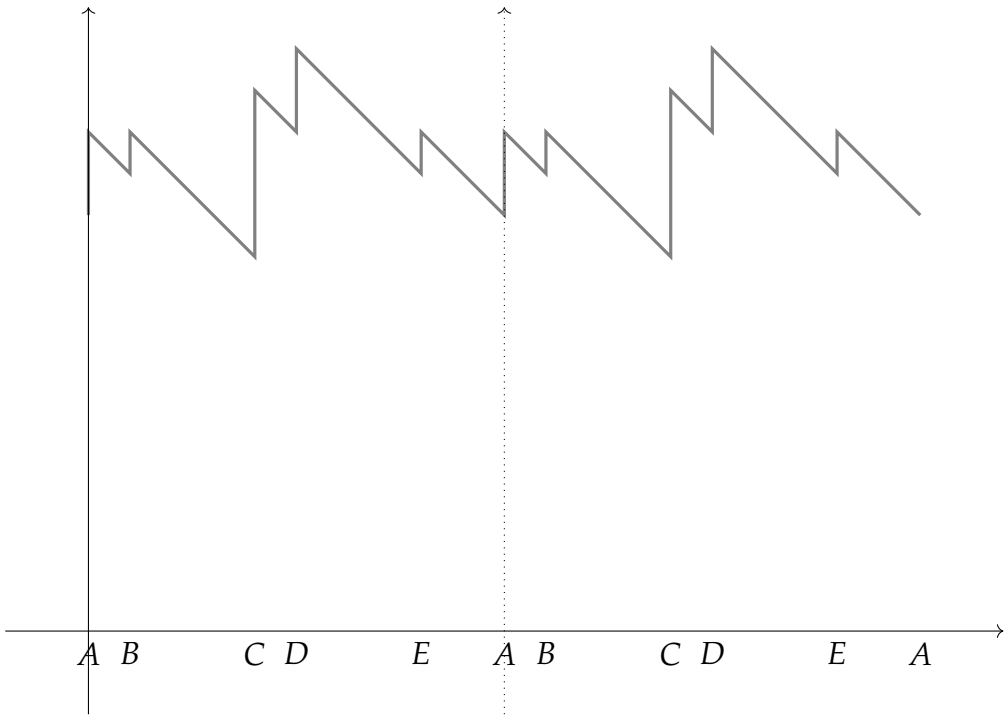


.

Clearly, the ghost car never runs out of gas (since it had enough even before refuelling); but there will be a point in its journey where its gas level is minimum. More precisely, there will be a gas station at which the ghost car's gas level (before refuelling) is minimum; let us pick such a gas station and call it P. Thus, the ghost car's gas level at P is $\leq$ to its gas level at any other gas station, and thus also $\leq$ to its gas level at any point on the road (because its gas level only declines between gas stations). In the example above, $P = C$.

After the ghost car has completed its loop around the track, it is back at the same station it started, and with the same amount of gas in its tank (since the total gas it has spent during the loop was precisely the amount it collected from the gas stations). At that point, we refill all gas stations (with the same amount of gas that they initially had) and we let the ghost car keep going around the circle (again collecting gas at the stations). Thus, the ghost car makes another loop. This second loop is an exact repetition of its previous loop (i.e., at each point, the ghost car has the same gas level on its second loop as it did on its first loop), because the ghost car started with the same gas level as initially and the gas stations had the same amount of gas as initially. Thus, if we plot the gas level in the car, the part of the plot that corresponds to the second loop is just a horizontally shifted copy of the

part corresponding to the first loop:

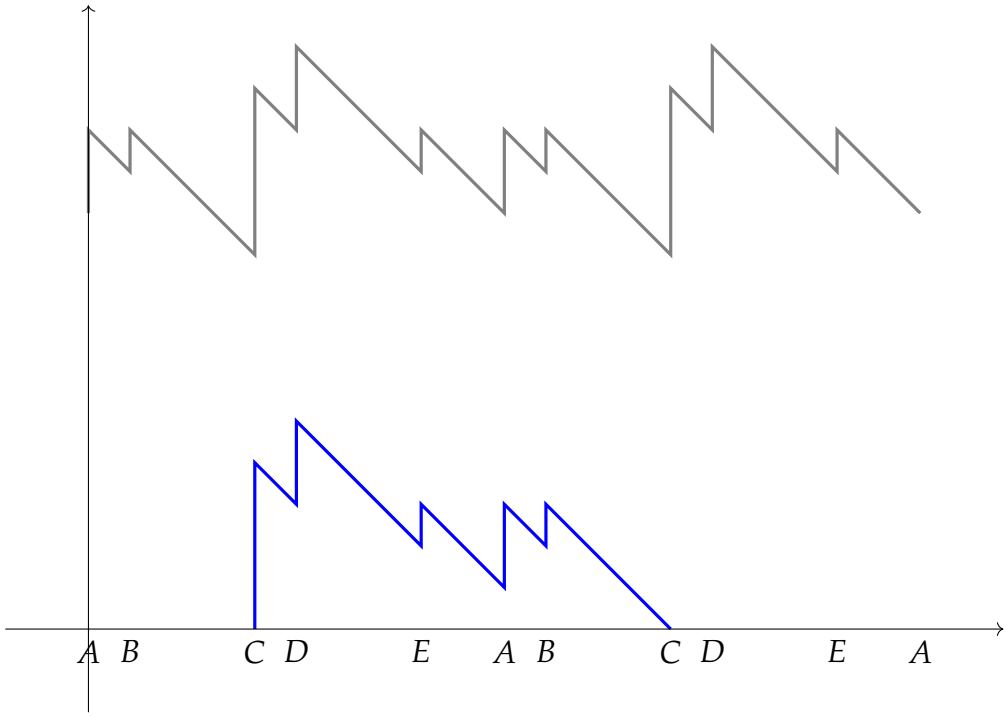


.

Thus, the minimum gas level (over both loops of the ghost car) is still achieved at station P. In other words, the ghost car never has less gas in its tank than it does at station P (before refuelling).

Now, let us start our original car (not the ghost car) from station P. We claim that it will complete the entire track without running out of gas. Indeed, the following picture shows (again in our above example) both the gas level of the ghost car (in

grey) and the gas level of the original car (in blue):



We see that the gas level plot of the original car is just a copy of a piece of the gas level plot of the ghost car, shifted downwards so it starts on the horizontal axis. Indeed, the original car makes the same journey as the ghost car from P to P, except that the original car starts with an empty tank whereas the ghost car starts with some gas.

But recall that the ghost car never has less gas in its tank than it does at station P. In other words, the gas level plot of the ghost car never goes below the level at station P. Therefore, the gas level plot of the original car never goes below the horizontal axis[147]. In other words, the original car (on its journey from P to P) never runs out of gas. This is precisely what we wanted to show. Thus, Exercise 5.2.4 is solved. □

5.2.5. n cowboys, n bullets

Here is another classic exercise (essentially [Engel98, Chapter 3, Exercise 7 (a)]), which illustrates how the Extremal Principle can be applied during an induction step:

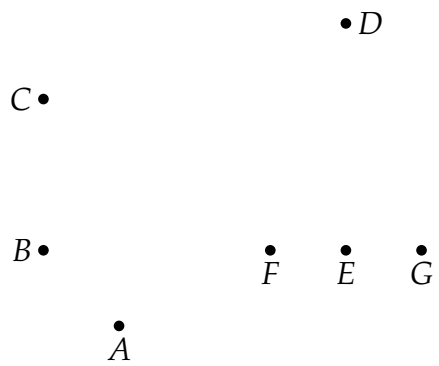
Exercise 5.2.5. Let $n \geq 1$ be an odd integer. Consider n cowboys, each standing at some point in space.

[147] since the gas level plot of the original car is just a copy of a piece of the gas level plot of the ghost car, shifted downwards so it starts on the horizontal axis

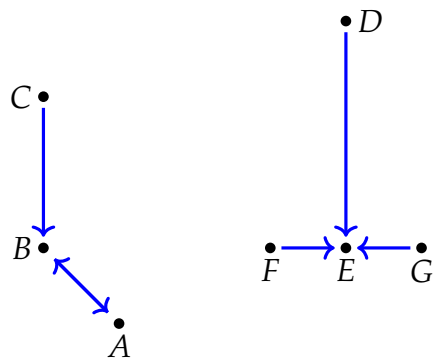
At high noon, each cowboy shoots his nearest neighbor, provided that he has a unique nearest neighbor. (If his nearest neighbor does not exist[148] or is not unique, then he shoots no one.)

Prove that at least one cowboy does not get shot.

Example 5.2.8. Let $n = 7$, and consider seven cowboys A, B, C, D, E, F, G placed in the plane as follows:



Then, the following picture shows the paths of the bullets:



(where an arrow from I to J means that I shoots J). (Note that cowboy E shoots no one, since he has two nearest neighbors F and G.) Thus, cowboys C, D, F and G do not get shot.

Solution to Exercise 5.2.5. We proceed by strong induction on n:

Induction step: Let $m \geq 1$ be an odd integer. Assume (as the induction hypothesis) that Exercise 5.2.5 holds for $n < m$. We must prove that Exercise 5.2.5 holds for $n = m$.

Thus, let $m \geq 1$ be an odd integer. Consider m cowboys, each standing at some point in space. At high noon, each cowboy shoots his nearest neighbor, provided

[148] Of course, this happens precisely when $n = 1$.

that he has a unique nearest neighbor. We must prove that at least one cowboy does not get shot.

Indeed, if $m = 1$, then there is only one cowboy, and he has no nearest neighbor and thus does not shoot anyone. Thus, if $m = 1$, then the lone cowboy does not got shot. Hence, we are done if $m = 1$. Thus, for the rest of this proof, we WLOG assume that $m \neq 1$. Hence, $m > 1$ (since $m \geq 1$), and thus $m \geq 3$ (since m is odd). Consequently, $m - 2 \geq 1$. Also, the integer $m - 2$ is odd (since m is odd).

Moreover, $m \geq 3 \geq 2$. Thus, there exists at least one pair (A, B) of two distinct cowboys A and B. Let us pick a pair (A, B) of two distinct cowboys A and B with **minimum** distance $|AB|$. Then, B is a nearest neighbor of A, so that A shoots either B or no one (depending on whether B is the **only** nearest neighbor of A). Likewise, B shoots either A or no one.

Now, let us consider the $m - 2$ cowboys distinct from A and B. We call these $m - 2$ cowboys *lucky* (because they might get to live a bit longer than A and B). The two cowboys A and B do not shoot any lucky cowboy (since A shoots either B or no one, and since B shoots either A or no one).

Now, we know that $m - 2 \geq 1$ is an odd integer satisfying $m - 2 < m$. Hence, our induction hypothesis yields that Exercise 5.2.5 holds for $n = m - 2$. Therefore, we can apply Exercise 5.2.5 to the $m - 2$ lucky cowboys. We thus conclude that if the two cowboys A and B did not exist (i.e., if the only cowboys around were the lucky ones), then at least one lucky cowboy would not get shot. In other words, if the two cowboys A and B did not exist, then there would be a lucky cowboy C that would not get shot. Consider this C.

Now, we claim that C does not get shot even in our original setup (i.e., even when A and B do exist). Indeed, assume the contrary. Thus, there is a cowboy D that shoots C. Then, D must be a lucky cowboy (since A and B do not shoot any lucky cowboy, and hence do not shoot C), and C must be his unique nearest neighbor (since any cowboy shoots his unique nearest neighbor). But this entails that D would shoot C even if A and B did not exist (because its unique nearest neighbor would not change when the two unrelated cowboys A and B are removed, unless this neighbor was one of A and B); but this is impossible, since we know that C would not get shot if A and B did not exist. This contradiction shows that our assumption was wrong.

Hence, we have showed that C does not get shot even in our original setup. Therefore, at least one cowboy does not get shot.

We have thus proved that Exercise 5.2.5 holds for $n = m$. This completes the induction step, so Exercise 5.2.5 is solved. □

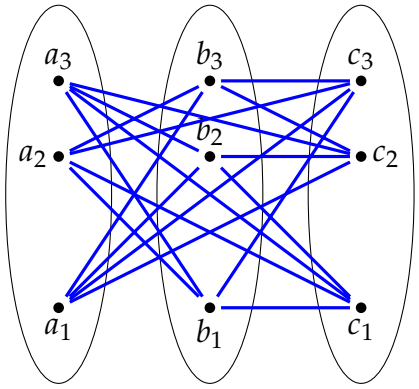
Another example of the Extremal Principle used in an induction is found in the solution of Exercise 3.7.9.

5.2.6. The three chess clubs problem

The following problem ([Engel98, Exercise 3.44], [Negut05, Chapter 3, problem 27]) requires a fairly subtle application of the extremal principle:

Exercise 5.2.6. Let n be a positive integer. Consider three chess clubs, each of which has n players. Assume that no two players in the same club have played against each other, but each player has played against at least $n+1$ other players. Prove that there exist three players a, b and c that have mutually played against each other (that is, a has played against b, and b has played against c, and c has played against a).

Example 5.2.9. Let us take $n = 3$, and consider the following constellation:



Here, the nodes are the players; the oval blobs are the three clubs; and the connecting lines show who has played against whom (viz., two players are connected by a line segment if and only if they have played against each other). We thus see that a_2, b_3 and c_3 have mutually played against each other.

Solution to Exercise 5.2.6. If a is any player, then we define the *score* of a to be the maximum number of players from a single club that a has played against. (For example, if A, B and C are the three clubs, and if a player a has played against 0 players from club A and against 6 players from club B and against 4 players from club C, then the score of a is 6.)

It is easy to see that the score of any player is positive[149]. (Actually, it is at least $\frac{n+1}{2}$, but we won't need this.)

The three clubs have $3n$ players altogether. Among these $3n$ players, let a be the player with maximum score. Let k be the score of a. Thus, a has played against k players from some club (by the definition of a score). Let B be this club, and let A

[149] *Proof.* Let a be a player. We must show that the score of a is positive.

We assumed that each player has played against at least $n+1$ other players. Hence, a has played against at least $n+1$ other players. Thus, a has played against at least one other player (since $n+1 \geq 1$). Hence, in particular, a has played against at least one player from a single club.

But the score of a was defined as the maximum number of players from a single club that a has played against. Hence, this score is at least 1 (since a has played against at least one player from a single club), and thus is positive. Hence we have shown that the score of a is positive. Qed.

be the club that contains a. It is clear that $A \neq B$ [150]. Thus, A and B are two of the three clubs. Let C be the remaining club.

Player a belongs to club A, and thus has played against no one from club A (since no two players in the same club have played against each other). Furthermore, player a has played against k players from club B. But we also know that player a has played against at least $n+1$ other players in total[151]. Hence, we can easily see that player a must have played against at least one player c from club C [152]. Consider this c.

Recall that a is a player with maximum score. Since the score of a is k, this entails that the maximum score of a player is k. Therefore, the score of c is at most k. Thus, player c has played against at most k players from club A (because otherwise, c would have played against more than k players from club A, and therefore the score of c would be greater than k). Therefore, player c must have played against at least $n+1-k$ players from club B [153].

Now, let X be the set of all players from club B that player a has played against. Then, $|X| = k$ (since a has played against k players from club B).

Also, let Z be the set of all players from club B that player c has played against. Then, $|Z| \geq n+1-k$ (since player c must have played against at least $n+1-k$

[150] *Proof.* Assume the contrary. Thus, $A = B$. Hence, B is the club that contains a (since A is the club that contains a). Thus, a has played against no players from club B (since no two players in the same club have played against each other). But we recall that player a has played against k players from club B. Comparing the previous two sentences, we conclude that $k = 0$. However, k is the score of A and therefore positive (since the score of any player is positive). This contradicts $k = 0$. This contradiction shows that our assumption was wrong. Qed.

[151] because each player has played against at least $n+1$ other players

[152] *Proof.* Assume the contrary. Thus, player a has played against no one from club C. We further know that a has played against no one from club A. Recall hat player a has played against at least $n+1$ other players in total. All these at least $n+1$ players must belong to club B (since player a has played against no one from club C and against no one from club A). Thus, club B must have at least $n+1$ players. But this contradicts the fact that club B has only n players (since any club has exactly n players). This contradiction shows that our assumption was wrong. Qed.

[153] *Proof.* Assume the contrary. Thus, player c has played against fewer than $n+1-k$ players from club B. In other words,

$$(\text{the number of players from club } B \text{ that } c \text{ has played against}) < n+1-k.$$

Moreover, player c has played against at most k players from club A (as we have already seen); in other words,

$$(\text{the number of players from club } A \text{ that } c \text{ has played against}) \leq k.$$

Finally, player c belongs to club C, and thus cannot have played against anyone from club C (because no two players in the same club have played against each other). Hence,

$$(\text{the number of players from club } C \text{ that } c \text{ has played against}) = 0.$$

Now, recall that each player has played against at least $n+1$ other players. Hence, c has played against at least $n+1$ other players. In other words,

$$(\text{the number of players that } c \text{ has played against}) \geq n+1.$$

players from club B).

Consider the club B as a set of players; thus, $|B| = n$ (since each club has exactly n players). The sets X and Z are subsets of B (since they consist of players from club B); therefore, their union $X \cup Z$ is a subset of B as well. Hence, $|X \cup Z| \leq |B| = n$.

If the sets X and Z were disjoint, then they would satisfy

$$\begin{aligned} |X \cup Z| &= \underbrace{|X|}_{=k} + \underbrace{|Z|}_{\geq n+1-k} \qquad \text{(by (27), applied to } P = X \text{ and } Q = Z) \\ &\geq k + (n+1-k) = n+1 > n, \end{aligned}$$

which would contradict $|X \cup Z| \leq n$. Thus, the sets X and Z cannot be disjoint. In other words, there exists some $b \in X \cap Z$.

Now, $b \in X \cap Z \subseteq X$; in other words, b is a player from club B that player a has played against (by the definition of X). Also, $b \in X \cap Z \subseteq Z$; in other words, b is a player from club B that player c has played against (by the definition of Z).

We now know that player a has played against player c, and we know that both players a and c have played against b. Therefore, the three players a, b and c have mutually played against each other. This shows that there exist three players that have mutually played against each other. This solves Exercise 5.2.6. □

It is worth highlighting a certain part of the above argument: namely, the part where we argued that the sets X and Z cannot be disjoint because they have too many elements among them (that is, $|X| + |Z| \geq n+1$). This is a relative of what is called the *pigeonhole principle*, which we shall discuss in the next chapter.

(A few years ago I unsuccessfully tried to generalize Exercise 5.2.6 to four chess clubs. See `https://mathoverflow.net/questions/247079` for details.)

5.3. Infinite descent

Infinite descent is a problem solving technique that is closely related to both induction and the extremal principle (and, in fact, can easily be restated as either of the two). We motivate it with an example that is perhaps too simple:

Thus,

$$\begin{aligned} n+1 &\leq (\text{the number of players that } c \text{ has played against}) \\ &= \underbrace{(\text{the number of players from club } A \text{ that } c \text{ has played against})}_{\leq k} \\ &\qquad + \underbrace{(\text{the number of players from club } B \text{ that } c \text{ has played against})}_{< n+1-k} \\ &\qquad + \underbrace{(\text{the number of players from club } C \text{ that } c \text{ has played against})}_{=0} \\ &\qquad\qquad (\text{since each player belongs to exactly one of the clubs } A, B \text{ and } C) \\ &< k + (n+1-k) + 0 = n+1. \end{aligned}$$

This is absurd. This contradiction shows that our assumption was wrong, qed.

Exercise 5.3.1. Prove that $\sqrt{3} \notin \mathbb{Q}$.

There are many simple ways to solve Exercise 5.3.1. In particular, Corollary 3.5.13 could be used to obtain an easy contradiction from an assumption of $\sqrt{3} \in \mathbb{Q}$. However, let us forget for a moment about Corollary 3.5.13, and discover a more complicated solution to Exercise 5.3.1 that illustrates how infinite descent works.

Discussion of Exercise 5.3.1. Assume the contrary. Thus, $\sqrt{3} \in \mathbb{Q}$. In other words, there exists a pair (a, b) of integers a and b satisfying $b \neq 0$ and $\sqrt{3} = a/b$. We shall call such a pair a *solution*.[154] Thus, by our assumption, there exists a solution.

Consider any solution (a, b). Thus, a and b are integers satisfying $b \neq 0$ and $\sqrt{3} = a/b$. Squaring the equality $\sqrt{3} = a/b$, we obtain $3 = (a/b)^2 = a^2/b^2$, so that $a^2 = 3b^2$.

Now, let us analyze the parity of b (that is, whether b is even or odd). Assume first that b is odd. Hence, Exercise 3.3.3 **(b)** (applied to $u = b$) yields $b^2 \equiv 1 \bmod 4$. Hence, $a^2 = 3 \underbrace{b^2}_{\equiv 1 \bmod 4} \equiv 3 \cdot 1 = 3 \bmod 4$. But if a was odd, then we would have $a^2 \equiv 1 \bmod 4$ (by Exercise 3.3.3 **(b)**, applied to $u = a$), which would entail $1 \equiv a^2 \equiv 3 \bmod 4$, which would in turn contradict $1 \not\equiv 3 \bmod 4$. Hence, a cannot be odd. Thus, a is even. Hence, $a^2 \equiv 0 \bmod 4$ (by Exercise 3.3.3 **(a)**, applied to $u = a$), so that $0 \equiv a^2 \equiv 3 \bmod 4$, which contradicts $0 \not\equiv 3 \bmod 4$. This contradiction shows that our assumption (that b is odd) was false. Hence, b is not odd. In other words, b is even.

Hence, $b^2 \equiv 0 \bmod 4$ (by Exercise 3.3.3 **(a)**, applied to $u = b$). If a was odd, then we would have $a^2 \equiv 1 \bmod 4$ (by Exercise 3.3.3 **(b)**, applied to $u = a$), which would entail $1 \equiv a^2 = 3 \underbrace{b^2}_{\equiv 0 \bmod 4} \equiv 3 \cdot 0 = 0 \bmod 4$, which would in turn contradict $1 \not\equiv 0 \bmod 4$. Thus, a cannot be odd. Hence, a must be even.

We have now shown that both a and b are even. In other words, $a/2$ and $b/2$ are integers. Furthermore, $b/2 \neq 0$ (since $b \neq 0$) and $\sqrt{3} = (a/2) / (b/2)$ (since $(a/2) / (b/2) = a/b = \sqrt{3}$). Therefore, the pair $(a/2, b/2)$ is a solution again (by the definition of a "solution").

Let us see what we have achieved. We have not obtained the contradiction that we were hoping for. However, proceeding from our original (assumed) solution (a, b), we have found a new solution $(a/2, b/2)$. This new solution is "smaller" than (a, b) in various regards: For example, $|b/2| < |b|$ (since $b \neq 0$). Let us define the *weight* of a solution (a, b) to be the nonnegative integer $|b|$. Thus, the new solution $(a/2, b/2)$ has smaller weight than (a, b) (since $b \neq 0$ and thus $|b/2| < |b|$). Hence, we have proved the following observation:

> *Observation 1:* If (a, b) is any solution, then there exists a solution of smaller weight than (a, b).

[154] The word "solution" here has been chosen in analogy to "solution of an equation", not to "solution of a problem". In particular, Exercise 5.3.1 will be solved once we have shown that there is no solution.

It turns out that this observation is almost as good as a contradiction. Here are three ways to complete the solution of Exercise 5.3.1 using Observation 1:

- *First way:* We have assumed that there exists a solution. Consider such a solution, and call it (a_0, b_0).

  Observation 1 (applied to $(a, b) = (a_0, b_0)$) thus yields that there exists a solution of smaller weight than (a_0, b_0). Consider such a solution, and call it (a_1, b_1).

  Observation 1 (applied to $(a, b) = (a_1, b_1)$) thus yields that there exists a solution of smaller weight than (a_1, b_1). Consider such a solution, and call it (a_2, b_2).

  Clearly, this argument (in which we apply Observation 1 to construct a new solution from an old one) can be applied over and over. Thus, we obtain an infinite sequence of solutions $(a_0, b_0), (a_1, b_1), (a_2, b_2), \ldots$, each of which has smaller weight than the previous one. Thus, the weights of these solutions form an infinite strictly decreasing sequence. In other words, if we denote the weight of the solution (a_i, b_i) by w_i, then the sequence $(w_0, w_1, w_2, \ldots)$ is strictly decreasing.

  However, the weight of a solution is a nonnegative integer. Thus, $w_0, w_1, w_2, \ldots$ are nonnegative integers; therefore, $(w_0, w_1, w_2, \ldots)$ is a strictly decreasing sequence of nonnegative integers. But there is no strictly decreasing sequence of nonnegative integers (in fact, if $(k_0, k_1, k_2, \ldots)$ is any strictly decreasing sequence of integers, then it is easy to see that k_i will be negative for any $i > k_0$). The preceding two sentences contradict each other; thus we have found the contradiction we wanted.

- *Second way:* We recall the Extremal Principle, and pick a solution (a, b) of smallest possible weight. This can be done, since the set of weights of all solutions is a nonempty set of nonnegative integers[155] and therefore has a minimum[156]. Now, Observation 1 yields that there exists a solution of smaller weight than (a, b). But this contradicts the fact that the solution (a, b) has the smallest possible weight. This gives us the contradiction we were looking for.

- *Third way:* Finally, we can also finish our argument by induction. Namely, we use strong induction to prove the following claim: For each $n \in \mathbb{N}$,

  $$\text{there exists no solution with weight } n. \tag{267}$$

  [*Proof of (267):* We proceed by strong induction on n:

  Induction step: Let $m \in \mathbb{N}$. Assume (as the induction hypothesis) that (267) holds for $n < m$. We must prove that (267) holds for $n = m$.

[155] It is nonempty because we have assumed that there exists a solution.

[156] by Theorem 5.1.2

Indeed, let (a, b) be a solution with weight m. Thus, Observation 1 yields that there exists a solution of smaller weight than (a, b). In other words, there exists a solution of weight $< m$ (since the weight of (a, b) is m). Let (a', b') be this latter solution, and let q be its weight. Thus, q is a nonnegative integer (since the weight of any solution is a nonnegative integer). That is, $q \in \mathbb{N}$. Also, $q < m$ (since q is the weight of the solution (a, b), which has weight $< m$). Our induction hypothesis says that (267) holds for $n < m$. Thus, (267) holds for $n = q$ (since $q < m$). In other words, there exists no solution with weight q. But this contradicts the fact that the solution (a', b') has weight q.

Forget that we fixed (a, b). We thus have found a contradiction for each solution (a, b) with weight m. Hence, there exists no solution with weight m. In other words, (267) holds for $n = m$. This completes the induction step. Thus, (267) is proved.]

Now, having proved (267), we easily see that there exists no solution whatsoever. (Indeed, if there was a solution, then it would have weight n for some $n \in \mathbb{N}$; but this would then contradict (267).) This contradicts our assumption that there is a solution. Thus, again, we have found a contradiction.

Thus, we have finished the solution of Exercise 5.3.1 in three different ways. It is fairly clear that these three ways formalize the same intuition: that a nonnegative integer cannot keep getting decreased indefinitely (without eventually falling below 0), and therefore a method to decrease the weight of a solution (no matter how small it originally was) gives a guarantee that no solutions can exist. This principle is known as the *principle of Infinite Descent*, and is often ascribed to Fermat even though it was known to the ancient Greeks. Note that our arguments (in all three ways) crucially rely on the fact that the weight of a solution was a nonnegative integer. If it was merely an integer or a nonnegative rational number, then we could not conclude much from Observation 1, since an infinite sequence of integers (or of nonnegative rational numbers) could decrease indefinitely. □

Here is a more sophisticated example for the Infinite Descent technique (Putnam 1973 problem B1):

Exercise 5.3.2. Let $n \in \mathbb{N}$. Assume that $a_1, a_2, \ldots, a_{2n+1}$ are $2n + 1$ integers with the following property:

> *Splitting property:* If any of the $2n + 1$ numbers $a_1, a_2, \ldots, a_{2n+1}$ is removed, then the remaining $2n$ numbers can be split into two equinumerous heaps with equal sum. ("Equinumerous" means that each heap contains exactly n numbers.)

Prove that all $2n + 1$ numbers $a_1, a_2, \ldots, a_{2n+1}$ are equal.

Example 5.3.1. Let $n = 1$, so that $2n + 1 = 3$. Then, in Exercise 5.3.2, we have three integers a_1, a_2, a_3. The splitting property boils down to requiring that $a_2 = a_3$ and $a_3 = a_1$ and $a_1 = a_2$ (because if one of the 3 numbers a_1, a_2, a_3 is removed, then the remaining 2 numbers can only be split into two equinumerous heaps in one way – up to the order of the heaps). Thus, for $n = 1$, Exercise 5.3.2 says that if $a_2 = a_3$ and $a_3 = a_1$ and $a_1 = a_2$, then all 3 numbers a_1, a_2, a_3 are equal. This is obvious.

Example 5.3.2. Let $n = 2$, so that $2n + 1 = 5$. Then, in Exercise 5.3.2, we have five integers a_1, a_2, a_3, a_4, a_5. The splitting property requires that if any of these five integers is removed, then the remaining four integers can be split into two equinumerous heaps with equal sum. For instance, if a_2 is removed, then the remaining four integers a_1, a_3, a_4, a_5 should satisfy $a_1 + a_3 = a_4 + a_5$ or $a_1 + a_4 = a_3 + a_5$ or $a_1 + a_5 = a_3 + a_4$. Again, Exercise 5.3.2 is saying that if the splitting property is satisfied (not just for a_2, but also for each of the other numbers), then all 5 numbers a_1, a_2, a_3, a_4, a_5 are equal. This time, there is no quick way to see this; there are too many ways in which the splitting property may be satisfied.

Example 5.3.3. Here is a non-example. Set $n = 2$ (so that $2n + 1 = 5$) and

$$a_1 = 1, \qquad a_2 = 2, \qquad a_3 = 3, \qquad a_4 = 4, \qquad a_5 = 5.$$

Then, the splitting property holds for removing a_1: Indeed, after removing a_1, we can split the remaining four numbers a_2, a_3, a_4, a_5 into the two equinumerous heaps a_2, a_5 and a_3, a_4 with equal sums ($a_2 + a_5 = a_3 + a_4$). Likewise, the splitting property holds for removing a_3 (since $a_1 + a_5 = a_2 + a_4$) and for removing a_5 (since $a_1 + a_4 = a_2 + a_3$), but not for removing a_2 or a_4.

Solution to Exercise 5.3.2 (sketched). Some notations first.

A *solution* will mean an $(2n + 1)$-tuple $(a_1, a_2, \ldots, a_{2n+1})$ of integers that has the splitting property.

A solution $(a_1, a_2, \ldots, a_{2n+1})$ will be called *flat* if all $2n + 1$ numbers $a_1, a_2, \ldots, a_{2n+1}$ are equal. Thus, our goal is to show that every solution is flat. In other words, our goal is to show that there is no non-flat solution.

A solution $(a_1, a_2, \ldots, a_{2n+1})$ will be called *nonnegative* if all $2n + 1$ numbers $a_1, a_2, \ldots, a_{2n+1}$ are nonnegative.

First, we observe the following:

> *Observation 1:* If there exists a non-flat solution, then there exists a nonnegative non-flat solution.

[*Proof of Observation 1:* Assume that there exists a non-flat solution $(a_1, a_2, \ldots, a_{2n+1})$. Set $b = \min\{a_1, a_2, \ldots, a_{2n+1}\}$. Then, all $2n + 1$ integers $a_1, a_2, \ldots, a_{2n+1}$ are $\geq b$; hence, all $2n + 1$ differences $a_1 - b, a_2 - b, \ldots, a_{2n+1} - b$ are nonnegative. Moreover, it is easy to see that the $(2n + 1)$-tuple $(a_1 - b, a_2 - b, \ldots, a_{2n+1} - b)$ is still

a solution (indeed, subtracting b from all $2n+1$ numbers does not invalidate the splitting property[157]) and still non-flat (since $a_1 - b = a_2 - b = \cdots = a_{2n+1} - b$ would imply $a_1 = a_2 = \cdots = a_{2n+1}$, which would contradict the assumption that $(a_1, a_2, \ldots, a_{2n+1})$ is non-flat). Thus, this $(2n+1)$-tuple $(a_1 - b, a_2 - b, \ldots, a_{2n+1} - b)$ is a nonnegative non-flat solution. Hence, there exists a nonnegative non-flat solution. This proves Observation 1.]

Thanks to Observation 1, we don't need to bother with negative integers if we don't want to. This will come useful later.

Next, let us study the parity of the integers in a solution:

> *Observation 2:* If $(a_1, a_2, \ldots, a_{2n+1})$ is a solution, then the $2n+1$ integers $a_1, a_2, \ldots, a_{2n+1}$ all have the same parity (i.e., are either all even or all odd).

[*Proof of Observation 2:* Let $(a_1, a_2, \ldots, a_{2n+1})$ be a solution. Let $t = a_1 + a_2 + \cdots + a_{2n+1}$ be the sum of all its entries.

The splitting property shows that if the integer a_{2n+1} is removed, then the remaining $2n$ numbers can be split into two equinumerous heaps with equal sum. In other words, the $2n$ numbers $a_1, a_2, \ldots, a_{2n}$ can be split into two equinumerous heaps with equal sum. Let s be the sum of either heap. Then, s is an integer (since $a_1, a_2, \ldots, a_{2n}$ are integers). But the sum $a_1 + a_2 + \cdots + a_{2n}$ of all the $2n$ numbers $a_1, a_2, \ldots, a_{2n}$ must be $s + s$ (since these $2n$ numbers can be split into two heaps, each of which has sum s). Thus,

$$a_1 + a_2 + \cdots + a_{2n} = s + s = 2s \equiv 0 \bmod 2 \qquad (\text{since } s \text{ is an integer}).$$

Hence,

$$t = a_1 + a_2 + \cdots + a_{2n+1} = \underbrace{(a_1 + a_2 + \cdots + a_{2n})}_{\equiv 0 \bmod 2} + a_{2n+1} \equiv a_{2n+1} \bmod 2,$$

so that $a_{2n+1} \equiv t \bmod 2$.

Thus, by removing a_{2n+1} and applying the splitting property, we have obtained $a_{2n+1} \equiv t \bmod 2$. But we can apply the same argument to any of our $2n+1$ integers $a_1, a_2, \ldots, a_{2n+1}$ in place of a_{2n+1}. Thus, we find that $a_i \equiv t \bmod 2$ for each $i \in \{1, 2, \ldots, 2n+1\}$. As a consequence,

- if t is even, then all of $a_1, a_2, \ldots, a_{2n+1}$ are even;
- if t is odd, then all of $a_1, a_2, \ldots, a_{2n+1}$ are odd.

[157] For example, in the case $2n+1 = 5$, if we had $a_2 + a_3 = a_1 + a_4$, then after subtraction of b we still have $(a_2 - b) + (a_3 - b) = (a_1 - b) + (a_4 - b)$. Here it is important that the two heaps in the splitting property were required to be equinumerous!

Thus, $a_1, a_2, \ldots, a_{2n+1}$ all have the same parity. This proves Observation 2.]

Now, Observation 2 helps us transform non-flat solutions into smaller non-flat solutions with an appropriate meaning of "smaller". To be more precise, we consider nonnegative solutions. If $(a_1, a_2, \ldots, a_{2n+1})$ is a nonnegative solution, then the *weight* of this solution is defined to be the nonnegative integer $a_1 + a_2 + \cdots + a_{2n+1}$. Now we claim:

> *Observation 3:* If $(a_1, a_2, \ldots, a_{2n+1})$ is a nonnegative non-flat solution, then there exists a nonnegative non-flat solution with smaller weight than $(a_1, a_2, \ldots, a_{2n+1})$.

[*Proof of Observation 3:* Let $(a_1, a_2, \ldots, a_{2n+1})$ be a nonnegative non-flat solution. Then, Observation 2 yields that $a_1, a_2, \ldots, a_{2n+1}$ all have the same parity. In other words, the numbers $a_1, a_2, \ldots, a_{2n+1}$ are either all even or all odd. Now, we construct a new nonnegative non-flat solution $(b_1, b_2, \ldots, b_{2n+1})$ as follows:

- If $a_1, a_2, \ldots, a_{2n+1}$ are all even, then we set
$$(b_1, b_2, \ldots, b_{2n+1}) = \left(\frac{a_1}{2}, \frac{a_2}{2}, \ldots, \frac{a_{2n+1}}{2}\right).$$
- If $a_1, a_2, \ldots, a_{2n+1}$ are all odd, then we set
$$(b_1, b_2, \ldots, b_{2n+1}) = \left(\frac{a_1 - 1}{2}, \frac{a_2 - 1}{2}, \ldots, \frac{a_{2n+1} - 1}{2}\right).$$

It is easy to check that (in either case) $(b_1, b_2, \ldots, b_{2n+1})$ is a nonnegative non-flat solution with smaller weight than $(a_1, a_2, \ldots, a_{2n+1})$.[158] Thus, there exists a nonnegative non-flat solution with smaller weight than $(a_1, a_2, \ldots, a_{2n+1})$. This proves Observation 3.]

Now, all we need is to reap our rewards. By the Principle of Infinite Descent, Observation 3 entails that there exists no nonnegative non-flat solution. Hence, by Observation 1, we conclude that there exists no non-flat solution either. In other words, any solution is flat. Exercise 5.3.2 is solved. □

There are some variants of this solution; in particular, it is possible to avoid Observation 1, at the expense of a more complicated notion of weight[159] and a more complicated proof of (the analogue of) Observation 3.

Can we replace the "integers" in Exercise 5.3.2 by some more general types of numbers? Let's first try rational numbers:

[158] Note that the weight of $(b_1, b_2, \ldots, b_{2n+1})$ is indeed **smaller** (and not just smaller-or-equal) than the weight of $(a_1, a_2, \ldots, a_{2n+1})$, because $a_1, a_2, \ldots, a_{2n+1}$ cannot all be 0 (since $(a_1, a_2, \ldots, a_{2n+1})$ is non-flat).

[159] Indeed, the way we defined the weight of a solution $(a_1, a_2, \ldots, a_{2n+1})$, it would not be nonnegative in general if we don't require $(a_1, a_2, \ldots, a_{2n+1})$ to be nonnegative. Thus, we would have to fix this – e.g., by redefining this weight as $|a_1| + |a_2| + \cdots + |a_{2n+1}|$ instead.

Exercise 5.3.3. Let $n \in \mathbb{N}$. Assume that $a_1, a_2, \ldots, a_{2n+1}$ are $2n+1$ **rational numbers** with the following property:

> *Splitting property:* If any of the $2n+1$ numbers $a_1, a_2, \ldots, a_{2n+1}$ is removed, then the remaining $2n$ numbers can be split into two equinumerous heaps with equal sum. ("Equinumerous" means that each heap contains exactly n numbers.)

Prove that all $2n+1$ numbers $a_1, a_2, \ldots, a_{2n+1}$ are equal.

Solution to Exercise 5.3.3 (sketched). The numbers $a_1, a_2, \ldots, a_{2n+1}$ are rational. Thus, there exists a positive integer d such that all the $2n+1$ products $da_1, da_2, \ldots, da_{2n+1}$ are integers. (Such a d is called a *common denominator* of $a_1, a_2, \ldots, a_{2n+1}$.) Consider such a d.

We assumed that the splitting property holds for our $2n+1$ numbers $a_1, a_2, \ldots, a_{2n+1}$; thus, it also holds for the $2n+1$ numbers $da_1, da_2, \ldots, da_{2n+1}$. (For example, in the case $2n+1 = 5$, if we had $a_2 + a_3 = a_1 + a_4$, then upon multiplication by d we get $da_2 + da_3 = da_1 + da_4$.) Hence, we can apply Exercise 5.3.2 to the $2n+1$ integers $da_1, da_2, \ldots, da_{2n+1}$ instead of $a_1, a_2, \ldots, a_{2n+1}$. Thus, we conclude that all $2n+1$ numbers $da_1, da_2, \ldots, da_{2n+1}$ are equal. Hence, all $2n+1$ numbers $a_1, a_2, \ldots, a_{2n+1}$ are equal as well (since $d \neq 0$). This solves Exercise 5.3.3. $\square$

Can we go further and replace the integers in Exercise 5.3.2 by real numbers? Yes, as the following shows:

Exercise 5.3.4. Let $n \in \mathbb{N}$. Assume that $a_1, a_2, \ldots, a_{2n+1}$ are $2n+1$ **real numbers** with the following property:

> *Splitting property:* If any of the $2n+1$ numbers $a_1, a_2, \ldots, a_{2n+1}$ is removed, then the remaining $2n$ numbers can be split into two equinumerous heaps with equal sum. ("Equinumerous" means that each heap contains exactly n numbers.)

Prove that all $2n+1$ numbers $a_1, a_2, \ldots, a_{2n+1}$ are equal.

We are not going to solve Exercise 5.3.4 now; but we may come back to it later when discussing the use of linear algebra. (Two solutions can be found in [GelAnd17, Problem 300].)

5.4. Homework set #5: More sequences and the extremal principle

This is a **regular** problem set. See Section 3.7 for details on grading.

This homework set covers the above parts of Chapter 4 and Chapter 5. Some of the problems may be unrelated.

Please solve **at most 5** problems. (No points will be given for further solutions.)

Exercise 5.4.1. Let $n \in \mathbb{N}$ and $a \in \mathbb{N}$. Prove that $an + 1 \mid \binom{(a+1)n}{n}$.

In the following exercise, we shall consider polynomials in a single variable x. The coefficients of these polynomials can be any kinds of numbers (rational, real or complex). Recall that a polynomial has degree $\leq m$ (for some integer m) if and only if it can be written as a sum $\sum_{i=0}^{m} c_i x^i$ with some constants $c_0, c_1, \ldots, c_m$. If $m < 0$, then the only polynomial that has degree $\leq m$ is the zero polynomial 0 (since the sum $\sum_{i=0}^{m} c_i x^i$ is an empty sum when $m < 0$).

If $P = P(x)$ is a polynomial, then ΔP shall denote the polynomial defined by

$$(\Delta P)(x) = P(x) - P(x-1).$$

For example, if $P(x) = x^3$, then $(\Delta P)(x) = P(x) - P(x-1) = x^3 - (x-1)^3 = 3x^2 - 3x + 1$.

If P is a polynomial, and if $n \in \mathbb{N}$, then $\Delta^n P$ shall denote the polynomial $\Delta(\Delta(\cdots(\Delta P)\cdots))$ with n copies of Δ in front of the P. (Formally speaking, this means that $\Delta^n P$ is defined recursively by setting $\Delta^0 P = P$ and $\Delta^n P = \Delta\left(\Delta^{n-1} P\right)$ for each $n \geq 1$.)

Exercise 5.4.2. Let $m \in \mathbb{Z}$. Let P be a polynomial in a single variable x. Assume that P has degree $\leq m$. Prove the following:

(a) The polynomial ΔP has degree $\leq m - 1$.

(b) The polynomial $\Delta^n P$ has degree $\leq m - n$ for each $n \in \mathbb{N}$.

(c) For each $n \in \mathbb{N}$, we have

$$(\Delta^n P)(x) = \sum_{k=0}^{n} (-1)^k \binom{n}{k} P(x-k).$$

(d) For each $n \in \mathbb{N}$ satisfying $n > m$, we have

$$\sum_{k=0}^{n} (-1)^k \binom{n}{k} P(x-k) = 0.$$

(e) Assume that $m \geq 0$. The sequence $(P(0), P(1), P(2), \ldots)$ is $(d_1, d_2, \ldots, d_{m+1})$-recurrent, where we set $d_i = (-1)^{i-1} \binom{m+1}{i}$ for each $i \in \{1, 2, \ldots, m+1\}$.

Exercise 5.4.3. Let $a, b, c \in \mathbb{N}$ be such that $c \leq b$ and $a \leq b$. Simplify

$$\sum_{k=c}^{b} \frac{\binom{a}{k}}{\binom{b}{k}}.$$

Exercise 5.4.4. Let $q \in \mathbb{R}$ and $s \in \mathbb{R}$. Define a sequence $(a_0, a_1, a_2, \ldots)$ of reals recursively by

$$\begin{aligned} a_0 &= s, \qquad \text{and} \\ a_n &= a_{n-1}\left(qa_{n-1} + 2\right) \qquad \text{for each integer } n \geq 1. \end{aligned}$$

Find an explicit formula for a_n.

Exercise 5.4.5. Define a sequence $(a_1, a_2, a_3, \ldots)$ of integers recursively by

$$\begin{aligned} &a_1 = 1, \qquad a_2 = 1, \qquad a_3 = 3, \qquad \text{and} \\ &a_n = \frac{a_{n-1}^2 - a_{n-2}^2}{a_{n-3}} \qquad \text{for each integer } n \geq 4. \end{aligned}$$

Compute a_n explicitly (in terms of sequences we already know).

The next exercise is a "Fibonacci analogue" of Theorem 5.2.1; it uses the notion of a lacunar subset (defined in Definition 2.3.3).

Exercise 5.4.6. Let $n \in \mathbb{N}$. Let $\mathbb{N}_2$ denote the set $\{2, 3, 4, \ldots\}$. Prove that there is a unique lacunar finite subset T of $\mathbb{N}_2$ such that $n = \sum\limits_{t \in T} f_t$.

(For example, $28 = f_3 + f_5 + f_8 = \sum\limits_{t \in \{3,5,8\}} f_t$.)

Exercise 5.4.7. Solve Exercise 5.2.1 if n is not known to the lecturer. That is, find a way to construct the moments a and b in Exercise 5.2.1 in such a way that the lecturer will know that these moments have arrived the very time they arrive (rather than only in hindsight).

The following exercise generalizes Exercise 5.2.2:

Exercise 5.4.8. Let n and k be positive integers with $k \geq 2$. Let $I_1, I_2, \ldots, I_n$ be n nonempty finite closed intervals on the real axis. Assume that for any k distinct elements $i_1, i_2, \ldots, i_k \in \{1, 2, \ldots, n\}$, at least two of the k intervals $I_{i_1}, I_{i_2}, \ldots, I_{i_k}$ intersect. Prove that there exist $k-1$ reals $a_1, a_2, \ldots, a_{k-1}$ such that each of the intervals $I_1, I_2, \ldots, I_n$ contains at least one of $a_1, a_2, \ldots, a_{k-1}$.

Exercise 5.4.9. Let n be a positive integer. Let $(x_0, x_1, x_2, \ldots)$ be an n-periodic sequence of reals such that $x_0 + x_1 + \cdots + x_{n-1} = 0$. Prove that there exists some $k \in \{0, 1, \ldots, n-1\}$ such that every $m \geq k$ satisfies $x_k + x_{k+1} + \cdots + x_m \geq 0$.

Exercise 5.4.10. A *golden pair* will mean a pair (x, y) of nonnegative integers such that $\left|x^2 - xy - y^2\right| = 1$. For example, $(3, 2)$ is a golden pair, since $\left|3^2 - 3 \cdot 2 - 2^2\right| = |-1| = 1$. Prove the following:

(a) If (x, y) is a golden pair such that $(x, y) \neq (0, 1)$, then $x - y \geq 0$.

(b) If (x, y) is a golden pair such that $(x, y) \neq (0, 1)$, then $(y, x - y)$ is a golden pair.

(c) If (x, y) is a golden pair such that $(x, y) \neq (1, 0)$, then $y > 0$.

(d) Find an explicit formula for all golden pairs different from $(0, 1)$.

[**Remark:** The word "golden" in "golden pair" refers to the resemblance of the equality $\left|x^2 - xy - y^2\right| = 1$ to the equation $\varphi^2 - \varphi - 1 = 0$ satisfied by the golden ratio $\varphi \approx 1.618$. Dividing the equation $\left|x^2 - xy - y^2\right| = 1$ by y^2 yields $\left|\left(\frac{x}{y}\right)^2 - \frac{x}{y} - 1\right| = \frac{1}{y^2}$, which is a way of saying that $\frac{x}{y}$ is a close rational approximation to φ.]

6. The Pigeonhole Principle

Next, we shall explore the *pigeonhole principle* (also known as the *box principle* or *Dirichlet's principle*) and its consequences. Most textbooks on problem solving devote a chapter or at least a section to it; examples are [Engel98, Chapter 4], [GelAnd17, §1.3], [Zeitz17, §3.3], [Grinbe08, Kapitel 2], [Carl17, Kapitel 3], [Galvin20, Week 2] and [Macgil17]. The pigeonhole principle is generally considered to be a part of combinatorics; thus, it tends to be explored in texts on combinatorics as well (e.g., [Bona17, Chapter 1], [Liu85, §4.8], [Brualdi09, Chapter 3]).

6.1. The principles

The words "*pigeonhole principle*" typically refer to one of several related theorems which we shall explore in this section. As usual, the theorems are simple; the art is to find a way to apply them to a given problem.

6.1.1. The Pigeonhole Principle for Injections

We begin with the best-known "pigeonhole principle". It is saying that if more than n pigeons are distributed into n pigeonholes (for a given $n \in \mathbb{N}$), then at least two pigeons must be in the same pigeonhole. This formulation is what gave this principle its common name[160]; however, we prefer a more formal way of stating it:[161]

Theorem 6.1.1. Let U and V be two finite sets such that $|U| > |V|$. Let $f : U \to V$ be any map. Then, f cannot be injective.

Indeed, our previous informal statement about pigeons is precisely Theorem 6.1.1 (applied to $U = \{\text{pigeons}\}$ and $V = \{\text{pigeonholes}\}$ and to f being the map that sends each pigeon to its pigeonhole).

Here is the contrapositive of Theorem 6.1.1:

Theorem 6.1.2. Let U and V be two finite sets. Let $f : U \to V$ be an injective map. Then, $|U| \leq |V|$.

In terms of pigeons and pigeonholes, this is saying that any injective distribution of pigeons into pigeonholes (where "injective" means "no two pigeons sharing a hole") requires at least as many pigeonholes as pigeons.

Most of the time, the words "*pigeonhole principle*" refer to one of Theorems 6.1.1 and 6.1.2. The next theorems in this section are closely related but less commonly used.

[160] The variant name "box principle" likewise results from stating the principle in terms of pearls and boxes (instead of pigeons and pigeonholes).

[161] We will give proofs of these theorems (or references to such proofs) later.

The "equality case" of Theorem 6.1.2 is also interesting: If we have equally many pigeons and pigeonholes, then any injective distribution of pigeons into pigeonholes must be surjective as well (i.e., each pigeonhole is inhabited), and thus must be bijective. Let us state this formally, too:

Theorem 6.1.3. Let U and V be two finite sets. Let $f : U \to V$ be an injective map. Assume that $|U| = |V|$. Then, f is bijective.

A few words about the proofs: Theorem 6.1.3 follows immediately from [Grinbe15, Lemma 1.5]. Theorem 6.1.1 is an easy consequence of [Grinbe15, Lemma 1.5] as well[162]. Theorem 6.1.2 is just the contrapositive of Theorem 6.1.1. All three theorems are basic facts about finite sets and can be used without proof on any contest.

A warning about infinite sets: Theorem 6.1.3 fails badly when U and V are allowed to be infinite. For example, the map $\mathbb{N} \to \mathbb{N}$, $i \mapsto i+1$ is injective, but not bijective.[163]

For future use, let us restate Theorem 6.1.1 in more explicit terms:

Corollary 6.1.4. Let $n, m \in \mathbb{N}$ satisfy $m > n$. Let V be an n-element set. Let $a_1, a_2, \ldots, a_m$ be m elements of V. Then, at least two of these m elements $a_1, a_2, \ldots, a_m$ must be equal.

Proof of Corollary 6.1.4. We have $|V| = n$ (since V is an n-element set). Let $f : \{1, 2, \ldots, m\} \to V$ be the map that sends each $i \in \{1, 2, \ldots, m\}$ to a_i. Then, Theorem 6.1.1 (applied to $U = \{1, 2, \ldots, m\}$) shows that this map f cannot be injective (since $|\{1, 2, \ldots, m\}| = m > n = |V|$). In other words, there exist two distinct elements i and j of $\{1, 2, \ldots, m\}$ such that $f(i) = f(j)$. In other words, there exist two distinct elements i and j of $\{1, 2, \ldots, m\}$ such that $a_i = a_j$ (since the definition of f yields $f(i) = a_i$ and $f(j) = a_j$). In other words, at least two of the m elements $a_1, a_2, \ldots, a_m$ must be equal. This proves Corollary 6.1.4. $\square$

[162] *Proof.* Let U and V be two finite sets such that $|U| > |V|$. Let $f : U \to V$ be any map. Assume that f is injective. Then, [Grinbe15, Lemma 1.5] yields the logical equivalence

$$(f \text{ is injective}) \Longleftrightarrow (f \text{ is bijective}).$$

Hence, f is bijective (since f is injective). Thus, f is a bijection. Hence, there exists a bijection from U to V (namely, f). This yields $|U| = |V|$, which contradicts $|U| > |V|$. This contradiction shows that our assumption (that f is injective) was wrong. Hence, f cannot be injective. This proves Theorem 6.1.1.

[163] Compare "Hilbert's Hotel" for various phenomena like this. (Hilbert seemed to prefer humans to pigeons in his thought experiments.)

6.1.2. The Pigeonhole Principle for "Multi-injections"

The pigeonhole principle can be generalized. Namely, if more than kn pigeons are distributed into n pigeonholes (for given $n, k \in \mathbb{N}$), then at least $k+1$ pigeons must be in the same pigeonhole. Let us restate this as a contrapositive: If we have n pigeonholes and each pigeonhole fits only k pigeons, then we cannot distribute more than kn pigeons into these holes. Let us restate this formally:

Theorem 6.1.5. Let U and V be two finite sets. Let $k \in \mathbb{N}$. Let $f : U \to V$ be a map. Assume that each $v \in V$ satisfies

$$(\text{the number of all } u \in U \text{ such that } f(u) = v) \leq k. \tag{268}$$

Then, $|U| \leq k|V|$.

Proof of Theorem 6.1.5. We have $\sum_{s \in U} 1 = |U| \cdot 1 = |U|$, so that

$$\begin{aligned}
|U| &= \sum_{s \in U} 1 = \sum_{w \in V} \sum_{\substack{s \in U; \\ f(s) = w}} 1 \qquad \text{(by (92), applied to } S = U \text{ and } W = V \text{ and } a_s = 1) \\
&= \sum_{v \in V} \underbrace{\sum_{\substack{u \in U; \\ f(u) = v}} 1}_{\substack{= (\text{the number of all } u \in U \text{ such that } f(u) = v) \cdot 1 \\ = (\text{the number of all } u \in U \text{ such that } f(u) = v) \\ \underset{(\text{by } (268))}{\leq k}}} \\
&\qquad (\text{here, we have renamed the summation indices } w \text{ and } s \text{ as } v \text{ and } u) \\
&\leq \sum_{v \in V} k = |V| \cdot k = k|V|.
\end{aligned}$$

This proves Theorem 6.1.5. □

The contrapositive of Theorem 6.1.5 says the following:

Theorem 6.1.6. Let U and V be two finite sets. Let $k \in \mathbb{N}$. Let $f : U \to V$ be a map. Assume that $|U| > k|V|$. Then, there exists a $v \in V$ such that

$$(\text{the number of all } u \in U \text{ such that } f(u) = v) > k.$$

Theorem 6.1.5, too, is known as the "pigeonhole principle". There is again an equality case statement, whose proof we leave to the reader:

Exercise 6.1.1. Let U and V be two finite sets. Let $k \in \mathbb{N}$. Let $f : U \to V$ be a map. Assume that each $v \in V$ satisfies (268). Assume that $|U| = k|V|$. Then, the inequality (268) is an equality for each $v \in V$.

6.1.3. The Dual Pigeonhole Principle for Surjections

The "dual pigeonhole principle" (my own name) says that if fewer than n pigeons are distributed into n pigeonholes (for a given $n \in \mathbb{N}$), then at least one pigeonhole is empty. Formally speaking:

Theorem 6.1.7. Let U and V be two finite sets such that $|U| < |V|$. Let $f : U \to V$ be any map. Then, f cannot be surjective.

Here is the contrapositive of Theorem 6.1.7:

Theorem 6.1.8. Let U and V be two finite sets. Let $f : U \to V$ be a surjective map. Then, $|U| \geq |V|$.

In terms of pigeons and pigeonholes, this is saying that any surjective distribution of pigeons into pigeonholes (where "surjective" means "no hole stays empty") requires at least as many pigeons as pigeonholes.

Again, the "equality case" is interesting: If we have equally many pigeons and pigeonholes, then any surjective distribution of pigeons into pigeonholes must be injective as well (i.e., no two pigeons share a hole). In formal terms, this is saying the following:

Theorem 6.1.9. Let U and V be two finite sets. Let $f : U \to V$ be a surjective map. Assume that $|U| = |V|$. Then, f is bijective.

A few words about the proofs: Theorem 6.1.9 follows immediately from [Grinbe15, Lemma 1.4]. Theorem 6.1.7 is an easy consequence of [Grinbe15, Lemma 1.4] as well[164]. Theorem 6.1.8 is just the contrapositive of Theorem 6.1.7. Again, these are basic results that need no proof in real mathematics.

Again, Theorem 6.1.9 fails when U and V are allowed to be infinite. For example, the map $\mathbb{N} \to \mathbb{N}$, $i \mapsto i//2$ is surjective, but not bijective.

6.2. Applications

6.2.1. Simple applications

There are many ways in which the pigeonhole principle (in its various forms) can be used. Let us first collect some simple examples (before moving on to more sophisticated ones):

[164] *Proof.* Let U and V be two finite sets such that $|U| < |V|$. Let $f : U \to V$ be any map. Assume that f is surjective. Then, [Grinbe15, Lemma 1.4] yields the logical equivalence

$$(f \text{ is surjective}) \iff (f \text{ is bijective}).$$

Hence, f is bijective (since f is surjective). Thus, f is a bijection. Hence, there exists a bijection from U to V (namely, f). This yields $|U| = |V|$, which contradicts $|U| < |V|$. This contradiction shows that our assumption (that f is surjective) was wrong. Hence, f cannot be surjective. This proves Theorem 6.1.7.

- Among any 13 persons, there are two that are born in the same month.

  This follows from Theorem 6.1.1 (applied to $U = \{\text{the 13 persons}\}$ and $V = \{\text{the 12 months}\}$ and to f being the map that sends each person to their birth month). Or, to say this in less formal terms: This follows from the pigeonhole principle, where the "pigeons" are the 13 persons and the "pigeonholes" are the 12 months.

- There are two people alive right now with the exact same number of hairs on their heads.

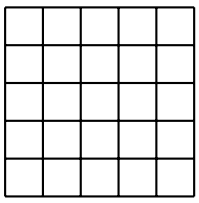
  Indeed, we can safely assume that any human has at most 5 000 000 hairs on their head (in fact, the typical number of hairs on a human head is around 100 000), but the world population is far larger (over 7 billion). Thus, we can apply Theorem 6.1.1 to $U = \{\text{humans alive right now}\}$ and $V = \{0, 1, \ldots, 5\,000\,000\}$ and to f being the map that sends each person to their number of hairs.

  (Alternatively, it suffices to find two baldies.)

  I have stolen this example from [Galvin20, Week 2], but in fact it is one of the oldest examples of the Pigeonhole Principle; it appeared in a 1622 book by Jean Leurechon (see [RitHee13]).

- If 26 mosquitoes are hanging from a $1\,\text{m} \times 1\,\text{m}$-sized rectangular window, and you have a $20\,\text{cm} \times 20\,\text{cm}$-sized rectangular flyswatter, then you can hit (at least) two mosquitoes with one strike of the flyswatter (if you are fast enough).

  Indeed, we can subdivide the window into 25 flyswatter-sized squares as follows:

  and then the pigeonhole principle (i.e., Theorem 6.1.1) will ensure that at least one of these 25 squares will contain at least 2 mosquitoes.

- Consider a triangle ABC, and a straight line ℓ in its plane. Assume that ℓ contains none of the vertices A, B and C. Then, the line ℓ cannot cut more than two sides of the triangle ABC.

  Indeed, if ℓ cuts a segment XY without passing through either of its endpoints X and Y, then the two points X and Y must lie on different sides[165] of the line ℓ. But the line ℓ has only two sides, and therefore the pigeonhole principle (i.e., Theorem 6.1.1) guarantees that at least two vertices of the triangle ABC will lie on one side of ℓ; therefore, the side connecting these two vertices cannot be cut by ℓ.

[165] formally speaking: half-planes

- Among any 25 persons, there are three that are born in the same month.

  This follows by applying Theorem 6.1.6 to $k = 2$ and $U = \{\text{the 25 persons}\}$ and $V = \{\text{the 12 months}\}$, because these two sets U and V satisfy $|U| > 2|V|$.

- If n is a positive integer, and if $x_1, x_2, \ldots, x_{n+1}$ are $n+1$ integers, then at least two of these $n+1$ integers $x_1, x_2, \ldots, x_{n+1}$ are congruent modulo n.

  Indeed, the remainders $x_1\%n, \ x_2\%n, \ \ldots, \ x_{n+1}\%n$ are $n+1$ elements of the n-element set $\{0, 1, \ldots, n-1\}$; thus, Corollary 6.1.4 (applied to $m = n+1$ and $V = \{0, 1, \ldots, n-1\}$ and $a_i = x_i\%n$) shows that at least two of these $n+1$ elements $x_1\%n, \ x_2\%n, \ \ldots, \ x_{n+1}\%n$ must be equal; but this entails that the corresponding two of the integers $x_1, x_2, \ldots, x_{n+1}$ are congruent modulo n (by Proposition 3.3.4).

- If $n+1$ distinct numbers are selected from the set $\{1, 2, \ldots, 2n\}$ (for some positive integer n), then some two of these $n+1$ numbers sum to $2n+1$. [166]

  Indeed, define the "pigeons" to be our $n+1$ selected numbers. Define the "pigeonholes" to be the n two-element sets

  $$\{1, 2n\}, \ \{2, 2n-1\}, \ \{3, 2n-2\}, \ \ldots, \ \{n, n+1\}$$

  (that is, the two-element sets $\{i, 2n+1-i\}$ for all $i \in \{1, 2, \ldots, n\}$). These n two-element sets are disjoint and together cover the whole set $\{1, 2, \ldots, 2n\}$. Thus, each of our $n+1$ selected numbers lies in precisely one of these n two-element sets. In other words, each of our $n+1$ "pigeons" lies in precisely one "pigeonhole". By the Pigeonhole Principle (Theorem 6.1.1), this entails that two of our "pigeons" share a "pigeonhole" – i.e., that two of our $n+1$ selected numbers lie in the same two-element set. Since the selected numbers are distinct, this means that these two numbers must be the two elements of the two-element set they lie in; therefore, they sum to $2n+1$.

- You have 10 pairs of shoes: 5 pairs of black shoes and 5 pairs of yellow shoes. You pick k shoes at random. How high must k be so you can be sure that among the shoes you pick, there will be at least one matching pair (i.e., one left and one right shoe of the same color)?

  The answer is that k must be at least 11. Indeed, $k \leq 10$ is insufficient, since picking $k \leq 10$ shoes may leave you with k left shoes. On the other hand, picking 11 shoes will suffice, because the Pigeonhole Principle (Theorem 6.1.5) guarantees that among these 11 shoes there will be at least 6 of the same color, and (by assumption) these 6 shoes cannot all be left or all be right (since there are only 5 left shoes of each color, and only 5 right shoes of each color).

[166] For example, for $n = 5$, if we select the 6 numbers $1, 3, 4, 5, 7, 9$ from the set $\{1, 2, \ldots, 10\}$, then some two of these 6 numbers (namely, 4 and 7) sum to 11.

More such examples can be found, e.g., in [Macgil17, Notes on the Pigeonhole Principle].

We have also seen an application of the Pigeonhole Principle (specifically, of Corollary 6.1.4) in the solution to Exercise 1.1.9. The third solution to Exercise 3.7.1 can also be viewed as a use of the Pigeonhole Principle (indeed, regard the elements of the k-lacunar subset S as pigeons, and regard the $q+1$ intervals $I_0, I_1, \ldots, I_q$ as pigeonholes).

Here is another simple consequence of the pigeonhole principles:

Corollary 6.2.1. Let U and V be two finite sets such that $|U| \leq |V|$. Let $f : U \to V$ and $g : V \to U$ be two maps such that $f \circ g = \mathrm{id}_V$. Then, the maps f and g are mutually inverse.

Note that Corollary 6.2.1 fails if the condition $|U| \leq |V|$ is removed[167], and also fails if U and V are allowed to be infinite[168].

Proof of Corollary 6.2.1. The assumption $f \circ g = \mathrm{id}_V$ shows that the map f is surjective[169]. Hence, Theorem 6.1.8 shows that $|U| \geq |V|$. Combining this with $|U| \leq |V|$, we obtain $|U| = |V|$. Thus, Theorem 6.1.9 yields that f is bijective. In other words, f is invertible (since a map is bijective if and only if it is invertible). Hence, the inverse map f^{-1} exists. Now, from $f \circ g = \mathrm{id}_V$, we conclude that $f^{-1} = g$ (because comparing $\underbrace{f^{-1} \circ f}_{=\mathrm{id}_U} \circ g = \mathrm{id}_U \circ g = g$ with $f^{-1} \circ \underbrace{f \circ g}_{=\mathrm{id}_V} = f^{-1} \circ \mathrm{id}_V = f^{-1}$ results in $f^{-1} = g$). Thus, the maps f and g are mutually inverse. This proves Corollary 6.2.1. □

Note that, just as we have proved Corollary 6.2.1 using Theorems 6.1.8 and 6.1.9, we can also prove Corollary 6.2.1 using Theorems 6.1.2 and 6.1.3 instead (by applying the latter theorems to V, U and g instead of U, V and f). The details of this alternative proof are left to the reader.

6.2.2. Handshakes

A slightly less obvious application of the pigeonhole principle is the following exercise ([Grinbe08, Aufgabe 2.10] or [Engel98, Exercise 4.13]):

[167] For example, if $U = \{0,1,2,3\}$ and $V = \{0,1\}$ and $f(u) = u//2$ and $g(v) = 2v$, then $f \circ g = \mathrm{id}_V$, but the maps f and g are not mutually inverse.

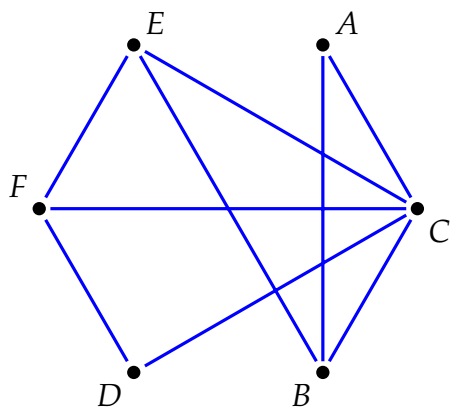
[168] For example, if $U = \mathbb{N}$ and $V = \mathbb{N}$ and $f(u) = u//2$ and $g(v) = 2v$, then $f \circ g = \mathrm{id}_V$, but the maps f and g are not mutually inverse.

[169] *Proof.* Let $v \in V$. Then, from $f \circ g = \mathrm{id}_V$, we obtain $(f \circ g)(v) = \mathrm{id}_V(v)$, so that $v = \mathrm{id}_V(v) = (f \circ g)(v) = f(g(v))$. Hence, v is a value of the map f (namely, the value of f on $g(v)$).

Forget that we fixed v. We thus have shown that each $v \in V$ is a value of the map f. In other words, the map f is surjective.

Exercise 6.2.1. Let $n \geq 2$ be an integer. At a conference long ago, n scientists have met; some of them have exchanged handshakes among each other[170]. Prove that two of these n scientists have shaken the same number of hands during the conference. (We assume that any pair of scientists shakes hands at most once. We also assume that no one shakes their own hands.)

Example 6.2.2. Let $n = 6$, and consider the following situation:



.

Here, A, B, C, D, E, F are the 6 scientists, and a line segment connects any pair of scientists that has exchanged a handshake. Thus, scientist A has shaken 2 hands; B has shaken 3 hands; C has shaken 5 hands; D has shaken 2 hands; etc.. Thus, A and D have shaken the same number of hands.

Example 6.2.3. The two assumptions in Exercise 6.2.1 (that any pair shakes hands at most once, and that no one shakes their own hands) are important. If a scientist is allowed to shake hands with themselves, then a counterexample can be obtained even for $n = 2$ (imagine two scientists, one of whom shakes their own hands while the other doesn't), independently of whether we count a "self-handshake" as one or two hands being shaken. If a pair of scientists is allowed to shake hands multiple times, then a counterexample can be found for $n = 3$ (imagine three scientists A, B and C, where A shakes hands with B once and with C twice). The language of graph theory will later give us faster ways to state such assumptions (in particular, we will be able to model the n scientists as a *loopless simple graph*).

Solution to Exercise 6.2.1. Assume the contrary. Thus, any two distinct scientists have shaken different numbers of hands.

Let U be the set of our n scientists. Let V be the set $\{0, 1, \ldots, n-1\}$. Thus, both U and V are n-element sets; hence, $|U| = |V|$. From $n \geq 2$, we obtain $n - 1 \neq 0$ and $0 \in V$ and $n - 1 \in V$.

[170] For readers from the future: Handshakes were a form of greeting popular until 2020. A handshake involves two people, each of whom shakes exactly one hand of the other. Thus, when person a shakes person b's hands, person b also shakes person a's hands.

Define a map $f : U \to V$ as follows: For each scientist $s \in U$, we let $f(s)$ be the number of hands that s has shaken. This is well-defined, because s has shaken at most $n-1$ hands[171] and thus we have $f(s) \in \{0, 1, \ldots, n-1\} = V$.

We have assumed that any two distinct scientists have shaken different numbers of hands. In other words, if s and t are two distinct scientists, then $f(s) \neq f(t)$. In other words, the map f is injective. Hence, Theorem 6.1.3 yields that f is bijective (since $|U| = |V|$). Thus, in particular, f is surjective. Hence, there exists a scientist $a \in U$ such that $f(a) = n-1$ (since $n-1 \in V$), and there exists a scientist $b \in U$ such that $f(b) = 0$ (since $0 \in V$). Consider these a and b. Note that $a \neq b$ (since $f(a) = n-1 \neq 0 = f(b)$).

The scientist a has shaken $n-1$ hands (since $f(a) = n-1$). Thus, a must have shaken everyone's hands (except for a themselves), because there are only $n-1$ hands in total for a to possibly shake. Hence, in particular, a must have shaken the hands of b (since $a \neq b$). Thus, b must have shaken the hands of a. But from $f(b) = 0$, we see that b has shaken no hands at all. The previous two sentences contradict each other. This contradiction shows that our assumption was false. Hence, Exercise 6.2.1 is solved. □

6.2.3. Back to Bezout

Let us next solve Exercise 3.7.7 again, this time using the pigeonhole principle. We begin with the particular case in which $a \perp b$ ([Engel98, Chapter 4, Example E5]):

Exercise 6.2.2. Let a and b be two positive integers such that $a \perp b$. Prove that there exist positive integers x and y such that $1 = xa - yb$.

Solution to Exercise 6.2.2. Let $U = \{2, 3, \ldots, b+1\}$ and $V = \{0, 1, \ldots, b-1\}$. Both U and V are b-element sets; thus, $|U| = |V|$.

Let $f : U \to V$ be the map that sends each $u \in U$ to the remainder $(ua)\,\%b$. (This is well-defined, since each $u \in U$ satisfies $(ua)\,\%b \in \{0, 1, \ldots, b-1\} = V$.)

We shall now show the following:

Claim 1: The map f is injective.

[*Proof of Claim 1:* Let u_1 and u_2 be two distinct elements of U. We shall show that $f(u_1) \neq f(u_2)$.

Indeed, assume the contrary. Thus, $f(u_1) = f(u_2)$, so that $f(u_2) = f(u_1)$. We WLOG assume that $u_1 \leq u_2$ (since otherwise, we can simply swap u_1 with u_2). Hence, $u_1 < u_2$ (since u_1 and u_2 are distinct), so that $u_2 - u_1 > 0$. Thus, $u_2 - u_1 \neq 0$.

From $u_1 \in U = \{2, 3, \ldots, b+1\}$, we obtain $u_1 \geq 2$. From $u_2 \in U = \{2, 3, \ldots, b+1\}$, we obtain $u_2 \leq b+1$. Hence,

$$\underbrace{u_2}_{\leq b+1} - \underbrace{u_1}_{\geq 2} \leq (b+1) - 2 = b - 1 < b.$$

[171] Here we use our assumptions that any pair of scientists shakes hands at most once, and that no one shakes their own hands.

The definition of f yields $f(u_1) = (u_1 a)\,\%b$ and $f(u_2) = (u_2 a)\,\%b$. Hence, $(u_2 a)\,\%b = f(u_2) = f(u_1) = (u_1 a)\,\%b$. But Proposition 3.3.4 (applied to $u_2 a$, $u_1 a$ and b instead of u, v and n) yields that $u_2 a \equiv u_1 a \bmod b$ if and only if $(u_2 a)\,\%b = (u_1 a)\,\%b$. Hence, we have $u_2 a \equiv u_1 a \bmod b$ (since we have $(u_2 a)\,\%b = (u_1 a)\,\%b$). Thus, $a u_2 = u_2 a \equiv u_1 a = a u_1 \bmod b$. Therefore, Lemma 3.5.11 (applied to u_2, u_1 and b instead of b, c and n) yields $u_2 \equiv u_1 \bmod b$ (since $a \perp b$). In other words, $b \mid u_2 - u_1$. Hence, Proposition 3.1.3 **(b)** (applied to b and $u_2 - u_1$ instead of a and b) yields $|b| \leq |u_2 - u_1| = u_2 - u_1$ (since $u_2 - u_1 > 0$). Hence, $u_2 - u_1 \geq |b| = b$ (since $b > 0$). This contradicts $u_2 - u_1 < b$.

This contradiction shows that our assumption was false. Hence, $f(u_1) \neq f(u_2)$ is proved.

Now, forget that we fixed u_1 and u_2. We thus have shown that $f(u_1) \neq f(u_2)$ for any two distinct elements u_1 and u_2 of U. In other words, the map f is injective. This proves Claim 1.]

Claim 1 says that f is injective. Thus, Theorem 6.1.3 yields that f is bijective (since $|U| = |V|$). Therefore, f is surjective. Hence, there exists some $u \in U$ such that $f(u) = 1\%b$ (since $1\%b \in \{0, 1, \ldots, b-1\} = V$). Consider this u. (Note that $1\%b = 1$ whenever $b > 1$; but I don't want to make an exception for the case $b = 1$.)

The definition of f yields $f(u) = (ua)\,\%b$. Hence, $(ua)\,\%b = f(u) = 1\%b$. But Proposition 3.3.4 (applied to ua, 1 and b instead of u, v and n) yields that $ua \equiv 1 \bmod b$ if and only if $(ua)\,\%b = 1\%b$. Hence, we have $ua \equiv 1 \bmod b$ (since we have $(ua)\,\%b = 1\%b$). In other words, $b \mid ua - 1$. In other words, there exists some integer v such that $ua - 1 = bv$. Consider this v. From $ua - 1 = bv$, we obtain $1 = ua - bv = ua - vb$.

We have $u \in U = \{2, 3, \ldots, b+1\}$; thus, u is a positive integer. Furthermore, it is easy to see that v is a positive integer[172]. Thus, u and v are two positive integers and satisfy $1 = ua - vb$. Therefore, there exist positive integers x and y such that $1 = xa - yb$ (namely, $x = u$ and $y = v$). This solves Exercise 6.2.2. □

It is now easy to reduce Exercise 3.7.7 to Exercise 6.2.2:

Second solution to Exercise 3.7.7. Both numbers a and b are positive and thus nonzero; hence, $(a, b) \neq (0, 0)$. Let $g = \gcd(a, b)$. Thus, Proposition 3.5.12 yields that $g > 0$ and $\dfrac{a}{g} \perp \dfrac{b}{g}$. In particular, the numbers $\dfrac{a}{g}$ and $\dfrac{b}{g}$ are integers. Moreover, these integers $\dfrac{a}{g}$ and $\dfrac{b}{g}$ are positive (since a, b and g are positive). Hence, Exercise 6.2.2 (applied to $\dfrac{a}{g}$ and $\dfrac{b}{g}$ instead of a and b) shows that there exist positive integers x and y such that $1 = x \cdot \dfrac{a}{g} - y \cdot \dfrac{b}{g}$. Consider these x and y, and denote them by u

[172] *Proof.* From $u \in U = \{2, 3, \ldots, b+1\}$, we obtain $u \geq 2 > 1$. Multiplying this inequality by a, we obtain $ua > 1a$ (since a is positive). Hence, $ua > 1a = a \geq 1$ (since a is a positive integer). However, $ua - 1 = bv$, so that $bv = ua - 1 > 0$ (since $ua > 1$). We can divide this equality by b (since b is positive), and obtain $v > 0$. Hence, v is a positive integer (since v is an integer).

and v. Thus, u and v are two positive integers and satisfy $1 = u \cdot \frac{a}{g} - v \cdot \frac{b}{g}$. Hence, $1 = u \cdot \frac{a}{g} - v \cdot \frac{b}{g} = \frac{1}{g}(ua - vb)$, so that $g = ua - vb$. In view of $g = \gcd(a, b)$, this rewrites as $\gcd(a, b) = ua - vb$. Hence, there exist positive integers x and y such that $\gcd(a, b) = xa - yb$ (namely, $x = u$ and $y = v$). This solves Exercise 3.7.7 again. $\square$

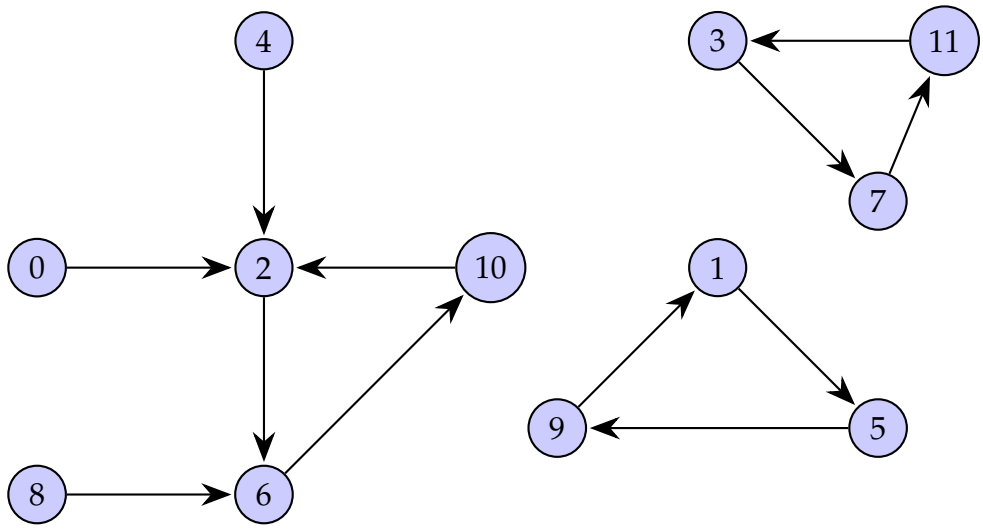
6.2.4. An endofunction of a finite set

An *endofunction* of a set X means a map from X to X. While we will not use this notation, its very existence suggests that there is something to say about these kinds of maps. Here is a first example:

Example 6.2.4. Let $X = \{0, 1, \ldots, 11\}$. Define a map $f : X \to X$ by setting

$$f(i) = \left(i^3 + 3i + 1\right) \%12 \qquad \text{for each } i \in X.$$

For example, $f(5) = \underbrace{\left(5^3 + 3 \cdot 5 + 1\right)}_{=141} \%12 = 141\%12 = 9$. The following diagram shows all elements of X (as nodes) and how the map f acts on them (by drawing an arrow from each node $i \in X$ to the node $f(i)$):



We observe on this diagram that if we start at any node and follow the arrows, we will eventually get stuck in a cycle (i.e., we will start walking around in circles). This observation holds in general whenever X is a finite set and $f : X \to X$ is a map (although the cycles in which we get stuck might be 1-node cycles – i.e., we may run into a fixed point). Here is a somewhat stronger statement:[173]

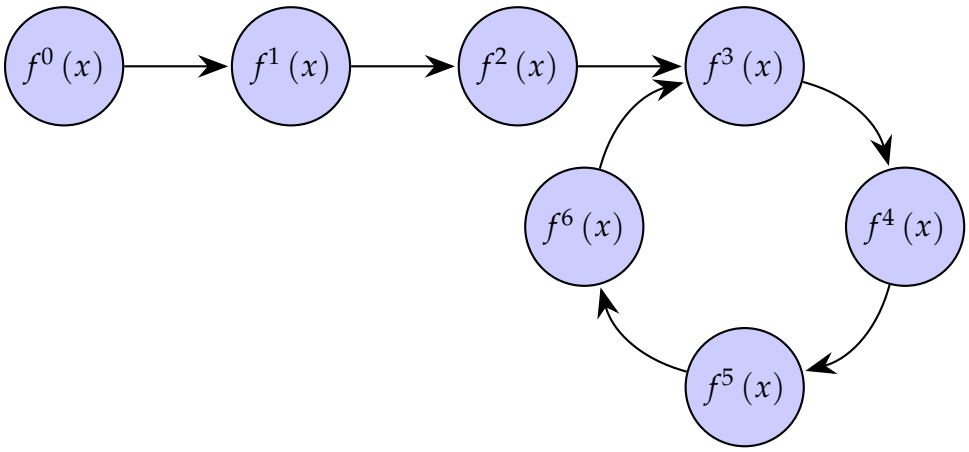
[173] If X is a set, and if $f : X \to X$ is a map, then the notation f^n (for $n \in \mathbb{N}$) denotes the n-fold

Theorem 6.2.5. Let X be a finite set. Let $n = |X|$. Let $f : X \to X$ be a map. Let $x \in X$. Then:

(a) There exist two integers i and j with $0 \leq i < j \leq n$ and $f^i(x) = f^j(x)$.

(b) Let i and j be two integers with $0 \leq i < j$ and $f^i(x) = f^j(x)$. Then, the sequence $\left(f^i(x), f^{i+1}(x), f^{i+2}(x), \ldots\right)$ is $(j-i)$-periodic.

Here is an illustration of this behavior (showing only an element $x \in X$ and its images $f^0(x), f^1(x), f^2(x), \ldots$ under repeated application of f):



(this shows a case when $f^3(x) = f^7(x)$, so that the i and j in Theorem 6.2.5 can be chosen to be 3 and 7). This kind of picture is known as a *rho* to number theorists (in honor of the Greek letter rho, which it vaguely resembles) or as a *lollipop* to graph theorists. Note that we can have $i = 0$ (in which the picture consists just of a cycle with no "running-in") or $j = i + 1$ (in which case the cycle is just a single node with an arrow directly to itself).

Proof of Theorem 6.2.5. **(a)** Consider the $n+1$ elements $f^0(x), f^1(x), \ldots, f^n(x)$ of the n-element set X. Corollary 6.1.4 (applied to $m = n+1$, $V = X$ and $a_i = f^{i-1}(x)$) yields that at least two of these $n+1$ elements $f^0(x), f^1(x), \ldots, f^n(x)$ must be equal (since $n+1 > n$). In other words, there exist two integers i and j with $0 \leq i < j \leq n$ and $f^i(x) = f^j(x)$. This proves Theorem 6.2.5 **(a)**.

(b) Let $d = j - i$. Then, $d = j - i > 0$ (since $i < j$); thus, d is a positive integer (since $i < j$). Let us denote the sequence $\left(f^i(x), f^{i+1}(x), f^{i+2}(x), \ldots\right)$ by $(u_0, u_1, u_2, \ldots)$. Thus,

$$u_k = f^{i+k}(x) \qquad \text{for each } k \in \mathbb{N}. \tag{269}$$

We shall now prove that every $k \in \mathbb{N}$ satisfies $u_k = u_{k+d}$.

composition $\underbrace{f \circ f \circ \cdots \circ f}_{n \text{ times}} : X \to X$. In particular, $f^0 = \mathrm{id}_X$ and $f^n = f \circ f^{n-1} = f^{n-1} \circ f$ for each positive integer n.

Indeed, let $k \in \mathbb{N}$. Set $\ell = k + d$. Then, $\ell = k + \underbrace{d}_{>0} > k \geq 0$, so that $\ell \in \mathbb{N}$. Also, from $\ell = k + \underbrace{d}_{=j-i} = k + j - i$, we obtain $k + j = i + \ell$.

It is well-known that $f^p \circ f^q = f^{p+q}$ for any $p, q \in \mathbb{N}$. Applying this to $p = k$ and $q = i$, we obtain $f^k \circ f^i = f^{k+i}$. Hence, $\left(f^k \circ f^i\right)(x) = f^{k+i}(x)$. But (269) yields $u_k = f^{i+k}(x) = f^{k+i}(x)$ (since $i + k = k + i$). Comparing these two equalities, we obtain $u_k = \left(f^k \circ f^i\right)(x) = f^k\left(f^i(x)\right)$.

Recall again that $f^p \circ f^q = f^{p+q}$ for any $p, q \in \mathbb{N}$. Applying this to $p = k$ and $q = j$, we obtain $f^k \circ f^j = f^{k+j}$. Hence, $\left(f^k \circ f^j\right)(x) = f^{k+j}(x) = f^{i+\ell}(x)$ (since $k + j = i + \ell$). But (269) (applied to ℓ instead of k) yields $u_\ell = f^{i+\ell}(x)$. Comparing these two equalities, we obtain $u_\ell = \left(f^k \circ f^j\right)(x) = f^k\left(f^j(x)\right)$.

Now, recall that

$$\begin{aligned} u_k &= f^k\left(\underbrace{f^i(x)}_{=f^j(x)}\right) = f^k\left(f^j(x)\right) = u_\ell \qquad \left(\text{since } u_\ell = f^k\left(f^j(x)\right)\right) \\ &= u_{k+d} \qquad (\text{since } \ell = k + d). \end{aligned}$$

Forget that we fixed k. We thus have proved that every $k \in \mathbb{N}$ satisfies $u_k = u_{k+d}$. However, d is a positive integer; thus, we know (from Definition 4.7.1 **(a)**) that d is a period of the sequence $(u_0, u_1, u_2, \ldots)$ if and only if every $k \in \mathbb{N}$ satisfies $u_k = u_{k+d}$. Therefore, d is a period of the sequence $(u_0, u_1, u_2, \ldots)$ (since every $k \in \mathbb{N}$ satisfies $u_k = u_{k+d}$). In other words, the sequence $(u_0, u_1, u_2, \ldots)$ is d-periodic (by Definition 4.7.1 **(c)**). In other words, the sequence $\left(f^i(x), f^{i+1}(x), f^{i+2}(x), \ldots\right)$ is $(j-i)$-periodic (since $(u_0, u_1, u_2, \ldots) = \left(f^i(x), f^{i+1}(x), f^{i+2}(x), \ldots\right)$ and $d = j - i$). This proves Theorem 6.2.5 **(b)**. □

Some of the most interesting endofunctions are the bijective ones. They are known as *permutations*:[174]

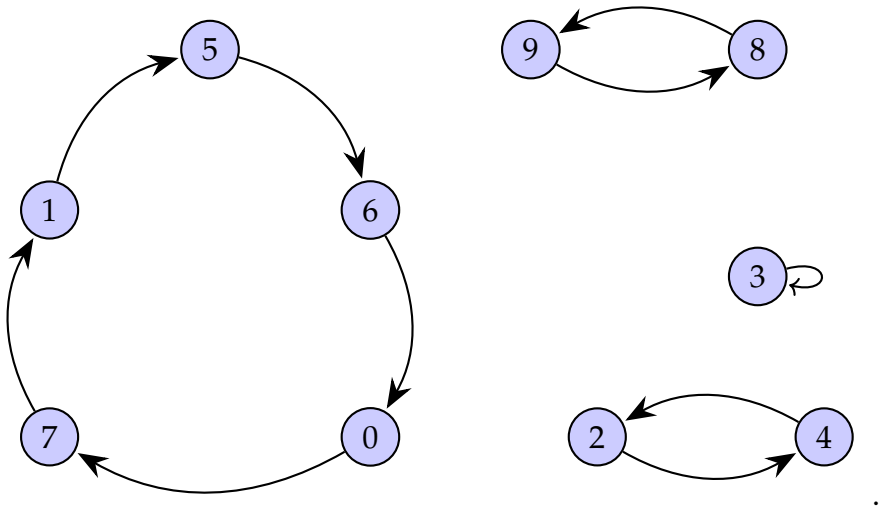
Definition 6.2.6. Let X be a set. A *permutation* of X means a bijective map $f : X \to X$.

The identity map $\mathrm{id}_X : X \to X$ of a set X is always a permutation of X. Here is another example of a permutation:

Example 6.2.7. Let $X = \{0, 1, \ldots, 9\}$. Let $f : X \to X$ be the map that sends the elements $0, 1, 2, 3, 4, 5, 6, 7, 8, 9$ to $7, 5, 4, 3, 2, 6, 0, 1, 9, 8$, respectively. The following

[174] This is only one of two commonly used meanings of the word "permutation"; see (e.g.) [Grinbe15, Remark 5.4] or the Wikipedia page for the distinction. In a nutshell, the second meaning of "permutation" is a list that contains each element of a set exactly once. In order to avoid confusion, we shall not use this second meaning.

diagram shows all elements of X (as nodes) and how the map f acts on them (by drawing an arrow from each node $i \in X$ to the node $f(i)$):



Note that our map f here is a permutation of X; thus, any node has exactly one arrow arriving at it.

The behavior seen in this diagram is a simpler particular case of that seen in Example 6.2.4: Every node is part of a cycle (rather than merely being connected to a cycle via some arrows). We can formalize this as follows:

Theorem 6.2.8. Let X be a finite set. Let $n = |X|$. Let $f : X \to X$ be a permutation of X. Let $x \in X$. Then:

(a) There exists a $k \in \{1, 2, \ldots, n\}$ such that $f^k(x) = x$.

(b) Let $k \in \{1, 2, \ldots, n\}$ be such that $f^k(x) = x$. Then, the sequence $\left(f^0(x), f^1(x), f^2(x), \ldots\right)$ is k-periodic.

Proof of Theorem 6.2.8. **(a)** Theorem 6.2.5 **(a)** shows that there exist two integers i and j with $0 \leq i < j \leq n$ and $f^i(x) = f^j(x)$. Consider these i and j. Let $d = j - i$. Then, $d = j - i > 0$ (since $i < j$) and $i + d = j$ (since $d = j - i$).

It is well-known that $f^p \circ f^q = f^{p+q}$ for any $p, q \in \mathbb{N}$. Applying this to $p = i$ and $q = d$, we obtain $f^i \circ f^d = f^{i+d} = f^j$ (since $i + d = j$). Hence, $f^j = f^i \circ f^d$, so that $f^j(x) = \left(f^i \circ f^d\right)(x) = f^i\left(f^d(x)\right)$. Therefore, $f^i(x) = f^j(x) = f^i\left(f^d(x)\right)$, so that $f^i\left(f^d(x)\right) = f^i(x)$.

The map f is a permutation, thus bijective (by the definition of a permutation). Hence, its power $f^i = \underbrace{f \circ f \circ \cdots \circ f}_{i \text{ times}}$ is bijective as well (since a composition of bijective maps is always bijective). Thus, in particular, this map f^i is injective. In other words, if u and v are two elements of X satisfying $f^i(u) = f^i(v)$, then $u = v$. Applying this to $u = f^d(x)$ and $v = x$, we conclude that $f^d(x) = x$ (since

$f^i\left(f^d(x)\right) = f^i(x)$). Moreover, combining $d = \underbrace{j}_{\leq n} - \underbrace{i}_{\geq 0} \leq n - 0 = n$ with $d > 0$, we obtain $d \in \{1, 2, \ldots, n\}$. Hence, there exists a $k \in \{1, 2, \ldots, n\}$ such that $f^k(x) = x$ (namely, $k = d$). This proves Theorem 6.2.8 **(a)**.

(b) We have $0 \leq 0 < k$ and $\underbrace{f^0}_{=\mathrm{id}_X}(x) = \mathrm{id}_X(x) = x = f^k(x)$ (since $f^k(x) = x$). Hence, Theorem 6.2.5 **(b)** (applied to $i = 0$ and $j = k$) yields that the sequence $\left(f^0(x), f^{0+1}(x), f^{0+2}(x), \ldots\right)$ is $(k-0)$-periodic. In other words, the sequence $\left(f^0(x), f^1(x), f^2(x), \ldots\right)$ is k-periodic (since $\left(f^0(x), f^{0+1}(x), f^{0+2}(x), \ldots\right) = \left(f^0(x), f^1(x), f^2(x), \ldots\right)$ and $k - 0 = k$). This proves Theorem 6.2.8 **(b)**. □

For an endofunction of a **finite** set X to be a permutation, it suffices to be injective **or** surjective:

Corollary 6.2.9. Let X be a finite set. Let $f : X \to X$ be a map. Then:
(a) If f is injective, then f is a permutation of X.
(b) If f is surjective, then f is a permutation of X.

Proof of Corollary 6.2.9. **(a)** Assume that f is injective. Hence, Theorem 6.1.3 (applied to $U = X$ and $V = X$) yields that f is bijective (since $|X| = |X|$). In other words, f is a permutation of X. This proves Corollary 6.2.9 **(a)**.

(b) Assume that f is surjective. Hence, Theorem 6.1.9 (applied to $U = X$ and $V = X$) yields that f is bijective (since $|X| = |X|$). In other words, f is a permutation of X. This proves Corollary 6.2.9 **(b)**. □

We also note that any composition of two permutations of a set X is again a permutation of X (since a composition of two bijections is always a bijection). Furthermore, the inverse of any permutation of a set X is again a permutation of X (since the inverse of a bijection is always a bijection). If you are familiar with the notion of a group, you will thus recognize the set of all permutations of a given set X to be a group (known as the *symmetric group* on X).

More can be said about endofunctions. Even the "rho" (or "lollipop") of an element bears some surprises:

Theorem 6.2.10. Let X be a finite set. Let $n = |X|$. Let $f : X \to X$ be a map. Let $x \in X$. Then, there exists some $k \in \{1, 2, \ldots, n\}$ such that $f^k(x) = f^{2k}(x)$.

For instance, in Example 6.2.4, we have $f^3(4) = f^{2\cdot 3}(4)$.

Theorem 6.2.10 is often known as the *tortoise-and-hare theorem*. (Imagine a tortoise and a hare starting at the node x on a diagram like the one we drew in Example 6.2.4. Both animals walk along the arrows. Each second, the tortoise makes a single step forward while the hare makes two. Theorem 6.2.10 then says that the tortoise and the hare will meet again after at most n seconds. This is used in *Floyd's tortoise-and-hare algorithm for cycle detection*.)

See Exercise 6.3.3 for a generalization of Theorem 6.2.10 (allowing both the tortoise and the hare to start with leads).

6.2.5. Periodicity of linear recurrences modulo m

Exercise 3.3.6 and Exercise 3.7.6 show that the Fibonacci numbers $f_0, f_1, f_2, \ldots$ have a periodic behavior modulo 2 and modulo 5. To wit, Exercise 3.3.6 shows that the sequence $(f_0\%2, f_1\%2, f_2\%2, \ldots)$ is 3-periodic, whereas Exercise 3.7.6 shows (fairly easily) that the sequence $(f_0\%5, f_1\%5, f_2\%5, \ldots)$ is 20-periodic (indeed, each $n \in \mathbb{N}$ satisfies $f_{n\%5} \equiv f_{(n+20)\%5} \bmod 5$ and $3^{n//5} \equiv 3^{(n+20)//5} \bmod 5$). These are parts of a pattern: For any positive integer m, the sequence $(f_0\%m, f_1\%m, f_2\%m, \ldots)$ is k-periodic for some $k \in \left\{1, 2, \ldots, m^2\right\}$. More generally, we can extend this to (a, b)-recurrent sequences whenever $b \perp m$:

Exercise 6.2.3. Let m be a positive integer. Let a and b be integers such that $b \perp m$. Let $(x_0, x_1, x_2, \ldots)$ be any (a, b)-recurrent sequence of integers. Prove that the sequence $(x_0\%m, x_1\%m, x_2\%m, \ldots)$ is k-periodic for some $k \in \left\{1, 2, \ldots, m^2\right\}$.

Solution to Exercise 6.2.3 (sketched). The matrix approach from Proposition 4.9.21 **(a)** will be useful here again – even though we don't really need matrices but can simply work with pairs of numbers. Recall that the idea behind this approach was that, even though an entry x_n of our sequence cannot be directly obtained from the previous entry x_{n-1}, each pair (x_n, x_{n+1}) of two consecutive entries can be obtained easily from the preceding such pair (x_{n-1}, x_n). It turns out that the same holds for the remainders $x_0\%m, x_1\%m, x_2\%m, \ldots$: Namely, each pair $(x_n\%m, x_{n+1}\%m)$ can be obtained from the preceding such pair $(x_{n-1}\%m, x_n\%m)$ (we shall soon see how).

First, let us introduce some notations.

Let M be the m-element set $\{0, 1, \ldots, m-1\}$. Thus, $|M \times M| = |M| \cdot |M| = |M|^2 = m^2$ (since $|M| = m$).

For each integer z, let $\overrightarrow{z}$ denote the remainder $z\%m$. Thus, every integer z satisfies $\overrightarrow{z} = z\%m \in \{0, 1, \ldots, m-1\} = M$ and

$$\overrightarrow{z} = z\%m \equiv z \bmod m \tag{270}$$

(by Proposition 3.3.2 **(a)**, applied to $u = z$ and $n = m$).

Define a map $f : M \times M \to M \times M$ by

$$f((p, q)) = \left(q, \overrightarrow{aq + bp}\right) \qquad \text{for each } (p, q) \in M \times M.$$

This is well-defined, since each $(p, q) \in M \times M$ satisfies $\left(q, \overrightarrow{aq + bp}\right) \in M \times M$ (because $q \in M$ and $\overrightarrow{aq + bp} = (aq + bp)\%m \in \{0, 1, \ldots, m-1\} = M$).

Let ω be the pair $\left(\overrightarrow{x_0}, \overrightarrow{x_1}\right) \in M \times M$.

Now, we claim the following:

Claim 1: The map f is bijective (thus a permutation of the set $M \times M$).

Claim 2: We have $f\left(\left(\overrightarrow{x_i}, \overrightarrow{x_{i+1}}\right)\right) = \left(\overrightarrow{x_{i+1}}, \overrightarrow{x_{i+2}}\right)$ for each $i \in \mathbb{N}$.

Claim 3: We have $\left(\overrightarrow{x_i}, \overrightarrow{x_{i+1}}\right) = f^i(\omega)$ for each $i \in \mathbb{N}$.

Proving these three claims will be an exercise (Exercise 6.3.6 below). Assume that they are proved for now. Now, we can apply Theorem 6.2.8 **(a)** to $X = M \times M$, $n = m^2$ and $x = \omega$. This yields that there exists a $k \in \left\{1, 2, \ldots, m^2\right\}$ such that $f^k(\omega) = \omega$. Consider this k. Applying Theorem 6.2.8 **(b)** to $X = M \times M$, $n = m^2$ and $x = \omega$, we conclude that the sequence $\left(f^0(\omega), f^1(\omega), f^2(\omega), \ldots\right)$ is k-periodic. In other words, k is a period of this sequence. In other words, every $i \in \mathbb{N}$ satisfies

$$f^i(\omega) = f^{i+k}(\omega) \tag{271}$$

(by the definition of a period).

Now, let $i \in \mathbb{N}$. Then, Claim 3 yields $\left(\overrightarrow{x_i}, \overrightarrow{x_{i+1}}\right) = f^i(\omega)$. Also, Claim 3 (applied to $i+k$ instead of i) yields $\left(\overrightarrow{x_{i+k}}, \overrightarrow{x_{i+k+1}}\right) = f^{i+k}(\omega)$. Hence,

$$\begin{aligned}\left(\overrightarrow{x_i}, \overrightarrow{x_{i+1}}\right) &= f^i(\omega) = f^{i+k}(\omega) \qquad \text{(by (271))} \\ &= \left(\overrightarrow{x_{i+k}}, \overrightarrow{x_{i+k+1}}\right) \qquad \left(\text{since } \left(\overrightarrow{x_{i+k}}, \overrightarrow{x_{i+k+1}}\right) = f^{i+k}(\omega)\right).\end{aligned}$$

In other words, $\overrightarrow{x_i} = \overrightarrow{x_{i+k}}$ and $\overrightarrow{x_{i+1}} = \overrightarrow{x_{i+k+1}}$. Thus, in particular, $\overrightarrow{x_i} = \overrightarrow{x_{i+k}}$.

Forget that we fixed i. We thus have shown that every $i \in \mathbb{N}$ satisfies $\overrightarrow{x_i} = \overrightarrow{x_{i+k}}$. In other words, k is a period of the sequence $\left(\overrightarrow{x_0}, \overrightarrow{x_1}, \overrightarrow{x_2}, \ldots\right)$ (since k is a positive integer). In other words, the sequence $\left(\overrightarrow{x_0}, \overrightarrow{x_1}, \overrightarrow{x_2}, \ldots\right)$ is k-periodic. In other words, the sequence $(x_0 \% m, x_1 \% m, x_2 \% m, \ldots)$ is k-periodic (since $\overrightarrow{x_i} = x_i \% m$ for each $i \in \mathbb{N}$). Thus, we have found a $k \in \left\{1, 2, \ldots, m^2\right\}$ such that the sequence $(x_0 \% m, x_1 \% m, x_2 \% m, \ldots)$ is k-periodic. This solves Exercise 6.2.3. $\square$

The condition $b \perp m$ in Exercise 6.2.3 is important. For example, if $(x_0, x_1, x_2, \ldots)$ is the $(1,2)$-recurrent sequence with starting values $x_0 = 0$ and $x_1 = 1$, then the sequence $(x_0 \% 2, x_1 \% 2, x_2 \% 2, \ldots) = (0, 1, 1, 1, \ldots)$ is not periodic at all. This does not mean that the claim of Exercise 6.2.3 will never hold when b is not coprime to m, but the precise conditions under which it can be ensured are nontrivial to ascertain.

6.2.6. The eventual image of an endofunction

We recall that if X and Y are two sets, and if $f : X \to Y$ is a map, and if U is a subset of X, then the set $f(U)$ is defined by

$$f(U) = \{f(u) \mid u \in U\}. \tag{272}$$

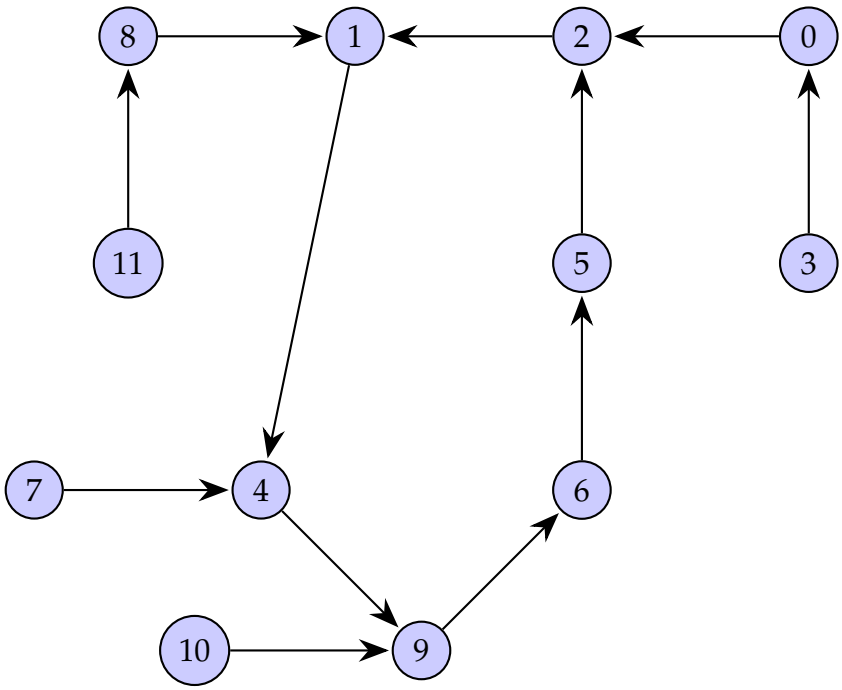
This set $f(U)$ is a subset of Y, and is called the *image* of U under f. In particular, $f(X)$ is called the *image* of f.

When f is an endofunction of a set X, we can now consider the images $f^i(X)$ for all $i \in \mathbb{N}$.

Example 6.2.11. Let $X = \{0, 1, \ldots, 11\}$. Define a map $f : X \to X$ by setting

$$f(i) = \left(i^3 + 2i + 1\right) \%12 \qquad \text{for each } i \in X.$$

The following diagram shows all elements of X (as nodes) and how the map f acts on them (by drawing an arrow from each node $i \in X$ to the node $f(i)$):



We have

$$\begin{aligned}
f^0(X) &= X = \{0,1,2,3,4,5,6,7,8,9,10,11\}, \\
f^1(X) &= f(X) = \{0,1,2,4,5,6,8,9\}, \\
f^2(X) &= \{1,2,4,5,6,9\}, \qquad \text{and} \\
f^i(X) &= \{1,2,4,5,6,9\} = f^2(X) \text{ for all } i \geq 2.
\end{aligned}$$

The pattern we might be seeing here is that the sets $f^i(X)$ get smaller and smaller as i gets larger, until they eventually stop changing at all. This and more is claimed in the following proposition:

Proposition 6.2.12. Let X be a finite set. Let $n = |X|$. Let $f : X \to X$ be a map. Then:

(a) We have $f^0(X) \supseteq f^1(X) \supseteq f^2(X) \supseteq \cdots$.

(b) If some $i \in \mathbb{N}$ satisfies $f^i(X) = f^{i+1}(X)$, then $f^i(X) = f^k(X)$ for each integer $k \geq i$.

(c) We have $f^n(X) = f^k(X)$ for each integer $k \geq n$.

(d) The map $f^n(X) \to f^n(X)$, $x \mapsto f(x)$ is well-defined and is a permutation of $f^n(X)$.

The set $f^n(X)$ in Proposition 6.2.12 is called the *eventual image* of the map f (since each element of X eventually ends up in $f^n(X)$ if we apply f to it often enough).

Proof of Proposition 6.2.12. **(a)** We need to show that $f^i(X) \supseteq f^{i+1}(X)$ for each $i \in \mathbb{N}$.

Indeed, let $i \in \mathbb{N}$. Fix some $p \in f^{i+1}(X)$. Thus, $p = f^{i+1}(x)$ for some $x \in X$. Consider this x. Now, $p = f^{i+1}(x) = f^i\left(\underbrace{f(x)}_{\in X}\right) \in f^i(X)$.

Forget that we fixed p. We thus have shown that $p \in f^i(X)$ for each $p \in f^{i+1}(X)$. In other words, $f^{i+1}(X) \subseteq f^i(X)$. In other words, $f^i(X) \supseteq f^{i+1}(X)$.

Now, forget that we fixed i. We thus have shown that $f^i(X) \supseteq f^{i+1}(X)$ for each $i \in \mathbb{N}$. In other words, $f^0(X) \supseteq f^1(X) \supseteq f^2(X) \supseteq \cdots$. This proves Proposition 6.2.12 **(a)**.

(b) Let $i \in \mathbb{N}$ satisfy $f^i(X) = f^{i+1}(X)$. We must show that

$$f^i(X) = f^k(X) \qquad \text{for each integer } k \geq i. \tag{273}$$

[*Proof of (273):* We shall prove (273) by induction on k:

Induction base: We have $f^i(X) = f^i(X)$. In other words, (273) holds for $k = i$.

Induction step: Let $m \geq i$ be an integer. Assume (as the induction hypothesis) that (273) holds for $k = m$. We must show that (273) holds for $k = m + 1$. In other words, we must show that $f^i(X) = f^{m+1}(X)$.

We have assumed that (273) holds for $k = m$. In other words, we have $f^i(X) = f^m(X)$.

Proposition 6.2.12 **(a)** yields $f^0(X) \supseteq f^1(X) \supseteq f^2(X) \supseteq \cdots$. Thus, $f^m(X) \supseteq f^{m+1}(X)$. We shall now show that $f^m(X) \subseteq f^{m+1}(X)$.

Set $j = m - i$. Then, $j = m - i \in \mathbb{N}$ (since $m \geq i$) and $m = j + i$ (since $j = m - i$). Now, let $p \in f^m(X)$. Then, $p = f^m(x)$ for some $x \in X$. Consider this x. From $m = j + i$, we obtain $f^m = f^{j+i} = f^j \circ f^i$. Thus, $p = \underbrace{f^m}_{=f^j \circ f^i}(x) = \left(f^j \circ f^i\right)(x) = f^j\left(f^i(x)\right)$. But $f^i(x) \in f^i(X) = f^{i+1}(X)$; in other words, there exists some $y \in X$ such that $f^i(x) = f^{i+1}(y)$. Consider this y. Now,

$$p = f^j\left(\underbrace{f^i(x)}_{=f^{i+1}(y)}\right) = f^j\left(f^{i+1}(y)\right) = \underbrace{\left(f^j \circ f^{i+1}\right)}_{=f^{j+(i+1)}}(y) = f^{j+(i+1)}(y) = f^{m+1}(y)$$

(since $j + (i+1) = \underbrace{j+i}_{=m} + 1 = m + 1$). Thus, $p = f^{m+1}(y) \in f^{m+1}(X)$ (since $y \in X$).

Forget that we fixed p. We thus have shown that $p \in f^{m+1}(X)$ for each $p \in f^m(X)$. In other words, $f^m(X) \subseteq f^{m+1}(X)$. Combining this with $f^m(X) \supseteq f^{m+1}(X)$, we obtain $f^m(X) = f^{m+1}(X)$. Thus, $f^i(X) = f^m(X) = f^{m+1}(X)$. This completes the induction step. Thus, (273) is proved by induction.]

Having proved (273), we have proved Proposition 6.2.12 **(b)**.

(c) Note that $f^0 = \mathrm{id}_X$, so that $f^0(X) = \mathrm{id}_X(X) = X$. Hence, $\left|f^0(X)\right| = |X| = n$.

We shall first show that there exists some $i \in \{0, 1, \ldots, n\}$ such that $f^i(X) = f^{i+1}(X)$.

Indeed, assume the contrary. Thus, each $i \in \{0, 1, \ldots, n\}$ satisfies

$$f^i(X) \neq f^{i+1}(X). \tag{274}$$

Now, let $i \in \{0, 1, \ldots, n\}$. Then, Proposition 6.2.12 **(a)** yields $f^0(X) \supseteq f^1(X) \supseteq f^2(X) \supseteq \cdots$. Thus, $f^i(X) \supseteq f^{i+1}(X)$. In other words, $f^{i+1}(X)$ is a subset of $f^i(X)$. Moreover, this subset is **proper**, because (274) shows that $f^i(X) \neq f^{i+1}(X)$.

It is a well-known fact that if U is a proper subset of a finite set V, then $|U| < |V|$. Applying this to $U = f^{i+1}(X)$ and $V = f^i(X)$, we obtain $\left|f^{i+1}(X)\right| < \left|f^i(X)\right|$ (since $f^{i+1}(X)$ is a proper subset of the finite set $f^i(X)$). This entails $\left|f^{i+1}(X)\right| \leq \left|f^i(X)\right| - 1$ (since $\left|f^{i+1}(X)\right|$ and $\left|f^i(X)\right|$ are integers). In other words, $\left|f^{i+1}(X)\right| - \left|f^i(X)\right| \leq -1$.

Now, forget that we fixed i. We thus have proved the inequality $\left|f^{i+1}(X)\right| - \left|f^i(X)\right| \leq -1$ for each $i \in \{0, 1, \ldots, n\}$. Summing these inequalities over all $i \in \{0, 1, \ldots, n\}$, we obtain

$$\sum_{i=0}^{n} \left(\left|f^{i+1}(X)\right| - \left|f^i(X)\right|\right) \leq \sum_{i=0}^{n} (-1) = (n+1) \cdot (-1) = -(n+1) < -n.$$

But this contradicts

$$\begin{aligned}
\sum_{i=0}^{n} \left(\left|f^{i+1}(X)\right| - \left|f^i(X)\right|\right) &= \underbrace{\left|f^{n+1}(X)\right|}_{\geq 0} - \underbrace{\left|f^0(X)\right|}_{=n} \\
&\qquad \left(\begin{array}{c}\text{by Corollary 4.1.17,} \\ \text{applied to } u = 0 \text{ and } v = n \text{ and } a_i = \left|f^i(X)\right|\end{array}\right) \\
&\geq 0 - n = -n.
\end{aligned}$$

This contradiction shows that our assumption was false.

Hence, we have shown that there exists some $i \in \{0, 1, \ldots, n\}$ such that $f^i(X) = f^{i+1}(X)$. Consider this i.

Now, $i \leq n$ (since $i \in \{0, 1, \ldots, n\}$), so that $n \geq i$. Thus, Proposition 6.2.12 **(b)** (applied to $k = n$) yields $f^i(X) = f^n(X)$.

Now, let $k \geq n$ be an integer. Then, $k \geq n \geq i$. Hence, Proposition 6.2.12 **(b)** yields $f^i(X) = f^k(X)$. Hence, $f^k(X) = f^i(X) = f^n(X)$. In other words, $f^n(X) = f^k(X)$. This proves Proposition 6.2.12 **(c)**.

(d) Let $x \in f^n(X)$. Thus, $x = f^n(y)$ for some $y \in X$. Consider this y. From $x = f^n(y)$, we obtain

$$f(x) = f(f^n(y)) = \underbrace{(f \circ f^n)}_{=f^{n+1}=f^n \circ f}(y) = (f^n \circ f)(y) = f^n\left(\underbrace{f(y)}_{\in X}\right) \in f^n(X).$$

Forget that we fixed x. We thus have proved that $f(x) \in f^n(X)$ for each $x \in f^n(X)$. Hence, the map $f^n(X) \to f^n(X)$, $x \mapsto f(x)$ is well-defined. It remains to prove that this map is a permutation of $f^n(X)$.

Let us denote this map by g. We must thus show that g is a permutation of $f^n(X)$.

Indeed, we shall show that this map g is surjective. To wit, let $z \in f^n(X)$. Proposition 6.2.12 **(c)** (applied to $k = n+1$) yields $f^n(X) = f^{n+1}(X)$ (since $n+1 \geq n$). Hence, $z \in f^n(X) = f^{n+1}(X)$. In other words, there exists a $w \in X$ such that $z = f^{n+1}(w)$. Consider this w. Now,

$$z = \underbrace{f^{n+1}}_{=f \circ f^n}(w) = (f \circ f^n)(w) = f(f^n(w)). \tag{275}$$

But $f^n(w) \in f^n(X)$ (since $w \in X$). Hence, $g(f^n(w))$ is well-defined (since g is a map defined on $f^n(X)$). The definition of g yields $g(f^n(w)) = f(f^n(w))$. Comparing this with (275), we obtain $z = g(f^n(w))$. Therefore, z lies in the image of g.

Forget that we fixed z. We thus have shown that each $z \in f^n(X)$ lies in the image of g. Hence, the image of g is the entire set $f^n(X)$. In other words, the map g is surjective.

Now, the set $f^n(X)$ is finite; the map $g : f^n(X) \to f^n(X)$ is surjective. Hence, Corollary 6.2.9 **(b)** (applied to $f^n(X)$ and g instead of X and f) shows that g is a permutation of $f^n(X)$. This completes our proof of Proposition 6.2.12 **(d)**. □

6.3. Homework set #6: Extremal and pigeonhole principles

This is a **regular** problem set. See Section 3.7 for details on grading.

This homework set covers the above parts of Chapter 4 and Chapter 5. Some of the problems may be unrelated.

Please solve **at most 5** problems. (No points will be given for further solutions.)

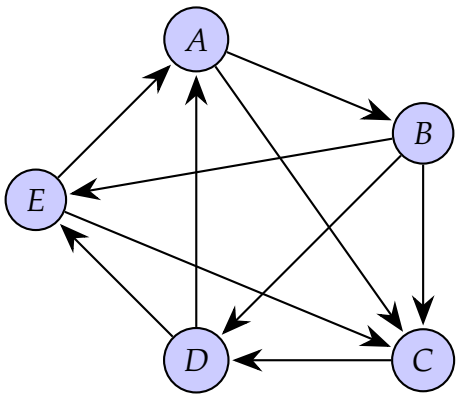
Exercise 6.3.1. Let n be a positive integer. Consider a round-robin tournament in which n players participate. ("Round-robin" means that each pair of distinct players play exactly one match against one another.) No match ends with a draw.

A player a is said to have *directly owned* a player b if a has won the match against b.

A player a is said to have *indirectly owned* a player b if there exists a player c such that a has won a match against c and c has won a match against b.

Prove that there exists a player who has (directly or indirectly) owned all other players.

[**Example:** Consider a tournament between 5 players A, B, C, D, E encoded by the following diagram:



(where an arrow from player p to player q means that p has won the match against q). Then, player A has directly owned players B and C and indirectly owned players D (via C) and E (via B). Actually, every player other than C has directly or indirectly owned all other players.]

The following exercise is a slight generalization of Exercise 5.3.2:

Exercise 6.3.2. Let $n \in \mathbb{N}$. Assume that $a_1, a_2, \ldots, a_{2n+1}$ are $2n+1$ integers with the following property:

> *Weaker splitting property:* If any of the first $2n$ numbers $a_1, a_2, \ldots, a_{2n}$ is removed (from our $2n+1$ numbers $a_1, a_2, \ldots, a_{2n+1}$), then the remaining $2n$ numbers (including a_{2n+1}) can be split into two equinumerous heaps with equal sum. ("Equinumerous" means that each heap contains exactly n numbers.)

Prove that all $2n+1$ numbers $a_1, a_2, \ldots, a_{2n+1}$ are equal.

The following exercise generalizes Theorem 6.2.10:

Exercise 6.3.3. Let X be a finite set. Let $n = |X|$. Let $f : X \to X$ be a map. Let $x \in X$. Let $p, q \in \mathbb{N}$. Prove that there exists some $k \in \{1, 2, \ldots, n\}$ such that $f^{k+p}(x) = f^{2k+q}(x)$.

(Theorem 6.2.10 is the particular case of Exercise 6.3.3 for $p = 0$ and $q = 0$.)

Exercise 6.3.4. Let X be a finite nonempty set. Let $n = |X|$. Let $f : X \to X$ be a map. Prove that $f^n(X) = f^{n-1}(X)$.

Exercise 6.3.5. Let X be a set. Let $f : X \to X$ be a map. Prove the following:

(a) If some $x \in X$ and $k \in \mathbb{N}$ satisfy $f^k(x) = f^{2k}(x)$, then $f^{ik}(x) = f^k(x)$ for every positive integer i.

Now, assume that X is finite, and let $n = |X|$. Then:

(b) We have $f^{n!} = f^{2n!}$.

(c) If f is a permutation of X, then $f^{n!} = \mathrm{id}_X$.

Exercise 6.3.6. Complete the above solution to Exercise 6.2.3 by proving Claims 1, 2 and 3.

Exercise 6.3.7. Improve Exercise 6.2.3 as follows:

Let $m > 1$ be an integer. Let a and b be integers such that $b \perp m$. Let $(x_0, x_1, x_2, \ldots)$ be any (a, b)-recurrent sequence. Prove that the sequence $(x_0 \% m, x_1 \% m, x_2 \% m, \ldots)$ is k-periodic for some $k \in \{1, 2, \ldots, m^2 - 1\}$.

Exercise 6.3.8. Let $n \geq 3$ be an integer. Let $a_1, a_2, \ldots, a_{n-1}$ be any $n-1$ integers. Assume that $n \nmid a_1 - a_2$. Prove that there exists a nonempty subset I of $\{1, 2, \ldots, n-1\}$ such that

$$n \mid \sum_{i \in I} a_i.$$

The following exercise uses a (very) slight bit of analysis:

Exercise 6.3.9. For every $x \in \mathbb{R}$, we define the *fractional part* $\operatorname{frac} x$ of x to be the number $x - \lfloor x \rfloor$. Note that $0 \leq \operatorname{frac} x < 1$ for each $x \in \mathbb{R}$. Prove the following:

(a) For each $x \in \mathbb{R}$ and each positive integer n, there exists a positive integer m such that $\operatorname{frac}(mx)$ is either $< \dfrac{1}{n}$ or $> \dfrac{n-1}{n}$. (The words "either/or" are meant non-exclusively here; i.e., it is allowed for $\operatorname{frac}(mx)$ to be both $< \dfrac{1}{n}$ and $> \dfrac{n-1}{n}$ at the same time. Of course, this will only happen for $n = 1$.)

(b) For each $x \in \mathbb{R}$ and each positive integer n, there exists a positive integer m such that $\operatorname{frac}(mx) < \dfrac{1}{n}$.

(c) For each $x \in \mathbb{R}$ and each positive real ε, there exists a positive integer m such that $\operatorname{frac}(mx) < \varepsilon$.

(d) For any positive real ε and any real z, there exists a positive integer m such that $0 \leq \sin(mz) < \varepsilon$.

[**Hint:** For part **(a)**, subdivide the half-open interval $[0, 1)$ into n intervals

$$\left[\frac{0}{n}, \frac{1}{n}\right), \quad \left[\frac{1}{n}, \frac{2}{n}\right), \quad \ldots, \quad \left[\frac{n-1}{n}, \frac{n}{n}\right),$$

and argue that two of the numbers $\operatorname{frac}(0x), \operatorname{frac}(1x), \ldots, \operatorname{frac}(nx)$ must lie in one of these n intervals. If these are $\operatorname{frac}(ix)$ and $\operatorname{frac}(jx)$ (with $i < j$), then what can you say about $\operatorname{frac}(jx - ix)$?

For part **(b)**, use part **(a)** as a stepping stone. If p is a positive integer satisfying $\operatorname{frac}(px) > \dfrac{n-1}{n}$, then find a $q > p$ such that $\operatorname{frac}(qx)$ is either $< \dfrac{1}{n}$ or $> \operatorname{frac}(px)$ (again using part **(a)**).]

Exercise 6.3.10. Let $R = \left\{\dfrac{1}{1}, \dfrac{1}{2}, \dfrac{1}{3}, \ldots\right\}$. Find the smallest positive real ε such that the entire set R can be covered with 5 **closed** intervals of length ε each.

7. Mostly Enumerative Combinatorics

We now come to the topic of *enumerative combinatorics* – the art of finding the cardinalities of finite sets (and some related quantities, like sums over finite sets), also known as "counting". We have seen some examples of this already – e.g., in Theorem 4.3.12, which states that the number of k-element subsets of a fixed n-element set is $\binom{n}{k}$. We will now see some more, even though we will merely scratch the surface.

For readers wishing to dive deeper, there is a lot to read. Several books have been written about enumerative combinatorics, such as Stanley's quasi-encyclopedic two-volume treatise [Stanle11] and [Stanle01] (with hundreds of problems). Some introductions to this subject are [Aigner07], [AndFen04], [Bogart17], [Bona17], [Wagner05], [Wagner08], [Galvin17], [Harju11], [Mazur10] and (my favorite, due to its readability and comprehensiveness) [Loehr11]. A bigger list of texts can be found at `https://math.stackexchange.com/questions/1454339` . Engel's book also has a chapter on enumerative combinatorics [Engel98, Chapter 5]. Finally, [19fco] (work in progress) aims to be a (probably over-)rigorous introduction with all detail one might want to see.

7.1. The basic principles

Let us first state the basic principles of counting that underlie most of the proofs in enumerative combinatorics:

Theorem 7.1.1 (Bijection principle)**.** Let X and Y be two sets. Then, $|X| = |Y|$ if and only if there exists a bijection from X to Y.

Theorem 7.1.2 (Sum rule)**.** If $S_1, S_2, \ldots, S_k$ are k disjoint finite sets, then the set $S_1 \cup S_2 \cup \cdots \cup S_k$ is finite and satisfies

$$|S_1 \cup S_2 \cup \cdots \cup S_k| = |S_1| + |S_2| + \cdots + |S_k| . \tag{276}$$

(Note that "disjoint" means "pairwise disjoint"; i.e., the k sets $S_1, S_2, \ldots, S_k$ are said to be disjoint if and only if every two distinct elements i and j of $\{1, 2, \ldots, k\}$ satisfy $S_i \cap S_j = \varnothing$.)

Remark 7.1.3. A weaker version of Theorem 7.1.2 holds even if we don't require $S_1, S_2, \ldots, S_k$ to be disjoint: If $S_1, S_2, \ldots, S_k$ are any k finite sets, then the set $S_1 \cup S_2 \cup \cdots \cup S_k$ is finite and satisfies

$$|S_1 \cup S_2 \cup \cdots \cup S_k| \leq |S_1| + |S_2| + \cdots + |S_k| . \tag{277}$$

This inequality (277) becomes an equality if and only if the k sets $S_1, S_2, \ldots, S_k$ are disjoint.

Theorem 7.1.4 (Product rule)**.** If $A_1, A_2, \ldots, A_n$ are finite sets, then the set $A_1 \times A_2 \times \cdots \times A_n$ is finite and satisfies

$$|A_1 \times A_2 \times \cdots \times A_n| = |A_1| \cdot |A_2| \cdot \cdots \cdot |A_n| . \tag{278}$$

Remark 7.1.5. The set $A_1 \times A_2 \times \cdots \times A_n$ in Theorem 7.1.4 is the *Cartesian product* of the n sets $A_1, A_2, \ldots, A_n$; it is defined to be the set of all n-tuples $(a_1, a_2, \ldots, a_n)$ with

$$a_1 \in A_1 \qquad \text{and} \qquad a_2 \in A_2 \qquad \text{and} \qquad \cdots \qquad \text{and} \qquad a_n \in A_n .$$

If $n = 0$, then this Cartesian product $A_1 \times A_2 \times \cdots \times A_n$ is (by convention) a one-element set, consisting only of the 0-tuple $()$. This latter 0-tuple $()$ is also known as the *empty list*.

Theorem 7.1.6 (Power rule)**.** If A is a finite set, and if $n \in \mathbb{N}$, then the set A^n is finite and satisfies

$$|A^n| = |A|^n . \tag{279}$$

Remark 7.1.7. The set A^n in Theorem 7.1.6 is defined as the Cartesian product $\underbrace{A \times A \times \cdots \times A}_{n \text{ times}}$; it is known as the *n-th Cartesian power* of A. It consists of the n-tuples $(a_1, a_2, \ldots, a_n)$ whose entries all belong to A.

Theorem 7.1.8 (Difference rule)**.** If A is a finite set, and if B is a subset of A, then the set $A \setminus B$ is finite and satisfies

$$|A \setminus B| = |A| - |B| . \tag{280}$$

All the above theorems are basic facts and can be used without proof.[175]

[175] For the sake of completeness, here are some references to their proofs (except for Theorem 7.1.1, which is truly fundamental and even used to define the size of a set by some authors):

- Theorem 7.1.2 is [Loehr11, 1.2]. It can easily be derived from Theorem 2.3.6 by induction on n. (Theorem 2.3.6, on the other hand, can be proved, e.g., by induction on $|Q|$; but the details boil down to how the size of a set is defined.)
- Remark 7.1.3 can also be proved by induction on k, using the fact that $|P \cup Q| \leq |P| + |Q|$ for any two finite sets P and Q. (The latter fact, in turn, can be shown by induction on $|Q|$.)
- Theorem 7.1.4 is [Loehr11, 1.5]. It can be proved by induction on n, using the fact that $|P \times Q| = |P| \cdot |Q|$ for any two finite sets P and Q. The latter fact can be obtained from Theorem 7.1.2 as follows: If we denote the elements of P by $p_1, p_2, \ldots, p_k$ (where $k = |P|$),

7.2. Notations

The following two conventions will be used throughout Chapter 7:

Convention 7.2.1. If $n \in \mathbb{N}$, then $[n]$ shall mean the n-element set $\{1, 2, \ldots, n\}$. (Thus, in particular, $[1] = \{1\}$ and $[0] = \{\} = \varnothing$.)

This convention might appear to clash with the Iverson bracket notation for truth values (Definition 4.3.19); but in practice, there is never a chance of confusion. (Convention 7.2.1 only defines $[n]$ when n is a number, whereas Definition 4.3.19 only defines $[\mathcal{A}]$ when $\mathcal{A}$ is a logical statement.)

Convention 7.2.2. The symbol "#" shall mean the word "number" (or "the number"). For example, "# of subsets of $\{1, 2, 3\}$" means "the number of subsets of $\{1, 2, 3\}$".

Example 7.2.3. We have $[7] = \{1, 2, 3, 4, 5, 6, 7\}$ and thus

$$\{\text{the odd numbers in } [7]\} = \{1, 3, 5, 7\}.$$

Hence,

$$(\text{\# of odd numbers in } [7]) = 4.$$

7.3. Elementary examples

We shall now solve some elementary counting problems to illustrate the use of the above theorems.

7.3.1. Subsets

We have already seen (in Theorem 4.3.12) how many k-element subsets a given n-element set has. Now, let us count all subsets (of all possible sizes) of a given n-element set:

then $P \times Q$ is the union of the k disjoint sets

$$\{p_1\} \times Q, \quad \{p_2\} \times Q, \quad \ldots, \quad \{p_k\} \times Q,$$

each of which has size $|Q|$ (by Theorem 7.1.1). The details are left to the reader.

- Theorem 7.1.6 follows by applying Theorem 7.1.4 to $A_i = A$.
- Theorem 7.1.8 is [Loehr11, 1.3] and [19fco, Theorem 1.4.7 **(a)**]. It is an easy consequence of Theorem 2.3.6.

Theorem 7.3.1. Let $n \in \mathbb{N}$. Let S be an n-element set. Then,

$$(\#\text{ of subsets of } S) = 2^n.$$

Proof of Theorem 7.3.1 (sketched). We shall only give a brief outline; see [19fco, Theorem 1.4.1] for details. (More precisely, the proof we are sketching here is the Third proof of [19fco, Theorem 1.4.1]; it appears in [19fco, §1.5.3]. Two other proofs are given in [19fco, solution to Exercise 1.4.1] and [19fco, §1.4.2].)

We denote the n elements of S by $s_1, s_2, \ldots, s_n$ (in some order).

If I is any subset of S, then the n-tuple[176]

$$([s_1 \in I], [s_2 \in I], \ldots, [s_n \in I]) \in \{0,1\}^n$$

"encodes" the set I: Indeed, we can read I off this n-tuple, since the i-th entry of this n-tuple tells us whether s_i lies in I or not. Thus, the map

$$\begin{aligned} f : \{\text{subsets of } S\} &\to \{0,1\}^n, \\ I &\mapsto ([s_1 \in I], [s_2 \in I], \ldots, [s_n \in I]) \end{aligned}$$

is injective. This map f is also surjective, because any n-tuple $(a_1, a_2, \ldots, a_n) \in \{0,1\}^n$ is taken as a value by f. (In fact, if $(a_1, a_2, \ldots, a_n) \in \{0,1\}^n$ is any n-tuple, then $(a_1, a_2, \ldots, a_n) = f(I)$, where I is the subset $\{s_i \mid i \in [n] \text{ satisfying } a_i = 1\}$ of S.)

Thus, the map f is bijective (since it is surjective and injective). In other words, f is a bijection. Hence, there exists a bijection from $\{\text{subsets of } S\}$ to $\{0,1\}^n$. The bijection principle (Theorem 7.1.1) therefore yields

$$\begin{aligned} |\{\text{subsets of } S\}| &= \left|\{0,1\}^n\right| = |\{0,1\}|^n \qquad (\text{by } (279)) \\ &= 2^n \qquad (\text{since } |\{0,1\}| = 2). \end{aligned}$$

In other words, $(\#\text{ of subsets of } S) = 2^n$. This proves Theorem 7.3.1. □

7.3.2. Integer compositions

Next, we will count some objects called *compositions* (or, to be more precise, *integer compositions*). These have nothing to do with compositions of maps; they are tuples of positive integers:[177]

Definition 7.3.2. **(a)** A *composition* (or, to be more precise, an *integer composition*) means a tuple of positive integers.

(b) Let $n \in \mathbb{N}$. A *composition of n* means a tuple of positive integers whose sum is n.

[176] See Definition 4.3.19 for the meaning of the square brackets.

[177] Recall that a *tuple* means a finite ordered list (of any kind of objects).

Example 7.3.3. **(a)** The tuple $(3,1,1,4)$ is a composition. It is a composition of 9, since $3+1+1+4=9$.

(b) The tuple $(3,0,1,4)$ is not a composition, since 0 is not a positive integer.

(c) The compositions of 3 are

$$(3), \qquad (2,1), \qquad (1,2), \qquad (1,1,1).$$

Note that $(2,1)$ and $(1,2)$ are not the same composition, since tuples are ordered.

Theorem 7.3.4. Let $n \in \mathbb{N}$. Then,

$$(\#\text{ of compositions of } n) = \begin{cases} 2^{n-1}, & \text{if } n \geq 1; \\ 1, & \text{if } n = 0. \end{cases}$$

Proof of Theorem 7.3.4 (sketched). We shall outline the proof; the details can be found in [19fco-hw0s, Exercise 1 **(b)**].

The case $n=0$ is easy (the only composition of 0 is the 0-tuple $()$). Thus, for the rest of this proof, we WLOG assume that $n \neq 0$. Hence, $n \geq 1$. We thus need to show that $(\#\text{ of compositions of } n) = 2^{n-1}$. Inspired by the bijection principle, we thus would like to define a bijection

$$\{\text{compositions of } n\} \to \{\text{subsets of } [n-1]\},$$

where (as we recall from Convention 7.2.1) we have $[n-1] = \{1,2,\ldots,n-1\}$. Indeed, once we have found such a bijection, we can apply the bijection principle to obtain

$$\begin{aligned} |\{\text{compositions of } n\}| &= |\{\text{subsets of } [n-1]\}| \\ &= (\#\text{ of subsets of } [n-1]) = 2^{n-1} \end{aligned}$$

(by Theorem 7.3.1 (applied to $[n-1]$ and $n-1$ instead of S and n), since $[n-1]$ is an $(n-1)$-element set), so that

$$(\#\text{ of compositions of } n) = |\{\text{compositions of } n\}| = 2^{n-1}.$$

This will clearly prove Theorem 7.3.4. Thus, in order to complete this proof, all that remains to be done is to find a bijection $\{\text{compositions of } n\} \to \{\text{subsets of } [n-1]\}$.

How do we do this? If $a = (a_1, a_2, \ldots, a_k)$ is a composition of n, then consider the set

$$\begin{aligned} C(a) &:= \{a_1 + a_2 + \cdots + a_i \mid i \in \{1,2,\ldots,k-1\}\} \\ &= \{a_1, \\ &\quad a_1 + a_2, \\ &\quad a_1 + a_2 + a_3, \\ &\quad \ldots, \\ &\quad a_1 + a_2 + \cdots + a_{k-1}\}. \end{aligned}$$

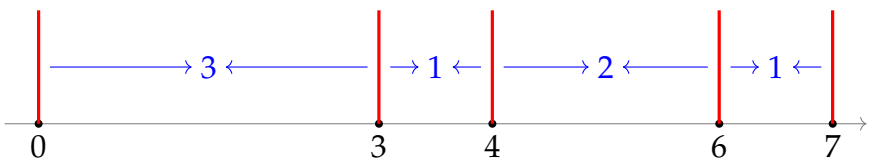
This set $C(a)$ consists of all partial sums of the tuple a (not counting the empty sum 0, and also not counting the full sum $a_1 + a_2 + \cdots + a_k$). Since $a = (a_1, a_2, \ldots, a_k)$ is a composition of n, the entries $a_1, a_2, \ldots, a_k$ are positive integers and their sum is $a_1 + a_2 + \cdots + a_k = n$; hence, all their partial sums $a_1 + a_2 + \cdots + a_i$ for $i \in \{1, 2, \ldots, k-1\}$ are larger than 0 and smaller than n. In other words, all these partial sums belong to the set $\{1, 2, \ldots, n-1\} = [n-1]$. That is, $C(a)$ is a subset of $[n-1]$ (since $C(a)$ is the set of these partial sums).

Thus, for each composition a of n, we have defined a subset $C(a)$ of $[n-1]$. We thus obtain a map

$$\begin{aligned} C : \{\text{compositions of } n\} &\to \{\text{subsets of } [n-1]\}, \\ a &\mapsto C(a). \end{aligned}$$

For example, if $n = 7$ and $a = (3, 1, 2, 1)$, then $C(a) = \{3,\ 3+1,\ 3+1+2\} = \{3, 4, 6\}$.

Now, we claim that this map C is a bijection. Indeed, its inverse map C^{-1} can be constructed rather easily: If I is any subset of $[n-1]$, then the elements of I subdivide the interval $[0, n]$ into several blocks; the lengths of these blocks (from left to right) form a composition of n (since they are positive integers summing to n), and this composition is precisely $C^{-1}(I)$. For example, if $n = 7$ and $I = \{3, 4, 6\}$, then the elements of I subdivide the interval $[0, n] = [0, 7]$ into 4 blocks as follows:



(where the lengths of the blocks are shown in blue, and where the red vertical lines separate the blocks), and the lengths of these 4 blocks are $3, 1, 2, 1$ (from left to right), so that $C^{-1}(I) = (3, 1, 2, 1)$. Proving that this construction really defines an inverse map to C is straightforward and left to the reader (see [19fco-hw0s, solution to Exercise 1 **(b)**], where it is done rigorously and with no reference to pictures).

Thus, the map $C : \{\text{compositions of } n\} \to \{\text{subsets of } [n-1]\}$ is a bijection. So we have found a bijection $\{\text{compositions of } n\} \to \{\text{subsets of } [n-1]\}$. As we explained above, this completes the proof of Theorem 7.3.4. $\square$

7.3.3. Maps

Let us next count maps (i.e., functions) between two sets.

Definition 7.3.5. Let A and B be two sets. Then, B^A shall mean the set of all maps from A to B.

Theorem 7.3.6. Let A and B be two finite sets. Then, the set B^A is finite, and satisfies

$$\left|B^A\right| = |B|^{|A|}. \tag{281}$$

The equality (281) can be rewritten as

$$(\text{\# of maps from } A \text{ to } B) = |B|^{|A|}. \tag{282}$$

Proof of Theorem 7.3.6 (sketched). This is an outline of the proof; see [19fco, proof of Theorem 1.5.7] for details.

Let us denote the elements of A by $a_1, a_2, \ldots, a_n$ (in any order). Thus, $|A| = n$. Now, any map $f : A \to B$ is uniquely determined by the n-tuple $(f(a_1), f(a_2), \ldots, f(a_n)) \in B^n$ (which is simply its list of values). Conversely, any n-tuple $(b_1, b_2, \ldots, b_n) \in B^n$ can be obtained as the list $(f(a_1), f(a_2), \ldots, f(a_n))$ of values of some map $f : A \to B$. Thus, we have a bijection

$$\begin{aligned} B^A &\to B^n, \\ f &\mapsto (f(a_1), f(a_2), \ldots, f(a_n)) \end{aligned}$$

(since B^A is the set of all maps from A to B). Hence, the bijection principle yields

$$\begin{aligned} \left|B^A\right| &= |B^n| = |B|^n \qquad \text{(by the power rule)} \\ &= |B|^{|A|} \qquad (\text{since } n = |A|). \end{aligned}$$

This proves Theorem 7.3.6. □

The next problem is similar to Putnam 1985 problem A1:[178]

Exercise 7.3.1. Let $n \in \mathbb{N}$. Find the # of all triples (A, B, C) of subsets of $[n]$ satisfying $A \cap B \cap C = \varnothing$.

Discussion of Exercise 7.3.1. The condition $A \cap B \cap C = \varnothing$ says that no element belongs to A, B and C at the same time.

Let us construct a triple (A, B, C) of subsets of $[n]$ satisfying $A \cap B \cap C = \varnothing$. It may appear natural to do this by first choosing A, then choosing B, then choosing C. But this is not a good way to count these triples, as the number of options for C depends on the choices of A and B. We thus take a different way. We construct a triple (A, B, C) of subsets of $[n]$ satisfying $A \cap B \cap C = \varnothing$ as follows:

- We decide which of the three sets A, B, C will contain 1. There are 7 options[179] for this decision; namely, the options are "only A", "only B", "only C", "both A and B", "both A and C", "both B and C" and "none of A, B and C". (The option "all three of A, B and C" is unavailable, since this would contradict $A \cap B \cap C = \varnothing$.)

[178] Recall that $[n] = \{1, 2, \ldots, n\}$ for $n \in \mathbb{N}$.

[179] What we call "option" here is often called "choice" in the literature. See Convention 7.3.7 below for a discussion of this concept.

- We decide which of the three sets A, B, C will contain 2. Again, there are 7 options for this decision.
- We decide which of the three sets A, B, C will contain 3. Again, there are 7 options for this decision.
- And so on, until we have placed each of the n elements $1, 2, \ldots, n$ of $[n]$ in the sets A, B, C (or, rather, in whichever of these sets A, B, C we want to place it in).

Altogether, we thus make n decisions, and each time there are 7 available options. Thus, the total # of possibilities (for the triple (A, B, C) we are constructing) is $\underbrace{7 \cdot 7 \cdot \cdots \cdot 7}_{n \text{ times}} = 7^n$. So the # of all triples (A, B, C) of subsets of $[n]$ satisfying $A \cap B \cap C = \varnothing$ is 7^n.

This wasn't very rigorous: what exactly is a "decision", what is an "option", what is a "possibility", and why are we multiplying the 7's? The answer is that our informal argument above was informal code for an application of the bijection principle. Here is an outline of how to formalize it: Define the map

$$\begin{aligned} \{\text{triples } (A, B, C) \text{ of subsets of } [n] \text{ satisfying } A \cap B \cap C = \varnothing\} &\to \{1, 2, \ldots, 7\}^n, \\ (A, B, C) &\mapsto (u_1, u_2, \ldots, u_n), \end{aligned}$$

where

$$u_i = \begin{cases} 1, & \text{if } i \notin A \text{ and } i \notin B \text{ and } i \notin C; \\ 2, & \text{if } i \in A \text{ and } i \notin B \text{ and } i \notin C; \\ 3, & \text{if } i \notin A \text{ and } i \in B \text{ and } i \notin C; \\ 4, & \text{if } i \notin A \text{ and } i \notin B \text{ and } i \in C; \\ 5, & \text{if } i \in A \text{ and } i \in B \text{ and } i \notin C; \\ 6, & \text{if } i \in A \text{ and } i \notin B \text{ and } i \in C; \\ 7, & \text{if } i \notin A \text{ and } i \in B \text{ and } i \in C \end{cases} \qquad \text{for each } i \in [n]$$

(so that u_i essentially encodes which of the sets A, B, C contain i). It is not hard to see that this map is a bijection. Thus, the bijection principle entails

$$\begin{aligned} &|\{\text{triples } (A, B, C) \text{ of subsets of } [n] \text{ satisfying } A \cap B \cap C = \varnothing\}| \\ &= \left|\{1, 2, \ldots, 7\}^n\right| = |\{1, 2, \ldots, 7\}|^n \qquad \text{(by the power rule)} \\ &= 7^n. \end{aligned}$$

Thus, we recover our answer to Exercise 7.3.1 with a formal proof. □

For the future, let me clarify my use of certain words:

Convention 7.3.7. (a) I am using the three words "decision", "option" and "possibility" for what most authors (including Loehr in [Loehr11]) call "choice". In my opinion, using the word "choice" for all three concepts is misleading, as they are not the same. Let me illustrate the differences on an example:

Imagine you are about to buy a car. You get to choose its color ("black", "grey" or "red") and its body style ("sedan", "minivan" or "convertible"). All pairs of a color and a body style are available. Thus:

- you make 2 *decisions* (namely, the color and the body style);
- you have 3 *options* in your first decision (namely, "black", "grey" and "red");
- you have 3 *options* in your second decision (namely, "sedan", "minivan" and "convertible");
- you have a total of 9 *possibilities* (namely, "black sedan", "black minivan", "black convertible", "grey sedan", etc.).

The general rule is that you are choosing between a number of *options* in each *decision*; and the total combination of options you have chosen after you made all decisions is a *possibility*.

By avoiding the word "choice", I am also eliminating the source of confusion that the standard idioms "You have one choice" and "You have no choice" cause. (In my terminology, these would translate as "You have one decision, which allows for at least 2 options" and "You have 1 option".) I hope the pedantry will pay off as we get to more complicated proofs where any bit of clarity will be helpful.

(b) Assume you are making several decisions (as in the example above, where you are buying a car). The decisions are said to be *independent* if the set of options you get to choose from in one decision does not depend on the options you have chosen in the others. Otherwise, they are *dependent*. For example, in the above example, the two decisions were independent; however, if a red sedan becomes unavailable, then they become dependent, since choosing "red" in the first decision makes the "sedan" option in the second decision unavailable. (The decision-making process in our above discussion of Exercise 7.3.1 was an example where we had n independent decisions, each allowing for 7 options.)

7.3.4. Injective maps

Counting all maps between two given sets was easy. What about injective maps?

Theorem 7.3.8. Let $m, n \in \mathbb{N}$. Let A be an m-element set. Let B be an n-element

set. Then,

$$\begin{aligned}
&(\text{\# of all injective maps from } A \text{ to } B) \\
&= n(n-1)(n-2)\cdots(n-m+1) \qquad (283)\\
&= m! \cdot \binom{n}{m}. \qquad (284)
\end{aligned}$$

Remark 7.3.9. Let m, n, A, B be as in Theorem 7.3.8. If $m > n$, then the product $n(n-1)(n-2)\cdots(n-m+1)$ contains the factor $n-n=0$, and thus must itself be 0. Thus, Theorem 7.3.8 shows that there are **no** injective maps from A to B if $m > n$. This is precisely what Theorem 6.1.1 tells us.

Proof of Theorem 7.3.8 (sketched). We shall give an informal proof, and later explain what it relies on.

First of all, we note that $m = |A| \in \mathbb{N}$ and thus

$$\binom{n}{m} = \frac{n(n-1)(n-2)\cdots(n-m+1)}{m!}$$

(by (117), applied to $k = m$). Hence,

$$m! \cdot \binom{n}{m} = n(n-1)(n-2)\cdots(n-m+1). \qquad (285)$$

We denote the m elements of A by $a_1, a_2, \ldots, a_m$ (in some order). Now, if $f : A \to B$ is an injective map, then the m values $f(a_1), f(a_2), \ldots, f(a_m)$ are distinct (i.e., no two of them are equal), so that they satisfy

$$\begin{aligned}
&f(a_1) \in B, \\
&f(a_2) \in B \setminus \{f(a_1)\}, \\
&f(a_3) \in B \setminus \{f(a_1), f(a_2)\}, \\
&f(a_4) \in B \setminus \{f(a_1), f(a_2), f(a_3)\}, \\
&\qquad \ldots.
\end{aligned}$$

Hence, we can construct an injective map $f : A \to B$ as follows:

- Choose the value $f(a_1) \in B$. We have n options for this (since B has n elements).
- Choose the value $f(a_2) \in B \setminus \{f(a_1)\}$. We have $n-1$ options for this (since $B \setminus \{f(a_1)\}$ has $n-1$ elements).
- Choose the value $f(a_3) \in B \setminus \{f(a_1), f(a_2)\}$. We have $n-2$ options for this (since $B \setminus \{f(a_1), f(a_2)\}$ has $n-2$ elements[180]).

[180] This is because $f(a_1)$ and $f(a_2)$ are distinct (because of how we chose $f(a_2)$).

- Choose the value $f(a_4) \in B \setminus \{f(a_1), f(a_2), f(a_3)\}$. We have $n-3$ options for this (since $B \setminus \{f(a_1), f(a_2), f(a_3)\}$ has $n-3$ elements[181]).
- And so on, until we have chosen the last value $f(a_m)$ (there are $n-m+1$ options for it).

Note that the options that are available to us at each step of this procedure depend on the previously made decisions (so our decisions are dependent); however, the **number** of options at each step does not. Namely, we have n options in our first decision, $n-1$ options in our second decision, $n-2$ options in our third decision, and so on. Hence, in total, we have $n(n-1)(n-2)\cdots(n-m+1)$ possibilities for how we can make these choices. Since each of these possibilities leads to a different injective map $f : A \to B$, and since each injective map $f : A \to B$ can be obtained from one of these possibilities, we thus conclude that

$$\begin{aligned}(\#\text{ of all injective maps from } A \text{ to } B) &= n(n-1)(n-2)\cdots(n-m+1) \\ &= m! \cdot \binom{n}{m} \qquad (\text{by (285)}).\end{aligned}$$

This proves Theorem 7.3.8 (if you believe in our informal argument using "choosing values"). □

Our above proof of Theorem 7.3.8 used the following fact, which we shall state in an informal way here:

Theorem 7.3.10 (Dependent product rule)**.** Consider a situation in which you have to make n decisions (in order). Assume that

- you have a_1 options in decision 1;
- you have a_2 options in decision 2 (no matter what option you chose in decision 1);
- you have a_3 options in decision 3 (no matter what options you chose in decisions 1 and 2);
- ...;
- you have a_n options in decision n (no matter what options you chose in the previous decisions).

Then, the total # of possibilities for how you can make these choices is $a_1 a_2 \cdots a_n$.

This can be made rigorous (see [Newste20, Theorem 7.2.19] or [Loehr11, §1.8]). Note that we have applied Theorem 7.3.10 to m instead of n in our proof of Theorem

[181] This is because $f(a_1)$, $f(a_2)$ and $f(a_3)$ are distinct (because of how we chose $f(a_2)$ and $f(a_3)$).

7.3.8 (since the procedure by which we constructed an injective map $f : A \to B$ required m decisions).

There is also an alternative way of formalizing the above proof of Theorem 7.3.8, without explicitly mentioning the Dependent product rule. Namely, instead of constructing an injective map $f : A \to B$ by choosing its values $f(a_1), f(a_2), \ldots, f(a_m)$ in order, we fix an element a of A (unless A is empty, in which case Theorem 7.3.8 is rather trivial), and we construct an injective map $f : A \to B$ by first choosing its value $f(a)$ and then choosing all remaining values at the same time (which is tantamount to choosing an injective map from $A \setminus \{a\}$ to $B \setminus \{f(a)\}$, because the values to be chosen must differ from $f(a)$). This procedure involves only two steps, and thus is easier to formalize. There are always exactly n options at the first step, whereas the # of options at the second step is the # of injective maps from $A \setminus \{a\}$ to $B \setminus \{f(a)\}$, which we can obtain from our induction hypothesis if we induct on $|A|$. See [19fco, §2.4.2] for the details of this argument.

7.3.5. Tuples with non-repetition requirements

The next definition codifies three different ways in which tuples can avoid having repeated elements:

Definition 7.3.11. A tuple $(a_1, a_2, \ldots, a_k)$ (of any kind of objects) is said to be

- *injective* if $a_1, a_2, \ldots, a_k$ are distinct;
- *Smirnov* (or *Carlitz* or *non-stuttering*) if it satisfies $a_i \neq a_{i+1}$ for each $i \in \{1, 2, \ldots, k-1\}$;
- *cyc-Smirnov* if it satisfies both
  - the relation $a_i \neq a_{i+1}$ for each $i \in \{1, 2, \ldots, k-1\}$ and
  - the relation $a_k \neq a_1$ (assuming that $k > 0$).

We are using the proper names "Smirnov" and "Carlitz" as adjectives here. "Cyc-Smirnov" is short for "cyclically Smirnov". One can think of Smirnov tuples as tuples that can have repeated entries, but never in consecutive positions. Cyc-Smirnov tuples are defined in the same way, except that the last and the first positions also count as consecutive (which is natural if one imagines the tuple being written on a round band).

Clearly, any injective tuple with at least one entry is cyc-Smirnov; also, every cyc-Smirnov tuple is Smirnov.

Example 7.3.12. **(a)** The tuple $(5, 1, 2)$ is injective (and thus Smirnov and cyc-Smirnov).

(b) The tuple $(2, 1, 2)$ is not injective and not cyc-Smirnov, but it is Smirnov.

(c) The tuple $(2, 1, 2, 1)$ is cyc-Smirnov (and thus Smirnov), but not injective.

(d) The tuple $(2,2,1)$ is not Smirnov (and thus neither cyc-Smirnov nor injective).

Let us now count these tuples:

Exercise 7.3.2. Let $n,k \in \mathbb{N}$. Let A be an n-element set.
(a) How many injective k-tuples $(a_1,a_2,\ldots,a_k) \in A^k$ exist?
(b) Assume that $k > 0$. How many Smirnov k-tuples $(a_1,a_2,\ldots,a_k) \in A^k$ exist?
(c) Assume that $k > 0$. How many cyc-Smirnov k-tuples $(a_1,a_2,\ldots,a_k) \in A^k$ exist?

Discussion of Exercise 7.3.2. **(a)** Recall the notation $[k] = \{1,2,\ldots,k\}$ (from Convention 7.2.1). Now, we claim that the injective k-tuples $(a_1,a_2,\ldots,a_k) \in A^k$ are "the same as" the injective maps from $[k]$ to A. To be more precise, there is a bijection

$$\begin{aligned}\{\text{injective maps from } [k] \text{ to } A\} &\to \left\{\text{injective } k\text{-tuples } (a_1,a_2,\ldots,a_k) \in A^k\right\}, \\ f &\mapsto (f(1), f(2), \ldots, f(k))\end{aligned}$$

(check this!). Hence, the bijection principle yields

$$\begin{aligned}&(\#\text{ of injective maps from } [k] \text{ to } A) \\ &= \left(\#\text{ of injective } k\text{-tuples } (a_1,a_2,\ldots,a_k) \in A^k\right),\end{aligned}$$

therefore

$$\begin{aligned}&\left(\#\text{ of injective } k\text{-tuples } (a_1,a_2,\ldots,a_k) \in A^k\right) \\ &= (\#\text{ of injective maps from } [k] \text{ to } A) \\ &= n(n-1)(n-2)\cdots(n-k+1)\end{aligned}$$

(by (283), applied to k, $[k]$ and A instead of m, A and B). This solves Exercise 7.3.2 **(a)**.

(b) To construct a Smirnov k-tuple $(a_1,a_2,\ldots,a_k) \in A^k$, we proceed as follows:

- We choose its first entry $a_1 \in A$. (There are n options.)
- We choose its second entry $a_2 \in A \setminus \{a_1\}$. (There are $n-1$ options.)
- We choose its third entry $a_3 \in A \setminus \{a_2\}$. (There are $n-1$ options.)
- We choose its fourth entry $a_4 \in A \setminus \{a_3\}$. (There are $n-1$ options.)
- And so on, until the last entry a_k has been chosen.

Thus, the Dependent product rule shows that the total # of possibilities for how we can make these choices is

$$n \underbrace{(n-1)(n-1)\cdots(n-1)}_{k-1 \text{ times}} = n(n-1)^{k-1}.$$

So the # of Smirnov k-tuples $(a_1, a_2, \ldots, a_k) \in A^k$ is $n(n-1)^{k-1}$. This solves Exercise 7.3.2 **(b)**.

(c) Forget that we fixed k. For each positive integer k, we let $c(n,k)$ denote the # of cyc-Smirnov k-tuples $(a_1, a_2, \ldots, a_k) \in A^k$. Thus, we want to compute $c(n,k)$.

First we deal with the case $k = 1$: A 1-tuple $(a_1) \in A^1$ cannot be cyc-Smirnov, since the condition "$a_k \neq a_1$" in the definition of a cyc-Smirnov tuple would boil down to the clearly impossible inequality $a_1 \neq a_1$ for a 1-tuple. Thus, there exist no cyc-Smirnov 1-tuples $(a_1) \in A^1$. In other words, $c(n,1) = 0$.

Assume now that $k > 1$. We agree to only consider tuples with entries in A in this solution. That is, "k-tuple" will always mean "k-tuple of elements of A" (and likewise for $(k-1)$-tuples).

We say that a k-tuple $(a_1, a_2, \ldots, a_k) \in A^k$ is *noncyc-Smirnov* if it is Smirnov and satisfies $a_k = a_1$. Thus, each Smirnov k-tuple $(a_1, a_2, \ldots, a_k) \in A^k$ is either cyc-Smirnov or noncyc-Smirnov (depending on whether it satisfies $a_k \neq a_1$ or $a_k = a_1$); it cannot be both at the same time. Hence, the sum rule yields

$$\begin{aligned}&(\text{\# of Smirnov } k\text{-tuples})\\ &= (\text{\# of cyc-Smirnov } k\text{-tuples}) + (\text{\# of noncyc-Smirnov } k\text{-tuples}).\end{aligned}$$

Hence,

$$\begin{aligned}&(\text{\# of cyc-Smirnov } k\text{-tuples})\\ &= (\text{\# of Smirnov } k\text{-tuples}) - (\text{\# of noncyc-Smirnov } k\text{-tuples}).\end{aligned}$$

This equality turns out to be useful, because we will soon see that both terms on its right hand side are easier to compute than the left hand side.

If $(a_1, a_2, \ldots, a_k) \in A^k$ is a noncyc-Smirnov k-tuple, then $a_{k-1} \neq a_k$ (since $(a_1, a_2, \ldots, a_k)$ is Smirnov) and therefore $a_{k-1} \neq a_k = a_1$ (since $(a_1, a_2, \ldots, a_k)$ is noncyc-Smirnov), so that the $(k-1)$-tuple $(a_1, a_2, \ldots, a_{k-1}) \in A^{k-1}$ is a cyc-Smirnov $(k-1)$-tuple. Thus, we obtain a map

$$\begin{aligned}f : \{\text{noncyc-Smirnov } k\text{-tuples}\} &\to \{\text{cyc-Smirnov } (k-1)\text{-tuples}\},\\ (a_1, a_2, \ldots, a_k) &\mapsto (a_1, a_2, \ldots, a_{k-1}).\end{aligned}$$

(All that this map f does is removing the last entry of a k-tuple.)

Conversely, if $(a_1, a_2, \ldots, a_{k-1}) \in A^{k-1}$ is a cyc-Smirnov $(k-1)$-tuple, then $(a_1, a_2, \ldots, a_{k-1}, a_1) \in A^k$ is a noncyc-Smirnov k-tuple (since its last entry equals its first entry). Thus, we obtain a map

$$\begin{aligned}g : \{\text{cyc-Smirnov } (k-1)\text{-tuples}\} &\to \{\text{noncyc-Smirnov } k\text{-tuples}\},\\ (a_1, a_2, \ldots, a_{k-1}) &\mapsto (a_1, a_2, \ldots, a_{k-1}, a_1).\end{aligned}$$

It is easy to see that the maps f and g are mutually inverse (indeed, $f \circ g = \mathrm{id}$ is clear, whereas $g \circ f = \mathrm{id}$ follows from the fact that each noncyc-Smirnov k-tuple has the form $(a_1, a_2, \ldots, a_{k-1}, a_1)$ for some $a_1, a_2, \ldots, a_{k-1} \in A$). Thus, the map f is invertible, i.e., a bijection. Hence, the bijection principle shows that

$$\begin{aligned}(\#\text{ of noncyc-Smirnov } k\text{-tuples}) &= (\#\text{ of cyc-Smirnov } (k-1)\text{-tuples}) \\ &= c(n, k-1)\end{aligned}$$

(since $c(n, k-1)$ was defined as the # of cyc-Smirnov $(k-1)$-tuples).

Now, the definition of $c(n, k)$ yields

$$\begin{aligned}c(n,k) &= (\#\text{ of cyc-Smirnov } k\text{-tuples}) \\ &= \underbrace{(\#\text{ of Smirnov } k\text{-tuples})}_{\substack{=n(n-1)^{k-1} \\ \text{(by Exercise 7.3.2 \textbf{(b)})}}} - \underbrace{(\#\text{ of noncyc-Smirnov } k\text{-tuples})}_{=c(n,k-1)} \\ &= n(n-1)^{k-1} - c(n, k-1). \qquad (286)\end{aligned}$$

Forget that we fixed k. We thus have proved the equality (286) for each integer $k > 1$. Moreover, recall that $c(n, 1) = 0$. This gives us a recurrent description of the sequence $(c(n,1), c(n,2), c(n,3), \ldots)$. We attempt to solve this recurrence by

plugging the recursive equation (286) into itself:

$$
\begin{aligned}
& c(n,k) \\
& = n(n-1)^{k-1} - \underbrace{c(n,k-1)}_{\substack{=n(n-1)^{k-2}-c(n,k-2) \\ \text{(by (286))}}} \qquad \text{(by (286))} \\
& = n(n-1)^{k-1} - \left(n(n-1)^{k-2} - \underbrace{c(n,k-2)}_{\substack{=n(n-1)^{k-3}-c(n,k-3) \\ \text{(by (286))}}} \right) \\
& = n(n-1)^{k-1} - \left(n(n-1)^{k-2} - \left(n(n-1)^{k-3} - \underbrace{c(n,k-3)}_{=\cdots} \right) \right) \\
& = \cdots \\
& = n(n-1)^{k-1} - \left(n(n-1)^{k-2} - \left(n(n-1)^{k-3} - \left(\cdots - \left(n(n-1)^{1} - \underbrace{c(n,1)}_{=0} \right) \right) \right) \right) \\
& = n(n-1)^{k-1} - \left(n(n-1)^{k-2} - \left(n(n-1)^{k-3} - \left(\cdots - \left(n(n-1)^{1} \right) \right) \right) \right) \\
& = n(n-1)^{k-1} - n(n-1)^{k-2} + n(n-1)^{k-3} \pm \cdots + (-1)^{k} n(n-1)^{1} \\
& = \sum_{i=0}^{k-2} \underbrace{(-1)^{k-i}}_{=(-1)^{k}(-1)^{i}} n \underbrace{(n-1)^{i+1}}_{=(n-1)(n-1)^{i}} = (-1)^{k} n(n-1) \cdot \sum_{i=0}^{k-2} \underbrace{(-1)^{i}(n-1)^{i}}_{\substack{=((-1)(n-1))^{i} \\ =(1-n)^{i}}} \\
& = (-1)^{k} n(n-1) \cdot \underbrace{\sum_{i=0}^{k-2} (1-n)^{i}}_{\substack{=(1-n)^{0}+(1-n)^{1}+\cdots+(1-n)^{k-2} \\ = \dfrac{1-(1-n)^{k-1}}{1-(1-n)} \\ \text{(by (4))}}} = (-1)^{k} n(n-1) \cdot \frac{1-(1-n)^{k-1}}{1-(1-n)} \\
& = (-1)^{k} n(n-1) \cdot \frac{1-(1-n)^{k-1}}{n} = (-1)^{k}(n-1) \cdot \left(1-(1-n)^{k-1} \right) \\
& = (-1)^{k}(n-1) - (-1)^{k}(n-1) \underbrace{(1-n)^{k-1}}_{=(-1)^{k-1}(n-1)^{k-1}} \\
& = (-1)^{k}(n-1) - \underbrace{(-1)^{k}(n-1)(-1)^{k-1}(n-1)^{k-1}}_{=-(n-1)^{k}} \\
& = (n-1)^{k} + (-1)^{k}(n-1).
\end{aligned}
$$

Thus our answer is

$$c(n,k) = (n-1)^k + (-1)^k (n-1) \qquad \text{for each } k \geq 1.$$

(If you found the way we derived this answer vertiginous, you can also prove it by induction on k using $c(n,1) = 0$ and (286). See [17f-hw3s, solution to Exercise 5 **(b)**] for the details of this straightforward proof.[182]) Thus, the # of cyc-Smirnov k-tuples is $c(n,k) = (n-1)^k + (-1)^k (n-1)$. This solves Exercise 7.3.2 **(c)**. □

(An aside for the graph-theoretically trained reader: Exercise 7.3.2 is equivalent to some known formulas for the chromatic polynomials of certain graphs. Indeed, the injective k-tuples can be regarded as proper colorings of the complete graph K_k; the Smirnov k-tuples as proper colorings of the path graph P_k; the cyc-Smirnov k-tuples as proper colorings of the cycle graph C_k. From this point of view, the answers to the three parts of Exercise 7.3.2 are really computing the chromatic polynomials of these graphs K_k, P_k and C_k, and our approach to Exercise 7.3.2 **(c)** is a particular case of the "deletion-contraction recurrence" for chromatic polynomials. See [White10a] or [Guicha20, §5.9] for brief introductions to chromatic polynomials. Four different solutions of Exercise 7.3.2 **(c)** – all stated in the language of chromatic polynomials – can be found in [LeeShi19].)

7.4. Permutations

7.4.1. All permutations

Next, let us count the permutations of a finite set ([19fco, §2.4.4]):

Theorem 7.4.1. Let $n \in \mathbb{N}$. Let X be an n-element set. Then,

$$(\text{\# of permutations of } X) = n!.$$

Example 7.4.2. Let $n = 3$ and $X = [n] = [3]$. Then, Theorem 7.4.1 says that $(\text{\# of permutations of } X) = n! = 3! = 6$. Here are the 6 permutations of X (written in two-line notation – i.e., the entries in the bottom row are the images of the entries in the top row):

$$\begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}, \qquad \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix}, \qquad \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix},$$
$$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}, \qquad \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix}, \qquad \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix}.$$

[182] I have denoted cyc-Smirnov tuples as "rounded smords" in [17f-hw3s, Exercise 5 **(b)**] (and also denoted Smirnov tuples as "smords" in [17f-hw3s, Exercise 5 **(a)**]). This was mainly done in order to make googling for solutions harder. ("Smord" is short for "Smirnov word".)

We shall prove Theorem 7.4.1 as a particular case of the following fact:

Theorem 7.4.3. Let $n \in \mathbb{N}$. Let U and V be two n-element sets. Then,

$$(\# \text{ of bijections from } U \text{ to } V) = n!.$$

Proof of Theorem 7.4.3. The sets U and V are n-element sets; hence, $|U| = n$ and $|V| = n$. Thus, $|U| = n = |V|$. Therefore, Theorem 6.1.3 shows that every injective map from U to V is bijective. Conversely, every bijective map from U to V is injective (by definition). Combining the preceding two sentences, we conclude that the bijective maps from U to V are precisely the injective maps from U to V. Thus,

$$\begin{aligned}
&(\# \text{ of bijective maps from } U \text{ to } V) \\
&= (\# \text{ of injective maps from } U \text{ to } V) \\
&= n(n-1)(n-2)\cdots\underbrace{(n-n+1)}_{=1} \\
&\qquad (\text{by (283), applied to } A = U,\ B = V \text{ and } m = n) \\
&= n(n-1)(n-2)\cdots 1 = 1\cdot 2\cdot\cdots\cdot n = n!.
\end{aligned}$$

But "bijection" means the same thing as "bijective map". Hence,

$$(\# \text{ of bijections from } U \text{ to } V) = (\# \text{ of bijective maps from } U \text{ to } V) = n!.$$

This proves Theorem 7.4.3. □

Proof of Theorem 7.4.1. The permutations of X are the same thing as the bijective maps from X to X (by the definition of "permutation"). In other words, they are the same thing as the bijections from X to X. Hence,

$$(\# \text{ of permutations of } X) = (\# \text{ of bijections from } X \text{ to } X) = n!$$

(by Theorem 7.4.3, applied to $U = X$ and $V = X$).This proves Theorem 7.4.1. □

7.4.2. Permutations σ with $\sigma(1) > \sigma(2)$

The following notation is fairly standard in mathematics:

Definition 7.4.4. Let $n \in \mathbb{N}$. The set of all permutations of $[n]$ is denoted by S_n, and is called the *n-th symmetric group*.

As we explained above, this set S_n is indeed a group (in the sense of abstract algebra) and thus deserves its name. In particular, the set S_n (for each $n \in \mathbb{N}$) is closed under composition (i.e., if $f, g \in S_n$, then $f \circ g \in S_n$) and closed under inversion (i.e., if $f \in S_n$, then $f^{-1} \in S_n$).

(Some authors use notations like $\mathfrak{S}_n$, Σ_n or $\operatorname{Sym}(n)$ instead of S_n; note that all of these are variations on the letter "S".)

Exercise 7.4.1. Let $n \geq 2$ be an integer. How many permutations $\sigma \in S_n$ satisfy $\sigma(1) > \sigma(2)$?

We shall outline two solutions to this exercise, as they illustrate two different counting strategies that are both rather useful.

First solution to Exercise 7.4.1 (sketched). This is an outline; see [18s-mt1s, proof of Proposition 0.5] for a more detailed writeup[183].

Recall that S_n is the set of all permutations of $[n]$. Thus,

$$|S_n| = (\text{\# of permutations of } [n]) = n!$$

(by Theorem 7.4.1, applied to $X = [n]$).

We shall say that a permutation $\sigma \in S_n$ is

- *green* if it satisfies $\sigma(1) > \sigma(2)$;
- *red* if it satisfies $\sigma(1) < \sigma(2)$.

Every permutation $\sigma \in S_n$ is either green or red[184], but no permutation $\sigma \in S_n$ can be both green and red at the same time[185]. Hence, the set S_n is the union of its two disjoint subsets $\{\text{green permutations } \sigma \in S_n\}$ and $\{\text{red permutations } \sigma \in S_n\}$. Thus, the sum rule yields

$$|S_n| = |\{\text{green permutations } \sigma \in S_n\}| + |\{\text{red permutations } \sigma \in S_n\}|. \tag{287}$$

On the other hand, I claim that "the colors are equidistributed": i.e., the number of green permutations $\sigma \in S_n$ equals the number of red permutations $\sigma \in S_n$.

To prove this, I will construct a bijection from $\{\text{green permutations } \sigma \in S_n\}$ to $\{\text{red permutations } \sigma \in S_n\}$.

Indeed, the idea is simple: If $\sigma \in S_n$ is a green permutation, then interchanging the first two values of σ [186] yields a new permutation $\tau \in S_n$, which is red (because interchanging the first two values $\sigma(1)$ and $\sigma(2)$ obviously inverts the inequality $\sigma(1) > \sigma(2)$). Thus, we obtain a map

$$\begin{aligned} f : \{\text{green permutations } \sigma \in S_n\} &\to \{\text{red permutations } \sigma \in S_n\}, \\ \sigma &\mapsto (\sigma \text{ with the first two values interchanged}). \end{aligned}$$

[183] Note that [18s-mt1s, Proposition 0.5] is not quite the same as our Exercise 7.4.1, but it is completely analogous; the only difference is that $\sigma(1) > \sigma(2)$ is replaced by $\sigma(3) > \sigma(4)$.

[184] Indeed, every permutation $\sigma \in S_n$ is injective, and thus satisfies $\sigma(1) \neq \sigma(2)$, so that it must satisfy either $\sigma(1) > \sigma(2)$ or $\sigma(1) < \sigma(2)$.

[185] since $\sigma(1) > \sigma(2)$ would contradict $\sigma(1) < \sigma(2)$

[186] We consider the values of a permutation $\sigma \in S_n$ to be implicitly listed in the order $\sigma(1), \sigma(2), \ldots, \sigma(n)$. Thus, the "first two values of σ" are understood to be $\sigma(1)$ and $\sigma(2)$. Interchanging these two values means that we change the permutation so that the values $\sigma(1)$ and $\sigma(2)$ become the values at 2 and 1, respectively (i.e., they switch their roles).

(More formally, "σ with the first two values interchanged" can be defined as follows: Let s_1 be the permutation of $[n]$ that swaps the numbers 1 and 2 while leaving all other numbers unchanged. That is, s_1 is given by

$$s_1(i) = \begin{cases} 2, & \text{if } i = 1; \\ 1, & \text{if } i = 2; \\ i, & \text{if } i \notin \{1, 2\} \end{cases} \qquad \text{for all } i \in [n].$$

Then, "σ with the first two values interchanged" is the permutation $\sigma \circ s_1$ of $[n]$. To wit, when we compose σ with s_1, the first two values $\sigma(1)$ and $\sigma(2)$ of σ get interchanged, while all other values of σ stay at their places.)

So we have obtained a map

$$f : \{\text{green permutations } \sigma \in S_n\} \to \{\text{red permutations } \sigma \in S_n\}.$$

Likewise, we obtain a map

$$g : \{\text{red permutations } \sigma \in S_n\} \to \{\text{green permutations } \sigma \in S_n\}$$

that is defined in the same way as f (viz., it interchanges the first two values of σ). The maps f and g are mutually inverse, because if we interchange the first two values of a permutation σ and subsequently interchange them again, we end up back at our original permutation σ. Hence, the map f is invertible, and thus is a bijection. So we have found a bijection from $\{\text{green permutations } \sigma \in S_n\}$ to $\{\text{red permutations } \sigma \in S_n\}$ (namely, f). Therefore, the bijection principle yields

$$|\{\text{green permutations } \sigma \in S_n\}| = |\{\text{red permutations } \sigma \in S_n\}|. \tag{288}$$

Now, (287) becomes

$$\begin{aligned} |S_n| &= |\{\text{green permutations } \sigma \in S_n\}| + \underbrace{|\{\text{red permutations } \sigma \in S_n\}|}_{\substack{=|\{\text{green permutations } \sigma \in S_n\}| \\ \text{(by (288))}}} \\ &= |\{\text{green permutations } \sigma \in S_n\}| + |\{\text{green permutations } \sigma \in S_n\}| \\ &= 2 \cdot |\{\text{green permutations } \sigma \in S_n\}|. \end{aligned}$$

Hence,

$$|\{\text{green permutations } \sigma \in S_n\}| = \frac{1}{2} \underbrace{|S_n|}_{=n!} = \frac{1}{2} n! = n!/2.$$

In other words, the number of all green permutations $\sigma \in S_n$ is $n!/2$. In other words, the number of all permutations $\sigma \in S_n$ satisfying $\sigma(1) > \sigma(2)$ is $n!/2$ (because these permutations are precisely the green permutations $\sigma \in S_n$). This solves Exercise 7.4.1. $\square$

Our above solution was an example of "counting by symmetry": We did not count the green permutations directly; instead, we showed that they are in bijection with the remaining (i.e., red) permutations $\sigma \in S_n$ (that is, we matched up each green permutation with a red one), from which we concluded that they make up exactly half of the set S_n; and this told us that there are $\frac{1}{2}|S_n| = n!/2$ of them. This sort of reasoning is not so infrequent in combinatorics; it often explains why answers to some counting problems turn out to be representable as fractions (like $n!/2$), even though of course they are integers.

Here is a more conventional way to solve Exercise 7.4.1:

Second solution to Exercise 7.4.1 (sketched). A permutation $\sigma \in S_n$ is the same as an injective map $\sigma : [n] \to [n]$ (as we have already seen in our proof of Theorem 7.4.1). Thus, the following is a way to construct a permutation $\sigma \in S_n$ satisfying $\sigma(1) > \sigma(2)$:

- First, we choose the **set** $\{\sigma(1), \sigma(2)\}$. This set must be a 2-element subset of $[n]$ (indeed, we must have $\sigma(1) \neq \sigma(2)$, since σ must be injective), but is otherwise arbitrary; thus, the number of options is the # of all 2-element subsets of $[n]$. But the latter number is $\binom{n}{2}$ (by Theorem 4.3.12, applied to $S = [n]$ and $k = 2$). Hence, the number of options at this step is $\binom{n}{2}$.

  Note that, once we have chosen the set $\{\sigma(1), \sigma(2)\}$, the values $\sigma(1)$ and $\sigma(2)$ are uniquely determined, because the inequality $\sigma(1) > \sigma(2)$ forces $\sigma(1)$ to be the largest element of this set and $\sigma(2)$ to be its smallest element.

- Next, we choose the value $\sigma(3)$. This value must be an element of the $(n-2)$-element set $[n] \setminus \{\sigma(1), \sigma(2)\}$; thus, there are $n-2$ options for it.

- Next, we choose the value $\sigma(4)$. This value must be an element of the $(n-3)$-element set $[n] \setminus \{\sigma(1), \sigma(2), \sigma(3)\}$ (this is an $(n-3)$-element set because $\sigma(1), \sigma(2), \sigma(3)$ are distinct); thus, there are $n-3$ options for it.

- Next, we choose the value $\sigma(5)$. This value must be an element of the $(n-4)$-element set $[n] \setminus \{\sigma(1), \sigma(2), \sigma(3), \sigma(4)\}$; thus, there are $n-4$ options for it.

- And so on, until all n values $\sigma(1), \sigma(2), \ldots, \sigma(n)$ are chosen. (Note that there will be $n-(n-1) = 1$ options for the last value $\sigma(n)$.)

Thus, the Dependent product rule shows that the total # of possibilities for how we

can make these choices is

$$\begin{aligned}
&\underbrace{\binom{n}{2}}_{\substack{=\frac{n(n-1)}{2} \\ \text{(by (121))}}} \cdot (n-2)(n-3)\cdots 1 \\
&= \frac{n(n-1)}{2} \cdot (n-2)(n-3)\cdots 1 = \frac{1}{2} \cdot \underbrace{n(n-1)\cdot(n-2)(n-3)\cdots 1}_{=n(n-1)\cdots 1 = 1\cdot 2\cdot\cdots\cdot n = n!} \\
&= \frac{1}{2} \cdot n! = n!/2.
\end{aligned}$$

Hence, the # of all permutations $\sigma \in S_n$ satisfying $\sigma(1) > \sigma(2)$ is $n!/2$. This solves Exercise 7.4.1 again.

(This solution followed the Hint in [18s-mt1s, Remark 0.7].) □

7.4.3. The average number of fixed points of a permutation

Recall that for any $n \in \mathbb{N}$, we let S_n denote the set of all permutations of $[n]$.

For our next exercise, we need the following definition:

Definition 7.4.5. Let X be a set. Let $f : X \to X$ be a map.

(a) A *fixed point* of f means an element $x \in X$ such that $f(x) = x$.

(b) We let $\operatorname{Fix} f$ denote the set of all fixed points of f. (Thus, $\operatorname{Fix} f = \{x \in X \mid f(x) = x\}$.)

Exercise 7.4.2. Let n be a positive integer. Prove that $\sum_{w \in S_n} |\operatorname{Fix} w| = n!$.

In other words, this exercise states that the average number of fixed points of a permutation of $[n]$ is 1.

Exercise 7.4.2 was Problem 1 at the International Mathematical Olympiad (IMO) 1987. We give a solution using some (very basic) probability theory (specifically, the linearity of expectation):

Solution to Exercise 7.4.2 (sketched). We shall use the language of probability (specifically, discrete random variables and their expected values). See [GriSne07, §6.1] for a quick refresher. Note that our solution can just as well be restated as a purely combinatorial argument (using interchange of summation signs instead of linearity of expectation).

We have

$$|S_n| = n! \tag{289}$$

(indeed, this can be shown as in the First solution to Exercise 7.4.1).

Our goal is to prove that $\sum_{w \in S_n} |\operatorname{Fix} w| = n!$. In other words, our goal is to prove that $\sum_{w \in S_n} |\operatorname{Fix} w| = |S_n|$ (since $|S_n| = n!$). In other words, our goal is to prove that $\frac{1}{|S_n|} \sum_{w \in S_n} |\operatorname{Fix} w| = 1$. The left hand side of this equality is the average of $|\operatorname{Fix} w|$ when w ranges over all of S_n. In other words, it is the expected value of $|\operatorname{Fix} \sigma|$ when σ is a discrete random variable uniformly distributed over S_n. This suggests that we use random variables.

We consider a discrete random variable σ that is uniformly distributed over S_n. We use the notation $\operatorname{Pr}(A)$ for the probability of an event A, and we use the notation $\operatorname{E}(X)$ for the expected value of a numerically-valued random variable X. It is known that any n numerically-valued random variables $X_1, X_2, \ldots, X_n$ satisfy

$$\operatorname{E}(X_1 + X_2 + \cdots + X_n) = \operatorname{E}(X_1) + \operatorname{E}(X_2) + \cdots + \operatorname{E}(X_n). \tag{290}$$

(In fact, this is part of the principle of *linearity of expectation*. See [GriSne07, Theorem 6.2] for the case $n = 2$; the general case follows by induction on n.)

Now, we claim that each $w \in S_n$ satisfies[187]

$$|\operatorname{Fix} w| = [w(1) = 1] + [w(2) = 2] + \cdots + [w(n) = n]. \tag{291}$$

[*Proof:* Let $w \in S_n$. Let $i = |\operatorname{Fix} w|$. Thus, w has i fixed points. If x is one of these i fixed points, then $w(x) = x$ and thus $[w(x) = x] = 1$; on the other hand, if x is not a fixed point of w, then $w(x) \neq x$ and thus $[w(x) = x] = 0$. Hence, the sum on the right hand side of (291) has i addends equal to 1 (one such addend for each of the i fixed points of w), while the remaining addends are 0. Therefore, the right hand side of (291) equals $i \cdot 1 + (n - i) \cdot 0 = i = |\operatorname{Fix} w|$, and thus (291) is proved.]

Now, applying (291) to our random permutation σ, we find

$$|\operatorname{Fix} \sigma| = [\sigma(1) = 1] + [\sigma(2) = 2] + \cdots + [\sigma(n) = n].$$

Hence,

$$\begin{aligned}
\operatorname{E}(|\operatorname{Fix} \sigma|) &= \operatorname{E}([\sigma(1) = 1] + [\sigma(2) = 2] + \cdots + [\sigma(n) = n]) \\
&= \operatorname{E}([\sigma(1) = 1]) + \operatorname{E}([\sigma(2) = 2]) + \cdots + \operatorname{E}([\sigma(n) = n]) \\
&\qquad \text{(by (290))} \\
&= \sum_{i=1}^{n} \operatorname{E}([\sigma(i) = i]). \qquad (292)
\end{aligned}$$

Now, let us fix an $i \in [n]$. What is the expected value $\operatorname{E}([\sigma(i) = i])$?

If E is any event, then $\operatorname{E}([E]) = \operatorname{Pr}(E)$ (since $[E] = 1$ if E holds, and $[E] = 0$ otherwise). Thus, $\operatorname{E}([\sigma(i) = i]) = \operatorname{Pr}(\sigma(i) = i)$. Now, what is the probability that $\sigma(i) = i$?

Here are two ways of answering this question:[188]

[187] We are using the Iverson bracket notation (see Definition 4.3.19) here.

[188] We are **not** doing a proof by double counting here, so we only need to answer this question in one way. I am showing two ways just for the sake of illustrating different approaches.

- One way is to use symmetry. Namely, since σ is uniformly distributed across S_n, the random variable $\sigma(i)$ will take each of the n possible values $1, 2, \ldots, n$ with equal probabilities (for symmetry reasons[189]). In other words, the n probabilities

$$\Pr(\sigma(i) = 1), \quad \Pr(\sigma(i) = 2), \quad \ldots, \quad \Pr(\sigma(i) = n)$$

are equal. But the sum of these n probabilities is

$$\Pr(\sigma(i) = 1) + \Pr(\sigma(i) = 2) + \cdots + \Pr(\sigma(i) = n) = 1$$

(since $\sigma(i)$ must take one of the n values $1, 2, \ldots, n$, and of course cannot take more than one of them simultaneously). Hence, these n probabilities must all

[189] For the skeptics, let me spell these symmetry reasons out:

We must show that if u and v are any two elements of $[n]$, then the probability of $\sigma(i) = u$ equals the probability of $\sigma(i) = v$.

So let $u, v \in [n]$. Then, there exists a permutation $t \in S_n$ that sends u to v. (For example, we can define t to be the permutation of $[n]$ that swaps u and v and leaves the remaining elements of $[n]$ unchanged.) Consider such a t. The inverse t^{-1} of this permutation t is then a permutation of $[n]$ that sends v to u.

Now, consider the map

$$\begin{aligned} \Phi : \{w \in S_n \mid w(i) = u\} &\to \{w \in S_n \mid w(i) = v\}, \\ w &\mapsto t \circ w. \end{aligned}$$

It is easy to see that this map Φ is well-defined (i.e., if $w \in S_n$ satisfies $w(i) = u$, then $t \circ w \in S_n$ and $(t \circ w)(i) = v$). Likewise, consider the map

$$\begin{aligned} \Psi : \{w \in S_n \mid w(i) = v\} &\to \{w \in S_n \mid w(i) = u\}, \\ w &\mapsto t^{-1} \circ w. \end{aligned}$$

Then, it is easy to see that the maps Φ and Ψ are mutually inverse (for example, $\Phi \circ \Psi = \mathrm{id}$ is because each permutation $w \in S_n$ satisfies $(\Phi \circ \Psi)(w) = \Phi(\Psi(w)) = t \circ \left(t^{-1} \circ w\right) = \underbrace{\left(t \circ t^{-1}\right)}_{=\mathrm{id}} \circ w = w = \mathrm{id}(w)$). Hence, the map Φ is invertible, i.e., is bijective. The bijection principle thus yields

$$|\{w \in S_n \mid w(i) = u\}| = |\{w \in S_n \mid w(i) = v\}|. \tag{293}$$

However, σ is uniformly distributed over S_n; thus,

$$(\text{the probability of } \sigma(i) = u) = \frac{|\{w \in S_n \mid w(i) = u\}|}{|S_n|}.$$

Likewise,

$$(\text{the probability of } \sigma(i) = v) = \frac{|\{w \in S_n \mid w(i) = v\}|}{|S_n|}.$$

The right hand sides of these two equalities are equal (because of (293)). Hence, the left hand sides are also equal. In other words, the probability of $\sigma(i) = u$ equals the probability of $\sigma(i) = v$. This is precisely what we wanted to show.

equal $\frac{1}{n}$ (because if we are given n equal numbers whose sum is 1, then each of these n numbers must be $\frac{1}{n}$). In other words, we have $\Pr(\sigma(i) = j) = \frac{1}{n}$ for each $j \in [n]$. Applying this to $j = i$, we obtain $\Pr(\sigma(i) = i) = \frac{1}{n}$.

- Alternatively, we can just count the permutations $w \in S_n$ satisfying $w(i) = i$. Indeed, we can construct such a permutation w by choosing its values $w(1), w(2), \ldots, w(i-1), w(i+1), w(i+2), \ldots, w(n)$ in order (skipping $w(i)$, since $w(i)$ is already predetermined to be i). Each value must differ both from i and from all previously chosen values (since w must be injective), so we have $n-1$ options for the first value, $n-2$ options for the second, $n-3$ for the third, and so on. Thus, by the dependent product rule, the total # of possibilities for how we can make these choices is $(n-1)(n-2)\cdots 1 = 1 \cdot 2 \cdot \cdots \cdot (n-1) = (n-1)!$. Therefore,

  $$(\text{\# of permutations } w \in S_n \text{ satisfying } w(i) = i) = (n-1)!.$$

  In other words,

  $$|\{w \in S_n \mid w(i) = i\}| = (n-1)!. \tag{294}$$

  But since σ is uniformly distributed over S_n, we have

  $$\begin{aligned} \Pr(\sigma(i) = i) &= \frac{|\{w \in S_n \mid w(i) = i\}|}{|S_n|} \\ &= \frac{(n-1)!}{n!} \qquad (\text{by (294) and (289)}) \\ &= \frac{1}{n} \qquad (\text{since } n! = n \cdot (n-1)!). \end{aligned}$$

Thus, we have shown that $\Pr(\sigma(i) = i) = \frac{1}{n}$, so that

$$\mathrm{E}([\sigma(i) = i]) = \Pr(\sigma(i) = i) = \frac{1}{n}. \tag{295}$$

Forget that we fixed i. We thus have proved (295) for each $i \in [n]$. Hence, (292) becomes

$$\mathrm{E}(|\operatorname{Fix}\sigma|) = \sum_{i=1}^{n} \underbrace{\mathrm{E}([\sigma(i) = i])}_{\substack{=\frac{1}{n} \\ (\text{by (295)})}} = \sum_{i=1}^{n} \frac{1}{n} = n \cdot \frac{1}{n} = 1.$$

But the definition of expected value yields $\mathrm{E}(|\operatorname{Fix}\sigma|) = \frac{1}{|S_n|} \sum_{w \in S_n} |\operatorname{Fix} w|$ (since σ is uniformly distributed over S_n). Hence,

$$\frac{1}{|S_n|} \sum_{w \in S_n} |\operatorname{Fix} w| = \mathrm{E}(|\operatorname{Fix}\sigma|) = 1,$$

so that $\sum_{w \in S_n} |\operatorname{Fix} w| = |S_n| = n!$. This solves Exercise 7.4.2. $\square$

See [17f-hw7s, solution to Exercise 2] for the same solution, rewritten in elementary combinatorial language. See also [19fco-mt2s, Exercise 4] for a solution to the following generalization:

Exercise 7.4.3. Let $k \in \mathbb{N}$. Let $n \geq k$ be an integer. Prove that

$$\sum_{w \in S_n} \binom{|\operatorname{Fix} w|}{k} = (n-k)! \binom{n}{k} = \frac{n!}{k!}.$$

7.5. Double counting

Counting is not just an end in itself; it can also be used to prove equalities. To wit, many counting problems can be solved in several ways, yielding different-looking results. (For example, the counting problem "how many subsets does the set $[n]$ have?" has answer 2^n because of Theorem 7.3.1; but it also has answer $\sum_{k=0}^{n} \binom{n}{k}$ because of Theorem 4.3.12[190].) Of course, these results must therefore be equal (so, for example, the previous parenthetical sentence entails $2^n = \sum_{k=0}^{n} \binom{n}{k}$, which is precisely the claim of Corollary 4.3.17). Hence, by answering a counting problem in two different ways, we obtain an equality "for free". Thus we obtain a powerful method for proving equalities: Try to find a counting problem such that both sides of the equality are answers to this problem. This strategy is known as *double counting*. Let us see a simple but important example of this strategy:

7.5.1. The Chu–Vandermonde identity for nonnegative integers

Proposition 7.5.1 (Chu–Vandermonde identity for nonnegative integers)**.** Let $n \in \mathbb{N}$ and $x, y \in \mathbb{N}$. Then,

$$\binom{x+y}{n} = \sum_{k=0}^{n} \binom{x}{k} \binom{y}{n-k}.$$

[190] In more detail: Any subset of $[n]$ has size k for some $k \in \{0, 1, \ldots, n\}$. Thus, the sum rule yields

$$(\#\text{ of subsets of } [n]) = \sum_{k=0}^{n} \underbrace{(\#\text{ of subsets of } [n] \text{ having size } k)}_{\substack{=(\#\text{ of } k\text{-element subsets of } [n]) \\ = \binom{n}{k} \\ \text{(by Theorem 4.3.12)}}} = \sum_{k=0}^{n} \binom{n}{k},$$

qed.

For example, for $n = 2$, this equality states that

$$\binom{x+y}{2} = \sum_{k=0}^{2} \binom{x}{k}\binom{y}{2-k} = \underbrace{\binom{x}{0}}_{=1}\binom{y}{2} + \underbrace{\binom{x}{1}}_{=x}\underbrace{\binom{y}{1}}_{=y} + \binom{x}{2}\underbrace{\binom{y}{0}}_{=1}$$
$$= \binom{y}{2} + xy + \binom{x}{2} = \binom{x}{2} + xy + \binom{y}{2}.$$

Proof of Proposition 7.5.1 (sketched). We apply double counting. The counting problem we want to solve in two ways is the following: Let S be the $(x+y)$-element set $\{1, 2, \ldots, x\} \cup \{-1, -2, \ldots, -y\}$. How many n-element subsets does S have? Let us answer this question in two ways:[191]

First way: The answer is $\binom{x+y}{n}$. Indeed, this follows from Theorem 4.3.12 (applied to $x+y$, S and n instead of n, S and k), since S is an $(x+y)$-element set.

Second way: An n-element subset of S has at least 0 and at most n positive elements. Thus, the sum rule yields

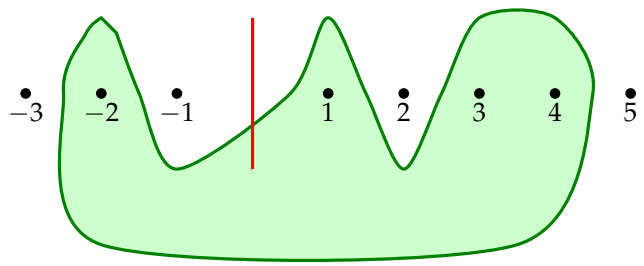
$$(\#\text{ of } n\text{-element subsets of } S)$$
$$= \sum_{k=0}^{n} (\#\text{ of } n\text{-element subsets of } S \text{ having exactly } k \text{ positive elements}).$$

Now, let us compute the addends on the right hand side of this equality. (Note that some of these addends can be 0, but this needs not worry us: The argument we will be making in the following paragraph applies to them just as well as it does to the others.)

Fix $k \in \{0, 1, \ldots, n\}$. How many n-element subsets I of S have exactly k positive elements? Such a subset I must have exactly k positive elements, and thus must have exactly $n-k$ negative elements (since it must have n elements in total, and 0 is not available because $0 \notin S$). Hence, we can construct such a subset I as follows:

- First, we choose the k positive elements of I. (There are $\binom{x}{k}$ options for them, by Theorem 4.3.12[192].)

[191] Here is an illustration of the case $x = 5$, $y = 3$ and $n = 4$ (with the green blob being the n-element subset):



[192] since they must belong to the x-element set $\{1, 2, \ldots, x\}$

- Then, we choose the $n-k$ negative elements of I. (There are $\binom{y}{n-k}$ options for them, by Theorem 4.3.12[193].)

Thus, the product rule shows that the total # of possibilities for how we can make these choices is $\binom{x}{k}\binom{y}{n-k}$. Thus,

$$\begin{aligned}&(\text{\# of } n\text{-element subsets of } S \text{ having exactly } k \text{ positive elements})\\&=\binom{x}{k}\binom{y}{n-k}.\end{aligned} \tag{296}$$

Now, forget that we fixed k. Hence, we have proved (296) for each $k \in \{0,1,\ldots,n\}$. Now,

$$\begin{aligned}&(\text{\# of } n\text{-element subsets of } S)\\&=\sum_{k=0}^{n}\underbrace{(\text{\# of } n\text{-element subsets of } S \text{ having exactly } k \text{ positive elements})}_{\substack{=\binom{x}{k}\binom{y}{n-k}\\ \text{(by (296))}}}\\&=\sum_{k=0}^{n}\binom{x}{k}\binom{y}{n-k}.\end{aligned}$$

So the answer to our counting problem is $\sum_{k=0}^{n}\binom{x}{k}\binom{y}{n-k}$.

Now, we have answered our counting problem in two ways. The first way yielded $\binom{x+y}{n}$ as an answer; the second way yielded $\sum_{k=0}^{n}\binom{x}{k}\binom{y}{n-k}$. Comparing these answers, we obtain $\binom{x+y}{n}=\sum_{k=0}^{n}\binom{x}{k}\binom{y}{n-k}$. This proves Proposition 7.5.1.

(See [19s, proof of Lemma 2.17.15] or [19fco, §2.6.1, Second proof of Theorem 2.6.1 for $x,y\in\mathbb{N}$] for different writeups of this proof.) □

The words "for nonnegative integers" in the name of Proposition 7.5.1 hint at the fact that there is a more general version of the Chu–Vandermonde identity; we shall see this soon.

7.5.2. The trinomial revision formula for nonnegative integers

Here is another identity that can be shown using double counting:

[193] since they must belong to the y-element set $\{-1,-2,\ldots,-y\}$

Proposition 7.5.2 (Trinomial revision formula for nonnegative integers)**.** Let $n, a, b \in \mathbb{N}$. Then,

$$\binom{n}{a}\binom{a}{b} = \binom{n}{b}\binom{n-b}{a-b}.$$

Proposition 7.5.2 is known as the *trinomial revision formula*, since the product $\binom{n}{a}\binom{a}{b}$ of two binomial coefficients is known as a "trinomial coefficient", and this proposition allows "revising" it (i.e., rewriting it as a different product of two binomial coefficients). It is a useful tool in manipulating binomial coefficients. Again, we have not stated it in the greatest possible generality, but we have to start somewhere. Note that particular cases of Proposition 7.5.2 have already been used twice in solutions of problems posed above[194].

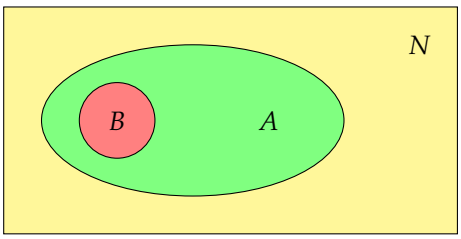
Proof of Proposition 7.5.2. We will use double counting again. Consider the following counting problem: Assume you have n people. How many ways are there to choose a *committee* consisting of a of these n people, and also choose a *subcommittee* consisting of b people from the committee?

Throwing away the unnecessary anthropomorphism, we can restate this problem in a more convenient language as follows: Fix an n-element set N (whose elements are our n people). We want to count all pairs (A, B), where A is an a-element subset of N (this will be our committee) and B is a b-element subset of A (this will be our subcommittee).[195] In other words, we want to count all pairs (A, B) of sets satisfying $B \subseteq A \subseteq N$ and $|A| = a$ and $|B| = b$. We shall refer to such pairs as *CS pairs*[196], and we shall count their # in two ways[197]:

First way: In order to construct a CS pair (A, B), we first choose the committee A and then choose its subcommittee B. We have $\binom{n}{a}$ many options for A (because A has to be an a-element subset of the n-element set N); then, after choosing A, we have $\binom{a}{b}$ many options for B (because B has to be a b-element subset of A). Thus,

[194]Namely, the equality (525) in the solution to Exercise 4.5.5 and the equality (667) in the solution to Exercise 5.4.3 are particular cases of Proposition 7.5.2.

[195]Here is a symbolic picture of this situation:



.

[196]short for "committee-subcommittee pairs"

[197]We will be using Theorem 4.3.12 in both of these ways.

by the dependent product rule, the total # of CS pairs is $\binom{n}{a}\binom{a}{b}$.

Second way: In order to construct a CS pair (A, B), we first choose the subcommittee B and then choose the committee A. We have $\binom{n}{b}$ many options for B (because B has to be a b-element subset of the n-element set N). After choosing B, how many options do we have for A ? The committee A has to be an a-element subset of N, but it is also required to contain the (already chosen) subcommittee B as a subset, so that b of its a elements are already decided. We only need to choose the remaining $a - b$ elements of A. These $a - b$ elements have to come from the $(n - b)$-element set $N \setminus B$ (because the elements of B have already been chosen to lie in A). Thus, we are choosing an $(a - b)$-element subset of the $(n - b)$-element set $N \setminus B$. The # of ways to do this is $\binom{n-b}{a-b}$. [198] Therefore, we have $\binom{n-b}{a-b}$ many options for choosing A. Hence, by the dependent product rule, the total # of CS pairs is $\binom{n}{b}\binom{n-b}{a-b}$.

Now, we have computed the # of CS pairs in two different ways. The first way gave us the result $\binom{n}{a}\binom{a}{b}$, while the second way gave us the result $\binom{n}{b}\binom{n-b}{a-b}$. Comparing these results, we find $\binom{n}{a}\binom{a}{b} = \binom{n}{b}\binom{n-b}{a-b}$. Thus, Proposition 7.5.2 is proven. □

See [19fco, §2.2.2] for a more formal version of the above proof. Because of this proof, Proposition 7.5.2 is sometimes called the "*committee-subcommittee identity*".

7.5.3. The polynomial identity trick

We have proved both the Chu–Vandermonde identity and the Trinomial revision formula by double counting. These proofs relied heavily on the combinatorial interpretation of binomial coefficients (Theorem 4.3.12). The latter interpretation makes sense only for binomial coefficients $\binom{n}{k}$ with $n \in \mathbb{N}$; it says nothing about binomial coefficients $\binom{n}{k}$ in which n is negative or non-integer. Nevertheless, the Chu–Vandermonde identity and the Trinomial revision formula can be generalized to non-integer values of some of the arguments:[199]

[198] Here, we need to know that $n - b \in \mathbb{N}$. Why is this the case? It does not follow from the assumptions of Proposition 7.5.2, since we have **not** assumed that $b \leq n$. However, it does follow from the fact that we are in the second step of our construction of (A, B) and therefore already have chosen a b-element subset B of N; indeed, this fact clearly entails that $b = |B| \leq |N| = n$. If we didn't have $b \leq n$, we could not have chosen a b-element subset B of N in the first place, and thus we would have never gotten to the point where we are choosing A.

[199] We are using the set $\mathbb{C}$ of complex numbers in stating the following two theorems. Readers unfamiliar with complex numbers can replace the symbol $\mathbb{C}$ by $\mathbb{R}$ throughout this section; this

Theorem 7.5.3 (Chu–Vandermonde identity)**.** Let $n \in \mathbb{N}$ and $x, y \in \mathbb{C}$. Then,

$$\binom{x+y}{n} = \sum_{k=0}^{n} \binom{x}{k} \binom{y}{n-k}.$$

Theorem 7.5.4 (Trinomial revision formula)**.** Let $n, a, b \in \mathbb{C}$. Then,

$$\binom{n}{a} \binom{a}{b} = \binom{n}{b} \binom{n-b}{a-b}.$$

Clearly, Proposition 7.5.1 and Proposition 7.5.2 are particular cases of Theorem 7.5.3 and Theorem 7.5.4 (which is why we have called them "propositions" rather than "theorems"). The proofs we gave for the two propositions no longer apply to the theorems that generalize them; how should we then prove the theorems?

One thing we can try to do is start from scratch and search for non-combinatorial proofs of Theorem 7.5.3 and Theorem 7.5.4. Such proofs indeed exist: An algebraic proof of Theorem 7.5.3 (completely unrelated to our above proof of Proposition 7.5.1) can be found in [Grinbe15, first proof of Theorem 3.29][200]. It is not a particularly enlightening or memorable proof, but it serves its purpose just fine (i.e., it proves Theorem 7.5.3 in full generality, without requiring $x, y \in \mathbb{N}$), and it even has a slight advantage in being easy to explain (as it requires little else besides standard manipulations of sums and products). Theorem 7.5.4 can also be proved by direct computation (using Definition 4.3.1); this proof can be found in [19fco, §2.2.1, proof of Proposition 1.3.35] (and is arguably even easier than the above combinatorial proof of Proposition 7.5.2).

However, this supplantation of combinatorial proofs by algebraic ones leaves a bad aftertaste. Is double counting just a toy that only works in particular cases (i.e., usually, when all variables lie in $\mathbb{N}$)? One of the hallmarks of a good mathematical proof is that it can be adapted to many natural generalizations of the result being proved; our double-counting proofs of Proposition 7.5.1 and Proposition 7.5.2 fail this criterion. There is also a more practical concern: Double counting can be used not just in proving these two propositions; there are many other results which need it far more (sometimes, the only known proofs use double counting). We cannot always expect to find an alternative proof that avoids double counting and thus can be generalized immediately.

Good news: this is a solvable problem. You **can** have your cake and eat it too! While our double-counting proofs cannot themselves be generalized to non-integer values of x, y or n, there is nevertheless a trick that allows to easily deduce the

will result in a slight loss of generality, but this generality is not important for what we are doing. (None of our arguments in this section depend on whether our numbers are real or complex.)

[200] The main idea is to induct on n, using the "absorption identity" $\binom{y}{n} = \frac{y}{n}\binom{y-1}{n-1}$ (which is easily proved).

general results (such as Theorem 7.5.3 and Theorem 7.5.4) from their particular cases that can be proved combinatorially (such as Proposition 7.5.1 and Proposition 7.5.2). I call this trick the *polynomial identity trick*, and it has far more applications than salvaging combinatorial proofs; but this is probably one of the simplest to explain. My exposition of this trick follows [19fco, §2.6.2].

The polynomial identity trick relies on the following well-known theorem:[201]

Theorem 7.5.5. Let P be a nonzero polynomial (with complex coefficients, in a single variable X). Then, P has at most $\deg P$ many roots.

For example:

- The degree-3 polynomial $3X^3 - 2X + 1$ has 3 roots, namely -1, $\frac{1}{2} + \frac{1}{6}i\sqrt{3}$ and $\frac{1}{2} - \frac{1}{6}i\sqrt{3}$ (where $i = \sqrt{-1}$ is the imaginary unit).
- The degree-3 polynomial $X^3 - X^2 - X + 1$ has 2 roots, namely 1 and -1.

Theorem 7.5.5 is often called the "easy half of the Fundamental Theorem of Algebra". We will not prove it here, but the reader can find a proof of Theorem 7.5.5 in any good textbook on abstract algebra (e.g., in [Goodma15, Corollary 1.8.24], [Joyce17, Theorem 1.58], [Walker87, Corollary 4.5.10], [CoLiOs15, Chapter 1, §5, Corollary 3], [19s, Theorem 7.6.11], [Elman20, Corollary 33.8] and [Knapp16, Corollary 1.14]).

We don't really need to use complex numbers in Theorem 7.5.5; the theorem remains true if we restrict ourselves to real or rational numbers (since any real or rational number is a complex number as well). Complex numbers are special in that they allow replacing the words "at most" in Theorem 7.5.5 by "exactly", as long as we count the roots with multiplicity (so, for example, the double root 1 of the polynomial $(X-1)^2 X$ counts as two roots $1, 1$ rather than as a single root 1). But this is a much deeper result (known as the *Fundamental Theorem of Algebra*), which we shall not need in our combinatorial applications and thus don't bother stating here.

Theorem 7.5.5 has two corollaries, which are used frequently in combinatorics:

Corollary 7.5.6. If a polynomial P (with complex coefficients, in a single variable X) has infinitely many roots, then P is the zero polynomial.

Proof of Corollary 7.5.6 (sketched). If P was nonzero, then Theorem 7.5.5 would yield that P has at most $\deg P$ many roots, which would contradict the assumption that P has infinitely many roots. Thus, P is zero, so that Corollary 7.5.6 follows. (See [19fco, proof of Corollary 2.6.9] for details.) □

[201] Recall that a *root* of a univariate polynomial P means a complex number r that satisfies $P(r) = 0$.

Also, the *degree* $\deg P$ of a nonzero polynomial P is the exponent in the largest power of the variable that appears in P with a nonzero coefficient. (That is, if we write P in the form $P = a_0 X^0 + a_1 X^1 + \cdots + a_n X^n$, where X is our variable and where $a_n \neq 0$, then the degree of P is n.)

Corollary 7.5.7. Let P and Q be polynomials (with complex coefficients, in a single variable X). Let S be an infinite set of complex numbers. Assume that

$$P(x) = Q(x) \qquad \text{for all } x \in S. \tag{297}$$

Then, $P = Q$.

Proof of Corollary 7.5.7. We claim that the polynomial $P - Q$ has infinitely many roots. Indeed, each $x \in S$ satisfies

$$(P - Q)(x) = P(x) - Q(x) = 0 \qquad \text{(by (297))}$$

and therefore is a root of $P - Q$. Thus, the polynomial $P - Q$ has each $x \in S$ as a root; hence, it has infinitely many roots (since S is an infinite set). Corollary 7.5.6 (applied to $P - Q$ instead of P) thus shows that $P - Q$ is the zero polynomial. In other words, $P - Q = 0$. Hence, $P = Q$. This proves Corollary 7.5.7. □

Corollary 7.5.7 reveals that an equality $P = Q$ between two (univariate) polynomials P and Q can be proved by showing that the polynomials agree on each element of a (fixed) infinite set S (i.e., that $P(x) = Q(x)$ for all $x \in S$); it is not necessary to check that they agree on each complex number (but, rather, this result is obtained "for free"). This is called the *polynomial identity trick*.

Before we use this trick to derive Theorem 7.5.3 and Theorem 7.5.4 from their particular cases, we illustrate it on a toy problem:

Exercise 7.5.1. Let $a \in \mathbb{C}$. Let $n \in \mathbb{N}$ be even. Prove that

$$\left(\sum_{i=0}^{n} (-1)^i a^i\right) \cdot \left(\sum_{i=0}^{n} a^i\right) = \sum_{i=0}^{n} a^{2i}. \tag{298}$$

For instance, for $n = 4$, the equality (298) says that

$$\left(1 - a + a^2 - a^3 + a^4\right) \cdot \left(1 + a + a^2 + a^3 + a^4\right) = 1 + a^2 + a^4 + a^6 + a^8.$$

There are many simple ways to solve Exercise 7.5.1 (for instance, induction on $n/2$ can be used); but let us give a solution that saves on computation by using the polynomial identity trick:

Solution to Exercise 7.5.1. Note that $n + 1$ is odd (since n is even).

Let us first assume that $a \notin \{-1, 1\}$. Thus, $a \neq -1$ and $a \neq 1$. From $a \neq -1$, we obtain $-a \neq 1$.

The number a is distinct from 1 (since $a \neq 1$). Hence, applying the equality (4) to a and $n + 1$ instead of b and n, we obtain[202]

$$a^0 + a^1 + \cdots + a^{(n+1)-1} = \frac{1 - a^{n+1}}{1 - a}.$$

[202] To be fully precise, the equality (4) was only stated for real b, whereas here we are applying it to the complex number a. But this does not really matter, since the equality (4) holds for all complex numbers b (and the same proof that we gave for it applies in this generality).

Comparing this with

$$\begin{aligned} a^0 + a^1 + \cdots + a^{(n+1)-1} &= a^0 + a^1 + \cdots + a^n \qquad (\text{since } (n+1)-1 = n) \\ &= \sum_{i=0}^{n} a^i, \end{aligned}$$

we obtain

$$\sum_{i=0}^{n} a^i = \frac{1 - a^{n+1}}{1 - a}. \tag{299}$$

The same argument (applied to $-a$ instead of a) yields

$$\sum_{i=0}^{n} (-a)^i = \frac{1 - (-a)^{n+1}}{1 - (-a)} \tag{300}$$

(since $-a \neq 1$). Multiplying this equality by (299), we obtain

$$\begin{aligned} &\left(\sum_{i=0}^{n} (-a)^i\right) \cdot \left(\sum_{i=0}^{n} a^i\right) \\ &= \frac{1 - (-a)^{n+1}}{1 - (-a)} \cdot \frac{1 - a^{n+1}}{1 - a} \\ &= \left(1 - \underbrace{(-a)^{n+1}}_{\substack{= -a^{n+1} \\ (\text{since } n+1 \text{ is odd})}}\right) \cdot \left(1 - a^{n+1}\right) / \left(\left(\underbrace{1 - (-a)}_{=1+a}\right) \cdot (1 - a)\right) \\ &= \underbrace{\left(1 - \left(-a^{n+1}\right)\right)}_{=1+a^{n+1}} \cdot \left(1 - a^{n+1}\right) / \underbrace{((1 + a) \cdot (1 - a))}_{=1-a^2} \\ &= \underbrace{\left(1 + a^{n+1}\right) \cdot \left(1 - a^{n+1}\right)}_{\substack{=1-\left(a^{n+1}\right)^2 \\ (\text{since } (1+b)\cdot(1-b)=1-b^2 \\ \text{for each } b \in \mathbb{C})}} / \left(1 - a^2\right) = \left(1 - \underbrace{\left(a^{n+1}\right)^2}_{=a^{(n+1)\cdot 2} = a^{2(n+1)} = \left(a^2\right)^{n+1}}\right) / \left(1 - a^2\right) \\ &= \left(1 - \left(a^2\right)^{n+1}\right) / \left(1 - a^2\right) = \frac{1 - \left(a^2\right)^{n+1}}{1 - a^2}. \end{aligned}$$

Comparing this with

$$\left(\sum_{i=0}^{n} \underbrace{(-a)^i}_{=(-1)^i a^i}\right) \cdot \left(\sum_{i=0}^{n} a^i\right) = \left(\sum_{i=0}^{n} (-1)^i a^i\right) \cdot \left(\sum_{i=0}^{n} a^i\right),$$

we obtain

$$\left(\sum_{i=0}^{n}(-1)^i a^i\right)\cdot\left(\sum_{i=0}^{n}a^i\right)=\frac{1-\left(a^2\right)^{n+1}}{1-a^2}. \tag{301}$$

On the other hand, a^2 is distinct from 1 (since $a\notin\{-1,1\}$). Now,

$$\begin{aligned}\sum_{i=0}^{n}\underbrace{a^{2i}}_{=\left(a^2\right)^i}&=\sum_{i=0}^{n}\left(a^2\right)^i=\left(a^2\right)^0+\left(a^2\right)^1+\cdots+\left(a^2\right)^n\\&=\left(a^2\right)^0+\left(a^2\right)^1+\cdots+\left(a^2\right)^{(n+1)-1}\qquad\text{(since } n=(n+1)-1\text{)}\\&=\frac{1-\left(a^2\right)^{n+1}}{1-a^2}\end{aligned}$$

(by (4), applied to a^2 and $n+1$ instead of b and n). [203] Comparing this equality with (301), we obtain

$$\left(\sum_{i=0}^{n}(-1)^i a^i\right)\cdot\left(\sum_{i=0}^{n}a^i\right)=\sum_{i=0}^{n}a^{2i}.$$

Thus, we have proved (298) under the assumption that $a\notin\{-1,1\}$. How can we complete this solution to a full (unconditional) proof of (298)?

One way would be to check the remaining two cases ($a=1$ and $a=-1$) by hand. This is easy and can easily be done by the reader. But I will instead show another way, which requires no computation whatsoever. Namely, forget that we fixed a. We thus have proved that the equality (298) holds for all complex numbers $a\notin\{-1,1\}$. In other words, we have proved that (298) holds for all $a\in\mathbb{C}\setminus\{-1,1\}$. Now, define two polynomials P and Q (with complex coefficients, in a single variable X) by

$$P=\left(\sum_{i=0}^{n}(-1)^i X^i\right)\cdot\left(\sum_{i=0}^{n}X^i\right)\qquad\text{and}\qquad Q=\sum_{i=0}^{n}X^{2i}.$$

Then, for each $a\in\mathbb{C}$, we have

$$P(a)=\left(\sum_{i=0}^{n}(-1)^i a^i\right)\cdot\left(\sum_{i=0}^{n}a^i\right) \tag{302}$$

(by the definition of P) and

$$Q(a)=\sum_{i=0}^{n}a^{2i} \tag{303}$$

[203] Again, we have applied (4) to a complex number b; again, this is fine because (4) could just as well have been stated for a complex number b.

(by the definition of Q). Now, each $a \in \mathbb{C} \setminus \{-1, 1\}$ satisfies

$$\begin{aligned} P(a) &= \left(\sum_{i=0}^{n} (-1)^i a^i\right) \cdot \left(\sum_{i=0}^{n} a^i\right) \qquad \text{(by (302))} \\ &= \sum_{i=0}^{n} a^{2i} \qquad \text{(since we have proved that (298) holds for all } a \in \mathbb{C} \setminus \{-1, 1\}) \\ &= Q(a) \qquad \text{(by (303))}. \end{aligned}$$

In other words, we have $P(a) = Q(a)$ for all $a \in \mathbb{C} \setminus \{-1, 1\}$. Renaming the variable a as x in this statement, we obtain the following: We have $P(x) = Q(x)$ for all $x \in \mathbb{C} \setminus \{-1, 1\}$. Hence, Corollary 7.5.7 (applied to $S = \mathbb{C} \setminus \{-1, 1\}$) yields that $P = Q$ (since the set $\mathbb{C} \setminus \{-1, 1\}$ is infinite). Thus, each $a \in \mathbb{C}$ satisfies $P(a) = Q(a)$. In view of (302) and (303), this rewrites as follows: Each $a \in \mathbb{C}$ satisfies

$$\left(\sum_{i=0}^{n} (-1)^i a^i\right) \cdot \left(\sum_{i=0}^{n} a^i\right) = \sum_{i=0}^{n} a^{2i}.$$

This solves Exercise 7.5.1. □

Now, let us use the polynomial identity trick to derive Theorem 7.5.4 from Proposition 7.5.2:

Proof of Theorem 7.5.4 (sketched). We must prove the equality

$$\binom{n}{a}\binom{a}{b} = \binom{n}{b}\binom{n-b}{a-b}. \tag{304}$$

First of all, it is easy to see that (304) holds if $b \notin \mathbb{N}$. Indeed, if $b \notin \mathbb{N}$, then both binomial coefficients $\binom{a}{b}$ and $\binom{n}{b}$ are 0 (by (118)). Thus, if $b \notin \mathbb{N}$, then the equality (304) boils down to $0 = 0$ (since each of its sides is a product with one factor equal to 0), which is clearly true. Thus, we WLOG assume that $b \in \mathbb{N}$ in the rest of this proof.

If we have $a \notin \mathbb{N}$, then we have $a - b \notin \mathbb{N}$ as well (since otherwise, we would have $a - b \in \mathbb{N}$ and therefore $a = \underbrace{a-b}_{\in\mathbb{N}} + \underbrace{b}_{\in\mathbb{N}} \in \mathbb{N}$, which would contradict $a \notin \mathbb{N}$), and therefore both binomial coefficients $\binom{n}{a}$ and $\binom{n-b}{a-b}$ are 0 (by (118)). Hence, if we have $a \notin \mathbb{N}$, then the equality (304) boils down to $0 = 0$, which is clearly true. Thus, we WLOG assume that $a \in \mathbb{N}$ in the rest of this proof.

Now, we have $a, b \in \mathbb{N}$, but we don't necessarily have $n \in \mathbb{N}$. If $n \in \mathbb{N}$, then (304) follows from Proposition 7.5.2, but how can we prove it for $n \notin \mathbb{N}$?

We take a broader view. We have defined the binomial coefficient $\binom{n}{k}$ for any **number** n; but we can just as well extend this definition by plugging a polynomial

variable (or even a polynomial) for n. That is, if F is any **polynomial** and k is a number, then we define a **polynomial** $\binom{F}{k}$ as follows:

- If $k \in \mathbb{N}$, then we set

$$\binom{F}{k} = \frac{F(F-1)(F-2)\cdots(F-k+1)}{k!}. \tag{305}$$

- If $k \notin \mathbb{N}$, then we set

$$\binom{F}{k} = 0. \tag{306}$$

This definition is precisely Definition 4.3.1, with n replaced by F. Thus, if F is any polynomial, and if x is any number, then the result of substituting x into the polynomial $\binom{F}{k}$ equals the number $\binom{F(x)}{k}$. This will be used without saying.

With this all said, we can define two polynomials P and Q (with complex coefficients, in a single variable X) by

$$P(X) = \binom{X}{a}\binom{a}{b} \qquad \text{and} \qquad Q(X) = \binom{X}{b}\binom{X-b}{a-b}.$$

[204] Then, for all $x \in \mathbb{N}$, we have

$$\begin{aligned} P(x) &= \binom{x}{a}\binom{a}{b} \qquad \left(\text{since } P(X) = \binom{X}{a}\binom{a}{b}\right) \\ &= \binom{x}{b}\binom{x-b}{a-b} \qquad \text{(by Proposition 7.5.2, applied to } x \text{ instead of } n\text{)} \\ &= Q(x) \end{aligned}$$

(since $Q(X) = \binom{X}{b}\binom{X-b}{a-b}$ and thus $Q(x) = \binom{x}{b}\binom{x-b}{a-b}$). Therefore, Corollary 7.5.7 (applied to $S = \mathbb{N}$) shows that $P = Q$ (since the set $\mathbb{N}$ is infinite). Evaluating both sides of this equality at the number n, we thus find $P(n) = Q(n)$

[204] Explicitly, these polynomials look as follows: We have

$$P(X) = \frac{X(X-1)(X-2)\cdots(X-a+1)}{a!}\binom{a}{b}.$$

Furthermore, if $a - b \in \mathbb{N}$, then

$$Q(X) = \frac{X(X-1)(X-2)\cdots(X-b+1)}{b!} \cdot \frac{(X-b)(X-b-1)(X-b-2)\cdots(X-b-(a-b)+1)}{(a-b)!}.$$

(On the other hand, if $a - b \notin \mathbb{N}$, then $Q(X) = \dfrac{X(X-1)(X-2)\cdots(X-b+1)}{b!} \cdot 0 = 0$.)

(no matter whether $n \in \mathbb{N}$ or $n \notin \mathbb{N}$). But the definitions of P and Q show that $P(n) = \binom{n}{a}\binom{a}{b}$ and $Q(n) = \binom{n}{b}\binom{n-b}{a-b}$. Hence, $\binom{n}{a}\binom{a}{b} = P(n) = Q(n) = \binom{n}{b}\binom{n-b}{a-b}$. This proves (304), and thus Theorem 7.5.4 is proven. □

Note that we have used the polynomial identity trick (in the shape of Corollary 7.5.7) to generalize the equality (304) from $n \in \mathbb{N}$ to $n \in \mathbb{C}$. However, we did **not** use this trick to generalize it from $a \in \mathbb{N}$ to $a \in \mathbb{C}$ (or from $b \in \mathbb{N}$ to $b \in \mathbb{C}$); instead, we have used ad-hoc arguments to prove that (304) holds in the cases $a \notin \mathbb{N}$ and $b \notin \mathbb{N}$. In fact, the polynomial identity trick would not have helped us here. The reason is that, while the binomial coefficient $\binom{n}{k}$ is a polynomial function in n [205], it is **not** a polynomial function in k (indeed, $\binom{n}{k}$ is zero for all negative k, but nonzero for $k = 0$; but this is not how a polynomial function in k could behave). Thus we needed a different argument to generalize from $a, b \in \mathbb{N}$ to $a, b \in \mathbb{C}$.

Next, let us derive Theorem 7.5.3 from Proposition 7.5.1. This will use the polynomial identity trick **twice** – once to generalize from $x \in \mathbb{N}$ to $x \in \mathbb{C}$, and once again to generalize from $y \in \mathbb{N}$ to $y \in \mathbb{C}$. (Alternatively, we could do it in sitting if we generalized Corollary 7.5.7 to polynomials in two variables; this is how it is done in [Grinbe15, §3.3.3, Second proof of Theorem 3.30].)

We begin by showing an "intermediate stage" between Proposition 7.5.1 and Theorem 7.5.3:

Lemma 7.5.8. Let $n \in \mathbb{N}$ and $x \in \mathbb{N}$ and $y \in \mathbb{C}$. Then,

$$\binom{x+y}{n} = \sum_{k=0}^{n} \binom{x}{k}\binom{y}{n-k}.$$

Lemma 7.5.8 is clearly more general than Proposition 7.5.1 but less than Theorem 7.5.3. We shall first deduce this lemma from Proposition 7.5.1, and then use this lemma to deduce Theorem 7.5.3 in turn. Both deductions will use the polynomial identity trick.

Proof of Lemma 7.5.8. It shall be convenient for us to rename the variables x and y as u and v. Thus, we let $n \in \mathbb{N}$ and $u \in \mathbb{N}$ and $v \in \mathbb{C}$. We must then prove that

$$\binom{u+v}{n} = \sum_{k=0}^{n} \binom{u}{k}\binom{v}{n-k}. \tag{307}$$

[205] This means that for any fixed number k, there is a polynomial P such that each number n satisfies $\binom{n}{k} = P(n)$. (Namely, this P is the polynomial $\binom{X}{k}$.)

Define two polynomials P and Q (with complex coefficients, in a single variable X) by

$$P(X) = \binom{u+X}{n} \qquad \text{and} \qquad Q(X) = \sum_{k=0}^{n} \binom{u}{k}\binom{X}{n-k}.$$

[206] Thus, each $x \in \mathbb{N}$ satisfies

$$P(x) = \binom{u+x}{n} \qquad \text{and} \qquad Q(x) = \sum_{k=0}^{n} \binom{u}{k}\binom{x}{n-k}$$

and therefore

$$\begin{aligned} P(x) &= \binom{u+x}{n} = \sum_{k=0}^{n} \binom{u}{k}\binom{x}{n-k} \\ &\qquad \left(\begin{array}{c}\text{by Proposition 7.5.1 (applied to } u \text{ and } x \text{ instead of } x \text{ and } y\text{),} \\ \text{since } u \in \mathbb{N} \text{ and } x \in \mathbb{N}\end{array}\right) \\ &= Q(x). \end{aligned}$$

That is, we have $P(x) = Q(x)$ for all $x \in \mathbb{N}$. Hence, Corollary 7.5.7 (applied to $S = \mathbb{N}$) yields that $P = Q$ (since $\mathbb{N}$ is infinite). Thus, $P(v) = Q(v)$. But the definitions of P and Q yield

$$P(v) = \binom{u+v}{n} \qquad \text{and} \qquad Q(v) = \sum_{k=0}^{n} \binom{u}{k}\binom{v}{n-k}.$$

Hence,

$$\binom{u+v}{n} = P(v) = Q(v) = \sum_{k=0}^{n} \binom{u}{k}\binom{v}{n-k}.$$

This proves (307). Thus, Lemma 7.5.8 is proven. □

Proof of Theorem 7.5.3. It shall again be convenient for us to rename the variables x and y as u and v. Thus, we let $n \in \mathbb{N}$ and $u \in \mathbb{C}$ and $v \in \mathbb{C}$. We must then prove that

$$\binom{u+v}{n} = \sum_{k=0}^{n} \binom{u}{k}\binom{v}{n-k}. \tag{308}$$

Define two polynomials P and Q (with complex coefficients, in a single variable X) by

$$P(X) = \binom{X+v}{n} \qquad \text{and} \qquad Q(X) = \sum_{k=0}^{n} \binom{X}{k}\binom{v}{n-k}.$$

[206] Here we are using binomial coefficients $\binom{F}{k}$ in which F is a polynomial. See our above proof of Theorem 7.5.4 for the definition of this kind of binomial coefficients.

[207] Thus, each $x \in \mathbb{N}$ satisfies

$$P(x) = \binom{x+v}{n} \qquad \text{and} \qquad Q(x) = \sum_{k=0}^{n} \binom{x}{k}\binom{v}{n-k}$$

and therefore

$$\begin{aligned} P(x) &= \binom{x+v}{n} = \sum_{k=0}^{n} \binom{x}{k}\binom{v}{n-k} \\ &\qquad \text{(by Lemma 7.5.8 (applied to } y = v\text{), since } x \in \mathbb{N}) \\ &= Q(x). \end{aligned}$$

That is, we have $P(x) = Q(x)$ for all $x \in \mathbb{N}$. Hence, Corollary 7.5.7 (applied to $S = \mathbb{N}$) yields that $P = Q$ (since $\mathbb{N}$ is infinite). Thus, $P(u) = Q(u)$. But the definitions of P and Q yield

$$P(u) = \binom{u+v}{n} \qquad \text{and} \qquad Q(u) = \sum_{k=0}^{n} \binom{u}{k}\binom{v}{n-k}.$$

Hence,

$$\binom{u+v}{n} = P(u) = Q(u) = \sum_{k=0}^{n} \binom{u}{k}\binom{v}{n-k}.$$

This proves (308). Thus, Theorem 7.5.3 is proven. □

After these proofs, it might appear that the polynomial identity trick is magic. In some ways, this is true, as it allowed us to extend combinatorial proofs to situations in which combinatorics has no say (such as properties of binomial coefficients $\binom{n}{k}$ with non-integer n). However, it has its limitations. For example, we cannot use it to extend Exercise 4.5.8 **(a)** to negative n (since $\sum_{i=0}^{n} \binom{i}{k}$ is not a polynomial function in n); nor can we use it to extend Theorem 4.3.10 to negative n (since $\binom{n}{n-k}$ is not a polynomial function in n); nor can we use it to extend Theorem 4.3.8 to negative n (indeed, $\dfrac{n!}{k! \cdot (n-k)!}$ does not even make sense when n is negative). See [19fco, §2.6.4] for more examples of what can and what cannot be done using the polynomial identity trick.

[207] Here we are using binomial coefficients $\binom{F}{k}$ in which F is a polynomial. See our above proof of Theorem 7.5.4 for the definition of this kind of binomial coefficients.

7.5.4. A probabilistic proof

The next exercise is a classical identity with binomial coefficients ([Engel98, Chapter 5, Example E18], [Engel98, Exercise 8.4], [GrKnPa94, (5.20)], [Grinbe15, Exercise 3.27 **(b)**]):

Exercise 7.5.2. Let $n \in \mathbb{N}$. Prove that

$$\sum_{k=0}^{n} \binom{n+k}{k} \frac{1}{2^k} = 2^n.$$

Several solutions to this exercise are known. In particular, [Grinbe15, Exercise 3.27 **(b)**] gives an algebraic proof by manipulating sums; an elegant proof by induction on n appears in `https://math.stackexchange.com/a/1874857` and [Tomesc85, solution to Problem 1.1 (b)]. More proofs can be found at `https://math.stackexchange.com/questions/1874816` and at `https://math.stackexchange.com/questions/1928040` and at `https://math.stackexchange.com/questions/1782432` and at `https://math.stackexchange.com/questions/3392981` . An equivalent problem is [YagYag64, Problem 73b]. We shall give a probabilistic proof – i.e., in essence, a proof by double counting, except that we formulate it in terms of probabilities rather than numbers. This proof is essentially the one given in [Engel98, Chapter 5, Example E18].

Solution to Exercise 7.5.2 (sketched). Consider a fair coin (i.e., a coin that is equally likely to come up heads as it is to come up tails when it is tossed). We toss this coin $2n+1$ times[208]. Let H be the event "the coin comes up heads at least $n+1$ times", and let T be the event "the coin comes up tails at least $n+1$ times". What can we say about the probabilities $\Pr H$ and $\Pr T$ of the events H and T ?

First of all, the events H and T cannot happen at the same time, because this would require at least $(n+1)+(n+1) = 2n+2$ tosses (but we are only tossing our coin $2n+1$ times). Furthermore, at least one of the events H and T must happen (since otherwise, the coin comes up heads at most n times and comes up tails at most n times, so that we must have tossed it at most $n+n = 2n$ times, which contradicts the fact that we are tossing it $2n+1$ times). Thus, **exactly one** of the events H and T must happen. In other words, the events H and T are complements of each other. Hence, their probabilities add up to 1; that is, we have $\Pr H + \Pr T = 1$.

Now, let us compute $\Pr H$. For each $k \in \{0, 1, \ldots, n\}$, we define the event

$$\begin{aligned} H_k = (&\text{"the coin comes up heads at least } n+1 \text{ times,} \\ &\text{and by the } (n+1)\text{-st time it comes up heads,} \\ &\text{it has come up tails exactly } k \text{ times"}). \end{aligned}$$

[208] The tosses are independent.

For example, if $n = 3$, and if the coin comes up tails first, then heads, then tails again, then heads, then heads, then heads, then tails again, then the event H_2 happens (because the coin has come up tails exactly twice by the 4-th time it comes up heads).

Now we claim that

$$\Pr H = \Pr(H_0) + \Pr(H_1) + \cdots + \Pr(H_n). \tag{309}$$

[*Proof:* If the event H happens, then exactly one of the $n+1$ events $H_0, H_1, \ldots, H_n$ must be happening (because by the $(n+1)$-st time the coin comes up heads, it must have come up tails some number of times between 0 and n [209]). Since the events $H_0, H_1, \ldots, H_n$ are disjoint, and since they are subsets of H, we thus conclude that (309) holds.]

Next, let us fix $k \in \{0, 1, \ldots, n\}$. We want to compute $\Pr(H_k)$. The event H_k can be rewritten as follows:

$$\begin{aligned} H_k = (&\text{"the first } n+k \text{ times that we toss the coin,} \\ &\text{it comes up tails exactly } k \text{ times; furthermore,} \\ &\text{it comes up heads the } (n+k+1)\text{-st time we toss it"}). \end{aligned} \tag{310}$$

Thus, its probability $\Pr(H_k)$ is easy to compute:[210]

$$\Pr(H_k) = \binom{n+k}{k} \cdot \frac{1}{2^{n+k}} \cdot \frac{1}{2}. \tag{311}$$

[*Proof:* Let us first find the probability of the event "the first $n+k$ times that we toss the coin, it comes up tails exactly k times".

Indeed, the first $n+k$ tosses of the coin can give 2^{n+k} different outcomes (where we treat an outcome as an $(n+k)$-tuple of heads/tails events[211]). All these 2^{n+k} outcomes are equally probable (since the coin is fair), and thus have probability $\frac{1}{2^{n+k}}$ each. How many of these 2^{n+k} outcomes have the property that the coin comes up tails exactly k times? An outcome O of the first $n+k$ tosses is uniquely determined by the set

$$I_O := \{i \in \{1, 2, \ldots, n+k\} \mid \text{the coin comes up tails in the } i\text{-th toss}\}.$$

The coin comes up tails exactly k times in this outcome O if and only if the set I_O has size k. Thus, we are looking for the # of k-element subsets of $\{1, 2, \ldots, n+k\}$; but Theorem 4.3.12 shows that this # is $\binom{n+k}{k}$. Thus, there are precisely $\binom{n+k}{k}$ many outcomes (of the first $n+k$ tosses) that have the property that the coin comes

[209] It cannot have come up tails more than n times, since we are only tossing the coin $2n+1$ times.

[210] Computing $\Pr(H_k)$ is known as *Banach's matchbox problem*.

[211] For example, the outcome "the coin comes up heads, then tails, then heads" is encoded as the 3-tuple ("heads", "tails", "heads").

up tails exactly k times. Since each outcome has probability $\dfrac{1}{2^{n+k}}$, we thus conclude that

$$\begin{aligned}&\Pr\left(\text{“the first } n+k \text{ times that we toss the coin,}\right.\\&\qquad\left.\text{it comes up tails exactly } k \text{ times”}\right)\\&=\binom{n+k}{k}\cdot\frac{1}{2^{n+k}}.\end{aligned}\tag{312}$$

Also, since our coin is fair, we have

$$\begin{aligned}&\Pr\left(\text{“the coin comes up heads the } (n+k+1)\text{-st time we toss it”}\right)\\&=\frac{1}{2}.\end{aligned}\tag{313}$$

But the different tosses of the coin are independent. Hence,

$$\begin{aligned}&\Pr\left(\text{“the first } n+k \text{ times that we toss the coin,}\right.\\&\qquad\text{it comes up tails exactly } k \text{ times; furthermore,}\\&\qquad\left.\text{it comes up heads the } (n+k+1)\text{-st time we toss it”}\right)\\&=\Pr\left(\text{“the first } n+k \text{ times that we toss the coin,}\right.\\&\qquad\left.\text{it comes up tails exactly } k \text{ times”}\right)\\&\qquad\cdot\Pr\left(\text{“the coin comes up heads the } (n+k+1)\text{-st time we toss it”}\right)\\&=\binom{n+k}{k}\cdot\frac{1}{2^{n+k}}\cdot\frac{1}{2}\qquad\left(\text{by (312) and (313)}\right).\end{aligned}$$

Since the event on the left hand side of this equality is precisely H_k, we can rewrite this equality as $\Pr(H_k)=\dbinom{n+k}{k}\cdot\dfrac{1}{2^{n+k}}\cdot\dfrac{1}{2}$. Thus, (311) is proved.]

Now, (309) becomes

$$\begin{aligned}\Pr H&=\Pr(H_0)+\Pr(H_1)+\cdots+\Pr(H_n)=\sum_{k=0}^{n}\underbrace{\Pr(H_k)}_{\substack{=\binom{n+k}{k}\cdot\frac{1}{2^{n+k}}\cdot\frac{1}{2}\\ \text{(by (311))}}}\\&=\sum_{k=0}^{n}\binom{n+k}{k}\cdot\underbrace{\frac{1}{2^{n+k}}}_{=\frac{1}{2^n}\cdot\frac{1}{2^k}}\cdot\frac{1}{2}=\sum_{k=0}^{n}\binom{n+k}{k}\cdot\frac{1}{2^n}\cdot\frac{1}{2^k}\cdot\frac{1}{2}\\&=\frac{1}{2}\cdot\frac{1}{2^n}\cdot\sum_{k=0}^{n}\binom{n+k}{k}\frac{1}{2^k}.\end{aligned}$$

The same argument (with the roles of heads and tails interchanged) yields

$$\Pr T = \frac{1}{2} \cdot \frac{1}{2^n} \cdot \sum_{k=0}^{n} \binom{n+k}{k} \frac{1}{2^k}.$$

Adding these two equalities together, we find

$$\begin{aligned}\Pr H + \Pr T &= \frac{1}{2} \cdot \frac{1}{2^n} \cdot \sum_{k=0}^{n} \binom{n+k}{k} \frac{1}{2^k} + \frac{1}{2} \cdot \frac{1}{2^n} \cdot \sum_{k=0}^{n} \binom{n+k}{k} \frac{1}{2^k} \\ &= 2 \cdot \frac{1}{2} \cdot \frac{1}{2^n} \cdot \sum_{k=0}^{n} \binom{n+k}{k} \frac{1}{2^k} = \frac{1}{2^n} \cdot \sum_{k=0}^{n} \binom{n+k}{k} \frac{1}{2^k}.\end{aligned}$$

Comparing this with $\Pr H + \Pr T = 1$, we obtain

$$\frac{1}{2^n} \cdot \sum_{k=0}^{n} \binom{n+k}{k} \frac{1}{2^k} = 1.$$

Hence,

$$\sum_{k=0}^{n} \binom{n+k}{k} \frac{1}{2^k} = 2^n.$$

This solves Exercise 7.5.2. □

The solution we just gave for Exercise 7.5.2 suggests a generalization, which can be obtained if we replace the fair coin by a biased coin. (See [Grinbe15, Exercise 3.27 **(a)**] for this generalization, which is known as the Daubechies identity.)

Results like Theorem 7.5.3, Theorem 7.5.4, Theorem 4.3.10, Proposition 4.3.6, Theorem 4.3.7, Exercise 4.5.8 and Exercise 7.5.2 are commonly called *binomial identities*, as they are identities involving binomial coefficients. We shall see several more such identities in the coming sections. Even more can be found in [GrKnPa94, Chapter 5], [Spivey19], [Riorda68] and [BenQui03]. (The book [BenQui03], in particular, is devoted to proving such identities by double counting.)

7.6. Recitation #7: More on counting and binomial coefficients

Here are some more exercises on counting and binomial coefficients.

7.6.1. More binomial identities

The first of these exercises ([19fco, Proposition 2.6.13]) is a binomial identity sometimes called the "upside-down Chu–Vandermonde identity" (since it looks like Proposition 7.5.1 with the binomial coefficients turned "upside down"):

Exercise 7.6.1. Let $n \in \mathbb{N}$ and $x, y \in \mathbb{N}$. Prove that

$$\binom{n+1}{x+y+1} = \sum_{k=0}^{n} \binom{k}{x} \binom{n-k}{y}.$$

Note that (unlike Proposition 7.5.1) this exercise **cannot** be generalized to $x, y \in \mathbb{C}$. (For instance, $n = 0$ and $x = -1$ and $y = 0$ would be a counterexample.)

First solution to Exercise 7.6.1 (sketched). The following proof is the same as [19fco, §2.6.5, second proof of Proposition 2.6.13].

We apply double counting. Specifically, we ask ourselves: How many $(x+y+1)$-element subsets does the set $[n+1]$ have? As always, we shall answer this question in two ways:

First way: The answer is $\binom{n+1}{x+y+1}$, since Theorem 4.3.12 (applied to $n+1$, $[n+1]$ and $x+y+1$ instead of n, S and k) shows that

$$\begin{aligned} &(\#\text{ of } (x+y+1)\text{-element subsets of } [n+1]) \\ &= \binom{n+1}{x+y+1}. \end{aligned} \tag{314}$$

Second way: If U is any set of integers, and $k \in \{1, 2, \ldots, |U|\}$, then we let $\min_k U$ denote the k-th smallest element of U. For example, $\min_2 \{3, 6, 8, 9\} = 6$ and $\min_4 \{3, 6, 8, 9\} = 9$.

If U is an $(x+y+1)$-element subset of $[n+1]$, then $\min_{x+1} U$ is well-defined (since $x+1 \in \{1, 2, \ldots, x+y+1\} = \{1, 2, \ldots, |U|\}$) and belongs to U and therefore to $[n+1]$ (since $U \subseteq [n+1]$). Hence, the sum rule yields

$$\begin{aligned} &(\#\text{ of } (x+y+1)\text{-element subsets of } [n+1]) \\ &= \sum_{j \in [n+1]} (\#\text{ of } (x+y+1)\text{-element subsets } U \text{ of } [n+1] \text{ satisfying } \min_{x+1} U = j). \end{aligned}$$

(Some addends on the right hand side of this equality will be 0, but as usual we are OK with this.)

Now, fix $j \in [n+1]$. Let us compute

$$(\#\text{ of } (x+y+1)\text{-element subsets } U \text{ of } [n+1] \text{ satisfying } \min_{x+1} U = j).$$

If U is an $(x+y+1)$-element subset of $[n+1]$ satisfying $\min_{x+1} U = j$, then U has precisely x elements smaller than j (by the definition of $\min_{x+1} U$) and therefore precisely $(x+y+1) - (x+1) = y$ elements larger than j. Hence, we can construct such a subset U through the following procedure:

- First, we choose the x elements of U that are smaller than j. We have $\binom{j-1}{x}$ options for this. (Indeed, the x elements we are choosing need to belong

to $[n+1]$ and be smaller than j; that is, they need to come from the set $\{1, 2, \ldots, j-1\}$. Thus, we are really choosing an x-element subset of the set $(j-1)$-element set $\{1, 2, \ldots, j-1\}$. But Theorem 4.3.12 says that there are precisely $\binom{j-1}{x}$ many such subsets.)

- Then, we choose the y elements of U that are larger than j. We have $\binom{n+1-j}{y}$ options for this. (Indeed, the y elements we are choosing need to belong to $[n+1]$ and be larger than j; that is, they need to come from the set $\{j+1, j+2, \ldots, n+1\}$. Thus, we are really choosing a y-element subset of the set $(n+1-j)$-element set $\{j+1, j+2, \ldots, n+1\}$. But Theorem 4.3.12 says that there are precisely $\binom{n+1-j}{y}$ many such subsets.)
- We don't need to decide whether j will be an element of U, since this is already predetermined: We must have $j \in U$ in order to have $\min_{x+1} U = j$.

According to the product rule, there are $\binom{j-1}{x}\binom{n+1-j}{y}$ possibilities for how these choices can be made. Thus,

$$\begin{aligned}&(\#\text{ of } (x+y+1)\text{-element subsets } U \text{ of } [n+1] \text{ satisfying } \min_{x+1} U = j)\\ &= \binom{j-1}{x}\binom{n+1-j}{y}. \end{aligned} \tag{315}$$

Now, forget that we fixed j. Recall that

$$\begin{aligned}&(\#\text{ of } (x+y+1)\text{-element subsets of } [n+1])\\ &= \sum_{j\in[n+1]} \underbrace{(\#\text{ of } (x+y+1)\text{-element subsets } U \text{ of } [n+1] \text{ satisfying } \min_{x+1} U = j)}_{\substack{=\binom{j-1}{x}\binom{n+1-j}{y} \\ \text{(by (315))}}}\\ &= \underbrace{\sum_{j\in[n+1]}}_{=\sum_{j=1}^{n+1}} \binom{j-1}{x} \underbrace{\binom{n+1-j}{y}}_{\substack{=\binom{n-(j-1)}{y} \\ \text{(since } n+1-j=n-(j-1))}} = \sum_{j=1}^{n+1} \binom{j-1}{x}\binom{n-(j-1)}{y}\\ &= \sum_{k=0}^{n} \binom{k}{x}\binom{n-k}{y} \end{aligned} \tag{316}$$

(here, we have substituted k for $j-1$ in the sum).

Comparing (316) with (314), we obtain

$$\binom{n+1}{x+y+1} = \sum_{k=0}^{n} \binom{k}{x} \binom{n-k}{y}.$$

This solves Exercise 7.6.1. □

There is also an alternative, algebraic solution to Exercise 7.6.1, which proceeds by applying the Chu–Vandermonde identity (Theorem 7.5.3) to $n-x-y$, $-x-1$ and $-y-1$ instead of n, x and y (this requires $n-x-y \in \mathbb{N}$, but the case when this is not so can easily be boiled down to $0=0$) and rewriting the result using symmetry (Theorem 4.3.10) and upper negation (Proposition 4.3.6). See [19fco, first proof of Proposition 2.6.13] for the details of this solution. This solution illustrates how useful it can be to plug in "combinatorially meaningless" values (like, in our case, the negative numbers $-x-1$ and $-y-1$) into binomial identities (presuming, of course, that said identities do hold for these values; this is why we have bothered to generalize the Chu–Vandermonde identity from $x, y \in \mathbb{N}$ to $x, y \in \mathbb{C}$).

Remark 7.6.1. Exercise 4.5.8 **(a)** (which is commonly known as the "hockey-stick identity") is a particular case of Exercise 7.6.1. Figure out why! (See [19fco, Exercise 2.6.4] for the answer.)

Here is another binomial identity:

Exercise 7.6.2. Let $a, b \in \mathbb{C}$ and $n, m \in \mathbb{N}$ satisfy $m < n$. Prove that

$$\sum_{j=0}^{n} (-1)^j \binom{n}{j} \binom{aj+b}{m} = 0.$$

Before we solve this exercise, let me quickly mention that Exercise 7.6.2 is a generalization of Proposition 4.3.18. Indeed, if n is a positive integer, then Exercise 7.6.2 (applied to $m=0$, $a=1$ and $b=0$) yields

$$\sum_{j=0}^{n} (-1)^j \binom{n}{j} \binom{1j+0}{0} = 0 = [n=0] \qquad (\text{since } n \neq 0)$$

and thus

$$[n=0] = \sum_{j=0}^{n} (-1)^j \binom{n}{j} \underbrace{\binom{1j+0}{0}}_{\substack{=1 \\ (\text{by } (119))}} = \sum_{j=0}^{n} (-1)^j \binom{n}{j} = \sum_{k=0}^{n} (-1)^k \binom{n}{k};$$

but this is precisely the claim of Proposition 4.3.18 in the case when $n > 0$. Thus, in the case when $n > 0$, Proposition 4.3.18 follows from Exercise 7.6.2. (The case when $n = 0$ is trivial.)

There are various ways to solve Exercise 7.6.2, but the following is perhaps the shortest (and has the advantage of illustrating the usefulness of Exercise 5.4.2):

Solution to Exercise 7.6.2. Define a polynomial P (with complex coefficients, in a single variable X) by

$$P(X) = \binom{-aX+b}{m}.$$

[212] Thus,

$$\begin{aligned} P(X) &= \binom{-aX+b}{m} \\ &= \frac{(-aX+b)(-aX+b-1)(-aX+b-2)\cdots(-aX+b-m+1)}{m!} \\ &\qquad \text{(by (305), since } m \in \mathbb{N}). \end{aligned}$$

Hence, the polynomial P has degree $\leq m$ (since each of the m factors $-aX+b$, $-aX+b-1$, $-aX+b-2$, ..., $-aX+b-m+1$ in the numerator has degree ≤ 1 [213], whereas the denominator is just a constant). Thus, Exercise 5.4.2 **(d)** yields that each $x \in \mathbb{N}$ satisfies

$$\sum_{k=0}^{n} (-1)^k \binom{n}{k} P(x-k) = 0$$

(since $n > m$ (because we assumed that $m < n$)). Applying this to $x = 0$, we obtain

$$\sum_{k=0}^{n} (-1)^k \binom{n}{k} P(0-k) = 0.$$

Thus,

$$\begin{aligned} 0 &= \sum_{k=0}^{n} (-1)^k \binom{n}{k} P\left(\underbrace{0-k}_{=-k}\right) = \sum_{k=0}^{n} (-1)^k \binom{n}{k} \underbrace{P(-k)}_{\substack{=\binom{-a(-k)+b}{m} \\ \left(\text{since } P(X)=\binom{-aX+b}{m}\right)}} \\ &= \sum_{k=0}^{n} (-1)^k \binom{n}{k} \underbrace{\binom{-a(-k)+b}{m}}_{\substack{=\binom{ak+b}{m} \\ (\text{since } -a(-k)=ak)}} = \sum_{k=0}^{n} (-1)^k \binom{n}{k} \binom{ak+b}{m} \\ &= \sum_{j=0}^{n} (-1)^j \binom{n}{j} \binom{aj+b}{m}. \end{aligned}$$

This solves Exercise 7.6.2. □

[212] Here we are using binomial coefficients $\binom{F}{k}$ in which F is a polynomial. See our above proof of Theorem 7.5.4 for the definition of this kind of binomial coefficients.

[213] You may be tempted to say "degree 1", but this would not be quite correct: If $a = 0$, then all these factors are constants and thus have degree 0.

7.6.2. Counting perfect matchings of a finite set

Let us count something else: set partitions.

Definition 7.6.2. Let S be a set.

(a) A *set partition* of S means a set $\{E_1, E_2, \ldots, E_k\}$ of disjoint nonempty subsets of S satisfying $S = E_1 \cup E_2 \cup \cdots \cup E_k$.

(b) If $P = \{E_1, E_2, \ldots, E_k\}$ is a set partition of S, then the sets $E_1, E_2, \ldots, E_k$ are called the *blocks* of P.

Example 7.6.3. (a) Here are three set partitions of the set $[6] = \{1,2,3,4,5,6\}$:

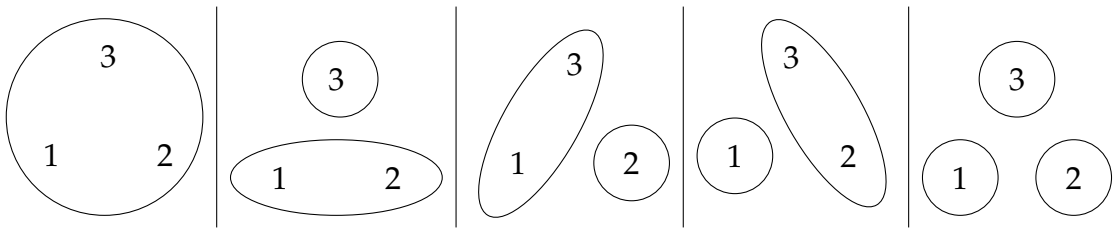
$$\begin{aligned} P_1 &= \{\{1,3\}, \{2,4,5\}, \{6\}\}; \\ P_2 &= \{\{1\}, \{2,4,5,6\}, \{3\}\}; \\ P_3 &= \{\{1\}, \{3\}, \{2,5,6,4\}\}. \end{aligned}$$

Note that the set partitions P_2 and P_3 are actually identical (since the order of elements in a set does not matter).

(b) Here are all set partitions of the set $[3] = \{1,2,3\}$:

$$\{\{1,2,3\}\}, \quad \{\{1,2\},\{3\}\}, \quad \{\{1,3\},\{2\}\}, \quad \{\{2,3\},\{1\}\}, \quad \{\{1\},\{2\},\{3\}\}.$$

And here are the same set partitions, drawn as pictures (each block of the set partition is drawn as a blob):



If S is a set, and if P is a set partition of S, then any element of S belongs to exactly one block of P. Readers familiar with equivalence relations will thus recognize the notion of a set partition as a different language for the notion of an equivalence relation.[214]

[214] Namely:

- If $\sim$ is an equivalence relation on a set S, then the set of all equivalence classes of $\sim$ is a set partition of S.
- If P is a set partition of a set S, then the binary relation "a and b belong to one and the same block of P" (on two elements a and b of S) is an equivalence relation on S.

Thus, we can transform an equivalence relation into a set partition and vice versa. These two transformations are mutually inverse, so they are bijections between

Remark 7.6.4. Let $n \in \mathbb{N}$. The # of all set partitions of $[n]$ is known as the *n-th Bell number* B_n. The sequence $(B_0, B_1, B_2, \ldots) = (1, 1, 2, 5, 15, 52, 203, \ldots)$ is Sequence A000110 in the OEIS, and has the following nice recursive formula:

$$B_{n+1} = \sum_{k=0}^{n} \binom{n}{k} B_k \qquad \text{for each } n \in \mathbb{N}.$$

(It is a neat exercise on counting to prove this. A proof can be found in [Guicha20, Theorem 1.4.3].) No closed form expression for B_n is known. (See [Gardne91, Chapter Two] for a popular-science survey of Bell numbers.)

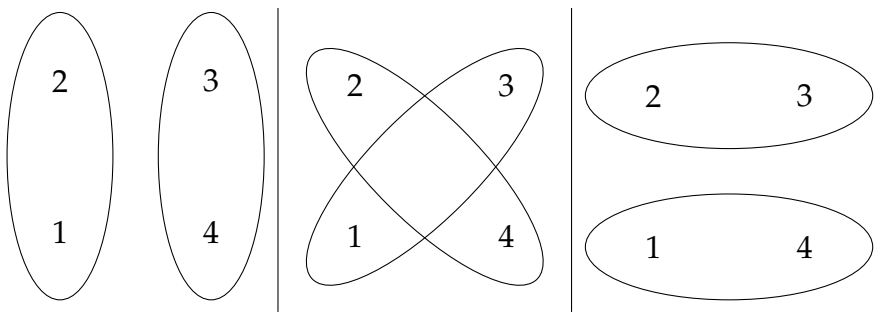
We want to count a certain special type of set partitions:

Definition 7.6.5. Let S be a set. A *perfect matching* of S shall mean a set partition $\{E_1, E_2, \ldots, E_k\}$ of S such that $E_1, E_2, \ldots, E_k$ are 2-element sets.

Example 7.6.6. **(a)** Here are all perfect matchings of the set $[4] = \{1, 2, 3, 4\}$:

$$\{\{1,2\}, \{3,4\}\}, \qquad \{\{1,3\}, \{2,4\}\}, \qquad \{\{1,4\}, \{2,3\}\}.$$

And here are the same perfect matchings, drawn as pictures (each block of the perfect matching is drawn as a blob):



(b) The set $[6]$ has 15 perfect matchings; three of them are

$$\{\{1,2\}, \{3,4\}, \{5,6\}\}, \qquad \{\{1,4\}, \{2,6\}, \{3,5\}\}, \qquad \{\{1,6\}, \{2,5\}, \{3,4\}\}.$$

(c) The set $[5]$ has no perfect matchings.

One can think of a perfect matching of a set S as a way to pair up the elements of S with each other in such a way that each element of S ends up in exactly one pair. (Here, the word "pair" is to be understood in its common-language meaning, not in its mathematical meaning; the correct mathematical word is "two-element set".)

$\{$equivalence relations on $S\}$ and $\{$set partitions on $S\}$ (for a given set S). The proof of this is completely straightforward axiom checking; you can find it spelled out in [Goodma15, Proposition 2.6.7].

Exercise 7.6.3. Let $n \in \mathbb{N}$. Let S be an n-element set. Find the # of all perfect matchings of S.

We shall sketch two solutions to Exercise 7.6.3, without ever going fully rigorous (which would be rather tedious in this particular case). Our first solution will use a decision procedure to construct a perfect matching of S (which will allow counting the perfect matchings of S via the dependent product rule), while our second will exemplify a counting strategy known as *overcounting*.

First solution to Exercise 7.6.3 (sketched). Let us first show the following:

Claim 1: If n is odd, then $(\#\text{ of perfect matchings of } S) = 0$.

[*Proof of Claim 1:* Let n be odd. We must prove that $(\#\text{ of perfect matchings of } S) = 0$.

Assume the contrary. Thus, there exists a perfect matching of S. Consider this perfect matching, and denote it by $\{E_1, E_2, \ldots, E_k\}$ (with $E_1, E_2, \ldots, E_k$ being distinct).

Now, $\{E_1, E_2, \ldots, E_k\}$ is a perfect matching of S. In other words, $\{E_1, E_2, \ldots, E_k\}$ is a set partition of S such that $E_1, E_2, \ldots, E_k$ are 2-element sets (by the definition of "perfect matching"). Since $E_1, E_2, \ldots, E_k$ are 2-element sets, we have $|E_1| = 2$, $|E_2| = 2$, ..., $|E_k| = 2$.

We know that $\{E_1, E_2, \ldots, E_k\}$ is a set partition of S. In other words, $E_1, E_2, \ldots, E_k$ are disjoint subsets of S satisfying $S = E_1 \cup E_2 \cup \cdots \cup E_k$ (by the definition of "set partition"). Now, since S is an n-element set, we have

$$\begin{aligned} n &= |S| = |E_1 \cup E_2 \cup \cdots \cup E_k| && (\text{since } S = E_1 \cup E_2 \cup \cdots \cup E_k) \\ &= |E_1| + |E_2| + \cdots + |E_k| && (\text{since } E_1, E_2, \ldots, E_k \text{ are disjoint}) \\ &= \underbrace{2 + 2 + \cdots + 2}_{k \text{ times}} && (\text{since } |E_1| = 2,\ |E_2| = 2,\ \ldots,\ |E_k| = 2) \\ &= k \cdot 2 = 2k. \end{aligned}$$

This entails that n is even. This contradicts the fact that n is odd. This contradiction shows that our assumption was false. Hence, Claim 1 is proven.]

Claim 1 solves Exercise 7.6.3 in the case when n is odd. Thus, for the rest of this solution, we WLOG assume that n is even. Thus, $n/2 \in \mathbb{N}$.

We WLOG assume that $S = [n]$. Indeed, S is an n-element set, so we can relabel the elements of S as $1, 2, \ldots, n$; this clearly does not change the # of perfect matchings of S (because, as we relabel the elements of S, we can relabel the elements of the blocks in all perfect matchings of S [215]).

If P is a perfect matching of S, then we shall say that two distinct elements of S are *partners* of each other (in P) if and only if they belong to one and the same block

[215] What we are actually using here is the "isomorphism principle". See [19fco, §1.7.2] for some more discussion about this principle.

of P. Given a perfect matching P of S, it is clear that any element of S has exactly one partner (in P), since each block of P is a 2-element set (and since any element of S belongs to exactly one block of P). For example, if P is the perfect matching $\{\{1,4\},\{2,3\}\}$ of the set $\{1,2,3,4\}$, then the partner of 1 (in P) is 4.

Clearly, if P is a perfect matching of S, then the partner of any $p \in S$ (in P) is an element $q \in S$ distinct from p, and the partner of this latter element q is p again. Moreover, a perfect matching of S is uniquely determined if we know the partner of each element. This suggests the following method for constructing a perfect matching P of S:

- Choose the partner of 1 (in P). There are $n-1$ options for this (since the partner of 1 must be distinct from 1). We declare both 1 and its partner to be *partnered*.
- Take the smallest unpartnered[216] element p_2 of S, and choose its partner. There are $n-3$ options (since neither p_2 nor the two partnered elements qualify as partners). We declare both p_2 and its partner to be *partnered*.
- Take the smallest unpartnered element p_3 of S, and choose its partner. There are $n-5$ options (since neither p_3 nor the four partnered elements qualify as partners). We declare both p_3 and its partner to be *partnered*.
- We go on like this until no more unpartnered elements of S remain. Since n is even, this will happen after precisely $n/2$ choices of partners, since every such choice removes exactly 2 elements from the set of unpartnered elements.

The dependent product rule shows that there are precisely

$$(n-1)(n-3)(n-5)\cdots(n-2(n/2)+1)$$

possibilities for how these choices can be made. Each possibility leads to a different perfect matching of S, and each perfect matching of S can be obtained in this way.

[216] We call an element of S *unpartnered* if we have not declared it to be partnered yet. (Of course, in the perfect matching we are constructing, it will eventually have a partner.) The smallest unpartnered element of S is 2 if the partner we have chosen for 1 is distinct from 2; otherwise, it is 3.

Hence,

$$\begin{aligned}
&(\text{\# of perfect matchings of } S) \\
&= (n-1)(n-3)(n-5)\cdots\underbrace{(n-2(n/2)+1)}_{=1} \\
&= (n-1)(n-3)(n-5)\cdots 1 \\
&= 1\cdot 3\cdot 5\cdot\cdots\cdot\left(\underbrace{n}_{=2(n/2)}-1\right) \qquad (317) \\
&= 1\cdot 3\cdot 5\cdot\cdots\cdot(2(n/2)-1) \\
&= \frac{(2(n/2))!}{2^{n/2}(n/2)!} \qquad (\text{by Exercise 4.2.3, applied to } n/2 \text{ instead of } n) \\
&= \frac{n!}{2^{n/2}\cdot(n/2)!} \qquad (\text{since } 2(n/2)=n). \qquad (318)
\end{aligned}$$

Combining this with Claim 1, we thus conclude that

$$(\text{\# of perfect matchings of } S) = \begin{cases} 0, & \text{if } n \text{ is odd;} \\ \dfrac{n!}{2^{n/2}\cdot(n/2)!}, & \text{if } n \text{ is even.} \end{cases}$$

[*Remark:* Instead of proving Claim 1 the way we did above, we could have used the same decision procedure that we used in the case when n is even. This would have given a different proof of Claim 1. Indeed, the last decision in our decision procedure requires choosing a partner for the smallest unpartnered element; if n is odd, then there are 0 options for it, because the smallest unpartnered element at that point is the last unpartnered element and thus cannot find a partner. This entails that (if n is odd) the number of possibilities is $(n-1)(n-3)(n-5)\cdots 0 = 0$. Thus, Claim 1 is proved again.] □

Second solution to Exercise 7.6.3 (sketched). As in the first solution above, we can show that (# of perfect matchings of S) $= 0$ when n is odd. Thus, for the rest of this solution, we WLOG assume that n is even.

Again, as in the first solution above, we WLOG assume that $S = [n]$. Thus, we need to find the # of perfect matchings of $[n]$.

To each permutation σ of $[n]$, we assign the perfect matching

$$M(\sigma) := \{\{\sigma(1), \sigma(2)\}, \ \{\sigma(3), \sigma(4)\}, \ \ldots, \ \{\sigma(n-1), \sigma(n)\}\} \text{ of } [n].$$

(Note that this is well-defined because n is even, and is a perfect matching because σ is bijective.) Thus, we have found a map

$$M : \{\text{permutations of } [n]\} \to \{\text{perfect matchings of } [n]\}.$$

We know that there are $n!$ permutations of $[n]$. What does this mean about the # of perfect matchings of $[n]$?

The map M is not a bijection (unless $n = 0$), since different permutations σ can lead to the same perfect matching $M(\sigma)$. Thus, we cannot apply the bijection principle. However, we might be able to salvage the idea if we can compute **how many** different permutations σ lead to a single perfect matching.

Thus, let us fix a perfect matching P of $[n]$. How many permutations σ of $[n]$ satisfy $M(\sigma) = P$?

Each block of P is a 2-element set (since P is a perfect matching), and the sum of the sizes of these 2-element sets is $|[n]|$ (since P is a set partition of $[n]$). Thus, the # of blocks of P is $\underbrace{|[n]|}_{=n}/2 = n/2$. In other words, P has exactly $n/2$ blocks. Hence, we can write P in the form

$$P = \{\{i_1, i_2\}, \ \{i_3, i_4\}, \ \ldots, \ \{i_{n-1}, i_n\}\},$$

where $i_1, i_2, \ldots, i_n$ are the elements of $[n]$ in some order. If σ is a permutation of $[n]$ that satisfies $M(\sigma) = P$, then we must thus have

$$\begin{aligned}
&\{\{\sigma(1), \sigma(2)\}, \ \{\sigma(3), \sigma(4)\}, \ \ldots, \ \{\sigma(n-1), \sigma(n)\}\} \\
&= M(\sigma) \qquad \text{(by the definition of } M(\sigma)) \\
&= P = \{\{i_1, i_2\}, \ \{i_3, i_4\}, \ \ldots, \ \{i_{n-1}, i_n\}\}.
\end{aligned}$$

This means that the $n/2$ two-element sets

$$\{\sigma(1), \sigma(2)\}, \ \{\sigma(3), \sigma(4)\}, \ \ldots, \ \{\sigma(n-1), \sigma(n)\}$$

must be precisely the $n/2$ two-element sets $\{i_1, i_2\}, \ \{i_3, i_4\}, \ \ldots, \ \{i_{n-1}, i_n\}$ in some order.

This suggests the following method for constructing a permutation σ of $[n]$ satisfying $M(\sigma) = P$:

- First, we decide which of the 2-element sets $\{\sigma(1), \sigma(2)\}, \ \{\sigma(3), \sigma(4)\}, \ \ldots, \ \{\sigma(n-1), \sigma(n)\}$ is which of the 2-element sets $\{i_1, i_2\}, \ \{i_3, i_4\}, \ \ldots, \ \{i_{n-1}, i_n\}$. There are $(n/2)!$ options for this decision, since we are (in effect) choosing a bijection from the set $\{1, 2, \ldots, n/2\}$ to the set $\{\{i_1, i_2\}, \ \{i_3, i_4\}, \ \ldots, \ \{i_{n-1}, i_n\}\}$ (and Theorem 7.4.3 shows that the # of such bijections is $(n/2)!$).

  After this decision, each of the 2-element sets

  $$\{\sigma(1), \sigma(2)\}, \ \{\sigma(3), \sigma(4)\}, \ \ldots, \ \{\sigma(n-1), \sigma(n)\}$$

  is determined.

- Then, we decide which of the two elements of the 2-element set $\{\sigma(1), \sigma(2)\}$ will be $\sigma(1)$. There are 2 options for this. The other (unchosen) element will then be $\sigma(2)$.

- Then, we decide which of the two elements of the 2-element set $\{\sigma(3), \sigma(4)\}$ will be $\sigma(3)$. There are 2 options for this. The other (unchosen) element will then be $\sigma(4)$.
- And so on, until all values $\sigma(1), \sigma(2), \ldots, \sigma(n)$ have been chosen.

The dependent product rule shows that the total # of possibilities for how these choices can be made is

$$(n/2)! \cdot \underbrace{2 \cdot 2 \cdot \cdots \cdot 2}_{n/2 \text{ times}} = (n/2)! \cdot 2^{n/2}.$$

Hence,

$$\begin{aligned} &(\text{\# of permutations } \sigma \text{ of } [n] \text{ satisfying } M(\sigma) = P) \\ &= (n/2)! \cdot 2^{n/2}. \end{aligned} \tag{319}$$

Now, forget that we fixed P. We thus have proved (319) for each perfect matching P of $[n]$. Note that the right hand side of (319) does not depend on P; it is the same for all P. This is what will now allow us to compute the # of perfect matchings of $[n]$.

For each permutation σ of $[n]$, we know that $M(\sigma)$ is a perfect matching of $[n]$. Hence, by the sum rule[217], we have

$$\begin{aligned} &(\text{\# of permutations } \sigma \text{ of } [n]) \\ &= \sum_{\substack{P \text{ is a perfect} \\ \text{matching of } [n]}} \underbrace{(\text{\# of permutations } \sigma \text{ of } [n] \text{ satisfying } M(\sigma) = P)}_{\substack{=(n/2)! \cdot 2^{n/2} \\ \text{(by (319))}}} \\ &= \sum_{\substack{P \text{ is a perfect} \\ \text{matching of } [n]}} (n/2)! \cdot 2^{n/2} \\ &= (\text{\# of perfect matchings of } [n]) \cdot (n/2)! \cdot 2^{n/2}. \end{aligned}$$

Therefore,

$$\begin{aligned} (\text{\# of perfect matchings of } [n]) &= \frac{1}{(n/2)! \cdot 2^{n/2}} \cdot \underbrace{(\text{\# of permutations } \sigma \text{ of } [n])}_{\substack{=n! \\ \text{(by Theorem 7.4.1)}}} \\ &= \frac{1}{(n/2)! \cdot 2^{n/2}} \cdot n! = \frac{n!}{2^{n/2} \cdot (n/2)!}. \end{aligned}$$

This recovers (318) and solves Exercise 7.6.3 again. □

Here is the "sum rule" we used in this solution:

[217] more precisely, by the slightly more flexible version of the sum rule we will state below (as Theorem 7.6.7)

Theorem 7.6.7 (Sum rule in map form)**.** Let S and W be two finite sets. Let $f : S \to W$ be a map. Then,

$$|S| = \sum_{w \in W} (\# \text{ of } s \in S \text{ satisfying } f(s) = w).$$

Theorem 7.6.7 is a more flexible restatement of Theorem 7.1.2 (which is why we also call it the "sum rule"). Indeed, both theorems are formalizing the same underlying principle: namely, that if a set is split into several disjoint subsets, then the size of the set is the sum of the sizes of these subsets. They differ only in how the subsets are encoded (namely, in Theorem 7.1.2, the subsets are $S_1, S_2, \ldots, S_k$; in Theorem 7.6.7, the subsets are $\{s \in S \mid f(s) = w\}$ for various $w \in W$).

The trick we used in our second solution to Exercise 7.6.3 deserves some more comment. We found a map

$$M : \{\text{permutations of } [n]\} \to \{\text{perfect matchings of } [n]\}$$

that is not (in general) a bijection, but nevertheless has the property that it takes all possible values the same number of times. (That is, the # of permutations σ of $[n]$ satisfying $M(\sigma) = P$ is the same for all perfect matchings P.) This allowed us to relate the # of perfect matchings of $[n]$ with the # of permutations of $[n]$. This strategy is known as the *shepherd's principle*: "If you want to count a flock of sheep, count the legs and divide by 4". In our case, the sheep were the perfect matchings of $[n]$, and the legs were the permutations of $[n]$. (So each sheep had $2^{n/2} \cdot (n/2)!$ legs, and the map M sent each leg to its sheep.) Uses of this principle abound in combinatorics; in particular, Theorem 4.3.12 can be proved using it (see [LeLeMe16, §15.5.1] or [19fco, §2.7] for this proof). The shepherd's principle is also known as the *division rule* or *rule of division*, although (as we have seen) it is a simple consequence of the sum rule and there is usually nothing gained by stating it as a separate theorem.

A different variant of our second solution to Exercise 7.6.3 appears in [18f-hw3s, solution to Exercise 3 **(c)**]. We also note that the case of Exercise 7.6.3 for n even appears in [Engel98, Chapter 5, Example E8] (our n is denoted by $2n$ there).

7.7. Homework set #7

This is a **regular** problem set. See Section 3.7 for details on grading.

This homework set covers the above parts of Chapter 7. Some of the problems may be unrelated. Keep in mind that identities between binomial coefficients can be proved in many ways (not just combinatorially)!

Please solve **at most 5** problems. (No points will be given for further solutions.)

Recall Definition 7.4.5 for the following exercise:

Exercise 7.7.1. Let U and V be two finite sets. Let $f : U \to V$ and $g : V \to U$ be two maps. Prove that $|\operatorname{Fix}(f \circ g)| = |\operatorname{Fix}(g \circ f)|$.

Our next exercise is an analogue of Exercise 3.8.3:

Exercise 7.7.2. Let n be a positive integer. Prove that the number of **even** positive divisors of n is even if and only if $n/2$ is not a perfect square.

The next exercise relies on Convention 7.2.1:

Exercise 7.7.3. Let n and r be positive integers. Prove that

$$\sum_{\substack{S \subseteq [n]; \\ |S| = r}} \min S = \binom{n+1}{r+1}.$$

(Recall that the summation sign "$\sum\limits_{\substack{S \subseteq [n]; \\ |S| = r}}$" means "sum over all subsets S of $[n]$ satisfying $|S| = r$", that is, "sum over all r-element subsets S of $[n]$".)

Exercise 7.7.4. Let n be a positive integer. Find the # of compositions of n that don't contain 1 as an entry.

(For example, if $n = 7$, then the compositions of n that don't contain 1 as an entry are (7), $(5,2)$, $(4,3)$, $(3,4)$, $(2,5)$, $(3,2,2)$, $(2,3,2)$ and $(2,2,3)$.)

Exercise 7.7.5. Let n be a positive integer. Let $k \in \mathbb{N}$.

(a) A *composition of n into k parts* means a composition of n that has exactly k entries (i.e., is a k-tuple). Find the # of compositions of n into k parts.

(b) A *weak composition of n into k parts* means a k-tuple of **nonnegative** integers whose sum is n. (So it is like a composition of n into k parts, but its entries are allowed to be 0.) Find the # of weak compositions of n into k parts.

Exercise 7.7.6. Let $n \in \mathbb{N}$. Prove that

$$\sum_{k=0}^{n-1} \binom{n-1}{k} \frac{(k+1)!}{n^k} = n.$$

Exercise 7.7.7. Let $n \in \mathbb{N}$. Prove that

$$\sum_{k=0}^{n} k \binom{2n}{n-k} = n \binom{2n-1}{n}.$$

Exercise 7.7.8. Let $(f_0, f_1, f_2, \ldots)$ and $(g_0, g_1, g_2, \ldots)$ be two sequences of numbers. Assume that

$$g_n = \sum_{i=0}^{n} (-1)^i \binom{n}{i} f_i \qquad \text{for every } n \in \mathbb{N}.$$

Prove that

$$f_n = \sum_{i=0}^{n} (-1)^i \binom{n}{i} g_i \qquad \text{for every } n \in \mathbb{N}.$$

Exercise 7.7.9. Let $n \in \mathbb{N}$.

(a) Prove that

$$\binom{-1/2}{n} = \left(\frac{-1}{4}\right)^n \binom{2n}{n}.$$

(b) Prove that

$$\sum_{k=0}^{n} \binom{2k}{k} \binom{2(n-k)}{n-k} = 4^n.$$

Exercise 7.7.10. Prove that there is a **unique** sequence $(u_0, u_1, u_2, \ldots)$ of positive integers such that

$$u_n^2 = \sum_{r=0}^{n} \binom{n+r}{r} u_{n-r} \qquad \text{for all } n \in \mathbb{N}.$$

7.8. Alternating sums and Inclusion/Exclusion

7.8.1. The Principle of Inclusion and Exclusion

Consider the following:

- If U is a finite set, and if A is a subset of U, then

$$|U \setminus A| = |U| - |A|. \tag{320}$$

  In fact, this is just Theorem 7.1.8 (applied to U and A instead of A and B).

- If U is a finite set, and if A and B are two subsets of U, then

$$|U \setminus (A \cup B)| = |U| - |A| - |B| + |A \cap B|. \tag{321}$$

  Indeed, this is not hard to see by applying (320) three times[218]. Alternatively, there is a simple intuitive way to convince yourself of (321): If A and B are two

[218] Here are the details: Let U be a finite set, and let A and B be two subsets of U. Then, it is known from set theory (or easy to verify by drawing Venn diagrams) that $U \setminus (A \cup B) =$

subsets of a finite set U, then you can try to count the elements of $U \setminus (A \cup B)$ by starting with $|U|$ and subtracting $|A|$ and $|B|$. However, this will leave the elements of $A \cap B$ subtracted twice instead of once; this flaw can be corrected by adding $|A \cap B|$ back in. This leads to the equality (321).

- If U is a finite set, and if A, B and C are three subsets of U, then

$$\begin{aligned}|U \setminus (A \cup B \cup C)| = |U| - |A| - |B| - |C| + |A \cap B| + |A \cap C| + |B \cap C| \\ - |A \cap B \cap C| . \end{aligned} \tag{322}$$

This, too, can be proved by applying (320) several times (or, more easily, by applying (320) once and (321) twice, using $U \setminus (A \cup B \cup C) = (U \setminus (A \cup B)) \setminus (C \setminus ((A \cap C) \cup (B \cap C)))$). We leave the proof to the interested reader. Again,

---

$(U \setminus A) \setminus (B \setminus (A \cap B))$. Hence,

$$\begin{aligned}|U \setminus (A \cup B)| &= |(U \setminus A) \setminus (B \setminus (A \cap B))| = \underbrace{|U \setminus A|}_{\substack{=|U|-|A| \\ \text{(by (320))}}} - \underbrace{|B \setminus (A \cap B)|}_{\substack{=|B|-|A\cap B| \\ \text{(by (320)} \\ \text{(applied to } B \text{ and } A\cap B \\ \text{instead of } U \text{ and } A\text{),} \\ \text{since } A\cap B \text{ is a subset of } B\text{)}}} \\ &\qquad \left(\begin{array}{c}\text{by (320) (applied to } U \setminus A \text{ and } B \setminus (A \cap B) \\ \text{instead of } U \text{ and } A\text{), since } B \setminus (A \cap B) \text{ is a subset of } U \setminus A\end{array}\right) \\ &= (|U| - |A|) - (|B| - |A \cap B|) = |U| - |A| - |B| + |A \cap B| .\end{aligned}$$

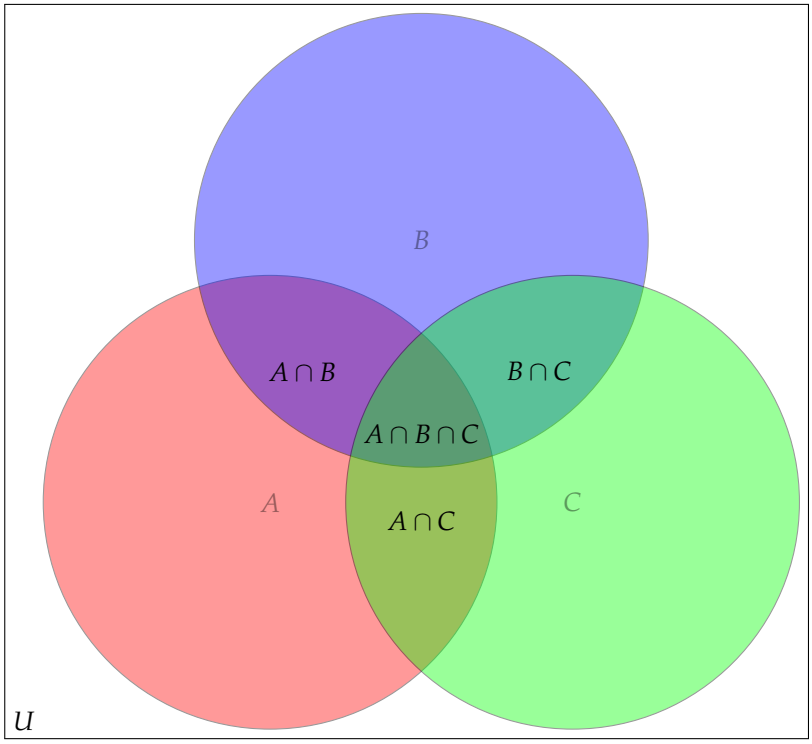
Thus, (321) is proved.

there also is a way to convince yourself of (322) by a counting argument[219]: If A, B and C are three subsets of a finite set U, then you can count the elements of $U \setminus (A \cup B \cup C)$ by starting with $|U|$, then subtracting $|A|$, $|B|$ and $|C|$, then correcting for the doubly subtracted common elements by adding $|A \cap B|$, $|A \cap C|$ and $|B \cap C|$, and finally correcting for the elements of $A \cap B \cap C$ (which have been added once, subtracted thrice and added back in thrice, but should not be counted) by subtracting $|A \cap B \cap C|$.

The three identities (320), (321) and (322) appear to be parts of a pattern: The left hand sides have the form $|U \setminus (A_1 \cup A_2 \cup \cdots \cup A_n)|$ for some subsets $A_1, A_2, \ldots, A_n$ of a finite set U, and the right hand sides are signed sums of the sizes of intersections of some of these subsets (including $|U|$, which we consider to be the "trivial intersection"). This pattern indeed persists; let us state it in its general form:

Theorem 7.8.1 (Principle of Inclusion and Exclusion (complement form))**.** Let

[219]Best done with a Venn diagram at hand (courtesy of Alain Matthes on tex.stackexchange):



$n \in \mathbb{N}$. Let U be a finite set. Let $A_1, A_2, \ldots, A_n$ be n subsets of U. Then,

$$\begin{aligned}
&|U \setminus (A_1 \cup A_2 \cup \cdots \cup A_n)| \\
&= |U| \\
&\qquad \underbrace{- |A_1| - |A_2| - \cdots - |A_n|}_{\text{all sizes } |A_i|} \\
&\qquad \underbrace{+ |A_1 \cap A_2| + |A_1 \cap A_3| + \cdots + |A_{n-1} \cap A_n|}_{\text{all sizes } |A_i \cap A_j| \text{ with } i<j} \\
&\qquad \underbrace{- |A_1 \cap A_2 \cap A_3| - |A_1 \cap A_2 \cap A_4| - \cdots - |A_{n-2} \cap A_{n-1} \cap A_n|}_{\text{all sizes } |A_i \cap A_j \cap A_k| \text{ with } i<j<k} \\
&\qquad \pm \cdots \\
&\qquad + (-1)^n |A_1 \cap A_2 \cap \cdots \cap A_n| \\
&= \sum_{m=0}^{n} (-1)^m \sum_{\substack{(i_1, i_2, \ldots, i_m) \in [n]^m; \\ i_1 < i_2 < \cdots < i_m}} \left| A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_m} \right| .
\end{aligned}$$

Here, the "empty" intersection $A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_m}$ for $m = 0$ is understood to mean the set U.

The right hand side of the equality in Theorem 7.8.1 is rather unwieldy, and the "middle hand side" (which is just a rewritten version of the right hand side without using the $\sum$ sign) is even worse. Thus, let us introduce a notation that will allow us to restate Theorem 7.8.1 in a simpler way:

Definition 7.8.2. Let I be a nonempty set. For each $i \in I$, let A_i be a set. Then, $\bigcap\limits_{i \in I} A_i$ denotes the set

$$\{x \mid x \in A_i \text{ for each } i \in I\} .$$

This set $\bigcap\limits_{i \in I} A_i$ is called the intersection of all A_i with $i \in I$.

It is easy to see that if $I = \{i_1, i_2, \ldots, i_m\}$ is a finite set, and if A_i is a set for each $i \in I$, then

$$\bigcap_{i \in I} A_i = A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_m}. \tag{323}$$

The notation $\bigcap\limits_{i \in I} A_i$ (for the intersection of a family of sets) is similar to the notation $\sum\limits_{i \in I} a_i$ (for the sum of a finite family of numbers). However, in the notation $\bigcap\limits_{i \in I} A_i$, the set I must be nonempty (unlike in $\sum\limits_{i \in I} a_i$) but can be infinite (unlike in $\sum\limits_{i \in I} a_i$, which is usually undefined when I is infinite).

Now, we can restate Theorem 7.8.1 as follows:

Theorem 7.8.3 (Principle of Inclusion and Exclusion (complement form))**.** Let $n \in \mathbb{N}$. Let U be a finite set. Let $A_1, A_2, \ldots, A_n$ be n subsets of U. Then,

$$|U \setminus (A_1 \cup A_2 \cup \cdots \cup A_n)| = \sum_{I \subseteq [n]} (-1)^{|I|} \left| \bigcap_{i \in I} A_i \right|. \tag{324}$$

Here, the "empty" intersection $\bigcap\limits_{i \in \varnothing} A_i$ is understood to mean the set U.

Remark 7.8.4. We have defined the "empty" intersection $\bigcap\limits_{i \in \varnothing} A_i$ in Theorem 7.8.3; why didn't we already do this back in Definition 7.8.2?

The reason is that we have had U available in Theorem 7.8.3, and, in a sense, were treating $A_1, A_2, \ldots, A_n$ as subsets of U rather than just as arbitrary sets. Had we tried to extend Definition 7.8.2 to the case when $I = \varnothing$, we would have obtained

$$\bigcap_{i \in \varnothing} A_i = \{x \mid x \in A_i \text{ for each } i \in \varnothing\} = \{\text{all } x\}$$

(since the condition "$x \in A_i$ for each $i \in \varnothing$" is vacuously true, whatever x is), which is not a well-defined set (it would be a so-called "universal set", which does not exist in set theory as it would lead to Russell's paradox). (And even if it was a well-defined set, its size would certainly be infinite.) However, in Theorem 7.8.3, we defined $\bigcap\limits_{i \in \varnothing} A_i$ to mean U, which is a reasonable interpretation of the (otherwise ill-defined) set "$\{\text{all } x\}$" if we consider x to be implicitly required to be an element of U.

(Note that defining $\bigcap\limits_{i \in \varnothing} A_i$ to mean U is slightly slippery, since it entails that $\bigcap\limits_{i \in \varnothing} A_i$ depends not just on the n sets $A_1, A_2, \ldots, A_n$ but also on U, even though U is not part of the notation. In practice, this causes no real confusion unless one works with two different U's at the same time – which one rarely does.)

Combining Theorem 7.8.3 with the difference rule, we also obtain the following:

Theorem 7.8.5 (Principle of Inclusion and Exclusion (union form))**.** Let $n \in \mathbb{N}$. Let $A_1, A_2, \ldots, A_n$ be n finite sets. Then,

$$|A_1 \cup A_2 \cup \cdots \cup A_n| = \sum_{\substack{I \subseteq [n]; \\ I \neq \varnothing}} (-1)^{|I|-1} \left| \bigcap_{i \in I} A_i \right|. \tag{325}$$

Theorem 7.8.3 and Theorem 7.8.5 are known as the *principle(s) of inclusion and exclusion* or as the *Sylvester sieve formulas*. (The word "sieve" refers to the metaphorical view that the sums on the right hand sides of (324) and (325) are "sieving" the

elements of U, leaving only those in $U \setminus (A_1 \cup A_2 \cup \cdots \cup A_n)$ (resp. $A_1 \cup A_2 \cup \cdots \cup A_n$) behind.)

We won't prove Theorem 7.8.1, Theorem 7.8.3 and Theorem 7.8.5 right now; the impatient reader can find their proofs all over the literature:

- Theorem 7.8.3 is [19fco, Theorem 2.9.7] (where it is proved using manipulation of finite sums and Iverson brackets). It also appears in [Smid09] (with a proof by induction), in [Galvin17, Theorem 16.1] (with two proofs) and in [Grinbe15, Theorem 3.42] (in a slightly more general form).

- Theorem 7.8.5 is [19fco, Theorem 2.9.6] (where it is derived from Theorem 7.8.3 by setting $U = A_1 \cup A_2 \cup \cdots \cup A_n$ and using the difference rule). It also appears in [White10b], in [Galvin17, (12)] and in [Grinbe15, Theorem 3.43].

- Theorem 7.8.1 is a restatement of Theorem 7.8.5. Indeed, the awkward double sum
$$\sum_{m=0}^{n} (-1)^m \sum_{\substack{(i_1,i_2,\ldots,i_m)\in[n]^m; \\ i_1<i_2<\cdots<i_m}} \left|A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_m}\right|$$
(when expanded) has the same addends as the sum $\sum_{I\subseteq[n]} (-1)^{|I|} \left|\bigcap_{i\in I} A_i\right|$; only the indexing of these addends is different. (See [19fco, proof of Proposition 2.9.4] for the details of this argument, although our sums have a $|U|$ addend whereas the sums in [19fco, proof of Proposition 2.9.4] do not.)

Several generalizations of the Principles of Inclusion and Exclusion can be found in [Grinbe15, Theorems 3.44, 3.45 and 3.46], in [Aigner07, §5.1 and §5.2] and in [Comtet74, Chapter IV].

In practice, the following restatement of Theorem 7.8.3 tends to be most convenient to apply:

Theorem 7.8.6 (Principle of Inclusion and Exclusion (complement form))**.** Let $n \in \mathbb{N}$. Let U be a finite set. Let $A_1, A_2, \ldots, A_n$ be n subsets of U. Then,
$$\begin{aligned}&(\#\text{ of } x \in U \text{ satisfying } x \notin A_i \text{ for all } i \in [n]) \\ &= \sum_{I\subseteq[n]} (-1)^{|I|} (\#\text{ of } x \in U \text{ satisfying } x \in A_i \text{ for all } i \in I)\,. \qquad (326)\end{aligned}$$

Theorem 7.8.6 is equivalent to Theorem 7.8.3 because each side of the equality (326) equals to the corresponding side of (324). (We shall explain this in more detail in Subsection 7.8.3 below.)

7.8.2. An example: counting surjections

The principle of inclusion and exclusion can be used to count many kinds of objects. See [19fco, §2.9], [Aigner07, §5.1], [AndFen04, Chapter 6], [Tomesc85, Chapter 2], [Galvin17, §16–§17] or [Guicha20, §2.2] for various examples. We shall show just one application here (which may be one of the most famous): We shall count the surjective maps between two given finite sets. Note that we have already counted all maps (Theorem 7.3.6), all injective maps (Theorem 7.3.8) and all bijective maps (Theorem 7.4.3 when the two sets have equal sizes; otherwise there are none). Counting the surjective maps is the hardest of these questions; here is the answer:

Theorem 7.8.7. Let $m, n \in \mathbb{N}$. Let A be an m-element set. Let B be an n-element set. Then,

$$(\text{\# of all surjective maps from } A \text{ to } B) = \sum_{i=0}^{n} (-1)^{n-i} \binom{n}{i} i^m.$$

Proof of Theorem 7.8.7 (sketched). (See [19fco, §2.4.5, proof of Proposition 2.4.11, and §2.9.4, proof of Theorem 2.4.17] for more details.[220]) We WLOG assume that $B = [n]$ (since otherwise, we can just relabel the n elements of B as $1, 2, \ldots, n$, without changing the # of all surjective maps from A to B).

We say that a map $f : A \to B$ *misses* an element $i \in B$ if f does not take i as a value (i.e., if there exists no $a \in A$ such that $i = f(a)$). (The terminology is motivated by the standard illustrations in which maps are drawn as collections of arrows.)

We let $U = B^A = \{\text{all maps from } A \text{ to } B\}$. For each $i \in [n]$, we define a subset A_i of U by

$$\begin{aligned} A_i &= \{\text{all maps } f : A \to B \text{ that miss } i\} \\ &= \{\text{all maps } f : A \to B \text{ that do not take } i \text{ as a value}\}. \end{aligned}$$

Thus, we have defined n subsets $A_1, A_2, \ldots, A_n$ of U. Hence, Theorem 7.8.6 yields that (326) holds.

[220] See also [Gunder10, Exercise 595] or [19fco, solution to Exercise 2.4.4] for alternative proofs.

However, the left hand side of (326) is

$$\begin{aligned}
&(\#\text{ of } x \in U \text{ satisfying } x \notin A_i \text{ for all } i \in [n]) \\
&= (\#\text{ of } f \in U \text{ satisfying } f \notin A_i \text{ for all } i \in [n]) \\
&= (\#\text{ of all maps } f : A \to B \text{ satisfying } f \notin A_i \text{ for all } i \in [n]) \\
&\qquad\qquad (\text{since } U = \{\text{all maps from } A \text{ to } B\}) \\
&= (\#\text{ of all maps } f : A \to B \text{ such that } f \text{ takes } i \text{ as a value for all } i \in [n]) \\
&\qquad\qquad \left(\begin{array}{c}\text{since the statement "}f \notin A_i\text{" is equivalent} \\ \text{to "}f\text{ takes }i\text{ as a value" (by the definition of } A_i)\end{array}\right) \\
&= (\#\text{ of all maps } f : A \to B \text{ that take each } i \in [n] \text{ as a value}) \\
&= (\#\text{ of all maps } f : A \to B \text{ that take each } i \in B \text{ as a value}) \\
&\qquad\qquad (\text{since } [n] = B) \\
&= (\#\text{ of all surjective maps from } A \text{ to } B)\,. \qquad (327)
\end{aligned}$$

Meanwhile, we can also rewrite the numbers on the right hand side of (326): Let I be any subset of $[n]$. Then,

$$\begin{aligned}
&(\#\text{ of } x \in U \text{ satisfying } x \in A_i \text{ for all } i \in I) \\
&= (\#\text{ of } f \in U \text{ satisfying } f \in A_i \text{ for all } i \in I) \\
&= (\#\text{ of all maps } f : A \to B \text{ satisfying } f \in A_i \text{ for all } i \in I) \\
&\qquad\qquad (\text{since } U = \{\text{all maps from } A \text{ to } B\}) \\
&= (\#\text{ of all maps } f : A \to B \text{ such that } f \text{ misses } i \text{ for all } i \in I) \\
&\qquad\qquad \left(\begin{array}{c}\text{since the statement "}f \in A_i\text{" is equivalent} \\ \text{to "}f\text{ misses }i\text{" (by the definition of } A_i)\end{array}\right) \\
&= (\#\text{ of all maps } f : A \to B \text{ such that } f \text{ misses all elements of } I) \\
&= (\#\text{ of all maps } f : A \to B \text{ such that no value of } f \text{ belongs to } I) \\
&= (\#\text{ of all maps } f : A \to B \text{ such that all values of } f \text{ belong to } B \setminus I) \\
&= (\#\text{ of all maps from } A \text{ to } B \setminus I)\,.
\end{aligned}$$

Here, the last equality sign is due to the fact that the maps $f : A \to B$ such that all values of f belong to $B \setminus I$ are "essentially the same as" the maps from A to $B \setminus I$ (or, to be more precise, there is a bijection from the set of the former maps to the set of the latter maps; but all this bijection does is change the target of the map, without changing any of its values). Hence,

$$\begin{aligned}
&(\#\text{ of } x \in U \text{ satisfying } x \in A_i \text{ for all } i \in I) \\
&= (\#\text{ of all maps from } A \text{ to } B \setminus I) = \left|(B \setminus I)^A\right| = |B \setminus I|^{|A|} \\
&\qquad (\text{by Theorem 7.3.6, applied to } B \setminus I \text{ instead of } B) \\
&= (|B| - |I|)^{|A|} \qquad (\text{since } |B \setminus I| = |B| - |I| \text{ (because } I \subseteq [n] = B)) \\
&= (n - |I|)^m \qquad (328)
\end{aligned}$$

(since $|B| = n$ and $|A| = m$).

Forget that we fixed I. We thus have proved (328) for each subset I of $[n]$. Now, (327) entails

$$\begin{aligned}
&(\text{\# of all surjective maps from } A \text{ to } B) \\
&= (\text{\# of } x \in U \text{ satisfying } x \notin A_i \text{ for all } i \in [n]) \\
&= \sum_{I \subseteq [n]} (-1)^{|I|} \underbrace{(\text{\# of } x \in U \text{ satisfying } x \in A_i \text{ for all } i \in I)}_{\substack{=(n-|I|)^m \\ \text{(by (328))}}} \qquad \text{(by (326))} \\
&= \sum_{I \subseteq [n]} (-1)^{|I|} (n-|I|)^m = \sum_{k=0}^{n} \sum_{\substack{I \subseteq [n]; \\ |I|=k}} \underbrace{(-1)^{|I|} (n-|I|)^m}_{\substack{=(-1)^k (n-k)^m \\ \text{(since } |I|=k)}} \\
&\qquad \left(\begin{array}{c} \text{here, we have split the sum according to the value of } |I|, \\ \text{since } |I| \in \{0,1,\ldots,n\} \text{ for each subset } I \text{ of } [n] \end{array} \right) \\
&= \sum_{k=0}^{n} \underbrace{\sum_{\substack{I \subseteq [n]; \\ |I|=k}} (-1)^k (n-k)^m}_{=(\text{\# of subsets } I \text{ of } [n] \text{ satisfying } |I|=k) \cdot (-1)^k (n-k)^m} \\
&= \sum_{k=0}^{n} \underbrace{(\text{\# of subsets } I \text{ of } [n] \text{ satisfying } |I| = k)}_{\substack{=(\text{\# of } k\text{-element subsets of } [n]) \\ = \binom{n}{k} \\ \text{(by Theorem 4.3.12)}}} \cdot (-1)^k (n-k)^m \\
&= \sum_{k=0}^{n} \binom{n}{k} \cdot (-1)^k (n-k)^m = \sum_{k=0}^{n} (-1)^k \underbrace{\binom{n}{k}}_{\substack{= \binom{n}{n-k} \\ \text{(by Theorem 4.3.10)}}} (n-k)^m \\
&= \sum_{k=0}^{n} (-1)^k \binom{n}{n-k} (n-k)^m = \sum_{i=0}^{n} (-1)^{n-i} \binom{n}{i} i^m
\end{aligned}$$

(here, we have substituted $n - i$ for k in the sum). This proves Theorem 7.8.7. $\square$

As a welcome byproduct of Theorem 7.8.7, we obtain two binomial identities:

Corollary 7.8.8. **(a)** If $n, m \in \mathbb{N}$ satisfy $m < n$, then

$$\sum_{i=0}^{n} (-1)^{n-i} \binom{n}{i} i^m = 0.$$

(b) For any $n \in \mathbb{N}$, we have

$$\sum_{i=0}^{n} (-1)^{n-i} \binom{n}{i} i^n = n!.$$

Proof of Corollary 7.8.8. **(a)** We could easily obtain Corollary 7.8.8 **(a)** by applying Exercise 5.4.2 **(d)** to the polynomial $P(x) = x^m$; however, let us instead derive it from Theorem 7.8.7.

Indeed, let $n, m \in \mathbb{N}$ satisfy $m < n$. Then, $|[m]| = m$ and $|[n]| = n$. Thus, $|[m]| = m < n = |[n]|$; therefore, Theorem 6.1.7 (applied to $U = [m]$ and $V = [n]$) shows that a map $f : [m] \to [n]$ cannot be surjective. In other words, there exists no surjective map $f : [m] \to [n]$. In other words,

$$(\text{\# of all surjective maps from } [m] \text{ to } [n]) = 0.$$

However, Theorem 7.8.7 (applied to $A = [m]$ and $B = [n]$) yields

$$(\text{\# of all surjective maps from } [m] \text{ to } [n]) = \sum_{i=0}^{n} (-1)^{n-i} \binom{n}{i} i^m$$

(since $[m]$ is an m-element set, whereas $[n]$ is an n-element set). Comparing these two equalities, we obtain $\sum_{i=0}^{n} (-1)^{n-i} \binom{n}{i} i^m = 0$. This proves Corollary 7.8.8 **(a)**.

(b) Let $n \in \mathbb{N}$. Corollary 6.2.9 **(b)** (applied to $X = [n]$) shows that any surjective map $f : [n] \to [n]$ is a permutation of $[n]$. In other words, any surjective map from $[n]$ to $[n]$ is a permutation of $[n]$. Conversely, any permutation of $[n]$ is a surjective map from $[n]$ to $[n]$ (because any permutation is bijective and thus surjective). Combining the results of the previous two sentences, we conclude that the permutations of $[n]$ are precisely the surjective maps from $[n]$ to $[n]$. Hence,

$$\begin{aligned} &(\text{\# of all permutations of } [n]) \\ &= (\text{\# of all surjective maps from } [n] \text{ to } [n]) = \sum_{i=0}^{n} (-1)^{n-i} \binom{n}{i} i^n \end{aligned}$$

(by Theorem 7.8.7, applied to $m = n$, $A = [n]$ and $B = [n]$). Comparing this with

$$(\text{\# of all permutations of } [n]) = n! \qquad (\text{by Theorem 7.4.1, applied to } X = [n]),$$

we obtain $\sum_{i=0}^{n} (-1)^{n-i} \binom{n}{i} i^n = n!$. This proves Corollary 7.8.8 **(b)**. □

7.8.3. A weighted version and a proof

The principle of inclusion and exclusion in all its forms we have seen above is far from being the pinnacle of generality. One of its most useful generalizations is

obtained by assigning *weights* (i.e., arbitrary numbers) to the elements of U. Then, the size $|T|$ of a subset T of U can be replaced by its "total weight" $\sum_{x\in T} w(x)$ (where $w(x)$ is the weight assigned to an $x \in U$). This suggests the following generalization of Theorem 7.8.6:

Theorem 7.8.9 (Weighted Principle of Inclusion and Exclusion (complement form))**.** Let $n \in \mathbb{N}$. Let U be a finite set. For each $x \in U$, let $w(x)$ be a number. Let $A_1, A_2, \ldots, A_n$ be n subsets of U. Then,

$$\sum_{\substack{x\in U;\\ x\notin A_i \text{ for all } i\in[n]}} w(x) = \sum_{I\subseteq[n]} (-1)^{|I|} \sum_{\substack{x\in U;\\ x\in A_i \text{ for all } i\in I}} w(x).$$

We shall prove Theorem 7.8.9 and then derive Theorem 7.8.6 and Theorem 7.8.3 from it. The proof will rely on an innocent little proposition which I occasionally call the *cancellation lemma*, as it is about a sum of $+1$s and -1s that cancel each other out[221]:

Proposition 7.8.10. Let S be a finite set. Then,

$$\sum_{I\subseteq S} (-1)^{|I|} = [S = \varnothing].$$

Once again, we are using the Iverson bracket notation here (see Definition 4.3.19).

Example 7.8.11. The subsets of $\{1,2\}$ are $\varnothing$, $\{1\}$, $\{2\}$ and $\{1,2\}$. Thus, applying Proposition 7.8.10 to $S = \{1,2\}$, we find

$$\underbrace{(-1)^{|\varnothing|}}_{=1} + \underbrace{(-1)^{|\{1\}|}}_{=-1} + \underbrace{(-1)^{|\{2\}|}}_{=-1} + \underbrace{(-1)^{|\{1,2\}|}}_{=1} = [\{1,2\} = \varnothing].$$

Indeed, both sides of this equality are 0 (the left hand side because the addends cancel; the right hand side because $\{1,2\} \neq \varnothing$).

[221] **Aside:** This is far from the only "cancellation lemma" in mathematics! There are various others, such as $\sum_{k=0}^{n-1} \cos\frac{2kg\pi}{n} = n\cdot[n \mid g]$ for any positive integer n and any $g \in \mathbb{Z}$. Here, it is no longer $+1$s and -1s cancelling each other, but rather different phases of the cosine wave. With complex numbers available, this can be improved even further to

$$\sum_{k=0}^{n-1} e^{2kg\pi i/n} = n\cdot[n \mid g],$$

which is the driving force behind the discrete Fourier transform.

There are philosophical similarities between the cancellation lemma (and its combinatorial uses to "extract" certain kinds of addends from sums) and the Fourier inversion formula (and its Fourier-analytic uses to extract certain frequencies from waves).

Proof of Proposition 7.8.10 (sketched). Proposition 7.8.10 appears in [19fco, Proposition 2.9.10] with two detailed proofs; we shall only outline one here.

First of all, we observe that $[|S| = 0] = [S = \varnothing]$. Indeed, if $S = \varnothing$, then both truth values $[S = \varnothing]$ and $[|S| = 0]$ equal 1; otherwise, they both equal 0.

Let $n = |S|$. Hence, S is an n-element set. Thus, if I is any subset of S, then $|I| \in \{0, 1, \ldots, n\}$. Hence, we can split the sum $\sum\limits_{I \subseteq S} (-1)^{|I|}$ according to the value of $|I|$ as follows:

$$\begin{aligned}
&\sum_{I \subseteq S} (-1)^{|I|} \\
&= \sum_{k \in \{0,1,\ldots,n\}} \sum_{\substack{I \subseteq S; \\ |I| = k}} \underbrace{(-1)^{|I|}}_{\substack{=(-1)^k \\ \text{(since } |I|=k)}} = \sum_{k \in \{0,1,\ldots,n\}} \underbrace{\sum_{\substack{I \subseteq S; \\ |I| = k}} (-1)^k}_{=(\#\text{ of subsets } I \text{ of } S \text{ satisfying } |I|=k) \cdot (-1)^k} \\
&= \underbrace{\sum_{k \in \{0,1,\ldots,n\}}}_{= \sum\limits_{k=0}^{n}} \underbrace{\underbrace{(\#\text{ of subsets } I \text{ of } S \text{ satisfying } |I| = k)}_{=(\#\text{ of } k\text{-element subsets of } S)}}_{\substack{= \binom{n}{k} \\ \text{(by Theorem 4.3.12)}}} \cdot (-1)^k \\
&= \sum_{k=0}^{n} \binom{n}{k} (-1)^k = \sum_{k=0}^{n} (-1)^k \binom{n}{k} = [n = 0] \qquad \text{(by Proposition 4.3.18)} \\
&= [|S| = 0] \qquad (\text{since } n = |S|) \\
&= [S = \varnothing] .
\end{aligned}$$

This proves Proposition 7.8.10. □

It is now easy to prove Theorems 7.8.9, 7.8.6 and 7.8.3:

Proof of Theorem 7.8.9. We have

$$\begin{aligned}
\sum_{I \subseteq [n]} (-1)^{|I|} \sum_{\substack{x \in U; \\ x \in A_i \text{ for all } i \in I}} w(x) &= \sum_{I \subseteq [n]} \sum_{\substack{x \in U; \\ x \in A_i \text{ for all } i \in I}} (-1)^{|I|} w(x) \\
&= \sum_{x \in U} \sum_{\substack{I \subseteq [n]; \\ x \in A_i \text{ for all } i \in I}} (-1)^{|I|} w(x) . \qquad (329)
\end{aligned}$$

(Here, for the last equality sign, we have interchanged the two summation signs; as usual, the "$x \in A_i$ for all $i \in I$" condition moved under the inner summation sign. This relied on Theorem 4.1.25.)

Now, fix $x \in U$. We shall compute the sum $\sum\limits_{\substack{I \subseteq [n]; \\ x \in A_i \text{ for all } i \in I}} (-1)^{|I|}$.

Indeed, define a subset S of $[n]$ by

$$S = \{i \in [n] \mid x \in A_i\}.$$

Then, for any subset I of $[n]$, we have the logical equivalence

$$(x \in A_i \text{ for all } i \in I) \iff (I \subseteq S)$$

(that is, the statement "$x \in A_i$ for all $i \in I$" holds if and only if the statement "$I \subseteq S$" holds)[222]. Hence, the condition "$x \in A_i$ for all $i \in I$" under the summation sign $\sum\limits_{\substack{I \subseteq [n]; \\ x \in A_i \text{ for all } i \in I}}$ can be replaced by "$I \subseteq S$". We thus obtain

$$\begin{aligned}
\sum_{\substack{I \subseteq [n]; \\ x \in A_i \text{ for all } i \in I}} (-1)^{|I|} &= \sum_{\substack{I \subseteq [n]; \\ I \subseteq S}} (-1)^{|I|} = \sum_{I \subseteq S} (-1)^{|I|} \\
&\qquad \left(\begin{array}{l}\text{since the subsets } I \text{ of } [n] \text{ satisfying } I \subseteq S \\ \text{are just the subsets of } S \text{ (because } S \subseteq [n]\text{)}\end{array}\right) \\
&= [S = \varnothing] \qquad \text{(by Proposition 7.8.10)} \\
&= [\{i \in [n] \mid x \in A_i\} = \varnothing] \qquad (331)
\end{aligned}$$

(since $S = \{i \in [n] \mid x \in A_i\}$).

On the other hand, we have the following chain of logical equivalences:

$$\begin{aligned}
&(\{i \in [n] \mid x \in A_i\} = \varnothing) \\
\iff &(\text{there exists no } i \in [n] \text{ such that } x \in A_i) \\
\iff &(x \notin A_i \text{ for all } i \in [n]). \qquad (332)
\end{aligned}$$

However, equivalent logical statements have equal truth values: i.e., if $\mathcal{A}$ and $\mathcal{B}$ are two equivalent logical statements, then $[\mathcal{A}] = [\mathcal{B}]$. Therefore, from the equivalence (332), we obtain the equality

$$[\{i \in [n] \mid x \in A_i\} = \varnothing] = [x \notin A_i \text{ for all } i \in [n]].$$

[222] *Proof.* Let I be a subset of $[n]$. For any $i \in [n]$, we have the logical equivalence

$$(i \in S) \iff (x \in A_i) \qquad (330)$$

(because $S = \{i \in [n] \mid x \in A_i\}$). Now, we have the following chain of equivalences:

$$(I \subseteq S) \iff \left(\underbrace{i \in S}_{\substack{\iff (x \in A_i) \\ \text{(by (330))}}} \text{ for all } i \in I\right) \iff (x \in A_i \text{ for all } i \in I).$$

Thus, we have the equivalence $(x \in A_i \text{ for all } i \in I) \iff (I \subseteq S)$, qed.

Thus, (331) rewrites as

$$\sum_{\substack{I\subseteq[n];\\ x\in A_i \text{ for all } i\in I}} (-1)^{|I|} = [x \notin A_i \text{ for all } i \in [n]]\,. \tag{333}$$

Now, forget that we fixed x. We thus have proved (333) for each $x \in U$.

Now, (329) becomes

$$\begin{aligned}
&\sum_{I\subseteq[n]} (-1)^{|I|} \sum_{\substack{x\in U;\\ x\in A_i \text{ for all } i\in I}} w(x) \\
&= \sum_{x\in U} \sum_{\substack{I\subseteq[n];\\ x\in A_i \text{ for all } i\in I}} (-1)^{|I|} w(x) \\
&= \sum_{x\in U} \underbrace{\left(\sum_{\substack{I\subseteq[n];\\ x\in A_i \text{ for all } i\in I}} (-1)^{|I|} \right)}_{\substack{=[x\notin A_i \text{ for all } i\in[n]] \\ \text{(by (333))}}} w(x) \\
&= \sum_{x\in U} [x \notin A_i \text{ for all } i \in [n]]\, w(x) \\
&= \sum_{\substack{x\in U;\\ x\notin A_i \text{ for all } i\in[n]}} \underbrace{[x \notin A_i \text{ for all } i \in [n]]}_{\substack{=1 \\ \text{(since we have } (x\notin A_i \text{ for all } i\in[n]))}} w(x) \\
&\qquad + \sum_{\substack{x\in U;\\ \text{not } (x\notin A_i \text{ for all } i\in[n])}} \underbrace{[x \notin A_i \text{ for all } i \in [n]]}_{\substack{=0 \\ \text{(since we don't have } (x\notin A_i \text{ for all } i\in[n]))}} w(x) \\
&\qquad \left(\begin{array}{c} \text{since each } x \in U \text{ either satisfies } (x \notin A_i \text{ for all } i \in [n]) \\ \text{or does not} \end{array} \right) \\
&= \sum_{\substack{x\in U;\\ x\notin A_i \text{ for all } i\in[n]}} \underbrace{1w(x)}_{=w(x)} + \underbrace{\sum_{\substack{x\in U;\\ \text{not } (x\notin A_i \text{ for all } i\in[n])}} 0w(x)}_{=0} = \sum_{\substack{x\in U;\\ x\notin A_i \text{ for all } i\in[n]}} w(x)\,.
\end{aligned}$$

This proves Theorem 7.8.9. □

Proof of Theorem 7.8.6. For each $x \in U$, set $w(x) = 1$. Then, Theorem 7.8.9 yields

$$\sum_{\substack{x\in U;\\ x\notin A_i \text{ for all } i\in[n]}} w(x) = \sum_{I\subseteq[n]} (-1)^{|I|} \sum_{\substack{x\in U;\\ x\in A_i \text{ for all } i\in I}} w(x)\,. \tag{334}$$

However,

$$\begin{aligned}
&\sum_{\substack{x\in U;\\ x\notin A_i \text{ for all } i\in[n]}} \underbrace{w(x)}_{\substack{=1\\ \text{(by the definition of } w(x))}} \\
&= \sum_{\substack{x\in U;\\ x\notin A_i \text{ for all } i\in[n]}} 1 = (\#\text{ of } x\in U \text{ satisfying } x\notin A_i \text{ for all } i\in[n])\cdot 1 \\
&= (\#\text{ of } x\in U \text{ satisfying } x\notin A_i \text{ for all } i\in[n]) . \qquad (335)
\end{aligned}$$

Moreover, each subset I of $[n]$ satisfies

$$\begin{aligned}
&\sum_{\substack{x\in U;\\ x\in A_i \text{ for all } i\in I}} \underbrace{w(x)}_{\substack{=1\\ \text{(by the definition of } w(x))}} \\
&= \sum_{\substack{x\in U;\\ x\in A_i \text{ for all } i\in I}} 1 = (\#\text{ of } x\in U \text{ satisfying } x\in A_i \text{ for all } i\in I)\cdot 1 \\
&= (\#\text{ of } x\in U \text{ satisfying } x\in A_i \text{ for all } i\in I) . \qquad (336)
\end{aligned}$$

In light of (335) and (336), we can rewrite the equality (334) as

$$\begin{aligned}
&(\#\text{ of } x\in U \text{ satisfying } x\notin A_i \text{ for all } i\in[n]) \\
&= \sum_{I\subseteq[n]} (-1)^{|I|} (\#\text{ of } x\in U \text{ satisfying } x\in A_i \text{ for all } i\in I) .
\end{aligned}$$

This proves Theorem 7.8.6. □

Proof of Theorem 7.8.3. The set $U\setminus(A_1\cup A_2\cup\cdots\cup A_n)$ consists of all $x\in U$ satisfying $x\notin A_1\cup A_2\cup\cdots\cup A_n$ (by the definition of a set difference). In other words, this set consists of all $x\in U$ satisfying $x\notin A_i$ for all $i\in[n]$ (because the condition "$x\notin A_1\cup A_2\cup\cdots\cup A_n$" is equivalent to "$x\notin A_i$ for all $i\in[n]$"). Hence,

$$\begin{aligned}
&|U\setminus(A_1\cup A_2\cup\cdots\cup A_n)| \\
&= (\#\text{ of } x\in U \text{ satisfying } x\notin A_i \text{ for all } i\in[n]) \\
&= \sum_{I\subseteq[n]} (-1)^{|I|} (\#\text{ of } x\in U \text{ satisfying } x\in A_i \text{ for all } i\in I) \qquad (337)
\end{aligned}$$

(by Theorem 7.8.6).

Now, we claim the following:

Claim 1: Let I be a subset of $[n]$. Then,

$$\bigcap_{i\in I} A_i = \{x\in U \mid x\in A_i \text{ for all } i\in I\} .$$

[*Proof of Claim 1:* Let us warm up by studying the empty set. We have

$$\{x \in U \mid x \in A_i \text{ for all } i \in \varnothing\} = U, \tag{338}$$

since the condition "$x \in A_i$ for all $i \in \varnothing$" is vacuously true for all $x \in U$. On the other hand,

$$\bigcap_{i \in \varnothing} A_i = U,$$

since we defined the "empty intersection" $\bigcap\limits_{i \in \varnothing} A_i$ to be U. Comparing this with (338), we obtain

$$\bigcap_{i \in \varnothing} A_i = \{x \in U \mid x \in A_i \text{ for all } i \in \varnothing\}.$$

In other words, Claim 1 holds for $I = \varnothing$. Thus, for the rest of this proof, we WLOG assume that $I \neq \varnothing$. This assumption is helpful, since we defined $\bigcap\limits_{i \in \varnothing} A_i$ differently than we defined $\bigcap\limits_{i \in I} A_i$ for nonempty I.

Now we know that I is nonempty (since $I \neq \varnothing$). Hence, $\bigcap\limits_{i \in I} A_i$ is literally the intersection of the sets A_i for $i \in I$. Since the latter sets A_i are subsets of U, we thus conclude that their intersection $\bigcap\limits_{i \in I} A_i$ is also a subset of U (since any intersection of subsets of U is a subset of U). Therefore,

$$\begin{aligned}
\bigcap_{i \in I} A_i &= U \cap \left(\bigcap_{i \in I} A_i \right) \\
&= U \cap \{x \mid x \in A_i \text{ for each } i \in I\} \\
&\qquad \left(\begin{array}{c} \text{since the definition of } \bigcap\limits_{i \in I} A_i \\ \text{yields } \bigcap\limits_{i \in I} A_i = \{x \mid x \in A_i \text{ for each } i \in I\} \end{array} \right) \\
&= \{x \in U \mid x \in A_i \text{ for each } i \in I\} \\
&= \{x \in U \mid x \in A_i \text{ for all } i \in I\}.
\end{aligned}$$

This proves Claim 1.]

Now,

$$\begin{aligned}
&\sum_{I\subseteq[n]}(-1)^{|I|}\left|\underbrace{\bigcap_{i\in I}A_i}_{\substack{=\{x\in U\ \mid\ x\in A_i \text{ for all } i\in I\}\\ \text{(by Claim 1)}}}\right| \\
&= \sum_{I\subseteq[n]}(-1)^{|I|}\underbrace{\left|\{x\in U\ \mid\ x\in A_i \text{ for all } i\in I\}\right|}_{=(\#\text{ of } x\in U \text{ satisfying } x\in A_i \text{ for all } i\in I)} \\
&= \sum_{I\subseteq[n]}(-1)^{|I|}\left(\#\text{ of } x\in U \text{ satisfying } x\in A_i \text{ for all } i\in I\right).
\end{aligned}$$

Comparing this with (337), we obtain

$$|U\setminus(A_1\cup A_2\cup\cdots\cup A_n)| = \sum_{I\subseteq[n]}(-1)^{|I|}\left|\bigcap_{i\in I}A_i\right|.$$

This proves Theorem 7.8.3. □

Finally, Theorem 7.8.5 follows easily from Theorem 7.8.3.[223]

As many other properties of sums, Theorem 7.8.9 has an analogue for integrals over measurable sets. It says that a (Lebesgue) measurable function w defined on a measurable subset $U\in\mathcal{A}$ of a measure space $(X,\mathcal{A},\mu)$ satisfies

$$\int_{U\setminus(A_1\cup A_2\cup\cdots\cup A_n)} w\,d\mu = \sum_{I\subseteq[n]}(-1)^{|I|}\int_{\bigcap_{i\in I}A_i} w\,d\mu$$

whenever $A_1,A_2,\ldots,A_n$ are n measurable subsets of U. As a particular case, if $E_1,E_2,\ldots,E_n$ are n events in a probability space, then

$$\Pr\left(\text{neither } E_1 \text{ nor } E_2 \text{ nor } E_3 \text{ nor } \cdots \text{ nor } E_n\right) = \sum_{I\subseteq[n]}(-1)^{|I|}\Pr\left(E_i \text{ for all } i\in I\right).$$

[223] Here is the main idea: Let $A_1,A_2,\ldots,A_n$ be n finite sets. Then, define $U=A_1\cup A_2\cup\cdots\cup A_n$. This set U is finite again, and the sets $A_1,A_2,\ldots,A_n$ are subsets of U. Thus, Theorem 7.8.3 shows that (324) holds. Subtract this equality (324) from the tautological equality $|U|=|U|$. The result is

$$|U|-|U\setminus(A_1\cup A_2\cup\cdots\cup A_n)| = |U| - \sum_{I\subseteq[n]}(-1)^{|I|}\left|\bigcap_{i\in I}A_i\right|.$$

But the left hand side of this equality is easily seen to be $|A_1\cup A_2\cup\cdots\cup A_n|$, whereas the right hand side is $\sum_{\substack{I\subseteq[n];\\ I\neq\varnothing}}(-1)^{|I|-1}\left|\bigcap_{i\in I}A_i\right|$ (since the $|U|$ term cancels the $I=\varnothing$ addend in the sum). Thus, we obtain (325), and thus Theorem 7.8.5 is proved.

7.8.4. Recitation #8: More subtractive counting

We have already seen (e.g., in our solution to Exercise 7.3.2 **(c)**) how the difference rule can help in solving enumerative problems. In essence, the idea is that in order to count some objects that satisfy a certain property, it can be better to first count the objects that **don't** satisfy it. The Principle of Inclusion and Exclusion can be viewed as an extension of this to several properties. Let us now see some other uses of negativity (differences and powers of -1) in enumerative combinatorics.

Recall our convention $[d] = \{1, 2, \ldots, d\}$ for each $d \in \mathbb{N}$. (This was Convention 7.2.1.)

Exercise 7.8.1. Let $n \in \mathbb{N}$. Let d be a positive integer.

An n-tuple $(x_1, x_2, \ldots, x_n) \in [d]^n$ will be called 1-*even* if the number 1 occurs in it an even number of times (i.e., the number of $i \in [n]$ satisfying $x_i = 1$ is even). (For example, the 3-tuples $(5, 1, 1)$ and $(2, 2, 3)$ are 1-even, while the 3-tuple $(4, 2, 1)$ is not.)

Compute the # of 1-even n-tuples in $[d]^n$.

[**Example:** If $d = 3$ and $n = 2$, then the 1-even n-tuples in $[d]^n$ are $(1, 1)$, $(2, 2)$, $(2, 3)$ and $(3, 2)$.]

This exercise is [18s-hw3s, Exercise 5], but we shall give two solutions different from the one given in [18s-hw3s] (although the second one is related).

First solution of Exercise 7.8.1 (sketched). Forget that we fixed n. For each $n \in \mathbb{N}$, we let e_n denote the # of 1-even n-tuples in $[d]^n$. We thus must compute e_n for each $n \in \mathbb{N}$.

We shall do this similarly to how we solved Exercise 7.3.2 **(c)**: We will first establish a recursive equation for e_n in terms of e_{n-1}, and then solve this equation to obtain an explicit formula.

We first notice that $e_0 = 1$. Indeed, there is only one 0-tuple in $[d]^0$ (namely, the empty list $()$), and this 0-tuple is 1-even (since the number 1 occurs 0 times in it).

Now, let n be a positive integer. Then, the definition of e_n yields[224]

$$\begin{aligned} e_n &= (\text{\# of 1-even } n\text{-tuples}) \\ &= \sum_{k=1}^{d} (\text{\# of 1-even } n\text{-tuples } (x_1, x_2, \ldots, x_n) \text{ satisfying } x_n = k) \end{aligned}$$

(by the sum rule, since $x_n \in [d]$ for each n-tuple $(x_1, x_2, \ldots, x_n)$).

We now want to compute the # of 1-even n-tuples $(x_1, x_2, \ldots, x_n) \in [d]^n$ satisfying $x_n = k$ for a given $k \in [d]$. The answer will depend on whether k is 1 or not.

[224] In the following argument, the word "n-tuple" will always mean "n-tuple in $[d]^n$", and similarly the word "$(n-1)$-tuple" will always mean "$(n-1)$-tuple in $[d]^{n-1}$". We will not need tuples of elements of sets other than $[d]$.

First, we can find the answer if k is not 1: For each $k \in [d]$ satisfying $k \neq 1$, we have

$$\begin{aligned}&(\text{\# of 1-even } n\text{-tuples } (x_1, x_2, \ldots, x_n) \text{ satisfying } x_n = k)\\&= e_{n-1}.\end{aligned} \tag{339}$$

[*Proof of (339):* Let $k \in [d]$ satisfy $k \neq 1$. If an n-tuple $(x_1, x_2, \ldots, x_n)$ satisfies $x_n = k$, then the number 1 occurs as often in the $(n-1)$-tuple $(x_1, x_2, \ldots, x_{n-1})$ as it does in the n-tuple $(x_1, x_2, \ldots, x_n)$ (since we have $x_n = k \neq 1$, and thus the removal of the last entry x_n from the n-tuple $(x_1, x_2, \ldots, x_n)$ does not affect the occurrences of 1 in this n-tuple). Thus, in particular, if an n-tuple $(x_1, x_2, \ldots, x_n)$ satisfying $x_n = k$ is 1-even, then the $(n-1)$-tuple $(x_1, x_2, \ldots, x_{n-1})$ is 1-even. Hence, we obtain a map

$$\begin{aligned}\{\text{1-even } n\text{-tuples } (x_1, x_2, \ldots, x_n) \text{ satisfying } x_n = k\} &\to \{\text{1-even } (n-1)\text{-tuples}\},\\ (x_1, x_2, \ldots, x_n) &\mapsto (x_1, x_2, \ldots, x_{n-1}).\end{aligned}$$

This map is easily seen to be a bijection[225]. Thus, the bijection principle yields

$$\begin{aligned}&|\{\text{1-even } n\text{-tuples } (x_1, x_2, \ldots, x_n) \text{ satisfying } x_n = k\}|\\&= |\{\text{1-even } (n-1)\text{-tuples}\}| = (\text{\# of 1-even } (n-1)\text{-tuples}) = e_{n-1}\end{aligned}$$

(since e_{n-1} was defined as the # of 1-even $(n-1)$-tuples). But this is clearly equivalent to (339). This proves (339).]

Next, let us answer our question for $k = 1$: We have

$$\begin{aligned}&(\text{\# of 1-even } n\text{-tuples } (x_1, x_2, \ldots, x_n) \text{ satisfying } x_n = 1)\\&= d^{n-1} - e_{n-1}.\end{aligned} \tag{340}$$

[*Proof of (340):* The set of all $(n-1)$-tuples is $[d]^{n-1}$. Thus, $(\text{\# of } (n-1)\text{-tuples}) = \left|[d]^{n-1}\right| = |[d]|^{n-1} = d^{n-1}$ (since $|[d]| = d$).

We need one more notion: An $(n-1)$-tuple will be called 1-*odd* if the number 1 occurs in it an odd number of times. Thus, any $(n-1)$-tuple is either 1-even or 1-odd (but not both at the same time). Thus, the sum rule yields

$$(\text{\# of } (n-1)\text{-tuples}) = (\text{\# of 1-even } (n-1)\text{-tuples}) + (\text{\# of 1-odd } (n-1)\text{-tuples}),$$

so that

$$\begin{aligned}(\text{\# of 1-odd } (n-1)\text{-tuples}) &= \underbrace{(\text{\# of } (n-1)\text{-tuples})}_{=d^{n-1}} - \underbrace{(\text{\# of 1-even } (n-1)\text{-tuples})}_{\substack{=e_{n-1}\\ \text{(by the definition of } e_{n-1})}}\\&= d^{n-1} - e_{n-1}.\end{aligned}$$

[225] Indeed, the map

$$\begin{aligned}\{\text{1-even } (n-1)\text{-tuples}\} &\to \{\text{1-even } n\text{-tuples } (x_1, x_2, \ldots, x_n) \text{ satisfying } x_n = k\},\\ (x_1, x_2, \ldots, x_{n-1}) &\mapsto (x_1, x_2, \ldots, x_{n-1}, k)\end{aligned}$$

is easily seen to be well-defined and inverse to it.

Now, what does this have to do with the 1-even n-tuples $(x_1, x_2, \ldots, x_n)$ satisfying $x_n = 1$?

If an n-tuple $(x_1, x_2, \ldots, x_n)$ satisfies $x_n = 1$, then the number 1 occurs one less time in the $(n-1)$-tuple $(x_1, x_2, \ldots, x_{n-1})$ than it does in the n-tuple $(x_1, x_2, \ldots, x_n)$ (since we have $x_n = 1$, and thus the removal of the last entry x_n from the n-tuple $(x_1, x_2, \ldots, x_n)$ removes exactly one occurrence of 1 from this n-tuple). Thus, in particular, if an n-tuple $(x_1, x_2, \ldots, x_n)$ satisfies $x_n = 1$, and if the number 1 occurs an even number of times in this n-tuple, then the number 1 will occur an odd number of times in the $(n-1)$-tuple $(x_1, x_2, \ldots, x_{n-1})$. In other words, if an n-tuple $(x_1, x_2, \ldots, x_n)$ satisfying $x_n = 1$ is 1-even, then the $(n-1)$-tuple $(x_1, x_2, \ldots, x_{n-1})$ is 1-odd. Hence, we obtain a map

$$\begin{aligned}\{\text{1-even } n\text{-tuples } (x_1, x_2, \ldots, x_n) \text{ satisfying } x_n = 1\} &\to \{\text{1-odd } (n-1)\text{-tuples}\}, \\ (x_1, x_2, \ldots, x_n) &\mapsto (x_1, x_2, \ldots, x_{n-1}).\end{aligned}$$

This map is easily seen to be a bijection[226]. Thus, the bijection principle yields

$$\begin{aligned}&|\{\text{1-even } n\text{-tuples } (x_1, x_2, \ldots, x_n) \text{ satisfying } x_n = 1\}| \\ &= |\{\text{1-odd } (n-1)\text{-tuples}\}| = (\text{\# of 1-odd } (n-1)\text{-tuples}) = d^{n-1} - e_{n-1}.\end{aligned}$$

But this is clearly equivalent to (340). This proves (340).]

Now, recall that

$$\begin{aligned}e_n &= \sum_{k=1}^{d} (\text{\# of 1-even } n\text{-tuples } (x_1, x_2, \ldots, x_n) \text{ satisfying } x_n = k) \\ &= \underbrace{(\text{\# of 1-even } n\text{-tuples } (x_1, x_2, \ldots, x_n) \text{ satisfying } x_n = 1)}_{\substack{= d^{n-1} - e_{n-1} \\ \text{(by (340))}}} \\ &\qquad + \sum_{k=2}^{d} \underbrace{(\text{\# of 1-even } n\text{-tuples } (x_1, x_2, \ldots, x_n) \text{ satisfying } x_n = k)}_{\substack{= e_{n-1} \\ \text{(by (339) (since } k \geq 2 > 1 \text{ and thus } k \neq 1))}} \\ &\qquad \left(\begin{array}{c}\text{here, we have split off the addend for } k = 1 \\ \text{from the sum}\end{array}\right) \\ &= d^{n-1} - e_{n-1} + \underbrace{\sum_{k=2}^{d} e_{n-1}}_{= (d-1) e_{n-1}} = d^{n-1} - e_{n-1} + (d-1) e_{n-1} \\ &= d^{n-1} + (d-2) e_{n-1}.\end{aligned}$$

[226] Indeed, the map

$$\begin{aligned}\{\text{1-odd } (n-1)\text{-tuples}\} &\to \{\text{1-even } n\text{-tuples } (x_1, x_2, \ldots, x_n) \text{ satisfying } x_n = 1\}, \\ (x_1, x_2, \ldots, x_{n-1}) &\mapsto (x_1, x_2, \ldots, x_{n-1}, 1)\end{aligned}$$

is easily seen to be well-defined and inverse to it.

Forget that we fixed n. We thus have shown that

$$e_n = d^{n-1} + (d-2)\, e_{n-1} \tag{341}$$

for every positive integer n. Set $u = d - 2$. Then, (341) rewrites as

$$e_n = d^{n-1} + u e_{n-1}. \tag{342}$$

We can attempt to solve this recurrence by substituting this equality back into itself:

$$\begin{aligned}
e_n &= d^{n-1} + u \underbrace{e_{n-1}}_{\substack{=d^{n-2}+ue_{n-2} \\ \text{(by (342))}}} \qquad \text{(by (342))} \\
&= d^{n-1} + u \left(d^{n-2} + u \underbrace{e_{n-2}}_{\substack{=d^{n-3}+ue_{n-3} \\ \text{(by (342))}}} \right) \\
&= d^{n-1} + u \left(d^{n-2} + u \left(d^{n-3} + u \underbrace{e_{n-3}}_{=\cdots} \right) \right) \\
&= \cdots \\
&= d^{n-1} + u \left(d^{n-2} + u \left(d^{n-3} + u \left(\cdots + u \left(d^0 + u \underbrace{e_0}_{=1} \right) \right) \right) \right) \\
&= d^{n-1} + u \left(d^{n-2} + u \left(d^{n-3} + u \left(\cdots + u \left(d^0 + u \right) \right) \right) \right) \\
&= \underbrace{d^{n-1} + u d^{n-2} + u^2 d^{n-3} + u^3 d^{n-4} + \cdots + u^{n-1} d^0}_{\substack{= \sum\limits_{i=0}^{n-1} u^i d^{n-1-i} \\ = \dfrac{u^n - d^n}{u-d} \\ \text{(since (82) (applied to } m=n \text{ and } a=u \text{ and } b=d) \\ \text{yields } (u-d) \sum\limits_{i=0}^{n-1} u^i d^{n-1-i} = u^n - d^n)}} + u^n \\
&= \frac{u^n - d^n}{u-d} + u^n = \frac{u^n - d^n}{-2} + u^n \qquad \left(\text{since } \underbrace{u}_{=d-2} - d = d - 2 - d = -2 \right) \\
&= \frac{1}{2} \left(d^n + u^n \right) = \frac{1}{2} \left(d^n + (d-2)^n \right) \qquad (\text{since } u = d - 2).
\end{aligned}$$

Thus we have obtained the explicit formula

$$e_n = \frac{1}{2} \left(d^n + (d-2)^n \right) \tag{343}$$

for each $n \in \mathbb{N}$.

As with most such formulas, once they are known, they are easy to prove by induction. For example, here is a quick inductive proof of (343):

[*Alternative proof of (343):* We shall prove (343) by induction on n:

Induction base: Comparing $e_0 = 1$ with $\frac{1}{2}\left(\underbrace{d^0}_{=1} + \underbrace{(d-2)^0}_{=1}\right) = \frac{1}{2}(1+1) = 1$, we obtain $e_0 = \frac{1}{2}\left(d^0 + (d-2)^0\right)$. In other words, (343) holds for $n = 0$.

Induction step: Let m be a positive integer. Assume (as the induction hypothesis) that (343) holds for $n = m-1$. We must show that (343) holds for $n = m$.

We have assumed that (343) holds for $n = m-1$. In other words, we have $e_{m-1} = \frac{1}{2}\left(d^{m-1} + (d-2)^{m-1}\right)$. Now, (341) (applied to $n = m$) yields

$$\begin{aligned}
e_m &= d^{m-1} + (d-2) \underbrace{e_{m-1}}_{=\frac{1}{2}\left(d^{m-1}+(d-2)^{m-1}\right)} \\
&= d^{m-1} + (d-2) \cdot \frac{1}{2}\left(d^{m-1} + (d-2)^{m-1}\right) \\
&= d^{m-1} + d \cdot \frac{1}{2}\left(d^{m-1} + (d-2)^{m-1}\right) - \left(d^{m-1} + (d-2)^{m-1}\right) \\
&= d \cdot \frac{1}{2}\left(d^{m-1} + (d-2)^{m-1}\right) - (d-2)^{m-1} \\
&= \frac{1}{2}\left(\underbrace{d \cdot \left(d^{m-1} + (d-2)^{m-1}\right)}_{=dd^{m-1}+d(d-2)^{m-1}} - 2(d-2)^{m-1}\right) \\
&= \frac{1}{2}\left(\underbrace{dd^{m-1}}_{=d^m} + \underbrace{d(d-2)^{m-1} - 2(d-2)^{m-1}}_{=(d-2)(d-2)^{m-1}=(d-2)^m}\right) \\
&= \frac{1}{2}\left(d^m + (d-2)^m\right).
\end{aligned}$$

In other words, (343) holds for $n = m$. This completes the induction step. Thus, (343) is proved.]

Exercise 7.8.1 is answered by (343). □

Second solution of Exercise 7.8.1 (sketched). Here is a more elegant argument.

Having introduced the notion of "1-even" tuples, let us also introduce the opposite notion: An n-tuple $(x_1, x_2, \ldots, x_n) \in [d]^n$ will be called 1-*odd* if the number 1 occurs in it an odd number of times. Thus, any n-tuple $(x_1, x_2, \ldots, x_n) \in [d]^n$ is either 1-even or 1-odd (but not both at the same time).

Let e_n denote the # of 1-even n-tuples in $[d]^n$. Let o_n denote the # of 1-odd n-tuples in $[d]^n$. Thus, we must compute e_n.

We notice that $\left|[d]^n\right| = |[d]|^n = d^n$ (since $|[d]| = d$), and thus

$$\begin{aligned} d^n &= \left|[d]^n\right| = \left(\text{\# of all } n\text{-tuples in } [d]^n\right) \\ &= \underbrace{\left(\text{\# of 1-even } n\text{-tuples in } [d]^n\right)}_{\substack{=e_n \\ \text{(by the definition of } e_n)}} + \underbrace{\left(\text{\# of 1-odd } n\text{-tuples in } [d]^n\right)}_{\substack{=o_n \\ \text{(by the definition of } o_n)}} \\ &\qquad \left(\begin{array}{l}\text{by the sum rule, since any } n\text{-tuple } (x_1, x_2, \ldots, x_n) \in [d]^n \\ \text{is either 1-even or 1-odd (but not both at the same time)}\end{array}\right) \\ &= e_n + o_n. \end{aligned}$$

In other words,

$$e_n + o_n = d^n. \tag{344}$$

Now, we shall compute $e_n - o_n$. Indeed, for each $k \in [d]$, let us define a number

$$s(k) = \begin{cases} -1, & \text{if } k = 1; \\ 1, & \text{if } k > 1. \end{cases} \tag{345}$$

Thus, $s(1) = -1$ and $s(2) = s(3) = \cdots = s(d) = 1$. Why are we defining these strange numbers? We will soon see.

If $(x_1, x_2, \ldots, x_n) \in [d]^n$ is an n-tuple, then

$$s(x_1)\, s(x_2) \cdots s(x_n) = \begin{cases} 1, & \text{if } (x_1, x_2, \ldots, x_n) \text{ is 1-even;} \\ -1, & \text{if } (x_1, x_2, \ldots, x_n) \text{ is 1-odd.} \end{cases} \tag{346}$$

[*Proof of (346):* Let $(x_1, x_2, \ldots, x_n) \in [d]^n$ be an n-tuple. Each of the n factors $s(x_1), s(x_2), \ldots, s(x_n)$ in the product $s(x_1)\, s(x_2) \cdots s(x_n)$ is either -1 or 1 (according to (345)). Thus, the entire product $s(x_1)\, s(x_2) \cdots s(x_n)$ is a product of 1's and (-1)'s. Therefore, this product equals 1 if an even number of its factors are (-1)'s, and equals -1 if an odd number of its factors are (-1)'s. But the number of factors of the product $s(x_1)\, s(x_2) \cdots s(x_n)$ that are (-1)'s is precisely the number of occurrences of the number 1 in the n-tuple $(x_1, x_2, \ldots, x_n)$, since each such occurrence contributes a factor that is -1 (because (345) says that $s(x_i) = -1$ if and only if $x_i = 1$). Thus, this number of factors that are (-1)'s is even if $(x_1, x_2, \ldots, x_n)$ is 1-even, and is odd if $(x_1, x_2, \ldots, x_n)$ is 1-odd. Combining these observations, we obtain (346).]

Now, consider the sum

$$\begin{aligned} &\sum_{(x_1, x_2, \ldots, x_n) \in [d]^n} \underbrace{s(x_1)\, s(x_2) \cdots s(x_n)}_{= \begin{cases} 1, & \text{if } (x_1, x_2, \ldots, x_n) \text{ is 1-even;} \\ -1, & \text{if } (x_1, x_2, \ldots, x_n) \text{ is 1-odd} \end{cases}} \\ &= \sum_{(x_1, x_2, \ldots, x_n) \in [d]^n} \begin{cases} 1, & \text{if } (x_1, x_2, \ldots, x_n) \text{ is 1-even;} \\ -1, & \text{if } (x_1, x_2, \ldots, x_n) \text{ is 1-odd.} \end{cases} \end{aligned}$$

This sum has e_n many addends equal to 1 (since there are e_n many 1-even n-tuples $(x_1, x_2, \ldots, x_n) \in [d]^n$) and o_n many addends equal to -1 (since there are o_n many 1-odd n-tuples $(x_1, x_2, \ldots, x_n) \in [d]^n$). Hence, this sum simplifies to

$$e_n \cdot 1 + o_n \cdot (-1) = e_n - o_n.$$

Thus, we obtain

$$\sum_{(x_1, x_2, \ldots, x_n) \in [d]^n} s(x_1) s(x_2) \cdots s(x_n) = e_n - o_n.$$

Hence,

$$e_n - o_n = \sum_{(x_1, x_2, \ldots, x_n) \in [d]^n} s(x_1) s(x_2) \cdots s(x_n) = \sum_{(k_1, k_2, \ldots, k_n) \in [d]^n} s(k_1) s(k_2) \cdots s(k_n)$$

(here, we have renamed the summation index). On the other hand, Theorem 4.2.14 (applied to $m_i = d$ and $p_{i,k} = s(k)$) yields

$$\prod_{i=1}^{n} \sum_{k=1}^{d} s(k) = \sum_{(k_1, k_2, \ldots, k_n) \in [d]^n} \underbrace{\prod_{i=1}^{n} s(k_i)}_{= s(k_1) s(k_2) \cdots s(k_n)} = \sum_{(k_1, k_2, \ldots, k_n) \in [d]^n} s(k_1) s(k_2) \cdots s(k_n).$$

Comparing these two equalities, we obtain

$$e_n - o_n = \prod_{i=1}^{n} \sum_{k=1}^{d} s(k) = \left(\sum_{k=1}^{d} s(k) \right)^n.$$

In view of

$$\begin{aligned} \sum_{k=1}^{d} s(k) &= \underbrace{s(1)}_{=-1} + \underbrace{s(2)}_{=1} + \underbrace{s(3)}_{=1} + \cdots + \underbrace{s(d)}_{=1} \\ &= (-1) + \underbrace{1 + 1 + \cdots + 1}_{d-1 \text{ times}} = (-1) + (d-1) \cdot 1 = d - 2, \end{aligned}$$

this rewrites as

$$e_n - o_n = (d-2)^n. \tag{347}$$

Adding this equality to (344), we obtain

$$(e_n + o_n) + (e_n - o_n) = d^n + (d-2)^n.$$

In view of $(e_n + o_n) + (e_n - o_n) = 2e_n$, this rewrites as $2e_n = d^n + (d-2)^n$. Dividing this equality by 2, we obtain $e_n = \frac{1}{2}\left(d^n + (d-2)^n\right)$. Hence, we have solved Exercise 7.8.1 again (and, reassuringly, obtained the same answer as the first time). $\square$

The trick used in the second solution above is not just slick; it has further applications. Superficially, what we did was finding e_n by first computing $e_n + o_n$ and $e_n - o_n$ and then adding up and dividing by 2. But the "reason" why $e_n - o_n$ was so much easier to find than e_n alone is that $e_n - o_n$ was a sum (of 1's and (-1)'s) over **all** (not just the 1-even) n-tuples $(x_1, x_2, \ldots, x_n) \in [d]^n$, and the addends could be decomposed as products $s(x_1)\, s(x_2) \cdots s(x_n)$.

A more sophisticated use of this strategy can be used to solve the following exercise ([18f-hw4s, Exercise 7]):

Exercise 7.8.2. Let $n \in \mathbb{N}$ and $d \in \mathbb{N}$.

An n-tuple $(x_1, x_2, \ldots, x_n) \in [d]^n$ will be called *all-even* if each number occurs in it an even number of times (i.e., for each $j \in [d]$, the number of $i \in [n]$ satisfying $x_i = j$ is even). (For example, the 4-tuples $(1,5,1,5)$ and $(2,2,3,3)$ are all-even, while the 4-tuple $(4,1,1,1)$ is not.)

Prove that the # of all-even n-tuples in $[d]^n$ is

$$\frac{1}{2^d} \sum_{k=0}^{d} \binom{d}{k} (d-2k)^n .$$

We omit the solution for now (see [18f-hw4s, Exercise 7]). (TODO: Insert it.)

7.9. A bit of extremal combinatorics

So far in this chapter, we have been doing enumerative combinatorics, which is mostly about questions of the form "how many things exist with this or that property". *Extremal combinatorics*, on the other hand, is the science of questions like "how big or small can things get while satisfying this or that property?". Here are some examples:

- Given $n \in \mathbb{N}$, how many distinct subsets of $[n]$ can you pick so that every two subsets you have picked have a nonempty intersection? (We shall answer this below.)
- Given $n, k \in \mathbb{N}$, how many distinct k-element subsets of $[n]$ can you pick so that every two subsets you have picked have a nonempty intersection? (We shall answer this below. Note that the case $2k > n$ is trivial – make sure you see why!)
- Given $n \in \mathbb{N}$, how many distinct subsets of $[n]$ can you pick so that no subset you have picked is a subset of another? (We shall answer this below.)
- Given $n \in \mathbb{N}$, how many distinct n-tuples $(i_1, i_2, \ldots, i_n) \in \{0,1\}^n$ can you pick so that no two n-tuples you have picked differ in just a single entry? (Such a pick is known as a *single-error-detecting code*. But this makes the question sound harder than it is!)

- Given $n, k \in \mathbb{N}$, how many distinct numbers can you pick from the set $[n]$ so that any two distinct numbers a, b you have picked satisfy $|a - b| \geq k$? (Using the terminology of Exercise 3.7.1, this is just asking for the largest size of a k-lacunar subset of $\{1, 2, \ldots, n\}$.) It is easy to see that the answer is $\left\lfloor \dfrac{n+k-1}{k} \right\rfloor$. (Indeed, you can certainly pick $\left\lfloor \dfrac{n+k-1}{k} \right\rfloor$ distinct numbers with this property[227], and Exercise 3.7.1 shows that you cannot pick more[228].)

The attentive reader will notice that the pigeonhole principles also belong to extremal combinatorics, as they answer questions like "how many distinct elements can you pick from a given n-element set?" or "how few elements can you pick from a given n-element set so that every element has been picked at least once?". Our point here is not to dwell on fundamentals like this, but to put them into a new context.

I shall not give a proper introduction to extremal combinatorics, but I shall solve three of the above questions to give an idea of how the subject (or at least a small part of it) feels like. The reader can find a lot more in [Galvin17, Chapter 43 onward], [Jukna11], [Engel97] and other places.

Convention 7.9.1. The word "*collection*" is a synonym for "set". For reasons of clarity, we shall speak of "collections of sets" instead of "sets of sets". For example, $\{\{1,2\}, \{1,3\}, \{5\}\}$ is a collection of subsets of $[5]$.

7.9.1. Sperner's theorem

We begin with answering the question "Given $n \in \mathbb{N}$, how many distinct subsets of $[n]$ can you pick so that no subset you have picked is a subset of another?":

Definition 7.9.2. Let S be a set. A *set antichain* of S means a collection $\mathbf{J}$ of subsets of S such that no two distinct sets $A, B \in \mathbf{A}$ satisfy $A \subseteq B$.

Example 7.9.3. **(a)** The collection $\{\{1,2\}, \{1,3\}, \{5\}\}$ is a set antichain of $[5]$.

(b) The collection $\{\{1,2\}, \{1,3\}, \{2\}\}$ is not a set antichain of $[5]$, since $\{2\} \subseteq \{1,2\}$.

[227] Namely, you can pick the numbers

$$0k+1, 1k+1, 2k+1, \ldots, (q-1)k+1,$$

where $q = \left\lfloor \dfrac{n+k-1}{k} \right\rfloor$. An easy computation shows that they all belong to $[n]$ and have the property required.

[228] More precisely, Exercise 3.7.1 shows that you cannot pick more than $\dfrac{n+k-1}{k}$ distinct numbers with this property. But you can only pick an integer number of numbers; thus, "no more than $\dfrac{n+k-1}{k}$" is tantamount to "no more than $\left\lfloor \dfrac{n+k-1}{k} \right\rfloor$".

Theorem 7.9.4 (Sperner's theorem)**.** Let $n \in \mathbb{N}$. Let S be an n-element set. Then, the maximum possible size of a set antichain of S is $\binom{n}{\lfloor n/2 \rfloor}$.

Before we prove this theorem, let us do two things. First, as a warmup, we shall prove a similar (but much simpler) result about *set chains* – a dual notion to set antichains. Then, we shall show an inequality saying that $\binom{n}{\lfloor n/2 \rfloor}$ is the largest of all binomial coefficients $\binom{n}{k}$ (for a fixed $n \in \mathbb{N}$).

Definition 7.9.5. Let S be a set. A *set chain* of S means a collection $\mathbf{C}$ of subsets of S such that any two distinct sets $A, B \in \mathbf{C}$ satisfy $A \subseteq B$ or $B \subseteq A$.

Example 7.9.6. (a) The collection $\{\{1,2\}, \{1,2,5\}, \{1,2,3,4,5\}\}$ is a set chain of $[5]$.

(b) The collection $\{\{1,2\}, \{1,3\}, \{1,2,3\}\}$ is not a set chain of $[5]$, since $\{1,2\}$ and $\{1,3\}$ satisfy neither $\{1,2\} \subseteq \{1,3\}$ nor $\{1,3\} \subseteq \{1,2\}$.

Theorem 7.9.7. Let $n \in \mathbb{N}$. Let S be an n-element set. Then, the maximum possible size of a set chain of S is $n+1$.

Proof of Theorem 7.9.7 (sketched). Let $s_1, s_2, \ldots, s_n$ be the n elements of S. The collection

$$\begin{aligned}
&\{\{s_1, s_2, \ldots, s_n\} \mid i \in \{0, 1, \ldots, n\}\} \\
&= \{\{\}, \\
&\quad \{s_1\}, \\
&\quad \{s_1, s_2\}, \\
&\quad \{s_1, s_2, s_3\}, \\
&\quad \ldots, \\
&\quad \{s_1, s_2, \ldots, s_n\}\}
\end{aligned}$$

is clearly a set chain of S, and has size $n+1$. Thus, it remains to prove that any set chain of S has size $\leq n+1$.

Indeed, let $\mathbf{C}$ be a set chain of S. We must thus prove that $|\mathbf{C}| \leq n+1$.

Each set $I \in \mathbf{C}$ is a subset of S, and thus satisfies $|I| \in \{0, 1, \ldots, n\}$ (since S is an n-element set). Hence, we can define a map

$$\begin{aligned}
f : \mathbf{C} &\to \{0, 1, \ldots, n\}, \\
I &\mapsto |I|.
\end{aligned}$$

This map f simply sends each set $I \in \mathbf{C}$ to its size. Now, we claim that this map f is injective. Indeed, if A and B are two distinct elements of $\mathbf{C}$, then we have $A \subseteq B$

or $B \subseteq A$ (since $\mathbf{C}$ is a set chain), and therefore we must have $|A| \neq |B|$ (because if we had $|A| = |B|$, then either of the two relations $A \subseteq B$ and $B \subseteq A$ would imply that $A = B$ [229], which would contradict the fact that A and B are distinct); but this rewrites as $f(A) \neq f(B)$ (since the definition of f yields $f(A) = |A|$ and $f(B) = |B|$). Thus, the map f is injective. Hence, Theorem 6.1.2 (applied to $U = \mathbf{C}$ and $V = \{0, 1, \ldots, n\}$) yields $|\mathbf{C}| \leq |\{0, 1, \ldots, n\}| = n + 1$. This is precisely what we had to show. Thus, the proof of Theorem 7.9.7 is complete. □

Next, as we promised, comes an inequality for binomial coefficients:

Lemma 7.9.8. Let $n \in \mathbb{N}$ and $k \in \mathbb{R}$. Then, $\binom{n}{k} \leq \binom{n}{\lfloor n/2 \rfloor}$.

Proof of Lemma 7.9.8 (sketched). There are many proofs for this; here is an outline of what might be the simplest one.

Theorem 4.3.10 yields $\binom{n}{k} = \binom{n}{n-k}$. Thus, if $k > n - k$, then we can replace k by $n - k$ without changing $\binom{n}{k}$. Hence, for the rest of this proof, we WLOG assume that $k \leq n - k$. Hence, $2k \leq n$, so that $k \leq n/2$ and therefore $k \leq \lfloor n/2 \rfloor$ (since k is an integer).

If $k \notin \mathbb{N}$, then $\binom{n}{k} = 0 \leq \binom{n}{\lfloor n/2 \rfloor}$ (since $\binom{n}{\lfloor n/2 \rfloor} \geq 0$). Thus, if $k \notin \mathbb{N}$, then Lemma 7.9.8 holds. Hence, for the rest of this proof, we WLOG assume that $k \in \mathbb{N}$. Combining $k \in \mathbb{N}$ with $k \leq \lfloor n/2 \rfloor$, we find $k \in \{0, 1, \ldots, \lfloor n/2 \rfloor\}$.

Now, we claim the following chain of inequalities:

$$\binom{n}{0} < \binom{n}{1} < \binom{n}{2} < \cdots < \binom{n}{\lfloor n/2 \rfloor}. \tag{348}$$

(In terms of Pascal's triangle, this is simply saying that the binomial coefficients in the n-th row of Pascal's triangle strictly increase until the middle of the row. For example, for $n = 4$, this is saying that $1 < 4 < 6$.) Clearly, if we can prove (348), then $\binom{n}{k} \leq \binom{n}{\lfloor n/2 \rfloor}$ will follow, since $\binom{n}{k}$ is one of the binomial coefficients in the chain (348) (because $k \in \{0, 1, \ldots, \lfloor n/2 \rfloor\}$). Thus, we only need to prove (348).

In order to prove (348), we need to show that $\binom{n}{i} < \binom{n}{i+1}$ for each $i \in \{0, 1, \ldots, \lfloor n/2 \rfloor - 1\}$. Let us do this: Let $i \in \{0, 1, \ldots, \lfloor n/2 \rfloor - 1\}$. Thus, $i \geq 0$ and $i \leq \underbrace{\lfloor n/2 \rfloor}_{\leq n/2} - 1 \leq n/2 - 1$. Hence, $i + 1 \leq n/2$, so that $2(i+1) \leq n$ and thus $i + 1 \leq n - \underbrace{(i+1)}_{> i} < n - i$, so that $n - i > i + 1$. Also, $i \in \{0, 1, \ldots, n\}$ (since

[229] Here, we are using the fact that if X and Y are two finite sets satisfying $X \subseteq Y$ and $|X| = |Y|$, then $X = Y$.

$i \geq 0$ and $i \leq n/2 - 1 < n/2 \leq n$); hence, Theorem 4.3.8 (applied to $k = i$) yields $\binom{n}{i} = \dfrac{n!}{i!\,(n-i)!} > 0$ (since factorials are positive).

It is not hard to prove that

$$(i+1)\binom{n}{i+1} = (n-i)\binom{n}{i}. \tag{349}$$

(In fact, this is Lemma A.4.9 **(b)**, applied to $k = i$.) Thus,

$$(i+1)\binom{n}{i+1} = \underbrace{(n-i)}_{>i+1}\binom{n}{i} > (i+1)\binom{n}{i}$$

(here, we have used $\binom{n}{i} > 0$). We can divide this inequality by $i+1$ (since $\underbrace{i}_{\geq 0} + \underbrace{1}_{>0} > 0$), and thus obtain $\binom{n}{i+1} > \binom{n}{i}$. In other words, $\binom{n}{i} < \binom{n}{i+1}$. Thus, we have shown that $\binom{n}{i} < \binom{n}{i+1}$ for each $i \in \{0, 1, \ldots, \lfloor n/2 \rfloor - 1\}$. [230] As we explained above, this proves the chain (348) and thus completes the proof of Lemma 7.9.8. □

Proof of Theorem 7.9.4 (sketched). (We are following [Galvin17, second proof of Theorem 53.2] here.)

For each $k \in \{0, 1, \ldots, n\}$, the collection

$$\{\text{all } k\text{-element subsets of } S\}$$

is a set antichain of S (because a k-element subset of S cannot be a subset of a different k-element subset of S). This collection has size $\binom{n}{k}$ (by Theorem 4.3.12). Thus, for each $k \in \{0, 1, \ldots, n\}$, we have found a set antichain of S that has size $\binom{n}{k}$. In particular, taking $k = \lfloor n/2 \rfloor$, we thus have found a set antichain of S that has size $\binom{n}{\lfloor n/2 \rfloor}$.

It thus remains to show that any set antichain of S has size $\leq \binom{n}{\lfloor n/2 \rfloor}$.

Indeed, let $\mathbf{A}$ be a set antichain of S. We must thus show that $|\mathbf{A}| \leq \binom{n}{\lfloor n/2 \rfloor}$.

We WLOG assume that $S = [n]$ (as we can always relabel the n elements of S by $1, 2, \ldots, n$). Recall the set S_n from Definition 7.4.4. (Permutations appear nowhere

[230] Another proof of this can be found in [Guicha20, Theorem 1.3.4].

in Theorem 7.9.4, but as we will soon see, they are the key to its proof.) For each permutation $\sigma \in S_n$, we let $\mathbf{C}_\sigma$ be the collection

$$\begin{aligned}
&\{\{\sigma(1), \sigma(2), \ldots, \sigma(i)\} \mid i \in \{0, 1, \ldots, n\}\} \\
&= \{ \{\}, \\
&\quad \{\sigma(1)\}, \\
&\quad \{\sigma(1), \sigma(2)\}, \\
&\quad \{\sigma(1), \sigma(2), \sigma(3)\}, \\
&\quad \ldots, \\
&\quad \{\sigma(1), \sigma(2), \ldots, \sigma(n)\} \}.
\end{aligned}$$

This collection $\mathbf{C}_\sigma$ is a set chain of $[n] = S$, and has size $|\mathbf{C}_\sigma| = n + 1$.

We now set [231]

$$s := \sum_{I \in \mathbf{A}} \sum_{\sigma \in S_n} [I \in \mathbf{C}_\sigma].$$

We shall find an upper bound and a lower bound for s, and use them to obtain an inequality. (This is a variant of the double counting technique.)

To obtain an upper bound, we observe the following:

Claim 1: For any $\sigma \in S_n$, we have $\sum_{I \in \mathbf{A}} [I \in \mathbf{C}_\sigma] \leq 1$.

[*Proof of Claim 1:* Let $\sigma \in S_n$. Recall that $\mathbf{C}_\sigma$ is a set chain; thus, any two distinct sets $A, B \in \mathbf{C}_\sigma$ satisfy

$$A \subseteq B \text{ or } B \subseteq A \tag{350}$$

(by the definition of "set chain"). On the other hand, $\mathbf{A}$ is a set antichain; thus, no two distinct sets $A, B \in \mathbf{A}$ satisfy $A \subseteq B$. In other words, any two distinct sets $A, B \in \mathbf{A}$ satisfy

$$A \not\subseteq B. \tag{351}$$

Now, each addend of the sum $\sum_{I \in \mathbf{A}} [I \in \mathbf{C}_\sigma]$ is a truth value and thus equals either 0 or 1; moreover, the # of addends that equal 1 is precisely the # of all $I \in \mathbf{A}$ satisfying $I \in \mathbf{C}_\sigma$. Hence,

$$\begin{aligned}
\sum_{I \in \mathbf{A}} [I \in \mathbf{C}_\sigma] &= (\text{\# of all } I \in \mathbf{A} \text{ satisfying } I \in \mathbf{C}_\sigma) \cdot 1 + (\text{some number of 0's}) \\
&= (\text{\# of all } I \in \mathbf{A} \text{ satisfying } I \in \mathbf{C}_\sigma) \cdot 1 \\
&= (\text{\# of all } I \in \mathbf{A} \text{ satisfying } I \in \mathbf{C}_\sigma).
\end{aligned} \tag{352}$$

In order to prove Claim 1, we need to show that $\sum_{I \in \mathbf{A}} [I \in \mathbf{C}_\sigma] \leq 1$. In view of (352), this boils down to showing that

$$(\text{\# of all } I \in \mathbf{A} \text{ satisfying } I \in \mathbf{C}_\sigma) \leq 1. \tag{353}$$

[231] We use the Iverson bracket notation (Definition 4.3.19).

In other words, we need to show that there is **at most one** set $I \in \mathbf{A}$ satisfying $I \in \mathbf{C}_\sigma$.

Assume the contrary. Thus, there exist two such sets I. In other words, there exist two **distinct** sets $A, B \in \mathbf{A}$ satisfying $A, B \in \mathbf{C}_\sigma$. Consider these A, B. From (350), we obtain $A \subseteq B$ or $B \subseteq A$. However, from (351), we obtain $A \not\subseteq B$. Hence, $B \subseteq A$ (since $A \subseteq B$ or $B \subseteq A$). However, we can also apply (351) to B and A instead of A and B. Thus, we obtain $B \not\subseteq A$. This contradicts $B \subseteq A$. This contradiction shows that our assumption was wrong, so we are done proving that there is **at most one** set $I \in \mathbf{A}$ satisfying $I \in \mathbf{C}_\sigma$. Hence, we have proved (353). As explained above, this yields Claim 1.]

Now,

$$\begin{aligned} s &= \underbrace{\sum_{I \in \mathbf{A}} \sum_{\sigma \in S_n}}_{= \sum_{\sigma \in S_n} \sum_{I \in \mathbf{A}}} [I \in \mathbf{C}_\sigma] = \sum_{\sigma \in S_n} \underbrace{\sum_{I \in \mathbf{A}} [I \in \mathbf{C}_\sigma]}_{\substack{\leq 1 \\ \text{(by Claim 1)}}} \\ &\leq \sum_{\sigma \in S_n} 1 = |S_n| \cdot 1 = |S_n| = (\text{\# of permutations of } [n]) \\ &= n! \end{aligned} \tag{354}$$

(by Theorem 7.4.1, applied to $X = [n]$). Thus we have obtained an upper bound for s.

In order to find a lower bound, we will need the following:

Claim 2: For any $I \in \mathbf{A}$, we have $\sum_{\sigma \in S_n} [I \in \mathbf{C}_\sigma] \geq \dfrac{n!}{\binom{n}{\lfloor n/2 \rfloor}}$.

[*Proof of Claim 2:* Let $I \in \mathbf{A}$. Let $k = |I|$; thus, $k = |I| \in \{0, 1, \ldots, n\}$ (since I is a subset of the n-element set S).

Now, each addend of the sum $\sum_{\sigma \in S_n} [I \in \mathbf{C}_\sigma]$ is a truth value and thus equals either 0 or 1; moreover, the # of addends that equal 1 is precisely the # of all $\sigma \in S_n$ satisfying $I \in \mathbf{C}_\sigma$. Hence,

$$\begin{aligned} \sum_{\sigma \in S_n} [I \in \mathbf{C}_\sigma] &= (\text{\# of all } \sigma \in S_n \text{ satisfying } I \in \mathbf{C}_\sigma) \cdot 1 + (\text{some number of 0's}) \\ &= (\text{\# of all } \sigma \in S_n \text{ satisfying } I \in \mathbf{C}_\sigma) \cdot 1 \\ &= (\text{\# of all } \sigma \in S_n \text{ satisfying } I \in \mathbf{C}_\sigma). \end{aligned} \tag{355}$$

Now, how many permutations $\sigma \in S_n$ satisfy $I \in \mathbf{C}_\sigma$?

Let $\sigma \in S_n$ be a permutation. We know that I is a k-element set (since $k = |I|$); but the only k-element set in the set chain $\mathbf{C}_\sigma$ is $\{\sigma(1), \sigma(2), \ldots, \sigma(k)\}$ (by the definition of $\mathbf{C}_\sigma$). Hence, we have $I \in \mathbf{C}_\sigma$ if and only if $I = \{\sigma(1), \sigma(2), \ldots, \sigma(k)\}$.

Forget that we fixed σ. We thus have shown that a permutation $\sigma \in S_n$ satisfies $I \in \mathbf{C}_\sigma$ if and only if $I = \{\sigma(1), \sigma(2), \ldots, \sigma(k)\}$. Hence,

$$\begin{aligned}&(\#\text{ of all } \sigma \in S_n \text{ satisfying } I \in \mathbf{C}_\sigma)\\ &= (\#\text{ of all } \sigma \in S_n \text{ satisfying } I = \{\sigma(1), \sigma(2), \ldots, \sigma(k)\}).\end{aligned} \tag{356}$$

But a permutation $\sigma \in S_n$ satisfying $I = \{\sigma(1), \sigma(2), \ldots, \sigma(k)\}$ can be constructed using the following decision procedure:

- First, we choose the value $\sigma(1)$ (provided that $1 \leq k$). There are k options for it, since $\sigma(1)$ needs to belong to the k-element set I (because we want $I = \{\sigma(1), \sigma(2), \ldots, \sigma(k)\}$).
- Next, we choose the value $\sigma(2)$ (provided that $2 \leq k$). There are $k-1$ options for it, since $\sigma(2)$ needs to belong to the k-element set I (because we want $I = \{\sigma(1), \sigma(2), \ldots, \sigma(k)\}$) and be distinct from $\sigma(1)$.
- Next, we choose the value $\sigma(3)$ (provided that $3 \leq k$). There are $k-2$ options for it, since $\sigma(3)$ needs to belong to the k-element set I (because we want $I = \{\sigma(1), \sigma(2), \ldots, \sigma(k)\}$) and be distinct from $\sigma(1)$ and $\sigma(2)$.
- And so on, until the first k values $\sigma(1), \sigma(2), \ldots, \sigma(k)$ of σ have been chosen. At this point, we have assigned k distinct elements of the set I as values $\sigma(1), \sigma(2), \ldots, \sigma(k)$. Thus, by the pigeonhole principle for injections, each of the k elements of I has been assigned as one of the k values $\sigma(1), \sigma(2), \ldots, \sigma(k)$ (since I is a k-element set). We have therefore ensured that $I = \{\sigma(1), \sigma(2), \ldots, \sigma(k)\}$. We still need to choose the remaining values $\sigma(k+1), \sigma(k+2), \ldots, \sigma(n)$ of σ, though.
- Next, we choose the value $\sigma(k+1)$. There are $n-k$ options for it, since $\sigma(k+1)$ needs to belong to the n-element set $[n]$ and be distinct from the k numbers $\sigma(1), \sigma(2), \ldots, \sigma(k)$.
- Next, we choose the value $\sigma(k+2)$. There are $n-k-1$ options for it, since $\sigma(k+2)$ needs to belong to the n-element set $[n]$ and be distinct from the $k+1$ numbers $\sigma(1), \sigma(2), \ldots, \sigma(k+1)$.
- And so on, until all n values $\sigma(1), \sigma(2), \ldots, \sigma(n)$ have been chosen.

According to the dependent product rule, the total # of possibilities for making these choices is

$$\underbrace{k(k-1)(k-2)\cdots 1}_{\substack{=1\cdot 2\cdot\cdots\cdot k\\ =k!}} \cdot \underbrace{(n-k)(n-k-1)\cdots 1}_{\substack{=1\cdot 2\cdot\cdots\cdot(n-k)\\ =(n-k)!}} = k! \cdot (n-k)! = \frac{n!}{\binom{n}{k}}$$

(since Theorem 4.3.8 yields $\binom{n}{k} = \frac{n!}{k! \cdot (n-k)!}$). Hence,

$$\begin{aligned}
&(\text{\# of all } \sigma \in S_n \text{ satisfying } I = \{\sigma(1), \sigma(2), \ldots, \sigma(k)\}) \\
&= \frac{n!}{\binom{n}{k}}. \qquad (357)
\end{aligned}$$

Now, (355) becomes

$$\begin{aligned}
\sum_{\sigma \in S_n} [I \in \mathbf{C}_\sigma] &= (\text{\# of all } \sigma \in S_n \text{ satisfying } I \in \mathbf{C}_\sigma) \\
&= (\text{\# of all } \sigma \in S_n \text{ satisfying } I = \{\sigma(1), \sigma(2), \ldots, \sigma(k)\}) \\
&\qquad (\text{by (356)}) \\
&= \frac{n!}{\binom{n}{k}} \qquad (\text{by (357)}) \\
&\geq \frac{n!}{\binom{n}{\lfloor n/2 \rfloor}} \qquad \left(\text{since Lemma 7.9.8 yields } \binom{n}{k} \leq \binom{n}{\lfloor n/2 \rfloor}\right).
\end{aligned}$$

This proves Claim 2.]

Now,

$$s = \sum_{I \in \mathbf{A}} \underbrace{\sum_{\sigma \in S_n} [I \in \mathbf{C}_\sigma]}_{\substack{\geq \frac{n!}{\binom{n}{\lfloor n/2 \rfloor}} \\ (\text{by Claim 2})}} \geq \sum_{I \in \mathbf{A}} \frac{n!}{\binom{n}{\lfloor n/2 \rfloor}} = |\mathbf{A}| \cdot \frac{n!}{\binom{n}{\lfloor n/2 \rfloor}}. \qquad (358)$$

Thus we have obtained a lower bound for s. From (358), we obtain

$$|\mathbf{A}| \cdot \frac{n!}{\binom{n}{\lfloor n/2 \rfloor}} \leq s \leq n! \qquad (\text{by (354)}).$$

Dividing this inequality by $n!$ and multiplying it by $\binom{n}{\lfloor n/2 \rfloor}$, we obtain $|\mathbf{A}| \leq \binom{n}{\lfloor n/2 \rfloor}$ (since $n!$ and $\binom{n}{\lfloor n/2 \rfloor}$ are nonnegative). This is precisely what we needed to show. Thus, Theorem 7.9.4 is proven. $\square$

Remark 7.9.9. Theorem 7.9.4 (applied to $n = 5$ and $S = [5]$) yields that the maximum possible size of a set antichain of $[5]$ is $\binom{5}{\lfloor 5/2 \rfloor} = \binom{5}{2} = 10$. As we have seen in the proof, an example of a set antichain of $[5]$ having this maximum size is

$$\begin{aligned}&\{\text{all 2-element subsets of } [5]\}\\ &= \{\{1,2\}, \ \{1,3\}, \ \{1,4\}, \ \{1,5\}, \ \{2,3\}, \ \{2,4\}, \\ &\qquad \{2,5\}, \ \{3,4\}, \ \{3,5\}, \ \{4,5\}\}.\end{aligned}$$

Another example of such an antichain is $\{\text{all 3-element subsets of } [5]\}$. It can be shown that these two examples are the only set antichains of $[5]$ having maximum size. More generally: If $n \in \mathbb{N}$, and if S is an n-element set, then the only set antichains of S having maximum size (that is, size $\binom{n}{\lfloor n/2 \rfloor}$) are

$$\begin{aligned}&\{\text{all } \lfloor n/2 \rfloor \text{-element subsets of } S\} \qquad \text{and}\\ &\{\text{all } (n - \lfloor n/2 \rfloor) \text{-element subsets of } S\}.\end{aligned}$$

(When n is even, these two antichains are identical.) See [Engel97, Theorem 1.1.1 (b)] for a proof of this.

7.9.2. Intersecting collections

Next, let us answer the question "Given $n \in \mathbb{N}$, how many distinct subsets of $[n]$ can you pick so that every two subsets you have picked have a nonempty intersection?". Such a pick will be called an *intersecting collection*:

Definition 7.9.10. Let S be a set. An *intersecting collection* of S means a collection $\mathbf{J}$ of subsets of S such that any $A, B \in \mathbf{J}$ satisfy $A \cap B \neq \varnothing$.

Example 7.9.11. (a) The collection $\{\{1,2\}, \{1,3\}, \{2,3\}\}$ is an intersecting collection of $[4]$.

(b) The collection $\{\{1,2\}, \{1,3\}, \{2\}\}$ is not an intersecting collection of $[4]$, since $\{1,3\} \cap \{2\} = \varnothing$.

Now we can answer our question ([Galvin17, Theorem 44.2]):

Theorem 7.9.12. Let n be a positive integer. Let S be an n-element set. Then, the maximum possible size of an intersecting collection of S is 2^{n-1}.

Proof of Theorem 7.9.12 (sketched). The set S is nonempty (since it is an n-element set, but n is positive). Thus, there exists some $t \in S$. Pick such a t.

The collection

$$\{\text{all subsets of } S \text{ that contain } t\}$$

is an intersecting collection of S (indeed, any two sets A, B in this collection have at least the element t in common and therefore satisfy $A \cap B \neq \varnothing$). This collection has size 2^{n-1} (check this!). Thus, we have found an intersecting collection of S that has size 2^{n-1}.

It now remains to show that any intersecting collection of S has size $\leq 2^{n-1}$.

So let $\mathbf{J}$ be an intersecting collection of S. We must show that $|\mathbf{J}| \leq 2^{n-1}$.

If T is any subset of S, then the complement $S \setminus T$ of T will be denoted by $\overline{T}$. This complement $\overline{T}$ is disjoint from T (this is a general property of complements) and also distinct from T (since S is nonempty).

If T is any subset of S, then we shall refer to the collection $\left\{T, \overline{T}\right\}$ (which consists of just two sets: T and its complement $\overline{T}$) as the *pigeonhole* of T. Note that a subset T of S and its complement $\overline{T}$ have the same pigeonhole, since

$$\left(\text{the pigeonhole of } \overline{T}\right) = \left\{\overline{T}, \underbrace{\overline{\overline{T}}}_{=T}\right\} = \left\{\overline{T}, T\right\} = \left\{T, \overline{T}\right\} = \left(\text{the pigeonhole of } T\right).$$

There are 2^n subsets of S in total, and each of them lies in a unique pigeonhole (namely, each subset T lies in the pigeonhole of T). Since each pigeonhole contains precisely two subsets of S (this is where we are using our observation that the complement $\overline{T}$ of a set T is distinct from T), we thus conclude that there are precisely 2^{n-1} pigeonholes.

Now, the crucial observation is the following: The two sets in any given pigeonhole are disjoint (because they are complements of one another). Hence, $\mathbf{J}$ cannot contain more than one set from any given pigeonhole (since $\mathbf{J}$ is an intersecting collection, so that no two sets in $\mathbf{J}$ are disjoint). Thus, by the pigeonhole principle for injections, the size of $\mathbf{J}$ is $\leq$ to the # of pigeonholes. In other words, $|\mathbf{J}| \leq 2^{n-1}$ (since there are precisely 2^{n-1} pigeonholes). This is precisely what we needed to show. Thus, Theorem 7.9.12 is proved. $\square$

Remark 7.9.13. The maximum size (2^{n-1}) in Theorem 7.9.12 is achieved not just for collections of the form $\{\text{all subsets of } S \text{ that contain } t\}$ for some $t \in S$. For example, if $n = 3$ and $S = [3]$, then the collection

$$\{\{1,2\}, \ \{1,3\}, \ \{2,3\}, \ \{1,2,3\}\}$$

also is an intersecting collection of S having maximum size (that is, in this case, $2^{n-1} = 2^{3-1} = 4$). This illustrates that the sets in an intersecting collection need not all have an element in common.

The last extremal combinatorics question we shall answer is "Given $n, k \in \mathbb{N}$, how many distinct k-element subsets of $[n]$ can you pick so that every two subsets you have picked have a nonempty intersection?". In our language, this is asking for the maximum size of an intersecting collection that consists entirely of k-element sets. We are lazy and give these collections a new name:

Definition 7.9.14. Let S be a set. Let $k \in \mathbb{N}$. An *intersecting k-collection* of S means a collection $\mathbf{J}$ of k-element subsets of S such that any $A, B \in \mathbf{J}$ satisfy $A \cap B \neq \varnothing$.

Example 7.9.15. The collection $\{\{1,2\},\{1,3\},\{2,3\}\}$ is an intersecting 2-collection of $[4]$.

Theorem 7.9.16 (Erdös–Ko–Rado theorem)**.** Let n and k be positive integers. Let S be an n-element set. Then:

(a) If $n \geq 2k$, then the maximum possible size of an intersecting k-collection of S is $\binom{n-1}{k-1}$.

(b) If $n < 2k$, then the maximum possible size of an intersecting k-collection of S is $\binom{n}{k}$.

It may appear strange that the answers in the cases $n \geq 2k$ and $n < 2k$ are so different, but the proof will reveal the reason for this.

Proof of Theorem 7.9.16 (sketched). We start with part **(b)**, which is more of a joke than a theorem.

(b) Assume that $n < 2k$. Then, **any** two k-element subsets A, B of S satisfy $A \cap B \neq \varnothing$ (because otherwise, A and B would be disjoint, and thus the sum rule would yield $|A \cup B| = \underbrace{|A|}_{=k} + \underbrace{|B|}_{=k} = k + k = 2k$, which would contradict

$$\begin{aligned} |A \cup B| &\leq |S| \qquad (\text{since } A \cup B \subseteq S) \\ &= n < 2k \end{aligned}$$

). Therefore, the collection $\{\text{all } k\text{-element subsets of } S\}$ is an intersecting k-collection of S. The size of this collection is $\binom{n}{k}$, and clearly there is no intersecting k-collection of S having higher size (since there are only $\binom{n}{k}$ many k-element subsets of S). Thus, the maximum possible size of an intersecting k-collection of S is $\binom{n}{k}$. This proves Theorem 7.9.16 **(b)**.

(a) Assume that $n \geq 2k$. Thus, $n \geq 2k > k$ (since k is positive), so that $k < n$.

The set S is nonempty (since it is an n-element set, but n is positive). Thus, there exists some $t \in S$. Pick such a t.

The collection

$$\{\text{all } k\text{-element subsets of } S \text{ that contain } t\}$$

is an intersecting k-collection of S (check this!), and has size $\binom{n-1}{k-1}$ (check this!). Thus, we have found an intersecting k-collection of S that has size $\binom{n-1}{k-1}$.

It now remains to show that any intersecting k-collection of S has size $\leq \binom{n-1}{k-1}$.

So let $\mathbf{J}$ be an intersecting k-collection of S. We must show that $|\mathbf{J}| \leq \binom{n-1}{k-1}$.

The following elegant argument (which I have taken from [Galvin17, first proof of Theorem 49.1] with only expositional changes) is known as "Katona's cycle proof", and has seen uses in various settings (see [BorMea16] and [Frankl20] for two surveys). Again (as already in the proof of Sperner's theorem above) we involve permutations for no obvious reason.

We WLOG assume that $S = [n]$ (since we can otherwise just relabel the elements of S). Recall that S_n denotes the set of all permutations of $[n]$. We recall that $|S_n| = n!$ (which we have seen, e.g., in the proof of Theorem 7.9.4).

We extend any permutation $\sigma \in S_n$ of $[n]$ to an n-periodic map $\sigma : \{1, 2, 3, \ldots\} \to [n]$. That is, if $\sigma \in S_n$ is a permutation, then we recursively define the extra values $\sigma(n+1), \sigma(n+2), \sigma(n+3), \ldots$ by setting

$$\begin{aligned} \sigma(n+1) &:= \sigma(1), \\ \sigma(n+2) &:= \sigma(2), \\ \sigma(n+3) &:= \sigma(3), \\ &\ldots. \end{aligned}$$

(Thus, explicitly, if i is a positive integer, then $\sigma(i) = \sigma(j)$, where j is the unique element of $[n]$ satisfying $i \equiv j \bmod n$.) Thus, for each permutation $\sigma \in S_n$, the sequence $(\sigma(1), \sigma(2), \sigma(3), \ldots)$ begins with the n "original" values of σ and then continues by repeating these n values over and over. Hence, this sequence $(\sigma(1), \sigma(2), \sigma(3), \ldots)$ is n-periodic. Let us also notice that if $\sigma \in S_n$ is a permutation, and if $p \in \{0, 1, \ldots, n\}$, then any p consecutive elements of this sequence $(\sigma(1), \sigma(2), \sigma(3), \ldots)$ (that is, any p elements of the form $\sigma(i+1), \sigma(i+2), \ldots, \sigma(i+p)$, where $i \in \mathbb{N}$) are distinct.

Fix a permutation $\sigma \in S_n$. For each $i \in \mathbb{N}$, we define a subset

$$W_i(\sigma) = \{\sigma(i+1), \sigma(i+2), \ldots, \sigma(i+k)\}$$

of $[n]$. Thus,

$$\begin{aligned} W_0(\sigma) &= \{\sigma(1), \sigma(2), \ldots, \sigma(k)\}, \\ W_1(\sigma) &= \{\sigma(2), \sigma(3), \ldots, \sigma(k+1)\}, \\ W_2(\sigma) &= \{\sigma(3), \sigma(4), \ldots, \sigma(k+2)\}, \\ &\ldots. \end{aligned}$$

These sets $W_0(\sigma), W_1(\sigma), W_2(\sigma), \ldots$ (that is, the sets $W_i(\sigma)$ for all $i \in \mathbb{N}$) will be called the *windows* of σ. [232] It is easy to see that these windows $W_i(\sigma)$ are k-element subsets of $[n]$ (indeed, for any $i \in \mathbb{N}$, the k elements

[232] The following way of visualizing the windows of σ might be useful: Imagine the values

$\sigma(i+1), \sigma(i+2), \ldots, \sigma(i+k)$ of $W_i(\sigma)$ are distinct, since σ is a permutation). Moreover, the first n windows

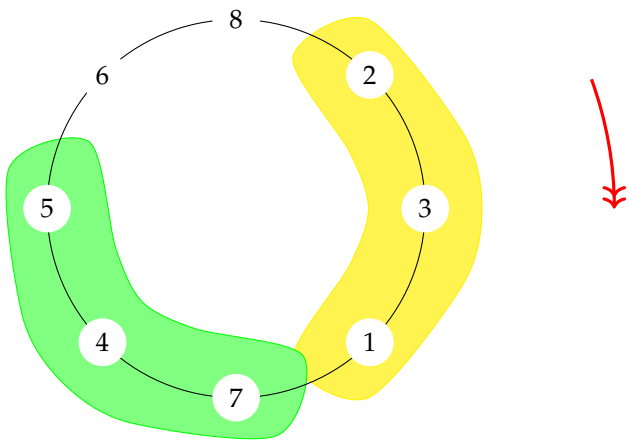
$$\begin{aligned} W_0(\sigma) &= \{\sigma(1), \sigma(2), \ldots, \sigma(k)\}, \\ W_1(\sigma) &= \{\sigma(2), \sigma(3), \ldots, \sigma(k+1)\}, \\ &\ldots, \\ W_{n-1}(\sigma) &= \{\sigma(n), \sigma(n+1), \ldots, \sigma(n+k-1)\} \end{aligned}$$

of σ are distinct[233], whereas all other windows of σ are repetitions of these n windows[234]. Hence, there are n windows of σ in total. We let $\mathbf{B}_\sigma$ denote the collection

$\sigma(1), \sigma(2), \ldots, \sigma(n)$ written on the circumference on a circle (in clockwise order); then, a window of σ is just a set of k values appearing consecutively on the circle. For example, if $n = 8$ and $k = 3$, and if $\sigma \in S_8$ is the permutation that sends $1,2,3,4,5,6,7,8$ to $3,1,7,4,5,6,8,2$, respectively, then the windows of σ are

$$\begin{aligned} W_0(\sigma) &= \{3,1,7\}, \\ W_1(\sigma) &= \{1,7,4\}, \\ W_2(\sigma) &= \{7,4,5\}, \\ W_3(\sigma) &= \{4,5,6\}, \\ W_4(\sigma) &= \{5,6,8\}, \\ W_5(\sigma) &= \{6,8,2\}, \\ W_6(\sigma) &= \{8,2,3\}, \\ W_7(\sigma) &= \{2,3,1\}. \end{aligned}$$

Two of these windows (namely, $W_2(\sigma)$ in green, and $W_7(\sigma)$ in yellow) are shown on the following picture:



[233]Check this! (Here, we need to recall that $0 < k < n$ and that σ is injective.)
[234]That is,

$$\begin{aligned} W_n(\sigma) &= W_0(\sigma), \\ W_{n+1}(\sigma) &= W_1(\sigma), \\ W_{n+2}(\sigma) &= W_2(\sigma), \\ &\ldots. \end{aligned}$$

$\{W_0(\sigma), W_1(\sigma), W_2(\sigma), \ldots\}$ of all windows of σ. Thus, $|\mathbf{B}_\sigma| = n$ (since there are n windows of σ in total).

Forget that we fixed σ. Thus, for each permutation $\sigma \in S_n$, we have defined a collection

$$\mathbf{B}_\sigma = \{W_0(\sigma), W_1(\sigma), W_2(\sigma), \ldots\} \tag{359}$$

of "windows of σ".

We now set [235]

$$s := \sum_{I \in \mathbf{J}} \sum_{\sigma \in S_n} [I \in \mathbf{B}_\sigma].$$

We shall find an upper bound and a lower bound for s, and use them to obtain an inequality. (If you are feeling a deja-vu here: yes, we are mimicking the above proof of Theorem 7.9.4.)

To obtain an upper bound, we observe the following:

Claim 1: For any $\sigma \in S_n$, we have $\sum_{I \in \mathbf{J}} [I \in \mathbf{B}_\sigma] \leq k$.

[*Proof of Claim 1:* Let $\sigma \in S_n$. Then,

$$\sum_{I \in \mathbf{J}} [I \in \mathbf{B}_\sigma] = (\text{\# of all } I \in \mathbf{J} \text{ satisfying } I \in \mathbf{B}_\sigma). \tag{360}$$

(Indeed, this can be proved in the same way as we proved (352) above, except that we are now using the collection $\mathbf{B}_\sigma$ instead of $\mathbf{C}_\sigma$.) Thus,

$$\begin{aligned} \sum_{I \in \mathbf{J}} [I \in \mathbf{B}_\sigma] &= (\text{\# of all } I \in \mathbf{J} \text{ satisfying } I \in \mathbf{B}_\sigma) \\ &= |\mathbf{J} \cap \mathbf{B}_\sigma|. \end{aligned} \tag{361}$$

In order to prove Claim 1, we need to show that $\sum_{I \in \mathbf{J}} [I \in \mathbf{B}_\sigma] \leq k$. In view of (361), this boils down to showing that

$$|\mathbf{J} \cap \mathbf{B}_\sigma| \leq k. \tag{362}$$

If $|\mathbf{J} \cap \mathbf{B}_\sigma| = 0$, then this is obvious (since $0 \leq k$). Hence, we WLOG assume that we don't have $|\mathbf{J} \cap \mathbf{B}_\sigma| = 0$. Thus, the set $\mathbf{J} \cap \mathbf{B}_\sigma$ is nonempty. Hence, there exists some $W \in \mathbf{J} \cap \mathbf{B}_\sigma$. Consider this W. Note that $W \in \mathbf{J} \cap \mathbf{B}_\sigma \subseteq \mathbf{B}_\sigma = \{\text{all windows of } \sigma\}$; in other words, W is a window of σ. Hence, W is a k-element set (since any window of σ is a k-element set). In other words, $|W| = k$, so that $W \neq \varnothing$ (since k is positive).

We shall say that two sets A and B *intersect* if $A \cap B \neq \varnothing$. We shall also use the verb "intersect" transitively – i.e., we shall say "A intersects B" (or "B intersects A") for "A and B intersect".

This follows from the fact that the sequence $(\sigma(1), \sigma(2), \sigma(3), \ldots)$ is n-periodic.

[235] We use the Iverson bracket notation (Definition 4.3.19).

If $A \in \mathbf{J}$ and $B \in \mathbf{J}$ are two distinct sets, then $A \cap B \neq \varnothing$ (since $\mathbf{J}$ is an intersecting k-collection). In other words, if $A \in \mathbf{J}$ and $B \in \mathbf{J}$ are two distinct sets, then A and B intersect. In other words, any two distinct sets belonging to $\mathbf{J}$ intersect. Thus, it is easy to see that

$$\mathbf{J} \cap \mathbf{B}_\sigma \subseteq \{\text{windows of } \sigma \text{ that intersect } W\}. \tag{363}$$

[*Proof of (363):* We must show that $W' \in \{\text{windows of } \sigma \text{ that intersect } W\}$ for each $W' \in \mathbf{J} \cap \mathbf{B}_\sigma$. So let $W' \in \mathbf{J} \cap \mathbf{B}_\sigma$. We must show that $W' \in \{\text{windows of } \sigma \text{ that intersect } W\}$. In other words, we must show that W' is a window of σ that intersects W. It is clear that W' is a window of σ (since $W' \in \mathbf{J} \cap \mathbf{B}_\sigma \subseteq \mathbf{B}_\sigma = \{\text{all windows of } \sigma\}$); hence, it only remains to show that W' intersects W. If $W' = W$, then this is obvious (because in this case, we have $W' \cap W = W \neq \varnothing$). Thus, we WLOG assume that $W' \neq W$. We have $W \in \mathbf{J} \cap \mathbf{B}_\sigma \subseteq \mathbf{J}$ and $W' \in \mathbf{J} \cap \mathbf{B}_\sigma \subseteq \mathbf{J}$ and $W' \neq W$; thus, W and W' are two distinct sets belonging to $\mathbf{J}$. Hence, W and W' intersect (since any two distinct sets belonging to $\mathbf{J}$ intersect). In other words, W' intersects W. This completes our proof of (363).]

So we have proved (363). In order to prove (362), it thus looks reasonable to ask: How many windows of σ intersect W ?

The answer is $2k-1$ (which is not hard to check, but too large to be useful), but it pays off to be more precise.

We have $W \in \mathbf{B}_\sigma = \{W_0(\sigma), W_1(\sigma), W_2(\sigma), \ldots\}$ (by the definition of $\mathbf{B}_\sigma$). In other words, there exists some $i \in \mathbb{N}$ satisfying $W = W_i(\sigma)$. Consider such an i. We note that if we replace i by $i+n$, then $W_i(\sigma)$ does not change[236]. Thus, we can WLOG assume that $i \geq n$ (since otherwise, we can replace i by $i+n$). We have

$$W = W_i(\sigma) = \{\sigma(i+1), \sigma(i+2), \ldots, \sigma(i+k)\}$$

(by the definition of $W_i(\sigma)$). Now, we define the $k-1$ windows

$$L_1 = W_{i-k+1}(\sigma), \quad L_2 = W_{i-k+2}(\sigma), \quad \ldots, \quad L_{k-1} = W_{i-1}(\sigma)$$

(that is, $L_j = W_{i-k+j}(\sigma)$ for each $j \in \{1, 2, \ldots, k-1\}$) as well as the $k-1$ windows

$$R_1 = W_{i+1}(\sigma), \quad R_2 = W_{i+2}(\sigma), \quad \ldots, \quad R_{k-1} = W_{i+k-1}(\sigma)$$

[236] because the definition of $W_{i+n}(\sigma)$ yields

$$\begin{aligned} W_{i+n}(\sigma) &= \{\sigma(i+n+1), \sigma(i+n+2), \ldots, \sigma(i+n+k)\} \\ &= \{\sigma(i+1+n), \sigma(i+2+n), \ldots, \sigma(i+k+n)\} \\ &= \{\sigma(i+1), \sigma(i+2), \ldots, \sigma(i+k)\} \\ &\qquad \left(\begin{array}{c}\text{since the sequence } (\sigma(1), \sigma(2), \sigma(3), \ldots) \text{ is } n\text{-periodic,} \\ \text{and thus } \sigma(u+n) = \sigma(u) \text{ for each } u \geq 1\end{array}\right) \\ &= W_i(\sigma) \end{aligned}$$

(that is, $R_j = W_{i+j}(\sigma)$ for each $j \in \{1, 2, \ldots, k-1\}$). [237] It is easy to see that the only windows of σ that intersect W are the $2(k-1)+1$ windows

$$L_1, L_2, \ldots, L_{k-1}, W, R_1, R_2, \ldots, R_{k-1}.$$

In other words,

$$\{\text{windows of } \sigma \text{ that intersect } W\} \subseteq \{L_1, L_2, \ldots, L_{k-1}, W, R_1, R_2, \ldots, R_{k-1}\}.$$

Now, (363) becomes

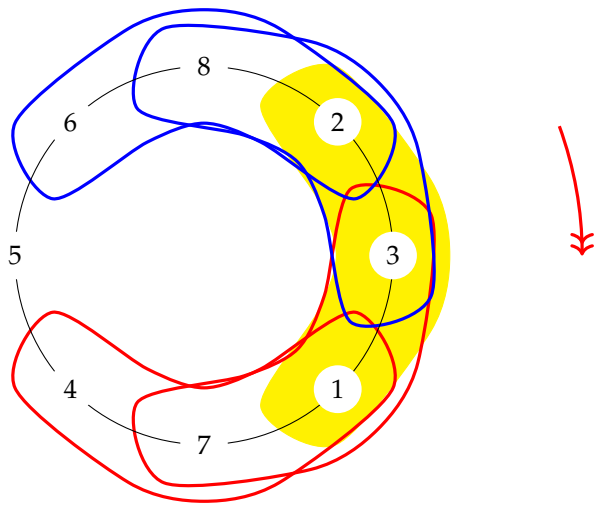
$$\begin{aligned} \mathbf{J} \cap \mathbf{B}_\sigma &\subseteq \{\text{windows of } \sigma \text{ that intersect } W\} \\ &\subseteq \{L_1, L_2, \ldots, L_{k-1}, W, R_1, R_2, \ldots, R_{k-1}\}. \end{aligned} \tag{364}$$

Now, in order to profit from this observation, let us see how many of the $2(k-1)+1$ windows $L_1, L_2, \ldots, L_{k-1}, W, R_1, R_2, \ldots, R_{k-1}$ can actually belong to $\mathbf{J} \cap \mathbf{B}_\sigma$.

If $j \in \{1, 2, \ldots, k-1\}$ is arbitrary, then the windows L_j and R_j are distinct and do not intersect[238], and thus cannot both belong to $\mathbf{J}$ at the same time (since any two distinct sets belonging to $\mathbf{J}$ intersect), and thus cannot both belong to $\mathbf{J} \cap \mathbf{B}_\sigma$ at

[237] Let us visualize these $2(k-1)$ windows on an example:

Let $n = 8$ and $k = 3$, and let $\sigma \in S_8$ be the permutation that sends $1, 2, 3, 4, 5, 6, 7, 8$ to $3, 1, 7, 4, 5, 6, 8, 2$, respectively. Let $W = W_7(\sigma) = \{2, 3, 1\}$. Then, the following picture shows the 2 windows L_1 and L_2 (in blue) and the 2 windows R_1 and R_2 (in red) as transparent blobs:



(The numbers on the circumference of the circle are the values of σ written in clockwise order. The yellow blob is the window W.) Note that the letter "L" has been chosen because the windows L_1 and L_2 are "left" (in the sense of "counterclockwise from", on the circle) of the window W; likewise the letter "R" stands for "right". (Furthermore, the colors blue and red have been chosen to hint at the letters L and R.)

[238] *Proof.* Let $j \in \{1, 2, \ldots, k-1\}$. The definitions of L_j and R_j yield

$$\begin{aligned} L_j &= W_{i-k+j}(\sigma) = \{\sigma(i-k+j+1), \sigma(i-k+j+2), \ldots, \sigma(i+j)\} \qquad \text{and} \\ R_j &= W_{i+j}(\sigma) = \{\sigma(i+j+1), \sigma(i+j+2), \ldots, \sigma(i+j+k)\}. \end{aligned}$$

the same time (since $\mathbf{J} \cap \mathbf{B}_\sigma \subseteq \mathbf{J}$). In other words, if $j \in \{1, 2, \ldots, k-1\}$ is arbitrary, then the collection $\mathbf{J} \cap \mathbf{B}_\sigma$ contains **at most one** of the two windows L_j and R_j. In other words:

- The collection $\mathbf{J} \cap \mathbf{B}_\sigma$ contains **at most one** of the two windows L_1 and R_1.
- The collection $\mathbf{J} \cap \mathbf{B}_\sigma$ contains **at most one** of the two windows L_2 and R_2.
- ...
- The collection $\mathbf{J} \cap \mathbf{B}_\sigma$ contains **at most one** of the two windows L_{k-1} and R_{k-1}.

Altogether, this shows that the collection $\mathbf{J} \cap \mathbf{B}_\sigma$ contains **at most** $k-1$ of the $2(k-1)$ windows $L_1, R_1, L_2, R_2, \ldots, L_{k-1}, R_{k-1}$. Since we further know that the collection $\mathbf{J} \cap \mathbf{B}_\sigma$ contains the window W, we thus conclude that the collection $\mathbf{J} \cap \mathbf{B}_\sigma$ contains **at most** $(k-1)+1$ of the $2(k-1)+1$ windows $L_1, R_1, L_2, R_2, \ldots, L_{k-1}, R_{k-1}, W$. In other words,

$$|(\mathbf{J} \cap \mathbf{B}_\sigma) \cap \{L_1, R_1, L_2, R_2, \ldots, L_{k-1}, R_{k-1}, W\}| \leq (k-1)+1 = k.$$

In view of

$$\begin{aligned}
&(\mathbf{J} \cap \mathbf{B}_\sigma) \cap \underbrace{\{L_1, R_1, L_2, R_2, \ldots, L_{k-1}, R_{k-1}, W\}}_{=\{L_1, L_2, \ldots, L_{k-1}, W, R_1, R_2, \ldots, R_{k-1}\}} \\
&= (\mathbf{J} \cap \mathbf{B}_\sigma) \cap \{L_1, L_2, \ldots, L_{k-1}, W, R_1, R_2, \ldots, R_{k-1}\} \\
&= \mathbf{J} \cap \mathbf{B}_\sigma \qquad \text{(by (364))},
\end{aligned}$$

this rewrites as $|\mathbf{J} \cap \mathbf{B}_\sigma| \leq k$. This proves (362). As we explained above, this proves Claim 1.]

Now,

$$\begin{aligned}
s &= \underbrace{\sum_{I \in \mathbf{J}} \sum_{\sigma \in S_n}}_{= \sum_{\sigma \in S_n} \sum_{I \in \mathbf{J}}} [I \in \mathbf{B}_\sigma] = \sum_{\sigma \in S_n} \underbrace{\sum_{I \in \mathbf{J}} [I \in \mathbf{B}_\sigma]}_{\substack{\leq k \\ \text{(by Claim 1)}}} \\
&\leq \sum_{\sigma \in S_n} k = \underbrace{|S_n|}_{=n!} \cdot k = n! \cdot k. \qquad (365)
\end{aligned}$$

Thus we have obtained an upper bound for s.

In order to find a lower bound (better yet: actually a precise value) for s, we will need the following:

But the $2k$ "consecutive" values

$$\sigma(i-k+j+1), \sigma(i-k+j+2), \ldots, \sigma(i+j+k)$$

of σ are distinct (since $n \geq 2k$); thus, the above equalities show that the sets L_j and R_j are disjoint. In other words, the sets L_j and R_j do not intersect. Qed.

Claim 2: For any $I \in \mathbf{J}$, we have $\sum_{\sigma \in S_n} [I \in \mathbf{B}_\sigma] = n \cdot k! \cdot (n-k)!$.

[*Proof of Claim 2:* Let $I \in \mathbf{J}$. Thus, I is a k-element subset of S (since $\mathbf{J}$ is a collection of k-element subsets of S). In other words, I is a k-element subset of $[n]$ (since $S = [n]$). We have

$$\sum_{\sigma \in S_n} [I \in \mathbf{B}_\sigma] = (\text{\# of all } \sigma \in S_n \text{ satisfying } I \in \mathbf{B}_\sigma). \tag{366}$$

(Indeed, this can be proved in the same way as we proved (355) above, except that we are now using the collection $\mathbf{B}_\sigma$ instead of $\mathbf{C}_\sigma$.) Now, how many permutations $\sigma \in S_n$ satisfy $I \in \mathbf{B}_\sigma$?

Let $\sigma \in S_n$ be a permutation. Then, $\mathbf{B}_\sigma$ is the collection of all windows of σ; thus,

$$\mathbf{B}_\sigma = \{W_0(\sigma), W_1(\sigma), \ldots, W_{n-1}(\sigma)\}$$

(since there are n windows of σ in total, and these n windows are precisely $W_0(\sigma), W_1(\sigma), \ldots, W_{n-1}(\sigma)$). Thus, we have $I \in \mathbf{B}_\sigma$ if and only if we have $I = W_i(\sigma)$ for some $i \in \{0, 1, \ldots, n-1\}$.

Forget that we fixed σ. We thus have shown that a permutation $\sigma \in S_n$ satisfies $I \in \mathbf{B}_\sigma$ if and only if it satisfies $I = W_i(\sigma)$ for some $i \in \{0, 1, \ldots, n-1\}$. Hence,

$$\begin{aligned}
&(\text{\# of all } \sigma \in S_n \text{ satisfying } I \in \mathbf{B}_\sigma) \\
&= (\text{\# of all } \sigma \in S_n \text{ satisfying } I = W_i(\sigma) \text{ for some } i \in \{0, 1, \ldots, n-1\}) \\
&= \sum_{i=0}^{n-1} (\text{\# of all } \sigma \in S_n \text{ satisfying } I = W_i(\sigma)).
\end{aligned} \tag{367}$$

(Here, the last equality sign is a consequence of the sum rule, because the sets $\{\sigma \in S_n \mid I = W_i(\sigma)\}$ for $i \in \{0, 1, \ldots, n-1\}$ are disjoint[239].)

Now, let us fix $i \in \{0, 1, \ldots, n-1\}$. Then, a permutation $\sigma \in S_n$ satisfying $I = W_i(\sigma)$ can be constructed using the following decision procedure:

- First, we choose the k values $\sigma(i+1), \sigma(i+2), \ldots, \sigma(i+k)$ of σ (in this order). These k values must belong to the k-element subset I of $[n]$ (since we want to have $I = W_i(\sigma) = \{\sigma(i+1), \sigma(i+2), \ldots, \sigma(i+k)\}$) and be distinct (since σ needs to be injective). Thus, we have k options for the first of these values, $k-1$ options for the second, $k-2$ for the third, and so on.

[239] *Proof.* Assume the contrary. Thus, these sets are not disjoint. In other words, there exists a permutation $\sigma \in S_n$ that satisfies $I = W_i(\sigma)$ for two different values of $i \in \{0, 1, \ldots, n-1\}$. Consider this σ. We have seen above that the first n windows $W_0(\sigma), W_1(\sigma), \ldots, W_{n-1}(\sigma)$ of σ are distinct. However, two of these n windows equal I (since σ satisfies $I = W_i(\sigma)$ for two different values of $i \in \{0, 1, \ldots, n-1\}$). Therefore, two of these n windows are equal. This contradicts the fact that these n windows are distinct. This contradiction shows that our assumption was false, qed.

Once these choices are made, we have assigned k distinct elements of the set I as values $\sigma(i+1), \sigma(i+2), \ldots, \sigma(i+k)$. Thus, by the pigeonhole principle for injections, each of the k elements of I has been assigned as one of the k values $\sigma(i+1), \sigma(i+2), \ldots, \sigma(i+k)$ (since I is a k-element set). We have therefore ensured that $I = \{\sigma(i+1), \sigma(i+2), \ldots, \sigma(i+k)\} = W_i(\sigma)$. We still need to choose the remaining $n-k$ values of σ, though.

- Next, we choose the remaining $n-k$ values $\sigma(i+k+1), \sigma(i+k+2), \ldots, \sigma(i+n)$ of σ (in this order). These $n-k$ values must be elements of $[n]$ that are distinct from each other and from the already chosen k values; thus, we have $n-k$ options for the first of them, $n-k-1$ options for the second, $n-k-2$ options for the third, and so on.

According to the dependent product rule, the total # of possibilities for making these choices is

$$\underbrace{k(k-1)(k-2)\cdots 1}_{\substack{=1\cdot 2\cdot\cdots\cdot k \\ =k!}} \cdot \underbrace{(n-k)(n-k-1)\cdots 1}_{\substack{=1\cdot 2\cdot\cdots\cdot(n-k) \\ =(n-k)!}} = k!\cdot(n-k)!.$$

Hence,

$$\begin{aligned}&(\#\text{ of all } \sigma \in S_n \text{ satisfying } I = W_i(\sigma)) \\ &= k!\cdot(n-k)!. \end{aligned} \tag{368}$$

Now, forget that we fixed i. We thus have proved (368) for each $i \in \{0, 1, \ldots, n-1\}$. Now, (366) becomes

$$\begin{aligned}\sum_{\sigma\in S_n} [I \in \mathbf{B}_\sigma] &= (\#\text{ of all } \sigma \in S_n \text{ satisfying } I \in \mathbf{B}_\sigma) \\ &= \sum_{i=0}^{n-1} \underbrace{(\#\text{ of all } \sigma \in S_n \text{ satisfying } I = W_i(\sigma))}_{\substack{=k!\cdot(n-k)! \\ \text{(by (368))}}} \qquad \text{(by (367))} \\ &= \sum_{i=0}^{n-1} k!\cdot(n-k)! = n\cdot k!\cdot(n-k)!.\end{aligned}$$

This proves Claim 2.]

Now,

$$s = \sum_{I\in\mathbf{J}} \underbrace{\sum_{\sigma\in S_n} [I \in \mathbf{B}_\sigma]}_{\substack{=n\cdot k!\cdot(n-k)! \\ \text{(by Claim 2)}}} = \sum_{I\in\mathbf{J}} n\cdot k!\cdot(n-k)! = |\mathbf{J}|\cdot n\cdot k!\cdot(n-k)!.$$

Hence,

$$|\mathbf{J}|\cdot n\cdot k!\cdot(n-k)! = s \leq n!\cdot k \qquad \text{(by (365))}.$$

Dividing this inequality by $n \cdot k! \cdot (n-k)!$, we obtain

$$|\mathbf{J}| \leq \frac{n! \cdot k}{n \cdot k! \cdot (n-k)!} = \frac{k}{n} \cdot \underbrace{\frac{n!}{k! \cdot (n-k)!}}_{\substack{= \binom{n}{k} \\ \text{(by Theorem 4.3.8)}}} = \frac{k}{n}\binom{n}{k} = \binom{n-1}{k-1}$$

(by Exercise 4.5.4 **(a)**, applied to $m = k$). This proves Theorem 7.9.16 **(a)**. $\square$

8. Invariants and Monovariants

Mathematicians frequently study "dynamical" objects, i.e., objects that change with time: games, algorithms, recursively defined sequences[240]. We shall refer to such things as *processes*. A process is called *deterministic* if its state at a given time is uniquely determined by its states at the previous times; otherwise it is called *nondeterministic*. For example, computing the Fibonacci numbers $f_0, f_1, f_2, \ldots$ one by one using their recursive definition is a deterministic process, whereas casting a die multiple times and summing the scores is a nondeterministic process. We shall restrict our attention to processes with precisely defined *states* (which the object can have at any given time) and precisely defined *moves* (i.e., ways in which the object can change from one moment to the other; i.e., allowable transitions between states). For example, let us formalize "the process of computing the Fibonacci numbers $f_0, f_1, f_2, \ldots$ one by one using their recursive definition" as a precise process:

- Its states are finite tuples $(a_1, a_2, \ldots, a_k)$ of integers with at least 2 entries (i.e., with $k \geq 2$).
- Its moves are to insert a new entry $a_k + a_{k-1}$ at the end of a given tuple $(a_1, a_2, \ldots, a_k)$.

Thus, if we start this process with the state $(0, 1)$, then we will arrive at the state $(0, 1, 1)$ after one move, then at the state $(0, 1, 1, 2)$ after another move, then at the state $(0, 1, 1, 2, 3)$ after a third move, and so on. After n moves, our state will be the $(n+2)$-tuple $(f_0, f_1, \ldots, f_{n+1})$. Note that the state will never be the entire Fibonacci sequence $(f_0, f_1, f_2, \ldots)$, since a state is always a **finite** tuple; but each Fibonacci number will eventually appear in the state. Of course, the same process (with a different starting state) can be used to compute not just the Fibonacci sequence, but any $(1, 1)$-recurrent sequence.

Given a process, we can ask several questions about it:

- Can a certain state ever be reached by the process from a given starting state?
- What states can be reached at all?
- Does the process exhibit periodic behavior (i.e., do its states form a periodic sequence)?
- Is the process reversible? That is, is there a way to reconstruct a state from the next state?
- What things stay unchanged throughout the process?
- What things change only in one direction (i.e., only get smaller or only get larger) throughout the process?

[240] or their continuous analogues: functions defined as solutions to differential equations

The last two questions in this list are sufficiently important that they have names. Things that stay unchanged during a process are called *invariants* (or *conserved quantities*) of this process; things that change only in one direction (e.g., only get smaller, or only get larger) are called *monovariants*. In many cases, finding invariants or monovariants (or both) is not just interesting in itself, but also helps answering some of the other questions (for example, knowing that some quantity decreases at each step of the process immediately guarantees that the process does not return to any previous state). Here are some examples that you may have seen:

- The Euclidean algorithm for computing the gcd of two nonnegative integers rests on the fact that if we subtract an integer from another (i.e., replace a pair (a, b) of integers by $(a - b, b)$ or by $(a, b - a)$), then the gcd of the two integers does not change. In other words, the gcd is an invariant of the subtraction operation that keeps getting used in the algorithm.[241] More generally, almost all nontrivial algorithms in mathematics are analyzed with the help of *loop invariants*, which are things (e.g., quantities or statements) that never change when the algorithm is performed.

- Proposition 2.2.4 can be restated as follows: If a positive integer can only change by 1 (upwards or downwards) but is not allowed to ever become 0, then it always remains positive. Thus, its sign is an invariant.

- In the Euclidean algorithm, the sum of the two nonnegative integers decreases at each step. Thus, it is a monovariant. (This is why the Euclidean algorithm terminates.)

- In Exercise 1.1.5, lemmings can fall off the cliff but never come back. Thus, the number of lemmings on the cliff is a monovariant: it can only decrease.

- The monotone convergence theorem says that a sequence of real numbers that is increasing and bounded from above converges to its supremum. The "increasing" condition here can be viewed as a monovariant (viz., the entries of the sequence increase as we move along the sequence), whereas the "bounded" condition can be viewed as an invariant (viz., the entries remain $\leq$ to the bound, no matter where we stand in the sequence). Thus, the monotone convergence theorem is one way in which monovariants and invariants can be useful.

- Our solutions to Exercise 3.7.10 illustrate various uses of invariants and monovariants. (For example, in the first solution, the comb number is a monovariant, whereas the number of 1s in the bitstring is an invariant. In the second solution, load a is a monovariant, and ones a is an invariant.)

[241] There are actually two versions of the Euclidean algorithm: In one version, the smaller integer keeps getting subtracted from the larger integer; in the other version, the larger integer keeps getting divided with remainder by the smaller integer. The gcd is an invariant for both of these versions.

Physics has its share of invariants and monovariants, too (e.g., energy and momentum are famous invariants, while entropy is a monovariant).

There is no general theory of how to construct invariants and monovariants, let alone find specific ones that are useful for a given problem. Doing so often requires ingenuity, experience and luck. The best we can do is show some examples and discuss some common threads. More can be found in [Engel98, Chapter 1], [GelAnd17, §1.5], [Grinbe08, Kapitel 4], [Carl17, Kapitel 5] and [Zeitz17, §3.4].[242]

8.1. Invariants

We begin by illustrating the use of invariants on some examples. See [LeLeMe16, §6.1] for a brief but rigorous introduction.

8.1.1. Simple examples

We begin with a classical simple exercise ([Grinbe08, Aufgabe 4.1]):

Exercise 8.1.1. A chunk of ice is floating in the sea. At each moment, a chunk can break into 3 or into 5 smaller chunks. Assuming that no chunks can ever melt, is it possible that there are precisely 100 chunks left after some time has passed?

Solution to Exercise 8.1.1 (sketched). No.

Proof. At the beginning, there is only 1 chunk of ice; thus, the total number of chunks is odd. Each time a chunk breaks apart, the number of chunks increases by 2 or by 4 (since the chunk breaks into 3 or into 5 smaller chunks); thus, the total number of chunks remains odd (because increasing an odd number by 2 or by 4 preserves its oddness). Thus, the total number of chunk can never become 100. This proves our negative answer. □

The next exercise is [Grinbe08, Aufgabe 4.2]:

Exercise 8.1.2. Fix a positive integer n.

(a) You have the n numbers $1, 2, \ldots, n$ written on a blackboard. In one move, you can erase two numbers a and b and write the number $a + b$ instead.

You keep making these moves until only one number is left on the blackboard. Prove that this will happen after precisely $n - 1$ moves. Will the remaining number depend on the specific moves you have taken? If not, what is this number?

(b) Answer the same questions if $a + b$ is replaced by $a - b$.

(c) Answer the same questions if $a + b$ is replaced by ab.

(d) Answer the same questions if $a + b$ is replaced by $a + b + 5$.

(e) Answer the same questions if $a + b$ is replaced by $ab + a + b$.

(f) Answer the same questions if $a + b$ is replaced by $a^2 + b^2$.

[242] Many of the following examples are taken from [Engel98, Chapter 1], [Grinbe08, Kapitel 4] and from the Russian problem database `http://www.problems.ru/` .

Solution to Exercise 8.1.2 (sketched). **(a)** Every move decreases the number of numbers written on the table by 1. Since this number was n at the onset, it is thus clear that it will take precisely $n-1$ moves to reduce it down to 1. Thus, after precisely $n-1$ moves, only one number will be left on the blackboard.

Now, we claim that this number is $1+2+\cdots+n$. In fact, each of our moves leaves the sum of all numbers on the blackboard unchanged (because it removes two addends a and b and inserts the addend $a+b$ instead). Since this sum was $1+2+\cdots+n$ at the onset, it will therefore remain $1+2+\cdots+n$ throughout the procedure, and thus will still be $1+2+\cdots+n$ at the end. But since there is only one number left on the blackboard at the end, this means that this number will have to be $1+2+\cdots+n$. Thus, in particular, this number will not depend on the specific moves taken. Thus, Exercise 8.1.2 **(a)** is solved.

(b) Again, we can show (as in our above solution to Exercise 8.1.2 **(a)**) that after precisely $n-1$ moves, only one number will be left on the blackboard. However, this time, this number may well depend on the specific moves taken. This does in fact happen already for $n=2$, since the two numbers 1 and 2 can be replaced by any of $1-2=-1$ and $2-1=+1$ (depending on which of them we take to be a and which we take to be b). Thus, Exercise 8.1.2 **(b)** is solved.

(c) This can be solved in the exact same way as Exercise 8.1.2 **(a)**, except that all sums need to be replaced by products. Thus, the number that remains on the blackboard after $n-1$ moves are made is not $1+2+\cdots+n$ but $1\cdot 2\cdot\cdots\cdot n=n!$. (But it still does not depend on the specific moves taken.) Thus, Exercise 8.1.2 **(c)** is solved.

(d) Again, we can show (as in our above solution to Exercise 8.1.2 **(a)**) that after precisely $n-1$ moves, only one number will be left on the blackboard. Again, this remaining number does not depend on the specific moves taken; but this time, this is a bit trickier to prove. Namely: We define the *weird-sum* of k numbers $a_1, a_2, \ldots, a_k$ to be the number $(a_1+5)+(a_2+5)+\cdots+(a_k+5)$. Then, each of our moves leaves the weird-sum of the numbers on the blackboard unchanged (indeed, the move removes two addends of the form $a+5$ and $b+5$ from this weird-sum, and replaces them with the addend $(a+b+5)+5$; but this new addend clearly has the same effect as the two addends that were removed, because $(a+5)+(b+5)=a+b+10=(a+b+5)+5$). Since this weird-sum was

$$(1+5)+(2+5)+\cdots+(n+5)=(1+2+\cdots+n)+5n$$

at the onset, it will therefore remain $(1+2+\cdots+n)+5n$ throughout the procedure, and thus will still be $(1+2+\cdots+n)+5n$ at the end. But since there is only one number left on the blackboard at the end, this means that this number will have to be $(1+2+\cdots+n)+5n-5$ (because the weird-sum of a single number a is $a+5$). Thus, in particular, this number will not depend on the specific moves taken. Thus, Exercise 8.1.2 **(d)** is solved.

(e) Again, we can show (as in our above solution to Exercise 8.1.2 **(a)**) that after precisely $n-1$ moves, only one number will be left on the blackboard. Again, this remaining number does not depend on the specific moves taken; but this time,

this is a bit trickier to prove. Namely: We define the *weird-product* of k numbers $a_1, a_2, \ldots, a_k$ to be the number $(a_1+1)(a_2+1)\cdots(a_k+1)$. Then, each of our moves leaves the weird-product of the numbers on the blackboard unchanged (indeed, the move removes two factors of the form $a+1$ and $b+1$ from this weird-product, and replaces them with the factor $(ab+a+b)+1$; but this new factor clearly has the same effect as the two factors that were removed, because $(a+1)(b+1) = ab+a+b+1 = (ab+a+b)+1$). Since this weird-product was

$$\begin{aligned}(1+1)(2+1)\cdots(n+1) &= 2\cdot 3\cdot\cdots\cdot(n+1) = \frac{1\cdot 2\cdot\cdots\cdot(n+1)}{1} \\ &= 1\cdot 2\cdot\cdots\cdot(n+1) = (n+1)!\end{aligned}$$

at the onset, it will therefore remain $(n+1)!$ throughout the procedure, and thus will still be $(n+1)!$ at the end. But since there is only one number left on the blackboard at the end, this means that this number will have to be $(n+1)!-1$ (because the weird-product of a single number a is $a+1$). Thus, in particular, this number will not depend on the specific moves taken. Thus, Exercise 8.1.2 **(e)** is solved.

(f) Again, we can show (as in our above solution to Exercise 8.1.2 **(a)**) that after precisely $n-1$ moves, only one number will be left on the blackboard. However, this time, this number may well depend on the specific moves taken. For example, if $n=3$, then we can get either $\left(1^2+2^2\right)^2+3^2 = 34$ or $1^2+\left(2^2+3^2\right)^2 = 170$ depending on what move we do first. Thus, Exercise 8.1.2 **(f)** is solved. □

Parity-related invariants are particularly common in contest problems. Here is an example:

Exercise 8.1.3. The numbers $1, 2, \ldots, 100$ are written in a row (in this order, from left to right). In a move, you can swap any two numbers at a distance of 2 (i.e., any two numbers that have exactly one number written between them). Can you end up with the numbers $100, 99, \ldots, 1$ (in this order, from left to right) by a sequence of such moves?

[**Example:** The analogous question for 5 instead of 100 has a positive answer:

$$(1,2,3,4,5) \xrightarrow{\substack{\text{swap first and}\\ \text{third numbers}}} (3,2,1,4,5) \xrightarrow{\substack{\text{swap third and}\\ \text{fifth numbers}}} (3,2,5,4,1)$$

$$\xrightarrow{\substack{\text{swap second and}\\ \text{fourth numbers}}} (3,4,5,2,1) \xrightarrow{\substack{\text{swap first and}\\ \text{third numbers}}} (5,4,3,2,1).$$

]

Solution to Exercise 8.1.3 (sketched). No, you cannot.

Proof. We define the *position* of a number $i \in [100]$ to be the (unique) $j \in [100]$ with the property that i is the j-th number from the left on our row. Thus, at the beginning, each number $i \in [100]$ stands in position i. Hence, in particular, the number 1 is in an odd position (namely, in position 1) at the beginning.

Of course, the positions of our numbers can change when you make moves. Namely, when you make a move, each number moves precisely 2 positions to the right or 2 positions to the left or not at all.[243] Hence, when you make a move, the position of each number does not change its parity (i.e., it stays even if it was even, and it stays odd if it was odd). Thus, in particular, the position of 1 does not change its parity under any move. Thus, after any sequence of moves, the number 1 will still be in an odd position (since 1 was in an odd position at the beginning). Hence, in particular, the number 1 will never be in position 100 (since 100 is not odd). Therefore, you will never end up with the numbers $100, 99, \ldots, 1$ (in this order, from left to right), because this would entail the number 1 being in position 100. This concludes our proof. □

8.1.2. More examples

Here are some trickier problems that can be solved with invariants. The following exercise ([Grinbe08, Aufgabe 4.6], somewhat generalized) is a Russian olympiad classic:

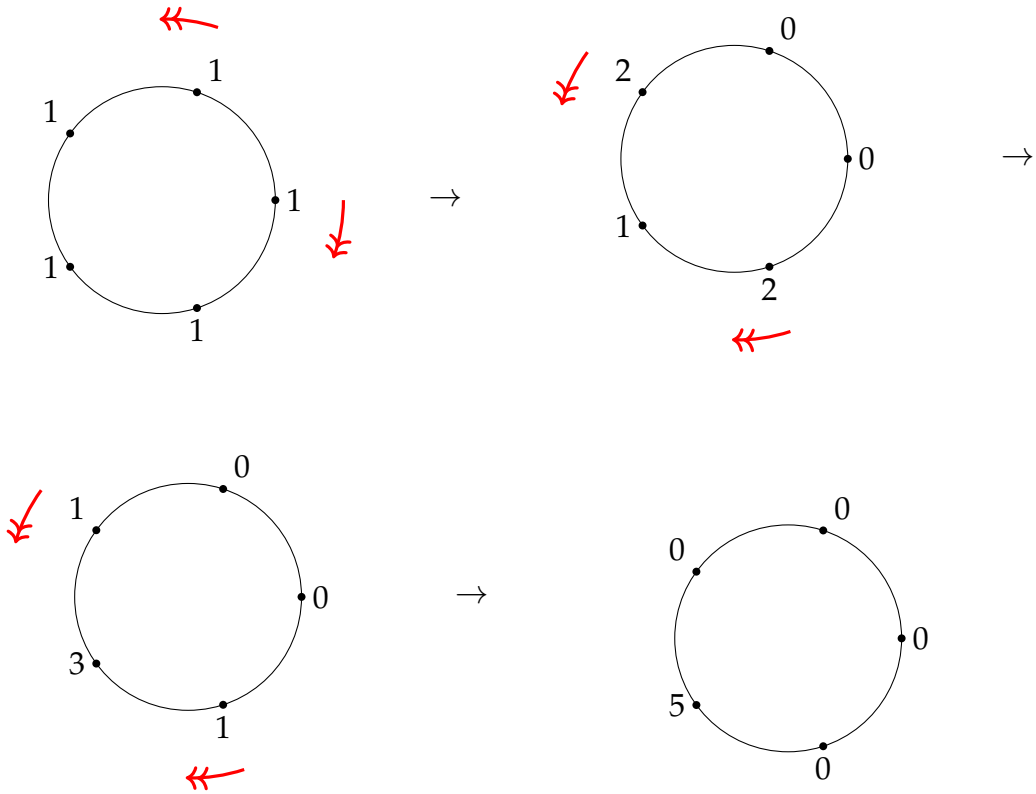
Exercise 8.1.4. Let $n \geq 2$ be an integer. Consider n trees arranged in a circle. Initially, there is one sparrow sitting on each tree. Every minute, two of the n sparrows move: Namely, one sparrow moves to the next tree in clockwise order, whereas the other sparrow moves to the next tree in counterclockwise order. (The two moves are simultaneous. Any tree can fit an arbitrary number of sparrows, including 0.)

Is it possible that, after some time, all sparrows end up on the same tree? Answer this question depending on n.

[**Example:** For $n = 5$, the answer is "yes". Indeed, here is a way how the

[243] For example, the move that transforms $(1, 2, 3, 4, 5)$ into $(3, 2, 1, 4, 5)$ causes the number 1 to move precisely 2 positions to the right and causes the number 3 to move precisely 2 positions to the left, and leaves all remaining numbers unmoved.

sparrows can end up on the same tree after 3 minutes:



(where the five points are the five trees, and the number at each point is counting the sparrows on that tree; furthermore, the red arrows are showing where the sparrows will move in the next step).]

Solution to Exercise 8.1.4 (sketched). Yes if n is odd; no if n is even.

Proof. The pair of simultaneous moves that happens every minute (i.e., one sparrow moving one tree clockwise, and another sparrow moving one tree counterclockwise) will be called a *bimove*. Thus, we must prove the following two claims:

> *Claim 1:* Assume that n is odd. Then, there is a sequence of bimoves that results in all sparrows sitting on the same tree.
>
> *Claim 2:* Assume that n is even. Then, there is no sequence of bimoves that results in all sparrows sitting on the same tree.

[*Proof of Claim 1:* The following is a generalization of the movements shown in the example in the statement of the exercise.

We write n in the form $n = 2m + 1$ for some $m \in \mathbb{N}$. (Such an m can be found, since n is an odd positive integer.) Thus, we have $2m + 1$ trees. We label these trees with the numbers $-m, -m + 1, \ldots, m - 1, m$ in clockwise order (starting at some randomly chosen tree). Initially, each tree has one sparrow sitting in it.

Now, the sparrows make the following bimoves:

- First, the sparrow on tree $-m$ and the sparrow on tree m move to the trees $-m+1$ and $m-1$, respectively[244]. Now, trees $-m$ and m are empty.
- Now, we have two sparrows on tree $-m+1$ and two sparrows on tree $m-1$. These four sparrows move to the trees $-m+2$ and $m-2$ (first one pair, then the other)[245]. Now, trees $-m+1$ and $m-1$ are empty.
- Now, we have three sparrows on tree $-m+2$ and three sparrows on tree $m-2$. These six sparrows move to the trees $-m+3$ and $m-3$ (first one pair, then another, then the last remaining pair). Now, trees $-m+2$ and $m-2$ are empty.
- Now, we have four sparrows on tree $-m+3$ and four sparrows on tree $m-3$. These eight sparrows move to the trees $-m+4$ and $m-4$ (in pairs of two). Now, trees $-m+3$ and $m-3$ are empty.
- And so on, until all trees other than tree 0 are empty. At this point, all sparrows have arrived on tree 0.

Thus, Claim 1 is proven.]

[*Proof of Claim 2:* Let us label the n trees with the numbers $1, 2, \ldots, n$ in clockwise order (starting at some randomly chosen tree). A placement of sparrows on trees (with each sparrow being on one tree) will be called a *state*. Thus, our initial state has one sparrow on each of the trees $1, 2, \ldots, n$.

If S is a state, then we define the *charge* $c(S)$ of S to be the sum

$$\sum_{s \text{ is a sparrow}} t_S(s),$$

where $t_S(s)$ is the number of the tree on which the sparrow s sits in the state S. Thus, the charge of the initial state is $1+2+\cdots+n$ (since the sparrows in the initial state sit on the trees $1, 2, \ldots, n$).

Clearly, a bimove transforms a state into another state. Now, the crucial observation is the following: When a state is transformed by a bimove, the charge of the state either does not change or changes by n (in one or the other direction). In order to convince ourselves of this, we let α and β be the two sparrows that move in the bimove (where α is the sparrow that moves clockwise, and β is the sparrow that moves counterclockwise), and we let i and j be the two trees on which they were sitting before the bimove. Now, we analyze four possible cases:

[244] These movements can be achieved by a single bimove, since tree $-m+1$ is the next tree from tree $-m$ in clockwise order, whereas tree $m-1$ is the next tree from tree m in counterclockwise order.

[245] These movements can be achieved by two bimoves, since tree $-m+2$ is the next tree from tree $-m+1$ in clockwise order, whereas tree $m-2$ is the next tree from tree $m-1$ in counterclockwise order.

- If $i < n$ and $j > 1$, then the sparrows α and β are sitting on trees $i+1$ and $j-1$ after the bimove. Thus, in this case, the charge of the state increases by $(i+1)+(j-1)-i-j=0$. In other words, the charge does not change.

- If $i = n$ and $j > 1$, then the sparrows α and β are sitting on trees 1 and $j-1$ after the bimove. Thus, in this case, the charge of the state increases by[246] $1+(j-1)-i-j=-i=-n$ (since $i=n$). In other words, the charge decreases by n.

- If $i < n$ and $j = 1$, then the sparrows α and β are sitting on trees $i+1$ and n after the bimove. Thus, in this case, the charge of the state increases by $(i+1)+n-i-j=n+1-\underbrace{j}_{=1}=n+1-1=n$. In other words, the charge increases by n.

- If $i = n$ and $j = 1$, then the sparrows α and β are sitting on trees 1 and n after the bimove. Thus, in this case, the charge of the state increases by $1+n-\underbrace{i}_{=n}-\underbrace{j}_{=1}=1+n-n-1=0$. In other words, the charge does not change.

Thus, in each of the four cases, the charge either does not change or changes by n. Therefore, the charge after the bimove is congruent to the charge before the bimove modulo n. In other words, the two charges leave the same remainder when divided by n.

Thus we have shown that a bimove does not change the remainder that the charge of the state leaves when divided by n. Since the charge of the initial state is $1+2+\cdots+n$, we thus conclude that the charge of any state that can be achieved by a sequence of bimoves will still leave the same remainder when divided by n as $1+2+\cdots+n$ does.

Now, let us look at a state where all sparrows sit on the same tree. If this tree is tree i, then the charge of the state is $\sum\limits_{s \text{ is a sparrow}} i = ni$ (since there are n sparrows). Thus, if we can achieve this state by a sequence of bimoves, then ni must leave the same remainder when divided by n as $1+2+\cdots+n$ does (according to the previous paragraph). In other words, if we can achieve this state by a sequence of bimoves, then we must have $ni \equiv 1+2+\cdots+n \bmod n$. Therefore, if we can achieve this state by a sequence of bimoves, then we must have $1+2+\cdots+n \equiv ni \equiv 0 \bmod n$ (since $n \mid ni$).

[246] The word "increase" is being understood in the wide sense here: To increase a number by k means to add k to the number, whether or not k is positive. Thus, "increase by -5" is simply another way to say "decrease by 5".

However, we do **not** have $1+2+\cdots+n \equiv 0 \bmod n$. Indeed, we have

$$1+2+\cdots+n = \frac{n(n+1)}{2} = \underbrace{\frac{n}{2}}_{\substack{\text{This is an integer} \\ \text{(since } n \text{ is even)}}} \cdot \left(\underbrace{n}_{\equiv 0 \bmod n} + 1 \right) \equiv \frac{n}{2} \cdot \underbrace{(0+1)}_{=1}$$

$$= \frac{n}{2} \not\equiv 0 \bmod n.$$

Thus, the previous paragraph lets us conclude that we **cannot** achieve a state where all sparrows sit on the same tree. This proves Claim 2.]

Thus, Exercise 8.1.4 is solved. □

Here is another classical exercise:

Exercise 8.1.5. Let $n \geq 3$ and $m \geq 3$ be two integers. You have a rectangular $n \times m$-grid of lamps (i.e., a table with n rows and m columns, with a lamp in each of its nm cells). Initially, all nm lamps are off. In a move, you can choose a row or a column of the grid, and flip all lamps in this row or column. (To "flip" a lamp means to turn it on if it was off, and to turn it off if it was on.) Can you, by some sequence of moves, obtain a state in which the four corner lamps (i.e., the lamps in the four corner cells of the grid) are on whereas all remaining lamps are off?

[**Example:** Let us represent a lamp turned off by the number 0, and a lamp turned on by the number 1. Then, for $n = 3$ and $m = 4$, here is the starting state:

0	0	0	0
0	0	0	0
0	0	0	0

.

One possible move is to flip all lamps in the third column (from the left). This results in

0	0	1	0
0	0	1	0
0	0	1	0

.

From this state, one possible further move is to flip all lamps in the second row. This results in

0	0	1	0
1	1	0	1
0	0	1	0

.

(Note that the third lamp in the second row is now turned off again, since it has been flipped twice.)]

Solution to Exercise 8.1.5 (sketched). No, you cannot.

Proof. Consider the four lamps in the upper-left corner (i.e., at the intersection of the topmost two rows with the leftmost two columns). Call these four lamps the *special lamps*. For example, in the case when $n = 3$ and $m = 4$, here are the four special lamps (marked with asterisks):

$*$	$*$		
$*$	$*$		

Each move either leaves the states of the four special lamps unchanged, or flips exactly two of them[247]. Thus, the number of special lamps that are on either remains unchanged or changes by 2 (in one or the other direction) when a move is made[248]. Therefore, the parity of this number never changes (i.e., if it was even, then it stays even; if it was odd, then it stays odd). But at the onset, this number was even (because at the onset, all four special lamps were off). Hence, after any sequence of moves, the number of special lamps that are on will remain even. Thus, in particular, it is impossible to achieve a state in which the four corner lamps are on whereas all remaining lamps are off (because such a state would have exactly 1 special lamp that is on[249]). This solves Exercise 8.1.5. □

Remark 8.1.1. Let us consider the lamp-flipping "game" from Exercise 8.1.5 a bit further. What is the total number of possible states that can be achieved by a sequence of moves (starting with the initial state in which all lamps are turned off)?

[247] Indeed, this can be verified case by case:

- If the move flips one of the topmost two rows, then exactly two special lamps get flipped.
- If the move flips any other row, then none of the special lamps get flipped, so the states of the four special lamps remain unchanged.
- If the move flips one of the leftmost two columns, then exactly two special lamps get flipped.
- If the move flips any other column, then none of the special lamps get flipped, so the states of the four special lamps remain unchanged.

[248] *Proof:* Consider a move. As we have just seen, this move either leaves the states of the four special lamps unchanged, or flips exactly two of them. If it leaves the states of the four special lamps unchanged, then the number of special lamps that are on clearly remains unchanged. On the other hand, if the move flips exactly two of the four special lamps, then the number of special lamps that are on increases by 2 (if the two special lamps it flips were both off) or decreases by 2 (if the two special lamps it flips were both on) or remains unchanged (if the two special lamps it flips were in different states – i.e., one of them was on and the other was off).

[249] Here we are using the assumptions $n \geq 3$ and $m \geq 3$. If you don't see why, ask yourself why only one of the four corner lamps is special!

Here is an outline of a solution to this question. We shall refer to rows and columns of our grid as *lines*. Thus, the grid has $n+m$ lines in total (viz., its n rows and its m columns). It is easy to see that the order of moves does not matter: If we are flipping two lines, it does not matter which of them we are flipping first. (Here, "flipping a line" means flipping every lamp on this line.) Furthermore, if we flip a line twice, we return to the state we had before the flipping (because every lamp on the line has been flipped twice). Thus, any state that can be achieved by a sequence of moves can also be achieved by a sequence of moves in which each line is flipped at most once. Moreover, the order of the moves does not matter. The total number of such sequences (without accounting for the order of moves) is 2^{n+m}, since we just need to decide which of the $n+m$ lines we flip. However, not all of the 2^{n+m} such sequences result in different states; for example, if we flip **every** line, the result is the original state (as if we hadn't flipped anything). It turns out that flipping the last column is unnecessary: The effect of flipping the last column is identical with the effect of flipping all remaining lines. Hence, we only need to consider sequences of moves in which each line is flipped at most once, and the last column is not flipped at all. There are 2^{n+m-1} such sequences, and moreover it can be shown that they all produce different states. (Why? Hint: Look at the effect on the last column.)

Hence, the total number of states that can be achieved by a sequence of moves is 2^{n+m-1}. This is a far cry from the total number of possible states (which is 2^{nm}, since each of the nm lamps can be either on or off).

The next exercise is [Engel98, Chapter 1, Example E2]:

Exercise 8.1.6. Let n be an odd positive integer. You have the $2n$ numbers $1, 2, \ldots, 2n$ written on a blackboard. In one move, you can erase two numbers a and b and write the number $|a-b|$ instead. You keep doing this until only one number remains. Prove that this remaining number will be odd.

First solution to Exercise 8.1.6 (sketched). In the initial state of the blackboard (i.e., before you make any moves), all numbers on the blackboard are integers (in fact, they are $1, 2, \ldots, 2n$ in the initial state). Thus, even after you start making moves, all numbers on the blackboard remain integers (because any move applied to integers only produces integers).

We define the *board sum* to be the sum of all numbers on the blackboard. In the initial state of the blackboard, this board sum is

$$\begin{aligned} 1+2+\cdots+2n &= \frac{2n(2n+1)}{2} \qquad \text{(by (9), applied to } 2n \text{ instead of } n\text{)} \\ &= \underbrace{n}_{\substack{\equiv 1 \bmod 2 \\ \text{(since } n \text{ is odd)}}} \left(\underbrace{2n}_{\equiv 0 \bmod 2} + 1 \right) \equiv 1(0+1) = 1 \bmod 2. \end{aligned}$$

In other words, in the initial state, the board sum is odd.

Now, it is not hard to see that (in general) a move will change the board sum. However, the **parity** of the board sum will remain unchanged. In order to see this, we need a few auxiliary observations: We first notice that every integer x satisfies

$$|x| \equiv x \bmod 2. \tag{369}$$

(Indeed, if x is an integer, then $|x|$ is always either x or $-x$; however, both x and $-x$ are congruent to x modulo 2. Thus, (369) follows.) Now, if a and b are any two integers, then we have

$$-a-b+\underbrace{|a-b|}_{\substack{\equiv a-b \bmod 2 \\ \text{(by (369)} \\ \text{(applied to } x=a-b))}} \equiv -a-b+(a-b) = -2b \equiv 0 \bmod 2,$$

and thus

$$\text{the integer } -a-b+|a-b| \text{ is even.} \tag{370}$$

Now, if you apply a move, then you erase two numbers a and b from the blackboard and write the number $|a-b|$ instead; as a result, the board sum increases[250] by $-a-b+|a-b|$. Since $-a-b+|a-b|$ is even (by (370), because a and b are integers), this entails that the board sum increases by an even integer. Hence, the parity of the board sum does not change.

Now, we have seen two things:

- In the initial state, the board sum is odd.
- The parity of the board sum does not change when we make a move.

Hence, after any number of moves, the board sum remains odd. Thus, the board sum will still be odd at the very end of the process, when there is only one number left on the blackboard. But this means that the number left on the blackboard is odd (because the board sum at that point will be that single number). Thus, Exercise 8.1.6 is solved. □

Second solution to Exercise 8.1.6 (sketched). Just as in the first solution above, we can see that all the numbers on the blackboard remain integers (even after you start making moves).

The *odd-count* will mean the number of odd numbers on the blackboard. In the initial state of the blackboard, this odd-count is n, since exactly n of the $2n$ numbers $1, 2, \ldots, 2n$ are odd. Now, how does the odd-count change when a move is applied?

- If you apply a move that erases an **even** number a and an **even** number b, then the new number $|a-b|$ written on the board in this move is even (since a and b are even, so that $a-b$ is even), and thus the odd-count remains the same (since no odd numbers were erased and no odd numbers were written).

[250] The word "increase" is being understood in the wide sense here: To increase a number by k means to add k to the number, whether or not k is positive. Thus, "increase by -5" is simply another way to say "decrease by 5".

- If you apply a move that erases an **even** number a and an **odd** number b, then the new number $|a-b|$ written on the board in this move is odd (since a is even and b is odd, so that $a-b$ is odd), and thus the odd-count remains the same (since one odd number was erased and one odd number was written).
- If you apply a move that erases an **odd** number a and an **even** number b, then the odd-count also remains the same (this is proved just as the preceding claim).
- If you apply a move that erases an **odd** number a and an **odd** number b, then the new number $|a-b|$ written on the board in this move is even (since a and b are odd, so that $a-b$ is even), and thus the odd-count decreases by 2 (since two odd numbers were erased but no odd numbers were written).

Combining these four observations, we see that if you apply a move, then the odd-count either remains the same or decreases by 2. Thus, the parity of the odd-count never changes. But we know that the odd-count was odd in the initial state (in fact, the odd-count was n in the initial state, but n is odd). Hence, the odd-count must remain odd throughout the process (since its parity never changes). Hence, at the end of the process, when there is only one number left on the blackboard, the odd-count will still be odd. But this means that the number left on the blackboard is odd (because if it was even, then there would be no odd number left on the blackboard; but this would cause the odd-count to be 0, which would contradict the fact that it is odd). This solves Exercise 8.1.6 again. □

Finally, here is a quickie problem (a popular puzzle in Russia, commonly ascribed to I. M. Gelfand) which shows that even the simplest processes can have useful invariants:

Exercise 8.1.7. A milk cup contains 200 ml of milk; a tea cup contains 200 ml of tea. You take a full spoon of milk from the milk cup and pour it into the tea cup; then you stir the latter cup. Then you take a full spoon of milk-tea mixture from the tea cup and pour it back into the milk cup. What is larger now: the amount of milk in the tea cup, or the amount of tea in the milk cup?

Solution to Exercise 8.1.7 (sketched). The two amounts are the same.

Proof. Let us introduce some notations:

- Let m be the amount of milk in the milk cup (after the pourings have been done).
- Let m' be the amount of milk in the tea cup (after the pourings have been done).
- Let t be the amount of tea in the tea cup (after the pourings have been done).

- Let t' be the amount of tea in the milk cup (after the pourings have been done).

We thus need to show that $m' = t'$.

Invariants come to our help here: The total quantity of milk does not change; thus, $m + m' = 200\,\text{ml}$ (since the total quantity of milk was 200 ml before the pourings, and is now $m + m'$ after the pourings). Similarly, $t + t' = 200\,\text{ml}$. Hence, $m + m' = 200\,\text{ml} = t + t'$. Also, the two cups have the same quantity of fluid (after the pourings have been done), because one full spoon was transferred from one to the other and then one full spoon was transferred backwards (so each cup has lost as much fluid as it has gained). Hence, $m + t' = t + m'$ (since $m + t'$ is the quantity of fluid in the milk cup, and $t + m'$ is the quantity of fluid in the tea cup). In other words, $m - m' = t - t'$. Subtracting this equality from $m + m' = t + t'$, we obtain $(m + m') - (m - m') = (t + t') - (t - t')$. This simplifies to $2m' = 2t'$. In other words, $m' = t'$. This proves our claim.

[*Remark:* Note that it doesn't matter whether the tea and the milk in the tea cup have mixed properly when we stirred them! Even if the spoon we have taken back from the tea cup was mostly milk, the argument still remains valid, and the answer is the same.] □

8.1.3. Applications to sequence integrality

Let us now briefly return to a topic we have already discussed before (particularly in Section 4.11): the phenomenon in which a sequence defined recursively turns out to consist of integers even though its recursive definition involves division. We shall now see a few more instances of this phenomenon; it turns out that invariants can be quite useful in proving these.

We begin with the following ([Gale98, Chapter 4, (2)]):

Exercise 8.1.8. Fix a positive integer $k \geq 2$. Define a sequence $(a_0, a_1, a_2, \ldots)$ of positive rational numbers recursively by setting

$$a_n = 1 \qquad \text{for each } n < k \tag{371}$$

and

$$a_n = \frac{a_{n-1}^2 + a_{n-2}^2 + \cdots + a_{n-k+1}^2}{a_{n-k}} \qquad \text{for each } n \geq k. \tag{372}$$

(For example, $a_k = \dfrac{a_{k-1}^2 + a_{k-2}^2 + \cdots + a_1^2}{a_0} = \dfrac{1^2 + 1^2 + \cdots + 1^2}{1} = k - 1$.)

Prove that a_n is a positive integer for each integer $n \geq 0$.

Solution to Exercise 8.1.8. Let us first play around with the recursive equation (372).

Let n be an integer such that $n \geq k+1$. Thus, $n \geq n-1 \geq k$ (since $n \geq k+1$). Hence, the equality (372) holds. Multiplying both sides of this equality by a_{n-k}, we obtain

$$a_n a_{n-k} = a_{n-1}^2 + a_{n-2}^2 + \cdots + a_{n-k+1}^2. \tag{373}$$

The same argument (applied to $n-1$ instead of n) yields

$$\begin{aligned} a_{n-1} a_{(n-1)-k} &= \underbrace{a_{(n-1)-1}^2}_{=a_{n-2}^2} + \underbrace{a_{(n-1)-2}^2}_{=a_{n-3}^2} + \cdots + \underbrace{a_{(n-1)-k+1}^2}_{=a_{n-k}^2} \qquad (\text{since } n-1 \geq k) \\ &= a_{n-2}^2 + a_{n-3}^2 + \cdots + a_{n-k}^2. \end{aligned}$$

In view of $(n-1)-k = n-k-1$, this rewrites as

$$a_{n-1} a_{n-k-1} = a_{n-2}^2 + a_{n-3}^2 + \cdots + a_{n-k}^2.$$

Subtracting this equality from (373), we obtain

$$\begin{aligned} a_n a_{n-k} - a_{n-1} a_{n-k-1} &= \underbrace{\left(a_{n-1}^2 + a_{n-2}^2 + \cdots + a_{n-k+1}^2\right)}_{=a_{n-1}^2 + \left(a_{n-2}^2 + a_{n-3}^2 + \cdots + a_{n-k+1}^2\right)} - \underbrace{\left(a_{n-2}^2 + a_{n-3}^2 + \cdots + a_{n-k}^2\right)}_{=\left(a_{n-2}^2 + a_{n-3}^2 + \cdots + a_{n-k+1}^2\right) + a_{n-k}^2} \\ &= \left(a_{n-1}^2 + \left(a_{n-2}^2 + a_{n-3}^2 + \cdots + a_{n-k+1}^2\right)\right) \\ &\qquad - \left(\left(a_{n-2}^2 + a_{n-3}^2 + \cdots + a_{n-k+1}^2\right) + a_{n-k}^2\right) \\ &= a_{n-1}^2 - a_{n-k}^2. \end{aligned}$$

Adding $a_{n-1} a_{n-k-1} + a_{n-k}^2$ to both sides of this equality, we obtain

$$a_n a_{n-k} + a_{n-k}^2 = a_{n-1}^2 + a_{n-1} a_{n-k-1}.$$

Thus,

$$a_{n-k}\left(a_n + a_{n-k}\right) = a_n a_{n-k} + a_{n-k}^2 = a_{n-1}^2 + a_{n-1} a_{n-k-1} = a_{n-1}\left(a_{n-1} + a_{n-k-1}\right).$$

Dividing both sides of this equality by $a_{n-1} a_{n-2} \cdots a_{n-k}$, we obtain

$$\frac{a_{n-k}\left(a_n + a_{n-k}\right)}{a_{n-1} a_{n-2} \cdots a_{n-k}} = \frac{a_{n-1}\left(a_{n-1} + a_{n-k-1}\right)}{a_{n-1} a_{n-2} \cdots a_{n-k}}.$$

This rewrites as

$$\frac{a_n + a_{n-k}}{a_{n-1} a_{n-2} \cdots a_{n-k+1}} = \frac{a_{n-1} + a_{n-k-1}}{a_{n-2} a_{n-3} \cdots a_{n-k}} \tag{374}$$

[251].

[251] because

$$\begin{aligned} \frac{a_{n-k}\left(a_n + a_{n-k}\right)}{a_{n-1} a_{n-2} \cdots a_{n-k}} &= \frac{a_{n-k}\left(a_n + a_{n-k}\right)}{\left(a_{n-1} a_{n-2} \cdots a_{n-k-1}\right) a_{n-k}} \\ &\qquad (\text{since } a_{n-1} a_{n-2} \cdots a_{n-k} = \left(a_{n-1} a_{n-2} \cdots a_{n-k-1}\right) a_{n-k}) \\ &= \frac{a_n + a_{n-k}}{a_{n-1} a_{n-2} \cdots a_{n-k+1}} \end{aligned}$$

Now, the equality (374) has a rather nice property: The subscripts on the right hand side are precisely by 1 lower than the corresponding subscripts on the left hand side! In other words, if we denote the left hand side by b_n, then the right hand side will be b_{n-1}. Thus, the equality (374) is, in a way, saying that the ratio $\dfrac{a_n + a_{n-k}}{a_{n-1}a_{n-2}\cdots a_{n-k+1}}$ is an "invariant" of the sequence $(a_0, a_1, a_2, \ldots)$: it does not change as we "move up" the sequence (i.e., it does not change from $n-1$ to n).

Let us now formalize this. Forget that we fixed n. We thus have proved the equality (374) for each integer $n \geq k+1$.

Now, let us define a number

$$b_n = \frac{a_n + a_{n-k}}{a_{n-1}a_{n-2}\cdots a_{n-k+1}}$$

for each integer $n \geq k$. Thus, we have defined a sequence $(b_k, b_{k+1}, b_{k+2}, \ldots)$ of rational numbers. It is easy to see that

$$b_k = k \tag{375}$$

[252].

Now, the equality (374) entails that

$$b_n = b_{n-1} \tag{377}$$

---

and

$$\begin{aligned}
\frac{a_{n-1}\left(a_{n-1} + a_{n-k-1}\right)}{a_{n-1}a_{n-2}\cdots a_{n-k}} &= \frac{a_{n-1}\left(a_{n-1} + a_{n-k-1}\right)}{a_{n-1}\left(a_{n-2}a_{n-3}\cdots a_{n-k}\right)} \\
&\qquad \left(\text{since } a_{n-1}a_{n-2}\cdots a_{n-k} = a_{n-1}\left(a_{n-2}a_{n-3}\cdots a_{n-k}\right)\right) \\
&= \frac{a_{n-1} + a_{n-k-1}}{a_{n-2}a_{n-3}\cdots a_{n-k}}
\end{aligned}$$

[252] *Proof of (375):* For each $i \in \{1, 2, \ldots, k-1\}$, we have $i \leq k-1 < k$ and therefore

$$a_i = 1 \tag{376}$$

(by (371), applied to $n = i$). Hence, $\prod_{i=1}^{k-1} \underbrace{a_i}_{\substack{=1 \\ \text{(by (376))}}} = \prod_{i=1}^{k-1} 1 = 1$. Moreover, $0 < k$ (since $k \geq 2 > 0$); thus, $a_0 = 1$ (by (371), applied to $n = 0$). Finally, (372) (applied to $n = k$) yields

$$\begin{aligned}
a_k &= \frac{a_{k-1}^2 + a_{k-2}^2 + \cdots + a_{k-k+1}^2}{a_{k-k}} = \frac{a_{k-1}^2 + a_{k-2}^2 + \cdots + a_{k-k+1}^2}{1} \qquad (\text{since } a_{k-k} = a_0 = 1) \\
&= a_{k-1}^2 + a_{k-2}^2 + \cdots + a_{k-k+1}^2 = a_{k-1}^2 + a_{k-2}^2 + \cdots + a_1^2 = a_1^2 + a_2^2 + \cdots + a_{k-1}^2 \\
&= \sum_{i=1}^{k-1} \underbrace{a_i^2}_{\substack{=1^2 \\ \text{(since (376) yields } a_i=1)}} = \sum_{i=1}^{k-1} \underbrace{1^2}_{=1} = \sum_{i=1}^{k-1} 1 = (k-1)\cdot 1 = k-1.
\end{aligned}$$

for every integer $n \geq k+1$ [253]. In other words, we have

$$b_{n-1} = b_n \qquad \text{for every integer } n \geq k+1.$$

Thus, we get a chain of equalities $b_k = b_{k+1} = b_{k+2} = b_{k+3} = \cdots$. Therefore, $b_n = b_k$ for each integer $n \geq k$. Hence, for each integer $n \geq k$, we have

$$b_n = b_k = k \tag{379}$$

(by (375)).

Now, if n is an integer satisfying $n \geq k$, then

$$\begin{aligned} a_n + a_{n-k} &= \underbrace{b_n}_{\substack{=k \\ \text{(by (379))}}} a_{n-1}a_{n-2}\cdots a_{n-k+1} \qquad \left(\text{since } b_n = \frac{a_n + a_{n-k}}{a_{n-1}a_{n-2}\cdots a_{n-k+1}}\right) \\ &= k a_{n-1}a_{n-2}\cdots a_{n-k+1} \end{aligned}$$

and thus

$$a_n = k a_{n-1}a_{n-2}\cdots a_{n-k+1} - a_{n-k}. \tag{380}$$

We are now almost there. The equality (380) that we just proved is a new recurrence equation for our sequence $(a_0, a_1, a_2, \ldots)$ that can be used to prove by a straightforward strong induction on n that all entries a_n of the sequence are integers. This is the same kind of argument that we have made in the solution to

---

Now, the definition of b_k yields

$$\begin{aligned} b_k &= \frac{a_k + a_{k-k}}{a_{k-1}a_{k-2}\cdots a_{k-k+1}} = \left(a_k + \underbrace{a_{k-k}}_{=a_0=1}\right) / \underbrace{\left(a_{k-1}a_{k-2}\cdots a_{k-k+1}\right)}_{\substack{=a_{k-1}a_{k-2}\cdots a_1 \\ =a_1a_2\cdots a_{k-1} = \prod_{i=1}^{k-1} a_i = 1}} = (a_k+1)/1 \\ &= a_k + 1 = k \qquad (\text{since } a_k = k-1). \end{aligned}$$

This proves (375).

[253] *Proof.* Let n be an integer such that $n \geq k+1$. The definition of b_n yields $b_n = \dfrac{a_n + a_{n-k}}{a_{n-1}a_{n-2}\cdots a_{n-k+1}}$.

The definition of b_{n-1} yields

$$b_{n-1} = \frac{a_{n-1} + a_{(n-1)-k}}{a_{(n-1)-1}a_{(n-1)-2}\cdots a_{(n-1)-k+1}} = \frac{a_{n-1} + a_{n-k-1}}{a_{n-2}a_{n-3}\cdots a_{n-k}} \tag{378}$$

(since $a_{(n-1)-k} = a_{n-k-1}$ and $\underbrace{a_{(n-1)-1}}_{=a_{n-2}}\underbrace{a_{(n-1)-2}}_{=a_{n-3}}\cdots\underbrace{a_{(n-1)-k+1}}_{=a_{n-k}} = a_{n-2}a_{n-3}\cdots a_{n-k}$). Hence,

$$\begin{aligned} b_n &= \frac{a_n + a_{n-k}}{a_{n-1}a_{n-2}\cdots a_{n-k+1}} = \frac{a_{n-1} + a_{n-k-1}}{a_{n-2}a_{n-3}\cdots a_{n-k}} \qquad (\text{by (374)}) \\ &= b_{n-1} \qquad (\text{by (378)}). \end{aligned}$$

This proves (377).

Exercise 1.1.2 **(b)**. But for the sake of completeness, let me give the details of this argument:

We claim that

$$a_m \in \mathbb{Z} \text{ for each } m \in \mathbb{N}. \tag{381}$$

[*Proof of (381):* We proceed by strong induction on m:

Induction step: Let $n \in \mathbb{N}$. Assume (as the induction hypothesis) that (381) holds for $m < n$. We must show that (381) holds for $m = n$. In other words, we must prove that $a_n \in \mathbb{Z}$.

If $n < k$, then this is clearly true (since (371) yields $a_n = 1 \in \mathbb{Z}$ in this case). Hence, we WLOG assume that $n \geq k$ for the rest of this proof.

Thus, (380) yields $a_n = ka_{n-1}a_{n-2}\cdots a_{n-k+1} - a_{n-k}$. But we have assumed that (381) holds for $m < n$. In other words, $a_m \in \mathbb{Z}$ holds for each $m \in \mathbb{N}$ satisfying $m < n$. In other words, $a_m \in \mathbb{Z}$ holds for each $m \in \{0, 1, \ldots, n-1\}$. In other words, the n elements $a_0, a_1, \ldots, a_{n-1}$ all belong to $\mathbb{Z}$. In other words, the n elements $a_0, a_1, \ldots, a_{n-1}$ are integers.

Hence, in particular, the $k-1$ elements $a_{n-1}, a_{n-2}, \ldots, a_{n-k+1}$ are integers (since these $k-1$ elements $a_{n-1}, a_{n-2}, \ldots, a_{n-k+1}$ are among the n elements $a_0, a_1, \ldots, a_{n-1}$), and the element a_{n-k} is an integer (since this element is, too, among the n elements $a_0, a_1, \ldots, a_{n-1}$). Therefore, the difference $ka_{n-1}a_{n-2}\cdots a_{n-k+1} - a_{n-k}$ is an integer as well (since it is formed by multiplying and subtracting the integers k, $a_{n-1}, a_{n-2}, \ldots, a_{n-k+1}$ and a_{n-k}). In other words, $ka_{n-1}a_{n-2}\cdots a_{n-k+1} - a_{n-k} \in \mathbb{Z}$. Hence, $a_n = ka_{n-1}a_{n-2}\cdots a_{n-k+1} - a_{n-k} \in \mathbb{Z}$. This completes the induction step. Thus, the proof of (381) is complete.]

Now, let $n \geq 0$ be an integer. Thus, $n \in \mathbb{N}$, so that $a_n \in \mathbb{Z}$ (by (381), applied to $m = n$). In other words, a_n is an integer. Furthermore, a_n is positive (since $(a_0, a_1, a_2, \ldots)$ is a sequence of positive rational numbers). Hence, a_n is a positive integer. This solves Exercise 8.1.8. □

Here is another example ([Gale98, Chapter 4, (4)]):[254]

Exercise 8.1.9. Fix a positive integer $k \geq 2$. Define a sequence $(a_0, a_1, a_2, \ldots)$ of positive rational numbers recursively by setting

$$a_n = 1 \qquad \text{for each } n < k \tag{382}$$

and

$$a_n = \frac{a_{n-1}a_{n-2} + a_{n-2}a_{n-3} + \cdots + a_{n-k+2}a_{n-k+1}}{a_{n-k}} \qquad \text{for each } n \geq k. \tag{383}$$

[254] We note that Exercise 8.1.9 generalizes Exercise 4.11.8 **(a)**, because if we take $k = 4$, then the sequence $(a_0, a_1, a_2, \ldots)$ defined in Exercise 8.1.9 will be precisely the sequence $(b_0, b_1, b_2, \ldots)$ in Exercise 4.11.8 **(a)**.

(The sum in the numerator on the right hand side of (383) is $a_{n-1}a_{n-2} + a_{n-2}a_{n-3} + \cdots + a_{n-k+2}a_{n-k+1} = \sum_{i=1}^{k-2} a_{n-i}a_{n-i-1}$.)

Prove that a_n is a positive integer for each integer $n \geq 0$.

Solution to Exercise 8.1.9 (sketched). This is similar to the above solution to Exercise 8.1.8, so we restrict ourselves to an outline.

Again, we begin by playing around with the recursive equation.

Let n be an integer satisfying $n \geq k+2$. Thus, $n \geq n-2 \geq k$ (since $n \geq k+2$); hence, (383) holds. Multiplying both sides of (383) by a_{n-k}, we obtain

$$a_n a_{n-k} = a_{n-1}a_{n-2} + a_{n-2}a_{n-3} + \cdots + a_{n-k+2}a_{n-k+1}. \tag{384}$$

The same argument (applied to $n-2$ instead of n) yields

$$a_{n-2}a_{n-k-2} = a_{n-3}a_{n-4} + a_{n-4}a_{n-5} + \cdots + a_{n-k}a_{n-k-1}.$$

[255] Subtracting this equality from (384), we find

$$\begin{aligned} a_n a_{n-k} - a_{n-2}a_{n-k-2} &= \left(a_{n-1}a_{n-2} + a_{n-2}a_{n-3} + \cdots + a_{n-k+2}a_{n-k+1}\right) \\ &\quad - \left(a_{n-3}a_{n-4} + a_{n-4}a_{n-5} + \cdots + a_{n-k}a_{n-k-1}\right) \\ &= a_{n-1}a_{n-2} + a_{n-2}a_{n-3} - a_{n-k+1}a_{n-k} - a_{n-k}a_{n-k-1}. \end{aligned}$$

Adding $a_{n-2}a_{n-k-2} + a_{n-k+1}a_{n-k} + a_{n-k}a_{n-k-1}$ to both sides of this equality, we obtain

$$a_n a_{n-k} + a_{n-k+1}a_{n-k} + a_{n-k}a_{n-k-1} = a_{n-1}a_{n-2} + a_{n-2}a_{n-3} + a_{n-2}a_{n-k-2}.$$

Adding $a_{n-k}a_{n-2}$ to both sides of this equality[256], we obtain

$$\begin{aligned} &a_n a_{n-k} + a_{n-k+1}a_{n-k} + a_{n-k}a_{n-k-1} + a_{n-k}a_{n-2} \\ &= a_{n-1}a_{n-2} + a_{n-2}a_{n-3} + a_{n-2}a_{n-k-2} + a_{n-k}a_{n-2}. \end{aligned}$$

In other words,

$$a_{n-k}\left(a_n + a_{n-2} + a_{n-k+1} + a_{n-k-1}\right) = a_{n-2}\left(a_{n-1} + a_{n-3} + a_{n-k} + a_{n-k-2}\right).$$

Dividing both sides of this equality by $a_{n-2}a_{n-3}\cdots a_{n-k}$, we obtain

$$\frac{a_n + a_{n-2} + a_{n-k+1} + a_{n-k-1}}{a_{n-2}a_{n-3}\cdots a_{n-k+1}} = \frac{a_{n-1} + a_{n-3} + a_{n-k} + a_{n-k-2}}{a_{n-3}a_{n-4}\cdots a_{n-k}}. \tag{385}$$

Now, forget that we fixed n. We thus have proved the equality (385) for each integer $n \geq k+2$.

[255] You might wonder why we used $n-2$ instead of $n-1$ here (unlike in the above solution to Exercise 8.1.8). The best answer I have is "I have tried using $n-1$, and it didn't lead me anywhere".

[256] This is somewhat unmotivated, but will serve to make the pattern below symmetric.

Now, let us define a number

$$b_n = \frac{a_n + a_{n-2} + a_{n-k+1} + a_{n-k-1}}{a_{n-2}a_{n-3}\cdots a_{n-k+1}}$$

for each integer $n \geq k+1$. Thus, we have defined a sequence $(b_{k+1}, b_{k+2}, b_{k+3}, \ldots)$ of rational numbers. It is easy to see that b_{k+1} is an integer (since $a_{k-1}, a_{k-2}, \ldots, a_0$ are all equal to 1, and thus all denominators involved in computing b_{k+1} are 1s).

Now, the equality (385) entails that

$$b_n = b_{n-1}$$

for each integer $n \geq k+2$ (because its left hand side is b_n, while its right hand side is b_{n-1}). In other words, we have $b_{n-1} = b_n$ for every integer $n \geq k+2$. Thus, we get a chain of equalities $b_{k+1} = b_{k+2} = b_{k+3} = b_{k+4} = \cdots$. From here, we can finish as in the above solution to Exercise 8.1.8. □

Here is a third example ([Gale98, Chapter 4, (3)]):

Exercise 8.1.10. Fix an odd positive integer $k \geq 2$. Define a sequence $(a_0, a_1, a_2, \ldots)$ of positive rational numbers recursively by setting

$$a_n = 1 \qquad \text{for each } n < k \tag{386}$$

and

$$a_n = \frac{a_{n-1}a_{n-2} + a_{n-3}a_{n-4} + \cdots + a_{n-k+2}a_{n-k+1}}{a_{n-k}} \qquad \text{for each } n \geq k. \tag{387}$$

(The sum in the numerator on the right hand side of (387) is $a_{n-1}a_{n-2} + a_{n-3}a_{n-4} + \cdots + a_{n-k+2}a_{n-k+1} = \sum_{i=1}^{(k-1)/2} a_{n-2i+1}a_{n-2i}$.)

Prove that a_n is a positive integer for each integer $n \geq 0$.

Solution to Exercise 8.1.10 (sketched). This is similar to the above solution to Exercise 8.1.8, so we restrict ourselves to an outline.

Again, we begin by playing around with the recursive equation.

Let n be an integer satisfying $n \geq k+2$. Thus, $n \geq n-2 \geq k$ (since $n \geq k+2$); hence, (387) holds. Multiplying both sides of (387) by a_{n-k}, we obtain

$$a_n a_{n-k} = a_{n-1}a_{n-2} + a_{n-3}a_{n-4} + \cdots + a_{n-k+2}a_{n-k+1}. \tag{388}$$

The same argument (applied to $n-2$ instead of n) yields

$$a_{n-2}a_{n-k-2} = a_{n-3}a_{n-4} + a_{n-5}a_{n-6} + \cdots + a_{n-k}a_{n-k-1}.$$

Subtracting this equality from (388), we find

$$\begin{aligned} a_n a_{n-k} - a_{n-2} a_{n-k-2} &= \left(a_{n-1} a_{n-2} + a_{n-3} a_{n-4} + \cdots + a_{n-k+2} a_{n-k+1}\right) \\ &\qquad - \left(a_{n-3} a_{n-4} + a_{n-5} a_{n-6} + \cdots + a_{n-k} a_{n-k-1}\right) \\ &= a_{n-1} a_{n-2} - a_{n-k} a_{n-k-1}. \end{aligned}$$

Adding $a_{n-2} a_{n-k-2} + a_{n-k} a_{n-k-1}$ to both sides of this equality, we obtain

$$a_n a_{n-k} + a_{n-k} a_{n-k-1} = a_{n-1} a_{n-2} + a_{n-2} a_{n-k-2}.$$

In other words,

$$a_{n-k} \left(a_n + a_{n-k-1}\right) = a_{n-2} \left(a_{n-1} + a_{n-k-2}\right).$$

Dividing both sides of this equality by $a_{n-2} a_{n-3} \cdots a_{n-k}$, we obtain

$$\frac{a_n + a_{n-k-1}}{a_{n-2} a_{n-3} \cdots a_{n-k+1}} = \frac{a_{n-1} + a_{n-k-2}}{a_{n-3} a_{n-4} \cdots a_{n-k}}. \tag{389}$$

Now, forget that we fixed n. We thus have proved the equality (389) for each integer $n \geq k+2$.

Now, let us define a number

$$b_n = \frac{a_n + a_{n-k-1}}{a_{n-2} a_{n-3} \cdots a_{n-k+1}}$$

for each integer $n \geq k+1$. Thus, we have defined a sequence $(b_{k+1}, b_{k+2}, b_{k+3}, \ldots)$ of rational numbers. It is easy to see that b_{k+1} is an integer (since $a_{k-1}, a_{k-2}, \ldots, a_0$ are all equal to 1, and thus all denominators involved in computing b_{k+1} are 1s).

Now, the equality (389) entails that

$$b_n = b_{n-1}$$

for each integer $n \geq k+2$ (because its left hand side is b_n, while its right hand side is b_{n-1}). In other words, we have $b_{n-1} = b_n$ for every integer $n \geq k+2$. Thus, we get a chain of equalities $b_{k+1} = b_{k+2} = b_{k+3} = b_{k+4} = \cdots$. From here, we can finish as in the above solution to Exercise 8.1.8. □

8.2. Monovariants

Next we shall see some exercises illustrating the use of monovariants (i.e., quantities that change only in one direction). We begin with a basic fact about sorting sequences. First, a definition:

Definition 8.2.1. (a) We say that an n-tuple $(b_1, b_2, \ldots, b_n)$ of real numbers is *weakly increasing* if $b_1 \leq b_2 \leq \cdots \leq b_n$. (For example, the 5-tuple $(1, 4, 4, 7, 9)$ is weakly increasing.)

(b) Let $(b_1, b_2, \ldots, b_n)$ be an n-tuple of real numbers. We say that two entries b_i and b_j (with $i < j$) of this n-tuple are *out of order* if $b_i > b_j$. (For example, the entries 4 and 2 in the 5-tuple $(1, 4, 3, 2, 5)$ are out of order; thus, this 5-tuple is not weakly increasing.)

It is clear that an n-tuple of real numbers is weakly increasing if and only if it has no two entries that are out of order. Now, the following exercise shows that we can sort any n-tuple of real numbers into weakly increasing order by repeatedly swapping out-of-order pairs of adjacent entries[257]:

Exercise 8.2.1. Let $n \in \mathbb{N}$. You start with an n-tuple $(a_1, a_2, \ldots, a_n)$ of real numbers. In one move, you are allowed to pick two adjacent entries a_i and a_{i+1} of this n-tuple that are out of order (i.e., satisfy $a_i > a_{i+1}$), and swap these two entries. Prove that after at most $\binom{n}{2}$ such moves, the n-tuple will become weakly increasing.

[**Example:** If $n = 5$ and if you start with the 5-tuple $(2, 5, 4, 1, 3)$, then one possible sequence of moves you can take is the following one:

$$(2,5,4,1,3) \overset{\text{swap 5 and 4}}{\longrightarrow} (2,4,5,1,3) \overset{\text{swap 5 and 1}}{\longrightarrow} (2,4,1,5,3) \overset{\text{swap 5 and 3}}{\longrightarrow} (2,4,1,3,5)$$

$$\overset{\text{swap 4 and 1}}{\longrightarrow} (2,1,4,3,5) \overset{\text{swap 2 and 1}}{\longrightarrow} (1,2,4,3,5) \overset{\text{swap 4 and 3}}{\longrightarrow} (1,2,3,4,5).$$

The 5-tuple has become weakly increasing after these moves. Other sequences of moves are also possible, but they all lead to the same final result.]

Solution to Exercise 8.2.1 (sketched). In this solution, the word "n-tuple" shall always mean "n-tuple of real numbers".

An *inversion* of an n-tuple $(a_1, a_2, \ldots, a_n)$ shall mean a pair (i, j) of elements of $[n]$ satisfying $i < j$ and $a_i > a_j$. Thus, the inversions of an n-tuple correspond to the pairs of entries of this n-tuple that are out of order. (However, an inversion is not the pair of these **entries**, but rather the pair of their **positions**.) For example, the inversions of the 4-tuple $(3, 1, 4, 2)$ are $(1, 2)$, $(1, 4)$ and $(3, 4)$. For another example, the inversions of the 3-tuple $(2, 2, 1)$ are $(1, 3)$ and $(2, 3)$. (Note that the pairs of entries corresponding to these inversions are both $(2, 1)$.)

The *inversion number* of an n-tuple shall be defined as the number of inversions of this n-tuple. For example, the inversion number of the 4-tuple $(3, 1, 4, 2)$ is 3, since it has 3 inversions. We notice that the inversion number of an n-tuple is always a nonnegative integer (by its definition).

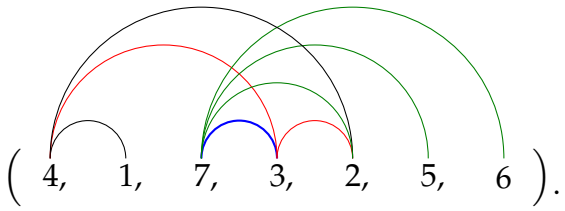
Now, we claim the following:

> *Claim 1:* Every move decreases the inversion number of our n-tuple precisely by 1.

[*Proof of Claim 1:* I shall illustrate the argument visually before discussing it in more technical detail.

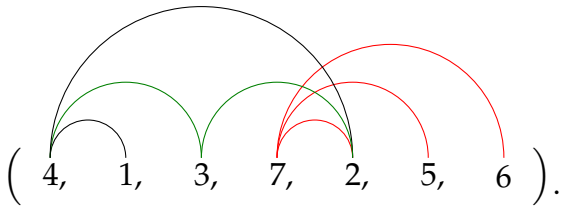
[257] This is the idea underlying the sorting algorithm called *bubble sort* ([TAoCP3, §5.2.2]).

Let us visually represent the inversions of the 7-tuple $(4,1,7,3,2,5,6)$.



(390)

The arcs here illustrate the inversions; more precisely, there is an arc between the i-th entry and the j-th entry whenever (i,j) is an inversion of the 7-tuple. (The colors will be explained later.) The two adjacent entries 7 and 3 (in positions 3 and 4) are out of order. The move that swaps these two entries transforms the 7-tuple into the following 7-tuple:



(391)

Again, we have represented the inversions as arcs. What do we see? The two black arcs have stayed in their positions; the red and the green arcs have slightly moved and changed their colors (from red to green, and from green to red); the blue arc has disappeared. The logic behind the colors is the following: The swap involved the 3-rd and the 4-th entry of our 7-tuple. The arc that connects these two entries was colored blue; the arcs that contain the 3-rd but not the 4-th entry were colored green; the arcs that contain the 4-th but not the 3-rd entry were colored red; all remaining arcs were colored black. The behavior is no longer strange:

- The black arcs from (390) remain arcs in (391), because the relevant entries do not change.
- The green arcs from (390) appear (slightly stretched or compressed) as red arcs in (391), since one of the relevant entries is moved (but not far enough to jump past the other entry and therefore destroy the inversion).
- The red arcs from (390) appear (slightly stretched or compressed) as green arcs in (391) (for the same reason).
- The blue arc from (390) disappears in (391) (since the two relevant entries are no longer out of order in (391)).

This analysis accounts for all arcs in (391). Thus, we see that there is precisely one less arc in (391) than in (390). In other words, the 7-tuple in (391) has precisely

one less inversion than that in (390). In other words, our move has decreased the inversion number by 1, exactly as Claim 1 predicted.

It is not hard (although somewhat laborious) to formalize this argument in full generality. Instead of assigning colors to arcs, let me assign colors to the inversions themselves. Consider a move that transforms an n-tuple $\mathbf{a} = (a_1, a_2, \ldots, a_n)$ into an n-tuple $\mathbf{b} = (b_1, b_2, \ldots, b_n)$ by swapping two adjacent entries a_i and a_{i+1} that are out of order. (Thus, $b_i = a_{i+1}$ and $b_{i+1} = a_i$ and $b_j = a_j$ for each $j \in [n] \setminus \{i, i+1\}$.) An inversion (u, v) of $\mathbf{a}$ or $\mathbf{b}$ will be called

- *blue* if both i and $i+1$ belong to $\{u, v\}$;
- *green* if $i \in \{u, v\}$ but $i+1 \notin \{u, v\}$;
- *red* if $i+1 \in \{u, v\}$ but $i \notin \{u, v\}$;
- *black* if neither i nor $i+1$ belongs to $\{u, v\}$.

(These are precisely the colors of the corresponding arcs in (390) and in (391).) It is clear that each inversion of $\mathbf{a}$ or $\mathbf{b}$ has exactly one of these four colors (blue, green, red and black). Moreover:

- The n-tuple $\mathbf{a}$ has exactly one blue inversion (namely, $(i, i+1)$), whereas the n-tuple $\mathbf{b}$ has no blue inversions (since its i-th and $(i+1)$-st entries are not out of order).
- The n-tuple $\mathbf{a}$ has as many green inversions as the n-tuple $\mathbf{b}$ has red inversions. Indeed, there is a bijection

  $$\{\text{green inversions of } \mathbf{a}\} \to \{\text{red inversions of } \mathbf{b}\}$$

  that sends each green inversion (u, i) of $\mathbf{a}$ to the red inversion $(u, i+1)$ of $\mathbf{b}$ and sends each green inversion (i, v) of $\mathbf{a}$ to the red inversion $(i+1, v)$ of $\mathbf{b}$. (Obviously, any green inversion of $\mathbf{a}$ has one of the forms (u, i) and (i, v).)
- The n-tuple $\mathbf{a}$ has as many red inversions as the n-tuple $\mathbf{b}$ has green inversions. Indeed, there is a bijection

  $$\{\text{red inversions of } \mathbf{a}\} \to \{\text{green inversions of } \mathbf{b}\}$$

  that sends each red inversion $(u, i+1)$ of $\mathbf{a}$ to the green inversion (u, i) of $\mathbf{b}$ and sends each red inversion $(i+1, v)$ of $\mathbf{a}$ to the green inversion (i, v) of $\mathbf{b}$. (Obviously, any red inversion of $\mathbf{a}$ has one of the forms $(u, i+1)$ and $(i+1, v)$.)
- The n-tuple $\mathbf{a}$ has as many black inversions as the n-tuple $\mathbf{b}$ has black inversions. Indeed, the black inversions of $\mathbf{a}$ are identical with the black inversions of $\mathbf{b}$.

This accounting shows that the n-tuple $\mathbf{a}$ has exactly one more inversion than the n-tuple $\mathbf{b}$ has. In other words, the inversion number of $\mathbf{a}$ equals the inversion number of $\mathbf{b}$ plus 1. Thus, the move that transformed $\mathbf{a}$ into $\mathbf{b}$ had the effect of decreasing the inversion number by 1. This proves Claim 1.[258]]

Now, Claim 1 shows that the inversion number of our n-tuple is a monovariant: it keeps decreasing. Let us make this more precise:

At the onset, the inversion number of our n-tuple is at most $\binom{n}{2}$ (since every inversion of the n-tuple is a pair (i,j) of elements of $[n]$ satisfying $i<j$; but there are only $\binom{n}{2}$ such pairs[259]). Claim 1 shows that this inversion number decreases by 1 every time you make a move. Hence, after k moves, this inversion number will be $\binom{n}{2}-k$. Therefore, if you make more than $\binom{n}{2}$ successive moves, then the inversion number becomes smaller than $\binom{n}{2}-\binom{n}{2}=0$, which is clearly impossible (since the inversion number of an n-tuple is always a nonnegative integer). Thus, you cannot make more than $\binom{n}{2}$ successive moves. Hence, after at most $\binom{n}{2}$ moves, you will obtain an n-tuple that does not allow for any further moves. But such an n-tuple must necessarily be weakly increasing (because it does not allow for any further moves, so it has no two adjacent entries that are out of order[260]; but this means that its entries weakly increase from left to right). Thus, after at most $\binom{n}{2}$ moves, you will obtain a weakly increasing n-tuple. This solves Exercise

[258] A more formal exposition of this proof can be found in [Grinbe15, §7.80, proof of (1072)]. (Note that the notation $\ell(\mathbf{a})$ used in [Grinbe15, §7.80, (1072)] is exactly what we call the inversion number of $\mathbf{a}$ here, whereas the notation $\mathbf{a}\circ s_k$ used in [Grinbe15, §7.80, (1072)] is the result of swapping the k-th and $(k+1)$-st entries of the n-tuple $\mathbf{a}$. Thus, [Grinbe15, (1072)] says precisely that a move (in our sense) decreases the inversion number of an n-tuple by 1. The fact that we are working with n-tuples of real numbers whereas [Grinbe15, §7.80, (1072)] is stated for n-tuples of integers is immaterial; this makes no difference for the proof.)

[259] *Proof.* There is a bijection

$$\begin{aligned}\{\text{pairs } (i,j) \text{ of elements of } [n] \text{ satisfying } i<j\} &\to \{\text{2-element subsets of } [n]\}, \\ (i,j) &\mapsto \{i,j\}.\end{aligned}$$

Hence, the bijection principle yields

$$\begin{aligned}&|\{\text{pairs } (i,j) \text{ of elements of } [n] \text{ satisfying } i<j\}| \\ &= |\{\text{2-element subsets of } [n]\}| = (\text{\# of 2-element subsets of } [n]) = \binom{n}{2}\end{aligned}$$

(by Theorem 4.3.12). In other words, the number of pairs (i,j) of elements of $[n]$ satisfying $i<j$ is $\binom{n}{2}$.

[260] because any two adjacent entries that are out of order would create an opportunity for a move

8.2.1. □

Our next exercise ([Grinbe08, Aufgabe 4.16]) is again about lamps in a rectangular grid (just like Exercise 8.1.5):

Exercise 8.2.2. Let n and m be two positive integers. You have a rectangular $n \times m$-grid of lamps (i.e., a table with n rows and m columns, with a lamp in each of its nm cells). A *line* shall mean a row or a column of the grid; thus, there are $n + m$ lines in total (namely, n rows and m columns). You say that a line is *bright* if at least half of all lamps in this line are turned on.

In a move, you can choose a line, and flip all lamps in this line. (To "flip" a lamp means to turn it on if it was off, and to turn it off if it was on. Note that our definition of a move is exactly the same as in Exercise 8.1.5.)

Prove that (starting with an arbitrary state of the lamps – not necessarily all turned off) you can always find a sequence of moves after which every line is bright.

[**Example:** Let us represent a lamp turned off by the number 0, and a lamp turned on by the number 1. Consider the following starting state (for $n = 3$ and $m = 4$):

$$\begin{array}{|c|c|c|c|}\hline 0 & 0 & 1 & 1 \\ \hline 1 & 1 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 \\ \hline \end{array} \quad . \tag{392}$$

In this state, the first two rows (from the top) are bright, and so is the third column (from the left); but the remaining lines are not bright. However, we can make all lines bright by the following sequence of moves: First, we flip all lamps in the first and second columns; then, we flip all lamps in the second row. The resulting state is

$$\begin{array}{|c|c|c|c|}\hline 1 & 1 & 1 & 1 \\ \hline 1 & 1 & 1 & 1 \\ \hline 1 & 1 & 1 & 0 \\ \hline \end{array} \quad ,$$

and this does have the property that all lines are bright.]

Solution to Exercise 8.2.2 (sketched). We define the *total brightness* of a state to be the # of all lamps that are turned on in this state. For example, the state shown in (392) has total brightness 5. Note that the total brightness of any state is a nonnegative integer that is $\leq nm$ (since there are only nm lamps in total).

Let me recall that any move consists in picking a line and flipping all lamps in this line. We shall call a move *lightbringing* if the line getting picked was not bright before the move. For example, if we start from the state (392), then flipping all lamps in the topmost row is not a lightbringing move (since the topmost row was already bright before the move), but flipping all lamps in the bottommost row is a lightbringing move (since the bottommost row was not bright before the move).

Now, we claim the following:

> *Claim 1:* Any lightbringing move increases the total brightness of a state by at least 1.

[*Proof of Claim 1:* Consider any lightbringing move. Let ℓ be the line that is picked in this move. Thus, the move consists in flipping all lamps in this line ℓ. Since this move is lightbringing, we thus conclude that the line ℓ is not bright before this move (by the definition of "lightbringing"). In other words, less than half of the lamps in line ℓ are turned on before the move (by the definition of a "bright" line). In other words, before the move, the line ℓ has fewer turned-on lamps than it has turned-off lamps. But the move turns all the former lamps off and turns all the latter lamps on (because the move flips all lamps in ℓ). Hence, the move turns fewer lamps off than it turns on (since the line ℓ has fewer turned-on lamps than it has turned-off lamps before the move). Hence, the move increases the total # of lamps that are turned on. In other words, the move increases the total brightness of the state (since the total brightness of a state was defined as the total # of lamps that are turned on). Therefore, the move increases the total brightness of the state by at least 1 (since the total brightness is an integer, and thus, if it is increased, it must necessarily be increased by at least 1). This proves Claim 1.]

Now, let us start with an arbitrary state, and keep making lightbringing moves until this is no longer possible. (That is, we keep making lightbringing moves as long as there is any line that is not bright; we stop when there are no such lines left any more.) I claim that this process will end after at most nm moves. Indeed, Claim 1 shows that any lightbringing move increases the total brightness of a state by at least 1. Hence, if we manage to make $nm+1$ moves, then the total brightness will be at least $nm+1$ (since the total brightness was nonnegative in the initial state, and has been increased by 1 in each of our $nm+1$ moves); but this would contradict the fact that the total brightness of any state is $\leq nm$ (and thus $< nm+1$). Thus, we cannot make $nm+1$ moves. Hence, our process will stop after at most nm moves. The resulting state (at the end of our process) will have the property that every line is bright (since otherwise, it would have a line that is not bright, and therefore we could apply one more lightbringing move to it). Thus, we have found a sequence of moves after which every line is bright. This solves Exercise 8.2.2. $\square$

Here is yet another lamp puzzle ([Grinbe08, Aufgabe 4.20]), albeit stated in terms of flags rather than lamps:

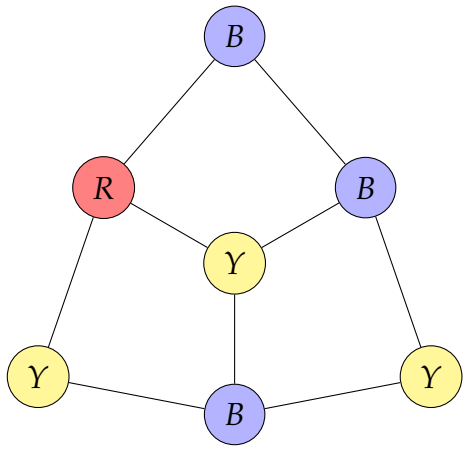
Exercise 8.2.3. A country has finitely many towns, some of which are connected by roads. We say that two towns are each others' *neighbors* if they are connected by a direct road (not passing through any other town). (No town counts as its own neighbor.)

Each town flies a flag, which has a certain color. We shall briefly say "A town T flies a color c" for "A town T flies a flag of color c".

Every once in a while, a revolution happens in a town: The town's flag is repainted in the color of the plurality of its neighbors. That is, if T is a town, and if a plurality of T's neighbors are flying a certain color c, then the revolution causes the flag of T to be repainted in color c. ("Plurality" means "relative majority", i.e., more neighbors than any other color. That is, "a plurality of T's neighbors are flying color c" means "for each color $d \neq c$, the number of T's neighbors flying color c is larger than the number of T's neighbors flying color d". If T was already flying color c right before the change, then nothing changes, and this does not count as a revolution.)

Assume that revolutions cannot happen simultaneously in several towns. Prove that the revolutions will eventually end – i.e., after sufficiently many revolutions, no more revolutions will happen.

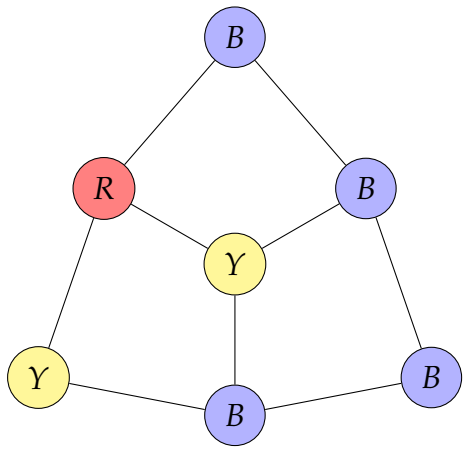
[**Example:** Here is an example of a country with 7 towns (with direct roads represented by line segments):



.

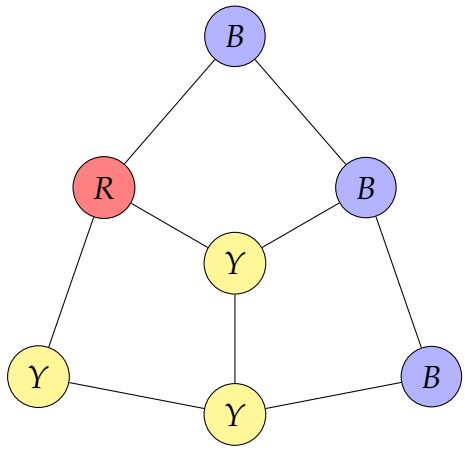
Here, the colors of the nodes represent the colors of the flags hoisted by the respective towns, and the letters inside the nodes represent them as well (for the convenience of anyone reading this in black-and-white). ("Y" means "yellow"; "R" means "red"; "B" means "blue".) Now, a revolution can happen (e.g.) in the yellow-flagged town at the bottom right of the country. The plurality of this town's neighbors are flying blue flags (in fact, both of its neighbors are), so the

revolution causes this town to fly a blue flag, too:



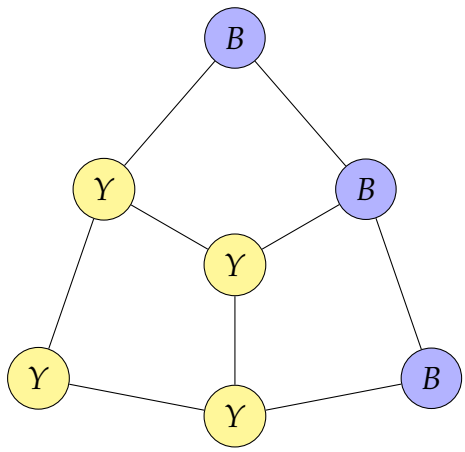
.

Next, a revolution can happen (e.g.) in the blue-flagged town at the bottom of the country. The plurality of this town's neighbors (namely, two of its three neighbors) are flying yellow flags; thus, this town becomes yellow too:



. (393)

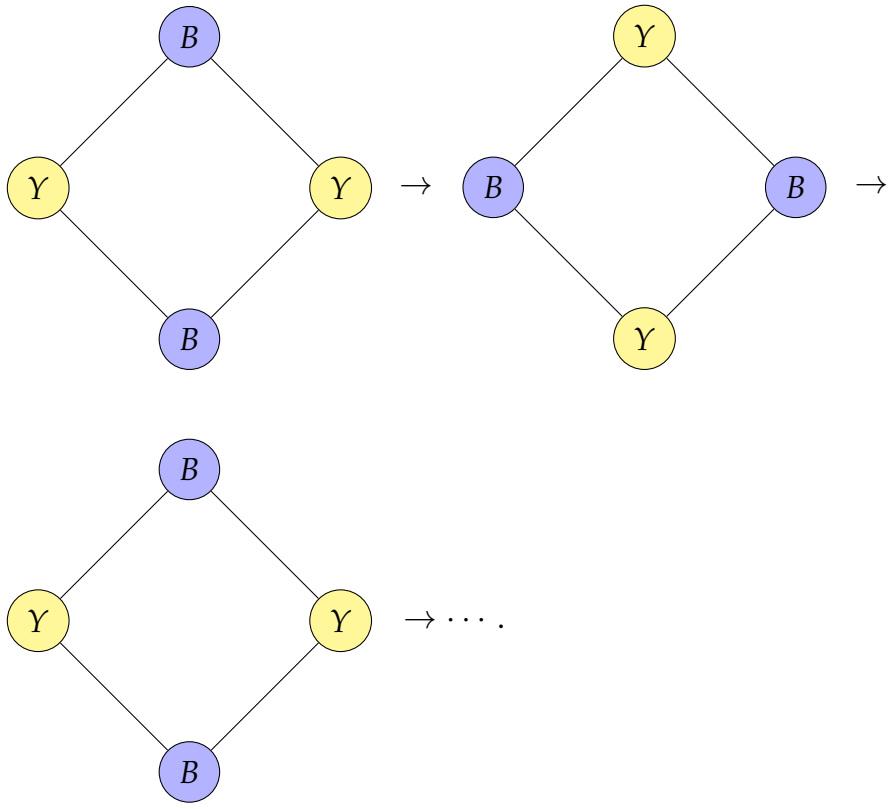
Next, a revolution can happen in the red town, resulting in



.

The country is now stable: No further revolutions can happen. (This does not mean that each town has the same flag color as a plurality of its neighbors; it only means that no other color has a plurality among its neighbors.)]

Remark 8.2.2. If revolutions were allowed to happen simultaneously in Exercise 8.2.3, then the claim of the exercise would not be true. For instance, here is an example where revolutions all happen simultaneously in all towns, causing every town to switch colors every time (and never stabilizing):



Some things can be said about the variant of the exercise in which **all possible** revolutions are happening simultaneously (e.g., each night, at midnight, every town whose flag color differs from that of a plurality of its neighbors undergoes a revolution). In this case, I suspect that the country will either become stable, or will end up oscillating between two states from a certain day on. A similar result (albeit not quite the same, as it includes a rule for tiebreaking between equinumerous pluralities) appears in [GolTch83, Theorem 2.1].

Solution to Exercise 8.2.3 (sketched). A *couple* shall mean a set $\{a, b\}$ of two towns a and b that are connected by a (direct) road. A couple $\{a, b\}$ shall be called *discordant* if the towns a and b fly different colors; otherwise, it shall be called *non-discordant*. (Of course, these concepts depend on the state of the country; a discordant couple can become non-discordant after a revolution, and vice versa.)

Define the *discord* of a state to be the # of all discordant couples in this state. (For example, the discord of the state shown in (393) is 5. The discord of a state can be read off its pictorial representation, as it is simply the # of segments whose two endpoints have different colors – at least if we don't have two towns connected by more than one direct road.)

The discord of a state is thus a nonnegative integer.

Now, we claim the following:

Claim 1: Each revolution decreases the discord of the state.

[*Proof of Claim 1:* Let us consider a revolution happening in a town t. We notice that the only flag that gets recolored during this revolution is the flag in town t. Let d be the color that t flies before the revolution, and let c be the color that t flies after the revolution. Due to the way revolutions work, the color c must be the flag color of a plurality of t's neighbors. Thus, in particular, more of t's neighbors are flying the color c than are flying the color d. (Here we are using the fact that $c \neq d$, which is because otherwise our "revolution" would not count as a revolution.) Let γ be the # of t's neighbors that are flying the color c, and let δ be the # of t's neighbors that are flying the color d. (We don't need to specify whether we are talking of "before the revolution" or "after the revolution", since the revolution only changes the color of t, not the color of t's neighbors.) Thus, $\gamma > \delta$ (since more of t's neighbors are flying the color c than are flying the color d). Hence, $\gamma - \delta > 0$.

Now, we consider the effect of the revolution on couples:

- If a couple does not include t, then the revolution does not change the discordant/non-discordant status of the couple (i.e., if the couple was discordant before the revolution, then it stays discordant; if it was non-discordant, then it stays non-discordant).

- If a couple has the form $\{t, x\}$, where x is a neighbor of t flying the color d, then this couple $\{t, x\}$ was non-discordant before the revolution (since t also had the color d before the revolution) but becomes discordant after the revolution (because the color of t changes). Note that there are exactly δ such couples (since δ is the # of t's neighbors that are flying the color d).

- If a couple has the form $\{t, y\}$, where y is a neighbor of t flying the color c, then this couple $\{t, y\}$ was discordant before the revolution (since t had the color d, and since $c \neq d$) but becomes non-discordant after the revolution (because the color of t changes to c). Note that there are exactly γ such couples (since γ is the # of t's neighbors that are flying the color c).

- If a couple has the form $\{t, z\}$, where z is a neighbor of t flying a color distinct from d and c, then this couple $\{t, z\}$ was discordant before the revolution and remains so after the revolution.

This accounts for all couples. Thus, the revolution causes exactly δ non-discordant couples to become discordant, and causes exactly γ discordant couples to become non-discordant. The total # of discordant couples thus decreases by $\gamma - \delta$. Since $\gamma - \delta > 0$, this shows that the total # of discordant couples decreases. Thus, Claim 1 is proven.]

The rest of the solution is now essentially a repetition of the "infinite descent" argument we have seen in the discussion of Exercise 5.3.1. Namely, Claim 1 shows that each revolution decreases the discord of the state. But the discord of any state is a nonnegative integer, and thus cannot decrease indefinitely. More concretely: If the discord of the original state was N, then the discord cannot decrease more than N times without becoming negative. Thus, we cannot have more than N successive revolutions. This solves Exercise 8.2.3. □

The following exercise is a variation on Exercise 8.1.3 – with a different answer, however:

Exercise 8.2.4. The numbers $1, 2, \ldots, 99$ are written in a row (in this order, from left to right). In a move, you can swap any two numbers at a distance of 2 (i.e., any two numbers that have exactly one number written between them). Can you end up with the numbers $99, 98, \ldots, 1$ (in this order, from left to right) by a sequence of such moves?

[See the Example in Exercise 8.1.3 for how such moves look like.]

Solution to Exercise 8.2.4 (sketched). Yes, you can.

Proof. The number 99 is odd, so we cannot proceed as in Exercise 8.1.3. However, we can argue as follows:[261]

We identify each state with the 99-tuple consisting of the numbers written on the row, in the order in which they appear (from left to right). Thus, the initial state is the 99-tuple $(1, 2, \ldots, 99)$, whereas the state we are trying to reach is $(99, 98, \ldots, 1)$.

First of all, we observe that our moves are reversible: More precisely, we can undo each move by making the same move (since swapping the same two numbers twice in a row is tantamount to doing nothing). Thus, if there is a sequence of moves that transforms a certain state A into a certain state B, then there also is a sequence of moves that transforms the state B into the state A (namely, the original sequence of moves, but made in backwards order). Hence, in particular, if there is a sequence of moves that transforms the state $(99, 98, \ldots, 1)$ into the state $(1, 2, \ldots, 99)$, then there also is a sequence of moves that transforms the state $(1, 2, \ldots, 99)$ into the state $(99, 98, \ldots, 1)$.

Hence, instead of solving the exercise as it is posed, let us solve the "inverse" exercise (i.e., the exercise with initial state and desired state interchanged): Let us find a sequence of moves that transforms the state $(99, 98, \ldots, 1)$ into the state $(1, 2, \ldots, 99)$. As we have just said, this is equivalent to the exercise as it is posed.

[261] Note that we are not going to use any "new" monovariants in this argument; instead, we shall use Exercise 8.2.1, thus in a way reusing the monovariant we already used to solve Exercise 8.2.1.

We define the *position* of a number $i \in [99]$ in a given state to be the (unique) $j \in [99]$ with the property that i is the j-th number from the left on our row.

In any state, we shall call a number $i \in [99]$ *odd-positioned* if its position is odd[262], and we shall call a number *even-positioned* if its position is even. In the initial state of our "inverse" exercise (which, as we recall, is the 99-tuple $(99, 98, \ldots, 1)$), the odd-positioned numbers are $99, 97, 95, \ldots, 1$ (indeed, their positions are $1, 3, 5, \ldots, 99$, respectively), and the even-positioned numbers are $98, 96, 94, \ldots, 2$ (indeed, their positions are $2, 4, 6, \ldots, 98$, respectively). A move can change the position of a number, but it does not change the **parity** of this position (as we have already seen in the solution to Exercise 8.1.3). Thus, odd-positioned numbers remain odd-positioned after any number of moves. Hence, the numbers $99, 97, 95, \ldots, 1$ are not just odd-positioned at the onset, but remain odd-positioned after any number of moves. Likewise, the numbers $98, 96, 94, \ldots, 2$ remain even-positioned.

Now, let us focus entirely on the odd-positioned numbers (while forgetting about the even-positioned numbers). Using our moves, we can swap any two adjacent odd-positioned numbers[263]. From Exercise 8.2.1, we know that there is a sequence of such swaps that rearranges these odd-positioned numbers in weakly increasing order. Each of these swaps corresponds to a move in the sense of Exercise 8.2.4; moreover, these moves do not affect the even-positioned numbers. Hence, after we perform all these moves, our odd-positioned numbers have been rearranged in increasing order (so that they are now $1, 3, 5, \ldots, 97, 99$ from left to right), whereas the even-positioned numbers are still as they were at the onset (i.e., they are still $98, 96, 94, \ldots, 2$ from left to right).

Next, let us perform the analogous operations with the even-positioned numbers: Our moves allow us to swap any two adjacent even-positioned numbers, and (with the help of Exercise 8.2.1) we can find a sequence of such swaps that rearranges these even-positioned numbers in weakly increasing order without affecting the odd-positioned numbers. Thus, after we perform all these moves, our even-positioned numbers have been rearranged in increasing order (so that they are now $2, 4, 6, \ldots, 96, 98$ from left to right), whereas the odd-positioned numbers are still as they were before this second set of moves (i.e., they are still $1, 3, 5, \ldots, 97, 99$ from left to right).

Thus, in the state we have now obtained after all these moves, the odd-positioned numbers are $1, 3, 5, \ldots, 97, 99$ from left to right, and the even-positioned numbers are $2, 4, 6, \ldots, 96, 98$ from left to right. Hence, this state is $(1, 2, \ldots, 99)$. Thus, we have found a sequence of moves that transforms the state $(99, 98, \ldots, 1)$ into the state $(1, 2, \ldots, 99)$. This solves the "inverse" exercise, and thus (as we have already explained) also solves the original Exercise 8.2.4. $\square$

We note that our above solution illustrates yet another useful tool in problem

[262] i.e., if it has an even # of numbers to its left

[263] Here, "adjacent" means "adjacent, if we ignore the even-positioned numbers". That is, two odd-positioned numbers are said to be *adjacent* if there are no odd-positioned numbers between them. In terms of the original row of 99 numbers, this means that they have exactly one number between them.

solving: the tactic of *inverting* a problem. In our case, this meant interchanging the starting state and the desired goal. This is a useful tactic in its own right (and can be useful even when the moves are not reversible; in such cases one will have to define a new kind of moves, which undo the original moves). It is known as "*working backwards*" and is briefly discussed in [Engel98, §14.3].

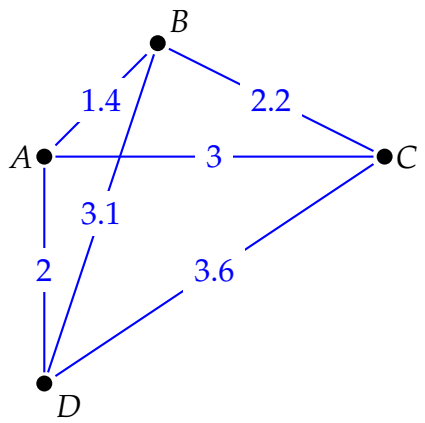
The next exercise is another quasi-geometric puzzle:

Exercise 8.2.5. Let $n \geq 2$ be an integer. A country has n towns, the distances between which are distinct.

(a) You start in a town A_1. From there you travel to the town A_2 that is farthest away from A_1. From there you travel to the town A_3 that is farthest away from A_2. You continue travelling in the same pattern. Prove the following: If $A_3 \neq A_1$, then you never come back to the town A_1.

(b) Prove the same claim if the words "farthest away from" are replaced by "closest to". (Of course, a town does not count as the closest town to itself.)

[**Example:** Assume that $n = 4$ and that the towns are the four points $A = (0,0)$, $B = (1,1)$, $C = (3,0)$ and $D = (0,-2)$ in the Euclidean plane. This is illustrated in the following picture:



(with the edges labeled by their respective lengths). Now, in Exercise 8.2.5 **(a)**, if you start in the town A, then you travel to the town C, then to D, then to C, and afterwards just keep moving around between C and D. In Exercise 8.2.5 **(b)**, if you start in the town D, then you travel to the town A, then to B, then to A, and afterwards just keep moving around between B and A.]

We notice that Exercise 8.2.5 **(b)** can be restated in terms of the cowboys from Exercise 5.2.5. Indeed, in the latter language, it says that the paths of the bullets do not form any closed polygonal path (other than "2-gons") whenever $n \geq 2$ (not necessarily odd).

Solution to Exercise 8.2.5 (sketched). **(a)** The trick is to realize that each leg of your journey is longer than the previous one, unless it leads you straight back to the town you just came from. Here are the details: We claim the following:

Claim 1: We have $|A_1A_2| \leq |A_2A_3| \leq |A_3A_4| \leq \cdots$.

Claim 2: If $A_3 \neq A_1$, then $|A_1A_2| < |A_2A_3|$.

[*Proof of Claim 1:* We must show that $|A_{i-1}A_i| \leq |A_iA_{i+1}|$ for each integer $i \geq 2$. So let $i \geq 2$ be an integer. We have $A_{i-1} \neq A_i$ (since you don't ever travel from a town directly to itself). From the way the journey $A_1A_2A_3\cdots$ was defined, we know that A_{i+1} is the town that is farthest away from A_i. Hence, $|A_iA_{i+1}| \geq |A_iX|$ for every town $X \neq A_i$. We can apply this to $X = A_{i-1}$ (since $A_{i-1} \neq A_i$), and conclude that $|A_iA_{i+1}| \geq |A_iA_{i-1}| = |A_{i-1}A_i|$. In other words, $|A_{i-1}A_i| \leq |A_iA_{i+1}|$. This is precisely what we needed to show. Thus, Claim 1 is proven.]

[*Proof of Claim 2:* Assume that $A_3 \neq A_1$. Then, the line segments A_1A_2 and A_2A_3 are distinct. Hence, $|A_1A_2| \neq |A_2A_3|$ (since we assumed that the distances between the towns are distinct). However, Claim 1 yields $|A_1A_2| \leq |A_2A_3|$. Combining this with $|A_1A_2| \neq |A_2A_3|$, we obtain $|A_1A_2| < |A_2A_3|$. Thus, Claim 2 follows.]

Now, assume that $A_3 \neq A_1$. We must prove that you never come back to the town A_1. Indeed, assume the contrary. Thus, you do come back to the town A_1 at some point. In other words, there exists an integer $i \geq 2$ such that $A_i = A_1$. Consider this i. Note that $A_i = A_1 \neq A_2$ (since you don't ever travel from a town directly to itself), and thus $i \neq 2$. Hence, $i \geq 3$ (since $i \geq 2$).

Now, Claim 1 yields $|A_1A_2| \leq |A_2A_3| \leq |A_3A_4| \leq \cdots$, so that $|A_2A_3| \leq |A_3A_4| \leq \cdots \leq |A_{i-1}A_i|$ [264]. Hence, $|A_2A_3| \leq |A_{i-1}A_i|$. But Claim 2 yields $|A_1A_2| < |A_2A_3|$. Thus,

$$|A_1A_2| < |A_2A_3| \leq |A_{i-1}A_i| = |A_iA_{i-1}| = |A_1A_{i-1}| \tag{394}$$

(since $A_i = A_1$). Also, $A_{i-1} \neq A_i$ (since you don't ever travel from a town directly to itself), and thus $A_{i-1} \neq A_i = A_1$.

However, from the way the journey $A_1A_2A_3\cdots$ was defined, we know that A_2 is the town that is farthest away from A_1. Hence, $|A_1A_2| \geq |A_1X|$ for every town $X \neq A_1$. Applying this to $X = A_{i-1}$, we obtain $|A_1A_2| \geq |A_1A_{i-1}|$ (since $A_{i-1} \neq A_1$). This contradicts (394). This contradiction shows that our assumption was wrong. Hence, Exercise 8.2.5 **(a)** is solved.

(b) Exercise 8.2.5 **(b)** can be solved by the exact same argument as Exercise 8.2.5 **(a)**, just with all inequality signs reversed (i.e., we have to replace every "$\leq$" by an "$\geq$" and likewise). □

The next exercise (which is similar in many ways to Exercise 8.2.3) models the typical failure of breakout rooms at a Zoom conference:

[264] If $i = 3$, then this chain of inequalities is a trivial chain with only one number and no inequality signs. But as you can easily see, this does not invalidate the argument that we are now going to make.

Exercise 8.2.6. Several people are in a building with several rooms. Each minute, one person leaves a room and moves to another that has more people (not counting the person who is moving). (That is: If a person was in a room with m other people, he moves to a room with more than m other people.) Prove that eventually, all of the people will end up in the same room.

[**Example:** Let us assume that the starting state is $(4,2,2)$, by which we mean that there are 4 people in one room, 2 in another, and 2 in a third room. Now, one of the two people from the second room may move into the third, resulting in the state $(4,1,3)$. Then, one person from the third room may move into the first, resulting in $(5,1,2)$. Then, the single inhabitant from the second room may move into the first, resulting in $(6,0,2)$. Two more moves later, we obtain the state $(8,0,0)$, which means that everybody is in the first room.]

Solution to Exercise 8.2.6 (sketched). A *dispersed couple* shall mean a 2-element set $\{u,v\}$ consisting of two people u and v that are in different rooms. (Of course, this notion depends on the state.) Define the *dispersion* of a state to be the # of dispersed couples in that state. (For example, the state $(4,2,2)$ in the above example has dispersion $4\cdot 2+4\cdot 2+2\cdot 2=20$.) Clearly, the dispersion of a state is a nonnegative integer.

Now, we claim the following:

Claim 1: Each minute, the dispersion of the state decreases.

[*Proof of Claim 1:* Consider what happens in a given minute. Namely, one person leaves a room and moves to another that has more people (not counting the person who is moving). Let p be this person; let M be the room that p is leaving; let N be the room that p is entering. Then, room N has more people than room M (not counting p), because p is moving to a room with more people.

Let m be the # of people in room M (not counting p), and let n be the # of people in room N (not counting p). Then, $n>m$ (since room N has more people than room M), so that $n-m>0$. It is now easy to see that this move decreases the dispersion of the state by $n-m$ (since n dispersed couples disappear[265], and m new dispersed couples appear[266]). Since $n-m>0$, this entails that the dispersion of the state decreases. Thus, Claim 1 is proven.]

The rest of the solution is now essentially a repetition of the "infinite descent" argument we have seen in the discussion of Exercise 5.3.1 (and again in the solution to Exercise 8.2.3). Namely, Claim 1 shows that the dispersion of the state decreases each minute. But the dispersion of any state is a nonnegative integer, and thus cannot decrease indefinitely. More concretely: If the dispersion of the original state was N, then the dispersion cannot decrease more than N times without becoming negative. Thus, the moves cannot go on for more than N minutes.

[265] Namely, the n couples of the form $\{p,x\}$ with $x\in N$ were dispersed before the move but are no longer dispersed after the move.

[266] Namely, the m couples of the form $\{p,x\}$ with $x\in M$ were not dispersed before the move but become dispersed after the move.

Hence, after at most N moves, we must arrive at a state where no person can move any more. But what does it mean that no person can move? If there is any dispersed couple $\{p, q\}$, then at least one of the two people p and q can move[267]. Hence, if no person can move, then there is no dispersed couple, and therefore all of the people must be in the same room. Thus, after at most N moves, we must arrive at a state where all of the people are in the same room (because we know that after at most N moves, we must arrive at a state where no person can move any more). This solves Exercise 8.2.6. $\square$

Next, a more advanced monovariant problem:

Exercise 8.2.7. Consider a hotel with an infinite number of rooms, arranged sequentially on the ground floor. The rooms are labelled (from left to right) by integers $i \in \mathbb{Z}$, with room i being adjacent to rooms $i-1$ and $i+1$. (To be specific, room $i-1$ is left of room i, while room $i+1$ is right of room i.) Thus, the hotel looks as follows:

$\cdots$							$\cdots$
	$\uparrow$	$\uparrow$	$\uparrow$	$\uparrow$	$\uparrow$	$\uparrow$	
	room -2	room -1	room 0	room 1	room 2	room 3	

(where the boxes of the table correspond to the rooms; of course, we are only showing a small part of the actually infinite hotel).

A finite number of violinists are staying in the hotel; each room has at most one violinist in it. Each night, some two violinists staying in adjacent rooms (if two such violinists exist) decide they cannot stand each other's noise, and move apart: One of them moves to the nearest unoccupied room to the left, while the other moves to the nearest unoccupied room to the right. (The two violinists move simultaneously, so one moving away does not prevent the other from moving. Only two violinists move in any given night.) This keeps happening for as long as there are two violinists in adjacent rooms.

Prove that this moving will stop after a finite number of days (i.e., there will be a day when no two violinists are in adjacent rooms any more).

[**Example:** Let us assume that initially, there are 5 violinists (named a, b, c, d, e), arranged as follows:

$\cdots$			a	b		c	d		e			$\cdots$

(395)

[267] *Proof:* Assume that there is a dispersed couple $\{p, q\}$. We claim that at least one of the two people p and q can move. Indeed, we WLOG assume that the room containing p has at least as many people in it as the room containing q (otherwise, we can just interchange p with q). Then, the room containing p has (strictly) more people than the room containing q **if we don't count** q (because not counting q decreases the head count of the latter room by 1). Therefore, q can move from the latter room into the former room (according to the rules of the game). Thus, at least one of the two people p and q can move (in this case, q). Qed.

(where the boxes of the table correspond to the rooms; of course, we are only showing a small part of the actually infinite hotel). On the first night, let's say the violinists a and b tire of each other's presence and move apart; this results in the following state:

$\cdots$		a			b	c	d		e			$\cdots$

Now, let's say that violinists b and c move apart on the next night, resulting in

$\cdots$		a		b			d	c	e			$\cdots$

On the next night, violinists c and e decide to move apart:

$\cdots$		a		b		c	d			e		$\cdots$

On the next night, violinists c and d decide to move apart:

$\cdots$		a		b	c			d		e		$\cdots$

On the next night, violinists b and c decide to move apart:

$\cdots$		a	b			c		d		e		$\cdots$

On the next night, violinists a and b decide to move apart:

$\cdots$	a			b		c		d		e		$\cdots$

Now, there are no more neighbors – so no further moves can happen.

(Of course, there were choices involved in this process; we have only shown one possibility. Other pairs of neighbors may have moved apart instead, and the resulting state could well be different. But the exercise claims that whatever way the violinists move, the moving will come to an end eventually.)]

Solution to Exercise 8.2.7 (sketched). Note that each room has at most one violinist staying in it in the initial state. This property remains valid in all later states as well, since violinists can only move into unoccupied rooms.

Let us regard the violinists as indistinguishable (i.e., we don't care which violinist is in which room, but only care about which rooms are occupied).[268]

Thus, we can identify any state with the finite subset $\{i \in \mathbb{Z} \mid \text{room } i \text{ is occupied}\}$ of $\mathbb{Z}$ (since each room has at most one violinist staying in it). Let us make this identification; thus, states are finite subsets of $\mathbb{Z}$. For example, the state shown in (395) is the subset $\{3, 4, 6, 7, 9\}$, provided that the rooms shown in (395) are the rooms

[268] Indeed, the identities of the violinists are clearly immaterial to the problem at hand.

$1, 2, \ldots, 11$.

Thus, any move removes two consecutive integers j and $j+1$ from the state and inserts two new integers into the state (one of which is $< j$ while the other is $> j+1$). More precisely: If S is a state and j is an integer such that both j and $j+1$ belong to S, then there is a move that transforms S into the state

$$\begin{aligned}&(S \setminus \{j, j+1\}) \\ &\qquad \cup \{\text{the largest integer } < j \text{ that does not belong to } S\} \\ &\qquad \cup \{\text{the smallest integer } > j+1 \text{ that does not belong to } S\}.\end{aligned}$$

We define the *entropy* of a state S to be the sum $\sum\limits_{i \in S} 2^i$ (recalling that S is a finite subset of $\mathbb{Z}$). This is a rational number (since the sum is finite).

It is easy to see that this entropy increases each night. In other words:

> *Claim 1:* Whenever two violinists move apart, the entropy of the state increases.

[*Proof of Claim 1:* Consider two violinists moving apart. Let j and $j+1$ be the rooms where they live before the move, and let p and q be the rooms where they live after the move, labelled in such a way that $p < j$ and $q > j+1$. Then, $q > j+1$, so that $q \geq j+2$ (since q and $j+1$ are integers) and thus

$$2^q \geq 2^{j+2} = 2 \cdot 2^{j+1} = \underbrace{2^{j+1}}_{>2^j} + 2^{j+1} > 2^j + 2^{j+1}$$

and therefore $2^p + 2^q > 2^q > 2^j + 2^{j+1}$. In other words, $2^p + 2^q - 2^j - 2^{j+1} > 0$.

Now, recall that the entropy of a state S was defined as the sum $\sum\limits_{i \in S} 2^i$. The move that we are currently considering removes the two addends 2^j and 2^{j+1} from this sum, and inserts the two addends 2^p and 2^q in their stead. Thus, this move increases this sum by $2^p + 2^q - 2^j - 2^{j+1}$. Since $2^p + 2^q - 2^j - 2^{j+1} > 0$, this shows that this move increases this sum. In other words, this move increases the entropy. In other words, the entropy after the move is greater than the entropy before the move. This proves Claim 1.]

This was a good step towards the solution, but we have still a long way to go. In theory, there are infinitely many states possible (after all, a state is just a finite subset of $\mathbb{Z}$, and there are infinitely many of those), so the entropy could keep increasing indefinitely without the moves ever coming to an end. If we could somehow show that only finitely many possible states can be reached, then Claim 1 would show that the moving must come to an end (because the set of possible entropies would then also be finite, and thus the entropy could not keep increasing forever).

Fortunately, we can indeed prove this. To do so, we let N be the # of violinists in the hotel. (Clearly, this # does not change during the process.) We shall show that

the violinists never move "too far" away from their original rooms; more precisely, they will always stay within $3N$ rooms of the interval between the leftmost and the rightmost room occupied in the initial state. We shall now make this precise.

Let S_0 be the initial state. Consider a sequence of g (successive) moves, starting from state S_0 and leading to states $S_1, S_2, S_3, \ldots, S_g$ in this order (i.e., the first move transforms state S_0 into state S_1; the next move transforms state S_1 into state S_2; and so on). Thus, in any one of the states $S_0, S_1, \ldots, S_g$, there are exactly N violinists in the hotel (since the # of violinists in the hotel never changes, and always equals N), and thus exactly N occupied rooms (since each occupied room has exactly one violinist in it). In other words, we have

$$N = |S_0| = |S_1| = \cdots = \left|S_g\right|. \tag{396}$$

We WLOG assume that $N > 0$ (since otherwise, there are no violinists at all in the hotel, and the exercise becomes trivial). Thus, the sets $S_0, S_1, \ldots, S_g$ are nonempty finite sets (because (396) shows that each of them has N elements); therefore, their minima and maxima are well-defined. Let $\alpha = \min(S_0)$ and $\omega = \max(S_0)$.

Now, we shall see that the violinists don't spread "too fast" through the hotel. To be more specific, we claim the following:

Claim 2: We have $\min(S_{i+1}) \geq \min(S_i) - 1$ for each $i \in \{0, 1, \ldots, g-1\}$.

[*Proof of Claim 2:* Let $i \in \{0, 1, \ldots, g-1\}$. We must show that $\min(S_{i+1}) \geq \min(S_i) - 1$.

Indeed, assume the contrary. Thus, $\min(S_{i+1}) < \min(S_i) - 1$. Hence, $\min(S_{i+1}) + 1 < \min(S_i)$, so that $\min(S_i) > \min(S_{i+1}) + 1$.

The room $\min(S_{i+1})$ is occupied in state S_{i+1} (since $\min(S_{i+1}) \in S_{i+1}$). However, all rooms[269] $k < \min(S_i)$ are unoccupied in state S_i (since $\min(S_i)$ is the **smallest** element of S_i). Thus, in particular, the room $\min(S_{i+1})$ is unoccupied in state S_i (since $\min(S_{i+1}) < \min(S_i) - 1 < \min(S_i)$). Thus, we have shown that the room $\min(S_{i+1})$ is occupied in state S_{i+1} but not in state S_i. Therefore, the move that transforms state S_i into state S_{i+1} must place a violinist in room $\min(S_{i+1})$. Let us denote this violinist by p, and let us refer to this move as "the move $S_i \to S_{i+1}$". Let r be the room which violinist p occupies in state S_i (that is, before this move). Then, $r \geq \min(S_i)$ (since all rooms $k < \min(S_i)$ are unoccupied in state S_i), so that $r \geq \min(S_i) > \min(S_i) - 1 > \min(S_{i+1})$. Hence, in the move $S_i \to S_{i+1}$, the violinist p moves left (since p moves from room r to room $\min(S_{i+1})$).

Recall again that all rooms $k < \min(S_i)$ are unoccupied in state S_i. Hence, the room $\min(S_{i+1}) + 1$ is unoccupied in state S_i (because $\min(S_{i+1}) + 1 < \min(S_i)$). Moreover, $\min(S_{i+1}) < \min(S_{i+1}) + 1 < r$ (since $r \geq \min(S_i) > \min(S_{i+1}) + 1$). In other words, the room $\min(S_{i+1}) + 1$ lies strictly between the two rooms $\min(S_{i+1})$ and r.

[269] We are identifying each integer with the corresponding room of our hotel. Thus, "rooms $k < \min(S_i)$" means "rooms k to the left of room $\min(S_i)$".

Now, recall the nature of moves: When two violinists move apart, only one of them moves left, and this violinist moves to the nearest unoccupied room to its left. Thus, in the move $S_i \to S_{i+1}$, the violinist p moves to the **nearest** unoccupied room to the left of room r (since he moves left, and since he is originally in room r). This **nearest** unoccupied room, however, cannot be the room $\min(S_{i+1})$, because the room $\min(S_{i+1}) + 1$ is also unoccupied in state S_i and lies strictly between the two rooms $\min(S_{i+1})$ and r (so that it is also to the left of room r but is nearer to r than $\min(S_{i+1})$). Hence, in the move $S_i \to S_{i+1}$, the violinist p cannot move to the room $\min(S_{i+1})$. But this contradicts the fact that (as we have seen above) the move $S_i \to S_{i+1}$ does place the violinist p in room $\min(S_{i+1})$. This contradiction shows that our assumption was wrong. Hence, $\min(S_{i+1}) \geq \min(S_i) - 1$ is proved. This proves Claim 2.]

We next introduce another notation. If S is any state (i.e., any finite subset of $\mathbb{Z}$), then we define the subset S^{-0+} of $\mathbb{Z}$ by

$$\begin{aligned} S^{-0+} &= S \cup \{s+1 \mid s \in S\} \cup \{s-1 \mid s \in S\} \\ &= \{i \in \mathbb{Z} \mid i \in S \text{ or } i-1 \in S \text{ or } i+1 \in S\}. \end{aligned}$$

In other words, S^{-0+} is the set of all integers i that belong to S themselves or have a neighbor (that is, $i-1$ or $i+1$) belong to S. If we identify each integer with the corresponding room in our hotel, then we can restate this as follows: S^{-0+} is the set of all rooms that are occupied or adjacent to an occupied room (assuming that S is the set of all occupied rooms)[270]. Clearly, for any state S, we have

$$\left|S^{-0+}\right| \leq 3 \cdot |S| \tag{397}$$

(this follows easily from (277)) and

$$S \subseteq S^{-0+}. \tag{398}$$

Our next claim shows that if at least one of three consecutive rooms is occupied at some point, then this property will remain valid in all future states (even though it will not always be the same room that is occupied, or even the same # of rooms):

Claim 3: We have $S_i^{-0+} \subseteq S_{i+1}^{-0+}$ for each $i \in \{0, 1, \ldots, g-1\}$.

[*Proof of Claim 3:* Let $i \in \{0, 1, \ldots, g-1\}$. We must show that $S_i^{-0+} \subseteq S_{i+1}^{-0+}$. In other words, we must show that every $k \in S_i^{-0+}$ satisfies $k \in S_{i+1}^{-0+}$. So, let us consider any $k \in S_i^{-0+}$. We shall show that $k \in S_{i+1}^{-0+}$.

We have $k \in S_i^{-0+}$. In other words, at least one of the three numbers $k-1$, k and $k+1$ belongs to S_i. In other words, at least one of the three rooms $k-1$, k and $k+1$ is occupied in state S_i.

[270] For example, if $S = \{2, 4, 8, 9\}$, then $S^{-0+} = \{2, 4, 8, 9\} \cup \{3, 5, 9, 10\} \cup \{1, 3, 7, 8\} = \{1, 2, 3, 4, 5, 7, 8, 9, 10\}$.

We must show that $k \in S_{i+1}^{-0+}$. In other words, we must show that at least one of the three numbers $k-1$, k and $k+1$ belongs to S_{i+1}. In other words, we must show that at least one of the three rooms $k-1$, k and $k+1$ is occupied in state S_{i+1}.

The state S_{i+1} is obtained from the state S_i by a move. Let us consider all possibilities for how this move can affect rooms $k-1$, k and $k+1$ (recalling that at least one of these three rooms must be occupied in state S_i):

- If there is a violinist in room k before the move but no violinist in room $k-1$, then there will be a violinist in (at least) one of the rooms $k-1$ and k after the move. (Indeed, the move will either leave the violinist in room k in place, or (if it does displace this violinist) it will cause a violinist to appear in room $k-1$.)

- If there is a violinist in room k before the move but no violinist in room $k+1$, then there will be a violinist in (at least) one of the rooms $k+1$ and k after the move. (Indeed, the move will either leave the violinist in room k in place, or (if it does displace this violinist) it will cause a violinist to appear in room $k+1$.)

- If there are violinists in all three rooms $k-1$, k and $k+1$ before the move, then at least one of these three violinists will remain in his room after the move (since a move only displaces two violinists).

- If there is a violinist in room $k-1$ before the move but no violinist in room k, then there will be a violinist in (at least) one of the rooms $k-1$ and k after the move. (Indeed, the move will either leave the violinist in room $k-1$ in place, or (if it does displace this violinist) it will cause a violinist to appear in room k.)

- If there is a violinist in room $k+1$ before the move but no violinist in room k, then there will be a violinist in (at least) one of the rooms $k+1$ and k after the move. (Indeed, the move will either leave the violinist in room $k+1$ in place, or (if it does displace this violinist) it will cause a violinist to appear in room k.)

These five possibilities cover all the possible cases (since we know that at least one of the three rooms $k-1$, k and $k+1$ is occupied in state S_i). Thus, in each possible case, we have shown that there is a violinist in at least one of the three rooms $k-1$, k and $k+1$ after the move. In other words, at least one of the three rooms $k-1$, k and $k+1$ is occupied in state S_{i+1}. As we explained above, this completes the proof of Claim 3.]

Let us draw some conclusions from Claim 3. It is **not** true that $S_0 \subseteq S_1 \subseteq \cdots \subseteq S_g$, since an occupied room can become unoccupied after a move. However, Claim 3 shows that we have

$$S_0^{-0+} \subseteq S_1^{-0+} \subseteq \cdots \subseteq S_g^{-0+}. \tag{399}$$

In words, this is saying that if a room is occupied or adjacent to an occupied room, then it will always remain occupied or adjacent to an occupied room.

In the rest of this solution, we shall use the notation $[p, q]$ for an **integer** interval. That is, if p and q are two integers, then the notation $[p, q]$ shall denote the set $\{p, p+1, \ldots, q\} = \{i \in \mathbb{Z} \mid p \leq i \leq q\}$ (rather than, as it commonly does, the real interval $\{i \in \mathbb{R} \mid p \leq i \leq q\}$). This will be convenient, since we will have to deal with integer intervals rather than real intervals in this solution.

Our next claim will show that the leftmost occupied room will always remain to the right of the room $\alpha - 3N$ (so it cannot "wander off" to the left too far):

Claim 4: Let $m \in \{0, 1, \ldots, g\}$. Then, $\min(S_m) > \alpha - 3N$.

[*Proof of Claim 4:* Before I prove this formally, let me sketch what is going on (as the formal proof does a good job of obscuring this). Consider how the state evolves from S_0 to S_m. With every move $S_i \to S_{i+1}$, the smallest element of the state (i.e., the leftmost occupied room) shifts to the left by at most 1 (because of Claim 2). For the initial state S_0, this smallest element is α (since $\alpha = \min(S_0)$). Hence, if the smallest element of S_m is some integer $\beta \leq \alpha$ (that is, we have $\beta = \min(S_m) \leq \alpha$), then each integer $j \in [\beta, \alpha]$ must be the smallest element of some intermediate state between S_0 and S_m (since otherwise, the smallest element would have "jumped over" j, which is impossible given that it only shifts to the left by at most 1 at each step). Hence, each integer $j \in [\beta, \alpha]$ must be occupied in one of these intermediate states. In other words, each integer $j \in [\beta, \alpha]$ belongs to some S_k (with $k \in \{0, 1, \ldots, m\}$), and therefore also to S_k^{-0+} (since $S_k \subseteq S_k^{-0+}$), and therefore also to S_m^{-0+} (since (399) yields $S_k^{-0+} \subseteq S_m^{-0+}$). This yields a lower bound on the size of S_m^{-0+}: Namely, it yields $\left|S_m^{-0+}\right| \geq \alpha - \beta + 1$. If $\beta \leq \alpha - 3N$, then this lower bound entails $\left|S_m^{-0+}\right| \geq 3N + 1 > 3N$, which contradicts the easily established fact that $\left|S_m^{-0+}\right| \leq 3 \cdot \underbrace{|S_m|}_{=N} = 3N$. Hence, we must have $\beta > \alpha - 3N$, and this is precisely what Claim 4 claims.

Here is the proof in all its boring detail. We must show that $\min(S_m) > \alpha - 3N$.

Assume the contrary. Thus, $\min(S_m) \leq \alpha - 3\underbrace{N}_{>0} < \alpha = \min(S_0)$.

Now, let $j \in [\min(S_m), \min(S_0)]$ be arbitrary. We shall show that $j \in S_m^{-0+}$.

Indeed, we have $j \in [\min(S_m), \min(S_0)]$. In other words, $j \in \mathbb{Z}$ and $\min(S_m) \leq j \leq \min(S_0)$.

There exists some $k \in \{0, 1, \ldots, m\}$ satisfying $\min(S_k) \leq j$ (for example, $k = m$ will work, since $\min(S_m) \leq j$). Consider the **smallest** such k. Then, $\min(S_k) \leq j$; however, if $k \neq 0$, then $\min(S_{k-1}) > j$ (since otherwise, k would not be the **smallest** element of $\{0, 1, \ldots, m\}$ to satisfy $\min(S_k) \leq j$). From this, it is easy to see that $\min(S_k) = j$. Indeed, assume the contrary. Thus, $\min(S_k) \neq j$, so that $\min(S_k) < j$ (since $\min(S_k) \leq j$) and therefore $k \neq 0$ (since otherwise, we would have $k = 0$, so that $\min(S_k) = \min(S_0) \geq j$, which would contradict $\min(S_k) < j$). Hence, $\min(S_{k-1}) > j$ (because we said that if $k \neq 0$, then $\min(S_{k-1}) > j$), so

that $\min(S_{k-1}) \geq j+1$ (since $\min(S_{k-1})$ and j are integers). But $k \in \{1,2,\ldots,m\}$ (since $k \in \{0,1,\ldots,m\}$ and $k \neq 0$), so that $k-1 \in \{0,1,\ldots,m-1\}$. Hence, Claim 2 (applied to $i = k-1$) yields $\min(S_k) \geq \min(S_{k-1}) - 1 \geq j$ (since $\min(S_{k-1}) \geq j+1$). This contradicts $\min(S_k) < j$. This contradiction shows that our assumption was false. Hence, we obtain $\min(S_k) = j$. Thus, $j = \min(S_k) \in S_k \subseteq S_k^{-0+}$ (by (398)).

However, (399) yields $S_k^{-0+} \subseteq S_m^{-0+}$ (since $k \in \{0,1,\ldots,m\}$). Hence, $j \in S_k^{-0+} \subseteq S_m^{-0+}$.

Now, forget that we fixed j. We thus have proved that $j \in S_m^{-0+}$ for each $j \in [\min(S_m), \min(S_0)]$. In other words, $[\min(S_m), \min(S_0)] \subseteq S_m^{-0+}$. Hence,

$$\begin{aligned} |[\min(S_m), \min(S_0)]| &\leq \left|S_m^{-0+}\right| \leq 3 \cdot \underbrace{|S_m|}_{\substack{=N \\ \text{(by (396))}}} \qquad \text{(by (397))} \\ &= 3N, \end{aligned}$$

so that

$$\begin{aligned} 3N &\geq \left| \left[\min(S_m), \underbrace{\min(S_0)}_{=\alpha} \right] \right| = |[\min(S_m), \alpha]| \\ &= \alpha - \min(S_m) + 1 \qquad (\text{since } \min(S_m) \leq \alpha) \\ &> \alpha - \min(S_m). \end{aligned}$$

In other words, $\min(S_m) > \alpha - 3N$. This proves Claim 4.]

Claim 5: Let $m \in \{0,1,\ldots,g\}$. Then, $\max(S_m) < \omega + 3N$.

[*Proof of Claim 5:* Claim 5 is an analogue of Claim 4: While Claim 4 says that $\min(S_m)$ cannot stray too far leftwards from $\min(S_0)$, Claim 5 says that $\max(S_m)$ cannot stray too far rightwards from $\max(S_0)$. The proofs are analogous, too; our above proof of Claim 4 thus can be easily transformed into a proof of Claim 5 (mutatis mutandis – e.g., all the signs need to be reversed, and similar changes).]

Now, Claim 4 and Claim 5 can be combined to the following:

Claim 6: Let $m \in \{0,1,\ldots,g\}$. Then, $S_m \subseteq [\alpha - 3N + 1, \omega + 3N - 1]$.

[*Proof of Claim 6:* Let $j \in S_m$. Thus, $j \geq \min(S_m) > \alpha - 3N$ (by Claim 4) and $j \leq \max(S_m) < \omega + 3N$ (by Claim 5). Combining these two inequalities, we obtain $j \in [\alpha - 3N + 1, \omega + 3N - 1]$ (since j is an integer).

Forget that we fixed j. We thus have shown that $j \in [\alpha - 3N + 1, \omega + 3N - 1]$ for each $j \in S_m$. In other words, $S_m \subseteq [\alpha - 3N + 1, \omega + 3N - 1]$. This proves Claim 6.]

We note that $\omega = \max(S_0) \geq \min(S_0) = \alpha$ and thus $\underbrace{\omega}_{\geq\alpha} + \underbrace{3N-1}_{\substack{>-3N+1 \\ \text{(since } N>0)}} > \alpha - 3N+1$. Hence, the interval $[\alpha - 3N+1, \omega + 3N - 1]$ is nonempty, and its size is

$$|[\alpha - 3N+1, \omega + 3N-1]| = (\omega + 3N - 1) - (\alpha - 3N + 1) + 1 = \omega - \alpha + 6N - 1.$$

Now, we can use (e.g.) the pigeonhole principle: Claim 1 entails that the entropy of the state increases with every move. Thus,

$$(\text{the entropy of } S_0) < (\text{the entropy of } S_1) < \cdots < \left(\text{the entropy of } S_g\right).$$

Hence, the entropies of the $g+1$ states $S_0, S_1, \ldots, S_g$ are distinct. Therefore, these $g+1$ states $S_0, S_1, \ldots, S_g$ must themselves be distinct. However, all these $g+1$ states $S_0, S_1, \ldots, S_g$ are subsets of the set $[\alpha - 3N+1, \omega + 3N - 1]$ (by Claim 6). Hence, we have found $g+1$ distinct subsets of the set $[\alpha - 3N+1, \omega + 3N - 1]$. By the pigeonhole principle, this entails that

$$\begin{aligned} g+1 &\leq (\text{\# of all subsets of } [\alpha - 3N+1, \omega + 3N - 1]) \\ &= 2^{|[\alpha-3N+1, \omega+3N-1]|} = 2^{\omega - \alpha + 6N - 1} \\ &\qquad (\text{since } |[\alpha - 3N+1, \omega + 3N - 1]| = \omega - \alpha + 6N - 1). \end{aligned}$$

In other words, $g \leq 2^{\omega-\alpha+6N-1} - 1$.

Now, forget that we fixed g and $S_1, S_2, \ldots, S_g$. We thus have shown that any sequence of g (successive) moves, starting from state S_0, must satisfy $g \leq 2^{\omega-\alpha+6N-1} - 1$. In other words, there is no sequence of (successive) moves, starting from state S_0, that has more than $2^{\omega-\alpha+6N-1} - 1$ moves. In other words, the moves cannot go on for more than $2^{\omega-\alpha+6N-1}$ nights (if the initial state is S_0). Hence, the moving will stop after a finite number of days. This solves Exercise 8.2.7. □

The next exercise (a slight generalization of [Engel98, Chapter 1, Example E4]) can be solved using a monovariant, even though no process or movement is inherently present in the exercise. The trick here is to find a process first:

Exercise 8.2.8. Let $m \in \mathbb{N}$. Each member of a parliament has **at most** m enemies. (The notion of an "enemy" is mutual: If A is an enemy of B, then B is an enemy of A.) Prove that it is possible to subdivide the parliament into two houses (with each member going into exactly one house)[271] such that each member has **at most** $\frac{m}{2}$ enemies in his own house.

Solution to Exercise 8.2.8 (sketched). We construct a desired subdivision algorithmically: Start with **some** arbitrary subdivision of the parliament into two houses. A member of the parliament will be called *unhappy* if he has more than $\frac{m}{2}$ enemies

[271] For the pedants: A house can be empty.

in his own house. As long as an unhappy member exists, we can pick an unhappy member and move him into the other house. We keep doing these moves until no more unhappy members exist. (We make these moves sequentially; we don't move several unhappy members simultaneously.)

I claim that these moves cannot go on forever (and thus, at some point, we will end up with no more unhappy members). Indeed, I define a *discordant couple* to be a two-element set $\{u, v\}$ of two parliament members that are mutual enemies but are in the same house. (Of course, this concept depends on the state.) I define the *discord* of a state to be the # of discordant couples in that state. Now, the discord of a state is always a nonnegative integer. Moreover, each of our moves decreases the discord (this is easy to check: if a member is unhappy, then he has more than $\frac{m}{2}$ enemies in his own house, and thus fewer than $\frac{m}{2}$ enemies in the other house[272]; the move thus decreases his contribution to the discord[273]). Hence, by the same kind of argument as in the above solution to Exercise 8.2.3, we see that the moves cannot go on forever (since the discord is a nonnegative integer and thus cannot keep decreasing forever). Hence, eventually, the moves will have stopped. In the resulting subdivision of the parliament, no member will be unhappy (since an unhappy member would enable a further move), which means that each member has **at most** $\frac{m}{2}$ enemies in his own house. Thus, Exercise 8.2.8 is solved. □

Remark 8.2.3. Exercise 8.2.8 would be false if the notion of an "enemy" was not mutual. For instance, if we take $m = 3$, and if $0, 1, 2, 3, 4$ are five parliament members with each $i \in \{0, 1, 2, 3, 4\}$ having enemies $(i+1)\,\%5$, $(i+2)\,\%5$ and $(i+3)\,\%5$, then there is no way to prevent one of them from having more than $\frac{3}{2}$ enemies in his own house, no matter how the parliament is subdivided into two houses.

8.3. Homework set #8

This problem set is **not to be handed in**.

This homework set covers the above parts of Chapter 8 as well as other chapters and unrelated ideas.[274]

[272] since he has at most m enemies in total

[273] In more detail: The member that is being moved belongs to more than $\frac{m}{2}$ discordant couples before the move, but belongs to fewer than $\frac{m}{2}$ discordant couples after the move. Thus, the total number of discordant couples that contain this member decreases when we make the move. Since the move does not change the discordant couples that don't contain this member, we can thus conclude that the move decreases the total number of discordant couples, i.e., the discord.

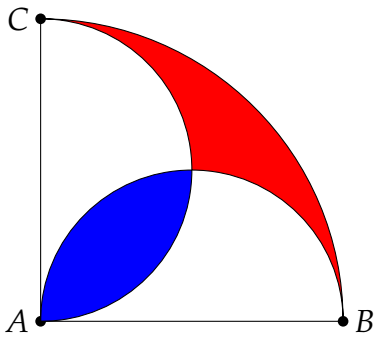
[274] Recall Definition 3.5.3, Convention 7.2.1 and Convention 7.2.2.

Exercise 8.3.1. Let n be a positive integer. Let Z be the set of all pairs $(x,y) \in [n]^2$ satisfying $x \perp y$ and $x+y>n$. (For example, if $n=5$, then

$$Z = \{(1,5),(2,5),(3,4),(3,5),(4,3),(4,5),(5,1),(5,2),(5,3),(5,4)\}.\ )$$

Find $\sum\limits_{(x,y)\in Z} \dfrac{1}{xy}$.

Exercise 8.3.2. **(a)** Consider the following picture:



.

Compare the area of the red region with that of the blue region.

(Here, ABC is an isosceles right-angled triangle with right angle at A. We have erected semicircles with diameters AB and AC, both pointing into the inside of triangle ABC. We have also drawn a quarter-circle with center A and radius $AB = AC$ bordered by the points B and C. The red region is formed by removing the two semicircles from the quarter-circle. The blue region is the (set-theoretical) intersection of the two semicircles.)

(b) Let $n \in \mathbb{N}$. Let σ be a permutation of $[n]$. Let $j \in \{0,1,\ldots,n\}$. Prove that

$$\begin{aligned}&(\#\text{ of all } i \in [n] \text{ satisfying } i \geq j > \sigma(i)) \\ &= (\#\text{ of all } i \in [n] \text{ satisfying } \sigma(i) \geq j > i).\end{aligned}$$

The next exercise uses a bit of analysis:

Exercise 8.3.3. What number is larger: e^{π} or π^{e} ? Here, π is the number pi (approximately 3.142), while e is Euler's number (approximately 2.718).

(You don't need a calculator.)

Exercise 8.3.4. Consider an $n \times m$-rectangle subdivided into 1×1-squares in the usual way (like a chessboard). Originally, fewer than $\dfrac{n+m}{2}$ of its nm squares are *infected*. Every minute, each square that has at least 2 infected neighbors becomes infected. (A *neighbor* of a square is a square that has an edge in common with

it. A corner in common does not suffice.) Show that at least one square remains uninfected no matter how long you wait.

[**Example:** If $n = 3$ and $m = 5$, then we may have the following initial state (where "X" stands for an infected square):

	X			
		X		
X				

.

Then, after one minute, we will have

	X	X		
	X	X		
X				

.

After one more minute, we will have

	X	X		
X	X	X		
X	X			

.

After one more minute, we will have

X	X	X		
X	X	X		
X	X	X		

.

There will be no more new infections after that.]

For the next exercise, imagine n students coming to a party; each student, on arrival, leaves his coat on the coatrack. When the party ends, the students leave one by one, each taking a (uniformly) random coat back from the coatrack. (We assume that no student leaves before all students have entered.) The probability that each student gets his original coat back is $\frac{1}{n!}$. What is the probability that **no** student gets his original coat back? The somewhat surprising answer is "approximately $\frac{1}{e}$, where e is Euler's number" (and the approximation becomes the better the larger n is). The following exercise makes this precise:

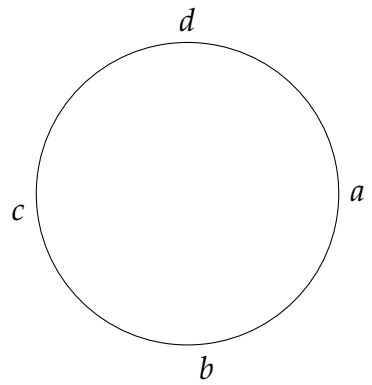
Exercise 8.3.5. Let $n \in \mathbb{N}$. A *derangement* of $[n]$ means a permutation of $[n]$ that has no fixed points (i.e., a permutation $\sigma \in S_n$ that satisfies $\sigma(i) \neq i$ for each $i \in [n]$). Let D_n denote the # of derangements of $[n]$.

(a) Prove that $D_n = \sum_{k=0}^{n} (-1)^k \frac{n!}{k!}$ for each $n \in \mathbb{N}$.

(b) Set $\operatorname{round}(x) = \left\lfloor x + \frac{1}{2} \right\rfloor$ for each $x \in \mathbb{R}$. Prove that

$$D_n = \operatorname{round}\left(\frac{n!}{e}\right) \qquad \text{for all } n \geq 1.$$

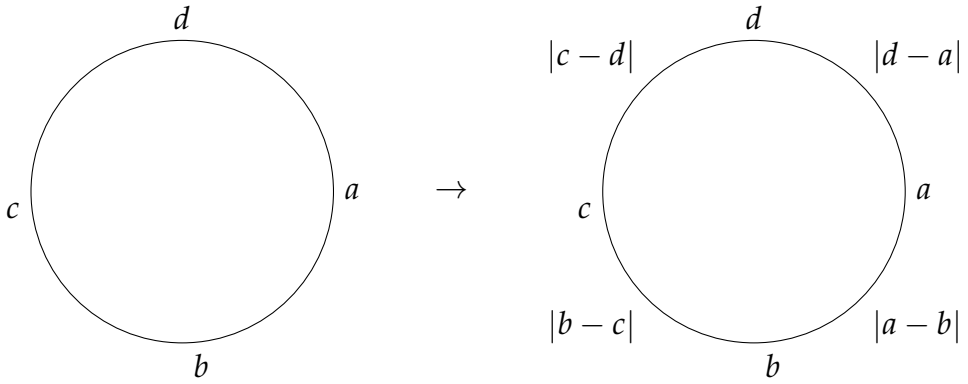
The next exercise is about the "four-numbers game". Assume you have four integers a, b, c, d written (in counterclockwise order) on the circumference of a circle:

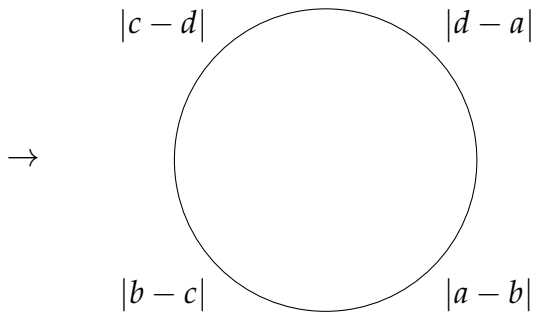


.

These four integers subdivide the circle into four arcs. To make a *difference move* means to replace these four integers by their consecutive (absolute-value) differences (i.e., on each of the four arcs of the circle, we write the absolute value of the difference between the two integers placed on the endpoints of the arc; then, we

erase the original integers). In other words, here is what a difference move does:





.

The result of a difference move is again a placement of four integers on a circle; thus, we can iterate difference moves again and again. Some experimenting suggests the following conjectures:

- After four difference moves, all four integers on the circle will be even.
- After $4k$ difference moves (where $k \in \mathbb{N}$), all four integers on the circle will be divisible by 2^k.
- If the four integers on the circle are nonnegative, then a difference move cannot increase the largest of them.
- No matter what four integers we start with, we will always end up with $0, 0, 0, 0$ if we make sufficiently many difference moves.

The following exercise (which redefines a difference move as a map D on the set of 4-tuples of integers) states these conjectures in a rigorous fashion and asks you to prove them:

Exercise 8.3.6. Consider the set $\mathbb{Z}^4$ of all 4-tuples of integers. Let $D : \mathbb{Z}^4 \to \mathbb{Z}^4$ be the map that sends each 4-tuple $(a, b, c, d) \in \mathbb{Z}^4$ to the 4-tuple

$$(|a - b|, |b - c|, |c - d|, |d - a|) \in \mathbb{Z}^4.$$

Prove the following:

(a) If $\mathbf{a} \in \mathbb{Z}^4$ is any 4-tuple, then all entries of the 4-tuple $D^4(\mathbf{a})$ are even.

(b) If $\mathbf{a} \in \mathbb{Z}^4$ is any 4-tuple, and if $k \in \mathbb{N}$, then all entries of the 4-tuple $D^{4k}(\mathbf{a})$ are divisible by 2^k.

(c) If $\mathbf{a} \in \mathbb{Z}^4$ is any 4-tuple whose all four entries are nonnegative, then

$$\max\left(D^0(\mathbf{a})\right) \geq \max\left(D^1(\mathbf{a})\right) \geq \max\left(D^2(\mathbf{a})\right) \geq \cdots.$$

Here, $\max \mathbf{b}$ denotes the largest entry of a 4-tuple $\mathbf{b} \in \mathbb{Z}^4$.

(d) If $\mathbf{a} \in \mathbb{Z}^4$ is any 4-tuple, then there exists some $k \in \mathbb{N}$ such that $D^k(\mathbf{a}) = (0,0,0,0)$.

The claim of Exercise 8.3.6 **(d)** cries out for generalization; the next exercise shows that this is not as easy as it may seem:

Exercise 8.3.7. **(a)** Prove that Exercise 8.3.6 **(d)** may fail if we replace $\mathbb{Z}$ by $\mathbb{R}$. That is, if we define a map $D : \mathbb{R}^4 \to \mathbb{R}^4$ in the same way as in Exercise 8.3.6 but using $\mathbb{R}$ instead of $\mathbb{Z}$, then there exists a 4-tuple $\mathbf{a} \in \mathbb{R}^4$ such that no $k \in \mathbb{N}$ satisfies $D^k(\mathbf{a}) = (0,0,0,0)$.

(b) Prove that the analogue of Exercise 8.3.6 **(d)** for 3-tuples instead of 4-tuples (with the map D sending the 3-tuple (a,b,c) to $(|a-b|, |b-c|, |c-a|)$) also fails: i.e., there exists a 3-tuple $\mathbf{a} \in \mathbb{Z}^3$ such that no $k \in \mathbb{N}$ satisfies $D^k(\mathbf{a}) = (0,0,0)$. Better yet, prove that if $\mathbf{a} \in \mathbb{Z}^3$ is a 3-tuple whose entries are not all equal, then there is **no** $k \in \mathbb{N}$ such that $D^k(\mathbf{a}) = (0,0,0)$.

[**Hint:** For part **(a)**, find two nonzero reals x, y such that $D\left(1, x, x^2, x^3\right) = \left(y, yx, yx^2, yx^3\right)$.]

Exercise 8.3.8. The *Leibniz triangle* is a distant relative of Pascal's triangle. Here are its first few rows:

																									$k=0$ ↙
$n=0$	$\to$																								$\frac{1}{1}$
																								$k=1$ ↙	
$n=1$	$\to$																						$\frac{1}{2}$		$\frac{1}{2}$
																						$k=2$ ↙			
$n=2$	$\to$																				$\frac{1}{3}$		$\frac{1}{6}$		$\frac{1}{3}$
																				$k=3$ ↙					
$n=3$	$\to$																		$\frac{1}{4}$		$\frac{1}{12}$		$\frac{1}{12}$		$\frac{1}{4}$
																		$k=4$ ↙							
$n=4$	$\to$																$\frac{1}{5}$		$\frac{1}{20}$		$\frac{1}{30}$		$\frac{1}{20}$		$\frac{1}{5}$
																$k=5$ ↙									
$n=5$	$\to$														$\frac{1}{6}$		$\frac{1}{30}$		$\frac{1}{60}$		$\frac{1}{60}$		$\frac{1}{30}$		$\frac{1}{6}$
														$k=6$ ↙											
$n=6$	$\to$												$\frac{1}{7}$		$\frac{1}{42}$		$\frac{1}{105}$		$\frac{1}{140}$		$\frac{1}{105}$		$\frac{1}{42}$		$\frac{1}{7}$

It is defined as a family of rational numbers $T(n,k)$ defined for all pairs of nonnegative integers n,k satisfying $k \leq n$ and satisfying the following two properties:

1. We have $T(n,0) = T(n,n) = \dfrac{1}{n+1}$ for each $n \in \mathbb{N}$. (That is, the entries on two sides of the triangle are $\dfrac{1}{1}, \dfrac{1}{2}, \dfrac{1}{3}, \ldots$.)
2. We have $T(n,k) = T(n+1,k) + T(n+1,k+1)$ for each $n \in \mathbb{N}$ and $k \in \{0,1,\ldots,n\}$. (That is, each entry is the sum of its two neighbors below.)

(a) Prove that the numbers $T(n,k)$ are uniquely determined by this.
(b) Find an explicit formula for $T(n,k)$.

Exercise 8.3.9. Let n and k be two positive integers. You have nk identical coins distributed in n piles. Whenever two piles have an even number of coins in total, you are allowed to move coins between these piles in such a way that the numbers of coins on these two piles become equal. Such a move will be called a *balancing move*. (For example, if you have a pile with 3 coins and a pile with 9 coins, then a balancing move will change these two piles into two piles with 6 coins each. Each balancing move only changes two piles.) The distribution of coins is said to be *level* if all piles have the same number of coins.

Prove the following:

(a) If n is a power of 2, then any distribution of nk coins in n piles can be made level by a sequence of balancing moves.

(b) If n is not a power of 2, then there is a distribution of nk coins in n piles (for an appropriate k) that cannot be made level by any sequence of balancing moves.

[**Example:** If $n = 5$ and $k = 3$, and if the original distribution is $(2,4,1,4,4)$ (that is, the first pile has 2 coins, the second has 4 coins, the third has 1 coin, etc.), then here is one sequence of balancing moves that can be applied to this distribution (where we are underlining the piles that are being balanced in the next step):

$$(\underline{2},\underline{4},1,4,4) \to (\underline{3},3,\underline{1},4,4) \to (\underline{2},3,2,\underline{4},4) \to (3,3,\underline{2},3,\underline{4}) \to (3,3,3,3,3).$$

The distribution at the end of this sequence is level.]

Exercise 8.3.10. Let $n \in \mathbb{N}$. A silo contains $2n+1$ boxes; each box contains an integer amount of apples and an integer amount of pears.

(a) Prove that you can select $n+1$ boxes that altogether contain at least half of all the apples in the silo and also at least half of all the pears in the silo.

(b) Now assume that the boxes can also contain peaches. Is it always possible to select $n+1$ boxes that altogether contain at least half of all the apples in the

silo, at least half of all the pears in the silo, and at least half of all the peaches in the silo?

Exercise 8.3.11. Let $a, b \in \mathbb{N}$. The "toads and frogs" puzzle is a one-player game played on a $1 \times (a+b+1)$ horizontal strip of squares. Initially, the leftmost a squares are occupied by toads (one toad per square); the rightmost b squares are occupied by frogs (one frog per square); the one remaining square is empty. Thus, for example, for $a = 3$ and $b = 5$, the strip of squares looks as follows:

T	T	T		F	F	F	F	F

(where "T" stands for a toad, and where "F" stands for a frog). A move in this game is one of the following four operations:

- A **toad** moves one square to the **right**, assuming that the square it moves to is empty. This is called a *toad slide*.
- A **toad** moves two squares to the **right**, assuming that the square it moves to is empty and the square inbetween is occupied by a frog. (That is, a toad jumps over a single frog into an empty square to its right.) This is called a *toad jump*.
- A **frog** moves one square to the **left**, assuming that the square it moves to is empty. This is called a *frog slide*.
- A **frog** moves two squares to the **left**, assuming that the square it moves to is empty and the square inbetween is occupied by a toad. (That is, a frog jumps over a single toad into an empty square to its left.) This is called a *frog jump*.

The objective of the game is to achieve (by a sequence of moves, starting with the initial state described above) the state in which the leftmost b squares are occupied by frogs, the rightmost a squares are occupied by toads, and the one remaining square is empty. For example, here is a way to achieve this objective when $a = 2$ and $b = 2$:

	T	T		F	F	toad slide $\longrightarrow$	T		T	F	F
frog jump $\longrightarrow$	T	F	T		F	frog slide $\longrightarrow$	T	F	T	F	
toad jump $\longrightarrow$	T	F		F	T	toad jump $\longrightarrow$		F	T	F	T
frog slide $\longrightarrow$	F		T	F	T	frog jump $\longrightarrow$	F	F	T		T
toad slide $\longrightarrow$	F	F		T	T	.					

Prove that this objective can always be achieved in $ab + a + b$ (strategically chosen) moves.

8.4. Homework set #9

This is a **regular** problem set. See Section 3.7 for details on grading.

This homework set covers Chapter 7 and Chapter 8. Some of the problems may be unrelated.

Please solve **at most 5** problems. (No points will be given for further solutions.)

The following exercise is a variation on Exercise 7.8.1:

Exercise 8.4.1. Let n and d be two positive integers.

An n-tuple $(x_1, x_2, \ldots, x_n) \in [d]^n$ will be called *first-even* if its first entry x_1 occurs in it an even number of times (i.e., the number of $i \in [n]$ satisfying $x_i = x_1$ is even). (For example, the 3-tuples $(1,5,1)$ and $(2,2,3)$ are first-even, while the 3-tuple $(4,1,1)$ is not.)

Compute the # of first-even n-tuples in $[d]^n$.

Next come two binomial identities:

Exercise 8.4.2. Let $n \in \mathbb{N}$ and $p \in \mathbb{Z}$. Prove that

$$\sum_{k=0}^{n} (-1)^k \binom{n}{k} \binom{2n-2k}{n+p} = 2^{n-p} \binom{n}{p}.$$

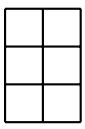
Exercise 8.4.3. Let n be a positive integer. Prove that

$$\sum_{i=0}^{n-1} \frac{(-1)^i}{n-i} \binom{n-i}{i} = \frac{1}{n} (-1)^{(n+1)//3} (1 + [3 \mid n]).$$

(See Definition 3.3.1 and Definition 4.3.19 for the notations used in Exercise 8.4.3.)

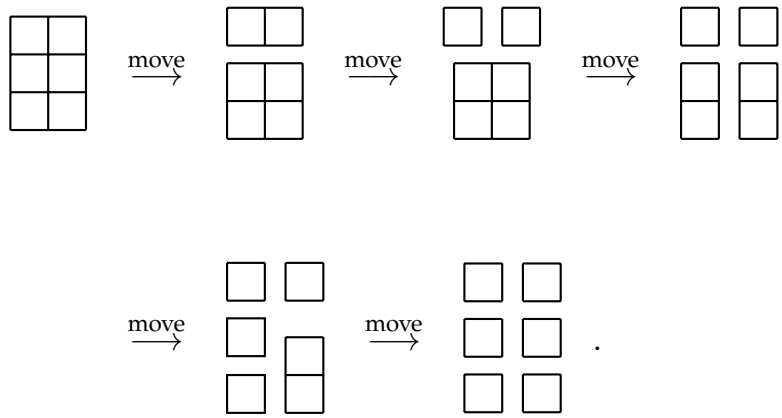
The next few exercises are about single-player games (aka nondeterministic processes):

Exercise 8.4.4. A chocolate bar has the shape of an $m \times n$-rectangle (subdivided into little 1×1-squares by horizontal and vertical lines, in case your eating habits are too healthy). For example, if $m = 3$ and $n = 2$, then it looks like this:



In one move, you can break a chocolate bar into two by splitting it along one of the (horizontal or vertical) lines that divide it (unless it is already a 1×1-square).

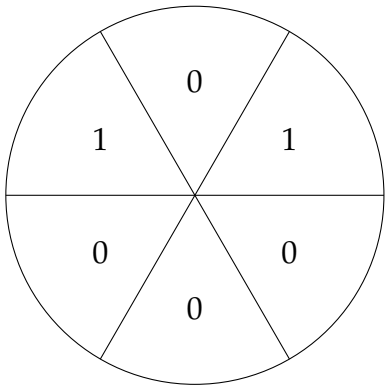
For example:



(a) What is the smallest number of moves necessary to break up the entire bar into 1×1-squares?

(b) What is the largest number of moves necessary to break up the entire bar into 1×1-squares?

Exercise 8.4.5. A circle is split into 6 sectors, and a number has been written in each sector. These numbers are $1, 0, 1, 0, 0, 0$ in clockwise order, as shown in the following picture:



In one move, you can add 1 to any two numbers written in adjacent sectors.

Can you ever ensure that all six sectors have the same number written in them?

Exercise 8.4.6. Fix a positive integer n. Consider n chips placed in a heap. In a move, you are allowed to split a heap H of chips into two smaller heaps H_1 and H_2; when doing so, you gain $|H_1| \cdot |H_2|$ cents. (We treat heaps as sets of chips; thus, $|H_1|$ is the # of chips in heap H_1.) After sufficiently many moves, you are left with n heaps, each containing exactly one chip. What is the maximum number of cents you can have made by that moment?

Exercise 8.4.7. Fix two positive integers a and b. Consider $a+b$ bowls, numbered $1, 2, \ldots, a+b$. Initially, each of the bowls $1, 2, \ldots, a$ contains an apple, and each of the bowls $a+1, a+2, \ldots, a+b$ contains a pear. A *move* consists of picking two numbers $i, j \in [a+b]$ satisfying $i < a+b$ and $j > 1$ and $i \equiv j \bmod 2$, and moving an apple from bowl i to bowl $i+1$ and a pear from bowl j to bowl $j-1$. (We assume that these fruits do exist in these bowls; otherwise, the move cannot be made. It is allowed for several fruits to lie in one bowl at the same time.)

The goal is to end up with each of the bowls $1, 2, \ldots, b$ containing a pear and each of the bowls $b+1, b+2, \ldots, b+a$ containing an apple. Show that this goal can be reached if and only if the product ab is even.

Exercise 8.4.8. Consider a hotel with an infinite number of rooms, arranged sequentially on the ground floor. The rooms are labelled by integers $i \in \mathbb{Z}$, with room i being adjacent to rooms $i-1$ and $i+1$. A finite number of violinists are staying in the hotel (it is possible for two violinists to be staying in the same room). Each night, two violinists staying in the same room decide they cannot stand each other's noise, and move to the two adjacent rooms (i.e., if they were in room i, they move to rooms $i-1$ and $i+1$). (Only two violinists move in any given night.) This keeps happening for as long as there are two violinists staying in the same room.

Prove that this moving will stop after a finite number of days (i.e., there will be a day when no two violinists share a room any more).

[**Example:** Let us assume that initially, there are 5 violinists (named a, b, c, d, e), arranged as follows:

$$\cdots \begin{array}{|c|c|c|c|c|c|c|c|c|c|c|} \hline & & a & c & d & & & & & & \\ & & b & & e & & & & & & \\ \hline \end{array} \cdots \tag{400}$$

(where the boxes of the table correspond to the rooms; of course, we are only showing a small part of the actually infinite hotel). On the first night, let's say the violinists a and b tire of each other's presence and move apart; this results in the following state:

$$\cdots \begin{array}{|c|c|c|c|c|c|c|c|c|c|c|} \hline & a & & b & d & & & & & & \\ & & & c & e & & & & & & \\ \hline \end{array} \cdots .$$

On the next night, the violinists b and c decide to move apart, and we obtain the following state:

$$\cdots \begin{array}{|c|c|c|c|c|c|c|c|c|c|c|} \hline & a & b & & c & & & & & & \\ & & & & d & & & & & & \\ & & & & e & & & & & & \\ \hline \end{array} \cdots .$$

On the next night, the violinists c and e decide to move apart, and we obtain the state

$$\cdots \begin{array}{|c|c|c|c|c|c|c|c|c|c|c|} \hline & a & b & c & d & e & & & & & \\ \hline \end{array} \cdots .$$

Now, no two violinists occupy the same room any more, so no further moves can happen.]

Our last two exercises on this homework set are about sequences of rational numbers that turn out to be integer sequences (not unlike the ones in Section 4.11 and Subsection 8.1.3):

Exercise 8.4.9. Define a sequence $(t_0, t_1, t_2, \ldots)$ of positive rational numbers recursively by setting

$$t_n = 1 \qquad \text{for each } n < 4$$

and

$$t_n = \frac{1 + t_{n-1}t_{n-3}}{t_{n-4}} \qquad \text{for each } n \geq 4.$$

(This is Sequence A217787 on the OEIS. Compare Example 4.11.5, which is similar but not the same.)

Prove that t_n is a positive integer for each integer $n \geq 0$.

Exercise 8.4.10. Let $q \in \mathbb{N}$. Define a sequence $(t_0, t_1, t_2, \ldots)$ of positive rational numbers recursively by setting

$$t_n = 1 \qquad \text{for each } n < 3$$

and

$$t_n = \frac{t_{n-1}^2 + qt_{n-1}t_{n-2} + t_{n-2}^2}{t_{n-3}} \qquad \text{for each } n \geq 3.$$

Prove that t_n is a positive integer for each integer $n \geq 0$.

9. Number Theory II: Primes

We covered some of the basics of number theory in Chapter 3; now we shall go deeper. The heroes of this chapter shall be the *prime numbers* (also known as *primes*); we will see both some of their intrinsic properties and their applications.

As with all other chapters, this one will be just an introduction; the primes are among the most well-studied and yet the most mysterious objects in mathematics, and the results that have been proved about them wouldn't fit into a book, let alone a chapter. Some books about primes are [Neale17] (a semi-popular survey with historical discussions and the occasional proof sketch), [CraPom05] (a "computational perspective", with special focus on algorithms and cryptographic applications), [FinRos16] (a somewhat advanced introduction aimed at graduate students) and [Narkie00] (a historical account with various classical proofs).

9.1. Primes

9.1.1. Definition and examples

We begin with the definition of a prime:

Definition 9.1.1. Let p be an integer greater than 1. We say that p is *prime* if the only positive divisors of p are 1 and p. A prime integer is often just called *a prime*.

Note that we required p to be greater than 1 here. Thus, 1 does not count as prime even though its only positive divisor is 1 itself.

Example 9.1.2. **(a)** The integer 5 is a prime, since its only positive divisors are 1 and 5.

(b) The integer 6 is not a prime, since it has positive divisors beyond just 1 and 6. (For example, 2 is one of its positive divisors.)

(c) None of the integers $4, 6, 8, 10, 12, 14, 16, \ldots$ (that is, the multiples of 2 that are larger than 2) is a prime. Indeed, if p is any of these numbers, then p has a positive divisor other than 1 and p (namely, 2), and therefore does not meet the definition of "prime".

(d) None of the integers $6, 9, 12, 15, 18, \ldots$ (that is, the multiples of 3 that are larger than 3) is a prime. The reason for this is similar to that in Example 9.1.2 **(c)**.

9.1.2. The infinitude of the primes

The first (i.e., smallest) 15 primes are

$$2,\ 3,\ 5,\ 7,\ 11,\ 13,\ 17,\ 19,\ 23,\ 29,\ 31,\ 37,\ 41,\ 43,\ 47.$$

One of the oldest famous theorems in mathematics is Euclid's result that there are infinitely many primes:

Theorem 9.1.3. There are infinitely many primes.

Many proofs of this theorem are known; for example, six proofs can be found in [AigZie14, Chapter 1]. We shall give two proofs here; but first, let us show that every integer $n > 1$ is divisible by at least one prime:

Proposition 9.1.4. Let $n > 1$ be an integer. Then, there exists at least one prime p such that $p \mid n$.

Proof of Proposition 9.1.4 (sketched). (See [19s, proof of Proposition 2.13.8] for details.) This is another application of the extremal principle: There exists a divisor of n that is larger than 1 (for example, n itself is such a divisor). Let q be the **smallest** such divisor. Then, $q \mid n$ and $q > 1$.

We now claim that q is a prime. Indeed, assume the contrary. Thus, q is not a prime; in other words, 1 and q are **not** the only positive divisors of q (by Definition 9.1.1). In other words, q has a positive divisor d distinct from 1 and q. Consider this d. We have $d \mid q \mid n$ and $d > 1$ (since d is a positive integer distinct from 1). Thus, d is a divisor of n that is larger than 1. Hence, $d \geq q$ (since q is the **smallest** such divisor). However, $d \mid q$ entails $d \leq q$ (since d and q are both positive). Combining this with $d \geq q$, we obtain $d = q$, which contradicts the fact that d is distinct from q. This contradiction shows that our assumption was wrong. Hence, q is a prime. Thus, there exists at least one prime p such that $p \mid n$ (namely, $p = q$). This proves Proposition 9.1.4. □

We can now step to the proofs of Theorem 9.1.3:

First proof of Theorem 9.1.3 (sketched). (See [19s, proof of Theorem 2.13.43] for details.) The following classical argument goes back to Euclid's *Elements*: Let $(p_1, p_2, \ldots, p_k)$ be any finite list of primes. We shall find a new prime p that is distinct from $p_1, p_2, \ldots, p_k$.

Indeed, set $n = p_1 p_2 \cdots p_k + 1$. It is easy to see that $n > 1$. Hence, Proposition 9.1.4 shows that there exists at least one prime p such that $p \mid n$. Consider this p. For each $i \in \{1, 2, \ldots, k\}$, we have $p_i \nmid 1$ (since p_i is a prime, so that $p_i > 1$) and thus

$$n = \underbrace{p_1 p_2 \cdots p_k}_{\substack{\equiv 0 \bmod p_i \\ (\text{since } p_i \mid p_1 p_2 \cdots p_k)}} + 1 \equiv 0 + 1 = 1 \not\equiv 0 \bmod p_i \qquad (\text{since } p_i \nmid 1 = 1 - 0).$$

In other words, for each $i \in \{1, 2, \ldots, k\}$, we have $p_i \nmid n$ and thus $p \neq p_i$ (since $p = p_i$ would yield $p = p_i \nmid n$, which would contradict $p \mid n$). In other words, p is distinct from $p_1, p_2, \ldots, p_k$.

Forget that we fixed $(p_1, p_2, \ldots, p_k)$. Thus, for any finite list $(p_1, p_2, \ldots, p_k)$ of primes, we have found a new prime p that is distinct from $p_1, p_2, \ldots, p_k$. In other words, for any finite list of primes, we have found a prime that is not in this list. Thus, there are infinitely many primes. This proves Theorem 9.1.3. □

Second proof of Theorem 9.1.3. Let $(F_0, F_1, F_2, \ldots)$ be the Fermat sequence, defined as in Exercise 3.7.3. Let $n \in \mathbb{N}$. Then, $F_n = \underbrace{2^{2^n}}_{>0} + 1 > 1$. Hence, Proposition 9.1.4 (applied to F_n instead of n) shows that there exists at least one prime p such that $p \mid F_n$. Consider this prime p, and denote it by p_n. Thus, p_n is a prime satisfying $p_n \mid F_n$.

Now, forget that we fixed n. Thus, for each $n \in \mathbb{N}$, we have constructed a prime p_n satisfying $p_n \mid F_n$. Consider all these primes $p_0, p_1, p_2, \ldots$.

We now claim that these primes $p_0, p_1, p_2, \ldots$ are distinct.

[*Proof:* Let n and m be two distinct nonnegative integers. We shall show that $p_n \neq p_m$.

Indeed, assume the contrary. Thus, $p_n = p_m$. Note that p_n is a prime; thus, $p_n > 1 > 0$, so that $|p_n| = p_n$. From $p_n = p_m$, we obtain $p_n \mid p_m$. Hence, Proposition 3.4.4 **(i)** (applied to $a = p_n$ and $b = p_m$) yields $\gcd(p_n, p_m) = |p_n| = p_n > 1$.

Now, Exercise 3.7.3 **(b)** yields $\gcd(F_n, F_m) = 1$. In other words, $F_n \perp F_m$ (by the definition of "coprime")[275]. But we have $p_n \mid F_n$ (by the construction of p_n) and $p_m \mid F_m$ (similarly). Thus, $p_n \perp p_m$ (by Proposition 3.5.8, applied to $a_1 = p_n$, $a_2 = p_m$, $b_1 = F_n$ and $b_2 = F_m$). In other words, $\gcd(p_n, p_m) = 1$ (by the definition of "coprime"). This contradicts $\gcd(p_n, p_m) > 1$. This contradiction shows that our assumption was wrong. Hence, $p_n \neq p_m$.

Forget that we fixed n and m. We thus have shown that $p_n \neq p_m$ whenever n and m are two distinct nonnegative integers. In other words, the primes $p_0, p_1, p_2, \ldots$ are distinct.]

We now know that the primes $p_0, p_1, p_2, \ldots$ are distinct. Hence, there are infinitely many primes (namely, these primes $p_0, p_1, p_2, \ldots$, and possibly others as well). This proves Theorem 9.1.3 again. □

9.1.3. Basic properties

We next review some of the most basic properties of primes. Their proofs are almost trivial, as we have already done the hard work in Chapter 3.

Proposition 9.1.5. Let p be a prime. Then, each $i \in \{1, 2, \ldots, p-1\}$ is coprime to p.

Proof of Proposition 9.1.5 (sketched). (See [19s, proof of Proposition 2.13.4] for details.) Let $i \in \{1, 2, \ldots, p-1\}$. We must prove that i is coprime to p. In other words, we must prove that $\gcd(i, p) = 1$.

From $i \in \{1, 2, \ldots, p-1\}$, we see that i is positive and satisfies $i \leq p - 1 < p$. Proposition 3.4.4 **(f)** yields $\gcd(i, p) \mid i$ and $\gcd(i, p) \mid p$. From $\gcd(i, p) \mid i$, we obtain $|\gcd(i, p)| \leq |i|$ (by Proposition 3.1.3 **(b)**). This rewrites as $\gcd(i, p) \leq i$ (since $\gcd(i, p)$ and i are positive). Hence, $\gcd(i, p) \leq i < p$.

[275] Here, we are using the notation from Definition 3.5.3.

But p is prime; thus, the only positive divisors of p are 1 and p. However, $\gcd(i,p)$ is a positive divisor of p (since $\gcd(i,p) \mid p$). Thus, $\gcd(i,p)$ must be either 1 or p. Since $\gcd(i,p) < p$, we thus conclude $\gcd(i,p) = 1$. Proposition 9.1.5 is proven. □

Proposition 9.1.6. Let p be a prime. Let $a \in \mathbb{Z}$. Then, either $p \mid a$ or $p \perp a$.

Proof of Proposition 9.1.6 (sketched). (See [19s, proof of Proposition 2.13.5] for a different proof.) We are in one of the following two cases:

Case 1: We have $a\%p = 0$.

Case 2: We have $a\%p \neq 0$.

Let us first consider Case 1. In this case, we have $a\%p = 0$. However, Proposition 3.3.2 **(b)** (applied to $n = p$ and $u = a$) yields that we have $p \mid a$ if and only if $a\%p = 0$. Thus, $p \mid a$ (since $a\%p = 0$). Hence, Proposition 9.1.6 is proved in Case 1.

Let us now consider Case 2. In this case, we have $a\%p \neq 0$. But Proposition 3.3.2 **(a)** (applied to $n = p$ and $u = a$) yields that $a\%p \in \{0, 1, \ldots, p-1\}$ and $a\%p \equiv a \bmod p$. Combining $a\%p \in \{0, 1, \ldots, p-1\}$ with $a\%p \neq 0$, we find $a\%p \in \{0, 1, \ldots, p-1\} \setminus \{0\} = \{1, 2, \ldots, p-1\}$. Hence, Proposition 9.1.5 (applied to $i = a\%p$) yields that $a\%p$ is coprime to p. In other words, $\gcd(a\%p, p) = 1$. But Proposition 3.4.4 **(e)** (applied to p and a instead of a and b) yields

$$\begin{aligned} \gcd(p,a) &= \gcd(p, a\%p) = \gcd(a\%p, p) \qquad \text{(by Proposition 3.4.4 \textbf{(b)})} \\ &= 1. \end{aligned}$$

In other words, $p \perp a$. Hence, Proposition 9.1.6 is proved in Case 2.

We have now proved Proposition 9.1.6 in both Cases 1 and 2. Hence, Proposition 9.1.6 always holds. □

Any two distinct primes are coprime:

Proposition 9.1.7. Let p and q be two distinct primes. Then, $p \perp q$.

This is easy to prove using the above propositions (see [19s, solution to Exercise 2.13.1] for a detailed proof).

The following property of primes ([19s, Theorem 2.13.6]) is probably one of the most important ones:

Theorem 9.1.8. Let p be a prime. Let $a, b \in \mathbb{Z}$ such that $p \mid ab$. Then, $p \mid a$ or $p \mid b$.

Proof of Theorem 9.1.8. Clearly, Theorem 9.1.8 is true if $p \mid a$. Hence, for the rest of this proof, we WLOG assume that we don't have $p \mid a$.

However, Proposition 9.1.6 shows that we have either $p \mid a$ or $p \perp a$. Thus, $p \perp a$ (since we don't have $p \mid a$). However, $p \mid ab$. Hence, Theorem 3.5.6 (applied to p, a and b instead of a, b and c) yields $p \mid b$. Thus, $p \mid a$ or $p \mid b$. This proves Theorem 9.1.8. □

Theorem 9.1.8 can be generalized to products of multiple factors:

Proposition 9.1.9. Let p be a prime. Let $a_1, a_2, \ldots, a_k$ be integers such that $p \mid a_1 a_2 \cdots a_k$. Then, $p \mid a_i$ for some $i \in \{1, 2, \ldots, k\}$.

Hint to the proof of Proposition 9.1.9. This is easy to prove by induction on k (with the induction step relying on Theorem 9.1.8). Alternatively, this can be derived from Exercise 3.5.4 and Proposition 9.1.6. (See [19s, proof of Proposition 2.13.7] for the latter proof.) □

We notice that Proposition 9.1.9 is, in some way, saying that primes "act like 0" in the following sense: It is well-known that a product of numbers is 0 if and only if one of the factors is 0. Proposition 9.1.9 says something similar: It says that a product of integers is divisible by a prime p if and only if one of its factors is divisible by p. (To be more precise, the proposition only claims the "only if" part of this sentence; but the "if" part is obvious.) This analogy becomes even more vivid if we replace the wording "divisible by p" by the (equivalent) wording "congruent to 0 modulo p". This foreshadows the existence of the *finite field* $\mathbb{F}_p$, which we will later get to define.

9.1.4. A few exercises

A number of contest problems about primes can be solved using just the basic properties above. As an example, let us prove what is perhaps one of the first properties one will notice when writing down a list of primes:

Exercise 9.1.1. **(a)** Prove that every prime $p > 2$ is odd.
(b) Prove that every prime $p > 3$ satisfies either $p \equiv 1 \bmod 6$ or $p \equiv 5 \bmod 6$.

Solution to Exercise 9.1.1. **(a)** Let $p > 2$ be a prime. We must show that p is odd.

Indeed, assume (for the sake of contradiction) that p is even. Hence, $2 \mid p$. Thus, 2 is a positive divisor of p. However, p is prime. In other words, the only positive divisors of p are 1 and p (by the definition of "prime"). Thus, 2 must be either 1 or p (since 2 is a positive divisor of p). Since 2 is not 1, we thus conclude that 2 is p. In other words, $2 = p$. This contradicts $p > 2$. This contradiction shows that our assumption was false.

Hence, p is not even. In other words, p is odd. This proves Exercise 9.1.1 **(a)**.

(b) Let $p > 3$ be a prime. We must prove that $p \equiv 1 \bmod 6$ or $p \equiv 5 \bmod 6$.

Indeed, $p > 3 > 2$. Hence, Exercise 9.1.1 **(a)** shows that p is odd. In other words, $p \equiv 1 \bmod 2$ (by Exercise 3.3.2 **(d)**, applied to $u = p$). In other words, $2 \mid p - 1$. In other words, there exists an integer c such that $p - 1 = 2c$. Consider this c. From $p - 1 = 2c$, we obtain $p = 2c + 1$.

In the above solution to Exercise 9.1.1 **(a)**, we have showed that p is not even. A similar argument can be used to show that we don't have $p \equiv 0 \bmod 3$ [276].

[276] Here are the details:

Now, Proposition 3.3.2 **(a)** (applied to $n = 3$ and $u = c$) yields that $c\%3 \in \{0, 1, \ldots, 3-1\}$ and $c\%3 \equiv c \bmod 3$. From $c\%3 \equiv c \bmod 3$, we obtain $c \equiv c\%3 \bmod 3$. Also, Proposition 3.3.2 **(d)** (applied to $n = 3$ and $u = c$) yields that $c = (c//3)\,3 + (c\%3)$. Hence,

$$p = 2 \underbrace{c}_{=(c//3)3+(c\%3)} + 1 = 2\left((c//3)\,3 + (c\%3)\right) + 1 = \underbrace{6\,(c//3)}_{\substack{\equiv 0 \bmod 6 \\ \text{(since } c//3 \text{ is an integer)}}} + 2\,(c\%3) + 1$$

$$\equiv 2\,(c\%3) + 1 \bmod 6. \tag{401}$$

We have $c\%3 \in \{0, 1, \ldots, 3-1\} = \{0, 1, 2\}$. If we had $c\%3 = 1$, then we would have

$$p = 2 \underbrace{c}_{\equiv c\%3 = 1 \bmod 3} + 1 \equiv 2 \cdot 1 + 1 = 3 \equiv 0 \bmod 3,$$

which would contradict the fact that we don't have $p \equiv 0 \bmod 3$. Hence, we cannot have $c\%3 = 1$. In other words, we have $c\%3 \neq 1$.

Combining $c\%3 \in \{0, 1, 2\}$ with $c\%3 \neq 1$, we obtain $c\%3 \in \{0, 1, 2\} \setminus \{1\} = \{0, 2\}$. In other words, we have $c\%3 = 0$ or $c\%3 = 2$. Hence, we must be in one of the following two cases:

Case 1: We have $c\%3 = 0$.

Case 2: We have $c\%3 = 2$.

Let us first consider Case 1. In this case, we have $c\%3 = 0$. Now, (401) becomes $p \equiv 2 \underbrace{(c\%3)}_{=0} + 1 = 1 \bmod 6$. Hence, $p \equiv 1 \bmod 6$ or $p \equiv 5 \bmod 6$. Thus, Exercise 9.1.1 **(b)** is solved in Case 1.

Let us now consider Case 2. In this case, we have $c\%3 = 2$. Now, (401) becomes $p \equiv 2 \underbrace{(c\%3)}_{=2} + 1 = 2 \cdot 2 + 1 = 5 \bmod 6$. Hence, $p \equiv 1 \bmod 6$ or $p \equiv 5 \bmod 6$. Thus, Exercise 9.1.1 **(b)** is solved in Case 2.

We have now solved Exercise 9.1.1 **(b)** in both Cases 1 and 2. Thus, Exercise 9.1.1 **(b)** always holds. □

Exercise 9.1.1 **(a)** shows that the primes larger than 2 are precisely the odd primes.

Our next exercise is, in a certain sense, an analogue of Exercise 3.2.7. Recall that Exercise 3.2.7 allows taking a congruence to the k-th power: If n, a, b are integers satisfying $a \equiv b \bmod n$, then $a^k \equiv b^k \bmod n$ for each $k \in \mathbb{N}$. One might hope for something similar to hold when the powers are replaced by binomial coefficients (i.e., when a^k and b^k are replaced by $\binom{a}{k}$ and $\binom{b}{k}$); but this is not generally the

Assume (for the sake of contradiction) that $p \equiv 0 \bmod 3$. Hence, $3 \mid p$. Thus, 3 is a positive divisor of p. However, p is prime. In other words, the only positive divisors of p are 1 and p (by the definition of "prime"). Thus, 3 must be either 1 or p (since 3 is a positive divisor of p). Since 3 is not 1, we thus conclude that 3 is p. In other words, $3 = p$. This contradicts $p > 3$. This contradiction shows that our assumption was false. In other words, we don't have $p \equiv 0 \bmod 3$.

case (for example, $2 \equiv 0 \bmod 2$ is true, but $\binom{2}{2} \equiv \binom{0}{2} \bmod 2$ is false). Nevertheless, it holds if n is prime and $k < n$, as the following exercise shows:[277]

Exercise 9.1.2. Let p be a prime. Let a and b be two integers such that $a \equiv b \bmod p$. Let $k \in \{0, 1, \ldots, p-1\}$. Prove that $\binom{a}{k} \equiv \binom{b}{k} \bmod p$.

Solution to Exercise 9.1.2. Our plan is the following: First, we shall prove the congruence $k!\binom{a}{k} \equiv k!\binom{b}{k} \bmod p$; then, we will argue that $k! \perp p$, and thus (using Lemma 3.5.11) we will be able to cancel the $k!$ from this congruence.

We begin by proving $k!\binom{a}{k} \equiv k!\binom{b}{k} \bmod p$. Indeed, (117) (applied to $n = a$) yields

$$\binom{a}{k} = \frac{a(a-1)(a-2)\cdots(a-k+1)}{k!}.$$

Multiplying this equality by $k!$, we find

$$\begin{aligned} k!\binom{a}{k} &= a(a-1)(a-2)\cdots(a-k+1) \\ &= \prod_{s \in \{0,1,\ldots,k-1\}} (a-s). \end{aligned} \tag{402}$$

The same argument (applied to b instead of a) yields

$$k!\binom{b}{k} = \prod_{s \in \{0,1,\ldots,k-1\}} (b-s). \tag{403}$$

However, $\underbrace{a}_{\equiv b \bmod p} - s \equiv b - s \bmod p$ for each $s \in \{0, 1, \ldots, k-1\}$. Hence, (41) (applied to $S = \{0, 1, \ldots, k-1\}$ and $a_s = a - s$ and $b_s = b - s$ and $n = p$) yields $\prod_{s \in \{0,1,\ldots,k-1\}} (a-s) \equiv \prod_{s \in \{0,1,\ldots,k-1\}} (b-s) \bmod p$. Hence,

$$k!\binom{a}{k} = \prod_{s \in \{0,1,\ldots,k-1\}} (a-s) \equiv \prod_{s \in \{0,1,\ldots,k-1\}} (b-s) = k!\binom{b}{k} \bmod p$$

(by (403)).

Now, $k \leq p-1$ (since $k \in \{0, 1, \ldots, p-1\}$) and $k! = 1 \cdot 2 \cdot \cdots \cdot k$. We shall next show that $k! \perp p$.

Let $i \in \{1, 2, \ldots, k\}$. Then, $i \in \{1, 2, \ldots, k\} \subseteq \{1, 2, \ldots, p-1\}$ (since $k \leq p-1$). Thus, i is coprime to p (by Proposition 9.1.5). In other words, $i \perp p$.

[277] Exercise 9.1.2 is [Grinbe19c, Proposition 5.5] (with u and v renamed as a and b).

Forget that we fixed i. We thus have shown that each $i \in \{1, 2, \ldots, k\}$ satisfies $i \perp p$. Thus, Exercise 3.5.4 (applied to $c = p$ and $a_i = i$) yields $1 \cdot 2 \cdot \cdots \cdot k \perp p$. In other words, $k! \perp p$ (since $k! = 1 \cdot 2 \cdot \cdots \cdot k$). Therefore, Lemma 3.5.11 (applied to $k!$, $\binom{a}{k}$, $\binom{b}{k}$ and p instead of a, b, c and n) yields $\binom{a}{k} \equiv \binom{b}{k} \bmod p$ (since $k!\binom{a}{k} \equiv k!\binom{b}{k} \bmod p$). This proves Exercise 9.1.2. □

As a further example, here is a problem from the 12th University of Michigan Undergraduate Mathematics Competition in 1995 ([VelWag20, Problem 28]):

Exercise 9.1.3. Let n be a positive integer. Prove that n is prime if and only if there is a **unique** pair (j, k) of positive integers satisfying $\frac{1}{j} - \frac{1}{k} = \frac{1}{n}$.

Solution to Exercise 9.1.3 (sketched). We shall prove the following four claims:

Claim 1: If $n > 1$, then there is **at least one** pair (j, k) of positive integers satisfying $\frac{1}{j} - \frac{1}{k} = \frac{1}{n}$.

Claim 2: If n is prime, then there is **a unique** pair (j, k) of positive integers satisfying $\frac{1}{j} - \frac{1}{k} = \frac{1}{n}$.

Claim 3: If there is **at least one** pair (j, k) of positive integers satisfying $\frac{1}{j} - \frac{1}{k} = \frac{1}{n}$, then $n > 1$.

Claim 4: If there is a **unique** pair (j, k) of positive integers satisfying $\frac{1}{j} - \frac{1}{k} = \frac{1}{n}$, then n is prime.

Once these four claims are proved, Exercise 9.1.3 will follow by combining Claim 2 with Claim 4. (Claims 1 and 3 serve ancillary roles in proving other claims.) So let us prove these four claims:

[*Proof of Claim 1:* Assume that $n > 1$. Then, $n - 1$ and $(n-1)n$ are positive integers, and an easy computation reveals that $\frac{1}{n-1} - \frac{1}{(n-1)n} = \frac{1}{n}$. Hence, there is **at least one** pair (j, k) of positive integers satisfying $\frac{1}{j} - \frac{1}{k} = \frac{1}{n}$ (namely, $(j, k) = (n-1, (n-1)n)$). This proves Claim 1.]

[*Proof of Claim 2:* Assume that n is prime. Thus, $n > 1$. Now, let (j, k) be a pair of positive integers satisfying $\frac{1}{j} - \frac{1}{k} = \frac{1}{n}$. We shall show that $(j, k) = (n-1, (n-1)n)$.

Indeed, we have $\frac{1}{j} - \frac{1}{k} = \frac{1}{n}$, so that $\frac{1}{n} = \frac{1}{j} - \underbrace{\frac{1}{k}}_{>0} < \frac{1}{j}$ and thus $j < n$. Hence, we cannot have $n \mid j$ [278]. Also, from $j < n$, we see that $n - j$ is a positive integer. Also, $n - j$ cannot be n (since $n - \underbrace{j}_{>0} < n$).

Also, $\frac{1}{n} = \frac{1}{j} - \frac{1}{k} = \frac{k-j}{jk}$, so that $jk = n(k-j)$. Now, $n \mid n(k-j) = jk$. Hence, Theorem 9.1.8 (applied to $p = n$, $a = j$ and $b = k$) yields that $n \mid j$ or $n \mid k$. Thus, we have $n \mid k$ (since we cannot have $n \mid j$). In other words, $k = na$ for some $a \in \mathbb{Z}$. Consider this a.

Now, recall that $jk = n(k-j)$. In view of $k = na$, this rewrites as $jna = n(na - j)$. We can cancel n from this equality (since $n > 0$), and thus obtain $ja = na - j$. Hence,

$$(n-j)(a+1) = na - \underbrace{ja}_{=na-j} + n - j = na - (na - j) + n - j = n.$$

Thus, $n - j \mid (n-j)(a+1) = n$. Hence, $n - j$ is a positive divisor of n (since $n - j$ is a positive integer).

But n is prime. Hence, the only positive divisors of are 1 and n (by the definition of a "prime"). Thus, $n - j$ must be 1 or n (since $n - j$ is a positive divisor of n). Since $n - j$ cannot be n, we thus conclude that $n - j$ must be 1. Hence, $n - j = 1$, so that $j = n - 1$. Now, recall that $(n-j)(a+1) = n$. Hence, $n = \underbrace{(n-j)}_{=1}(a+1) = a + 1$, so that $a = n - 1$ and therefore $k = n \underbrace{a}_{=n-1} = n(n-1)$.

From $j = n - 1$ and $k = n(n-1)$, we obtain $(j,k) = (n-1, n(n-1))$.

Forget that we fixed (j,k). Thus, we have proved that every pair (j,k) of positive integers satisfying $\frac{1}{j} - \frac{1}{k} = \frac{1}{n}$ satisfies $(j,k) = (n-1, n(n-1))$. Hence, there is **at most one** such pair (j,k). Since we also know (from Claim 1) that there is **at least one** such pair (j,k), we can thus conclude that there is a **unique** such pair (j,k). This proves Claim 2.]

[*Proof of Claim 3:* Assume that there is **at least one** pair (j,k) of positive integers satisfying $\frac{1}{j} - \frac{1}{k} = \frac{1}{n}$. Consider this pair (j,k). Thus, $\frac{1}{j} - \frac{1}{k} = \frac{1}{n}$, so that $\frac{1}{n} = \frac{1}{j} - \underbrace{\frac{1}{k}}_{>0} < \frac{1}{j}$. Thus, $n > j \geq 1$. This proves Claim 3.]

[*Proof of Claim 4:* Assume that there is a **unique** pair (j,k) of positive integers satisfying $\frac{1}{j} - \frac{1}{k} = \frac{1}{n}$. Thus, in particular, there is **at least one** such pair (j,k).

[278] *Proof.* Assume the contrary. Thus, $n \mid j$. Hence, from Proposition 3.1.3 **(b)**, we obtain $|n| \leq |j|$ (since $j \neq 0$). But j and n are positive integers; thus, $|j| = j$ and $|n| = n$. Hence, $n = |n| \leq |j| = j < n$, which is absurd. This contradiction shows that our assumption was false, qed.

Hence, Claim 3 yields that $n > 1$.

However, we must prove that n is prime.

Assume the contrary. Thus, n is not a prime. Since $n > 1$, we thus conclude that 1 and n are **not** the only positive divisors of n (by the definition of a "prime"). In other words, there exists a positive divisor d of n that is distinct from 1 and n. Consider this d.

Set $e = n/d$. Then, $e = n/d \in \mathbb{Z}$ (since d is a divisor of n) and $e = n/d > 0$ (since n and d are positive). Hence, e is a positive integer. Moreover, from $e = n/d$, we obtain $ed = n$. We have $d > 1$ (since d is a positive integer distinct from 1) and thus $d - 1 > 0$. Hence, $e(d-1)$ and $ed(d-1)$ are positive integers. Thus, $(e(d-1), ed(d-1))$ is a pair (j,k) of positive integers satisfying $\frac{1}{j} - \frac{1}{k} = \frac{1}{n}$ (indeed, it is easy to see that $\frac{1}{e(d-1)} - \frac{1}{ed(d-1)} = \frac{1}{ed} = \frac{1}{n}$ (since $ed = n$)). However, we also know (from the proof of Claim 1) that $(n-1, (n-1)n)$ is a pair (j,k) of positive integers satisfying $\frac{1}{j} - \frac{1}{k} = \frac{1}{n}$.

Thus, we have found two pairs (j,k) of positive integers satisfying $\frac{1}{j} - \frac{1}{k} = \frac{1}{n}$: namely, $(e(d-1), ed(d-1))$ and $(n-1, (n-1)n)$. Since we have assumed that there is a **unique** such pair (j,k), we thus conclude that these two pairs must be equal. In other words, we have $(e(d-1), ed(d-1)) = (n-1, (n-1)n)$. Hence, $e(d-1) = n-1$ and $ed(d-1) = (n-1)n$. Now, comparing $e(d-1) = n-1$ with $e(d-1) = \underbrace{ed}_{=n} - e = n - e$, we obtain $n - 1 = n - e$. Thus, $1 = e = n/d$, so that $d = n$. This contradicts the fact that d is distinct from n. This contradiction shows that our assumption was false. Hence, Claim 4 is proved.]

Combining Claim 2 with Claim 4, we conclude the claim of Exercise 9.1.3. □

9.1.5. Homework set #10A: Elementary properties of primes

This homework set is optional. I will grade your solutions if you choose to write them up, but there won't be any points to gain.

Exercise 9.1.4. Let n be an integer. Prove that n can be represented in the form $n = u^2 - v^2$ for some $u, v \in \mathbb{Z}$ if and only if $n \not\equiv 2 \bmod 4$.

Exercise 9.1.5. Let n be a positive integer. Prove that n can be represented in the form $n = uv - \binom{u}{2}$ for some $u, v \in \mathbb{Z}$ satisfying $v \geq u \geq 3$ if and only if n is neither a prime nor a power of 2.

The following exercise is among the first results in the deep topic of interplays between primes and Pascal's triangle:

Exercise 9.1.6. Let p be a prime.

(a) Prove that $p \mid \binom{p}{k}$ for each $k \in \{1, 2, \ldots, p-1\}$.

(b) Prove that $\binom{p-1}{k} \equiv (-1)^k \bmod p$ for each $k \in \{0, 1, \ldots, p-1\}$.

Exercise 9.1.7. Let p be a prime. Let $k \in \{0, 1, \ldots, p-2\}$. Prove that $\sum_{i=0}^{p-1} i^k \equiv 0 \bmod p$.

[**Hint:** Exercise 9.1.6 **(b)** might help.]

The next exercise extends Exercise 9.1.3:

Exercise 9.1.8. Let n be a positive integer. Let u be the # of pairs (j, k) of positive integers satisfying $\frac{1}{j} - \frac{1}{k} = \frac{1}{n}$.

Prove that u is the # of all integers $i \in [n-1]$ satisfying $i \mid n^2$.

9.1.6. Fermat's little theorem

The following theorem ([19s, Theorem 2.15.1]) is known as *Fermat's Little Theorem* (often abbreviated as "FLT"):

Theorem 9.1.10 (Fermat's Little Theorem)**.** Let p be a prime. Let $a \in \mathbb{Z}$.

(a) If $p \nmid a$, then $a^{p-1} \equiv 1 \bmod p$.

(b) We always have $a^p \equiv a \bmod p$.

Theorem 9.1.10 is often abbreviated "FLT" or "Little Fermat". The word "little" in the name of this theorem is to distinguish it from "Fermat's Last Theorem", a much more difficult result only proven in the 1990s.

Note that Exercise 3.3.5 is the particular case of Theorem 9.1.10 **(a)** for $p = 3$ (and $a = n$).

We will outline two proofs of Theorem 9.1.10. The first relies on the following lemma:

Lemma 9.1.11. Let p be a prime. Then:

(a) We have $(a+b)^p \equiv a^p + b^p \bmod p$ for any $a, b \in \mathbb{Z}$.

(b) We have $(-a)^p \equiv -a^p \bmod p$ for any $a \in \mathbb{Z}$.

Proof of Lemma 9.1.11. **(a)** Let $a, b \in \mathbb{Z}$. Theorem 4.3.16 (applied to $x = a$, $y = b$ and

$n = p$) yields

$$\begin{aligned}
(a+b)^p &= \sum_{k=0}^{p} \binom{p}{k} a^k b^{p-k} \\
&= \underbrace{\binom{p}{0}}_{\substack{=1 \\ \text{(by (119))}}} \underbrace{a^0}_{=1} \underbrace{b^{p-0}}_{=b^p} + \sum_{k=1}^{p-1} \underbrace{\binom{p}{k}}_{\substack{\equiv 0 \bmod p \\ \text{(since Exercise 9.1.6 \textbf{(a)}} \\ \text{yields } p \mid \binom{p}{k}\text{)}}} a^k b^{p-k} + \underbrace{\binom{p}{p}}_{\substack{=1 \\ \text{(by (124))}}} a^p \underbrace{b^{p-p}}_{=b^0=1} \\
&\qquad \left(\begin{array}{c} \text{here, we have split off the addends for } k=0 \text{ and} \\ \text{for } k=p \text{ from the sum} \end{array} \right) \\
&\equiv b^p + \underbrace{\sum_{k=1}^{p-1} 0 a^k b^{p-k}}_{=0} + a^p = b^p + a^p = a^p + b^p \bmod p.
\end{aligned}$$

This proves Lemma 9.1.11 **(a)**.

(b) Let $a \in \mathbb{Z}$. Applying Lemma 9.1.11 **(a)** to $b = -a$, we obtain $(a + (-a))^p \equiv a^p + (-a)^p \bmod p$. Hence,

$$a^p + (-a)^p \equiv \left(\underbrace{a + (-a)}_{=0} \right)^p = 0^p = 0 \bmod p \qquad (\text{since } p > 0).$$

Thus, $(-a)^p \equiv 0 - a^p = -a^p \bmod p$. This proves Lemma 9.1.11 **(b)**. □

First proof of Theorem 9.1.10. **(b)** Forget that we fixed a. We shall first show that Theorem 9.1.10 **(b)** holds for $a \geq 0$. That is, we shall prove the following claim:

Claim 1: We have $a^p \equiv a \bmod p$ for each $a \in \mathbb{N}$.

[*Proof of Claim 1:* We proceed by induction on a:

Induction base: We have $0^p = 0$ (since p is positive) and thus $0^p \equiv 0 \bmod p$. In other words, Claim 1 holds for $a = 0$.

Induction step: Let $k \in \mathbb{N}$. Assume (as the induction hypothesis) that Claim 1 holds for $a = k$. We must prove that Claim 1 holds for $a = k + 1$.

We have assumed that Claim 1 holds for $a = k$. In other words, we have $k^p \equiv k \bmod p$.

Now, Lemma 9.1.11 **(a)** (applied to $a = k$ and $b = 1$) yields $(k+1)^p \equiv \underbrace{k^p}_{\equiv k \bmod p} + \underbrace{1^p}_{=1} \equiv k + 1 \bmod p$. In other words, Claim 1 holds for $a = k + 1$. This completes the induction step. Thus, Claim 1 is proved by induction.]

We can now prove Theorem 9.1.10 **(b)** in full generality:

Let $a \in \mathbb{Z}$. We must show that $a^p \equiv a \bmod p$. If $a \in \mathbb{N}$, then this follows from Claim 1. Hence, for the rest of this proof, we WLOG assume that $a \notin \mathbb{N}$. Combining $a \in \mathbb{Z}$ with $a \notin \mathbb{N}$, we find $a \in \mathbb{Z} \setminus \mathbb{N} = \{-1, -2, -3, \ldots\}$. Hence, $-a \in \{1, 2, 3, \ldots\} \subseteq \mathbb{N}$. Thus, we can apply Claim 1 to $-a$ instead of a. As a result, we obtain $(-a)^p \equiv -a \bmod p$. However, Lemma 9.1.11 **(b)** yields $(-a)^p \equiv -a^p \bmod p$. Thus, $-a^p \equiv (-a)^p \equiv -a \bmod p$. Hence,

$$a^p = (-1) \underbrace{(-a^p)}_{\equiv -a \bmod p} \equiv (-1)(-a) = a \bmod p.$$

This proves Theorem 9.1.10 **(b)**.

(a) Assume that $p \nmid a$. Theorem 9.1.10 **(b)** yields $a^p \equiv a \bmod p$. In other words, $p \mid a^p - a$. In view of $a^p - a = a\left(a^{p-1} - 1\right)$, this rewrites as $p \mid a\left(a^{p-1} - 1\right)$. Hence, Theorem 9.1.8 (applied to $b = a^{p-1} - 1$) yields that $p \mid a$ or $p \mid a^{p-1} - 1$. Since we don't have $p \mid a$ (because we have assumed that $p \nmid a$), we thus obtain $p \mid a^{p-1} - 1$. In other words, $a^{p-1} \equiv 1 \bmod p$. Thus, Theorem 9.1.10 **(a)** is proved. □

Remark 9.1.12. We have used Lemma 9.1.11 to prove Theorem 9.1.10 **(b)** above; conversely, Lemma 9.1.11 can easily be derived from Theorem 9.1.10 **(b)**. This might suggest that Lemma 9.1.11 is just a one-trick tool for the proof of Theorem 9.1.10 **(b)**. However, when properly generalized, Lemma 9.1.11 becomes much stronger and more useful than Theorem 9.1.10 **(b)** can ever get! Indeed, there is a "grown-up" version of Lemma 9.1.11, which no longer requires a and b to be integers but rather allows them to be any two commuting elements of a ring (see any course on abstract algebra for the meanings of these words). A down-to-earth example of this is when a and b are two $n \times n$-matrices with integer entries satisfying $ab = ba$; the "grown-up" version of Lemma 9.1.11 **(a)** then says that $(a+b)^p \equiv a^p + b^p \bmod p$, in the sense that each entry of the matrix $(a+b)^p$ is congruent to the corresponding entry of $a^p + b^p$ modulo p.

Thus, Lemma 9.1.11 in the form we stated it above is merely the tip of an iceberg. Fortunately, the proof we gave for it applies almost verbatim to the generalization (which is known as "Idiot's Binomial Formula" or "Freshman's Dream", as it allows replacing the p-th power of a sum by a sum of p-th powers in certain situations).

9.1.7. Euler's totient function

Our second proof of Theorem 9.1.10 will derive it from a more general result, known as *Euler's theorem* ([19s, Theorem 2.15.3]):

Theorem 9.1.13 (Euler's theorem)**.** Let n be a positive integer. We let $\phi(n)$ denote the number of all $i \in \{1, 2, \ldots, n\}$ satisfying $i \perp n$. (This notation was already used in Exercise 4.5.6.)

Let $a \in \mathbb{Z}$ be coprime to n. Then, $a^{\phi(n)} \equiv 1 \bmod n$.

Theorem 9.1.13 yields Theorem 9.1.10 **(a)**, since $\phi(p) = p - 1$ when p is prime (check this!). Theorem 9.1.10 **(b)** follows from Theorem 9.1.10 **(a)** easily (one just needs to handle the cases $p \mid a$ and $p \nmid a$ separately). Thus, in order to obtain a second proof of Theorem 9.1.10, we only need to establish Theorem 9.1.13. Before we do so, let us first illustrate how it can be used on a typical contest problem:

Exercise 9.1.9. What is the last digit of 7^{7^7}?

Notational remark: An expression of the form "a^{b^c}" always means $a^{(b^c)}$, not $\left(a^b\right)^c$.

Discussion of Exercise 9.1.9. The last digit of a positive integer n is $n\%10$ (that is, the remainder of n upon division by 10). So we need to compute $7^{7^7}\%10$.

Since 7 is coprime to 10, we can apply Theorem 9.1.13 to $n = 10$ and $a = 7$. We thus get $7^{\phi(10)} \equiv 1 \bmod 10$. Since $\phi(10) = 4$, this rewrites as $7^4 \equiv 1 \bmod 10$. (This is also not hard to check directly, using $7^2 = 49 \equiv -1 \bmod 10$.) Hence, if we write 7^7 in the form $7^7 = 4a + b$ for some $a, b \in \mathbb{N}$, then

$$7^{7^7} = 7^{4a+b} = \underbrace{7^{4a}}_{=\left(7^4\right)^a} \cdot 7^b = \left(\underbrace{7^4}_{\equiv 1 \bmod 10}\right)^a \cdot 7^b \equiv \underbrace{1^a}_{=1} \cdot 7^b = 7^b \bmod 10$$

and therefore $7^{7^7}\%10 = 7^b\%10$. This gives us an easy way to compute $7^{7^7}\%10$ provided that b is small enough (so we can compute $7^b\%10$).

Now, how do we write 7^7 in the form $7^7 = 4a + b$ for some $a, b \in \mathbb{N}$, with smallest possible b ? The answer is, of course, that we take $a = 7^7//4$ and $b = 7^7\%4$ (because Proposition 3.3.2 **(d)** yields $7^7 = \left(7^7//4\right) \cdot 4 + \left(7^7\%4\right) = 4\left(7^7//4\right) + \left(7^7\%4\right)$). Thus, let us set $a = 7^7//4$ and $b = 7^7\%4$. As we have seen, we have $7^{7^7}\%10 = 7^b\%10$. It remains to compute $7^b\%10$.

We first compute b: We have $7 \equiv -1 \bmod 4$ and thus $7^7 \equiv (-1)^7 = -1 \bmod 4$, so that $7^7\%4 = (-1)\%4 = 3$. (Of course, we could also have used Theorem 9.1.13 to obtain this, in the same way as we used it to show $7^{7^7}\%10 = 7^b\%10$. But powers of -1 are easy enough to take by hand!) Hence, $b = 7^7\%4 = 3$ and thus

$$7^b = 7^3 = \underbrace{7^2}_{=49 \equiv -1 \bmod 10} \cdot 7 \equiv (-1) \cdot 7 \equiv 3 \bmod 10,$$

so that $7^b\%10 = 3\%10 = 3$. Hence,

$$7^{7^7}\%10 = 7^b\%10 = 3.$$

In other words, the last digit of 7^{7^7} is 3. □

See [19s, Exercise 2.15.1] for another similar application of Euler's theorem.

Theorem 9.1.13 can also be used to explain why certain rational numbers (such as $\frac{2}{7} = 0.\overline{285714}$ [279]) have purely periodic decimal expansions, while others (such as $\frac{1}{12} = 0.08\overline{3} = 0.0833333\ldots$ or $\frac{1}{2} = 0.5\overline{0} = 0.50000\ldots$) have their periods start only after some initial nonrepeating block. We refer [ConradE, §4] to the details of this.[280]

Let us now outline a proof of Theorem 9.1.13:

Proof of Theorem 9.1.13 (sketched). (See [19s, proof of Theorem 2.15.3] for details.) It is easy to see (using Proposition 3.4.4) that $\gcd(0,n) = \gcd(n,n)$. Hence, $0 \perp n$ holds if and only if $n \perp n$.

Set

$$C = \{i \in \{0,1,\ldots,n-1\} \mid i \perp n\}.$$

A moment's thought reveals that $|C| = \phi(n)$ [281].

Now, define the integer

$$z = \prod_{i \in C} i. \tag{404}$$

All factors i in the product $\prod_{i \in C} i$ are coprime to n (by the definition of C); thus, the product itself is also coprime to n (by Exercise 3.5.4). In other words, $z \perp n$.

[279] The bar ($\overline{}$) over the "285714" means that we are repeating 285714 over and over. So $0.\overline{285714} = 0.285714285714285714\ldots$.

[280] Here is the rule, in a nutshell: A fraction $\frac{a}{b}$ (where a and b are two integers with $b \neq 0$) has a purely periodic decimal expansion if and only if $b \perp 10$ (in other words, $2 \nmid b$ and $5 \nmid b$); otherwise, it has an eventually periodic decimal expansion (which may be a finite decimal expansion, such as $\frac{1}{2} = 0.5 = 0.5\overline{0}$). This can be proven using Theorem 9.1.13.

[281] *Proof.* The definition of C yields

$$\begin{aligned}
|C| &= (\#\text{ of all } i \in \{0,1,\ldots,n-1\} \text{ satisfying } i \perp n) \\
&= \underbrace{(\#\text{ of all } i \in \{0\} \text{ satisfying } i \perp n)}_{\substack{= \begin{cases} 1, & \text{if } 0 \perp n; \\ 0, & \text{else} \end{cases} \\ = \begin{cases} 1, & \text{if } n \perp n; \\ 0, & \text{else} \end{cases} \\ \text{(since } 0 \perp n \text{ holds if and only if } n \perp n)}} + (\#\text{ of all } i \in \{1,2,\ldots,n-1\} \text{ satisfying } i \perp n) \\
&\qquad \left(\begin{array}{c} \text{by the sum rule, since each } i \in \{0,1,\ldots,n-1\} \\ \text{satisfies either } i \in \{0\} \text{ or } i \in \{1,2,\ldots,n-1\} \text{ (but not both)} \end{array}\right) \\
&= \begin{cases} 1, & \text{if } n \perp n; \\ 0, & \text{else} \end{cases} + (\#\text{ of all } i \in \{1,2,\ldots,n-1\} \text{ satisfying } i \perp n) \\
&= (\#\text{ of all } i \in \{1,2,\ldots,n-1\} \text{ satisfying } i \perp n) + \begin{cases} 1, & \text{if } n \perp n; \\ 0, & \text{else.} \end{cases}
\end{aligned}$$

Our plan is to prove the congruence $za^{\phi(n)} \equiv z \bmod n$. Once it is shown, we will be able to cancel z from this congruence (by Lemma 3.5.11, since $z \perp n$), thus obtaining $a^{\phi(n)} \equiv 1 \bmod n$.

In order to implement this plan, we will compute the product $\prod_{i \in C} (ai)$ in two ways.

On one hand, we have

$$\prod_{i \in C} (ai) = a^{|C|} \underbrace{\prod_{i \in C} i}_{=z} = a^{|C|} z = za^{|C|} = za^{\phi(n)} \tag{405}$$

(since $|C| = \phi(n)$).

On the other hand, every $i \in C$ satisfies $ai \equiv (ai) \%n \bmod n$ (since Proposition 3.3.2 **(a)** (applied to $u = ai$) yields $(ai) \%n \equiv ai \bmod n$). Multiplying these congruences over all $i \in C$, we obtain

$$\prod_{i \in C} (ai) \equiv \prod_{i \in C} ((ai) \%n) \bmod n. \tag{406}$$

Now, we claim that the product $\prod_{i \in C} ((ai) \%n)$ on the right hand side of this equality has the same factors as the product $\prod_{i \in C} i$, just in a different order. In other words, the map

$$\begin{aligned} C &\to C, \\ i &\mapsto (ai) \%n \end{aligned}$$

is bijective. In order to prove this, we must first show that this map is well-defined:

Claim 1: We have $(ai) \%n \in C$ for each $i \in C$.

Comparing this with

$$\begin{aligned} \phi(n) &= (\#\text{ of all } i \in \{1, 2, \ldots, n\} \text{ satisfying } i \perp n) \qquad \text{(by the definition of } \phi(n)) \\ &= (\#\text{ of all } i \in \{1, 2, \ldots, n-1\} \text{ satisfying } i \perp n) + \underbrace{(\#\text{ of all } i \in \{n\} \text{ satisfying } i \perp n)}_{= \begin{cases} 1, & \text{if } n \perp n; \\ 0, & \text{else} \end{cases}} \\ &\qquad \text{(by the sum rule)} \\ &= (\#\text{ of all } i \in \{1, 2, \ldots, n-1\} \text{ satisfying } i \perp n) + \begin{cases} 1, & \text{if } n \perp n; \\ 0, & \text{else} \end{cases}, \end{aligned}$$

we obtain

$$|C| = \phi(n).$$

[*Proof of Claim 1:* Let $i \in C$. Proposition 3.3.2 **(a)** (applied to $u = ai$) yields that $(ai)\,\%n \in \{0, 1, \ldots, n-1\}$.

From $a \perp n$ and $i \perp n$, we obtain $ai \perp n$ (by Theorem 3.5.10). In other words, $n \perp ai$. However, Proposition 3.4.4 **(e)** yields $\gcd(n, ai) = \gcd(n, (ai)\,\%n)$, so that $\gcd(n, (ai)\,\%n) = \gcd(n, ai) = 1$ (since $n \perp ai$) and thus $n \perp (ai)\,\%n$. In other words, $(ai)\,\%n \perp n$. Combining this with $(ai)\,\%n \in \{0, 1, \ldots, n-1\}$, we obtain $(ai)\,\%n \in C$ (by the definition of C). This proves Claim 1.]

Thus, we can define a map

$$\begin{aligned} f : C &\to C, \\ i &\mapsto (ai)\,\%n. \end{aligned}$$

Consider this map f.

Claim 2: The map f is injective.

[*Proof of Claim 2:* Let i and j be two elements of C such that $f(i) = f(j)$. We must prove that $i = j$.

We have $f(i) = f(j)$. In view of $f(i) = (ai)\,\%n$ (by the definition of f) and $f(j) = (aj)\,\%n$, this rewrites as $(ai)\,\%n = (aj)\,\%n$. Because of Proposition 3.3.4 (applied to $u = ai$ and $v = aj$), this entails that $ai \equiv aj \bmod n$. By Lemma 3.5.11, we can "cancel" a from this congruence (since $a \perp n$), and obtain $i \equiv j \bmod n$. However, both i and j belong to C and thus belong to $\{0, 1, \ldots, n-1\}$ (by the definition of C). Hence, from $i \equiv j \bmod n$, we can easily obtain that $i = j$ [282].

Now, forget that we fixed i and j. We thus have proven that if i and j are two elements of C such that $f(i) = f(j)$, then $i = j$. In other words, f is injective.]

Claim 3: The map f is a bijection.

[*Proof:* Claim 2 shows that f is injective. Hence, Corollary 6.2.9 **(a)** (applied to $X = C$) shows that f is a permutation of C. Thus, f is bijective. This proves Claim 3.]

Now, we know that f is a bijection from C to C. Thus, we can substitute $f(i)$ for i in the product $\prod_{i \in C} i$. So we obtain

$$\prod_{i \in C} i = \prod_{i \in C} \underbrace{f(i)}_{\substack{=(ai)\%n \\ \text{(by the definition of } f)}} = \prod_{i \in C} ((ai)\,\%n). \tag{407}$$

Hence, (406) becomes

$$\begin{aligned} \prod_{i \in C} (ai) &\equiv \prod_{i \in C} ((ai)\,\%n) = \prod_{i \in C} i \qquad \text{(by (407))} \\ &= z \bmod n. \end{aligned}$$

[282] *Proof.* Proposition 3.3.2 **(c)** (applied to $u = j$ and $c = i$) yields $i = j\%n$ (since $i \equiv j \bmod n$ and $i \in \{0, 1, \ldots, n-1\}$). Also, Proposition 3.3.2 **(c)** (applied to $u = j$ and $c = j$) yields $j = j\%n$ (since $j \equiv j \bmod n$ and $j \in \{0, 1, \ldots, n-1\}$). Hence, $i = j\%n = j$.

Now, (405) leads to

$$za^{\phi(n)} = \prod_{i \in C} (ai) \equiv z = z \cdot 1 \bmod n.$$

We can cancel z from this congruence (by Lemma 3.5.11, since $z \perp n$); thus, we obtain $a^{\phi(n)} \equiv 1 \bmod n$. This proves Theorem 9.1.13. □

Having proved Euler's theorem, we can now derive Fermat's Little Theorem from it:

Second proof of Theorem 9.1.10 (sketched). (See [19s, §2.15.2, Proof of Theorem 2.15.1] for details.)

(a) Let $\phi(p)$ denote the number of all $i \in \{1, 2, \ldots, p\}$ satisfying $i \perp p$. Then, $\phi(p) = p - 1$.

[*Proof.* We must show that there are precisely $p - 1$ numbers $i \in \{1, 2, \ldots, p\}$ satisfying $i \perp p$. In other words, we must show that exactly $p - 1$ of the p numbers $1, 2, \ldots, p$ are coprime to p. But this is easy: The $p - 1$ numbers $1, 2, \ldots, p - 1$ are coprime to p (by Proposition 9.1.5), whereas the remaining number p is not coprime to p (since $\gcd(p, p) = p > 1$). Thus, $\phi(p) = p - 1$ is proven.]

Now, assume that $p \nmid a$. Thus, $p \perp a$ (by Proposition 9.1.6). In other words, $a \perp p$. Hence, Theorem 9.1.13 (applied to $n = p$) yields $a^{\phi(p)} \equiv 1 \bmod p$. In other words, $a^{p-1} \equiv 1 \bmod p$ (since $\phi(p) = p - 1$). This proves Theorem 9.1.10 **(a)**.

(b) We must prove that $a^p \equiv a \bmod p$. In other words, we must prove that $p \mid a^p - a$. In other words, we must prove that $p \mid a\left(a^{p-1} - 1\right)$ (since $a^p - a = a\left(a^{p-1} - 1\right)$). If $p \mid a$, then this is obvious (since $p \mid a$ entails $p \mid a \mid a\left(a^{p-1} - 1\right)$). Hence, for the rest of this proof, we WLOG assume that $p \nmid a$. Thus, Theorem 9.1.10 **(a)** yields $a^{p-1} \equiv 1 \bmod p$. In other words, $p \mid a^{p-1} - 1$. Hence, $p \mid a^{p-1} - 1 \mid a\left(a^{p-1} - 1\right)$. But as we have explained, this completes the proof of Theorem 9.1.10 **(b)**. □

The number $\phi(n)$ defined in Theorem 9.1.13 has further properties; thus, let us introduce a name for it:

Definition 9.1.14. We define a function $\phi : \{1, 2, 3, \ldots\} \to \mathbb{N}$ as follows: For each $n \in \{1, 2, 3, \ldots\}$, we let $\phi(n)$ be the number of all $i \in \{1, 2, \ldots, n\}$ that are coprime to n. (This is precisely how $\phi(n)$ was defined in Theorem 9.1.13 and in Exercise 4.5.6.)

This function ϕ is called *Euler's totient function* or just *ϕ-function*.

Example 9.1.15. **(a)** We have $\phi(12) = 4$, since the number of all $i \in \{1, 2, \ldots, 12\}$ that are coprime to 12 is 4 (indeed, these i are 1, 5, 7 and 11).

(b) We have $\phi(13) = 12$, since the number of all $i \in \{1, 2, \ldots, 13\}$ that are coprime to 13 is 12 (indeed, these i are $1, 2, \ldots, 12$).

(c) We have $\phi(14) = 6$, since the number of all $i \in \{1, 2, \ldots, 14\}$ that are coprime to 14 is 6 (indeed, these i are $1, 3, 5, 9, 11, 13$).

(d) We have $\phi(1) = 1$, since the number of all $i \in \{1, 2, \ldots, 1\}$ that are coprime to 1 is 1 (indeed, the only such i is 1).

We will later come back to Euler's totient function ϕ, to give an "explicit" formula for it (using the prime factorization). For now, we observe that the sequence $(\phi(1), \phi(2), \phi(3), \ldots)$ is Sequence A000010 in the OEIS (where, as usual, lots of its properties and references can be found).

9.2. The Fundamental Theorem of Arithmetic

One of the leitmotifs in mathematics is the study of decompositions of objects into smaller, simpler objects. For example, here are a few ways to decompose the positive integer 14 into a sum of smaller positive integers:

$$14 = 1 + 13 = 7 + 7 = 3 + 5 + 6 = 2 + 3 + 4 + 5.$$

Almost all such decompositions can be decomposed further ("refined") by breaking up some of their addends into even smaller numbers: For example, $14 = 7 + 7$ can be "refined" by breaking up the first 7 as $2 + 5$, thus obtaining the decomposition $14 = 2 + 5 + 7$. The only decomposition of 14 that cannot be decomposed further is $14 = \underbrace{1 + 1 + \cdots + 1}_{14 \text{ times}}$, since 1 cannot be written as a sum of more than one positive integer. This phenomenon clearly holds not just for 14, but for any other positive integer n: If we decompose n into a sum of positive integers, then we can "refine" this decomposition until we arrive at $n = \underbrace{1 + 1 + \cdots + 1}_{n \text{ times}}$.

Things become more interesting if we instead try to decompose a positive integer n into a **product** of positive integers. For example, for $n = 2020$, we have the decompositions

$$2020 = 20 \cdot 101 = 4 \cdot 505 = 2 \cdot 2 \cdot 505 = 1 \cdot 2 \cdot 2 \cdot 505$$

and many others. In particular, we can always insert a factor equal to 1 into any such decomposition (since $1a = a$ for each $a \in \mathbb{Z}$), so we can find infinitely many such decompositions. Thus, let us restrict ourselves to considering only the decompositions with no factor equal to 1. By trial and error, we eventually find the decomposition

$$2020 = 2 \cdot 2 \cdot 5 \cdot 101,$$

which cannot be "refined" since the numbers $2, 2, 5, 101$ are prime and thus cannot be further decomposed. Recall that a prime number is precisely an integer $p > 1$ that cannot be decomposed into a product of two integers > 1. Thus, a decomposition of n into a product of positive integers > 1 cannot be further decomposed if and only if all its factors are primes. Such a decomposition (or, to be pedantic, the list of its factors) is called a *prime factorization* of n. Thus, for example, the decomposition $2020 = 2 \cdot 2 \cdot 5 \cdot 101$ (or, to be pedantic, the list $(2, 2, 5, 101)$) is a prime factorization of 2020. Likewise, $2021 = 43 \cdot 47$ is a prime factorization of 2021.

It is worth memorizing these prime factorizations if you plan on participating in any mathematical contests in the years 2020 and 2021, as there is a tradition in contests to

involve the current year as a number in problems. For example, problem B6 on the Putnam contest 2017 asked for the # of all 64-tuples $(x_0, x_1, \ldots, x_{63})$ with $x_0, x_1, \ldots, x_{63}$ being distinct elements of the set $[2017]$ and satisfying $2017 \mid x_0 + x_1 + 2x_2 + 3x_3 + \cdots + 63x_{63}$. The solution relied heavily on the fact that 2017 is prime. Problem A3 on the Putnam contest 2015 relied on the fact that $2015 = 5 \cdot 13 \cdot 31$ is a prime factorization of 2015. Problem B1 on the Putnam contest 2013 involved the number 2013, but only used the fact that it is odd. While there are many problems in which the year number could be replaced by any positive integer, my experience suggests that whenever some specific property of the number is used, it is usually easy to read off that property from the prime factorization of the number.

Now, it is natural to ask two questions: Does any positive integer n have a prime factorization, and is it unique? It is easy to see that the answer to the first question is "yes" (essentially because you can start with the trivial decomposition $n = n$ and then keep "refining" it until it is no longer possible; all you need to show is that this cannot go on forever, but this can easily be done using a monovariant). If you are pedantic, then the answer to the second question is "no": For example, the number 12 has the three prime factorizations $2 \cdot 2 \cdot 3$, $2 \cdot 3 \cdot 2$ and $3 \cdot 2 \cdot 2$. But these three prime factorizations only differ in the order of their factors; they are not distinct in any useful sense. Formally speaking, these factorizations (written as lists $(2, 2, 3)$, $(2, 3, 2)$ and $(3, 2, 2)$) are *permutations* of each other. Let us define this notion formally:[283]

Definition 9.2.1. Let $(p_1, p_2, \ldots, p_k)$ be a k-tuple. A *permutation* of $(p_1, p_2, \ldots, p_k)$ means a k-tuple of the form $\left(p_{\sigma(1)}, p_{\sigma(2)}, \ldots, p_{\sigma(k)}\right)$ where σ is a permutation of the set $[k]$. A permutation of $(p_1, p_2, \ldots, p_k)$ is also known as a *rearrangement* of $(p_1, p_2, \ldots, p_k)$.

Informally speaking, a permutation of a k-tuple $\mathbf{p}$ means a k-tuple obtained from $\mathbf{p}$ by rearranging its entries. Thus, two k-tuples $\mathbf{p}$ and $\mathbf{q}$ are permutations of each other if and only if they are "equal up to order", i.e., if they differ only in the order of their entries.

Example 9.2.2. (a) The 4-tuple $(2, 1, 1, 3)$ is a permutation of the 4-tuple $(3, 1, 2, 1)$. In fact, if we denote the 4-tuple $(3, 1, 2, 1)$ by (p_1, p_2, p_3, p_4), then there exists a permutation σ of the set $[4]$ such that $(2, 1, 1, 3) = \left(p_{\sigma(1)}, p_{\sigma(2)}, p_{\sigma(3)}, p_{\sigma(4)}\right)$. (Actually, there exist two such permutations σ: One of them sends $1, 2, 3, 4$ to $3, 2, 4, 1$, while the other sends $1, 2, 3, 4$ to $3, 4, 2, 1$.)

(b) Any k-tuple is a permutation of itself. Indeed, if $(p_1, p_2, \ldots, p_k)$ is any k-tuple, then $(p_1, p_2, \ldots, p_k) = \left(p_{\sigma(1)}, p_{\sigma(2)}, \ldots, p_{\sigma(k)}\right)$ if we let σ be the identity map $\mathrm{id} : [k] \to [k]$.

[283] We will use the notation from Definition 7.2.1; thus, $[k]$ means the set $\{1, 2, \ldots, k\}$.

(c) Let $\mathbf{p}$ and $\mathbf{q}$ be two k-tuples. Then, $\mathbf{p}$ is a permutation of $\mathbf{q}$ if and only if $\mathbf{q}$ is a permutation of $\mathbf{p}$. (This follows easily from the fact that the inverse of a permutation of $[k]$ is again a permutation of $[k]$.)

Next, for the sake of consistency, let us formally repeat our above definition of a prime factorization:

Definition 9.2.3. Let n be a positive integer. A *prime factorization* of n means a tuple $(p_1, p_2, \ldots, p_k)$ of primes such that $n = p_1 p_2 \cdots p_k$.

Example 9.2.4. **(a)** The prime factorizations of 12 are

$$(2,2,3), \qquad (2,3,2), \qquad (3,2,2).$$

(b) If p is a prime, then the only prime factorization of p is the 1-tuple (p).

(c) If p is a prime and $i \in \mathbb{N}$, then the only prime factorization of p^i is the i-tuple $\left(\underbrace{p, p, \ldots, p}_{i \text{ times}}\right)$. Indeed, this i-tuple clearly is a prime factorization of p^i; the fact that it is the only such factorization follows readily from Theorem 9.2.5 **(b)** below.

(d) The only prime factorization of 1 is the 0-tuple $()$.

We are finally ready to state and prove the so-called *Fundamental Theorem of Arithmetic*:

Theorem 9.2.5. Let n be a positive integer.

(a) There exists a prime factorization of n.

(b) Any two prime factorizations of n are permutations of each other.

We shall soon sketch a proof of this theorem; first, let us introduce a notation (which will only be used in this section):

Definition 9.2.6. Let L be the set of all finite lists of positive integers. We define a binary relation $\sim$ on L as follows: Given two lists $\mathbf{p}$ and $\mathbf{q}$ of positive integers, we set $\mathbf{p} \sim \mathbf{q}$ if and only if the list $\mathbf{p}$ is a permutation of $\mathbf{q}$.

Thus, Example 9.2.2 **(a)** yields $(2,1,1,3) \sim (3,1,2,1)$.

The following is easy to see:

Proposition 9.2.7. **(a)** The relation $\sim$ defined in Definition 9.2.6 is an equivalence relation (i.e., it is transitive, reflexive and symmetric).

(b) Let $(a_1, a_2, \ldots, a_k)$ and $(b_1, b_2, \ldots, b_\ell)$ be two lists of positive integers. If $(a_1, a_2, \ldots, a_k) \sim (b_1, b_2, \ldots, b_\ell)$, then $k = \ell$.

(c) Let $(a_1, a_2, \ldots, a_k)$ and $(b_1, b_2, \ldots, b_\ell)$ be two lists of positive integers satisfying $(a_1, a_2, \ldots, a_k) \sim (b_1, b_2, \ldots, b_\ell)$. Then, $(a_1, a_2, \ldots, a_k, c) \sim (b_1, b_2, \ldots, b_\ell, c)$ for each positive integer c.

Of course, nothing in Proposition 9.2.7 is specific to lists of positive integers; all claims here are general properties of lists of any objects (and follow from general properties of permutations, such as the fact that the composition of two permutations is a permutation).

We can now prove Theorem 9.2.5:

Proof of Theorem 9.2.5 (sketched). **(a)** We already outlined one way to prove Theorem 9.2.5 **(a)** above (viz., we start with the decomposition $n = n$ and keep "refining" it by decomposing factors into products, until this is no longer possible). Let me now show a slicker proof (taken from [19s, proof of Proposition 2.13.10]). We proceed by strong induction on n:

Induction step: Let m be a positive integer. Assume (as the induction hypothesis) that there exists a prime factorization of n for each positive integer $n < m$. We must prove that there exists a prime factorization of m.

If $m = 1$, then this is obvious (indeed, the 0-tuple $()$ is a prime factorization of m in this case, because $m = 1 = (\text{empty product})$). Thus, for the rest of this proof, we WLOG assume that $m \neq 1$. Hence, $m > 1$. Thus, Proposition 9.1.4 (applied to $n = m$) shows that there exists at least one prime p such that $p \mid m$. Consider this p. Now, $p > 1$ (since p is prime), so that $m/p < m/1 = m$. Furthermore, m/p is an integer (since $p \mid m$) and positive (since $m > 1 > 0$ and $p > 1 > 0$). Thus, m/p is a positive integer satisfying $m/p < m$. Hence, our induction hypothesis yields that there exists a prime factorization of m/p. Let $(q_1, q_2, \ldots, q_k)$ be this prime factorization. Thus, $q_1, q_2, \ldots, q_k$ are primes satisfying $m/p = q_1 q_2 \cdots q_k$. Multiplying both sides of the latter equality by p, we obtain $m = p q_1 q_2 \cdots q_k$. Hence, $(p, q_1, q_2, \ldots, q_k)$ is a prime factorization of m (since p is a prime, and since $q_1, q_2, \ldots, q_k$ are primes). Thus, there exists a prime factorization of m. This completes the induction step; thus, Theorem 9.2.5 **(a)** is proved.

(b) We proceed by strong induction on n:

Induction step: Let m be a positive integer. Assume (as the induction hypothesis) that Theorem 9.2.5 **(b)** holds for each positive integer $n < m$. We must prove that Theorem 9.2.5 **(b)** holds for $n = m$.

Let $(p_1, p_2, \ldots, p_k)$ and $(q_1, q_2, \ldots, q_\ell)$ be two prime factorizations of m. We shall show that $(p_1, p_2, \ldots, p_k) \sim (q_1, q_2, \ldots, q_\ell)$ (using the notation of Definition 9.2.6).

The two lists $(p_1, p_2, \ldots, p_k)$ and $(q_1, q_2, \ldots, q_\ell)$ are playing symmetric roles in this claim (since the relation $\sim$ is symmetric). Hence, we can WLOG assume that $k \leq \ell$ (since we can otherwise swap these two lists). Assume this.

We must prove that $(p_1, p_2, \ldots, p_k) \sim (q_1, q_2, \ldots, q_\ell)$. If $\ell = 0$, then this is obviously true[284]. Hence, for the rest of this proof, we WLOG assume that $\ell \neq 0$. Therefore, ℓ is a positive integer. Thus, q_ℓ is well-defined.

[284] *Proof.* Assume that $\ell = 0$. Then, $(q_1, q_2, \ldots, q_\ell) = (q_1, q_2, \ldots, q_0) = ()$. Also, $k \leq \ell = 0$ and thus $k = 0$. Hence, $(p_1, p_2, \ldots, p_k) = (p_1, p_2, \ldots, p_0) = ()$. But obviously, $() \sim ()$. In other words, $(p_1, p_2, \ldots, p_k) \sim (q_1, q_2, \ldots, q_\ell)$ (since $(p_1, p_2, \ldots, p_k) = ()$ and $(q_1, q_2, \ldots, q_\ell) = ()$). Hence, we have proved that $(p_1, p_2, \ldots, p_k) \sim (q_1, q_2, \ldots, q_\ell)$ under the assumption that $\ell = 0$.

We know that $(p_1, p_2, \ldots, p_k)$ is a prime factorization of m. In other words, $(p_1, p_2, \ldots, p_k)$ is a k-tuple of primes such that $p_1 p_2 \cdots p_k = m$. In other words, $p_1, p_2, \ldots, p_k$ are primes and satisfy $p_1 p_2 \cdots p_k = m$. Similarly, $q_1, q_2, \ldots, q_\ell$ are primes and satisfy $q_1 q_2 \cdots q_\ell = m$.

The number q_ℓ is a prime (since $q_1, q_2, \ldots, q_\ell$ are primes). Thus, $q_\ell > 1$, so that $q_\ell \neq 1$. Moreover, q_ℓ is a factor of the product $q_1 q_2 \cdots q_\ell$. Thus, $q_\ell \mid q_1 q_2 \cdots q_\ell = m = p_1 p_2 \cdots p_k$ (since $p_1 p_2 \cdots p_k = m$). Hence, Proposition 9.1.9 (applied to $p = q_\ell$ and $a_i = p_i$) yields that $q_\ell \mid p_i$ for some $i \in \{1, 2, \ldots, k\}$. Consider this i. It is easy to see that $q_\ell = p_i$ [285].

Now, consider the k-tuple $(p_1, p_2, \ldots, p_{i-1}, p_{i+1}, p_{i+2}, \ldots, p_k, p_i)$ (that is, the k-tuple obtained from $(p_1, p_2, \ldots, p_k)$ by moving the i-th entry to the very end). This k-tuple is a permutation of the k-tuple $(p_1, p_2, \ldots, p_k)$ (since it is obtained by rearranging entries of the latter k-tuple). In other words,

$$(p_1, p_2, \ldots, p_{i-1}, p_{i+1}, p_{i+2}, \ldots, p_k, p_i) \sim (p_1, p_2, \ldots, p_k).$$

Since the relation $\sim$ is symmetric, this entails

$$(p_1, p_2, \ldots, p_k) \sim (p_1, p_2, \ldots, p_{i-1}, p_{i+1}, p_{i+2}, \ldots, p_k, p_i). \tag{408}$$

On the other hand, m/q_ℓ is an integer (since $q_\ell \mid m$) and is positive (since m and q_ℓ are positive) and satisfies $m/\underbrace{q_\ell}_{>1} < m/1 = m$. Thus, our induction hypothesis shows that Theorem 9.2.5 **(b)** holds for $n = m/q_\ell$. In other words, any two prime factorizations of m/q_ℓ are permutations of each other.

Let us now find two such prime factorizations to apply this fact to. We notice the following:

- The $(k-1)$-tuple $(p_1, p_2, \ldots, p_{i-1}, p_{i+1}, p_{i+2}, \ldots, p_k)$ (that is, the $(k-1)$-tuple obtained from $(p_1, p_2, \ldots, p_k)$ by removing the i-th entry) is a $(k-1)$-tuple of primes (since $p_1, p_2, \ldots, p_k$ are primes) and satisfies

  $$p_1 p_2 \cdots p_{i-1} p_{i+1} p_{i+2} \cdots p_k = m/q_\ell$$

  (since

  $$\begin{aligned} m = p_1 p_2 \cdots p_k &= \underbrace{p_i}_{=q_\ell} (p_1 p_2 \cdots p_{i-1} p_{i+1} p_{i+2} \cdots p_k) \\ &= q_\ell (p_1 p_2 \cdots p_{i-1} p_{i+1} p_{i+2} \cdots p_k) \end{aligned}$$

). In other words, this $(k-1)$-tuple $(p_1, p_2, \ldots, p_{i-1}, p_{i+1}, p_{i+2}, \ldots, p_k)$ is a prime factorization of m/q_ℓ.

[285] *Proof.* We know that q_ℓ is positive (since $q_\ell > 1 > 0$) and divides p_i (since $q_\ell \mid p_i$). Hence, q_ℓ is a positive divisor of p_i.

The number p_i is a prime (since $p_1, p_2, \ldots, p_k$ are primes). Hence, the only positive divisors of p_i are 1 and p_i (by the definition of a prime). Thus, q_ℓ must be either 1 or p_i (since q_ℓ is a positive divisor of p_i). Since we know that $q_\ell \neq 1$, we thus conclude that $q_\ell = p_i$.

- The $(\ell-1)$-tuple $(q_1, q_2, \ldots, q_{\ell-1})$ is an $(\ell-1)$-tuple of primes (since $q_1, q_2, \ldots, q_\ell$ are primes) and satisfies

$$q_1 q_2 \cdots q_{\ell-1} = m/q_\ell$$

(since $m = q_1 q_2 \cdots q_\ell = q_\ell (q_1 q_2 \cdots q_{\ell-1})$). In other words, this $(\ell-1)$-tuple $(q_1, q_2, \ldots, q_{\ell-1})$ is a prime factorization of m/q_ℓ.

Thus, we have found two prime factorizations of m/q_ℓ: namely,

$$(p_1, p_2, \ldots, p_{i-1}, p_{i+1}, p_{i+2}, \ldots, p_k) \qquad \text{and} \qquad (q_1, q_2, \ldots, q_{\ell-1}).$$

Hence, we conclude that these two prime factorizations must be permutations of each other (since any two prime factorizations of m/q_ℓ are permutations of each other). In other words,

$$(p_1, p_2, \ldots, p_{i-1}, p_{i+1}, p_{i+2}, \ldots, p_k) \sim (q_1, q_2, \ldots, q_{\ell-1}).$$

Hence, Proposition 9.2.7 **(c)** (applied to $(p_1, p_2, \ldots, p_{i-1}, p_{i+1}, p_{i+2}, \ldots, p_k)$, $(q_1, q_2, \ldots, q_{\ell-1})$ and p_i instead of $(a_1, a_2, \ldots, a_k)$, $(b_1, b_2, \ldots, b_\ell)$ and c) yields that

$$(p_1, p_2, \ldots, p_{i-1}, p_{i+1}, p_{i+2}, \ldots, p_k, p_i) \sim (q_1, q_2, \ldots, q_{\ell-1}, p_i).$$

Now, (408) becomes[286]

$$\begin{aligned}
(p_1, p_2, \ldots, p_k) &\sim (p_1, p_2, \ldots, p_{i-1}, p_{i+1}, p_{i+2}, \ldots, p_k, p_i) \\
&\sim (q_1, q_2, \ldots, q_{\ell-1}, p_i) \\
&= (q_1, q_2, \ldots, q_{\ell-1}, q_\ell) \qquad (\text{since } p_i = q_\ell) \\
&= (q_1, q_2, \ldots, q_\ell).
\end{aligned}$$

In other words, the two prime factorizations $(p_1, p_2, \ldots, p_k)$ and $(q_1, q_2, \ldots, q_\ell)$ are permutations of each other.

Forget that we fixed $(p_1, p_2, \ldots, p_k)$ and $(q_1, q_2, \ldots, q_\ell)$. We thus have shown that any two prime factorizations $(p_1, p_2, \ldots, p_k)$ and $(q_1, q_2, \ldots, q_\ell)$ of m are permutations of each other. In other words, Theorem 9.2.5 **(b)** holds for $n = m$. This completes the induction proof of Theorem 9.2.5 **(b)**. □

In Section 9.3, we will introduce p-valuations, which will allow us to get real mileage out of Theorem 9.2.5. For now, let us show a simple application of Theorem 9.2.5 **(a)**:

Exercise 9.2.1. Prove that there are infinitely many primes p that satisfy $p \equiv 2 \bmod 3$.

[286] The following manipulations tacitly use the fact that the relation $\sim$ is an equivalence relation.

Solution to Exercise 9.2.1 (sketched). The following solution is rather similar to our first proof of Theorem 9.1.3 above, but is somewhat complicated by the need to keep track of the $p \equiv 2 \bmod 3$ condition.

We define a $2 \bmod 3$-*prime* to mean a prime p that satisfies $p \equiv 2 \bmod 3$. Thus, we need to show that there are infinitely many $2 \bmod 3$-primes.

Let $(p_1, p_2, \ldots, p_k)$ be any finite list of $2 \bmod 3$-primes. We shall find a new $2 \bmod 3$-prime p that is distinct from $p_1, p_2, \ldots, p_k$.

Indeed, set $n = 3p_1p_2 \cdots p_k - 1$. It is easy to see that $n > 0$. Hence, Theorem 9.2.5 **(a)** shows that there exists a prime factorization of n. Let $(q_1, q_2, \ldots, q_\ell)$ be this prime factorization. Thus, $(q_1, q_2, \ldots, q_\ell)$ is an ℓ-tuple of primes such that $n = q_1q_2 \cdots q_\ell$. In other words, $q_1, q_2, \ldots, q_\ell$ are primes and satisfy $n = q_1q_2 \cdots q_\ell$.

We now claim that at least one of the primes $q_1, q_2, \ldots, q_\ell$ must be a $2 \bmod 3$-prime.

[*Proof:* Assume the contrary. Thus, none of the primes $q_1, q_2, \ldots, q_\ell$ is a $2 \bmod 3$-prime.

Note that $n = \underbrace{3p_1p_2 \cdots p_k}_{\equiv 0 \bmod 3} - 1 \equiv 0 - 1 = -1 \equiv 2 \bmod 3$. Hence, $n \equiv 2 \not\equiv 0 \bmod 3$, so that $3 \nmid n$.

Let $i \in \{1, 2, \ldots, \ell\}$. Then, q_i is not a $2 \bmod 3$-prime (since none of the primes $q_1, q_2, \ldots, q_\ell$ is a $2 \bmod 3$-prime). In other words, $q_i \not\equiv 2 \bmod 3$ (since q_i is a prime). Moreover, $q_i \mid n$ (since $n = q_1q_2 \cdots q_\ell$) and thus $q_i \not\equiv 0 \bmod 3$ (because if we had $q_i \equiv 0 \bmod 3$, then we would have $3 \mid q_i \mid n$, which would contradict $3 \nmid n$).

However, the remainder $q_i \% 3$ is either 0, 1 or 2 (since $q_i \% 3 \in \{0, 1, 2\}$). Thus, q_i must be congruent to one of the three numbers 0, 1 and 2 modulo 3 (by Proposition 3.3.4, because $0\%3 = 0$ and $1\%3 = 1$ and $2\%3 = 2$). In other words, we have $q_i \equiv 0 \bmod 3$ or $q_i \equiv 1 \bmod 3$ or $q_i \equiv 2 \bmod 3$. Since we know that $q_i \not\equiv 0 \bmod 3$ and $q_i \not\equiv 2 \bmod 3$, we thus conclude that $q_i \equiv 1 \bmod 3$.

Now, forget that we fixed i. We thus have shown that $q_i \equiv 1 \bmod 3$ for each $i \in \{1, 2, \ldots, \ell\}$. Hence, $\prod_{i=1}^{\ell} \underbrace{q_i}_{\equiv 1 \bmod 3} \equiv \prod_{i=1}^{\ell} 1 = 1 \bmod 3$. Therefore, $n = q_1q_2 \cdots q_\ell = \prod_{i=1}^{\ell} q_i \equiv 1 \bmod 3$. This contradicts $n \equiv 2 \not\equiv 1 \bmod 3$. This contradiction shows that our assumption was false. Hence, at least one of the primes $q_1, q_2, \ldots, q_\ell$ must be a $2 \bmod 3$-prime.]

Now we have shown that at least one of the primes $q_1, q_2, \ldots, q_\ell$ must be a $2 \bmod 3$-prime. In other words, there exists some $j \in \{1, 2, \ldots, \ell\}$ such that q_j is a $2 \bmod 3$-prime. Consider this j. Clearly, $q_j \mid n$ (since $n = q_1q_2 \cdots q_\ell$).

Set $p = q_j$. Thus, p is a $2 \bmod 3$-prime (since q_j is a $2 \bmod 3$-prime) and satisfies $p = q_j \mid n$.

For each $i \in \{1, 2, \ldots, k\}$, we have $p_i \nmid -1$ (since p_i is a prime, so that $p_i > 1$) and

thus

$$n = 3 \underbrace{p_1 p_2 \cdots p_k}_{\substack{\equiv 0 \bmod p_i \\ (\text{since } p_i \mid p_1 p_2 \cdots p_k)}} - 1 \equiv 0 - 1 = -1 \not\equiv 0 \bmod p_i \qquad (\text{since } p_i \nmid -1 = (-1) - 0).$$

In other words, for each $i \in \{1, 2, \ldots, k\}$, we have $p_i \nmid n$ and thus $p \neq p_i$ (since $p = p_i$ would yield $p = p_i \nmid n$, which would contradict $p \mid n$). In other words, p is distinct from $p_1, p_2, \ldots, p_k$.

Forget that we fixed $(p_1, p_2, \ldots, p_k)$. Thus, for any finite list $(p_1, p_2, \ldots, p_k)$ of $2 \bmod 3$-primes, we have found a new $2 \bmod 3$-prime p that is distinct from $p_1, p_2, \ldots, p_k$. In other words, for any finite list of $2 \bmod 3$-primes, we have found a $2 \bmod 3$-prime that is not in this list. Thus, there are infinitely many $2 \bmod 3$-primes. This solves Exercise 9.2.1. □

9.3. p-valuations

9.3.1. The p-valuation of an integer

The prime factorization of a positive integer n provides an easy way to tell how often n can be divided by p without remainder (i.e., how high a power of p divides n): namely, as often as p appears in this prime factorization. This number is commonly known as the *p-valuation* of n. Let us define it in slightly greater generality (for any integer n, not just positive integers) and in more detail ([19s, Definition 2.13.23]):

Definition 9.3.1. Let p be a prime.

(a) Let n be a nonzero integer. Then, $v_p(n)$ shall denote the largest $m \in \mathbb{N}$ such that $p^m \mid n$. This is well-defined (see [19s, Lemma 2.13.22] for a detailed proof). This nonnegative integer $v_p(n)$ will be called the *p-valuation* (or the *p-adic valuation*) of n.

(b) We extend this definition of $v_p(n)$ to the case of $n = 0$ as follows: Set $v_p(0) = \infty$, where ∞ is a new symbol. This symbol ∞ is supposed to model the concept of "positive infinity"; in particular, we extend some of the standard arithmetic operations to ∞ according to the following rules:

- We set $k + \infty = \infty + k = \infty$ for all integers k.
- We set $\infty + \infty = \infty$.
- For each integer k, we declare the inequalities $k < \infty$ and $\infty > k$ and $k \leq \infty$ and $\infty \geq k$ to be true, and the inequalities $k \geq \infty$ and $\infty \leq k$ and $k > \infty$ and $\infty < k$ to be false.
- If S is a nonempty set of integers, then we set $\min(S \cup \{\infty\}) = \min S$ (provided that $\min S$ exists).

- We set $\min\{\infty\} = \infty$.
- If S is any set of integers, then we set $\max(S \cup \{\infty\}) = \infty$.

(This being said, ∞ is not supposed to be a "first class citizen" of the number system. In particular, differences of the form $k - \infty$ are not defined, since any definition of $k - \infty$ would break some of the familiar rules of arithmetic. The only operations that we shall subject ∞ to are addition, minimum and maximum.)

Note that the rules for the symbol ∞ yield that

$$k + \infty = \infty + k = \max\{k, \infty\} = \infty \qquad \text{and} \qquad \min\{k, \infty\} = k$$

for each $k \in \mathbb{Z} \cup \{\infty\}$. It is not hard to see that basic properties of inequalities (such as "if $a \leq b$ and $b \leq c$, then $a \leq c$") and of addition (such as "$(a+b)+c = a+(b+c)$") and of the interaction between inequalities and addition (such as "if $a \leq b$, then $a + c \leq b + c$") are still valid in $\mathbb{Z} \cup \{\infty\}$ (that is, they still hold if we plug ∞ for one or more of the variables).[287] However, of course, we cannot "cancel" ∞ from equalities (i.e., we cannot cancel ∞ from $a + \infty = b + \infty$ to obtain $a = b$) or inequalities.

Example 9.3.2. **(a)** We have $v_3(18) = 2$. Indeed, 2 is the largest $m \in \mathbb{N}$ such that $3^m \mid 18$ (because $3^2 = 9 \mid 18$ but $3^3 = 27 \nmid 18$).

(b) We have $v_3(14) = 0$. Indeed, 0 is the largest $m \in \mathbb{N}$ such that $3^m \mid 14$ (because $3^0 = 1 \mid 14$ but $3^1 = 3 \nmid 14$).

(c) We have $v_3(51) = 1$. Indeed, 1 is the largest $m \in \mathbb{N}$ such that $3^m \mid 51$ (because $3^1 = 3 \mid 51$ but $3^2 = 9 \nmid 51$).

(d) We have $v_3(0) = \infty$ (by Definition 9.3.1 **(b)**).

Definition 9.3.1 **(a)** can be restated in the following more intuitive way: Given a prime p and a nonzero integer n, we let $v_p(n)$ be the number of times we can divide n by p without leaving $\mathbb{Z}$. Definition 9.3.1 **(b)** is consistent with this restatement, because we can clearly divide 0 by p infinitely often without leaving $\mathbb{Z}$. From this point of view, the following lemma should be obvious:

Lemma 9.3.3. Let p be a prime. Let $i \in \mathbb{N}$. Let $n \in \mathbb{Z}$. Then, $p^i \mid n$ if and only if $v_p(n) \geq i$.

Proof of Lemma 9.3.3 (sketched). (See [19s, Lemma 2.13.25] for a detailed proof.) In the $n = 0$ case, we have both $p^i \mid n$ and $v_p(n) \geq i$ (since $v_p(n) = v_p(0) = \infty \geq i$). Thus, WLOG assume that $n \neq 0$. Hence, $v_p(n)$ is defined as the largest $m \in \mathbb{N}$

[287] For algebraists: The set $\mathbb{Z} \cup \{\infty\}$ equipped with the operations max and $+$ as "addition" and "multiplication" and with the relation $\leq$ is a totally ordered commutative semiring. The same is true if max is replaced by min. These two semirings are known as the *tropical semirings* over $\mathbb{Z}$. See [SpeStu09] for more about them.

such that $p^m \mid n$. Therefore, $p^{v_p(n)} \mid n$ holds, but $p^{v_p(n)+1} \mid n$ does not. From this, it is easy to see that $p^i \mid n$ holds if $v_p(n) \geq i$, but not if $v_p(n) < i$. But this is precisely the claim of Lemma 9.3.3. □

Corollary 9.3.4. Let p be a prime. Let $n \in \mathbb{Z}$. Then, $v_p(n) = 0$ if and only if $p \nmid n$.

Proof of Corollary 9.3.4 (sketched). (See [19s, Corollary 2.13.26] for a detailed proof.) Lemma 9.3.3 (applied to $i = 0$) shows that $p \mid n$ if and only if $v_p(n) \geq 1$. In other words, $p \mid n$ if and only if $v_p(n) \neq 0$ (since $v_p(n) \geq 1$ is equivalent to $v_p(n) \neq 0$). Taking the contrapositive of this claim, we obtain precisely Corollary 9.3.4. □

Using the p-valuation, we can decompose a nonzero integer into the product of a power of p and an integer coprime to p:

Lemma 9.3.5. Let p be a prime. Let $n \in \mathbb{Z}$ be nonzero. Then:
(a) There exists a nonzero integer u such that $u \perp p$ and $n = up^{v_p(n)}$.
(b) If $i \in \mathbb{N}$ and $w \in \mathbb{Z}$ are such that $w \perp p$ and $n = wp^i$, then $v_p(n) = i$.

Proof of Lemma 9.3.5 (sketched). (See [19s, Lemma 2.13.27] for a detailed proof.) The definition of $v_p(n)$ yields $v_p(n) \in \mathbb{N}$ and $p^{v_p(n)} \mid n$ and $p^{v_p(n)+1} \nmid n$.

(a) Set $u = n/p^{v_p(n)}$; thus, u is nonzero (since n is nonzero) and satisfies $n = up^{v_p(n)}$. It remains to show that $u \perp p$. But we have $pp^{v_p(n)} = p^{v_p(n)+1} \nmid n = up^{v_p(n)}$. Cancelling the (nonzero) factor $p^{v_p(n)}$ from this non-divisibility, we find $p \nmid u$, and thus $p \perp u$ (by Proposition 9.1.6, applied to $a = u$). In other words, $u \perp p$. This proves Lemma 9.3.5 **(a)**.

(b) Let $i \in \mathbb{N}$ and $w \in \mathbb{Z}$ be such that $w \perp p$ and $n = wp^i$. We must prove that $v_p(n) = i$.

Assume the contrary. Thus, $v_p(n) \neq i$. But $p^i \mid n$ (since $n = wp^i$) and thus $v_p(n) \geq i$ (by Lemma 9.3.3). Hence, $v_p(n) > i$ (since $v_p(n) \neq i$) and therefore $v_p(n) \geq i + 1$. Equivalently, we have $p^{i+1} \mid n$ (by Lemma 9.3.3, applied to $i + 1$ instead of i). In other words, $pp^i \mid wp^i$ (since $p^{i+1} = pp^i$ and $n = wp^i$). Cancelling the (nonzero) factor p^i from this divisibility, we find $p \mid w$. Hence, $\gcd(p, w) = |p| = p > 1$. On the other hand, $\gcd(p, w) = \gcd(w, p) = 1$ (since $w \perp p$). These two facts clearly contradict each other.

This contradiction shows that our assumption was false. Hence, $v_p(n) = i$ must hold. This proves Lemma 9.3.5 **(b)**. □

The next theorem is crucial for computing and bounding p-valuations:

Theorem 9.3.6. Let p be a prime.
(a) We have $v_p(ab) = v_p(a) + v_p(b)$ for any two integers a and b.
(b) We have $v_p(a + b) \geq \min\{v_p(a), v_p(b)\}$ for any two integers a and b.
(c) We have $v_p(1) = 0$.

(d) We have $v_p(q) = \begin{cases} 1, & \text{if } q = p; \\ 0, & \text{if } q \neq p \end{cases}$ for any prime q.

Note that Theorem 9.3.6 **(a)** determines $v_p(ab)$ exactly (in terms of $v_p(a)$ and $v_p(b)$), but Theorem 9.3.6 **(b)** merely gives a lower bound on $v_p(a+b)$. There is no way to improve on this, since $v_p(a)$ and $v_p(b)$ do not uniquely determine $v_p(a+b)$.

Proof of Theorem 9.3.6 (sketched). (See [19s, Theorem 2.13.28] for a more detailed proof.)

(a) Let a and b be two integers. We must prove that $v_p(ab) = v_p(a) + v_p(b)$.

If $a = 0$, then this boils down to $\infty = \infty$, which is obvious. Thus, we WLOG assume that $a \neq 0$. Likewise, we WLOG assume that $b \neq 0$. Hence, $ab \neq 0$.

Lemma 9.3.5 **(a)** (applied to $n = a$) shows that there exists a nonzero integer x such that $x \perp p$ and $a = xp^{v_p(a)}$. Likewise, there exists a nonzero integer y such that $y \perp p$ and $b = yp^{v_p(b)}$. Consider these x and y. From $x \perp p$ and $y \perp p$, we obtain $xy \perp p$ (by Theorem 3.5.10). Furthermore, multiplying the equalities $a = xp^{v_p(a)}$ and $b = yp^{v_p(b)}$, we obtain

$$ab = \left(xp^{v_p(a)}\right)\left(yp^{v_p(b)}\right) = (xy)\, p^{v_p(a)+v_p(b)}.$$

Thus, Lemma 9.3.5 **(b)** (applied to $n = ab$, $i = v_p(a) + v_p(b)$ and $w = xy$) shows that $v_p(ab) = v_p(a) + v_p(b)$. This proves Theorem 9.3.6 **(a)**.

(b) Let a and b be two integers. We must prove that $v_p(a+b) \geq \min\{v_p(a), v_p(b)\}$.

We WLOG assume that $a \neq 0$ and $b \neq 0$ (as the other cases are easy). Thus, $v_p(a) \in \mathbb{N}$ and $v_p(b) \in \mathbb{N}$.

Let $m = \min\{v_p(a), v_p(b)\}$. Thus, $m \in \mathbb{N}$ and $v_p(a) \geq m$. Hence, Lemma 9.3.3 (applied to $n = a$ and $i = m$) yields $p^m \mid a$. In other words, $a \equiv 0 \bmod p^m$. Similarly, $b \equiv 0 \bmod p^m$. Adding these two congruences together, we obtain $a + b \equiv 0 + 0 = 0 \bmod p^m$. In other words, $p^m \mid a + b$. This, in turn, leads to $v_p(a+b) \geq m$ (by Lemma 9.3.3, applied to $n = a + b$ and $i = m$). That is, $v_p(a+b) \geq m = \min\{v_p(a), v_p(b)\}$. This proves Theorem 9.3.6 **(b)**.

(c) This follows from Lemma 9.3.4 (applied to $n = 1$), since $p \nmid 1$ (because $|p| = p > 1 = |1|$).

(d) Let q be a prime. We must prove that $v_p(q) = \begin{cases} 1, & \text{if } q = p; \\ 0, & \text{if } q \neq p. \end{cases}$

This requires us to prove two things: First, we must show that $v_p(p) = 1$; second, we must show that $v_p(q) = 0$ if $q \neq p$.

Let us prove $v_p(p) = 1$ first. Indeed, $1 \perp p$ (by Exercise 3.5.1 **(a)**) and $p = 1 \cdot p^1$. Thus, Lemma 9.3.5 **(b)** (applied to $n = p$, $i = 1$ and $w = 1$) yields $v_p(p) = 1$.

It remains to show that $v_p(q) = 0$ if $q \neq p$. Thus, let us assume that $q \neq p$. Hence, Proposition 9.1.7 (applied to q and p instead of p and q) yields $q \perp p$. Also, $q = q \cdot p^0$ (since $p^0 = 1$). Therefore, Lemma 9.3.5 **(b)** (applied to $n = q$, $i = 0$ and $w = q$) yields $v_p(q) = 0$. This completes our proof of Theorem 9.3.6 **(d)**. $\square$

Corollary 9.3.7. Let p be a prime. Let $a_1, a_2, \ldots, a_k$ be k integers. Then, $v_p(a_1 a_2 \cdots a_k) = v_p(a_1) + v_p(a_2) + \cdots + v_p(a_k)$.

Proof of Corollary 9.3.7 (sketched). (See [19s, Corollary 2.13.29] for details.) This follows straightforwardly by induction on k, using Theorem 9.3.6 **(a)** (as well as Theorem 9.3.6 **(c)** for the induction base). □

The following simple properties of p-valuations are left to the reader to prove:[288]

Proposition 9.3.8. Let p be a prime. Let $n \in \mathbb{Z}$. Then, $v_p(|n|) = v_p(n)$.

Corollary 9.3.9. Let p be a prime. Let $a \in \mathbb{Z}$ and $k \in \mathbb{N}$. Then, $v_p(a^k) = k v_p(a)$.

Corollary 9.3.10. Let $p_1, p_2, \ldots, p_u$ be finitely many distinct primes. Let $a_1, a_2, \ldots, a_u$ be nonnegative integers.

(a) We have $v_{p_i}\left(p_1^{a_1} p_2^{a_2} \cdots p_u^{a_u}\right) = a_i$ for each $i \in \{1, 2, \ldots, u\}$.

(b) We have $v_p\left(p_1^{a_1} p_2^{a_2} \cdots p_u^{a_u}\right) = 0$ for each prime p satisfying $p \notin \{p_1, p_2, \ldots, p_u\}$.

Another useful property of p-valuations is how they transform gcds into minima:

Proposition 9.3.11. Let p be a prime. Let k be a positive integer. Let $n_1, n_2, \ldots, n_k$ be k integers. Then,

$$v_p(\gcd(n_1, n_2, \ldots, n_k)) = \min\{v_p(n_1), v_p(n_2), \ldots, v_p(n_k)\}.$$

We shall prove this proposition using the following innocent-looking lemma:

Lemma 9.3.12. Let a and b be two elements of $\mathbb{N} \cup \{\infty\}$. Assume that the logical equivalence

$$(a \geq j) \iff (b \geq j) \tag{409}$$

holds for each $j \in \mathbb{N}$. Then, $a = b$.

Proof of Lemma 9.3.12. Assume the contrary. Thus, $a \neq b$.

Clearly, a and b play symmetric roles in our setting (indeed, they play symmetric roles in (409), since logical equivalence is a symmetric relation). Thus, we WLOG assume that $a \geq b$ (since otherwise, we can swap a with b). Combining $a \geq b$ with $a \neq b$, we obtain $a > b$.

Hence, $b < a \leq \infty$, so that $b \neq \infty$. Hence, $b \in \mathbb{N}$ and thus $b + 1 > b$. Hence, the statement "$b \geq b + 1$" is false. However, we have $b + 1 \in \mathbb{N}$ (since $b \in \mathbb{N}$), and thus (409) (applied to $j = b + 1$) yields the logical equivalence $(a \geq b + 1) \iff$

[288] Proposition 9.3.8 is [19s, Exercise 2.13.5]. Corollary 9.3.9 is [19s, Exercise 2.13.6]. Corollary 9.3.10 is [19s, Exercise 2.13.7].

$(b \geq b+1)$. Therefore, the statement "$a \geq b+1$" is false (since the statement "$b \geq b+1$" is false). In other words, we have $a < b+1$. This entails that $a \in \mathbb{N}$ (since $b+1 \in \mathbb{N}$). Thus, a and b are integers. Hence, from $a > b$, we obtain $a \geq b+1$. However, this contradicts $a < b+1$. This contradiction shows that our assumption was false; this proves Lemma 9.3.12. $\square$

Proof of Proposition 9.3.11. Set

$$g = \gcd(n_1, n_2, \ldots, n_k) \qquad \text{and} \qquad s = \min\left\{v_p(n_1), v_p(n_2), \ldots, v_p(n_k)\right\}.$$

Note that $v_p(g) \in \mathbb{N} \cup \{\infty\}$ (since the p-valuation of any integer is an element of $\mathbb{N} \cup \{\infty\}$) and $s \in \mathbb{N} \cup \{\infty\}$ (for the same reason). Our end goal is to show that $v_p(g) = s$. First, we shall show the following:

Claim 1: Let $j \in \mathbb{N}$. Then, we have the logical equivalence

$$\left(v_p(g) \geq j\right) \iff (s \geq j).$$

[*Proof of Claim 1:* Lemma 9.3.3 (applied to $n = g$ and $i = j$) yields that $p^j \mid g$ if and only if $v_p(g) \geq j$. In other words, we have the logical equivalence

$$\left(p^j \mid g\right) \iff \left(v_p(g) \geq j\right). \tag{410}$$

On the other hand, Theorem 3.4.14 (applied to $b_i = n_i$ and $m = p^j$) yields that we have the following logical equivalence:

$$\begin{aligned} &\left(p^j \mid n_i \text{ for all } i \in \{1, 2, \ldots, k\}\right) \\ \iff &\left(p^j \mid \gcd(n_1, n_2, \ldots, n_k)\right). \end{aligned} \tag{411}$$

Hence, we have the following chain of logical equivalences:

$$\begin{aligned} \left(v_p(g) \geq j\right) &\iff \left(p^j \mid g\right) \qquad \text{(by (410))} \\ &\iff \left(p^j \mid \gcd(n_1, n_2, \ldots, n_k)\right) \qquad (\text{since } g = \gcd(n_1, n_2, \ldots, n_k)) \\ &\iff \left(p^j \mid n_i \text{ for all } i \in \{1, 2, \ldots, k\}\right) \end{aligned} \tag{412}$$

(by (411)).

Now, let $i \in \{1, 2, \ldots, k\}$. Then, Lemma 9.3.3 (applied to n_i and j instead of n and i) yields that $p^j \mid n_i$ if and only if $v_p(n_i) \geq j$. In other words, we have the logical equivalence

$$\left(p^j \mid n_i\right) \iff \left(v_p(n_i) \geq j\right). \tag{413}$$

Forget that we fixed i. We thus have proved the equivalence (413) for each $i \in \{1, 2, \ldots, k\}$.

Now, from (412), we obtain the following chain of logical equivalences:

$$\begin{aligned}
&\left(v_p(g) \geq j\right) \\
&\iff \left(\underbrace{p^j \mid n_i}_{\substack{\iff \left(v_p(n_i) \geq j\right) \\ \text{(by (413))}}} \quad \text{for all } i \in \{1, 2, \ldots, k\}\right) \\
&\iff \left(v_p(n_i) \geq j \text{ for all } i \in \{1, 2, \ldots, k\}\right) \\
&\iff \left(\text{each of the } k \text{ elements } v_p(n_1), v_p(n_2), \ldots, v_p(n_k) \text{ is } \geq j\right) \\
&\iff \left(\text{the } \textbf{smallest} \text{ of the } k \text{ elements } v_p(n_1), v_p(n_2), \ldots, v_p(n_k) \text{ is } \geq j\right) \\
&\qquad\qquad \left(\begin{array}{c}\text{because for any } k \text{ elements } a_1, a_2, \ldots, a_k \text{ of } \mathbb{N} \cup \{\infty\}, \\ \text{the statement "each of the } k \text{ elements } a_1, a_2, \ldots, a_k \text{ is } \geq j\text{"} \\ \text{is equivalent to} \\ \text{"the } \textbf{smallest} \text{ of the } k \text{ elements } a_1, a_2, \ldots, a_k \text{ is } \geq j\text{"}\end{array}\right) \\
&\iff \left(\underbrace{\min\left\{v_p(n_1), v_p(n_2), \ldots, v_p(n_k)\right\}}_{\substack{=s \\ \text{(by the definition of } s)}} \geq j\right) \\
&\iff (s \geq j).
\end{aligned}$$

This proves Claim 1.]

Now, we are almost there. Claim 1 shows that the logical equivalence

$$\left(v_p(g) \geq j\right) \iff (s \geq j)$$

holds for each $j \in \mathbb{N}$. Thus, Lemma 9.3.12 (applied to $a = v_p(g)$ and $b = s$) yields $v_p(g) = s$. In view of $g = \gcd(n_1, n_2, \ldots, n_k)$ and $s = \min\left\{v_p(n_1), v_p(n_2), \ldots, v_p(n_k)\right\}$, this rewrites as

$$v_p(\gcd(n_1, n_2, \ldots, n_k)) = \min\left\{v_p(n_1), v_p(n_2), \ldots, v_p(n_k)\right\}.$$

This proves Proposition 9.3.11. □

The technique we just used in our proof of Proposition 9.3.11 is worth crystallizing out, as it is often helpful in proving properties of p-valuations: In order to show that $v_p(g) = s$, we proved that any $j \in \mathbb{N}$ satisfies the equivalence $\left(v_p(g) \geq j\right) \iff (s \geq j)$ (this was our Claim 1). In other words, in order to show that the "numbers" $v_p(g)$ and s are equal[289], we showed that the $j \in \mathbb{N}$ that don't exceed one of them are the same $j \in \mathbb{N}$ that don't exceed the other. Thus, instead of reasoning about these "numbers" $v_p(g)$ and s themselves, we have argued about what it means for a $j \in \mathbb{N}$ to satisfy $v_p(g) \geq j$ or $s \geq j$.

[289] We are putting the word "numbers" in quotation marks, since $v_p(g)$ and s can be ∞.

9.3.2. p-valuations and prime factorizations

As already mentioned, the p-valuation of a positive integer n tells us how often the prime p appears in a prime factorization of n. Let us state this as a proposition:

Proposition 9.3.13. Let n be a positive integer. Let $(a_1, a_2, \ldots, a_k)$ be a prime factorization of n. Let p be a prime. Then,

$$\begin{aligned}&(\text{the number of times } p \text{ appears in the tuple } (a_1, a_2, \ldots, a_k))\\ &= (\text{the number of } i \in \{1, 2, \ldots, k\} \text{ such that } a_i = p)\\ &= v_p(n).\end{aligned}$$

Proof of Proposition 9.3.13 (sketched). (See [19s, Proposition 2.13.30] for details.)

From the definition of a prime factorization, we know that $(a_1, a_2, \ldots, a_k)$ is a tuple of primes such that $n = a_1 a_2 \cdots a_k$. Now, from $n = a_1 a_2 \cdots a_k$, we obtain

$$\begin{aligned}v_p(n) &= v_p(a_1 a_2 \cdots a_k) = v_p(a_1) + v_p(a_2) + \cdots + v_p(a_k) \qquad \text{(by Corollary 9.3.7)}\\ &= \sum_{i \in \{1,2,\ldots,k\}} \underbrace{v_p(a_i)}_{\substack{= \begin{cases} 1, & \text{if } a_i = p; \\ 0, & \text{if } a_i \neq p \end{cases} \\ \text{(by Theorem 9.3.6 \textbf{(d)},} \\ \text{applied to } q = a_i \text{ (since } a_i \text{ is a prime))}}} = \sum_{i \in \{1,2,\ldots,k\}} \begin{cases} 1, & \text{if } a_i = p; \\ 0, & \text{if } a_i \neq p \end{cases}\\ &= (\text{the number of } i \in \{1, 2, \ldots, k\} \text{ such that } a_i = p) \cdot 1\\ &\qquad + (\text{the number of } i \in \{1, 2, \ldots, k\} \text{ such that } a_i \neq p) \cdot 0\\ &= (\text{the number of } i \in \{1, 2, \ldots, k\} \text{ such that } a_i = p)\\ &= (\text{the number of times } p \text{ appears in the tuple } (a_1, a_2, \ldots, a_k)).\end{aligned}$$

This proves Proposition 9.3.13. □

9.3.3. The canonical factorization

Proposition 9.3.13 allows us to expand each positive integer n "explicitly" as a product of primes:

$$n = \prod_{p \text{ prime}} p^{v_p(n)} \tag{414}$$

(where the product sign "$\prod_{p \text{ prime}}$" means a product over all primes p). But first, let me explain why the product in this equality makes sense. This product is infinite (as there are infinitely many primes), so this is not a priori obvious. Fortunately, it is the simplest possible kind of an infinite product – one that has only finitely many factors different from 1. Let me quickly explain the meaning of such products. I begin by defining their additive analogues – i.e., sums that have only finitely many addends different from 0:

Definition 9.3.14. Let S be a set. Let a_s be a number for each $s \in S$. Assume that only finitely many $s \in S$ satisfy $a_s \neq 0$. (In other words, assume that the set $\{s \in S \mid a_s \neq 0\}$ is finite.) Then, the sum $\sum\limits_{s \in S} a_s$ is defined to be the finite sum $\sum\limits_{\substack{s \in S; \\ a_s \neq 0}} a_s$. (The latter sum is already defined according to Definition 4.1.1, since it is a finite sum.)

For example, the infinite sum $\sum\limits_{s \in \{1,2,3,\ldots\}} \lfloor 4/s \rfloor$ is well-defined by this definition, because only finitely many $s \in \{1,2,3,\ldots\}$ satisfy $\lfloor 4/s \rfloor \neq 0$ (indeed, $\lfloor 4/s \rfloor \neq 0$ holds only for $s \in \{1,2,3,4\}$); its value is

$$\begin{aligned}\sum_{s \in \{1,2,3,\ldots\}} \lfloor 4/s \rfloor &= \sum_{\substack{s \in \{1,2,3,\ldots\}; \\ \lfloor 4/s \rfloor \neq 0}} \lfloor 4/s \rfloor = \sum_{s \in \{1,2,3,4\}} \lfloor 4/s \rfloor \\ &= \lfloor 4/1 \rfloor + \lfloor 4/2 \rfloor + \lfloor 4/3 \rfloor + \lfloor 4/4 \rfloor = 8.\end{aligned}$$

A pedantic reader may observe that Definition 9.3.14 above might possibly clash with our old definition of finite sums (Definition 4.1.1), since it applies both to finite and infinite sets S. However, this clash is harmless, because when S is finite, Definition 9.3.14 agrees with Definition 4.1.1 (since a sum does not change when vanishing addends are removed from it).

Definition 4.1.1 allows us to make sense of infinite sums when all but finitely many of their addends are zero. Such sums are called *finitely supported sums*, and satisfy many of the nice properties of finite sums (see [Grinbe15, Subsection 2.14.15] for details). They are, in many ways, much simpler than the convergent infinite sums of analysis; in particular, they can be computed in finite time (simply by throwing away all the zero addends and summing the finitely many addends that remain).

Let us now state the multiplicative analogue of Definition 9.3.14:

Definition 9.3.15. Let S be a set. Let a_s be a number for each $s \in S$. Assume that only finitely many $s \in S$ satisfy $a_s \neq 1$. (In other words, assume that the set $\{s \in S \mid a_s \neq 1\}$ is finite.) Then, the product $\prod\limits_{s \in S} a_s$ is defined to be the finite product $\prod\limits_{\substack{s \in S; \\ a_s \neq 1}} a_s$. (The latter product is already defined according to Definition 4.2.1, since it is a finite product.)

For example,

$$2 \cdot 9 \cdot 3 \cdot \underbrace{1 \cdot 1 \cdot 1 \cdot 1 \cdot \cdots}_{\text{infinitely many 1's}} = 2 \cdot 9 \cdot 3 = 54$$

(where, of course, we are using the notation $a_1 a_2 a_3 \cdots$ for an infinite product $\prod\limits_{s \in \{1,2,3,\ldots\}} a_s$).

Convention 9.3.16. Here and in the following, the symbol "$\prod\limits_{p \text{ prime}}$" means a product ranging over all primes p. This is an infinite product, but can sometimes make sense (according to Definition 9.3.15).

Now, we can make sense of the equality (414):

Theorem 9.3.17. Let n be a nonzero integer.

(a) We have $v_p(n) = 0$ for every prime $p > |n|$. (Note that "for every prime $p > |n|$" is shorthand for "for every prime p satisfying $p > |n|$".)

(b) The product $\prod\limits_{p \text{ prime}} p^{v_p(n)}$ has only finitely many factors different from 1, and thus is well-defined.

(c) We have

$$|n| = \prod_{p \text{ prime}} p^{v_p(n)}.$$

(d) If n is positive, then

$$n = \prod_{p \text{ prime}} p^{v_p(n)}.$$

This expression $n = \prod\limits_{p \text{ prime}} p^{v_p(n)}$ is called the *canonical factorization* of n.

Proof of Theorem 9.3.17 (sketched). **(a)** (See [19s, Lemma 2.13.32 **(a)**] for details.) Let p be a prime such that $p > |n|$. Then, Proposition 3.1.3 **(b)** yields $p \nmid n$ (since $|p| = p > |n|$), and thus Corollary 9.3.4 yields $v_p(n) = 0$. This proves Theorem 9.3.17 **(a)**.

(b) (See [19s, Lemma 2.13.32 **(b)**] for details.) For every prime $p > |n|$, we have $v_p(n) = 0$ (by Theorem 9.3.17 **(a)**) and thus $p^{v_p(n)} = p^0 = 1$. Thus, a prime p can satisfy $v_p(n) \neq 0$ only if $p \leq |n|$. Therefore, only finitely many primes p satisfy $v_p(n) \neq 0$ (since only finitely many primes p satisfy $p \leq |n|$). In other words, the product $\prod\limits_{p \text{ prime}} p^{v_p(n)}$ has only finitely many factors different from 1. This proves Theorem 9.3.17 **(b)**.

(d) (See [19s, Corollary 2.13.33] for details.) Assume that n is positive. Theorem 9.2.5 **(a)** shows that there exists a prime factorization $(a_1, a_2, \ldots, a_k)$ of n. Consider this $(a_1, a_2, \ldots, a_k)$. Thus, $a_1, a_2, \ldots, a_k$ are primes satisfying $n = a_1 a_2 \cdots a_k$. For each prime p, we have

$$(\text{the number of } i \in \{1, 2, \ldots, k\} \text{ such that } a_i = p) = v_p(n) \tag{415}$$

(by Proposition 9.3.13). Now, consider the product $a_1 a_2 \cdots a_k$. This product $a_1 a_2 \cdots a_k$ is a product of primes, and each prime p appears in it precisely $v_p(n)$ times (by (415)). Hence, this product equals $\prod\limits_{p \text{ prime}} p^{v_p(n)}$. Thus, $a_1 a_2 \cdots a_k = \prod\limits_{p \text{ prime}} p^{v_p(n)}$, so that $n = a_1 a_2 \cdots a_k = \prod\limits_{p \text{ prime}} p^{v_p(n)}$. This proves Theorem 9.3.17 **(d)**.

(c) (See [19s, Corollary 2.13.34] for details.) Apply Theorem 9.3.17 **(d)** to $|n|$ instead of n, and use Proposition 9.3.8 to simplify the result. □

Theorem 9.3.17 **(c)** greatly demystifies the "multiplicative structure" of the positive integers (i.e., the way positive integers behave under multiplication), as long as one is not interested in addition and subtraction. A rule of thumb is that, if an exercise about integers involves no $+$ and $-$ signs, then Theorem 9.3.17 **(c)** and the properties of p-valuations might simplify it significantly. We will see several examples of this below. One of the basic tools here is that divisibility relations can be reduced to inequalities between p-valuations:

Proposition 9.3.18. Let n and m be integers. Then, $n \mid m$ if and only if each prime p satisfies $v_p(n) \leq v_p(m)$.

Proof of Proposition 9.3.18 (sketched). (See [19s, Proposition 2.13.35] for details.) We must prove the following two claims:

Claim 1: If $n \mid m$, then each prime p satisfies $v_p(n) \leq v_p(m)$.

Claim 2: If each prime p satisfies $v_p(n) \leq v_p(m)$, then $n \mid m$.

[*Proof of Claim 1:* Assume that $n \mid m$. In other words, there exists some integer b such that $m = nb$. Consider this b. Now, each prime p satisfies

$$\begin{aligned} v_p(m) &= v_p(nb) \qquad (\text{since } m = nb) \\ &= v_p(n) + \underbrace{v_p(b)}_{\geq 0} \qquad (\text{by Theorem 9.3.6 } \textbf{(a)}\text{, applied to } a = n) \\ &\geq v_p(n), \end{aligned}$$

so that $v_p(n) \leq v_p(m)$. This proves Claim 1.]

[*Proof of Claim 2:* Assume that each prime p satisfies $v_p(n) \leq v_p(m)$. We must prove that $n \mid m$.

If $m = 0$, then this is obvious. Thus, we WLOG assume that $m \neq 0$. Hence, for each prime p, we have $v_p(m) \in \mathbb{N}$ and thus $v_p(m) < \infty$. In particular, $v_2(m) < \infty$.

We assumed that each prime p satisfies $v_p(n) \leq v_p(m)$. Thus, in particular, $v_2(n) \leq v_2(m) < \infty$, so that $n \neq 0$.

The statement of Claim 2 does not change if we replace n and m by $|n|$ and $|m|$, respectively (because of Proposition 3.1.3 **(a)** and Proposition 9.3.8). Hence, we WLOG assume that n and m are nonnegative. Assume this. Then, n and m are positive (since $n \neq 0$ and $m \neq 0$). Thus, Theorem 9.3.17 **(d)** yields

$$n = \prod_{p \text{ prime}} p^{v_p(n)}. \tag{416}$$

Similarly,

$$m = \prod_{p \text{ prime}} p^{v_p(m)}. \tag{417}$$

We have assumed that each prime p satisfies $v_p(n) \leq v_p(m)$. In other words, each prime p satisfies $v_p(m) - v_p(n) \geq 0$ and therefore $p^{v_p(m)-v_p(n)} \in \mathbb{Z}$. Now, set

$$c = \prod_{p \text{ prime}} p^{v_p(m)-v_p(n)}. \tag{418}$$

The infinite product in this equality has only finitely many factors different from 1 (for essentially the same reason as in Theorem 9.3.17 **(b)**), and thus is well-defined. Moreover, this product is an integer (since we have shown that each prime p satisfies $p^{v_p(m)-v_p(n)} \in \mathbb{Z}$; but this is saying that all factors of this product are integers). Thus, c is an integer. Multiplying the equalities (416) and (418), we obtain

$$nc = \left(\prod_{p \text{ prime}} p^{v_p(n)}\right)\left(\prod_{p \text{ prime}} p^{v_p(m)-v_p(n)}\right) = \prod_{p \text{ prime}} p^{v_p(m)} = m \qquad (\text{by } (417)).$$

In other words, $m = nc$. Hence, $n \mid m$. This completes the proof of Claim 2.]

Now, Claim 1 and Claim 2 are both proved; hence, Proposition 9.3.18 follows. □

The next corollary says that an integer n is determined up to sign by the family $(v_p(n))_{p \text{ prime}}$ of its p-valuations for all primes p:

Corollary 9.3.19. Let n and m be two integers. Assume that

$$v_p(n) = v_p(m) \qquad \text{for every prime } p. \tag{419}$$

(a) Then, $|n| = |m|$.
(b) If n and m are nonnegative, then $n = m$.

Proof of Corollary 9.3.19. **(a)** Each prime p satisfies $v_p(n) \leq v_p(m)$ (since (419) yields $v_p(n) = v_p(m)$). Hence, Proposition 9.3.18 shows that $n \mid m$. However, the same argument (with the roles of n and m interchanged) shows that $m \mid n$ (since n and m play symmetric roles in Corollary 9.3.19 **(a)**). Thus, Proposition 3.1.3 **(c)** (applied to $a = n$ and $b = m$) yields $|n| = |m|$. This proves Corollary 9.3.19 **(a)**.

(b) Assume that n and m are nonnegative. Thus, $|n| = n$ and $|m| = m$. However, Corollary 9.3.19 **(a)** yields $|n| = |m|$. Hence, $n = |n| = |m| = m$. This proves Corollary 9.3.19 **(b)**. □

We can also describe the gcd of several integers via p-valuations:

Proposition 9.3.20. Let $n_1, n_2, \ldots, n_k$ be finitely many integers. Assume that not all of $n_1, n_2, \ldots, n_k$ are zero. Then,

$$\gcd(n_1, n_2, \ldots, n_k) = \prod_{p \text{ prime}} p^{\min\{v_p(n_1), v_p(n_2), \ldots, v_p(n_k)\}}. \tag{420}$$

Proof of Proposition 9.3.20. We know (from Proposition 3.4.3 **(b)**, applied to $b_i = n_i$) that $\gcd(n_1, n_2, \ldots, n_k)$ is a positive integer. Hence, Theorem 9.3.17 **(d)** (applied to $n = \gcd(n_1, n_2, \ldots, n_k)$) yields

$$\begin{aligned}\gcd(n_1, n_2, \ldots, n_k) &= \prod_{p \text{ prime}} \underbrace{p^{v_p(\gcd(n_1,n_2,\ldots,n_k))}}_{\substack{=p^{\min\{v_p(n_1),v_p(n_2),\ldots,v_p(n_k)\}} \\ \text{(since Proposition 9.3.11} \\ \text{yields } v_p(\gcd(n_1,n_2,\ldots,n_k))=\min\{v_p(n_1),v_p(n_2),\ldots,v_p(n_k)\})}} \\ &= \prod_{p \text{ prime}} p^{\min\{v_p(n_1),v_p(n_2),\ldots,v_p(n_k)\}}.\end{aligned}$$

This proves Proposition 9.3.20.

(See [19s, Proposition 2.13.40] for a different proof in the case when $n_1, n_2, \ldots, n_k$ are all nonzero.) □

Another consequence of the above is saying that a congruence modulo a nonzero integer n can be reduced to the corresponding congruences modulo all prime powers $p^{v_p(n)}$:

Corollary 9.3.21. Let n be a nonzero integer. Let a and b be two integers. Assume that

$$a \equiv b \bmod p^{v_p(n)} \qquad \text{for every prime } p. \tag{421}$$

Then, $a \equiv b \bmod n$.

Proof of Corollary 9.3.21 (sketched). (See [19s, Exercise 2.13.9] for details and for an alternative proof.) We must prove that $a \equiv b \bmod n$. If $a = b$, then this is obvious. Thus, we WLOG assume that $a \neq b$. Hence, $a - b \neq 0$, so that $v_p(a-b) \in \mathbb{N}$ for every prime p.

Let p be any prime. Then, (421) yields $a \equiv b \bmod p^{v_p(n)}$. In other words, $p^{v_p(n)} \mid a - b$. In view of Lemma 9.3.3 (applied to $v_p(n)$ and $a - b$ instead of i and n), this entails that $v_p(a-b) \geq v_p(n)$. In other words, $v_p(n) \leq v_p(a-b)$.

Now, forget that we fixed p. We thus have proven that each prime p satisfies $v_p(n) \leq v_p(a-b)$. In view of Proposition 9.3.18 (applied to $m = a - b$), this entails that $n \mid a - b$. In other words, $a \equiv b \bmod n$. This proves Corollary 9.3.21. □

The p-valuation of a positive integer n is easily read off its canonical factorization:

Corollary 9.3.22. For each prime p, let b_p be a nonnegative integer. Assume that only finitely many primes p satisfy $b_p \neq 0$. Let $n = \prod_{p \text{ prime}} p^{b_p}$. Then,

$$v_q(n) = b_q \qquad \text{for each prime } q.$$

Proof of Corollary 9.3.22 (sketched). (See [19s, Corollary 2.13.37] for the details.) We have $n = \prod_{p \text{ prime}} p^{b_p}$; thus, the number n has a prime factorization

$$\left(\underbrace{2,2,\ldots,2}_{b_2 \text{ times}}, \underbrace{3,3,\ldots,3}_{b_3 \text{ times}}, \underbrace{5,5,\ldots,5}_{b_5 \text{ times}}, \ldots\right)$$

in which each prime p appears b_p times. (This is a finite list, since only finitely many primes p satisfy $b_p \neq 0$.) Hence, Corollary 9.3.22 follows from Proposition 9.3.13 (applied to $p = q$). □

9.3.4. Some applications

As mentioned above, Theorem 9.3.17 reduces multiplicative questions about integers to additive questions about their p-valuations. The following exercise (which is in itself an important result) illustrates this:

Exercise 9.3.1. Let $n \in \mathbb{N}$. Let a and b be positive integers satisfying $a \perp b$. Assume that ab is the n-th power of a positive integer. Prove that a and b are n-th powers of positive integers.

Note that the word "positive" in Exercise 9.3.1 cannot be dispensed with (for a counterexample, set $a = -1$ and $b = -1$ and $n = 2$).

Solution to Exercise 9.3.1. If $n = 0$, then Exercise 9.3.1 is easy to solve[290]. Thus, for the rest of this solution, we WLOG assume that $n \neq 0$. Hence, $n > 0$.

We shall first show the following:

> *Claim 1:* Let p be a prime. Then, $v_p(a)/n$ and $v_p(b)/n$ are nonnegative integers.

[*Proof of Claim 1:* We have $a \perp b$; in other words, $\gcd(a,b) = 1$ (by the definition of "coprime"). Thus, $v_p(\gcd(a,b)) = v_p(1) = 0$ (by Theorem 9.3.6 **(c)**). However, Proposition 9.3.11 (applied to $k = 2$ and $n_1 = a$ and $n_2 = b$) yields

$$v_p(\gcd(a,b)) = \min\{v_p(a), v_p(b)\}.$$

Hence, $\min\{v_p(a), v_p(b)\} = v_p(\gcd(a,b)) = 0$. Thus, one of the two numbers $v_p(a)$ and $v_p(b)$ is 0 (since $\min\{v_p(a), v_p(b)\}$ is one of the two numbers $v_p(a)$

[290] *Proof.* Assume that $n = 0$. We have assumed that ab is the n-th power of a positive integer. In other words, there exists a positive integer c such that $ab = c^n$. Consider this c. Thus, $ab = c^n = c^0$ (since $n = 0$), so that $ab = c^0 = 1$ and thus $a \mid ab = 1$. Hence, Proposition 3.1.3 **(b)** (applied to 1 instead of b) yields $|a| \leq |1| = 1$. Since a is positive, we have $|a| = a$, so that $a = |a| \leq 1$. Thus, $a = 1$ (since a is a positive integer), so that $a = 1 = 1^0$. Hence, a is the n-th power of a positive integer (namely, of 1). Similarly, b is the n-th power of a positive integer. Thus, a and b are n-th powers of positive integers. Hence, we have solved Exercise 9.3.1 under the assumption that $n = 0$.

and $v_p(b)$). In other words, we have $v_p(a) = 0$ or $v_p(b) = 0$. Thus, we WLOG assume that $v_p(a) = 0$ (since we can otherwise achieve this by swapping a with b).

However, we assumed that ab is the n-th power of a positive integer. In other words, there exists a positive integer c such that $ab = c^n$. Consider this c. From $ab = c^n$, we obtain $v_p(ab) = v_p(c^n) = nv_p(c)$ (by Corollary 9.3.4, applied to c and n instead of a and k). Also, $v_p(c) \in \mathbb{N}$ (since c is a positive integer, thus a nonzero integer). However, Theorem 9.3.6 **(a)** yields $v_p(ab) = \underbrace{v_p(a)}_{=0} + v_p(b) = v_p(b)$, so that $v_p(b) = v_p(ab) = nv_p(c)$ and therefore $v_p(b)/n = v_p(c) \in \mathbb{N}$. Also, $\underbrace{v_p(a)}_{=0}/n = 0/n = 0 \in \mathbb{N}$.

Now, we see that $v_p(a)/n$ and $v_p(b)/n$ are nonnegative integers (since $v_p(a)/n \in \mathbb{N}$ and $v_p(b)/n \in \mathbb{N}$). This proves Claim 1.]

Now, recall that the integer a is positive and thus nonzero. Hence, Theorem 9.3.17 **(d)** (applied to a instead of n) yields

$$a = \prod_{p \text{ prime}} p^{v_p(a)}. \tag{422}$$

However, if p is any prime, then $v_p(a)/n$ is a nonnegative integer (by Claim 1), and therefore $p^{v_p(a)/n}$ is a well-defined positive integer (since p is a positive integer). Moreover, if p is any prime satisfying $p > |a|$, then $v_p(a) = 0$ (by Theorem 9.3.17 **(a)**, applied to a instead of n) and thus $\underbrace{v_p(a)}_{=0}/n = 0/n = 0$ and therefore $p^{v_p(a)/n} = p^0 = 1$. Hence, the product $\prod_{p \text{ prime}} p^{v_p(a)/n}$ has only finitely many factors different from 1 (since only finitely many primes p do not satisfy $p > |a|$). Thus, this product is well-defined. Moreover, all factors of this product are positive integers (because we have shown that if p is any prime, then $p^{v_p(a)/n}$ is a positive integer). Thus, this product itself is a positive integer. In other words, $\prod_{p \text{ prime}} p^{v_p(a)/n}$ is a positive integer. However,

$$\left(\prod_{p \text{ prime}} p^{v_p(a)/n}\right)^n = \prod_{p \text{ prime}} \underbrace{\left(p^{v_p(a)/n}\right)^n}_{=p^{v_p(a)}} = \prod_{p \text{ prime}} p^{v_p(a)}.$$

Comparing this with (422), we obtain $a = \left(\prod_{p \text{ prime}} p^{v_p(a)/n}\right)^n$. This shows that a is the n-th power of a positive integer (since $\prod_{p \text{ prime}} p^{v_p(a)/n}$ is a positive integer). Similarly, b is the n-th power of a positive integer. This solves Exercise 9.3.1. $\square$

Another application of p-valuations is the following simple but useful fact:

Exercise 9.3.2. Let k be a positive integer. Let w be a rational number such that w^k is an integer. Prove that w is an integer.

Exercise 9.3.2 can be restated as follows: If the k-th root of an integer is rational, then this root is an integer. In other words, if an integer is not a k-th power of an integer, then its k-th root is irrational. This makes it completely straightforward to check (e.g.) that the numbers $\sqrt{2}$ and $\sqrt[3]{5}$ and $\sqrt[97]{35}$ are irrational.

Solution to Exercise 9.3.2. If $w = 0$, then this is obvious. Thus, for the rest of this solution, we WLOG assume that $w \neq 0$.

The number w is rational. Thus, we can write w in the form $w = m/n$ for some integers m and n with $n \neq 0$. Consider these m and n. Note that $m = nw$ (since $w = m/n$) and thus $m \neq 0$ (since $n \neq 0$ and $w \neq 0$).

From $w = m/n$, we obtain $w^k = (m/n)^k = m^k/n^k$. Hence, m^k/n^k is an integer (since w^k is an integer). Thus, $n^k \mid m^k$.

However, Proposition 9.3.18 (applied to n^k and m^k instead of n and m) shows that $n^k \mid m^k$ if and only if each prime p satisfies $v_p\left(n^k\right) \leq v_p\left(m^k\right)$. Hence,

$$\text{each prime } p \text{ satisfies } v_p\left(n^k\right) \leq v_p\left(m^k\right) \tag{423}$$

(since we have $n^k \mid m^k$).

Now, let p be a prime. Then, $v_p(n)$ and $v_p(m)$ are integers (since $n \neq 0$ and $m \neq 0$). We have $v_p\left(n^k\right) = kv_p(n)$ (by Corollary 9.3.9, applied to $a = n$) and $v_p\left(m^k\right) = kv_p(m)$ (similarly). Hence,

$$\begin{aligned} kv_p(n) &= v_p\left(n^k\right) \leq v_p\left(m^k\right) \qquad \text{(by (423))} \\ &= kv_p(m). \end{aligned}$$

Dividing this inequality by k, we obtain $v_p(n) \leq v_p(m)$ (since k is positive).

Forget that we fixed p. Thus, we have shown that each prime p satisfies $v_p(n) \leq v_p(m)$. According to Proposition 9.3.18, this entails that $n \mid m$. Hence, m/n is an integer (since n is nonzero). In other words, w is an integer (since $w = m/n$). This solves Exercise 9.3.2. □

Another solution to Exercise 9.3.2 can be given using Corollary 3.5.17 (hint: w is a root of the polynomial $x^k - w^k$). We leave the details to the reader.

Here is another, slightly goofy, application of Theorem 9.3.17 (based on an idea of Kurt Gödel):

Exercise 9.3.3. A sequence $(a_1, a_2, a_3, \ldots) \in \mathbb{N}^\infty$ of nonnegative integers will be called *finitary* if only finitely many positive integers n satisfy $a_n \neq 0$. (For example, the sequence $\left(1, 3, 0, 2, \underbrace{0, 0, 0, \ldots}_{\text{infinitely many zeroes}}\right)$ is finitary, whereas the sequence $(1, 0, 1, 0, 1, 0, \ldots)$ (which alternates between 1s and 0s) is not.)

Prove that the set of all finitary sequences of nonnegative integers is countable.

Solution to Exercise 9.3.3 (sketched). There are various ways of solving Exercise 9.3.3, but a particularly short one can be obtained using prime factorization:

Let $\mathbb{P} = \{1, 2, 3, \ldots\}$ be the set of all positive integers. Let F be the set of all finitary sequences of nonnegative integers. We must prove that F is countable. We shall achieve this by constructing a bijection from F to $\mathbb{P}$.

Let $(p_1, p_2, p_3, \ldots) = (2, 3, 5, 7, 11, \ldots)$ be the sequence of all primes, listed in increasing order with no repetitions. (This is indeed a well-defined infinite sequence, because Theorem 9.1.3 shows that there are infinitely many primes.) Now, define a map

$$\begin{aligned} \alpha : F &\to \mathbb{P}, \\ (a_1, a_2, a_3, \ldots) &\mapsto p_1^{a_1} p_2^{a_2} p_3^{a_3} \cdots . \end{aligned}$$

(Note that the product $p_1^{a_1} p_2^{a_2} p_3^{a_3} \cdots$ in this definition is a well-defined positive integer, since only finitely many of its factors are different from 1 (because only finitely many positive integers n satisfy $a_n \neq 0$).) Conversely, define a map

$$\begin{aligned} \beta : \mathbb{P} &\to F, \\ n &\mapsto \left(v_{p_1}(n), v_{p_2}(n), v_{p_3}(n), \ldots \right). \end{aligned}$$

(Note that the sequence $\left(v_{p_1}(n), v_{p_2}(n), v_{p_3}(n), \ldots \right)$ in this definition is indeed finitary, because Theorem 9.3.17 **(a)** entails that $v_{p_i}(n) = 0$ for all primes $p_i > |n|$.)

Now, the maps α and β are mutually inverse. (Indeed, $\alpha \circ \beta = \mathrm{id}$ follows easily from Theorem 9.3.17 **(d)**, whereas $\beta \circ \alpha = \mathrm{id}$ follows easily from Corollary 9.3.22.) Thus, the map α is invertible, i.e., a bijection. Hence, we have found a bijection from F to $\mathbb{P}$. Since the set $\mathbb{P}$ is countable, we thus conclude that F is also countable. This solves Exercise 9.3.3. □

Here is another exercise in which primes do not appear, yet are crucial for its solution:

Exercise 9.3.4. Let a, b and c be three integers such that $\gcd(a, b, c) = 1$ and $c \neq 0$. Prove that there is a positive integer x such that $a + bx \perp c$.

Note that the claim of Exercise 9.3.4 can also be restated as "there exists an entry of the arithmetic progression $(a + b, a + 2b, a + 3b, \ldots)$ that is coprime to c". This stands to reason, since the assumption $\gcd(a, b, c) = 1$ rules out the existence of a nontrivial divisor shared by all entries of this arithmetic progression and c (where "nontrivial" means "larger than 1"). However, this heuristic argument is not a proof, since it is not hard to imagine that each entry of this arithmetic progression could have some nontrivial divisor in common with c, but a different divisor depending on the entry. Nevertheless, this theoretical possibility does not actually happen, and Exercise 9.3.4 is true. However, the reasons for this are subtler than it may appear. Before we solve this exercise, let us prove two simple but useful lemmas:

Lemma 9.3.23. Let n and m be two integers that don't satisfy $n \perp m$. Then, there exists a prime q that divides both n and m.

Proof of Lemma 9.3.23. Proposition 3.4.3 **(a)** shows that the number $\gcd(n,m)$ is a nonnegative integer. Proposition 3.4.4 **(f)** (applied to $a = n$ and $b = m$) yields $\gcd(n,m) \mid n$ and $\gcd(n,m) \mid m$.

We are in one of the following two cases:

Case 1: We have $\gcd(n,m) \leq 1$.

Case 2: We have $\gcd(n,m) > 1$.

Let us first consider Case 1. In this case, we have $\gcd(n,m) \leq 1$. However, if we had $\gcd(n,m) = 1$, then we would have $n \perp m$ (by the definition of "coprime"), which would contradict the assumption that we don't have $n \perp m$. Hence, we cannot have $\gcd(n,m) = 1$. Thus, $\gcd(n,m) \neq 1$. Combining this with $\gcd(n,m) \leq 1$, we obtain $\gcd(n,m) < 1$. Therefore, $\gcd(n,m) = 0$ (since $\gcd(n,m)$ is a nonnegative integer). Thus, $0 = \gcd(n,m)$, so that $2 \mid 0 = \gcd(n,m) \mid n$ and $2 \mid 0 = \gcd(n,m) \mid m$. In other words, 2 divides both n and m. Hence, there exists a prime q that divides both n and m (namely, $q = 2$). This proves Lemma 9.3.23 in Case 1.

Let us now consider Case 2. In this case, we have $\gcd(n,m) > 1$. Hence, Proposition 9.1.4 (applied to $\gcd(n,m)$ instead of n) yields that there exists at least one prime p such that $p \mid \gcd(n,m)$. Consider this p. Thus, $p \mid \gcd(n,m) \mid n$ and $p \mid \gcd(n,m) \mid m$. In other words, p divides both n and m. Hence, there exists a prime q that divides both n and m (namely, $q = p$). This proves Lemma 9.3.23 in Case 2.

We have now proved Lemma 9.3.23 in both Cases 1 and 2. This shows that Lemma 9.3.23 always holds. □

We note that the converse of Lemma 9.3.23 is also true: If there exists a prime q that divides both n and m, then n and m don't satisfy $n \perp m$. (This is pretty easy to show; do it!)

Our next lemma is no less simple:

Lemma 9.3.24. Let S be a finite set of primes. Let q be a prime such that $q \notin S$. Then, $q \perp \prod_{p \in S} p$.

For example, Lemma 9.3.24 (applied to $S = \{2,7,11\}$ and $q = 3$) yields that $3 \perp 2 \cdot 7 \cdot 11$.

Proof of Lemma 9.3.24. Write the finite set S in the form $S = \{p_1, p_2, \ldots, p_k\}$ (where $p_1, p_2, \ldots, p_k$ are distinct). Thus, $\prod_{p \in S} p = p_1 p_2 \cdots p_k$.

Now, let $i \in \{1, 2, \ldots, k\}$. Then, $p_i \in \{p_1, p_2, \ldots, p_k\} = S$. Hence, p_i is a prime (since S is a set of primes). Moreover, we cannot have $p_i = q$ (since $p_i = q$ would yield $p_i = q \notin S$, which would contradict $p_i \in S$). Thus, the primes p_i and q are distinct. Hence, Proposition 9.1.7 (applied to $p = p_i$) yields $p_i \perp q$.

Forget that we fixed i. We thus have shown that each $i \in \{1, 2, \ldots, k\}$ satisfies $p_i \perp q$. Hence, Exercise 3.5.4 (applied to $a_i = p_i$ and $c = q$) yields $p_1 p_2 \cdots p_k \perp q$. In other words, $q \perp p_1 p_2 \cdots p_k$ (by Proposition 3.5.4). In other words, $q \perp \prod_{p \in S} p$ (since $\prod_{p \in S} p = p_1 p_2 \cdots p_k$). This proves Lemma 9.3.24. $\square$

We are now ready to solve Exercise 9.3.4:

Solution to Exercise 9.3.4. (The following solution is the second solution to Problem 2 in `https://math.stackexchange.com/a/3128111/` ; I learnt it from a student.)

Let C be the set of all primes that divide c. This is a finite set[291]. Moreover, C is a set of primes (by the definition of C). Hence, C is a finite set of primes.

Let A be the set of all primes that divide a. The set $C \setminus A$ is a subset of C, and thus is a finite set of primes (since C is a finite set of primes).

Now, let

$$x = \prod_{p \in C \setminus A} p.$$

Clearly, x is a positive integer. Now, I claim that $a + bx \perp c$. Obviously, once this is proven, Exercise 9.3.4 will be solved.

So we must prove that $a + bx \perp c$.

Indeed, assume the contrary. Thus, we don't have $a + bx \perp c$. Hence, Lemma 9.3.23 (applied to $n = a + bx$ and $m = c$) shows that there exists a prime q that divides both $a + bx$ and c. Consider this q.

We know that q is a prime that divides c. In other words, $q \in C$ (by the definition of C). Moreover, q divides $a + bx$; in other words, $q \mid a + bx$. Furthermore, $q > 1$ (since q is a prime).

Now, we must be in one of the following two cases:

Case 1: We have $q \in A$.

Case 2: We have $q \notin A$.

Let us consider Case 1 first. In this case, we have $q \in A$. Thus, q is a prime that divides a (by the definition of A). In particular, $q \mid a$. Now,

$$xb = bx = \underbrace{(a + bx)}_{\substack{\equiv 0 \bmod q \\ (\text{since } q \mid a+bx)}} - \underbrace{a}_{\substack{\equiv 0 \bmod q \\ (\text{since } q \mid a)}} \equiv 0 - 0 = 0 \bmod q.$$

In other words, $q \mid xb$.

We cannot have $q \in C \setminus A$ (since this would entail $q \notin A$, which would contradict $q \in A$). In other words, we have $q \notin C \setminus A$. Hence, Lemma 9.3.24 (applied to

[291] *Proof.* Let $p \in C$. Then, p is a prime that divides c (by the definition of C). Hence, $p \mid c$. Thus, Proposition 3.1.3 **(b)** (applied to p and c instead of a and b) yields $|p| \leq |c|$ (since $c \neq 0$). However, p is prime; thus, $p > 1 > 0$ and therefore $|p| = p$. Hence, $p = |p| \leq |c|$ and thus $p \in \{1, 2, \ldots, |c|\}$ (since p is a positive integer).

Forget that we fixed p. We thus have shown that $p \in \{1, 2, \ldots, |c|\}$ for each $p \in C$. In other words, $C \subseteq \{1, 2, \ldots, |c|\}$. Hence, the set C is finite (since the set $\{1, 2, \ldots, |c|\}$ is finite).

$S = C \setminus A$) yields $q \perp \prod\limits_{p \in C \setminus A} p$. This rewrites as $q \perp x$ (since $x = \prod\limits_{p \in C \setminus A} p$). Hence, Theorem 3.5.6 (applied to q, x and b instead of a, b and c) yields $q \mid b$ (since $q \mid xb$). Note that we also have $q \mid a$ and $q \mid c$ (since q divides c).

Now, for any $b_1, b_2, b_3 \in \mathbb{Z}$ and $m \in \mathbb{Z}$, we have the logical equivalence

$$(m \mid b_1 \text{ and } m \mid b_2 \text{ and } m \mid b_3) \iff (m \mid \gcd(b_1, b_2, b_3))$$

[292]. Applying this to $b_1 = a$ and $b_2 = b$ and $b_3 = c$ and $m = q$, we obtain the logical equivalence

$$(q \mid a \text{ and } q \mid b \text{ and } q \mid c) \iff (q \mid \gcd(a, b, c)).$$

Hence, we have $q \mid \gcd(a, b, c)$ (since we have $q \mid a$ and $q \mid b$ and $q \mid c$). In other words, $q \mid 1$ (since $\gcd(a, b, c) = 1$). Thus, q is a divisor of 1. This shows that q is either 1 or -1 (since the only divisors of 1 are 1 and -1). But this contradicts the fact that $q > 1$. Thus, we have obtained a contradiction in Case 1.

Let us now consider Case 2. In this case, we have $q \notin A$. Combining this with $q \in C$, we obtain $q \in C \setminus A$. Hence, q is a factor in the product $\prod\limits_{p \in C \setminus A} p$. Therefore, q divides this product. Thus, $q \mid \prod\limits_{p \in C \setminus A} p = x$. Now,

$$a = \underbrace{(a + bx)}_{\substack{\equiv 0 \bmod q \\ (\text{since } q \mid a+bx)}} - b \underbrace{x}_{\substack{\equiv 0 \bmod q \\ (\text{since } q \mid x)}} \equiv 0 - b \cdot 0 = 0 \bmod q.$$

In other words, $q \mid a$. Hence, q divides a. Therefore, q is a prime that divides a (since q is a prime). In other words, $q \in A$ (by the definition of A). But this contradicts $q \notin A$. Thus, we have obtained a contradiction in Case 2.

We have now found contradictions in both Cases 1 and 2. Since these two cases cover all possibilities, this shows that we always get a contradiction. Thus, our assumption was wrong, and therefore $a + bx \perp c$ is proven. This solves Exercise 9.3.4. □

Next, let us illustrate the use of Corollary 9.3.21 by proving a variant of Theorem 9.1.13 (Euler's theorem):

[292] *Proof.* Let $b_1, b_2, b_3 \in \mathbb{Z}$ and $m \in \mathbb{Z}$ be arbitrary. Then, we have the following chain of logical equivalences:

$$\begin{aligned}
& (m \mid b_1 \text{ and } m \mid b_2 \text{ and } m \mid b_3) \\
\iff & (m \mid b_i \text{ for all } i \in \{1, 2, 3\}) \\
\iff & (m \mid b_i \text{ for all } i \in \{1, 2, \ldots, 3\}) \qquad (\text{since } \{1, 2, 3\} = \{1, 2, \ldots, 3\}) \\
\iff & (m \mid \gcd(b_1, b_2, \ldots, b_3)) \qquad (\text{by Theorem 3.4.14 (applied to } k = 3)) \\
\iff & (m \mid \gcd(b_1, b_2, b_3)) \qquad (\text{since } \gcd(b_1, b_2, \ldots, b_3) = \gcd(b_1, b_2, b_3)).
\end{aligned}$$

Thus, we have proved the equivalence $(m \mid b_1 \text{ and } m \mid b_2 \text{ and } m \mid b_3) \iff (m \mid \gcd(b_1, b_2, b_3))$.

Exercise 9.3.5. Let a be an integer, and let n be a positive integer. Let $\phi(n)$ be defined as in Theorem 9.1.13. Prove that $a^n \equiv a^{n-\phi(n)} \bmod n$.

Note that Exercise 9.3.5 competes with Theorem 9.1.13 over the claim of being the natural generalization of Fermat's Little Theorem: Indeed, Exercise 9.3.5 generalizes Theorem 9.1.10 **(b)** just like Theorem 9.1.13 generalizes Theorem 9.1.10 **(a)**.

Solution to Exercise 9.3.5 (sketched). We shall only give the main steps; see [19s, Exercise 2.16.3] for a similar solution written out in full detail.

First, let us show that $a^{n-\phi(n)}$ is an integer (in order to be sure that the congruence $a^n \equiv a^{n-\phi(n)} \bmod n$ makes sense in the first place!). Indeed, in Theorem 9.1.13, we have defined $\phi(n)$ to be the # of all $i \in \{1, 2, \ldots, n\}$ that satisfy $i \perp n$. Hence,

$$\begin{aligned} \phi(n) &= (\text{\# of all } i \in \{1, 2, \ldots, n\} \text{ that satisfy } i \perp n) \\ &\leq (\text{\# of all } i \in \{1, 2, \ldots, n\}) = |\{1, 2, \ldots, n\}| = n. \end{aligned} \tag{424}$$

Thus, $n - \phi(n) \in \mathbb{N}$. This shows that $a^{n-\phi(n)}$ is an integer.

We must prove the congruence $a^n \equiv a^{n-\phi(n)} \bmod n$. According to Corollary 9.3.21 (applied to a^n and $a^{n-\phi(n)}$ instead of a and b), it suffices therefor to show that $a^n \equiv a^{n-\phi(n)} \bmod p^{v_p(n)}$ for every prime p.

So let us fix a prime p. Our goal is thus to show that

$$a^n \equiv a^{n-\phi(n)} \bmod p^{v_p(n)}. \tag{425}$$

Proposition 9.1.6 yields that we have either $p \mid a$ or $p \perp a$. Hence, we are in one of the following two cases:

Case 1: We have $p \mid a$.

Case 2: We have $p \perp a$.

Let us first consider Case 1. In this case, we have $p \mid a$. Our goal, as we remember, is to prove (425). We shall achieve this by showing that both a^n and $a^{n-\phi(n)}$ are congruent to 0 modulo $p^{v_p(n)}$.

Indeed, the difference rule (Theorem 7.1.8) yields

$$\begin{aligned} &(\text{\# of all } i \in \{1, 2, \ldots, n\} \text{ that } \textbf{don't} \text{ satisfy } i \perp n) \\ &= \underbrace{(\text{\# of all } i \in \{1, 2, \ldots, n\})}_{=|\{1,2,\ldots,n\}|=n} - \underbrace{(\text{\# of all } i \in \{1, 2, \ldots, n\} \text{ that satisfy } i \perp n)}_{\substack{=\phi(n) \\ \text{(by (424))}}} \\ &= n - \phi(n). \end{aligned} \tag{426}$$

However, let us set $j = v_p(n)$. Then, $v_p(n) \geq j$, so that $p^j \mid n$ (by Lemma 9.3.3, applied to $i = j$). Hence, Proposition 3.1.3 **(b)** readily yields $\left|p^j\right| \leq |n|$. In other words, $p^j \leq n$ (since p^j and n are positive). Thus,

$$1 \leq p^1 \leq p^2 \leq \cdots \leq p^j \leq n.$$

This shows that the j numbers $p^1, p^2, \ldots, p^j$ all belong to the set $\{1, 2, \ldots, n\}$. Moreover, none of these j numbers is coprime to n (since they are all divisible by p, and thus have the divisor p in common with n, which shows that their gcd with n is $\geq p$). Hence, these j numbers $p^1, p^2, \ldots, p^j$ all belong to the set of all $i \in \{1, 2, \ldots, n\}$ that **don't** satisfy $i \perp n$. Hence, this set has at least j many elements (because the j numbers $p^1, p^2, \ldots, p^j$ are distinct). In other words,

$$(\# \text{ of all } i \in \{1, 2, \ldots, n\} \text{ that } \textbf{don't} \text{ satisfy } i \perp n) \geq j.$$

In view of (426), this rewrites as $n - \phi(n) \geq j$.

Now, $p^1 = p \mid a$, so that $v_p(a) \geq 1$ (by Lemma 9.3.3, applied to 1 and a instead of i and n). Corollary 9.3.9 (applied to $k = n - \phi(n)$) yields

$$v_p\left(a^{n-\phi(n)}\right) = (n - \phi(n)) \underbrace{v_p(a)}_{\geq 1} \geq n - \phi(n) \geq j.$$

In view of Lemma 9.3.3 (applied to j and $a^{n-\phi(n)}$ instead of i and n), this entails that $p^j \mid a^{n-\phi(n)}$. In other words, $a^{n-\phi(n)} \equiv 0 \bmod p^j$. Hence,

$$a^n = \underbrace{a^{n-\phi(n)}}_{\equiv 0 \bmod p^j} \cdot a^{\phi(n)} \equiv 0 \cdot a^{\phi(n)} = 0 \equiv a^{n-\phi(n)} \bmod p^j.$$

In other words, $a^n \equiv a^{n-\phi(n)} \bmod p^{v_p(n)}$ (since $j = v_p(n)$). Thus, (425) is proved in Case 1.

Let us now consider Case 2. In this case, we have $p \perp a$. In other words, $a \perp p$.

We shall now show that $a^{\phi(n)} \equiv 1 \bmod p^{v_p(n)}$. Once this is proved, the claim (425) will quickly follow.

It is tempting to obtain $a^{\phi(n)} \equiv 1 \bmod p^{v_p(n)}$ by applying Theorem 9.1.13. However, this is not so easy, since we don't know whether $a \perp n$; we only know that the weaker statement $a \perp p$ is true. We need to work around this.

Fortunately, Exercise 9.3.4 is here to help. We set $i = v_p(n)$. Then, $v_p(n) \geq i$, so that $p^i \mid n$ (by Lemma 9.3.3). Furthermore, Exercise 3.5.5 yields $a^1 \perp p^i$ (since $a \perp p$). In other words, $a \perp p^i$. Hence, it is easy to see that $\gcd\left(a, p^i, n\right) = 1$ [293]. Thus, Exercise 9.3.4 (applied to $b = p^i$ and $c = n$) yields that there is a positive integer x such that $a + p^i x \perp n$. Consider this x. Set $b = a + p^i x$. Thus, $b - a = p^i x$

[293] *Proof.* Let $m = \gcd\left(a, p^i, n\right)$. Then, m is a common divisor of a, p^i and n (by the definition of a gcd). Hence, $m \mid a$ and $m \mid p^i$ and $m \mid n$. However, Theorem 3.4.7 (applied to $b = p^i$) yields that we have the following logical equivalence:

$$\left(m \mid a \text{ and } m \mid p^i\right) \iff \left(m \mid \gcd\left(a, p^i\right)\right).$$

Thus, $m \mid \gcd\left(a, p^i\right)$ (since $m \mid a$ and $m \mid p^i$). Thus, we have $m \mid \gcd\left(a, p^i\right) = 1$ (since $a \perp p^i$). Since m is nonnegative (because we defined m as a gcd), this entails $m = 1$. Thus, $\gcd\left(a, p^i, n\right) = m = 1$.

is divisible by p^i; in other words, $b \equiv a \bmod p^i$. Hence, $b^{\phi(n)} \equiv a^{\phi(n)} \bmod p^i$, so that $a^{\phi(n)} \equiv b^{\phi(n)} \bmod p^i$.

Now, $b = a + p^i x \perp n$; hence, Theorem 9.1.13 (applied to b instead of a) yields $b^{\phi(n)} \equiv 1 \bmod n$. This entails $b^{\phi(n)} \equiv 1 \bmod p^i$ (by Proposition 3.2.6 **(e)**, since $p^i \mid n$). However, $a^{\phi(n)} \equiv b^{\phi(n)} \equiv 1 \bmod p^i$. In other words, $a^{\phi(n)} \equiv 1 \bmod p^{v_p(n)}$ (since $i = v_p(n)$).

Now,

$$a^n = a^{n-\phi(n)} \cdot \underbrace{a^{\phi(n)}}_{\equiv 1 \bmod p^{v_p(n)}} \equiv a^{n-\phi(n)} \cdot 1 = a^{n-\phi(n)} \bmod p^{v_p(n)}.$$

Thus, (425) is proved in Case 2.

Hence, we have proved (425) in both Cases 1 and 2. This shows that (425) always holds. As explained above, this completes the solution to Exercise 9.3.5. □

The above solution to Exercise 9.3.5 might appear clunky with its two cases, its rough (analysis-style) estimates of $n - \phi(n)$ (in Case 1), and its somewhat unmotivated use of Exercise 9.3.4. However, it illustrates several important techniques.

The first technique is the passage from proving a congruence modulo n to proving a congruence modulo $p^{v_p(n)}$ (via Corollary 9.3.21). This is called *working locally* (where "locally" means "focusing on a single prime"), and has the advantage that $p^{v_p(n)}$ behaves "nicer" than the (a priori) arbitrary positive integer n. In our specific case, we profited from this passage by being able to apply Proposition 9.1.6 and use p-valuations (whereas there is no analogue of Proposition 9.1.6 for n instead of p). There are many other ways in which working locally can be useful; for example, if we want to show that some two integers are coprime, then (by Lemma 9.3.23) it suffices to show that every prime fails to divide at least one of them. Other applications of working locally are enabled by the *Chinese remainder theorem*, which we shall discuss further below.

The second technique is the trick that allowed us to apply Theorem 9.1.13 in Case 2: We could not apply Theorem 9.1.13 directly to a, since a may fail to satisfy $a \perp n$; thus, we instead applied Theorem 9.1.13 to a new integer b that we constructed tactically to satisfy $b \perp n$ (so that it satisfies the assumption of Theorem 9.1.13) while being congruent to a modulo p^i (so that it could act as a stand-in for a for the purpose of proving a congruence modulo p^i). Thus, so to speak, we have "snuck" a into Theorem 9.1.13 under the "disguise" of b. The "moral" of the story is that all is not lost when a theorem does not immediately apply; often it is possible to tweak the objects until it does apply to them. (Of course, we were lucky to have Exercise 9.3.4 available when it came to tweaking a. It would have been much harder otherwise...)

9.3.5. Factorials and their p-valuations

Next, we shall study a very special case of p-valuations: those of the factorials. In other words, we shall compute $v_p(n!)$ for any $n \in \mathbb{N}$ and any prime p:

Theorem 9.3.25 (de Polignac's formula)**.** Let p be a prime. Let $n \in \mathbb{N}$. Then,

$$v_p(n!) = \sum_{i \geq 1} \left\lfloor \frac{n}{p^i} \right\rfloor .$$

(The summation sign "$\sum\limits_{i \geq 1}$" is shorthand for "$\sum\limits_{i \in \{1,2,3,\ldots\}}$". The sum $\sum\limits_{i \geq 1} \left\lfloor \frac{n}{p^i} \right\rfloor$ in this equality is well-defined according to Definition 9.3.14, since it has only finitely many nonzero addends.)

For example, applied to $n = 5$ and $p = 2$, Theorem 9.3.25 predicts that

$$\begin{aligned} v_2(5!) &= \sum_{i \geq 1} \left\lfloor \frac{5}{2^i} \right\rfloor = \underbrace{\left\lfloor \frac{5}{2^1} \right\rfloor}_{=2} + \underbrace{\left\lfloor \frac{5}{2^2} \right\rfloor}_{=1} + \underbrace{\left\lfloor \frac{5}{2^3} \right\rfloor}_{=0} + \underbrace{\left\lfloor \frac{5}{2^4} \right\rfloor}_{=0} + \cdots \\ &= 2 + 1 + \underbrace{0 + 0 + \cdots}_{=0} = 2 + 1 = 3. \end{aligned}$$

This is easily confirmed (since $5! = 120 = 2^3 \cdot 3 \cdot 5$).

We shall give only a brief outline of the proof of Theorem 9.3.25, since this proof can be found in many places. (In particular, Theorem 9.3.25 is [Grinbe16, Theorem 1.3.3]. Moreover, it appears in an equivalent form in [19s, Exercise 2.17.2 **(c)**][294].)

The proof of Theorem 9.3.25 hinges on two lemmas:[295]

Lemma 9.3.26. Let $n \in \mathbb{N}$. Let k be a positive integer. Then, $\left\lfloor \frac{n}{k} \right\rfloor = \sum\limits_{i=1}^{n} [k \mid i]$.

Proof of Lemma 9.3.26 (sketched). (See [19s, Exercise 2.17.2 **(a)**] for details[296].) Each addend of the sum $\sum\limits_{i=1}^{n} [k \mid i]$ is either 0 or 1. Namely, an addend $[k \mid i]$ is 1 if i is a

[294] Indeed, [19s, Exercise 2.17.2 **(c)**] says that $v_p(n!) = \sum\limits_{i \geq 1} n//p^i$ instead of $v_p(n!) = \sum\limits_{i \geq 1} \left\lfloor \frac{n}{p^i} \right\rfloor$. However, this is equivalent, since an easy application of Proposition 3.3.5 yields that $n//p^i = \left\lfloor \frac{n}{p^i} \right\rfloor$ for any $i \in \mathbb{N}$.

[295] We are using the Iverson bracket notation (Definition 4.3.19) again.

[296] but notice that [19s, Exercise 2.17.2 **(a)**] talks about $n//k$ instead of $\left\lfloor \frac{n}{k} \right\rfloor$ (which, however, is the same number, since an easy application of Proposition 3.3.5 shows that $n//k = \left\lfloor \frac{n}{k} \right\rfloor$)

multiple of k, and 0 otherwise. Hence,

$$\begin{aligned}\sum_{i=1}^{n}[k \mid i] &= (\#\text{ of integers } i \in \{1,2,\ldots,n\} \text{ that are multiples of } k)\cdot 1 \\ &\qquad + (\#\text{ of integers } i \in \{1,2,\ldots,n\} \text{ that are not multiples of } k)\cdot 0 \\ &= (\#\text{ of integers } i \in \{1,2,\ldots,n\} \text{ that are multiples of } k) \\ &= n//k \qquad \left(\begin{array}{c}\text{since there are precisely } n//k \text{ many} \\ \text{multiples of } k \text{ among the integers } i \in \{1,2,\ldots,n\}\end{array}\right) \\ &= \left\lfloor \frac{n}{k} \right\rfloor\end{aligned}$$

(by an easy application of Proposition 3.3.5). This proves Lemma 9.3.26. □

Lemma 9.3.27. Let p be a prime. Let n be a nonzero integer. Then, $v_p(n) = \sum_{i\geq 1}\left[p^i \mid n\right]$.

Proof of Lemma 9.3.27 (sketched). (See [19s, Exercise 2.17.2 **(b)**] for details.) Let $k = v_p(n)$. Then, k is the largest $m \in \mathbb{N}$ such that $p^m \mid n$ (by the definition of $v_p(n)$). Hence, $p^k \mid n$ but $p^{k+1} \nmid n$. Thus, the k numbers $p^1, p^2, \ldots, p^k$ divide n, but none of the numbers $p^{k+1}, p^{k+2}, p^{k+3}, \ldots$ does. In other words, the statement $p^i \mid n$ holds for each $i \in \{1,2,\ldots,k\}$, but is false for each $i \in \{k+1, k+2, k+3, \ldots\}$. Hence, the truth value $\left[p^i \mid n\right]$ equals 1 for each $i \in \{1,2,\ldots,k\}$, but equals 0 for each $i \in \{k+1, k+2, k+3, \ldots\}$. Therefore, the sum $\sum_{i\geq 1}\left[p^i \mid n\right]$ has k addends equal to 1, while all remaining addends are 0. Therefore, this sum equals k. In other words, $\sum_{i\geq 1}\left[p^i \mid n\right] = k = v_p(n)$. This proves Lemma 9.3.27. □

It is now easy to prove Theorem 9.3.25:

Proof of Theorem 9.3.25 (sketched). (See [19s, Exercise 2.17.2 **(c)**] for details.) From $n! = 1\cdot 2\cdot\cdots\cdot n$, we obtain

$$\begin{aligned}v_p(n!) &= v_p(1\cdot 2\cdot\cdots\cdot n) = v_p(1) + v_p(2) + \cdots + v_p(n) \\ &\qquad \text{(by Corollary 9.3.7, applied to } k = n \text{ and } a_i = i) \\ &= \sum_{m=1}^{n} \underbrace{v_p(m)}_{\substack{=\sum_{i\geq 1}\left[p^i \mid m\right] \\ \text{(by Lemma 9.3.27)}}} = \sum_{m=1}^{n}\sum_{i\geq 1}\left[p^i \mid m\right] = \sum_{m=1}^{n}\sum_{j\geq 1}\left[p^j \mid m\right] \\ &\qquad \text{(here, we have renamed the summation index } i \text{ as } j) \\ &= \sum_{j\geq 1}\sum_{m=1}^{n}\left[p^j \mid m\right]. \qquad (427)\end{aligned}$$

(In the last step of this computation, we have interchanged the two summation signs "$\sum\limits_{m=1}^{n}$" and "$\sum\limits_{j\geq 1}$". Make sure you understand why this is legitimate![297]) But each positive integer j satisfies

$$\begin{aligned}\left\lfloor \frac{n}{p^j} \right\rfloor &= \sum_{i=1}^{n} \left[p^j \mid i \right] \qquad \left(\text{by Lemma 9.3.26, applied to } k = p^j\right) \\ &= \sum_{m=1}^{n} \left[p^j \mid m \right] \qquad (428)\end{aligned}$$

(here, we have renamed the summation index i as m). Thus, (427) becomes

$$v_p(n!) = \sum_{j\geq 1} \underbrace{\sum_{m=1}^{n} \left[p^j \mid m \right]}_{\substack{= \left\lfloor \frac{n}{p^j} \right\rfloor \\ \text{(by (428))}}} = \sum_{j\geq 1} \left\lfloor \frac{n}{p^j} \right\rfloor = \sum_{i\geq 1} \left\lfloor \frac{n}{p^i} \right\rfloor$$

(here, we have renamed the summation index j as i). This proves Theorem 9.3.25. □

Theorem 9.3.25 tends to be useful whenever it comes to proving divisibilities involving factorials. The following is perhaps the simplest application:

Exercise 9.3.6. Let $n, m \in \mathbb{N}$. Prove that $n!m! \mid (n+m)!$.

We shall give two solutions to Exercise 9.3.6: one using binomial coefficients, and one using p-valuations. The first solution is shorter and neater, but the second illustrates a generalizable method.

First solution to Exercise 9.3.6. We have $m \leq n+m$ (since $(n+m) - m = n \geq 0$). Hence, Theorem 4.3.8 (applied to $n+m$ and m instead of n and k) yields

$$\binom{n+m}{m} = \frac{(n+m)!}{m! \cdot ((n+m)-m)!} = \frac{(n+m)!}{m! \cdot n!} \qquad (\text{since } (n+m) - m = n).$$

However, Theorem 4.3.15 (applied to $n+m$ and m instead of n and k) yields $\binom{n+m}{m} \in \mathbb{Z}$. In view of $\binom{n+m}{m} = \frac{(n+m)!}{m! \cdot n!}$, this rewrites as $\frac{(n+m)!}{m! \cdot n!} \in \mathbb{Z}$. In other words, $m! \cdot n! \mid (n+m)!$. In other words, $n!m! \mid (n+m)!$ (since $m! \cdot n! = n!m!$). This solves Exercise 9.3.6. □

[297] It is not always legitimate to interchange two infinite sums, even if both are well-defined. Fortunately, one of our two summation signs – viz., "$\sum\limits_{m=1}^{n}$" – is a finite sum.

Second solution to Exercise 9.3.6 (sketched). Proposition 9.3.18 (applied to $n!m!$ and $(n+m)!$ instead of n and m) shows that we have $n!m! \mid (n+m)!$ if and only if each prime p satisfies $v_p(n!m!) \leq v_p((n+m)!)$. Thus, it suffices to show that each prime p satisfies $v_p(n!m!) \leq v_p((n+m)!)$.

Let us show this. Let p be a prime. We must prove that $v_p(n!m!) \leq v_p((n+m)!)$. Theorem 9.3.25 (applied to $n+m$ instead of n) yields

$$v_p((n+m)!) = \sum_{i \geq 1} \left\lfloor \frac{n+m}{p^i} \right\rfloor. \tag{429}$$

Also, Theorem 9.3.6 **(a)** (applied to $a = n!$ and $b = m!$) yields

$$\begin{aligned} v_p(n!m!) &= \underbrace{v_p(n!)}_{\substack{= \sum_{i \geq 1} \left\lfloor \frac{n}{p^i} \right\rfloor \\ \text{(by Theorem 9.3.25)}}} + \underbrace{v_p(m!)}_{\substack{= \sum_{i \geq 1} \left\lfloor \frac{m}{p^i} \right\rfloor \\ \text{(by Theorem 9.3.25)}}} \\ &= \sum_{i \geq 1} \left\lfloor \frac{n}{p^i} \right\rfloor + \sum_{i \geq 1} \left\lfloor \frac{m}{p^i} \right\rfloor. \end{aligned} \tag{430}$$

Recall that our goal is to prove that $v_p(n!m!) \leq v_p((n+m)!)$. In view of (429) and (430), this is equivalent to proving that

$$\sum_{i \geq 1} \left\lfloor \frac{n}{p^i} \right\rfloor + \sum_{i \geq 1} \left\lfloor \frac{m}{p^i} \right\rfloor \leq \sum_{i \geq 1} \left\lfloor \frac{n+m}{p^i} \right\rfloor.$$

This would be easy if we knew that we have

$$\left\lfloor \frac{n}{p^i} \right\rfloor + \left\lfloor \frac{m}{p^i} \right\rfloor \leq \left\lfloor \frac{n+m}{p^i} \right\rfloor \qquad \text{for each } i \in \mathbb{N}.$$

Fortunately, this is indeed the case: We can show that $\lfloor u \rfloor + \lfloor v \rfloor \leq \lfloor u+v \rfloor$ for any two reals u and v [298]. Applying this to $u = \frac{n}{p^i}$ and $v = \frac{m}{p^i}$, we obtain the inequality

$$\left\lfloor \frac{n}{p^i} \right\rfloor + \left\lfloor \frac{m}{p^i} \right\rfloor \leq \left\lfloor \frac{n}{p^i} + \frac{m}{p^i} \right\rfloor = \left\lfloor \frac{n+m}{p^i} \right\rfloor \qquad \left(\text{since } \frac{n}{p^i} + \frac{m}{p^i} = \frac{n+m}{p^i} \right)$$

for each $i \in \mathbb{N}$. Summing these inequalities over all $i \in \{1, 2, 3, \ldots\}$, we obtain

$$\sum_{i \geq 1} \left(\left\lfloor \frac{n}{p^i} \right\rfloor + \left\lfloor \frac{m}{p^i} \right\rfloor \right) \leq \sum_{i \geq 1} \left\lfloor \frac{n+m}{p^i} \right\rfloor.$$

[298] *Proof.* (See [Grinbe16, Proposition 1.1.13] for a detailed proof.) Let u and v be two reals. We have $\lfloor u \rfloor \leq u$ (by the definition of $\lfloor u \rfloor$) and $\lfloor v \rfloor \leq v$. Adding these two inequalities together, we obtain $\lfloor u \rfloor + \lfloor v \rfloor \leq u + v$. Hence, $\lfloor u \rfloor + \lfloor v \rfloor$ is an integer that is $\leq u + v$ (since $\lfloor u \rfloor + \lfloor v \rfloor$ is clearly an integer). But $\lfloor u+v \rfloor$ is the **largest** such integer (by the definition of $\lfloor u+v \rfloor$). Therefore, $\lfloor u \rfloor + \lfloor v \rfloor \leq \lfloor u+v \rfloor$, qed.

In view of

$$\sum_{i\geq 1}\left(\left\lfloor\frac{n}{p^i}\right\rfloor+\left\lfloor\frac{m}{p^i}\right\rfloor\right)=\sum_{i\geq 1}\left\lfloor\frac{n}{p^i}\right\rfloor+\sum_{i\geq 1}\left\lfloor\frac{m}{p^i}\right\rfloor=v_p\left(n!m!\right)\qquad\text{(by (430))}$$

and

$$\sum_{i\geq 1}\left\lfloor\frac{n+m}{p^i}\right\rfloor=v_p\left((n+m)!\right)\qquad\text{(by (429))},$$

this rewrites as $v_p\left(n!m!\right)\leq v_p\left((n+m)!\right)$. But this is precisely what we wanted to show. Hence, Exercise 9.3.6 is solved again. □

Let us prove something less straightforward using de Polignac's formula. For example, let us prove the following ([Grinbe15, Exercise 3.25]):

Exercise 9.3.7. Let $a,b\in\mathbb{N}$. Prove that $\dfrac{(2a)!\,(2b)!}{a!b!\,(a+b)!}\in\mathbb{Z}$.

Better yet, let us prove a more general result:

Exercise 9.3.8. Let $n\in\mathbb{N}$. Let $q_1,q_2,\ldots,q_n$ be n positive integers such that $\dfrac{1}{q_1}+\dfrac{1}{q_2}+\cdots+\dfrac{1}{q_n}\leq 1$. Let $a_1,a_2,\ldots,a_n$ be n nonnegative integers. Prove that

$$\frac{(q_1a_1)!\cdot(q_2a_2)!\cdot\cdots\cdot(q_na_n)!}{(a_1!)^{q_1-1}\cdot(a_2!)^{q_2-1}\cdot\cdots\cdot(a_n!)^{q_n-1}\cdot(a_1+a_2+\cdots+a_n)!}\in\mathbb{Z}.$$

Note that Exercise 9.3.7 is a particular case of Exercise 9.3.8 (namely, the case for $n=2$ and $q_1=2$ and $q_2=2$ and $a_1=a$ and $a_2=b$). Thus, solving Exercise 9.3.8 will automatically solve Exercise 9.3.7.

Before we solve Exercise 9.3.8, we pave our way with a lemma about floors (which will play a similar role to the inequality $\lfloor u\rfloor+\lfloor v\rfloor\leq\lfloor u+v\rfloor$ in our above second solution to Exercise 9.3.6):

Lemma 9.3.28. Let $n\in\mathbb{N}$. Let $q_1,q_2,\ldots,q_n$ be n positive integers such that $\dfrac{1}{q_1}+\dfrac{1}{q_2}+\cdots+\dfrac{1}{q_n}\leq 1$. Let $u_1,u_2,\ldots,u_n$ be n reals. Then,

$$\sum_{i=1}^{n}\lfloor q_iu_i\rfloor\geq\sum_{i=1}^{n}(q_i-1)\lfloor u_i\rfloor+\lfloor u_1+u_2+\cdots+u_n\rfloor.$$

There are various ways to prove this lemma. A particularly easy one proceeds by WLOG assuming that each of $u_1,u_2,\ldots,u_n$ belongs to the half-open interval $[0,1)$ (because subtracting an integer from any single u_i changes both sides of the

inequality in question by the same amount[299]). In this case, Lemma 9.3.28 quickly boils down to the following:

Lemma 9.3.29. Let $n \in \mathbb{N}$. Let $q_1, q_2, \ldots, q_n$ be n positive integers such that $\frac{1}{q_1} + \frac{1}{q_2} + \cdots + \frac{1}{q_n} \leq 1$. Let $u_1, u_2, \ldots, u_n$ be n nonnegative reals. Then,

$$\sum_{i=1}^{n} \lfloor q_i u_i \rfloor \geq \lfloor u_1 + u_2 + \cdots + u_n \rfloor .$$

Proof of Lemma 9.3.29. It is easy to see that Lemma 9.3.29 holds for $n = 0$ [300]. Hence, for the rest of this proof, we WLOG assume that we don't have $n = 0$. Hence, $n > 0$ (since $n \in \mathbb{N}$).

Let $k = \lfloor u_1 + u_2 + \cdots + u_n \rfloor$. Thus, k is an integer (since the floor of any real is an integer). Furthermore, the sum $u_1 + u_2 + \cdots + u_n$ is nonnegative (since its addends $u_1, u_2, \ldots, u_n$ are nonnegative). In other words, $u_1 + u_2 + \cdots + u_n \geq 0$. From this, we easily see that $k \geq 0$ [301].

Now, I claim that at least one $j \in \{1, 2, \ldots, n\}$ satisfies $u_j \geq \frac{k}{q_j}$.

[*Proof:* Assume the contrary. Hence, no $j \in \{1, 2, \ldots, n\}$ satisfies $u_j \geq \frac{k}{q_j}$. In other words, each $j \in \{1, 2, \ldots, n\}$ satisfies $u_j < \frac{k}{q_j}$. Adding these n inequalities $u_j < \frac{k}{q_j}$

[299] This technique is rather similar to the periodicity argument we used in our second solution to Exercise 1.1.3 above (in Subsection 4.7.2).

[300] *Proof.* Assume that $n = 0$. Then, $u_1 + u_2 + \cdots + u_n = u_1 + u_2 + \cdots + u_0 = (\text{empty sum}) = 0$, so that $\lfloor u_1 + u_2 + \cdots + u_n \rfloor = \lfloor 0 \rfloor = 0$. Also, from $n = 0$, we obtain

$$\sum_{i=1}^{n} \lfloor q_i u_i \rfloor = \sum_{i=1}^{0} \lfloor q_i u_i \rfloor = (\text{empty sum}) = 0 \geq 0 = \lfloor u_1 + u_2 + \cdots + u_n \rfloor$$

(since $\lfloor u_1 + u_2 + \cdots + u_n \rfloor = 0$). Thus, we have proved Lemma 9.3.29 under the assumption that $n = 0$. Qed.

[301] *Proof.* We have $0 \leq u_1 + u_2 + \cdots + u_n$ (since $u_1 + u_2 + \cdots + u_n \geq 0$). Thus, 0 is an integer that is $\leq u_1 + u_2 + \cdots + u_n$.

Recall that $\lfloor u_1 + u_2 + \cdots + u_n \rfloor$ was defined as the **largest** integer that is $\leq u_1 + u_2 + \cdots + u_n$. Hence, if m is an integer that is $\leq u_1 + u_2 + \cdots + u_n$, then $\lfloor u_1 + u_2 + \cdots + u_n \rfloor \geq m$. Applying this to $m = 0$, we obtain $\lfloor u_1 + u_2 + \cdots + u_n \rfloor \geq 0$ (since 0 is an integer that is $\leq u_1 + u_2 + \cdots + u_n$). Hence, $k = \lfloor u_1 + u_2 + \cdots + u_n \rfloor \geq 0$.

(for $j \in \{1, 2, \ldots, n\}$) together, we obtain

$$\begin{aligned}
\sum_{j=1}^{n} u_j &< \sum_{j=1}^{n} \frac{k}{q_j} \qquad (\text{since } n > 0) \\
&= \frac{k}{q_1} + \frac{k}{q_2} + \cdots + \frac{k}{q_n} = k \cdot \underbrace{\left(\frac{1}{q_1} + \frac{1}{q_2} + \cdots + \frac{1}{q_n} \right)}_{\leq 1} \\
&\leq k \cdot 1 \qquad (\text{since } k \geq 0) \\
&= k.
\end{aligned}$$

However, (1) (applied to $x = \sum_{j=1}^{n} u_j$) yields $\left\lfloor \sum_{j=1}^{n} u_j \right\rfloor \leq \sum_{j=1}^{n} u_j < \left\lfloor \sum_{j=1}^{n} u_j \right\rfloor + 1$. Hence, $\left\lfloor \sum_{j=1}^{n} u_j \right\rfloor \leq \sum_{j=1}^{n} u_j < k$. In view of $\sum_{j=1}^{n} u_j = u_1 + u_2 + \cdots + u_n$, this rewrites as $\lfloor u_1 + u_2 + \cdots + u_n \rfloor < k$. But this contradicts $\lfloor u_1 + u_2 + \cdots + u_n \rfloor = k$. This contradiction shows that our assumption was false. Hence, we have proved that at least one $j \in \{1, 2, \ldots, n\}$ satisfies $u_j \geq \dfrac{k}{q_j}$.]

Thus, we know that at least one $j \in \{1, 2, \ldots, n\}$ satisfies $u_j \geq \dfrac{k}{q_j}$. Consider this j. Multiplying both sides of the inequality $u_j \geq \dfrac{k}{q_j}$ by q_j, we obtain $q_j u_j \geq k$ (since q_j is positive). That is, $k \leq q_j u_j$. Hence, k is an integer that is $\leq q_j u_j$ (since k is an integer). Hence, it easily follows that $\lfloor q_j u_j \rfloor \geq k$ [302].

Furthermore, recall again that $u_1, u_2, \ldots, u_n$ are nonnegative. Hence, it is easy to see that

$$\lfloor q_i u_i \rfloor \geq 0 \qquad \text{for each } i \in \{1, 2, \ldots, n\} \tag{431}$$

[302] *Proof.* Recall that $\lfloor q_j u_j \rfloor$ was defined as the **largest** integer that is $\leq q_j u_j$. Hence, if m is an integer that is $\leq q_j u_j$, then $\lfloor q_j u_j \rfloor \geq m$. Applying this to $m = k$, we obtain $\lfloor q_j u_j \rfloor \geq k$ (since k is an integer that is $\leq q_j u_j$).

[303]. From this, it easily follows that $\sum_{i=1}^{n} \lfloor q_i u_i \rfloor \geq \lfloor q_j u_j \rfloor$ [304]. Hence,

$$\sum_{i=1}^{n} \lfloor q_i u_i \rfloor \geq \lfloor q_j u_j \rfloor \geq k = \lfloor u_1 + u_2 + \cdots + u_n \rfloor .$$

This proves Lemma 9.3.29. □

Let us now derive Lemma 9.3.28 from Lemma 9.3.29:

Proof of Lemma 9.3.28. For each $i \in \{1, 2, \ldots, n\}$, we set $v_i = \lfloor u_i \rfloor$ and $w_i = u_i - \lfloor u_i \rfloor$. Thus, $v_1, v_2, \ldots, v_n$ are n integers[305]. Furthermore, the reals $w_1, w_2, \ldots, w_n$ are nonnegative[306]. Hence, Lemma 9.3.29 (applied to w_i instead of u_i) yields

$$\sum_{i=1}^{n} \lfloor q_i w_i \rfloor \geq \lfloor w_1 + w_2 + \cdots + w_n \rfloor . \tag{432}$$

However, it is easy to see that each $i \in \{1, 2, \ldots, n\}$ satisfies the equality

$$\lfloor q_i u_i \rfloor = \lfloor q_i w_i \rfloor + q_i v_i \tag{433}$$

[303] *Proof of (431):* Let $i \in \{1, 2, \ldots, n\}$. Then, u_i is nonnegative (since $u_1, u_2, \ldots, u_n$ are nonnegative). Hence, $q_i u_i$ is nonnegative (since q_i is positive (because $q_1, q_2, \ldots, q_n$ are n positive integers)). In other words, $q_i u_i \geq 0$. That is, $0 \leq q_i u_i$. Thus, 0 is an integer that is $\leq q_i u_i$.

Recall that $\lfloor q_i u_i \rfloor$ was defined as the **largest** integer that is $\leq q_i u_i$. Hence, if m is an integer that is $\leq q_i u_i$, then $\lfloor q_i u_i \rfloor \geq m$. Applying this to $m = 0$, we obtain $\lfloor q_i u_i \rfloor \geq 0$ (since 0 is an integer that is $\leq q_i u_i$). This proves (431).

[304] *Proof.* We have

$$\underbrace{\sum_{i=1}^{n}}_{= \sum_{i \in \{1,2,\ldots,n\}}} \lfloor q_i u_i \rfloor = \sum_{i \in \{1,2,\ldots,n\}} \lfloor q_i u_i \rfloor = \lfloor q_j u_j \rfloor + \sum_{\substack{i \in \{1,2,\ldots,n\}; \\ i \neq j}} \underbrace{\lfloor q_i u_i \rfloor}_{\substack{\geq 0 \\ \text{(by (431))}}}$$

(here, we have split off the addend for $i = j$ from the sum)

$$\geq \lfloor q_j u_j \rfloor + \underbrace{\sum_{\substack{i \in \{1,2,\ldots,n\}; \\ i \neq j}} 0}_{=0} = \lfloor q_j u_j \rfloor .$$

[305] *Proof.* Let $i \in \{1, 2, \ldots, n\}$. It is clear that $\lfloor x \rfloor$ is an integer for each $x \in \mathbb{R}$. Applying this to $x = u_i$, we conclude that $\lfloor u_i \rfloor$ is an integer. In other words, v_i is an integer (since $v_i = \lfloor u_i \rfloor$).

Forget that we fixed i. We thus have shown that v_i is an integer for each $i \in \{1, 2, \ldots, n\}$. In other words, $v_1, v_2, \ldots, v_n$ are n integers.

[306] *Proof.* Let $i \in \{1, 2, \ldots, n\}$. Applying (1) to $x = u_i$, we obtain $\lfloor u_i \rfloor \leq u_i < \lfloor u_i \rfloor + 1$. Hence, $\lfloor u_i \rfloor \leq u_i$, so that $u_i - \lfloor u_i \rfloor \geq 0$. Now, $w_i = u_i - \lfloor u_i \rfloor \geq 0$. In other words, w_i is nonnegative.

Forget that we fixed i. We thus have shown that w_i is nonnegative for each $i \in \{1, 2, \ldots, n\}$. In other words, $w_1, w_2, \ldots, w_n$ are nonnegative.

[307]. Hence,

$$\sum_{i=1}^{n} \underbrace{\lfloor q_i u_i \rfloor}_{\substack{=\lfloor q_i w_i \rfloor + q_i v_i \\ \text{(by (433))}}} = \sum_{i=1}^{n} \left(\lfloor q_i w_i \rfloor + q_i v_i \right) = \underbrace{\sum_{i=1}^{n} \lfloor q_i w_i \rfloor}_{\substack{\geq \lfloor w_1 + w_2 + \cdots + w_n \rfloor \\ \text{(by (432))}}} + \sum_{i=1}^{n} q_i v_i$$

$$\geq \lfloor w_1 + w_2 + \cdots + w_n \rfloor + \sum_{i=1}^{n} q_i v_i. \tag{434}$$

On the other hand, it is easy to see that

$$\lfloor u_1 + u_2 + \cdots + u_n \rfloor = \lfloor w_1 + w_2 + \cdots + w_n \rfloor + \sum_{i=1}^{n} v_i \tag{435}$$

[307] *Proof of (433):* Let $i \in \{1, 2, \ldots, n\}$. Adding the two equalities $w_i = u_i - \lfloor u_i \rfloor$ and $v_i = \lfloor u_i \rfloor$ together, we obtain $w_i + v_i = (u_i - \lfloor u_i \rfloor) + \lfloor u_i \rfloor = u_i$.

However, q_i is an integer (since $q_1, q_2, \ldots, q_n$ are n integers), and v_i is an integer (since $v_1, v_2, \ldots, v_n$ are n integers). Thus, the product $q_i v_i$ of these two integers must also be an integer. In other words, $q_i v_i \in \mathbb{Z}$. Therefore, (163) (applied to $y = q_i w_i$ and $k = q_i v_i$) yields $\lfloor q_i w_i + q_i v_i \rfloor = \lfloor q_i w_i \rfloor + q_i v_i$. In view of $q_i w_i + q_i v_i = q_i \underbrace{(w_i + v_i)}_{=u_i} = q_i u_i$, this rewrites as $\lfloor q_i u_i \rfloor = \lfloor q_i w_i \rfloor + q_i v_i$. Qed.

[308]. On the other hand,

$$\sum_{i=1}^{n}(q_i-1)\underbrace{\lfloor u_i\rfloor}_{\substack{=v_i\\ \text{(since } v_i=\lfloor u_i\rfloor\\ \text{(by the definition of } v_i))}}=\sum_{i=1}^{n}(q_i-1)\,v_i.$$

[308] *Proof of (435):* The sum $\sum_{i=1}^{n} v_i$ is an integer (since its addends $v_1, v_2, \ldots, v_n$ are n integers). In other words, $\sum_{i=1}^{n} v_i \in \mathbb{Z}$. Therefore, (163) (applied to $y = \sum_{i=1}^{n} w_i$ and $k = \sum_{i=1}^{n} v_i$) yields

$$\left\lfloor \sum_{i=1}^{n} w_i + \sum_{i=1}^{n} v_i \right\rfloor = \left\lfloor \sum_{i=1}^{n} w_i \right\rfloor + \sum_{i=1}^{n} v_i.$$

In view of

$$\begin{aligned}\sum_{i=1}^{n}\underbrace{w_i}_{\substack{=u_i-\lfloor u_i\rfloor\\ \text{(by the definition of } w_i)}}+\sum_{i=1}^{n}\underbrace{v_i}_{\substack{=\lfloor u_i\rfloor\\ \text{(by the definition of } v_i)}}&=\sum_{i=1}^{n}(u_i-\lfloor u_i\rfloor)+\sum_{i=1}^{n}\lfloor u_i\rfloor\\ &=\sum_{i=1}^{n}\underbrace{(u_i-\lfloor u_i\rfloor+\lfloor u_i\rfloor)}_{=u_i}=\sum_{i=1}^{n}u_i\\ &=u_1+u_2+\cdots+u_n,\end{aligned}$$

this rewrites as

$$\lfloor u_1+u_2+\cdots+u_n\rfloor=\left\lfloor\sum_{i=1}^{n}w_i\right\rfloor+\sum_{i=1}^{n}v_i.$$

In view of $\sum_{i=1}^{n} w_i = w_1 + w_2 + \cdots + w_n$, this rewrites as

$$\lfloor u_1+u_2+\cdots+u_n\rfloor=\lfloor w_1+w_2+\cdots+w_n\rfloor+\sum_{i=1}^{n}v_i.$$

Qed.

Adding (435) to this equality, we obtain

$$\begin{aligned}
&\sum_{i=1}^{n}(q_i-1)\lfloor u_i\rfloor+\lfloor u_1+u_2+\cdots+u_n\rfloor \\
&=\sum_{i=1}^{n}(q_i-1)v_i+\lfloor w_1+w_2+\cdots+w_n\rfloor+\sum_{i=1}^{n}v_i \\
&=\lfloor w_1+w_2+\cdots+w_n\rfloor+\underbrace{\sum_{i=1}^{n}v_i+\sum_{i=1}^{n}(q_i-1)v_i}_{=\sum_{i=1}^{n}(v_i+(q_i-1)v_i)} \\
&=\lfloor w_1+w_2+\cdots+w_n\rfloor+\sum_{i=1}^{n}\underbrace{(v_i+(q_i-1)v_i)}_{=q_iv_i} \\
&=\lfloor w_1+w_2+\cdots+w_n\rfloor+\sum_{i=1}^{n}q_iv_i.
\end{aligned}$$

In light of this equality, we can rewrite (434) as

$$\sum_{i=1}^{n}\lfloor q_iu_i\rfloor\geq\sum_{i=1}^{n}(q_i-1)\lfloor u_i\rfloor+\lfloor u_1+u_2+\cdots+u_n\rfloor.$$

This proves Lemma 9.3.28. □

At last, we can solve Exercise 9.3.8:

Solution to Exercise 9.3.8. Let p be a prime. We shall prove that

$$\begin{aligned}
&v_p\left((a_1!)^{q_1-1}\cdot(a_2!)^{q_2-1}\cdot\cdots\cdot(a_n!)^{q_n-1}\cdot(a_1+a_2+\cdots+a_n)!\right) \\
&\leq v_p\left((q_1a_1)!\cdot(q_2a_2)!\cdot\cdots\cdot(q_na_n)!\right).
\end{aligned}$$

Indeed, the n numbers $q_1-1,q_2-1,\ldots,q_n-1$ are nonnegative integers (since $q_1,q_2,\ldots,q_n$ are n positive integers). Hence, $(a_1!)^{q_1-1},(a_2!)^{q_2-1},\ldots,(a_n!)^{q_n-1}$ are n integers. Thus, Corollary 9.3.7 (applied to n and $(a_j!)^{q_j-1}$ instead of k and a_j) yields

$$\begin{aligned}
&v_p\left((a_1!)^{q_1-1}\cdot(a_2!)^{q_2-1}\cdot\cdots\cdot(a_n!)^{q_n-1}\right) \\
&=v_p\left((a_1!)^{q_1-1}\right)+v_p\left((a_2!)^{q_2-1}\right)+\cdots+v_p\left((a_n!)^{q_n-1}\right) \\
&=\sum_{j=1}^{n}v_p\left((a_j!)^{q_j-1}\right). \qquad (436)
\end{aligned}$$

Now, let $j\in\{1,2,\ldots,n\}$. Then, q_j is a positive integer (since $q_1,q_2,\ldots,q_n$ are n positive integers). Hence, $q_j-1\in\mathbb{N}$. Thus, Corollary 9.3.9 (applied to $a_j!$ and

$q_j - 1$ instead of a and k) yields

$$\begin{aligned} v_p\left(\left(a_j!\right)^{q_j-1}\right) &= \left(q_j-1\right) \underbrace{v_p\left(a_j!\right)}_{\substack{=\sum\limits_{i\geq 1}\left\lfloor \frac{a_j}{p^i}\right\rfloor \\ \text{(by Theorem 9.3.25,} \\ \text{applied to } a_j \text{ instead of } n)}} \\ &= \left(q_j-1\right)\sum_{i\geq 1}\left\lfloor \frac{a_j}{p^i}\right\rfloor = \sum_{i\geq 1}\left(q_j-1\right)\left\lfloor \frac{a_j}{p^i}\right\rfloor . \end{aligned} \tag{437}$$

Forget that we fixed j. We thus have proved (437) for each $j \in \{1, 2, \ldots, n\}$. Thus, (436) becomes

$$\begin{aligned} & v_p\left(\left(a_1!\right)^{q_1-1}\cdot\left(a_2!\right)^{q_2-1}\cdot\cdots\cdot\left(a_n!\right)^{q_n-1}\right) \\ &= \sum_{j=1}^{n}\underbrace{v_p\left(\left(a_j!\right)^{q_j-1}\right)}_{\substack{=\sum\limits_{i\geq 1}\left(q_j-1\right)\left\lfloor \frac{a_j}{p^i}\right\rfloor \\ \text{(by (437))}}} = \underbrace{\sum_{j=1}^{n}\sum_{i\geq 1}}_{=\sum\limits_{i\geq 1}\sum\limits_{j=1}^{n}}\left(q_j-1\right)\left\lfloor \frac{a_j}{p^i}\right\rfloor \\ &= \sum_{i\geq 1}\sum_{j=1}^{n}\left(q_j-1\right)\left\lfloor \frac{a_j}{p^i}\right\rfloor . \end{aligned} \tag{438}$$

Theorem 9.3.6 **(a)** (applied to $a = \left(a_1!\right)^{q_1-1}\cdot\left(a_2!\right)^{q_2-1}\cdot\cdots\cdot\left(a_n!\right)^{q_n-1}$ and $b = \left(a_1+a_2+\cdots+a_n\right)!$) yields

$$\begin{aligned} & v_p\left(\left(a_1!\right)^{q_1-1}\cdot\left(a_2!\right)^{q_2-1}\cdot\cdots\cdot\left(a_n!\right)^{q_n-1}\cdot\left(a_1+a_2+\cdots+a_n\right)!\right) \\ &= \underbrace{v_p\left(\left(a_1!\right)^{q_1-1}\cdot\left(a_2!\right)^{q_2-1}\cdot\cdots\cdot\left(a_n!\right)^{q_n-1}\right)}_{\substack{=\sum\limits_{i\geq 1}\sum\limits_{j=1}^{n}\left(q_j-1\right)\left\lfloor \frac{a_j}{p^i}\right\rfloor \\ \text{(by (438))}}} + \underbrace{v_p\left(\left(a_1+a_2+\cdots+a_n\right)!\right)}_{\substack{=\sum\limits_{i\geq 1}\left\lfloor \frac{a_1+a_2+\cdots+a_n}{p^i}\right\rfloor \\ \text{(by Theorem 9.3.25,} \\ \text{applied to } a_1+a_2+\cdots+a_n \text{ instead of } n)}} \\ &= \sum_{i\geq 1}\sum_{j=1}^{n}\left(q_j-1\right)\left\lfloor \frac{a_j}{p^i}\right\rfloor + \sum_{i\geq 1}\left\lfloor \frac{a_1+a_2+\cdots+a_n}{p^i}\right\rfloor \\ &= \sum_{i\geq 1}\left(\sum_{j=1}^{n}\left(q_j-1\right)\left\lfloor \frac{a_j}{p^i}\right\rfloor + \left\lfloor \frac{a_1+a_2+\cdots+a_n}{p^i}\right\rfloor\right) \\ &= \sum_{s\geq 1}\left(\sum_{i=1}^{n}\left(q_i-1\right)\left\lfloor \frac{a_i}{p^s}\right\rfloor + \left\lfloor \frac{a_1+a_2+\cdots+a_n}{p^s}\right\rfloor\right) \end{aligned} \tag{439}$$

(here, we have renamed the summation indices i and j of the two sums as s and i, respectively).

On the other hand, $(q_1a_1)!, (q_2a_2)!, \ldots, (q_na_n)!$ are n integers. Thus, Corollary 9.3.7 (applied to n and $(q_ja_j)!$ instead of k and a_j) yields

$$\begin{aligned}
&v_p\left((q_1a_1)! \cdot (q_2a_2)! \cdot \cdots \cdot (q_na_n)!\right) \\
&= v_p\left((q_1a_1)!\right) + v_p\left((q_2a_2)!\right) + \cdots + v_p\left((q_na_n)!\right) \\
&= \sum_{j=1}^{n} \underbrace{v_p\left((q_ja_j)!\right)}_{\substack{= \sum\limits_{i\geq 1} \left\lfloor \frac{q_ja_j}{p^i} \right\rfloor \\ \text{(by Theorem 9.3.25,} \\ \text{applied to } q_ja_j \text{ instead of } n)}} = \underbrace{\sum_{j=1}^{n} \sum_{i\geq 1}}_{= \sum\limits_{i\geq 1} \sum\limits_{j=1}^{n}} \underbrace{\left\lfloor \frac{q_ja_j}{p^i} \right\rfloor}_{\substack{= \left\lfloor q_j \cdot \frac{a_j}{p^i} \right\rfloor \\ \text{(since } \frac{q_ja_j}{p^i} = q_j \cdot \frac{a_j}{p^i}\text{)}}} \\
&= \sum_{i\geq 1} \sum_{j=1}^{n} \left\lfloor q_j \cdot \frac{a_j}{p^i} \right\rfloor = \sum_{s\geq 1} \underbrace{\sum_{i=1}^{n} \left\lfloor q_i \cdot \frac{a_i}{p^s} \right\rfloor}_{\substack{\geq \sum\limits_{i=1}^{n} (q_i - 1) \left\lfloor \frac{a_i}{p^s} \right\rfloor + \left\lfloor \frac{a_1}{p^s} + \frac{a_2}{p^s} + \cdots + \frac{a_n}{p^s} \right\rfloor \\ \text{(by Lemma 9.3.28, applied to } u_i = \frac{a_i}{p^s}\text{)}}} \\
&\qquad \left(\begin{array}{c}\text{here, we have renamed the summation indices } i \text{ and } j \\ \text{of the two sums as } s \text{ and } i\text{, respectively}\end{array}\right) \\
&\geq \sum_{s\geq 1} \left(\sum_{i=1}^{n} (q_i - 1) \left\lfloor \frac{a_i}{p^s} \right\rfloor + \left\lfloor \frac{a_1}{p^s} + \frac{a_2}{p^s} + \cdots + \frac{a_n}{p^s} \right\rfloor \right) \\
&= \sum_{s\geq 1} \left(\sum_{i=1}^{n} (q_i - 1) \left\lfloor \frac{a_i}{p^s} \right\rfloor + \left\lfloor \frac{a_1 + a_2 + \cdots + a_n}{p^s} \right\rfloor \right) \\
&\qquad \left(\text{since } \frac{a_1}{p^s} + \frac{a_2}{p^s} + \cdots + \frac{a_n}{p^s} = \frac{a_1 + a_2 + \cdots + a_n}{p^s} \text{ for any } s \geq 1\right) \\
&= v_p\left((a_1!)^{q_1 - 1} \cdot (a_2!)^{q_2 - 1} \cdot \cdots \cdot (a_n!)^{q_n - 1} \cdot (a_1 + a_2 + \cdots + a_n)!\right)
\end{aligned}$$

(by (439)).

Forget that we fixed p. We thus have shown that each prime p satisfies

$$\begin{aligned}
&v_p\left((a_1!)^{q_1 - 1} \cdot (a_2!)^{q_2 - 1} \cdot \cdots \cdot (a_n!)^{q_n - 1} \cdot (a_1 + a_2 + \cdots + a_n)!\right) \\
&\leq v_p\left((q_1a_1)! \cdot (q_2a_2)! \cdot \cdots \cdot (q_na_n)!\right).
\end{aligned}$$

Thus, Proposition 9.3.18 (applied to $(a_1!)^{q_1 - 1} \cdot (a_2!)^{q_2 - 1} \cdot \cdots \cdot (a_n!)^{q_n - 1} \cdot (a_1 + a_2 + \cdots + a_n)!$ and $(q_1a_1)! \cdot (q_2a_2)! \cdot \cdots \cdot (q_na_n)!$ instead of n and m) entails that

$$\begin{aligned}
&(a_1!)^{q_1 - 1} \cdot (a_2!)^{q_2 - 1} \cdot \cdots \cdot (a_n!)^{q_n - 1} \cdot (a_1 + a_2 + \cdots + a_n)! \\
&\mid (q_1a_1)! \cdot (q_2a_2)! \cdot \cdots \cdot (q_na_n)!.
\end{aligned}$$

In other words,

$$\frac{(q_1a_1)! \cdot (q_2a_2)! \cdot \cdots \cdot (q_na_n)!}{(a_1!)^{q_1-1} \cdot (a_2!)^{q_2-1} \cdot \cdots \cdot (a_n!)^{q_n-1} \cdot (a_1+a_2+\cdots+a_n)!} \in \mathbb{Z}$$

(since $(a_1!)^{q_1-1} \cdot (a_2!)^{q_2-1} \cdot \cdots \cdot (a_n!)^{q_n-1} \cdot (a_1+a_2+\cdots+a_n)!$ is positive and thus nonzero). This solves Exercise 9.3.8. □

As we already mentioned, Exercise 9.3.7 is a particular case of Exercise 9.3.8. However, Exercise 9.3.7 can also be proved algebraically, via binomial identities. (See [Grinbe15, Exercise 3.25 **(b)**] for such a proof[309], and [Grinbe15, Exercise 3.25] for further properties of $\frac{(2a)!\,(2b)!}{a!b!\,(a+b)!}$.) I am not aware of a similar solution to Exercise 9.3.8.

9.3.6. Homework set #10B: More about primes and numbers

Again, this homework set is optional.

Exercise 9.3.9. Let $n \in \mathbb{Z}$ be nonzero. Prove the following:

(a) The product $\prod_{p \text{ prime}} (v_p(n)+1)$ is well-defined, since only finitely many of its factors are different from 1.

(b) We have

$$(\text{the number of positive divisors of } n) = \prod_{p \text{ prime}} (v_p(n)+1)$$

and

$$(\text{the number of divisors of } n) = 2 \prod_{p \text{ prime}} (v_p(n)+1).$$

The next exercise extends the notion of the p-adic valuation of an integer to a rational number:

Exercise 9.3.10. Fix a prime p. For each nonzero rational number r, define an integer $w_p(r)$ (called the *extended p-adic valuation* of r) as follows: We write r in the form $r = a/b$ for two nonzero integers a and b, and we set $w_p(r) = v_p(a) - v_p(b)$. (It also makes sense to set $w_p(0) = \infty$, but we shall not concern ourselves with this border case in this exercise.)

[309] In more detail: The number $\frac{(2a)!\,(2b)!}{a!b!\,(a+b)!}$ is what is called $T(a,b)$ in [Grinbe15, Exercise 3.25]. Thus, [Grinbe15, Exercise 3.25 **(b)**] (applied to $m=a$ and $n=b$) shows that $\frac{(2a)!\,(2b)!}{a!b!\,(a+b)!} \in \mathbb{N} \subseteq \mathbb{Z}$.

(a) Prove that this is well-defined – i.e., that $w_p(r)$ does not depend on the precise choice of a and b satisfying $r = a/b$.

(b) Prove that $w_p(n) = v_p(n)$ for each nonzero integer n.

(c) Prove that $w_p(ab) = w_p(a) + w_p(b)$ for any two nonzero rational numbers a and b.

(d) Prove that $w_p(a+b) \geq \min\{w_p(a), w_p(b)\}$ for any two nonzero rational numbers a and b if $a + b \neq 0$.

(e) Prove that $r = \prod_{p \text{ prime}} p^{w_p(r)}$ for any positive rational number r. (This is a generalization of the canonical factorization to rational numbers.)

The following exercise is a generalization of Exercise 1.1.4:

Exercise 9.3.11. Let n and u be positive integers. Let $a_1, a_2, \ldots, a_u$ be any integers. Set $a_{u+1} = a_1$. Assume that

$$a_i \mid a_{i+1}^n \qquad \text{for each } i \in [u].$$

Set $m = n^{u-1} + n^{u-2} + \cdots + n^0$. Prove that

$$a_1 a_2 \cdots a_u \mid (a_1 + a_2 + \cdots + a_u)^m.$$

The next two exercises are parts of what is known as the *lifting-the-exponent lemma*:

Exercise 9.3.12. Let p be a nonnegative integer (not necessarily prime). Let a and b be two integers such that $a \equiv b \bmod p$.

(a) Prove that $a^p \equiv b^p \bmod p^2$.

(b) Prove that $a^{p^i} \equiv b^{p^i} \bmod p^{i+1}$ for each $i \in \mathbb{N}$. (Keep in mind that a^{p^i} means $a^{(p^i)}$.)

Exercise 9.3.13. Let p be a prime. Let a and b be two integers such that $a \equiv b \not\equiv 0 \bmod p$. Let n be a positive integer.

(a) Prove that $v_p(a^n - b^n) \geq v_p(a-b) + v_p(n)$.

(b) Prove that $v_p(a^n - b^n) = v_p(a-b) + v_p(n)$ if $p \neq 2$.

(c) Prove that $v_p(a^n - b^n) = v_p(a-b) + v_p(n)$ if $p = 2$ and $a \equiv b \bmod 4$.

(d) Find an example where $p = 2$ and $a \equiv b \bmod 2$ and $v_p(a^n - b^n) > v_p(a-b) + v_p(n)$.

See `https://brilliant.org/wiki/lifting-the-exponent` or [AndDos12, §3.5] or [Chen20, OTIS Excerpts, Chapter 13] or [Parvar11] for applications of Exercise 9.3.13.

The next two exercises are connected to p-valuations of factorials:

Exercise 9.3.14. Let p be a prime. Let $k \in \mathbb{N}$. Let $N = p^k - 1$. Let $c \in \{1, 2, \ldots, p^k\}$. Prove that

$$v_p((Nc)!) = c\left(p^0 + p^1 + \cdots + p^{k-1}\right) - k.$$

Exercise 9.3.15. Let $n \in \mathbb{N}$ be odd. Let $m = (n-1)/2$. Let $k \in \mathbb{N}$. Let $a_1, a_2, \ldots, a_k$ be nonnegative integers.

(a) Prove that the number

$$\frac{(na_1)!\,(na_2)!\cdots(na_k)!}{a_1!a_2!\cdots a_k!\cdot((a_1+a_2)!)^m((a_2+a_3)!)^m\cdots((a_{k-1}+a_k)!)^m((a_k+a_1)!)^m}$$

is an integer.

(b) Show that this integer furthermore is divisible by $n^{k-\lfloor k/2\rfloor}$ when n is prime and $a_1, a_2, \ldots, a_k$ are positive.

The next exercise gives yet another variation on Fermat's Little Theorem:

Exercise 9.3.16. Let $a \in \mathbb{Z}$ and $n \in \mathbb{N}$. Prove that

$$n! \mid \prod_{k=0}^{n-1}\left(a^n - a^k\right).$$

This might sound familiar to algebraists: Indeed, if a is prime, then $\prod_{k=0}^{n-1}\left(a^n - a^k\right)$ is the order of the general linear group $\mathrm{GL}_n(\mathbb{F}_a)$ over the finite field $\mathbb{F}_a$, whereas $n!$ is the order of the symmetric group S_n. Thus, if a is prime, then Exercise 9.3.16 follows from Lagrange's theorem, as the symmetric group S_n embeds as a subgroup into $\mathrm{GL}_n(\mathbb{F}_a)$ (via the embedding that identifies each permutation with its permutation matrix). More generally, this argument can be made when a is a power of a prime (not necessarily a prime itself), since there is a finite field of size a in this case. However, this slick argument does not work when a is not a power of a prime. I am not aware of any way how to generalize it to arbitrary a; it appears to be a dead end, at least if solving Exercise 9.3.16 for general a is the goal.

The following exercise is an analogue of Proposition 9.3.11 for lowest common multiples[310]:

Exercise 9.3.17. Let p be a prime. Let k be a positive integer. Let $n_1, n_2, \ldots, n_k$ be k integers. Then,

$$v_p(\operatorname{lcm}(n_1, n_2, \ldots, n_k)) = \max\{v_p(n_1), v_p(n_2), \ldots, v_p(n_k)\}.$$

[310] See Section 3.6 for the definition of lowest common multiples and their most important properties.

The following exercise is Problem B3 from the Putnam contest 2003:

Exercise 9.3.18. Let $n \in \mathbb{N}$. Prove that

$$n! = \prod_{k=1}^{n} \operatorname{lcm}\left(1, 2, \ldots, \left\lfloor \frac{n}{k} \right\rfloor\right).$$

See [Chen20, OTIS Excerpts, Chapter 13] and [AndDos10, Chapter 3] and [AndDos12, Chapter 3] for more about p-valuations.

The next exercise expands on the idea of Exercise 9.2.1:

Exercise 9.3.19. Prove the following:

(a) There are infinitely many primes p that satisfy $p \equiv 3 \bmod 4$.

(b) There are infinitely many primes p that satisfy $p \equiv 5 \bmod 6$.

We note that Exercise 9.2.1 and both parts of Exercise 9.3.19 are particular cases of *Dirichlet's theorem*, which says that if a and $b > 0$ are two coprime integers, then there are infinitely many primes p that satisfy $p \equiv a \bmod b$. This theorem is deep and difficult (all known proofs use some kind of analysis, often complex); the three particular cases we just mentioned are among the few cases that have known simple proofs.

The next exercise generalizes Exercise 9.1.7:

Exercise 9.3.20. Let p be a prime. Let $k \in \mathbb{N}$ be such that k is not a positive multiple of $p-1$. Prove that $\sum_{i=0}^{p-1} i^k \equiv 0 \bmod p$.

The next exercise is a fundamental property of the Euler totient function, originally found by Gauss:

Exercise 9.3.21. For any positive integer n, we let $\phi(n)$ denote the number of all $i \in \{1, 2, \ldots, n\}$ satisfying $i \perp n$.

Prove that

$$\sum_{d|n} \phi(d) = n \qquad \text{for any positive integer } n.$$

Here, the summation sign "$\sum_{d|n}$" means a sum over all **positive** divisors d of n. (For example, $\sum_{d|15} \phi(d) = \phi(1) + \phi(3) + \phi(5) + \phi(15)$.)

The next exercise is somewhat similar to Exercise 9.3.1: It says that any equality of the form $ab = cd$, where a, b, c, d are integers, can be "explained" by decomposing a, b, c, d further into products.

Exercise 9.3.22. Let a, b, c, d be four integers such that $ab = cd$. Prove that there exist four integers x, y, z, w such that

$$a = xy, \qquad b = zw, \qquad c = xz, \qquad d = yw.$$

The next exercise is one of the simplest forms of what is known as the *Chinese Remainder Theorem*:

Exercise 9.3.23. Let m and n be two positive integers. Prove that the map

$$\begin{aligned} \{0, 1, \ldots, mn-1\} &\to \{0, 1, \ldots, m-1\} \times \{0, 1, \ldots, n-1\}, \\ i &\mapsto (i\%m,\ i\%n) \end{aligned}$$

is a bijection if and only if m and n are coprime.

The Chinese Remainder Theorem has several more advanced forms, but even Exercise 9.3.23 has many uses. Here is just one application:

Exercise 9.3.24. For any positive integer n, we let $\phi(n)$ denote the number of all $i \in \{1, 2, \ldots, n\}$ satisfying $i \perp n$.

Prove that $\phi(mn) = \phi(m) \cdot \phi(n)$ for any two coprime positive integers m and n.

Exercise 9.3.24 is known as the *multiplicativity of the Euler totient function*. One of its consequences is the following "explicit" formula for $\phi(n)$:

Exercise 9.3.25. Let n be a positive integer. Let $\phi(n)$ denote the number of all $i \in \{1, 2, \ldots, n\}$ satisfying $i \perp n$. Prove that

$$\phi(n) = \prod_{\substack{p \text{ prime;} \\ p \mid n}} \left((p-1)\, p^{v_p(n)-1} \right) = n \cdot \prod_{\substack{p \text{ prime;} \\ p \mid n}} \left(1 - \frac{1}{p} \right).$$

Finally, a complex-looking but not particularly difficult exercise:

Exercise 9.3.26. Assume that p is no longer required to be a prime in Definition 9.3.1, but is merely assumed to be an integer > 1. Which of the above properties of p-valuations remain true, and which become false? Specifically, analyze Lemma 9.3.3, Corollary 9.3.4, Lemma 9.3.5, Theorem 9.3.6, Proposition 9.3.11 and Exercise 9.3.17.

A. Discussion of homework questions

A.1. Homework set #0 discussion

The following are discussions of the problems on homework set #0 (Section 1.1). I distinguish "discussion" from "solution" in that a solution is a (usually rather rigorous) mathematical proof while a "discussion" is an (often informal) piece of writing that (at least hopefully) explains how a solution can be obtained. Thus, a discussion can be both less than a solution (as it can omit certain straightforward arguments that a solution would have to include) and more than a solution (as it can talk about the motivation behind the solution and about how similar problems can be solved). I hope my discussions will be more instructive than pure (unmotivated) solutions would be, while still being sufficiently rigorous that you won't have much trouble turning them into full solutions. As always, there is no guarantee that I will succeed at these goals. Let me know of any unclarities and errors!

A.1.1. Discussion of Exercise 1.1.1

Discussion of Exercise 1.1.1. The obvious thing to do against such a problem is to check small values of n and see what the corresponding q's are. This is somewhat complicated by the fact that for some n's, there are several q's that work. Here is a table:

n	values of q
2	2
4	3, 4
6	6
8	8, 9
10	10
12	12

(You don't need a computer to create this table; it is easy to do these calculations by "casting out squares": Indeed, basic number theory (of which we will probably see more as this course progresses) says that for any two nonzero integers a and b, we have the equivalence

$$\left(a^2 b \text{ is a perfect square}\right) \Longleftrightarrow (b \text{ is a perfect square}).$$

[311] Hence, if we have a product of nonzero integers and we want to know whether this product is a square, we can "cancel" any squares from this product. Thus, for example, we can see that $\dfrac{1! \cdot 2! \cdot \cdots \cdot (2 \cdot 2)!}{3!}$ is not a perfect square by observing

[311] To wit, the "$\Longleftarrow$" direction of this equivalence is obvious, while the "$\Longrightarrow$" direction is easiest to prove using the uniqueness of prime decomposition.

that

$$\frac{1!\cdot 2!\cdot\cdots\cdot(2\cdot 2)!}{2!} = \frac{1!\cdot 2!\cdot 3!\cdot 4!}{3!} = \underbrace{1!}_{=1}\cdot\underbrace{2!}_{=1\cdot 2}\cdot\underbrace{4!}_{=1\cdot 2\cdot 3\cdot 4} = 1\cdot 1\cdot 2\cdot 1\cdot 2\cdot 3\cdot 4$$
$$= 2^2\cdot 3\cdot 4^2 \sim 3,$$

where in the last step we have "cancelled" the perfect squares 2^2 and 4^2. If you "cancel" strategically, you can do these computations fairly quickly, as any two successive factorials $k!$ and $(k+1)!$ mostly "cancel" each other.)

Based on the above table, it appears that $q = n$ always does the job (although sometimes there is a second value that also does it). So we need to prove that

$$\frac{1!\cdot 2!\cdot\cdots\cdot(2n)!}{n!} \text{ is a perfect square.} \tag{440}$$

To do this, we shall try to combine as many factors as we can into perfect squares (inspired by our above "cancelling" strategy). To wit, the numerator of our fraction is

$$\begin{aligned}
1!\cdot 2!\cdot\cdots\cdot(2n)! &= (1!\cdot 2!)\cdot(3!\cdot 4!)\cdot\cdots\cdot((2n-1)!\cdot(2n)!) \\
&= \prod_{k=1}^{n}\left((2k-1)!\cdot\underbrace{(2k)!}_{=(2k-1)!\cdot 2k}\right) = \prod_{k=1}^{n}\left(\underbrace{(2k-1)!\cdot(2k-1)!}_{=((2k-1)!)^2}\cdot 2k\right) \\
&= \prod_{k=1}^{n}\left(((2k-1)!)^2\cdot 2k\right) = \left(\prod_{k=1}^{n}((2k-1)!)^2\right)\cdot\underbrace{\prod_{k=1}^{n}(2k)}_{=2^n\prod_{k=1}^{n}k} \\
&= \left(\prod_{k=1}^{n}((2k-1)!)^2\right)\cdot 2^n\underbrace{\prod_{k=1}^{n}k}_{=n!} = \left(\prod_{k=1}^{n}((2k-1)!)^2\right)\cdot 2^n n!.
\end{aligned}$$

Dividing this by $n!$, we find

$$\frac{1!\cdot 2!\cdot\cdots\cdot(2n)!}{n!} = \left(\prod_{k=1}^{n}((2k-1)!)^2\right)\cdot 2^n = \left(\left(\prod_{k=1}^{n}(2k-1)!\right)\cdot 2^{n/2}\right)^2.$$

The right hand side is a perfect square (since n is even, so that $2^{n/2}\in\mathbb{Z}$). Thus, (440) is proved, and the problem is solved.[312]

[*Remark:* This exercise is a puzzle I found on reddit.

The claim (440) can also be proved by a straightforward induction on n.

[312]See [Grinbe15, Exercise 3.5 **(c)**] for this argument in more detail.

If neither n nor $n+1$ is a perfect square, then n is the **only** $q \in \{1, 2, \ldots, 2n\}$ for which $\dfrac{1! \cdot 2! \cdot \cdots \cdot (2n)!}{q!}$ is a perfect square. I don't know of a nice and simple proof of this, but a somewhat laborious proof (using a result of Erdös) can be found at `https://math.stackexchange.com/a/3325448/` .] □

A.1.2. Discussion of Exercise 1.1.2

Discussion of Exercise 1.1.2. This exercise has appeared in various places (e.g., it was Exercise 8 on the 5th Virginia Tech Regional Mathematics Contest). I gave a detailed solution in [Grinbe15, Proposition 2.64], so here I will mostly comment on how the sausage is made. Note that the sequence $(t_0, t_1, t_2, \ldots)$ appears in the OEIS as Sequence A005246.

Since the sequence $(t_0, t_1, t_2, \ldots)$ is defined recursively, it is reasonable to guess that its properties are best proved by induction. More precisely, strong induction is the best thing to try, since the recursive definition of t_n refers not just to t_{n-1} but also to the previous two values t_{n-2} and t_{n-3}.

A strong induction needs no induction base – although this often comes at the price that a few "edge" cases have to be considered separately within the induction step. We shall see this now.

(a) We proceed by strong induction on n.

Induction step: Let $m \geq 2$ be an integer. Assume (as induction hypothesis) that the claim of Exercise 1.1.2 **(a)** holds for all $n < m$ (meaning, of course, for all integers $n \geq 2$ satisfying $n < m$). We must prove that the claim of Exercise 1.1.2 **(a)** also holds for $n = m$. In other words, we must prove that $t_{m+2} = 4t_m - t_{m-2}$.

We want to reduce this to our induction hypothesis, which says things like $t_{m+1} = 4t_{m-1} - t_{m-3}$ and $t_m = 4t_{m-2} - t_{m-4}$ and so on. So we need to get rid of "ts with high indices" – that is, in our case, t_{m+2}. We achieve this by rewriting t_{m+2} as $\dfrac{1 + t_{m+1}t_m}{t_{m-1}}$ (a consequence of the recursive definition of our sequence). So, the equality we must prove (namely, $t_{m+2} = 4t_m - t_{m-2}$) rewrites as follows:

$$\begin{aligned}
&(t_{m+2} = 4t_m - t_{m-2}) \\
\iff &\left(\frac{1 + t_{m+1}t_m}{t_{m-1}} = 4t_m - t_{m-2}\right) \\
\iff &(1 + t_{m+1}t_m = (4t_m - t_{m-2})\, t_{m-1}) \\
\iff &(1 + t_{m+1}t_m = 4t_m t_{m-1} - t_{m-2}t_{m-1}) \\
\iff &(1 = 4t_m t_{m-1} - t_{m+1}t_m - t_{m-2}t_{m-1})
\end{aligned}$$

(where in the last step we brought the $t_{m+1}t_m$ term to the right hand side so it could join the rather similar terms there). The $4t_m t_{m-1} - t_{m+1}t_m$ on the right hand side rewrites as $t_m(4t_{m-1} - t_{m+1})$; but this looks familiar! Indeed, our induction hypothesis (applied to $n = m-1$) yields $t_{m+1} = 4t_{m-1} - t_{m-3}$, so that $4t_{m-1} -$

$t_{m+1} = t_{m-3}$. We can now continue our above chain of equivalent transformations of our goal:

$$\begin{aligned}
&(1 = 4t_m t_{m-1} - t_{m+1}t_m - t_{m-2}t_{m-1}) \\
\iff &\left(1 = t_m \underbrace{(4t_{m-1} - t_{m+1})}_{=t_{m-3}} - t_{m-2}t_{m-1}\right) \\
\iff &(1 = t_m t_{m-3} - t_{m-2}t_{m-1}).
\end{aligned}$$

But the equality $1 = t_m t_{m-3} - t_{m-2}t_{m-1}$ follows from the very definition of t_m, which (as we recall) says $t_m = \dfrac{1 + t_{m-1}t_{m-2}}{t_{m-3}}$. So we have reduced the equality in question (that is, $t_{m+2} = 4t_m - t_{m-2}$) to an equality we know to be true. We now need to walk this argument backwards and we have our induction step.

Did you spot the gap? You can see that there is a gap, since we have never used the base values ($t_0 = t_1 = t_2 = 1$). But it is not hard to see that the problem would not remain true if the base values were replaced by arbitrary numbers. Something has to give.

The problem is that we applied the induction hypothesis to $n = m - 1$. But the induction hypothesis said that Exercise 1.1.2 **(a)** holds for all integers $n \geq 2$ satisfying $n < m$. Thus, in order to apply it to $n = m - 1$, we need to know that $m - 1$ is an integer satisfying $m - 1 \geq 2$ and $m - 1 < m$. Of course, $m - 1$ is an integer and satisfies $m - 1 < m$. But $m - 1 \geq 2$ is not automatically satisfied; it holds for $m \geq 3$, but not for $m = 2$. Thus, we need to treat the $m = 2$ case separately. So much for a strong induction not needing an induction base! (Of course, technically it is true: We are not doing an induction base; we are just special-casing a special case in the induction step. But the difference is just organizational.)

So we need to verify the claim of Exercise 1.1.2 **(a)** for $m = 2$. This means verifying that $t_4 = 4t_2 - t_0$. This is not the most challenging part of the problem, so I stop here.

One last thing about **(a)**: If you look again at the above induction, you might notice that we only ever applied the induction hypothesis to $n = m - 1$. This means that we did not need the "strength" of strong induction; we can just as well restate the argument as a (standard) induction on n. The case $m = 2$ then becomes an actual (de-jure, not just de-facto) induction base. I think the resulting proof is a little bit easier to write up.

(b) We will separately prove the two statements that

$$t_n \text{ is positive for each integer } n \geq 0 \tag{441}$$

and that

$$t_n \text{ is an integer for each integer } n \geq 0. \tag{442}$$

To be pedantic, (441) does not even have to be proved, since it is implicit in the problem statement ("a sequence $(t_0, t_1, t_2, \ldots)$ of **positive** rational numbers"). (The

reason why I put the "positive" in the statement is to pre-empt any question about "what happens if the expression $\dfrac{1+t_{n-1}t_{n-2}}{t_{n-3}}$ has a zero in the denominator?". If we know that all entries of the sequence are positive, then of course there cannot be any zero denominators.) Just for the sake of completeness: The reason that t_n is positive for each integer $n \geq 0$ is that the sequence $(t_0, t_1, t_2, \ldots)$ was defined recursively with three positive base values (t_0, t_1, t_2 are all positive) and a recursive rule ($t_n = \dfrac{1+t_{n-1}t_{n-2}}{t_{n-3}}$) that produces a positive t_n when given positive $t_{n-1}, t_{n-2}, t_{n-3}$. So to speak, the sequence $(t_0, t_1, t_2, \ldots)$ never has a chance to escape the positive numbers. To make this argument formal (which you don't need to do on any contests, given how trivial it is), you can argue by strong induction, similarly to how we are soon going to prove (442) (but using the recursive rule $t_n = \dfrac{1+t_{n-1}t_{n-2}}{t_{n-3}}$ instead of part **(a)**).

Let us now prove (442). We proceed by strong induction on n.

Induction step: Fix an integer $m \geq 0$, and we assume (as induction hypothesis) that (442) holds for all $n < m$. We must now prove that (442) holds for $n = m$. In other words, we must prove that t_m is an integer.

First assume that $m \geq 4$. Then, $m - 2 \geq 2$; thus, we can apply Exercise 1.1.2 **(a)** to $n = m - 2$, and thus obtain $t_m = 4t_{m-2} - t_{m-4}$. The induction hypothesis then yields that t_{m-2} and t_{m-4} are integers (since $m-2$ and $m-4$ are integers ≥ 0 and satisfy $m - 2 < m$ and $m - 4 < m$), and so t_m is an integer as well (since $t_m = 4t_{m-2} - t_{m-4}$). So we are done with the induction step in the case when $m \geq 4$. It remains to cover the remaining cases. These are the cases when $m < 4$, that is, $m \in \{0, 1, 2, 3\}$. In other words, we need to show that t_0, t_1, t_2, t_3 are integers. This is straightforward (and has been already done in the problem statement). Thus the induction step is complete. Hence, (442) is proved.

(Note that this was a real strong induction: We applied the induction hypothesis to two different values of n, none of which was $m - 1$.)

Having proved (441) and (442), we are done with Exercise 1.1.2 **(b)**. □

A.1.3. Discussion of Exercise 1.1.3

Discussion of Exercise 1.1.3. This equality is known as *Hermite's identity for floor functions*. The "trick" is simply to express every addend of the sum $\sum_{k=0}^{n-1} \left\lfloor x + \dfrac{k}{n} \right\rfloor$ explicitly in terms of as few parameters as possible. Here is the solution in full detail:

Solution to Exercise 1.1.3. Let $g = \lfloor nx \rfloor$. Then, (1) (applied to nx instead of x) yields $\lfloor nx \rfloor \leq nx < \lfloor nx \rfloor + 1$. In view of $g = \lfloor nx \rfloor$, this rewrites as $g \leq nx < g + 1$.

Let q and r be the quotient and the remainder obtained when g is divided by n. Thus, $g = qn + r$ and $q \in \mathbb{Z}$ and $r \in \{0, 1, \ldots, n-1\}$.

Now, we shall prove explicit formulas for the values $\left\lfloor x + \frac{k}{n} \right\rfloor$ for $k \in \{0, 1, \ldots, n-1\}$. Namely:

- For any $k \in \{0, 1, \ldots, n-r-1\}$, we have
$$\left\lfloor x + \frac{k}{n} \right\rfloor = q. \tag{443}$$
- For any $k \in \{n-r, n-r+1, \ldots, n-1\}$, we have
$$\left\lfloor x + \frac{k}{n} \right\rfloor = q+1. \tag{444}$$

[*Proof of (443):* Let $k \in \{0, 1, \ldots, n-r-1\}$. Thus, $k \geq 0$ and $k \leq n-r-1$. From $k \leq n-r-1$, we obtain $k+r+1 \leq n$.

We shall prove that $q \leq x + \frac{k}{n} < q+1$ (since this will quickly entail $\left\lfloor x + \frac{k}{n} \right\rfloor = q$). In order to do so, we shall show that $qn \leq nx + k < (q+1)\,n$.

Indeed, $r \geq 0$, so that $qn \leq qn + r = g \leq nx \leq nx + k$ (since $k \geq 0$). Also,
$$\underbrace{nx}_{<g+1} + k < \underbrace{g}_{=qn+r} + 1 + k = qn + r + 1 + k = qn + \underbrace{k+r+1}_{\leq n} \leq qn + n = (q+1)\,n.$$
Combining $qn \leq nx + k$ with this, we obtain the chain of inequalities $qn \leq nx + k < (q+1)\,n$. Dividing this chain by n, we find $q \leq \frac{nx+k}{n} < q+1$. Hence, the integer q is $\leq \frac{nx+k}{n}$, but the next integer $q+1$ no longer is. Thus, q is the largest integer $\leq \frac{nx+k}{n}$. In other words, $q = \left\lfloor \frac{nx+k}{n} \right\rfloor$ (by the definition of $\left\lfloor \frac{nx+k}{n} \right\rfloor$). In other words, $q = \left\lfloor x + \frac{k}{n} \right\rfloor$ (since $\frac{nx+k}{n} = x + \frac{k}{n}$). This proves (443).]

[*Proof of (444):* Let $k \in \{n-r, n-r+1, \ldots, n-1\}$. Thus, $k \geq n-r$ and $k \leq n-1$. From $k \geq n-r$, we obtain $k+r \geq n$, so that $n \leq k+r$.

We shall prove that $q+1 \leq x + \frac{k}{n} < q+2$ (since this will quickly entail $\left\lfloor x + \frac{k}{n} \right\rfloor = q+1$). In order to do so, we shall show that $(q+1)\,n \leq nx + k < (q+2)\,n$.

Indeed, combining
$$(q+1)\,n = qn + \underbrace{n}_{\leq k+r} \leq qn + k + r = \underbrace{qn + r}_{=g \leq nx} + k \leq nx + k$$
with
$$\underbrace{nx}_{<g+1} + \underbrace{k}_{\leq n-1} < g + 1 + n - 1 = \underbrace{g}_{=qn+r} + n = qn + \underbrace{r}_{\leq n-1 < n} + n < qn + n + n = (q+2)\,n,$$

we obtain the chain of inequalities $(q+1)\,n \leq nx+k < (q+2)\,n$. Dividing this chain by n, we find $q+1 \leq \dfrac{nx+k}{n} < q+2$. Hence, the integer $q+1$ is $\leq \dfrac{nx+k}{n}$, but the next integer $q+2$ no longer is. Thus, $q+1$ is the largest integer $\leq \dfrac{nx+k}{n}$. In other words, $q+1 = \left\lfloor \dfrac{nx+k}{n} \right\rfloor$ (by the definition of $\left\lfloor \dfrac{nx+k}{n} \right\rfloor$). In other words, $q+1 = \left\lfloor x + \dfrac{k}{n} \right\rfloor$ (since $\dfrac{nx+k}{n} = x + \dfrac{k}{n}$). This proves (444).]

Now, $r \in \{0,1,\ldots,n-1\}$, so that $n-r \in \{1,2,\ldots,n\}$. Thus, we can split the sum $\sum\limits_{k=0}^{n-1} \left\lfloor x + \dfrac{k}{n} \right\rfloor$ at $k = n-r$. We thus obtain[313]

$$\begin{aligned}
\sum_{k=0}^{n-1} \left\lfloor x + \frac{k}{n} \right\rfloor &= \sum_{k=0}^{n-r-1} \underbrace{\left\lfloor x + \frac{k}{n} \right\rfloor}_{\substack{=q \\ \text{(by (443))}}} + \sum_{k=n-r}^{n-1} \underbrace{\left\lfloor x + \frac{k}{n} \right\rfloor}_{\substack{=q+1 \\ \text{(by (444))}}} \\
&= \underbrace{\sum_{k=0}^{n-r-1} q}_{\substack{=(n-r)q \\ \text{(since this sum has } n-r \text{ addends)}}} + \underbrace{\sum_{k=n-r}^{n-1} (q+1)}_{\substack{=r(q+1) \\ \text{(since this sum has } r \text{ addends)}}} \\
&= (n-r)\,q + r\,(q+1) = qn + r = g = \lfloor nx \rfloor .
\end{aligned}$$

This solves the exercise. □

With some experience, the above solution will appear straightforward. The decisive step was to introduce q and r; but this was not unmotivated. Indeed, the first addend of the sum $\sum\limits_{k=0}^{n-1} \left\lfloor x + \dfrac{k}{n} \right\rfloor$ is $\left\lfloor x + \dfrac{0}{n} \right\rfloor = \lfloor x \rfloor = q$, and all other addends are either q or $q+1$. The question that remains is which of them are q and which are $q+1$; but it soon becomes clear that the remainder of g divided by n decides this. Thus the above solution. □

A.1.4. Discussion of Exercise 1.1.4

Discussion of Exercise 1.1.4. The following solution is probably the easiest one to find, but not the nicest or the most generalizable one. More on this after the solution.

Solution to Exercise 1.1.4. Expanding $(a+b+c)^{n^2+n+1}$ will yield a sum of monomials of the form $a^i b^j c^k$ with i,j,k being nonnegative integers satisfying $i+j+k =$

[313] Keep in mind that the sum $\sum\limits_{k=n-r}^{n-1} \left\lfloor x + \dfrac{k}{n} \right\rfloor$ will be empty if $r = 0$. An empty sum equals 0 by definition.

$n^2 + n + 1$. (Indeed, the multinomial theorem yields the exact result, but we don't need it; all we need is that it is some sum of such monomials.) Thus, in order to prove that abc divides $(a+b+c)^{n^2+n+1}$, it will suffice to show that abc divides each of these monomials $a^i b^j c^k$. So let us do this.

Let us fix three nonnegative integers i, j, k satisfying $i + j + k = n^2 + n + 1$. We must prove that $abc \mid a^i b^j c^k$.

If all of i, j, k are nonzero (thus ≥ 1), then this is true for obvious reasons. Thus, we WLOG assume that i, j, k are not all nonzero. So at least one of i, j, k is zero.

Now $abc \mid a^i b^j c^k$ is no longer obvious, but we can try to use our assumptions $a \mid b^n$ and $b \mid c^n$ and $c \mid a^n$ to our advantage. These assumptions allow us to "trade" a b^n for an a, or a c^n for a b, or an a^n for a c in the term $a^i b^j c^k$. We hope that by a few such "trades", we can transform $a^i b^j c^k$ into a monomial where each of a, b, c appears at least once (whence the monomial is divisible by abc).

Which "trades" we need to make depends on the values of i, j, k. We distinguish between two cases:

Case 1: Exactly one of i, j, k is zero.

Case 2: Exactly two of i, j, k are zero.

(These are the only possibilities, because we already know that at least one of i, j, k is zero, whereas the condition $i + j + k = n^2 + n + 1$ rules out the possibility that all of i, j, k are zero).

Let us first consider Case 1. In this case, exactly one of i, j, k is zero.

Notice the cyclic symmetry in the problem: If we cyclically permute a, b, c (that is, replace a, b, c by b, c, a, respectively) and simultaneously cyclically permute i, j, k, then neither the assumptions nor the claim of the problem change. Thus, we can cyclically permute the numbers i, j, k at will (as long as we remember to permute a, b, c along with them). By doing so, we can always guarantee that $i = 0$ (because we know that at least one of the three numbers i, j, k is zero, and we can bring this number to the front by a cyclic permutation). Thus, WLOG assume that $i = 0$. Hence, j and k are nonzero (since exactly one of i, j, k is zero). In other words, $j \geq 1$ and $k \geq 1$. Our monomial $a^i b^j c^k$ thus has no a, but at least one b and at least one c in it.

We now want to "trade" a b^n for an a, or, if this is impossible (because we don't have a b^n), to "trade" n copies of c^n for b's and then trade the resulting b^n for an a. We need to be careful not to "sell" all our b's in the trade, so we should "trade" b^n only if we have at least one b left after that – i.e., if $j \geq n + 1$. Thus, we distinguish between the following two subcases:

Subcase 1.1: We have $j \geq n + 1$.

Subcase 1.2: We have $j < n + 1$.

In Subcase 1.1, we have $j \geq n + 1$. Hence, $b^{n+1} \mid b^j$. But $a \mid b^n$. Multiplying this divisibility by b, we find[314] $ab \mid b^n b = b^{n+1} \mid b^j$. Also, $c \mid c^k$ (since $k \geq 1$).

[314] We are using the following basic fact of number theory here: If x and y are two integers satisfying $x \mid y$, then $xb \mid yb$.

Multiplying these two divisibilities[315], we obtain $(ab)\,c \mid b^j c^k$. Hence, $abc = (ab)\,c \mid b^j c^k \mid a^i b^j c^k$. Thus, $abc \mid a^i b^j c^k$ is proved in Subcase 1.1.

Let us next consider Subcase 1.2. In this case, we have $j < n+1$. Hence, $j \leq n$ (since j and n are integers). Now, recall that $i+j+k = n^2+n+1$, so that $n^2+n+1 = \underbrace{i}_{=0} + \underbrace{j}_{\leq n} + k \leq n+k$, so that $k \geq n^2+1$. Therefore, $c^{n^2+1} \mid c^k$. But we can take the divisibility $b \mid c^n$ to the n-th power[316], and obtain $b^n \mid (c^n)^n = c^{n^2}$. Hence, $a \mid b^n \mid c^{n^2}$. Multiplying this divisibility by c, we obtain $ac \mid c^{n^2} c = c^{n^2+1} \mid c^k$. Also, $b \mid b^j$ (since $j \geq 1$). Multiplying the preceding two divisibilities, we obtain $(ac)\,b \mid c^k b^j$. Thus, $abc = (ac)\,b \mid c^k b^j = b^j c^k \mid a^i b^j c^k$. Thus, $abc \mid a^i b^j c^k$ is proved in Subcase 1.2.

We have now proved $abc \mid a^i b^j c^k$ in all of Case 1. It remains to deal with Case 2.

So let us consider Case 2. In this case, exactly two of i, j, k are zero. The same "cyclic symmetry" argument as in Case 1 lets us WLOG assume that i and j are zero. Thus, $i+j+k = 0+0+k = k$, so that $k = i+j+k = n^2+n+1$. Hence, $c^k = c^{n^2+n+1} = c^{n^2} c^n c$. As in Subcase 1.2, we can see that $a \mid c^{n^2}$. Also, we know that $b \mid c^n$. Multiplying these two divisibilities, we find $ab \mid c^{n^2} c^n$. Multiplying this further by c, we find $abc \mid c^{n^2} c^n c = c^k \mid a^i b^j c^k$. Thus, $abc \mid a^i b^j c^k$ is proved in Case 2.

This completes our proof of $abc \mid a^i b^j c^k$. As explained above, this solves the problem. □

Now, a few words about the ideas behind the problem. It is a slight generalization of an olympiad problem from Norway (Abelkonkurransen 1998-99 final problem 2b). As the above solution reveals, it is not much of a number theory problem; it is a combinatorial problem about "trading exponents" in a monomial, wrapped in number-theoretical packaging. At the core is the following fact: If you have a monomial $a^i b^j c^k$ with $i+j+k = n^2+n+1$ (treating a, b, c as formal variables rather than specific numbers), and you are allowed to trade a b^n for an a, a c^n for a b, or an a^n for a c, then you can always (choosing your trades strategically) wind up with a monomial that contains each of a, b, c at least once. (You can think of a, b, c as three different currencies that can only be exchanged for one another at a loss.)

This fact can be generalized to k (rather than 3) currencies; Fedor Petrov has proved this generalization in `https://mathoverflow.net/questions/198605/`. □

A.1.5. Discussion of Exercise 1.1.5

Discussion of Exercise 1.1.5. Let me outline two solutions:

First solution to Exercise 1.1.5 (sketched). This is the "gotcha" solution.

[315] This means we are using the following fact from number theory: If x, y, u, v are four integers satisfying $x \mid y$ and $u \mid v$, then $xu \mid yv$.

[316] This is, again, a consequence of a standard fact from number theory: If x and y are two integers such that $x \mid y$, then $x^n \mid y^n$.

Let M_1 be our mountain ridge. Imagine, in a parallel universe, a second mountain ridge M_2, which has the same shape and length as M_1 and initially has the same lemmings as M_1 at the same positions as on M_1, walking in the same directions as on M_1 and with the same speed as on M_1. However, they interact differently: When two M_2-lemmings meet, they just walk through one another like ghosts (as opposed to M_1-lemmings, which "bounce off" one another). For example, if M_1 and M_2 initially look like this:

$$\underline{\quad \overrightarrow{1} \quad \overleftarrow{2} \quad \overleftarrow{3} \qquad \overrightarrow{4} \qquad \overleftarrow{5} \quad}$$

then after the first "lemming collision" they will look as follows:

$$\begin{array}{ll} M_1 : & \underline{\quad \overleftarrow{1}\,\overrightarrow{2} \quad \overleftarrow{3} \qquad \overrightarrow{4} \qquad \overleftarrow{5} \quad} \ ; \\ M_2 : & \underline{\quad \overleftarrow{2}\,\overrightarrow{1} \quad \overleftarrow{3} \qquad \overrightarrow{4} \qquad \overleftarrow{5} \quad} \ . \end{array}$$

Now the key observation is the following:

> *Observation 1:* At any moment in time, the multiset of the positions of all M_1-lemmings on the M_1-ridge is the same as the multiset of the positions of all M_2-lemmings on the M_2-ridge.

In other words, if we don't distinguish between the specific lemmings, then we never see a difference between M_1 and M_2.

Why is Observation 1 true? Roughly speaking[317], it is true because it is true at the onset (by definition of M_2), it remains true as long as there are no collisions (since the only difference between M_1 and M_2 is the rule for collisions), and it remains true through each collision (since the collision rules on M_1 and on M_2 do not depend on the identities of the individual lemmings colliding, and their results differ only if one distinguishes between the lemmings[318]). Thus, Observation 1 never has any chance to become false.

[317] We shall not formalize this argument, as it would require a lot of busywork even to formally define what it means for a lemming to walk along the ridge, and to ensure that the positions of the lemmings are actually well-defined at any time. From a rigorous point of view, these things do need to be verified!

(If you are wondering why, ponder the following variant of the exercise. Assume that the ridge is a circle instead of a line segment, and that the lemmings double their speeds after every "collision". Assume further that you start out with two lemmings that have just bounced off one another. Say they collide after 1 minute. After this collision, their speeds double, so they collide again after another $\frac{1}{2}$ minute. Then their speeds double again, so they collide after another $\frac{1}{4}$ minute. And so on... Thus, the lemmings keep running around the ridge all the way up to the 2-minute mark (since $1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots = 2$). But what happens afterwards – say, after 3 minutes? This has no well-defined answer; our rules of collision simply do not define any behavior outside of the time interval $[0, 2)$.

This is rather similar to finite-time blow-up of solutions to some different equations.)

[318] To be fully honest, we need to first observe that the speeds of all lemmings remain constant and equal throughout the ordeal, both on M_1 and M_2 (because collisions never change the speeds).

But the M_2-lemmings are much easier to analyze than the M_1-lemmings. After all, collisions don't matter on M_2. Thus, each M_2-lemming keeps walking in the direction it has always been walking. As a consequence, each M_2-lemming eventually falls off the cliff (since the speeds are constant and nonzero, and the ridge is finite). In other words, there are eventually no lemmings on M_2 any more. According to Observation 1, this means that there are eventually no lemmings on M_1 any more. In other words, each M_1-lemming eventually falls off the cliff. This solves the exercise. □

Note that the above solution shows something better than the exercise asked for: Namely, it shows that if the ridge has length d and the speed of each lemming is v, then there will be no lemmings on the ridge after time d/v. Indeed, this is clearly what happens on M_2 (since each M_2-lemming just walks with speed v in a fixed direction), and therefore (by Observation 1) is also what happens on M_1.

The next proof does not (at least directly) show this stronger result, but is probably easier to come up with:

Second solution to Exercise 1.1.5 (sketched). Consider the initial state. If the leftmost lemming is walking left, then it will reach the left cliff undisturbed and fall off it. If the leftmost lemming is walking right, then it will either reach the right cliff undisturbed and fall off it, or it will collide with another lemming and bounce off, which will then send it moving left and eventually falling off the left cliff (undisturbed, since there won't be any other lemming for it to collide with; it will stay the leftmost lemming on the ridge until it falls). In either case, the leftmost lemming will eventually fall off the cliff. After this happens, another lemming will become the leftmost lemming, and thus (by the same argument) will also eventually fall off the cliff. Repeat this argument once for each lemming (strictly speaking, this means arguing by induction). This shows that eventually, all lemmings will have fallen off the cliff (although it does not offer a good bound on how soon). □

□

A.1.6. Discussion of Exercise 1.1.6

Discussion of Exercise 1.1.6. This is a particular case of [18s-hw2s, Exercise 7]. (More precisely, part **(a)** is the particular case of [18s-hw2s, Exercise 7] for $n = 40$ and $s = 10$, whereas part **(b)** is the particular case of [18s-hw2s, Exercise 7] for $n = 26$ and $s = 10$.)

Let us introduce notations that cover both parts **(a)** and **(b)**. Let n and s be two even positive integers. Assume that we have $n/2$ white socks and $n/2$ black socks hanging on a clothesline, in some order. We are looking to pick s consecutive socks

When two lemmings collide at equal speeds, the result is the same on M_1 and on M_2 as long as you don't distinguish between the lemmings: You still have two lemmings walking in the same directions at the same speeds. The only difference is that the two lemmings have traded places on M_1.

on the clothesline such that $s/2$ of them are black and the other $s/2$ white. We shall refer to such a pick as a *balanced pick*. (Let's not call it "color-balanced" for brevity.) Thus, part **(a)** claims that we can find a balanced pick if $n = 40$ and $s = 10$, whereas part **(b)** claims that we can find a balanced pick if $n = 26$ and $s = 10$.

We shall prove a more general claim ([18s-hw2s, Exercise 7]). Namely, we shall prove the following:

> *Claim 1:* Set $q = n//s$ and $r = n\%s$ (where we are using the notations $//$ and $\%$ introduced in Definition 3.3.1). Assume that there is **no** balanced pick. Then, $r \geq 2q$ and $s > 2q + r$.

Claim 1 will solve part **(a)** of our problem, since $r \geq 2q$ is violated for $n = 40$ and $s = 10$; it will also solve part **(b)**, since $s > 2q + r$ is violated for $n = 26$ and $s = 10$.

[*Proof of Claim 1:* We have $s/2 \in \mathbb{Z}$ (since s is even). Also, q and r are the quotient and the remainder of division of n by s (since $q = n//s$ and $r = n\%s$). Thus, $q \in \mathbb{Z}$, $r \in \{0, 1, \ldots, s-1\}$ and $n = qs + r$. From $r \in \{0, 1, \ldots, s-1\}$, we obtain $r < s$. Thus, $s > r$.

We must prove that $r \geq 2q$ and $s > 2q + r$. If $q = 0$, then this is obvious (because $r \geq 0$ and $s > r$). Hence, we WLOG assume that $q \neq 0$. Thus, $q \geq 1$. Therefore, $n = \underbrace{q}_{\geq 1} s + \underbrace{r}_{\geq 0} \geq s$, so that $n - s + 1 \geq 1$.

Number the socks by $1, 2, \ldots, n$ in the order in which they appear on the clothesline. For any $m \in \mathbb{N}$, let $[m]$ denote the m-element set $\{1, 2, \ldots, m\}$.

For each $i \in [n - s + 1]$, define the integer

$$b_i = (\text{the number of black socks among socks } i, i+1, \ldots, i+s-1) - s/2. \quad (445)$$

(This is indeed an integer, because $s/2 \in \mathbb{Z}$.) Then, each $i \in [n - s + 1]$ satisfies

$$\begin{aligned} b_i = \frac{1}{2} &(\text{the number of black socks among socks } i, i+1, \ldots, i+s-1) \\ &- \frac{1}{2} (\text{the number of white socks among socks } i, i+1, \ldots, i+s-1). \end{aligned} \quad (446)$$

(This comes from the fact that each sock is either black or white; see [18s-hw2s, solution to Exercise 7] for a detailed proof.)

The equality (446) shows that if we invert the colors of all socks (simultaneously), then all the numbers $b_1, b_2, \ldots, b_{n-s+1}$ change signs. Hence, we can WLOG assume that $b_1 \geq 0$ (since otherwise, we can invert the colors of all socks, and then b_1 will change sign). Assume this.

Note that each $i \in [n - s + 1]$ satisfies

$$(\text{the number of black socks among socks } i, i+1, \ldots, i+s-1) = b_i + s/2 \quad (447)$$

(by (445)).

For each $i \in [n-s+1]$, we have $b_i \neq 0$ (because if we had $b_i = 0$, then the s consecutive socks $i, i+1, \ldots, i+s-1$ would form a balanced pick; but this would contradict our assumption that there is no balanced pick). Thus, $b_1, b_2, \ldots, b_{n-s+1}$ are nonzero integers. Furthermore,

$$|b_{i+1} - b_i| \leq 1 \qquad \text{for all } i \in [(n-s+1)-1]$$

[319]. Hence, Proposition 2.2.4 (applied to $g = 1$ and $h = n-s+1$) shows that $b_i > 0$ for all $i \in [n-s+1]$. Since the b_i are integers, this shows that

$$b_i \geq 1 \text{ for all } i \in [n-s+1]. \tag{448}$$

Now, let g be the number of black socks among the r socks $qs+1, qs+2, \ldots, qs+r$. Thus, clearly, $0 \leq g \leq r$.

Recall that the total number of black socks on the clothesline is $n/2$. Thus,

$$\begin{aligned}
n/2 &= (\text{the total number of black socks}) \\
&= (\text{the number of black socks among socks } 1, 2, \ldots, n) \\
&= (\text{the number of black socks among socks } 1, 2, \ldots, qs+r) \\
&\qquad (\text{since } n = qs+r) \\
&= (\text{the number of black socks among socks } 1, 2, \ldots, s) \\
&\qquad + (\text{the number of black socks among socks } s+1, s+2, \ldots, 2s) \\
&\qquad + (\text{the number of black socks among socks } 2s+1, 2s+2, \ldots, 3s) \\
&\qquad + \cdots \\
&\qquad + (\text{the number of black socks among socks } (q-1)s+1, (q-1)s+2, \ldots, qs) \\
&\qquad + (\text{the number of black socks among socks } qs+1, qs+2, \ldots, qs+r) \\
&= \sum_{h=0}^{q-1} \underbrace{(\text{the number of black socks among socks } hs+1, hs+2, \ldots, (h+1)s)}_{\substack{= b_{hs+1} + s/2 \\ (\text{by } (447))}} \\
&\qquad + \underbrace{(\text{the number of black socks among socks } qs+1, qs+2, \ldots, qs+r)}_{= g} \\
&= \sum_{h=0}^{q-1} \left(\underbrace{b_{hs+1}}_{\substack{\geq 1 \\ (\text{by } (448))}} + s/2 \right) + g \\
&\geq \underbrace{\sum_{h=0}^{q-1} (1 + s/2)}_{\substack{= q(1+s/2) \\ = q + qs/2}} + g = q + qs/2 + g.
\end{aligned}$$

[319] *Proof.* Compare the definitions of b_i and b_{i+1}, and observe that the s socks $i, i+1, \ldots, i+s-1$ differ from the s socks $i+1, i+2, \ldots, i+s$ in at most one sock. (See [18s-hw2s, solution to Exercise 7] for the details.)

Hence,

$$q + qs/2 + g \leq \underbrace{n}_{=qs+r} /2 = (qs + r)/2 = qs/2 + r/2.$$

Subtracting $qs/2$ from both sides of this inequality, we find

$$q + g \leq r/2. \tag{449}$$

Hence, $r/2 \geq q + \underbrace{g}_{\geq 0} \geq q$, so that $r \geq 2q$.

It remains to prove that $s > 2q + r$. From $n = qs + r$, we obtain $n - r = qs$. Thus,

$$\begin{aligned}
&(\text{the number of black socks among socks } n-r+1, n-r+2, \ldots, n) \\
&= (\text{the number of black socks among socks } qs+1, qs+2, \ldots, n) \\
&= (\text{the number of black socks among socks } qs+1, qs+2, \ldots, qs+r) \\
&\qquad (\text{since } n = qs + r) \\
&= g.
\end{aligned}$$

But there is a sock $n - s + 1$ on our clothesline (since $n - s + 1 \geq 1$). Let p be the number of black socks among the $s - r$ socks $n - s + 1, n - s + 2, \ldots, n - r$. Thus, clearly, $p \leq s - r$.

From (447), we obtain

$$\begin{aligned}
&(\text{the number of black socks among socks } n-s+1, n-s+2, \ldots, n) \\
&= \underbrace{b_{n-s+1}}_{\substack{\geq 1 \\ (\text{by } (448))}} + s/2 \geq 1 + s/2.
\end{aligned}$$

Hence,

$$\begin{aligned}
&1 + s/2 \\
&\leq (\text{the number of black socks among socks } n-s+1, n-s+2, \ldots, n) \\
&= \underbrace{(\text{the number of black socks among socks } n-s+1, n-s+2, \ldots, n-r)}_{=p} \\
&\qquad + \underbrace{(\text{the number of black socks among socks } n-r+1, n-r+2, \ldots, n)}_{=g} \\
&= \underbrace{p}_{\leq s-r} + \underbrace{g}_{\substack{\leq r/2-q \\ (\text{by } (449))}} \leq (s-r) + (r/2 - q) = s - r/2 - q.
\end{aligned}$$

Subtracting $s/2$ from both sides of this inequality, we find $1 \leq s/2 - r/2 - q$, so that $s/2 - r/2 - q \geq 1 > 0$. Multiplying this inequality by 2, we obtain $s - r - 2q > 0$, so that $s > 2q + r$. This completes the proof of Claim 1.]

Thus, Exercise 1.1.6 is solved.

[*Remark:* I have heard that Claim 1 is an "if and only if": If we **don't** have $r \geq 2q$ and $s > 2q + r$, then there exists a way to place $n/2$ black and $n/2$ white socks on a clothesline such that there is no balanced pick. I don't currently remember a proof.] □

A.1.7. Discussion of Exercise 1.1.7

Discussion of Exercise 1.1.7. The solution is short:

$$\begin{aligned}&bc\left(b-c\right)\left(b+c\right)+ca\left(c-a\right)\left(c+a\right)+ab\left(a-b\right)\left(a+b\right)\\&=\left(a-b\right)\left(a-c\right)\left(b-c\right)\left(a+b+c\right).\end{aligned}$$

But how to find this? There are several options. The easiest approach is by identifying and factoring out divisors. The most elementary implementation of this approach is the following. We treat b and c as constants, and consider our polynomial as a polynomial in the single variable a. Recall that if a polynomial in a variable x has a root x_0, then this polynomial is divisible by the linear polynomial $x-x_0$. Now, the polynomial

$$P\left(a\right):=bc\left(b-c\right)\left(b+c\right)+ca\left(c-a\right)\left(c+a\right)+ab\left(a-b\right)\left(a+b\right)$$

has root b, since

$$\begin{aligned}P\left(b\right)&=bc\left(b-c\right)\left(b+c\right)+cb\left(c-b\right)\left(c+b\right)+bb\underbrace{\left(b-b\right)}_{=0}\left(b+b\right)\\&=bc\left(b-c\right)\left(b+c\right)+cb\left(c-b\right)\left(c+b\right)=bc\underbrace{\left(\left(b-c\right)+\left(c-b\right)\right)}_{=0}\left(b+c\right)=0.\end{aligned}$$

Thus, it is divisible by $a-b$. For the same reason, it is divisible by $a-c$. Occam's Razor thus suggests that it is divisible by the product $\left(a-b\right)\left(a-c\right)$ as well[320]. By polynomial long division, we can find the quotient $\dfrac{P\left(a\right)}{\left(a-b\right)\left(a-c\right)}$ to be

$$b^2+ab-c^2-ac=a\left(b-c\right)+\underbrace{\left(b^2-c^2\right)}_{=\left(b+c\right)\left(b-c\right)}=\left(a+b+c\right)\left(b-c\right).$$

And thus $P\left(a\right)=\left(a-b\right)\left(a-c\right)\left(a+b+c\right)\left(b-c\right)$.

There are other ways to find this as well. The probably nicest one is using determinants: Recall that the determinant of a matrix does not change if we subtract a

[320] Algebraists will not need Occam's Razor here: They know that the ring of polynomials in the three variables a,b,c over (say) $\mathbb{Q}$ is a unique factorization domain (this holds more generally for any polynomial ring over a field), and therefore a polynomial divisible by two coprime polynomials like $a-b$ and $a-c$ must necessarily be divisible by their product. But justifying this without using abstract algebra is not easy. (Our approach of treating b and c as constants, in particular, makes this argument rather slippery, because b and c might be equal and then the polynomials $a-b$ and $a-c$ are no longer coprime.) Fortunately, we don't need to justify this step in order to solve the problem – it is merely used to find a factorization of a specific polynomial, which, once found, can be proved by direct computation.

multiple of a column from another column. Now,

$$bc\underbrace{(b-c)(b+c)}_{=b^2-c^2}+ca\underbrace{(c-a)(c+a)}_{=c^2-a^2}+ab\underbrace{(a-b)(a+b)}_{=a^2-b^2}$$

$$= bc\left(b^2-c^2\right)+ca\left(c^2-a^2\right)+ab\left(a^2-b^2\right)$$

$$= b^3c-bc^3+c^3a-ca^3+a^3b-ab^3 = \det\begin{pmatrix} a^3 & a & 1 \\ b^3 & b & 1 \\ c^3 & c & 1 \end{pmatrix}$$

(by the definition of the determinant)

$$= \det\begin{pmatrix} a^3-c^2a & a & 1 \\ b^3-c^2b & b & 1 \\ c^3-c^2c & c & 1 \end{pmatrix} \qquad \left(\begin{array}{c}\text{here, we subtracted the } c^2\text{-multiple} \\ \text{of the second column} \\ \text{from the first column}\end{array}\right)$$

$$= \det\begin{pmatrix} a^3-c^2a & a-c & 1 \\ b^3-c^2b & b-c & 1 \\ c^3-c^2c & c-c & 1 \end{pmatrix} \qquad \left(\begin{array}{c}\text{here, we subtracted the } c\text{-multiple} \\ \text{of the third column} \\ \text{from the second column}\end{array}\right)$$

$$= \det\begin{pmatrix} a(a+c)(a-c) & a-c & 1 \\ b(b+c)(b-c) & b-c & 1 \\ 0 & 0 & 1 \end{pmatrix} \qquad \left(\begin{array}{c}\text{here we just rewrote the} \\ \text{entries of the matrix}\end{array}\right)$$

$$= \det\begin{pmatrix} a(a+c)(a-c)-b(b+c)(a-c) & a-c & 1 \\ b(b+c)(b-c)-b(b+c)(b-c) & b-c & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

$$\left(\begin{array}{c}\text{here, we subtracted the } b(b+c)\text{-multiple} \\ \text{of the second column from the first column}\end{array}\right)$$

$$= \det\begin{pmatrix} (a+b+c)(a-b)(a-c) & a-c & 1 \\ 0 & b-c & 1 \\ 0 & 0 & 1 \end{pmatrix} \qquad \left(\begin{array}{c}\text{here we just rewrote the} \\ \text{entries of the matrix}\end{array}\right)$$

$$= (a+b+c)(a-b)(a-c)\cdot(b-c)\cdot 1$$

$$\left(\begin{array}{c}\text{since the determinant of a triangular matrix equals} \\ \text{the product of its diagonal entries}\end{array}\right)$$

$$= (a-b)(a-c)(b-c)(a+b+c).$$

This way of solving the exercise is suggestive of some generalizations (see [Grinbe15, Exercise 6.16] for one of them).

Finally, let me notice that factoring a polynomial over $\mathbb{Q}$ (that is, into irreducible polynomials with rational coefficients) is an algorithmically solvable problem – in the sense that there are algorithms that can be used to solve it mechanically. The first algorithm was probably found by Kronecker [Edward05, Essay 1.4]. Most computer algebra systems have this algorithm implemented. Thus, the problem becomes trivial if one has access to a computer. However, the manual solution-

finding strategies shown above have other uses. □

A.1.8. Discussion of Exercise 1.1.8

Discussion of Exercise 1.1.8. **(a)** This is problem A1 in the 4th QEDMO (2007); the first of the following two solutions is copied from the model solutions (German), while the second (simpler) solution has been suggested by one of you (name withheld for FERPA).

Both solutions rely on the following two important properties of absolute values:

- If $x, y \in \mathbb{R}$, then
$$|xy| = |x| \cdot |y| . \tag{450}$$
- The *triangle inequality*: If $x, y \in \mathbb{R}$, then
$$|x| + |y| \geq |x + y| . \tag{451}$$

Both of these properties are easily verified by case distinction.

First solution to Exercise 1.1.8 ***(a)***. The inequality (451) (applied to $x = b - c$ and $y = c - a$) yields

$$\begin{aligned} |b-c| + |c-a| &\geq |(b-c) + (c-a)| = |(-1)(a-b)| \\ &\qquad \text{(since } (b-c)+(c-a) = b-a = (-1)(a-b)) \\ &= \underbrace{|-1|}_{=1} \cdot |a-b| \qquad \text{(by (450))} \\ &= |a-b| . \end{aligned} \tag{452}$$

The same reasoning (but with the variables cyclically permuted) yields

$$|c-a| + |a-b| \geq |b-c| \qquad \text{and} \tag{453}$$
$$|a-b| + |b-c| \geq |c-a| . \tag{454}$$

(Note that at least one of these three inequalities is an equality. The best way to see this is to imagine the three numbers a, b, c as points on the real axis; then, one of the three lies between the two others, and then the corresponding inequality is an equality. But we won't need this.)

Now, $b^2 - c^2 = (b+c)(b-c)$, so that

$$\begin{aligned} \left|b^2 - c^2\right| &= |(b+c)(b-c)| = \underbrace{|b+c|}_{\substack{=b+c \\ \text{(since } b+c \geq 0)}} \cdot |b-c| \qquad \text{(by (450))} \\ &= (b+c) \cdot |b-c| = b \cdot |b-c| + c \cdot |b-c| . \end{aligned}$$

Again, the same reasoning (but with the variables cyclically permuted) yields

$$\left|c^2 - a^2\right| = c \cdot |c - a| + a \cdot |c - a| \qquad \text{and}$$
$$\left|a^2 - b^2\right| = a \cdot |a - b| + b \cdot |a - b| .$$

Adding up these three equalities, we find

$$\begin{aligned}
&\left|b^2 - c^2\right| + \left|c^2 - a^2\right| + \left|a^2 - b^2\right| \\
&= (b \cdot |b - c| + c \cdot |b - c|) + (c \cdot |c - a| + a \cdot |c - a|) + (a \cdot |a - b| + b \cdot |a - b|) \\
&= a \cdot \underbrace{(|c - a| + |a - b|)}_{\substack{\geq |b-c| \\ \text{(by (453))}}} + b \cdot \underbrace{(|a - b| + |b - c|)}_{\substack{\geq |c-a| \\ \text{(by (454))}}} + c \cdot \underbrace{(|b - c| + |c - a|)}_{\substack{\geq |a-b| \\ \text{(by (452))}}} \\
&\geq a \cdot |b - c| + b \cdot |c - a| + c \cdot |a - b|
\end{aligned}$$

(here we have multiplied three inequalities with a, b, c, respectively; this was allowed since a, b, c are nonnegative). In view of

$$\begin{aligned}
&\left|\underbrace{ca - ab}_{=(-a)(b-c)}\right| + \left|\underbrace{ab - bc}_{=(-b)(c-a)}\right| + \left|\underbrace{bc - ca}_{=(-c)(a-b)}\right| \\
&= \underbrace{|(-a)\,(b - c)|}_{\substack{=|-a|\cdot|b-c| \\ \text{(by (450))}}} + \underbrace{|(-b)\,(c - a)|}_{\substack{=|-b|\cdot|c-a| \\ \text{(by (450))}}} + \underbrace{|(-c)\,(a - b)|}_{\substack{=|-c|\cdot|a-b| \\ \text{(by (450))}}} \\
&= \underbrace{|-a|}_{=a} \cdot |b - c| + \underbrace{|-b|}_{=b} \cdot |c - a| + \underbrace{|-c|}_{=c} \cdot |a - b| \\
&= a \cdot |b - c| + b \cdot |c - a| + c \cdot |a - b| ,
\end{aligned}$$

this rewrites as

$$\left|b^2 - c^2\right| + \left|c^2 - a^2\right| + \left|a^2 - b^2\right| \geq |ca - ab| + |ab - bc| + |bc - ca| .$$

This solves Exercise 1.1.8 **(a)**. □

Second solution to Exercise 1.1.8 ***(a)***. The inequality we are trying to prove is symmetric in a, b and c, in the sense that permuting a, b and c in any way does not change it (check this!). Thus, we can WLOG assume that $a \leq b \leq c$ (since we can always achieve this by permuting a, b and c so that the smallest of a, b and c becomes a, the second-smallest becomes b and the largest becomes c). Assume this. From $a \leq b \leq c$, we obtain $a^2 \leq b^2 \leq c^2$ (since a, b, c are nonnegative). Furthermore, multiplying both sides of the inequality $b \leq c$ by a, we obtain $ab \leq ac$ (since a is nonnegative). Similarly, from $a \leq b$, we obtain $ca \leq cb$. Now, $ab \leq ac = ca \leq cb = bc$, so that $ab \leq ca \leq bc$.

Now, we can get rid of the absolute value signs in the problem. Indeed, we have $b^2 \leq c^2$, so that $b^2 - c^2 \leq 0$ and thus $\left|b^2 - c^2\right| = -\left(b^2 - c^2\right) = c^2 - b^2$. Likewise, $\left|c^2 - a^2\right| = c^2 - a^2$ (since $a^2 \leq c^2$) and $\left|a^2 - b^2\right| = b^2 - a^2$ (since $a^2 \leq b^2$) and $|ca - ab| = ca - ab$ (since $ab \leq ca$) and $|ab - bc| = bc - ab$ (since $ab \leq bc$) and $|bc - ca| = bc - ca$ (since $ca \leq bc$). Hence,

$$\begin{aligned}
&\underbrace{\left|b^2 - c^2\right|}_{=c^2-b^2} + \underbrace{\left|c^2 - a^2\right|}_{=c^2-a^2} + \underbrace{\left|a^2 - b^2\right|}_{=b^2-a^2} \\
&= c^2 - b^2 + c^2 - a^2 + b^2 - a^2 = 2\left(c^2 - a^2\right) \\
&= 2(c+a)(c-a) \qquad (455)
\end{aligned}$$

and

$$\begin{aligned}
&\underbrace{|ca - ab|}_{=ca-ab} + \underbrace{|ab - bc|}_{=bc-ab} + \underbrace{|bc - ca|}_{=bc-ca} \\
&= ca - ab + bc - ab + bc - ca = 2(bc - ab) \\
&= 2b(c-a). \qquad (456)
\end{aligned}$$

But we need to prove that

$$|ca - ab| + |ab - bc| + |bc - ca| \leq \left|b^2 - c^2\right| + \left|c^2 - a^2\right| + \left|a^2 - b^2\right|.$$

In view of (456) and (455), this rewrites as

$$2b(c-a) \leq 2(c+a)(c-a). \qquad (457)$$

Thus, it remains to prove (457). But this is easy to check directly: We have $b \leq c \leq c + a$ (since $a \geq 0$), thus $2b \leq 2(c+a)$. We can multiply both sides of this inequality by $c - a$ (since $c - \underbrace{a}_{\leq c} \geq c - c = 0$), and thus obtain $2b(c-a) \leq 2(c+a)(c-a)$. This proves (457); thus, Exercise 1.1.8 **(a)** is solved. □

[*Remark:* The first solution to Exercise 1.1.8 **(a)** is longer, but it has a redeeming quality: It generalizes to n variables instead of 3. The corresponding result is that if $a_1, a_2, \ldots, a_n$ are n nonnegative reals, then

$$\sum_{i=1}^{n} |a_{i-1}a_i - a_i a_{i+1}| \leq \sum_{i=1}^{n} \left|a_i^2 - a_{i+1}^2\right|,$$

where we set $a_0 = a_n$ and $a_{n+1} = a_1$.]

(b) No. For a simple counterexample, set $a = b = 1$ and $c = -1$. In this case, the left hand side becomes 4 while the right hand side is 0 (since $a^2 = b^2 = c^2$). □

A.1.9. Discussion of Exercise 1.1.9

Discussion of Exercise 1.1.9. This is an exercise on modular arithmetic and the pigeonhole principle.[321] Indeed, let $b_i = a_1 + a_2 + \cdots + a_i$ for each $i \in \{0, 1, \ldots, n\}$. Thus, b_0 is an empty sum, so that $b_0 = 0$. Recall that the remainder of an integer u upon division by n is denoted by $u\%n$. Now, the $n+1$ remainders

$$b_0\%n, \quad b_1\%n, \quad b_2\%n, \quad \ldots, \quad b_n\%n$$

are $n+1$ elements of the n-element set $\{0, 1, \ldots, n-1\}$ (since they are remainders upon division by n), and thus at least two of them must be equal (by the Pigeonhole Principle)[322]. In other words, there exist two integers u and v with $0 \leq u < v \leq n$ and $b_u\%n = b_v\%n$. Consider these u and v. From $u < v$, we obtain $u + 1 \leq v$ (since u and v are integers). Also, $0 \leq u$, thus $1 \leq u + 1 \leq v \leq n$. Hence, both $u + 1$ and v belong to $\{1, 2, \ldots, n\}$.

The integers b_v and b_u leave the same remainder when divided by n (since $b_v\%n = b_u\%n$); thus, $b_v \equiv b_u \bmod n$ (by Proposition 3.3.4, applied to b_v and b_u instead of u and v). In other words, $n \mid b_v - b_u$. In view of

$$\begin{aligned} \underbrace{b_v}_{\substack{=a_1+a_2+\cdots+a_v \\ \text{(by the definition of } b_v)}} - \underbrace{b_u}_{\substack{=a_1+a_2+\cdots+a_u \\ \text{(by the definition of } b_u)}} &= (a_1 + a_2 + \cdots + a_v) - (a_1 + a_2 + \cdots + a_u) \\ &= a_{u+1} + a_{u+2} + \cdots + a_v \qquad (\text{since } u < v), \end{aligned}$$

this rewrites as $n \mid a_{u+1} + a_{u+2} + \cdots + a_v$. Hence, there exist some $p, q \in \{1, 2, \ldots, n\}$ with $p \leq q$ and $n \mid a_p + a_{p+1} + \cdots + a_q$ (namely, $p = u + 1$ and $q = v$). This solves Exercise 1.1.9. □

A.2. Homework set #1 discussion

The following are discussions of the problems on homework set #1 (Section 3.7).

A.2.1. Discussion of Exercise 3.7.1

Discussion of Exercise 3.7.1. We shall give two solutions, and then sketch two more. The first is a fairly straightforward strong induction argument whose induction step proceeds by removing the largest element w from S and observing that the remaining part $S \setminus \{w\}$ of S is a subset of $\{1, 2, \ldots, n-k\}$ (because the k-lacunarity condition forces any element of $S \setminus \{w\}$ to be $\leq w - k$ and therefore $\leq n - k$). There is a bit of busywork involved in ensuring that $n - k$ is still an element of $\mathbb{N}$ (so that the induction hypothesis can be applied). Here are the full details, for the skeptics:

[321] And it appears in many places as an example for the use of the pigeonhole principle (e.g., in [Brualdi09, §3.1, Application 3] or – in a weaker form – in [Engel98, Chapter 4, Example E3]).

[322] We are using the following form of the Pigeonhole Principle here: If $g_1, g_2, \ldots, g_m$ are m elements of a (fixed) n-element set, and if we have $m > n$, then at least two of the m elements $g_1, g_2, \ldots, g_m$ must be equal. (This is our Corollary 6.1.4 with slightly different notations.)

First solution to Exercise 3.7.1. Forget that we fixed n and S. We thus need to prove the following claim:

> *Claim 1:* Let $n \in \mathbb{N}$. Let S be a k-lacunar subset of $\{1, 2, \ldots, n\}$. Then, $|S| \leq \dfrac{n+k-1}{k}$.

[*Proof of Claim 1:* We apply strong induction on n:

Induction step: Let $m \in \mathbb{N}$. Assume (as the induction hypothesis) that Claim 1 holds for all $n < m$. We must prove that Claim 1 holds for $n = m$.

Note that $k \geq 1$ (since k is a positive integer), so that $m + \underbrace{k}_{\geq 1} - 1 \geq m + 1 - 1 = m \geq 0$. Hence, $\dfrac{m+k-1}{k} \geq 0$, so that $0 \leq \dfrac{m+k-1}{k}$. Also, $m - k < m$ (since $k > 0$).

We have assumed (as the induction hypothesis) that Claim 1 holds for all $n < m$. In other words, if $n \in \mathbb{N}$ satisfies $n < m$, and if S is a k-lacunar subset of $\{1, 2, \ldots, n\}$, then

$$|S| \leq \frac{n+k-1}{k}. \tag{458}$$

Now, let us prove that Claim 1 holds for $n = m$. Let S be a k-lacunar subset of $\{1, 2, \ldots, m\}$. We shall show that

$$|S| \leq \frac{m+k-1}{k}. \tag{459}$$

If $|S| = 0$, then this is definitely true (since we have $0 \leq \dfrac{m+k-1}{k}$). Thus, we WLOG assume that $|S| \neq 0$. Hence, the set S is nonempty. Also, the set S is finite (since it is a subset of the finite set $\{1, 2, \ldots, m\}$). Therefore, Proposition 2.1.2 shows that the set S has a maximum. Let w be this maximum. Thus, $w \in S \subseteq \{1, 2, \ldots, m\}$. Hence, $1 \leq w \leq m$, so that $m \geq 1$. Now, $\underbrace{m}_{\geq 1} + k - 1 \geq 1 + k - 1 = k$, so that $\dfrac{m+k-1}{k} \geq 1$. Therefore, if $|S| = 1$, then $|S| = 1 \leq \dfrac{m+k-1}{k}$ (since $\dfrac{m+k-1}{k} \geq 1$), and thus (459) is proved in this case. Hence, we WLOG assume that $|S| \neq 1$. But $w \in S$, so that $|S \setminus \{w\}| = |S| - 1 \neq 0$ (since $|S| \neq 1$). In other words, the set $S \setminus \{w\}$ is nonempty.

Now, each $g \in S \setminus \{w\}$ satisfies $g \leq w - k$ [323] and therefore $g \leq \underbrace{w}_{\leq m} - k \leq m - k$ and consequently $g \in \{1, 2, \ldots, m-k\}$ (since $g \in S \setminus \{w\} \subseteq S \subseteq \{1, 2, \ldots, m\}$

[323] *Proof.* Let $g \in S \setminus \{w\}$. Thus, $g \neq w$ and $g \in S \setminus \{w\} \subseteq S$. But each $s \in S$ satisfies $s \leq w$ (since w is the maximum of S). Applying this to $s = g$, we obtain $g \leq w$ (since $g \in S$). Hence, $w - g \geq 0$, so that $|w - g| = w - g$.

Recall that the set S is k-lacunar. In other words, every two distinct elements $u, v \in S$ satisfy $|u - v| \geq k$ (by the definition of "k-lacunar"). Applying this to $u = w$ and $v = g$, we obtain $|w - g| \geq k$ (since w and g are distinct (because $g \neq w$)). In other words, $w - g \geq k$ (since $|w - g| = w - g$). Thus, $w \geq g + k$ and therefore $g \leq w - k$. Qed.

entails $g \geq 1$) [324]. In other words, $S \setminus \{w\} \subseteq \{1, 2, \ldots, m-k\}$. But recall that the set $S \setminus \{w\}$ is nonempty; thus, there exists some $x \in S \setminus \{w\}$. Consider this x. We have $x \in S \setminus \{w\} \subseteq \{1, 2, \ldots, m-k\}$, so that $1 \leq x \leq m-k$, thus $m-k \geq 1 \geq 0$ and therefore $m-k \in \mathbb{N}$.

Furthermore, it is easy to see that any subset of a k-lacunar subset is itself k-lacunar; hence, the set $S \setminus \{w\}$ is k-lacunar (because it is a subset of the k-lacunar set S). Thus, $S \setminus \{w\}$ is a k-lacunar subset of $\{1, 2, \ldots, m-k\}$ (since $S \setminus \{w\} \subseteq \{1, 2, \ldots, m-k\}$). Hence, (458) (applied to $m-k$ and $S \setminus \{w\}$ instead of n and S) yields

$$|S \setminus \{w\}| \leq \frac{(m-k)+k-1}{k} = \frac{m-1}{k}.$$

Now, recall that $|S \setminus \{w\}| = |S| - 1$, so that

$$|S| = \underbrace{|S \setminus \{w\}|}_{\leq \frac{m-1}{k}} + 1 \leq \frac{m-1}{k} + 1 = \frac{m+k-1}{k}.$$

Thus, (459) is proved.

Forget that we fixed S. We thus have showed that if S is a k-lacunar subset of $\{1, 2, \ldots, m\}$, then $|S| \leq \dfrac{m+k-1}{k}$. In other words, Claim 1 holds for $n = m$. This completes the induction step; thus, Claim 1 is proved.] $\square$

The second solution of Exercise 3.7.1 is more direct, and relies on the following basic fact of enumerative combinatorics (known as the *sum rule*): If $S_1, S_2, \ldots, S_k$ are k disjoint finite sets, then the set $S_1 \cup S_2 \cup \cdots \cup S_k$ is finite and satisfies

$$|S_1 \cup S_2 \cup \cdots \cup S_k| = |S_1| + |S_2| + \cdots + |S_k|. \tag{460}$$

(Note that "disjoint" means "pairwise disjoint"; i.e., the k sets $S_1, S_2, \ldots, S_k$ are said to be disjoint if and only if every two distinct elements i and j of $\{1, 2, \ldots, k\}$ satisfy $S_i \cap S_j = \varnothing$.) Note that this is a generalization of Theorem 2.3.6 to multiple (not just 2) disjoint sets.

Second solution to Exercise 3.7.1. For each $i \in \{1, 2, \ldots, k\}$, we define a set

$$S_i = \{s + i \mid s \in S\}.$$

Roughly speaking, S_i is just the set S shifted to the right by a distance of i (on the real axis). Hence, S_i has the same size as S; that is, we have

$$|S_i| = |S| \qquad \text{for each } i \in \{1, 2, \ldots, k\}. \tag{461}$$

[324] We are here following the convention that $\{1, 2, \ldots, m-k\}$ is the empty set when $m-k \leq 0$. (Although we shall soon see that we don't have $m-k \leq 0$.)

Furthermore, the sets $S_1, S_2, \ldots, S_k$ are subsets of the set $\{2, 3, \ldots, n+k\}$ [325]. Hence, their union $S_1 \cup S_2 \cup \cdots \cup S_k$ is a subset of $\{2, 3, \ldots, n+k\}$ as well. Therefore,

$$\begin{aligned} |S_1 \cup S_2 \cup \cdots \cup S_k| &\leq |\{2, 3, \ldots, n+k\}| \\ &\qquad \text{(because if } A \text{ is a subset of a finite set } B\text{, then } |A| \leq |B|) \\ &= n + k - 1. \end{aligned}$$

On the other hand, since S is k-lacunar, it is easy to see that the k sets $S_1, S_2, \ldots, S_k$ are disjoint[326]. Hence, (460) yields

$$|S_1 \cup S_2 \cup \cdots \cup S_k| = |S_1| + |S_2| + \cdots + |S_k| = \sum_{i=1}^{k} \underbrace{|S_i|}_{\substack{=|S| \\ \text{(by (461))}}} = \sum_{i=1}^{k} |S| = k \cdot |S|.$$

Therefore,

$$k \cdot |S| = |S_1 \cup S_2 \cup \cdots \cup S_k| \leq n + k - 1.$$

Dividing both sides of this inequality by k, we find $|S| \leq \dfrac{n+k-1}{k}$. This solves Exercise 3.7.1 again. $\square$

[325] *Proof.* We must show that $S_i \subseteq \{2, 3, \ldots, n+k\}$ for each $i \in \{1, 2, \ldots, k\}$. Let us do this.

So let $i \in \{1, 2, \ldots, k\}$. Thus, $1 \leq i \leq k$.

Let $z \in S_i$. Thus, $z \in S_i = \{s + i \mid s \in S\}$, so that $z = s + i$ for some $s \in S$. Consider this s. We have $s \in S \subseteq \{1, 2, \ldots, n\}$, so that $1 \leq s \leq n$. Now, $z = \underbrace{s}_{\geq 1} + \underbrace{i}_{\geq 1} \geq 1 + 1 = 2$ and $z = \underbrace{s}_{\leq n} + \underbrace{i}_{\leq k} \leq n + k$. Combining these two inequalities, we obtain $2 \leq z \leq n + k$ and thus $z \in \{2, 3, \ldots, n+k\}$.

Forget that we fixed z. We thus have shown that $z \in \{2, 3, \ldots, n+k\}$ for each $z \in S_i$. In other words, $S_i \subseteq \{2, 3, \ldots, n+k\}$. This completes our proof.

[326] *Proof.* We must prove that $S_i \cap S_j = \varnothing$ for every two distinct elements i and j of $\{1, 2, \ldots, k\}$. Thus, let i and j be two distinct elements of $\{1, 2, \ldots, k\}$. We must prove that $S_i \cap S_j = \varnothing$.

We WLOG assume that $i \leq j$ (since otherwise, we can swap i with j). Hence, $j - i \geq 0$. Thus, $|j - i| = j - i$.

Let $t \in S_i \cap S_j$. Thus, $t \in S_i \cap S_j \subseteq S_i = \{s + i \mid s \in S\}$. In other words, there exists some $s \in S$ such that $t = s + i$. Consider this s. From $t = s + i$, we obtain $t - i = s \in S$. Similarly, $t - j \in S$. Moreover, $i \neq j$ (since i and j are distinct) and thus $t - i \neq t - j$. Hence, the elements $t - i$ and $t - j$ of S are distinct.

Recall that the set S is k-lacunar. In other words, every two distinct elements $u, v \in S$ satisfy $|u - v| \geq k$ (by the definition of "k-lacunar"). Applying this to $u = t - i$ and $v = t - j$, we obtain $|(t - i) - (t - j)| \geq k$. In other words, $|j - i| \geq k$ (since $(t - i) - (t - j) = j - i$). In other words, $j - i \geq k$ (since $|j - i| = j - i$). Therefore, $j \geq i + k > k$ (since $i > 0$). On the other hand, from $j \in \{1, 2, \ldots, k\}$, we obtain $j \leq k$. This contradicts $j > k$.

Forget that we fixed t. We thus have obtained a contradiction for each $t \in S_i \cap S_j$. Hence, there exists no $t \in S_i \cap S_j$. In other words, $S_i \cap S_j$ is an empty set. That is, $S_i \cap S_j = \varnothing$. This completes our proof.

Third solution to Exercise 3.7.1 (sketched). This was suggested by one of the students. We let $q = (n-1)//k$, and we subdivide the set $\{1,2,\ldots,n\}$ into $q+1$ integer intervals

$$\begin{aligned}
I_0 &= \{1,2,\ldots,k\}, \\
I_1 &= \{k+1,k+2,\ldots,2k\}, \\
I_2 &= \{2k+1,2k+2,\ldots,3k\}, \\
&\ldots, \\
I_{q-1} &= \{(q-1)k+1,(q-1)k+2,\ldots,qk\}, \\
I_q &= \{qk+1,qk+2,\ldots,n\}.
\end{aligned}$$

Each of these $q+1$ intervals $I_0, I_1, \ldots, I_q$ has length $\leq k$ (indeed, all but I_q have length exactly k, whereas I_q has length $n-qk \leq k$). Therefore, each $i \in \{0,1,\ldots,q\}$ satisfies

$$|S \cap I_i| \leq 1 \tag{462}$$

(because otherwise, the set $S \cap I_i$ would contain at least two distinct elements u and v, but then these two elements would satisfy $u, v \in S$ and $|u-v| < k$, which would contradict the assumption that S is k-lacunar). But the sum rule (460) yields

$$\begin{aligned}
&\left|(S\cap I_0)\cup(S\cap I_1)\cup\cdots\cup\left(S\cap I_q\right)\right| \\
&= |S\cap I_0| + |S\cap I_1| + \cdots + \left|S\cap I_q\right| = \sum_{i=0}^{q} \underbrace{|S\cap I_i|}_{\substack{\leq 1 \\ \text{(by (462))}}} \leq \sum_{i=0}^{q} 1 \\
&= q+1 \leq \frac{n-1}{k}+1 \qquad \left(\text{since } q = (n-1)//k = \left\lfloor \frac{n-1}{k} \right\rfloor \leq \frac{n-1}{k}\right) \\
&= \frac{n+k-1}{k}.
\end{aligned}$$

But $S \subseteq \{1,2,\ldots,n\}$ and therefore $S = (S\cap I_0)\cup(S\cap I_1)\cup\cdots\cup\left(S\cap I_q\right)$ (since the intervals $I_0, I_1, \ldots, I_q$ cover the entire set $\{1,2,\ldots,n\}$). Hence,

$$|S| = \left|(S\cap I_0)\cup(S\cap I_1)\cup\cdots\cup\left(S\cap I_q\right)\right| \leq \frac{n+k-1}{k}.$$

This solves Exercise 3.7.1 again. □

Hint to a fourth solution to Exercise 3.7.1. This was suggested by one of the students. This time I will be really terse, since formalizing this would take a while. We say that the k-lacunar subset S is *left-flush* if it has the form

$$\{1, k+1, 2k+1, 3k+1, \ldots, pk+1\}$$

for some $p \in \mathbb{N} \cup \{-1\}$ (this allows $p = -1$, in which case S will be the empty set). In other words, the k-lacunar subset S is left-flush if its smallest element is

as small as possible (that is, 1), its second-smallest element is as small as possible (that is, $k+1$, because S has to be k-lacunar), its third-smallest element is as small as possible (that is, $2k+1$), etc.. It is easy to see that the inequality $|S| \leq \dfrac{n+k-1}{k}$ holds when the k-lacunar subset S is left-flush. Now, it remains to show that we can transform **any** k-lacunar subset into a left-flush k-lacunar subset without changing its size. But it is fairly clear how to do this: Just keep decreasing the smallest element until no longer possible (i.e., until it hits 1); then do the same with the second-smallest element (which will end up at $k+1$ because the subset must remain k-lacunar); then do the same with the third-smallest element; and so on. □

□

A.2.2. Discussion of Exercise 3.7.2

Discussion of Exercise 3.7.2. This is a classical problem (see, e.g., [Vorobi02, Chapter 2, §12] or [Gunder10, Exercise 376] or `https://artofproblemsolving.com/community/c6h63078p376755`).

The solution is rather similar to the solution we gave for Exercise 3.4.1 **(b)**; but first we need a lemma:

Lemma A.2.1. Let $(f_0, f_1, f_2, \ldots)$ be the Fibonacci sequence. Let $a, b \in \mathbb{N}$ be such that $a > 0$ and $a \leq b$. Then, $\gcd(f_a, f_b) = \gcd(f_a, f_{b-a})$.

Proof of Lemma A.2.1. From $a > 0$, we obtain $a \geq 1$ and thus $a - 1 \in \mathbb{N}$. Hence, Exercise 2.2.3 (applied to $n = b - a$ and $m = a - 1$) yields

$$\begin{aligned} f_{(b-a)+(a-1)+1} &= f_{b-a} f_{a-1} + f_{(b-a)+1} \underbrace{f_{(a-1)+1}}_{\substack{=f_a \\ \equiv 0 \bmod f_a}} \\ &\equiv f_{b-a} f_{a-1} + f_{(b-a)+1} 0 = f_{b-a} f_{a-1} \bmod f_a. \end{aligned}$$

In view of $(b-a)+(a-1)+1 = b$, this rewrites as $f_b \equiv f_{b-a} f_{a-1} \bmod f_a$. Hence, Proposition 3.4.4 **(d)** (applied to f_a, f_b and $f_{b-a} f_{a-1}$ instead of a, b and c) yields

$$\gcd(f_a, f_b) = \gcd(f_a, f_{b-a} f_{a-1}). \tag{463}$$

But Exercise 3.5.2 (applied to $n = a - 1$) yields $f_{a-1} \perp f_{(a-1)+1}$ (since $a - 1 \in \mathbb{N}$). In other words, $f_{a-1} \perp f_a$. According to Proposition 3.5.4, this yields $f_a \perp f_{a-1}$. Hence, Proposition 3.5.18 (applied to f_a, f_{b-a} and f_{a-1} instead of a, b and c) yields $\gcd(f_a, f_{b-a} f_{a-1}) = \gcd(f_a, f_{b-a})$. Thus, (463) becomes

$$\gcd(f_a, f_b) = \gcd(f_a, f_{b-a} f_{a-1}) = \gcd(f_a, f_{b-a}).$$

This proves Lemma A.2.1. □

We can now step to the actual solution to Exercise 3.7.2:

Solution to Exercise 3.7.2 (sketched). We use strong induction on $n+m$:

Induction step: Let $k \in \mathbb{N}$. Assume (as the induction hypothesis) that Exercise 3.7.2 is true for $n+m<k$. We must prove that Exercise 3.7.2 is true for $n+m=k$. Let us do this now, renaming n, m as a, b in order to match the notations from the proof of Theorem 3.4.5.

So let $a, b \in \mathbb{N}$ be such that $a+b=k$. We must show that $\gcd(f_a, f_b) = f_{\gcd(a,b)}$.

Note that a and b play symmetric roles in this claim[327], and thus can be swapped at will. By swapping a and b if necessary, we can ensure that $a \leq b$. Hence, we WLOG assume that $a \leq b$. Thus, $b-a \in \mathbb{N}$.

Definition 2.2.1 easily shows that all the Fibonacci numbers $f_0, f_1, f_2, \ldots$ are nonnegative[328]. Thus, $f_{\gcd(a,b)}$ is nonnegative.

It is easy to see that our claim $\gcd(f_a, f_b) = f_{\gcd(a,b)}$ holds if $a=0$ [329]. Thus, we are done if $a=0$. Hence, we WLOG assume that $a \neq 0$. Therefore, $a>0$ (since $a \in \mathbb{N}$). Thus, $a+b>b$, so that $b<a+b=k$.

But our induction hypothesis says that Exercise 3.7.2 is true for $a+b<k$. Hence, we can apply Exercise 3.7.2 to $b-a$ instead of b (since $b-a \in \mathbb{N}$ and $a+(b-a) = b<k$). We thus obtain

$$\gcd(f_a, f_{b-a}) = f_{\gcd(a,b-a)}. \tag{464}$$

But we have $\gcd(a, b-a) = \gcd(a,b)$ (this has already been proved during our proof of Theorem 3.4.5). Furthermore, Lemma A.2.1 yields

$$\begin{aligned} \gcd(f_a, f_b) &= \gcd(f_a, f_{b-a}) = f_{\gcd(a,b-a)} \qquad \text{(by (464))} \\ &= f_{\gcd(a,b)} \qquad (\text{since } \gcd(a, b-a) = \gcd(a,b)). \end{aligned}$$

Now, forget that we fixed a, b. We thus have shown that any $a, b \in \mathbb{N}$ satisfying $a+b=k$ satisfy $\gcd(f_a, f_b) = f_{\gcd(a,b)}$. Renaming the variables a and b as n and m in this statement, we obtain the following: Any $n, m \in \mathbb{N}$ satisfying $n+m=k$ satisfy $\gcd(f_n, f_m) = f_{\gcd(n,m)}$. In other words, Exercise 3.7.2 is true for $n+m=k$. This completes the induction step. Thus, Exercise 3.7.2 is solved. □

[327] because Proposition 3.4.4 **(b)** yields $\gcd(a,b) = \gcd(b,a)$ and $\gcd(f_a, f_b) = \gcd(f_b, f_a)$

[328] Strictly speaking, this can be proved by strong induction.

[329] *Proof.* Assume that $a=0$. Then, $\gcd(a,b) = b$ (this has already been proved during our above proof of Theorem 3.4.5) and thus $b = \gcd(a,b)$. Furthermore, from $a=0$, we obtain $f_a = f_0 = 0$ and therefore

$$\begin{aligned} \gcd(f_a, f_b) &= \gcd(0, f_b) \\ &= \gcd(f_b, 0) \qquad \text{(by Proposition 3.4.4 } \textbf{(b)}) \\ &= |f_b| \qquad \text{(by Proposition 3.4.4 } \textbf{(a)}) \\ &= \left| f_{\gcd(a,b)} \right| \qquad (\text{since } b = \gcd(a,b)) \\ &= f_{\gcd(a,b)} \qquad \left(\text{since } f_{\gcd(a,b)} \text{ is nonnegative}\right), \end{aligned}$$

qed.

□

A.2.3. Discussion of Exercise 3.7.3

Discussion of Exercise 3.7.3. Exercise 3.7.3 is a classic (see, e.g., [Gunder10, Exercise 22 and the paragraph thereafter]). Part **(a)** is a simple induction, while part **(b)** follows from part **(a)**. Here is the solution in detail:

Solution to Exercise 3.7.3. **(a)** We proceed by induction on n:

Induction base: The definition of F_0 yields $F_0 = 2^{2^0} + 1 = 2^1 + 1 = 3$. Comparing this with $\underbrace{F_0 F_1 \cdots F_{-1}}_{=(\text{empty product})=1} + 2 = 1 + 2 = 3$, we obtain $F_0 = F_0 F_1 \cdots F_{-1} + 2$. In other words, Exercise 3.7.3 **(a)** holds for $n = 0$.

Induction step: Let $m \in \mathbb{N}$. Assume (as the induction hypothesis) that Exercise 3.7.3 **(a)** holds for $n = m$. We must prove that Exercise 3.7.3 **(a)** holds for $n = m + 1$. In other words, we must prove that $F_{m+1} = F_0 F_1 \cdots F_m + 2$.

Our induction hypothesis says that Exercise 3.7.3 **(a)** holds for $n = m$. In other words, $F_m = F_0 F_1 \cdots F_{m-1} + 2$. Solving this for $F_0 F_1 \cdots F_{m-1}$, we obtain

$$\begin{aligned} F_0 F_1 \cdots F_{m-1} &= \underbrace{F_m}_{\substack{=2^{2^m}+1 \\ (\text{by the definition of } F_m)}} - 2 = \left(2^{2^m} + 1\right) - 2 \\ &= 2^{2^m} - 1. \end{aligned} \tag{465}$$

Now, the definition of F_{m+1} yields $F_{m+1} = 2^{2^{m+1}} + 1$, so that

$$F_{m+1} - 2 = \left(2^{2^{m+1}} + 1\right) - 2 = 2^{2^{m+1}} - 1 = 2^{2^m \cdot 2} - 1 \qquad \left(\text{since } 2^{m+1} = 2^m \cdot 2\right).$$

But we know that $a^{b \cdot c} = \left(a^b\right)^c$ for any $a, b, c \in \mathbb{N}$. Applying this to $a = 2$, $b = 2^m$ and $c = 2$, we find $2^{2^m \cdot 2} = \left(2^{2^m}\right)^2$. Hence,

$$2^{2^m \cdot 2} - 1 = \left(2^{2^m}\right)^2 - 1 = \left(2^{2^m} - 1\right)\left(2^{2^m} + 1\right)$$

(since $a^2 - 1 = (a - 1)(a + 1)$ for any $a \in \mathbb{R}$). Hence,

$$F_{m+1} - 2 = 2^{2^m \cdot 2} - 1 = \left(2^{2^m} - 1\right)\left(2^{2^m} + 1\right).$$

Comparing this with

$$F_0 F_1 \cdots F_m = \underbrace{F_0 F_1 \cdots F_{m-1}}_{\substack{=2^{2^m}-1 \\ (\text{by } (465))}} \cdot \underbrace{F_m}_{=2^{2^m}+1} = \left(2^{2^m} - 1\right)\left(2^{2^m} + 1\right),$$

we find $F_{m+1} - 2 = F_0 F_1 \cdots F_m$. In other words, $F_{m+1} = F_0 F_1 \cdots F_m + 2$. This is precisely what we needed to prove. Thus, the induction step is complete, and Exercise 3.7.3 **(a)** is solved.

(b) Let n and m be two distinct nonnegative integers. We must prove that $\gcd(F_n, F_m) = 1$.

Note that $\gcd(F_n, F_m) = \gcd(F_m, F_n)$ (by Proposition 3.4.4 **(b)**). Hence, our situation is symmetric in n and m. Thus, we can WLOG assume that $n \geq m$ (since otherwise, we can swap n with m). Assume this. Then, $n > m$ (since n and m are distinct and satisfy $n \geq m$). Hence, $m < n$, so that $m \leq n - 1$ (since m and n are integers). Thus, $m \in \{0, 1, \ldots, n-1\}$, so that F_m is one of the factors in the product $F_0 F_1 \cdots F_{n-1}$. Hence, $F_m \mid F_0 F_1 \cdots F_{n-1}$; in other words, $F_0 F_1 \cdots F_{n-1} \equiv 0 \bmod F_m$.

Exercise 3.7.3 **(a)** yields $F_n = \underbrace{F_0 F_1 \cdots F_{n-1}}_{\equiv 0 \bmod F_m} + 2 \equiv 0 + 2 = 2 \bmod F_m$. Thus, Proposition 3.4.4 **(d)** (applied to $a = F_m$, $b = F_n$ and $c = 2$) yields $\gcd(F_m, F_n) = \gcd(F_m, 2)$.

But the definition of F_m yields $F_m = 2^{2^m} + 1$. From $2^m > 0$, we conclude that 2^{2^m} is divisible by 2, so that $2^{2^m} \equiv 0 \bmod 2$. Hence, $F_m = \underbrace{2^{2^m}}_{\equiv 0 \bmod 2} + 1 \equiv 0 + 1 = 1 \bmod 2$. In view of Exercise 3.3.2 **(d)**, this shows that F_m is odd. In other words, $2 \nmid F_m$. From this, it easily follows that $\gcd(F_m, 2) = 1$ [330]. Hence,

$$\gcd(F_n, F_m) = \gcd(F_m, F_n) = \gcd(F_m, 2) = 1.$$

This solves Exercise 3.7.3 **(b)**. □

□

A.2.4. Discussion of Exercise 3.7.4

Discussion of Exercise 3.7.4. This is clearly a followup to Exercise 1.1.1, meant to illustrate that $\dfrac{1! \cdot 2! \cdot \cdots \cdot (2n)!}{q!}$ can be a perfect square not just when n is even but also for some odd values of n. In order to solve this, we fix some $n \in \mathbb{N}$, and we recall the equality

$$1! \cdot 2! \cdot \cdots \cdot (2n)! = \left(\prod_{k=1}^{n} ((2k-1)!)^2 \right) \cdot 2^n n! \tag{466}$$

[330] *Proof.* Proposition 3.4.4 **(f)** yields $\gcd(F_m, 2) \mid F_m$ and $\gcd(F_m, 2) \mid 2$. But $\gcd(F_m, 2)$ is a nonnegative integer (by Proposition 3.4.3 **(a)**). Hence, from $\gcd(F_m, 2) \mid 2$, we conclude that $\gcd(F_m, 2)$ is a nonnegative divisor of 2. Since the only nonnegative divisors of 2 are 1 and 2, we thus have either $\gcd(F_m, 2) = 1$ or $\gcd(F_m, 2) = 2$. But $\gcd(F_m, 2) = 2$ is impossible, since this would yield $2 = \gcd(F_m, 2) \mid F_m$, which would contradict $2 \nmid F_m$. Hence, the only option that remains is $\gcd(F_m, 2) = 1$.

that was proved in our solution to Exercise 1.1.1. (Our proof of this equality made no use of the assumption that n be even.) Thus, for each integer $n \geq 2$, we have

$$\begin{aligned}
\frac{1! \cdot 2! \cdot \cdots \cdot (2n)!}{(n+1)!} &= \frac{\left(\prod_{k=1}^{n} ((2k-1)!)^2\right) \cdot 2^n n!}{(n+1)!} \qquad \text{(by (466))} \\
&= \left(\prod_{k=1}^{n} ((2k-1)!)^2\right) \cdot \underbrace{2^n}_{=\left(2^{(n-1)/2}\right)^2 \cdot 2} \cdot \underbrace{\frac{n!}{(n+1)!}}_{=\frac{1}{n+1}} \\
&= \left(\prod_{k=1}^{n} ((2k-1)!)^2\right) \cdot \left(2^{(n-1)/2}\right)^2 \cdot 2 \cdot \frac{1}{n+1} \\
&= \left(\left(\prod_{k=1}^{n} ((2k-1)!)\right) \cdot 2^{(n-1)/2}\right)^2 \cdot \frac{2}{n+1} \\
&= \underbrace{\left(\left(\prod_{k=1}^{n} ((2k-1)!)\right) \cdot \frac{2^{(n-1)/2}}{n+1}\right)^2}_{\substack{\text{a perfect square when } n \text{ is odd} \\ \text{(indeed, the denominator } n+1 \\ \text{is cancelled by the } (2n-1)! \\ \text{factor in the product)}}} \cdot 2\,(n+1)\,. \qquad (467)
\end{aligned}$$

Now, we want to find odd positive integers n for which $\dfrac{1! \cdot 2! \cdot \cdots \cdot (2n)!}{(n+1)!}$ is a perfect square. If $n \geq 2$ is such an integer, then (466) suggests that $2\,(n+1)$ should be a perfect square[331]. In other words, we should have $2\,(n+1) = u^2$ for some $u \in \mathbb{Z}$. Solving this equation for n, we obtain $n = \dfrac{u^2}{2} - 1$. For $\dfrac{u^2}{2}$ to be an integer, u should be even (check this!), so that $u = 2v$ for some $v \in \mathbb{Z}$. Thus, $n = \dfrac{u^2}{2} - 1 = \dfrac{(2v)^2}{2} - 1 = 2v^2 - 1$.

So we have shown that $\dfrac{1! \cdot 2! \cdot \cdots \cdot (2n)!}{(n+1)!}$ is a perfect square whenever $n \geq 2$ is an integer of the form $n = 2v^2 - 1$ for some $v \in \mathbb{Z}$. It is clear that any integer of this form is odd, and furthermore there are infinitely many integers of this form that are ≥ 2. Thus, there are infinitely many odd positive integers n for which $\dfrac{1! \cdot 2! \cdot \cdots \cdot (2n)!}{(n+1)!}$ is a perfect square (namely, all integers of the form $n = 2v^2 - 1$ for $v \in \mathbb{Z}$ satisfying $v \geq 2$). $\square$

[331] At least this is a sufficient condition. (It is also necessary, but this needs a bit more thought.)

A.2.5. Discussion of Exercise 3.7.5

Discussion of Exercise 3.7.5. This is a generalization of Exercise 1.1.3 (which can be recovered by setting $m = 1$). But we will solve this using Exercise 1.1.3 (so we will not get a new solution to Exercise 1.1.3). We will use the following basic property of sums:

Proposition A.2.2. Let S be a finite set. Let $S_1, S_2, \ldots, S_m$ be m subsets of S, where $m \in \mathbb{N}$. Assume that these subsets $S_1, S_2, \ldots, S_m$ are disjoint (i.e., we have $S_i \cap S_j = \varnothing$ for any two distinct elements i and j of $\{1, 2, \ldots, m\}$) and their union is S. Let a_s be a number for each $s \in S$. Then,

$$\sum_{s \in S} a_s = \sum_{s \in S_1} a_s + \sum_{s \in S_2} a_s + \cdots + \sum_{s \in S_m} a_s.$$

Proposition A.2.2 is a slightly rewritten version of [Grinbe15, (26)] (with n renamed as m).

We will also use the following simple property of floors:

Proposition A.2.3. Let $x \in \mathbb{R}$ and $k \in \mathbb{Z}$. Then, $\lfloor x + k \rfloor = \lfloor x \rfloor + k$.

The proof of Proposition A.2.3 can be found in [Grinbe16, proof of Proposition 1.1.10]. (The definition of the floor of a number given in [Grinbe16] is a bit different from the one we gave; but [Grinbe16, Corollary 1.1.6] shows that the two definitions are equivalent.)

Solution to Exercise 3.7.5. The interval $\{0, 1, \ldots, mn - 1\}$ can be written as the union of the m disjoint intervals

$$\begin{aligned}
&\{0, 1, \ldots, n-1\}, \\
&\{n, n+1, \ldots, 2n-1\}, \\
&\{2n, 2n+1, \ldots, 3n-1\}, \\
&\ldots, \\
&\{(m-1)\,n, (m-1)\,n+1, \ldots, mn-1\}
\end{aligned}$$

(each containing exactly n numbers). Hence, Proposition A.2.2 shows that the sum

$\sum\limits_{k=0}^{mn-1} \left\lfloor x + \dfrac{k}{n} \right\rfloor$ can be split into m smaller sums as follows:

$$\begin{aligned}
&\sum_{k=0}^{mn-1} \left\lfloor x + \frac{k}{n} \right\rfloor \\
&= \sum_{k=0}^{n-1} \left\lfloor x + \frac{k}{n} \right\rfloor + \sum_{k=n}^{2n-1} \left\lfloor x + \frac{k}{n} \right\rfloor + \sum_{k=2n}^{3n-1} \left\lfloor x + \frac{k}{n} \right\rfloor + \cdots + \sum_{k=(m-1)n}^{mn-1} \left\lfloor x + \frac{k}{n} \right\rfloor \\
&= \sum_{i=0}^{m-1} \underbrace{\sum_{k=in}^{(i+1)n-1} \left\lfloor x + \frac{k}{n} \right\rfloor}_{\substack{= \sum\limits_{k=0}^{n-1} \left\lfloor x + \frac{k+in}{n} \right\rfloor \\ \text{(here, we have substituted } k+in \\ \text{for } k \text{ in the sum)}}} = \sum_{i=0}^{m-1} \sum_{k=0}^{n-1} \left\lfloor \underbrace{x + \frac{k+in}{n}}_{= x+i+\frac{k}{n}} \right\rfloor \\
&= \sum_{i=0}^{m-1} \underbrace{\sum_{k=0}^{n-1} \left\lfloor x + i + \frac{k}{n} \right\rfloor}_{\substack{= \lfloor n(x+i) \rfloor \\ \text{(by Exercise 1.1.3,} \\ \text{applied to } x+i \text{ instead of } x)}} = \sum_{i=0}^{m-1} \underbrace{\lfloor nx + ni \rfloor}_{\substack{= \lfloor nx \rfloor + ni \\ \text{(by Proposition A.2.3,} \\ \text{applied to } nx \text{ and } ni \text{ instead of } x \text{ and } k)}} \\
&= \sum_{i=0}^{m-1} \left(\lfloor nx \rfloor + ni \right) = \underbrace{\sum_{i=0}^{m-1} \lfloor nx \rfloor}_{= m \lfloor nx \rfloor} + \underbrace{\sum_{i=0}^{m-1} ni}_{= n \sum\limits_{i=0}^{m-1} i} = m \lfloor nx \rfloor + n \underbrace{\sum_{i=0}^{m-1} i}_{\substack{= 0+1+\cdots+(m-1) \\ = 1+2+\cdots+(m-1)}} \\
&= m \lfloor nx \rfloor + n \underbrace{\left(1 + 2 + \cdots + (m-1) \right)}_{\substack{= \frac{(m-1)((m-1)+1)}{2} \\ \text{(by (9), applied to } m-1 \text{ instead of } n)}} \\
&= m \lfloor nx \rfloor + \underbrace{n \frac{(m-1)((m-1)+1)}{2}}_{= n \frac{(m-1)m}{2} = m \frac{n(m-1)}{2}} = m \lfloor nx \rfloor + m \frac{n(m-1)}{2} \\
&= m \left(\lfloor nx \rfloor + \frac{n(m-1)}{2} \right).
\end{aligned}$$

This solves Exercise 3.7.5. □

□

A.2.6. Discussion of Exercise 3.7.6

Discussion of Exercise 3.7.6. This is a straightforward strong induction proof, similar to the solution of Exercise 3.3.6. We apply strong induction on n:

Induction step: Let $m \in \mathbb{N}$. Assume (as the induction hypothesis) that Exercise 3.7.6 holds for all $n < m$. We must prove that Exercise 3.7.6 for $n = m$. In other words, we must prove that

$$f_m \equiv f_{m\%5} \cdot 3^{m//5} \bmod 5. \tag{468}$$

If $m < 5$, then we can see this easily from $m\%5 = m$ and $m//5 = 0$. Thus, we WLOG assume that $m \geq 5$. Hence, $m-2$ and $m-1$ are nonnegative integers. Since these two nonnegative integers $m-2$ and $m-1$ are $< m$, we can thus apply Exercise 3.7.6 to $n = m-2$ and to $n = m-1$ (by our induction hypothesis). We thus obtain

$$f_{m-1} \equiv f_{(m-1)\%5} \cdot 3^{(m-1)//5} \bmod 5$$

and

$$f_{m-2} \equiv f_{(m-2)\%5} \cdot 3^{(m-2)//5} \bmod 5.$$

Adding these two congruences, we obtain

$$f_{m-1} + f_{m-2} \equiv f_{(m-1)\%5} \cdot 3^{(m-1)//5} + f_{(m-2)\%5} \cdot 3^{(m-2)//5} \bmod 5.$$

This rewrites as

$$f_m \equiv f_{(m-1)\%5} \cdot 3^{(m-1)//5} + f_{(m-2)\%5} \cdot 3^{(m-2)//5} \bmod 5 \tag{469}$$

(since the recursive definition of the Fibonacci sequence yields $f_m = f_{m-1} + f_{m-2}$). Our goal is now to deduce (468) from this congruence. In order to do so, it suffices to show that

$$f_{(m-1)\%5} \cdot 3^{(m-1)//5} + f_{(m-2)\%5} \cdot 3^{(m-2)//5} \equiv f_{m\%5} \cdot 3^{m//5} \bmod 5 \tag{470}$$

(because then, combining (469) with (470) will immediately yield (468) by the transitivity of congruence).

The proof of (470) is an easy case distinction. Indeed, Proposition 3.3.2 **(a)** (applied to $n = 5$ and $u = m$) yields that $m\%5 \in \{0,1,2,3,4\}$ and $m\%5 \equiv m \bmod 5$. Symmetry of congruence yields $m \equiv m\%5 \bmod 5$ (since $m\%5 \equiv m \bmod 5$). Since $m\%5 \in \{0,1,2,3,4\}$, we are in one of the following five cases:

Case 1: We have $m\%5 = 0$.

Case 2: We have $m\%5 = 1$.

Case 3: We have $m\%5 = 2$.

Case 4: We have $m\%5 = 3$.

Case 5: We have $m\%5 = 4$.

Before I get to any of these cases, let me say a few generalities. We want to know how the numbers $k//5$ and $k\%5$ change when an integer k is incremented (i.e., increased by 1). The answer is fairly easy:

Observation 1: When an integer k is incremented, its quotient $k//5$ upon division by 5 stays the same unless $5 \mid k+1$, in which case $k//5$ also increases by 1.

Observation 2: When an integer k is incremented, its remainder $k\%5$ upon division by 5 increases by 1 unless $5 \mid k+1$, in which case $k\%5$ changes from 4 to 0.

We leave the proofs of these two observations to the reader.[332] Observation 1 entails that

$$m//5 = \begin{cases} (m-1)//5, & \text{if } 5 \nmid m; \\ (m-1)//5+1, & \text{if } 5 \mid m \end{cases} \tag{471}$$

and

$$(m-1)//5 = \begin{cases} (m-2)//5, & \text{if } 5 \nmid m-1; \\ (m-2)//5+1, & \text{if } 5 \mid m-1. \end{cases} \tag{472}$$

With these equalities in hand, let us analyze the above five cases.

We begin with Case 1. In this case, we have $m\%5 = 0$. Thus, $m \equiv m\%5 = 0 \bmod 5$. Thus, $\underbrace{m}_{\equiv 0 \bmod 5} - 1 \equiv 0 - 1 \equiv 4 \bmod 5$ and $\underbrace{m}_{\equiv 0 \bmod 5} - 2 \equiv 0 - 2 \equiv 3 \bmod 5$. From these three congruences, it easily follows (using Proposition 3.3.2 **(c)**) that $m\%5 = 0$ and $(m-1)\%5 = 4$ and $(m-2)\%5 = 3$. Let us next express $(m-1)//5$ and $(m-2)//5$ through $m//5$. Indeed, $m\%5 = 0$ entails $5 \mid m$, hence $m//5 = (m-1)//5+1$ (because of (471)). Solving this for $(m-1)//5$, we obtain $(m-1)//5 = m//5-1$. Furthermore, $(m-1)\%5 = 4 \neq 0$ entails $5 \nmid m-1$ and therefore $(m-1)//5 = (m-2)//5$ (by (472)). Thus, $(m-2)//5 = (m-1)//5 = m//5 - 1$. Now,

$$\begin{aligned} & \underbrace{f_{(m-1)\%5}}_{\substack{=f_4 \\ (\text{since } (m-1)\%5=4)}} \cdot \underbrace{3^{(m-1)//5}}_{\substack{=3^{m//5-1} \\ (\text{since } (m-1)//5=m//5-1)}} + \underbrace{f_{(m-2)\%5}}_{\substack{=f_3 \\ (\text{since } (m-2)\%5=3)}} \cdot \underbrace{3^{(m-2)//5}}_{\substack{=3^{m//5-1} \\ (\text{since } (m-2)//5=m//5-1)}} \\ & = \underbrace{f_4}_{=3} \cdot 3^{m//5-1} + \underbrace{f_3}_{=2} \cdot 3^{m//5-1} = 3 \cdot 3^{m//5-1} + 2 \cdot 3^{m//5-1} = \underbrace{(3+2)}_{\equiv 0 \bmod 5} \cdot 3^{m//5-1} \\ & \equiv 0 \bmod 5. \end{aligned}$$

Comparing this with

$$\underbrace{f_{m\%5}}_{\substack{=f_0 \\ (\text{since } m\%5=0)}} \cdot 3^{m//5} = \underbrace{f_0}_{=0} \cdot 3^{m//5} = 0 \equiv 0 \bmod 5,$$

we obtain $f_{(m-1)\%5} \cdot 3^{(m-1)//5} + f_{(m-2)\%5} \cdot 3^{(m-2)//5} \equiv f_{m\%5} \cdot 3^{m//5} \bmod 5$. Thus, we have proved (470) in Case 1.

[332] We won't actually use Observation 2; we stated it only for the sake of completeness.

Next, let us handle Case 2. In this case, we have $m\%5 = 1$. Thus, $m \equiv m\%5 = 1 \bmod 5$. Thus, $\underbrace{m}_{\equiv 1 \bmod 5} - 1 \equiv 1 - 1 = 0 \bmod 5$ and $\underbrace{m}_{\equiv 1 \bmod 5} - 2 \equiv 1 - 2 \equiv 4 \bmod 5$. From these three congruences, it easily follows (using Proposition 3.3.2 **(c)**) that $m\%5 = 1$ and $(m-1)\%5 = 0$ and $(m-2)\%5 = 4$. Let us next express $(m-1)//5$ and $(m-2)//5$ through $m//5$. Indeed, $m\%5 = 1 \neq 0$ entails $5 \nmid m$, hence $m//5 = (m-1)//5$ (because of (471)). Solving this for $(m-1)//5$, we obtain $(m-1)//5 = m//5$. Furthermore, $(m-1)\%5 = 0$ entails $5 \mid m-1$ and therefore $(m-1)//5 = (m-2)//5 + 1$ (by (472)). Thus, $(m-2)//5 = \underbrace{(m-1)//5}_{=m//5} - 1 = m//5 - 1$. Now,

$$\underbrace{f_{(m-1)\%5}}_{\substack{=f_0 \\ \text{(since } (m-1)\%5=0)}} \cdot \underbrace{3^{(m-1)//5}}_{\substack{=3^{m//5} \\ \text{(since } (m-1)//5=m//5)}} + \underbrace{f_{(m-2)\%5}}_{\substack{=f_4 \\ \text{(since } (m-2)\%5=4)}} \cdot \underbrace{3^{(m-2)//5}}_{\substack{=3^{m//5-1} \\ \text{(since } (m-2)//5=m//5-1)}}$$
$$= \underbrace{f_0}_{=0} \cdot 3^{m//5} + \underbrace{f_4}_{=3} \cdot 3^{m//5-1} = 3 \cdot 3^{m//5-1} = 3^{m//5} \equiv 3^{m//5} \bmod 5.$$

Comparing this with

$$\underbrace{f_{m\%5}}_{\substack{=f_1 \\ \text{(since } m\%5=1)}} \cdot 3^{m//5} = \underbrace{f_1}_{=1} \cdot 3^{m//5} = 3^{m//5} \equiv 3^{m//5} \bmod 5,$$

we obtain $f_{(m-1)\%5} \cdot 3^{(m-1)//5} + f_{(m-2)\%5} \cdot 3^{(m-2)//5} \equiv f_{m\%5} \cdot 3^{m//5} \bmod 5$. (Actually, this holds even as an equality, not just as a congruence; i.e., the two sides are equal.) Thus, we have proved (470) in Case 2.

We leave it to the reader to prove (470) in the remaining three cases. (In all of these three cases, just as in Case 2, the congruence (470) holds as an equality, not just as a congruence.) □

A.2.7. Discussion of Exercise 3.7.7

Discussion of Exercise 3.7.7. Theorem 3.4.5 yields that there exist integers $x \in \mathbb{Z}$ and $y \in \mathbb{Z}$ such that $\gcd(a,b) = xa + yb$. Consider these x and y, and denote them by x' and y'. (We do not want to call them x and y, since they are not the x and y we are looking for.) Thus, x' and y' are integers satisfying $\gcd(a,b) = x'a + y'b$. Hence, $\gcd(a,b) = x'a + y'b = x'a - (-y')b$.

If x' and $-y'$ are positive, then we are already done (since we can just take $x = x'$ and $y = -y'$). But x' and $-y'$ might not be positive yet. The trick is now to modify x' and $-y'$ in such a way that they become positive but the difference $x'a - (-y')b$ is unchanged. (This is the same technique that we used in the solution to Exercise 3.8.5.)

How can we modify x' and $-y'$ in a way that $x'a - (-y')b$ is unchanged? A simple way to do so is to add b to x' and subtract a from y' (because this causes

$x'a - (-y')\,b$ to become $(x'+b)\,a - (-(y'-a))\,b$, which is still the same as $x'a - (-y')\,b$). More generally, we can pick any $d \in \mathbb{Z}$ and add db to x' and subtract da from y'. Obviously, if we pick d large enough, then this will cause x' and $-y'$ to become positive (since b and a are positive), and we will be done.

Let us make this more explicit: Pick any integer d such that $d > \max\left\{\frac{-x'}{b}, \frac{y'}{a}\right\}$. (Such a d clearly exists, since the integers are unbounded from above.) Then, we have $d > \max\left\{\frac{-x'}{b}, \frac{y'}{a}\right\} \geq \frac{-x'}{b}$. Multiplying this inequality by b, we obtain $bd > -x'$ (since $b > 0$), so that $x' + bd > 0$. Furthermore, $d > \max\left\{\frac{-x'}{b}, \frac{y'}{a}\right\} \geq \frac{y'}{a}$. Multiplying this inequality by a, we obtain $ad > y'$ (since $a > 0$), so that $-y' + ad > 0$. Now, we know that $x' + bd$ and $-y' + ad$ are positive integers (since $x' + bd > 0$ and $-y' + ad > 0$) and satisfy

$$\gcd(a,b) = \left(x' + bd\right) a - \left(-y' + ad\right) b$$

(because $(x'+bd)\,a - (-y'+ad)\,b = x'a + y'b = \gcd(a,b)$). Hence, there exist positive integers x and y such that $\gcd(a,b) = xa - yb$ (namely, $x = x' + bd$ and $y = -y' + ad$). This solves Exercise 3.7.7. □

A.2.8. Discussion of Exercise 3.7.8

Discussion of Exercise 3.7.8. Exercise 3.7.8 is a twist on the well-known fact that every nonnegative integer has a unique base-3 representation (i.e., a unique representation in the form $b_m 3^m + b_{m-1} 3^{m-1} + \cdots + b_0 3^0$ for some $m \in \mathbb{N}$ and some $b_0, b_1, \ldots, b_m \in \{0,1,2\}$ with $b_m \neq 0$). A proof of the latter can be found, e.g., in [Newste20, §7.3, subsection on "Application: tests for divisibility"] or in [Dudley12, Section 13, Theorem 3]. (See also [BecGeo10, §7.1], where the same proof is given for the more familiar base-10 representation.) The solution to Exercise 3.7.8 that we shall outline below is inspired by the proof of this fact.

Solution to Exercise 3.7.8 (sketched). In the following, a *balanced ternary expression* of an integer a will mean a way to express a in the form

$$a = 3^m + b_{m-1} 3^{m-1} + b_{m-2} 3^{m-2} + \cdots + b_0 3^0$$

with $m \in \mathbb{N}$ and $b_0, b_1, \ldots, b_{m-1} \in \{0, 1, -1\}$. Thus, we must prove that any positive integer a has a unique balanced ternary expression.

Let us first prove that no positive integer has more than one such expression. To do so, we shall show the following lemma:

Lemma A.2.4. Let $k \in \mathbb{N}$. If $(c_0, c_1, \ldots, c_k)$ and $(d_0, d_1, \ldots, d_k)$ are two $(k+1)$-tuples of elements of $\{0, 1, -1\}$ satisfying

$$c_k 3^k + c_{k-1} 3^{k-1} + \cdots + c_0 3^0 = d_k 3^k + d_{k-1} 3^{k-1} + \cdots + d_0 3^0, \tag{473}$$

then $(c_0, c_1, \ldots, c_k) = (d_0, d_1, \ldots, d_k)$.

Proof of Lemma A.2.4. We shall prove Lemma A.2.4 by induction over k.

Induction base: Lemma A.2.4 holds for $k = 0$ (because if $k = 0$, then (473) simplifies to $c_0 3^0 = d_0 3^0$, which clearly implies $c_0 = d_0$ and thus $(c_0) = (d_0)$).

Induction step: Let $m \in \mathbb{N}$. Assume (as the induction hypothesis) that Lemma A.2.4 holds for $k = m - 1$. We must prove that Lemma A.2.4 holds for $k = m$.

Our induction hypothesis says that Lemma A.2.4 holds for $k = m - 1$. In other words, if $(c_0, c_1, \ldots, c_{m-1})$ and $(d_0, d_1, \ldots, d_{m-1})$ are two m-tuples of elements of $\{0, 1, -1\}$ satisfying

$$c_{m-1}3^{m-1} + c_{m-2}3^{m-2} + \cdots + c_0 3^0 = d_{m-1}3^{m-1} + d_{m-2}3^{m-2} + \cdots + d_0 3^0,$$

then

$$(c_0, c_1, \ldots, c_{m-1}) = (d_0, d_1, \ldots, d_{m-1}). \tag{474}$$

Now, we need to prove that Lemma A.2.4 holds for $k = m$. So let $(c_0, c_1, \ldots, c_m)$ and $(d_0, d_1, \ldots, d_m)$ be two $(m+1)$-tuples of elements of $\{0, 1, -1\}$ satisfying

$$c_m 3^m + c_{m-1}3^{m-1} + \cdots + c_0 3^0 = d_m 3^m + d_{m-1}3^{m-1} + \cdots + d_0 3^0. \tag{475}$$

Then,

$$\begin{aligned} d_m 3^m + d_{m-1}3^{m-1} + \cdots + d_0 3^0 &= \underbrace{d_m 3^m + d_{m-1}3^{m-1} + \cdots + d_1 3^1}_{\substack{=3\cdot\left(d_m 3^{m-1} + d_{m-1}3^{m-2} + \cdots + d_1 3^0\right) \\ \equiv 0 \bmod 3}} + d_0 \underbrace{3^0}_{=1} \\ &\equiv 0 + d_0 \cdot 1 = d_0 \bmod 3 \end{aligned}$$

and similarly

$$c_m 3^m + c_{m-1}3^{m-1} + \cdots + c_0 3^0 \equiv c_0 \bmod 3.$$

Hence,

$$\begin{aligned} c_0 &\equiv c_m 3^m + c_{m-1}3^{m-1} + \cdots + c_0 3^0 \\ &= d_m 3^m + d_{m-1}3^{m-1} + \cdots + d_0 3^0 \qquad \text{(by (475))} \\ &\equiv d_0 \bmod 3. \end{aligned}$$

But c_0 and d_0 both are elements of $\{0, 1, -1\}$, and thus can only be congruent modulo 3 if they are equal[333]. Thus, from $c_0 \equiv d_0 \bmod 3$, we obtain $c_0 = d_0$. Thus, $c_0 3^0 = d_0 3^0$. Subtracting the latter equality from the equality (475), we obtain

$$c_m 3^m + c_{m-1}3^{m-1} + \cdots + c_1 3^1 = d_m 3^m + d_{m-1}3^{m-1} + \cdots + d_1 3^1.$$

Dividing both sides of this new equality by 3, we find

$$c_m 3^{m-1} + c_{m-1}3^{m-2} + \cdots + c_1 3^0 = d_m 3^{m-1} + d_{m-1}3^{m-2} + \cdots + d_1 3^0.$$

[333] Indeed, it is easy to check that no two distinct elements of $\{0, 1, -1\}$ are congruent modulo 3.

Consequently, we can apply (474) to $(c_1, c_2, \ldots, c_m)$ and $(d_1, d_2, \ldots, d_m)$ instead of $(c_0, c_1, \ldots, c_{m-1})$ and $(d_0, d_1, \ldots, d_{m-1})$. As a result, we obtain $(c_1, c_2, \ldots, c_m) = (d_1, d_2, \ldots, d_m)$. Together with $c_0 = d_0$, this yields $(c_0, c_1, \ldots, c_m) = (d_0, d_1, \ldots, d_m)$.

Forget that we fixed $(c_0, c_1, \ldots, c_m)$ and $(d_0, d_1, \ldots, d_m)$. We thus have shown that if $(c_0, c_1, \ldots, c_m)$ and $(d_0, d_1, \ldots, d_m)$ are two $(m+1)$-tuples of elements of $\{0, 1, -1\}$ satisfying (475), then $(c_0, c_1, \ldots, c_m) = (d_0, d_1, \ldots, d_m)$. In other words, Lemma A.2.4 holds for $k = m$. The induction step is thus complete, and Lemma A.2.4 is proven. $\square$

Now, using Lemma A.2.4, we can easily conclude that no positive integer has

more than one balanced ternary expression.[334]

It remains to prove that every positive integer a has some balanced ternary expression. We shall prove this by strong induction over a.

Induction step: Let b be a positive integer. Assume (as the induction hypothesis) that every positive integer a satisfying $a < b$ has some balanced ternary expression. We must now show that b has some balanced ternary expression.

---

[334] *Proof.* Assume the contrary. Thus, some positive integer a has two different balanced ternary expressions

$$a = 3^k + c_{k-1}3^{k-1} + c_{k-2}3^{k-2} + \cdots + c_0 3^0 \tag{476}$$

and

$$a = 3^m + d_{m-1}3^{m-1} + d_{m-2}3^{m-2} + \cdots + d_0 3^0. \tag{477}$$

Consider this a and these two expressions. We WLOG assume that $m \leq k$ (otherwise, we could just switch the roles of the two expressions). We set $c_k = 1$; thus, (476) rewrites as

$$a = c_k 3^k + c_{k-1}3^{k-1} + \cdots + c_0 3^0. \tag{478}$$

We also set $d_m = 1$; thus, (477) rewrites as

$$a = d_m 3^m + d_{m-1}3^{m-1} + \cdots + d_0 3^0. \tag{479}$$

We furthermore set $d_{m+1} = 0$, $d_{m+2} = 0$, ..., $d_k = 0$ (that is, we set $d_i = 0$ for each $i \in \{m+1, m+2, \ldots, k\}$). Thus,

$$\begin{aligned}
& d_k 3^k + d_{k-1}3^{k-1} + \cdots + d_0 3^0 \\
&= \left(\underbrace{d_k}_{=0} 3^k + \underbrace{d_{k-1}}_{=0} 3^{k-1} + \cdots + \underbrace{d_{m+1}}_{=0} 3^{m+1} \right) + \left(d_m 3^m + d_{m-1}3^{m-1} + \cdots + d_0 3^0 \right) \\
&\qquad (\text{since } m \leq k) \\
&= \underbrace{\left(0 \cdot 3^k + 0 \cdot 3^{k-1} + \cdots + 0 \cdot 3^{m+1} \right)}_{=0} + \left(d_m 3^m + d_{m-1}3^{m-1} + \cdots + d_0 3^0 \right) \\
&= d_m 3^m + d_{m-1}3^{m-1} + \cdots + d_0 3^0.
\end{aligned}$$

Hence, (479) rewrites as

$$a = d_k 3^k + d_{k-1}3^{k-1} + \cdots + d_0 3^0. \tag{480}$$

Thus, from (478), we have

$$c_k 3^k + c_{k-1}3^{k-1} + \cdots + c_0 3^0 = a = d_k 3^k + d_{k-1}3^{k-1} + \cdots + d_0 3^0$$

(by (480)). Lemma A.2.4 thus shows that $(c_0, c_1, \ldots, c_k) = (d_0, d_1, \ldots, d_k)$ (since $(c_0, c_1, \ldots, c_k)$ and $(d_0, d_1, \ldots, d_k)$ are two $(k+1)$-tuples of elements of $\{0, 1, -1\}$). Hence, in particular, we have $c_k = d_k$. In other words, $d_k = c_k = 1 \neq 0$. If we had $m < k$, then we would have $d_k = 0$ (since we have set $d_{m+1} = 0$, $d_{m+2} = 0$, ..., $d_k = 0$), which would contradict $d_k \neq 0$. Thus, we cannot have $m < k$. Hence, we have $m \geq k$, and therefore $m = k$ (since $m \leq k$). Now,

$$(c_0, c_1, \ldots, c_k) = (d_0, d_1, \ldots, d_k) = (d_0, d_1, \ldots, d_m) \qquad (\text{since } k = m).$$

As a consequence, the two expressions (476) and (477) are identical; this contradicts our assumption that they be different. This contradiction proves that our assumption was wrong, qed.

If $b < 3$, then b clearly has a balanced ternary expression (since $1 = 3^0$ is a balanced ternary expression of 1, whereas $2 = 3^1 - 3^0$ is a balanced ternary expression of 2). Thus, for the rest of this proof, we WLOG assume that $b \geq 3$.

Theorem 3.1.8 (applied to $n = 3$ and $u = b$) yields that there exists a unique pair $(q, r) \in \mathbb{Z} \times \{0, 1, \ldots, 3-1\}$ such that $b = q \cdot 3 + r$. Consider this pair (q, r). From $(q, r) \in \mathbb{Z} \times \{0, 1, \ldots, 3-1\}$, we obtain $q \in \mathbb{Z}$ and $r \in \{0, 1, \ldots, 3-1\} = \{0, 1, 2\}$, so that $0 \leq r \leq 2 < 3$. Now,

$$b = q \cdot 3 + r = 3q + \underbrace{r}_{<3} < 3q + 3 = 3(q+1),$$

so that $3(q+1) > b \geq 3$. Dividing this inequality by 3, we find $q + 1 > 1$, thus $q > 0$. Hence, $q \geq 1$ (since $q \in \mathbb{Z}$). Thus, q is a positive integer. Hence, $q + 1$ is a positive integer, too. Moreover, $b = 3q + \underbrace{r}_{\geq 0} \geq 3q > q$ (since $q > 0$), thus $q < b$.

Recall that our induction hypothesis says that every positive integer a satisfying $a < b$ has some balanced ternary expression. We can apply this to $a = q$ (since q is a positive integer and satisfies $q < b$). Thus, we conclude that q has some balanced ternary expression. Let

$$q = 3^k + c_{k-1}3^{k-1} + c_{k-2}3^{k-2} + \cdots + c_0 3^0 \tag{481}$$

be this balanced ternary expression. But recall that $r \in \{0, 1, 2\}$; thus we are in one of the following three cases:

Case 1: We have $r = 0$.

Case 2: We have $r = 1$.

Case 3: We have $r = 2$.

Let us first consider Case 1. In this case, we have $r = 0$. Hence,

$$\begin{aligned} b &= 3q + \underbrace{r}_{=0} = 3q = 3\left(3^k + c_{k-1}3^{k-1} + c_{k-2}3^{k-2} + \cdots + c_0 3^0\right) \qquad \text{(by (481))} \\ &= 3^{k+1} + c_{k-1}3^k + c_{k-2}3^{k-1} + \cdots + c_0 3^1 \\ &= 3^{k+1} + c_{k-1}3^k + c_{k-2}3^{k-1} + \cdots + c_0 3^1 + 0 \cdot 3^0. \end{aligned}$$

This is clearly a balanced ternary expression of b. Hence, we have shown that b has some balanced ternary expression in Case 1.

Let us next consider Case 2. In this case, we have $r = 1$. Hence,

$$\begin{aligned} b &= 3q + \underbrace{r}_{=1} = 3q + 1 \\ &= \underbrace{3\left(3^k + c_{k-1}3^{k-1} + c_{k-2}3^{k-2} + \cdots + c_0 3^0\right)}_{=3^{k+1} + c_{k-1}3^k + c_{k-2}3^{k-1} + \cdots + c_0 3^1} + 1 \qquad \text{(by (481))} \\ &= 3^{k+1} + c_{k-1}3^k + c_{k-2}3^{k-1} + \cdots + c_0 3^1 + \underbrace{1}_{=1 \cdot 3^0} \\ &= 3^{k+1} + c_{k-1}3^k + c_{k-2}3^{k-1} + \cdots + c_0 3^1 + 1 \cdot 3^0. \end{aligned}$$

This is clearly a balanced ternary expression of b. Hence, we have shown that b has some balanced ternary expression in Case 2.

Finally, let us consider Case 3. In this case, we have $r = 2$. Thus, $b = 3q + \underbrace{r}_{=2} = 3q + 2$, so that $b + 1 = (3q + 2) + 1 = 3q + 3 = 3(q + 1)$. But now, from $q > 0$, we obtain $q + 1 < 3q + 1 < 3q + 2 = b$. Recall that our induction hypothesis says that every positive integer a satisfying $a < b$ has some balanced ternary expression. We can apply this to $a = q + 1$ (since $q + 1$ is a positive integer and satisfies $q + 1 < b$). Thus, we conclude that $q + 1$ has some balanced ternary expression. Let

$$q + 1 = 3^\ell + d_{\ell-1}3^{\ell-1} + d_{\ell-2}3^{\ell-2} + \cdots + d_0 3^0 \tag{482}$$

be this balanced ternary expression. Then,

$$\begin{aligned} b + 1 &= 3(q + 1) = 3\left(3^\ell + d_{\ell-1}3^{\ell-1} + d_{\ell-2}3^{\ell-2} + \cdots + d_0 3^0\right) \qquad \text{(by (482))} \\ &= 3^{\ell+1} + d_{\ell-1}3^\ell + d_{\ell-2}3^{\ell-1} + \cdots + d_0 3^1, \end{aligned}$$

so that

$$\begin{aligned} b &= 3^{\ell+1} + d_{\ell-1}3^\ell + d_{\ell-2}3^{\ell-1} + \cdots + d_0 3^1 - 1 \\ &= 3^{\ell+1} + d_{\ell-1}3^\ell + d_{\ell-2}3^{\ell-1} + \cdots + d_0 3^1 + (-1) \cdot 3^0. \end{aligned}$$

This is again a balanced ternary expression of b. Hence, we have shown that b has some balanced ternary expression in Case 3.

Thus, in all three cases, we have shown that b has some balanced ternary expression. Hence, this statement always holds. This completes the induction step.

Now, we have shown that every positive integer a has some balanced ternary expression. But previously, we have shown that no positive integer has more than one balanced ternary expression. Combining these two results, we conclude that every positive integer a has a unique balanced ternary expression. In other words, every positive integer a can be uniquely expressed in the form

$$a = 3^m + b_{m-1}3^{m-1} + b_{m-2}3^{m-2} + \cdots + b_0 3^0$$

where m is a nonnegative integer, and where $b_0, b_1, \ldots, b_{m-1} \in \{0, 1, -1\}$. This solves Exercise 3.7.8. □

[*Remark:* The above proof of existence of a balanced ternary expression is one of several possible approaches. Let me sketch two others, found by students in a different class I taught a while ago:

- *Second approach:* Prove that if $k \in \mathbb{N}$ and if a is a positive integer satisfying $a \leq \frac{1}{2} \cdot 3^k$, then a has a balanced ternary expression which begins[335] with a

[335] We say that a balanced ternary expression of the form $a = 3^m + b_{m-1}3^{m-1} + b_{m-2}3^{m-2} + \cdots + b_0 3^0$ *begins* with 3^m.

power of 3 that is smaller than 3^k. This can be proven by induction on k. In the induction step, you assume that $a \leq \frac{1}{2} \cdot 3^{k+1}$. If $a \leq \frac{1}{2} \cdot 3^k$ as well, then the induction hypothesis finishes the proof; otherwise, you have $\frac{1}{2} \cdot 3^k < a \leq \frac{1}{2} \cdot 3^{k+1}$, and thus $\left|a - 3^k\right| \leq \frac{1}{2} \cdot 3^k$, which means that you can apply the induction hypothesis to $\left|a - 3^k\right|$ instead of a (unless $\left|a - 3^k\right| = 0$, but in this case you are done anyway). This gives you a balanced ternary expression of $\left|a - 3^k\right|$ which begins with a power of 3 that is smaller than 3^k. Now, either add it to 3^k or subtract it from 3^k (depending on whether $\left|a - 3^k\right| = a - 3^k$ or $\left|a - 3^k\right| = -\left(a - 3^k\right)$) to obtain a balanced ternary expression of a that begins with 3^k. This finishes the induction step.

- *Third approach:* Let a be a positive integer. We want to find a balanced ternary expression of a.

  Start with **any** way of writing a in the form $a = c_n 3^n + c_{n-1} 3^{n-1} + \cdots + c_0 3^0$, where $c_n, c_{n-1}, \ldots, c_0$ are some integers which are ≥ -1. (For instance, you can take $n = 0$ and $c_0 = a$. Alternatively, you can take the usual ternary representation[336] of a.) Now, we shall gradually transform this representation of a into one where $c_n, c_{n-1}, \ldots, c_0$ are all in $\{0, 1, -1\}$.

  How do we do this? We pick any i for which $c_i \geq 2$, and then we subtract 3 from c_i while at the same time adding 1 to c_{i+1} [337]. (For instance, if we had $a = \cdots + 4 \cdot 3^7 + 5 \cdot 3^6 + \cdots$, and we take $i = 6$, then it becomes $a = \cdots + 5 \cdot 3^7 + 2 \cdot 3^6 + \cdots$.) Of course, a single step like this does not guarantee us that all c_i are in $\{0, 1, -1\}$. However, we can perform steps like this repeatedly, and none of the steps changes the value of the sum $c_n 3^n + c_{n-1} 3^{n-1} + \cdots + c_0$ (because we have added 1 to c_{i+1} while subtracting 3 from c_i, and thus $c_n 3^n + c_{n-1} 3^{n-1} + \cdots + c_0$ has changed by $1 \cdot 3^{i+1} - 3 \cdot 3^i = 3^{i+1} - 3^{i+1} = 0$). Each of these steps **decreases the sum of the c_i by** 2 (indeed, we subtract 3 from one of them and add 1 to another). Since this sum cannot decrease by 2 indefinitely (after all, it is bounded from below by $(n+1) \cdot (-1)$, since there are $n+1$ of the c_i's and each of them is ≥ -1), this means that our process will come to an end – at some point, we just won't be able to make a step anymore; this means that none of our c_i will be ≥ 2, and that means that what we will have then is a balanced ternary expression of a.

  Did you spot the mistake? It is subtle, because it is true that our process will come to an end. But my argument is not correct: I claimed that the sum is bounded from below by $(n+1) \cdot (-1)$. But n is not fixed; sometimes it grows during the algorithm (when we gain a new "digit"), and "being

[336] "Ternary representation" means "base-3 representation".

[337] If c_{i+1} does not exist (that is, if $i = n$), then we change n to $n+1$ and set $c_{i+1} = 1$. (That is, we essentially extend the representation by one more "digit", similarly to what happens to the usual decimal representation of 999 when we add 1 to 999.)

bounded from below by a bound that keeps moving down" is not "being bounded from below". I hope that this tripwire I inserted makes clear why the termination of the algorithm is **not obvious**. Here is one way to fix the argument: Each of the steps decreases the sum of the c_i by 2, while increasing n by at most 1. So it decreases $n + (c_0 + c_1 + \cdots + c_n)$ by at least $2 - 1 = 1$. And $n + (c_0 + c_1 + \cdots + c_n)$ is really bounded from below, namely by -1 (because each of the c_i is ≥ -1), which shows that it cannot decrease by at least 1 indefinitely.

On another note, balanced ternary representation once seemed to have a bright future in computing.] □

A.2.9. Discussion of Exercise 3.7.9

Discussion of Exercise 3.7.9. This exercise is Exercise 26 in Chapter 8 of Engel's book [Engel98] (although Engel forgets to require the entries of the table to be distinct, which is important for the validity of the exercise). The following solution is essentially taken from [Engel98]; I am not aware of any other solution.

Solution to Exercise 3.7.9 (sketched). We proceed by induction on $m + n$.

Induction base: Exercise 3.7.9 holds for $m + n = 0$ (because in this case we must have $m = 0$ and $n = 0$ and therefore $p = 0$ and $q = 0$, so that $pq = 0$).

Induction step: Let $k \in \mathbb{N}$. Assume (as the induction hypothesis) that Exercise 3.7.9 holds for $m + n = k$. We must prove that Exercise 3.7.9 holds for $m + n = k + 1$.

So let $p, q, m, n \in \mathbb{N}$ be such that $p \leq m$ and $q \leq n$ and $m + n = k + 1$. Let T be an $m \times n$-table of integers, with all entries distinct. Mark some of the entries in T with a cyan marker and some with a red marker, as described in the statement of the exercise. We must show that at least pq entries of T are marked twice (i.e., with both colors).

Let us first simplify our language. We shall call an entry *cyan* if it is marked cyan, and *red* if it is marked red. We shall say that an entry is 1-*marked* if it is marked exactly once (i.e., either cyan or red, but not both). We shall say that an entry is 2-*marked* if it is marked twice. Thus, we must show that at least pq entries of T are 2-marked.

If T has no 1-marked entries, then this is easy to check[338]. Hence, we WLOG assume that T has at least one 1-marked entry. Hence, the set of all 1-marked entries of T is nonempty; this set is also finite (obviously). Thus, this set has a

[338] *Proof.* Assume that T has no 1-marked entries. Hence, every cyan entry of T is 2-marked (since otherwise it would be 1-marked, but T has no 1-marked entries). However, exactly pn entries of the table T are cyan (since T has n columns, and exactly p entries in any column are cyan). All of these pn entries are therefore 2-marked (since every cyan entry of T is 2-marked). This entails that at least pn entries of T are 2-marked. Therefore, at least pq entries of T are 2-marked (since $p \underbrace{n}_{\geq q} \geq pq$). But this is precisely what we wanted to show.

maximum[339]. In other words, the **largest** 1-marked entry of T exists. Let M be this largest 1-marked entry. Since M is 1-marked, this entry M is either cyan or red but not both. In other words, we are in one of the following two cases:

Case 1: The entry M is cyan but not red.

Case 2: The entry M is red but not cyan.

Let us consider Case 1. In this case, the entry M is cyan but not red. Recalling the definitions of the cyan and red markings, we can restate this as follows: The entry M is one of the p largest entries in its column, but not one of the q largest entries in its row. Let R denote the row of T that contains M. Thus, M is not one of the q largest entries of R (since M is not one of the q largest entries in its row). Hence, the q largest entries of R are larger than M, and therefore cannot be 1-marked (since M is the largest 1-marked entry of T). But the q largest entries of R are red (because they are the q largest entries in a row of T), and thus must be either 1-marked or 2-marked. Since we have just shown that they cannot be 1-marked, we conclude that they must be 2-marked. So we have shown that the q largest entries of R are 2-marked.

Note that $p > 0$ (since otherwise, T would not have any cyan entries at all; but we know that the entry M is cyan). Thus, $p \geq 1$, so that $p - 1 \in \mathbb{N}$. Also, $p - 1 \leq m - 1$ (since $p \leq m$).

Let us now remove the row R from the $m \times n$-table T. The result is an $(m-1) \times n$-table T'. Note that the table T' and the row R have no entries in common (since the entries of T are all distinct). Also, all entries of the table T' are distinct (for the same reason). Let us pause this proof for an example:

Example A.2.5. For this example, let $m = 4$ and $n = 4$ and $p = 2$ and $q = 2$, and let T be the $m \times n$-table $\begin{pmatrix} 1 & 2 & 11 & 9 \\ 12 & 4 & 3 & 8 \\ 5 & 6 & 13 & 7 \\ 10 & 14 & 16 & 15 \end{pmatrix}$. Then, the cyan entries are $10, 12, 6, 14, 13, 16, 9, 15$, whereas the red entries are $9, 11, 8, 12, 7, 13, 15, 16$. Thus, the 2-marked entries are $9, 12, 13, 15, 16$, whereas the 1-marked entries are $6, 7, 8, 10, 11, 14$. Hence, M (the largest 1-marked entry) is 14. This entry $M = 14$ is cyan, so we are in Case 1. The row R containing the entry $M = 14$ is the last row of T. Hence, $T' = \begin{pmatrix} 1 & 2 & 11 & 9 \\ 12 & 4 & 3 & 8 \\ 5 & 6 & 13 & 7 \end{pmatrix}$.

We do not copy the cyan and red markings from T to T', but instead we mark some of the entries in T' as follows: In each column of T', we mark the $p - 1$ largest entries with a *cyan* marker. In each row of T', we mark the q largest entries with a *red* marker. Since we have $(m-1) + n = \underbrace{m+n}_{=k+1} - 1 = k + 1 - 1 = k$, our induction

[339] Here we are using the *Extremal Principle* – specifically, the version thereof that says that any nonempty finite set of integers has a largest element (i.e., a maximum). This is exactly the claim of Proposition 2.1.2.

hypothesis shows that we can apply Exercise 3.7.9 to $m-1$, $p-1$ and T' instead of m, p and T. Thus we conclude that at least $(p-1)\,q$ entries of T' are marked twice. In other words, at least $(p-1)\,q$ entries of T' are 2-marked.

Now, we shall show that every entry of T' that is 2-marked in T' is also 2-marked in T. Indeed, we have the following two observations:

- Every entry of T' that is cyan in T' is also cyan in T. (Indeed, if an entry of T' is cyan in T', then it is one of the $p-1$ largest entries in its column in T'. Therefore, it must be one of the p largest entries in its column in T (since its column in T differs from its column in T' only in having one extra entry). In other words, it must be cyan in T.)

- Every entry of T' that is red in T' is also red in T. (Indeed, if an entry of T' is red in T', then it is one of the q largest entries in its row in T'. Therefore, it must be one of the q largest entries in its row in T (since its row in T is a copy of its row in T'). In other words, it must be red in T.)

Combining these two observations, we conclude that every entry of T' that is 2-marked in T' is also 2-marked in T. But recall that at least $(p-1)\,q$ entries of T' are 2-marked in T'. All these entries are therefore also 2-marked in T (since every entry of T' that is 2-marked in T' is also 2-marked in T). Thus, we have found $(p-1)\,q$ entries that are 2-marked in T. Additionally, we know of another set of q entries that are 2-marked in T: namely, the q largest entries of R are 2-marked in T. These two sets of entries are disjoint, since the entries in the former set all come from T' while the entries in the latter set all come from the row R (and since the table T' and the row R have no entries in common). Hence, by combining these two sets of entries, we obtain at least $(p-1)\,q+q$ entries that are 2-marked in T. Thus, at least $(p-1)\,q+q$ entries of T are 2-marked. In other words, at least pq entries of T are 2-marked (since $(p-1)\,q+q=pq$). Thus, in Case 1, we have shown that at least pq entries of T are 2-marked.

An analogous argument (with the roles of m and n interchanged, the roles of p and q interchanged, and the roles of rows and columns interchanged) proves the same result in Case 2.

Thus, in both cases, we have shown that at least pq entries of T are 2-marked. In other words, at least pq entries of T are marked twice.

Now, forget that we fixed m, n, p, q and T. We thus have shown that if $p,q,m,n \in \mathbb{N}$ and T are as defined in the exercise, and if $m+n=k+1$, then at least pq entries of T are marked twice (where the coloring is defined as in the statement of the exercise). In other words, Exercise 3.7.9 holds for $m+n=k+1$. This completes the induction step. Thus, Exercise 3.7.9 is solved. $\square$

$\square$

A.2.10. Discussion of Exercise 3.7.10

Discussion of Exercise 3.7.10. I have taken Exercise 3.7.10 **(a)** from [Grinbe08, Aufgabe 1.27], where it is solved by a double induction (i.e., an induction nested within the induction step of another induction)[340]. I will sketch two solutions to Exercise 3.7.10.

Before we come to the solutions, let us agree on a few notations that will be used in all of them:

- A *bit* shall mean an element of $\{0,1\}$.
- We will never use the notation ab for the product of two numbers a and b. Instead, we will denote this product by $a \cdot b$. This will allow us to use the notation ab for things like the bitstring (a,b) without worrying about ambiguity.

First solution to Exercise 3.7.10 (sketched). A *subsequence* of a bitstring $a_1 a_2 \ldots a_n$ is defined to be a bitstring of the form $a_{i_1} a_{i_2} \ldots a_{i_m}$, where $i_1, i_2, \ldots, i_m$ are elements of $\{1, 2, \ldots, n\}$ satisfying $i_1 < i_2 < \cdots < i_m$. In other words, a *subsequence* of a bitstring a is a bitstring obtained from a by removing some (possibly none or all) entries. For example, the bitstring 011 has eight subsequences: the empty bitstring, the one-letter bitstrings 0, 1, 1, the two-letter bitstrings 01, 01 and 11, and the entire bitstring 011. Some of these subsequences are equal as bitstrings (e.g., the two 01s), but we shall nevertheless distinguish between them. More generally: We shall distinguish two subsequences $a_{i_1} a_{i_2} \ldots a_{i_m}$ and $a_{j_1} a_{j_2} \ldots a_{j_m}$ of a bitstring $a_1 a_2 \ldots a_n$ whenever the respective m-tuples $(i_1, i_2, \ldots, i_m)$ and $(j_1, j_2, \ldots, j_m)$ are distinct (even if the bitstrings $a_{i_1} a_{i_2} \ldots a_{i_m}$ and $a_{j_1} a_{j_2} \ldots a_{j_m}$ are equal).

A bitstring will be called a *comb* if it has the form $0 \underbrace{11\ldots1}_{\substack{\text{some nonzero} \\ \text{number of 1's}}}$. In other words, it is called a comb if it has at least 2 entries, its first entry is a 0, and all its remaining entries are 1s.

[340] Here is a very brief outline of this proof: Given a bitstring a, let n_a denote the number of 0s in a (that is, the number of entries of a equal to 0), and let k_a denote the number of 1s in a. Induct on k_a. Within the induction step, induct on n_a. We must prove that any sequence of moves that can be applied successively to a must have an end. To this end, we consider the first entry of a. If this first entry is 1, then this 1 will never be involved in any move (since a move can only involve a 1 that has a 0 in front of it) and will never move away from its first position; thus, we can as well pretend it does not exist, and thus our claim reduces to the situation with one fewer 1 (to which we can apply the induction hypothesis). On the other hand, if the first entry of a is 0, then our sequence of moves (if long enough) will eventually result in this 0 being involved in a move (since otherwise, we can as well pretend it does not exist, which again reduces the problem to the induction hypothesis). Once this happens, the first entry becomes 1, so we are back in the first case. When fleshing out this proof, keep in mind that we are doing an induction within an induction; make sure to clarify which of the two induction hypotheses is being used at which point!

If a is a bitstring, then an *a-comb* will mean a subsequence of a that is a comb. The *comb number* of a bitstring a is defined as the number of a-combs.[341] For example, the comb number of the bitstring 001011 is 17, because there are exactly seventeen 001011-combs (namely, eight copies of 01, seven copies of 011 and two copies of 0111).

We now make the following crucial observation:

> *Claim 1:* When we apply a move to an bitstring a, the comb number of a decreases by 1.

[*Proof of Claim 1 (sketched):* Let us consider a move that changes a bitstring a to a new bitstring a'. We must prove that the comb number of a' equals the comb number of a minus 1.

The move that changes a to a' looks as follows:

$$a = \ldots \underline{01} \ldots \quad \to \quad \ldots \underline{100} \ldots = a'$$

(where the two "…" parts remain unchanged). We shall refer to the two consecutive entries 01 of a that are being replaced by the move, as well as to the three consecutive entries 100 of a' that replace them, as *mobile*; all other entries of the bitstrings a and a' will be called *frozen*. We furthermore notice that the two mobile entries of a form an a-comb (of the form 01); we call this specific a-comb *trivial*, and we call all other a-combs (including all other a-combs of the form 01) *nontrivial*.

Now, to each nontrivial a-comb, we shall assign a certain a'-comb. We do this as follows:

- If a nontrivial a-comb contains none of the two mobile entries, then it remains an a'-comb after the move (even though its entries can shift their positions: namely, each entry that is to the right of the mobile entries gets moved by one step).
- If a nontrivial a-comb contains the mobile 1 but not the mobile 0, then it remains an a'-comb after the move (even though the mobile 1 moves one step to the left, and the frozen entries to its right move one step to the right).
- If a nontrivial a-comb contains the mobile 0 but not the mobile 1 (and thus begins with the mobile 0, because a 0 in a comb must necessarily be the first entry of the comb), then we turn it into an a'-comb by replacing the mobile 0 in a by the **first** of the two mobile 0s in a'.
- If a nontrivial a-comb contains both mobile entries (and thus begins with them, because a 0 in a comb must necessarily be the first entry of the comb), then we turn it into an a'-comb by replacing these two entries with the **second** of the two mobile 0s in a'. We note that the result is indeed a valid a'-comb (because the a-comb we started with was nontrivial, and thus contained not only the two mobile entries but also at least one further entry).

[341] Recall that we are distinguishing two subsequences that differ in their positions in a, even if they are equal as bitstrings.

Thus, we have transformed each nontrivial a-comb into an a'-comb. This transformation is easily seen to be a 1-to-1 correspondence, because an a'-comb cannot contain more than one mobile entry (check this!). Hence,

$$(\text{the number of nontrivial } a\text{-combs}) = \left(\text{the number of } a'\text{-combs}\right).$$

Therefore,

$$\begin{aligned}\left(\text{the number of } a'\text{-combs}\right) &= (\text{the number of nontrivial } a\text{-combs}) \\ &= (\text{the number of } a\text{-combs}) - 1.\end{aligned}$$

In other words, the comb number of a' equals the comb number of a minus 1. This proves Claim 1.]

We can now easily solve part **(a)** of Exercise 3.7.10: If we apply a sequence of moves (successively) to a bitstring a, then the comb number of the bitstring decreases by 1 during each move (by Claim 1). But the comb number (being a nonnegative integer) cannot keep decreasing by 1 indefinitely[342]. Thus, our sequence of moves cannot be infinite. This solves Exercise 3.7.10 **(a)**.

(b) Let a be any bitstring. Let c_a denote the comb number of a. Let k_a denote the number of 1s in a (that is, the number of entries of a that are equal to 1). Let n_a denote the number of 0s in a.

Let a° be an immovable bitstring obtained from a by performing moves until no more moves are possible. We shall show that a° is uniquely determined by a (without knowing the specific sequence of moves used to construct it).

Indeed, we first observe that a° must have the form $11\ldots100\ldots0$ (that is, some number of 1s followed by some number of 0s). Indeed, if it didn't, then it would have two consecutive entries 01 (check this!), and thus a move could be applied to it; but this would contradict the fact that it is immovable.

Furthermore, any move leaves the number of 1s in a bitstring unchanged. Hence, the number of 1s in a° equals the number of 1s in a (since a° was obtained from a by a sequence of moves). But the latter number is k_a (by definition of k_a). Hence, the number of 1s in a° is k_a.

The bitstring a° has no comb (since a° has the form $11\ldots100\ldots0$). In other words, the comb number of a° is 0. But the comb number of a is c_a. But we know that a° was obtained from a by a sequence of moves, and we know (from Claim 1) that each move decreases the comb number of the bitstring by 1. Thus, the sequence of moves that transformed a into a° must have contained precisely c_a many moves (because the comb number of a was c_a, but the comb number of a° is 0). Since each move increases the number of 0s in the bitstring by 1, we thus conclude that a° has precisely c_a more 0s than a has. In other words, the number of 0s in a° is $n_a + c_a$ (since a has n_a many 0s).

Now we know that a° has the form $11\ldots100\ldots0$, but we also know that the number of 1s in a° is k_a and that the number of 0s in a° is $n_a + c_a$. Hence, we know

[342] In fact, any strictly decreasing sequence of nonnegative integers is finite (and its length cannot surpass its first element plus 1).

exactly how the bitstring a° looks like: It consists of k_a many 1s, followed by $n_a + c_a$ many 0s. This shows that a° is uniquely determined by a (without knowing the specific sequence of moves used to construct it).

It remains to show that the number of moves needed to reach a° is also uniquely determined by a. But this is clear, since we have already seen that this number of moves is c_a. Exercise 3.7.10 **(b)** is thus completely solved. □

Second solution to Exercise 3.7.10 (sketched). If $a = a_1 a_2 \ldots a_n$ is a bitstring and b is an element of $\{0, 1\}$, then ba shall denote the bitstring $ba_1 a_2 \ldots a_n$ (that is, the bitstring obtained by inserting b at the front of a). For example, if $a = 1011$ and $b = 0$, then $ba = 01011$.

For any bitstring a, we let $\operatorname{ones} a$ denote the number of 1s in a (that is, the number of entries of a that are equal to 1). (This was denoted by k_a in the previous proof.) For example, $\operatorname{ones}(01001) = 2$. It is clear that any bitstring c satisfies

$$\operatorname{ones}(1c) = 1 + \operatorname{ones} c \tag{483}$$

and

$$\operatorname{ones}(0c) = \operatorname{ones} c. \tag{484}$$

Also, we let ε denote the empty bitstring; then,

$$\operatorname{ones} \varepsilon = 0. \tag{485}$$

Notice that the formulas (485), (484) and (483) (taken together) can be used to compute the integers $\operatorname{ones} a$ for all bitstrings a. In fact, a bitstring a is either empty (in which case $\operatorname{ones} a$ is determined by (485)), or begins with a 0 (in which case we can use (484) to express $\operatorname{ones} a$ through $\operatorname{ones} c$ for a shorter bitstring c), or begins with a 1 (in which case we can use (483) to express $\operatorname{ones} a$ through $\operatorname{ones} c$ for a shorter bitstring c). Moreover, these formulas provide a unique way of computing $\operatorname{ones} a$ (because for each bitstring a, only one of these three formulas has $\operatorname{ones} a$ on its left hand side). Thus, if we forget how we originally defined $\operatorname{ones} a$, then we can use the formulas (485), (484) and (483) as a recursive definition of the integers $\operatorname{ones} a$ for every bitstring a.

We notice that the two formulas (484) and (483) can be combined into a single formula using the Iverson bracket notation (Definition 4.3.19): For any bitstring c and any $b \in \{0, 1\}$, we have

$$\operatorname{ones}(bc) = [b = 1] + \operatorname{ones} c. \tag{486}$$

(Indeed, this boils down to (484) when $b = 0$, and boils down to (483) when $b = 1$. Since b is either 0 or 1, we thus conclude that (486) always holds.)

For any bitstring a, we let $\operatorname{zeros} a$ denote the number of 0s in a (that is, the number of entries of a that are equal to 0). (This was denoted by n_a in the previous proof.) For example, $\operatorname{zeros}(01001) = 3$. It is clear that any bitstring c satisfies

$$\operatorname{zeros}(1c) = \operatorname{zeros} c \tag{487}$$

and

$$\operatorname{zeros}(0c) = 1 + \operatorname{zeros} c. \tag{488}$$

Also, the empty bitstring ε satisfies

$$\operatorname{zeros} \varepsilon = 0. \tag{489}$$

Again, the formulas (489), (488) and (487) (taken together) can be used as a recursive definition of the integers $\operatorname{zeros} a$ for every bitstring a.

We notice that the two formulas (488) and (487) can be combined into a single formula using the Iverson bracket notation: For any bitstring c and any $b \in \{0, 1\}$, we have

$$\operatorname{zeros}(bc) = [b = 0] + \operatorname{zeros} c. \tag{490}$$

(Indeed, this boils down to (488) when $b = 0$, and boils down to (487) when $b = 1$. Since b is either 0 or 1, we thus conclude that (490) always holds.)

We shall now define another integer $\operatorname{load} a$ for any bitstring a. We define it recursively, by setting

$$\operatorname{load} \varepsilon = 0 \tag{491}$$

and

$$\operatorname{load}(0c) = \operatorname{load} c + 2^{\operatorname{ones} c} - 1 \tag{492}$$

and

$$\operatorname{load}(1c) = \operatorname{load} c \tag{493}$$

for any bitstring c. This is a valid definition of $\operatorname{load} a$, for the same reason as why the formulas (485), (484) and (483) provide a valid recursive definition of the integers $\operatorname{ones} a$ for every bitstring a.

Example A.2.6. Here is an example of how load a is computed:

$$\begin{aligned}
\operatorname{load}(010001) &= \underbrace{\operatorname{load}(10001)}_{\substack{=\operatorname{load}(0001)\\ \text{(by (493))}}} + \underbrace{2^{\operatorname{ones}(10001)}}_{=2^2=4} - 1 \qquad \text{(by (492))} \\
&= \underbrace{\operatorname{load}(0001)}_{\substack{=\operatorname{load}(001)+2^{\operatorname{ones}(001)}-1\\ \text{(by (492))}}} + 4 - 1 \\
&= \underbrace{\operatorname{load}(001)}_{\substack{=\operatorname{load}(01)+2^{\operatorname{ones}(01)}-1\\ \text{(by (492))}}} + \underbrace{2^{\operatorname{ones}(001)}}_{=2^1=2} - 1 + 4 - 1 \\
&= \underbrace{\operatorname{load}(01)}_{\substack{=\operatorname{load}(1)+2^{\operatorname{ones}(1)}-1\\ \text{(by (492))}}} + \underbrace{2^{\operatorname{ones}(01)}}_{=2^1=2} - 1 + 2 - 1 + 4 - 1 \\
&= \underbrace{\operatorname{load}(1)}_{\substack{=\operatorname{load}(1\varepsilon)\\ =\operatorname{load}\varepsilon\\ \text{(by (493))}}} + \underbrace{2^{\operatorname{ones}(1)}}_{=2^1=2} - 1 + 2 - 1 + 2 - 1 + 4 - 1 \\
&= \underbrace{\operatorname{load}\varepsilon}_{=0} + 2 - 1 + 2 - 1 + 2 - 1 + 4 - 1 \\
&= 2 - 1 + 2 - 1 + 2 - 1 + 4 - 1 = 6.
\end{aligned}$$

We notice that

$$\operatorname{load} a \text{ is a nonnegative integer for every bitstring } a. \tag{494}$$

Indeed, this is clear from the recursive definition of load a, since the $2^{\operatorname{ones} c} - 1$ on the right hand side of (492) is always a nonnegative integer (because $\operatorname{ones} c \geq 0$ and thus $2^{\operatorname{ones} c} - 1 \geq 2^0 - 1 = 0$).

We notice that the two formulas (492) and (493) can be combined into a single formula using the Iverson bracket notation: For any bitstring c and any $b \in \{0, 1\}$, we have

$$\operatorname{load}(bc) = \operatorname{load} c + [b = 0] \cdot (2^{\operatorname{ones} c} - 1). \tag{495}$$

(Indeed, this boils down to (492) when $b = 0$, and boils down to (493) when $b = 1$. Since b is either 0 or 1, we thus conclude that (495) always holds.)

We now claim that when we apply a move to a bitstring a,

- the number load a decreases by 1,
- the number ones a stays unchanged, and
- the number zeros a increases by 1.

We shall state this claim in a slightly more formal way. First, we define a notation: We shall use the notation "$a \to b$" to say that a bitstring b is obtained by applying a move to a bitstring a. Now, we claim the following:

> *Claim 1:* Let a and b be two bitstrings such that $a \to b$. Then, $\operatorname{load} b = \operatorname{load} a - 1$ and $\operatorname{ones} b = \operatorname{ones} a$ and $\operatorname{zeros} b = \operatorname{zeros} a + 1$.

We will prove this in the following rewritten form:

> *Claim 2:* Let d be a bitstring. Let $k \in \mathbb{N}$, and let $b_1, b_2, \ldots, b_k$ be any elements of $\{0, 1\}$. Then,
>
> $$\operatorname{load}(b_k b_{k-1} \ldots b_1 100d) = \operatorname{load}(b_k b_{k-1} \ldots b_1 01d) - 1$$
>
> and
>
> $$\operatorname{ones}(b_k b_{k-1} \ldots b_1 100d) = \operatorname{ones}(b_k b_{k-1} \ldots b_1 01d)$$
>
> and
>
> $$\operatorname{zeros}(b_k b_{k-1} \ldots b_1 100d) = \operatorname{zeros}(b_k b_{k-1} \ldots b_1 01d) + 1.$$

[*Proof of Claim 2:* We shall prove Claim 2 by induction on k:

Induction base: For the base case, we need to prove that Claim 2 holds for $k = 0$. In other words, we need to prove that $\operatorname{load}(100d) = \operatorname{load}(01d) - 1$ and $\operatorname{ones}(100d) = \operatorname{ones}(01d)$ and $\operatorname{zeros}(100d) = \operatorname{zeros}(01d) + 1$ (since the list $b_k b_{k-1} \cdots b_1$ is empty when $k = 0$). But this is an easy matter of computation: The bitstrings $100d$ and $01d$ clearly have the same number of 1s (namely, each of them has exactly one more 1 than d). In other words, $\operatorname{ones}(100d) = \operatorname{ones}(01d)$. Furthermore, the bitstring $100d$ has exactly one more 0 than the bitstring $01d$ (since the former bitstring has two more 0s than d, while the latter bitstring has one more 0 than d). In other words, $\operatorname{zeros}(100d) = \operatorname{zeros}(01d) + 1$. Next, we notice that the bitstrings $0d$ and d have the same number of 1s; in other words, $\operatorname{ones}(0d) = \operatorname{ones} d$. On the other hand, the number of 1s in the bitstring $1d$ is larger than the number of 1s in the bitstring d by exactly 1; in other words, $\operatorname{ones}(1d) = \operatorname{ones} d + 1$. Now,

$$\begin{aligned}
\operatorname{load}(01d) &= \underbrace{\operatorname{load}(1d)}_{\substack{=\operatorname{load} d \\ \text{(by (493),} \\ \text{applied to } c=d)}} + \underbrace{2^{\operatorname{ones}(1d)}}_{\substack{=2^{\operatorname{ones} d+1} \\ (\text{since } \operatorname{ones}(1d)=\operatorname{ones} d+1)}} - 1 \qquad \text{(by (492), applied to } c = 1d) \\
&= \operatorname{load} d + 2^{\operatorname{ones} d+1} - 1. \qquad (496)
\end{aligned}$$

But

$$\begin{aligned}
\operatorname{load}(100d) &= \operatorname{load}(00d) \qquad \text{(by (493), applied to } c = 00d) \\
&= \underbrace{\operatorname{load}(0d)}_{\substack{=\operatorname{load} d + 2^{\operatorname{ones} d} - 1 \\ \text{(by (492),} \\ \text{applied to } c = d)}} + \underbrace{2^{\operatorname{ones}(0d)}}_{\substack{=2^{\operatorname{ones} d} \\ (\text{since } \operatorname{ones}(0d) = \operatorname{ones} d)}} - 1 \qquad \text{(by (492), applied to } c = 0d) \\
&= \operatorname{load} d + 2^{\operatorname{ones} d} - 1 + 2^{\operatorname{ones} d} - 1 = \operatorname{load} d + \underbrace{2 \cdot 2^{\operatorname{ones} d}}_{=2^{\operatorname{ones} d + 1}} - 2 \\
&= \operatorname{load} d + 2^{\operatorname{ones} d + 1} - 2 = \underbrace{\left(\operatorname{load} d + 2^{\operatorname{ones} d + 1} - 1\right)}_{\substack{=\operatorname{load}(01d) \\ \text{(by (496))}}} - 1 = \operatorname{load}(01d) - 1.
\end{aligned}$$

Thus, we have shown that $\operatorname{load}(100d) = \operatorname{load}(01d) - 1$ and $\operatorname{ones}(100d) = \operatorname{ones}(01d)$ and $\operatorname{zeros}(100d) = \operatorname{zeros}(01d) + 1$. In other words, Claim 2 holds for $k = 0$.

Induction step: Let $\ell \in \mathbb{N}$. Assume (as the induction hypothesis) that Claim 2 holds for $k = \ell$. We must prove that Claim 2 holds for $k = \ell + 1$.

Let $b_1, b_2, \ldots, b_{\ell+1}$ be any elements of $\{0, 1\}$. We must show that

$$\operatorname{load}(b_{\ell+1} b_\ell \ldots b_1 100d) = \operatorname{load}(b_{\ell+1} b_\ell \ldots b_1 01d) - 1 \tag{497}$$

and

$$\operatorname{ones}(b_{\ell+1} b_\ell \ldots b_1 100d) = \operatorname{ones}(b_{\ell+1} b_\ell \ldots b_1 01d) \tag{498}$$

and

$$\operatorname{zeros}(b_{\ell+1} b_\ell \ldots b_1 100d) = \operatorname{zeros}(b_{\ell+1} b_\ell \ldots b_1 01d) + 1. \tag{499}$$

The induction hypothesis says that Claim 2 holds for $k = \ell$. Hence, we have

$$\operatorname{load}(b_\ell b_{\ell-1} \ldots b_1 100d) = \operatorname{load}(b_\ell b_{\ell-1} \ldots b_1 01d) - 1 \tag{500}$$

and

$$\operatorname{ones}(b_\ell b_{\ell-1} \ldots b_1 100d) = \operatorname{ones}(b_\ell b_{\ell-1} \ldots b_1 01d) \tag{501}$$

and

$$\operatorname{zeros}(b_\ell b_{\ell-1} \ldots b_1 100d) = \operatorname{zeros}(b_\ell b_{\ell-1} \ldots b_1 01d) + 1. \tag{502}$$

Let $c = b_\ell b_{\ell-1} \ldots b_1 100d$ and $c' = b_\ell b_{\ell-1} \ldots b_1 01d$. Thus, the three equalities (500), (501) and (502) can be rewritten as

$$\operatorname{load} c = \operatorname{load}(c') - 1 \tag{503}$$

and

$$\operatorname{ones} c = \operatorname{ones}(c') \tag{504}$$

and

$$\operatorname{zeros} c = \operatorname{zeros}(c') + 1. \tag{505}$$

Now, $b_{\ell+1} \in \{0,1\}$. Hence, (490) (applied to $b = b_{\ell+1}$) yields

$$\operatorname{zeros}(b_{\ell+1}c) = [b_{\ell+1} = 0] + \operatorname{zeros} c. \tag{506}$$

The same argument (applied to c' instead of c) yields

$$\operatorname{zeros}\left(b_{\ell+1}c'\right) = [b_{\ell+1} = 0] + \operatorname{zeros}\left(c'\right). \tag{507}$$

Hence, (506) becomes

$$\begin{aligned}\operatorname{zeros}(b_{\ell+1}c) &= [b_{\ell+1} = 0] + \underbrace{\operatorname{zeros} c}_{\substack{=\operatorname{zeros}(c')+1 \\ \text{(by (505))}}} = \underbrace{[b_{\ell+1} = 0] + \operatorname{zeros}\left(c'\right)}_{\substack{=\operatorname{zeros}(b_{\ell+1}c') \\ \text{(by (507))}}} + 1 \\ &= \operatorname{zeros}\left(b_{\ell+1}c'\right) + 1. \end{aligned} \tag{508}$$

Furthermore, (486) (applied to $b = b_{\ell+1}$) yields

$$\operatorname{ones}(b_{\ell+1}c) = [b_{\ell+1} = 1] + \operatorname{ones} c. \tag{509}$$

The same argument (applied to c' instead of c) yields

$$\operatorname{ones}\left(b_{\ell+1}c'\right) = [b_{\ell+1} = 1] + \operatorname{ones}\left(c'\right). \tag{510}$$

Hence, (509) becomes

$$\begin{aligned}\operatorname{ones}(b_{\ell+1}c) &= [b_{\ell+1} = 1] + \underbrace{\operatorname{ones} c}_{\substack{=\operatorname{ones}(c') \\ \text{(by (504))}}} = [b_{\ell+1} = 1] + \operatorname{ones}\left(c'\right) \\ &= \operatorname{ones}\left(b_{\ell+1}c'\right) \end{aligned} \tag{511}$$

(by (510)).

Finally, (495) (applied to $b = b_{\ell+1}$) yields

$$\operatorname{load}(b_{\ell+1}c) = \operatorname{load} c + [b_{\ell+1} = 0] \cdot \left(2^{\operatorname{ones} c} - 1\right). \tag{512}$$

The same argument (applied to c' instead of c) yields

$$\operatorname{load}\left(b_{\ell+1}c'\right) = \operatorname{load}\left(c'\right) + [b_{\ell+1} = 0] \cdot \left(2^{\operatorname{ones}(c')} - 1\right). \tag{513}$$

Hence, (512) becomes

$$\begin{aligned}\operatorname{load}(b_{\ell+1}c) &= \underbrace{\operatorname{load} c}_{\substack{=\operatorname{load}(c')-1 \\ \text{(by (503))}}} + [b_{\ell+1} = 0] \cdot \underbrace{\left(2^{\operatorname{ones} c} - 1\right)}_{\substack{=2^{\operatorname{ones}(c')}-1 \\ \text{(by (504))}}} \\ &= \operatorname{load}\left(c'\right) - 1 + [b_{\ell+1} = 0] \cdot \left(2^{\operatorname{ones}(c')} - 1\right) \\ &= \underbrace{\operatorname{load}\left(c'\right) + [b_{\ell+1} = 0] \cdot \left(2^{\operatorname{ones}(c')} - 1\right)}_{\substack{=\operatorname{load}(b_{\ell+1}c') \\ \text{(by (513))}}} - 1 \\ &= \operatorname{load}\left(b_{\ell+1}c'\right) - 1. \end{aligned} \tag{514}$$

Now, recall that

$$b_{\ell+1} \underbrace{c}_{=b_\ell b_{\ell-1}\ldots b_1 100d} = b_{\ell+1} b_\ell b_{\ell-1} \ldots b_1 100d = b_{\ell+1} b_\ell \ldots b_1 100d$$

and

$$b_{\ell+1} \underbrace{c'}_{=b_\ell b_{\ell-1}\ldots b_1 01d} = b_{\ell+1} b_\ell b_{\ell-1} \ldots b_1 01d = b_{\ell+1} b_\ell \ldots b_1 01d.$$

In light of these two equalities, we can rewrite the three equalities (514), (511) and (508) as

$$\operatorname{load}(b_{\ell+1} b_\ell \ldots b_1 100d) = \operatorname{load}(b_{\ell+1} b_\ell \ldots b_1 01d) - 1$$

and

$$\operatorname{ones}(b_{\ell+1} b_\ell \ldots b_1 100d) = \operatorname{ones}(b_{\ell+1} b_\ell \ldots b_1 01d)$$

and

$$\operatorname{zeros}(b_{\ell+1} b_\ell \ldots b_1 100d) = \operatorname{zeros}(b_{\ell+1} b_\ell \ldots b_1 01d) + 1.$$

But these are precisely the three equalities (497), (498) and (499). So the latter three equalities are proved.

Now, forget that we fixed $b_1, b_2, \ldots, b_{\ell+1}$. We thus have proved the equalities (497), (498) and (499) for any $\ell+1$ elements $b_1, b_2, \ldots, b_{\ell+1}$ of $\{0,1\}$. In other words, Claim 2 holds for $k = \ell+1$. This completes the induction step, so that Claim 2 is proved.]

[*Proof of Claim 1:* We have $a \to b$. In other words, the bitstring b is obtained by applying a move to a (by the definition of the notation "$a \to b$"). In other words, the bitstring b is obtained from a by picking two consecutive entries 01 and replacing them by 100. In other words, there exist some $k, \ell \in \mathbb{N}$ and some elements $b_1, b_2, \ldots, b_k, d_1, d_2, \ldots, d_\ell$ of $\{0,1\}$ such that

$$a = b_k b_{k-1} \ldots b_1 01 d_1 d_2 \ldots d_\ell \tag{515}$$

and

$$b = b_k b_{k-1} \ldots b_1 100 d_1 d_2 \ldots d_\ell \tag{516}$$

(indeed, $b_k, b_{k-1}, \ldots, b_1$ are the entries of a to the left of the two consecutive entries 01 that are being replaced, whereas $d_1, d_2, \ldots, d_\ell$ are the entries of a to the right of these two consecutive entries 01). Consider these k, ℓ and these $b_1, b_2, \ldots, b_k, d_1, d_2, \ldots, d_\ell$. Claim 2 (applied to $d = d_1 d_2 \ldots d_\ell$) yields

$$\operatorname{load}(b_k b_{k-1} \ldots b_1 100 d_1 d_2 \ldots d_\ell) = \operatorname{load}(b_k b_{k-1} \ldots b_1 01 d_1 d_2 \ldots d_\ell) - 1$$

and

$$\operatorname{ones}(b_k b_{k-1} \ldots b_1 100 d_1 d_2 \ldots d_\ell) = \operatorname{ones}(b_k b_{k-1} \ldots b_1 01 d_1 d_2 \ldots d_\ell)$$

and

$$\operatorname{zeros}(b_k b_{k-1} \ldots b_1 100 d_1 d_2 \ldots d_\ell) = \operatorname{zeros}(b_k b_{k-1} \ldots b_1 01 d_1 d_2 \ldots d_\ell) + 1.$$

In view of (515) and (516), these three equalities rewrite as $\operatorname{load} b = \operatorname{load} a - 1$ and $\operatorname{ones} b = \operatorname{ones} a$ and $\operatorname{zeros} b = \operatorname{zeros} a + 1$. This proves Claim 1.]

Next, we introduce a notation for chains of moves. Namely, if $a_0, a_1, \ldots, a_k$ are several bitstrings[343], then the notation "$a_0 \to a_1 \to \cdots \to a_k$" shall mean that every $i \in \{1, 2, \ldots, k\}$ satisfies $a_{i-1} \to a_i$ (that is, every bitstring in the chain $a_0 \to a_1 \to \cdots \to a_k$ is obtained from the preceding one by a move). Now, from Claim 1, we can easily derive the following:

> *Claim 3:* Let $a_0, a_1, \ldots, a_k$ be bitstrings such that $a_0 \to a_1 \to \cdots \to a_k$. Then, $\operatorname{load}(a_k) = \operatorname{load}(a_0) - k$ and $\operatorname{ones}(a_k) = \operatorname{ones}(a_0)$ and $\operatorname{zeros}(a_k) = \operatorname{zeros}(a_0) + k$.

[*Proof of Claim 3:* This is straightforward to prove by induction on k (using Claim 1 in the induction step). Alternatively, we can proceed as follows: For each $i \in \{1, 2, \ldots, k\}$, we have $a_{i-1} \to a_i$ (since $a_0 \to a_1 \to \cdots \to a_k$), and therefore Claim 1 (applied to $a = a_{i-1}$ and $b = a_i$) yields

$$\operatorname{load}(a_i) = \operatorname{load}(a_{i-1}) - 1 \tag{517}$$

and

$$\operatorname{ones}(a_i) = \operatorname{ones}(a_{i-1}) \tag{518}$$

and

$$\operatorname{zeros}(a_i) = \operatorname{zeros}(a_{i-1}) + 1. \tag{519}$$

Now, combining the equalities (518) for all $i \in \{1, 2, \ldots, k\}$, we obtain

$$\operatorname{ones}(a_k) = \operatorname{ones}(a_{k-1}) = \operatorname{ones}(a_{k-2}) = \cdots = \operatorname{ones}(a_0).$$

Hence, $\operatorname{ones}(a_k) = \operatorname{ones}(a_0)$ is proved. Furthermore, the telescope principle (specifically, an application of Theorem 4.1.16) yields

$$\sum_{i=1}^{k} (\operatorname{load}(a_i) - \operatorname{load}(a_{i-1})) = \operatorname{load}(a_k) - \operatorname{load}(a_0),$$

so that

$$\begin{aligned}\operatorname{load}(a_k) - \operatorname{load}(a_0) &= \sum_{i=1}^{k} \underbrace{(\operatorname{load}(a_i) - \operatorname{load}(a_{i-1}))}_{\substack{=-1 \\ \text{(by (517))}}} = \sum_{i=1}^{k} (-1) \\ &= k \cdot (-1) = -k.\end{aligned}$$

Therefore, $\operatorname{load}(a_k) = \operatorname{load}(a_0) - k$. Finally, the telescope principle (specifically, an application of Theorem 4.1.16) yields

$$\sum_{i=1}^{k} (\operatorname{zeros}(a_i) - \operatorname{zeros}(a_{i-1})) = \operatorname{zeros}(a_k) - \operatorname{zeros}(a_0),$$

[343] Note that the notation a_i stands for a bitstring, not for a single entry of a bitstring here.

so that

$$\begin{aligned}\operatorname{zeros}(a_k) - \operatorname{zeros}(a_0) &= \sum_{i=1}^{k} \underbrace{(\operatorname{zeros}(a_i) - \operatorname{zeros}(a_{i-1}))}_{\substack{=1 \\ \text{(by (519))}}} = \sum_{i=1}^{k} 1 \\ &= k \cdot 1 = k.\end{aligned}$$

In other words, $\operatorname{zeros}(a_k) = \operatorname{zeros}(a_0) + k$. This completes the proof of Claim 3.]

We can now reap the consequences of Claim 3:

> *Claim 4:* Let a be any bitstring. Let $a_0, a_1, \ldots, a_k$ be bitstrings such that $a_0 \to a_1 \to \cdots \to a_k$ and $a_0 = a$. Then, $k \leq \operatorname{load} a$.

[*Proof of Claim 4:* Claim 3 yields $\operatorname{load}(a_k) = \operatorname{load}(a_0) - k$ and $\operatorname{ones}(a_k) = \operatorname{ones}(a_0)$ and $\operatorname{zeros}(a_k) = \operatorname{zeros}(a_0) + k$. But (494) (applied to a_k instead of a) shows that $\operatorname{load}(a_k)$ is a nonnegative integer. Hence, $\operatorname{load}(a_k) \geq 0$. In view of $\operatorname{load}(a_k) = \operatorname{load}(a_0) - k = \operatorname{load} a - k$ (since $a_0 = a$), this rewrites as $\operatorname{load} a - k \geq 0$. In other words, $k \leq \operatorname{load} a$. This proves Claim 4.]

Next, a few more notations. A bitstring shall be called *immovable* if no move can be applied to it. A bitstring shall be called *sorted* if it has the form

$$\underbrace{11\ldots1}_{\text{some number of 1s}} \quad \underbrace{00\ldots0}_{\text{some number of 0s}} .$$

Here, "some number" is allowed to mean "0"; thus, for example, 111 and 00 and the empty bitstring ε are sorted.

It is easy to see that each sorted bitstring is immovable (since it contains no two consecutive entries 01). Now, we claim that the converse is true as well:

> *Claim 5:* Let a be any immovable bitstring. Then, a is sorted and satisfies $\operatorname{load} a = 0$.

[*Proof of Claim 5:* This is easy to see in a few moments of thought, but since we have made a habit of proving everything by induction in this solution, let us prove Claim 5 by induction as well.

The *length* $\ell(c)$ of a bitstring c is defined as the number of its entries. For example, the bitstring 010 has length $\ell(010) = 3$. Now, we shall prove Claim 5 by induction on $\ell(a)$:

Induction base: It is easy to see that Claim 5 holds for $\ell(a) = 0$. [344] This completes the induction base.

Induction step: Let $m \in \mathbb{N}$. Assume (as the induction hypothesis) that Claim 5 holds for $\ell(a) = m$. We must prove that Claim 5 holds for $\ell(a) = m + 1$.

[344] Indeed, if $\ell(a) = 0$, then a is the empty bitstring ε, and thus a is sorted (since ε is sorted) and satisfies $\operatorname{load} a = \operatorname{load} \varepsilon = 0$.

Our induction hypothesis says that Claim 5 holds for $\ell(a) = m$. In other words, if a is any immovable bitstring satisfying $\ell(a) = m$, then

$$a \text{ is sorted and satisfies } \operatorname{load} a = 0. \tag{520}$$

Now, let us prove that Claim 5 holds for $\ell(a) = m + 1$. Let a be an immovable bitstring such that $\ell(a) = m + 1$. Then, a is not the empty bitstring (since a has length $\ell(a) = m + 1 > m \geq 0$). Hence, a has a well-defined first entry. In other words, we can write a in the form $a = ba'$ for some $b \in \{0, 1\}$ and some bitstring a'. Moreover, from $a = ba'$, we see that the length of a is exactly by 1 larger than the length of a' (since b is a single entry). In other words, $\ell(a) = \ell(a') + 1$. Hence, $\ell(a') = \ell(a) - 1 = m$ (since $\ell(a) = m + 1$).

Recall that the bitstring a is immovable. In other words, no move can be applied to a (by the definition of "immovable"). Hence, the bitstring a contains no two consecutive entries 01 (because if it would contain two such consecutive entries, then we could apply a move to a that would replace these entries by 100; but this would contradict the preceding sentence). Thus, in particular, the bitstring a cannot begin with 01.

But the bitstring a contains the bitstring a' as a contiguous segment (since $a = ba'$). Hence, the bitstring a' is immovable[345]. Thus, a' is an immovable bitstring satisfying $\ell(a') = m$. Hence, (520) (applied to a' instead of a) yields that a' is sorted and satisfies $\operatorname{load}(a') = 0$.

Now, $b \in \{0, 1\}$. Hence, we are in one of the following two cases:

Case 1: We have $b = 0$.

Case 2: We have $b = 1$.

Let us first consider Case 1. In this case, we have $b = 0$. Thus, $a = \underbrace{b}_{=0} a' = 0a'$. Recall that the bitstring a cannot begin with 01. In other words, the bitstring $0a'$ cannot begin with 01 (since $a = 0a'$). However, if the bitstring a' would begin with a 1, then the bitstring $0a'$ would begin with 01, which would contradict the previous sentence. Thus, the bitstring a' cannot begin with a 1.

But recall that the bitstring a' is sorted. Thus, if a' contains any 1, then a' must begin with a 1. Since a' cannot begin with a 1 (as we have just shown), we thus conclude that a' cannot contain any 1. Therefore, all entries of a' are 0s. Thus, $a' = \underbrace{00\ldots0}_{\text{some number of 0s}}$. Therefore,

$$a = 0 \underbrace{a'}_{= \underbrace{00\ldots0}_{\text{some number of 0s}}} = 0 \underbrace{00\ldots0}_{\text{some number of 0s}} = \underbrace{00\ldots0}_{\text{some number of 0s}}.$$

[345] *Proof.* Assume the contrary. Thus, the string a' is not immovable. In other words, some move can be applied to a' (by the definition of "immovable"). Hence, the bitstring a' contains two consecutive entries 01 (since any move requires two consecutive entries 01). Therefore, the bitstring a contains two consecutive entries 01 as well (since the bitstring a contains the bitstring a' as a contiguous segment). This contradicts the fact that the bitstring a contains no two consecutive entries 01. Hence, our assumption was wrong, qed.

Therefore, the bitstring a is sorted. Furthermore, $\operatorname{ones}(a') = 0$ (since a' cannot contain any 1). From $a = 0a'$, we obtain

$$\begin{aligned}
\operatorname{load} a &= \operatorname{load}(0a') = \operatorname{load}(a') + 2^{\operatorname{ones}(a')} - 1 \qquad \left(\text{by (492), applied to } c = a'\right) \\
&= \underbrace{\operatorname{load}(a')}_{=0} + \underbrace{2^0 - 1}_{=0} \qquad \left(\text{since } \operatorname{ones}(a') = 0\right) \\
&= 0 + 0 = 0.
\end{aligned}$$

Thus, in Case 1, we have shown that the bitstring a is sorted and satisfies $\operatorname{load} a = 0$.

Let us next consider Case 2. In this case, we have $b = 1$. Thus, $a = \underbrace{b}_{=1} a' = 1a'$. Hence, the bitstring a is sorted[346]. Moreover, from $a = 1a'$, we obtain

$$\begin{aligned}
\operatorname{load} a &= \operatorname{load}(1a') = \operatorname{load}(a') \qquad \left(\text{by (493), applied to } c = a'\right) \\
&= 0.
\end{aligned}$$

Thus, in Case 2, we have shown that the bitstring a is sorted and satisfies $\operatorname{load} a = 0$.

Hence, in each of the two Cases 1 and 2, we have shown that the bitstring a is sorted and satisfies $\operatorname{load} a = 0$. Thus, this always holds.

Now, forget that we fixed a. We thus have proved that if a is any immovable bitstring satisfying $\ell(a) = m + 1$, then a is sorted and satisfies $\operatorname{load} a = 0$. In other words, Claim 5 holds for $\ell(a) = m + 1$. This completes the induction step. Thus, Claim 5 is proved.]

Using Claim 3 and Claim 5, we can now easily see the following:

> *Claim 6:* Let a be any bitstring. Let $a_0, a_1, \ldots, a_k$ be bitstrings such that $a_0 \to a_1 \to \cdots \to a_k$ and $a_0 = a$. Assume that the bitstring a_k is immovable. Then,
>
> $$k = \operatorname{load} a$$
>
> and
>
> $$a_k = \underbrace{11\ldots1}_{\operatorname{ones} a \text{ many 1s}} \underbrace{00\ldots0}_{\operatorname{zeros} a + \operatorname{load} a \text{ many 0s}}.$$

[346] *Proof.* The bitstring a' is sorted. In other words, a' has the form $a' = \underbrace{11\ldots1}_{\text{some number of 1s}} \underbrace{00\ldots0}_{\text{some number of 0s}}$ (by the definition of "sorted"). Hence,

$$\begin{aligned}
a &= 1 \underbrace{a'}_{= \underbrace{11\ldots1}_{\text{some number of 1s}} \underbrace{00\ldots0}_{\text{some number of 0s}}} = 1 \underbrace{11\ldots1}_{\text{some number of 1s}} \underbrace{00\ldots0}_{\text{some number of 0s}}.
\end{aligned}$$

Therefore, the bitstring a also has the form $\underbrace{11\ldots1}_{\text{some number of 1s}} \underbrace{00\ldots0}_{\text{some number of 0s}}$. In other words, a is sorted (by the definition of "sorted").

[*Proof of Claim 6:* Claim 3 yields $\operatorname{load}(a_k) = \operatorname{load}(a_0) - k$ and $\operatorname{ones}(a_k) = \operatorname{ones}(a_0)$ and $\operatorname{zeros}(a_k) = \operatorname{zeros}(a_0) + k$. Furthermore, Claim 5 (applied to a_k instead of a) yields that a_k is sorted and satisfies $\operatorname{load}(a_k) = 0$. Comparing $\operatorname{load}(a_k) = \operatorname{load}(a_0) - k$ with $\operatorname{load}(a_k) = 0$, we obtain $\operatorname{load}(a_0) - k = 0$. Thus, $\operatorname{load}(a_0) = k$. Hence, $k = \operatorname{load}(a_0) = \operatorname{load} a$ (since $a_0 = a$).

The bitstring a_k is sorted. In other words, a_k has the form

$$\underbrace{11\ldots1}_{\text{some number of 1s}} \quad \underbrace{00\ldots0}_{\text{some number of 0s}}$$

(by the definition of "sorted"). In other words, there exist numbers $p, q \in \mathbb{N}$ such that

$$a_k = \underbrace{11\ldots1}_{p \text{ many 1s}} \quad \underbrace{00\ldots0}_{q \text{ many 0s}}. \tag{521}$$

Consider these p and q.

From (521), we see that the bitstring a_k has exactly p many 1s. In other words, the number of 1s in a_k is p. In other words, $\operatorname{ones}(a_k) = p$. Hence, $p = \operatorname{ones}(a_k) = \operatorname{ones}(a_0) = \operatorname{ones} a$ (since $a_0 = a$).

Furthermore, from (521), we see that the bitstring a_k has exactly q many 0s. In other words, the number of 0s in a_k is q. In other words, $\operatorname{zeros}(a_k) = q$. Hence, $q = \operatorname{zeros}(a_k) = \operatorname{zeros}(a_0) + k = \operatorname{zeros} a + \operatorname{load} a$ (since $a_0 = a$ and $k = \operatorname{load} a$).

Thus, we know that $p = \operatorname{ones} a$ and $q = \operatorname{zeros} a + \operatorname{load} a$. Hence, the equality (521) rewrites as

$$a_k = \underbrace{11\ldots1}_{\operatorname{ones} a \text{ many 1s}} \quad \underbrace{00\ldots0}_{\operatorname{zeros} a + \operatorname{load} a \text{ many 0s}}.$$

This completes the proof of Claim 6.]

Solving the exercise is now just a matter of translating Claims 4 and 6 back:

(a) Let us start with a bitstring a, and consider a sequence of moves that can be applied successively to it. We must prove that this sequence must have an end. We shall prove a stronger claim: namely, that this sequence cannot have more than $\operatorname{load} a$ moves.

Indeed, assume the contrary. Thus, this sequence has more than $\operatorname{load} a$ moves. This means that we can apply more than $\operatorname{load} a$ moves consecutively starting at a. Hence, in particular, we can apply $\operatorname{load} a + 1$ moves consecutively starting at a. Let $a_0, a_1, \ldots, a_{\operatorname{load} a + 1}$ denote the bitstrings obtained in this process (starting with a itself). Thus, $a_0 \to a_1 \to \cdots \to a_{\operatorname{load} a + 1}$ and $a_0 = a$. Now, Claim 4 (applied to $k = \operatorname{load} a + 1$) yields $\operatorname{load} a + 1 \leq \operatorname{load} a$, which is absurd. This contradiction shows that our assumption was wrong. Hence, our claim (that our sequence of moves cannot have more than $\operatorname{load} a$ moves) is proved. This solves Exercise 3.7.10 **(a)**.

(b) Let a° be an immovable bitstring obtained from a by performing moves until no more moves are possible. We must prove that this bitstring a° is uniquely determined by a (independently of the sequence of moves used to obtain it). We

must also show that the number of moves used to obtain a° from a is uniquely determined by a.

We have assumed that a° is obtained from a by performing moves. Let $a_0, a_1, \ldots, a_k$ be the bitstrings obtained in this sequence of moves (starting with a and ending with a°). Thus, $a_0 \to a_1 \to \cdots \to a_k$ and $a_0 = a$ and $a_k = a^\circ$. Note that the number of moves used to obtain a° from a through this sequence is k.

Now, the bitstring a° is immovable. In other words, the bitstring a_k is immovable (since $a_k = a^\circ$). Hence, Claim 6 yields that

$$k = \operatorname{load} a$$

and

$$a_k = \underbrace{11\ldots1}_{\text{ones } a \text{ many 1s}} \underbrace{00\ldots0}_{\text{zeros } a + \operatorname{load} a \text{ many 0s}} .$$

Thus,

$$a^\circ = a_k = \underbrace{11\ldots1}_{\text{ones } a \text{ many 1s}} \underbrace{00\ldots0}_{\text{zeros } a + \operatorname{load} a \text{ many 0s}} .$$

This equality shows that a° is uniquely determined by a (since the right hand side of this equality is clearly uniquely determined by a). Furthermore, k is uniquely determined by a (since $k = \operatorname{load} a$). In other words, the number of moves used to obtain a° from a is uniquely determined by a (since the number of moves used to obtain a° from a is k). Thus, Exercise 3.7.10 **(b)** is solved. □

[*Remark:* The first and the second solution to Exercise 3.7.10 given above differ mostly in their presentation. At their core, they are doing the same thing. For example, the comb number of a bitstring a (as defined in the first solution) is precisely the number $\operatorname{load} a$ (as defined in the second solution). Thus, Claim 1 in the first solution is equivalent to Claim 1 in the second. The proofs of the two Claims 1 are different, but the way they are used is also the same except for the presentation. The main difference is that the first solution involved some combinatorial handwaving (in defining a 1-to-1 correspondence between nontrivial a-combs and a'-combs) while the second instead relied on rigorous induction arguments. This allowed the first solution to be shorter, at the (probable) expense of readability. I think of the two solutions as more or less the same argument, with the first solution being how a combinatorialist would write it up, while the second solution is how a computer scientist would write it up.] □

A.3. Homework set #2 discussion

The following are discussions of the problems on homework set #2 (Section 4.5).

A.3.1. Discussion of Exercise 4.5.1

Discussion of Exercise 4.5.1. Here is the probably most direct solution:

Solution to Exercise 4.5.1. Let $i \in \{1, 2, \ldots, n\}$. Then, a_i is an odd integer (since $a_1, a_2, \ldots, a_n$ are n odd integers). Thus, Exercise 3.3.2 **(d)** (applied to $u = a_i$) yields that $a_i \equiv 1 \bmod 2$. In other words, $2 \mid a_i - 1$. In other words, there exists an integer c_i such that $a_i - 1 = 2c_i$. Consider this c_i.

Forget that we fixed i. Thus, for each $i \in \{1, 2, \ldots, n\}$, we have constructed an integer c_i such that $a_i - 1 = 2c_i$. Hence, for each $i \in \{1, 2, \ldots, n\}$, we have

$$a_i = 2c_i + 1 \tag{522}$$

(since $a_i - 1 = 2c_i$). Now,

$$\begin{aligned}
&\underbrace{a_1a_2 + a_2a_3 + \cdots + a_{n-1}a_n}_{=\sum\limits_{k=1}^{n-1} a_k a_{k+1}} + a_n a_1 \\
&= \sum_{k=1}^{n-1} \underbrace{a_k}_{\substack{=2c_k+1 \\ \text{(by (522))}}} \underbrace{a_{k+1}}_{\substack{=2c_{k+1}+1 \\ \text{(by (522))}}} + \underbrace{a_n}_{\substack{=2c_n+1 \\ \text{(by (522))}}} \underbrace{a_1}_{\substack{=2c_1+1 \\ \text{(by (522))}}} \\
&= \sum_{k=1}^{n-1} \underbrace{(2c_k+1)(2c_{k+1}+1)}_{=4c_kc_{k+1}+2c_k+2c_{k+1}+1} + \underbrace{(2c_n+1)(2c_1+1)}_{=4c_nc_1+2c_n+2c_1+1} \\
&= \sum_{k=1}^{n-1} (4c_kc_{k+1} + 2c_k + 2c_{k+1} + 1) + 4c_nc_1 + 2c_n + 2c_1 + 1 \\
&= \sum_{k=1}^{n-1} 4c_kc_{k+1} + \sum_{k=1}^{n-1} 2c_k + \underbrace{\sum_{k=1}^{n-1} 2c_{k+1}}_{\substack{=\sum\limits_{k=2}^{n} 2c_k \\ \text{(here, we have substituted } k \\ \text{for } k+1 \text{ in the sum)}}} + \sum_{k=1}^{n-1} 1 + 4c_nc_1 + 2c_n + 2c_1 + 1 \\
&= \sum_{k=1}^{n-1} 4c_kc_{k+1} + \sum_{k=1}^{n-1} 2c_k + \sum_{k=2}^{n} 2c_k + 4c_nc_1 + 2c_n + 2c_1 + \underbrace{\sum_{k=1}^{n-1} 1 + 1}_{=\sum\limits_{k=1}^{n} 1 = n\cdot 1 = n} \\
&= \underbrace{\sum_{k=1}^{n-1} 4c_kc_{k+1}}_{=4\sum\limits_{k=1}^{n-1} c_kc_{k+1}} + 4c_nc_1 + \underbrace{\sum_{k=1}^{n-1} 2c_k + 2c_n}_{\substack{=\sum\limits_{k=1}^{n} 2c_k \\ =2\sum\limits_{k=1}^{n} c_k}} + \underbrace{\sum_{k=2}^{n} 2c_k + 2c_1}_{\substack{=\sum\limits_{k=1}^{n} 2c_k \\ =2\sum\limits_{k=1}^{n} c_k}} + n \\
&= 4\sum_{k=1}^{n-1} c_kc_{k+1} + 4c_nc_1 + 2\sum_{k=1}^{n} c_k + 2\sum_{k=1}^{n} c_k + n \\
&= 4\sum_{k=1}^{n-1} c_kc_{k+1} + 4c_nc_1 + 4\sum_{k=1}^{n} c_k + n \\
&= \underbrace{4\left(\sum_{k=1}^{n-1} c_kc_{k+1} + c_nc_1 + \sum_{k=1}^{n} c_k\right)}_{\equiv 0 \bmod 4} + n \equiv 0 + n = n \bmod 4.
\end{aligned}$$

This solves Exercise 4.5.1. □

Several variants of the above solution exist. One way to simplify it would be to

discard multiples of 4 the moment they appear (rather than, as we did, at the end of the computation). This is allowed because of (40). Another simplification can be obtained by "making the sum cyclic": Namely, we WLOG assume that $n \geq 1$ (since the case $n = 0$ is trivial). We set $a_{n+1} = a_1$ and $c_{n+1} = c_1$. Then,

$$\begin{aligned}\sum_{k=1}^{n} c_{k+1} &= \sum_{k=2}^{n+1} c_k \qquad \text{(here, we have substituted } k \text{ for } k+1 \text{ in the sum)} \\ &= \sum_{k=2}^{n} c_k + \underbrace{c_{n+1}}_{=c_1} = \sum_{k=2}^{n} c_k + c_1 = \sum_{k=1}^{n} c_k. \end{aligned} \tag{523}$$

Furthermore, the equality (522) holds not only for all $i \in \{1, 2, \ldots, n\}$, but also for $i = n+1$ (since $a_{n+1} = a_1$ and $c_{n+1} = c_1$). Therefore, we can apply (522) to each $i \in \{1, 2, \ldots, n+1\}$. Now,

$$\begin{aligned}& a_1 a_2 + a_2 a_3 + \cdots + a_{n-1} a_n + a_n \underbrace{a_1}_{=a_{n+1}} \\ &= a_1 a_2 + a_2 a_3 + \cdots + a_{n-1} a_n + a_n a_{n+1} \\ &= \sum_{k=1}^{n} \underbrace{a_k}_{\substack{=2c_k+1 \\ \text{(by (522))}}} \underbrace{a_{k+1}}_{\substack{=2c_{k+1}+1 \\ \text{(by (522))}}} = \sum_{k=1}^{n} \underbrace{(2c_k+1)(2c_{k+1}+1)}_{=4c_k c_{k+1}+2c_k+2c_{k+1}+1} \\ &= \sum_{k=1}^{n} (4c_k c_{k+1} + 2c_k + 2c_{k+1} + 1) = \underbrace{\sum_{k=1}^{n} 4c_k c_{k+1}}_{=4\sum_{k=1}^{n} c_k c_{k+1}} + \underbrace{\sum_{k=1}^{n} 2c_k}_{=2\sum_{k=1}^{n} c_k} + \underbrace{\sum_{k=1}^{n} 2c_{k+1}}_{\substack{=2\sum_{k=1}^{n} c_{k+1} \\ =2\sum_{k=1}^{n} c_k \\ \text{(by (523))}}} + \underbrace{\sum_{k=1}^{n} 1}_{=n\cdot 1=n} \\ &= 4\sum_{k=1}^{n} c_k c_{k+1} + 2\sum_{k=1}^{n} c_k + 2\sum_{k=1}^{n} c_k + n = 4\sum_{k=1}^{n} c_k c_{k+1} + 4\sum_{k=1}^{n} c_k + n \\ &= \underbrace{4\left(\sum_{k=1}^{n} c_k c_{k+1} + \sum_{k=1}^{n} c_k\right)}_{\equiv 0 \bmod 4} + n \equiv 0 + n = n \bmod 4.\end{aligned}$$

Note how much simpler the computation has become after we integrated the last addend $a_n a_1$ into the sum. □

A.3.2. Discussion of Exercise 4.5.2

Solution to Exercise 4.5.2. **(a)** Let x and y be positive integers satisfying $ab = xa + yb$. We shall derive a contradiction.

We have $a \mid a(b-x) = ab - ax = by$ (since $ab = xa + yb = ax + by$). But $a \perp b$ (since a and b are coprime). Hence, Theorem 3.5.6 (applied to $c = y$) yields $a \mid y$.

Hence, $y \geq a$ [347]. Multiplying this inequality by b, we obtain $yb \geq ab$ (since b is positive).

Also, xa is positive (since x and a are positive). Hence, $xa > 0$. Therefore, $\underbrace{xa}_{>0} + yb > yb \geq ab$. This contradicts $ab = xa + yb$.

Forget that we fixed x and y. We thus have obtained a contradiction for any two positive integers x and y satisfying $ab = xa + yb$. Hence, there do **not** exist such integers. This solves Exercise 4.5.2 **(a)**.

(b) Exercise 4.5.2 **(a)** shows that there do **not** exist any positive integers x and y satisfying $ab = xa + yb$. In other words:

> *Claim 1:* If x and y are two positive integers satisfying $ab = xa + yb$, then a contradiction ensues.

Now, let $x, y \in \mathbb{N}$ be such that $ab - a - b = xa + yb$. We shall derive a contradiction.

Indeed, $x + 1$ and $y + 1$ are two positive integers (since $x, y \in \mathbb{N}$) and satisfy

$$ab = \underbrace{ab - a - b}_{=xa+yb} + a + b = xa + yb + a + b = (x+1)\,a + (y+1)\,b.$$

Thus, Claim 1 (applied to $x + 1$ and $y + 1$ instead of x and y) yields a contradiction.

Forget that we fixed x, y. We thus have found a contradiction for any two integers $x, y \in \mathbb{N}$ satisfying $ab - a - b = xa + yb$. Thus, there do **not** exist such integers. This solves Exercise 4.5.2 **(b)**. □

A.3.3. Discussion of Exercise 4.5.3

Solution to Exercise 4.5.3. We have $\gcd(n, m) = 1$ (since n and m are coprime). Thus, $u^{\gcd(n,m)} = u^1 = u$.

Exercise 3.4.1 **(b)** (applied to $a = n$ and $b = m$) yields $\gcd(u^n - 1, u^m - 1) = \left|u^{\gcd(n,m)} - 1\right| = |u - 1|$ (since $u^{\gcd(n,m)} = u$). But each integer v satisfies $|v| = \pm v$. Applying this to $v = u - 1$, we find $|u - 1| = \pm(u - 1)$.

On the other hand, it is easy to see that $u^n - 1 \mid u^{nm} - 1$. Indeed, the simplest way to prove this is as follows: Clearly, $u^n \equiv 1 \bmod u^n - 1$ (since $u^n - 1 \mid u^n - 1$). Hence, Proposition 3.2.7 (applied to u^n, 1, $u^n - 1$ and m instead of a, b, n and k) yields $(u^n)^m \equiv 1^m \bmod u^n - 1$. In view of $(u^n)^m = u^{nm}$ and $1^m = 1$, this rewrites as $u^{nm} \equiv 1 \bmod u^n - 1$. In other words, $u^n - 1 \mid u^{nm} - 1$.

The same argument (with the roles of n and m interchanged) yields $u^m - 1 \mid u^{mn} - 1$. In other words, $u^m - 1 \mid u^{nm} - 1$ (since $mn = nm$).

[347] *Proof.* We have $a \mid y$ and $y \neq 0$ (since y is positive). Hence, Proposition 3.1.3 (applied to y instead of b) yields that $|a| \leq |y|$. But $|y| = y$ (since y is positive) and $|a| = a$ (since a is positive). Hence, $a = |a| \leq |y| = y$. In other words, $y \geq a$.

Hence, Theorem 3.4.9 (applied to $a = u^n - 1$, $b = u^m - 1$ and $c = u^{nm} - 1$) yields

$$\begin{aligned}(u^n - 1)(u^m - 1) &\mid \underbrace{\gcd(u^n - 1, u^m - 1)}_{=|u-1|=\pm(u-1)} \cdot (u^{nm} - 1) \\ &= \pm(u-1) \cdot (u^{nm} - 1) \mid (u-1) \cdot (u^{nm} - 1).\end{aligned}$$

This solves Exercise 4.5.3. □

A.3.4. Discussion of Exercise 4.5.4

Exercise 4.5.4 **(b)** is a well-known result (due to Hermite in 1889) and rather popular on mathematical contests; in particular, it appeared as Problem B2 on the Putnam contest 2000, and as Problem 4 on the 40th Virginia Tech Regional Mathematics Contest 2018.

Solution to Exercise 4.5.4. **(a)** We are in one of the following two cases:

Case 1: We have $m = 0$.

Case 2: We have $m > 0$.

Let us first consider Case 1. In this case, we have $m = 0$. Hence, $m - 1 = -1 \notin \mathbb{N}$, so that $\binom{n-1}{m-1} = 0$ (by an application of (118)). Comparing this with

$$\frac{m}{n}\binom{n}{m} = \underbrace{m}_{=0} \cdot \frac{1}{n}\binom{n}{m} = 0,$$

we obtain $\frac{m}{n}\binom{n}{m} = \binom{n-1}{m-1}$. Thus, Exercise 4.5.4 **(a)** is solved in Case 1.

Let us now consider Case 2. In this case, we have $m > 0$. Hence, $m \geq 1$ (since $m \in \mathbb{N}$), so that $m - 1 \in \mathbb{N}$. Hence, (117) yields

$$\begin{aligned}\binom{n-1}{m-1} &= \frac{(n-1)((n-1)-1)((n-1)-2)\cdots((n-1)-(m-1)+1)}{(m-1)!} \\ &= \frac{(n-1)(n-2)(n-3)\cdots(n-m+1)}{(m-1)!}. \qquad (524)\end{aligned}$$

On the other hand, (117) yields

$$\binom{n}{m} = \frac{n(n-1)(n-2)\cdots(n-m+1)}{m!} = \frac{n(n-1)(n-2)\cdots(n-m+1)}{m \cdot (m-1)!}$$

(since $m! = m \cdot (m-1)!$). Multiplying both sides of this equality by $\frac{m}{n}$, we find

$$\begin{aligned}
\frac{m}{n}\binom{n}{m} &= \frac{m}{n} \cdot \frac{n(n-1)(n-2)\cdots(n-m+1)}{m \cdot (m-1)!} \\
&= \frac{1}{(m-1)!} \cdot \underbrace{\frac{n(n-1)(n-2)\cdots(n-m+1)}{n}}_{=(n-1)(n-2)(n-3)\cdots(n-m+1)} \\
&= \frac{1}{(m-1)!} \cdot (n-1)(n-2)(n-3)\cdots(n-m+1) \\
&= \frac{(n-1)(n-2)(n-3)\cdots(n-m+1)}{(m-1)!}.
\end{aligned}$$

Comparing this with (524), we obtain $\frac{m}{n}\binom{n}{m} = \binom{n-1}{m-1}$. Thus, Exercise 4.5.4 **(a)** is solved in Case 2.

We have now solved Exercise 4.5.4 **(a)** in both Cases 1 and 2. This completes the solution to Exercise 4.5.4 **(a)**.

(b) Exercise 4.5.4 **(a)** yields $\frac{m}{n}\binom{n}{m} = \binom{n-1}{m-1}$. Multiplying both sides of this equality by n, we obtain $m\binom{n}{m} = n\binom{n-1}{m-1}$. But Theorem 4.3.15 (applied to $n-1$ and $m-1$ instead of n and k) yields $\binom{n-1}{m-1} \in \mathbb{Z}$. In other words, $\binom{n-1}{m-1}$ is an integer. Also, $\binom{n}{m}$ is an integer (for similar reasons). Now, from $m\binom{n}{m} = n\binom{n-1}{m-1}$, we obtain $n \mid m\binom{n}{m}$ (since $\binom{n-1}{m-1}$ is an integer). Hence, Theorem 3.4.11 (applied to $a = n$, $b = m$ and $c = \binom{n}{m}$) yields $n \mid \gcd(n,m) \cdot \binom{n}{m}$.

But $n \neq 0$ (since $n > 0$). Hence, Proposition 3.1.3 **(d)** (applied to $a = n$ and $b = \gcd(n,m) \cdot \binom{n}{m}$) yields that $n \mid \gcd(n,m) \cdot \binom{n}{m}$ if and only if $\frac{\gcd(n,m) \cdot \binom{n}{m}}{n} \in \mathbb{Z}$. Therefore, $\frac{\gcd(n,m) \cdot \binom{n}{m}}{n} \in \mathbb{Z}$ (since $n \mid \gcd(n,m) \cdot \binom{n}{m}$). Therefore,

$$\frac{\gcd(n,m)}{n}\binom{n}{m} = \frac{\gcd(n,m) \cdot \binom{n}{m}}{n} \in \mathbb{Z}.$$

This solves Exercise 4.5.4 **(b)**. $\square$

A.3.5. Discussion of Exercise 4.5.5

Exercise 4.5.5 is a result of Erdös and Szekeres [ErdSze78, (1)]; it later re-appeared in Crux Mathematicorum as problem 1915 and in `https://artofproblemsolving.com/community/c6h770` .

Solution to Exercise 4.5.5. We note that i and j play symmetric roles in Exercise 4.5.5 (since $\gcd\left(\binom{n}{i}, \binom{n}{j}\right) = \gcd\left(\binom{n}{j}, \binom{n}{i}\right)$). Hence, we can WLOG assume that $i \leq j$ (since otherwise, we can swap i with j). Furthermore, $j \leq n$ (since $j < n$) and $i \geq 0$ (since i is positive). Now, we claim that

$$\binom{n}{j}\binom{j}{i} = \binom{n}{i}\binom{n-i}{j-i}. \tag{525}$$

[*Proof of (525):* The equality (525) is a particular case of the *trinomial revision formula* ([Grinbe15, Proposition 3.23]), which we will eventually see in this course. However, let us give a quick proof for it here: We have $i \leq j$, thus $j - i \geq 0$ and therefore $j - i \in \mathbb{N}$. Also, $\underbrace{j}_{\leq n} - i \leq n - i \leq n$ (since $i \geq 0$) and $n - i \in \mathbb{N}$ (since $i \leq j \leq n$). Now, comparing

$$\underbrace{\binom{n}{i}}_{\substack{= \frac{n!}{i! \cdot (n-i)!} \\ \text{(by Theorem 4.3.8,} \\ \text{applied to } k=i)}} \underbrace{\binom{n-i}{j-i}}_{\substack{= \frac{(n-i)!}{(j-i)! \cdot ((n-i)-(j-i))!} \\ \text{(by Theorem 4.3.8,} \\ \text{applied to } n-i \text{ and } j-i \\ \text{instead of } n \text{ and } k)}} = \frac{n!}{i! \cdot (n-i)!} \cdot \frac{(n-i)!}{(j-i)! \cdot ((n-i)-(j-i))!}$$

$$= \frac{n!}{i! \cdot (j-i)! \cdot ((n-i)-(j-i))!}$$

$$= \frac{n!}{i! \cdot (j-i)! \cdot (n-j)!} \quad (\text{since } (n-i)-(j-i) = n-j)$$

with

$$\underbrace{\binom{n}{j}}_{\substack{= \frac{n!}{j! \cdot (n-j)!} \\ \text{(by Theorem 4.3.8,} \\ \text{applied to } k=j)}} \underbrace{\binom{j}{i}}_{\substack{= \frac{j!}{i! \cdot (j-i)!} \\ \text{(by Theorem 4.3.8,} \\ \text{applied to } j \text{ and } i \\ \text{instead of } n \text{ and } k)}} = \frac{n!}{j! \cdot (n-j)!} \cdot \frac{j!}{i! \cdot (j-i)!} = \frac{n!}{i! \cdot (j-i)! \cdot (n-j)!},$$

we obtain $\binom{n}{j}\binom{j}{i} = \binom{n}{i}\binom{n-i}{j-i}$. This proves (525).]

Next, we observe that all four binomial coefficients $\binom{n}{j}$, $\binom{j}{i}$, $\binom{n}{i}$ and $\binom{n-i}{j-i}$ that appear in (525) belong to $\mathbb{Z}$ (because of Theorem 4.3.15). Hence, (525) entails

$$\binom{n}{i} \mid \binom{n}{j}\binom{j}{i}. \tag{526}$$

Furthermore, (117) yields

$$\begin{aligned}\binom{n}{i} &= \frac{n(n-1)(n-2)\cdots(n-i+1)}{i!} = \frac{1}{i!}\cdot n(n-1)(n-2)\cdots(n-i+1) \\ &= \frac{1}{i!}\cdot\prod_{k=0}^{i-1}(n-k). \end{aligned}\tag{527}$$

The same argument (applied to j instead of n) yields

$$\binom{j}{i} = \frac{1}{i!}\cdot\prod_{k=0}^{i-1}(j-k). \tag{528}$$

But we have $\underbrace{j}_{<n} - k < n-k$ for each $k \in \{0,1,\ldots,i-1\}$. We can multiply these i inequalities, since each of the factors involved is positive (indeed, each $k \in \{0,1,\ldots,i-1\}$ satisfies $j - \underbrace{k}_{<i} > j-i \geq 0$ and $\underbrace{n}_{>j} - k > j-k > 0$). Thus, we obtain

$$\prod_{k=0}^{i-1}(j-k) < \prod_{k=0}^{i-1}(n-k). \tag{529}$$

(Note that this is a **strict** inequality, because $i > 0$. In fact, if i was 0, then both sides of (529) would be empty products and thus would be equal. Trivial as it sounds, but this is a nontrivial pitfall in working with strict inequalities.)

Now, (528) yields

$$\binom{j}{i} = \frac{1}{i!}\cdot\underbrace{\prod_{k=0}^{i-1}(j-k)}_{\substack{<\prod_{k=0}^{i-1}(n-k) \\ \text{(by (529))}}} < \frac{1}{i!}\cdot\prod_{k=0}^{i-1}(n-k) = \binom{n}{i} \tag{530}$$

(by (528)). (Note that we have tacitly used the fact that $\frac{1}{i!} > 0$ here.)

Also, each of the factors $j-k$ on the right hand side of (528) is positive (because for each $k \in \{0,1,\ldots,i-1\}$, we have $j - \underbrace{k}_{<i} > j-i \geq 0$). Thus, the entire right hand side of (528) is positive. Therefore, $\binom{j}{i} > 0$, so that $\binom{j}{i} \neq 0$. From (530), we

obtain $\binom{n}{i} > \binom{j}{i} > 0$, so that $\binom{n}{i} \neq 0$. Hence, $\gcd\left(\binom{n}{i}, \binom{n}{j}\right)$ is a positive integer (by Proposition 3.4.3 **(b)**).

Recall that we must prove that $\gcd\left(\binom{n}{i}, \binom{n}{j}\right) > 1$. Indeed, assume the contrary. Hence, $\gcd\left(\binom{n}{i}, \binom{n}{j}\right) \leq 1$, so that $\gcd\left(\binom{n}{i}, \binom{n}{j}\right) = 1$ (since $\gcd\left(\binom{n}{i}, \binom{n}{j}\right)$ is a positive integer). In other words, $\binom{n}{i} \perp \binom{n}{j}$. Combining this with (526), we conclude that $\binom{n}{i} \mid \binom{j}{i}$ (by Theorem 3.5.6, applied to $a = \binom{n}{i}$, $b = \binom{n}{j}$ and $c = \binom{j}{i}$). Hence, Proposition 3.1.3 **(b)** (applied to $a = \binom{n}{i}$ and $b = \binom{j}{i}$) yields $\left|\binom{n}{i}\right| \leq \left|\binom{j}{i}\right|$ (since $\binom{j}{i} \neq 0$). Since $\binom{n}{i} > \binom{j}{i} > 0$, we have $\left|\binom{n}{i}\right| = \binom{n}{i}$ and $\left|\binom{j}{i}\right| = \binom{j}{i}$. Thus,

$$\binom{n}{i} = \left|\binom{n}{i}\right| \leq \left|\binom{j}{i}\right| = \binom{j}{i},$$

which contradicts $\binom{n}{i} > \binom{j}{i}$. This contradiction shows that our assumption was false. Hence, Exercise 4.5.5 is solved. □

A.3.6. Discussion of Exercise 4.5.6

Discussion of Exercise 4.5.6. We shall be very terse; a detailed solution can be found in [19s]. (Exercise 4.5.6 **(a)** is [19s, Exercise 2.14.4], while Exercise 4.5.6 **(b)** is [19s, Exercise 2.14.5].)

(a) Assume that $n > 2$. Let T be the set of all $i \in \{1, 2, \ldots, n\}$ satisfying $i \perp n$. Then, $|T| = \phi(n)$ (by the definition of $\phi(n)$). We thus must prove that $|T|$ is even.

The main idea is to match each $i \in T$ with $n - i$, in a similar vein as we matched every positive divisor of n with its complement in the solution to Exercise 3.8.3. Here is how this works in a bit more detail: We first notice that $n > 1$ (since $n > 2$); hence, n does not satisfy $n \perp n$. Thus, $n \notin T$, so we could just as well have defined T as the set of all $i \in \{1, 2, \ldots, n-1\}$ satisfying $i \perp n$. Furthermore, $n/2 \notin T$ (because we have $n > 2$, and thus, even if $n/2$ is an integer, then $n/2$ is not coprime to n). Hence, each $i \in T$ satisfies $i \neq n/2$. Therefore, we can split the set T into two disjoint subsets X and Y defined by

$$X = \{i \in T \mid i < n/2\} \qquad \text{and} \qquad Y = \{i \in T \mid i > n/2\}.$$

These subsets X and Y are disjoint and their union is $X \cup Y = T$; thus, $|T| = |X| + |Y|$. Hence, if we can show that $|X| = |Y|$, then it will follow that $|T|$ is even (as desired).

In order to show that $|X| = |Y|$, we need to find a bijection from X to Y. But such a bijection is easy to construct: Namely, the two maps

$$X \to Y, \qquad i \mapsto n - i$$

and

$$Y \to X, \qquad i \mapsto n - i$$

are easily seen to be well-defined (because $\gcd(i, n) = \gcd(n - i, n)$ for each $i \in \mathbb{Z}$, and because $i < n/2$ holds if and only if $n - i > n/2$) and mutually inverse (since $n - (n - i) = i$ for each $i \in \mathbb{Z}$); thus, they are bijections. Hence, we conclude that $|X| = |Y|$, so that $|T| = \underbrace{|X|}_{=|Y|} + |Y| = |Y| + |Y| = 2 \cdot |Y|$, and therefore $|T|$ is even. In other words, $\phi(n)$ is even; this solves Exercise 4.5.6 **(a)**.

(b) This can be solved using Gauss's "doubling trick", which we have already seen when solving Exercise 4.1.2. To wit, let us assume that $n > 1$. Let T be the set of all $i \in \{1, 2, \ldots, n\}$ satisfying $i \perp n$. Then, we must prove that $\sum_{i \in T} i = \frac{1}{2} n\phi(n)$. But notice that $n \notin T$ (as we already saw in the solution to Exercise 4.5.6 **(a)** above); thus, the map

$$T \to T, \qquad i \mapsto n - i$$

is easily seen to be well-defined (because $\gcd(i, n) = \gcd(n - i, n)$ for each $i \in \mathbb{Z}$) and therefore a bijection (since it is its own inverse). Hence, we can substitute $n - i$ for i in the sum $\sum_{i \in T} i$. We thus obtain $\sum_{i \in T} i = \sum_{i \in T} (n - i)$. Now, Gauss's "doubling trick" tells us that

$$\begin{aligned} 2 \cdot \sum_{i \in T} i &= \sum_{i \in T} i + \underbrace{\sum_{i \in T} i}_{= \sum_{i \in T} (n-i)} = \sum_{i \in T} i + \sum_{i \in T} (n - i) \\ &= \sum_{i \in T} \underbrace{(i + (n - i))}_{= n} = \sum_{i \in T} n = n|T| = n\phi(n) \end{aligned}$$

(since the definition of $\phi(n)$ yields $|T| = \phi(n)$). Therefore, $\sum_{i \in T} i = \frac{1}{2} n\phi(n)$. This solves Exercise 4.5.6 **(b)**. □

A.3.7. Discussion of Exercise 4.5.7

Discussion of Exercise 4.5.7. Exercise 4.5.7 is [AndTet18, Problem E25]. Here is a sketch of the solution:

Let us fix a positive integer n, and compute the finite sum $S(n) := \sum_{k=2}^{n} \frac{f_k}{f_{k-1} f_{k+1}}$.

Indeed, each $k \geq 1$ satisfies $f_{k+1} = f_k + f_{k-1}$ (by the definition of the Fibonacci sequence) and thus $f_k = f_{k+1} - f_{k-1}$ and therefore

$$\frac{f_k}{f_{k-1}f_{k+1}} = \frac{f_{k+1} - f_{k-1}}{f_{k-1}f_{k+1}} = \frac{1}{f_{k-1}} - \frac{1}{f_{k+1}}. \tag{531}$$

Now,

$$\begin{aligned}
S(n) &= \sum_{k=2}^{n} \underbrace{\frac{f_k}{f_{k-1}f_{k+1}}}_{\substack{= \frac{1}{f_{k-1}} - \frac{1}{f_{k+1}} \\ \text{(by (531))}}} = \sum_{k=2}^{n} \left(\frac{1}{f_{k-1}} - \frac{1}{f_{k+1}} \right) = \sum_{k=2}^{n} \frac{1}{f_{k-1}} - \sum_{k=2}^{n} \frac{1}{f_{k+1}} \\
&= \underbrace{\sum_{k=1}^{n-1} \frac{1}{f_k}}_{= \sum_{k=1}^{n+1} \frac{1}{f_k} - \frac{1}{f_n} - \frac{1}{f_{n+1}}} - \underbrace{\sum_{k=3}^{n+1} \frac{1}{f_k}}_{= \sum_{k=1}^{n+1} \frac{1}{f_k} - \frac{1}{f_1} - \frac{1}{f_2}} \\
&\qquad \left(\begin{array}{c} \text{here, we have substituted } k \text{ for } k-1 \text{ in the first sum,} \\ \text{and substituted } k \text{ for } k+1 \text{ in the second sum} \end{array} \right) \\
&= \left(\sum_{k=1}^{n+1} \frac{1}{f_k} - \frac{1}{f_n} - \frac{1}{f_{n+1}} \right) - \left(\sum_{k=1}^{n+1} \frac{1}{f_k} - \frac{1}{f_1} - \frac{1}{f_2} \right) \\
&= \frac{1}{f_1} + \frac{1}{f_2} - \frac{1}{f_n} - \frac{1}{f_{n+1}}.
\end{aligned} \tag{532}$$

The sum we are looking for is $\sum_{k=2}^{\infty} \frac{f_k}{f_{k-1}f_{k+1}} = \lim_{n\to\infty} S(n)$. But (532) makes this limit easy to compute: As n goes to ∞, the Fibonacci numbers f_n go to ∞ as well[348], and thus $\frac{1}{f_n}$ goes to 0, whence $\frac{1}{f_{n+1}}$ also goes to 0. Now,

$$\begin{aligned}
\sum_{k=2}^{\infty} \frac{f_k}{f_{k-1}f_{k+1}} &= \lim_{n\to\infty} S(n) = \lim_{n\to\infty} \left(\frac{1}{f_1} + \frac{1}{f_2} - \frac{1}{f_n} - \frac{1}{f_{n+1}} \right) \qquad \text{(by (532))} \\
&= \frac{1}{f_1} + \frac{1}{f_2} \qquad \left(\text{since } \frac{1}{f_n} \text{ and } \frac{1}{f_{n+1}} \text{ go to 0 as } n \text{ goes to } \infty \right) \\
&= \frac{1}{1} + \frac{1}{1} \qquad (\text{since } f_1 = 1 \text{ and } f_2 = 1) \\
&= 2.
\end{aligned}$$

[348] There are various ways to see this. For example, it is easy to check that the sequence $(f_2, f_3, f_4, \ldots)$ is a strictly increasing sequence of positive integers, and thus goes to ∞. Alternatively, one can use Theorem 2.3.1 to see not only that f_n goes to ∞ (since $\varphi^n \to \infty$ whereas $\psi^n \to 0$), but also to see how exactly f_n grows (viz., exponentially).

This solves the exercise.

A few remarks about the computation (532) are in order. First, why did we rewrite $\sum_{k=1}^{n-1} \frac{1}{f_k}$ as $\sum_{k=1}^{n+1} \frac{1}{f_k} - \frac{1}{f_n} - \frac{1}{f_{n+1}}$ instead of $\sum_{k=3}^{n-1} \frac{1}{f_k} + \frac{1}{f_1} + \frac{1}{f_2}$? Because the former works for every positive integer n, whereas the latter would only work for $n \geq 3$. Second, there is a reason why we introduced the finite sum $S(n)$ instead of working with the infinite sum $\sum_{k=2}^{\infty} \frac{f_k}{f_{k-1} f_{k+1}}$; indeed, while it is possible to perform a computation analogous to (532) on the level of infinite sums, this would require some nontrivial justification (e.g., in order to rewrite $\sum_{k=2}^{\infty} \left(\frac{1}{f_{k-1}} - \frac{1}{f_{k+1}} \right)$ as $\sum_{k=2}^{\infty} \frac{1}{f_{k-1}} - \sum_{k=2}^{\infty} \frac{1}{f_{k+1}}$, we would need to prove that both sums $\sum_{k=2}^{\infty} \frac{1}{f_{k-1}}$ and $\sum_{k=2}^{\infty} \frac{1}{f_{k+1}}$ converge). In general, infinite sums are not worth the headache when they can be avoided this easily. □

A.3.8. Discussion of Exercise 4.5.8

Discussion of Exercise 4.5.8. Exercise 4.5.8 **(a)** is (part of) [Grinbe15, Exercise 3.3 **(b)**] (and known as the "hockey-stick identity" or the "upper summation formula for binomial coefficients"), whereas Exercise 4.5.8 **(b)** is [18f-hw2s, Exercise 4] (more precisely, it is a generalization of the latter exercise to arbitrary complex n, but the first two solutions of the latter exercise given in [18f-hw2s] are still valid in this generality).

However, you can probably solve both parts by the telescope principle[349] or by induction[350] before you have opened these references. □

A.3.9. Discussion of Exercise 4.5.9

Discussion of Exercise 4.5.9. Here is the idea: Rewrite (16) as

$$f_n = \frac{1}{\sqrt{5}} \left(\frac{1+\sqrt{5}}{2} \right)^n - \frac{1}{\sqrt{5}} \left(\frac{1-\sqrt{5}}{2} \right)^n .$$

Now apply the binomial formula on the right hand side, and cancel the addends with odd powers of $\sqrt{5}$.

Here is the argument in detail:

[349] To wit, part **(a)** uses $\binom{i}{k} = \binom{i+1}{k+1} - \binom{i}{k+1}$, while part **(b)** uses $(-1)^k \binom{n}{k} = (-1)^k \binom{n-1}{k} - (-1)^{k-1} \binom{n-1}{k-1}$. (Both of these equalities follow from Theorem 4.3.7.)

[350] Induction on n for part **(a)**, and induction on m for part **(b)**.

Solution to Exercise 4.5.9. Let $n \in \mathbb{N}$. Let $\varphi = \dfrac{1+\sqrt{5}}{2}$ and $\psi = \dfrac{1-\sqrt{5}}{2}$ be the two solutions of the quadratic equation $X^2 - X - 1 = 0$. Then, Theorem 2.3.1 yields

$$\begin{aligned}
f_n &= \frac{1}{\sqrt{5}}\varphi^n - \frac{1}{\sqrt{5}}\psi^n = \frac{1}{\sqrt{5}}\underbrace{\left(\frac{1+\sqrt{5}}{2}\right)^n}_{=\frac{\left(1+\sqrt{5}\right)^n}{2^n}} - \frac{1}{\sqrt{5}}\underbrace{\left(\frac{1-\sqrt{5}}{2}\right)^n}_{=\frac{\left(1-\sqrt{5}\right)^n}{2^n}} \\
&\qquad \left(\text{since } \varphi = \frac{1+\sqrt{5}}{2} \text{ and } \psi = \frac{1-\sqrt{5}}{2}\right) \\
&= \underbrace{\frac{1}{\sqrt{5}} \cdot \frac{\left(1+\sqrt{5}\right)^n}{2^n}}_{=\frac{1}{\sqrt{5}\cdot 2^n}\cdot\left(1+\sqrt{5}\right)^n} - \underbrace{\frac{1}{\sqrt{5}} \cdot \frac{\left(1-\sqrt{5}\right)^n}{2^n}}_{=\frac{1}{\sqrt{5}\cdot 2^n}\cdot\left(1-\sqrt{5}\right)^n} \\
&= \frac{1}{\sqrt{5}\cdot 2^n} \cdot \left(\underbrace{1+\sqrt{5}}_{=\sqrt{5}+1}\right)^n - \frac{1}{\sqrt{5}\cdot 2^n} \cdot \left(\underbrace{1-\sqrt{5}}_{=-\sqrt{5}+1}\right)^n \\
&= \frac{1}{\sqrt{5}\cdot 2^n} \cdot \underbrace{\left(\sqrt{5}+1\right)^n}_{\substack{=\sum_{k=0}^{n}\binom{n}{k}\left(\sqrt{5}\right)^k 1^{n-k} \\ \text{(by Theorem 4.3.16,} \\ \text{applied to } x=\sqrt{5} \text{ and } y=1)}} - \frac{1}{\sqrt{5}\cdot 2^n} \cdot \underbrace{\left(-\sqrt{5}+1\right)^n}_{\substack{=\sum_{k=0}^{n}\binom{n}{k}\left(-\sqrt{5}\right)^k 1^{n-k} \\ \text{(by Theorem 4.3.16,} \\ \text{applied to } x=-\sqrt{5} \text{ and } y=1)}} \\
&= \frac{1}{\sqrt{5}\cdot 2^n} \cdot \sum_{k=0}^{n}\binom{n}{k}\left(\sqrt{5}\right)^k \underbrace{1^{n-k}}_{=1} - \frac{1}{\sqrt{5}\cdot 2^n} \cdot \sum_{k=0}^{n}\binom{n}{k} \underbrace{\left(-\sqrt{5}\right)^k}_{=(-1)^k\left(\sqrt{5}\right)^k} \underbrace{1^{n-k}}_{=1} \\
&= \frac{1}{\sqrt{5}\cdot 2^n} \cdot \sum_{k=0}^{n}\binom{n}{k}\left(\sqrt{5}\right)^k - \frac{1}{\sqrt{5}\cdot 2^n} \cdot \sum_{k=0}^{n}\binom{n}{k}(-1)^k\left(\sqrt{5}\right)^k \\
&= \frac{1}{\sqrt{5}\cdot 2^n} \cdot \left(\sum_{k=0}^{n}\binom{n}{k}\left(\sqrt{5}\right)^k - \sum_{k=0}^{n}\binom{n}{k}(-1)^k\left(\sqrt{5}\right)^k\right). \qquad (533)
\end{aligned}$$

But

$$\begin{aligned}
&\sum_{k=0}^{n}\binom{n}{k}\left(\sqrt{5}\right)^{k}-\sum_{k=0}^{n}\binom{n}{k}(-1)^{k}\left(\sqrt{5}\right)^{k}\\
&=\sum_{k=0}^{n}\underbrace{\left(\binom{n}{k}\left(\sqrt{5}\right)^{k}-\binom{n}{k}(-1)^{k}\left(\sqrt{5}\right)^{k}\right)}_{=\binom{n}{k}\left(1-(-1)^{k}\right)\left(\sqrt{5}\right)^{k}}\\
&=\sum_{k=0}^{n}\binom{n}{k}\left(1-(-1)^{k}\right)\left(\sqrt{5}\right)^{k}.
\end{aligned}$$

Comparing this with

$$\begin{aligned}
&\sum_{k=0}^{2n+1}\binom{n}{k}\left(1-(-1)^{k}\right)\left(\sqrt{5}\right)^{k}\\
&=\sum_{k=0}^{n}\binom{n}{k}\left(1-(-1)^{k}\right)\left(\sqrt{5}\right)^{k}+\sum_{k=n+1}^{2n+1}\underbrace{\binom{n}{k}}_{\substack{=0\\ \text{(by Proposition 4.3.4}\\ \text{(since } k\geq n+1>n))}}\left(1-(-1)^{k}\right)\left(\sqrt{5}\right)^{k}\\
&\qquad\text{(here, we have split the sum at } n\text{, using (84))}\\
&=\sum_{k=0}^{n}\binom{n}{k}\left(1-(-1)^{k}\right)\left(\sqrt{5}\right)^{k}+\underbrace{\sum_{k=n+1}^{2n+1}0\left(1-(-1)^{k}\right)\left(\sqrt{5}\right)^{k}}_{=0}\\
&=\sum_{k=0}^{n}\binom{n}{k}\left(1-(-1)^{k}\right)\left(\sqrt{5}\right)^{k},
\end{aligned}$$

we obtain

$$\begin{aligned}
&\sum_{k=0}^{n}\binom{n}{k}\left(\sqrt{5}\right)^{k}-\sum_{k=0}^{n}\binom{n}{k}(-1)^{k}\left(\sqrt{5}\right)^{k}\\
&=\underbrace{\sum_{k=0}^{2n+1}}_{=\sum_{k\in\{0,1,\ldots,2n+1\}}}\binom{n}{k}\left(1-(-1)^{k}\right)\left(\sqrt{5}\right)^{k}\\
&=\sum_{k\in\{0,1,\ldots,2n+1\}}\binom{n}{k}\left(1-(-1)^{k}\right)\left(\sqrt{5}\right)^{k}\\
&=\sum_{\substack{k\in\{0,1,\ldots,2n+1\};\\ k\text{ is even}}}\binom{n}{k}\left(1-\underbrace{(-1)^{k}}_{\substack{=1\\ (\text{since } k \text{ is even})}}\right)\left(\sqrt{5}\right)^{k}+\sum_{\substack{k\in\{0,1,\ldots,2n+1\};\\ k\text{ is odd}}}\binom{n}{k}\left(1-\underbrace{(-1)^{k}}_{\substack{=-1\\ (\text{since } k \text{ is odd})}}\right)\left(\sqrt{5}\right)^{k}\\
&\qquad(\text{here, we used Theorem 4.1.20})\\
&=\sum_{\substack{k\in\{0,1,\ldots,2n+1\};\\ k\text{ is even}}}\binom{n}{k}\underbrace{(1-1)}_{=0}\left(\sqrt{5}\right)^{k}+\sum_{\substack{k\in\{0,1,\ldots,2n+1\};\\ k\text{ is odd}}}\binom{n}{k}\underbrace{(1-(-1))}_{=2}\left(\sqrt{5}\right)^{k}\\
&=\underbrace{\sum_{\substack{k\in\{0,1,\ldots,2n+1\};\\ k\text{ is even}}}\binom{n}{k}\cdot 0\left(\sqrt{5}\right)^{k}}_{=0}+\sum_{\substack{k\in\{0,1,\ldots,2n+1\};\\ k\text{ is odd}}}\binom{n}{k}\cdot 2\left(\sqrt{5}\right)^{k}\\
&=\sum_{\substack{k\in\{0,1,\ldots,2n+1\};\\ k\text{ is odd}}}\binom{n}{k}\cdot 2\left(\sqrt{5}\right)^{k}=2\sum_{\substack{k\in\{0,1,\ldots,2n+1\};\\ k\text{ is odd}}}\binom{n}{k}\left(\sqrt{5}\right)^{k}\\
&=2\sum_{i\in\{0,1,\ldots,n\}}\binom{n}{2i+1}\left(\sqrt{5}\right)^{2i+1}
\end{aligned}\tag{534}$$

(here, we have substituted $2i+1$ for k in the sum, since the map

$$\begin{aligned}\{0,1,\ldots,n\}&\to\{k\in\{0,1,\ldots,2n+1\}\mid k\text{ is odd}\},\\ i&\mapsto 2i+1\end{aligned}$$

is a bijection[351]).

[351] This is just saying that the odd integers in $\{0,1,\ldots,2n+1\}$ are $1,3,5,\ldots,2n+1$ (and these numbers are all distinct).

Hence, (533) becomes

$$\begin{aligned}
f_n &= \frac{1}{\sqrt{5}\cdot 2^n}\cdot \underbrace{\left(\sum_{k=0}^{n}\binom{n}{k}\left(\sqrt{5}\right)^k - \sum_{k=0}^{n}\binom{n}{k}(-1)^k\left(\sqrt{5}\right)^k\right)}_{\substack{=2\sum\limits_{i\in\{0,1,\ldots,n\}}\binom{n}{2i+1}\left(\sqrt{5}\right)^{2i+1} \\ \text{(by (534))}}} \\
&= \frac{1}{\sqrt{5}\cdot 2^n}\cdot 2\sum_{i\in\{0,1,\ldots,n\}}\binom{n}{2i+1}\left(\sqrt{5}\right)^{2i+1} \\
&= \underbrace{\frac{2}{2^n}}_{=\frac{1}{2^{n-1}}}\underbrace{\sum_{i\in\{0,1,\ldots,n\}}}_{=\sum\limits_{i=0}^{n}}\binom{n}{2i+1}\underbrace{\frac{\left(\sqrt{5}\right)^{2i+1}}{\sqrt{5}}}_{=\left(\sqrt{5}\right)^{2i}=\left(\left(\sqrt{5}\right)^2\right)^i=5^i} \\
&= \frac{1}{2^{n-1}}\sum_{i=0}^{n}\binom{n}{2i+1}5^i = \frac{1}{2^{n-1}}\sum_{k=0}^{n}\binom{n}{2k+1}5^k
\end{aligned}$$

(here, we have renamed the summation index i as k). Multiplying both sides of this equality by 2^{n-1}, we obtain

$$2^{n-1}\cdot f_n = \sum_{k=0}^{n}\binom{n}{2k+1}5^k.$$

This solves Exercise 4.5.9. □

□

A.3.10. Discussion of Exercise 4.5.10

Discussion of Exercise 4.5.10. The quickest way to prove part **(a)** is by defining the "Fibonacci factorial" $m!_F$ of any $m \in \mathbb{N}$ to be the product $f_1 f_2 \cdots f_m$, and then showing that if $k \leq n$, then $\dbinom{n}{k}_F = \dfrac{n!_F}{k!_F\,(n-k)!_F}$ (an analogue of Theorem 4.3.8). Parts **(b)** and **(c)** are [18f-hw3s, Exercise 4].

The numbers $\dbinom{n}{k}_F$ are known as *Fibonomial coefficients* (entering these words in Google Scholar will produce a noticeable amount of literature), and are one of several analogues of binomial coefficients known.

A combinatorial solution to Exercise 4.5.10 **(c)** can be found in [BenPlo08].

For the sake of completeness, let us give a detailed solution to Exercise 4.5.10 **(a)** (while detailed solutions to parts **(b)** and **(c)** can be found in [18f-hw3s, Exercise 4]):

Solution to Exercise 4.5.10 ***(a)***. For each $m \in \mathbb{N}$, define the positive integer $m!_F$ by

$$m!_F = f_1 f_2 \cdots f_m. \tag{535}$$

(This is indeed a positive integer, because it is easy to see that the Fibonacci numbers $f_1, f_2, \ldots, f_m$ all are positive[352].)

We first show the following auxiliary fact:

Claim 1: Let $n \in \mathbb{N}$ and $k \in \mathbb{N}$ satisfy $k \leq n$. Then,

$$\binom{n}{k}_F = \frac{n!_F}{k!_F\,(n-k)!_F}.$$

[*Proof of Claim 1:* We have $k \leq n$, thus $n \geq k \geq 0$ (since $k \in \mathbb{N}$). The definition of $\binom{n}{k}_F$ yields

$$\begin{aligned}\binom{n}{k}_F &= \begin{cases} \dfrac{f_n f_{n-1} \cdots f_{n-k+1}}{f_k f_{k-1} \cdots f_1}, & \text{if } n \geq k \geq 0; \\ 0, & \text{otherwise} \end{cases} \\ &= \frac{f_n f_{n-1} \cdots f_{n-k+1}}{f_k f_{k-1} \cdots f_1} \qquad (\text{since } n \geq k \geq 0)\,.\end{aligned}$$

In view of

$$\begin{aligned} f_n f_{n-1} \cdots f_{n-k+1} &= \underbrace{(f_n f_{n-1} \cdots f_1)}_{\substack{=f_1 f_2 \cdots f_n \\ =n!_F \\ (\text{since } n!_F \text{ was defined to be } f_1 f_2 \cdots f_n)}} \Big/ \underbrace{(f_{n-k} f_{n-k-1} \cdots f_1)}_{\substack{=f_1 f_2 \cdots f_{n-k} \\ =(n-k)!_F \\ (\text{since } (n-k)!_F \text{ was defined to be } f_1 f_2 \cdots f_{n-k})}} \\ &\qquad (\text{since } (f_n f_{n-1} \cdots f_{n-k+1}) \cdot (f_{n-k} f_{n-k-1} \cdots f_1) = f_n f_{n-1} \cdots f_1) \\ &= n!_F / (n-k)!_F \end{aligned}$$

and

$$f_k f_{k-1} \cdots f_1 = f_1 f_2 \cdots f_k = k!_F \qquad (\text{since } k!_F \text{ was defined to be } f_1 f_2 \cdots f_k)\,,$$

this rewrites as

$$\binom{n}{k}_F = \frac{n!_F / (n-k)!_F}{k!_F} = \frac{n!_F}{k!_F\,(n-k)!_F}.$$

This proves Claim 1.]

Now, let $n \in \mathbb{N}$ and $k \in \mathbb{N}$. We must prove that

$$\binom{n}{k}_F = \binom{n}{n-k}_F. \tag{536}$$

[352] because the Fibonacci sequence $(f_0, f_1, f_2, \ldots)$ is weakly increasing (this is easily seen from its recursive definition) and its entry $f_1 = 1$ is already positive

We are in one of the following two cases:

Case 1: We have $k \leq n$.

Case 2: We have $k > n$.

Let us first consider Case 1. In this case, we have $k \leq n$. Hence, $n - k \in \mathbb{N}$. Also, $n - k \leq n$ (since $k \geq 0$). Now, Claim 1 yields

$$\binom{n}{k}_F = \frac{n!_F}{k!_F\,(n-k)!_F} \qquad (\text{since } k \leq n)\,.$$

On the other hand, Claim 1 (applied to $n - k$ instead of k) yields

$$\begin{aligned}
\binom{n}{n-k}_F &= \frac{n!_F}{(n-k)!_F\,(n-(n-k))!_F} \qquad (\text{since } n-k \in \mathbb{N} \text{ and } n-k \leq n) \\
&= \frac{n!_F}{(n-k)!_F k!_F} \qquad (\text{since } n-(n-k) = k) \\
&= \frac{n!_F}{k!_F\,(n-k)!_F}.
\end{aligned}$$

Comparing these two equalities, we obtain $\binom{n}{k}_F = \binom{n}{n-k}_F$. This proves (536) in Case 1.

Let us now consider Case 2. In this case, we have $k > n$. Hence, we don't have $n \geq k \geq 0$ (since $n \geq k$ would contradict $k > n$). Also, $n - k < 0$ (since $k > n$). Hence, we don't have $n \geq n - k \geq 0$ (since $n - k \geq 0$ would contradict $n - k < 0$). The definition of $\binom{n}{k}_F$ yields

$$\begin{aligned}
\binom{n}{k}_F &= \begin{cases} \dfrac{f_n f_{n-1} \cdots f_{n-k+1}}{f_k f_{k-1} \cdots f_1}, & \text{if } n \geq k \geq 0; \\ 0, & \text{otherwise} \end{cases} \\
&= 0 \qquad (\text{since we don't have } n \geq k \geq 0)\,.
\end{aligned}$$

On the other hand, the definition of $\binom{n}{n-k}_F$ yields

$$\begin{aligned}
\binom{n}{n-k}_F &= \begin{cases} \dfrac{f_n f_{n-1} \cdots f_{n-(n-k)+1}}{f_{n-k} f_{(n-k)-1} \cdots f_1}, & \text{if } n \geq n-k \geq 0; \\ 0, & \text{otherwise} \end{cases} \\
&= 0 \qquad (\text{since we don't have } n \geq n-k \geq 0)\,.
\end{aligned}$$

Comparing these two equalities, we find $\binom{n}{k}_F = \binom{n}{n-k}_F$. This proves (536) in Case 2.

We have now proved (536) in each of the two Cases 1 and 2. Thus, (536) always holds. This solves Exercise 4.5.10 **(a)**. □

□

A.4. Homework set #3 discussion

The following are discussions of the problems on homework set #3 (Section 4.8).

A.4.1. Discussion of Exercise 4.8.1

Discussion of Exercise 4.8.1. We shall first prove part **(a)** of Exercise 4.8.1. Then, we will show a general result (Theorem A.4.1) that will allow us to derive parts **(b)** and **(c)** from part **(a)**.

Recall that

$$a_n = 1 + a_{n-1}a_{n-2} \qquad \text{for each integer } n \geq 2. \tag{537}$$

(a) Fix $n \in \mathbb{N}$. We claim that

$$a_{k+n} \equiv a_k \bmod a_n \qquad \text{for each } k \in \mathbb{N}. \tag{538}$$

[*Proof of (538):* We proceed by strong induction on k:

Induction step: Let $p \in \mathbb{N}$. Assume (as the induction hypothesis) that (538) holds for $k < p$. We must now prove that (538) holds for $k = p$. In other words, we must prove that $a_{p+n} \equiv a_p \bmod a_n$.

If $n = 0$, then this is obvious (because if $n = 0$, then $p + \underbrace{n}_{=0} = p$ and thus $a_{p+n} = a_p \equiv a_p \bmod a_n$). Hence, for the rest of this proof, we WLOG assume that $n \neq 0$. Hence, $n \geq 1$ (since $n \in \mathbb{N}$).

We have $p \in \mathbb{N}$. Hence, we are in one of the following three cases:

Case 1: We have $p = 0$.

Case 2: We have $p = 1$.

Case 3: We have $p \geq 2$.

Let us first consider Case 1. In this case, we have $p = 0$. Thus, $a_p = a_0 = 0$. On the other hand, $\underbrace{p}_{=0} + n = 0 + n = n$, so that $a_{p+n} = a_n \equiv 0 \bmod a_n$. This rewrites as $a_{p+n} \equiv a_p \bmod a_n$ (since $a_p = 0$). Thus, $a_{p+n} \equiv a_p \bmod a_n$ is proved in Case 1.

Let us next consider Case 2. In this case, we have $p = 1$. Thus, $a_p = a_1 = 1$. On the other hand, $\underbrace{p}_{=1} + n = 1 + n = n + 1$, so that $a_{p+n} = a_{n+1}$. Note that $n + 1 \geq 2$ (since $n \geq 1$). Hence, (537) (applied to $n + 1$ instead of n) yields

$$a_{n+1} = 1 + a_{(n+1)-1}a_{(n+1)-2} = 1 + \underbrace{a_n}_{\equiv 0 \bmod a_n} a_{n-1} \equiv 1 + 0a_{n-1} = 1 \bmod a_n.$$

In view of $a_{p+n} = a_{n+1}$ and $a_p = 1$, this rewrites as $a_{p+n} \equiv a_p \bmod a_n$. Thus, $a_{p+n} \equiv a_p \bmod a_n$ is proved in Case 2.

At last, let us consider Case 3. In this case, we have $p \geq 2$. Hence, $p - 1 \in \mathbb{N}$ and $p - 2 \in \mathbb{N}$.

We have assumed that (538) holds for $k < p$. Hence, we can apply (538) to $k = p - 1$ (since $p - 1 \in \mathbb{N}$ and $p - 1 < p$). Thus, we obtain

$$a_{(p-1)+n} \equiv a_{p-1} \bmod a_n. \tag{539}$$

We have assumed that (538) holds for $k < p$. Hence, we can apply (538) to $k = p - 2$ (since $p - 2 \in \mathbb{N}$ and $p - 2 < p$). Thus, we obtain

$$a_{(p-2)+n} \equiv a_{p-2} \bmod a_n. \tag{540}$$

However, $p \geq 2$. Hence, (537) (applied to p instead of n) yields

$$a_p = 1 + a_{p-1}a_{p-2}. \tag{541}$$

Also, $n \geq 0$, so that $p + n \geq p \geq 2$. Hence, (537) (applied to $p + n$ instead of n) yields

$$\begin{aligned} a_{p+n} &= 1 + a_{(p+n)-1}a_{(p+n)-2} = 1 + \underbrace{a_{(p-1)+n}}_{\substack{\equiv a_{p-1} \bmod a_n \\ \text{(by (539))}}} \underbrace{a_{(p-2)+n}}_{\substack{\equiv a_{p-2} \bmod a_n \\ \text{(by (540))}}} \\ &\qquad (\text{since } (p+n)-1 = (p-1)+n \text{ and } (p+n)-2 = (p-2)+n) \\ &\equiv 1 + a_{p-1}a_{p-2} = a_p \bmod a_n \qquad (\text{by (541)}). \end{aligned}$$

Thus, $a_{p+n} \equiv a_p \bmod a_n$ is proved in Case 3.

We have now proved $a_{p+n} \equiv a_p \bmod a_n$ in each of the three Cases 1, 2 and 3. Hence, $a_{p+n} \equiv a_p \bmod a_n$ always holds. In other words, (538) holds for $k = p$. This completes the induction step. Thus, (538) is proven.]

This solves Exercise 4.8.1 **(a)**.

Let us now forget the sequence $(a_0, a_1, a_2, \ldots)$ defined in Exercise 4.8.1. Instead, we shall prove a general property of a wider class of sequences:

Theorem A.4.1. Let $(a_0, a_1, a_2, \ldots)$ be **any** sequence of integers. Assume that we have

$$a_{k+n} \equiv a_k \bmod a_n \qquad \text{for any } k \in \mathbb{N} \text{ and } n \in \mathbb{N}. \tag{542}$$

Then:

(a) If $u, v \in \mathbb{N}$ satisfy $u \mid v$, then $a_u \mid a_v$.

(b) For any $n, m \in \mathbb{N}$, we have $\gcd(a_n, a_m) = \left|a_{\gcd(n,m)}\right|$.

Proof of Theorem A.4.1. We shall prove Theorem A.4.1 **(b)** first, since Theorem A.4.1 **(a)** follows easily from it. Of course, it is not much harder to prove Theorem A.4.1 **(a)** independently, by strong induction on v (or induction on v/u).

(b) This will be a calque of our solution to Exercise 3.4.1 **(b)**. Why innovate when you can just imitate?

We shall prove Theorem A.4.1 **(b)** by strong induction on $n + m$:

Induction step: Let $k \in \mathbb{N}$. Assume (as the induction hypothesis) that Theorem A.4.1 **(b)** is true for $n + m < k$. We must prove that Theorem A.4.1 **(b)** is true for $n + m = k$.

So let $n, m \in \mathbb{N}$ be such that $n + m = k$. We must show that $\gcd(a_n, a_m) = \left|a_{\gcd(n,m)}\right|$.

Note that n and m play symmetric roles in this claim[353], and thus can be swapped

[353] because Proposition 3.4.4 **(b)** yields $\gcd(a_n, a_m) = \gcd(a_m, a_n)$ and $\gcd(n, m) = \gcd(m, n)$

at will. By swapping n and m if necessary, we can ensure that $n \leq m$. Hence, we WLOG assume that $n \leq m$. Thus, $m - n \in \mathbb{N}$.

It is easy to see that our claim $\gcd(a_n, a_m) = \left|a_{\gcd(n,m)}\right|$ holds if $n = 0$ [354]. Thus, we are done if $n = 0$. Hence, we WLOG assume that $n \neq 0$. Therefore, $n > 0$ (since $n \in \mathbb{N}$). Thus, $n + m > m$, so that $m < n + m = k$.

But our induction hypothesis says that Theorem A.4.1 **(b)** is true for $n + m < k$. Hence, we can apply Theorem A.4.1 **(b)** to $m - n$ instead of m (since $m - n \in \mathbb{N}$ and $n + (m - n) = m < k$). We thus obtain

$$\gcd(a_n, a_{m-n}) = \left|a_{\gcd(n,m-n)}\right|. \tag{543}$$

But we have $m - \underbrace{n}_{\equiv 0 \bmod n} \equiv m \bmod n$, and thus $\gcd(n, m-n) = \gcd(n, m)$ (by Proposition 3.4.4 **(d)**, applied to $a = n$, $b = m - n$ and $c = m$). Furthermore, (542) (applied to $m - n$ instead of k) yields $a_{(m-n)+n} \equiv a_{m-n} \bmod a_n$ (since $m - n \in \mathbb{N}$). In other words, $a_m \equiv a_{m-n} \bmod a_n$ (since $(m - n) + n = m$). Hence, Proposition 3.4.4 **(d)** (applied to a_n, a_m and a_{m-n} instead of a, b and c) yields

$$\begin{aligned} \gcd(a_n, a_m) &= \gcd(a_n, a_{m-n}) = \left|a_{\gcd(n,m-n)}\right| \qquad \text{(by (543))} \\ &= \left|a_{\gcd(n,m)}\right| \qquad \left(\text{since } \gcd(n, m-n) = \gcd(n, m)\right). \end{aligned}$$

Now, forget that we fixed n, m. We thus have shown that any $n, m \in \mathbb{N}$ satisfying $n + m = k$ satisfy $\gcd(a_n, a_m) = \left|a_{\gcd(n,m)}\right|$. In other words, Theorem A.4.1 **(b)** is true for $n + m = k$. This completes the induction step. Thus, Theorem A.4.1 **(b)** is proved.

(a) Let us now prove Theorem A.4.1 **(a)**. This is easy after part **(b)** has already been shown:

Let $u, v \in \mathbb{N}$ be such that $u \mid v$. We must show that $a_u \mid a_v$.

We have $u \mid v$. Thus, Proposition 3.4.4 **(i)** (applied to $a = u$ and $b = v$) yields $\gcd(u, v) = |u| = u$ (since $u \geq 0$). But Theorem A.4.1 **(b)** (applied to $n = u$ and

[354] *Proof.* Assume that $n = 0$. Thus, $a_n = a_0 = 0$. But Proposition 3.4.4 **(a)** (applied to m instead of a) yields $\gcd(m, 0) = \gcd(m) = |m| = m$ (since $m \geq 0$). But Proposition 3.4.4 **(b)** yields $\gcd(n, m) = \gcd\left(m, \underbrace{n}_{=0}\right) = \gcd(m, 0) = m$. Thus, $m = \gcd(n, m)$.

Also, Proposition 3.4.4 **(a)** (applied to a_m instead of a) yields $\gcd(a_m, 0) = \gcd(a_m) = |a_m|$. Now, Proposition 3.4.4 **(b)** yields

$$\begin{aligned} \gcd(a_n, a_m) &= \gcd(a_m, a_n) = \gcd(a_m, 0) \qquad (\text{since } a_n = 0) \\ &= |a_m| = \left|a_{\gcd(n,m)}\right| \qquad (\text{since } m = \gcd(n, m)), \end{aligned}$$

qed.

$m = v$) yields

$$\begin{aligned} \gcd(a_u, a_v) &= \left|a_{\gcd(u,v)}\right| = |a_u| \qquad (\text{since } \gcd(u,v) = u) \\ &= \pm a_u, \end{aligned}$$

so that $a_u \mid \gcd(a_u, a_v)$. However, Proposition 3.4.4 **(f)** (applied to $a = a_u$ and $b = a_v$) yields $\gcd(a_u, a_v) \mid a_u$ and $\gcd(a_u, a_v) \mid a_v$. Hence, $a_u \mid \gcd(a_u, a_v) \mid a_v$. This proves Theorem A.4.1 **(a)**. □

Let us now resume the solution of Exercise 4.8.1. Consider again the sequence $(a_0, a_1, a_2, \ldots)$ defined in Exercise 4.8.1. We claim that

$$a_n \geq 0 \qquad \text{for each } n \in \mathbb{N}. \tag{544}$$

[*Proof of (544):* This is straightforward by now. We proceed by strong induction on n:

Induction step: Let $k \in \mathbb{N}$. Assume (as the induction hypothesis) that (544) holds for $n < k$. We must prove that (544) holds for $n = k$. In other words, we must prove that $a_k \geq 0$.

This is clearly true if $k = 0$ (since $a_0 = 0 \geq 0$), and also is clearly true if $k = 1$ (since $a_1 = 1 \geq 0$). Hence, for the rest of this proof, we can WLOG assume that k equals neither 0 nor 1. Hence, $k \geq 2$ (since $k \in \mathbb{N}$). Therefore, $k - 2 \in \mathbb{N}$ and $k - 1 \in \mathbb{N}$.

But we assumed that (544) holds for $n < k$. Hence, we can apply (544) to $n = k - 2$ (since $k - 2 \in \mathbb{N}$ and $k - 2 < k$). Thus we obtain $a_{k-2} \geq 0$. Similarly, $a_{k-1} \geq 0$.

However, $k \geq 2$. Thus, (537) (applied to $n = k$) yields $a_k = 1 + \underbrace{a_{k-1}}_{\geq 0} \underbrace{a_{k-2}}_{\geq 0} \geq 1 \geq 0$.

This completes our induction step. Thus, (537) is proved.]

Now, we can solve parts **(b)** and **(c)** of Exercise 4.8.1:

(b) We have $a_{k+n} \equiv a_k \bmod a_n$ for any $k \in \mathbb{N}$ and $n \in \mathbb{N}$ (by Exercise 4.8.1 **(a)**). Hence, Theorem A.4.1 **(a)** shows that if $u, v \in \mathbb{N}$ satisfy $u \mid v$, then $a_u \mid a_v$. This solves Exercise 4.8.1 **(b)**.

(c) We have $a_{k+n} \equiv a_k \bmod a_n$ for any $k \in \mathbb{N}$ and $n \in \mathbb{N}$ (by Exercise 4.8.1 **(a)**). Hence, Theorem A.4.1 **(b)** shows that for any $n, m \in \mathbb{N}$, we have

$$\gcd(a_n, a_m) = \left|a_{\gcd(n,m)}\right| = a_{\gcd(n,m)}$$

(since $a_{\gcd(n,m)} \geq 0$ (by (544), applied to $\gcd(n,m)$ instead of n)). This solves Exercise 4.8.1 **(c)**. □

A.4.2. Discussion of Exercise 4.8.2

Discussion of Exercise 4.8.2. Exercise 4.8.2 is a result of Dirichlet (published in 1849). Let us give a solution using manipulation of sums.

We will need the following fact:

Proposition A.4.2. Let $n \in \mathbb{N}$, and let b be a positive integer. Then,

$$\sum_{\substack{k \in \{1,2,\ldots,n\}; \\ b \mid k}} 1 = \left\lfloor \frac{n}{b} \right\rfloor . \tag{545}$$

Proof of Proposition A.4.2 (sketched). Clearly, the sum $\sum_{\substack{k \in \{1,2,\ldots,n\}; \\ b \mid k}} 1$ equals the number of all $k \in \{1, 2, \ldots, n\}$ that are multiples of b (by (60)). In other words, this sum equals the number of all multiples of b in the set $\{1, 2, \ldots, n\}$. But these multiples are precisely $1b, 2b, 3b, \ldots, (n//b)\, b$. Hence, their number is $n//b$ (see Definition 3.3.1 **(a)** for the meaning of this notation). Thus, we obtain $\sum_{\substack{k \in \{1,2,\ldots,n\}; \\ b \mid k}} 1 = n//b = \left\lfloor \frac{n}{b} \right\rfloor$ (by Proposition 3.3.5, applied to n and b instead of u and n), so Proposition A.4.2 is proven. (See [Grinbe16, Proposition 1.1.11] for a detailed proof.) □

The next lemma is even simpler:

Lemma A.4.3. Let n be a positive integer. Let $k \in \{1, 2, \ldots, n\}$. Then,

$$\sum_{\substack{b \in \{1,2,\ldots,n\}; \\ b \mid k}} 1 = d(k) . \tag{546}$$

Proof of Lemma A.4.3. Using (60), we see that

$$\begin{aligned} &\sum_{\substack{b \in \{1,2,\ldots,n\}; \\ b \mid k}} 1 \\ &= (\text{the number of all } b \in \{1, 2, \ldots, n\} \text{ satisfying } b \mid k) \cdot 1 \\ &= (\text{the number of all } b \in \{1, 2, \ldots, n\} \text{ satisfying } b \mid k) . \end{aligned} \tag{547}$$

However, the numbers $b \in \{1, 2, \ldots, n\}$ satisfying $b \mid k$ are precisely the positive

divisors of k [355]. Hence,

$$\begin{aligned}&(\text{the number of all } b \in \{1,2,\ldots,n\} \text{ satisfying } b \mid k)\\ &= (\text{the number of all positive divisors of } k) = d(k)\end{aligned}$$

(since $d(k)$ was defined to be the number of all positive divisors of k). Thus, (547) becomes

$$\sum_{\substack{b\in\{1,2,\ldots,n\};\\ b\mid k}} 1 = (\text{the number of all } b \in \{1,2,\ldots,n\} \text{ satisfying } b \mid k) = d(k).$$

This proves Lemma A.4.3. □

Now,

$$\begin{aligned}&\left\lfloor \frac{n}{1} \right\rfloor + \left\lfloor \frac{n}{2} \right\rfloor + \cdots + \left\lfloor \frac{n}{n} \right\rfloor\\ &= \sum_{b\in\{1,2,\ldots,n\}} \underbrace{\left\lfloor \frac{n}{b} \right\rfloor}_{\substack{= \sum_{k\in\{1,2,\ldots,n\};\, b\mid k} 1 \\ \text{(by (545))}}} = \underbrace{\sum_{b\in\{1,2,\ldots,n\}} \sum_{\substack{k\in\{1,2,\ldots,n\};\\ b\mid k}}}_{\substack{= \sum_{k\in\{1,2,\ldots,n\}} \sum_{\substack{b\in\{1,2,\ldots,n\};\\ b\mid k}}}} 1\\ &\qquad \text{(here, we have used Theorem 4.1.25 to interchange the two summation signs)}\\ &= \sum_{k\in\{1,2,\ldots,n\}} \underbrace{\sum_{\substack{b\in\{1,2,\ldots,n\};\\ b\mid k}} 1}_{\substack{=d(k)\\ \text{(by Lemma A.4.3)}}} = \sum_{k\in\{1,2,\ldots,n\}} d(k) = d(1) + d(2) + \cdots + d(n).\end{aligned}$$

This solves Exercise 4.8.2. □

A.4.3. Discussion of Exercise 4.8.3

Discussion of Exercise 4.8.3. **(a)** The answer is

$$\sum_{k=0}^{n} k \binom{n}{k} = n \cdot 2^{n-1}. \tag{548}$$

[355] *Proof.* It is clear that each number $b \in \{1,2,\ldots,n\}$ satisfying $b \mid k$ is a positive divisor of k. Thus, we only need to prove the converse statement – i.e., we need to prove that each positive divisor of k is a $b \in \{1,2,\ldots,n\}$ satisfying $b \mid k$. In other words, we need to show that if b is a positive divisor of k, then $b \in \{1,2,\ldots,n\}$ and $b \mid k$.

So let us do this. Let b be a positive divisor of k. We must show that $b \in \{1,2,\ldots,n\}$ and $b \mid k$.

We have $b \mid k$ (since b is a divisor of k) and $k \neq 0$ (since $k \in \{1,2,\ldots,n\}$). Hence, Proposition 3.1.3 **(b)** (applied to b and k instead of a and b) yields $|b| \leq |k|$. Hence, $|k| \geq |b| = b$ (since b is positive). Also, k is positive (since $k \in \{1,2,\ldots,n\}$), so that $|k| = k$. Hence, $k = |k| \geq b$, thus $b \leq k \leq n$ (since $k \in \{1,2,\ldots,n\}$). Since b is a positive integer, we thus conclude that $b \in \{1,2,\ldots,n\}$. Thus, we have shown that $b \in \{1,2,\ldots,n\}$ and $b \mid k$. As explained above, this completes our proof.

The formula (548) appears twice in [19fco]: once as [19fco, Exercise 1.3.6] (where two proofs are given) and once again as [19fco, Corollary 1.6.5] (with a different, combinatorial proof). We shall briefly outline the first two proofs as well as another using derivatives of polynomials:

First proof of (548) (sketched). We WLOG assume that $n \neq 0$ (since the case $n = 0$ is easily checked by hand). Hence, $n \geq 1$, so that $n - 1 \in \mathbb{N}$. Now,

$$\sum_{k=0}^{n} k\binom{n}{k} = \underbrace{0\binom{n}{0}}_{=0} + \sum_{k=1}^{n} \underbrace{k}_{=n\cdot\frac{k}{n}}\binom{n}{k} = \sum_{k=1}^{n} n \cdot \underbrace{\frac{k}{n}\binom{n}{k}}_{\substack{=\binom{n-1}{k-1} \\ \text{(by Exercise 4.5.4 \textbf{(a)}, } \\ \text{applied to } m=k)}}$$

$$= \sum_{k=1}^{n} n \cdot \binom{n-1}{k-1} = n \sum_{k=1}^{n} \binom{n-1}{k-1} = n \underbrace{\sum_{k=0}^{n-1} \binom{n-1}{k}}_{\substack{=2^{n-1} \\ \text{(by Corollary 4.3.17,} \\ \text{applied to } n-1 \text{ instead of } n)}}$$

$$\text{(here, we have substituted } k \text{ for } k-1 \text{ in the sum)}$$

$$= n \cdot 2^{n-1}.$$

This proves (548). □

Second proof of (548) (sketched). We use Gauss's "doubling trick" as in the solution to Exercise 4.1.2: We have

$$2 \cdot \sum_{k=0}^{n} k\binom{n}{k} = \sum_{k=0}^{n} k\binom{n}{k} + \sum_{k=0}^{n} k\binom{n}{k} = \sum_{k=0}^{n} k\binom{n}{k} + \sum_{k=0}^{n} (n-k) \underbrace{\binom{n}{n-k}}_{\substack{=\binom{n}{k} \\ \text{(by Theorem 4.3.10)}}}$$

$$\text{(here, we have substituted } n-k \text{ for } k \text{ in the second sum)}$$

$$= \sum_{k=0}^{n} k\binom{n}{k} + \sum_{k=0}^{n} (n-k)\binom{n}{k} = \sum_{k=0}^{n} \underbrace{\left(k\binom{n}{k} + (n-k)\binom{n}{k}\right)}_{=(k+(n-k))\binom{n}{k}}$$

$$= \sum_{k=0}^{n} \underbrace{(k+(n-k))}_{=n}\binom{n}{k} = \sum_{k=0}^{n} n\binom{n}{k} = n \underbrace{\sum_{k=0}^{n} \binom{n}{k}}_{\substack{=2^n \\ \text{(by Corollary 4.3.17)}}} = n \cdot 2^n.$$

Dividing both sides of this equality by 2, we obtain (548). This proves (548) again. □

Third proof of (548) (sketched). Consider polynomials in a single indeterminate x over the real numbers. Applying Theorem 4.3.16 to $y = 1$, we obtain

$$(x+1)^n = \sum_{k=0}^{n} \binom{n}{k} x^k \underbrace{1^{n-k}}_{=1} = \sum_{k=0}^{n} \binom{n}{k} x^k. \tag{549}$$

(Strictly speaking, we did not literally apply Theorem 4.3.16 here, because the x in Theorem 4.3.16 is required to be a number, not a polynomial. But Theorem 4.3.16 holds for polynomials just as well as for numbers (i.e., the x and the y in Theorem 4.3.16 can be polynomials rather than numbers), and can be proved in this generality just as it was proved for numbers. It is this generalized version of Theorem 4.3.16 that we have applied above.)

Now, (549) is an equality between two polynomials in x; thus, we can take derivatives on both sides. The derivative of $(x+1)^n$ is easily seen to be $n(x+1)^{n-1}$, whereas the derivative of $\sum_{k=0}^{n} \binom{n}{k} x^k$ is $\sum_{k=0}^{n} \binom{n}{k} kx^{k-1}$ (since the derivative of each x^k is kx^{k-1}). Hence, by taking derivatives on both sides of (549), we obtain

$$n(x+1)^{n-1} = \sum_{k=0}^{n} \binom{n}{k} kx^{k-1}.$$

Substituting 1 for x in this equality, we obtain

$$n(1+1)^{n-1} = \sum_{k=0}^{n} \underbrace{\binom{n}{k} k}_{=k\binom{n}{k}} \underbrace{1^{k-1}}_{=1} = \sum_{k=0}^{n} k \binom{n}{k}.$$

Hence,

$$\sum_{k=0}^{n} k \binom{n}{k} = n \left(\underbrace{1+1}_{=2} \right)^{n-1} = n \cdot 2^{n-1}.$$

This again proves (548). □

Now, (548) has been proved (three times to boot), so Exercise 4.8.3 **(a)** is solved.

(b) This is similar to Exercise 4.4.3, except that the odd n has been replaced by an even $2n$. The solution is just a bit more complicated, because the $2n$-th row of Pascal's triangle splits into a left "half", a right "half", and a lone entry $\binom{2n}{n}$ in the middle. Here is the argument in detail:

Solution to Exercise 4.8.3 ***(b)***. Corollary 4.3.17 (applied to $2n$ instead of n) yields $\sum_{k=0}^{2n} \binom{2n}{k} = 2^{2n}$. Hence,

$$2^{2n} = \sum_{k=0}^{2n} \binom{2n}{k} = \sum_{k=0}^{n} \binom{2n}{k} + \sum_{k=n+1}^{2n} \binom{2n}{k} \tag{550}$$

(here, we have split the sum at $k = n$; that is, we have applied (84) to 0, n and $2n$ instead of u, v and w). But

$$\sum_{k=n+1}^{2n} \binom{2n}{k} = \sum_{k=0}^{n-1} \underbrace{\binom{2n}{2n-k}}_{\substack{=\binom{2n}{k} \\ \text{(since Theorem 4.3.10} \\ \text{(applied to } 2n \text{ instead of } n) \\ \text{yields } \binom{2n}{k} = \binom{2n}{2n-k})}}$$

$$\left(\begin{array}{c}\text{here, we have substituted } 2n-k \text{ for } k \text{ in the sum,} \\ \text{since the map } \{0,1,\ldots,n-1\} \to \{n+1,n+2,\ldots,2n\} \\ \text{that sends each } k \text{ to } 2n-k \text{ is a bijection}\end{array}\right)$$

$$= \sum_{k=0}^{n-1} \binom{2n}{k} = \sum_{k=0}^{n} \binom{2n}{k} - \binom{2n}{n}$$

(since $\sum_{k=0}^{n} \binom{2n}{k} = \sum_{k=0}^{n-1} \binom{2n}{k} + \binom{2n}{n}$). Hence, (550) becomes

$$2^{2n} = \sum_{k=0}^{n} \binom{2n}{k} + \underbrace{\sum_{k=n+1}^{2n} \binom{2n}{k}}_{=\sum_{k=0}^{n} \binom{2n}{k} - \binom{2n}{n}}$$

$$= \sum_{k=0}^{n} \binom{2n}{k} + \sum_{k=0}^{n} \binom{2n}{k} - \binom{2n}{n} = 2 \cdot \sum_{k=0}^{n} \binom{2n}{k} - \binom{2n}{n}.$$

Solving this equality for $\sum_{k=0}^{n} \binom{2n}{k}$, we obtain

$$\sum_{k=0}^{n} \binom{2n}{k} = \frac{1}{2} \cdot \left(2^{2n} + \binom{2n}{n}\right). \tag{551}$$

This already solves Exercise 4.8.3 **(b)**. But let us simplify this a bit further (getting rid of the $\frac{1}{2}$) when n is positive. Namely, assume that n is positive. Then, Exercise 4.5.4 **(a)** (applied to $2n$ and n instead of n and m) yields $\frac{n}{2n}\binom{2n}{n} = \binom{2n-1}{n-1}$. This

rewrites as $\frac{1}{2}\binom{2n}{n} = \binom{2n-1}{n-1}$ (since $\frac{n}{2n} = \frac{1}{2}$). Now, (551) becomes

$$\sum_{k=0}^{n}\binom{2n}{k} = \frac{1}{2}\cdot\left(2^{2n}+\binom{2n}{n}\right) = \underbrace{\frac{1}{2}\cdot 2^{2n}}_{=2^{2n-1}} + \underbrace{\frac{1}{2}\binom{2n}{n}}_{=\binom{2n-1}{n-1}}$$
$$= 2^{2n-1} + \binom{2n-1}{n-1}. \tag{552}$$

Note that this is **not** true for $n = 0$. Division by 0 is not a good idea! □

□

A.4.4. Discussion of Exercise 4.8.4

Discussion of Exercise 4.8.4. The two claims of Exercise 4.8.4 are known as "polarization identities" (although the name is often shared with other similar-looking formulas). Exercise 4.8.4 is a particular case of [Grinbe15, Exercise 6.51 parts **(a)** and **(b)**].[356] Let us give a self-contained solution here:

Solution to Exercise 4.8.4. For each $p \in \mathbb{N}$, we let $[p]$ be the p-element set $\{1, 2, \ldots, p\}$. (This generalizes the definition $[n] = \{1, 2, \ldots, n\}$ made in Exercise 4.8.4.)

For any $p \in \{0, 1, \ldots, n\}$ and $m \in \mathbb{N}$, we define a number

$$S_{p,m} := \sum_{I\subseteq[p]} (-1)^{p-|I|}\left(y + \sum_{i\in I} x_i\right)^m. \tag{553}$$

Thus, Exercise 4.8.4 **(a)** is saying that $S_{n,m} = 0$ for each $m \in \{0, 1, \ldots, n-1\}$, whereas Exercise 4.8.4 **(b)** is saying that $S_{n,n} = n!x_1x_2\cdots x_n$. The reason why we have introduced $S_{p,m}$ for general p (not just for $p = n$) is that we want to induct on p. (We could just as well induct on n, but it is somewhat easier to keep n fixed.)

Now, we claim that the $S_{p,m}$ satisfy the following recurrence relation:

Claim 1: Let $p \in \{1, 2, \ldots, n\}$ and $m \in \mathbb{N}$. Then,

$$S_{p,m} = \sum_{k=0}^{m-1}\binom{m}{k} x_p^{m-k} S_{p-1,k}.$$

[356] We are working with numbers here rather than elements of a noncommutative ring, so the complicated sum $\sum_{\sigma\in S_n} v_{\sigma(1)}v_{\sigma(2)}\cdots v_{\sigma(n)}$ from [Grinbe15, Exercise 6.51 **(b)**] can be simplified to $n!v_1v_2\cdots v_n$.

[*Proof of Claim 1:* First of all, both p and $p-1$ belong to the set $\{0,1,\ldots,n\}$ (since $p \in \{1,2,\ldots,n\}$). Thus, $[p]$ and $[p-1]$ are subsets of $[n]$.

If I is any subset of $[n]$, then we define a number z_I by

$$z_I := y + \sum_{i \in I} x_i. \tag{554}$$

(Thus, in particular, $z_\varnothing = y + \underbrace{(\text{empty sum})}_{=0} = y$.) The definition of $S_{p,m}$ now yields

$$\begin{aligned} S_{p,m} &= \sum_{I \subseteq [p]} (-1)^{p-|I|} \left(\underbrace{y + \sum_{i \in I} x_i}_{\substack{=z_I \\ \text{(by (554))}}} \right)^m \\ &= \sum_{I \subseteq [p]} (-1)^{p-|I|} z_I^m. \end{aligned} \tag{555}$$

Likewise, we can see that

$$S_{p-1,k} = \sum_{I \subseteq [p-1]} (-1)^{(p-1)-|I|} z_I^k \qquad \text{for each } k \in \mathbb{N}. \tag{556}$$

(Indeed, this follows from the same argument that we used to prove (555), but now applied to $p-1$ and k instead of p and m.)

We shall call a subset of $[p]$

- *red* if it contains p;
- *green* if it does not contain p.

Thus, each subset of $[p]$ is either red or green (but not both). Hence, the sum on the right hand side of (555) can be split up as follows:

$$\sum_{I \subseteq [p]} (-1)^{p-|I|} z_I^m = \sum_{\substack{I \subseteq [p]; \\ I \text{ is red}}} (-1)^{p-|I|} z_I^m + \sum_{\substack{I \subseteq [p]; \\ I \text{ is green}}} (-1)^{p-|I|} z_I^m.$$

We shall now take a closer look at the two sums on the right hand side of this.

The green subsets I of $[p]$ are the subsets of $[p]$ that do not contain p (by the definition of "green"). In other words, they are just the subsets of $[p] \setminus \{p\}$. Since $[p] \setminus \{p\} = [p-1]$, this means that they are just the subsets of $[p-1]$. Hence, we can rewrite the summation sign $\sum\limits_{\substack{I \subseteq [p]; \\ I \text{ is green}}}$ as $\sum\limits_{I \subseteq [p-1]}$. So we obtain

$$\sum_{\substack{I \subseteq [p]; \\ I \text{ is green}}} (-1)^{p-|I|} z_I^m = \sum_{I \subseteq [p-1]} (-1)^{p-|I|} z_I^m. \tag{557}$$

The red subsets are a bit more complicated. If I is a red subset of $[p]$, then I contains p (by the definition of "red"), and thus is not a subset of $[p-1]$. However, we can obtain a subset of $[p-1]$ by removing p from I. Conversely, if J is a subset of $[p-1]$, then we can obtain a red subset of $[p]$ by inserting p into J. Let us make this more formal: We have two maps

$$\begin{aligned} \{\text{red subsets of } [p]\} &\to \{\text{subsets of } [p-1]\}, \\ I &\mapsto I \setminus \{p\} \end{aligned}$$

and

$$\begin{aligned} \{\text{subsets of } [p-1]\} &\to \{\text{red subsets of } [p]\}, \\ J &\mapsto J \cup \{p\}. \end{aligned}$$

These two maps are mutually inverse; thus, they are invertible, i.e., they are bijections. In particular, the map

$$\begin{aligned} \{\text{subsets of } [p-1]\} &\to \{\text{red subsets of } [p]\}, \\ J &\mapsto J \cup \{p\} \end{aligned}$$

is a bijection. Hence, we can substitute $J \cup \{p\}$ for I in the sum

$$\sum_{\substack{I \subseteq [p]; \\ I \text{ is red}}} (-1)^{p-|I|} z_I^m.$$

As a result, we obtain

$$\begin{aligned} \sum_{\substack{I \subseteq [p]; \\ I \text{ is red}}} (-1)^{p-|I|} z_I^m &= \sum_{J \subseteq [p-1]} (-1)^{p-|J \cup \{p\}|} z_{J \cup \{p\}}^m \\ &= \sum_{I \subseteq [p-1]} (-1)^{p-|I \cup \{p\}|} z_{I \cup \{p\}}^m \end{aligned} \tag{558}$$

(here, we have renamed the summation index J as I).

We can rewrite this somewhat. Indeed, if I is a subset of $[p-1]$, then $p \notin I$ (since having $p \in I$ would entail $p \in I \subseteq [p-1] = \{1, 2, \ldots, p-1\}$, which is absurd) and therefore $|I \cup \{p\}| = |I| + 1$ and thus

$$(-1)^{p-|I \cup \{p\}|} = (-1)^{p-(|I|+1)} = (-1)^{p-|I|-1} = -(-1)^{p-|I|} \tag{559}$$

and

$$\begin{aligned} z_{I \cup \{p\}} &= y + \underbrace{\sum_{i \in I \cup \{p\}} x_i}_{\substack{= \sum_{i \in I} x_i + x_p \\ (\text{since } p \notin I)}} \qquad \left(\text{by the definition of } z_{I \cup \{p\}}\right) \\ &= \underbrace{y + \sum_{i \in I} x_i}_{\substack{= z_I \\ (\text{by } (554))}} + x_p = z_I + x_p. \end{aligned} \tag{560}$$

Thus, (558) becomes

$$\begin{aligned}
&\sum_{\substack{I\subseteq[p];\\ I \text{ is red}}} (-1)^{p-|I|} z_I^m \\
&= \sum_{I\subseteq[p-1]} \underbrace{(-1)^{p-|I\cup\{p\}|}}_{\substack{=-(-1)^{p-|I|}\\ \text{(by (559))}}} \underbrace{z_{I\cup\{p\}}^m}_{\substack{=\left(z_I+x_p\right)^m\\ \text{(by (560))}}} \\
&= \sum_{I\subseteq[p-1]} \left(-\left(-1\right)^{p-|I|}\right)\left(z_I+x_p\right)^m \\
&= -\sum_{I\subseteq[p-1]} (-1)^{p-|I|}\left(z_I+x_p\right)^m . \qquad (561)
\end{aligned}$$

Now, (555) yields

$$\begin{aligned}
S_{p,m} &= \sum_{I\subseteq[p]} (-1)^{p-|I|} z_I^m \\
&= \underbrace{\sum_{\substack{I\subseteq[p];\\ I \text{ is red}}} (-1)^{p-|I|} z_I^m}_{\substack{=-\sum_{I\subseteq[p-1]}(-1)^{p-|I|}\left(z_I+x_p\right)^m\\ \text{(by (561))}}} + \underbrace{\sum_{\substack{I\subseteq[p];\\ I \text{ is green}}} (-1)^{p-|I|} z_I^m}_{\substack{=\sum_{I\subseteq[p-1]}(-1)^{p-|I|}z_I^m\\ \text{(by (557))}}} \\
&= -\sum_{I\subseteq[p-1]} (-1)^{p-|I|}\left(z_I+x_p\right)^m + \sum_{I\subseteq[p-1]} (-1)^{p-|I|} z_I^m \\
&= \sum_{I\subseteq[p-1]} (-1)^{p-|I|} z_I^m - \sum_{I\subseteq[p-1]} (-1)^{p-|I|}\left(z_I+x_p\right)^m \\
&= \sum_{I\subseteq[p-1]} (-1)^{p-|I|}\left(z_I^m - \left(z_I+x_p\right)^m\right). \qquad (562)
\end{aligned}$$

Now, let us simplify the difference inside the big parentheses on the right hand side. Fix a subset I of $[p-1]$. Then, Theorem 4.3.16 (applied to m, z_I and x_p instead of n, x and y) yields

$$\begin{aligned}
\left(z_I+x_p\right)^m &= \sum_{k=0}^{m}\binom{m}{k} z_I^k x_p^{m-k} = \sum_{k=0}^{m-1}\binom{m}{k} z_I^k x_p^{m-k} + \underbrace{\binom{m}{m}}_{\substack{=1\\ \text{(by (124))}}} z_I^m \underbrace{x_p^{m-m}}_{=x_p^0=1} \\
&= \sum_{k=0}^{m-1}\binom{m}{k} z_I^k x_p^{m-k} + z_I^m .
\end{aligned}$$

Hence,

$$\begin{aligned} z_I^m - \left(z_I + x_p\right)^m &= z_I^m - \left(\sum_{k=0}^{m-1} \binom{m}{k} z_I^k x_p^{m-k} + z_I^m\right) \\ &= -\sum_{k=0}^{m-1} \binom{m}{k} z_I^k x_p^{m-k}. \end{aligned} \tag{563}$$

Forget that we fixed I. We thus have proved (563) for each subset I of $[p-1]$. Now, (562) becomes

$$\begin{aligned} S_{p,m} &= \sum_{I \subseteq [p-1]} (-1)^{p-|I|} \underbrace{\left(z_I^m - \left(z_I + x_p\right)^m\right)}_{\substack{= -\sum_{k=0}^{m-1} \binom{m}{k} z_I^k x_p^{m-k} \\ \text{(by (563))}}} \\ &= \sum_{I \subseteq [p-1]} (-1)^{p-|I|} \left(-\sum_{k=0}^{m-1} \binom{m}{k} z_I^k x_p^{m-k}\right) = -\sum_{I \subseteq [p-1]} (-1)^{p-|I|} \sum_{k=0}^{m-1} \binom{m}{k} z_I^k x_p^{m-k} \\ &= -\underbrace{\sum_{I \subseteq [p-1]} \sum_{k=0}^{m-1}}_{\substack{= \sum_{k=0}^{m-1} \sum_{I \subseteq [p-1]} \\ \text{(since } p-|I|-1=(p-1)-|I|)}} \underbrace{(-1)^{p-|I|}}_{\substack{= -(-1)^{p-|I|-1} \\ = -(-1)^{(p-1)-|I|}}} \binom{m}{k} z_I^k x_p^{m-k} \\ &= -\sum_{k=0}^{m-1} \sum_{I \subseteq [p-1]} \left(-(-1)^{(p-1)-|I|}\right) \binom{m}{k} z_I^k x_p^{m-k} \\ &= \sum_{k=0}^{m-1} \sum_{I \subseteq [p-1]} (-1)^{(p-1)-|I|} \binom{m}{k} z_I^k x_p^{m-k} \\ &= \sum_{k=0}^{m-1} \binom{m}{k} x_p^{m-k} \underbrace{\sum_{I \subseteq [p-1]} (-1)^{(p-1)-|I|} z_I^k}_{\substack{= S_{p-1,k} \\ \text{(by (556))}}} = \sum_{k=0}^{m-1} \binom{m}{k} x_p^{m-k} S_{p-1,k}. \end{aligned}$$

This proves Claim 1.]

With Claim 1 in hand, we can prove properties of the $S_{p,m}$ like the following by a straightforward induction:

> *Claim 2:* We have $S_{p,m} = 0$ for each $p \in \{0, 1, \ldots, n\}$ and each $m \in \{0, 1, \ldots, p-1\}$.

[*Proof of Claim 2:* We proceed by induction on p:

Induction base: If $p = 0$, then there exists no $m \in \{0, 1, \ldots, p-1\}$ (since the set $\{0, 1, \ldots, p-1\}$ is empty in this case). Thus, Claim 2 is vacuously true for $p = 0$.

Induction step: Let $q \in \{0, 1, \ldots, n-1\}$. Assume (as the induction hypothesis) that Claim 2 holds for $p = q$. We must prove that Claim 2 holds for $p = q + 1$.

We have assumed that Claim 2 holds for $p = q$. In other words, we have

$$S_{q,m} = 0 \qquad \text{for each } m \in \{0, 1, \ldots, q-1\}. \tag{564}$$

Now, let $m \in \{0, 1, \ldots, q\}$. Thus, $0 \leq m \leq q$. Note that $q + 1 \in \{1, 2, \ldots, n\}$ (since $q \in \{0, 1, \ldots, n-1\}$). Hence, Claim 1 (applied to $p = q + 1$) yields

$$S_{q+1,m} = \sum_{k=0}^{m-1} \binom{m}{k} x_{q+1}^{m-k} S_{(q+1)-1,k} = \sum_{k=0}^{m-1} \binom{m}{k} x_{q+1}^{m-k} S_{q,k} \tag{565}$$

(since $(q+1) - 1 = q$). However, if $k \in \{0, 1, \ldots, m-1\}$, then $k \leq \underbrace{m}_{\leq q} - 1 \leq q - 1$ and therefore $k \in \{0, 1, \ldots, q-1\}$ (since k is a nonnegative integer) and thus

$$S_{q,k} = 0 \tag{566}$$

(by (564), applied to k instead of m). Hence, (565) becomes

$$S_{q+1,m} = \sum_{k=0}^{m-1} \binom{m}{k} x_{q+1}^{m-k} \underbrace{S_{q,k}}_{\substack{=0 \\ \text{(by (566))}}} = \sum_{k=0}^{m-1} \binom{m}{k} x_{q+1}^{m-k} 0 = 0.$$

Forget that we fixed m. We thus have shown that $S_{q+1,m} = 0$ for each $m \in \{0, 1, \ldots, q\}$. In other words, Claim 2 holds for $p = q + 1$. This completes the induction step. Thus, Claim 2 is proved.]

Claim 2 yields part **(a)** of Exercise 4.8.4 rather quickly (see below for the details). In order to solve part **(b)** as well, we need another claim:

Claim 3: We have $S_{p,p} = p! x_1 x_2 \cdots x_p$ for each $p \in \{0, 1, \ldots, n\}$.

[*Proof of Claim 3:* We proceed by induction on p:

Induction base: If $p = 0$, then

$$\begin{aligned}
S_{p,p} = S_{0,0} &= \sum_{I \subseteq [0]} (-1)^{0-|I|} \underbrace{\left(y + \sum_{i \in I} x_i\right)^0}_{=1} \qquad \text{(by the definition of } S_{0,0}) \\
&= \sum_{I \subseteq [0]} (-1)^{0-|I|} = \sum_{I \subseteq \varnothing} (-1)^{0-|I|} \qquad (\text{since } [0] = \varnothing) \\
&= (-1)^{0-|\varnothing|} \qquad (\text{since the only subset } I \text{ of } \varnothing \text{ is } \varnothing) \\
&= 1 \qquad (\text{since } 0 - |\varnothing| = 0 - 0 = 0 \text{ is even}) \\
&= p! x_1 x_2 \cdots x_p
\end{aligned}$$

(because $p = 0$ entails $p!x_1x_2\cdots x_p = \underbrace{0!}_{=1} \underbrace{x_1x_2\cdots x_0}_{=(\text{empty product})=1} = 1$). In other words, Claim 3 holds for $p = 0$.

Induction step: Let $q \in \{0, 1, \ldots, n-1\}$. Assume (as the induction hypothesis) that Claim 3 holds for $p = q$. We must prove that Claim 3 holds for $p = q + 1$.

We have assumed that Claim 3 holds for $p = q$. In other words, we have

$$S_{q,q} = q!x_1x_2\cdots x_q. \tag{567}$$

Furthermore, $q \in \{0, 1, \ldots, n-1\} \subseteq \{0, 1, \ldots, n\}$. Hence, each $k \in \{0, 1, \ldots, q-1\}$ satisfies

$$S_{q,k} = 0 \tag{568}$$

(by Claim 2, applied to $p = q$ and $m = k$).

Now, $q + 1 \in \{1, 2, \ldots, n\}$ (since $q \in \{0, 1, \ldots, n-1\}$). Hence, Claim 1 (applied to $p = q + 1$ and $m = q + 1$) yields

$$\begin{aligned}
S_{q+1,q+1} &= \sum_{k=0}^{(q+1)-1} \binom{q+1}{k} x_{q+1}^{(q+1)-k} S_{(q+1)-1,k} \\
&= \sum_{k=0}^{q} \binom{q+1}{k} x_{q+1}^{(q+1)-k} S_{q,k} \qquad (\text{since } (q+1)-1 = q) \\
&= \sum_{k=0}^{q-1} \binom{q+1}{k} x_{q+1}^{(q+1)-k} \underbrace{S_{q,k}}_{\substack{=0 \\ (\text{by } (568))}} + \binom{q+1}{q} \underbrace{x_{q+1}^{(q+1)-q}}_{=x_{q+1}^1 = x_{q+1}} S_{q,q} \\
&= \underbrace{\sum_{k=0}^{q-1} \binom{q+1}{k} x_{q+1}^{(q+1)-k} 0}_{=0} + \binom{q+1}{q} x_{q+1} S_{q,q} \\
&= \binom{q+1}{q} x_{q+1} S_{q,q}.
\end{aligned} \tag{569}$$

Theorem 4.3.10 (applied to $q + 1$ and q instead of n and k) yields

$$\binom{q+1}{q} = \binom{q+1}{(q+1)-q} = \binom{q+1}{1} = q + 1 \qquad (\text{by } (120)).$$

Hence, (569) becomes

$$\begin{aligned}
S_{q+1,q+1} &= \underbrace{\binom{q+1}{q}}_{=q+1} x_{q+1} \underbrace{S_{q,q}}_{\substack{=q!x_1x_2\cdots x_q \\ (\text{by } (567))}} = (q+1)\, x_{q+1} \cdot q!x_1x_2\cdots x_q \\
&= \underbrace{(q+1)\cdot q!}_{=(q+1)!} \cdot \underbrace{(x_1x_2\cdots x_q)\, x_{q+1}}_{=x_1x_2\cdots x_{q+1}} = (q+1)!x_1x_2\cdots x_{q+1}.
\end{aligned}$$

In other words, Claim 3 holds for $p = q + 1$. This completes the induction step. Thus, Claim 3 is proved.]

Let us now finish the solution of Exercise 4.8.4:

(a) Let $m \in \{0, 1, \ldots, n-1\}$. Then, Claim 2 (applied to $p = n$) yields

$$S_{n,m} = 0.$$

But the definition of $S_{n,m}$ yields

$$S_{n,m} = \sum_{I \subseteq [n]} (-1)^{n-|I|} \left(y + \sum_{i \in I} x_i \right)^m .$$

Comparing these two equalities, we find

$$\sum_{I \subseteq [n]} (-1)^{n-|I|} \left(y + \sum_{i \in I} x_i \right)^m = 0.$$

This solves Exercise 4.8.4 **(a)**.

(b) Claim 3 (applied to $p = n$) yields

$$S_{n,n} = n! x_1 x_2 \cdots x_n.$$

But the definition of $S_{n,n}$ yields

$$S_{n,n} = \sum_{I \subseteq [n]} (-1)^{n-|I|} \left(y + \sum_{i \in I} x_i \right)^n .$$

Comparing these two equalities, we find

$$\sum_{I \subseteq [n]} (-1)^{n-|I|} \left(y + \sum_{i \in I} x_i \right)^n = n! x_1 x_2 \cdots x_n.$$

This solves Exercise 4.8.4 **(b)**. □

There is an alternative solution to Exercise 4.8.4, which proceeds by expanding the products

$$\left(y + \sum_{i \in I} x_i \right)^m = \left(y + \sum_{i \in I} x_i \right) \left(y + \sum_{i \in I} x_i \right) \cdots \left(y + \sum_{i \in I} x_i \right)$$

(for all $I \subseteq [n]$) as sums of products of y's and x_i's and then observing that the "$\sum_{I \subseteq [n]} (-1)^{n-|I|}$ operator" causes all such products other than (permutations of) $x_1 x_2 \cdots x_n$ to cancel each other. This argument (albeit only in the particular case $y = 0$, which is a bit simpler) can be found in [18f-hw3s, solution to Exercise 6]. □

A.4.5. Discussion of Exercise 4.8.5

Discussion of Exercise 4.8.5. Exercise 4.8.5 is an identity originally found by Euler in 1753 ([Euler190, §11]). It has recently reappeared as the first part of AMM problem #12022 (see [MerJoh19] for a solution) and (in a slightly rewritten form) in [BiFoFo95, Proposition]. The solution we give below is a restatement of the solution from [MerJoh19] (with all details expanded and with the induction replaced by a use of the telescope principle).

Solution to Exercise 4.8.5. First of all, let us see why the fractions appearing in Exercise 4.8.5 are well-defined in the first place. Indeed, all powers $x^1, x^2, x^3, \ldots$ of x are distinct from 1 (since $x \in \mathbb{R} \setminus \{1, -1\}$). Hence, all the numbers $1 - x^1, 1 - x^2, 1 - x^3, \ldots$ are nonzero. In other words, all the numbers $y_1, y_2, y_3, \ldots$ are nonzero (since $y_i = 1 - x^i$ for each integer $i \geq 1$). Therefore, the denominators of the fractions $\dfrac{y_n y_{n-1} \cdots y_{n-k}}{y_{k+1}}$ in Exercise 4.8.5 are nonzero, so the fractions are well-defined. This shows that Exercise 4.8.5 makes sense.

For any two numbers $p \in \{0, 1, \ldots, n\}$ and $k \in \{0, 1, \ldots, n\}$ satisfying $p \geq k$, we set

$$b_{p,k} = \frac{y_p y_{p-1} \cdots y_{p-k}}{y_{k+1}} \tag{570}$$

and

$$c_{p,k} = y_{p-1} y_{p-2} \cdots y_{p-k}. \tag{571}$$

Now, an easy computation reveals the following:

> *Claim 1:* For any two numbers $p \in \{0, 1, \ldots, n\}$ and $k \in \{0, 1, \ldots, n\}$ satisfying $p - 1 \geq k$, we have
>
> $$b_{p,k} - b_{p-1,k} = c_{p,k} - c_{p,k+1}.$$

[*Proof of Claim 1:* Let $p \in \{0, 1, \ldots, n\}$ and $k \in \{0, 1, \ldots, n\}$ be two numbers satisfying $p - 1 \geq k$. Then, $p - 1 \geq k \geq 0$ and $p - 1 \leq p \leq n$ (since $p \in \{0, 1, \ldots, n\}$). Combining these two inequalities, we obtain $p - 1 \in \{0, 1, \ldots, n\}$. Hence, $b_{p-1,k}$ is well-defined. Also, $p > p - 1 \geq k$; hence, $b_{p,k}$ and $c_{p,k}$ are well-defined. Finally, $p - 1 \geq k$ entails $p \geq k + 1$, so that $k + 1 \leq p \leq n$ and therefore $k + 1 \in \{0, 1, \ldots, n\}$. Hence, $c_{p,k+1}$ is well-defined.

Now, we have

$$\begin{aligned} b_{p,k} &= \frac{y_p y_{p-1} \cdots y_{p-k}}{y_{k+1}} \qquad \left(\text{by the definition of } b_{p,k}\right) \\ &= \frac{1}{y_{k+1}} \cdot \underbrace{y_p y_{p-1} \cdots y_{p-k}}_{= y_p \cdot y_{p-1} y_{p-2} \cdots y_{p-k}} = \frac{1}{y_{k+1}} \cdot y_p \cdot y_{p-1} y_{p-2} \cdots y_{p-k} \end{aligned}$$

and

$$\begin{aligned} b_{p-1,k} &= \frac{y_{p-1}y_{(p-1)-1}\cdots y_{(p-1)-k}}{y_{k+1}} \qquad \left(\text{by the definition of } b_{p-1,k}\right) \\ &= \frac{y_{p-1}y_{p-2}\cdots y_{p-k-1}}{y_{k+1}} = \frac{1}{y_{k+1}}\cdot \underbrace{y_{p-1}y_{p-2}\cdots y_{p-k-1}}_{=y_{p-1}y_{p-2}\cdots y_{p-k}\cdot y_{p-k-1}} \\ &= \frac{1}{y_{k+1}}\cdot y_{p-1}y_{p-2}\cdots y_{p-k}\cdot y_{p-k-1} = \frac{1}{y_{k+1}}\cdot y_{p-k-1}\cdot y_{p-1}y_{p-2}\cdots y_{p-k}. \end{aligned}$$

Subtracting these two equalities from one another, we obtain

$$\begin{aligned} b_{p,k} - b_{p-1,k} &= \frac{1}{y_{k+1}}\cdot y_p\cdot y_{p-1}y_{p-2}\cdots y_{p-k} - \frac{1}{y_{k+1}}\cdot y_{p-k-1}\cdot y_{p-1}y_{p-2}\cdots y_{p-k} \\ &= \frac{1}{y_{k+1}}\cdot\left(y_p - y_{p-k-1}\right)\cdot y_{p-1}y_{p-2}\cdots y_{p-k}. \end{aligned} \tag{572}$$

On the other hand, $y_p = 1 - x^p$ (by the definition of y_p) and $y_{p-k-1} = 1 - x^{p-k-1}$ (by the definition of y_{p-k-1}). Subtracting these two equalities from one another, we obtain

$$\begin{aligned} y_p - y_{p-k-1} &= \left(1 - x^p\right) - \left(1 - x^{p-k-1}\right) = x^{p-k-1} - x^p \\ &= \underbrace{\left(1 - x^{k+1}\right)}_{\substack{=y_{k+1} \\ \text{(since } y_{k+1} \text{ is} \\ \text{defined as } 1-x^{k+1})}} \underbrace{x^{p-k-1}}_{\substack{=1-y_{p-k-1} \\ \text{(since } y_{p-k-1}=1-x^{p-k-1})}} = y_{k+1}\left(1 - y_{p-k-1}\right). \end{aligned}$$

Thus, (572) becomes

$$\begin{aligned} b_{p,k} - b_{p-1,k} &= \frac{1}{y_{k+1}}\cdot \underbrace{\left(y_p - y_{p-k-1}\right)}_{=y_{k+1}\left(1-y_{p-k-1}\right)}\cdot y_{p-1}y_{p-2}\cdots y_{p-k} \\ &= \frac{1}{y_{k+1}}\cdot y_{k+1}\left(1 - y_{p-k-1}\right)\cdot y_{p-1}y_{p-2}\cdots y_{p-k} \\ &= \left(1 - y_{p-k-1}\right)\cdot y_{p-1}y_{p-2}\cdots y_{p-k}. \end{aligned}$$

Comparing this with

$$\begin{aligned} &\underbrace{c_{p,k}}_{\substack{=y_{p-1}y_{p-2}\cdots y_{p-k} \\ \text{(by the definition of } c_{p,k})}} - \underbrace{c_{p,k+1}}_{\substack{=y_{p-1}y_{p-2}\cdots y_{p-(k+1)} \\ \text{(by the definition of } c_{p,k+1})}} \\ &= y_{p-1}y_{p-2}\cdots y_{p-k} - \underbrace{y_{p-1}y_{p-2}\cdots y_{p-(k+1)}}_{\substack{=y_{p-1}y_{p-2}\cdots y_{p-k-1} \\ =y_{p-1}y_{p-2}\cdots y_{p-k}\cdot y_{p-k-1}}} \\ &= y_{p-1}y_{p-2}\cdots y_{p-k} - y_{p-1}y_{p-2}\cdots y_{p-k}\cdot y_{p-k-1} \\ &= y_{p-1}y_{p-2}\cdots y_{p-k}\cdot\left(1 - y_{p-k-1}\right) = \left(1 - y_{p-k-1}\right)\cdot y_{p-1}y_{p-2}\cdots y_{p-k}, \end{aligned}$$

we obtain $b_{p,k} - b_{p-1,k} = c_{p,k} - c_{p,k+1}$. This proves Claim 1.]

Another easy computation yields the following:

Claim 2: For each $p \in \{1, 2, \ldots, n\}$, we have

$$b_{p,p-1} = c_{p,p-1}.$$

[*Proof of Claim 2:* Let $p \in \{1, 2, \ldots, n\}$. Thus, both p and $p-1$ belong to $\{0, 1, \ldots, n\}$. Hence, both $b_{p,p-1}$ and $c_{p,p-1}$ are well-defined (since $p \geq p-1$). Comparing

$$\begin{aligned} b_{p,p-1} &= \frac{y_p y_{p-1} \cdots y_{p-(p-1)}}{y_{(p-1)+1}} \qquad \left(\text{by the definition of } b_{p,p-1}\right) \\ &= \frac{y_p y_{p-1} \cdots y_1}{y_p} = y_{p-1} y_{p-2} \cdots y_1 \\ &\qquad \left(\text{since } p \geq 1 \text{ and thus } y_p y_{p-1} \cdots y_1 = y_p \cdot y_{p-1} y_{p-2} \cdots y_1\right) \end{aligned}$$

with

$$\begin{aligned} c_{p,p-1} &= y_{p-1} y_{p-2} \cdots y_{p-(p-1)} \qquad \left(\text{by the definition of } c_{p,p-1}\right) \\ &= y_{p-1} y_{p-2} \cdots y_1, \end{aligned}$$

we obtain $b_{p,p-1} = c_{p,p-1}$. This proves Claim 2.]

For each $p \in \{1, 2, \ldots, n\}$, set

$$S_p = \sum_{k=0}^{p-1} b_{p,k}. \tag{573}$$

Now, we claim the following:

Claim 3: For any $p \in \{1, 2, \ldots, n\}$, we have

$$S_p - S_{p-1} = 1.$$

[*Proof of Claim 3:* Fix $p \in \{1, 2, \ldots, n\}$. Then, $p \geq 1$, so that $p - 1 \geq 0$. The definition of S_p yields

$$S_p = \sum_{k=0}^{p-1} b_{p,k} = b_{p,p-1} + \sum_{k=0}^{p-2} b_{p,k} \qquad (\text{since } p - 1 \geq 0).$$

Meanwhile, the definition of S_{p-1} yields

$$S_{p-1} = \sum_{k=0}^{(p-1)-1} b_{p-1,k} = \sum_{k=0}^{p-2} b_{p-1,k} \qquad (\text{since } (p-1) - 1 = p - 2).$$

Subtracting these two equalities from one another, we obtain

$$S_p - S_{p-1} = \left(b_{p,p-1} + \sum_{k=0}^{p-2} b_{p,k}\right) - \sum_{k=0}^{p-2} b_{p-1,k} = b_{p,p-1} + \underbrace{\sum_{k=0}^{p-2} b_{p,k} - \sum_{k=0}^{p-2} b_{p-1,k}}_{=\sum_{k=0}^{p-2}\left(b_{p,k}-b_{p-1,k}\right)}$$

$$= b_{p,p-1} + \sum_{k=0}^{p-2} \left(b_{p,k} - b_{p-1,k}\right).$$

In view of

$$\begin{aligned}
&\sum_{k=0}^{p-2} \underbrace{\left(b_{p,k} - b_{p-1,k}\right)}_{\substack{=c_{p,k}-c_{p,k+1} \\ \text{(by Claim 1} \\ \text{(since } k\leq p-2<p-1 \text{ and thus } p-1\geq k))}} \\
&= \sum_{k=0}^{p-2} \underbrace{\left(c_{p,k} - c_{p,k+1}\right)}_{=\left(-c_{p,k+1}\right)-\left(-c_{p,k}\right)} = \sum_{k=0}^{p-2} \left(\left(-c_{p,k+1}\right) - \left(-c_{p,k}\right)\right) \\
&= \sum_{i=0}^{p-2} \left(\left(-c_{p,i+1}\right) - \left(-c_{p,i}\right)\right) \qquad \left(\begin{array}{c}\text{here, we have renamed the} \\ \text{summation index } k \text{ as } i\end{array}\right) \\
&= \left(-c_{p,(p-2)+1}\right) - \left(-c_{p,0}\right) \\
&\qquad \left(\text{by Corollary 4.1.17, applied to } u = 0 \text{ and } v = p-2 \text{ and } a_i = -c_{p,i}\right) \\
&= c_{p,0} - c_{p,(p-2)+1} = c_{p,0} - c_{p,p-1} \qquad (\text{since } (p-2)+1 = p-1),
\end{aligned}$$

this becomes

$$\begin{aligned}
S_p - S_{p-1} &= \underbrace{b_{p,p-1}}_{\substack{=c_{p,p-1} \\ \text{(by Claim 2)}}} + \underbrace{\sum_{k=0}^{p-2} \left(b_{p,k} - b_{p-1,k}\right)}_{=c_{p,0}-c_{p,p-1}} = c_{p,p-1} + c_{p,0} - c_{p,p-1} \\
&= c_{p,0} = y_{p-1} y_{p-2} \cdots y_{p-0} \qquad \left(\text{by the definition of } c_{p,0}\right) \\
&= (\text{empty product}) = 1.
\end{aligned}$$

This proves Claim 3.]

Now,

$$\sum_{p=1}^{n} \underbrace{\left(S_p - S_{p-1}\right)}_{\substack{=1 \\ \text{(by Claim 3)}}} = \sum_{p=1}^{n} 1 = n \cdot 1 = n.$$

Hence,

$$\begin{aligned}
n &= \sum_{p=1}^{n} \left(S_p - S_{p-1}\right) = \sum_{s=1}^{n} \left(S_s - S_{s-1}\right) \qquad \left(\begin{array}{c}\text{here, we have renamed the} \\ \text{summation index } p \text{ as } s\end{array}\right) \\
&= \underbrace{S_n}_{\substack{=\sum_{k=0}^{n-1} b_{n,k} \\ \text{(by the definition of } S_n)}} - \underbrace{S_{1-1}}_{\substack{=S_0 = \sum_{k=0}^{0-1} b_{0,k} \\ \text{(by the definition of } S_0)}} \\
&\qquad \left(\text{by Theorem 4.1.16, applied to } u = 1 \text{ and } v = n \text{ and } a_s = S_s\right) \\
&= \sum_{k=0}^{n-1} \underbrace{b_{n,k}}_{\substack{= \frac{y_n y_{n-1} \cdots y_{n-k}}{y_{k+1}} \\ \text{(by the definition of } b_{n,k})}} - \underbrace{\sum_{k=0}^{0-1} b_{0,k}}_{=(\text{empty sum})=0} = \sum_{k=0}^{n-1} \frac{y_n y_{n-1} \cdots y_{n-k}}{y_{k+1}}.
\end{aligned}$$

This solves Exercise 4.8.5. □

Let us note in passing that we can prove the following generalization of Exercise 4.8.5:

Proposition A.4.4. Let $n \in \mathbb{N}$ and $x \in \mathbb{R} \setminus \{1, -1\}$. For each $i \in \mathbb{N}$, set $y_i = 1 - x^i$. Let $m \in \{0, 1, \ldots, n\}$. Then,

$$\sum_{k=m}^{n-1} \frac{y_n y_{n-1} \cdots y_{n-k}}{y_{k+1}} = \sum_{p=m+1}^{n} y_{p-1} y_{p-2} \cdots y_{p-m}. \tag{574}$$

Exercise 4.8.5 is easily seen to be the particular case of Proposition A.4.4 for $m = 0$. Our below proof of Proposition A.4.4 is a lazy adaptation of our above solution to Exercise 4.8.5.

Proof of Proposition A.4.4. We proceed precisely as in the above solution to Exercise 4.8.5 until the point where we define S_p. From that point on, we proceed as follows:

For each $p \in \{m, m+1, \ldots, n\}$, set

$$T_p = \sum_{k=m}^{p-1} b_{p,k}.$$

(Note that if $m = 0$, then T_p is the S_p from our above solution to Exercise 4.8.5.)

Now, we claim the following:

Claim 4: For any $p \in \{m+1, m+2, \ldots, n\}$, we have

$$T_p - T_{p-1} = c_{p,m}.$$

[*Proof of Claim 4:* Fix $p \in \{m+1, m+2, \ldots, n\}$. Then, $p \geq m+1$, so that $p-1 \geq m$. Both p and $p-1$ belong to the set $\{m, m+1, \ldots, n\}$ (since $p \in \{m+1, m+2, \ldots, n\}$). Hence, T_p and T_{p-1} are well-defined. The definition of T_p yields

$$T_p = \sum_{k=m}^{p-1} b_{p,k} = b_{p,p-1} + \sum_{k=m}^{p-2} b_{p,k} \qquad (\text{since } p-1 \geq m).$$

Meanwhile, the definition of T_{p-1} yields

$$T_{p-1} = \sum_{k=m}^{(p-1)-1} b_{p-1,k} = \sum_{k=m}^{p-2} b_{p-1,k} \qquad (\text{since } (p-1)-1 = p-2).$$

Subtracting these two equalities from one another, we obtain

$$\begin{aligned} T_p - T_{p-1} &= \left(b_{p,p-1} + \sum_{k=m}^{p-2} b_{p,k} \right) - \sum_{k=m}^{p-2} b_{p-1,k} = b_{p,p-1} + \underbrace{\sum_{k=m}^{p-2} b_{p,k} - \sum_{k=m}^{p-2} b_{p-1,k}}_{= \sum_{k=m}^{p-2} \left(b_{p,k} - b_{p-1,k} \right)} \\ &= b_{p,p-1} + \sum_{k=m}^{p-2} \left(b_{p,k} - b_{p-1,k} \right). \end{aligned}$$

In view of

$$\begin{aligned} &\sum_{k=m}^{p-2} \underbrace{\left(b_{p,k} - b_{p-1,k} \right)}_{\substack{= c_{p,k} - c_{p,k+1} \\ \text{(by Claim 1} \\ \text{(since } k \leq p-2 < p-1 \text{ and thus } p-1 \geq k))}} \\ &= \sum_{k=m}^{p-2} \underbrace{\left(c_{p,k} - c_{p,k+1} \right)}_{= \left(-c_{p,k+1} \right) - \left(-c_{p,k} \right)} = \sum_{k=m}^{p-2} \left(\left(-c_{p,k+1} \right) - \left(-c_{p,k} \right) \right) \\ &= \sum_{i=m}^{p-2} \left(\left(-c_{p,i+1} \right) - \left(-c_{p,i} \right) \right) \qquad \left(\begin{array}{c} \text{here, we have renamed the} \\ \text{summation index } k \text{ as } i \end{array} \right) \\ &= \left(-c_{p,(p-2)+1} \right) - \left(-c_{p,m} \right) \\ &\qquad \left(\text{by Corollary 4.1.17, applied to } u = m \text{ and } v = p-2 \text{ and } a_i = -c_{p,i} \right) \\ &= c_{p,m} - c_{p,(p-2)+1} = c_{p,m} - c_{p,p-1} \qquad (\text{since } (p-2)+1 = p-1), \end{aligned}$$

this becomes

$$T_p - T_{p-1} = \underbrace{b_{p,p-1}}_{\substack{= c_{p,p-1} \\ \text{(by Claim 2)}}} + \underbrace{\sum_{k=m}^{p-2} \left(b_{p,k} - b_{p-1,k} \right)}_{= c_{p,m} - c_{p,p-1}} = c_{p,p-1} + c_{p,m} - c_{p,p-1} = c_{p,m}.$$

This proves Claim 4.]

Now,

$$\sum_{p=m+1}^{n} \underbrace{\left(T_p - T_{p-1}\right)}_{\substack{=c_{p,m} \\ \text{(by Claim 4)}}} = \sum_{p=m+1}^{n} \underbrace{c_{p,m}}_{\substack{=y_{p-1}y_{p-2}\cdots y_{p-m} \\ \text{(by the definition of } c_{p,m})}} = \sum_{p=m+1}^{n} y_{p-1}y_{p-2}\cdots y_{p-m}.$$

Hence,

$$\begin{aligned}
&\sum_{p=m+1}^{n} y_{p-1}y_{p-2}\cdots y_{p-m} \\
&= \sum_{p=m+1}^{n} \left(T_p - T_{p-1}\right) = \sum_{s=m+1}^{n} \left(T_s - T_{s-1}\right) \qquad \left(\begin{array}{c}\text{here, we have renamed the} \\ \text{summation index } p \text{ as } s\end{array}\right) \\
&= \underbrace{T_n}_{\substack{=\sum_{k=m}^{n-1} b_{n,k} \\ \text{(by the definition of } T_n)}} - \underbrace{T_{(m+1)-1}}_{\substack{=T_m=\sum_{k=m}^{m-1} b_{m,k} \\ \text{(by the definition of } T_m)}} \\
&\qquad \text{(by Theorem 4.1.16, applied to } u = m+1 \text{ and } v = n \text{ and } a_s = T_s) \\
&= \sum_{k=m}^{n-1} \underbrace{b_{n,k}}_{\substack{=\frac{y_n y_{n-1}\cdots y_{n-k}}{y_{k+1}} \\ \text{(by the definition of } b_{n,k})}} - \underbrace{\sum_{k=m}^{m-1} b_{m,k}}_{=(\text{empty sum})=0} = \sum_{k=m}^{n-1} \frac{y_n y_{n-1}\cdots y_{n-k}}{y_{k+1}}.
\end{aligned}$$

This proves Proposition A.4.4. □

□

A.4.6. Discussion of Exercise 4.8.6

Discussion of Exercise 4.8.6. It is easy to realize that a_n is a multiple of $n!$ for each $n \in \mathbb{N}$. (Indeed, this is clearly true for $n = 0$ and for $n = 1$, while for higher n it follows by induction using the recursion $a_n = n\left(a_{n-1} + (n-1)\, a_{n-2}\right)$). In other words, $\dfrac{a_n}{n!}$ is an integer for each $n \in \mathbb{N}$. This suggests that we should study the sequence $\left(\dfrac{a_0}{0!}, \dfrac{a_1}{1!}, \dfrac{a_2}{2!}, \ldots\right)$ instead of the sequence $(a_0, a_1, a_2, \ldots)$. The recursion $a_n = n\left(a_{n-1} + (n-1)\, a_{n-2}\right)$ (which holds for each $n \geq 2$) can now be easily rewritten in terms of the new sequence $\left(\dfrac{a_0}{0!}, \dfrac{a_1}{1!}, \dfrac{a_2}{2!}, \ldots\right)$ as $\dfrac{a_n}{n!} = \dfrac{a_{n-1}}{(n-1)!} + \dfrac{a_{n-2}}{(n-2)!}$ (because $n! = n \cdot (n-1)! = n\,(n-1) \cdot (n-2)!$); but this is precisely the recursion of the Fibonacci sequence. Its starting values $\dfrac{a_0}{0!} = 0$ and $\dfrac{a_1}{1!} = 1$, too, are identical with those of the Fibonacci sequence. Thus, we conclude that the sequence

$\left(\frac{a_0}{0!}, \frac{a_1}{1!}, \frac{a_2}{2!}, \ldots\right)$ must be the Fibonacci sequence. In other words,

$$\frac{a_n}{n!} = f_n \qquad \text{for each } n \in \mathbb{N}$$

(where $(f_0, f_1, f_2, \ldots)$ is the Fibonacci sequence). In other words,

$$a_n = n! \cdot f_n \qquad \text{for each } n \in \mathbb{N}. \tag{575}$$

It is easy to restate the above derivation of (575) as a rigorous proof by strong induction on n. □

A.4.7. Discussion of Exercise 4.8.7

Discussion of Exercise 4.8.7. There are many approaches to Exercise 4.8.7. The most "systematic" approach is by showing that both sequences $(a_0, a_1, a_2, \ldots)$ and $(b_0, b_1, b_2, \ldots)$ are $\left(2u, v^2 - u^2\right)$-recurrent[357], and then applying Theorem 4.9.11 to obtain an explicit formula for them. But this is laborious. A simpler solution can be given using an utterly trivial lemma:

[357] Indeed, for each $n \geq 2$, we have the three equalities

$$\begin{cases} a_n = ua_{n-1} + vb_{n-1}; \\ a_{n-1} = ua_{n-2} + vb_{n-2}; \\ b_{n-1} = ub_{n-2} + va_{n-2}. \end{cases}$$

Eliminating b_{n-1} and b_{n-2} from them (for example, by using the first equality to express b_{n-1} through a_n and a_{n-1}, then using the second equality to express b_{n-2} through a_{n-1} and a_{n-2}, then substituting the resulting two expressions into the third equality), we obtain

$$a_n = 2ua_{n-1} + \left(v^2 - u^2\right) a_{n-2}.$$

Thus, the sequence $(a_0, a_1, a_2, \ldots)$ is $\left(2u, v^2 - u^2\right)$-recurrent. Similarly, the same holds for the sequence $(b_0, b_1, b_2, \ldots)$.

A more conceptual way to show this uses the Cayley–Hamilton theorem. Namely, the recursions

$$a_n = ua_{n-1} + vb_{n-1} \qquad \text{and} \qquad b_n = ub_{n-1} + va_{n-1}$$

can be combined into a single matrix equation

$$\begin{pmatrix} a_n \\ b_n \end{pmatrix} = \begin{pmatrix} u & v \\ v & u \end{pmatrix} \begin{pmatrix} a_{n-1} \\ b_{n-1} \end{pmatrix}.$$

We can rewrite this equation as

$$w_n = Aw_{n-1}, \qquad \text{where } A = \begin{pmatrix} u & v \\ v & u \end{pmatrix} \text{ and } w_i = \begin{pmatrix} a_i \\ b_i \end{pmatrix} \text{ for each } i \in \mathbb{N}.$$

Now, the characteristic polynomial of A is $X^2 - 2uX - \left(v^2 - u^2\right)$; thus, the Cayley–Hamilton theorem yields $A^2 - 2uA - \left(v^2 - u^2\right) I_2 = 0$ (where I_2 is the 2×2 identity matrix). Thus,

$$\underbrace{\left(A^2 - 2uA - \left(v^2 - u^2\right) I_2\right)}_{=0} w_n = 0 \qquad \text{for each } n \in \mathbb{N}.$$

Lemma A.4.5. Let a and b be two numbers. Let $s = a + b$ and $d = a - b$. Then, $a = \dfrac{s+d}{2}$ and $b = \dfrac{s-d}{2}$.

Proof of Lemma A.4.5. From $s = a + b$ and $d = a - b$, we obtain

$$\frac{s+d}{2} = \frac{(a+b)+(a-b)}{2} = a \qquad \text{and} \qquad \frac{s-d}{2} = \frac{(a+b)-(a-b)}{2} = b.$$

This proves Lemma A.4.5. □

Harmless as it is, Lemma A.4.5 suggests a useful trick: Any two numbers can be reconstructed from their sum and their difference. Sometimes, these sums and differences are easier to handle than the original two numbers, so it can be a good idea to work with the sum and the difference for as long as possible. Nothing is lost when doing so, since Lemma A.4.5 allows recovering the original numbers.

Let us now solve Exercise 4.8.7 using this trick:

Solution to Exercise 4.8.7. For each $n \in \mathbb{N}$, we define two numbers

$$s_n = a_n + b_n \qquad \text{and} \qquad d_n = a_n - b_n.$$

Then, for each $n \in \mathbb{N}$, we have

$$a_n = \frac{s_n + d_n}{2} \qquad \text{and} \qquad b_n = \frac{s_n - d_n}{2} \tag{576}$$

(by Lemma A.4.5, applied to a_n, b_n, s_n and d_n instead of a, b, s and d).

In view of

$$\begin{aligned}
\left(A^2 - 2uA - \left(v^2 - u^2\right) I_2\right) w_n &= \underbrace{A^2 w_n}_{\substack{=AAw_n = Aw_{n+1} \\ (\text{since } Aw_n = w_{n+1})}} - 2u \underbrace{Aw_n}_{=w_{n+1}} - \left(v^2 - u^2\right) \underbrace{I_2 w_n}_{=w_n} \\
&= \underbrace{Aw_{n+1}}_{=w_{n+2}} - 2uw_{n+1} - \left(v^2 - u^2\right) w_n \\
&= w_{n+2} - 2uw_{n+1} - \left(v^2 - u^2\right) w_n \\
&= \begin{pmatrix} a_{n+2} \\ b_{n+2} \end{pmatrix} - 2u \begin{pmatrix} a_{n+1} \\ b_{n+1} \end{pmatrix} - \left(v^2 - u^2\right) \begin{pmatrix} a_n \\ b_n \end{pmatrix} \\
&= \begin{pmatrix} a_{n+2} - 2ua_{n+1} - \left(v^2 - u^2\right) a_n \\ b_{n+2} - 2ub_{n+1} - \left(v^2 - u^2\right) b_n \end{pmatrix},
\end{aligned}$$

this rewrites as $\begin{pmatrix} a_{n+2} - 2ua_{n+1} - \left(v^2 - u^2\right) a_n \\ b_{n+2} - 2ub_{n+1} - \left(v^2 - u^2\right) b_n \end{pmatrix} = 0$. In other words, $a_{n+2} - 2ua_{n+1} - \left(v^2 - u^2\right) a_n = 0$ and $b_{n+2} - 2ub_{n+1} - \left(v^2 - u^2\right) b_n = 0$. Since we have shown this for each $n \in \mathbb{N}$, we thus conclude that both sequences $(a_0, a_1, a_2, \ldots)$ and $(b_0, b_1, b_2, \ldots)$ are $\left(2u, v^2 - u^2\right)$-recurrent. The main advantage of this argument is its generalizability (to more than two sequences and higher-order recurrences).

Now, each integer $n \geq 1$ satisfies

$$a_n = ua_{n-1} + vb_{n-1} \qquad \text{and} \qquad b_n = ub_{n-1} + va_{n-1}$$

(by the recursive definition of the sequences $(a_0, a_1, a_2, \ldots)$ and $(b_0, b_1, b_2, \ldots)$) and therefore

$$\begin{aligned} s_n &= \underbrace{a_n}_{=ua_{n-1}+vb_{n-1}} + \underbrace{b_n}_{=ub_{n-1}+va_{n-1}} = ua_{n-1} + vb_{n-1} + ub_{n-1} + va_{n-1} \\ &= (u+v) \underbrace{(a_{n-1} + b_{n-1})}_{\substack{=s_{n-1} \\ \text{(since } s_{n-1} \text{ was defined as } a_{n-1}+b_{n-1})}} = (u+v)\, s_{n-1} \end{aligned} \tag{577}$$

and

$$\begin{aligned} d_n &= \underbrace{a_n}_{=ua_{n-1}+vb_{n-1}} - \underbrace{b_n}_{=ub_{n-1}+va_{n-1}} = ua_{n-1} + vb_{n-1} - (ub_{n-1} + va_{n-1}) \\ &= (u-v) \underbrace{(a_{n-1} - b_{n-1})}_{\substack{=d_{n-1} \\ \text{(since } d_{n-1} \text{ was defined as } a_{n-1}-b_{n-1})}} = (u-v)\, d_{n-1}. \end{aligned} \tag{578}$$

The equality (577) holding for each $n \geq 1$ shows that the sequence $(s_0, s_1, s_2, \ldots)$ is a geometric progression with ratio $u + v$. Therefore,

$$s_n = (u+v)^n s_0 \qquad \text{for each } n \in \mathbb{N}. \tag{579}$$

[358] Likewise, we find that

$$d_n = (u-v)^n d_0 \qquad \text{for each } n \in \mathbb{N}. \tag{580}$$

[359]

[358] For the sake of completeness, here is the (entirely straightforward) *proof of (579):*

We shall prove (579) by induction on n:

Induction base: We have $s_0 = (u+v)^0 s_0$ (since $\underbrace{(u+v)^0}_{=1} s_0 = s_0$). In other words, (579) holds for $n = 0$.

Induction step: Let $m \in \mathbb{N}$. Assume (as the induction hypothesis) that (579) holds for $n = m$. We must prove that (579) holds for $n = m + 1$.

We have assumed that (579) holds for $n = m$. In other words, we have $s_m = (u+v)^m s_0$. But $m + 1 \geq 1$; hence, (577) (applied to $n = m + 1$) yields $s_{m+1} = (u+v) \underbrace{s_{(m+1)-1}}_{=s_m=(u+v)^m s_0} = \underbrace{(u+v)(u+v)^m}_{=(u+v)^{m+1}} s_0 = (u+v)^{m+1} s_0$. In other words, (579) holds for $n = m + 1$. This completes the induction step. Thus, (579) is proved.

[359] The proof of (580) is analogous to the proof of (579), once the obvious changes are made (viz., we must use (578) instead of (577), and we must replace s_k and $u + v$ by d_k and $u - v$, respectively).

Now, for each $n \in \mathbb{N}$, we have

$$\begin{aligned}
a_n &= \frac{s_n + d_n}{2} \qquad \text{(by (576))} \\
&= \frac{1}{2}\left(\underbrace{s_n}_{\substack{=(u+v)^n s_0 \\ \text{(by (579))}}} + \underbrace{d_n}_{\substack{=(u-v)^n d_0 \\ \text{(by (580))}}}\right) \\
&= \frac{1}{2}\left((u+v)^n \underbrace{s_0}_{\substack{=a_0+b_0 \\ \text{(by the definition of } s_0)}} + (u-v)^n \underbrace{d_0}_{\substack{=a_0-b_0 \\ \text{(by the definition of } d_0)}}\right) \\
&= \frac{1}{2}\left((u+v)^n (a_0+b_0) + (u-v)^n (a_0-b_0)\right)
\end{aligned}$$

and

$$\begin{aligned}
b_n &= \frac{s_n - d_n}{2} \qquad \text{(by (576))} \\
&= \frac{1}{2}\left(\underbrace{s_n}_{\substack{=(u+v)^n s_0 \\ \text{(by (579))}}} - \underbrace{d_n}_{\substack{=(u-v)^n d_0 \\ \text{(by (580))}}}\right) \\
&= \frac{1}{2}\left((u+v)^n \underbrace{s_0}_{\substack{=a_0+b_0 \\ \text{(by the definition of } s_0)}} - (u-v)^n \underbrace{d_0}_{\substack{=a_0-b_0 \\ \text{(by the definition of } d_0)}}\right) \\
&= \frac{1}{2}\left((u+v)^n (a_0+b_0) - (u-v)^n (a_0-b_0)\right).
\end{aligned}$$

This solves Exercise 4.8.7. □

□

A.4.8. Discussion of Exercise 4.8.8

Discussion of Exercise 4.8.8. Our solution to Exercise 4.8.8 will rely on the following two simple lemmas:

Lemma A.4.6. Let $q \geq -1$ be a real number. Then,

$$\lfloor q+1 \rfloor = \min\{k \in \mathbb{N} \mid k > q\}. \tag{581}$$

(In words: $\lfloor q+1 \rfloor$ is the smallest nonnegative integer that is greater than q.)

Proof of Lemma A.4.6. Applying (1) to $x = q+1$, we find $\lfloor q+1 \rfloor \leq q+1 < \lfloor q+1 \rfloor + 1$. Thus, $q+1 < \lfloor q+1 \rfloor + 1$, so that $q < \lfloor q+1 \rfloor$. Therefore, $\lfloor q+1 \rfloor > q \geq -1$. Also, $\lfloor q+1 \rfloor$ is an integer (since the floor of any real number is an integer). Therefore, from $\lfloor q+1 \rfloor > -1$, we obtain $\lfloor q+1 \rfloor \geq 0$. Hence, $\lfloor q+1 \rfloor \in \mathbb{N}$ (since $\lfloor q+1 \rfloor$ is an integer). Thus, $\lfloor q+1 \rfloor$ is a $k \in \mathbb{N}$ satisfying $k > q$ (since $\lfloor q+1 \rfloor > q$). In other words,

$$\lfloor q+1 \rfloor \in \{k \in \mathbb{N} \mid k > q\}. \tag{582}$$

Now, let s be any element of the set $\{k \in \mathbb{N} \mid k > q\}$. Thus, s is a $k \in \mathbb{N}$ satisfying $k > q$. In other words, $s \in \mathbb{N}$ and $s > q$. If we had $s < \lfloor q+1 \rfloor$, then we would have $s \leq \lfloor q+1 \rfloor - 1$ (since s and $\lfloor q+1 \rfloor$ are integers), which would entail $s+1 \leq \lfloor q+1 \rfloor \leq q+1$, which would contradict $\underbrace{s}_{>q} + 1 > q+1$. Hence, we cannot have $s < \lfloor q+1 \rfloor$. Thus, we have $s \geq \lfloor q+1 \rfloor$.

Forget that we fixed s. We thus have shown that $s \geq \lfloor q+1 \rfloor$ for every $s \in \{k \in \mathbb{N} \mid k > q\}$. In other words, every element of the set $\{k \in \mathbb{N} \mid k > q\}$ is $\geq \lfloor q+1 \rfloor$. In other words, $\lfloor q+1 \rfloor$ is smaller or equal to any element of the set $\{k \in \mathbb{N} \mid k > q\}$. Since $\lfloor q+1 \rfloor$ itself is an element of this set (by (582)), we thus conclude that $\lfloor q+1 \rfloor$ is the smallest element of this set. In other words, $\lfloor q+1 \rfloor = \min\{k \in \mathbb{N} \mid k > q\}$. Thus, (581) is proved. In other words, we have proved Lemma A.4.6. □

Lemma A.4.7. Let $m \in \mathbb{N}$. Then,

$$m = \min\{k \in \mathbb{N} \mid k \geq m\}.$$

Proof of Lemma A.4.7. This is even more obvious than Lemma A.4.6. The number m is an element of the set $\{k \in \mathbb{N} \mid k \geq m\}$ (since $m \in \mathbb{N}$ and $m \geq m$), but every element of this set is $\geq m$ (because every element of this set is a $k \in \mathbb{N}$ satisfying $k \geq m$). Hence, the number m is the smallest element of the set $\{k \in \mathbb{N} \mid k \geq m\}$. In other words, $m = \min\{k \in \mathbb{N} \mid k \geq m\}$. This proves Lemma A.4.7. □

Solution to Exercise 4.8.8. Let us first fix $n \in \mathbb{N}$. Then, $\sqrt{8n+1} > 0$; therefore, we have $\dfrac{\sqrt{8n+1}-1}{2} > \dfrac{0-1}{2} > -1$. Hence, (581) (applied to $q = \dfrac{\sqrt{8n+1}-1}{2}$) yields

$$\left\lfloor \frac{\sqrt{8n+1}-1}{2} + 1 \right\rfloor = \min\left\{k \in \mathbb{N} \mid k > \frac{\sqrt{8n+1}-1}{2}\right\}. \tag{583}$$

Let us now fix $k \in \mathbb{N}$. We shall prove the following logical equivalence:

$$(a_n \leq k) \iff \left(k > \frac{\sqrt{8n+1}-1}{2}\right). \tag{584}$$

In order to do so, let us take a look at the sequence $(a_0, a_1, a_2, \ldots)$. The sequence $(a_0, a_1, a_2, \ldots)$ contains each positive integer i exactly i times. Thus, it contains

the integer 1 exactly once, the integer 2 exactly twice, the integer 3 exactly thrice, and so on. Hence, this sequence contains a total of $1+2+\cdots+k$ entries that are $\leq k$. Since this sequence is furthermore weakly increasing, we know that these $1+2+\cdots+k$ entries are precisely the first $1+2+\cdots+k$ entries of the sequence $(a_0, a_1, a_2, \ldots)$ (since otherwise, the sequence would have an entry $a_i > k$ to the left of an entry $a_j \leq k$; but this would contradict the fact that the sequence is weakly increasing). In other words, they are the entries $a_0, a_1, \ldots, a_{(1+2+\cdots+k)-1}$.

Thus we have shown that the entries of the sequence $(a_0, a_1, a_2, \ldots)$ that are $\leq k$ are precisely the entries $a_0, a_1, \ldots, a_{(1+2+\cdots+k)-1}$. In other words, an entry a_m of the sequence $(a_0, a_1, a_2, \ldots)$ is $\leq k$ if and only if $m \in \{0, 1, \ldots, (1+2+\cdots+k)-1\}$. In other words, for each $m \in \mathbb{N}$, we have the logical equivalence

$$(a_m \leq k) \iff (m \in \{0, 1, \ldots, (1+2+\cdots+k)-1\}).$$

Applying this to $m = n$, we obtain the following chain of equivalences:

$$\begin{aligned}
& (a_n \leq k) \\
\iff & (n \in \{0, 1, \ldots, (1+2+\cdots+k)-1\}) \\
\iff & (n \leq (1+2+\cdots+k)-1) \qquad (\text{since } n \in \mathbb{N}) \\
\iff & (n < 1+2+\cdots+k) \qquad (\text{since } n \text{ and } 1+2+\cdots+k \text{ are integers}) \\
\iff & \left(n < \frac{k(k+1)}{2}\right) \qquad \left(\text{since } 1+2+\cdots+k = \frac{k(k+1)}{2}\right) \\
\iff & (8n < 4k(k+1)) \\
& \qquad (\text{here, we have multiplied both sides of our inequality by } 8) \\
\iff & (8n+1 < 4k(k+1)+1) \\
& \qquad (\text{here, we have added } 1 \text{ to both sides of our inequality}) \\
\iff & \left(8n+1 < (2k+1)^2\right) \qquad \left(\text{since } 4k(k+1)+1 = (2k+1)^2\right) \\
\iff & \left(\sqrt{8n+1} < 2k+1\right) \\
& \qquad \left(\begin{array}{c}\text{here, we have taken the square root on both sides} \\ \text{of the inequality; this is legitimate because } \sqrt{8n+1} \\ \text{and } 2k+1 \text{ are positive reals}\end{array}\right) \\
\iff & \left(\sqrt{8n+1} - 1 < 2k\right) \\
& \qquad (\text{here, we have subtracted } 1 \text{ from both sides of the inequality}) \\
\iff & \left(\frac{\sqrt{8n+1}-1}{2} < k\right) \\
& \qquad (\text{here, we have divided both sides of the inequality by } 2) \\
\iff & \left(k > \frac{\sqrt{8n+1}-1}{2}\right).
\end{aligned}$$

Thus, the equivalence (584) is proved.

Now, forget that we fixed k. We thus have proved the equivalence (584) for each $k \in \mathbb{N}$. Now, a_n is a positive integer (since $(a_0, a_1, a_2, \ldots)$ is a sequence of positive integers). Hence, $a_n \in \mathbb{N}$. Thus, Lemma A.4.7 (applied to $m = a_n$) yields

$$\begin{aligned}
a_n &= \min\{k \in \mathbb{N} \mid k \geq a_n\} \\
&= \min\{k \in \mathbb{N} \mid a_n \leq k\} \\
&\qquad\qquad \text{(since the statement "}k \geq a_n\text{" is equivalent to "}a_n \leq k\text{")} \\
&= \min\left\{k \in \mathbb{N} \mid k > \frac{\sqrt{8n+1}-1}{2}\right\} \\
&\qquad\qquad \left(\begin{array}{c}\text{since the statement "}a_n \leq k\text{" is equivalent} \\ \text{to "}k > \dfrac{\sqrt{8n+1}-1}{2}\text{" (by (584))}\end{array}\right) \\
&= \left\lfloor \frac{\sqrt{8n+1}-1}{2} + 1 \right\rfloor \qquad \text{(by (583))} \\
&= \left\lfloor \frac{1}{2}\sqrt{8n+1} + \frac{1}{2} \right\rfloor \qquad \left(\text{since } \frac{\sqrt{8n+1}-1}{2} + 1 = \frac{1}{2}\sqrt{8n+1} + \frac{1}{2}\right).
\end{aligned}$$

This solves Exercise 4.8.8. □

See also [Engel98, Chapter 5, Exercise 29, Second solution] for a result closely connected to Exercise 4.8.8 (but notice that our sequence begins with a_0 rather than a_1). □

A.4.9. Discussion of Exercise 4.8.9

Discussion of Exercise 4.8.9. **(a)** Assume that $a/b \in \mathbb{Q}$. We must show that $f + g$ is a periodic function.

We know that a and b are positive reals. Hence, a/b is a positive real as well. In other words, $a/b > 0$. But we also have $a/b \in \mathbb{Q}$; hence, we can write a/b in the form $a/b = n/m$ for some integers n and $m \neq 0$. Consider these n and m. From $n/m = a/b > 0$, we conclude that the two integers n and m are nonzero and have the same sign (i.e., are either both positive or both negative). Hence, we can WLOG assume that n and m are positive (since otherwise, we can just replace n and m by $-n$ and $-m$, without changing n/m). Assume this. From $a/b = n/m$, we obtain $ma = nb$. Note that ma is a positive real (since m and a are positive).

The function f is a-periodic. In other words, a is a period of f (by the definition of "a-periodic"). Hence, Theorem 4.7.14 **(c)** (applied to f and m instead of u and n) yields that ma is a period of f. But ma is a period of f if and only if every $x \in \mathbb{R}$ satisfies $f(x) = f(x + ma)$ (by Definition 4.7.10 **(a)**). Hence,

$$\text{every } x \in \mathbb{R} \text{ satisfies } f(x) = f(x + ma) \tag{585}$$

(since ma is a period of f). The same argument (applied to g, b and n instead of f, a and m) shows that

$$\text{every } x \in \mathbb{R} \text{ satisfies } g(x) = g(x+nb). \tag{586}$$

Now, every $x \in \mathbb{R}$ satisfies

$$\begin{aligned}(f+g)(x) &= \underbrace{f(x)}_{\substack{=f(x+ma)\\ \text{(by (585))}}} + \underbrace{g(x)}_{\substack{=g(x+nb)\\ \text{(by (586))}}} = f(x+ma) + g\left(x + \underbrace{nb}_{=ma}\right) \\ &= f(x+ma) + g(x+ma) = (f+g)(x+ma).\end{aligned}$$

Recall that ma is a positive real. Hence, ma is a period of $f+g$ if and only if every $x \in \mathbb{R}$ satisfies $(f+g)(x) = (f+g)(x+ma)$ (by Definition 4.7.10 **(a)**). Hence, ma is a period of $f+g$ (since we have just shown that every $x \in \mathbb{R}$ satisfies $(f+g)(x) = (f+g)(x+ma)$). Therefore, the function $f+g$ has a period (namely, ma). In other words, this function $f+g$ is periodic (by Definition 4.7.10 **(b)**). This solves Exercise 4.8.9 **(a)**.

(b) Assume that $a/b \notin \mathbb{Q}$. We shall find an a-periodic function $f : \mathbb{R} \to \mathbb{R}$ and a b-periodic function $g : \mathbb{R} \to \mathbb{R}$ such that $f+g$ is not periodic.

The easiest way to find such f and g is to define them by setting[360]

$$f(x) = \left[\frac{x}{a} \in \mathbb{Z}\right] \qquad \text{and} \qquad g(x) = \left[\frac{x}{b} \in \mathbb{Z}\right]$$

for each $x \in \mathbb{R}$. Note that each of the functions f and g only takes two different values (namely, 0 and 1). The function f is a-periodic, since each $x \in \mathbb{R}$ satisfies

$$\begin{aligned}f(x+a) &= \left[\frac{x+a}{a} \in \mathbb{Z}\right] = \left[\frac{x}{a} + 1 \in \mathbb{Z}\right] \qquad \left(\text{since } \frac{x+a}{a} = \frac{x}{a} + 1\right) \\ &= \left[\frac{x}{a} \in \mathbb{Z}\right] \qquad \left(\text{since } \frac{x}{a} + 1 \in \mathbb{Z} \text{ holds if and only if } \frac{x}{a} \in \mathbb{Z}\right) \\ &= f(x).\end{aligned}$$

Likewise, the function g is b-periodic.

We now claim that the function $f+g$ is not periodic. Indeed, assume the contrary. Thus, $f+g$ is periodic. In other words, there exists a period d of $f+g$. Consider this d. Since d is a period of $f+g$, we have $(f+g)(0) = (f+g)(0+d) = (f+g)(d) = f(d) + g(d)$. Hence,

$$f(d) + g(d) = (f+g)(0) = \underbrace{f(0)}_{=1} + \underbrace{g(0)}_{=1} = 1 + 1 = 2.$$

Since each of $f(d)$ and $g(d)$ is either 0 or 1, we thus conclude that both $f(d)$ and $g(d)$ must equal 1. Thus, in particular, $f(d) = 1$. But the definition of f yields $f(d) = \left[\frac{d}{a} \in \mathbb{Z}\right]$. Hence, $\left[\frac{d}{a} \in \mathbb{Z}\right] = f(d) = 1$, so that $\frac{d}{a} \in \mathbb{Z}$. Likewise, $\frac{d}{b} \in \mathbb{Z}$.

[360] We are using Definition 4.3.19.

Thus, we have shown that $\frac{d}{b}$ and $\frac{d}{a}$ are integers. These integers are furthermore positive (since d, a and b are positive). Hence, their ratio $\frac{d}{b}/\frac{d}{a}$ is a well-defined rational number. In other words, $\frac{d}{b}/\frac{d}{a} \in \mathbb{Q}$. But this contradicts $\frac{d}{b}/\frac{d}{a} = \frac{a}{b} \notin \mathbb{Q}$. This contradiction shows that our assumption was wrong. Hence, we have shown that the function $f+g$ is not periodic. This solves Exercise 4.8.9 **(b)**.

(An alternative choice of functions f and g that results in the same conclusion can be obtained by setting

$$f(x) = \cos\frac{2\pi x}{a} \qquad \text{and} \qquad g(x) = \cos\frac{2\pi x}{b}.$$

Again, these f and g have the property that $(f+g)(d)$ cannot equal $(f+g)(0) = 2$ for $d > 0$, because this would cause $\frac{d}{b}$ and $\frac{d}{a}$ to be integers.) □

A.4.10. Discussion of Exercise 4.8.10

Discussion of Exercise 4.8.10. Exercise 4.8.10 is known as the *Star of David theorem* (conjectured by Henry W. Gould in 1971 – see [Gould72] – and proved soon after by A. P. Hillman and V. E. Hoggatt, Jr. [HilHog72]). It owes its name to the fact that the six binomial coefficients it involves are arranged as follows on Pascal's triangle:

$$\begin{tikzcd}
 & \binom{n-1}{k-1} \arrow[dr, dash] \arrow[dd, dash] & \binom{n-1}{k} \arrow[dl, dash] \arrow[dd, dash] & \\
\binom{n}{k-1} \arrow[ur, dash] \arrow[dr, dash] & & & \binom{n}{k+1} \arrow[ul, dash] \arrow[dl, dash] \\
 & \binom{n+1}{k} \arrow[ur, dash] & \binom{n+1}{k+1} \arrow[ul, dash] &
\end{tikzcd}$$

(where edges have been drawn to connect binomial coefficients that appear on the same side of the equation). It can be contrasted with the similar identity

$$\binom{n-1}{k-1}\binom{n}{k+1}\binom{n+1}{k} = \binom{n-1}{k}\binom{n}{k-1}\binom{n+1}{k+1} \tag{587}$$

which also holds for all $n, k \in \mathbb{Z}$. However, the latter identity (which is sometimes also called "Star of David theorem", as it involves the exact same six binomial coefficients as Exercise 4.8.10) can be proved by straightforward computations (see [19fco-hw2s, Exercise 5] for the proof), and can be generalized to

$$\binom{n-r}{k-s}\binom{n}{k+r}\binom{n+s}{k} = \binom{n-r}{k}\binom{n}{k-s}\binom{n+s}{k+r}$$

for any $r, s \in \mathbb{Z}$, whereas Exercise 4.8.10 has neither a straightforward proof nor such a generalization. This should not be very surprising: gcds are more complicated than products.

This being said, let me give a solution to Exercise 4.8.10 now. It will rely on the following lemmas:

Lemma A.4.8. Let $u, v, w \in \mathbb{Z}$ and $m \in \mathbb{Z}$. Then, we have the following logical equivalence:

$$(m \mid u \text{ and } m \mid v \text{ and } m \mid w) \iff (m \mid \gcd(u, v, w)).$$

Proof of Lemma A.4.8. Set $b_1 = u$ and $b_2 = v$ and $b_3 = w$. Thus, we have defined three integers b_1, b_2, b_3. Now, we have the following chain of equivalences:

$$\begin{aligned}
&\left(m \mid \underbrace{u}_{=b_1} \text{ and } m \mid \underbrace{v}_{=b_2} \text{ and } m \mid \underbrace{w}_{=b_3}\right) \\
&\iff (m \mid b_1 \text{ and } m \mid b_2 \text{ and } m \mid b_3) \\
&\iff \left(m \mid \gcd\left(\underbrace{b_1}_{=u}, \underbrace{b_2}_{=v}, \underbrace{b_3}_{=w}\right)\right) \qquad \text{(by Theorem 3.4.14, applied to } k = 3) \\
&\iff (m \mid \gcd(u, v, w)).
\end{aligned}$$

This proves Lemma A.4.8. □

The next lemma collects some useful relations between adjacent entries in Pascal's triangle:

Lemma A.4.9. Let n and k be any numbers. Then:

(a) We have $n\binom{n-1}{k-1} = k\binom{n}{k}$.

(b) We have $(k+1)\binom{n}{k+1} = (n-k)\binom{n}{k}$.

(c) We have $n\binom{n-1}{k} = (n-k)\binom{n}{k}$.

Proof of Lemma A.4.9. There are many ways to prove Lemma A.4.9; in particular, each of the three parts of this lemma can be proved separately. The following line of argument is probably the quickest:

(c) We must prove that $n\binom{n-1}{k} = (n-k)\binom{n}{k}$. If $k \notin \mathbb{N}$, then this equality is

obvious[361]. Hence, for the rest of this proof, we WLOG assume that $k \in \mathbb{N}$. Thus, (117) (applied to $n-1$ instead of n) yields

$$\begin{aligned}\binom{n-1}{k} &= \frac{(n-1)\left((n-1)-1\right)\left((n-2)-2\right)\cdots\left((n-1)-k+1\right)}{k!} \\ &= \frac{(n-1)(n-2)(n-3)\cdots(n-k)}{k!} = \frac{1}{k!}\cdot(n-1)(n-2)(n-3)\cdots(n-k).\end{aligned}$$

Multiplying both sides of this equality by n, we find

$$\begin{aligned}n\binom{n-1}{k} &= n\cdot\frac{1}{k!}\cdot(n-1)(n-2)(n-3)\cdots(n-k) \\ &= \frac{1}{k!}\cdot\underbrace{n\cdot(n-1)(n-2)(n-3)\cdots(n-k)}_{\substack{=n(n-1)(n-2)\cdots(n-k) \\ =n(n-1)(n-2)\cdots(n-k+1)\cdot(n-k)}} \\ &= \frac{1}{k!}\cdot n(n-1)(n-2)\cdots(n-k+1)\cdot(n-k) \\ &= (n-k)\cdot\underbrace{\frac{n(n-1)(n-2)\cdots(n-k+1)}{k!}}_{\substack{=\binom{n}{k} \\ \text{(by (117))}}} = (n-k)\binom{n}{k}.\end{aligned}$$

This proves Lemma A.4.9 **(c)**.

(a) Theorem 4.3.7 yields $\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$. Thus,

$$\binom{n-1}{k-1} = \binom{n}{k} - \binom{n-1}{k}.$$

Multiplying both sides of this equality by n, we obtain

$$\begin{aligned}n\binom{n-1}{k-1} &= n\left(\binom{n}{k} - \binom{n-1}{k}\right) = n\binom{n}{k} - \underbrace{n\binom{n-1}{k}}_{\substack{=(n-k)\binom{n}{k} \\ \text{(by Lemma A.4.9 (c))}}} \\ &= n\binom{n}{k} - (n-k)\binom{n}{k} = \underbrace{(n-(n-k))}_{=k}\binom{n}{k} = k\binom{n}{k}.\end{aligned}$$

[361] *Proof.* Assume that $k \notin \mathbb{N}$. Then, $\binom{n}{k} = 0$ (by (118)) and $\binom{n-1}{k} = 0$ (likewise). Hence, $n\underbrace{\binom{n-1}{k}}_{=0} = 0$ and $(n-k)\underbrace{\binom{n}{k}}_{=0} = 0$. Comparing these two equalities, we obtain $n\binom{n-1}{k} = (n-k)\binom{n}{k}$. Thus, we have proved $n\binom{n-1}{k} = (n-k)\binom{n}{k}$ under the assumption that $k \notin \mathbb{N}$.

This proves Lemma A.4.9 **(a)**.

(b) Lemma A.4.9 **(a)** (applied to $k+1$ instead of k) yields

$$n\binom{n-1}{(k+1)-1} = (k+1)\binom{n}{k+1}.$$

Hence,

$$\begin{aligned}(k+1)\binom{n}{k+1} &= n\binom{n-1}{(k+1)-1} = n\binom{n-1}{k} \qquad (\text{since } (k+1)-1=k) \\ &= (n-k)\binom{n}{k} \qquad (\text{by Lemma A.4.9 } \textbf{(c)}).\end{aligned}$$

This proves Lemma A.4.9 **(b)**. □

Solution to Exercise 4.8.10. Define seven numbers A, B, C, D, E, F and M by

$$\begin{aligned}A &= \binom{n-1}{k-1}, \qquad B = \binom{n-1}{k}, \qquad C = \binom{n}{k+1}, \\ D &= \binom{n}{k-1}, \qquad E = \binom{n+1}{k}, \qquad F = \binom{n+1}{k+1}, \qquad M = \binom{n}{k}.\end{aligned}$$

All these seven numbers A, B, C, D, E, F and M belong to $\mathbb{Z}$ (because of Theorem 4.3.15); in other words, they are integers. Hence, speaking of $\gcd(A,C,E)$ and $\gcd(B,D,F)$ makes sense.

We set $x = \gcd(A,C,E)$ and $y = \gcd(B,D,F)$. Then, x and y are nonnegative integers (by Proposition 3.4.3 **(a)**). Our goal is to show that $x = y$. We shall achieve this by proving that $x \mid y$ and $y \mid x$ (just as in our proof of Theorem 3.4.8).

We have $x \mid x = \gcd(A,C,E)$. But Lemma A.4.8 (applied to A, C, E and x instead of u, v, w and m) shows that we have the following logical equivalence:

$$(x \mid A \text{ and } x \mid C \text{ and } x \mid E) \iff (x \mid \gcd(A,C,E)).$$

Hence, we have $(x \mid A \text{ and } x \mid C \text{ and } x \mid E)$ (since we have $x \mid \gcd(A,C,E)$). The same argument (applied to y, B, D and F instead of x, A, C and E) shows that we have $(y \mid B \text{ and } y \mid D \text{ and } y \mid F)$.

Now, we shall prove a few relations between A, B, C, D, E, F and M. First, we claim that

$$M - A = B; \tag{588}$$

$$E - M = D, \tag{589}$$

$$C + M = F. \tag{590}$$

These three equalities are clear from Theorem 4.3.7 upon a look at a picture of Pascal's triangle, in which the entries A, B, C, D, E, F and M are arranged as

follows:

$$\begin{array}{ccc} & A & & B & \\ \\ D & & M & & C \\ \\ & E & & F & \end{array}$$

(where each of the entries A, B, C, D, E, F is adjacent to M). [362]

We also need two less trivial equalities. Namely, we claim that

$$-nA - (k+1)\,C + (n+1-k)\,E = M. \tag{591}$$

[*Proof of (591):* From $A = \binom{n-1}{k-1}$, we obtain

$$\begin{aligned} nA &= n\binom{n-1}{k-1} = k\underbrace{\binom{n}{k}}_{=M} \qquad \text{(by Lemma A.4.9 } \textbf{(a)}\text{)} \\ &= kM. \end{aligned}$$

[362] For the sake of completeness, here are detailed proofs of (588), (589) and (590):

Proof of (588): Theorem 4.3.7 yields

$$\binom{n}{k} = \underbrace{\binom{n-1}{k-1}}_{=A} + \underbrace{\binom{n-1}{k}}_{=B} = A + B.$$

Hence, $M = \binom{n}{k} = A + B$, so that $M - A = B$. This proves (588).

Proof of (589): Theorem 4.3.7 (applied to $n+1$ instead of n) yields

$$\binom{n+1}{k} = \underbrace{\binom{(n+1)-1}{k-1}}_{=\binom{n}{k-1}=D} + \underbrace{\binom{(n+1)-1}{k}}_{=\binom{n}{k}=M} = D + M.$$

Hence, $E = \binom{n+1}{k} = D + M$, so that $E - M = D$. This proves (589).

Proof of (590): Theorem 4.3.7 (applied to $n+1$ and $k+1$ instead of n and k) yields

$$\binom{n+1}{k+1} = \underbrace{\binom{(n+1)-1}{(k+1)-1}}_{=\binom{n}{k}=M} + \underbrace{\binom{(n+1)-1}{k+1}}_{=\binom{n}{k+1}=C} = M + C = C + M.$$

Hence, $F = \binom{n+1}{k+1} = C + M$. This proves (590).

From $C = \binom{n}{k+1}$, we obtain

$$(k+1)\,C = (k+1)\binom{n}{k+1} = (n-k)\underbrace{\binom{n}{k}}_{=M} \qquad \text{(by Lemma A.4.9 \textbf{(b)})}$$
$$= (n-k)\,M.$$

Also, we have $M = \binom{n}{k} = \binom{(n+1)-1}{k}$ (since $n = (n+1)-1$), and therefore

$$(n+1)\,M = (n+1)\binom{(n+1)-1}{k} = (n+1-k)\underbrace{\binom{n+1}{k}}_{=E}$$
$$\text{(by Lemma A.4.9 \textbf{(c)}, applied to } n+1 \text{ instead of } n\text{)}$$
$$= (n+1-k)\,E,$$

so that

$$(n+1-k)\,E = (n+1)\,M.$$

Now,

$$-\underbrace{nA}_{=kM} - \underbrace{(k+1)\,C}_{=(n-k)M} + \underbrace{(n+1-k)\,E}_{=(n+1)M} = -kM - (n-k)\,M + (n+1)\,M = M.$$

This proves (591).]

Next, we claim that

$$-nB - (n+1-k)\,D + (k+1)\,F = M. \tag{592}$$

[*Proof of (592):* We have

$$\begin{aligned}
&-n\underbrace{B}_{\substack{=M-A\\ \text{(by (588))}}} - (n+1-k)\underbrace{D}_{\substack{=E-M\\ \text{(by (589))}}} + (k+1)\underbrace{F}_{\substack{=C+M\\ \text{(by (590))}}} \\
&= -n\,(M-A) - (n+1-k)\,(E-M) + (k+1)\,(C+M) \\
&= -nM + nA - (n+1-k)\,E + (n+1-k)\,M + (k+1)\,C + (k+1)\,M \\
&= \underbrace{\left(-nM + (n+1-k)\,M + (k+1)\,M\right)}_{=2M} - \underbrace{\left(-nA - (k+1)\,C + (n+1-k)\,E\right)}_{\substack{=M\\ \text{(by (591))}}} \\
&= 2M - M = M.
\end{aligned}$$

This proves (592).]

Now, we have some modular arithmetic to do.

Recall that $(x \mid A \text{ and } x \mid C \text{ and } x \mid E)$. Hence, $A \equiv 0 \bmod x$ (since $x \mid A$) and $C \equiv 0 \bmod x$ (since $x \mid C$) and $E \equiv 0 \bmod x$ (since $x \mid E$). Now, (591) yields

$$\begin{aligned} M &= -n \underbrace{A}_{\equiv 0 \bmod x} - (k+1) \underbrace{C}_{\equiv 0 \bmod x} + (n+1-k) \underbrace{E}_{\equiv 0 \bmod x} \\ &\equiv -n \cdot 0 - (k+1) \cdot 0 + (n+1-k) \cdot 0 = 0 \bmod x. \end{aligned}$$

Now, (588) yields $B = \underbrace{M}_{\equiv 0 \bmod x} - \underbrace{A}_{\equiv 0 \bmod x} \equiv 0 - 0 = 0 \bmod x$, so that $x \mid B$. Also, (589) yields $D = \underbrace{E}_{\equiv 0 \bmod x} - \underbrace{M}_{\equiv 0 \bmod x} \equiv 0 - 0 = 0 \bmod x$, so that $x \mid D$. Finally, (590) yields $F = \underbrace{C}_{\equiv 0 \bmod x} + \underbrace{M}_{\equiv 0 \bmod x} \equiv 0 + 0 = 0 \bmod x$, so that $x \mid F$. But Lemma A.4.8 (applied to B, D, F and x instead of u, v, w and m) shows that we have the following logical equivalence:

$$(x \mid B \text{ and } x \mid D \text{ and } x \mid F) \iff (x \mid \gcd(B, D, F)).$$

Hence, we have $x \mid \gcd(B, D, F)$ (since we have $(x \mid B \text{ and } x \mid D \text{ and } x \mid F)$). In other words, $x \mid y$ (since $y = \gcd(B, D, F)$).

On the other hand, recall that $(y \mid B \text{ and } y \mid D \text{ and } y \mid F)$. Hence, $B \equiv 0 \bmod y$ (since $y \mid B$) and $D \equiv 0 \bmod y$ (since $y \mid D$) and $F \equiv 0 \bmod y$ (since $y \mid F$). Now, (592) yields

$$\begin{aligned} M &= -n \underbrace{B}_{\equiv 0 \bmod y} - (n+1-k) \underbrace{D}_{\equiv 0 \bmod y} + (k+1) \underbrace{F}_{\equiv 0 \bmod y} \\ &\equiv -n \cdot 0 - (n+1-k) \cdot 0 + (k+1) \cdot 0 = 0 \bmod y. \end{aligned}$$

Now, (588) yields $B = M - A$, so that $A = \underbrace{M}_{\equiv 0 \bmod y} - \underbrace{B}_{\equiv 0 \bmod y} \equiv 0 - 0 = 0 \bmod y$, so that $y \mid A$. Also, (589) yields $D = E - M$, so that $E = \underbrace{D}_{\equiv 0 \bmod y} + \underbrace{M}_{\equiv 0 \bmod y} \equiv 0 + 0 = 0 \bmod y$, so that $y \mid E$. Finally, (590) yields $F = C + M$, so that $C = \underbrace{F}_{\equiv 0 \bmod y} - \underbrace{M}_{\equiv 0 \bmod y} \equiv 0 - 0 = 0 \bmod y$, so that $y \mid C$. But Lemma A.4.8 (applied to A, C, E and y instead of u, v, w and m) shows that we have the following logical equivalence:

$$(y \mid A \text{ and } y \mid C \text{ and } y \mid E) \iff (y \mid \gcd(A, C, E)).$$

Hence, we have $y \mid \gcd(A, C, E)$ (since we have $(y \mid A \text{ and } y \mid C \text{ and } y \mid E)$). In other words, $y \mid x$ (since $x = \gcd(A, C, E)$).

Thus we know that $x \mid y$ and $y \mid x$. Hence, Proposition 3.1.3 **(c)** (applied to $a = x$ and $b = y$) yields $|x| = |y|$. But x is nonnegative; thus, $|x| = x$. Similarly, $|y| = y$.

Hence, $x = |x| = |y| = y$. In view of

$$x = \gcd\left(\underbrace{A}_{=\binom{n-1}{k-1}}, \underbrace{C}_{=\binom{n}{k+1}}, \underbrace{E}_{=\binom{n+1}{k}}\right) = \gcd\left(\binom{n-1}{k-1}, \binom{n}{k+1}, \binom{n+1}{k}\right)$$

and

$$y = \gcd\left(\underbrace{B}_{=\binom{n-1}{k}}, \underbrace{D}_{=\binom{n}{k-1}}, \underbrace{F}_{=\binom{n+1}{k+1}}\right) = \gcd\left(\binom{n-1}{k}, \binom{n}{k-1}, \binom{n+1}{k+1}\right),$$

this rewrites as

$$\gcd\left(\binom{n-1}{k-1}, \binom{n}{k+1}, \binom{n+1}{k}\right) = \gcd\left(\binom{n-1}{k}, \binom{n}{k-1}, \binom{n+1}{k+1}\right).$$

This solves Exercise 4.8.10. □

□

A.5. Homework set #4 discussion

The following are discussions of the problems on homework set #4 (Section 4.10).

A.5.1. Discussion of Exercise 4.10.1

Discussion of Exercise 4.10.1. We shall use the following simple lemma:

Lemma A.5.1. Let a, b, c, n be four integers such that $c \neq 0$ and $ac \equiv bc \bmod nc$. Then, $a \equiv b \bmod n$.

Proof of Lemma A.5.1. We have $ac \equiv bc \bmod nc$. In other words, $nc \mid ac - bc$. In other words, $nc \mid (a-b)\,c$ (since $ac - bc = (a-b)\,c$). But Proposition 3.1.5 (applied to n and $a-b$ instead of a and b) yields that $n \mid a-b$ holds if and only if $nc \mid (a-b)\,c$. Thus, $n \mid a-b$ holds (since $nc \mid (a-b)\,c$ holds). In other words, $a \equiv b \bmod n$. This proves Lemma A.5.1. □

Solution to Exercise 4.10.1. We know that $k!$ is a positive integer; hence, $k! \neq 0$.

Theorem 4.3.15 yields that $\binom{n}{k} \in \mathbb{Z}$ for each $n \in \mathbb{Z}$. Hence, the sequence $(a_0, a_1, a_2, \ldots)$ is well-defined.

Let $v = uk!$. Then, v is a positive integer (since u and $k!$ are positive integers). We shall now prove that every $i \in \mathbb{N}$ satisfies $a_i = a_{i+v}$.

Indeed, let $i \in \mathbb{N}$. Set $j = i + v$. Thus, $j = i + \underbrace{v}_{\equiv 0 \bmod v} \equiv i \bmod v$.

Applying (117) to $n = i$, we obtain

$$\binom{i}{k} = \frac{i(i-1)(i-2)\cdots(i-k+1)}{k!}.$$

Multiplying both sides of this equality by $k!$, we find

$$\binom{i}{k} \cdot k! = i(i-1)(i-2)\cdots(i-k+1) = \prod_{s=0}^{k-1}(i-s). \tag{593}$$

The same argument (applied to j instead of i) yields

$$\binom{j}{k} \cdot k! = \prod_{s=0}^{k-1}(j-s). \tag{594}$$

However, for each $s \in \{0, 1, \ldots, k-1\}$, we have $\underbrace{j}_{\equiv i \bmod v} - s \equiv i - s \bmod v$. Multiplying these congruences for all $s \in \{0, 1, \ldots, k-1\}$ (that is, applying (41) to $S = \{0, 1, \ldots, k-1\}$ and $n = v$ and $a_s = j - s$ and $b_s = i - s$), we obtain

$$\prod_{s=0}^{k-1}(j-s) \equiv \prod_{s=0}^{k-1}(i-s) \bmod v.$$

In view of (593) and (594), we can rewrite this as

$$\binom{j}{k} \cdot k! \equiv \binom{i}{k} \cdot k! \bmod v.$$

In view of $v = uk!$, this rewrites as

$$\binom{j}{k} \cdot k! \equiv \binom{i}{k} \cdot k! \bmod uk!.$$

Hence, Lemma A.5.1 (applied to $a = \binom{j}{k}$, $b = \binom{i}{k}$, $c = k!$ and $n = u$) yields that $\binom{j}{k} \equiv \binom{i}{k} \bmod u$ (since $k! \neq 0$).

Proposition 3.3.4 (applied to u, $\binom{j}{k}$ and $\binom{i}{k}$ instead of n, u and v) shows that $\binom{j}{k} \equiv \binom{i}{k} \bmod u$ if and only if $\binom{j}{k}\%u = \binom{i}{k}\%u$. Thus, we have $\binom{j}{k}\%u = \binom{i}{k}\%u$ (since $\binom{j}{k} \equiv \binom{i}{k} \bmod u$). In other words, $\binom{i}{k}\%u = \binom{j}{k}\%u$.

But the definition of a_i yields $a_i = \binom{i}{k} \% u$. Similarly, $a_j = \binom{j}{k} \% u$. Now, $a_i = \binom{i}{k} \% u = \binom{j}{k} \% u = a_j = a_{i+v}$ (since $j = i + v$).

Forget that we fixed i. We thus have shown that every $i \in \mathbb{N}$ satisfies $a_i = a_{i+v}$. In other words, the sequence $(a_0, a_1, a_2, \ldots)$ is v-periodic (by the definition of "v-periodic"). In other words, the sequence $(a_0, a_1, a_2, \ldots)$ is $uk!$-periodic (since $v = uk!$). This solves Exercise 4.10.1. □

□

A.5.2. Discussion of Exercise 4.10.2

Discussion of Exercise 4.10.2. Exercise 4.10.2 generalizes Exercise 4.9.7 by removing the condition $x_1 = 1$. This extra generality is ridiculously shallow: If $(y_0, y_1, y_2, \ldots)$ is the unique (u, v)-recurrent sequence of integers with $y_0 = 0$ and $y_1 = 1$, then any (u, v)-recurrent sequence $(x_0, x_1, x_2, \ldots)$ satisfying $x_0 = 0$ can be written as $(\lambda y_0, \lambda y_1, \lambda y_2, \ldots)$ for an appropriate number λ (namely, for $\lambda = x_1$), and therefore Exercise 4.10.2 follows from Exercise 4.9.7 (applied to y_i instead of x_i). Here are the details of this argument:

Solution to Exercise 4.10.2. Define a sequence $(y_0, y_1, y_2, \ldots)$ of integers recursively by setting

$$\begin{aligned} &y_0 = 0, \qquad y_1 = 1, \qquad \text{and} \\ &y_n = u y_{n-1} + v y_{n-2} \qquad \text{for each } n \geq 2. \end{aligned}$$

Then, the sequence $(y_0, y_1, y_2, \ldots)$ is (u, v)-recurrent[363] and satisfies $y_0 = 0$ and $y_1 = 1$. Hence, Exercise 4.9.7 (applied to y_i instead of x_i) yields that all $a, b \in \mathbb{N}$ satisfying $a \mid b$ satisfy

$$y_a \mid y_b. \tag{595}$$

On the other hand, let $\lambda = x_1$. We claim that

$$x_i = \lambda y_i \qquad \text{for each } i \in \mathbb{N}. \tag{596}$$

[*Proof of (596):* Let us prove (596) by strong induction on i:

Induction step: Let $n \in \mathbb{N}$. Assume (as the induction hypothesis) that (596) holds for $i < n$. We must prove that (596) holds for $i = n$. In other words, we must prove that $x_n = \lambda y_n$.

Comparing $x_0 = 0$ with $\lambda \underbrace{y_0}_{=0} = 0$, we obtain $x_0 = \lambda y_0$. Hence, $x_n = \lambda y_n$ is proved if $n = 0$. Thus, for the rest of this proof, we WLOG assume that $n \neq 0$.

[363] since every $n \geq 2$ satisfies $y_n = u y_{n-1} + v y_{n-2}$

From $\lambda \underbrace{y_1}_{=1} = \lambda = x_1$, we obtain $x_1 = \lambda y_1$. Hence, $x_n = \lambda y_n$ is proved if $n = 1$. Thus, for the rest of this proof, we WLOG assume that $n \neq 1$.

We now have $n \neq 0$ and $n \neq 1$. Therefore, $n \geq 2$ (since $n \in \mathbb{N}$). Hence, $y_n = uy_{n-1} + vy_{n-2}$ (since the sequence $(y_0, y_1, y_2, \ldots)$ is (u, v)-recurrent) and $x_n = ux_{n-1} + vx_{n-2}$ (since the sequence $(x_0, x_1, x_2, \ldots)$ is (u, v)-recurrent).

But from $n \geq 2$, we obtain $n - 2 \in \mathbb{N}$. Also, $n - 1 \in \mathbb{N}$ (since $n \geq 2 \geq 1$).

Recall our induction hypothesis, which says that (596) holds for $i < n$. Hence, (596) holds for $i = n - 2$ (since $n - 2 \in \mathbb{N}$ and $n - 2 < n$). In other words, we have $x_{n-2} = \lambda y_{n-2}$. The same argument (applied to $n - 1$ instead of $n - 2$) yields $x_{n-1} = \lambda y_{n-1}$ (since $n - 1 \in \mathbb{N}$ and $n - 1 < n$). Hence,

$$x_n = u \underbrace{x_{n-1}}_{=\lambda y_{n-1}} + v \underbrace{x_{n-2}}_{=\lambda y_{n-2}} = u\lambda y_{n-1} + v\lambda y_{n-2} = \lambda \underbrace{(uy_{n-1} + vy_{n-2})}_{\substack{=y_n \\ (\text{since } y_n = uy_{n-1} + vy_{n-2})}} = \lambda y_n.$$

In other words, (596) holds for $i = n$. This completes the induction step. Thus, (596) is proved.]

Now, let $a, b \in \mathbb{N}$ satisfy $a \mid b$. We must prove that $x_a \mid x_b$. But (596) (applied to $i = a$) yields $x_a = \lambda y_a$. Similarly, $x_b = \lambda y_b$. However, (595) yields $y_a \mid y_b$. In other words, there exists an integer c such that $y_b = y_a c$ (by Definition 3.1.2). Consider this c. We have $x_b = \lambda \underbrace{y_b}_{=y_a c} = \underbrace{\lambda y_a}_{=x_a} c = x_a c$. Since c is an integer, we thus have $x_a \mid x_b$. This solves Exercise 4.10.2. □

□

A.5.3. Discussion of Exercise 4.10.3

Discussion of Exercise 4.10.3. Here is one possible generalization of Exercise 4.9.2:

Proposition A.5.2. Let a and b be two numbers. Let $(x_0, x_1, x_2, \ldots)$ and $(y_0, y_1, y_2, \ldots)$ be two (a, b)-recurrent sequences. Then,

$$x_{n+1}y_{n-1} - x_n y_n = (-b)^{n-1} (x_2 y_0 - x_1 y_1)$$

for any positive integer n.

Exercise 4.9.2 is the particular case of Proposition A.5.2 for $y_i = x_i$.

We can generalize Proposition A.5.2 even further:

Proposition A.5.3. Let a and b be two numbers. Let $(x_0, x_1, x_2, \ldots)$ and $(y_0, y_1, y_2, \ldots)$ be two (a, b)-recurrent sequences. Then,

$$x_{p+1}y_q - x_p y_{q+1} = (-b)^q \left(x_{p-q+1}y_0 - x_{p-q}y_1\right) \tag{597}$$

for any $p, q \in \mathbb{N}$ satisfying $p \geq q$.

Proposition A.5.2 is the particular case of Proposition A.5.3 for $p = n$ and $q = n - 1$.

Proof of Proposition A.5.3. We shall prove (597) by induction on q (without fixing p):

Induction base: For any $p \in \mathbb{N}$ satisfying $p \geq 0$, we have $x_{p+1}y_0 - x_p y_{0+1} = (-b)^0 \left(x_{p-0+1}y_0 - x_{p-0}y_1\right)$ (since $\underbrace{(-b)^0}_{=1} \left(\underbrace{x_{p-0+1}}_{=x_{p+1}} y_0 - \underbrace{x_{p-0}}_{=x_p} \underbrace{y_1}_{=y_{0+1}} \right) = x_{p+1}y_0 - x_p y_{0+1}$).

In other words, (597) holds for $q = 0$ (and any $p \in \mathbb{N}$ satisfying $p \geq 0$).

Induction step: Let m be a positive integer. Assume (as the induction hypothesis) that (597) holds for $q = m - 1$ (and any $p \in \mathbb{N}$ satisfying $p \geq m - 1$). We must prove that (597) holds for $q = m$ (and any $p \in \mathbb{N}$ satisfying $p \geq m$).

We have assumed that (597) holds for $q = m - 1$ (and any $p \in \mathbb{N}$ satisfying $p \geq m - 1$). In other words, we have

$$x_{p+1}y_{m-1} - x_p y_{(m-1)+1} = (-b)^{m-1} \left(x_{p-(m-1)+1}y_0 - x_{p-(m-1)}y_1\right) \tag{598}$$

for any $p \in \mathbb{N}$ satisfying $p \geq m - 1$.

Now, let $p \in \mathbb{N}$ be such that $p \geq m$. The integer m is positive; thus, $m \geq 1$. Hence, $m + 1 \geq 2$. But the sequence $(y_0, y_1, y_2, \ldots)$ is (a, b)-recurrent. In other words, every $n \geq 2$ satisfies $y_n = ay_{n-1} + by_{n-2}$ (by the definition of "(a, b)-recurrent"). Applying this equality to $n = m + 1$, we obtain

$$y_{m+1} = ay_{(m+1)-1} + by_{(m+1)-2} = ay_m + by_{m-1} \tag{599}$$

(since $m + 1 \geq 2$).

Also, $p \geq m \geq 1$. Hence, $p + 1 \geq 2$. But the sequence $(x_0, x_1, x_2, \ldots)$ is (a, b)-recurrent. In other words, every $n \geq 2$ satisfies $x_n = ax_{n-1} + bx_{n-2}$ (by the definition of "(a, b)-recurrent"). Applying this equality to $n = p + 1$, we obtain

$$x_{p+1} = ax_{(p+1)-1} + bx_{(p+1)-2} = ax_p + bx_{p-1}$$

(since $p + 1 \geq 2$). Thus,

$$x_{p+1} - ax_p = bx_{p-1}. \tag{600}$$

Now,

$$\begin{aligned}
& x_{p+1}y_m - x_p \underbrace{y_{m+1}}_{\substack{=ay_m+by_{m-1} \\ \text{(by (599))}}} \\
&= x_{p+1}y_m - x_p\left(ay_m + by_{m-1}\right) = x_{p+1}y_m - x_p a y_m - x_p b y_{m-1} \\
&= \underbrace{\left(x_{p+1} - ax_p\right)}_{\substack{=bx_{p-1} \\ \text{(by (600))}}} y_m - x_p b y_{m-1} = bx_{p-1}y_m - x_p b y_{m-1} \\
&= (-b)\left(x_p y_{m-1} - x_{p-1}y_m\right). \qquad (601)
\end{aligned}$$

However, $p \geq m$ and thus $p-1 \geq m-1 \geq 0$ (since $m \geq 1$). In other words, $p-1 \in \mathbb{N}$. Also, $p-1 \geq m-1$. Hence, we can apply (598) to $p-1$ instead of p. We thus obtain

$$x_{(p-1)+1}y_{m-1} - x_{p-1}y_{(m-1)+1} = (-b)^{m-1}\left(x_{(p-1)-(m-1)+1}y_0 - x_{(p-1)-(m-1)}y_1\right).$$

This rewrites as

$$x_p y_{m-1} - x_{p-1}y_m = (-b)^{m-1}\left(x_{p-m+1}y_0 - x_{p-m}y_1\right)$$

(since $(p-1)+1 = p$ and $(m-1)+1 = m$ and $(p-1)-(m-1) = p-m$). Thus, (601) becomes

$$\begin{aligned}
&x_{p+1}y_m - x_p y_{m+1} \\
&= (-b)\underbrace{\left(x_p y_{m-1} - x_{p-1}y_m\right)}_{=(-b)^{m-1}\left(x_{p-m+1}y_0 - x_{p-m}y_1\right)} = \underbrace{(-b)(-b)^{m-1}}_{=(-b)^m}\left(x_{p-m+1}y_0 - x_{p-m}y_1\right) \\
&= (-b)^m\left(x_{p-m+1}y_0 - x_{p-m}y_1\right).
\end{aligned}$$

Forget now that we fixed p. We thus have shown that

$$x_{p+1}y_m - x_p y_{m+1} = (-b)^m\left(x_{p-m+1}y_0 - x_{p-m}y_1\right)$$

holds for any $p \in \mathbb{N}$ satisfying $p \geq m$. In other words, (597) holds for $q = m$ (and any $p \in \mathbb{N}$ satisfying $p \geq m$). This completes the induction step. Thus, (597) is proved. In other words, Proposition A.5.3 is proved. □

This solves Exercise 4.10.3. (An alternative proof of Proposition A.5.3 can be given using the matrix approach, similar to our second solution to Exercise 4.9.2.) □

A.5.4. Discussion of Exercise 4.10.4

Discussion of Exercise 4.10.4. There are several ways of solving this. Here is one:

Solution to Exercise 4.10.4. Fix $m \in \mathbb{N}$. We must prove that

$$x_{n-m}y_{n+m} = x_n y_n - (-1)^{n+m} x_m y_m \tag{602}$$

for each $n \in \mathbb{N}$ satisfying $n \geq m$. In other words, we must prove that (602) holds for each integer $n \geq m$.

We shall prove this by induction on n:

Induction base: Comparing $\underbrace{x_{m-m}}_{=x_0=0} y_{m+m} = 0$ with $x_m y_m - \underbrace{(-1)^{m+m}}_{\substack{=1 \\ \text{(since } m+m=2m \text{ is even)}}} x_m y_m = x_m y_m - x_m y_m = 0$, we obtain $x_{m-m}y_{m+m} = x_m y_m - (-1)^{m+m} x_m y_m$. In other words, (602) holds for $n = m$.

Induction step: Let k be an integer satisfying $k \geq m$. Assume (as the induction hypothesis) that (602) holds for $n = k$. We must prove that (602) holds for $n = k+1$.

Exercise 4.9.4 (applied to 1, k and k instead of b, n and m) yields

$$1x_0y_{k+k} + x_1y_{k+k+1} = 1x_ky_k + x_{k+1}y_{k+1} = x_ky_k + x_{k+1}y_{k+1}.$$

Hence,

$$\begin{aligned} x_ky_k + x_{k+1}y_{k+1} &= 1x_0y_{k+k} + x_1y_{k+k+1} = x_0y_{k+k} + x_1y_{k+k+1} \\ &= x_0y_{2k} + x_1y_{2k+1} \end{aligned} \tag{603}$$

(since $k + k = 2k$).

From $k \geq m$, we obtain $k - m \in \mathbb{N}$. Hence, Exercise 4.9.4 (applied to 1, $k-m$ and $k+m$ instead of b, n and m) yields

$$\begin{aligned} 1x_0y_{(k-m)+(k+m)} + x_1y_{(k-m)+(k+m)+1} &= 1x_{k-m}y_{k+m} + x_{k-m+1}y_{k+m+1} \\ &= x_{k-m}y_{k+m} + x_{k-m+1}y_{k+m+1}. \end{aligned}$$

Hence,

$$\begin{aligned} x_{k-m}y_{k+m} + x_{k-m+1}y_{k+m+1} &= 1x_0y_{(k-m)+(k+m)} + x_1y_{(k-m)+(k+m)+1} \\ &= x_0y_{(k-m)+(k+m)} + x_1y_{(k-m)+(k+m)+1} \\ &= x_0y_{2k} + x_1y_{2k+1} \end{aligned}$$

(since $(k-m) + (k+m) = 2k$). Comparing this with (603), we obtain

$$x_{k-m}y_{k+m} + x_{k-m+1}y_{k+m+1} = x_ky_k + x_{k+1}y_{k+1}.$$

In other words,

$$x_{k-m+1}y_{k+m+1} = x_ky_k + x_{k+1}y_{k+1} - x_{k-m}y_{k+m}. \tag{604}$$

We have assumed that (602) holds for $n = k$. In other words, we have

$$x_{k-m}y_{k+m} = x_ky_k - (-1)^{k+m} x_my_m.$$

Hence, (604) becomes

$$\begin{aligned} x_{k-m+1}y_{k+m+1} &= x_ky_k + x_{k+1}y_{k+1} - \underbrace{x_{k-m}y_{k+m}}_{=x_ky_k-(-1)^{k+m}x_my_m} \\ &= x_ky_k + x_{k+1}y_{k+1} - \left(x_ky_k - (-1)^{k+m} x_my_m\right) \\ &= x_{k+1}y_{k+1} - (-1)^{k+m} x_my_m = x_{k+1}y_{k+1} + \underbrace{\left(-(-1)^{k+m}\right)}_{=(-1)^{k+m+1}} x_my_m \\ &= x_{k+1}y_{k+1} + (-1)^{k+m+1} x_my_m. \end{aligned}$$

In view of $k-m+1=(k+1)-m$ and $k+m+1=(k+1)+m$, we can rewrite this as

$$x_{(k+1)-m}y_{(k+1)+m}=x_{k+1}y_{k+1}+(-1)^{(k+1)+m}x_my_m.$$

In other words, (602) holds for $n=k+1$. This completes the induction step. Hence, (602) is proved. In other words, Exercise 4.10.4 is solved. □

□

A.5.5. Discussion of Exercise 4.10.5

Discussion of Exercise 4.10.5. The solution to Exercise 4.10.5 is rather similar to that of Exercise 3.7.2, which it generalizes. We will take the lazy route and just copy the latter solution, modifying whatever we need to modify. We begin with a lemma that generalizes Lemma A.2.1:

Lemma A.5.4. Let u and v be two integers such that $u\perp v$. Let $(x_0,x_1,x_2,\ldots)$ be a (u,v)-recurrent sequence of integers with $x_0=0$ and $x_1=1$. Let $a,b\in\mathbb{N}$ be such that $a>0$ and $a\leq b$. Then, $\gcd(x_a,x_b)=\gcd(x_a,x_{b-a})$.

Proof of Lemma A.5.4. From $a>0$, we obtain $a\geq 1$ and thus $a-1\in\mathbb{N}$. Hence, Exercise 4.9.3 (applied to u, v, $b-a$, $a-1$ and x_i instead of a, b, n, m and y_i) yields

$$\begin{aligned}x_{(b-a)+(a-1)+1}&=\underbrace{vx_{b-a}}_{=x_{b-a}v}x_{a-1}+x_{(b-a)+1}\underbrace{x_{(a-1)+1}}_{\substack{=x_a\\ \equiv 0\bmod x_a}}\\&\equiv x_{b-a}vx_{a-1}+x_{(b-a)+1}0=x_{b-a}vx_{a-1}\bmod x_a.\end{aligned}$$

In view of $(b-a)+(a-1)+1=b$, this rewrites as $x_b\equiv x_{b-a}vx_{a-1}\bmod x_a$. Hence, Proposition 3.4.4 **(d)** (applied to x_a, x_b and $x_{b-a}vx_{a-1}$ instead of a, b and c) yields

$$\gcd(x_a,x_b)=\gcd(x_a,x_{b-a}vx_{a-1}).\tag{605}$$

But recall the claim (216) that was proved during our solution to Exercise 4.9.8. Applying this claim (216) to $n=a-1$, we obtain

$$x_{a-1}\perp x_{(a-1)+1}\qquad\text{and}\qquad v\perp x_{(a-1)+1}.$$

In view of $(a-1)+1=a$, this rewrites as

$$x_{a-1}\perp x_a\qquad\text{and}\qquad v\perp x_a.$$

Hence, Theorem 3.5.10 (applied to v, x_{a-1} and x_a instead of a, b and c) yields $vx_{a-1}\perp x_a$. According to Proposition 3.5.4, this yields $x_a\perp vx_{a-1}$. Hence, Proposition 3.5.18 (applied to x_a, x_{b-a} and vx_{a-1} instead of a, b and c) yields $\gcd(x_a,x_{b-a}vx_{a-1})=\gcd(x_a,x_{b-a})$. Thus, (605) becomes

$$\gcd(x_a,x_b)=\gcd(x_a,x_{b-a}vx_{a-1})=\gcd(x_a,x_{b-a}).$$

This proves Lemma A.5.4. □

We can now step to the actual solution to Exercise 4.10.5:

Solution to Exercise 4.10.5 (sketched). We use strong induction on $a+b$:

Induction step: Let $k \in \mathbb{N}$. Assume (as the induction hypothesis) that Exercise 4.10.5 is true for $a+b<k$. We must prove that Exercise 4.10.5 is true for $a+b=k$. Let us do this now.

So let $a, b \in \mathbb{N}$ be such that $a+b=k$. We must show that $\gcd(x_a, x_b) = \left|x_{\gcd(a,b)}\right|$.

Note that a and b play symmetric roles in this claim[364], and thus can be swapped at will. By swapping a and b if necessary, we can ensure that $a \leq b$. Hence, we WLOG assume that $a \leq b$. Thus, $b-a \in \mathbb{N}$.

It is easy to see that our claim $\gcd(x_a, x_b) = \left|x_{\gcd(a,b)}\right|$ holds if $a=0$ [365]. Thus, we are done if $a=0$. Hence, we WLOG assume that $a \neq 0$. Therefore, $a>0$ (since $a \in \mathbb{N}$). Thus, $a+b>b$, so that $b<a+b=k$.

But our induction hypothesis says that Exercise 4.10.5 is true for $a+b<k$. Hence, we can apply Exercise 4.10.5 to $b-a$ instead of b (since $b-a \in \mathbb{N}$ and $a+(b-a)=b<k$). We thus obtain

$$\gcd(x_a, x_{b-a}) = \left|x_{\gcd(a,b-a)}\right|. \tag{606}$$

But we have $\gcd(a, b-a) = \gcd(a,b)$ (this has already been proved during our proof of Theorem 3.4.5). Furthermore, Lemma A.5.4 yields

$$\begin{aligned} \gcd(x_a, x_b) &= \gcd(x_a, x_{b-a}) = \left|x_{\gcd(a,b-a)}\right| \qquad \text{(by (606))} \\ &= \left|x_{\gcd(a,b)}\right| \qquad (\text{since } \gcd(a, b-a) = \gcd(a,b)). \end{aligned}$$

Now, forget that we fixed a, b. We thus have shown that any $a, b \in \mathbb{N}$ satisfying $a+b=k$ satisfy $\gcd(x_a, x_b) = \left|x_{\gcd(a,b)}\right|$. In other words, Exercise 4.10.5 is true for $a+b=k$. This completes the induction step. Thus, Exercise 4.10.5 is solved. □

□

[364] because Proposition 3.4.4 **(b)** yields $\gcd(a,b) = \gcd(b,a)$ and $\gcd(x_a, x_b) = \gcd(x_b, x_a)$

[365] *Proof.* Assume that $a=0$. Then, $\gcd(a,b)=b$ (this has already been proved during our above proof of Theorem 3.4.5) and thus $b = \gcd(a,b)$. Furthermore, from $a=0$, we obtain $x_a = x_0 = 0$ and therefore

$$\begin{aligned} \gcd(x_a, x_b) &= \gcd(0, x_b) \\ &= \gcd(x_b, 0) \qquad \text{(by Proposition 3.4.4 } \textbf{(b)}) \\ &= |x_b| \qquad \text{(by Proposition 3.4.4 } \textbf{(a)}) \\ &= \left|x_{\gcd(a,b)}\right| \qquad (\text{since } b = \gcd(a,b)), \end{aligned}$$

qed.

A.5.6. Discussion of Exercise 4.10.6

Discussion of Exercise 4.10.6. Exercise 4.10.6 is a classic exercise (see, e.g., [Engel98, Exercise 14.38] or [Strasz65, Problem 31] or [Polya81, Problem 15.53]) that can be used to illustrate various ideas. We shall generalize it by replacing $\sqrt{2}-1$ by $\sqrt{g+1}-\sqrt{g}$, where g is an arbitrary nonnegative integer. That is, we shall prove that there exists some $m \in \mathbb{N}$ such that

$$\left(\sqrt{g+1}-\sqrt{g}\right)^n = \sqrt{m+1}-\sqrt{m}. \tag{607}$$

This generalization does not simplify the problem, but it is so nice I couldn't leave it unexplored.[366]

So let us try to solve the generalized problem. The most straightforward approach is to try to find the m that satisfies (607) explicitly (in terms of n and g). The function that sends each $m \in \mathbb{N}$ to $\sqrt{m+1}-\sqrt{m} \in \mathbb{R}$ is strictly decreasing (check this!) and thus injective; hence, if there is an $m \in \mathbb{N}$ satisfying (607), then this m is unique. Furthermore, it can be computed explicitly by solving the equation (607) for m. More generally, for any fixed nonnegative real $a \leq 1$, we can solve the equation

$$a = \sqrt{x+1}-\sqrt{x} \tag{608}$$

in the unknown $x \geq 0$; the solution is always

$$x = \left(\frac{a-1/a}{2}\right)^2 \tag{609}$$

(check this!). We can apply this to $a = \left(\sqrt{g+1}-\sqrt{g}\right)^n$ (indeed, it is easy to see that the number $\sqrt{g+1}-\sqrt{g}$ is positive and ≤ 1; therefore, the same holds for its n-th power a), and thus we find that the solution m is given by

$$\begin{aligned} m &= \left(\frac{\left(\sqrt{g+1}-\sqrt{g}\right)^n - 1/\left(\sqrt{g+1}-\sqrt{g}\right)^n}{2}\right)^2 \\ &= \left(\frac{\left(\sqrt{g+1}-\sqrt{g}\right)^n - \left(\sqrt{g+1}+\sqrt{g}\right)^n}{2}\right)^2 \end{aligned} \tag{610}$$

(by fairly simple computations[367]). It thus remains to prove that this m is an element of $\mathbb{N}$. Thus, we must prove the following theorem:

[366] See also [GelAnd17, Problem 941] or [Tomesc85, Problem 1.4] for a variant of this generalization in which the minus signs are replaced by plus signs.

[367] Specifically, we have to notice that $1/\left(\sqrt{g+1}-\sqrt{g}\right) = \sqrt{g+1}+\sqrt{g}$ and thus $1/\left(\sqrt{g+1}-\sqrt{g}\right)^n = \left(\sqrt{g+1}+\sqrt{g}\right)^n$.

Theorem A.5.5. Let $g \in \mathbb{N}$ and $n \in \mathbb{N}$. Let $\lambda = \sqrt{g+1} - \sqrt{g}$ and $\mu = \sqrt{g+1} + \sqrt{g}$. Let

$$m = \left(\frac{\mu^n - \lambda^n}{2} \right)^2 .$$

Then, we have $m \in \mathbb{N}$ and $\lambda^n = \sqrt{m+1} - \sqrt{m}$.

We will give two proofs of Theorem A.5.5 and sketch a third one. First, let us prove an auxiliary lemma:

Lemma A.5.6. Let u and v be two numbers. Let $n \in \mathbb{N}$. Then,

$$(u+v)^n + (u-v)^n = 2 \sum_{\substack{k \in \{0,1,\ldots,n\}; \\ k \text{ is even}}} \binom{n}{k} v^k u^{n-k} \tag{611}$$

and

$$(u+v)^n - (u-v)^n = 2 \sum_{\substack{k \in \{0,1,\ldots,n\}; \\ k \text{ is odd}}} \binom{n}{k} v^k u^{n-k}. \tag{612}$$

Proof of Lemma A.5.6. We have $u + v = v + u$ and thus

$$(u+v)^n = (v+u)^n = \sum_{k=0}^{n} \binom{n}{k} v^k u^{n-k} \tag{613}$$

(by Theorem 4.3.16, applied to $x = v$ and $y = u$). Also, $u - v = (-v) + u$ and thus

$$(u-v)^n = ((-v)+u)^n = \sum_{k=0}^{n} \binom{n}{k} (-v)^k u^{n-k} \tag{614}$$

(by Theorem 4.3.16, applied to $x = -v$ and $y = u$). Adding this equality to the

equality (613), we obtain

$$
\begin{aligned}
&(u+v)^n + (u-v)^n \\
&= \sum_{k=0}^{n} \binom{n}{k} v^k u^{n-k} + \sum_{k=0}^{n} \binom{n}{k} \underbrace{(-v)^k}_{=(-1)^k v^k} u^{n-k} \\
&= \sum_{k=0}^{n} \binom{n}{k} v^k u^{n-k} + \sum_{k=0}^{n} \binom{n}{k} (-1)^k v^k u^{n-k} \\
&= \underbrace{\sum_{k=0}^{n}}_{=\sum_{k\in\{0,1,\ldots,n\}}} \underbrace{\left(\binom{n}{k} v^k u^{n-k} + \binom{n}{k} (-1)^k v^k u^{n-k} \right)}_{=\binom{n}{k}\left(1+(-1)^k\right) v^k u^{n-k}} \\
&= \sum_{k\in\{0,1,\ldots,n\}} \binom{n}{k} \left(1+(-1)^k\right) v^k u^{n-k} \\
&= \sum_{\substack{k\in\{0,1,\ldots,n\}; \\ k \text{ is even}}} \binom{n}{k} \left(1 + \underbrace{(-1)^k}_{\substack{=1 \\ (\text{since } k \text{ is even})}} \right) v^k u^{n-k} \\
&\qquad + \sum_{\substack{k\in\{0,1,\ldots,n\}; \\ k \text{ is odd}}} \binom{n}{k} \left(1 + \underbrace{(-1)^k}_{\substack{=-1 \\ (\text{since } k \text{ is odd})}} \right) v^k u^{n-k} \\
&\qquad\qquad (\text{since each } k \in \{0,1,\ldots,n\} \text{ is either even or odd}) \\
&= \sum_{\substack{k\in\{0,1,\ldots,n\}; \\ k \text{ is even}}} \binom{n}{k} \underbrace{(1+1)}_{=2} v^k u^{n-k} + \sum_{\substack{k\in\{0,1,\ldots,n\}; \\ k \text{ is odd}}} \binom{n}{k} \underbrace{(1+(-1))}_{=0} v^k u^{n-k} \\
&= \sum_{\substack{k\in\{0,1,\ldots,n\}; \\ k \text{ is even}}} \binom{n}{k} \cdot 2 v^k u^{n-k} + \underbrace{\sum_{\substack{k\in\{0,1,\ldots,n\}; \\ k \text{ is odd}}} \binom{n}{k} \cdot 0 v^k u^{n-k}}_{=0} \\
&= \sum_{\substack{k\in\{0,1,\ldots,n\}; \\ k \text{ is even}}} \binom{n}{k} \cdot 2 v^k u^{n-k} = 2 \sum_{\substack{k\in\{0,1,\ldots,n\}; \\ k \text{ is even}}} \binom{n}{k} v^k u^{n-k}.
\end{aligned}
$$

This proves (611).

Subtracting the equality (614) from the equality (613), we obtain

$$
\begin{aligned}
&(u+v)^n - (u-v)^n \\
&= \sum_{k=0}^{n} \binom{n}{k} v^k u^{n-k} - \sum_{k=0}^{n} \binom{n}{k} \underbrace{(-v)^k}_{=(-1)^k v^k} u^{n-k} \\
&= \sum_{k=0}^{n} \binom{n}{k} v^k u^{n-k} - \sum_{k=0}^{n} \binom{n}{k} (-1)^k v^k u^{n-k} \\
&= \underbrace{\sum_{k=0}^{n}}_{= \sum_{k\in\{0,1,\ldots,n\}}} \underbrace{\left(\binom{n}{k} v^k u^{n-k} - \binom{n}{k} (-1)^k v^k u^{n-k} \right)}_{= \binom{n}{k} \left(1-(-1)^k\right) v^k u^{n-k}} \\
&= \sum_{k\in\{0,1,\ldots,n\}} \binom{n}{k} \left(1-(-1)^k\right) v^k u^{n-k} \\
&= \sum_{\substack{k\in\{0,1,\ldots,n\}; \\ k \text{ is even}}} \binom{n}{k} \left(1 - \underbrace{(-1)^k}_{\substack{=1 \\ (\text{since } k \text{ is even})}} \right) v^k u^{n-k} \\
&\qquad + \sum_{\substack{k\in\{0,1,\ldots,n\}; \\ k \text{ is odd}}} \binom{n}{k} \left(1 - \underbrace{(-1)^k}_{\substack{=-1 \\ (\text{since } k \text{ is odd})}} \right) v^k u^{n-k} \\
&\qquad\qquad (\text{since each } k \in \{0,1,\ldots,n\} \text{ is either even or odd}) \\
&= \sum_{\substack{k\in\{0,1,\ldots,n\}; \\ k \text{ is even}}} \binom{n}{k} \underbrace{(1-1)}_{=0} v^k u^{n-k} + \sum_{\substack{k\in\{0,1,\ldots,n\}; \\ k \text{ is odd}}} \binom{n}{k} \underbrace{(1-(-1))}_{=2} v^k u^{n-k} \\
&= \underbrace{\sum_{\substack{k\in\{0,1,\ldots,n\}; \\ k \text{ is even}}} \binom{n}{k} \cdot 0 v^k u^{n-k}}_{=0} + \sum_{\substack{k\in\{0,1,\ldots,n\}; \\ k \text{ is odd}}} \binom{n}{k} \cdot 2 v^k u^{n-k} \\
&= \sum_{\substack{k\in\{0,1,\ldots,n\}; \\ k \text{ is odd}}} \binom{n}{k} \cdot 2 v^k u^{n-k} = 2 \sum_{\substack{k\in\{0,1,\ldots,n\}; \\ k \text{ is odd}}} \binom{n}{k} v^k u^{n-k}.
\end{aligned}
$$

This proves (612). Thus, Lemma A.5.6 is proven. □

First proof of Theorem A.5.5. We have $m = \left(\dfrac{\mu^n - \lambda^n}{2} \right)^2 \geq 0$ (since the square of any real is nonnegative) and thus $m+1 \geq 1 \geq 0$. Hence, $\sqrt{m+1}$ and $\sqrt{m}$ are well-defined.

Multiplying the equalities $\lambda = \sqrt{g+1} - \sqrt{g}$ and $\mu = \sqrt{g+1} + \sqrt{g}$, we obtain

$$\begin{aligned}\lambda\mu &= \left(\sqrt{g+1} - \sqrt{g}\right)\left(\sqrt{g+1} + \sqrt{g}\right) = \underbrace{\left(\sqrt{g+1}\right)^2}_{=g+1} - \underbrace{\left(\sqrt{g}\right)^2}_{=g} \\ &= (g+1) - g = 1. \qquad (615)\end{aligned}$$

Hence, the numbers λ and μ are nonzero and are each other's inverses; in particular, $\lambda = 1/\mu$. But $\mu = \sqrt{g+1} + \sqrt{g} \geq \sqrt{g+1} \geq 1$ (since $g+1 \geq 1$); hence, $1/\mu \leq 1/1 = 1$ and thus $\lambda = 1/\mu \leq 1$. Also, $\mu \geq 1 > 0$ and thus $1/\mu > 0$ and therefore $\lambda = 1/\mu > 0$. Now, $\mu \geq 1 \geq \lambda$ (since $\lambda \leq 1$), so that $\mu^n \geq \lambda^n$ (since $n \geq 0$ and since $\mu > 0$ and $\lambda > 0$). Therefore, $\mu^n - \lambda^n \geq 0$, so that $\dfrac{\mu^n - \lambda^n}{2} \geq 0$. Now, from $m = \left(\dfrac{\mu^n - \lambda^n}{2}\right)^2$, we obtain

$$\begin{aligned}\sqrt{m} &= \sqrt{\left(\frac{\mu^n - \lambda^n}{2}\right)^2} = \left|\frac{\mu^n - \lambda^n}{2}\right| \qquad \left(\text{since } \sqrt{x^2} = |x| \text{ for each } x \in \mathbb{R}\right) \\ &= \frac{\mu^n - \lambda^n}{2} \qquad \left(\text{since } \frac{\mu^n - \lambda^n}{2} \geq 0\right). \qquad (616)\end{aligned}$$

Furthermore, from $m = \left(\dfrac{\mu^n - \lambda^n}{2}\right)^2$, we obtain

$$m + 1 = \left(\frac{\mu^n - \lambda^n}{2}\right)^2 + 1 = \left(\frac{\mu^n + \lambda^n}{2}\right)^2$$

(by the equality $\left(\dfrac{a-b}{2}\right)^2 + 1 = \left(\dfrac{a+b}{2}\right)^2$, which holds for any two numbers a and b). Hence,

$$\begin{aligned}\sqrt{m+1} &= \sqrt{\left(\frac{\mu^n + \lambda^n}{2}\right)^2} = \left|\frac{\mu^n + \lambda^n}{2}\right| \qquad \left(\text{since } \sqrt{x^2} = |x| \text{ for each } x \in \mathbb{R}\right) \\ &= \frac{\mu^n + \lambda^n}{2} \qquad \left(\text{since } \frac{\mu^n + \lambda^n}{2} \geq 0 \text{ (because } \mu > 0 \text{ and } \lambda > 0)\right).\end{aligned}$$

Subtracting the equality (616) from this equality, we find

$$\sqrt{m+1} - \sqrt{m} = \frac{\mu^n + \lambda^n}{2} - \frac{\mu^n - \lambda^n}{2} = \lambda^n.$$

Thus, $\lambda^n = \sqrt{m+1} - \sqrt{m}$ is proven. It remains to show that $m \in \mathbb{N}$.

From $\lambda = \sqrt{g+1} - \sqrt{g}$ and $\mu = \sqrt{g+1} + \sqrt{g}$, we obtain

$$\begin{aligned}
\mu^n - \lambda^n &= \left(\sqrt{g+1} + \sqrt{g}\right)^n - \left(\sqrt{g+1} - \sqrt{g}\right)^n \\
&= 2 \sum_{\substack{k \in \{0,1,\ldots,n\}; \\ k \text{ is odd}}} \binom{n}{k} \underbrace{\left(\sqrt{g}\right)^k}_{=g^{k/2}} \underbrace{\left(\sqrt{g+1}\right)^{n-k}}_{=(g+1)^{(n-k)/2}} \\
&\qquad \left(\text{by (612), applied to } u = \sqrt{g+1} \text{ and } v = \sqrt{g}\right) \\
&= 2 \sum_{\substack{k \in \{0,1,\ldots,n\}; \\ k \text{ is odd}}} \binom{n}{k} g^{k/2} (g+1)^{(n-k)/2}.
\end{aligned}$$

Dividing both sides of this equality by 2, we find

$$\frac{\mu^n - \lambda^n}{2} = \sum_{\substack{k \in \{0,1,\ldots,n\}; \\ k \text{ is odd}}} \binom{n}{k} g^{k/2} (g+1)^{(n-k)/2}. \tag{617}$$

Thus, (616) becomes

$$\sqrt{m} = \frac{\mu^n - \lambda^n}{2} = \sum_{\substack{k \in \{0,1,\ldots,n\}; \\ k \text{ is odd}}} \binom{n}{k} g^{k/2} (g+1)^{(n-k)/2}. \tag{618}$$

Now, we make the following three claims:

Claim 1: If n is even, then there is an integer u such that $\sqrt{m} = \sqrt{g} \cdot \sqrt{g+1} \cdot u$.

Claim 2: If n is odd, then there is an integer u such that $\sqrt{m} = \sqrt{g} \cdot u$.

Claim 3: We always have $m \in \mathbb{Z}$.

[*Proof of Claim 1:* Assume that n is even. Thus, $n = 2r$ for some $r \in \mathbb{Z}$. Consider this r. We have $2r = n \geq 0$ and thus $r \geq 0$. The equality (618) rewrites as

$$\sqrt{m} = \sum_{\substack{k \in \{0,1,\ldots,2r\}; \\ k \text{ is odd}}} \binom{2r}{k} g^{k/2} (g+1)^{(2r-k)/2} \tag{619}$$

(since $n = 2r$).

Now, the odd elements of the set $\{0, 1, \ldots, 2r\}$ are $1, 3, 5, \ldots, 2r-1$. In other words, the odd elements of the set $\{0, 1, \ldots, 2r\}$ are the elements $2i - 1$ for $i \in \{1, 2, \ldots, r\}$. Hence, the map

$$\begin{aligned}
\{1, 2, \ldots, r\} &\to \{k \in \{0, 1, \ldots, 2r\} \mid k \text{ is odd}\}, \\
i &\mapsto 2i - 1
\end{aligned}$$

is a bijection. Thus, we can substitute $2i-1$ for k in the sum

$$\sum_{\substack{k\in\{0,1,\ldots,2r\};\\ k \text{ is odd}}} \binom{2r}{k} g^{k/2} (g+1)^{(2r-k)/2}.$$

We thus obtain

$$\begin{aligned}
&\sum_{\substack{k\in\{0,1,\ldots,2r\};\\ k \text{ is odd}}} \binom{2r}{k} g^{k/2} (g+1)^{(2r-k)/2} \\
&= \sum_{i\in\{1,2,\ldots,r\}} \binom{2r}{2i-1} \underbrace{g^{(2i-1)/2}}_{\substack{=g^{(i-1)+1/2} \\ \text{(since } (2i-1)/2=(i-1)+1/2)}} \underbrace{(g+1)^{(2r-(2i-1))/2}}_{\substack{=(g+1)^{(r-i)+1/2} \\ \text{(since } (2r-(2i-1))/2=(r-i)+1/2)}} \\
&= \sum_{i\in\{1,2,\ldots,r\}} \binom{2r}{2i-1} \underbrace{g^{(i-1)+1/2}}_{\substack{=g^{i-1}g^{1/2} \\ \text{(because } i-1 \text{ and } 1/2 \text{ are} \\ \text{nonnegative (since } i\in\{1,2,\ldots,r\} \\ \text{entails } i\geq 1 \text{ and thus } i-1\geq 0))}} \underbrace{(g+1)^{(r-i)+1/2}}_{\substack{=(g+1)^{r-i}(g+1)^{1/2} \\ \text{(because } r-i \text{ and } 1/2 \text{ are} \\ \text{nonnegative (since } i\in\{1,2,\ldots,r\} \\ \text{entails } i\leq r \text{ and thus } r-i\geq 0))}} \\
&= \sum_{i\in\{1,2,\ldots,r\}} \binom{2r}{2i-1} g^{i-1} \underbrace{g^{1/2}}_{=\sqrt{g}} (g+1)^{r-i} \underbrace{(g+1)^{1/2}}_{=\sqrt{g+1}} \\
&= \sum_{i\in\{1,2,\ldots,r\}} \binom{2r}{2i-1} g^{i-1} \sqrt{g} (g+1)^{r-i} \sqrt{g+1} \\
&= \sqrt{g}\cdot\sqrt{g+1}\cdot \sum_{i\in\{1,2,\ldots,r\}} \binom{2r}{2i-1} g^{i-1} (g+1)^{r-i}.
\end{aligned}$$

Hence, (619) becomes

$$\begin{aligned}
\sqrt{m} &= \sum_{\substack{k\in\{0,1,\ldots,2r\};\\ k \text{ is odd}}} \binom{2r}{k} g^{k/2} (g+1)^{(2r-k)/2} \\
&= \sqrt{g}\cdot\sqrt{g+1}\cdot \sum_{i\in\{1,2,\ldots,r\}} \binom{2r}{2i-1} g^{i-1} (g+1)^{r-i}. \qquad (620)
\end{aligned}$$

Note that the sum $\sum_{i\in\{1,2,\ldots,r\}} \binom{2r}{2i-1} g^{i-1} (g+1)^{r-i}$ on the right hand side of this equality is an integer[368]. Thus, the equality (620) shows that there is an integer u such that $\sqrt{m} = \sqrt{g}\cdot\sqrt{g+1}\cdot u$ (namely, $u = \sum_{i\in\{1,2,\ldots,r\}} \binom{2r}{2i-1} g^{i-1} (g+1)^{r-i}$). This proves Claim 1.]

[368]Indeed, for each $i \in \{1,2,\ldots,r\}$, the binomial coefficient $\binom{2r}{2i-1}$ is an integer (since Theo-

[*Proof of Claim 2:* Assume that n is odd. Thus, $n = 2r+1$ for some $r \in \mathbb{Z}$. Consider this r. We have $2r+1 = n \geq 0$ and thus $r \geq -\frac{1}{2}$, so that $r \geq 0$ (since $r \in \mathbb{Z}$). The equality (618) rewrites as

$$\sqrt{m} = \sum_{\substack{k \in \{0,1,\ldots,2r+1\}; \\ k \text{ is odd}}} \binom{2r+1}{k} g^{k/2} (g+1)^{(2r+1-k)/2} \tag{621}$$

(since $n = 2r+1$).

Now, the odd elements of the set $\{0,1,\ldots,2r+1\}$ are $1,3,5,\ldots,2r+1$. In other words, the odd elements of the set $\{0,1,\ldots,2r+1\}$ are the elements $2i+1$ for $i \in \{0,1,\ldots,r\}$. Hence, the map

$$\begin{aligned} \{0,1,\ldots,r\} &\to \{k \in \{0,1,\ldots,2r+1\} \mid k \text{ is odd}\}, \\ i &\mapsto 2i+1 \end{aligned}$$

is a bijection. Thus, we can substitute $2i+1$ for k in the sum

$$\sum_{\substack{k \in \{0,1,\ldots,2r+1\}; \\ k \text{ is odd}}} \binom{2r+1}{k} g^{k/2} (g+1)^{(2r+1-k)/2}.$$

rem 4.3.15 yields that $\binom{2r}{2i-1} \in \mathbb{Z}$), and the powers g^{i-1} and $(g+1)^{r-i}$ are integers as well (since $i \in \{1,2,\ldots,r\}$ entails $i-1 \in \mathbb{N}$ and $r-i \in \mathbb{N}$). Therefore, each addend of the sum $\sum_{i \in \{1,2,\ldots,r\}} \binom{2r}{2i-1} g^{i-1} (g+1)^{r-i}$ is an integer. Hence, the sum itself is an integer as well.

We thus obtain

$$\begin{aligned}
&\sum_{\substack{k\in\{0,1,\ldots,2r+1\};\\ k\text{ is odd}}} \binom{2r+1}{k} g^{k/2}\left(g+1\right)^{(2r+1-k)/2} \\
&= \sum_{i\in\{0,1,\ldots,r\}} \binom{2r+1}{2i+1} \underbrace{g^{(2i+1)/2}}_{\substack{=g^{i+1/2}\\ \text{(since } (2i+1)/2=i+1/2)}} \underbrace{\left(g+1\right)^{(2r+1-(2i+1))/2}}_{\substack{=(g+1)^{r-i}\\ \text{(since } (2r+1-(2i+1))/2=r-i)}} \\
&= \sum_{i\in\{0,1,\ldots,r\}} \binom{2r+1}{2i+1} \underbrace{g^{i+1/2}}_{\substack{=g^i g^{1/2}\\ \text{(because } i \text{ and } 1/2 \text{ are}\\ \text{nonnegative)}}} \left(g+1\right)^{r-i} \\
&= \sum_{i\in\{0,1,\ldots,r\}} \binom{2r+1}{2i+1} g^i \underbrace{g^{1/2}}_{=\sqrt{g}} \left(g+1\right)^{r-i} \\
&= \sum_{i\in\{0,1,\ldots,r\}} \binom{2r+1}{2i+1} g^i \sqrt{g}\left(g+1\right)^{r-i} \\
&= \sqrt{g}\cdot \sum_{i\in\{0,1,\ldots,r\}} \binom{2r+1}{2i+1} g^i \left(g+1\right)^{r-i}.
\end{aligned}$$

Hence, (621) becomes

$$\begin{aligned}
\sqrt{m} &= \sum_{\substack{k\in\{0,1,\ldots,2r+1\};\\ k\text{ is odd}}} \binom{2r+1}{k} g^{k/2}\left(g+1\right)^{(2r+1-k)/2} \\
&= \sqrt{g}\cdot \sum_{i\in\{0,1,\ldots,r\}} \binom{2r+1}{2i+1} g^i \left(g+1\right)^{r-i}. \qquad (622)
\end{aligned}$$

Note that the sum $\sum_{i\in\{0,1,\ldots,r\}} \binom{2r+1}{2i+1} g^i \left(g+1\right)^{r-i}$ on the right hand side of this equality is an integer[369]. Thus, the equality (622) shows that there is an integer u such that $\sqrt{m} = \sqrt{g}\cdot u$ (namely, $u = \sum_{i\in\{0,1,\ldots,r\}} \binom{2r+1}{2i+1} g^i \left(g+1\right)^{r-i}$). This proves Claim 2.]

[369] Indeed, for each $i \in \{0,1,\ldots,r\}$, the binomial coefficient $\binom{2r+1}{2i+1}$ is an integer (since Theorem 4.3.15 yields that $\binom{2r+1}{2i+1} \in \mathbb{Z}$), and the powers g^i and $\left(g+1\right)^{r-i}$ are integers as well (since $i \in \{0,1,\ldots,r\}$ entails $i \in \mathbb{N}$ and $r-i \in \mathbb{N}$). Therefore, each addend of the sum $\sum_{i\in\{0,1,\ldots,r\}} \binom{2r+1}{2i+1} g^i \left(g+1\right)^{r-i}$ is an integer. Hence, the sum itself is an integer as well.

[*Proof of Claim 3:* The integer n is either even or odd. We thus are in one of the following two cases:

Case 1: The integer n is even.

Case 2: The integer n is odd.

Let us first consider Case 1. In this case, the integer n is even. Hence, Claim 1 yields that there is an integer u such that $\sqrt{m} = \sqrt{g} \cdot \sqrt{g+1} \cdot u$. Consider this u. Squaring both sides of the equality $\sqrt{m} = \sqrt{g} \cdot \sqrt{g+1} \cdot u$, we obtain $m = \left(\sqrt{g} \cdot \sqrt{g+1} \cdot u\right)^2 = g(g+1)u^2 \in \mathbb{Z}$ (since g, $g+1$ and u are integers). Thus, Claim 3 is proved in Case 1.

Let us now consider Case 2. In this case, the integer n is odd. Hence, Claim 2 yields that there is an integer u such that $\sqrt{m} = \sqrt{g} \cdot u$. Consider this u. Squaring both sides of the equality $\sqrt{m} = \sqrt{g} \cdot u$, we obtain $m = \left(\sqrt{g} \cdot u\right)^2 = gu^2 \in \mathbb{Z}$ (since g and u are integers). Thus, Claim 3 is proved in Case 2.

We have now proved Claim 3 in both Cases 1 and 2. Hence, Claim 3 always holds.]

Claim 3 yields $m \in \mathbb{Z}$. Hence, $m \in \mathbb{N}$ (since $m \geq 0$). This completes the proof of Theorem A.5.5. $\square$

Second proof of Theorem A.5.5. As in the first proof of Theorem A.5.5 above, we can show that $m \geq 0$ and $\lambda\mu = 1$ and $\lambda^n = \sqrt{m+1} - \sqrt{m}$. It remains to show that $m \in \mathbb{N}$.

Multiplying both sides of the equality $m = \left(\dfrac{\mu^n - \lambda^n}{2}\right)^2$ by 4, we find

$$\begin{aligned} 4m = (\mu^n - \lambda^n)^2 &= \underbrace{(\mu^n)^2}_{=\mu^{2n}} + \underbrace{(\lambda^n)^2}_{=\lambda^{2n}} - 2 \underbrace{\mu^n \lambda^n}_{\substack{=\lambda^n\mu^n=(\lambda\mu)^n=1^n \\ (\text{since } \lambda\mu=1)}} \\ &= \mu^{2n} + \lambda^{2n} - 2 \cdot \underbrace{1^n}_{=1} = \mu^{2n} + \lambda^{2n} - 2. \end{aligned} \tag{623}$$

Now, from $\lambda = \sqrt{g+1} - \sqrt{g}$ and $\mu = \sqrt{g+1} + \sqrt{g}$, we obtain

$$\begin{aligned} \mu^{2n} + \lambda^{2n} &= \left(\sqrt{g+1} + \sqrt{g}\right)^{2n} + \left(\sqrt{g+1} - \sqrt{g}\right)^{2n} \\ &= 2 \sum_{\substack{k \in \{0,1,\ldots,2n\}; \\ k \text{ is even}}} \binom{2n}{k} \underbrace{\left(\sqrt{g}\right)^k}_{=g^{k/2}} \underbrace{\left(\sqrt{g+1}\right)^{2n-k}}_{=(g+1)^{(2n-k)/2}} \\ &\qquad \left(\begin{array}{c} \text{by (611), applied to } \sqrt{g+1}, \sqrt{g} \text{ and } 2n \\ \text{instead of } u, v \text{ and } n \end{array}\right) \\ &= 2 \sum_{\substack{k \in \{0,1,\ldots,2n\}; \\ k \text{ is even}}} \binom{2n}{k} g^{k/2} (g+1)^{(2n-k)/2}. \end{aligned} \tag{624}$$

Now, the even elements of the set $\{0,1,\ldots,2n\}$ are $0,2,4,\ldots,2n$. In other words, the even elements of the set $\{0,1,\ldots,2n\}$ are the elements $2i$ for $i \in \{0,1,\ldots,n\}$. Hence, the map

$$\begin{aligned} \{0,1,\ldots,n\} &\to \{k \in \{0,1,\ldots,2n\} \mid k \text{ is even}\}, \\ i &\mapsto 2i \end{aligned}$$

is a bijection. Thus, we can substitute $2i$ for k in the sum

$$\sum_{\substack{k\in\{0,1,\ldots,2n\}; \\ k \text{ is even}}} \binom{2n}{k} g^{k/2} (g+1)^{(2n-k)/2}.$$

We thus obtain

$$\begin{aligned} &\sum_{\substack{k\in\{0,1,\ldots,2n\}; \\ k \text{ is even}}} \binom{2n}{k} g^{k/2} (g+1)^{(2n-k)/2} \\ &= \underbrace{\sum_{i\in\{0,1,\ldots,n\}}}_{=\sum_{i=0}^{n}} \binom{2n}{2i} \underbrace{g^{2i/2}}_{\substack{=g^i \\ (\text{since } 2i/2=i)}} \underbrace{(g+1)^{(2n-2i)/2}}_{\substack{=(g+1)^{n-i} \\ (\text{since } (2n-2i)/2=n-i)}} \\ &= \sum_{i=0}^{n} \binom{2n}{2i} g^i (g+1)^{n-i}. \end{aligned} \tag{625}$$

The sum on the right hand side of this equality is an integer[370]; but we shall furthermore show that it is an odd integer. In other words, we shall show that

$$\sum_{i=0}^{n} \binom{2n}{2i} g^i (g+1)^{n-i} \text{ is odd.} \tag{626}$$

[*Proof of (626):* We are in one of the following two cases:
Case 1: The integer g is even.
Case 2: The integer g is odd.
Let us first consider Case 1. In this case, the integer g is even. Hence, $2 \mid g$ and

[370]Indeed, for each $i \in \{0,1,\ldots,n\}$, the binomial coefficient $\binom{2n}{2i}$ is an integer (since Theorem 4.3.15 yields that $\binom{2n}{2i} \in \mathbb{Z}$), and the powers g^i and $(g+1)^{n-i}$ are integers as well (since $i \in \{0,1,\ldots,n\}$ entails $i \in \mathbb{N}$ and $n-i \in \mathbb{N}$). Therefore, each addend of the sum $\sum_{i=0}^{n} \binom{2n}{2i} g^i (g+1)^{n-i}$ is an integer. Hence, the sum itself is an integer as well.

$g \equiv 0 \bmod 2$. Now,

$$\begin{aligned}
&\sum_{i=0}^{n} \binom{2n}{2i} g^i (g+1)^{n-i} \\
&= \sum_{i=1}^{n} \binom{2n}{2i} \underbrace{g^i}_{\substack{\equiv 0 \bmod 2 \\ \text{(since } 2 \mid g \mid g^i \text{)} \\ \text{(because } i \geq 1\text{))}}} (g+1)^{n-i} + \underbrace{\binom{2n}{2 \cdot 0}}_{\substack{= \binom{2n}{0} = 1 \\ \text{(by (119))}}} \underbrace{g^0}_{=1} \underbrace{(g+1)^{n-0}}_{=(g+1)^n} \\
&\equiv \underbrace{\sum_{i=1}^{n} \binom{2n}{2i} 0 (g+1)^{n-i}}_{=0} + \left(\underbrace{g}_{\equiv 0 \bmod 2} + 1 \right)^n \equiv (0+1)^n = 1^n = 1 \bmod 2.
\end{aligned}$$

In other words, $\sum_{i=0}^{n} \binom{2n}{2i} g^i (g+1)^{n-i}$ is odd. Thus, (626) is proved in Case 1.

Let us now consider Case 2. In this case, the integer g is odd. Hence, $g \equiv 1 \bmod 2$, so that $\underbrace{g}_{\equiv 1 \bmod 2} + 1 \equiv 1 + 1 = 2 \equiv 0 \bmod 2$. Now,

$$\begin{aligned}
\sum_{i=0}^{n} \binom{2n}{2i} g^i \left(\underbrace{g+1}_{\equiv 0 \bmod 2} \right)^{n-i} &\equiv \sum_{i=0}^{n} \binom{2n}{2i} g^i 0^{n-i} \\
&= \sum_{i=0}^{n-1} \binom{2n}{2i} g^i \underbrace{0^{n-i}}_{\substack{=0 \\ \text{(since } n-i \geq 1 \\ \text{(because } i \leq n-1\text{))}}} + \underbrace{\binom{2n}{2n}}_{\substack{=1 \\ \text{(by (124))}}} g^n \underbrace{0^{n-n}}_{=0^0=1} \\
&= \underbrace{\sum_{i=0}^{n-1} \binom{2n}{2i} g^i 0}_{=0} + g^n = g^n = \left(\underbrace{g}_{\equiv 1 \bmod 2} \right)^n \equiv 1^n = 1 \bmod 2.
\end{aligned}$$

In other words, $\sum_{i=0}^{n} \binom{2n}{2i} g^i (g+1)^{n-i}$ is odd. Thus, (626) is proved in Case 2.

We have now proved (626) in both Cases 1 and 2. Hence, (626) always holds.]

We have thus shown that the integer $\sum_{i=0}^{n} \binom{2n}{2i} g^i (g+1)^{n-i}$ is odd. In other words,

$$\sum_{i=0}^{n} \binom{2n}{2i} g^i (g+1)^{n-i} = 2u + 1 \tag{627}$$

for some integer u. Consider this u. Now, (624) becomes

$$\begin{aligned}\mu^{2n}+\lambda^{2n} &= 2\underbrace{\sum_{\substack{k\in\{0,1,\ldots,2n\};\\ k\text{ is even}}}\binom{2n}{k}g^{k/2}\left(g+1\right)^{(2n-k)/2}}_{\substack{=\sum_{i=0}^{n}\binom{2n}{2i}g^{i}(g+1)^{n-i}\\ \text{(by (625))}}} = 2\underbrace{\sum_{i=0}^{n}\binom{2n}{2i}g^{i}\left(g+1\right)^{n-i}}_{\substack{=2u+1\\ \text{(by (627))}}}\\ &= 2\left(2u+1\right).\end{aligned}$$

Now, (623) becomes

$$4m=\underbrace{\mu^{2n}+\lambda^{2n}}_{=2(2u+1)}-2=2\left(2u+1\right)-2=4u.$$

Cancelling the factor 4 from this equality, we find $m=u$. Hence, m is an integer (since u is an integer). Since $m\geq 0$, we thus obtain $m\in\mathbb{N}$. This again completes the proof of Theorem A.5.5. □

Third proof of Theorem A.5.5 (sketched). As in the first proof of Theorem A.5.5 above, we can show that $m\geq 0$ and $\lambda\mu=1$ and $\lambda^n=\sqrt{m+1}-\sqrt{m}$. It remains to show that $m\in\mathbb{N}$.

Let us rename m as m_n, in order to stress its dependence on n. Thus, we need to show that $m_n\in\mathbb{N}$. Forget that we fixed n.

Now, it is not hard to see that $m_0=0$ and $m_1=g$ and $m_2=4g\left(g+1\right)$. Also, for each $n\in\mathbb{N}$, we have

$$\begin{aligned}m_n &= \left(\frac{\mu^n-\lambda^n}{2}\right)^2=\frac{1}{4}\left(\mu^n-\lambda^n\right)^2=\frac{1}{4}\left(\underbrace{\left(\mu^n\right)^2}_{=\mu^{2n}}+\underbrace{\left(\lambda^n\right)^2}_{=\lambda^{2n}}-2\underbrace{\mu^n\lambda^n}_{=\lambda^n\mu^n=(\lambda\mu)^n}\right)\\ &= \frac{1}{4}\left(\underbrace{\mu^{2n}}_{=(\mu^2)^n}+\underbrace{\lambda^{2n}}_{=(\lambda^2)^n}-2\cdot\left(\underbrace{\lambda\mu}_{=1}\right)^n\right)=\frac{1}{4}\left(\left(\mu^2\right)^n+\left(\lambda^2\right)^n-2\cdot 1^n\right)\\ &= \frac{1}{4}\left(\mu^2\right)^n+\frac{1}{4}\left(\lambda^2\right)^n+\frac{-1}{2}\cdot 1^n. \qquad (628)\end{aligned}$$

This formula looks much like (224): The right hand side is a sum of several n-th powers multiplied by constant coefficients. This strongly suggests that our sequence $(m_0,m_1,m_2,\ldots)$ should be (a_1,a_2,a_3)-recurrent for some three numbers a_1,a_2,a_3 (three because there are three n-th powers on the right hand side of (628)). How do we find these a_1,a_2,a_3 ? If the formula (628) is to be an instance of (224), then the bases of the three powers on the right hand side (628) (that is, the three

numbers μ^2, λ^2 and 1) should be the roots of the polynomial $X^3 - a_1X^2 - a_2X - a_3$. This means that we should have

$$\left(X - \mu^2\right)\left(X - \lambda^2\right)(X - 1) = X^3 - a_1X^2 - a_2X - a_3.$$

But a straightforward computation shows that

$$\left(X - \mu^2\right)\left(X - \lambda^2\right)(X - 1) = X^3 - (4g + 3)X^2 + (4g + 3)X - 1.$$

Comparing these two equations yields $a_1 = 4g + 3$ and $a_2 = -(4g + 3)$ and $a_3 = 1$. Thus, we guess that our sequence $(m_0, m_1, m_2, \ldots)$ should be $(4g + 3, -(4g + 3), 1)$-recurrent. In other words, we guess that

$$m_n = (4g + 3)m_{n-1} - (4g + 3)m_{n-2} + m_{n-3} \tag{629}$$

for each $n \geq 3$.

Proving this guess is easy: Just rewrite the four values $m_n, m_{n-1}, m_{n-2}, m_{n-3}$ on both sides of (629) using (628), and then use the equalities

$$\begin{aligned}
\left(\mu^2\right)^n &= (4g + 3)\left(\mu^2\right)^{n-1} - (4g + 3)\left(\mu^2\right)^{n-2} + \left(\mu^2\right)^{n-3}, \\
\left(\lambda^2\right)^n &= (4g + 3)\left(\lambda^2\right)^{n-1} - (4g + 3)\left(\lambda^2\right)^{n-2} + \left(\lambda^2\right)^{n-3}, \\
1^n &= (4g + 3)1^{n-1} - (4g + 3)1^{n-2} + 1^{n-3}
\end{aligned}$$

(which follow from the fact that μ^2, λ^2 and 1 are roots of the polynomial $X^3 - (4g + 3)X^2 + (4g + 3)X - 1$).

Thus, the sequence $(m_0, m_1, m_2, \ldots)$ is $(4g + 3, -(4g + 3), 1)$-recurrent. Since the first three entries of this sequence are integers (indeed, $m_0 = 0 \in \mathbb{Z}$ and $m_1 = g \in \mathbb{Z}$ and $m_2 = 4g(g + 1) \in \mathbb{Z}$), we thus conclude (by strong induction on n) that all its entries m_n are integers (since $4g + 3$, $-(4g + 3)$ and 1 are integers). In other words, m_n is an integer for all $n \in \mathbb{N}$. Since we also know that $m_n \geq 0$, we thus conclude that $m_n \in \mathbb{N}$ for all $n \in \mathbb{N}$. This proves Theorem A.5.5. □

Now, let us solve Exercise 4.10.6 at last. Indeed, set $g = 1$. Define λ, μ and m as in Theorem A.5.5. Then, Theorem A.5.5 yields that $m \in \mathbb{N}$ and $\lambda^n = \sqrt{m + 1} - \sqrt{m}$. But

$$\begin{aligned}
\lambda &= \sqrt{g + 1} - \sqrt{g} = \underbrace{\sqrt{1 + 1}}_{=\sqrt{2}} - \underbrace{\sqrt{1}}_{=1} \qquad (\text{since } g = 1) \\
&= \sqrt{2} - 1,
\end{aligned}$$

so that $\lambda^n = \left(\sqrt{2} - 1\right)^n$. Hence, $\left(\sqrt{2} - 1\right)^n = \lambda^n = \sqrt{m + 1} - \sqrt{m}$. Thus, we have found an $m \in \mathbb{N}$ such that $\left(\sqrt{2} - 1\right)^n = \sqrt{m + 1} - \sqrt{m}$. This shows that such an m exists. Thus, Exercise 4.10.6 is solved. □

A.5.7. Discussion of Exercise 4.10.7

Discussion of Exercise 4.10.7. Exercise 4.10.7 is Exercise 2 from the 33rd Virginia Tech Regional Mathematics Contest. The trick is to guess the explicit formula for a_n; proving it afterwards is straightforward. The formula is the following: For each $n \in \mathbb{N}$, we have

$$a_n = n^2 - 1. \tag{630}$$

For the sake of completeness, let us give the straightforward proof of (630):

[*Proof of (630):* We shall prove (630) by strong induction on n:

Induction step: Let $m \in \mathbb{N}$. Assume (as the induction hypothesis) that (630) holds for $n < m$. We must prove that (630) holds for $n = m$. In other words, we must prove that $a_m = m^2 - 1$.

This is certainly true if $m = 0$ (since $a_0 = -1 = 0^2 - 1$), and is also true if $m = 1$ (since $a_1 = 0 = 1^2 - 1$). Hence, for the rest of this proof, we WLOG assume that m is neither 0 nor 1. Therefore, $m \geq 2$ (since $m \in \mathbb{N}$). Hence, $m - 1 \in \mathbb{N}$ and $m - 2 \in \mathbb{N}$.

We have assumed that (630) holds for $n < m$. Thus, (630) holds for $n = m - 1$ (since $m - 1 \in \mathbb{N}$ and $m - 1 < m$). In other words, we have $a_{m-1} = (m-1)^2 - 1$. The same argument (applied to $m - 2$ instead of $m - 1$) shows that $a_{m-2} = (m-2)^2 - 1$ (since $m - 2 \in \mathbb{N}$ and $m - 2 < m$).

Now, recall that $a_n = a_{n-1}^2 - n^2 a_{n-2} - 1$ for all $n \geq 2$ (by the definition of our sequence). Applying this to $n = m$, we find

$$\begin{aligned}
a_m &= a_{m-1}^2 - m^2 a_{m-2} - 1 = \left(\underbrace{(m-1)^2 - 1}_{=m(m-2)}\right)^2 - m^2 \left(\underbrace{(m-2)^2 - 1}_{=(m-3)(m-1)}\right) - 1 \\
&\qquad \left(\text{since } a_{m-1} = (m-1)^2 - 1 \text{ and } a_{m-2} = (m-2)^2 - 1\right) \\
&= (m(m-2))^2 - m^2 (m-3)(m-1) - 1 = m^2 \underbrace{\left((m-2)^2 - (m-3)(m-1)\right)}_{=1} - 1 \\
&= m^2 - 1.
\end{aligned}$$

This completes the induction step. Thus, (630) is proved by induction.]

The formula (630) can now be applied to $n = 100$, resulting in $a_{100} = 100^2 - 1 = 9999$.

A better question is: How would one come up with a formula like (630)? One way is to recognize the first few entries $a_0 = -1$, $a_1 = 0$, $a_2 = 3$, $a_3 = 8$ and $a_4 = 15$ as $0^2 - 1$, $1^2 - 1$ and so on. This is arguably easier with the benefit of hindsight. How else could one discover (630)? I think a natural way to do so is to observe that the sequence $(a_0, a_1, a_2, \ldots)$ grows "a lot slower than expected". To wit, the recurrence equation $a_n = a_{n-1}^2 - n^2 a_{n-2} - 1$ for computing an entry a_n involves squaring the preceding entry a_{n-1}; this suggests a doubly exponential growth (which would certainly be the case if the "$-$" signs were replaced by "$+$"

signs). Yet a look at the first 10 or so entries of the sequence suggests a much slower growth – polynomial perhaps. This is not a rigorous argument – the growth behavior of a sequence is not determined by its first 10 entries, or by its first 1000 entries for that matter – but it is a remarkable observation that calls for explanation. Guided by this observation, one might look for a polynomial formula for a_n – that is, a polynomial P such that $a_n = P(n)$ for all $n \in \mathbb{N}$. Making the reasonable assumption that $\deg P$ shouldn't be too large, one can then find P using polynomial interpolation. The result is the formula (630). Of course, this does not substitute for a rigorous proof of (630). □

A.5.8. Discussion of Exercise 4.10.8

Discussion of Exercise 4.10.8. Exercise 4.10.8 is [Galvin20, Chapter 7, Problem 2]. Fairly similar exercises appear all over the literature; see, e.g., [Engel98, Problem 14.30]. The trick is always the same: Whenever you see a quadratic irrationality like $a + b\sqrt{q}$, try to also get its conjugate $a - b\sqrt{q}$ (that is, the other root of the quadratic equation $X^2 = 2aX + \left(b^2q - a^2\right)$, whose first root is $a + b\sqrt{q}$) involved. In the case of this exercise, the conjugate of $1 + \sqrt{2}$ is $1 - \sqrt{2}$, and (unlike $1 + \sqrt{2}$) this is a real number between -1 and 0. Thus, its n-th power

$$\left(1 - \sqrt{2}\right)^n \text{ lies } \begin{cases} \text{between } 0 \text{ and } 1, & \text{if } n \text{ is even;} \\ \text{between } -1 \text{ and } 0, & \text{if } n \text{ is odd.} \end{cases} \tag{631}$$

But the binomial formula can easily be used to show that $\left(1 + \sqrt{2}\right)^n + \left(1 - \sqrt{2}\right)^n$ is an even integer – let's call it $2u$. From (631), we thus conclude that

$$\left(1 + \sqrt{2}\right)^n \text{ lies } \begin{cases} \text{between } 2u - 1 \text{ and } 2u, & \text{if } n \text{ is even;} \\ \text{between } 2u \text{ and } 2u + 1, & \text{if } n \text{ is odd.} \end{cases}$$

Hence,

$$\left\lfloor \left(1 + \sqrt{2}\right)^n \right\rfloor = \begin{cases} 2u - 1, & \text{if } n \text{ is even;} \\ 2u, & \text{if } n \text{ is odd.} \end{cases}$$

Since $2u$ is even and $2u - 1$ is odd, the claim of the exercise thus follows.

The same argument can be used to prove the following more general result:

Theorem A.5.7. Let $m, q \in \mathbb{N}$ satisfy $m^2 < q < (m+1)^2$. Let $n \in \mathbb{N}$. Then, $\left\lfloor \left(m + \sqrt{q}\right)^n \right\rfloor$ is even if and only if n is odd.

For the sake of completeness, here is a proof of this theorem (following the above plan):

Proof of Theorem A.5.7. Let $\lambda = m + \sqrt{q}$ and $\mu = m - \sqrt{q}$.

Taking square roots on all sides of the inequality chain $m^2 < q < (m+1)^2$, we obtain $m < \sqrt{q} < m+1$ (since all of m, q and $m+1$ are nonnegative reals). Hence, $\mu = m - \sqrt{q} < 0$ (since $m < \sqrt{q}$) and $\mu = m - \sqrt{q} > -1$ (since $\sqrt{q} < m+1$). From $\mu < 0$, we see that μ is negative, so that $|\mu| = -\mu$. Hence, $|\mu| = -\mu < 1$ (since $\mu > -1$). Thus, $|\mu|^n \leq 1$ (since $|\mu|$ is nonnegative), so that $|\mu^n| = |\mu|^n \leq 1$.

From $\lambda = m + \sqrt{q}$ and $\mu = m - \sqrt{q}$, we obtain

$$\begin{aligned}\lambda^n + \mu^n &= (m+\sqrt{q})^n + (m-\sqrt{q})^n = 2 \sum_{\substack{k \in \{0,1,\ldots,n\}; \\ k \text{ is even}}} \binom{n}{k} \underbrace{(\sqrt{q})^k}_{=q^{k/2}} m^{n-k} \\ &\qquad \text{(by (611), applied to } u = m \text{ and } v = \sqrt{q}) \\ &= 2 \sum_{\substack{k \in \{0,1,\ldots,n\}; \\ k \text{ is even}}} \binom{n}{k} q^{k/2} m^{n-k}. \qquad (632)\end{aligned}$$

The sum $\sum_{\substack{k \in \{0,1,\ldots,n\}; \\ k \text{ is even}}} \binom{n}{k} q^{k/2} m^{n-k}$ is an integer[371]. Let us denote this integer by u. Thus,

$$u = \sum_{\substack{k \in \{0,1,\ldots,n\}; \\ k \text{ is even}}} \binom{n}{k} q^{k/2} m^{n-k}.$$

Hence, (632) rewrites as $\lambda^n + \mu^n = 2u$. Hence, $2u = \lambda^n + \mu^n$.

Now, we are in one of the following two cases:

Case 1: The integer n is even.

Case 2: The integer n is odd.

Let us first consider Case 1. In this case, the integer n is even. Hence, $\alpha^n \geq 0$ for each positive real α. Applying this to $\alpha = \mu$, we obtain $\mu^n \geq 0$. Thus, $|\mu^n| = \mu^n$, so that $\mu^n = |\mu^n| \leq 1$. Also, $\mu \neq 0$ (since $\mu < 0$), so that $\mu^n \neq 0$ and thus $\mu^n > 0$ (since $\mu^n \geq 0$).

The integer $2u - 1$ satisfies $2u - 1 \leq \lambda^n$ (since $2u = \lambda^n + \underbrace{\mu^n}_{\leq 1} \leq \lambda^n + 1$) and

[371] *Proof.* Let $k \in \{0,1,\ldots,n\}$ be even. Then, $k \in \{0,1,\ldots,n\} \subseteq \mathbb{N}$, so that $k \geq 0$ and thus $k/2 \geq 0$. The binomial coefficient $\binom{n}{k}$ is an integer (since Theorem 4.3.15 yields that $\binom{n}{k} \in \mathbb{Z}$). Moreover, $k/2 \in \mathbb{Z}$ (since k is even), so that $k/2 \in \mathbb{N}$ (since $k/2 \geq 0$). Hence, $q^{k/2}$ is an integer. Also, $k \leq n$ (since $k \in \{0,1,\ldots,n\}$) and thus $n - k \in \mathbb{N}$. Hence, m^{n-k} is an integer. Now, we conclude that $\binom{n}{k} q^{k/2} m^{n-k}$ is an integer (since $\binom{n}{k}$, $q^{k/2}$ and m^{n-k} are integers).

Forget that we fixed k. We thus have proved that $\binom{n}{k} q^{k/2} m^{n-k}$ is an integer for each even $k \in \{0,1,\ldots,n\}$. In other words, each addend of the sum $\sum_{\substack{k \in \{0,1,\ldots,n\}; \\ k \text{ is even}}} \binom{n}{k} q^{k/2} m^{n-k}$ is an integer. Hence, the whole sum is an integer as well.

$\lambda^n < 2u$ (since $2u = \lambda^n + \underbrace{\mu^n}_{>0} > \lambda^n$). Thus, the integer $2u - 1$ is $\leq \lambda^n$ (since $2u - 1 \leq \lambda^n$), while the next integer $2u$ is no longer $\leq \lambda^n$ (since $\lambda^n < 2u$). Therefore, the largest integer that is $\leq \lambda^n$ is $2u - 1$. In other words, $\lfloor \lambda^n \rfloor$ is $2u - 1$ (since $\lfloor \lambda^n \rfloor$ was defined as the largest integer that is $\leq \lambda^n$ (by the definition of a floor)). In other words, $\lfloor \lambda^n \rfloor = 2u - 1$. But $2u - 1$ is odd (since u is an integer). In other words, $\lfloor \lambda^n \rfloor$ is odd (since $\lfloor \lambda^n \rfloor = 2u - 1$). In view of $\lambda = m + \sqrt{q}$, this rewrites as follows: $\left\lfloor \left(m + \sqrt{q}\right)^n \right\rfloor$ is odd. Thus, $\left\lfloor \left(m + \sqrt{q}\right)^n \right\rfloor$ is not even. Also, n is not odd (since n is even). Thus, we conclude that $\left\lfloor \left(m + \sqrt{q}\right)^n \right\rfloor$ is even if and only if n is odd (because we know that $\left\lfloor \left(m + \sqrt{q}\right)^n \right\rfloor$ is not even, and that n is not odd). In other words, Theorem A.5.7 is proved in Case 1.

Let us now consider Case 2. In this case, the integer n is odd. Hence, $\alpha^n < 0$ for each negative real α. Applying this to $\alpha = \mu$, we obtain $\mu^n < 0$ (since μ is negative). Hence, $|\mu^n| = -\mu^n$, so that $-\mu^n = |\mu^n| = |\mu|^n$. But we have $|\mu| < 1$ and therefore $|\mu|^n < 1$ (since $n \geq 1$ (because n is an odd nonnegative integer)). Hence, $-\mu^n = |\mu|^n < 1$. In other words, $\mu^n > -1$.

The integer $2u$ satisfies $2u \leq \lambda^n$ (since $2u = \lambda^n + \underbrace{\mu^n}_{<0} < \lambda^n$) and $\lambda^n < 2u + 1$ (since $2u = \lambda^n + \underbrace{\mu^n}_{>-1} > \lambda^n + (-1) = \lambda^n - 1$). Thus, the integer $2u$ is $\leq \lambda^n$ (since $2u \leq \lambda^n$), while the next integer $2u + 1$ is no longer $\leq \lambda^n$ (since $\lambda^n < 2u + 1$). Therefore, the largest integer that is $\leq \lambda^n$ is $2u$. In other words, $\lfloor \lambda^n \rfloor$ is $2u$ (since $\lfloor \lambda^n \rfloor$ was defined as the largest integer that is $\leq \lambda^n$ (by the definition of a floor)). In other words, $\lfloor \lambda^n \rfloor = 2u$. But $2u$ is even (since u is an integer). In other words, $\lfloor \lambda^n \rfloor$ is even (since $\lfloor \lambda^n \rfloor = 2u$). In view of $\lambda = m + \sqrt{q}$, this rewrites as follows: $\left\lfloor \left(m + \sqrt{q}\right)^n \right\rfloor$ is even. Thus, we conclude that $\left\lfloor \left(m + \sqrt{q}\right)^n \right\rfloor$ is even if and only if n is odd (because we know that $\left\lfloor \left(m + \sqrt{q}\right)^n \right\rfloor$ is even, and that n is odd). In other words, Theorem A.5.7 is proved in Case 2.

We have now proved Theorem A.5.7 in both Cases 1 and 2. Hence, the proof of Theorem A.5.7 is complete. □

Exercise 4.10.8 follows by applying Theorem A.5.7 to $m = 1$ and $q = 2$ (since $1^2 < 2 < (1 + 1)^2$). □

A.5.9. Discussion of Exercise 4.10.9

Discussion of Exercise 4.10.9. Exercise 4.10.9 has appeared in the thread `https://artofproblemsolving.com/community/c6h49408p314696` . The following solution is taken from said thread.

Solution to Exercise 4.10.9. We assumed that $m \mid k + 1$. Thus, there exists an integer q such that $k + 1 = mq$. Consider this q.

For each integer $n \geq 3$, we have $a_n = \dfrac{k + a_{n-1}a_{n-2}}{a_{n-3}}$ (by the definition of the sequence $(a_0, a_1, a_2, \ldots)$) and thus

$$a_n a_{n-3} = k + a_{n-1}a_{n-2}$$

and therefore

$$a_n a_{n-3} - a_{n-1}a_{n-2} = k. \tag{633}$$

Applying this to $n = 3$, we find $a_3 a_{3-3} - a_{3-1}a_{3-2} = k$. In view of $a_{3-3} = a_0 = 1$ and $a_{3-1} = a_2 = m$ and $a_{3-2} = a_1 = 1$, this rewrites as $a_3 \cdot 1 - m \cdot 1 = k$. Hence, $a_3 \cdot 1 = m \cdot 1 + k = m + k$, so that $a_3 = a_3 \cdot 1 = m + k$.

Now, we will show the following:

Observation 1: Let u be an integer such that $u \geq 4$. Then, $(a_u + a_{u-2})\, a_{u-3} = (a_{u-2} + a_{u-4})\, a_{u-1}$.

[*Proof of Observation 1:* We have $u - 1 \geq 3$ (since $u \geq 4$). Hence, (633) (applied to $n = u - 1$) yields

$$a_{u-1}a_{(u-1)-3} - a_{(u-1)-1}a_{(u-1)-2} = k.$$

Hence,

$$k = a_{u-1}\underbrace{a_{(u-1)-3}}_{=a_{u-4}} - \underbrace{a_{(u-1)-1}}_{=a_{u-2}}\underbrace{a_{(u-1)-2}}_{=a_{u-3}} = a_{u-1}a_{u-4} - a_{u-2}a_{u-3}. \tag{634}$$

Also, $u \geq 4 \geq 3$. Hence, (633) (applied to $n = u$) yields

$$a_u a_{u-3} - a_{u-1}a_{u-2} = k = a_{u-1}a_{u-4} - a_{u-2}a_{u-3}$$

(by (634)). In other words,

$$a_u a_{u-3} + a_{u-2}a_{u-3} = a_{u-1}a_{u-4} + a_{u-1}a_{u-2}.$$

In view of

$$a_{u-1}a_{u-4} + a_{u-1}a_{u-2} = a_{u-1}\left(a_{u-4} + a_{u-2}\right) = \left(a_{u-4} + a_{u-2}\right) a_{u-1} = \left(a_{u-2} + a_{u-4}\right) a_{u-1}$$

and

$$a_u a_{u-3} + a_{u-2}a_{u-3} = \left(a_u + a_{u-2}\right) a_{u-3},$$

this rewrites as $(a_u + a_{u-2})\, a_{u-3} = (a_{u-2} + a_{u-4})\, a_{u-1}$. This proves Observation 1.]

The core of our argument will be the following two observations:

Observation 2: We have $a_n = (m+1)\, a_{n-1} - a_{n-2}$ for each even integer $n \geq 2$.

Observation 3: We have $a_n = (q+1)\, a_{n-1} - a_{n-2}$ for each odd integer $n \geq 2$.

[*Proof of Observation 2:* We shall prove Observation 2 by strong induction on n:

Induction step: Let u be an even integer such that $u \geq 2$. Assume (as the induction hypothesis) that Observation 2 holds for $n < u$. We must prove that Observation 2 holds for $n = u$. In other words, we must prove that $a_u = (m+1)\, a_{u-1} - a_{u-2}$.

If $u = 2$, then this is easy to check[372]. Hence, we WLOG assume that $u \neq 2$. Hence, $u > 2$ (since $u \geq 2$) and therefore $u \geq 4$ (since u is an even integer). In other words, $u - 2 \geq 2$. Moreover, $u - 2$ is an even integer (since u is an even integer). But our induction hypothesis says that Observation 2 holds for $n < u$. Hence, we can apply Observation 2 to $n = u - 2$ (since $u - 2$ is an even integer satisfying $u - 2 \geq 2$ and $u - 2 < u$). Thus, we obtain

$$a_{u-2} = (m+1)\, \underbrace{a_{(u-2)-1}}_{=a_{u-3}} - \underbrace{a_{(u-2)-2}}_{=a_{u-4}} = (m+1)\, a_{u-3} - a_{u-4}.$$

In other words, $a_{u-2} + a_{u-4} = (m+1)\, a_{u-3}$. Now, Observation 1 yields

$$(a_u + a_{u-2})\, a_{u-3} = \underbrace{(a_{u-2} + a_{u-4})}_{=(m+1)a_{u-3}}\, a_{u-1} = (m+1)\, a_{u-3} a_{u-1}. \tag{635}$$

But the number a_{u-3} is positive (since $(a_0, a_1, a_2, \ldots)$ is a sequence of positive integers) and thus nonzero; hence, we can cancel a_{u-3} from the equality (635). We thus find $a_u + a_{u-2} = (m+1)\, a_{u-1}$. In other words, $a_u = (m+1)\, a_{u-1} - a_{u-2}$. This completes our induction step. Thus, Observation 2 is proved.]

[*Proof of Observation 3:* We shall prove Observation 3 by strong induction on n:

Induction step: Let u be an odd integer such that $u \geq 2$. Assume (as the induction hypothesis) that Observation 3 holds for $n < u$. We must prove that Observation 3 holds for $n = u$. In other words, we must prove that $a_u = (q+1)\, a_{u-1} - a_{u-2}$.

If $u = 3$, then this is easy to check[373]. Hence, we WLOG assume that $u \neq 3$. Also, $u \neq 2$ (since u is odd), so that $u > 2$ (since $u \geq 2$) and therefore $u \geq 3$ (since u is an integer). Combining this with $u \neq 3$, we obtain $u > 3$, and thus $u \geq 4$ (since u is an integer). In other words, $u - 2 \geq 2$. Moreover, $u - 2$ is an odd integer (since u is an odd integer). But our induction hypothesis says that Observation 3 holds for $n < u$. Hence, we can apply Observation 3 to $n = u - 2$ (since $u - 2$ is an odd integer satisfying $u - 2 \geq 2$ and $u - 2 < u$). Thus, we obtain

$$a_{u-2} = (q+1)\, \underbrace{a_{(u-2)-1}}_{=a_{u-3}} - \underbrace{a_{(u-2)-2}}_{=a_{u-4}} = (q+1)\, a_{u-3} - a_{u-4}.$$

[372] *Proof.* Assume that $u = 2$. Thus, $a_u = a_2 = m$ and $a_{u-1} = a_{2-1} = a_1 = 1$ and $a_{u-2} = a_{2-2} = a_0 = 1$. Hence, $(m+1)\, \underbrace{a_{u-1}}_{=1} - \underbrace{a_{u-2}}_{=1} = (m+1) - 1 = m$. Comparing this with $a_u = m$, we find $a_u = (m+1)\, a_{u-1} - a_{u-2}$. Thus, $a_u = (m+1)\, a_{u-1} - a_{u-2}$ is proved under the assumption that $u = 2$.

[373] *Proof.* Assume that $u = 3$. Thus, $a_u = a_3 = m + k$ and $a_{u-1} = a_{3-1} = a_2 = m$ and $a_{u-2} = a_{3-2} = a_1 = 1$. Hence, $(q+1)\, \underbrace{a_{u-1}}_{=m} - \underbrace{a_{u-2}}_{=1} = (q+1)\, m - 1 = m + \underbrace{mq}_{=k+1} - 1 = m + (k+1) - 1 = m + k$. Comparing this with $a_u = m + k$, we find $a_u = (q+1)\, a_{u-1} - a_{u-2}$. Thus, $a_u = (q+1)\, a_{u-1} - a_{u-2}$ is proved under the assumption that $u = 3$.

In other words, $a_{u-2} + a_{u-4} = (q+1)\, a_{u-3}$. Now, Observation 1 yields

$$(a_u + a_{u-2})\, a_{u-3} = \underbrace{(a_{u-2} + a_{u-4})}_{=(q+1)a_{u-3}}\, a_{u-1} = (q+1)\, a_{u-3} a_{u-1}. \tag{636}$$

But the number a_{u-3} is positive (since $(a_0, a_1, a_2, \ldots)$ is a sequence of positive integers) and thus nonzero; hence, we can cancel a_{u-3} from the equality (636). We thus find $a_u + a_{u-2} = (q+1)\, a_{u-1}$. In other words, $a_u = (q+1)\, a_{u-1} - a_{u-2}$. This completes our induction step. Thus, Observation 3 is proved.]

It is now straightforward to solve Exercise 4.10.9 by induction. Indeed, we must show that

$$a_n \text{ is an integer} \tag{637}$$

for each $n \in \mathbb{N}$. We shall prove (637) by strong induction on n:

Induction step: Let $u \in \mathbb{N}$. Assume (as the induction hypothesis) that (637) holds for $n < u$. We must prove that (637) holds for $n = u$. In other words, we must prove that a_u is an integer. If $u = 0$ or $u = 1$, then this is clear (because $a_0 = 1$ and $a_1 = 1$ are integers). Thus, we WLOG assume that we have neither $u = 0$ nor $u = 1$. Hence, we must have $u \geq 2$ (since $u \in \mathbb{N}$). Thus, $u - 1 \in \mathbb{N}$ and $u - 2 \in \mathbb{N}$. Our induction hypothesis readily yields that a_{u-1} and a_{u-2} are integers[374]. Now, we are in one of the following two cases:

Case 1: The integer u is even.

Case 2: The integer u is odd.

Let us first consider Case 1. In this case, the integer u is even. Hence, Observation 2 (applied to $n = u$) yields $a_u = (m+1)\, a_{u-1} - a_{u-2}$. The right hand side of this equality is an integer (since m, a_{u-1} and a_{u-2} are integers); thus, the left hand side is an integer as well. In other words, a_u is an integer. Thus, we have shown that a_u is an integer in Case 1.

Let us now consider Case 2. In this case, the integer u is odd. Hence, Observation 3 (applied to $n = u$) yields $a_u = (q+1)\, a_{u-1} - a_{u-2}$. The right hand side of this equality is an integer (since q, a_{u-1} and a_{u-2} are integers); thus, the left hand side is an integer as well. In other words, a_u is an integer. Thus, we have shown that a_u is an integer in Case 2.

We have now proven that a_u is an integer in both Cases 1 and 2. Hence, a_u always is an integer. This completes the induction step. Thus, (637) is proved. In other words, Exercise 4.10.9 is solved. □

Note that we have not actually used the Hint in the above solution: We did show that the sequence $(a_0, a_1, a_2, \ldots)$ satisfies an (a, b)-recurrence-like equation in which the coefficients depend on the parity of the position (Observations 2 and 3); but this is not the same as proving that each of the two subsequences $(a_0, a_2, a_4, a_6, \ldots)$ and $(a_1, a_3, a_5, a_7, \ldots)$ is (a, b)-recurrent for some integers a and b (as suggested

[374] *Proof.* Our induction hypothesis says that (637) holds for $n < u$. Hence, (637) holds for $n = u - 2$ (since $u - 2 \in \mathbb{N}$ and $u - 2 < u$). In other words, a_{u-2} is an integer. The same argument (applied to $u - 1$ instead of $u - 2$) shows that a_{u-1} is an integer.

in the Hint). However, the latter statement is true and not hard to prove: Namely, both subsequences $(a_0, a_2, a_4, a_6, \ldots)$ and $(a_1, a_3, a_5, a_7, \ldots)$ are $(mq + m + q - 1, -1)$-recurrent. In other words, we have

$$a_n = (mq + m + q - 1)\, a_{n-2} - a_{n-4}$$

for each integer $n \geq 4$. The proof of this fact (best made using Observations 2 and 3 alone, forgetting the original definition of our sequence) is left to the reader. $\square$

A.5.10. Discussion of Exercise 4.10.10

Discussion of Exercise 4.10.10. Exercise 4.10.10 was proposed by N. B. Vasiljev in the *Kvant* journal (Kvant problem M202, posed in Kvant 5/1973, solved in Kvant 1/1974). The following solution sketch is taken from that journal (see also `https://artofproblemsolving.com/community/c6h16467p114645` for a discussion of the problem):

Solution to Exercise 4.10.10 (sketched). Let $A_{a,d}$ denote the arithmetic progression $(a, a+d, a+2d, a+3d, \ldots)$ (that is, the infinite arithmetic progression with difference d and starting entry a). We thus must show that this sequence $A_{a,d}$ contains an infinite geometric progression as a subsequence if and only if $\frac{a}{d} \in \mathbb{Q}$. In other words, we must prove the following two claims:

> *Claim 1:* If $\frac{a}{d} \in \mathbb{Q}$, then the arithmetic progression $A_{a,d}$ contains an infinite geometric progression as a subsequence.
>
> *Claim 2:* If the arithmetic progression $A_{a,d}$ contains an infinite geometric progression as a subsequence, then $\frac{a}{d} \in \mathbb{Q}$.

[*Proof of Claim 1:* Assume that $\frac{a}{d} \in \mathbb{Q}$. Pick an $m \in \mathbb{N}$ satisfying $m > -\frac{a}{d}$. (Such an m clearly exists.) Set $b = a + md$. Thus, $\frac{b}{d} = \frac{a+md}{d} = \underbrace{\frac{a}{d}}_{\in\mathbb{Q}} + \underbrace{m}_{\in\mathbb{N}\subseteq\mathbb{Q}} \in \mathbb{Q}$. Furthermore, $\frac{b}{d} = \frac{a}{d} + m > 0$ (since $m > -\frac{a}{d}$), so that $b \neq 0$.

We know that $\frac{b}{d}$ is a positive rational number (since $\frac{b}{d} > 0$ and $\frac{b}{d} \in \mathbb{Q}$). Hence, we can write $\frac{b}{d}$ in the form $\frac{b}{d} = \frac{s}{r}$ for two **positive** integers s and r. Consider these s and r. From $\frac{b}{d} = \frac{s}{r}$, we obtain $ds = br$.

Let $q = 1 + r$. Then, $q = 1 + r > 1$ (since r is positive), so that q is a positive integer.

Now, I claim that the geometric progression $\left(bq^0, bq^1, bq^2, bq^3, \ldots\right)$ is a subsequence of $A_{a,d}$. Indeed, in order to show this, I set

$$i_n = m + s\left(q^0 + q^1 + \cdots + q^{n-1}\right) \qquad \text{for each } n \in \mathbb{N}.$$

Then, $i_n \in \mathbb{N}$ for each $n \in \mathbb{N}$ (since $m, s, q \in \mathbb{N}$) and we have $i_0 < i_1 < i_2 < \cdots$ (this follows easily from the fact that s and q are positive). Hence, the sequence

$$\left(a + i_0 d, a + i_1 d, a + i_2 d, a + i_3 d, \ldots\right)$$

is a subsequence of the sequence $A_{a,d}$ (since the sequence $A_{a,d}$ consists of the numbers $a + id$ for all $i \in \mathbb{N}$).

However, for each $n \in \mathbb{N}$, we have

$$\begin{aligned}
a + \underbrace{i_n}_{=m+s\left(q^0+q^1+\cdots+q^{n-1}\right)} d &= a + \left(m + s\left(q^0 + q^1 + \cdots + q^{n-1}\right)\right) d \\
&= \underbrace{a + md}_{=b} + \underbrace{s\left(q^0 + q^1 + \cdots + q^{n-1}\right) d}_{=ds\left(q^0+q^1+\cdots+q^{n-1}\right)} \\
&= b + \underbrace{ds}_{=br}\left(q^0 + q^1 + \cdots + q^{n-1}\right) \\
&= b + b \underbrace{r}_{\substack{=q-1 \\ \text{(since } q=1+r\text{)}}} \left(q^0 + q^1 + \cdots + q^{n-1}\right) \\
&= b + b \underbrace{(q-1)\left(q^0 + q^1 + \cdots + q^{n-1}\right)}_{\substack{=q^n-1 \\ \text{(by Exercise 2.1.1, applied to } q \text{ instead of } b\text{)}}} \\
&= b + b\left(q^n - 1\right) = bq^n
\end{aligned}$$

and thus

$$bq^n = a + i_n d.$$

In other words,

$$\left(bq^0, bq^1, bq^2, bq^3, \ldots\right) = \left(a + i_0 d, a + i_1 d, a + i_2 d, a + i_3 d, \ldots\right). \tag{638}$$

Now recall that $\left(a + i_0 d, a + i_1 d, a + i_2 d, a + i_3 d, \ldots\right)$ is a subsequence of $A_{a,d}$. In view of (638), this rewrites as follows: The geometric progression $\left(bq^0, bq^1, bq^2, bq^3, \ldots\right)$ is a subsequence of $A_{a,d}$. Hence, the arithmetic progression $A_{a,d}$ contains an infinite geometric progression as a subsequence. This proves Claim 1.]

[*Proof of Claim 2:* We shall show a stronger statement: If the arithmetic progression $A_{a,d}$ contains a **three-term** geometric progression $\left(bq^0, bq^1, bq^2\right)$ as a subsequence, then $\dfrac{a}{d} \in \mathbb{Q}$.

Indeed, assume that the arithmetic progression $A_{a,d}$ contains a three-term geometric progression $\left(bq^0, bq^1, bq^2\right)$ as a subsequence. In other words, there exist reals b and q and nonnegative integers i, j, k satisfying $i < j < k$ and

$$bq^0 = a + id, \qquad bq^1 = a + jd, \qquad bq^2 = a + kd.$$

Consider these b, q, i, j and k. Now,

$$\left(bq^0\right)\left(bq^2\right) - \left(bq^1\right)^2 = b^2q^2 - b^2q^2 = 0,$$

so that

$$\begin{aligned} 0 &= \underbrace{\left(bq^0\right)}_{=a+id}\underbrace{\left(bq^2\right)}_{=a+kd} - \left(\underbrace{bq^1}_{=a+jd}\right)^2 = (a+id)(a+kd) - (a+jd)^2 \\ &= (i-2j+k)\,ad + \left(ik - j^2\right)d^2 = d \cdot \left((i-2j+k)\,a + \left(ik-j^2\right)d\right). \end{aligned}$$

We can cancel d from this equality (since $d \neq 0$), and thus obtain

$$0 = (i-2j+k)\,a + \left(ik - j^2\right)d.$$

In other words,

$$(i-2j+k)\,a = -\left(ik - j^2\right)d. \tag{639}$$

If we knew that $i - 2j + k \neq 0$, then we could divide the equality (639) by $d\,(i-2j+k)$ (since $d \neq 0$), and thus obtain

$$\frac{a}{d} = \frac{-\left(ik-j^2\right)}{i-2j+k} \in \mathbb{Q} \qquad (\text{since } i, j, k \text{ are integers}),$$

which would complete the proof of Claim 2. Thus, it remains to show that $i - 2j + k \neq 0$.

Indeed, assume the contrary. Thus, $i - 2j + k = 0$, so that $2j = i + k$ and therefore $j = \dfrac{i+k}{2}$. Hence,

$$ik - j^2 = ik - \left(\frac{i+k}{2}\right)^2 = -\left(\frac{k-i}{2}\right)^2 > 0$$

(since $k - i > 0$ (because $i < k$)). Hence, $ik - j^2 \neq 0$. Thus, both factors $ik - j^2$ and d on the right hand side of (639) are nonzero (since $d \neq 0$). Hence, the entire right hand side of (639) is nonzero. Thus, the left hand side is nonzero as well. That is, $(i-2j+k)\,a \neq 0$. But this contradicts $\underbrace{(i-2j+k)}_{=0}a = 0$. This contradiction shows that our assumption was false. Hence, $i - 2j + k \neq 0$ is proved, and our proof of Claim 2 is complete.] $\square$

A few words are in order about the somewhat baroque construction of m and b in the proof of Claim 1. When the numbers a and d have the same sign and a is nonzero, we don't need to go through this ordeal; we can just set $m = 0$ and $b = a$ and start our geometric subsequence directly at the first entry of $A_{a,d}$ (that is, start it with a). However, if a and d have different signs, then an infinite geometric subsequence of $A_{a,d}$ cannot start with a, since only finitely many entries of $A_{a,d}$ have the same sign as a (but a geometric subsequence starting with a would have infinitely many such entries). Thus, we need to throw away the first few entries of $A_{a,d}$ until we are only left with entries that are nonzero and have the same sign as d. This is what m and b are for. Namely, m is the number of entries we need to throw away, and b is the new first entry after throwing away. □

A.5.11. Discussion of Exercise 4.11.6

Discussion of Exercise 4.11.6. As usual, part **(b)** of the exercise is hard on its own, but becomes easy once part **(a)** has been solved. Part **(a)**, meanwhile, is fairly easy to prove by induction. Here is a detailed solution:

Solution to Exercise 4.11.6. We begin by computing the first 9 entries of our sequence.

From (246), we obtain $x_0 = 1$ and $x_1 = 1$ and $x_2 = 1$ and $x_3 = 1$ and $x_4 = 1$ and $x_5 = 1$. Applying (247) to $n = 6$, we find

$$\begin{aligned} x_6 &= \frac{x_{6-3}\left(x_{6-1} + x_{6-5}\right)}{x_{6-6}} = \frac{x_3\left(x_5 + x_1\right)}{x_0} \\ &= \frac{1\left(1+1\right)}{1} \qquad \left(\text{since } x_3 = 1 \text{ and } x_5 = 1 \text{ and } x_1 = 1 \text{ and } x_0 = 1\right) \\ &= 2. \end{aligned}$$

Applying (247) to $n = 7$, we find

$$\begin{aligned} x_7 &= \frac{x_{7-3}\left(x_{7-1} + x_{7-5}\right)}{x_{7-6}} = \frac{x_4\left(x_6 + x_2\right)}{x_1} \\ &= \frac{1\left(2+1\right)}{1} \qquad \left(\text{since } x_4 = 1 \text{ and } x_6 = 2 \text{ and } x_2 = 1 \text{ and } x_1 = 1\right) \\ &= 3. \end{aligned}$$

Applying (247) to $n = 8$, we find

$$\begin{aligned} x_8 &= \frac{x_{8-3}\left(x_{8-1} + x_{8-5}\right)}{x_{8-6}} = \frac{x_5\left(x_7 + x_3\right)}{x_2} \\ &= \frac{1\left(3+1\right)}{1} \qquad \left(\text{since } x_5 = 1 \text{ and } x_7 = 3 \text{ and } x_3 = 1 \text{ and } x_2 = 1\right) \\ &= 4. \end{aligned}$$

We have thus computed the values of $x_0, x_1, \ldots, x_8$. From these values, we see immediately that $x_0, x_1, \ldots, x_8$ are integers.

We now proceed to solving both parts of the exercise.

(a) We proceed by induction on n:

Induction base: A straightforward computation shows that Exercise 4.11.6 **(a)** holds for $n = 8$ [375].

Induction step: Let $m \geq 9$ be an integer. Assume (as the induction hypothesis) that Exercise 4.11.6 **(a)** holds for $n = m - 1$. We must prove that Exercise 4.11.6 **(a)** holds for $n = m$.

Our induction hypothesis says that Exercise 4.11.6 **(a)** holds for $n = m - 1$. In other words, we have

$$x_{m-1} + x_{(m-1)-4} + x_{(m-1)-8} = 6x_{(m-1)-3}x_{(m-1)-4}x_{(m-1)-5}.$$

We can rewrite this as

$$x_{m-1} + x_{m-5} + x_{m-9} = 6x_{m-4}x_{m-5}x_{m-6} \tag{640}$$

(since $(m-1)-4 = m-5$ and $(m-1)-8 = m-9$ and $(m-1)-3 = m-4$ and $(m-1)-5 = m-6$).

We have $m \geq 9 \geq 6$. Hence, from (247) (applied to $n = m$), we obtain $x_m = \dfrac{x_{m-3}\left(x_{m-1} + x_{m-5}\right)}{x_{m-6}}$, so that

$$x_{m-6}x_m = x_{m-3}\left(x_{m-5} + x_{m-1}\right). \tag{641}$$

We have $m - 3 \geq 6$ (since $m \geq 9 = 6 + 3$). Thus, from (247) (applied to $n = m - 3$), we obtain

$$x_{m-3} = \frac{x_{(m-3)-3}\left(x_{(m-3)-1} + x_{(m-3)-5}\right)}{x_{(m-3)-6}} = \frac{x_{m-6}\left(x_{m-4} + x_{m-8}\right)}{x_{m-9}}$$

(since $(m-3)-3 = m-6$ and $(m-3)-1 = m-4$ and $(m-3)-5 = m-8$ and $(m-3)-6 = m-9$). Thus,

$$x_{m-3}x_{m-9} = x_{m-6}\left(x_{m-4} + x_{m-8}\right). \tag{642}$$

[375] Here are the details of this computation: Comparing

$$\underbrace{x_8}_{=4} + \underbrace{x_{8-4}}_{=x_4=1} + \underbrace{x_{8-8}}_{=x_0=1} = 4 + 1 + 1 = 6$$

with

$$6\,\underbrace{x_{8-3}}_{=x_5=1}\,\underbrace{x_{8-4}}_{=x_4=1}\,\underbrace{x_{8-5}}_{=x_3=1} = 6,$$

we obtain $x_8 + x_{8-4} + x_{8-8} = 6x_{8-3}x_{8-4}x_{8-5}$. In other words, Exercise 4.11.6 **(a)** holds for $n = 8$.

Now,

$$\begin{aligned}
& x_{m-6}\left(x_m + x_{m-4} + x_{m-8}\right) \\
&= \underbrace{x_{m-6}x_m}_{\substack{=x_{m-3}(x_{m-5}+x_{m-1}) \\ \text{(by (641))}}} + \underbrace{x_{m-6}\left(x_{m-4} + x_{m-8}\right)}_{\substack{=x_{m-3}x_{m-9} \\ \text{(by (642))}}} \\
&= x_{m-3}\left(x_{m-5} + x_{m-1}\right) + x_{m-3}x_{m-9} = x_{m-3}\underbrace{\left(x_{m-1} + x_{m-5} + x_{m-9}\right)}_{\substack{=6x_{m-4}x_{m-5}x_{m-6} \\ \text{(by (640))}}} \\
&= x_{m-3} \cdot 6x_{m-4}x_{m-5}x_{m-6} = 6x_{m-3}x_{m-4}x_{m-5}x_{m-6}. \qquad (643)
\end{aligned}$$

However, x_{m-6} is a positive rational number (since $(x_0, x_1, x_2, \ldots)$ is a sequence of positive rational numbers), and thus is nonzero. Hence, we can cancel x_{m-6} from the equality (643). As a result, we obtain $x_m + x_{m-4} + x_{m-8} = 6x_{m-3}x_{m-4}x_{m-5}$. In other words, Exercise 4.11.6 **(a)** holds for $n = m$. This completes the induction step; thus, Exercise 4.11.6 **(a)** is proved.

(b) This is a straightforward strong induction on n. [376] □

□

A.5.12. Discussion of Exercise 4.11.7

Discussion of Exercise 4.11.7. On its own, part **(a)** of the exercise would be pretty hard (indeed, it stayed unsolved on math.stackexchange for a few years). But part **(b)** provides the scaffolding that makes it easy. Thus, we will start by solving part **(b)** this time.

[376] Here is the proof in detail:

We proceed by strong induction on n.

Induction step: Fix $k \in \mathbb{N}$. Assume (as the induction hypothesis) that Exercise 4.11.6 **(b)** holds for each $n \in \mathbb{N}$ satisfying $n < k$. We must prove that Exercise 4.11.6 **(b)** holds for $n = k$. In other words, we must prove that x_k is an integer.

If $k \leq 8$, then this is clearly true (since we have shown above that $x_0, x_1, \ldots, x_8$ are integers). Thus, for the rest of this induction step, we WLOG assume that $k > 8$. Hence, $k \geq 8$. Thus, Exercise 4.11.6 **(a)** (applied to $n = k$) yields

$$x_k + x_{k-4} + x_{k-8} = 6x_{k-3}x_{k-4}x_{k-5}. \qquad (644)$$

From $k \geq 8$, we obtain $k - 8 \in \mathbb{N}$ and $k - 4 \in \mathbb{N}$ and $k - 3 \in \mathbb{N}$ and $k - 5 \in \mathbb{N}$. Now, recall that Exercise 4.11.6 **(b)** holds for each $n \in \mathbb{N}$ satisfying $n < k$. Hence, in particular, Exercise 4.11.6 **(b)** holds for $n = k - 8$ (since $k - 8 \in \mathbb{N}$ and $k - 8 < k$). In other words, x_{k-8} is an integer. Similarly, we can see that x_{k-4} is an integer (since $k - 4 \in \mathbb{N}$ and $k - 4 < k$). Similarly, we can see that x_{k-3} is an integer (since $k - 3 \in \mathbb{N}$ and $k - 3 < k$). Similarly, we can see that x_{k-5} is an integer (since $k - 5 \in \mathbb{N}$ and $k - 5 < k$). Now, solving the equality (644) for x_k, we obtain

$$x_k = 6x_{k-3}x_{k-4}x_{k-5} - x_{k-4} - x_{k-8}.$$

This shows that x_k is an integer (since all the numbers 6, x_{k-3}, x_{k-4}, x_{k-5} and x_{k-8} on the right hand side are integers). In other words, Exercise 4.11.6 **(b)** holds for $n = k$. This completes the induction step. Thus, Exercise 4.11.6 **(b)** is solved.

Solution to Exercise 4.11.7. From (248), we know that $a_0 = 2$ and $a_1 = 1$ and $a_2 = 1$ and $a_3 = 1$. Thus, a_0 is an integer (since $a_0 = 2 \in \mathbb{Z}$). Applying (249) to $n = 4$, we find

$$\begin{aligned} a_4 &= \frac{(a_{4-1} + a_{4-2})(a_{4-2} + a_{4-3})}{a_{4-4}} = \frac{(a_3 + a_2)(a_2 + a_1)}{a_0} \\ &= \frac{(1+1)(1+1)}{2} \qquad (\text{since } a_3 = 1 \text{ and } a_2 = 1 \text{ and } a_1 = 1 \text{ and } a_0 = 2) \\ &= 2. \end{aligned}$$

Now, let $(x_0, x_1, x_2, \ldots)$ be the sequence defined in Exercise 4.11.6. We shall now solve part **(b)** of Exercise 4.11.7 (and then derive part **(a)** from it).

(b) We must prove that

$$a_n = x_{n+2} x_{n+1} x_n x_{n-1} \tag{645}$$

for each $n \geq 1$. We shall prove this by strong induction on n:

Induction step: Let $m \geq 1$. Assume (as the induction hypothesis) that (645) holds for each $n \geq 1$ satisfying $n < m$. We must prove that (645) holds for $n = m$. In other words, we must prove that $a_m = x_{m+2} x_{m+1} x_m x_{m-1}$.

If $m < 5$, then this can easily be checked by hand[377]. Thus, for the rest of this induction step, we WLOG assume that $m \geq 5$. Hence, $m - 4 \geq 5 - 4 = 1$ and $m - 2 \geq m - 4 \geq 1$ and $m - 3 \geq m - 4 \geq 1$ and $m - 1 \geq m - 4 \geq 1$.

Our induction hypothesis says that (645) holds for each $n \geq 1$ satisfying $n < m$. Thus, in particular, (645) holds for $n = m - 4$ (since $m - 4 \geq 1$ and $m - 4 < m$). In other words, we have $a_{m-4} = x_{(m-4)+2} x_{(m-4)+1} x_{m-4} x_{(m-4)-1}$. In other words,

$$a_{m-4} = x_{m-2} x_{m-3} x_{m-4} x_{m-5} \tag{646}$$

[377] Here are the details of this verification:

- We know (from the solution to Exercise 4.11.6) that $x_0 = 1$ and $x_1 = 1$ and $x_2 = 1$ and $x_3 = 1$ and $x_4 = 1$ and $x_5 = 1$ and $x_6 = 2$.
- Comparing $\underbrace{x_{1+2}}_{=x_3=1} \underbrace{x_{1+1}}_{=x_2=1} \underbrace{x_1}_{=1} \underbrace{x_{1-1}}_{=x_0=1} = 1$ with $a_1 = 1$, we obtain $a_1 = x_{1+2} x_{1+1} x_1 x_{1-1}$. In other words, $a_m = x_{m+2} x_{m+1} x_m x_{m-1}$ holds for $m = 1$.
- Comparing $\underbrace{x_{2+2}}_{=x_4=1} \underbrace{x_{2+1}}_{=x_3=1} \underbrace{x_2}_{=1} \underbrace{x_{2-1}}_{=x_1=1} = 1$ with $a_2 = 1$, we obtain $a_2 = x_{2+2} x_{2+1} x_2 x_{2-1}$. In other words, $a_m = x_{m+2} x_{m+1} x_m x_{m-1}$ holds for $m = 2$.
- Comparing $\underbrace{x_{3+2}}_{=x_5=1} \underbrace{x_{3+1}}_{=x_4=1} \underbrace{x_3}_{=1} \underbrace{x_{3-1}}_{=x_2=1} = 1$ with $a_3 = 1$, we obtain $a_3 = x_{3+2} x_{3+1} x_3 x_{3-1}$. In other words, $a_m = x_{m+2} x_{m+1} x_m x_{m-1}$ holds for $m = 3$.
- Comparing $\underbrace{x_{4+2}}_{=x_6=2} \underbrace{x_{4+1}}_{=x_5=1} \underbrace{x_4}_{=1} \underbrace{x_{4-1}}_{=x_3=1} = 1$ with $a_4 = 2$, we obtain $a_4 = x_{4+2} x_{4+1} x_4 x_{4-1}$. In other words, $a_m = x_{m+2} x_{m+1} x_m x_{m-1}$ holds for $m = 4$.

Thus, we have showed that $a_m = x_{m+2} x_{m+1} x_m x_{m-1}$ holds for $m = 1$, for $m = 2$, for $m = 3$ and for $m = 4$. In other words, $a_m = x_{m+2} x_{m+1} x_m x_{m-1}$ holds if $m < 5$ (because the only integers $m \geq 1$ satisfying $m < 5$ are 1, 2, 3 and 4).

(since $(m-4)+2=m-2$ and $(m-4)+1=m-3$ and $(m-4)-1=m-5$).

Our induction hypothesis says that (645) holds for each $n \geq 1$ satisfying $n < m$. Thus, in particular, (645) holds for $n = m-3$ (since $m-3 \geq 1$ and $m-3 < m$). In other words, we have $a_{m-3} = x_{(m-3)+2} x_{(m-3)+1} x_{m-3} x_{(m-3)-1}$. In other words,

$$a_{m-3} = x_{m-1} x_{m-2} x_{m-3} x_{m-4} \tag{647}$$

(since $(m-3)+2=m-1$ and $(m-3)+1=m-2$ and $(m-3)-1=m-4$).

Our induction hypothesis says that (645) holds for each $n \geq 1$ satisfying $n < m$. Thus, in particular, (645) holds for $n = m-2$ (since $m-2 \geq 1$ and $m-2 < m$). In other words, we have $a_{m-2} = x_{(m-2)+2} x_{(m-2)+1} x_{m-2} x_{(m-2)-1}$. In other words,

$$a_{m-2} = x_m x_{m-1} x_{m-2} x_{m-3} \tag{648}$$

(since $(m-2)+2=m$ and $(m-2)+1=m-1$ and $(m-2)-1=m-3$).

Our induction hypothesis says that (645) holds for each $n \geq 1$ satisfying $n < m$. Thus, in particular, (645) holds for $n = m-1$ (since $m-1 \geq 1$ and $m-1 < m$). In other words, we have $a_{m-1} = x_{(m-1)+2} x_{(m-1)+1} x_{m-1} x_{(m-1)-1}$. In other words,

$$a_{m-1} = x_{m+1} x_m x_{m-1} x_{m-2} \tag{649}$$

(since $(m-1)+2=m+1$ and $(m-1)+1=m$ and $(m-1)-1=m-2$).

However, $m \geq 5$ entails $m+2 \geq 5+2 = 7 \geq 6$. Thus, (247) (applied to $n = m+2$) yields

$$x_{m+2} = \frac{x_{(m+2)-3}\left(x_{(m+2)-1} + x_{(m+2)-5}\right)}{x_{(m+2)-6}} = \frac{x_{m-1}\left(x_{m+1} + x_{m-3}\right)}{x_{m-4}}$$

(since $(m+2)-3=m-1$ and $(m+2)-1=m+1$ and $(m+2)-5=m-3$ and $(m+2)-6=m-4$). Therefore,

$$x_{m+2} x_{m-4} = x_{m-1}\left(x_{m+1} + x_{m-3}\right). \tag{650}$$

Furthermore, $m \geq 5$ entails $m+1 \geq 5+1 = 6$. Thus, (247) (applied to $n = m+1$) yields

$$x_{m+1} = \frac{x_{(m+1)-3}\left(x_{(m+1)-1} + x_{(m+1)-5}\right)}{x_{(m+1)-6}} = \frac{x_{m-2}\left(x_m + x_{m-4}\right)}{x_{m-5}}$$

(since $(m+1)-3 = m-2$ and $(m+1)-1 = m$ and $(m+1)-5 = m-4$ and $(m+1)-6 = m-5$). Therefore,

$$x_{m+1} x_{m-5} = x_{m-2}\left(x_m + x_{m-4}\right). \tag{651}$$

Now,

$$\begin{aligned}
\underbrace{a_{m-1}}_{\substack{=x_{m+1}x_m x_{m-1}x_{m-2} \\ \text{(by (649))}}} + \underbrace{a_{m-2}}_{\substack{=x_m x_{m-1}x_{m-2}x_{m-3} \\ \text{(by (648))}}} &= x_{m+1}x_m x_{m-1}x_{m-2} + x_m x_{m-1}x_{m-2}x_{m-3} \\
&= x_m x_{m-2} \underbrace{x_{m-1}\left(x_{m+1} + x_{m-3}\right)}_{\substack{=x_{m+2}x_{m-4} \\ \text{(by (650))}}} \\
&= x_m x_{m-2} x_{m+2} x_{m-4}
\end{aligned}$$

and

$$\begin{aligned}
\underbrace{a_{m-2}}_{\substack{=x_m x_{m-1}x_{m-2}x_{m-3} \\ \text{(by (648))}}} + \underbrace{a_{m-3}}_{\substack{=x_{m-1}x_{m-2}x_{m-3}x_{m-4} \\ \text{(by (647))}}} &= x_m x_{m-1}x_{m-2}x_{m-3} + x_{m-1}x_{m-2}x_{m-3}x_{m-4} \\
&= x_{m-1}x_{m-3} \underbrace{x_{m-2}\left(x_m + x_{m-4}\right)}_{\substack{=x_{m+1}x_{m-5} \\ \text{(by (651))}}} \\
&= x_{m-1}x_{m-3}x_{m+1}x_{m-5}.
\end{aligned}$$

Now, $m \geq 5 \geq 4$. Hence, (249) (applied to $n = m$) yields

$$\begin{aligned}
a_m &= \frac{\left(a_{m-1} + a_{m-2}\right)\left(a_{m-2} + a_{m-3}\right)}{a_{m-4}} = \frac{x_m x_{m-2}x_{m+2}x_{m-4} \cdot x_{m-1}x_{m-3}x_{m+1}x_{m-5}}{x_{m-2}x_{m-3}x_{m-4}x_{m-5}} \\
&\qquad \left(\begin{array}{c} \text{since } a_{m-1} + a_{m-2} = x_m x_{m-2}x_{m+2}x_{m-4} \\ \text{and } a_{m-2} + a_{m-3} = x_{m-1}x_{m-3}x_{m+1}x_{m-5} \\ \text{and } a_{m-4} = x_{m-2}x_{m-3}x_{m-4}x_{m-5} \end{array}\right) \\
&= x_{m+2}x_{m+1}x_m x_{m-1}.
\end{aligned}$$

In other words, (645) holds for $n = m$. This completes the induction step. Thus, (645) is proved. In other words, Exercise 4.11.7 **(b)** is solved.

(a) Let $n \in \mathbb{N}$. We must prove that a_n is an integer. If $n = 0$, then this is obvious (since we know that a_0 is an integer). Thus, for the rest of this proof, we WLOG assume that $n \neq 0$. Hence, $n \geq 1$ (since $n \in \mathbb{N}$).

Hence, Exercise 4.11.7 **(b)** yields $a_n = x_{n+2}x_{n+1}x_n x_{n-1}$. However, for each $m \in \mathbb{N}$, the number x_m is an integer (by Exercise 4.11.6 **(b)**, applied to m instead of n). In other words, all entries of the sequence $(x_0, x_1, x_2, x_3, \ldots)$ are integers. Hence, in particular, the four numbers $x_{n+2}, x_{n+1}, x_n, x_{n-1}$ are integers (since these four numbers are entries of the sequence $(x_0, x_1, x_2, x_3, \ldots)$). Thus, their product $x_{n+2}x_{n+1}x_n x_{n-1}$ is an integer as well. In other words, a_n is an integer (since $a_n = x_{n+2}x_{n+1}x_n x_{n-1}$). This solves Exercise 4.11.7 **(a)**. $\square$

$\square$

A.5.13. Discussion of Exercise 4.11.8

Discussion of Exercise 4.11.8. Once again, the "active ingredient" in the exercise is part **(b)**; the other two parts will follow easily once part **(b)** is solved.

Solution to Exercise 4.11.8. Let $(x_0, x_1, x_2, \ldots)$ be the sequence defined in Exercise 4.11.6.

(b) We must prove that

$$b_n = x_{n+2} x_n \tag{652}$$

for each $n \geq 0$. We shall prove this by strong induction on n:

Induction step: Let $m \geq 0$. Assume (as the induction hypothesis) that (652) holds for each $n \geq 0$ satisfying $n < m$. We must prove that (652) holds for $n = m$. In other words, we must prove that $b_m = x_{m+2} x_m$.

If $m < 4$, then this is easy to verify directly[378]. Thus, for the rest of this induction step, we WLOG assume that $m \geq 4$. Hence, $m - 4 \geq 0$ and $m - 3 \geq m - 4 \geq 0$ and $m - 2 \geq m - 4 \geq 0$ and $m - 1 \geq m - 4 \geq 0$.

Our induction hypothesis says that (652) holds for each $n \geq 0$ satisfying $n < m$. Thus, in particular, (652) holds for $n = m - 4$ (since $m - 4 \geq 0$ and $m - 4 < m$). In other words, we have $b_{m-4} = x_{(m-4)+2} x_{m-4}$. In other words,

$$b_{m-4} = x_{m-2} x_{m-4}$$

(since $(m - 4) + 2 = m - 2$).

Our induction hypothesis says that (652) holds for each $n \geq 0$ satisfying $n < m$. Thus, in particular, (652) holds for $n = m - 3$ (since $m - 3 \geq 0$ and $m - 3 < m$). In other words, we have $b_{m-3} = x_{(m-3)+2} x_{m-3}$. In other words,

$$b_{m-3} = x_{m-1} x_{m-3}$$

(since $(m - 3) + 2 = m - 1$).

Our induction hypothesis says that (652) holds for each $n \geq 0$ satisfying $n < m$. Thus, in particular, (652) holds for $n = m - 2$ (since $m - 2 \geq 0$ and $m - 2 < m$). In other words, we have $b_{m-2} = x_{(m-2)+2} x_{m-2}$. In other words,

$$b_{m-2} = x_m x_{m-2}$$

(since $(m - 2) + 2 = m$).

[378] *Proof.* Assume that $m < 4$. We must verify that $b_m = x_{m+2} x_m$.

We have $m \in \{0, 1, 2, 3\}$ (since m is an integer satisfying $m \geq 0$ and $m < 4$).

The equalities (250) say that each of the four numbers b_0, b_1, b_2, b_3 equals 1. In other words, $b_k = 1$ for each $k \in \{0, 1, 2, 3\}$. Applying this to $k = m$, we find $b_m = 1$ (since $m \in \{0, 1, 2, 3\}$). On the other hand, $m < 4 < 6$ and therefore $x_m = 1$ (by (246), applied to $n = m$). Furthermore, $m + 2 < 6$ (since $m < 4 = 6 - 2$) and therefore $x_{m+2} = 1$ (by (246), applied to $n = m + 2$). Now, comparing $b_m = 1$ with $\underbrace{x_m}_{=1} \underbrace{x_{m+2}}_{=1} = 1$, we obtain $b_m = x_{m+2} x_m$. Qed.

Our induction hypothesis says that (652) holds for each $n \geq 0$ satisfying $n < m$. Thus, in particular, (652) holds for $n = m - 1$ (since $m - 1 \geq 0$ and $m - 1 < m$). In other words, we have $b_{m-1} = x_{(m-1)+2} x_{m-1}$. In other words,

$$b_{m-1} = x_{m+1} x_{m-1}$$

(since $(m-1) + 2 = m + 1$).

From $m \geq 4$, we obtain $m + 2 \geq 4 + 2 = 6$. Hence, (247) (applied to $n = m + 2$) yields

$$\begin{aligned} x_{m+2} &= \frac{x_{(m+2)-3}\left(x_{(m+2)-1} + x_{(m+2)-5}\right)}{x_{(m+2)-6}} \\ &= \frac{x_{m-1}\left(x_{m+1} + x_{m-3}\right)}{x_{m-4}} \end{aligned} \tag{653}$$

(since $(m+2) - 3 = m - 1$ and $(m+2) - 1 = m + 1$ and $(m+2) - 5 = m - 3$ and $(m+2) - 6 = m - 4$).

However, $m \geq 4$. Thus, (251) (applied to $n = m$) yields

$$\begin{aligned} b_m &= \frac{b_{m-2}\left(b_{m-1} + b_{m-3}\right)}{b_{m-4}} = \frac{x_m x_{m-2}\left(x_{m+1} x_{m-1} + x_{m-1} x_{m-3}\right)}{x_{m-2} x_{m-4}} \\ &\qquad \left(\begin{array}{c} \text{since } b_{m-1} = x_{m+1} x_{m-1} \text{ and } b_{m-2} = x_m x_{m-2} \\ \text{and } b_{m-3} = x_{m-1} x_{m-3} \text{ and } b_{m-4} = x_{m-2} x_{m-4} \end{array}\right) \\ &= \underbrace{\frac{x_{m-1}\left(x_{m+1} + x_{m-3}\right)}{x_{m-4}}}_{\substack{= x_{m+2} \\ \text{(by (653))}}} \cdot x_m = x_{m+2} x_m. \end{aligned}$$

In other words, (652) holds for $n = m$. This completes the induction step. Thus, (652) is proven. In other words, Exercise 4.11.8 **(b)** is solved.

(a) Let $n \in \mathbb{N}$. We must prove that b_n is an integer.

Exercise 4.11.8 **(b)** yields $b_n = x_{n+2} x_n$. However, for each $m \in \mathbb{N}$, the number x_m is an integer (by Exercise 4.11.6 **(b)**, applied to m instead of n). In other words, all entries of the sequence $(x_0, x_1, x_2, x_3, \ldots)$ are integers. Hence, in particular, the two numbers x_{n+2} and x_n are integers (since these two numbers are entries of the sequence $(x_0, x_1, x_2, x_3, \ldots)$). Thus, their product $x_{n+2} x_n$ is an integer as well. In other words, b_n is an integer (since $b_n = x_{n+2} x_n$). This solves Exercise 4.11.8 **(a)**.

(c) Let $(a_0, a_1, a_2, \ldots)$ be the sequence defined in Exercise 4.11.7. Let $n \geq 1$ be an integer.

Exercise 4.11.8 **(b)** yields $b_n = x_{n+2} x_n$. Also, we have $n - 1 \geq 0$ (since $n \geq 1$). Hence, Exercise 4.11.8 **(b)** (applied to $n - 1$ instead of n) yields $b_{n-1} = x_{(n-1)+2} x_{n-1} = x_{n+1} x_{n-1}$ (since $(n-1) + 2 = n + 1$). Thus,

$$\underbrace{b_n}_{= x_{n+2} x_n} \underbrace{b_{n-1}}_{= x_{n+1} x_{n-1}} = x_{n+2} x_n x_{n+1} x_{n-1} = x_{n+2} x_{n+1} x_n x_{n-1}.$$

On the other hand, Exercise 4.11.7 **(b)** yields

$$a_n = x_{n+2}x_{n+1}x_n x_{n-1}.$$

Comparing these two equalities, we obtain $a_n = b_n b_{n-1}$. This solves Exercise 4.11.8 **(c)**. □

□

A.6. Homework set #5 discussion

The following are discussions of the problems on homework set #5 (Section 5.4).

A.6.1. Discussion of Exercise 5.4.1

Discussion of Exercise 5.4.1. We shall show a more general result (due to Hermite):

Proposition A.6.1. Let $n, m \in \mathbb{Z}$. Then:

(a) We have $n+1-m \mid \gcd(n+1, m) \cdot \binom{n}{m}$.

(b) If $m \neq n+1$, then $\dfrac{\gcd(n+1, m)}{n+1-m}\binom{n}{m} \in \mathbb{Z}$.

This proposition is a sort of counterpart to Exercise 4.5.4 **(b)**.

Proof of Proposition A.6.1. **(a)** Applying Lemma A.4.9 **(c)** to $n+1$ and m instead of n and k, we obtain

$$(n+1)\binom{(n+1)-1}{m} = \underbrace{((n+1)-m)}_{=n+1-m}\binom{n+1}{m} = (n+1-m)\binom{n+1}{m}.$$

In view of $(n+1)-1 = n$, this rewrites as

$$(n+1)\binom{n}{m} = (n+1-m)\binom{n+1}{m}. \tag{654}$$

But $\binom{n+1}{m}$ is an integer (since Theorem 4.3.15 yields that $\binom{n+1}{m} \in \mathbb{Z}$). Hence, the equality (654) reveals that

$$n+1-m \mid (n+1)\binom{n}{m}.$$

Thus, Theorem 3.4.11 (applied to $a = n+1-m$, $b = n+1$ and $c = \binom{n}{m}$) yields

$$n+1-m \mid \gcd(n+1-m, n+1) \cdot \binom{n}{m}. \tag{655}$$

But Proposition 3.4.4 **(b)** (applied to $a = n+1-m$ and $b = n+1$) yields

$$\begin{aligned}
\gcd(n+1-m, n+1) &= \gcd\left(n+1, \underbrace{n+1-m}_{=1(n+1)+(-m)}\right) \\
&= \gcd(n+1, 1(n+1)+(-m)) \\
&= \gcd(n+1, -m) \\
&\qquad \left(\begin{array}{c}\text{by Proposition 3.4.4 } \textbf{(c)}, \\ \text{applied to } a = n+1,\ b = -m \text{ and } u = 1\end{array}\right) \\
&= \gcd(n+1, m)
\end{aligned}$$

(by Proposition 3.4.4 **(h)**, applied to $a = n+1$ and $b = m$). Therefore, (655) rewrites as

$$n+1-m \mid \gcd(n+1, m) \cdot \binom{n}{m}.$$

This proves Proposition A.6.1 **(a)**.

(b) Assume that $m \neq n+1$. Thus, $n+1-m \neq 0$. Hence, the fraction $\dfrac{\gcd(n+1,m)}{n+1-m}$ is well-defined.

But $n+1-m \neq 0$. Hence, Proposition 3.1.3 **(d)** (applied to $a = n+1-m$ and $b = \gcd(n+1,m) \cdot \binom{n}{m}$) yields that $n+1-m \mid \gcd(n+1,m) \cdot \binom{n}{m}$ if and only if $\dfrac{\gcd(n+1,m) \cdot \binom{n}{m}}{n+1-m} \in \mathbb{Z}$. Therefore, $\dfrac{\gcd(n+1,m) \cdot \binom{n}{m}}{n+1-m} \in \mathbb{Z}$ (since $n+1-m \mid \gcd(n+1,m) \cdot \binom{n}{m}$). Therefore,

$$\frac{\gcd(n+1,m)}{n+1-m}\binom{n}{m} = \frac{\gcd(n+1,m) \cdot \binom{n}{m}}{n+1-m} \in \mathbb{Z}.$$

This proves Proposition A.6.1 **(b)**. □

Proposition A.6.1 **(a)** quickly yields Exercise 5.4.1:

Solution to Exercise 5.4.1. Exercise 3.5.1 **(a)** (applied to $a = n$) yields $1 \perp n$. According to Proposition 3.5.4, this entails $n \perp 1$.

Proposition A.6.1 **(a)** (applied to $(a+1)n$ and n instead of n and m) yields

$$(a+1)n+1-n \mid \gcd((a+1)n+1, n) \cdot \binom{(a+1)n}{n}.$$

In view of

$$\begin{aligned}
\gcd\left((a+1)\,n+1, n\right) &= \gcd\left(n, (a+1)\,n+1\right) \\
&\qquad \left(\begin{array}{c}\text{by Proposition 3.4.4 } \textbf{(b)}, \\ \text{applied to } (a+1)\,n+1 \text{ and } n \text{ instead of } a \text{ and } b\end{array}\right) \\
&= \gcd\left(n, 1\right) \\
&\qquad \left(\begin{array}{c}\text{by Proposition 3.4.4 } \textbf{(c)}, \\ \text{applied to } n, 1 \text{ and } a+1 \text{ instead of } a, b \text{ and } u\end{array}\right) \\
&= 1 \qquad (\text{since } n \perp 1)
\end{aligned}$$

and

$$(a+1)\,n+1-n = an+1,$$

this rewrites as

$$an+1 \mid 1 \cdot \binom{(a+1)\,n}{n}.$$

In other words, $an+1 \mid \binom{(a+1)\,n}{n}$. This solves Exercise 5.4.1. □

A more explicit solution of Exercise 5.4.1 is also possible: Namely, a simple computation (using Lemma A.4.9 **(c)**) reveals that

$$\frac{1}{an+1}\binom{(a+1)\,n}{n} = (a+1)\binom{(a+1)\,n}{n} - a\binom{(a+1)\,n+1}{n}.$$

But the right hand side of this equality is an integer. Thus, the left hand side is an integer as well; in other words, $an+1 \mid \binom{(a+1)\,n}{n}$.

Note that Exercise 5.4.1 shows that the number $\frac{1}{an+1}\binom{(a+1)\,n}{n}$ is an integer for any $a, n \in \mathbb{N}$. This number is known as a *Fuss-Catalan number* (see, e.g., [Stanle15, Additional problem A14]). □

A.6.2. Discussion of Exercise 5.4.2

Discussion of Exercise 5.4.2. Exercise 5.4.2 is an introduction to the so-called *finite differences* of polynomials. Given a polynomial P, it is common to refer to ΔP as the *first difference* of P, and to $\Delta^m P$ as the *m-th difference* of P. These differences ΔP and $\Delta^m P$ of a polynomial P are discrete analogues of the derivatives f' and $f^{(m)}$ of a function f. Exercise 5.4.2 barely scratches the surface of the concept; more can be found in [GrKnPa94, §2.6 and §5.3, Trick 2] and in [19s-mt3s, Exercise 6][379].

[379] Beware of unusual notations: In [19s-mt3s, Exercise 6], I write $P[a]$ (rather than $P(a)$) for the value of a polynomial P at a number a (this serves to avoid it being confused for a product), and I write P^{Δ} for ΔP.

Finite differences can be applied not just to polynomials, but also to any functions, and in some more general form than in Exercise 5.4.2 (for example, instead of $P(x) - P(x-1)$ we can consider $P(x) - P(x-r)$ for any fixed number r); they are used a lot in numerical mathematics.

We shall give what is likely the shortest solution to Exercise 5.4.2. Various other arguments are possible.

Solution to Exercise 5.4.2. **(a)** Write the polynomial P in the form $P = \sum_{i=0}^{m} c_i x^i$ for some constants $c_0, c_1, \ldots, c_m$. Thus, $P(x) = P = \sum_{i=0}^{m} c_i x^i$. Substituting $x-1$ for x in this equality, we obtain $P(x-1) = \sum_{i=0}^{m} c_i (x-1)^i$. Now, the definition of ΔP yields

$$\begin{aligned}
(\Delta P)(x) &= \underbrace{P(x)}_{=\sum_{i=0}^{m} c_i x^i} - \underbrace{P(x-1)}_{=\sum_{i=0}^{m} c_i (x-1)^i} \\
&= \sum_{i=0}^{m} c_i x^i - \sum_{i=0}^{m} c_i (x-1)^i \\
&= \sum_{i=0}^{m} c_i \left(x^i - (x-1)^i \right). \qquad (656)
\end{aligned}$$

Now, let us rewrite the differences $x^i - (x-1)^i$ on the right hand side of this equality. Namely, for each $i \in \{0, 1, \ldots, m\}$, we have

$$\begin{aligned}
(x-1)^i &= (x + (-1))^i \qquad (\text{since } x - 1 = x + (-1)) \\
&= \sum_{k=0}^{i} \binom{i}{k} x^k (-1)^{i-k} \qquad (\text{by Theorem 4.3.16, applied to } n = i \text{ and } y = -1) \\
&= \sum_{k=0}^{i} (-1)^{i-k} \binom{i}{k} x^k = \sum_{k=0}^{i-1} (-1)^{i-k} \binom{i}{k} x^k + \underbrace{(-1)^{k-k}}_{=(-1)^0=1} \underbrace{\binom{k}{k}}_{\substack{=1 \\ (\text{by } (124))}} x^i \\
&= \sum_{k=0}^{i-1} (-1)^{i-k} \binom{i}{k} x^k + x^i
\end{aligned}$$

and therefore

$$\begin{aligned}
x^i - (x-1)^i &= -\sum_{k=0}^{i-1} (-1)^{i-k} \binom{i}{k} x^k = \sum_{k=0}^{i-1} \underbrace{\left(-(-1)^{i-k} \right)}_{=(-1)^{i-k+1}} \binom{i}{k} x^k \\
&= \sum_{k=0}^{i-1} (-1)^{i-k+1} \binom{i}{k} x^k. \qquad (657)
\end{aligned}$$

Hence, (656) becomes

$$(\Delta P)(x) = \sum_{i=0}^{m} c_i \underbrace{\left(x^i - (x-1)^i\right)}_{\substack{= \sum_{k=0}^{i-1} (-1)^{i-k+1} \binom{i}{k} x^k \\ \text{(by (657))}}}$$

$$= \sum_{i=0}^{m} c_i \sum_{k=0}^{i-1} (-1)^{i-k+1} \binom{i}{k} x^k. \tag{658}$$

Now, we notice that the right hand side of (658) is a sum of powers x^k of x (multiplied by constants)[380]. Moreover, all powers x^k that appear in this sum satisfy $k \leq m-1$ [381]. Thus, the right hand side of (658) is a sum of powers x^k (multiplied by constants) with $k \leq m-1$. In other words, the right hand side of (658) is a polynomial in x of degree $\leq m-1$. Therefore, so is the left hand side of (658). In other words, $(\Delta P)(x)$ is a polynomial in x of degree $\leq m-1$. In other words, the polynomial ΔP has degree $\leq m-1$. This solves Exercise 5.4.2 **(a)**.

(b) This follows by induction on n, using the (already solved) Exercise 5.4.2 **(a)** in the induction step. Here are the details:

Forget that we fixed m and P. The following claim follows from Exercise 5.4.2 **(a)**:

> *Claim 1:* Let $m \in \mathbb{Z}$. Let P be a polynomial in a single variable x. Assume that P has degree $\leq m$. Then, the polynomial ΔP has degree $\leq m-1$.

Now, we must prove the following claim:

> *Claim 2:* Let $m \in \mathbb{Z}$ and $n \in \mathbb{N}$. Let P be a polynomial in a single variable x. Assume that P has degree $\leq m$. Then, the polynomial $\Delta^n P$ has degree $\leq m-n$.

[*Proof of Claim 2:* We proceed by induction on n:

Induction base: It is easy to see that Claim 2 holds for $n = 0$ [382].

[380] Note that a single power x^k may appear multiple times in the sum.

[381] Indeed, if x^k appears in this sum, then $k \in \{0, 1, \ldots, i-1\}$ for some $i \in \{0, 1, \ldots, m\}$; but this implies

$$\begin{aligned} k &\leq \underbrace{i}_{\substack{\leq m \\ \text{(since } i \in \{0,1,\ldots,m\})}} - 1 \qquad \text{(since } k \in \{0, 1, \ldots, i-1\}) \\ &\leq m-1. \end{aligned}$$

[382] *Proof.* The polynomial $\Delta^0 P$ is just equal to P (by its definition) and therefore has degree $\leq m$ (by assumption). In other words, the polynomial $\Delta^0 P$ has degree $\leq m - 0$ (since $m = m - 0$). In other words, Claim 2 holds for $n = 0$.

Induction step: Let $k \in \mathbb{N}$. Assume (as the induction hypothesis) that Claim 2 holds for $n = k$. We must prove that Claim 2 holds for $n = k + 1$.

We have assumed that Claim 2 holds for $n = k$. In other words, the polynomial $\Delta^k P$ has degree $\leq m - k$. The definition of $\Delta^{k+1} P$ yields $\Delta^{k+1} P = \Delta\left(\Delta^k P\right)$. But Claim 1 (applied to $m - k$ and $\Delta^k P$ instead of m and P) shows that the polynomial $\Delta\left(\Delta^k P\right)$ has degree $\leq m - k - 1$ (since the polynomial $\Delta^k P$ has degree $\leq m - k$). Since $\Delta^{k+1} P = \Delta\left(\Delta^k P\right)$ and $m - k - 1 = m - (k + 1)$, we can rewrite this as follows: The polynomial $\Delta^{k+1} P$ has degree $\leq m - (k + 1)$. In other words, Claim 2 holds for $n = k + 1$. This completes the induction step. Thus, Claim 2 is proven.]

Having proved Claim 2, we thus have solved Exercise 5.4.2 **(b)**.

(c) This is similar to the standard inductive proof of the Binomial Theorem (as done, e.g., in [Grinbe15, solution to Exercise 3.6]):

We shall prove Exercise 5.4.2 **(c)** by induction on n:

Induction base: We have $\Delta^0 P = P$ (by the definition of $\Delta^0 P$) and thus

$$\left(\Delta^0 P\right)(x) = P(x) = \sum_{k=0}^{0} (-1)^k \binom{0}{k} P(x - k)$$

(since $\sum_{k=0}^{0} (-1)^k \binom{0}{k} P(x - k) = \underbrace{(-1)^0}_{=1} \underbrace{\binom{0}{0}}_{=1} P\left(\underbrace{x - 0}_{=x}\right) = P(x)$). In other words, Exercise 5.4.2 **(c)** holds for $n = 0$.

Induction step: Let $i \in \mathbb{N}$. Assume (as the induction hypothesis) that Exercise 5.4.2 **(c)** holds for $n = i$. We must prove that Exercise 5.4.2 **(c)** holds for $n = i + 1$.

The definition of $\Delta^{i+1} P$ yields $\Delta^{i+1} P = \Delta\left(\Delta^{(i+1)-1} P\right) = \Delta\left(\Delta^i P\right)$ (since $(i + 1) - 1 = i$). Thus,

$$\begin{aligned} \left(\Delta^{i+1} P\right)(x) &= \left(\Delta\left(\Delta^i P\right)\right)(x) \\ &= \left(\Delta^i P\right)(x) - \left(\Delta^i P\right)(x - 1) \end{aligned} \tag{659}$$

(by the definition of $\Delta\left(\Delta^i P\right)$). We now want to get explicit formulas for the two terms on the right hand side of this equality.

We have assumed that Exercise 5.4.2 **(c)** holds for $n = i$. In other words, we have

$$\left(\Delta^i P\right)(x) = \sum_{k=0}^{i} (-1)^k \binom{i}{k} P(x - k). \tag{660}$$

Substituting $x-1$ for x in this equality, we obtain

$$\begin{aligned}
\left(\Delta^i P\right)(x-1) &= \sum_{k=0}^{i} (-1)^k \binom{i}{k} P(x-1-k) \\
&= \sum_{k=1}^{i+1} (-1)^{k-1} \binom{i}{k-1} P\left(\underbrace{x-1-(k-1)}_{=x-k}\right) \\
&\qquad \text{(here, we have substituted } k-1 \text{ for } k \text{ in the sum)} \\
&= \sum_{k=1}^{i+1} (-1)^{k-1} \binom{i}{k-1} P(x-k). \qquad (661)
\end{aligned}$$

We want to subtract the equality (661) from the equality (660). In order for the result to be simplifiable, it would be good to have the two summations have the same limits. Fortunately, this is not hard to achieve: We just need to stretch the summations to range from 0 to $i+1$.

To wit, we have $i+1 > i$ and thus $\binom{i}{i+1} = 0$ (by Proposition 4.3.4, applied to i and $i+1$ instead of n and k). Now,

$$\begin{aligned}
&\sum_{k=0}^{i+1} (-1)^k \binom{i}{k} P(x-k) \\
&= (-1)^{i+1} \underbrace{\binom{i}{i+1}}_{=0} P(x-(i+1)) + \sum_{k=0}^{i} (-1)^k \binom{i}{k} P(x-k) \\
&= \underbrace{(-1)^{i+1} 0 P(x-(i+1))}_{=0} + \sum_{k=0}^{i} (-1)^k \binom{i}{k} P(x-k) \\
&= \sum_{k=0}^{i} (-1)^k \binom{i}{k} P(x-k).
\end{aligned}$$

Comparing this with (660), we obtain

$$\left(\Delta^i P\right)(x) = \sum_{k=0}^{i+1} (-1)^k \binom{i}{k} P(x-k). \qquad (662)$$

Furthermore, we have $-1 \notin \mathbb{N}$ and thus $\binom{i}{-1} = 0$ (by (118), applied to i and

-1 instead of n and k). Now,

$$\begin{aligned}
&\sum_{k=0}^{i+1}(-1)^{k-1}\binom{i}{k-1}P(x-k)\\
&=(-1)^{0-1}\underbrace{\binom{i}{0-1}}_{=\binom{i}{-1}=0}P(x-0)+\sum_{k=1}^{i+1}(-1)^{k-1}\binom{i}{k-1}P(x-k)\\
&=\underbrace{(-1)^{0-1}0P(x-0)}_{=0}+\sum_{k=1}^{i+1}(-1)^{k-1}\binom{i}{k-1}P(x-k)\\
&=\sum_{k=1}^{i+1}(-1)^{k-1}\binom{i}{k-1}P(x-k).
\end{aligned}$$

Comparing this with (661), we obtain

$$\left(\Delta^i P\right)(x-1)=\sum_{k=0}^{i+1}(-1)^{k-1}\binom{i}{k-1}P(x-k). \tag{663}$$

Now, (659) becomes

$$\begin{aligned}
\left(\Delta^{i+1}P\right)(x)&=\left(\Delta^i P\right)(x)-\left(\Delta^i P\right)(x-1)\\
&=\sum_{k=0}^{i+1}(-1)^k\binom{i}{k}P(x-k)-\sum_{k=0}^{i+1}(-1)^{k-1}\binom{i}{k-1}P(x-k)\\
&\qquad\left(\begin{array}{c}\text{here, we have subtracted the equality (663)}\\ \text{from the equality (662)}\end{array}\right)\\
&=\sum_{k=0}^{i+1}\left((-1)^k\binom{i}{k}P(x-k)-\underbrace{(-1)^{k-1}}_{=-(-1)^k}\binom{i}{k-1}P(x-k)\right)\\
&=\sum_{k=0}^{i+1}\underbrace{\left((-1)^k\binom{i}{k}P(x-k)-\left(-(-1)^k\right)\binom{i}{k-1}P(x-k)\right)}_{=(-1)^k\left(\binom{i}{k}+\binom{i}{k-1}\right)P(x-k)}\\
&=\sum_{k=0}^{i+1}(-1)^k\left(\binom{i}{k}+\binom{i}{k-1}\right)P(x-k). \qquad (664)
\end{aligned}$$

But each $k \in \mathbb{Z}$ satisfies

$$\begin{aligned}\binom{i+1}{k} &= \binom{i+1-1}{k-1} + \binom{i+1-1}{k} \\ &\qquad \text{(by Theorem 4.3.7, applied to } n = i+1\text{)} \\ &= \binom{i}{k-1} + \binom{i}{k} \qquad \text{(since } i+1-1 = i\text{)} \\ &= \binom{i}{k} + \binom{i}{k-1}. \end{aligned} \tag{665}$$

Hence, (664) becomes

$$\begin{aligned}\left(\Delta^{i+1}P\right)(x) &= \sum_{k=0}^{i+1} (-1)^k \underbrace{\left(\binom{i}{k} + \binom{i}{k-1}\right)}_{\substack{=\binom{i+1}{k} \\ \text{(by (665))}}} P(x-k) \\ &= \sum_{k=0}^{i+1} (-1)^k \binom{i+1}{k} P(x-k).\end{aligned}$$

In other words, Exercise 5.4.2 **(c)** holds for $n = i+1$. This completes the induction step. Hence, Exercise 5.4.2 **(c)** is solved.

(d) Let $n \in \mathbb{N}$ satisfy $n > m$. Then, Exercise 5.4.2 **(b)** shows that the polynomial $\Delta^n P$ has degree $\leq m-n$. But $m-n < 0$ (since $n > m$); hence, the only polynomial that has degree $\leq m-n$ is the zero polynomial 0. Therefore, the polynomial $\Delta^n P$ is the zero polynomial 0 (since $\Delta^n P$ has degree $\leq m-n$). In other words, $\Delta^n P = 0$. Thus, $(\Delta^n P)(x) = 0$. But Exercise 5.4.2 **(c)** yields $(\Delta^n P)(x) = \sum_{k=0}^{n} (-1)^k \binom{n}{k} P(x-k)$. Hence,

$$\sum_{k=0}^{n} (-1)^k \binom{n}{k} P(x-k) = (\Delta^n P)(x) = 0.$$

This solves Exercise 5.4.2 **(d)**.

(e) Set $d_i = (-1)^{i-1} \binom{m+1}{i}$ for each $i \in \{1, 2, \ldots, m+1\}$.

We have $m+1 > m$. Hence, Exercise 5.4.2 **(d)** (applied to $n = m+1$) yields

$$\sum_{k=0}^{m+1} (-1)^k \binom{m+1}{k} P(x-k) = 0. \tag{666}$$

Now, let $n \geq m+1$ be an integer. Then, substituting n for x in (666), we obtain

$$\sum_{k=0}^{m+1} (-1)^k \binom{m+1}{k} P(n-k) = 0.$$

Hence,

$$\begin{aligned}
0 &= \sum_{k=0}^{m+1} (-1)^k \binom{m+1}{k} P(n-k) \\
&= \underbrace{(-1)^0}_{=1} \underbrace{\binom{m+1}{0}}_{\substack{=1 \\ \text{(by (119),} \\ \text{applied to } m+1 \text{ instead of } n)}} P\left(\underbrace{n-0}_{=n}\right) + \sum_{k=1}^{m+1} \underbrace{(-1)^k}_{=-(-1)^{k-1}} \binom{m+1}{k} P(n-k) \\
&= P(n) + \sum_{k=1}^{m+1} \left(-(-1)^{k-1}\right) \binom{m+1}{k} P(n-k) \\
&= P(n) - \sum_{k=1}^{m+1} (-1)^{k-1} \binom{m+1}{k} P(n-k),
\end{aligned}$$

so that

$$\begin{aligned}
P(n) &= \sum_{k=1}^{m+1} \underbrace{(-1)^{k-1} \binom{m+1}{k}}_{\substack{=d_k \\ \left(\text{since } d_k = (-1)^{k-1}\binom{m+1}{k} \text{ (by the definition of } d_k\text{)}\right)}} P(n-k) = \sum_{k=1}^{m+1} d_k P(n-k) \\
&= d_1 P(n-1) + d_2 P(n-2) + \cdots + d_{m+1} P(n-(m+1)).
\end{aligned}$$

Now, forget that we fixed n. We thus have shown that

$$P(n) = d_1 P(n-1) + d_2 P(n-2) + \cdots + d_{m+1} P(n-(m+1))$$

for each integer $n \geq m+1$. In other words, the sequence $(P(0), P(1), P(2), \ldots)$ is $(d_1, d_2, \ldots, d_{m+1})$-recurrent (by the definition of "$(d_1, d_2, \ldots, d_{m+1})$-recurrent"). This solves Exercise 5.4.2 **(e)**. □

Let us briefly state three more properties of finite differences. We will need a notation: If Q is a polynomial in a single variable, and if $i \in \mathbb{Z}$, then $\left[x^i\right](Q)$ shall denote the coefficient of x^i in Q. (For example, if $Q(x) = x^4 + 7x - 3$, then $\left[x^5\right](Q) = 0$ and $\left[x^4\right](Q) = 1$ and $\left[x^3\right](Q) = 0$ and $\left[x^0\right](Q) = -3$ and $\left[x^{-1}\right](Q) = 0$.) Now, here are a few more properties of finite differences:

Proposition A.6.2. Let $m \in \mathbb{N}$. Let P be a polynomial in a single variable x. Assume that P has degree $\leq m$. Then:

(a) We have $\left[x^{m-1}\right](\Delta P) = m \cdot \left[x^m\right](P)$.

(b) We have $\left[x^{m-n}\right](\Delta^n P) = n! \binom{m}{n} \cdot \left[x^m\right](P)$.

(c) We have

$$\sum_{k=0}^{m} (-1)^k \binom{m}{k} P(x-k) = m! \cdot [x^m](P).$$

Proposition A.6.2 **(a)** can be proved by retracing the above solution to Exercise 5.4.2 **(a)** and analyzing the appearances of x^{m-1} on the right hand side of (658) (namely, x^{m-1} appears only once on this right hand side, with coefficient $\underbrace{c_m}_{=[x^m](P)} \underbrace{(-1)^{m-(m-1)+1}}_{=1} \underbrace{\binom{m}{m-1}}_{=m} = [x^m](P) \cdot m = m \cdot [x^m](P)$). Proposition A.6.2 **(b)** can be proved by induction on n using Proposition A.6.2 **(a)**, just as we proved Exercise 5.4.2 **(b)** by induction using Exercise 5.4.2 **(a)**. Finally, Proposition A.6.2 **(c)** follows by applying both Proposition A.6.2 **(b)** and Exercise 5.4.2 **(c)** to $n = m$ (and recalling that a polynomial having degree ≤ 0 must be a constant polynomial). □

A.6.3. Discussion of Exercise 5.4.3

Discussion of Exercise 5.4.3. Here is the answer:

Proposition A.6.3. Let $a, b, c \in \mathbb{N}$ be such that $c \leq b$ and $a \leq b$. Then,

$$\sum_{k=c}^{b} \frac{\binom{a}{k}}{\binom{b}{k}} = \frac{\binom{a}{c}(b+1-c)}{\binom{b}{c}(b+1-a)}.$$

In order to prove this, we will need the following identity: If $n, i, j \in \mathbb{N}$ satisfy $j \leq n$, then

$$\binom{n}{j}\binom{j}{i} = \binom{n}{i}\binom{n-i}{j-i}. \tag{667}$$

This fact is a particular case of the *trinomial revision formula* ([Grinbe15, Proposition 3.23]), but let us give a proof of it here, seeing that we have already done most of the work:

[*Proof of (667):* In the case when $i \leq j$, we have already proved (667) in the solution of Exercise 4.5.5 above (indeed, the equality (667) becomes precisely the equality (525) in this case). Thus, for the rest of this proof, we WLOG assume that we don't have $i \leq j$. Hence, $i > j$. Therefore, Proposition 4.3.4 (applied to i and j instead of k and n) yields $\binom{j}{i} = 0$. Also, $j - i < 0$ (since $i > j$) and thus $j - i \notin \mathbb{N}$. Hence, (118) (applied to $n - i$ and $j - i$ instead of n and k) yields $\binom{n-i}{j-i} = 0$. Now, comparing

$\binom{n}{j}\underbrace{\binom{j}{i}}_{=0} = 0$ with $\binom{n}{i}\underbrace{\binom{n-i}{j-i}}_{=0} = 0$, we obtain $\binom{n}{j}\binom{j}{i} = \binom{n}{i}\binom{n-i}{j-i}$. This proves (667).]

Proof of Proposition A.6.3. First, we shall show that certain denominators are nonzero. Indeed, we have $a \leq b < b+1$ and thus $b+1-a > 0$, so that $b+1-a \neq 0$. Furthermore, each $k \in \{c, c+1, \ldots, b\}$ satisfies $\binom{b}{k} \neq 0$ [383]. Applying this to $k = c$, we obtain $\binom{b}{c} \neq 0$ (since $c \in \{c, c+1, \ldots, b\}$ (because $c \leq b$)). Hence, we have shown that all the three denominators $b+1-a$, $\binom{b}{k}$ and $\binom{b}{c}$ appearing in Proposition A.6.3 are nonzero. The fractions are therefore well-defined.

Now, let $k \in \{c, c+1, \ldots, b\}$. Then, $c \leq k \leq b$ and thus $k \geq c \geq 0$ (since $c \in \mathbb{N}$), so that $k \in \mathbb{N}$. Hence, (667) (applied to $n = b$, $i = k$ and $j = a$) yields that

$$\binom{b}{a}\binom{a}{k} = \binom{b}{k}\binom{b-k}{a-k} \tag{668}$$

(since $a \leq b$).

It is easy to see that both numbers $\binom{b}{a}$ and $\binom{b}{k}$ are nonzero[384]. Hence, their product $\binom{b}{a}\binom{b}{k}$ is nonzero. We can therefore divide the equality (668) by this product $\binom{b}{a}\binom{b}{k}$. As a result, we obtain

$$\frac{\binom{a}{k}}{\binom{b}{k}} = \frac{\binom{b-k}{a-k}}{\binom{b}{a}} = \frac{1}{\binom{b}{a}}\binom{b-k}{a-k}. \tag{669}$$

Furthermore, $b-k \in \mathbb{N}$ (since $k \leq b$). Thus, Theorem 4.3.10 (applied to $b-k$ and $a-k$ instead of n and k) yields $\binom{b-k}{a-k} = \binom{b-k}{(b-k)-(a-k)} = \binom{b-k}{b-a}$ (since

[383] *Proof.* Let $k \in \{c, c+1, \ldots, b\}$. Then, $k \geq c$ and $k \leq b$. From $k \geq c \geq 0$ (since $c \in \mathbb{N}$), we obtain $k \in \mathbb{N}$. Hence, Theorem 4.3.8 (applied to $n = b$) yields $\binom{b}{k} = \frac{b!}{k! \cdot (b-k)!} \neq 0$ (since $b! \neq 0$ (because $b!$ is a positive integer)).

[384] *Proof.* We have $a \in \mathbb{N}$ and $b \in \mathbb{N}$ and $a \leq b$. Hence, Theorem 4.3.8 (applied to b and a instead of n and k) yields $\binom{b}{a} = \frac{b!}{a! \cdot (b-a)!} \neq 0$ (since $b! \neq 0$ (because $b!$ is a positive integer)). Also, we have $k \in \mathbb{N}$ and $k \leq b$. Hence, Theorem 4.3.8 (applied to $n = b$) yields $\binom{b}{k} = \frac{b!}{k! \cdot (b-k)!} \neq 0$ (since $b! \neq 0$ (because $b!$ is a positive integer)). Thus, we have now seen that both $\binom{b}{a}$ and $\binom{b}{k}$ are nonzero.

$(b-k)-(a-k)=b-a$). Hence, (669) rewrites as

$$\frac{\binom{a}{k}}{\binom{b}{k}} = \frac{1}{\binom{b}{a}}\binom{b-k}{b-a}. \tag{670}$$

Forget that we fixed k. We thus have proved (670) for each $k \in \{c, c+1, \ldots, b\}$.

We have $c \in \{c, c+1, \ldots, b\}$ (since $c \leq b$). Hence, applying (670) to $k = c$, we obtain

$$\frac{\binom{a}{c}}{\binom{b}{c}} = \frac{1}{\binom{b}{a}}\binom{b-c}{b-a}. \tag{671}$$

Now, summing the equality (669) over all $k \in \{c, c+1, \ldots, b\}$, we obtain

$$\begin{aligned}\sum_{k=c}^{b} \frac{\binom{a}{k}}{\binom{b}{k}} &= \sum_{k=c}^{b} \frac{1}{\binom{b}{a}}\binom{b-k}{b-a} = \frac{1}{\binom{b}{a}}\sum_{k=c}^{b}\binom{b-k}{b-a} \\ &= \frac{1}{\binom{b}{a}}\sum_{i=0}^{b-c}\binom{i}{b-a} \qquad (672) \\ &\qquad \left(\text{here, we have substituted } i \text{ for } b-k \text{ in the sum}\right).\end{aligned}$$

But we have $b-c \in \mathbb{N}$ (since $c \leq b$) and $b-a \in \mathbb{N}$ (since $a \leq b$). Thus, Exercise 4.5.8 **(a)** (applied to $b-c$ and $b-a$ instead of n and k) yields

$$\sum_{i=0}^{b-c}\binom{i}{b-a} = \binom{b-c+1}{b-a+1}. \tag{673}$$

Meanwhile, $b - \underbrace{c}_{\leq b} + 1 \geq b - b + 1 = 1 > 0$, so that $b - c + 1 \in \mathbb{N}$. Also, $b - \underbrace{a}_{\leq b} + 1 \geq b - b + 1 = 1 > 0$ and thus $b - a + 1 \in \mathbb{N}$. Hence, Exercise 4.5.4 **(a)** (applied to $b-c+1$ and $b-a+1$ instead of n and m) yields

$$\frac{b-a+1}{b-c+1}\binom{b-c+1}{b-a+1} = \binom{b-c}{b-a}.$$

Solving this equality for $\binom{b-c+1}{b-a+1}$, we obtain

$$\binom{b-c+1}{b-a+1} = \frac{b-c+1}{b-a+1}\binom{b-c}{b-a}. \tag{674}$$

(Here, we have been able to divide by $b-a+1$ and by $b-c+1$, because both numbers $b-c+1$ and $b-a+1$ are nonzero (since $b-c+1>0$ and $b-a+1>0$).)

Now, (673) becomes

$$\sum_{i=0}^{b-c}\binom{i}{b-a}=\binom{b-c+1}{b-a+1}=\frac{b-c+1}{b-a+1}\binom{b-c}{b-a}$$

(by (674)). Hence, (672) becomes

$$\sum_{k=c}^{b}\frac{\binom{a}{k}}{\binom{b}{k}}=\frac{1}{\binom{b}{a}}\underbrace{\sum_{i=0}^{b-c}\binom{i}{b-a}}_{=\frac{b-c+1}{b-a+1}\binom{b-c}{b-a}}=\frac{1}{\binom{b}{a}}\cdot\frac{b-c+1}{b-a+1}\binom{b-c}{b-a}$$

$$=\underbrace{\frac{b-c+1}{b-a+1}}_{=\frac{b+1-c}{b+1-a}}\cdot\underbrace{\frac{1}{\binom{b}{a}}\binom{b-c}{b-a}}_{\substack{=\frac{\binom{a}{c}}{\binom{b}{c}}\\ \text{(by (671))}}}=\frac{b+1-c}{b+1-a}\cdot\frac{\binom{a}{c}}{\binom{b}{c}}=\frac{\binom{a}{c}(b+1-c)}{\binom{b}{c}(b+1-a)}.$$

This proves Proposition A.6.3. □

We note that the condition $a\leq b$ in Proposition A.6.3 is surprisingly crucial. If $a>b$, then the claim of Proposition A.6.3 may be false even when the denominator is nonzero. Here are some examples of formulas for $\sum_{k=c}^{b}\frac{\binom{a}{k}}{\binom{b}{k}}$ when $a>b$:

- For $a=b+1$, we have

$$\sum_{k=c}^{b}\frac{\binom{a}{k}}{\binom{b}{k}}=\sum_{k=c}^{b}\frac{\binom{b+1}{k}}{\binom{b}{k}}=\sum_{k=c}^{b}\frac{b+1}{b-k+1}=\sum_{i=0}^{b-c}\frac{b+1}{i+1}=(b+1)\sum_{i=0}^{b-c}\frac{1}{i+1}.$$

  Thus, an explicit formula for $\sum_{k=c}^{b}\frac{\binom{b+1}{k}}{\binom{b}{k}}$ (without summation signs) would imply an explicit formula for the so-called *harmonic numbers* $H_n:=\sum_{i=0}^{n-1}\frac{1}{i+1}=$

$\frac{1}{1}+\frac{1}{2}+\cdots+\frac{1}{n}$. Is there an explicit formula for the harmonic numbers? The answer depends on how you define "explicit formula". For some meanings of this word, it has been proved that an explicit formula does not exist (see, e.g., `https://math.stackexchange.com/a/52579/` and the references therein). But if you accept, e.g., floor functions, then you can show that

$$H_n = G_n - (n+1)\left\lfloor \frac{G_n}{n+1} \right\rfloor, \qquad \text{where}$$
$$G_n = \frac{\binom{n+(n+1)!}{n} - 1}{(n+1)!}.$$

This is not hard to prove; but is there any chance for H_n to be easier to compute using this formula than using the definition?

- For $a = b+2$, we have

$$\begin{aligned}
\sum_{k=c}^{b} \frac{\binom{a}{k}}{\binom{b}{k}} &= \sum_{k=c}^{b} \frac{\binom{b+2}{k}}{\binom{b}{k}} = \sum_{k=c}^{b} \frac{(b+1)(b+2)}{(b-k+1)(b-k+2)} \\
&= \sum_{i=0}^{b-c} \frac{(b+1)(b+2)}{(i+1)(i+2)} = \sum_{i=1}^{b-c+1} \frac{(b+1)(b+2)}{i(i+1)} \\
&= (b+1)(b+2) \underbrace{\sum_{i=1}^{b-c+1} \frac{1}{i(i+1)}}_{\substack{= \frac{b-c+1}{b-c+2} \\ \text{(by (29))}}} = (b+1)(b+2) \cdot \frac{b-c+1}{b-c+2}.
\end{aligned}$$

Let us also list formulas for some higher values of a, in the particular case when

$c = 0$:

$$\sum_{k=0}^{b} \frac{\binom{b+2}{k}}{\binom{b}{k}} = (b+1)^2;$$

$$\sum_{k=0}^{b} \frac{\binom{b+3}{k}}{\binom{b}{k}} = \frac{1}{4}(b+1)^2(b+4);$$

$$\sum_{k=0}^{b} \frac{\binom{b+4}{k}}{\binom{b}{k}} = \frac{1}{18}(b+1)^2\left(b^2+8b+18\right);$$

$$\sum_{k=0}^{b} \frac{\binom{b+5}{k}}{\binom{b}{k}} = \frac{1}{96}(b+1)^2(b+6)\left(b^2+7b+16\right);$$

$$\sum_{k=0}^{b} \frac{\binom{b+6}{k}}{\binom{b}{k}} = \frac{1}{600}(b+1)^2\left(b^4+19b^3+136b^2+444b+600\right).$$

Do you see any patterns? (I'm seeing some, but no general rule.) □

A.6.4. Discussion of Exercise 5.4.4

Discussion of Exercise 5.4.4. Here is the answer: For each $n \in \mathbb{N}$, we have

$$a_n = \begin{cases} q^{2^n-1}\left(s+\frac{1}{q}\right)^{2^n} - \frac{1}{q}, & \text{if } q \neq 0; \\ 2^n s, & \text{if } q = 0. \end{cases} \tag{675}$$

Proof of (675): We are in one of the following two cases:
Case 1: We have $q = 0$.
Case 2: We have $q \neq 0$.

Let us first consider Case 1. In this case, we have $q = 0$. We must show that $a_n = 2^n s$ for each $n \in \mathbb{N}$. The definition of our sequence $(a_0, a_1, a_2, \ldots)$ shows that $a_n = a_{n-1}(qa_{n-1}+2)$ for each integer $n \geq 1$. Thus, for each integer $n \geq 1$, we have

$$a_n = a_{n-1}\left(\underbrace{q}_{=0} a_{n-1} + 2\right) = a_{n-1}\underbrace{(0a_{n-1}+2)}_{=2} = a_{n-1}\cdot 2 = 2a_{n-1}.$$

In other words, the sequence $(a_0, a_1, a_2, \ldots)$ is a geometric progression with ratio 2. Hence, for each $n \in \mathbb{N}$, we have $a_n = 2^n \underbrace{a_0}_{=s} = 2^n s$. Thus, (675) is proved in Case 1.

Let us now consider Case 2. In this case, we have $q \neq 0$. We must prove that

$$a_n = q^{2^n - 1} \left(s + \frac{1}{q} \right)^{2^n} - \frac{1}{q} \tag{676}$$

for each $n \in \mathbb{N}$.

Set

$$b_n = a_n + \frac{1}{q} \tag{677}$$

for each $n \in \mathbb{N}$. Thus, $b_0 = \underbrace{a_0}_{=s} + \frac{1}{q} = s + \frac{1}{q}$. The definition of our sequence $(a_0, a_1, a_2, \ldots)$ shows that $a_n = a_{n-1}(qa_{n-1} + 2)$ for each integer $n \geq 1$. Thus, for each integer $n \geq 1$, we have

$$\begin{aligned}
b_n &= a_n + \frac{1}{q} \qquad \text{(by the definition of } b_n\text{)} \\
&= a_{n-1}(qa_{n-1} + 2) + \frac{1}{q} \qquad \text{(since } a_n = a_{n-1}(qa_{n-1} + 2)\text{)} \\
&= \frac{1}{q} \underbrace{\left(q^2 a_{n-1}^2 + 2qa_{n-1} + 1 \right)}_{=(qa_{n-1}+1)^2} = \frac{1}{q}(qa_{n-1} + 1)^2 = q \cdot \left(a_{n-1} + \frac{1}{q} \right)^2 \\
&= q \cdot b_{n-1}^2
\end{aligned} \tag{678}$$

(since the definition of b_{n-1} yields $a_{n-1} + \frac{1}{q} = b_{n-1}$).

From this, we can easily see that

$$b_n = q^{2^n - 1} b_0^{2^n} \tag{679}$$

for each $n \in \mathbb{N}$. [385] Now, for each $n \in \mathbb{N}$, we have

$$\begin{aligned} a_n &= \underbrace{b_n}_{\substack{=q^{2^n-1}b_0^{2^n} \\ \text{(by (679))}}} - \frac{1}{q} \qquad \text{(by (677))} \\ &= q^{2^n-1}b_0^{2^n} - \frac{1}{q} = q^{2^n-1}\left(s+\frac{1}{q}\right)^{2^n} - \frac{1}{q} \qquad \left(\text{since } b_0 = s + \frac{1}{q}\right). \end{aligned}$$

This proves (676). In other words, (675) is proved in Case 1.

Hence, (675) is proved in both Cases 1 and 2, thus holds in full generality.

The formula (675) can be rewritten in two further ways:

$$\begin{aligned} a_n &= \begin{cases} \dfrac{(sq+1)^{2^n}-1}{q}, & \text{if } q \neq 0; \\ 2^n s, & \text{if } q = 0 \end{cases} \\ &= \sum_{k=1}^{2^n} \binom{2^n}{k} s^k q^{k-1}. \end{aligned}$$

(The last equality sign is easy to prove using the binomial theorem.)

[*Remark:* The above solution is not unmotivated. The reason for defining $b_n = a_n + \frac{1}{q}$ was to complete the square on the right hand side of $a_n = a_{n-1}(qa_{n-1}+2)$; indeed, $a_{n-1}(qa_{n-1}+2) + \frac{1}{q} = \frac{1}{q}(qa_{n-1}+1)^2$ ensures that the right hand side

[385] *Proof of (679):* Let us prove (679) by induction on n:

Induction base: We have $2^0 = 1$ and thus $q^{2^0-1}b_0^{2^0} = \underbrace{q^{1-1}}_{=q^0=1} b_0^1 = b_0^1 = b_0$. In other words, $b_0 = q^{2^0-1}b_0^{2^0}$. In other words, (679) holds for $n = 0$.

Induction step: Let m be a positive integer. Assume (as the induction hypothesis) that (679) holds for $n = m-1$. We must prove that (679) holds for $n = m$.

We have assumed that (679) holds for $n = m-1$. In other words, $b_{m-1} = q^{2^{m-1}-1}b_0^{2^{m-1}}$. Now, (678) (applied to $n = m$) yields

$$\begin{aligned} b_m &= q \cdot b_{m-1}^2 = q \cdot \left(q^{2^{m-1}-1}b_0^{2^{m-1}}\right)^2 \qquad \left(\text{since } b_{m-1} = q^{2^{m-1}-1}b_0^{2^{m-1}}\right) \\ &= q \cdot \underbrace{\left(q^{2^{m-1}-1}\right)^2}_{\substack{=q^{2\cdot\left(2^{m-1}-1\right)}=q^{2\cdot 2^{m-1}-2}=q^{2^m-2} \\ \text{(since } 2\cdot 2^{m-1}=2^m\text{)}}} \cdot \underbrace{\left(b_0^{2^{m-1}}\right)^2}_{\substack{=b_0^{2\cdot 2^{m-1}}=b_0^{2^m} \\ \text{(since } 2\cdot 2^{m-1}=2^m\text{)}}} \\ &= \underbrace{q \cdot q^{2^m-2}}_{=q^{1+(2^m-2)}=q^{2^m-1}} \cdot b_0^{2^m} = q^{2^m-1}b_0^{2^m}. \end{aligned}$$

In other words, (679) holds for $n = m$. This completes the induction step. Thus, (679) is proved.

becomes a square after addition of $\frac{1}{q}$, and this reveals that the numbers $b_n = a_n + \frac{1}{q}$ behave more predictably than the numbers a_n.] $\square$

A.6.5. Discussion of Exercise 5.4.5

Discussion of Exercise 5.4.5. The answer is:

$$a_n = f_{2f_n} \qquad \text{for each } n \geq 1, \tag{680}$$

where $(f_0, f_1, f_2, \ldots)$ is the Fibonacci sequence (as defined in Definition 2.2.1).

In order to prove this, we need a lemma:

Lemma A.6.4. Let $(f_0, f_1, f_2, \ldots)$ be the Fibonacci sequence. Let $n, m \in \mathbb{N}$ satisfy $n \geq m$. Then,

$$f_{n-m} f_{n+m} = f_n^2 - (-1)^{n+m} f_m^2.$$

Proof of Lemma A.6.4. The Fibonacci sequence $(f_0, f_1, f_2, \ldots)$ is $(1,1)$-recurrent and satisfies $f_0 = 0$. Thus, Exercise 4.10.4 (applied to $a = 1$ and $x_i = f_i$ and $y_i = f_i$) yields that

$$f_{n-m} f_{n+m} = \underbrace{f_n f_n}_{=f_n^2} - (-1)^{n+m} \underbrace{f_m f_m}_{=f_m^2} = f_n^2 - (-1)^{n+m} f_m^2.$$

This proves Lemma A.6.4. $\square$

Let us now prove (680):

[*Proof of (680):* We proceed by strong induction on n:

Induction step: Let $m \geq 1$ be an integer. Assume (as the induction hypothesis) that (680) holds for $n < m$. We must prove that (680) holds for $n = m$. In other words, we must prove that $a_m = f_{2f_m}$.

This is easily checked by hand if $m \in \{1, 2, 3\}$ (indeed, $a_1 = 1 = f_2 = f_{2 \cdot f_1}$ and $a_2 = 1 = f_2 = f_{2 \cdot f_2}$ and $a_3 = 3 = f_4 = f_{2 \cdot f_3}$). Thus, for the rest of this proof, we WLOG assume that $m \notin \{1, 2, 3\}$. Hence, $m \geq 4$. Thus, the recursive definition of our sequence $(a_1, a_2, a_3, \ldots)$ yields

$$a_m = \frac{a_{m-1}^2 - a_{m-2}^2}{a_{m-3}}. \tag{681}$$

But $m \geq 4$ entails that $m - 3 \geq 1$. Also, $m - 3 < m$. Hence, our induction hypothesis shows that (680) holds for $n = m - 3$. In other words, we have $a_{m-3} = f_{2f_{m-3}}$. Similarly, $a_{m-2} = f_{2f_{m-2}}$ and $a_{m-1} = f_{2f_{m-1}}$. In view of these three equalities, we can rewrite (681) as

$$a_m = \frac{f_{2f_{m-1}}^2 - f_{2f_{m-2}}^2}{f_{2f_{m-3}}}. \tag{682}$$

However, the definition of the Fibonacci sequence yields that $f_n = f_{n-1} + f_{n-2}$ for each $n \geq 2$. Thus, $f_m = f_{m-1} + f_{m-2}$ and $f_{m-1} = f_{m-2} + f_{m-3}$. The latter equality entails $f_{m-3} = f_{m-1} - f_{m-2}$, so that $f_{m-1} - f_{m-2} = f_{m-3} \geq 0$ (since all Fibonacci numbers are nonnegative) and therefore $f_{m-1} \geq f_{m-2}$, so that $2f_{m-1} \geq 2f_{m-2}$. Thus, we can apply Lemma A.6.4 to $2f_{m-1}$ and $2f_{m-2}$ instead of n and m. We thus obtain

$$f_{2f_{m-1}-2f_{m-2}} f_{2f_{m-1}+2f_{m-2}} = f_{2f_{m-1}}^2 - \underbrace{(-1)^{2f_{m-1}+2f_{m-2}}}_{\substack{=1 \\ \text{(since } 2f_{m-1}+2f_{m-2} \\ \text{is clearly even)}}} f_{2f_{m-2}}^2 = f_{2f_{m-1}}^2 - f_{2f_{m-2}}^2.$$

Therefore,

$$f_{2f_{m-1}}^2 - f_{2f_{m-2}}^2 = f_{2f_{m-1}-2f_{m-2}} f_{2f_{m-1}+2f_{m-2}} = f_{2f_{m-3}} f_{2f_m}$$

(since $2f_{m-1} + 2f_{m-2} = 2\underbrace{(f_{m-1} + f_{m-2})}_{=f_m} = 2f_m$ and $2f_{m-1} - 2f_{m-2} = 2\underbrace{(f_{m-1} - f_{m-2})}_{=f_{m-3}} = 2f_{m-3}$). Solving this equation for f_{2f_m}, we obtain $f_{2f_m} = \dfrac{f_{2f_{m-1}}^2 - f_{2f_{m-2}}^2}{f_{2f_{m-3}}}$ (here, we are using the fact that $f_{2f_{m-3}} \neq 0$, which follows from $2\underbrace{f_{m-3}}_{\geq 1} \geq 2$). Comparing this with (682), we obtain $a_m = f_{2f_m}$. This completes the induction step. Thus, (680) is proved.]

The problem is thus solved. We note that the sequence $(a_1, a_2, a_3, \ldots)$ is Sequence A101361 in the OEIS. □

A.6.6. Discussion of Exercise 5.4.6

Discussion of Exercise 5.4.6. Exercise 5.4.6 is the *Zeckendorf theorem*; it appears, e.g., in [Grinbe18, Theorem 2.4]. The proof given in [Grinbe18, §2] is rather similar to our proof of Theorem 5.2.1 above. (See [Grinbe18, detailed version] for a more detailed writeup of this proof. A brief sketch also appears in [GrKnPa94, §6.6], and a fairly similar proof is found in [Hender16]. A different – combinatorial – proof is found in `https://planetmath.org/combinatorialproofofzeckendorfstheorem` . Yet another proof is found in [Chen08].) See [Hoggat72] and [Lengye06] for two (different!) generalizations. □

A.6.7. Discussion of Exercise 5.4.7

Discussion of Exercise 5.4.7. Let us sketch the solution.

The lecturer makes the first announcement at the moment the first student is leaving. Any student who, at this moment, has not entered the classroom yet will be called *late*. The lecturer makes the second announcement when the first **late**

student leaves[386]. We claim that these two announcements are sufficient – i.e., each student hears (at least) one of them.

Indeed, assume the contrary. Thus, there exists a student s who hears neither the first nor the second announcement. Consider this student s. Let f be the first student to leave the classroom, and let ℓ be the first **late** student to leave the classroom (we assume that ℓ exists; the other case is even easier[387]). Now, we shall prove that no two of the three students s, f and ℓ are ever together in the room.

Indeed, the student ℓ is late, and thus does not enter the classroom until after f has left (by the definition of "late"). Hence, f and ℓ are never together in the room. Moreover, the student s cannot have left before f (since f was first to leave), and thus cannot have entered before f has left (since otherwise, s would have witnessed the leaving of f, and thus would have heard the first announcement, contradicting our definition of s). In other words, s is late. Hence, s cannot have left before ℓ (since ℓ was the first **late** student to leave), and thus cannot have entered before ℓ has left (since otherwise, s would have witnessed the leaving of ℓ, and thus would have heard the second announcement, contradicting our definition of s). Hence, s has entered after ℓ has left. Therefore, s and ℓ are never together in the room. Moreover, s and f are never together in the room (since s cannot have entered before f has left, as we have seen above).

We have now shown that f and ℓ are never together in the room; that s and ℓ are never together in the room; and that s and f are never together in the room. In other words, no two of the three students s, f and ℓ are ever together in the room. This contradicts the assumption that among any three (distinct) students, there are at least two that are together in the room at some moment.[388] This contradiction shows that our assumption was false. Hence, we have shown that each student hears (at least) one of the two announcements. Thus, we have solved Exercise 5.4.7 (and, with it, solved Exercise 5.2.1 again). □

A.6.8. Discussion of Exercise 5.4.8

Discussion of Exercise 5.4.8. We shall first sketch the solution, then formalize it in detail.

Let us first restate the exercise in terms of n students and a lecturer:[389]

> *Restated exercise:* Let n and k be positive integers with $k \geq 2$. A lecture is attended by n students. Each student enters the classroom once and leaves it once (and does not come back). We know that among any k

[386] or, in the case when there are no late students, at the end of class

[387] In fact, if ℓ does not exist, then there are no late students at all, and therefore every student hears the first announcement.

[388] Here we have used the fact that the three students s, f and ℓ are indeed distinct (which is clear, since no two of them are ever together in the room).

[389] The assumption that "for any k distinct elements $i_1, i_2, \ldots, i_k \in \{1, 2, \ldots, n\}$, at least two of the k intervals $I_{i_1}, I_{i_2}, \ldots, I_{i_k}$ intersect" thus becomes "among any k distinct students, there are at least two that are together in the room at some moment".

distinct students, there are at least two that are together in the room at some moment. The lecturer wants to make an announcement that every student will hear. Prove that the lecturer can pick $k-1$ moments at which to make the announcement so that each student will hear it. (We assume that the announcement takes no time – i.e., if a student leaves at the same moment that another student enters, the lecturer can make the announcement at this moment and both students will hear it.)

We solve this restated exercise using the following method (which generalizes our above solution to Exercise 5.4.7):

- The lecturer makes the first announcement when the first student leaves. Any student who, at this moment, has not entered the classroom yet will be called *late*.
- The lecturer makes the second announcement when the first **late** student leaves[390]. Any student who, at this moment, has not entered the classroom yet will be called *doubly late*.
- The lecturer makes the second announcement when the first **doubly late** student leaves[391]. Any student who, at this moment, has not entered the classroom yet will be called *triply late*.
- And so on.

The lecturer stops this process after $k-1$ announcements. We claim that each student will have heard (at least) one of these $k-1$ announcements. Indeed, if there was a student s who hears none of them, then we could show (similarly to how we did this in the above solution of Exercise 5.4.7) that there would be k students no two of whom had ever been in the classroom at the same time (namely, the student s, the first student to leave, the first late student to leave, the first doubly late student to leave, etc.). To be fully precise, this argument would have to be tweaked for the case when (e.g.) there are no doubly late students at all; but this is fairly easy.

A conceptually simpler (but essentially equivalent) way to solve Exercise 5.4.8 is by induction on k. The idea (again stated in terms of students) is to make the first announcement when the first student leaves, and then forget about all students who have heard this announcement. The remaining students have the property that among any $k-1$ of them, there are at least two that are together in the room at some moment (because the first student that has left has not overlapped with any of them); thus, by the induction hypothesis, $k-2$ announcements will suffice for them. Altogether, the lecturer will thus make $1+(k-2)=k-1$ announcements and be heard by each student.

[390] or, in the case when there are no late students, at the end of class

[391] Again, if there are no doubly late students, then this announcement will be made at the end of class (and likewise for all future announcements).

Here is a rigorous way to state this proof:[392]

Solution to Exercise 5.4.8 (formal version). Forget that we fixed n, k and $I_1, I_2, \ldots, I_n$. We must solve Exercise 5.4.8. We shall proceed by induction on k.

Induction base: Let us show that Exercise 5.4.8 holds for $k = 2$. In other words, let us verify the following claim:

> *Claim 1:* Let n be a positive integer. Let $I_1, I_2, \ldots, I_n$ be n nonempty finite closed intervals on the real axis. Assume that for any 2 distinct elements $i_1, i_2 \in \{1, 2, \ldots, n\}$, at least two of the 2 intervals I_{i_1}, I_{i_2} intersect. Then, there exist $2 - 1$ reals $a_1, a_2, \ldots, a_{2-1}$ such that each of the intervals $I_1, I_2, \ldots, I_n$ contains at least one of $a_1, a_2, \ldots, a_{2-1}$.

[*Proof of Claim 1:* Write each interval I_m as $[p_m, q_m]$ for two reals p_m and q_m. Thus, each $m \in \{1, 2, \ldots, n\}$ satisfies

$$I_m = [p_m, q_m] \tag{683}$$

and therefore $[p_m, q_m] = I_m \neq \varnothing$ (since I_m is nonempty), so that

$$p_m \leq q_m. \tag{684}$$

Let

$$b = \min \{q_1, q_2, \ldots, q_n\}. \tag{685}$$

We shall show that each of the intervals $I_1, I_2, \ldots, I_n$ contains b.

Indeed, let $j \in \{1, 2, \ldots, n\}$. We shall show that $b \in I_j$.

We have

$$b = \min \{q_1, q_2, \ldots, q_n\} \in \{q_1, q_2, \ldots, q_n\}$$

(since the minimum of a set always lies in this set). In other words, there exists some $k \in \{1, 2, \ldots, n\}$ such that $b = q_k$. Consider this k.

Applying (683) to $m = k$, we obtain $I_k = [p_k, q_k]$. But applying (684) to $m = k$, we obtain $p_k \leq q_k$. Hence, $q_k \in [p_k, q_k] = I_k$, so that $b = q_k \in I_k$.

Now, our goal is to show that $b \in I_j$. If $j = k$, then this follows directly from $b \in I_k$. Thus, for the rest of this proof, we WLOG assume that $j \neq k$. Thus, j and k are two distinct elements of $\{1, 2, \ldots, n\}$.

Recall our assumption that for any 2 distinct elements $i_1, i_2 \in \{1, 2, \ldots, n\}$, at least two of the 2 intervals I_{i_1}, I_{i_2} intersect. Applying this to $i_1 = j$ and $i_2 = k$, we conclude that at least two of the 2 intervals I_j, I_k intersect (since j, k are 2 distinct elements of $\{1, 2, \ldots, n\}$). In other words, the two intervals I_j and I_k intersect (since the only way to pick two of the 2 intervals I_j, I_k is to pick I_j and I_k). In other words, $I_j \cap I_k \neq \varnothing$.

[392] Unfortunately, this is a good example of a proof that heavily increases in length and decreases in readability when it is formalized. Arguably, it could be that it is just **me** formalizing it badly (suggestions for improvement are greatly encouraged!); but it is a sad empirical fact that this happens to most writers with most proofs in combinatorics.

Recall that $I_k = [p_k, q_k]$. Also, (683) (applied to $m = j$) yields $I_j = \left[p_j, q_j\right]$. Recall that $I_j \cap I_k \neq \varnothing$. In other words, there exists some $c \in I_j \cap I_k$. Consider this c. From $c \in I_j \cap I_k \subseteq I_j = \left[p_j, q_j\right]$, we obtain $c \geq p_j$. From $c \in I_j \cap I_k \subseteq I_k = [p_k, q_k]$, we obtain $c \leq q_k$. Hence, $q_k \geq c \geq p_j$. Recalling that $b = q_k$, we thus obtain $b = q_k \geq p_j$.

We have $b = \min\{q_1, q_2, \ldots, q_n\}$. In other words, b is the **smallest** of the n numbers $q_1, q_2, \ldots, q_n$. Hence, $b \leq q_i$ for each $i \in \{1, 2, \ldots, n\}$. Applying this to $i = j$, we find $b \leq q_j$. Combining this with $b \geq p_j$, we obtain $b \in \left[p_j, q_j\right] = I_j$.

Forget that we fixed j. We thus have shown that $b \in I_j$ for each $j \in \{1, 2, \ldots, n\}$. In other words, each of the intervals $I_1, I_2, \ldots, I_n$ contains b. Thus, there exists a real a_1 such that each of the intervals $I_1, I_2, \ldots, I_n$ contains at least one of a_1 (namely, $a_1 = b$). In other words, there exist $2 - 1$ reals $a_1, a_2, \ldots, a_{2-1}$ such that each of the intervals $I_1, I_2, \ldots, I_n$ contains at least one of $a_1, a_2, \ldots, a_{2-1}$ (because $2 - 1$ reals $a_1, a_2, \ldots, a_{2-1}$ are the same as a single real a_1). This proves Claim 1.]

Claim 1 is precisely the claim of Exercise 5.4.8 for $k = 2$; thus, Exercise 5.4.8 holds for $k = 2$ (since we have proved Claim 1). This completes the induction base.

Induction step: Fix an integer $\ell \geq 2$. Assume (as the induction hypothesis) that Exercise 5.4.8 holds for $k = \ell$. We must prove that Exercise 5.4.8 holds for $k = \ell + 1$.

We have assumed that Exercise 5.4.8 holds for $k = \ell$. In other words, the following claim holds:

> *Claim 2:* Let n be a positive integer. Let $I_1, I_2, \ldots, I_n$ be n nonempty finite closed intervals on the real axis. Assume that for any ℓ distinct elements $i_1, i_2, \ldots, i_\ell \in \{1, 2, \ldots, n\}$, at least two of the ℓ intervals $I_{i_1}, I_{i_2}, \ldots, I_{i_\ell}$ intersect. Then, there exist $\ell - 1$ reals $a_1, a_2, \ldots, a_{\ell-1}$ such that each of the intervals $I_1, I_2, \ldots, I_n$ contains at least one of $a_1, a_2, \ldots, a_{\ell-1}$.

We must prove that Exercise 5.4.8 holds for $k = \ell + 1$. In other words, we must prove the following claim:

> *Claim 3:* Let n be a positive integer. Let $I_1, I_2, \ldots, I_n$ be n nonempty finite closed intervals on the real axis. Assume that for any $\ell + 1$ distinct elements $i_1, i_2, \ldots, i_{\ell+1} \in \{1, 2, \ldots, n\}$, at least two of the $\ell + 1$ intervals $I_{i_1}, I_{i_2}, \ldots, I_{i_{\ell+1}}$ intersect. Then, there exist ℓ reals $a_1, a_2, \ldots, a_\ell$ such that each of the intervals $I_1, I_2, \ldots, I_n$ contains at least one of $a_1, a_2, \ldots, a_\ell$.

[*Proof of Claim 3:* Write each interval I_m as $[p_m, q_m]$ for two reals p_m and q_m. Thus, each $m \in \{1, 2, \ldots, n\}$ satisfies

$$I_m = [p_m, q_m] \tag{686}$$

and therefore $[p_m, q_m] = I_m \neq \varnothing$ (since I_m is nonempty), so that

$$p_m \leq q_m. \tag{687}$$

Let

$$b = \min\{q_1, q_2, \ldots, q_n\}. \tag{688}$$

Thus, b is the **smallest** among the n numbers $q_1, q_2, \ldots, q_n$. Hence,

$$b \leq q_j \qquad \text{for each } j \in \{1, 2, \ldots, n\}. \tag{689}$$

We note that our situation is symmetric in the n intervals $I_1, I_2, \ldots, I_n$; that is, we can permute these n intervals arbitrarily (as long as we accordingly permute the n numbers $p_1, p_2, \ldots, p_n$ and the n numbers $q_1, q_2, \ldots, q_n$) without changing our assumptions or the claim we are trying to prove. Hence, we can WLOG assume that we have

$$p_1 \geq p_2 \geq \cdots \geq p_n \tag{690}$$

(since we can always achieve this by appropriately permuting $I_1, I_2, \ldots, I_n$). Assume this.

We have $b = \min\{q_1, q_2, \ldots, q_n\} \in \{q_1, q_2, \ldots, q_n\}$ (since the minimum of a set always belongs to this set). In other words, there exists some $k \in \{1, 2, \ldots, n\}$ such that $b = q_k$. Consider this k. Applying (687) to $m = k$, we obtain $p_k \leq q_k = b$ (since $b = q_k$). Thus, there exists some $i \in \{1, 2, \ldots, n\}$ such that $p_i \leq b$ (namely, $i = k$). Let s be the **smallest** such i. Thus, we have $p_s \leq b$, but none of the numbers $p_1, p_2, \ldots, p_{s-1}$ is $\leq b$.

It is now easy to see that each of the intervals $I_s, I_{s+1}, \ldots, I_n$ contains b [393]. If $s = 1$, then Claim 3 now easily follows[394]. Hence, for the rest of the proof, we WLOG assume that $s \neq 1$. Hence, $s > 1$ (since $s \in \{1, 2, \ldots, n\}$). Thus, $s - 1 > 0$, so that $s - 1$ is a positive integer. Also, $s \in \{1, 2, \ldots, n\}$, so that $s \leq n$ and thus $s - 1 \leq s \leq n$. Hence, $\{1, 2, \ldots, s-1\} \subseteq \{1, 2, \ldots, n\}$.

If $u \in \{1, 2, \ldots, s-1\}$, then

$$\text{the intervals } I_u \text{ and } I_k \text{ do not intersect.} \tag{691}$$

[*Proof of (691):* Let $u \in \{1, 2, \ldots, s-1\}$. We must show that the intervals I_u and I_k do not intersect.

Assume the contrary. Thus, the intervals I_u and I_k intersect. In other words, $I_u \cap I_k \neq \varnothing$. Hence, there exists some $c \in I_u \cap I_k$. Consider this c. We have $c \in I_u \cap I_k \subseteq I_u = [p_u, q_u]$ (by (686), applied to $m = u$). Thus, $c \geq p_u$, so that

[393] *Proof.* Let $j \in \{s, s+1, \ldots, n\}$. We shall show that $b \in I_j$.

Indeed, (686) (applied to $m = j$) yields $I_j = [p_j, q_j]$. Now, $j \in \{s, s+1, \ldots, n\}$, so that $s \leq j \leq n$. But (690) shows that $p_\alpha \geq p_\beta$ for any two elements α and β of $\{1, 2, \ldots, n\}$ satisfying $\alpha \leq \beta$. Applying this to $\alpha = s$ and $\beta = j$, we find $p_s \geq p_j$ (since $s \leq j$), so that $p_j \leq p_s \leq b$. Hence, $b \geq p_j$. But (689) yields $b \leq q_j$. Combining this with $b \geq p_j$, we find $b \in [p_j, q_j] = I_j$.

Forget that we fixed j. We thus have shown that $b \in I_j$ for each $j \in \{s, s+1, \ldots, n\}$. In other words, each of the intervals $I_s, I_{s+1}, \ldots, I_n$ contains b.

[394] *Proof.* Assume that $s = 1$. We have just shown that each of the intervals $I_s, I_{s+1}, \ldots, I_n$ contains b. In view of $s = 1$, we can rewrite this as follows: Each of the intervals $I_1, I_2, \ldots, I_n$ contains b. In other words, each of the intervals $I_1, I_2, \ldots, I_n$ contains at least one of the ℓ reals $\underbrace{b, b, \ldots, b}_{\ell \text{ times}}$ (since $\ell \geq 2 \geq 1$). Hence, there exist ℓ reals $a_1, a_2, \ldots, a_\ell$ such that each of the intervals $I_1, I_2, \ldots, I_n$ contains at least one of $a_1, a_2, \ldots, a_\ell$ (namely, $a_i = b$ for each $i \in \{1, 2, \ldots, \ell\}$). This proves Claim 3. Thus, Claim 3 is proved under the assumption that $s = 1$.

$p_u \leq c$. We have $c \in I_u \cap I_k \subseteq I_k = [p_k, q_k]$ (by (686), applied to $m = k$). Thus, $c \leq q_k = b$ (since $b = q_k$). Hence, $p_u \leq c \leq b$. However, p_u is one of the numbers $p_1, p_2, \ldots, p_{s-1}$ (since $u \in \{1, 2, \ldots, s-1\}$). Hence, p_u is not $\leq b$ (since we have shown that none of the numbers $p_1, p_2, \ldots, p_{s-1}$ is $\leq b$). This contradicts $p_u \leq b$. This contradiction shows that our assumption was false. Hence, (691) is proven.]

We recall our assumption that

$$\left(\begin{array}{c} \text{for any } \ell+1 \text{ distinct elements } i_1, i_2, \ldots, i_{\ell+1} \in \{1, 2, \ldots, n\}, \\ \text{at least two of the } \ell+1 \text{ intervals } I_{i_1}, I_{i_2}, \ldots, I_{i_{\ell+1}} \text{ intersect} \end{array} \right). \tag{692}$$

We shall now prove that

$$\left(\begin{array}{c} \text{for any } \ell \text{ distinct elements } i_1, i_2, \ldots, i_\ell \in \{1, 2, \ldots, s-1\}, \\ \text{at least two of the } \ell \text{ intervals } I_{i_1}, I_{i_2}, \ldots, I_{i_\ell} \text{ intersect} \end{array} \right). \tag{693}$$

[*Proof of (693):* Let $i_1, i_2, \ldots, i_\ell \in \{1, 2, \ldots, s-1\}$ be ℓ distinct elements. We must prove that at least two of the ℓ intervals $I_{i_1}, I_{i_2}, \ldots, I_{i_\ell}$ intersect.

Assume the contrary. Thus, no two of the ℓ intervals $I_{i_1}, I_{i_2}, \ldots, I_{i_\ell}$ intersect. Moreover, if $g \in \{1, 2, \ldots, \ell\}$, then $i_g \in \{1, 2, \ldots, s-1\}$ (since $i_1, i_2, \ldots, i_\ell \in \{1, 2, \ldots, s-1\}$), and therefore the intervals I_{i_g} and I_k do not intersect (by (691), applied to $u = i_g$). In other words, none of the intervals $I_{i_1}, I_{i_2}, \ldots, I_{i_\ell}$ intersects I_k.

We note that $i_1, i_2, \ldots, i_\ell$ are ℓ distinct elements of $\{1, 2, \ldots, s-1\}$ and therefore are ℓ distinct elements of $\{1, 2, \ldots, n\}$ (since $\{1, 2, \ldots, s-1\} \subseteq \{1, 2, \ldots, n\}$). We extend the ℓ-tuple $(i_1, i_2, \ldots, i_\ell)$ to an $(\ell+1)$-tuple $(i_1, i_2, \ldots, i_{\ell+1})$ by setting $i_{\ell+1} = k$. Thus, $i_1, i_2, \ldots, i_{\ell+1}$ are $\ell+1$ elements of $\{1, 2, \ldots, n\}$ (because $i_1, i_2, \ldots, i_\ell$ are ℓ elements of $\{1, 2, \ldots, n\}$, and because $i_{\ell+1} = k \in \{1, 2, \ldots, n\}$). Furthermore, these $\ell+1$ elements $i_1, i_2, \ldots, i_{\ell+1}$ are distinct[395]. Hence, (692) yields that at least two of the $\ell+1$ intervals $I_{i_1}, I_{i_2}, \ldots, I_{i_{\ell+1}}$ intersect. In other words, there exist two numbers

[395] *Proof.* Assume the contrary. Thus, two of the $\ell+1$ elements $i_1, i_2, \ldots, i_{\ell+1}$ are equal. In other words, there exist two numbers $\alpha, \beta \in \{1, 2, \ldots, \ell+1\}$ with $\alpha < \beta$ and $i_\alpha = i_\beta$. Consider these α and β.

We have $\alpha < \beta \leq \ell+1$ (since $\beta \in \{1, 2, \ldots, \ell+1\}$), so that $\alpha \leq (\ell+1) - 1$ (since α and $\ell+1$ are integers). Thus, $\alpha \leq (\ell+1) - 1 = \ell$, so that $\alpha \in \{1, 2, \ldots, \ell\}$ (since $\alpha \in \{1, 2, \ldots, \ell+1\} \subseteq \{1, 2, 3, \ldots\}$). Therefore, i_α is one of the ℓ elements $i_1, i_2, \ldots, i_\ell$. Hence, $i_\alpha \in \{1, 2, \ldots, s-1\}$ (since $i_1, i_2, \ldots, i_\ell$ are elements of $\{1, 2, \ldots, s-1\}$). Hence, (691) (applied to $u = i_\alpha$) yields that the intervals I_{i_α} and I_k do not intersect. In other words, $I_{i_\alpha} \cap I_k = \varnothing$.

If we had $\beta = \ell+1$, then we would have

$$\begin{aligned} i_\alpha = i_\beta = i_{\ell+1} & \qquad (\text{since } \beta = \ell+1) \\ = k, & \end{aligned}$$

which would entail $I_{i_\alpha} \cap I_k = I_k \cap I_k = I_k \neq \varnothing$ (since the interval I_k is nonempty), which would contradict $I_{i_\alpha} \cap I_k = \varnothing$. Thus, we cannot have $\beta = \ell+1$. Hence, $\beta \neq \ell+1$. From $\beta \in \{1, 2, \ldots, \ell+1\}$ and $\beta \neq \ell+1$, we obtain $\beta \in \{1, 2, \ldots, \ell+1\} \setminus \{\ell+1\} = \{1, 2, \ldots, \ell\}$.

Thus, we know that α and β are two distinct elements of $\{1, 2, \ldots, \ell\}$ (since $\alpha \in \{1, 2, \ldots, \ell\}$ and $\beta \in \{1, 2, \ldots, \ell\}$ and $\alpha < \beta$). Hence, $i_\alpha \neq i_\beta$ (since the ℓ elements $i_1, i_2, \ldots, i_\ell$ are distinct). This contradicts $i_\alpha = i_\beta$. This contradiction shows that our assumption was wrong, qed.

$\alpha, \beta \in \{1, 2, \ldots, \ell + 1\}$ with $\alpha < \beta$ and with the property that the intervals I_{i_α} and I_{i_β} intersect. Consider these α and β.

We have $\alpha < \beta \leq \ell + 1$ (since $\beta \in \{1, 2, \ldots, \ell + 1\}$), so that $\alpha \leq (\ell + 1) - 1$ (since α and $\ell + 1$ are integers). Thus, $\alpha \leq (\ell + 1) - 1 = \ell$, so that $\alpha \in \{1, 2, \ldots, \ell\}$ (since $\alpha \in \{1, 2, \ldots, \ell + 1\} \subseteq \{1, 2, 3, \ldots\}$). Therefore, i_α is one of the ℓ elements $i_1, i_2, \ldots, i_\ell$. Hence, $i_\alpha \in \{1, 2, \ldots, s - 1\}$ (since $i_1, i_2, \ldots, i_\ell$ are elements of $\{1, 2, \ldots, s - 1\}$). Hence, (691) (applied to $u = i_\alpha$) yields that the intervals I_{i_α} and I_k do not intersect.

However, the intervals I_{i_α} and I_{i_β} intersect. If we had $i_\beta = k$, then this would entail that the intervals I_{i_α} and I_k intersect; but this would contradict the fact that the intervals I_{i_α} and I_k do not intersect. Hence, we cannot have $i_\beta = k$. Thus, we have $i_\beta \neq k$.

If we had $\beta = \ell + 1$, then we would have $i_\beta = i_{\ell+1} = k$, which would contradict $i_\beta \neq k$. Thus, we cannot have $\beta = \ell + 1$. Hence, $\beta \neq \ell + 1$. From $\beta \in \{1, 2, \ldots, \ell + 1\}$ and $\beta \neq \ell + 1$, we obtain $\beta \in \{1, 2, \ldots, \ell + 1\} \setminus \{\ell + 1\} = \{1, 2, \ldots, \ell\}$.

Thus, we know that α and β are two distinct elements of $\{1, 2, \ldots, \ell\}$ (since $\alpha \in \{1, 2, \ldots, \ell\}$ and $\beta \in \{1, 2, \ldots, \ell\}$ and $\alpha < \beta$). Hence, the intervals I_{i_α} and I_{i_β} do not intersect (since no two of the ℓ intervals $I_{i_1}, I_{i_2}, \ldots, I_{i_\ell}$ intersect). This contradicts the fact that the intervals I_{i_α} and I_{i_β} intersect. This contradiction shows that our assumption was wrong. Hence, we have shown that at least two of the ℓ intervals $I_{i_1}, I_{i_2}, \ldots, I_{i_\ell}$ intersect. This proves (693).]

Now, (693) shows that for any ℓ distinct elements $i_1, i_2, \ldots, i_\ell \in \{1, 2, \ldots, s - 1\}$, at least two of the ℓ intervals $I_{i_1}, I_{i_2}, \ldots, I_{i_\ell}$ intersect. Hence, we can apply Claim 2 to $s - 1$ instead of n (since $s - 1$ is a positive integer, and since $I_1, I_2, \ldots, I_{s-1}$ are $s - 1$ nonempty finite closed intervals on the real axis). As a result, we conclude that there exist $\ell - 1$ reals $a_1, a_2, \ldots, a_{\ell-1}$ such that each of the intervals $I_1, I_2, \ldots, I_{s-1}$ contains at least one of $a_1, a_2, \ldots, a_{\ell-1}$. Consider these $\ell - 1$ reals $a_1, a_2, \ldots, a_{\ell-1}$. Extend the $(\ell - 1)$-tuple $(a_1, a_2, \ldots, a_{\ell-1})$ to an ℓ-tuple $(a_1, a_2, \ldots, a_\ell)$ by setting $a_\ell = b$.

Now, it is easy to see that for each $h \in \{1, 2, \ldots, n\}$,

$$\text{the interval } I_h \text{ contains at least one of } a_1, a_2, \ldots, a_\ell. \tag{694}$$

[*Proof of (694):* Let $h \in \{1, 2, \ldots, n\}$. We must prove that the interval I_h contains at least one of $a_1, a_2, \ldots, a_\ell$. We are in one of the following two cases:

Case 1: We have $h < s$.

Case 2: We have $h \geq s$.

Let us first consider Case 1. In this case, we have $h < s$. Thus, $h \leq s - 1$ (since h and s are integers), so that $h \in \{1, 2, \ldots, s - 1\}$ (since $h \in \{1, 2, \ldots, n\} \subseteq \{1, 2, 3, \ldots\}$). Thus, I_h is one of the intervals $I_1, I_2, \ldots, I_{s-1}$.

Recall that each of the intervals $I_1, I_2, \ldots, I_{s-1}$ contains at least one of $a_1, a_2, \ldots, a_{\ell-1}$. Hence, I_h contains at least one of $a_1, a_2, \ldots, a_{\ell-1}$ (since I_h is one of the intervals $I_1, I_2, \ldots, I_{s-1}$). Therefore, I_h contains at least one of $a_1, a_2, \ldots, a_\ell$. Thus, (694) is proved in Case 1.

Let us now consider Case 2. In this case, we have $h \geq s$. Thus, $h \in \{s, s+1, \ldots, n\}$ (since $h \in \{1, 2, \ldots, n\}$ entails $h \leq n$). Therefore, I_h is one of the intervals $I_s, I_{s+1}, \ldots, I_n$.

Recall that each of the intervals $I_s, I_{s+1}, \ldots, I_n$ contains b. Hence, I_h contains b (since I_h is one of the intervals $I_s, I_{s+1}, \ldots, I_n$). In other words, I_h contains a_ℓ (since $a_\ell = b$). Hence, I_h contains at least one of $a_1, a_2, \ldots, a_\ell$. Thus, (694) is proved in Case 2.

We have now proved (694) in each of the two Cases 1 and 2. Thus, (694) always holds.]

So we have shown that for each $h \in \{1, 2, \ldots, n\}$, the interval I_h contains at least one of $a_1, a_2, \ldots, a_\ell$. In other words, each of the intervals $I_1, I_2, \ldots, I_n$ contains at least one of $a_1, a_2, \ldots, a_\ell$.

Now, we have found ℓ reals $a_1, a_2, \ldots, a_\ell$ with the property that each of the intervals $I_1, I_2, \ldots, I_n$ contains at least one of $a_1, a_2, \ldots, a_\ell$. Hence, such ℓ reals exist. This proves Claim 3.]

Now, Claim 3 is proved; in other words, we have proved that Exercise 5.4.8 holds for $k = \ell + 1$. This completes the induction step. Thus, Exercise 5.4.8 is solved. □

We note that, if Exercise 5.4.8 is restated in terms of students and lecturers, then the ℓ numbers $a_1, a_2, \ldots, a_\ell$ constructed in the above solution (i.e., the ℓ moments at which the lecturer needs to make their announcement) will be known to the lecturer as soon as they arrive, at least if the number n is known to the lecturer. (If the number n is not known, a little tweak has to be made.)

We note furthermore that Exercise 5.4.8 has a generalization, in which the words "at least two" are replaced by "at least q" for some fixed integer $q \geq 2$. See [Gunder10, Exercise 437] for this generalization. (Note that it is not hard to recover this generalization from Exercise 5.4.8.) □

A.6.9. Discussion of Exercise 5.4.9

Discussion of Exercise 5.4.9. Exercise 5.4.9 can easily be reduced to Exercise 5.2.4:

First solution to Exercise 5.4.9 (sketched). Let C be a positive real number large enough that all of the n numbers $C + x_0, C + x_1, \ldots, C + x_{n-1}$ are positive.

Consider a circular track with n gas stations on it. The n gas stations are labelled $0, 1, \ldots, n-1$ (in the order in which a car would encounter them when driving around the track). Assume that the gas stations are placed at equal distances (i.e., the distance between any two consecutive gas stations is the same), and driving from any gas station to the next consumes C gallons of gas. Thus, a car needs nC gallons of gas in total to complete the entire track.

Assume that gas station i has $C + x_i$ gallons of gas available. Then, the n gas stations have just enough gas for a car to complete the entire track (indeed, the

total amount of gas in the stations is

$$\begin{aligned}&(C+x_0)+(C+x_1)+\cdots+(C+x_{n-1})\\&=nC+\underbrace{(x_0+x_1+\cdots+x_{n-1})}_{=0}=nC+0=nC,\end{aligned}$$

which is exactly what the car needs to complete the track).

Now, Exercise 5.2.4 yields that at least one of these n gas stations has the property that if the car starts at this gas station with an initially empty gas tank, then it can traverse the entire track without ever running out of gas[396]. Let k be this gas station. Then, for each $m \in \{k+1, k+2, \ldots, k+n\}$, we have

$$(C+x_k)+(C+x_{k+1})+\cdots+(C+x_{m-1}) \geq (m-k)\,C,$$

since the car has made it from gas station k to gas station m (collecting $(C+x_k)+(C+x_{k+1})+\cdots+(C+x_{m-1})$ gallons of gas along the way, and using up $(m-k)\,C$ gallons) without running out of gas. This inequality rewrites as

$$(m-k)\,C+(x_k+x_{k+1}+\cdots+x_{m-1}) \geq (m-k)\,C.$$

Subtracting $(m-k)\,C$ from both sides of this inequality, we obtain

$$x_k+x_{k+1}+\cdots+x_{m-1} \geq 0.$$

Thus, we have shown that the inequality $x_k+x_{k+1}+\cdots+x_{m-1} \geq 0$ holds for each $m \in \{k+1, k+2, \ldots, k+n\}$. Renaming m as $m+1$, we thus conclude that the inequality $x_k+x_{k+1}+\cdots+x_m \geq 0$ holds for each $m \in \{k, k+1, \ldots, k+n-1\}$.

However, recall that the sequence $(x_0, x_1, x_2, \ldots)$ is n-periodic and satisfies $x_0+x_1+\cdots+x_{n-1}=0$. Thus, it is easily seen that the sum of **any** n consecutive entries of the sequence $(x_0, x_1, x_2, \ldots)$ is 0. Thus,

$$x_{m+1}+x_{m+2}+\cdots+x_{m+n}=0 \tag{695}$$

for each $m \in \mathbb{N}$. From this, we can easily conclude that the sequence

$$(x_k, x_k+x_{k+1}, x_k+x_{k+1}+x_{k+2}, x_k+x_{k+1}+x_{k+2}+x_{k+3}, \ldots)$$

(that is, the sequence of all sums of the form $x_k+x_{k+1}+\cdots+x_m$ with $m \geq k$) is n-periodic as well (because any $m \geq k$ satisfies

$$\begin{aligned}x_k+x_{k+1}+\cdots+x_{m+n}&=(x_k+x_{k+1}+\cdots+x_m)+\underbrace{(x_{m+1}+x_{m+2}+\cdots+x_{m+n})}_{\substack{=0\\(\text{by }(695))}}\\&=x_k+x_{k+1}+\cdots+x_m\end{aligned}$$

). But the first n entries of this sequence are nonnegative (since $x_k+x_{k+1}+\cdots+x_m \geq 0$ holds for each $m \in \{k, k+1, \ldots, k+n-1\}$). Hence, all its entries are nonnegative (because it is n-periodic). In other words, the inequality $x_k+x_{k+1}+\cdots+x_m \geq 0$ holds for each $m \geq k$. This solves Exercise 5.4.9. □

[396] provided that it refuels at every gas station it comes across (including the one at which it starts)

The following second solution may appear very different from the first, but in truth is closely related: It simply formalizes and applies the idea of our solution to Exercise 5.2.4 straight to our sequence $(x_0, x_1, x_2, \ldots)$, without the detour of restating it in terms of cars and gas stations. The resulting argument is more rigorous and perhaps also more readable, although less vivid and intuitive.

Second solution to Exercise 5.4.9. For each $m \in \mathbb{N}$, we define a real number

$$s_m = x_0 + x_1 + \cdots + x_{m-1}. \tag{696}$$

The sequence $(x_0, x_1, x_2, \ldots)$ is n-periodic. In other words, n is a period of this sequence $(x_0, x_1, x_2, \ldots)$ (by the definition of "n-periodic"). In other words, every $i \in \mathbb{N}$ satisfies

$$x_i = x_{i+n} \tag{697}$$

(by the definition of a "period").

Next, we claim that every $i \in \mathbb{N}$ satisfies

$$s_i = s_{i+n}. \tag{698}$$

[*Proof of (698):* We shall prove (698) by induction on i:

Induction base: The definition of s_0 yields $s_0 = x_0 + x_1 + \cdots + x_{0-1} = (\text{empty sum}) = 0$. The definition of s_n yields $s_n = x_0 + x_1 + \cdots + x_{n-1} = 0$. Comparing these two equalities, we obtain $s_0 = s_n = s_{0+n}$ (since $n = 0 + n$). In other words, (698) holds for $i = 0$.

Induction step: Let $j \in \mathbb{N}$. Assume (as the induction hypothesis) that (698) holds for $i = j$. We must prove that (698) holds for $i = j + 1$. In other words, we must prove that $s_{j+1} = s_{(j+1)+n}$.

We have assumed that (698) holds for $i = j$. In other words, we have $s_j = s_{j+n}$.

The definition of s_j yields $s_j = x_0 + x_1 + \cdots + x_{j-1}$. The definition of s_{j+1} yields

$$\begin{aligned} s_{j+1} &= x_0 + x_1 + \cdots + x_{(j+1)-1} \\ &= x_0 + x_1 + \cdots + x_j \qquad (\text{since } (j+1) - 1 = j) \\ &= \underbrace{\left(x_0 + x_1 + \cdots + x_{j-1}\right)}_{=s_j = s_{j+n}} + \underbrace{x_j}_{\substack{=x_{j+n} \\ \text{(by (697),} \\ \text{applied to } i=j)}} = s_{j+n} + x_{j+n}. \end{aligned} \tag{699}$$

The definition of s_{j+n} yields $s_{j+n} = x_0 + x_1 + \cdots + x_{j+n-1}$. The definition of s_{j+n+1} yields

$$\begin{aligned} s_{j+n+1} &= x_0 + x_1 + \cdots + x_{(j+n+1)-1} \\ &= x_0 + x_1 + \cdots + x_{j+n} \qquad (\text{since } (j+n+1) - 1 = j+n) \\ &= \underbrace{\left(x_0 + x_1 + \cdots + x_{j+n-1}\right)}_{=s_{j+n}} + x_{j+n} = s_{j+n} + x_{j+n}. \end{aligned}$$

Comparing this with (699), we obtain $s_{j+1} = s_{j+n+1} = s_{(j+1)+n}$ (since $j+n+1 = (j+1)+n$). This completes the induction step. Thus, (698) is proven.]

Thus, we have shown that every $i \in \mathbb{N}$ satisfies $s_i = s_{i+n}$. In other words, n is a period of the sequence $(s_0, s_1, s_2, \ldots)$ (by the definition of a "period"). In other words, the sequence $(s_0, s_1, s_2, \ldots)$ is n-periodic (by the definition of "n-periodic").

Pick an element $u \in \{0, 1, \ldots, n-1\}$ for which the number s_u is minimum[397]. Thus,

$$s_u \leq s_v \qquad \text{for each } v \in \{0, 1, \ldots, n-1\}. \tag{700}$$

Hence, we can easily see that

$$s_u \leq s_m \qquad \text{for each } m \in \mathbb{N}. \tag{701}$$

[*Proof of (701):* Let $m \in \mathbb{N}$. Proposition 3.3.2 **(a)** (applied to m instead of u) yields $m\%n \in \{0, 1, \ldots, n-1\}$ and $m\%n \equiv m \bmod n$. Hence, (700) (applied to $v = m\%n$) yields $s_u \leq s_{m\%n}$. However, n is a period of the sequence $(s_0, s_1, s_2, \ldots)$, whereas $m\%n$ and m are two nonnegative integers satisfying $m\%n \equiv m \bmod n$. Therefore, Theorem 4.7.8 **(e)** (applied to $(s_0, s_1, s_2, \ldots)$, s_i, n, $m\%n$ and m instead of u, u_i, a, p and q) yields that $s_{m\%n} = s_m$. Hence, $s_u \leq s_{m\%n} = s_m$. This proves (701).]

The definition of s_u yields $s_u = x_0 + x_1 + \cdots + x_{u-1}$.

Now, let $m \geq u$ be an integer. Thus, $m \geq u \geq 0$, so that $m \in \mathbb{N}$. Hence, $m+1 \in \mathbb{N}$. Thus, (701) (applied to $m+1$ instead of m) shows that

$$\begin{aligned}
s_u &\leq s_{m+1} = x_0 + x_1 + \cdots + x_{(m+1)-1} \qquad \text{(by the definition of } s_{m+1}\text{)} \\
&= x_0 + x_1 + \cdots + x_m \qquad \text{(since } (m+1)-1 = m\text{)} \\
&= \underbrace{(x_0 + x_1 + \cdots + x_{u-1})}_{=s_u} + (x_u + x_{u+1} + \cdots + x_m) \\
&\qquad \left(\begin{array}{c}\text{here, we have split the sum after its } u\text{-th addend,} \\ \text{since } 0 \leq u \leq m+1\end{array}\right) \\
&= s_u + (x_u + x_{u+1} + \cdots + x_m).
\end{aligned}$$

Subtracting s_u from both sides of this inequality, we obtain $0 \leq x_u + x_{u+1} + \cdots + x_m$. In other words, $x_u + x_{u+1} + \cdots + x_m \geq 0$.

Forget that we fixed m. We thus have shown that every $m \geq u$ satisfies $x_u + x_{u+1} + \cdots + x_m \geq 0$. Hence, there exists some $k \in \{0, 1, \ldots, n-1\}$ such that every $m \geq k$ satisfies $x_k + x_{k+1} + \cdots + x_m \geq 0$ (namely, $k = u$). This solves Exercise 5.4.9. □

□

[397] This is, again, an application of Theorem 5.1.1: Indeed, the set $\{s_u \mid u \in \{0, 1, \ldots, n-1\}\}$ is nonempty (since n is positive) and finite; thus, it has a minimum (by Theorem 5.1.1). If we denote this minimum by g, then $g \in \{s_u \mid u \in \{0, 1, \ldots, n-1\}\}$ (since the minimum of a set always belongs to this set), and thus there exists some $u \in \{0, 1, \ldots, n-1\}$ such that $g = s_u$. We pick such a u.

A.6.10. Discussion of Exercise 5.4.10

Discussion of Exercise 5.4.10. Here is one possible solution:

Solution to Exercise 5.4.10. **(a)** Let (x,y) be a golden pair such that $(x,y) \neq (0,1)$. We must prove that $x-y \geq 0$.

Indeed, (x,y) is a golden pair. In other words, (x,y) is a pair of nonnegative integers such that $\left|x^2-xy-y^2\right| = 1$ (by the definition of a "golden pair").

Each $z \in \mathbb{R}$ satisfies $z \geq -|z|$. Applying this to $z = x^2-xy-y^2$, we obtain $x^2-xy-y^2 \geq -\underbrace{\left|x^2-xy-y^2\right|}_{=1} = -1$.

Now, our goal is to show that $x-y \geq 0$.

Assume the contrary. Thus, $x-y<0$, so that $x<y$. Hence, $x^2<y^2$ (since x and y are nonnegative) and therefore $x^2 \leq y^2-1$ (since x and y are integers). Moreover, from $x<y$, we obtain $y>x\geq 0$ (since x is nonnegative). If we had $x>0$, we would thus obtain $xy>0$ (since $x>0$ and $y>0$) and therefore $\underbrace{x^2}_{\leq y^2-1} - \underbrace{xy}_{>0} - y^2 < y^2-1-y^2 = -1$, which would contradict $x^2-xy-y^2 \geq -1$. Hence, we cannot have $x>0$. Thus, $x \leq 0$. Combining this with $x \geq 0$, we find $x=0$. Hence, $x^2-xy-y^2 = 0^2-0\cdot y-y^2 = -y^2$, so that $\left|x^2-xy-y^2\right| = \left|-y^2\right| = \left|y^2\right| = y^2$ (since $y^2 \geq 0$). Therefore, $y^2 = \left|x^2-xy-y^2\right| = 1$, so that $y=1$ (since y is nonnegative). Combining $x=0$ with $y=1$, we find $(x,y)=(0,1)$; but this contradicts $(x,y) \neq (0,1)$. This contradiction shows that our assumption was false. Hence, we conclude that $x-y \geq 0$. This solves Exercise 5.4.10 **(a)**.

(b) Let (x,y) be a golden pair such that $(x,y) \neq (0,1)$. We must prove that $(y, x-y)$ is a golden pair.

Indeed, (x,y) is a golden pair. In other words, (x,y) is a pair of nonnegative integers such that $\left|x^2-xy-y^2\right| = 1$ (by the definition of a "golden pair"). Moreover, $x-y \geq 0$ (by Exercise 5.4.10 **(a)**), so that $x-y$ is a nonnegative integer. Thus, $(y, x-y)$ is a pair of nonnegative integers. A straightforward computation shows that

$$y^2 - y(x-y) - (x-y)^2 = -\left(x^2-xy-y^2\right),$$

so that

$$\begin{aligned}\left|y^2-y(x-y)-(x-y)^2\right| &= \left|-\left(x^2-xy-y^2\right)\right| = \left|x^2-xy-y^2\right| \\ &\qquad \text{(since } |-z| = |z| \text{ for each } z \in \mathbb{R}\text{)} \\ &= 1.\end{aligned}$$

Hence, $(y, x-y)$ is a pair of nonnegative integers satisfying $\left|y^2-y(x-y)-(x-y)^2\right| = 1$. In other words, $(y, x-y)$ is a golden pair (by the definition of a "golden pair"). This solves Exercise 5.4.10 **(b)**.

(c) Let (x,y) be a golden pair such that $(x,y) \neq (1,0)$. We must prove that $y>0$.

Indeed, assume the contrary. Thus, $y \leq 0$.

However, (x, y) is a golden pair. In other words, (x, y) is a pair of nonnegative integers such that $\left|x^2 - xy - y^2\right| = 1$ (by the definition of a "golden pair"). Now, $y \geq 0$ (since y is nonnegative). Combining this with $y \leq 0$, we find $y = 0$.

Hence, $x^2 - xy - y^2 = x^2 - x \cdot 0 - 0^2 = x^2$, so that $\left|x^2 - xy - y^2\right| = \left|x^2\right| = x^2$ (since $x^2 \geq 0$). Therefore, $x^2 = \left|x^2 - xy - y^2\right| = 1$, so that $x = 1$ (since x is nonnegative). Combining $x = 1$ with $y = 0$, we find $(x, y) = (1, 0)$; but this contradicts $(x, y) \neq (1, 0)$. This contradiction shows that our assumption was false. Hence, we conclude that $y > 0$. This solves Exercise 5.4.10 **(c)**.

(d) Let $(f_0, f_1, f_2, \ldots)$ be the Fibonacci sequence (defined as in Definition 2.2.1). We claim the following:

> *Claim 1:* The golden pairs different from $(0, 1)$ are the pairs (f_{n+1}, f_n) for $n \in \mathbb{N}$ (that is, the pairs (f_1, f_0), (f_2, f_1), (f_3, f_2), $\ldots$).

We shall prove Claim 1 after some preparatory work. First, we prove the easy direction:

> *Observation 2:* Let $n \in \mathbb{N}$. Then, the pair (f_{n+1}, f_n) is a golden pair different from $(0, 1)$.

[*Proof of Observation 2:* Clearly, $n + 1$ is a positive integer (since $n \in \mathbb{N}$). Hence, we can apply the Cassini identity (i.e., Exercise 2.2.2) to $n + 1$ instead of n. We thus find $f_{(n+1)+1} f_{(n+1)-1} - f_{n+1}^2 = (-1)^{n+1}$. In view of $(n + 1) + 1 = n + 2$ and $(n + 1) - 1 = n$, this rewrites as $f_{n+2} f_n - f_{n+1}^2 = (-1)^{n+1}$. But the recursive definition of the Fibonacci sequence yields $f_m = f_{m-1} + f_{m-2}$ for each integer $m \geq 2$. Applying this to $m = n + 2$, we obtain $f_{n+2} = \underbrace{f_{(n+2)-1}}_{=f_{n+1}} + \underbrace{f_{(n+2)-2}}_{=f_n} = f_{n+1} + f_n$. Hence,

$$\underbrace{f_{n+2}}_{=f_{n+1}+f_n} f_n - f_{n+1}^2 = (f_{n+1} + f_n) f_n - f_{n+1}^2 = -\left(f_{n+1}^2 - f_{n+1} f_n - f_n^2\right).$$

Therefore,

$$\left|f_{n+2} f_n - f_{n+1}^2\right| = \left|-\left(f_{n+1}^2 - f_{n+1} f_n - f_n^2\right)\right| = \left|f_{n+1}^2 - f_{n+1} f_n - f_n^2\right|$$

(since $|-z| = |z|$ for each $z \in \mathbb{R}$). Comparing this with

$$\left|\underbrace{f_{n+2} f_n - f_{n+1}^2}_{=(-1)^{n+1}}\right| = \left|(-1)^{n+1}\right| = 1 \qquad \left(\text{since } (-1)^{n+1} \text{ is either } 1 \text{ or } -1\right),$$

we obtain $\left|f_{n+1}^2 - f_{n+1} f_n - f_n^2\right| = 1$.

It is easy to see that the Fibonacci numbers $f_1, f_2, f_3, \ldots$ are positive integers. In other words, f_m is a positive integer whenever m is a positive integer. Applying this

to $m = n+1$, we conclude that f_{n+1} is a positive integer (since $n+1$ is a positive integer). Hence, $f_{n+1} \neq 0$. Thus, $(f_{n+1}, f_n) \neq (0,1)$. Also, f_n and f_{n+1} are nonnegative integers (since all Fibonacci numbers $f_0, f_1, f_2, \ldots$ are nonnegative integers). Thus, (f_{n+1}, f_n) is a pair of nonnegative integers such that $\left|f_{n+1}^2 - f_{n+1}f_n - f_n^2\right| = 1$. In other words, (f_{n+1}, f_n) is a golden pair (by the definition of a "golden pair"). Since we know that $(f_{n+1}, f_n) \neq (0,1)$, we thus conclude that (f_{n+1}, f_n) is a golden pair different from $(0,1)$. This proves Observation 2.]

Now, let us say that a pair (x,y) of two nonnegative integers is *fibonacci* (this word should be understood as an adjective) if it has the form (f_{n+1}, f_n) for some $n \in \mathbb{N}$. Thus, Observation 2 can be restated as follows:

Observation 3: Every fibonacci pair is a golden pair different from $(0,1)$.

We shall now prove the converse statement:

Observation 4: If (x,y) is a golden pair different from $(0,1)$, then (x,y) is fibonacci.

[*Proof of Observation 4:* If (x,y) is a golden pair, then $x+y \in \mathbb{N}$ (since x and y are nonnegative integers). Thus, we can prove Observation 4 by strong induction on $x+y$:

Induction step: Let $m \in \mathbb{N}$. Assume (as the induction hypothesis) that Observation 4 holds for $x+y < m$. We must prove that Observation 4 holds for $x+y = m$.

We have assumed that Observation 4 holds for $x+y < m$. In other words, if (x,y) is a golden pair different from $(0,1)$ and satisfying $x+y < m$, then

$$(x,y) \text{ is fibonacci.} \tag{702}$$

Now, let (x,y) be a golden pair different from $(0,1)$ and satisfying $x+y = m$. We shall prove that (x,y) is fibonacci.

Indeed, if $(x,y) = (1,0)$, this is clear (because the pair $(1,0)$ is fibonacci[398]). Thus, for the rest of this proof, we WLOG assume that $(x,y) \neq (1,0)$. Hence, Exercise 5.4.10 **(c)** yields $y > 0$. Thus, $x+y > x$, so that $x < x+y = m$. Furthermore, we have $(x,y) \neq (0,1)$ (since (x,y) is different from $(0,1)$). Hence, Exercise 5.4.10 **(b)** yields that $(y, x-y)$ is a golden pair. We have $(y, x-y) \neq (0,1)$ [399] and $y + (x-y) = x < m$.

Now, we know that $(y, x-y)$ is a golden pair different from $(0,1)$ (since $(y, x-y) \neq (0,1)$) and satisfying $y + (x-y) < m$. Hence, we can apply (702) to $(y, x-y)$ instead of (x,y). As a consequence, we conclude that the pair $(y, x-y)$ is fibonacci.

[398] since $(1,0) = (f_{n+1}, f_n)$ for $n = 0$

[399] *Proof.* Assume the contrary. Thus, $(y, x-y) = (0,1)$. In other words, $y = 0$ and $x - y = 1$. Hence, $x = \underbrace{y}_{=0} + \underbrace{x-y}_{=1} = 0 + 1 = 1$. Combined with $y = 0$, this yields $(x,y) = (1,0)$; but this contradicts $(x,y) \neq (1,0)$. This contradiction shows that our assumption was false. Qed.

In other words, $(y, x-y)$ has the form (f_{n+1}, f_n) for some $n \in \mathbb{N}$ (by the definition of "fibonacci"). Consider this n, and denote it by u. Thus, $u \in \mathbb{N}$ and $(y, x-y) = (f_{u+1}, f_u)$.

From $(y, x-y) = (f_{u+1}, f_u)$, we obtain $y = f_{u+1}$ and $x - y = f_u$. Now, $x = \underbrace{y}_{=f_{u+1}} + \underbrace{x-y}_{=f_u} = f_{u+1} + f_u$.

But the recursive definition of the Fibonacci sequence yields $f_p = f_{p-1} + f_{p-2}$ for each integer $p \geq 2$. Applying this to $p = u+2$, we obtain $f_{u+2} = \underbrace{f_{(u+2)-1}}_{=f_{u+1}} + \underbrace{f_{(u+2)-2}}_{=f_u} = f_{u+1} + f_u$. Hence, $x = f_{u+1} + f_u = f_{u+2} = f_{(u+1)+1}$. Combining this with $y = f_{u+1}$, we obtain $(x, y) = \left(f_{(u+1)+1}, f_{u+1}\right)$. Hence, the pair (x, y) has the form (f_{n+1}, f_n) for some $n \in \mathbb{N}$ (namely, for $n = u+1$). In other words, the pair (x, y) is fibonacci (by the definition of "fibonacci").

Forget that we fixed (x, y). We thus have shown that if (x, y) is a golden pair different from $(0, 1)$ and satisfying $x + y = m$, then (x, y) is fibonacci. In other words, Observation 4 holds for $x + y = m$. This completes the induction step. Thus, Observation 4 is proved.]

Observation 4 can be restated as follows:

Observation 5: Every golden pair different from $(0, 1)$ is a fibonacci pair.

Combining this with Observation 3, we conclude that the golden pairs different from $(0, 1)$ are precisely the fibonacci pairs. In other words, the golden pairs different from $(0, 1)$ are precisely the pairs of the form (f_{n+1}, f_n) for some $n \in \mathbb{N}$ (because this is how the fibonacci pairs are defined). This proves Claim 1. Thus, Exercise 5.4.10 **(d)** is solved. □

[*Remark:* Exercise 5.4.10 **(d)** is a result of Lucas (1876). A proof similar to ours is given in [Jones75, Lemmas 1, 2, 3].

Note that our proof is an example of an argument by infinite descent, even though we have formulated it as a strong induction argument. To see this, it suffices to rewrite Observation 4 above as "there exists no non-fibonacci golden pair different from $(0, 1)$", and reframe our proof of Observation 4 as transforming an (ostensible) non-fibonacci golden pair (x, y) different from $(0, 1)$ into a smaller non-fibonacci golden pair $(y, x-y)$ which is also different from $(0, 1)$. ("Smaller" here means "smaller sum of entries".) We chose to instead organize our proof as a strong induction argument, in order to avoid the detour through the contrapositive; but it is probably easiest to find this proof by attempting an infinite descent.

We have thus solved the equation $\left|x^2 - xy - y^2\right| = 1$ in nonnegative integers x and y. This is an example of a *quadratic Diophantine equation*, closely connected to the *Pell equation* $\left|x^2 - 5y^2\right| = 1$ (see [AndAnd14] for the connection). Similar arguments can be used for solving other equations like this; it is not a coincidence that the solutions in our case were given by the Fibonacci numbers. Usually, solutions (when they exist) are given by (u, v)-recurrent sequences for some values of u and

v. For example, the solutions (x, y) of the equation $\left|x^2 - 2y^2\right| = 1$ in nonnegative integers x and y are precisely the pairs of the form (a_n, b_n) for $n \in \mathbb{N}$, where $(a_0, a_1, a_2, \ldots)$ and $(b_0, b_1, b_2, \ldots)$ are the two $(2, 1)$-recurrent sequences with starting values $a_0 = 1$, $a_1 = 1$, $b_0 = 0$ and $b_1 = 1$. (These two sequences are A001333 and A000129 in the OEIS.) See [Emerso69] for the proof.

Finally, note that the golden pair $(0, 1)$ is not really an exception from the fibonacci formula (f_{n+1}, f_n), even though we have treated it as such in our above solution. Indeed, if we define f_{-1} as in Exercise 4.4.4, then $(0, 1) = (f_{n+1}, f_n)$ for $n = -1$, and thus the pair $(0, 1)$ becomes fibonacci if we extend the definition of fibonacci pairs to allow for $n = -1$.] □

A.7. Homework set #6 discussion

The following are discussions of the problems on homework set #6 (Section 6.3).

A.7.1. Discussion of Exercise 6.3.1

Discussion of Exercise 6.3.1. Exercise 6.3.1 is [Engel98, Exercise 3.13] (and also easily seen to be equivalent to [Engel98, Chapter 3, Example E11]). Its solution is a copybook example of a use of the Extremal Principle:

Solution to Exercise 6.3.1. First, we WLOG assume that no player plays a match against himself. (Indeed, this is arguably clear from common sense or perhaps implicit in the statement of the problem. But even if we don't consider this to be clear, we can make this true by forgetting about all self-matches[400]. That is, throughout the argument below, we pretend that self-matches do not happen (and, in particular, we ignore their results). It is clear that if the problem has been solved without taking the self-matches into account, then it has also been solved with the self-matches, because a player who has (directly or indirectly) owned all other players will not lose this property if we add some self-matches. Thus, we can WLOG assume that self-matches don't exist – i.e., that no player plays a match against himself.)

Define the *score* of a player d to be the number of all players that d has directly owned. Consider a player a with maximum score. (Such a player clearly exists, since the set $\{\text{scores of players}\}$ is a finite nonempty set of integers and thus has a maximum.) We shall show that a has (directly or indirectly) owned all other players.

Indeed, let b be a player different from a. We shall show that a has (directly or indirectly) owned b.

Indeed, assume the contrary. Thus, a has neither directly owned b nor indirectly owned b.

Hence, in particular, a has not directly owned b. In other words, a has not won a match against b (since "directly owning" b means winning a match against b).

[400] i.e., matches that a player plays against himself

However, the players a and b are distinct (since b is different from a), and thus there must be a match between a and b (since each pair of distinct players play exactly one match against one another). This match cannot have been won by a (since a has not won a match against b), and thus must have been won by b (since no match ends with a draw). Hence, b has won a match against a. In other words, b has directly owned a (since "directly owning" a means winning a match against a).

Let Y be the set of all players that b has directly owned. Thus,

$$\begin{aligned} |Y| &= (\text{the number of all players that } b \text{ has directly owned}) \\ &= (\text{the score of } b) \end{aligned} \tag{703}$$

(since the score of b is defined as the number of all players that b has directly owned). Moreover, we recall that b has directly owned a; in other words, $a \in Y$ (since Y is the set of all players that b has directly owned). Hence, $|Y \setminus \{a\}| = |Y| - 1$.

Let X be the set of all players that a has directly owned. Thus,

$$\begin{aligned} |X| &= (\text{the number of all players that } a \text{ has directly owned}) \\ &= (\text{the score of } a) \end{aligned}$$

(since the score of a is defined as the number of all players that a has directly owned). Thus,

$$\begin{aligned} |X| &= (\text{the score of } a) \geq (\text{the score of } b) \qquad (\text{since } a \text{ has maximum score}) \\ &= |Y| \qquad (\text{by (703)}) \\ &> |Y| - 1 = |Y \setminus \{a\}| \qquad (\text{since } |Y \setminus \{a\}| = |Y| - 1). \end{aligned}$$

Thus, we have proved the inequality $|X| > |Y \setminus \{a\}|$. Next, we will show that $X \subseteq Y \setminus \{a\}$, which will clearly contradict this inequality (and thus give us the contradiction that we want).

Indeed, let $x \in X$. We shall show that $x \in Y \setminus \{a\}$.

We have $x \in X$. In other words, x is a player that a has directly owned (since X is the set of all players that a has directly owned). In other words, x is a player that a has won a match against (since "directly owning" a player means winning a match against this player). Thus, a has won a match against x. Therefore, $x \neq a$ (since no player plays a match against himself).

Now, we shall show that $x \in Y$. Indeed, assume the contrary.[401] Thus, $x \notin Y$. In other words, b has not directly owned x (since Y is the set of all players that b has directly owned). In other words, b has not won a match against x (since "directly owning" x means winning a match against x).

Recall that a has not directly owned b. In other words, $b \notin X$ (since X is the set of all players that a has directly owned). If we had $x = b$, then we would thus

[401] You are reading right: We are doing a proof by contradiction inside a proof by contradiction. This is perhaps not particularly elegant, but perfectly valid.

obtain $x = b \notin X$, which would contradict $x \in X$. Hence, we cannot have $x = b$. Thus, we have $x \neq b$. Therefore, the players x and b are distinct, so that there must be a match between x and b (since each pair of distinct players play exactly one match against one another). This match cannot have been won by b (since b has not won a match against x), and thus must have been won by x (since no match ends with a draw). Hence, x has won a match against b. Recall also that a has won a match against x. Hence, there exists a player c such that a has won a match against c and c has won a match against b (namely, $c = x$). In other words, a has indirectly owned b (by the definition of "indirectly owned"). This contradicts the fact that a has neither directly owned b nor indirectly owned b. This contradiction shows that our latest assumption was false. Thus, $x \in Y$ is proved.

Combining $x \in Y$ with $x \neq a$, we find $x \in Y \setminus \{a\}$.

Forget that we fixed x. We thus have shown that $x \in Y \setminus \{a\}$ for each $x \in X$. In other words, $X \subseteq Y \setminus \{a\}$. Hence, $|X| \leq |Y \setminus \{a\}|$. But this contradicts $|X| > |Y \setminus \{a\}|$. This contradiction shows that our assumption was false. Hence, we have shown that a has (directly or indirectly) owned b.

Forget that we fixed b. We thus have shown that if b is any player different from a, then a has (directly or indirectly) owned b. In other words, a has (directly or indirectly) owned all other players. Thus, there exists a player who has (directly or indirectly) owned all other players (namely, a). This solves Exercise 6.3.1. □

□

A.7.2. Discussion of Exercise 6.3.2

Discussion of Exercise 6.3.2. Exercise 6.3.2 is a generalization of Exercise 5.3.2 (in that its assumptions are weaker, but its conclusion is the same). We shall solve it by modifying our solution to Exercise 5.3.2 as follows:

Solution to Exercise 6.3.2 (sketched). Some notations first.

A *solution* will mean an $(2n+1)$-tuple $(a_1, a_2, \ldots, a_{2n+1})$ of integers that has the weaker splitting property.

A solution $(a_1, a_2, \ldots, a_{2n+1})$ will be called *flat* if all $2n+1$ numbers $a_1, a_2, \ldots, a_{2n+1}$ are equal. Thus, our goal is to show that every solution is flat. In other words, our goal is to show that there is no non-flat solution.

A solution $(a_1, a_2, \ldots, a_{2n+1})$ will be called *nonnegative* if all $2n+1$ numbers $a_1, a_2, \ldots, a_{2n+1}$ are nonnegative.

First, we observe the following:

> *Observation 1′:* If there exists a non-flat solution, then there exists a nonnegative non-flat solution.

[*Proof of Observation 1′:* This can be proved in the same way as Observation 1 was proved in our solution to Exercise 5.3.2.]

Thanks to Observation 1′, we don't need to bother with negative integers if we don't want to. This will come useful later.

Next, let us study the parity of the integers in a solution:

> *Observation 2′:* If $(a_1, a_2, \ldots, a_{2n+1})$ is a solution, then the $2n+1$ integers $a_1, a_2, \ldots, a_{2n+1}$ all have the same parity (i.e., are either all even or all odd).

[*Proof of Observation 2′:* This is similar to the proof of Observation 2 in our solution to Exercise 5.3.2, but somewhat more complicated because the weaker splitting property is (as its name suggests) weaker than the original splitting property. Here are the details:

Let $(a_1, a_2, \ldots, a_{2n+1})$ be a solution. Let $t = a_1 + a_2 + \cdots + a_{2n+1}$ be the sum of all its entries. We WLOG assume that $2n+1 > 1$ (since otherwise, the claim of Observation 2′ is obvious). Hence, $2n > 0$, so that $n > 0$ and thus $n \geq 1$ (since $n \in \mathbb{N}$). Therefore, $1 \in \{1, 2, \ldots, 2n\}$.

Thus, the weaker splitting property shows that if the integer a_1 is removed (from our $2n+1$ numbers $a_1, a_2, \ldots, a_{2n+1}$), then the remaining $2n$ numbers can be split into two equinumerous heaps with equal sum. In other words, the $2n$ numbers $a_2, a_3, \ldots, a_{2n+1}$ can be split into two equinumerous heaps with equal sum. Let s be the sum of either heap. Then, s is an integer (since $a_2, a_3, \ldots, a_{2n+1}$ are integers). But the sum $a_2 + a_3 + \cdots + a_{2n+1}$ of all the $2n$ numbers $a_2, a_3, \ldots, a_{2n+1}$ must be $s + s$ (since these $2n$ numbers can be split into two heaps, each of which has sum s). Thus,

$$a_2 + a_3 + \cdots + a_{2n+1} = s + s = 2s \equiv 0 \bmod 2 \qquad (\text{since } s \text{ is an integer}).$$

Hence,

$$t = a_1 + a_2 + \cdots + a_{2n+1} = a_1 + \underbrace{(a_2 + a_3 + \cdots + a_{2n+1})}_{\equiv 0 \bmod 2} \equiv a_1 \bmod 2,$$

so that $a_1 \equiv t \bmod 2$.

Thus, by removing a_1 and applying the weaker splitting property, we have obtained $a_1 \equiv t \bmod 2$. But we can apply the same argument to any of the $2n$ integers $a_1, a_2, \ldots, a_{2n}$ in place of a_1. Thus, we find that

$$a_i \equiv t \bmod 2 \tag{704}$$

for each $i \in \{1, 2, \ldots, 2n\}$. Using (40), we now have

$$\sum_{i=1}^{2n} \underbrace{a_i}_{\substack{\equiv t \bmod 2 \\ (\text{by } (704))}} \equiv \sum_{i=1}^{2n} t = 2nt \equiv 0 \bmod 2$$

(since $2 \mid 2nt$). Now,

$$t = a_1 + a_2 + \cdots + a_{2n+1} = \underbrace{(a_1 + a_2 + \cdots + a_{2n})}_{= \sum_{i=1}^{2n} a_i \equiv 0 \bmod 2} + a_{2n+1} \equiv a_{2n+1} \bmod 2.$$

Hence, $a_{2n+1} \equiv t \bmod 2$. Thus, the congruence (704) holds not only for each $i \in \{1, 2, \ldots, 2n\}$, but also for $i = 2n+1$. Hence, this congruence (704) holds for each $i \in \{1, 2, \ldots, 2n+1\}$. As a consequence,

- if t is even, then all of $a_1, a_2, \ldots, a_{2n+1}$ are even;
- if t is odd, then all of $a_1, a_2, \ldots, a_{2n+1}$ are odd.

Thus, $a_1, a_2, \ldots, a_{2n+1}$ all have the same parity. This proves Observation 2′.]

Now, Observation 2′ helps us transform non-flat solutions into smaller non-flat solutions with an appropriate meaning of "smaller". To be more precise, we consider nonnegative solutions. If $(a_1, a_2, \ldots, a_{2n+1})$ is a nonnegative solution, then the *weight* of this solution is defined to be the nonnegative integer $a_1 + a_2 + \cdots + a_{2n+1}$. Now we claim:

> *Observation 3′:* If $(a_1, a_2, \ldots, a_{2n+1})$ is a nonnegative non-flat solution, then there exists a nonnegative non-flat solution with smaller weight than $(a_1, a_2, \ldots, a_{2n+1})$.

[*Proof of Observation 3′:* This can be proved in the same way as Observation 3 was proved in our solution to Exercise 5.3.2. (Of course, we now need to use Observation 2′ instead of Observation 2.)]

Now, all we need is to reap our rewards. By the Principle of Infinite Descent, Observation 3′ entails that there exists no nonnegative non-flat solution. Hence, by Observation 1′, we conclude that there exists no non-flat solution either. In other words, any solution is flat. Exercise 6.3.2 is solved. □

□

A.7.3. Discussion of Exercise 6.3.3

Discussion of Exercise 6.3.3. Exercise 6.3.3 is an easy consequence of Theorem 6.2.5:

Solution to Exercise 6.3.3. Theorem 6.2.5 **(a)** yields that there exist two integers i and j with $0 \leq i < j \leq n$ and $f^i(x) = f^j(x)$. Consider these i and j. Set $a = j - i$. We have $a = j - i > 0$ (since $i < j$); thus, a is a positive integer.

Theorem 6.2.5 **(b)** yields that the sequence $\left(f^i(x), f^{i+1}(x), f^{i+2}(x), \ldots\right)$ is $(j-i)$-periodic. In other words, the sequence $\left(f^i(x), f^{i+1}(x), f^{i+2}(x), \ldots\right)$ is a-periodic (since $a = j - i$). In other words, a is a period of the sequence $\left(f^i(x), f^{i+1}(x), f^{i+2}(x), \ldots\right)$ (by the definition of "a-periodic"). Thus, if u and v are two nonnegative integers satisfying $u \equiv v \bmod a$, then

$$f^{i+u}(x) = f^{i+v}(x) \tag{705}$$

(by Theorem 4.7.8 **(e)**, applied to $\left(f^i(x), f^{i+1}(x), f^{i+2}(x), \ldots\right)$, $f^{i+m}(x)$, u and v instead of u, u_m, p and q).

Define an integer s by

$$s = 1 + (p - q - i - 1)\,\%a.$$

Proposition 3.3.2 **(a)** (applied to a and $p - q - i - 1$ instead of n and u) yields that $(p - q - i - 1)\,\%a \in \{0, 1, \ldots, a - 1\}$ and $(p - q - i - 1)\,\%a \equiv p - q - i - 1 \bmod a$. Now,

$$s = 1 + (p - q - i - 1)\,\%a \in \{1, 2, \ldots, a\}$$

(since $(p - q - i - 1)\,\%a \in \{0, 1, \ldots, a - 1\}$), so that $s \geq 1$ and $s \leq a$. Hence, $i + \underbrace{s}_{\leq a = j - i} \leq i + (j - i) = j \leq n$. Also, $\underbrace{i}_{\geq 0} + s \geq s \geq 1$. Combining $i + s \geq 1$ with $i + s \leq n$, we find $i + s \in \{1, 2, \ldots, n\}$ (since $i + s$ is an integer). Moreover,

$$s = 1 + \underbrace{(p - q - i - 1)\,\%a}_{\equiv p - q - i - 1 \bmod a} \equiv 1 + (p - q - i - 1) = p - q - i \bmod a.$$

Hence,

$$(i + 2s + q) - (s + p) = \underbrace{s}_{\equiv p - q - i \bmod a} + i + q - p \equiv p - q - i + i + q - p \equiv 0 \bmod a.$$

In other words, $a \mid (i + 2s + q) - (s + p)$. In other words,

$$i + 2s + q \equiv s + p \bmod a.$$

Thus, (705) (applied to $u = i + 2s + q$ and $v = s + p$) yields $f^{i + (i + 2s + q)}(x) = f^{i + (s + p)}(x)$. In view of $i + (i + 2s + q) = 2(i + s) + q$ and $i + (s + p) = (i + s) + p$, this rewrites as $f^{2(i + s) + q}(x) = f^{(i + s) + p}(x)$. In other words, $f^{(i + s) + p}(x) = f^{2(i + s) + q}(x)$. Since we know that $i + s \in \{1, 2, \ldots, n\}$, we thus conclude that there exists some $k \in \{1, 2, \ldots, n\}$ such that $f^{k + p}(x) = f^{2k + q}(x)$ (namely, $k = i + s$). This solves Exercise 6.3.3. $\square$

The above solution may appear unmotivated in places (in particular, how did we think of defining s as $1 + (p - q - i - 1)\,\%a$?); however, it can be found by fairly straightforward reasoning. Indeed, once i, j and a have been defined, it is clear that the best bet for a $k \in \{1, 2, \ldots, n\}$ satisfying $f^{k + p}(x) = f^{2k + q}(x)$ is to have $k \geq 1$ and $k \geq i$ and $k + p \equiv 2k + q \bmod a$ (since Theorem 6.2.5 **(b)** will then yield $f^{k + p}(x) = f^{2k + q}(x)$). The easiest way to satisfy $k \geq 1$ and $k \geq i$ is by taking $k = i + s$ for some positive integer s. Rewriting the congruence $k + p \equiv 2k + q \bmod a$ in terms of this latter s, we obtain $i + s + p \equiv 2(i + s) + q \bmod a$, which is equivalent to $s \equiv p - q - i \bmod a$. This suggests setting $s = (p - q - i)\,\%a$, but this would sometimes result in $s = 0$, which we don't want (as we want s to be positive). Thus, we instead set $s = 1 + (p - q - i - 1)\,\%a$, in order for s to be positive. It remains to verify that the resulting $k = i + s$ does indeed belong to $\{1, 2, \ldots, n\}$. Thus, the above solution is found.

Let us notice that Theorem 6.2.10 is a particular case of Exercise 6.3.3:

Proof of Theorem 6.2.10. Exercise 6.3.3 (applied to $p = 0$ and $q = 0$) yields that there exists some $k \in \{1, 2, \ldots, n\}$ such that $f^{k+0}(x) = f^{2k+0}(x)$. In other words, there exists some $k \in \{1, 2, \ldots, n\}$ such that $f^k(x) = f^{2k}(x)$ (since $k + 0 = k$ and $2k + 0 = 2k$). This proves Theorem 6.2.10. □

□

A.7.4. Discussion of Exercise 6.3.4

Discussion of Exercise 6.3.4. The solution to Exercise 6.3.4 is similar to our proof of Proposition 6.2.12 **(c)**, except that we need to use $|f^n(X)| \geq 1$ instead of $\left|f^{n+1}(X)\right| \geq 0$. Here are the details:

Solution to Exercise 6.3.4. Note that $f^0 = \mathrm{id}_X$, so that $f^0(X) = \mathrm{id}_X(X) = X$. Hence, $\left|f^0(X)\right| = |X| = n$.

Also, the set $f^n(X)$ is nonempty[402]. Hence, $|f^n(X)| \geq 1$.

We shall first show that there exists some $i \in \{0, 1, \ldots, n-1\}$ such that $f^i(X) = f^{i+1}(X)$.

Indeed, assume the contrary. Thus, each $i \in \{0, 1, \ldots, n-1\}$ satisfies

$$f^i(X) \neq f^{i+1}(X). \tag{706}$$

Now, let $i \in \{0, 1, \ldots, n-1\}$. Then, Proposition 6.2.12 **(a)** yields $f^0(X) \supseteq f^1(X) \supseteq f^2(X) \supseteq \cdots$. Thus, $f^i(X) \supseteq f^{i+1}(X)$. In other words, $f^{i+1}(X)$ is a subset of $f^i(X)$. Moreover, this subset is **proper**, because (706) shows that $f^i(X) \neq f^{i+1}(X)$.

It is a well-known fact that if U is a proper subset of a finite set V, then $|U| < |V|$. Applying this to $U = f^{i+1}(X)$ and $V = f^i(X)$, we obtain $\left|f^{i+1}(X)\right| < \left|f^i(X)\right|$ (since $f^{i+1}(X)$ is a proper subset of the finite set $f^i(X)$). This entails $\left|f^{i+1}(X)\right| \leq \left|f^i(X)\right| - 1$ (since $\left|f^{i+1}(X)\right|$ and $\left|f^i(X)\right|$ are integers). In other words, $\left|f^{i+1}(X)\right| - \left|f^i(X)\right| \leq -1$.

Now, forget that we fixed i. We thus have proved the inequality $\left|f^{i+1}(X)\right| - \left|f^i(X)\right| \leq -1$ for each $i \in \{0, 1, \ldots, n-1\}$. Summing these inequalities over all $i \in \{0, 1, \ldots, n-1\}$, we obtain

$$\sum_{i=0}^{n} \left(\left|f^{i+1}(X)\right| - \left|f^i(X)\right|\right) \leq \sum_{i=0}^{n-1} (-1) = n \cdot (-1) = -n.$$

[402] *Proof.* The set X is nonempty. Hence, there exists some $x \in X$. Consider this x. Thus, $f^n(x) \in f^n(X)$ (since $x \in X$). Therefore, the set $f^n(X)$ has an element (namely, $f^n(x)$), and thus is nonempty. Qed.

But this contradicts

$$\sum_{i=0}^{n-1}\left(\left|f^{i+1}(X)\right|-\left|f^{i}(X)\right|\right)=\left|\underbrace{f^{(n-1)+1}}_{=f^n}(X)\right|-\underbrace{\left|f^{0}(X)\right|}_{=n}$$

$$\left(\begin{array}{c}\text{by Corollary 4.1.17,}\\ \text{applied to } u=0 \text{ and } v=n-1 \text{ and } a_i=\left|f^{i}(X)\right|\end{array}\right)$$

$$=\underbrace{\left|f^{n}(X)\right|}_{\geq 1}-n\geq 1-n>-n.$$

This contradiction shows that our assumption was false.

Hence, we have shown that there exists some $i\in\{0,1,\ldots,n-1\}$ such that $f^{i}(X)=f^{i+1}(X)$. Consider this i.

Now, $i\leq n-1$ (since $i\in\{0,1,\ldots,n-1\}$), so that $n-1\geq i$. Thus, Proposition 6.2.12 **(b)** (applied to $k=n-1$) yields $f^{i}(X)=f^{n-1}(X)$. However, we also have $n\geq n-1\geq i$. Thus, Proposition 6.2.12 **(b)** (applied to $k=n$) yields $f^{i}(X)=f^{n}(X)$. Hence, $f^{n}(X)=f^{i}(X)=f^{n-1}(X)$. This solves Exercise 6.3.4. □

□

A.7.5. Discussion of Exercise 6.3.5

Discussion of Exercise 6.3.5. Here is the simplest way to solve Exercise 6.3.5:

Solution to Exercise 6.3.5. **(a)** Let $x\in X$ and $k\in\mathbb{N}$ be such that $f^{k}(x)=f^{2k}(x)$. We must prove that

$$f^{ik}(x)=f^{k}(x) \tag{707}$$

for every positive integer i.

[*Proof of (707):* We shall prove (707) by induction on i:

Induction base: We have $f^{1k}(x)=f^{k}(x)$ (since $1k=k$). In other words, (707) holds for $i=1$.

Induction step: Let j be a positive integer. Assume (as the induction hypothesis) that (707) holds for $i=j$. We must prove that (707) holds for $i=j+1$. In other words, we must prove that $f^{(j+1)k}(x)=f^{k}(x)$.

We have assumed that (707) holds for $i=j$. In other words, we have $f^{jk}(x)=f^{k}(x)$. Now, $(j+1)k=k+jk$, so that

$$f^{(j+1)k}=f^{k+jk}=f^{k}\circ f^{jk}$$

(since $f^{p+q}=f^{p}\circ f^{q}$ for any $p,q\in\mathbb{N}$). Therefore,

$$f^{(j+1)k}(x)=\left(f^{k}\circ f^{jk}\right)(x)=f^{k}\left(\underbrace{f^{jk}(x)}_{=f^{k}(x)}\right)=f^{k}\left(f^{k}(x)\right). \tag{708}$$

On the other hand, $2k = k + k$, so that $f^{2k} = f^{k+k} = f^k \circ f^k$ (since $f^{p+q} = f^p \circ f^q$ for any $p, q \in \mathbb{N}$). Hence,

$$f^{2k}(x) = \left(f^k \circ f^k\right)(x) = f^k\left(f^k(x)\right).$$

Comparing this with (708), we obtain $f^{(j+1)k}(x) = f^{2k}(x) = f^k(x)$ (since $f^k(x) = f^{2k}(x)$). This completes the induction step. Thus, (707) is proved.]

Exercise 6.3.5 **(a)** is thus solved.

(b) Let $x \in X$. We shall show that $f^{n!}(x) = f^{2n!}(x)$.

Indeed, Theorem 6.2.10 yields that there exists some $k \in \{1, 2, \ldots, n\}$ such that $f^k(x) = f^{2k}(x)$. Consider this k.

We have

$$n! = 1 \cdot 2 \cdot \cdots \cdot n = \prod_{i \in \{1,2,\ldots,n\}} i = k \cdot \prod_{\substack{i \in \{1,2,\ldots,n\}; \\ i \neq k}} i \tag{709}$$

(here, we have split off the factor k from the product, because $k \in \{1, 2, \ldots, n\}$). Set $s = \prod_{\substack{i \in \{1,2,\ldots,n\}; \\ i \neq k}} i$. Then, s is a positive integer (since s is defined as a product of positive integers). Hence, $2s$ is a positive integer as well.

The equality (709) rewrites as $n! = k \cdot s$ (since $s = \prod_{\substack{i \in \{1,2,\ldots,n\}; \\ i \neq k}} i$). Thus, $n! = k \cdot s = sk$, so that

$$f^{n!}(x) = f^{sk}(x) = f^k(x) \tag{710}$$

(by Exercise 6.3.5 **(a)**, applied to $i = s$). Also, $2\underbrace{n!}_{=sk} = 2sk$, so that

$$f^{2n!}(x) = f^{2sk}(x) = f^k(x)$$

(by Exercise 6.3.5 **(a)**, applied to $i = 2s$). Comparing this with (710), we obtain $f^{n!}(x) = f^{2n!}(x)$.

Now, forget that we fixed x. We thus have proved that $f^{n!}(x) = f^{2n!}(x)$ for each $x \in X$. In other words, $f^{n!} = f^{2n!}$. This solves Exercise 6.3.5 **(b)**.

(c) Assume that f is a permutation of X.

The map f is a permutation, thus bijective (by the definition of a permutation). Hence, its power $f^{n!} = \underbrace{f \circ f \circ \cdots \circ f}_{n! \text{ times}}$ is bijective as well (since a composition of bijective maps is always bijective). Thus, in particular, this map $f^{n!}$ is injective.

Now, let $x \in X$. Exercise 6.3.5 **(b)** yields $f^{n!} = f^{2n!}$. But $2n! = n! + n!$ and thus $f^{2n!} = f^{n!+n!} = f^{n!} \circ f^{n!}$. Hence, $f^{n!} = f^{2n!} = f^{n!} \circ f^{n!}$, so that $f^{n!}(x) = \left(f^{n!} \circ f^{n!}\right)(x) = f^{n!}\left(f^{n!}(x)\right)$. Therefore, $x = f^{n!}(x)$ (since the map $f^{n!}$ is injective). Consequently, $f^{n!}(x) = x = \mathrm{id}_X(x)$.

Forget that we fixed x. We thus have shown that $f^{n!}(x) = \mathrm{id}_X(x)$ for each $x \in X$. In other words, $f^{n!} = \mathrm{id}_X$. This solves Exercise 6.3.5 **(c)**. $\square$

[*Remark:* Our above solution to Exercise 6.3.5 **(c)** is not how this exercise is usually solved. A more common approach to Exercise 6.3.5 **(c)** uses the notion of a *group* (in the sense of group theory), and the following facts:

- The group of all permutations of a given n-element set X has size $n!$.
- If G is a finite group with identity element e, then $g^{|G|} = e$ for each $g \in G$.

The first of these two facts is a basic fact in enumerative combinatorics (see, e.g., our Theorem 7.4.1, or [Loehr11, Theorem 1.28] or [Grinbe15, Corollary 7.82]); the second is a known result in elementary group theory (see, e.g., [Loehr11, Theorem 9.119] or [Elman20, Corollary 10.12] or [Goodma15, Corollary 2.5.9] or [Steinb06, Corollary 3.2.10]). Combining these two facts, we conclude that $f^{n!} = \mathrm{id}_X$ for each permutation f of X (since the permutations of X form a finite group with size $n!$ and identity element id_X). This solves Exercise 6.3.5 **(c)** again.] □

A.7.6. Discussion of Exercise 6.3.6

Discussion of Exercise 6.3.6. This would become really straightforward if we had used residue classes instead of remainders (i.e., the ring $\mathbb{Z}/m$ instead of the set $\{0, 1, \ldots, m-1\}$).

Solution to Exercise 6.3.6. We shall use all the notations that we have introduced in our above solution to Exercise 6.2.3. We must prove Claims 1, 2 and 3 from this solution. Let us do this now:

[*Proof of Claim 1:* We have $b \perp m$ (by assumption). In other words, $\gcd(b, m) = 1$ (by the definition of "coprime"). But Theorem 3.4.5 (applied to b and m instead of a and b) yields that there exist integers $x \in \mathbb{Z}$ and $y \in \mathbb{Z}$ such that

$$\gcd(b, m) = xb + ym.$$

Consider these x and y. Comparing the equalities $\gcd(b, m) = xb + ym$ and $\gcd(b, m) = 1$, we obtain

$$1 = xb + y \underbrace{m}_{\equiv 0 \bmod m} \equiv xb + y \cdot 0 = xb = bx \bmod m.$$

In other words,

$$bx \equiv 1 \bmod m. \tag{711}$$

Now, we define a map $g : M \times M \to M \times M$ by

$$g((p, q)) = \left(\overrightarrow{x(q - ap)}, p\right) \qquad \text{for each } (p, q) \in M \times M.$$

This is well-defined, since each $(p, q) \in M \times M$ satisfies $\left(\overrightarrow{x(q - ap)}, p\right) \in M \times M$ (because $\overrightarrow{x(q - ap)} = (x(q - ap)) \,\%m \in \{0, 1, \ldots, m-1\} = M$ and $p \in M$).

We shall now show that $f \circ g = \mathrm{id}_{M \times M}$.

Indeed, let $\alpha \in M \times M$ be arbitrary. Thus, $\alpha = (p, q)$ for some $p \in M$ and $q \in M$. Consider these p and q. From $\alpha = (p, q)$, we obtain $g(\alpha) = g((p, q)) = \left(\overrightarrow{x(q - ap)}, p\right)$ (by the definition of g). Set $w = \overrightarrow{x(q - ap)}$. Thus,

$$w = \overrightarrow{x(q - ap)} \equiv x(q - ap) \bmod m \tag{712}$$

(by (270), applied to $z = x(q - ap)$). Also,

$$g(\alpha) = \left(\overrightarrow{x(q - ap)}, p\right) = (w, p)$$

(since $\overrightarrow{x(q - ap)} = w$). Now,

$$(f \circ g)(\alpha) = f\left(\underbrace{g(\alpha)}_{=(w,p)}\right) = f((w, p)) = \left(p, \overrightarrow{ap + bw}\right) \tag{713}$$

(by the definition of f). We will now show that $\overrightarrow{ap + bw} = q$.

Indeed, (270) (applied to $z = ap + bw$) yields

$$\begin{aligned}\overrightarrow{ap + bw} &\equiv ap + b \underbrace{w}_{\substack{\equiv x(q-ap) \bmod m \\ \text{(by (712))}}} \equiv ap + \underbrace{bx}_{\substack{\equiv 1 \bmod m \\ \text{(by (711))}}} (q - ap) \\ &\equiv ap + (q - ap) = q \bmod m.\end{aligned}$$

Moreover, $\overrightarrow{ap + bw} \in \{0, 1, \ldots, m - 1\}$ (since every integer z satisfies $\overrightarrow{z} = z\%m \in \{0, 1, \ldots, m - 1\}$). Hence, Proposition 3.3.2 **(c)** (applied to m, q and $\overrightarrow{ap + bw}$ instead of n, u and c) yields $\overrightarrow{ap + bw} = q\%m$ (since $\overrightarrow{ap + bw} \equiv q \bmod m$).

But from $(p, q) \in M \times M$, we obtain $q \in M = \{0, 1, \ldots, m - 1\}$. Hence, Proposition 3.3.2 **(c)** (applied to m, q and q instead of n, u and c) yields $q = q\%m$ (since $q \equiv q \bmod m$). Comparing this with $\overrightarrow{ap + bw} = q\%m$, we obtain $\overrightarrow{ap + bw} = q$. Thus, (713) becomes

$$(f \circ g)(\alpha) = \left(p, \underbrace{\overrightarrow{ap + bw}}_{=q}\right) = (p, q) = \alpha = \mathrm{id}_{M \times M}(\alpha).$$

Forget that we fixed α. Thus we have shown that $(f \circ g)(\alpha) = \mathrm{id}_{M \times M}(\alpha)$ for each $\alpha \in M \times M$. In other words, $f \circ g = \mathrm{id}_{M \times M}$.

We could now prove $g \circ f = \mathrm{id}_{M \times M}$ by a similar argument, and then conclude that the maps f and g are mutually inverse, whence f is bijective. However, let us instead take a shortcut: The set $M \times M$ is finite (since M is finite) and clearly

satisfies $|M \times M| = |M \times M|$. Hence, Corollary 6.2.1 (applied to $U = M \times M$ and $V = M \times M$) yields that the maps f and g are mutually inverse (since $f \circ g = \mathrm{id}_{M \times M}$). Thus, the map f is invertible, i.e., bijective. That is, f is a permutation of the set $M \times M$. This proves Claim 1.]

[*Proof of Claim 2:* Let $i \in \mathbb{N}$. Applying (270) to $z = x_{i+1}$, we obtain $\overrightarrow{x_{i+1}} \equiv x_{i+1} \bmod m$. Hence, $x_{i+1} \equiv \overrightarrow{x_{i+1}} \bmod m$. The same argument (applied to x_i instead of x_{i+1}) shows that $x_i \equiv \overrightarrow{x_i} \bmod m$.

Recall that the sequence $(x_0, x_1, x_2, \ldots)$ is (a, b)-recurrent. In other words, every $n \geq 2$ satisfies $x_n = a x_{n-1} + b x_{n-2}$ (by the definition of "(a, b)-recurrent"). Applying this to $n = i + 2$, we obtain $x_{i+2} = a \underbrace{x_{(i+2)-1}}_{= x_{i+1}} + b \underbrace{x_{(i+2)-2}}_{= x_i} = a x_{i+1} + b x_i$. Now, (270) (applied to $z = x_{i+2}$) yields

$$x_{i+2} = a \underbrace{x_{i+1}}_{\equiv \overrightarrow{x_{i+1}} \bmod m} + b \underbrace{x_i}_{\equiv \overrightarrow{x_i} \bmod m} \equiv a \overrightarrow{x_{i+1}} + b \overrightarrow{x_i} \bmod m.$$

Now, Proposition 3.3.4 (applied to m, x_{i+2} and $a \overrightarrow{x_{i+1}} + b \overrightarrow{x_i}$ instead of n, u and v) yields that $x_{i+2} \equiv a \overrightarrow{x_{i+1}} + b \overrightarrow{x_i} \bmod m$ if and only if $x_{i+2} \% m = \left(a \overrightarrow{x_{i+1}} + b \overrightarrow{x_i}\right) \% m$. Hence, we have $x_{i+2} \% m = \left(a \overrightarrow{x_{i+1}} + b \overrightarrow{x_i}\right) \% m$ (since we have $x_{i+2} \equiv a \overrightarrow{x_{i+1}} + b \overrightarrow{x_i} \bmod m$). The definition of $\overrightarrow{x_{i+2}}$ now yields

$$\overrightarrow{x_{i+2}} = x_{i+2} \% m = \left(a \overrightarrow{x_{i+1}} + b \overrightarrow{x_i}\right) \% m = \overrightarrow{a \overrightarrow{x_{i+1}} + b \overrightarrow{x_i}} \tag{714}$$

(since the definition of $\overrightarrow{a \overrightarrow{x_{i+1}} + b \overrightarrow{x_i}}$ yields $\overrightarrow{a \overrightarrow{x_{i+1}} + b \overrightarrow{x_i}} = \left(a \overrightarrow{x_{i+1}} + b \overrightarrow{x_i}\right) \% m$).

Now, the definition of f yields

$$f\left(\left(\overrightarrow{x_i}, \overrightarrow{x_{i+1}}\right)\right) = \left(\overrightarrow{x_{i+1}}, \underbrace{\overrightarrow{a \overrightarrow{x_{i+1}} + b \overrightarrow{x_i}}}_{\substack{= \overrightarrow{x_{i+2}} \\ \text{(by (714))}}}\right) = \left(\overrightarrow{x_{i+1}}, \overrightarrow{x_{i+2}}\right).$$

This proves Claim 2.]

[*Proof of Claim 3:* We shall prove Claim 3 by induction on i:

Induction base: We have $f^0 = \mathrm{id}_{M \times M}$ and thus $f^0(\omega) = \mathrm{id}_{M \times M}(\omega) = \omega = \left(\overrightarrow{x_0}, \overrightarrow{x_1}\right)$ (by the definition of ω). Now, $\left(\overrightarrow{x_0}, \overrightarrow{x_{0+1}}\right) = \left(\overrightarrow{x_0}, \overrightarrow{x_1}\right) = f^0(\omega)$ (since $f^0(\omega) = \left(\overrightarrow{x_0}, \overrightarrow{x_1}\right)$). In other words, Claim 3 holds for $i = 0$.

Induction step: Let $j \in \mathbb{N}$. Assume (as the induction hypothesis) that Claim 3 holds for $i = j$. We must show that Claim 3 holds for $i = j + 1$. In other words, we must show that $\left(\overrightarrow{x_{j+1}}, \overrightarrow{x_{(j+1)+1}}\right) = f^{j+1}(\omega)$.

We have assumed that Claim 3 holds for $i = j$. In other words, we have

$\left(\overrightarrow{x_j}, \overrightarrow{x_{j+1}}\right) = f^j(\omega)$. But $f^{j+1} = f \circ f^j$, so that

$$\begin{aligned} f^{j+1}(\omega) &= \left(f \circ f^j\right)(\omega) = f\left(\underbrace{f^j(\omega)}_{=\left(\overrightarrow{x_j}, \overrightarrow{x_{j+1}}\right)}\right) = f\left(\left(\overrightarrow{x_j}, \overrightarrow{x_{j+1}}\right)\right) \\ &= \left(\overrightarrow{x_{j+1}}, \overrightarrow{x_{j+2}}\right) \qquad \text{(by Claim 2, applied to } i = j) \\ &= \left(\overrightarrow{x_{j+1}}, \overrightarrow{x_{(j+1)+1}}\right) \qquad \text{(since } j+2 = (j+1)+1). \end{aligned}$$

In other words, $\left(\overrightarrow{x_{j+1}}, \overrightarrow{x_{(j+1)+1}}\right) = f^{j+1}(\omega)$. This completes the induction step. Thus, Claim 3 is proven.]

We have now proved Claims 1, 2 and 3; thus, Exercise 6.3.6 is solved. □

□

A.7.7. Discussion of Exercise 6.3.7

Discussion of Exercise 6.3.7. The slickest way to solve Exercise 6.3.7 is probably by using the following variant of Theorem 6.2.8 **(a)**:

Theorem A.7.1. Let X be a finite set. Let $n = |X|$. Assume that $n > 1$. Let $p \in X$. Let $f : X \to X$ be a permutation of X satisfying $f(p) = p$. Let $x \in X$. Then, there exists a $k \in \{1, 2, \ldots, n-1\}$ such that $f^k(x) = x$.

Proof of Theorem A.7.1. Informally, the idea is the following: If $x = p$, then this is clear (just set $k = 1$). But if $x \neq p$, then we can "throw away" the element p from the set X (indeed, $f(p) = p$ shows that p does not "interfere" with the remaining elements of X), and obtain an $(n-1)$-element set $X \setminus \{p\}$, to which we can apply Theorem 6.2.8 **(a)**.

Here are the details of this argument: It is easy to see that Theorem A.7.1 holds if $x = p$ [403]. Hence, for the rest of this proof, we WLOG assume that $x \neq p$.

The map f is a permutation of X; thus, f is a bijection. Hence, f is bijective, thus injective.

Set $Y = X \setminus \{p\}$. Thus, $|Y| = |X \setminus \{p\}| = |X| - 1$ (since $p \in X$), so that $|Y| = \underbrace{|X|}_{=n} - 1 = n - 1$. In particular, the set Y is finite.

Combining $x \in X$ with $x \neq p$, we obtain $x \in X \setminus \{p\} = Y$ (since $Y = X \setminus \{p\}$).

[403] *Proof.* Assume that $x = p$. Note that $n - 1 > 0$ (since $n > 1$), so that $n - 1$ is a positive integer. Hence, $1 \in \{1, 2, \ldots, n-1\}$. Furthermore, $\underbrace{f^1}_{=f}(p) = f(p) = p$. In other words, $f^1(x) = x$ (since $x = p$). So the element $1 \in \{1, 2, \ldots, n-1\}$ satisfies $f^1(x) = x$. Hence, there exists a $k \in \{1, 2, \ldots, n-1\}$ such that $f^k(x) = x$ (namely, $k = 1$). In other words, the claim of Theorem A.7.1 holds. Thus, Theorem A.7.1 is proved under the assumption that $x = p$.

For each $y \in Y$, we have $f(y) \in Y$ [404]. Thus, we can define the map

$$\begin{aligned} g : Y &\to Y, \\ y &\mapsto f(y). \end{aligned}$$

This map g is "essentially just" the map f, except with the element p removed from its domain and target.

It is easy to see (by a straightforward induction) that

$$g^k(x) = f^k(x) \qquad \text{for each } k \in \mathbb{N}. \tag{715}$$

[405]

The map g is injective[406]. Thus, $g : Y \to Y$ is an injective map. Hence, Corollary 6.2.9 **(a)** (applied to Y and g instead of X and f) yields that g is a permutation of Y. Hence, Theorem 6.2.8 **(a)** (applied to Y, $n-1$ and g instead of X, n and f) yields that there exists a $k \in \{1, 2, \ldots, n-1\}$ such that $g^k(x) = x$ (since $n-1 = |Y|$ and $x \in Y$). In view of (715), this rewrites as follows: There exists a $k \in \{1, 2, \ldots, n-1\}$ such that $f^k(x) = x$. This proves Theorem A.7.1. □

Now, we can solve Exercise 6.3.7:

Solution to Exercise 6.3.7. We shall use all notations introduced in our above solution to Exercise 6.2.3.

[404] *Proof.* Let $y \in Y$. Then, $y \in Y = X \setminus \{p\}$; in other words, $y \in X$ and $y \neq p$.

But the map f is injective. Therefore, if we had $f(y) = f(p)$, then we would have $y = p$, which would contradict $y \neq p$. Hence, we cannot have $f(y) = f(p)$. Therefore, $f(y) \neq f(p) = p$. Combining $f(y) \in X$ with $f(y) \neq p$, we obtain $f(y) \in X \setminus \{p\} = Y$. Qed.

[405] *Proof of (715):* We shall prove (715) by induction on k:

Induction base: We have $g^0 = \mathrm{id}_Y$ and thus $g^0(x) = \mathrm{id}_Y(x) = x$. Also, $f^0 = \mathrm{id}_X$ and thus $f^0(x) = \mathrm{id}_X(x) = x$. Hence, $g^0(x) = x = f^0(x)$. In other words, (715) holds for $k = 0$.

Induction step: Let $\ell \in \mathbb{N}$. Assume (as the induction hypothesis) that (715) holds for $k = \ell$. We must prove that (715) holds for $k = \ell + 1$. In other words, we must prove that $g^{\ell+1}(x) = f^{\ell+1}(x)$.

We have assumed that (715) holds for $k = \ell$. In other words, $g^\ell(x) = f^\ell(x)$. Now,

$$\begin{aligned} \underbrace{g^{\ell+1}}_{=g \circ g^\ell}(x) &= \left(g \circ g^\ell\right)(x) = g\left(\underbrace{g^\ell(x)}_{=f^\ell(x)}\right) = g\left(f^\ell(x)\right) \\ &= f\left(f^\ell(x)\right) \qquad \text{(by the definition of } g\text{)} \\ &= \underbrace{\left(f \circ f^\ell\right)}_{=f^{\ell+1}}(x) = f^{\ell+1}(x). \end{aligned}$$

This completes the induction step. Thus, (715) is proved.

[406] *Proof.* This is a straightforward consequence of the fact that f is injective. In more detail: Let $a, b \in Y$ satisfy $g(a) = g(b)$. The definition of g yields $g(a) = f(a)$ and $g(b) = f(b)$. Hence, $f(a) = g(a) = g(b) = f(b)$. Since f is injective, this entails $a = b$. Now, forget that we fixed a, b. We thus have shown that if $a, b \in Y$ satisfy $g(a) = g(b)$, then $a = b$. In other words, the map g is injective.

The definition of $\overrightarrow{0}$ yields $\overrightarrow{0} = 0\%m = 0$. Now, the definition of f yields

$$\begin{aligned} f((0,0)) &= \left(0, \overrightarrow{a \cdot 0 + b \cdot 0}\right) = \left(0, \overrightarrow{0}\right) \qquad (\text{since } a \cdot 0 + b \cdot 0 = 0) \\ &= (0,0) \qquad \left(\text{since } \overrightarrow{0} = 0\right). \end{aligned}$$

We also know (from Claim 1 in our solution to Exercise 6.2.3) that the map f is bijective. Hence, f is a permutation of the set $M \times M$. Moreover, from $m > 1$, we obtain $m^2 > 1$. Hence, we can apply Theorem A.7.1 to $X = M \times M$, $n = m^2$, $p = (0,0)$ and $x = \omega$ (since $|M \times M| = m^2$). This yields that there exists a $k \in \{1, 2, \ldots, m^2 - 1\}$ such that $f^k(\omega) = \omega$. Consider this k.

Now, it is easy to see that the sequence $(x_0\%m, x_1\%m, x_2\%m, \ldots)$ is k-periodic (indeed, this can be shown exactly as in the solution to Exercise 6.2.3). Thus, we have found a $k \in \{1, 2, \ldots, m^2 - 1\}$ such that the sequence $(x_0\%m, x_1\%m, x_2\%m, \ldots)$ is k-periodic. This solves Exercise 6.3.7. □

□

A.7.8. Discussion of Exercise 6.3.8

Discussion of Exercise 6.3.8. Exercise 6.3.8 was problem 9 on the Baltic Way mathematical contest 2004; it has been discussed in [Grinbe08, Aufgabe 2.20] and in `https://artofproblemsolving.com/community/c6h20213` . We can regard it as a more complicated variant of Exercise 1.1.9 (the solution uses the same idea, but more work is needed to apply it). The following is probably the simplest solution:

Solution to Exercise 6.3.8. Define $n+1$ integers $b_0, b_1, \ldots, b_n$ as follows:

- For each $i \in \{0, 1, \ldots, n-1\}$, set
$$b_i = a_1 + a_2 + \cdots + a_i. \tag{716}$$
(Thus, $b_1 = a_1$ and $b_0 = (\text{empty sum}) = 0$.)
- Set
$$b_n = a_2. \tag{717}$$

Recall that the remainder of an integer u upon division by n is denoted by $u\%n$. Now, the $n+1$ remainders

$$b_0\%n, \quad b_1\%n, \quad b_2\%n, \quad \ldots, \quad b_n\%n$$

are $n+1$ elements of the n-element set $\{0, 1, \ldots, n-1\}$ (since they are remainders upon division by n), and thus at least two of them must be equal (by the Pigeonhole Principle[407]). In other words, there exist two integers u and v with $0 \leq u < v \leq n$

[407] To be more specific: by Corollary 6.1.4 (applied to $n+1$, $\{0, 1, \ldots, n-1\}$ and $b_{i-1}\%n$ instead of m, V and a_i).

and $b_u\%n = b_v\%n$. Consider these u and v. From $u < v$, we obtain $u+1 \leq v$. Also, $0 \leq u$, thus $1 \leq u+1 \leq v \leq n$. Hence, both $u+1$ and v belong to $\{1,2,\ldots,n\}$. Moreover, from $u+1 \leq v$, we see that the set $\{u+1, u+2, \ldots, v\}$ is nonempty.

We have $u < v \leq n$ and thus $u \in \{0,1,\ldots,n-1\}$ (since u is a nonnegative integer). Therefore, (716) (applied to $i = u$) yields $b_u = a_1 + a_2 + \cdots + a_u$. Furthermore, $\{1,2,\ldots,u\} \subseteq \{1,2,\ldots,n-1\}$ (since $u \in \{0,1,\ldots,n-1\}$).

The integers b_v and b_u leave the same remainder when divided by n (since $b_v\%n = b_u\%n$); thus, $b_v \equiv b_u \bmod n$ (by Proposition 3.3.4, applied to b_v and b_u instead of u and v). In other words, $n \mid b_v - b_u$.

Now, we are in one of the following two cases:

Case 1: We have $v \neq n$.

Case 2: We have $v = n$.

Let us consider Case 1 first. In this case, we have $v \neq n$. Hence, $v < n$ (since $v \leq n$), so that $v \leq n-1$ (since v and n are integers). Thus, $v \in \{0,1,\ldots,n-1\}$ (since $0 \leq v$). Therefore, (716) (applied to $i = v$) yields $b_v = a_1 + a_2 + \cdots + a_v$. Also, the set $\{u+1, u+2, \ldots, v\}$ is nonempty (as we have seen) and is a subset of $\{1,2,\ldots,n-1\}$ (since $1 \leq u+1$ and $v \leq n-1$).

Now, recall that $n \mid b_v - b_u$. In view of

$$\begin{aligned}
\underbrace{b_v}_{=a_1+a_2+\cdots+a_v} - \underbrace{b_u}_{=a_1+a_2+\cdots+a_u} &= (a_1 + a_2 + \cdots + a_v) - (a_1 + a_2 + \cdots + a_u) \\
&= a_{u+1} + a_{u+2} + \cdots + a_v \qquad \text{(since } u < v\text{)} \\
&= \sum_{i \in \{u+1, u+2, \ldots, v\}} a_i,
\end{aligned}$$

this rewrites as $n \mid \sum_{i \in \{u+1, u+2, \ldots, v\}} a_i$. Since we know that $\{u+1, u+2, \ldots, v\}$ is a nonempty subset of $\{1,2,\ldots,n-1\}$, we can thus conclude that there exists a nonempty subset I of $\{1,2,\ldots,n-1\}$ such that $n \mid \sum_{i \in I} a_i$ (namely, $I = \{u+1, u+2, \ldots, v\}$). Hence, Exercise 6.3.8 is solved in Case 1.

Let us now consider Case 2. In this case, we have $v = n$. Thus, $b_v = b_n = a_2$ (by (717)). Recall that

$$n \mid \underbrace{b_v}_{=a_2} - b_u = a_2 - b_u \mid -(a_2 - b_u) = b_u - a_2.$$

Note that $n \geq 3$, so that $n-1 \geq 2 \geq 1$ and thus $1 \in \{0,1,\ldots,n-1\}$. Hence, (716) (applied to $i = 1$) yields $b_1 = a_1$. If we had $u = 1$, then we would have $b_u = b_1 = a_1$ and therefore

$$n \mid \underbrace{b_u}_{=a_1} - a_2 = a_1 - a_2,$$

which would contradict the assumption $n \nmid a_1 - a_2$. Hence, we cannot have $u = 1$. Thus, we have $u \neq 1$.

We are now in one of the following two subcases:

Subcase 2.1: We have $u = 0$.

Subcase 2.2: We have $u \neq 0$.

Let us consider Subcase 2.1 first. In this subcase, we have $u = 0$. But $0 \in \{0, 1, \ldots, n-1\}$ (since $n - 1 \geq 1 \geq 0$); therefore, (716) (applied to $i = 0$) yields $b_0 = a_1 + a_2 + \cdots + a_0 = (\text{empty sum}) = 0$. Now, from $u = 0$, we obtain $b_u = b_0 = 0$. But

$$n \mid \underbrace{b_v}_{=a_2} - \underbrace{b_u}_{=0} = a_2 = \sum_{i \in \{2\}} a_i.$$

Note that $2 \in \{1, 2, \ldots, n-1\}$ (because $n - 1 \geq 2$), and therefore $\{2\}$ is a nonempty subset of $\{1, 2, \ldots, n-1\}$. Thus, we conclude that there exists a nonempty subset I of $\{1, 2, \ldots, n-1\}$ such that $n \mid \sum_{i \in I} a_i$ (namely, $I = \{2\}$), because we have $n \mid \sum_{i \in \{2\}} a_i$. Hence, Exercise 6.3.8 is solved in Subcase 2.1.

Let us finally consider Subcase 2.2. In this subcase, we have $u \neq 0$. But u is a nonnegative integer (since $0 \leq u$) and satisfies $u \neq 0$ and $u \neq 1$. Hence, $u \geq 2$. Thus, $2 \in \{1, 2, \ldots, u\}$. Moreover, the set $\{1, 2, \ldots, u\} \setminus \{2\}$ contains (at least) the element 1, and thus is nonempty. This nonempty set $\{1, 2, \ldots, u\} \setminus \{2\}$ is furthermore a subset of $\{1, 2, \ldots, n-1\}$ (since $\{1, 2, \ldots, u\} \setminus \{2\} \subseteq \{1, 2, \ldots, u\} \subseteq \{1, 2, \ldots, n-1\}$).

Recall that

$$b_u = a_1 + a_2 + \cdots + a_u = \sum_{i \in \{1, 2, \ldots, u\}} a_i = a_2 + \sum_{i \in \{1, 2, \ldots, u\} \setminus \{2\}} a_i$$

(here, we have split off the addend for $i = 2$ from the sum, since $2 \in \{1, 2, \ldots, u\}$). Therefore,

$$b_u - a_2 = \sum_{i \in \{1, 2, \ldots, u\} \setminus \{2\}} a_i. \tag{718}$$

Now, recall that $n \mid b_u - a_2$. In view of (718), this rewrites as $n \mid \sum_{i \in \{1, 2, \ldots, u\} \setminus \{2\}} a_i$. Since we know that $\{1, 2, \ldots, u\} \setminus \{2\}$ is a nonempty subset of $\{1, 2, \ldots, n-1\}$, we can thus conclude that there exists a nonempty subset I of $\{1, 2, \ldots, n-1\}$ such that $n \mid \sum_{i \in I} a_i$ (namely, $I = \{1, 2, \ldots, u\} \setminus \{2\}$). Hence, Exercise 6.3.8 is solved in Subcase 2.2.

We have now solved Exercise 6.3.8 in both Subcases 2.1 and 2.2. Hence, Exercise 6.3.8 holds in Case 2.

We have now solved Exercise 6.3.8 in both Cases 1 and 2. Thus, the solution to Exercise 6.3.8 is complete. □

□

A.7.9. Discussion of Exercise 6.3.9

Discussion of Exercise 6.3.9. Exercise 6.3.9 **(c)** is one of the simplest parts of the *equidistribution theorem* (and Exercise 6.3.9 **(a)** and Exercise 6.3.9 **(b)** are weaker versions of it, which are useful as stepping stones in the solution of Exercise 6.3.9 **(c)**).

A deeper part of the equidistribution theorem claims that if $x \in \mathbb{R}$ is irrational, then the sequence $(\operatorname{frac}(0x), \operatorname{frac}(1x), \operatorname{frac}(2x), \ldots)$ is uniformly distributed on the half-open interval $[0,1)$ (that is, roughly speaking, its entries are equally likely to end up near any point of this interval[408]). However, Exercise 6.3.9 **(c)** is already useful enough for some applications. Exercise 6.3.9 **(d)** is a sample application of Exercise 6.3.9 **(c)**.

Our solution to Exercise 6.3.9 **(a)** will rely on the following two simple properties of floors and fractional parts:

Lemma A.7.2. Let $x \in \mathbb{R}$ and $r \in \mathbb{Z}$ satisfy $r \leq x < r+1$. Then, $\lfloor x \rfloor = r$ and $\operatorname{frac} x = x - r$.

Proof of Lemma A.7.2. We know that r is an integer (since $r \in \mathbb{Z}$). This integer r is $\leq x$ (since $r \leq x$), whereas the next integer $r+1$ is no longer $\leq x$ (since $x < r+1$). Hence, r is the largest integer that is $\leq x$. In other words, r is $\lfloor x \rfloor$ (since $\lfloor x \rfloor$ was defined to be the largest integer that is $\leq x$). In other words, $\lfloor x \rfloor = r$. Furthermore, the definition of $\operatorname{frac} x$ yields $\operatorname{frac} x = x - \underbrace{\lfloor x \rfloor}_{=r} = x - r$. Thus, Lemma A.7.2 is proved. □

Lemma A.7.3. Let $a, b \in \mathbb{R}$. Then:
(a) If $\operatorname{frac} a \geq \operatorname{frac} b$, then $\operatorname{frac}(a-b) = \operatorname{frac} a - \operatorname{frac} b$.
(b) If $\operatorname{frac} a < \operatorname{frac} b$, then $\operatorname{frac}(a-b) = \operatorname{frac} a - \operatorname{frac} b + 1$.

Proof of Lemma A.7.3. Let $p = \lfloor a \rfloor$ and $q = \lfloor b \rfloor$. Thus, p and q are integers (since the floor of any real number is an integer). In other words, $p \in \mathbb{Z}$ and $q \in \mathbb{Z}$. Hence, $p - q \in \mathbb{Z}$ and $p - q - 1 \in \mathbb{Z}$.

The definition of $\operatorname{frac} a$ yields $\operatorname{frac} a = a - \underbrace{\lfloor a \rfloor}_{=p} = a - p$. The definition of $\operatorname{frac} b$ yields $\operatorname{frac} b = b - \underbrace{\lfloor b \rfloor}_{=q} = b - q$.

The chain of inequalities (1) (applied to $x = a$) says that $\lfloor a \rfloor \leq a < \lfloor a \rfloor + 1$. In view of $p = \lfloor a \rfloor$, this rewrites as $p \leq a < p+1$. The same argument (applied to b and q instead of a and p) yields $q \leq b < q+1$.

(a) Assume that $\operatorname{frac} a \geq \operatorname{frac} b$. This rewrites as $a - p \geq b - q$ (since $\operatorname{frac} a = a - p$ and $\operatorname{frac} b = b - q$). In other words, $a + q \geq b + p$. In other words, $a - b \geq p - q$. That is, $p - q \leq a - b$. Moreover, $\underbrace{a}_{<p+1} - \underbrace{b}_{\substack{\geq q \\ (\text{since } q \leq b)}} < p + 1 - q = p - q + 1$. Hence, $p - q \leq a - b < p - q + 1$. Therefore, Lemma A.7.2 (applied to $x = a - b$ and $r = p - q$) yields that $\lfloor a - b \rfloor = p - q$ and $\operatorname{frac}(a-b) = (a-b) - (p-q)$. Hence,

$$\operatorname{frac}(a-b) = (a-b) - (p-q) = \underbrace{(a-p)}_{\substack{=\operatorname{frac} a \\ (\text{since } \operatorname{frac} a = a-p)}} - \underbrace{(b-q)}_{\substack{=\operatorname{frac} b \\ (\text{since } \operatorname{frac} b = b-q)}} = \operatorname{frac} a - \operatorname{frac} b.$$

[408] Don't mistake this for a rigorous statement.

This proves Lemma A.7.3 **(a)**.

(b) Assume that $\operatorname{frac} a < \operatorname{frac} b$. This rewrites as $a - p < b - q$ (since $\operatorname{frac} a = a - p$ and $\operatorname{frac} b = b - q$). In other words, $a + q < b + p$. In other words, $a - b < p - q$. Hence, $a - b < p - q = (p - q - 1) + 1$. Moreover, $\underbrace{a}_{\substack{\geq p \\ (\text{since } p \leq a)}} - \underbrace{b}_{<q+1} > p - (q+1) = p - q - 1$. Hence, $p - q - 1 \leq a - b < (p - q - 1) + 1$. Therefore, Lemma A.7.2 (applied to $x = a - b$ and $r = p - q - 1$) yields that $\lfloor a - b \rfloor = p - q - 1$ and $\operatorname{frac}(a - b) = (a - b) - (p - q - 1)$. Hence,

$$\begin{aligned}\operatorname{frac}(a-b) &= (a-b) - (p-q-1) = \underbrace{(a-p)}_{\substack{=\operatorname{frac} a \\ (\text{since } \operatorname{frac} a = a-p)}} - \underbrace{(b-q)}_{\substack{=\operatorname{frac} b \\ (\text{since } \operatorname{frac} b = b-q)}} + 1 \\ &= \operatorname{frac} a - \operatorname{frac} b + 1.\end{aligned}$$

This proves Lemma A.7.3 **(b)**. □

Solution to Exercise 6.3.9. **(a)** Let $x \in \mathbb{R}$. Let n be a positive integer.

The numbers $\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}$ subdivide the half-open interval $[0, 1)$ into n equally sized segments. Denote these segments (themselves half-open intervals) by $I_1, I_2, \ldots, I_n$. More precisely: Define n half-open intervals $I_1, I_2, \ldots, I_n$ by

$$I_1 = \left[\frac{0}{n}, \frac{1}{n}\right), \qquad I_2 = \left[\frac{1}{n}, \frac{2}{n}\right), \qquad \ldots, \qquad I_n = \left[\frac{n-1}{n}, \frac{n}{n}\right)$$

(that is, $I_k = \left[\frac{k-1}{n}, \frac{k}{n}\right)$ for each $k \in \{1, 2, \ldots, n\}$). The union of these n intervals is $I_1 \cup I_2 \cup \cdots \cup I_n = [0, 1)$; furthermore, these n intervals $I_1, I_2, \ldots, I_n$ are disjoint. Hence, each number $y \in [0, 1)$ lies in exactly one of these n intervals $I_1, I_2, \ldots, I_n$.

Now, consider the $n + 1$ numbers $\operatorname{frac}(0x), \operatorname{frac}(1x), \ldots, \operatorname{frac}(nx)$. Each of these $n + 1$ numbers belongs to $[0, 1)$ (since $\operatorname{frac} z \in [0, 1)$ for each $z \in \mathbb{R}$), and thus must lie in one of the n intervals $I_1, I_2, \ldots, I_n$ (since we know that each number $y \in [0, 1)$ lies in exactly one of these n intervals $I_1, I_2, \ldots, I_n$). Hence, by the Pigeonhole Principle, we see that at least two of these $n + 1$ numbers $\operatorname{frac}(0x), \operatorname{frac}(1x), \ldots, \operatorname{frac}(nx)$ must belong to one and the same of the n intervals $I_1, I_2, \ldots, I_n$ [409]. In other words,

[409] Here is this argument in more detail:

Let $i \in \{0, 1, \ldots, n\}$. Then,

$$\begin{aligned}\operatorname{frac}(ix) &\in [0, 1) \qquad (\text{since } \operatorname{frac} z \in [0, 1) \text{ for each } z \in \mathbb{R}) \\ &= I_1 \cup I_2 \cup \cdots \cup I_n \qquad (\text{since } I_1 \cup I_2 \cup \cdots \cup I_n = [0, 1)).\end{aligned}$$

Hence, there exists some $k \in \{1, 2, \ldots, n\}$ such that $\operatorname{frac}(ix) \in I_k$. Fix such a k, and denote it by b_i. (Note that this k is unique, since the n intervals $I_1, I_2, \ldots, I_n$ are disjoint; but we will not need this in our argument.)

Forget that we fixed i. We thus have defined an element $b_i \in \{1, 2, \ldots, n\}$ for each $i \in \{0, 1, \ldots, n\}$. In other words, we have defined $n + 1$ elements $b_0, b_1, \ldots, b_n \in \{1, 2, \ldots, n\}$.

there exist two elements i and j of $\{0,1,\ldots,n\}$ such that $i<j$ and such that the two numbers $\operatorname{frac}(ix)$ and $\operatorname{frac}(jx)$ belong to one and the same of the n intervals $I_1, I_2, \ldots, I_n$. Consider these i and j.

The two numbers $\operatorname{frac}(ix)$ and $\operatorname{frac}(jx)$ belong to one and the same of the n intervals $I_1, I_2, \ldots, I_n$. In other words, there exists some $k \in \{1,2,\ldots,n\}$ such that $\operatorname{frac}(ix)$ and $\operatorname{frac}(jx)$ belong to I_k. Consider this k.

We know that $\operatorname{frac}(ix)$ belongs to I_k. Hence, $\operatorname{frac}(ix) \in I_k = \left[\frac{k-1}{n}, \frac{k}{n}\right)$ (by the definition of I_k). In other words, $\operatorname{frac}(ix) \geq \frac{k-1}{n}$ and $\operatorname{frac}(ix) < \frac{k}{n}$. The same argument (applied to j instead of i) shows that $\operatorname{frac}(jx) \geq \frac{k-1}{n}$ and $\operatorname{frac}(jx) < \frac{k}{n}$.

Now, $j-i$ is a positive integer (since i and j are integers satisfying $i<j$). We now claim that

$$\operatorname{frac}((j-i)x) \text{ is either } < \frac{1}{n} \text{ or } > \frac{n-1}{n}. \tag{719}$$

[*Proof of (719):* We are in one of the following two cases:

Case 1: We have $\operatorname{frac}(jx) \geq \operatorname{frac}(ix)$.

Case 2: We have $\operatorname{frac}(jx) < \operatorname{frac}(ix)$.

Let us first consider Case 1. In this case, we have $\operatorname{frac}(jx) \geq \operatorname{frac}(ix)$. Hence, Lemma A.7.3 **(a)** (applied to $a = jx$ and $b = ix$) yields

$$\operatorname{frac}(jx-ix) = \underbrace{\operatorname{frac}(jx)}_{<\frac{k}{n}} - \underbrace{\operatorname{frac}(ix)}_{\geq\frac{k-1}{n}} < \frac{k}{n} - \frac{k-1}{n} = \frac{1}{n}.$$

In view of $jx - ix = (j-i)x$, this rewrites as $\operatorname{frac}((j-i)x) < \frac{1}{n}$. Hence, $\operatorname{frac}((j-i)x)$ is either $< \frac{1}{n}$ or $> \frac{n-1}{n}$. Thus, (719) is proved in Case 1.

Let us now consider Case 2. In this case, we have $\operatorname{frac}(jx) < \operatorname{frac}(ix)$. Hence,

Hence, Corollary 6.1.4 (applied to $\{1,2,\ldots,n\}$, $n+1$ and b_{i-1} instead of V, m and a_i) yields that at least two of these $n+1$ elements $b_0, b_1, \ldots, b_n$ are equal. In other words, there exist two numbers $i, j \in \{0,1,\ldots,n\}$ such that $i<j$ and $b_i = b_j$. Consider these i, j.

Our definition of b_i shows that b_i is a $k \in \{1,2,\ldots,n\}$ satisfying $\operatorname{frac}(ix) \in I_k$. Hence, $\operatorname{frac}(ix) \in I_{b_i}$. The same argument (applied to j instead of i) yields $\operatorname{frac}(jx) \in I_{b_j}$. In view of $b_i = b_j$, this rewrites as $\operatorname{frac}(jx) \in I_{b_i}$.

Now, both $\operatorname{frac}(ix)$ and $\operatorname{frac}(jx)$ belong to I_{b_i}. Hence, $\operatorname{frac}(ix)$ and $\operatorname{frac}(jx)$ belong to one and the same of the n intervals $I_1, I_2, \ldots, I_n$ (namely, to I_{b_i}). Since $i<j$, this shows that at least two of these $n+1$ numbers $\operatorname{frac}(0x), \operatorname{frac}(1x), \ldots, \operatorname{frac}(nx)$ must belong to one and the same of the n intervals $I_1, I_2, \ldots, I_n$.

Note that we have not used the fact that the intervals $I_1, I_2, \ldots, I_n$ are disjoint.

Lemma A.7.3 **(b)** (applied to $a = jx$ and $b = ix$) yields

$$\operatorname{frac}(jx - ix) = \underbrace{\operatorname{frac}(jx)}_{\geq \frac{k-1}{n}} - \underbrace{\operatorname{frac}(ix)}_{< \frac{k}{n}} + 1 > \frac{k-1}{n} - \frac{k}{n} + 1 = \frac{n-1}{n}.$$

In view of $jx - ix = (j-i)x$, this rewrites as $\operatorname{frac}((j-i)x) > \frac{n-1}{n}$. Hence, $\operatorname{frac}((j-i)x)$ is either $< \frac{1}{n}$ or $> \frac{n-1}{n}$. Thus, (719) is proved in Case 2.

We have now proved (719) in both Cases 1 and 2. Hence, (719) always holds.]

Now, recall that $j - i$ is a positive integer. Furthermore, (719) shows that $\operatorname{frac}((j-i)x)$ is either $< \frac{1}{n}$ or $> \frac{n-1}{n}$. Hence, there exists a positive integer m such that $\operatorname{frac}(mx)$ is either $< \frac{1}{n}$ or $> \frac{n-1}{n}$ (namely, $m = j - i$). This solves Exercise 6.3.9 **(a)**.

(b) Let $x \in \mathbb{R}$. Let n be a positive integer. Exercise 6.3.9 **(a)** yields that there exists a positive integer m such that $\operatorname{frac}(mx)$ is either $< \frac{1}{n}$ or $> \frac{n-1}{n}$. Consider this m, and denote it by p. Thus, p is a positive integer such that $\operatorname{frac}(px)$ is either $< \frac{1}{n}$ or $> \frac{n-1}{n}$.

We must prove that there exists a positive integer m such that $\operatorname{frac}(mx) < \frac{1}{n}$. If $\operatorname{frac}(px) < \frac{1}{n}$, then this is immediately clear (since we can just take $m = p$). Thus, for the rest of this proof, we WLOG assume that we don't have $\operatorname{frac}(px) < \frac{1}{n}$. In other words, $\operatorname{frac}(px)$ is not $< \frac{1}{n}$. Hence, $\operatorname{frac}(px)$ is $> \frac{n-1}{n}$ (since we know that $\operatorname{frac}(px)$ is either $< \frac{1}{n}$ or $> \frac{n-1}{n}$). In other words, $\operatorname{frac}(px) > \frac{n-1}{n}$.

But every $z \in \mathbb{R}$ satisfies $\operatorname{frac} z < 1$. Hence, $\operatorname{frac}(px) < 1$, so that $1 - \operatorname{frac}(px) > 0$. Thus, $\frac{1}{1 - \operatorname{frac}(px)}$ is a well-defined positive real. Hence, there exists an integer u such that $u > \frac{1}{1 - \operatorname{frac}(px)}$ (for example, we can take $u = \left\lfloor \frac{1}{1 - \operatorname{frac}(px)} \right\rfloor + 1$). Consider this u. Clearly, $u > \frac{1}{1 - \operatorname{frac}(px)} > 0$ (since $1 - \operatorname{frac}(px)$ is positive), so that u is positive. Also, from $u > \frac{1}{1 - \operatorname{frac}(px)}$, we obtain $\frac{1}{u} < 1 - \operatorname{frac}(px)$ (since u and $1 - \operatorname{frac}(px)$ are positive) and thus

$$\frac{1}{u} < 1 - \underbrace{\operatorname{frac}(px)}_{> \frac{n-1}{n}} < 1 - \frac{n-1}{n} = \frac{1}{n}.$$

Now, let us apply Exercise 6.3.9 **(a)** to px and u instead of x and n. We thus conclude that there exists a positive integer m such that $\operatorname{frac}(mpx)$ is either $< \frac{1}{u}$ or $> \frac{u-1}{u}$. Consider this m, and denote it by r. Thus, r is a positive integer such that $\operatorname{frac}(rpx)$ is either $< \frac{1}{u}$ or $> \frac{u-1}{u}$.

The product rp is a positive integer (since r and p are positive integers). We must prove that there exists a positive integer m such that $\operatorname{frac}(mx) < \frac{1}{n}$. If $\operatorname{frac}(rpx) < \frac{1}{n}$, then this is immediately clear (since we can just take $m = rp$). Thus, for the rest of this proof, we WLOG assume that we don't have $\operatorname{frac}(rpx) < \frac{1}{n}$. Hence, we have $\operatorname{frac}(rpx) \geq \frac{1}{n} > \frac{1}{u}$ (since $\frac{1}{u} < \frac{1}{n}$). Therefore, $\operatorname{frac}(rpx)$ is not $< \frac{1}{u}$. Hence, $\operatorname{frac}(rpx)$ is $> \frac{u-1}{u}$ (since we know that $\operatorname{frac}(rpx)$ is either $< \frac{1}{u}$ or $> \frac{u-1}{u}$). In other words, $\operatorname{frac}(rpx) > \frac{u-1}{u}$.

Now,

$$\begin{aligned} \operatorname{frac}(rpx) &> \frac{u-1}{u} = 1 - \frac{1}{u} > 1 - (1 - \operatorname{frac}(px)) \qquad \left(\text{since } \frac{1}{u} < 1 - \operatorname{frac}(px)\right) \\ &= \operatorname{frac}(px). \end{aligned} \tag{720}$$

If we had $r = 1$, then we would have $\operatorname{frac}\left(\underbrace{r}_{=1} px\right) = \operatorname{frac}(px)$, which would contradict (720). Hence, we cannot have $r = 1$. Thus, $r > 1$ (since r is a positive integer), so that $r - 1$ is a positive integer. Thus, $(r-1)p$ is a positive integer (since p is a positive integer). Moreover, (720) entails $\operatorname{frac}(rpx) \geq \operatorname{frac}(px)$. Hence, Lemma A.7.3 **(a)** (applied to $a = rpx$ and $b = px$) yields

$$\operatorname{frac}(rpx - px) = \underbrace{\operatorname{frac}(rpx)}_{\substack{<1 \\ (\text{since } \operatorname{frac} z < 1 \\ \text{for each } z \in \mathbb{R})}} - \underbrace{\operatorname{frac}(px)}_{> \frac{n-1}{n}} < 1 - \frac{n-1}{n} = \frac{1}{n}.$$

In view of $rpx - px = (r-1)px$, this rewrites as $\operatorname{frac}((r-1)px) < \frac{1}{n}$.

Thus, $(r-1)p$ is a positive integer and satisfies $\operatorname{frac}((r-1)px) < \frac{1}{n}$. Hence, there exists a positive integer m such that $\operatorname{frac}(mx) < \frac{1}{n}$ (namely, $m = (r-1)p$). This solves Exercise 6.3.9 **(b)**.

(c) Let $x \in \mathbb{R}$. Let ε be a positive real. Thus, $\frac{1}{\varepsilon}$ is a well-defined positive real. Hence, there exists some integer n such that $n > \frac{1}{\varepsilon}$ (for example, we can take

$n = \left\lfloor \frac{1}{\varepsilon} \right\rfloor + 1$). Consider this n. Note that $n > \frac{1}{\varepsilon} > 0$ (since ε is positive); thus, n is positive. Furthermore, from $n > \frac{1}{\varepsilon}$, we obtain $\frac{1}{n} < \varepsilon$ (since n and ε are positive).

Exercise 6.3.9 **(b)** shows that there exists a positive integer m such that $\operatorname{frac}(mx) < \frac{1}{n}$. This positive integer m must then also satisfy $\operatorname{frac}(mx) < \varepsilon$ (since it satisfies $\operatorname{frac}(mx) < \frac{1}{n} < \varepsilon$). Hence, there exists a positive integer m such that $\operatorname{frac}(mx) < \varepsilon$. This solves Exercise 6.3.9 **(c)**.

(d) Here is where we will need a bit of analysis. We will use the following two well-known facts about the sine function:

Fact 1: If x is a nonnegative real, then $\sin x \leq x$.

Fact 2: If $x \in \mathbb{R}$ and $j \in \mathbb{Z}$, then $\sin(x + 2\pi j) = \sin x$.

Note that Fact 2 is a consequence of the fact that the function sin has period 2π. We can easily use these two facts to derive the following:

Fact 3: If $z \in \mathbb{R}$, then $\sin z = \sin\left(2\pi \operatorname{frac}\left(\frac{z}{2\pi}\right)\right)$.

[*Proof of Fact 3:* Let $z \in \mathbb{R}$. Then, $\left\lfloor \frac{z}{2\pi} \right\rfloor$ is an integer[410]. Now, the definition of $\operatorname{frac}\left(\frac{z}{2\pi}\right)$ yields $\operatorname{frac}\left(\frac{z}{2\pi}\right) = \frac{z}{2\pi} - \left\lfloor \frac{z}{2\pi} \right\rfloor$. Hence, $\frac{z}{2\pi} = \operatorname{frac}\left(\frac{z}{2\pi}\right) + \left\lfloor \frac{z}{2\pi} \right\rfloor$. Multiplying both sides of this equality by 2π, we find

$$z = 2\pi\left(\operatorname{frac}\left(\frac{z}{2\pi}\right) + \left\lfloor \frac{z}{2\pi} \right\rfloor\right) = 2\pi \operatorname{frac}\left(\frac{z}{2\pi}\right) + 2\pi \left\lfloor \frac{z}{2\pi} \right\rfloor.$$

Hence,

$$\sin z = \sin\left(2\pi \operatorname{frac}\left(\frac{z}{2\pi}\right) + 2\pi \left\lfloor \frac{z}{2\pi} \right\rfloor\right) = \sin\left(2\pi \operatorname{frac}\left(\frac{z}{2\pi}\right)\right)$$

(by Fact 2, applied to $x = 2\pi \operatorname{frac}\left(\frac{z}{2\pi}\right)$ and $j = \left\lfloor \frac{z}{2\pi} \right\rfloor$), since $\left\lfloor \frac{z}{2\pi} \right\rfloor$ is an integer. This proves Fact 3.]

Let ε be a positive real. Let z be a real. We must show that there exists a positive integer m such that $0 \leq \sin(mz) < \varepsilon$.

Indeed, we WLOG assume that $\varepsilon \leq \pi$ (since otherwise, we can simply replace ε by the positive real $\min\{\varepsilon, \pi\}$, which is no larger than ε).

The real $\frac{\varepsilon}{2\pi}$ is positive (since ε and 2π are positive). Hence, Exercise 6.3.9 **(c)** (applied to $\frac{z}{2\pi}$ and $\frac{\varepsilon}{2\pi}$ instead of x and ε) yields that there exists a positive integer m such that $\operatorname{frac}\left(m \cdot \frac{z}{2\pi}\right) < \frac{\varepsilon}{2\pi}$. Consider this m. We have

$$\operatorname{frac}\left(\frac{mz}{2\pi}\right) = \operatorname{frac}\left(m \cdot \frac{z}{2\pi}\right) < \frac{\varepsilon}{2\pi}.$$

[410] since a floor of a real number is always an integer

Multiplying both sides of this inequality by 2π, we obtain

$$2\pi \operatorname{frac}\left(\frac{mz}{2\pi}\right) < \varepsilon \leq \pi. \tag{721}$$

But $\operatorname{frac}\left(\frac{mz}{2\pi}\right) \geq 0$ (since $\operatorname{frac} y \geq 0$ for each $y \in \mathbb{R}$) and thus $2\pi \operatorname{frac}\left(\frac{mz}{2\pi}\right) \geq 0$. Combining this with (721), we obtain $2\pi \operatorname{frac}\left(\frac{mz}{2\pi}\right) \in [0, \pi)$, so that

$$\sin\left(2\pi \operatorname{frac}\left(\frac{mz}{2\pi}\right)\right) \geq 0. \tag{722}$$

Now, Fact 3 (applied to mz instead of z) yields

$$\begin{aligned} \sin(mz) &= \sin\left(2\pi \operatorname{frac}\left(\frac{mz}{2\pi}\right)\right) \\ &\geq 0 \qquad (\text{by (722)}). \end{aligned} \tag{723}$$

Also, recall that $2\pi \operatorname{frac}\left(\frac{mz}{2\pi}\right) \geq 0$; thus, Fact 1 (applied to $x = 2\pi \operatorname{frac}\left(\frac{mz}{2\pi}\right)$) yields

$$\sin\left(2\pi \operatorname{frac}\left(\frac{mz}{2\pi}\right)\right) \leq 2\pi \operatorname{frac}\left(\frac{mz}{2\pi}\right) < \varepsilon.$$

Thus, (723) becomes

$$\sin(mz) = \sin\left(2\pi \operatorname{frac}\left(\frac{mz}{2\pi}\right)\right) < \varepsilon.$$

Combining this with $\sin(mz) \geq 0$, we obtain $0 \leq \sin(mz) < \varepsilon$.

Thus, we have found a positive integer m such that $0 \leq \sin(mz) < \varepsilon$. Hence, such an m exists. This solves Exercise 6.3.9 **(d)**. □

See `https://artofproblemsolving.com/community/c4h15415p109278` for slightly different solution to Exercise 6.3.9 **(d)** (using the continuity of sin instead of Fact 1). □

A.7.10. Discussion of Exercise 6.3.10

Discussion of Exercise 6.3.10. I have taken Exercise 6.3.10 from `https://artofproblemsolving.com/community/c6h15141` . Here is my solution:

Solution to Exercise 6.3.10 (sketched). We claim that the smallest positive real ε such that the entire set R can be covered with 5 **closed** intervals of length ε each is $\frac{1}{10}$. In order to prove this, we need to prove the following two claims:

> *Claim 1:* The entire set R can be covered with 5 closed intervals of length $\frac{1}{10}$ each.

Claim 2: Let ε be a positive real such that the entire set R can be covered with 5 closed intervals of length ε each. Then, $\varepsilon \geq \dfrac{1}{10}$.

[*Proof of Claim 1:* Define five closed intervals A, B, C, D, E by

$$A = \left[0, \frac{1}{10}\right], \qquad B = \left[\frac{1}{10}, \frac{1}{5}\right], \qquad C = \left[\frac{1}{4}, \frac{7}{20}\right], \qquad D = \left[\frac{1}{2}, \frac{3}{5}\right], \qquad E = \left[\frac{9}{10}, 1\right].$$

These five intervals A, B, C, D, E have length $\dfrac{1}{10}$ each, and together cover the entire set R (indeed, the interval A contains all numbers $\dfrac{1}{s}$ with $s \geq 10$; the interval B contains the numbers $\dfrac{1}{9}, \dfrac{1}{8}, \dfrac{1}{7}, \dfrac{1}{6}, \dfrac{1}{5}$; the interval C contains the numbers $\dfrac{1}{4}, \dfrac{1}{3}$; the interval D contains the number $\dfrac{1}{2}$; the interval E contains the number 1). Thus, the entire set R can be covered with 5 closed intervals of length $\dfrac{1}{10}$ each (namely, with the 5 intervals A, B, C, D, E). This proves Claim 1.]

[*Proof of Claim 2:* Assume the contrary. Thus, $\varepsilon < \dfrac{1}{10}$. In other words, $\dfrac{1}{10} - \varepsilon > 0$. Hence, $\dfrac{1}{\frac{1}{10} - \varepsilon}$ is a well-defined positive real. Hence, there exists an integer N such that $N > \dfrac{1}{\frac{1}{10} - \varepsilon}$ (for example, $N = \left\lfloor \dfrac{1}{\frac{1}{10} - \varepsilon} \right\rfloor + 1$). Consider this N. We have $N > \dfrac{1}{\frac{1}{10} - \varepsilon} > 0$ (since $\dfrac{1}{10} - \varepsilon > 0$). Also, from $N > \dfrac{1}{\frac{1}{10} - \varepsilon}$, we obtain $\dfrac{1}{N} < \dfrac{1}{10} - \varepsilon$ (since N and $\dfrac{1}{10} - \varepsilon$ are positive). Hence, $\varepsilon < \dfrac{1}{10} - \dfrac{1}{N}$, so that $\dfrac{1}{10} - \dfrac{1}{N} > \varepsilon$.

We have assumed that the entire set R can be covered with 5 closed intervals of length ε each. In other words, there exist 5 closed intervals A, B, C, D, E of length ε each such that these 5 intervals cover the entire set R. Consider these 5 intervals A, B, C, D, E.

Now, consider the 6 numbers

$$\frac{1}{1}, \quad \frac{1}{2}, \quad \frac{1}{3}, \quad \frac{1}{5}, \quad \frac{1}{10}, \quad \frac{1}{N}.$$

These 6 numbers all belong to R, and therefore each of them belongs to (at least) one of the 5 intervals A, B, C, D, E (since these 5 intervals cover the entire set R). Hence, by the Pigeonhole Principle, we conclude that at least two of these 6 numbers belong to one and the same of the 5 intervals A, B, C, D, E. These two numbers therefore belong to a common interval of length ε (since each of the 5 intervals

A, B, C, D, E has length ε), and thus their difference (if we subtract the smaller one from the larger) must be $\leq \varepsilon$. But this contradicts the fact that no two of the 6 numbers

$$\frac{1}{1}, \quad \frac{1}{2}, \quad \frac{1}{3}, \quad \frac{1}{5}, \quad \frac{1}{10}, \quad \frac{1}{N}$$

have a difference that is $\leq \varepsilon$ (indeed, the differences between consecutive numbers in this list are

$$\begin{aligned}
&\frac{1}{1} - \frac{1}{2} = \frac{1}{2} > \frac{1}{10} > \varepsilon, \\
&\frac{1}{2} - \frac{1}{3} = \frac{1}{6} > \frac{1}{10} > \varepsilon, \\
&\frac{1}{3} - \frac{1}{5} = \frac{2}{15} > \frac{1}{10} > \varepsilon, \\
&\frac{1}{5} - \frac{1}{10} = \frac{1}{10} > \varepsilon, \\
&\frac{1}{10} - \frac{1}{N} > \varepsilon,
\end{aligned}$$

and therefore the differences between non-consecutive numbers in this list are even larger, since the list is in decreasing order). This contradiction shows that our assumption was false. Thus, Claim 2 is proved.]

With Claim 1 and Claim 2 both proved, we thus have solved Exercise 6.3.10. □

□

A.8. Homework set #7 discussion

The following are discussions of the problems on homework set #7 (Section 7.7).

A.8.1. Discussion of Exercise 7.7.1

Discussion of Exercise 7.7.1. Exercise 7.7.1 is a minor generalization of [17f-hw6s, Exercise 2 **(a)**]. The solution (essentially copied from [17f-hw6s, Exercise 2 **(a)**]) is a textbook application of the bijection principle:

Solution to Exercise 7.7.1 (sketched). We have $g(x) \in \operatorname{Fix}(g \circ f)$ for each $x \in \operatorname{Fix}(f \circ g)$ [411]. The same argument (applied to V, U, g and f instead of U, V, f and g) shows that $f(x) \in \operatorname{Fix}(f \circ g)$ for each $x \in \operatorname{Fix}(g \circ f)$.

[411] *Proof.* Let $x \in \operatorname{Fix}(f \circ g)$. We must show that $g(x) \in \operatorname{Fix}(g \circ f)$.

We know that $x \in \operatorname{Fix}(f \circ g)$. In other words, x is a fixed point of $f \circ g$ (since $\operatorname{Fix}(f \circ g)$ is defined as the set of all fixed points of $f \circ g$). In other words, $x \in V$ and $(f \circ g)(x) = x$. Thus, $f(g(x)) = (f \circ g)(x) = x$. Hence,

$$(g \circ f)(g(x)) = g\left(\underbrace{f(g(x))}_{=x}\right) = g(x).$$

Let γ be the map

$$\operatorname{Fix}(f \circ g) \to \operatorname{Fix}(g \circ f), \qquad x \mapsto g(x).$$

(This is well-defined, because $g(x) \in \operatorname{Fix}(g \circ f)$ for each $x \in \operatorname{Fix}(f \circ g)$.)

Let φ be the map

$$\operatorname{Fix}(g \circ f) \to \operatorname{Fix}(f \circ g), \qquad x \mapsto f(x).$$

(This is well-defined, because $f(x) \in \operatorname{Fix}(f \circ g)$ for each $x \in \operatorname{Fix}(g \circ f)$.)

We have $\varphi \circ \gamma = \operatorname{id}$ [412]. The same argument (applied to V, U, g, f, γ and φ instead of U, V, f, g, φ and γ) shows that $\gamma \circ \varphi = \operatorname{id}$.

The maps φ and γ are mutually inverse (since $\varphi \circ \gamma = \operatorname{id}$ and $\gamma \circ \varphi = \operatorname{id}$), and thus are bijections. Hence, we have found a bijection $\operatorname{Fix}(f \circ g) \to \operatorname{Fix}(g \circ f)$ (namely, γ). Hence, $|\operatorname{Fix}(f \circ g)| = |\operatorname{Fix}(g \circ f)|$. This solves Exercise 7.7.1. $\square$

A few remarks are in order:

- The assumption that U and V are finite in Exercise 7.7.1 is unnecessary (we have never used it in the above solution).
- It is **not** true that every three maps f, g, h from a finite set S to S satisfy $|\operatorname{Fix}(f \circ g \circ h)| = |\operatorname{Fix}(g \circ f \circ h)|$. See [17f-hw6s, Exercise 2 **(b)**] for a counterexample.
- Exercise 7.7.1 is a combinatorial analogue of the following fact from linear algebra: If A is an $n \times m$-matrix and B is an $m \times n$-matrix (so that both products AB and BA are well-defined), then

  $$\operatorname{Tr}(AB) = \operatorname{Tr}(BA) \tag{724}$$

  (where $\operatorname{Tr} M$ denotes the *trace* of a square matrix M, that is, the sum of all diagonal entries of M). This is not just an analogy; it is in fact possible to derive Exercise 7.7.1 by applying (724) to two appropriate matrices! See [17f-hw6s, Remark 0.11] for this derivation. This yields another solution to Exercise 7.7.1.

$\square$

In other words, $g(x)$ is a fixed point of $g \circ f$. In other words, $g(x) \in \operatorname{Fix}(g \circ f)$ (since $\operatorname{Fix}(g \circ f)$ is defined as the set of all fixed points of $g \circ f$). This completes our proof.

[412] *Proof.* Let $x \in \operatorname{Fix}(f \circ g)$. Thus, $\gamma(x) = g(x)$ (by the definition of γ). But $\gamma(x) \in \operatorname{Fix}(g \circ f)$, so that $\varphi(\gamma(x)) = f(\gamma(x))$ (by the definition of φ).

We have $x \in \operatorname{Fix}(f \circ g)$. In other words, x is a fixed point of $f \circ g$ (since $\operatorname{Fix}(f \circ g)$ is defined as the set of all fixed points of $f \circ g$). In other words, $x \in V$ and $(f \circ g)(x) = x$. Now,

$$(\varphi \circ \gamma)(x) = \varphi(\gamma(x)) = f\left(\underbrace{\gamma(x)}_{=g(x)}\right) = f(g(x)) = (f \circ g)(x) = x = \operatorname{id}(x).$$

Now, forget that we fixed x. We thus have proven that $(\varphi \circ \gamma)(x) = \operatorname{id}(x)$ for each $x \in \operatorname{Fix}(f \circ g)$. In other words, $\varphi \circ \gamma = \operatorname{id}$.

A.8.2. Discussion of Exercise 7.7.2

Discussion of Exercise 7.7.2. Exercise 7.7.2 is a somewhat more sophisticated variant of Exercise 3.8.3, and can easily be reduced to the latter:

Solution to Exercise 7.7.2. We are in one of the following two cases:

Case 1: The integer n is even.

Case 2: The integer n is odd.

Let us first consider Case 1. In this case, the integer n is even. Thus, $2 \mid n$, so that $n/2 \in \mathbb{Z}$.

If d is an even positive divisor of n, then $d/2$ is a positive divisor of $n/2$ [413]. Hence, the map

$$\begin{aligned} \Phi : \{\text{even positive divisors of } n\} &\to \{\text{positive divisors of } n/2\}, \\ d &\mapsto d/2 \end{aligned}$$

is well-defined. Consider this map Φ.

If u is a positive divisor of $n/2$, then $2u$ is an even positive divisor of n [414]. Hence, the map

$$\begin{aligned} \Psi : \{\text{positive divisors of } n/2\} &\to \{\text{even positive divisors of } n\}, \\ u &\mapsto 2u \end{aligned}$$

is well-defined. Consider this map Ψ.

The maps Φ and Ψ are mutually inverse. (Indeed, the map Φ divides its argument by 2, whereas the map Ψ multiplies its argument by 2. Thus, clearly these two maps undo each other.)

Now, the map Φ is invertible (since the maps Φ and Ψ are mutually inverse), and thus is a bijection. Hence, we have found a bijection

$$\Phi : \{\text{even positive divisors of } n\} \to \{\text{positive divisors of } n/2\}.$$

The bijection principle thus yields

$$|\{\text{even positive divisors of } n\}| = |\{\text{positive divisors of } n/2\}|.$$

[413] *Proof.* Let d be an even positive divisor of n. Then, $2 \mid d$ (since d is even), so that $d/2 \in \mathbb{Z}$. Moreover, there exists an integer c such that $n = dc$ (since d is a divisor of n). Consider this c. We have $\underbrace{n}_{=dc}/2 = dc/2 = (d/2)\,c$. Hence, $d/2 \mid n/2$ (since c is an integer). That is, $d/2$ is a divisor of $n/2$. Moreover, $d/2$ is positive (since d is positive). We have thus shown that $d/2$ is a positive divisor of $n/2$. Qed.

[414] *Proof.* Let u be a positive divisor of $n/2$. Then, u is an integer, so that $2u$ is an even integer. Moreover, $2u$ is positive (since u is positive). Furthermore, there exists some integer c such that $n/2 = uc$ (since u is a divisor of $n/2$). Consider this c. From $n/2 = uc$, we obtain $n = 2uc = (2u)\,c$. Hence, $2u$ is a divisor of n (since c is an integer). Thus, $2u$ is an even positive divisor of n (since $2u$ is even and positive). Qed.

In other words, the number of even positive divisors of n equals the number of positive divisors of $n/2$.

However, $n/2$ is positive (since n is positive) and is an integer (since $n/2 \in \mathbb{Z}$). Thus, Exercise 3.8.3 (applied to $n/2$ instead of n) yields that the number of positive divisors of $n/2$ is even if and only if $n/2$ is not a perfect square. Since the number of even positive divisors of n equals the number of positive divisors of $n/2$, we can rewrite this as follows: The number of even positive divisors of n is even if and only if $n/2$ is not a perfect square. Thus, Exercise 7.7.2 is solved in Case 1.

Let us now consider Case 2. In this case, the integer n is odd. Thus, $2 \nmid n$, so that $n/2 \notin \mathbb{Z}$. In other words, $n/2$ is not an integer. Hence, $n/2$ is not a perfect square (since each perfect square is an integer).

Moreover, the number n has no even positive divisors[415]. Thus, the number of even positive divisors of n is 0. Hence, the number of even positive divisors of n is even (since 0 is even).

We thus conclude that the number of even positive divisors of n is even if and only if $n/2$ is not a perfect square (because we know that the number of even positive divisors of n is even, and we also know that $n/2$ is not a perfect square). Thus, Exercise 7.7.2 is solved in Case 2.

We have now solved Exercise 7.7.2 in both Cases 1 and 2. This completes the solution to Exercise 7.7.2. □

□

A.8.3. Discussion of Exercise 7.7.3

Discussion of Exercise 7.7.3. Exercise 7.7.3 is essentially Problem 2 on the International Mathematical Olympiad 1981 (except that the latter problem asked for the average instead of the sum of the $\min S$; but one question is easily reduced to the other). We give two solutions:

First solution to Exercise 7.7.3. We have $r \neq 0$ (since r is positive). If S is a subset of $[n]$ satisfying $|S| = r$, then the set S is nonempty (since $|S| = r \neq 0$) and therefore its minimum $\min S$ exists. This minimum $\min S$ is an element of S and therefore an element of $[n]$ (since S is a subset of $[n]$). Hence, we can split the sum $\sum\limits_{\substack{S \subseteq [n]; \\ |S| = r}} \min S$

[415] *Proof.* Let d be an even positive divisor of n. We shall derive a contradiction.

Indeed, $2 \mid d$ (since d is even) and $d \mid n$ (since d is a divisor of n). Hence, $2 \mid d \mid n$. But this contradicts $2 \nmid n$.

Forget that we fixed d. We thus have found a contradiction for each even positive divisor d of n. Hence, there exist no even positive divisors of n. In other words, the number n has no even positive divisors.

according to the value of $\min S$ as follows:

$$\sum_{\substack{S\subseteq[n];\\|S|=r}} \min S = \sum_{k\in[n]} \sum_{\substack{S\subseteq[n];\\|S|=r;\\\min S=k}} \underbrace{\min S}_{=k} = \sum_{k\in[n]} \underbrace{\sum_{\substack{S\subseteq[n];\\|S|=r;\\\min S=k}} k}_{=(\#\text{ of subsets } S \text{ of } [n] \text{ satisfying } |S|=r \text{ and } \min S=k)\cdot k}$$

$$= \sum_{k\in[n]} (\#\text{ of subsets } S \text{ of } [n] \text{ satisfying } |S| = r \text{ and } \min S = k) \cdot k.$$

Now, let us fix $k \in [n]$. We shall now compute the # of subsets S of $[n]$ satisfying $|S| = r$ and $\min S = k$.

For the sake of brevity, let us introduce some terminology:

- A *red set* will mean a subset S of $[n]$ satisfying $|S| = r$ and $\min S = k$.
- A *blue set* will mean an $(r-1)$-element subset of $\{k+1, k+2, \ldots, n\}$.

The blue sets are easy to count (see the next paragraph), whereas the red sets are what we want to count. We shall soon find a bijection between {red sets} and {blue sets}, and therefore we will know how many red sets there are.

First, let us count the blue sets. Note that $k \in [n]$ entails $1 \leq k \leq n$; hence, we have $n - k \in \mathbb{N}$, and the set $\{k+1, k+2, \ldots, n\}$ is an $(n-k)$-element set. Theorem 4.3.12 (applied to $n-k$, $\{k+1, k+2, \ldots, n\}$ and $r-1$ instead of n, S and k) thus yields

$$\begin{aligned}\binom{n-k}{r-1} &= (\text{the number of } (r-1)\text{-element subsets of } \{k+1, k+2, \ldots, n\}) \\ &= (\#\text{ of } (r-1)\text{-element subsets of } \{k+1, k+2, \ldots, n\}).\end{aligned}$$

However, from the definition of a "blue set", we see that

$$(\#\text{ of blue sets}) = (\#\text{ of } (r-1)\text{-element subsets of } \{k+1, k+2, \ldots, n\}).$$

Comparing these two equalities, we thus find

$$(\#\text{ of blue sets}) = \binom{n-k}{r-1}. \tag{725}$$

Now, let us construct a bijection between {red sets} and {blue sets}. We shall present the proof in maximum possible detail, in order to show at least once what one needs to check when claiming the existence of a bijection; in practice, the actual checking can often be done easily in one's head (certainly this is true in the case at hand), and thus the definition of the bijection may well be sufficient.

If U is a red set, then $U \setminus \{k\}$ is a blue set[416]. Hence, we can define a map

$$\begin{aligned}\Phi : \{\text{red sets}\} &\to \{\text{blue sets}\}, \\ U &\mapsto U \setminus \{k\}.\end{aligned}$$

[416] *Proof.* Let U be a red set. We must show that $U \setminus \{k\}$ is a blue set.

Consider this map Φ.

If V is a blue set, then $V \cup \{k\}$ is a red set[417]. Hence, we can define a map

$$\begin{aligned} \Psi : \{\text{blue sets}\} &\to \{\text{red sets}\}, \\ V &\mapsto V \cup \{k\}. \end{aligned}$$

Consider this map Ψ.

We know that U is a red set. In other words, U is a subset S of $[n]$ satisfying $|S| = r$ and $\min S = k$ (by the definition of a "red set"). In other words, U is a subset of $[n]$ and satisfies $|U| = r$ and $\min U = k$. Hence, $k = \min U \in U$ (since the minimum of a set always belongs to this set) and therefore $|U \setminus \{k\}| = \underbrace{|U|}_{=r} - 1 = r - 1$. In other words, $U \setminus \{k\}$ is an $(r-1)$-element set.

Next, let $u \in U \setminus \{k\}$. Thus, $u \in U$ and $u \neq k$. We have $u \in U \subseteq [n]$ (since U is a subset of $[n]$); that is, u is an integer satisfying $1 \leq u \leq n$. But k is the minimum of the set U (since $k = \min U$), and therefore is $\leq$ to any element of U. In other words, $k \leq v$ for each $v \in U$. Applying this to $v = u$, we obtain $k \leq u$. Thus, $u \geq k$, so that $u > k$ (since $u \neq k$) and therefore $u \geq k + 1$ (since u and k are integers). Combining this with $u \leq n$, we obtain $u \in \{k+1, k+2, \ldots, n\}$ (since u is an integer).

Forget that we fixed u. We thus have shown that $u \in \{k+1, k+2, \ldots, n\}$ for each $u \in U \setminus \{k\}$. In other words, $U \setminus \{k\}$ is a subset of $\{k+1, k+2, \ldots, n\}$. Hence, we have shown that $U \setminus \{k\}$ is an $(r-1)$-element subset of $\{k+1, k+2, \ldots, n\}$. In other words, $U \setminus \{k\}$ is a blue set (by the definition of a "blue set"). Qed.

[417] *Proof.* Let V be a blue set. We must show that $V \cup \{k\}$ is a red set.

We know that V is a blue set. In other words, V is an $(r-1)$-element subset of $\{k+1, k+2, \ldots, n\}$ (by the definition of a "blue set"). In other words, $V \subseteq \{k+1, k+2, \ldots, n\}$ and $|V| = r - 1$.

If we had $k \in V$, then we would have $k \in V \subseteq \{k+1, k+2, \ldots, n\}$ and thus $k \geq k+1 > k$, which is absurd. Thus, we cannot have $k \in V$. Hence, $k \notin V$, so that $|V \cup \{k\}| = |V| + 1 = r$ (since $|V| = r - 1$). Moreover, the two sets V and $\{k\}$ are subsets of $\{k, k+1, \ldots, n\}$ (since $V \subseteq \{k+1, k+2, \ldots, n\} \subseteq \{k, k+1, \ldots, n\}$ and $\{k\} \subseteq \{k, k+1, \ldots, n\}$ (because $k \leq n$)); therefore, their union $V \cup \{k\}$ is a subset of $\{k, k+1, \ldots, n\}$ as well. Hence, each element of $V \cup \{k\}$ belongs to the set $\{k, k+1, \ldots, n\}$ and therefore is $\geq k$. Moreover, k is an element of $V \cup \{k\}$ (since $k \in \{k\} \subseteq V \cup \{k\}$). Thus, we have shown that k is an element of $V \cup \{k\}$ and has the property that each element of $V \cup \{k\}$ is $\geq k$. In other words, k is the **smallest** element of $V \cup \{k\}$. In other words, $k = \min(V \cup \{k\})$. Hence, $\min(V \cup \{k\}) = k$.

Furthermore, the two sets V and $\{k\}$ are subsets of $[n]$ (since $V \subseteq \{k+1, k+2, \ldots, n\} \subseteq \{k, k+1, \ldots, n\} \subseteq \{1, 2, \ldots, n\} = [n]$ and $\{k\} \subseteq [n]$ (since $k \in [n]$)); therefore, their union $V \cup \{k\}$ is a subset of $[n]$ as well.

Now, we have shown that $V \cup \{k\}$ is a subset of $[n]$ and satisfies $|V \cup \{k\}| = r$ and $\min(V \cup \{k\}) = k$. In other words, $V \cup \{k\}$ is a subset S of $[n]$ satisfying $|S| = r$ and $\min S = k$. In other words, $V \cup \{k\}$ is a red set (by the definition of a "red set"). Qed.

We have $\Phi \circ \Psi = \mathrm{id}$ [418] and $\Psi \circ \Phi = \mathrm{id}$ [419]. Thus, the two maps Φ and Ψ are mutually inverse. Hence, the map Φ is invertible, i.e., is a bijection. Now, the bijection principle yields $|\{\text{red sets}\}| = |\{\text{blue sets}\}|$ (since $\Phi : \{\text{red sets}\} \to \{\text{blue sets}\}$ is a bijection). Hence,

$$(\#\text{ of red sets}) = |\{\text{red sets}\}| = |\{\text{blue sets}\}| = (\#\text{ of blue sets}) = \binom{n-k}{r-1}$$

(by (725)). Comparing this with

$$(\#\text{ of red sets}) = (\#\text{ of subsets } S \text{ of } [n] \text{ satisfying } |S| = r \text{ and } \min S = k) \\ (\text{by the definition of a "red set"}),$$

we obtain

$$(\#\text{ of subsets } S \text{ of } [n] \text{ satisfying } |S| = r \text{ and } \min S = k) \\ = \binom{n-k}{r-1}. \tag{726}$$

Now, forget that we fixed k. We thus have proved (726) for each $k \in [n]$. Now,

[418] *Proof.* Let $V \in \{\text{blue sets}\}$. Thus, V is a blue set. In other words, V is an $(r-1)$-element subset of $\{k+1, k+2, \ldots, n\}$ (by the definition of a "blue set"). In other words, $V \subseteq \{k+1, k+2, \ldots, n\}$ and $|V| = r-1$.

If we had $k \in V$, then we would have $k \in V \subseteq \{k+1, k+2, \ldots, n\}$ and thus $k \geq k+1 > k$, which is absurd. Thus, we cannot have $k \in V$. Hence, $k \notin V$, so that $(V \cup \{k\}) \setminus \{k\} = V$. Now, we have $\Psi(V) = V \cup \{k\}$ (by the definition of Ψ), and furthermore

$$(\Phi \circ \Psi)(V) = \Phi(\Psi(V)) = \underbrace{(\Psi(V))}_{=V\cup\{k\}} \setminus \{k\} \qquad (\text{by the definition of } \Phi) \\ = (V \cup \{k\}) \setminus \{k\} = V = \mathrm{id}(V).$$

Forget that we fixed V. We thus have shown that $(\Phi \circ \Psi)(V) = \mathrm{id}(V)$ for each $V \in \{\text{blue sets}\}$. In other words, $\Phi \circ \Psi = \mathrm{id}$.

[419] *Proof.* Let $U \in \{\text{red sets}\}$. Thus, U is a red set. In other words, U is a subset S of $[n]$ satisfying $|S| = r$ and $\min S = k$ (by the definition of a "red set"). In other words, U is a subset of $[n]$ and satisfies $|U| = r$ and $\min U = k$. Hence, $k = \min U \in U$ (since the minimum of a set always belongs to this set). Therefore, $(U \setminus \{k\}) \cup \{k\} = U$. Now, we have $\Phi(U) = U \setminus \{k\}$ (by the definition of Φ), and furthermore

$$(\Psi \circ \Phi)(U) = \Psi(\Phi(U)) = \underbrace{(\Phi(U))}_{=U\setminus\{k\}} \cup \{k\} \qquad (\text{by the definition of } \Psi) \\ = (U \setminus \{k\}) \cup \{k\} = U = \mathrm{id}(U).$$

Forget that we fixed U. We thus have shown that $(\Psi \circ \Phi)(U) = \mathrm{id}(U)$ for each $U \in \{\text{red sets}\}$. In other words, $\Psi \circ \Phi = \mathrm{id}$.

recall that

$$\sum_{\substack{S\subseteq[n];\\|S|=r}} \min S = \sum_{k\in[n]} \underbrace{(\#\text{ of subsets } S \text{ of } [n] \text{ satisfying } |S|=r \text{ and } \min S = k)}_{\substack{=\binom{n-k}{r-1}\\ \text{(by (726))}}} \cdot k$$

$$= \sum_{k\in[n]} \binom{n-k}{r-1} \cdot k. \tag{727}$$

Now, let us compute the sum on the right hand side of this equality.

We have $r-1 \in \mathbb{N}$ (since r is a positive integer). Hence, Exercise 7.6.1 (applied to $x=1$ and $y=r-1$) yields

$$\binom{n+1}{1+(r-1)+1} = \sum_{k=0}^{n} \underbrace{\binom{k}{1}}_{\substack{=k\\ \text{(by (120),}\\ \text{applied to } k\\ \text{instead of } n)}} \binom{n-k}{r-1} = \sum_{k=0}^{n} k\binom{n-k}{r-1}$$

$$= \underbrace{0\binom{n-0}{r-1}}_{=0} + \sum_{k=1}^{n} k\binom{n-k}{r-1} = \underbrace{\sum_{k=1}^{n}}_{\substack{=\sum_{k\in[n]}\\ \text{(since } \{1,2,\ldots,n\}=[n])}} \underbrace{k\binom{n-k}{r-1}}_{=\binom{n-k}{r-1}\cdot k}$$

$$= \sum_{k\in[n]} \binom{n-k}{r-1} \cdot k.$$

Comparing this with (727), we obtain

$$\sum_{\substack{S\subseteq[n];\\|S|=r}} \min S = \binom{n+1}{1+(r-1)+1} = \binom{n+1}{r+1}$$

(since $1+(r-1)=r$). This solves Exercise 7.7.3. □

Second solution to Exercise 7.7.3 (sketched). We proceed by double counting. The things we shall count are the $(r+1)$-element subsets of the set $\{0,1,\ldots,n\}$. How many such subsets are there? Here are two ways to answer this question:

First way: The set $\{0,1,\ldots,n\}$ is an $(n+1)$-element set; thus, an easy application of Theorem 4.3.12 yields

$$(\#\text{ of } (r+1)\text{-element subsets of } \{0,1,\ldots,n\})$$
$$= \binom{n+1}{r+1}. \tag{728}$$

Second way: If U is an $(r+1)$-element subset of $\{0,1,\ldots,n\}$, then we define the *tail* of U to be the set $U \setminus \{\min U\}$. That is, the tail of U is defined to be U with the smallest element removed.[420] It is easy to see that if U is an $(r+1)$-element subset of $\{0,1,\ldots,n\}$, then the tail of U is an r-element subset of $[n]$ (indeed, removing the smallest element from U clearly reduces the size of U by 1, and furthermore all the remaining elements are positive[421] and therefore belong to $[n]$). Hence, the sum rule yields

$$\begin{aligned}&(\#\text{ of }(r+1)\text{-element subsets of }\{0,1,\ldots,n\})\\&=\sum_{\substack{S\subseteq[n];\\|S|=r}}(\#\text{ of }(r+1)\text{-element subsets of }\{0,1,\ldots,n\}\text{ with tail }S).\end{aligned}$$

Now, let S be an r-element subset of $[n]$. What is the # of $(r+1)$-element subsets of $\{0,1,\ldots,n\}$ with tail S ? Clearly, any such subset must contain all r elements of S as well as one extra element; this extra element must furthermore be smaller than $\min S$ (in order for S to be the tail of the subset). Thus, any $(r+1)$-element subset of $\{0,1,\ldots,n\}$ with tail S must have the form $S\cup\{i\}$ for some $i\in\{0,1,\ldots,\min S-1\}$. Conversely, any set of the latter form is an $(r+1)$-element subset of $\{0,1,\ldots,n\}$ with tail S (check this!). Thus, the $(r+1)$-element subsets of $\{0,1,\ldots,n\}$ with tail S are the $\min S$ sets

$$S\cup\{0\},\quad S\cup\{1\},\quad S\cup\{2\},\quad\ldots,\quad S\cup\{\min S-1\}.$$

Hence,

$$\begin{aligned}&(\#\text{ of }(r+1)\text{-element subsets of }\{0,1,\ldots,n\}\text{ with tail }S)\\&=\min S.\end{aligned}\tag{729}$$

Now, forget that we fixed S. We thus have proved (729) for each r-element subset S of $[n]$. Now, recall that

$$\begin{aligned}&(\#\text{ of }(r+1)\text{-element subsets of }\{0,1,\ldots,n\})\\&=\sum_{\substack{S\subseteq[n];\\|S|=r}}\underbrace{(\#\text{ of }(r+1)\text{-element subsets of }\{0,1,\ldots,n\}\text{ with tail }S)}_{\substack{=\min S\\(\text{by }(729))}}\\&=\sum_{\substack{S\subseteq[n];\\|S|=r}}\min S.\end{aligned}$$

Comparing this with (728), we obtain

$$\sum_{\substack{S\subseteq[n];\\|S|=r}}\min S=\binom{n+1}{r+1}.$$

[420] For example, if $U=\{2,4,6,7\}$, then the tail of U is $\{4,6,7\}$.

[421] because if one of them was 0, then it would be the smallest element of U and thus would have been removed

This solves Exercise 7.7.3 again. □

□

A.8.4. Discussion of Exercise 7.7.4

Discussion of Exercise 7.7.4. We shall call a tuple *onefree* if it does not contain 1 as an entry. Thus, Exercise 7.7.4 asks for the # of onefree compositions of n. Here is the answer:

Proposition A.8.1. Let $(f_0, f_1, f_2, \ldots)$ be the Fibonacci sequence (defined as in Definition 2.2.1). Then,

$$(\text{\# of onefree compositions of } n) = f_{n-1} \tag{730}$$

for each positive integer n.

We shall outline three proofs of Proposition A.8.1:

First proof of Proposition A.8.1 (sketched). We proceed by strong induction on n:

Induction step: Let m be a positive integer. Assume (as the induction hypothesis) that Proposition A.8.1 holds for $n < m$. We must prove that Proposition A.8.1 holds for $n = m$. In other words, we must prove that (# of onefree compositions of m) $= f_{m-1}$.

If $m = 1$ or $m = 2$, then this is easy to check[422]. Thus, for the rest of this proof, we WLOG assume that we have neither $m = 1$ nor $m = 2$. Hence, $m \geq 3$ (since m is a positive integer). Therefore, $m - 1$ and $m - 2$ are positive integers. Moreover, from $m \geq 3$, we obtain $m - 1 \geq 2$.

Definition 2.2.1 yields that $f_n = f_{n-1} + f_{n-2}$ for all $n \geq 2$. We can apply this to $n = m - 1$ (since $m - 1 \geq 2$), and thus obtain

$$f_{m-1} = \underbrace{f_{(m-1)-1}}_{=f_{m-2}} + \underbrace{f_{(m-1)-2}}_{=f_{m-3}} = f_{m-2} + f_{m-3}. \tag{731}$$

We know that $m - 1$ is a positive integer satisfying $m - 1 < m$. Hence, Proposition A.8.1 holds for $n = m - 1$ (since our induction hypothesis says that Proposition

[422] *Proof.* There are no onefree compositions of 1 (since the only composition of 1 is (1), which is clearly not onefree). That is, we have (# of onefree compositions of 1) $= 0$. Comparing this with $f_{1-1} = f_0 = 0$, we obtain (# of onefree compositions of 1) $= f_{1-1}$. Hence, (# of onefree compositions of m) $= f_{m-1}$ holds for $m = 1$.

The only onefree composition of 2 is (2) (since the only compositions of 2 are $(1,1)$ and (2), but $(1,1)$ is not onefree). Thus, we have (# of onefree compositions of 2) $= 1$. Comparing this with $f_{2-1} = f_1 = 1$, we obtain (# of onefree compositions of 2) $= f_{2-1}$. Hence, (# of onefree compositions of m) $= f_{m-1}$ holds for $m = 2$.

We have thus proved that (# of onefree compositions of m) $= f_{m-1}$ holds for $m = 1$ and for $m = 2$.

A.8.1 holds for $n < m$). In other words, we have

$$(\text{\# of onefree compositions of } m-1) = f_{(m-1)-1}. \tag{732}$$

The same argument (applied to $m-2$ instead of $m-1$) yields

$$(\text{\# of onefree compositions of } m-2) = f_{(m-2)-1}. \tag{733}$$

Now, if $(a_1, a_2, \ldots, a_k)$ is a onefree composition of m, then we necessarily have $k \geq 1$ [423], and therefore its first entry a_1 is well-defined. We say that a onefree composition $(a_1, a_2, \ldots, a_k)$ of m is

- *green* if $a_1 = 2$;
- *red* if $a_1 \neq 2$.

Each onefree composition $(a_1, a_2, \ldots, a_k)$ of m is either green or red (since it satisfies either $a_1 = 2$ or $a_1 \neq 2$), but cannot be both at the same time. Hence, the sum rule yields

$$\begin{aligned}&(\text{\# of onefree compositions of } m)\\ &= (\text{\# of green onefree compositions of } m)\\ &\qquad + (\text{\# of red onefree compositions of } m)\,.\end{aligned} \tag{734}$$

We shall now compute the two addends on the right hand side.

If $(a_1, a_2, \ldots, a_k)$ is a green onefree composition of m, then $(a_2, a_3, \ldots, a_k)$ is a onefree composition of $m-2$ [424]. Hence, we can define a map

$$\begin{aligned}\Phi : \{\text{green onefree compositions of } m\} &\to \{\text{onefree compositions of } m-2\}\,,\\ (a_1, a_2, \ldots, a_k) &\mapsto (a_2, a_3, \ldots, a_k)\,.\end{aligned}$$

[423] *Proof.* Let $(a_1, a_2, \ldots, a_k)$ be a onefree composition of m. Then, $(a_1, a_2, \ldots, a_k)$ is a composition of m, that is, a tuple of positive integers whose sum is m (by the definition of a "composition"). In other words, $a_1, a_2, \ldots, a_k$ are positive integers satisfying $a_1 + a_2 + \cdots + a_k = m$.

If we had $k = 0$, then we would have $a_1 + a_2 + \cdots + a_k = a_1 + a_2 + \cdots + a_0 = (\text{empty sum}) = 0$, which would contradict $a_1 + a_2 + \cdots + a_k = m > 0$. Hence, we cannot have $k = 0$. Thus, $k \neq 0$, so that $k \geq 1$ (since $k \in \mathbb{N}$). Qed.

[424] *Proof.* Let $(a_1, a_2, \ldots, a_k)$ be a green onefree composition of m. Then, $a_1 = 2$ (since $(a_1, a_2, \ldots, a_k)$ is green). Moreover, $(a_1, a_2, \ldots, a_k)$ is a composition of m, that is, a tuple of positive integers whose sum is m (by the definition of a "composition"). In other words, $a_1, a_2, \ldots, a_k$ are positive integers satisfying $a_1 + a_2 + \cdots + a_k = m$. Hence,

$$m = a_1 + a_2 + \cdots + a_k = \underbrace{a_1}_{=2} + (a_2 + a_3 + \cdots + a_k) = 2 + (a_2 + a_3 + \cdots + a_k)\,,$$

so that $a_2 + a_3 + \cdots + a_k = m - 2$.

The positive integers $a_1, a_2, \ldots, a_k$ are distinct from 1 (since the composition $(a_1, a_2, \ldots, a_k)$ is onefree, i.e., does not contain 1 as an entry). Thus, in particular, the positive integers $a_2, a_3, \ldots, a_k$ are distinct from 1. Now, $(a_2, a_3, \ldots, a_k)$ is a tuple of positive integers whose sum is $m-2$ (since $a_2 + a_3 + \cdots + a_k = m - 2$); in other words, $(a_2, a_3, \ldots, a_k)$ is a composition of $m-2$ (by the definition of a "composition"). This composition $(a_2, a_3, \ldots, a_k)$ doesn't contain 1 as an entry (since the positive integers $a_2, a_3, \ldots, a_k$ are distinct from 1); in other words, it is onefree. Hence, $(a_2, a_3, \ldots, a_k)$ is a onefree composition of $m-2$, qed.

It is easy to see that this map Φ is a bijection[425]. Hence, the bijection principle yields

$$|\{\text{green onefree compositions of } m\}| = |\{\text{onefree compositions of } m-2\}|.$$

Thus,

$$\begin{aligned}
&(\text{\# of green onefree compositions of } m) \\
&= |\{\text{green onefree compositions of } m\}| \\
&= |\{\text{onefree compositions of } m-2\}| \\
&= (\text{\# of onefree compositions of } m-2) \\
&= f_{(m-2)-1} \qquad (\text{by (733)}) \\
&= f_{m-3}. \qquad (735)
\end{aligned}$$

Next, let us count the red onefree compositions of m.

If $(a_1, a_2, \ldots, a_k)$ is a red onefree composition of m, then $(a_1 - 1, a_2, a_3, \ldots, a_k)$ is

[425] Indeed, we can easily construct a map inverse to Φ: Namely, if $(b_1, b_2, \ldots, b_\ell)$ is a onefree composition of $m-2$, then $(2, b_1, b_2, \ldots, b_\ell)$ is a green onefree composition of m (check this!). Thus, the map

$$\begin{aligned}
\Psi : \{\text{onefree compositions of } m-2\} &\to \{\text{green onefree compositions of } m\}, \\
(b_1, b_2, \ldots, b_\ell) &\mapsto (2, b_1, b_2, \ldots, b_\ell)
\end{aligned}$$

is well-defined. It is straightforward to see that the maps Φ and Ψ are mutually inverse. Thus, the map Φ is invertible, i.e., is a bijection.

a onefree composition of $m-1$ [426]. Hence, we can define a map

$$\Phi' : \{\text{red onefree compositions of } m\} \to \{\text{onefree compositions of } m-1\}, \\ (a_1, a_2, \ldots, a_k) \mapsto (a_1 - 1, a_2, a_3, \ldots, a_k).$$

It is easy to see that this map Φ' is a bijection[427]. Hence, the bijection principle yields

$$|\{\text{red onefree compositions of } m\}| = |\{\text{onefree compositions of } m-1\}|.$$

Thus,

$$\begin{aligned} &(\#\text{ of red onefree compositions of } m) \\ &= |\{\text{red onefree compositions of } m\}| \\ &= |\{\text{onefree compositions of } m-1\}| \\ &= (\#\text{ of onefree compositions of } m-1) \\ &= f_{(m-1)-1} \qquad (\text{by (732)}) \\ &= f_{m-2}. \end{aligned} \tag{736}$$

[426] *Proof.* Let $(a_1, a_2, \ldots, a_k)$ be a red onefree composition of m. Then, $a_1 \neq 2$ (since $(a_1, a_2, \ldots, a_k)$ is red). Moreover, $(a_1, a_2, \ldots, a_k)$ is a composition of m, that is, a tuple of positive integers whose sum is m (by the definition of a "composition"). In other words, $a_1, a_2, \ldots, a_k$ are positive integers satisfying $a_1 + a_2 + \cdots + a_k = m$. Comparing $a_1 + a_2 + \cdots + a_k = m$ with

$$a_1 + a_2 + \cdots + a_k = \underbrace{a_1}_{=1+(a_1-1)} + (a_2 + a_3 + \cdots + a_k) = 1 + (a_1 - 1) + (a_2 + a_3 + \cdots + a_k),$$

we obtain $1 + (a_1 - 1) + (a_2 + a_3 + \cdots + a_k) = m$. Hence, $(a_1 - 1) + (a_2 + a_3 + \cdots + a_k) = m - 1$.

The positive integers $a_1, a_2, \ldots, a_k$ are distinct from 1 (since the composition $(a_1, a_2, \ldots, a_k)$ is onefree, i.e., does not contain 1 as an entry). Thus, in particular, the positive integer a_1 is distinct from 1, and the positive integers $a_2, a_3, \ldots, a_k$ are distinct from 1.

The positive integer a_1 is distinct from 1, and thus must be ≥ 2. Hence, $a_1 - 1 \geq 1$. This shows that $a_1 - 1$ is a positive integer. Moreover, $a_1 - 1 \neq 1$ (since $a_1 \neq 2$); thus, the number $a_1 - 1$ is distinct from 1. Since we also know that the numbers $a_2, a_3, \ldots, a_k$ are distinct from 1, we thus conclude that the numbers $a_1 - 1, a_2, a_3, \ldots, a_k$ are distinct from 1.

We now know that $a_1 - 1$ is a positive integer, and that $a_2, a_3, \ldots, a_k$ are positive integers as well. Hence, $(a_1 - 1, a_2, a_3, \ldots, a_k)$ is a tuple of positive integers. Moreover, $(a_1 - 1, a_2, a_3, \ldots, a_k)$ is a tuple of positive integers whose sum is $m - 1$ (since $(a_1 - 1) + (a_2 + a_3 + \cdots + a_k) = m - 1$); in other words, $(a_1 - 1, a_2, a_3, \ldots, a_k)$ is a composition of $m - 1$ (by the definition of a "composition"). This composition $(a_1 - 1, a_2, a_3, \ldots, a_k)$ doesn't contain 1 as an entry (since the numbers $a_1 - 1, a_2, a_3, \ldots, a_k$ are distinct from 1); in other words, it is onefree. Hence, $(a_1 - 1, a_2, a_3, \ldots, a_k)$ is a onefree composition of $m - 1$, qed.

[427] Indeed, we can easily construct a map inverse to Φ': Namely, if $(b_1, b_2, \ldots, b_\ell)$ is a onefree composition of $m - 1$, then the tuple $(b_1 + 1, b_2, b_3, \ldots, b_\ell)$ is well-defined (this requires proving that $\ell \geq 1$; check this!) and is a red onefree composition of m (check this!). Thus, the map

$$\Psi' : \{\text{onefree compositions of } m-1\} \to \{\text{red onefree compositions of } m\}, \\ (b_1, b_2, \ldots, b_\ell) \mapsto (b_1 + 1, b_2, b_3, \ldots, b_\ell)$$

is well-defined. It is straightforward to see that the maps Φ' and Ψ' are mutually inverse. Thus, the map Φ' is invertible, i.e., is a bijection.

Now, (734) becomes

$$\begin{aligned}
&(\#\text{ of onefree compositions of } m) \\
&= \underbrace{(\#\text{ of green onefree compositions of } m)}_{\substack{=f_{m-3} \\ \text{(by (735))}}} + \underbrace{(\#\text{ of red onefree compositions of } m)}_{\substack{=f_{m-2} \\ \text{(by (736))}}} \\
&= f_{m-3} + f_{m-2} = f_{m-2} + f_{m-3} = f_{m-1} \qquad \text{(by (731))}.
\end{aligned}$$

In other words, Proposition A.8.1 holds for $n = m$. This completes the induction step; thus, Proposition A.8.1 is proved. □

Second proof of Proposition A.8.1 (sketched). We proceed by strong induction on n:

Induction step: Let m be a positive integer. Assume (as the induction hypothesis) that Proposition A.8.1 holds for $n < m$. We must prove that Proposition A.8.1 holds for $n = m$. In other words, we must prove that (# of onefree compositions of m) $= f_{m-1}$.

If $m = 1$ or $m = 2$, then this is easy to check. Thus, for the rest of this proof, we WLOG assume that we have neither $m = 1$ nor $m = 2$. Hence, $m \geq 3$ (since m is a positive integer), so that $m - 3 \geq 0$.

If $(a_1, a_2, \ldots, a_k)$ is a onefree composition of m, then we necessarily have $k \geq 1$ [428], and therefore its first entry a_1 is well-defined. This first entry a_1 must be a positive integer (since $a_1, a_2, \ldots, a_k$ are positive integers) and satisfies $a_1 \neq 1$ (since the composition $(a_1, a_2, \ldots, a_k)$ is onefree). Hence, it satisfies $a_1 \geq 2$ (since any positive integer that is $\neq 1$ must be ≥ 2). It furthermore satisfies $a_1 \leq m$ [429], and thus $a_1 \in \{2, 3, \ldots, m\}$ (since $a_1 \geq 2$ and $a_1 \leq m$). Hence, the sum rule yields

$$\begin{aligned}
&(\#\text{ of onefree compositions of } m) \\
&= \sum_{r \in \{2,3,\ldots,m\}} (\#\text{ of onefree compositions } (a_1, a_2, \ldots, a_k) \text{ of } m \text{ such that } a_1 = r).
\end{aligned}$$

Now, let us fix some $r \in \{2, 3, \ldots, m\}$. We shall compute

$$(\#\text{ of onefree compositions } (a_1, a_2, \ldots, a_k) \text{ of } m \text{ such that } a_1 = r).$$

Indeed, it is not hard to see that the map

$$\begin{aligned}
&\{\text{onefree compositions } (a_1, a_2, \ldots, a_k) \text{ of } m \text{ such that } a_1 = r\} \\
&\to \{\text{onefree compositions of } m - r\}
\end{aligned}$$

[428] This follows from the positivity of m. See the first proof of Proposition A.8.1 above for the details of the argument.

[429] *Proof.* The numbers $a_1, a_2, \ldots, a_k$ are positive integers; hence, none of them cannot exceed their sum. In particular, this shows that $a_1 \leq a_1 + a_2 + \cdots + a_k = m$ (since $(a_1, a_2, \ldots, a_k)$ is a composition of m, that is, a tuple of positive integers whose sum is m).

that sends each $(a_1, a_2, \ldots, a_k)$ to $(a_2, a_3, \ldots, a_k)$ is a bijection[430]. Thus, the bijection principle yields

$$\begin{aligned}&\left|\{\text{onefree compositions } (a_1, a_2, \ldots, a_k) \text{ of } m \text{ such that } a_1 = r\}\right| \\ &= \left|\{\text{onefree compositions of } m - r\}\right|.\end{aligned}$$

In other words,

$$\begin{aligned}&(\#\text{ of onefree compositions } (a_1, a_2, \ldots, a_k) \text{ of } m \text{ such that } a_1 = r) \\ &= (\#\text{ of onefree compositions of } m - r). \qquad (737)\end{aligned}$$

Forget that we fixed r. We thus have proved (737) for each $r \in \{2, 3, \ldots, m\}$. Now,

$$\begin{aligned}&(\#\text{ of onefree compositions of } m) \\ &= \underbrace{\sum_{r \in \{2,3,\ldots,m\}}}_{= \sum_{r=2}^{m}} \underbrace{(\#\text{ of onefree compositions } (a_1, a_2, \ldots, a_k) \text{ of } m \text{ such that } a_1 = r)}_{\substack{=(\#\text{ of onefree compositions of } m-r) \\ \text{(by (737))}}} \\ &= \sum_{r=2}^{m} (\#\text{ of onefree compositions of } m - r) \\ &= \underbrace{\sum_{r=2}^{m-1} (\#\text{ of onefree compositions of } m - r)}_{\substack{= \sum_{n=1}^{m-2} (\#\text{ of onefree compositions of } n) \\ \text{(here, we have substituted } n \text{ for } m-r \text{ in the sum)}}} \\ &\qquad + \underbrace{(\#\text{ of onefree compositions of } m - m)}_{\substack{=1 \\ \text{(since } m-m=0\text{, and thus the only onefree composition of } m-m \text{ is the empty tuple } ()\text{)}}} \\ &\qquad \left(\begin{array}{c}\text{here, we have split off the addend for } r = m \text{ from the sum} \\ \text{(this addend was indeed in the sum, because } m \geq 3 \geq 2\text{)}\end{array}\right) \\ &= \sum_{n=1}^{m-2} \underbrace{(\#\text{ of onefree compositions of } n)}_{\substack{=f_{n-1} \\ \text{(by our induction hypothesis, since } n \leq m-2 < m\text{)}}} + 1 \\ &= \underbrace{\sum_{n=1}^{m-2} f_{n-1}}_{\substack{=f_0+f_1+\cdots+f_{m-3} \\ =f_{m-1}-1 \\ \text{(by Exercise 2.2.1 (applied to } n=m-3\text{),} \\ \text{since } m-3 \geq 0\text{)}}} + 1 = f_{m-1} - 1 + 1 = f_{m-1}.\end{aligned}$$

[430] Its inverse map, of course, sends each onefree composition $(b_1, b_2, \ldots, b_\ell)$ of $m - r$ to the onefree composition $(r, b_1, b_2, \ldots, b_\ell)$ of m.

This is precisely what we needed to show. Thus, the induction step is complete, so that Proposition A.8.1 is proved. □

Our third proof of Proposition A.8.1 will be bijective (i.e., it will rely on a bijection). Crucial to this proof will be the following combinatorial interpretation of Fibonacci numbers:

Proposition A.8.2. Let $(f_0, f_1, f_2, \ldots)$ be the Fibonacci sequence. Let $n \in \mathbb{N}$ and $a \in \mathbb{Z}$. Then,

$$(\#\text{ of lacunar subsets of } [a+1, a+n]) = f_{n+2},$$

where $[a+1, a+n]$ denotes the n-element set $\{a+1, a+2, \ldots, a+n\}$.

Hints to the proof of Proposition A.8.2. (See [19fco, proof of Proposition 1.4.18] for details.) Proposition A.8.2 can easily be derived from Theorem 2.3.4. Indeed, the set $[a+1, a+n]$ (where we are using the notations of Proposition A.8.2) is just the set $[n]$ shifted to the right by a units (if we visualize these sets on the real axis); in more formal language, there is a bijection from $[n]$ to $[a+1, a+n]$ that sends each $i \in [n]$ to $a+i$. This bijection gives rise to a bijection from $\{\text{lacunar subsets of } [n]\}$ to $\{\text{lacunar subsets of } [a+1, a+n]\}$ (namely, we apply the same shift to each element of our lacunar subset). Thus, the bijection principle yields

$$\begin{aligned}&(\#\text{ of lacunar subsets of } [a+1, a+n]) \\ &= (\#\text{ of lacunar subsets of } [n]) = f_{n+2} \qquad (\text{by Theorem 2.3.4}).\end{aligned}$$

This proves Proposition A.8.2. □

Third proof of Proposition A.8.1 (sketched). Let n be a positive integer. We WLOG assume that $n \geq 3$ (since the other cases can easily be checked by hand). Thus, $n-3 \in \mathbb{N}$.

Consider the map

$$C : \{\text{compositions of } n\} \to \{\text{subsets of } [n-1]\}$$

defined in the proof of Theorem 7.3.4. We recall (from said proof) that this map C is a bijection; thus, it has an inverse map C^{-1}.

Now, the following two claims are crucial:[431]

> *Claim 1:* Let a be a onefree composition of n. Then, $C(a)$ is a lacunar subset of $[2, n-2]$.
>
> *Claim 2:* Let U be a lacunar subset of $[2, n-2]$. Then, $C^{-1}(U)$ is a onefree composition of n.

[431] In the following, the notation $[p, q]$ for two integers p and q shall denote the set $\{p, p+1, \ldots, q\} \subseteq \mathbb{Z}$.

[*Proof of Claim 1:* The composition a is onefree; in other words, it does not contain 1 as an entry (by the definition of "onefree").

Write the composition a in the form $a = (a_1, a_2, \ldots, a_k)$. Then, $(a_1, a_2, \ldots, a_k)$ is a composition of n. In other words, $(a_1, a_2, \ldots, a_k)$ is a tuple of positive integers whose sum is n. Hence, the numbers $a_1, a_2, \ldots, a_k$ are positive integers and satisfy $a_1 + a_2 + \cdots + a_k = n$.

The numbers $a_1, a_2, \ldots, a_k$ are the entries of the composition a (since $a = (a_1, a_2, \ldots, a_k)$), and therefore are distinct from 1 (since this composition a does not contain 1 as an entry). Hence, these numbers $a_1, a_2, \ldots, a_k$ are ≥ 2 (since they are positive integers).

Now, the definition of $C(a)$ yields

$$C(a) = \{a_1 + a_2 + \cdots + a_i \mid i \in \{1, 2, \ldots, k-1\}\}.$$

Hence, it is easy to see that $C(a)$ is a subset of $[2, n-2]$ [432]. Furthermore, the set $C(a)$ is lacunar[433]. This completes the proof of Claim 1.]

---

[432] *Proof.* Let $p \in C(a)$. We shall show that $p \in [2, n-2]$.

We have $p \in C(a) = \{a_1 + a_2 + \cdots + a_i \mid i \in \{1, 2, \ldots, k-1\}\}$. In other words, $p = a_1 + a_2 + \cdots + a_i$ for some $i \in \{1, 2, \ldots, k-1\}$. Consider this i.

We have $i \geq 1$ (since $i \in \{1, 2, \ldots, k-1\}$). Thus, a_1 is an addend of the sum $a_1 + a_2 + \cdots + a_i$. But $a_1 + a_2 + \cdots + a_i$ is a sum of positive integers (since $a_1, a_2, \ldots, a_k$ are positive integers), and thus is $\geq$ to any of its addends. Hence, $a_1 + a_2 + \cdots + a_i \geq a_1$ (since a_1 is an addend of this sum). Now, $p = a_1 + a_2 + \cdots + a_i \geq a_1 \geq 2$ (since the numbers $a_1, a_2, \ldots, a_k$ are ≥ 2).

We have $i \leq k-1$ (since $i \in \{1, 2, \ldots, k-1\}$). Thus, the sum $a_1 + a_2 + \cdots + a_{k-1}$ contains all i addends of the sum $a_1 + a_2 + \cdots + a_i$ and possibly a few more. Since all these extra addends are positive (because $a_1, a_2, \ldots, a_k$ are positive integers), we thus conclude that $a_1 + a_2 + \cdots + a_{k-1} \geq a_1 + a_2 + \cdots + a_i$ (because a sum can only get larger if we add some extra positive addends). Hence,

$$\begin{aligned} a_1 + a_2 + \cdots + a_i &\leq a_1 + a_2 + \cdots + a_{k-1} \\ &= \underbrace{(a_1 + a_2 + \cdots + a_k)}_{=n} - \underbrace{a_k}_{\substack{\geq 2 \\ \text{(since the numbers } a_1, a_2, \ldots, a_k \text{ are } \geq 2)}} \\ &\leq n - 2. \end{aligned}$$

Hence, $p = a_1 + a_2 + \cdots + a_i \leq n - 2$.

Combining $p \geq 2$ with $p \leq n-2$, we obtain $p \in [2, n-2]$.

Forget that we fixed p. We thus have shown that $p \in [2, n-2]$ for each $p \in C(a)$. Hence, $C(a)$ is a subset of $[2, n-2]$.

[433] *Proof.* Assume the contrary. Thus, the set $C(a)$ is not lacunar; in other words, the set $C(a)$ contains two consecutive integers (by the definition of "lacunar"). Let p and q be these two consecutive integers (in some order). Thus, $p \in C(a)$ and $q \in C(a)$.

We have $p \in C(a) = \{a_1 + a_2 + \cdots + a_i \mid i \in \{1, 2, \ldots, k-1\}\}$. In other words, there exists some $u \in \{1, 2, \ldots, k-1\}$ such that $p = a_1 + a_2 + \cdots + a_u$. Consider this u.

We have $q \in C(a) = \{a_1 + a_2 + \cdots + a_i \mid i \in \{1, 2, \ldots, k-1\}\}$. In other words, there exists some $v \in \{1, 2, \ldots, k-1\}$ such that $q = a_1 + a_2 + \cdots + a_v$. Consider this v.

We WLOG assume that $u \leq v$ (since otherwise, we can just swap u with v, while simultaneously swapping p with q). Note that $p \neq q$ (since p and q are consecutive integers), so that $u \neq v$ (because if we had $u = v$, then we would have $a_1 + a_2 + \cdots + a_u = a_1 + a_2 + \cdots + a_v$, which would contradict $a_1 + a_2 + \cdots + a_u = p \neq q = a_1 + a_2 + \cdots + a_v$). Combining this with $u \leq v$,

[*Proof of Claim 2:* I shall be somewhat brief and handwavy, but I trust the reader to be able to formalize this argument by now.

The set U is a subset of $[2, n-2]$ and thus a subset of $[n-1]$ (since $[2, n-2] \subseteq [n-1]$). Hence, $C^{-1}(U)$ is a well-defined composition of n.

Recall how the inverse map C^{-1} of C was described in the proof of Theorem 7.3.4: If I is any subset of $[n-1]$, then the elements of I subdivide the interval $[0, n]$ into several blocks; the composition $C^{-1}(I)$ is formed by the lengths of these blocks (from left to right).

Applying this to $I = U$, we see that the composition $C^{-1}(U)$ is formed by the lengths of the blocks into which the elements of U subdivide the interval $[0, n]$. [434] We shall show that each of these blocks has length ≥ 2.

Indeed, this is obvious if $U = \varnothing$ (because in this case, there is only one block (namely, $[0, n]$), and it has length $n \geq 3 \geq 2$). Hence, for the rest of this proof, we WLOG assume that $U \neq \varnothing$; thus, the set U has a smallest element $\min U$ and a largest element $\max U$. Note that $\min U \in U \subseteq [2, n-2]$, so that $\min U \geq 2$. Also, $\max U \in U \subseteq [2, n-2]$ and thus $\max U \leq n-2$. Note that the set U is lacunar; thus, any two distinct elements of U are at least a distance of 2 apart (on the real axis).

Now, consider the blocks into which the elements of U subdivide the interval $[0, n]$. The first of these blocks is $[0, \min U]$, and thus has length $\min U - 0 =$

---

we obtain $u < v$. Now,

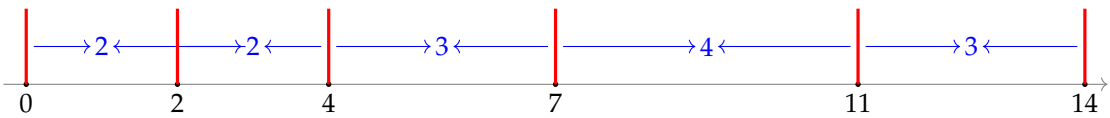
$$q = a_1 + a_2 + \cdots + a_v = \underbrace{(a_1 + a_2 + \cdots + a_u)}_{=p} + (a_{u+1} + a_{u+2} + \cdots + a_v) \qquad (\text{since } u \leq v)$$
$$= p + (a_{u+1} + a_{u+2} + \cdots + a_v).$$

But a_v is an addend of the sum $a_{u+1} + a_{u+2} + \cdots + a_v$ (since $u < v$). However, $a_{u+1} + a_{u+2} + \cdots + a_v$ is a sum of positive integers (since $a_1, a_2, \ldots, a_k$ are positive integers), and thus is $\geq$ to any of its addends. Hence, $a_{u+1} + a_{u+2} + \cdots + a_v \geq a_v$ (since a_v is an addend of this sum). Also, $a_v \geq 2$ (since the numbers $a_1, a_2, \ldots, a_k$ are ≥ 2). Now,

$$q = p + \underbrace{(a_{u+1} + a_{u+2} + \cdots + a_v)}_{\geq a_v \geq 2} \geq p + 2.$$

That is, the integer q is at least by 2 larger than p. Hence, p and q cannot be consecutive integers. This contradicts the fact that p and q are consecutive integers. This contradiction shows that our assumption was false. Qed.

[434] Let us again illustrate this on an example, this time making sure to pick a lacunar subset of $[2, n-2]$. Namely, set $n = 14$ and $U = \{2, 4, 7, 11\}$. Then, the elements of U subdivide the interval $[0, n] = [0, 14]$ into 5 blocks as follows:



The lengths of these blocks are $2, 2, 3, 4, 3$ (from left to right). Thus, the composition $C^{-1}(U)$ is $(2, 2, 3, 4, 3)$.

$\min U \geq 2$. The last of these blocks is $[\max U, n]$, and thus has length $n - \underbrace{\max U}_{\leq n-2} \geq n - (n-2) = 2$. Each of the remaining blocks begins and ends at two (distinct) elements of U, and thus has length ≥ 2 (since any two distinct elements of U are at least a distance of 2 apart (on the real axis)). Combining these results, we conclude that each of our blocks has length ≥ 2. Hence, none of our blocks has length 1.

Now, recall that the composition $C^{-1}(U)$ is formed by the lengths of these blocks. Since none of our blocks has length 1, we thus conclude that the composition $C^{-1}(U)$ does not contain 1 as an entry. In other words, the composition $C^{-1}(U)$ is onefree. This proves Claim 2.]

Now, Claim 1 shows that we can define a map

$$\begin{aligned} \overline{C} : \{\text{onefree compositions of } n\} &\to \{\text{lacunar subsets of } [2, n-2]\}, \\ a &\mapsto C(a). \end{aligned}$$

Claim 2 shows that we can define a map

$$\begin{aligned} \overline{D} : \{\text{lacunar subsets of } [2, n-2]\} &\to \{\text{onefree compositions of } n\}, \\ U &\mapsto C^{-1}(U). \end{aligned}$$

These two maps $\overline{C}$ and $\overline{D}$ are restrictions of the maps C and C^{-1} (respectively), and thus are mutually inverse (since the maps C and C^{-1} are mutually inverse). Hence, the map $\overline{C}$ is invertible, i.e., is a bijection. The bijection principle thus yields

$$\begin{aligned} &|\{\text{onefree compositions of } n\}| \\ &= |\{\text{lacunar subsets of } [2, n-2]\}| \\ &= \left(\# \text{ of lacunar subsets of } \left[\underbrace{2}_{=1+1}, \underbrace{n-2}_{=1+(n-3)} \right] \right) \\ &= (\# \text{ of lacunar subsets of } [1+1, 1+(n-3)]) \\ &= f_{(n-3)+2} \qquad (\text{by Proposition A.8.2 (applied to 1 and } n-3 \text{ instead of } a \text{ and } n)) \\ &= f_{n-1}. \end{aligned}$$

In other words, $(\# \text{ of onefree compositions of } n) = f_{n-1}$. This proves Proposition A.8.1 again. □

□

A.8.5. Discussion of Exercise 7.7.5

Discussion of Exercise 7.7.5. Exercise 7.7.5 is a well-known problem, appearing in some form in almost any text on combinatorics. See, for instance, [YagYag64, Problem 31], [Stanle11, §1.2] or [AndFen04, Theorem 4.2]. We note that Exercise 7.7.5 **(b)** is commonly used in commutative algebra, since the weak compositions of n into

k parts are in bijection with the degree-n monomials in k variables $x_1, x_2, \ldots, x_k$ (namely, a weak composition $(a_1, a_2, \ldots, a_k)$ of n into k parts corresponds to the monomial $x_1^{a_1} x_2^{a_2} \cdots x_k^{a_k}$).

(a) We claim that

$$(\text{\# of compositions of } n \text{ into } k \text{ parts}) = \binom{n-1}{n-k} = \binom{n-1}{k-1}.$$

Even more generally, we can allow n to be 0 and get the following:

Theorem A.8.3. Let $n \in \mathbb{N}$ and $k \in \mathbb{N}$. Then,

$$\begin{aligned} &(\text{\# of compositions of } n \text{ into } k \text{ parts}) \\ &= \binom{n-1}{n-k} \qquad (738) \\ &= \begin{cases} \binom{n-1}{k-1}, & \text{if } n > 0; \\ [k=0], & \text{if } n = 0. \end{cases} \qquad (739) \end{aligned}$$

We refer to [19fco, §2.10.1] for two proofs of Theorem A.8.3. Here, we shall just briefly outline one of the proofs:

Hints to the proof of Theorem A.8.3. The case when $n = 0$ is straightforward busywork (and is not part of Exercise 7.7.5 anyway), so we WLOG assume that $n \neq 0$. Thus, n is a positive integer.

Likewise, we leave the case $k = 0$ to the reader (again, noting that this case is straightforward), so we WLOG assume that $k \neq 0$. Hence, k is a positive integer.

Consider the map

$$C : \{\text{compositions of } n\} \to \{\text{subsets of } [n-1]\}$$

defined in the proof of Theorem 7.3.4. We recall (from said proof) that this map C is a bijection; thus, it has an inverse map C^{-1}.

Now, the following two claims are crucial:

> *Claim 1:* Let a be a composition of n into k parts. Then, $C(a)$ is a $(k-1)$-element subset of $[n-1]$.
>
> *Claim 2:* Let U be a $(k-1)$-element subset of $[n-1]$. Then, $C^{-1}(U)$ is a composition of n into k parts.

The reader will have no trouble proving these two claims. Now, Claim 1 shows that we can define a map

$$\begin{aligned} \overline{C} : \{\text{compositions of } n \text{ into } k \text{ parts}\} &\to \{(k-1)\text{-element subsets of } [n-1]\}, \\ a &\mapsto C(a). \end{aligned}$$

Claim 2 shows that we can define a map

$$\overline{D} : \{(k-1)\text{-element subsets of } [n-1]\} \to \{\text{compositions of } n \text{ into } k \text{ parts}\}, \\ U \mapsto C^{-1}(U).$$

These two maps $\overline{C}$ and $\overline{D}$ are restrictions of the maps C and C^{-1} (respectively), and thus are mutually inverse. Hence, the map $\overline{C}$ is invertible, i.e., is a bijection. The bijection principle thus yields

$$\begin{aligned}
&|\{\text{compositions of } n \text{ into } k \text{ parts}\}| \\
&= |\{(k-1)\text{-element subsets of } [n-1]\}| \\
&= (\#\text{ of } (k-1)\text{-element subsets of } [n-1]) \\
&= \binom{n-1}{k-1} \qquad \text{(by an application of Theorem 4.3.12)} \\
&= \binom{n-1}{(n-1)-(k-1)} \qquad \text{(by an application of Theorem 4.3.10)} \\
&= \binom{n-1}{n-k}.
\end{aligned}$$

Theorem A.8.3 follows. □

(b) We claim that

$$\begin{aligned}
&(\#\text{ of weak compositions of } n \text{ into } k \text{ parts}) \\
&= \binom{n+k-1}{n} = \begin{cases} \dbinom{n+k-1}{k-1}, & \text{if } k > 0; \\ 0, & \text{if } k = 0. \end{cases}
\end{aligned}$$

Even more generally, we can allow n to be 0 and get the following:

Theorem A.8.4. Let $n \in \mathbb{N}$ and $k \in \mathbb{N}$. Then,

$$\begin{aligned}
&(\#\text{ of weak compositions of } n \text{ into } k \text{ parts}) \\
&= \binom{n+k-1}{n} \qquad (740) \\
&= \begin{cases} \dbinom{n+k-1}{k-1}, & \text{if } k > 0; \\ [n=0], & \text{if } k = 0. \end{cases} \qquad (741)
\end{aligned}$$

We refer to [19fco, §2.10.3] for a detailed proof of Theorem A.8.4. Here, let us just briefly outline it:

Hints to the proof of Theorem A.8.4. If $(a_1, a_2, \ldots, a_k)$ is a weak composition of n into k parts, then $(a_1+1, a_2+1, \ldots, a_k+1)$ is a composition of $n+k$ into k parts[435]. Thus, there is a map

$$\begin{aligned}\{\text{weak compositions of } n \text{ into } k \text{ parts}\} &\to \{\text{compositions of } n+k \text{ into } k \text{ parts}\}, \\ (a_1, a_2, \ldots, a_k) &\mapsto (a_1+1, a_2+1, \ldots, a_k+1).\end{aligned}$$

It is easy to see that this map is a bijection. The bijection principle therefore shows that

$$\begin{aligned}&(\#\text{ of weak compositions of } n \text{ into } k \text{ parts}) \\ &= (\#\text{ of compositions of } n+k \text{ into } k \text{ parts}) \\ &= \binom{n+k-1}{n+k-k} \qquad (\text{by (738) (applied to } n+k \text{ instead of } n)) \\ &= \binom{n+k-1}{n} = \begin{cases} \dbinom{n+k-1}{k-1}, & \text{if } k > 0; \\ [n=0], & \text{if } k = 0 \end{cases}\end{aligned}$$

(where the last equality sign follows from an application of Theorem 4.3.10 in the case when $k > 0$, and can be checked by hand in the case when $k = 0$). Theorem A.8.4 thus follows. □

The equality (740) can also be proved by induction (see [Jarvin20, §19.3] or [Grinbe15, Exercise 3.15] for such proofs) or by a direct counting argument (see [AndFen04, proof of Theorem 4.2 (b)] or [Jarvin20, §19.3] for this). □

A.8.6. Discussion of Exercise 7.7.6

Discussion of Exercise 7.7.6. Exercise 7.7.6 is [GrKnPa94, Exercise 5.65] or [Tomesc85, Problem 1.14]; I have also seen it called *Riordan's identity*. We first give an algebraic solution (using the telescope principle), then a combinatorial solution (via double counting).

First solution to Exercise 7.7.6. For each $k \in \{0, 1, \ldots, n\}$, we define a number

$$b_k = n\binom{n-1}{k}\frac{k!}{n^k}. \tag{742}$$

[435] because adding 1 to a nonnegative integer yields a positive integer, and because

$$(a_1+1)+(a_2+1)+\cdots+(a_k+1) = \underbrace{(a_1+a_2+\cdots+a_k)}_{=n} + \underbrace{\underbrace{(1+1+\cdots+1)}_{k \text{ times}}}_{=k} = n+k$$

Now, we claim that

$$\binom{n-1}{k}\frac{(k+1)!}{n^k} = b_k - b_{k+1} \tag{743}$$

for each $k \in \{0, 1, \ldots, n-1\}$.

[*Proof of (743):* Let $k \in \{0, 1, \ldots, n-1\}$. The definition of b_{k+1} yields

$$\begin{aligned} b_{k+1} &= n\binom{n-1}{k+1}\frac{(k+1)!}{n^{k+1}} = n\binom{n-1}{k+1}\frac{(k+1)\cdot k!}{n^k \cdot n} \\ &\qquad \left(\text{since } (k+1)! = (k+1)\cdot k! \text{ and } n^{k+1} = n^k \cdot n\right) \\ &= \underbrace{(k+1)\binom{n-1}{k+1}}_{\substack{=((n-1)-k)\binom{n-1}{k} \\ \text{(by Lemma A.4.9 } \textbf{(b)} \\ \text{(applied to } n-1 \text{ instead of } n))}} \frac{k!}{n^k} = ((n-1)-k)\binom{n-1}{k}\frac{k!}{n^k}. \end{aligned}$$

Subtracting this equality from (742), we find

$$\begin{aligned} b_k - b_{k+1} &= n\binom{n-1}{k}\frac{k!}{n^k} - ((n-1)-k)\binom{n-1}{k}\frac{k!}{n^k} \\ &= \underbrace{(n-((n-1)-k))}_{=k+1}\binom{n-1}{k}\frac{k!}{n^k} = (k+1)\binom{n-1}{k}\frac{k!}{n^k} \\ &= \binom{n-1}{k}\frac{(k+1)\cdot k!}{n^k} = \binom{n-1}{k}\frac{(k+1)!}{n^k} \end{aligned}$$

(since $(k+1)\cdot k! = (k+1)!$). This proves (743).]

Now,

$$\begin{aligned} &\sum_{k=0}^{n-1}\underbrace{\binom{n-1}{k}\frac{(k+1)!}{n^k}}_{\substack{=b_k - b_{k+1} \\ \text{(by (743))}}} \\ &= \sum_{k=0}^{n-1}\underbrace{(b_k - b_{k+1})}_{=(-b_{k+1})-(-b_k)} = \sum_{k=0}^{n-1}((-b_{k+1})-(-b_k)) = \sum_{i=0}^{n-1}((-b_{i+1})-(-b_i)) \\ &\qquad \text{(here, we have renamed the summation index } k \text{ as } i) \\ &= \left(-b_{(n-1)+1}\right) - (-b_0) \\ &\qquad \text{(by Corollary 4.1.17 (applied to } u = 0,\ v = n-1 \text{ and } a_i = -b_i)) \\ &= b_0 - b_{(n-1)+1} \\ &= b_0 - b_n \qquad (\text{since } (n-1)+1 = n). \end{aligned} \tag{744}$$

However, the definition of b_0 yields

$$b_0 = n \underbrace{\binom{n-1}{0}}_{\substack{=1 \\ \text{(by (119)} \\ \text{(applied to } n-1 \text{ instead of } n))}} \frac{0!}{n^0} = n \cdot \frac{0!}{n^0} = n \cdot \underbrace{0!}_{=1} / \underbrace{n^0}_{=1} = n.$$

Furthermore, it is easy to see that $b_n = 0$ [436]. Hence, (744) becomes

$$\sum_{k=0}^{n-1} \binom{n-1}{k} \frac{(k+1)!}{n^k} = \underbrace{b_0}_{=n} - \underbrace{b_n}_{=0} = n.$$

This solves Exercise 7.7.6. □

Second solution to Exercise 7.7.6 (sketched). It is easy to solve Exercise 7.7.6 in the case when $n = 0$ (indeed, it boils down to $0 = 0$ in this case). Thus, for the rest of this solution, we WLOG assume that $n \neq 0$. Hence, $n \geq 1$.

If $a = (a_1, a_2, \ldots, a_n) \in [n]^n$ is an n-tuple, then there exists an $i \in [n]$ such that the i entries $a_1, a_2, \ldots, a_i$ are distinct (indeed, $i = 1$ always fits the bill, since the single entry a_1 is distinct). We define the *wit* of an n-tuple $a = (a_1, a_2, \ldots, a_n) \in [n]^n$ to be the largest $i \in [n]$ such that the i entries $a_1, a_2, \ldots, a_i$ are distinct. Thus, roughly speaking, the wit of an n-tuple measures "how long the n-tuple can avoid repeating its own entries". For example, the 9-tuple $(3, 1, 6, 2, 4, 1, 5, 2)$ has wit 5 (since its first 5 entries $3, 1, 6, 2, 4$ are distinct, while its first 6 entries $3, 1, 6, 2, 4, 1$ are not). Likewise, the 7-tuple $(3, 1, 3, 2, 4, 5, 3)$ has wit 2.

For any $k \in [n]$, we have

$$\begin{aligned} &\left(\# \text{ of } n\text{-tuples } a \in [n]^n \text{ with wit } k\right) \\ &= \binom{n-1}{k-1} k! \cdot n^{n-k}. \end{aligned} \tag{745}$$

[*Proof of (745):* Let $k \in [n]$. We need to prove the equality (745).

We have $k \leq n$ (since $k \in [n]$). Thus, we are in one of the two cases $k < n$ and $k = n$. We shall only handle the case $k < n$ in the following; the case $k = n$ is similar (but easier) and we leave it to the reader.

[436] *Proof.* If $n = 0$, then $b_n = b_0 = n = 0$. Thus, for the rest of this proof of $b_n = 0$, we WLOG assume that $n \neq 0$. Hence, n is a positive integer (since $n \in \mathbb{N}$), so that $n - 1 \in \mathbb{N}$. Therefore, Proposition 4.3.4 (applied to $n - 1$ and n instead of n and k) yields $\binom{n-1}{n} = 0$ (since $n > n - 1$). Now, the definition of b_n yields

$$b_n = n \underbrace{\binom{n-1}{n}}_{=0} \frac{n!}{n^n} = 0.$$

Qed.

So let us assume that $k < n$. An n-tuple $a = (a_1, a_2, \ldots, a_n) \in [n]^n$ has wit k if and only if its first k entries $a_1, a_2, \ldots, a_k$ are distinct while its $(k+1)$-st entry a_{k+1} equals one of these first k entries[437]. Thus, the following decision procedure can be used to construct an n-tuple $a = (a_1, a_2, \ldots, a_n) \in [n]^n$ with wit k:

- First, we choose the first k entries $a_1, a_2, \ldots, a_k$ of a, one after the other (keeping in mind that they have to be distinct). This procedure involves k decisions, and it is easy to count how many options there are in each decision: There are n options for choosing a_1 (since a_1 has to be an element of $[n]$); there are $n-1$ options for choosing a_2 (since a_2 has to be an element of $[n]$ distinct from a_1); there are $n-2$ options for choosing a_3 (since a_3 has to be an element of $[n]$ distinct from both the two distinct elements a_1 and a_2); and so on.

  After these k decisions, the first k entries $a_1, a_2, \ldots, a_k$ of our n-tuple a have been chosen.

- Next, we choose the $(k+1)$-st entry a_{k+1} of a. This entry has to equal one of the first k entries $a_1, a_2, \ldots, a_k$. Since these first k entries $a_1, a_2, \ldots, a_k$ are distinct, we thus have k options for choosing a_{k+1}.

- Finally, we choose the remaining $n-k-1$ entries $a_{k+2}, a_{k+3}, \ldots, a_n$ of a. These entries can be arbitrary elements of $[n]$ (they need not be distinct from anything or equal to anything); thus, there are n options for choosing each of them.

The dependent product rule shows that the total # of possibilities for making these choices is

$$\begin{aligned} & n(n-1)(n-2)\cdots(n-k+1)\cdot k\cdot \underbrace{nn\cdots n}_{n-k-1 \text{ times}} \\ &= n(n-1)(n-2)\cdots(n-k+1)\cdot k\cdot n^{n-k-1}. \end{aligned}$$

Hence,

$$\begin{aligned} & \left(\#\text{ of } n\text{-tuples } a \in [n]^n \text{ with wit } k\right) \\ &= n(n-1)(n-2)\cdots(n-k+1)\cdot k\cdot n^{n-k-1}. \end{aligned} \tag{746}$$

[437] This is where we need $k < n$: If we had $k = n$, then there would be no $(k+1)$-st entry a_{k+1}.

On the other hand, from $k \geq 1$, we have $k-1 \in \mathbb{N}$ and thus

$$\begin{aligned}
&\underbrace{\binom{n-1}{k-1}}_{\substack{=\frac{(n-1)(n-2)(n-3)\cdots((n-1)-(k-1)+1)}{(k-1)!} \\ \text{(by (117) (applied to } n-1 \text{ and } k-1 \text{ instead of } n \text{ and } k))}} \underbrace{k!}_{=k\cdot(k-1)!} \cdot \underbrace{n^{n-k}}_{=n\cdot n^{n-k-1}} \\
&= \frac{(n-1)(n-2)(n-3)\cdots((n-1)-(k-1)+1)}{(k-1)!} \cdot k(k-1)! \cdot n \cdot n^{n-k-1} \\
&= (n-1)(n-2)(n-3)\cdots \underbrace{((n-1)-(k-1)+1)}_{=n-k+1} \cdot kn \cdot n^{n-k-1} \\
&= (n-1)(n-2)(n-3)\cdots(n-k+1) \cdot kn \cdot n^{n-k-1} \\
&= \underbrace{n \cdot (n-1)(n-2)(n-3)\cdots(n-k+1)}_{=n(n-1)(n-2)\cdots(n-k+1)} \cdot k \cdot n^{n-k-1} \\
&= n(n-1)(n-2)\cdots(n-k+1) \cdot k \cdot n^{n-k-1}.
\end{aligned}$$

Comparing this with (746), we find

$$\left(\# \text{ of } n\text{-tuples } a \in [n]^n \text{ with wit } k\right) = \binom{n-1}{k-1} k! \cdot n^{n-k}.$$

This proves (745).]

Now, the wit of an n-tuple $a \in [n]^n$ is an element of $[n]$. Hence, the sum rule yields

$$\begin{aligned}
&\left(\# \text{ of } n\text{-tuples } a \in [n]^n\right) \\
&= \sum_{k=1}^{n} \underbrace{\left(\# \text{ of } n\text{-tuples } a \in [n]^n \text{ with wit } k\right)}_{\substack{=\binom{n-1}{k-1} k! \cdot n^{n-k} \\ \text{(by (745))}}} \\
&= \sum_{k=1}^{n} \binom{n-1}{k-1} k! \cdot n^{n-k} = \sum_{k=0}^{n-1} \underbrace{\binom{n-1}{(k+1)-1}}_{=\binom{n-1}{k}} (k+1)! \cdot \underbrace{n^{n-(k+1)}}_{=n^{n-1-k}=\frac{n^{n-1}}{n^k}} \\
&\qquad \text{(here, we have substituted } k+1 \text{ for } k \text{ in the sum)} \\
&= \sum_{k=0}^{n-1} \binom{n-1}{k} (k+1)! \cdot \frac{n^{n-1}}{n^k} = n^{n-1} \cdot \sum_{k=0}^{n-1} \binom{n-1}{k} \frac{(k+1)!}{n^k}.
\end{aligned}$$

Thus,

$$n^{n-1} \cdot \sum_{k=0}^{n-1} \binom{n-1}{k} \frac{(k+1)!}{n^k} = \left(\# \text{ of } n\text{-tuples } a \in [n]^n\right) = \left|[n]^n\right| = n^n.$$

Dividing both sides of this equality by n^{n-1}, we obtain

$$\sum_{k=0}^{n-1} \binom{n-1}{k} \frac{(k+1)!}{n^k} = n.$$

Thus, Exercise 7.7.6 is solved again. □

□

A.8.7. Discussion of Exercise 7.7.7

Discussion of Exercise 7.7.7. Exercise 7.7.7 is part of [Riorda68, Problem 1.10]; it also appears implicitly in [Shapir76, Proposition 3.1] (since it is easily seen that $\frac{1}{2}\binom{2n}{n} = \binom{2n-1}{n}$ when n is a positive integer). We shall give a purely algebraic solution:

Solution to Exercise 7.7.7. For each $k \in \{0, 1, \ldots, n\}$, we have

$$\begin{aligned}
&\underbrace{k}_{=n-(n-k)} \binom{2n}{n-k} \\
&= (n-(n-k)) \binom{2n}{n-k} \\
&= n \underbrace{\binom{2n}{n-k}}_{\substack{=\binom{2n-1}{n-k-1}+\binom{2n-1}{n-k} \\ \text{(by Theorem 4.3.7} \\ \text{(applied to } 2n \text{ and } n-k \text{ instead of } n \text{ and } k))}} - \underbrace{(n-k)\binom{2n}{n-k}}_{\substack{=2n\binom{2n-1}{n-k-1} \\ \text{(since Lemma A.4.9 \textbf{(a)}} \\ \text{(applied to } 2n \text{ and } n-k \text{ instead of } n \text{ and } k) \\ \text{yields } 2n\binom{2n-1}{n-k-1}=(n-k)\binom{2n}{n-k})}} \\
&= n\left(\binom{2n-1}{n-k-1} + \binom{2n-1}{n-k}\right) - 2n\binom{2n-1}{n-k-1} \\
&= n\left(\binom{2n-1}{n-k} - \binom{2n-1}{n-k-1}\right) = n\left(\binom{2n-1}{n-k} - \binom{2n-1}{n-(k+1)}\right)
\end{aligned}$$

(since $n-k-1 = n-(k+1)$). Summing this equality over all $k \in \{0, 1, \ldots, n\}$, we obtain

$$\begin{aligned}
&\sum_{k=0}^{n} k\binom{2n}{n-k} \\
&= \sum_{k=0}^{n} n\left(\binom{2n-1}{n-k} - \binom{2n-1}{n-(k+1)}\right) \\
&= n\sum_{k=0}^{n}\left(\binom{2n-1}{n-k} - \binom{2n-1}{n-(k+1)}\right). \qquad (747)
\end{aligned}$$

Now,

$$\begin{aligned}
&\sum_{k=0}^{n} \underbrace{\left(\binom{2n-1}{n-k} - \binom{2n-1}{n-(k+1)}\right)}_{=\left(-\binom{2n-1}{n-(k+1)}\right)-\left(-\binom{2n-1}{n-k}\right)} \\
&= \sum_{k=0}^{n} \left(\left(-\binom{2n-1}{n-(k+1)}\right) - \left(-\binom{2n-1}{n-k}\right)\right) \\
&= \sum_{i=0}^{n} \left(\left(-\binom{2n-1}{n-(i+1)}\right) - \left(-\binom{2n-1}{n-i}\right)\right) \\
&\qquad \text{(here, we have renamed the summation index } k \text{ as } i\text{)} \\
&= \left(-\binom{2n-1}{n-(n+1)}\right) - \left(-\binom{2n-1}{n-0}\right) \\
&\qquad \left(\begin{array}{c} \text{by Corollary 4.1.17} \\ \text{(applied to } u=0 \text{ and } v=n \text{ and } a_i = -\binom{2n-1}{n-i}\text{)} \end{array}\right) \\
&= \binom{2n-1}{n-0} - \binom{2n-1}{n-(n+1)} = \binom{2n-1}{n} - \underbrace{\binom{2n-1}{-1}}_{\substack{=0 \\ \text{(by (118)} \\ \text{(since } -1 \notin \mathbb{N}\text{))}}} \\
&\qquad \text{(since } n-0=n \text{ and } n-(n+1)=-1\text{)} \\
&= \binom{2n-1}{n}.
\end{aligned}$$

Hence, (747) rewrites as

$$\sum_{k=0}^{n} k\binom{2n}{n-k} = n\binom{2n-1}{n}.$$

This solves Exercise 7.7.7. □

□

A.8.8. Discussion of Exercise 7.7.8

Discussion of Exercise 7.7.8. Exercise 7.7.8 is one of the facts known as the *binomial inversion formula*. An equivalent fact appears in [Grinbe15, Exercise 3.18 **(a)**] and in [17f-hw4s, Exercise 6][438]. For convenience, let me show an adaptation of the proof of [17f-hw4s, Exercise 6].

First, I will need the following lemma:

[438]The differences between [Grinbe15, Exercise 3.18 **(a)**], [17f-hw4s, Exercise 6] and our Exercise 7.7.8 are insubstantial: Our Exercise 7.7.8 is stated for two **infinite** sequences $(f_0, f_1, f_2, \ldots)$ and

Lemma A.8.5. Let $n \in \mathbb{N}$. Let $i \in \{0, 1, \ldots, n\}$. Then,

$$\sum_{j=i}^{n} (-1)^{j+i} \binom{n}{j} \binom{j}{i} = [i = n].$$

Proof of Lemma A.8.5. We have $i \in \{0, 1, \ldots, n\}$, so that $i \leq n$ and $i \in \mathbb{N}$. From $i \leq n$, we obtain $n - i \geq 0$. Thus, $n - i \in \mathbb{N}$.

Let $j \in \{i, i+1, \ldots, n\}$. Then, $j \geq i \geq 0$. Therefore, Theorem 7.5.2 (applied to $a = j$ and $b = i$) yields

$$\binom{n}{j} \binom{j}{i} = \binom{n}{i} \binom{n-i}{j-i}. \tag{748}$$

Also, $j + i \equiv j - i \bmod 2$ (since $(j+i) - (j-i) = 2i$ is even). Thus, $(-1)^{j+i} = (-1)^{j-i}$. Multiplying this equality by (748), we obtain

$$(-1)^{j+i} \binom{n}{j} \binom{j}{i} = (-1)^{j-i} \binom{n}{i} \binom{n-i}{j-i}. \tag{749}$$

Now, forget that we fixed j. We thus have proven the equality (749) for each $j \in \{i, i+1, \ldots, n\}$. Summing this equality over all $j \in \{i, i+1, \ldots, n\}$, we obtain

$$\begin{aligned}
\sum_{j=i}^{n} (-1)^{j+i} \binom{n}{j} \binom{j}{i} &= \sum_{j=i}^{n} (-1)^{j-i} \binom{n}{i} \binom{n-i}{j-i} = \sum_{k=0}^{n-i} (-1)^{k} \binom{n}{i} \binom{n-i}{k} \\
&\qquad \text{(here, we have substituted } k \text{ for } j-i \text{ in the sum)} \\
&= \binom{n}{i} \underbrace{\sum_{k=0}^{n-i} (-1)^{k} \binom{n-i}{k}}_{\substack{=[n-i=0] \\ \text{(by Proposition 4.3.18 (applied to } n-i \text{ instead of } n))}} \\
&= \binom{n}{i} [n - i = 0]. \qquad (750)
\end{aligned}$$

But it is easy to see that

$$\binom{n}{i} [n - i = 0] = [i = n] \tag{751}$$

[439]. Thus, (750) becomes

$$\sum_{j=i}^{n} (-1)^{j+i} \binom{n}{j} \binom{j}{i} = \binom{n}{i} [n - i = 0] = [i = n].$$

$(g_0, g_1, g_2, \ldots)$, whereas [Grinbe15, Exercise 3.18 **(a)**] and [17f-hw4s, Exercise 6] are stated for two **finite** sequences (i.e., tuples) $(f_0, f_1, \ldots, f_N)$ and $(g_0, g_1, \ldots, g_N)$ (which are called $(a_0, a_1, \ldots, a_N)$ and $(b_0, b_1, \ldots, b_N)$ in [17f-hw4s, Exercise 6]). The proofs given in [Grinbe15] and [17f-hw4s] for [Grinbe15, Exercise 3.18 **(a)**] and [17f-hw4s, Exercise 6] apply to Exercise 7.7.8 (once some obvious changes are made).

[439] *Proof of (751):* We are in one of the following two cases:

Case 1: We have $i \neq n$.

This proves Lemma A.8.5. □

We are now ready for the solution to Exercise 7.7.8:

Solution to Exercise 7.7.8. We have assumed that

$$g_n = \sum_{i=0}^{n} (-1)^i \binom{n}{i} f_i \tag{752}$$

for each $n \in \mathbb{N}$.

Case 2: We have $i = n$.

Let us consider Case 1 first. In this case, we have $i \neq n$. In other words, $n \neq i$. Hence, $n - i \neq 0$. Thus, $[n - i = 0] = 0$, so that $\binom{n}{i} \underbrace{[n - i = 0]}_{=0} = 0$. Comparing this with $[i = n] = 0$ (since $i \neq n$), we obtain $\binom{n}{i} [n - i = 0] = [i = n]$. Hence, (751) is proven in Case 1.

Now, let us consider Case 2. In this case, we have $i = n$. In other words, $n = i$. Thus, $n - i = 0$. Thus, $[n - i = 0] = 1$. Also, from $i = n$, we obtain $\binom{n}{i} = \binom{n}{n} = 1$ (by (124)). Hence, $\underbrace{\binom{n}{i}}_{=1} \underbrace{[n - i = 0]}_{=1} = 1$. Comparing this with $[i = n] = 1$ (since $i = n$), we obtain $\binom{n}{i} [n - i = 0] = [i = n]$. Hence, (751) is proven in Case 2.

We thus have proven (751) in each of the two Cases 1 and 2. Thus, (751) always holds.

Now, let $n \in \mathbb{N}$. Then,

$$\sum_{i=0}^{n} (-1)^i \binom{n}{i} g_i = \sum_{j=0}^{n} (-1)^j \binom{n}{j} \underbrace{g_j}_{\substack{= \sum_{i=0}^{j} (-1)^i \binom{j}{i} f_i \\ \text{(by (752), applied to } j \\ \text{instead of } n)}}$$

(here, we have renamed the summation index i as j)

$$= \sum_{j=0}^{n} (-1)^j \binom{n}{j} \sum_{i=0}^{j} (-1)^i \binom{j}{i} f_i$$

$$= \underbrace{\sum_{j=0}^{n} \sum_{i=0}^{j}}_{\substack{= \sum_{i=0}^{n} \sum_{j=i}^{n} \\ \text{(by (102))}}} (-1)^j \underbrace{\binom{n}{j} (-1)^i}_{= (-1)^i \binom{n}{j}} \binom{j}{i} f_i$$

$$= \sum_{i=0}^{n} \sum_{j=i}^{n} \underbrace{(-1)^j (-1)^i}_{= (-1)^{j+i}} \binom{n}{j} \binom{j}{i} f_i = \sum_{i=0}^{n} \sum_{j=i}^{n} (-1)^{j+i} \binom{n}{j} \binom{j}{i} f_i$$

$$= \sum_{i=0}^{n} \underbrace{\left(\sum_{j=i}^{n} (-1)^{j+i} \binom{n}{j} \binom{j}{i} \right)}_{\substack{= [i=n] \\ \text{(by Lemma A.8.5)}}} f_i$$

$$= \sum_{i=0}^{n} [i = n] f_i = \underbrace{[n = n]}_{\substack{=1 \\ \text{(since } n=n)}} f_n + \sum_{i=0}^{n-1} \underbrace{[i = n]}_{\substack{=0 \\ \text{(since } i \neq n \\ \text{(because } i \leq n-1 < n))}} f_i$$

$$\left(\begin{array}{c} \text{here, we have split off the addend for } i = n \\ \text{from the sum} \end{array} \right)$$

$$= f_n + \underbrace{\sum_{i=0}^{n-1} 0 f_i}_{=0} = f_n.$$

In other words, $f_n = \sum_{i=0}^{n} (-1)^i \binom{n}{i} g_i$. This solves Exercise 7.7.8. $\square$

$\square$

A.8.9. Discussion of Exercise 7.7.9

Discussion of Exercise 7.7.9. Exercise 7.7.9 is [18f-hw3s, Exercise 3]. Its part **(a)** also appears in [Grinbe15, Exercise 3.2 **(b)**] and in [GrKnPa94, (5.37)], whereas its part **(b)** also appears in [Grinbe15, Exercise 3.23 **(a)**], in [Spivey19, Identity 154] and in [GrKnPa94, (5.39)]. The solution I shall give below is copied almost verbatim from [18f-hw3s]. Note that the solution to part **(b)** is probably one of the nicest illustrations of the "substitute non-integer values into the Chu–Vandermonde identity, and simplify the result" strategy. Combinatorial solutions to Exercise 7.7.9 **(b)** also exist, but are much more complicated (some such proofs are sketched in [Sved84] as well as in the math.stackexchange discussions `https://math.stackexchange.com/questions/72367` and `https://math.stackexchange.com/a/360780`).

Solution to Exercise 7.7.9. **(a)** From $n \in \mathbb{N}$, we obtain $2n \geq n$ and $2n \in \mathbb{N}$. Hence, Theorem 4.3.8 (applied to $2n$ and n instead of n and k) yields

$$\begin{aligned}
\binom{2n}{n} &= \frac{(2n)!}{n! \cdot (2n-n)!} = \frac{(2n)!}{n! \cdot n!} = \frac{1}{n!n!} \underbrace{(2n)!}_{\substack{=1\cdot 2\cdot\cdots\cdot(2n) \\ =(1\cdot 3\cdot 5\cdot\cdots\cdot(2n-1))\cdot(2\cdot 4\cdot 6\cdot\cdots\cdot(2n)) \\ \text{(here, we have split the product into} \\ \text{the product of its odd factors and} \\ \text{the product of its even factors)}}} \\
&= \frac{1}{n!n!} \underbrace{(1\cdot 3\cdot 5\cdot\cdots\cdot(2n-1))}_{=\prod_{i=0}^{n-1}(2i+1)} \cdot \underbrace{(2\cdot 4\cdot 6\cdot\cdots\cdot(2n))}_{=\prod_{i=1}^{n}(2i)=2^n\prod_{i=1}^{n} i} \\
&= \frac{1}{n!n!}\left(\prod_{i=0}^{n-1}(2i+1)\right)\cdot 2^n \underbrace{\prod_{i=1}^{n} i}_{=n!} = \frac{1}{n!n!}\left(\prod_{i=0}^{n-1}(2i+1)\right)\cdot 2^n n! \\
&= \frac{2^n}{n!}\prod_{i=0}^{n-1}(2i+1).
\end{aligned}$$

Solving this equality for $\prod_{i=0}^{n-1}(2i+1)$, we obtain

$$\prod_{i=0}^{n-1}(2i+1) = \binom{2n}{n} \Big/ \frac{2^n}{n!} = \frac{n!}{2^n}\binom{2n}{n}. \tag{753}$$

For each $a \in \mathbb{Q}$, we have

$$\begin{aligned}
\binom{a}{n} &= \frac{a(a-1)(a-2)\cdots(a-n+1)}{n!} \qquad \left(\begin{array}{c}\text{by (117) (applied to } a \text{ and } n \\ \text{instead of } n \text{ and } k)\end{array}\right) \\
&= \frac{\prod_{i=0}^{n-1}(a-i)}{n!} = \frac{1}{n!}\prod_{i=0}^{n-1}(a-i).
\end{aligned}$$

Applying this to $a = -1/2$, we obtain

$$\begin{aligned}
\binom{-1/2}{n} &= \frac{1}{n!}\prod_{i=0}^{n-1}\underbrace{(-1/2-i)}_{=\frac{2i+1}{-2}} = \frac{1}{n!}\prod_{i=0}^{n-1}\frac{2i+1}{-2} = \frac{1}{n!}\cdot\frac{\prod_{i=0}^{n-1}(2i+1)}{(-2)^n} \\
&= \frac{1}{n!}\cdot\frac{1}{(-2)^n}\underbrace{\prod_{i=0}^{n-1}(2i+1)}_{\substack{=\frac{n!}{2^n}\binom{2n}{n} \\ \text{(by (753))}}} = \frac{1}{n!}\cdot\frac{1}{(-2)^n}\cdot\frac{n!}{2^n}\binom{2n}{n} = \underbrace{\frac{1}{(-2)^n\cdot 2^n}}_{=\left(\frac{-1}{4}\right)^n}\binom{2n}{n} \\
&= \left(\frac{-1}{4}\right)^n\binom{2n}{n}.
\end{aligned}$$

This solves Exercise 7.7.9 **(a)**.

(b) Theorem 7.5.3 (applied to $x = -1/2$ and $y = -1/2$) yields

$$\binom{(-1/2)+(-1/2)}{n} = \sum_{k=0}^{n}\binom{-1/2}{k}\binom{-1/2}{n-k}.$$

Comparing this with

$$\binom{(-1/2)+(-1/2)}{n} = \binom{-1}{n} = (-1)^n \qquad \text{(by (122), applied to } k = n\text{)},$$

we obtain

$$\begin{aligned}
(-1)^n &= \sum_{k=0}^{n}\underbrace{\binom{-1/2}{k}}_{\substack{=\left(\frac{-1}{4}\right)^k\binom{2k}{k} \\ \text{(by Exercise 7.7.9 \textbf{(a)}, applied to } k \text{ instead of } n\text{)}}}\underbrace{\binom{-1/2}{n-k}}_{\substack{=\left(\frac{-1}{4}\right)^{n-k}\binom{2(n-k)}{n-k} \\ \text{(by Exercise 7.7.9 \textbf{(a)}, applied to } n-k \text{ instead of } n\text{)}}} \\
&= \sum_{k=0}^{n}\underbrace{\left(\frac{-1}{4}\right)^k\left(\frac{-1}{4}\right)^{n-k}}_{=\left(\frac{-1}{4}\right)^n}\binom{2k}{k}\binom{2(n-k)}{n-k} = \left(\frac{-1}{4}\right)^n\sum_{k=0}^{n}\binom{2k}{k}\binom{2(n-k)}{n-k}.
\end{aligned}$$

Multiplying both sides of this equality by $(-4)^n$, we obtain

$$(-4)^n(-1)^n = \underbrace{(-4)^n\left(\frac{-1}{4}\right)^n}_{=1}\sum_{k=0}^{n}\binom{2k}{k}\binom{2(n-k)}{n-k} = \sum_{k=0}^{n}\binom{2k}{k}\binom{2(n-k)}{n-k}.$$

Hence,

$$\sum_{k=0}^{n} \binom{2k}{k} \binom{2(n-k)}{n-k} = (-4)^n (-1)^n = 4^n.$$

This solves Exercise 7.7.9 **(b)**. □

We end this discussion by mentioning two binomial identities similar to Exercise 7.7.9 **(b)**:

- For each $n \in \mathbb{N}$, we have

$$\sum_{k=0}^{n} (-1)^k \binom{2k}{k} \binom{2(n-k)}{n-k} = \begin{cases} 2^n \binom{n}{n/2}, & \text{if } n \text{ is even;} \\ 0, & \text{if } n \text{ is odd.} \end{cases}$$

  This appears in [Grinbe15, Exercise 3.23 **(b)**].

- For each $n \in \mathbb{N}$ and $u \in \mathbb{C}$, we have

$$\sum_{k=0}^{n} \binom{2k-u}{k} \binom{2(n-k)+u}{n-k} = 4^n. \tag{754}$$

  This is a result of Duarte and De Oliveira ([DuaOli13, Theorem]). I give a detailed proof in [19fco-mt2s, Exercise 1] (where I require $u \in \mathbb{R}$ instead of $u \in \mathbb{C}$, but this makes no difference). Note that Exercise 7.7.9 **(b)** is the particular case of this identity (754) for $u = 0$. However, our above solution to Exercise 7.7.9 **(b)** does not appear to adapt to the more general identity (754).

□

A.8.10. Discussion of Exercise 7.7.10

Discussion of Exercise 7.7.10. Exercise 7.7.10 was on the IMO Longlist 1983 (which means it was proposed for the IMO in 1983 but did not get into the shortlist). To us, it is an easy consequence of Exercise 7.5.2. For convenience, we front-load part of our solution as a self-contained lemma:

Lemma A.8.6. Let x and y be two positive integers such that $x^2 - x = y^2 - y$. Then, $x = y$.

Proof of Lemma A.8.6. We have $x \geq 1$ (since x is a positive integer) and $y > 0$ (since y is a positive integer), thus $\underbrace{x}_{\geq 1} + \underbrace{y}_{>0} > 1 + 0 = 1$. Hence, $x + y - 1 > 0$, so that $x + y - 1 \neq 0$.

Now, recall that $x^2 - x = y^2 - y$. Hence, $\left(x^2 - x\right) - \left(y^2 - y\right) = 0$. Comparing this with

$$\begin{aligned}\left(x^2 - x\right) - \left(y^2 - y\right) &= \underbrace{x^2 - y^2}_{=(x+y)(x-y)} - (x-y) = (x+y)(x-y) - (x-y) \\ &= (x+y-1)(x-y),\end{aligned}$$

we obtain $(x+y-1)(x-y) = 0$. We can divide this equality by $x+y-1$ (since $x+y-1 \neq 0$), and thus obtain $x - y = 0$. In other words, $x = y$. This proves Lemma A.8.6. $\square$

(An alternative proof of Lemma A.8.6 could be given using Viete's theorem, as the $x^2 - x = y^2 - y$ condition means that x and y are two roots of the polynomial $X^2 - X - c$ for $c = x^2 - x = y^2 - y$. However, this would be overkill compared to the completely pedestrian proof given above.)

Solution to Exercise 7.7.10. We need to show the following two claims:

Claim 1: There is **at least one** sequence $(u_0, u_1, u_2, \ldots)$ of positive integers such that

$$u_n^2 = \sum_{r=0}^{n} \binom{n+r}{r} u_{n-r} \qquad \text{for all } n \in \mathbb{N}.$$

Claim 2: There is **at most one** sequence $(u_0, u_1, u_2, \ldots)$ of positive integers such that

$$u_n^2 = \sum_{r=0}^{n} \binom{n+r}{r} u_{n-r} \qquad \text{for all } n \in \mathbb{N}. \tag{755}$$

[*Proof of Claim 1:* We shall show that the sequence $\left(2^0, 2^1, 2^2, \ldots\right)$ fits the bill.

Indeed, this sequence $\left(2^0, 2^1, 2^2, \ldots\right)$ is clearly a sequence of positive integers. For each $n \in \mathbb{N}$, we have

$$\begin{aligned}\sum_{r=0}^{n} \binom{n+r}{r} \underbrace{2^{n-r}}_{=\frac{2^n}{2^r}} &= \sum_{r=0}^{n} \binom{n+r}{r} \frac{2^n}{2^r} = 2^n \cdot \sum_{r=0}^{n} \binom{n+r}{r} \frac{1}{2^r} = 2^n \cdot \underbrace{\sum_{k=0}^{n} \binom{n+k}{k} \frac{1}{2^k}}_{\substack{=2^n \\ \text{(by Exercise 7.5.2)}}} \\ &\qquad \text{(here, we have renamed the summation index } r \text{ as } k\text{)} \\ &= 2^n \cdot 2^n = \left(2^n\right)^2.\end{aligned}$$

Thus, $\left(2^n\right)^2 = \sum_{r=0}^{n} \binom{n+r}{r} 2^{n-r}$ for all $n \in \mathbb{N}$. Hence, there is **at least one** sequence $(u_0, u_1, u_2, \ldots)$ of positive integers such that

$$u_n^2 = \sum_{r=0}^{n} \binom{n+r}{r} u_{n-r} \qquad \text{for all } n \in \mathbb{N}$$

(namely, $(u_0, u_1, u_2, \ldots) = \left(2^0, 2^1, 2^2, \ldots\right)$). This proves Claim 1.]

[*Proof of Claim 2:* Let $(v_0, v_1, v_2, \ldots)$ and $(w_0, w_1, w_2, \ldots)$ be two sequences $(u_0, u_1, u_2, \ldots)$ of positive integers satisfying (755). We shall show that $(v_0, v_1, v_2, \ldots) = (w_0, w_1, w_2, \ldots)$.

Indeed, we claim that

$$v_j = w_j \qquad \text{for each } j \in \mathbb{N}. \tag{756}$$

[*Proof of (756):* We shall prove (756) by strong induction on j:

Induction step: Let $n \in \mathbb{N}$. Assume (as the induction hypothesis) that (756) holds for $j < n$. We must prove that (756) holds for $j = n$. In other words, we must prove that $v_n = w_n$.

We know that $(v_0, v_1, v_2, \ldots)$ is a sequence $(u_0, u_1, u_2, \ldots)$ satisfying (755). Thus, (755) holds for $u_i = v_i$. Therefore,

$$v_n^2 = \sum_{r=0}^{n} \binom{n+r}{r} v_{n-r} = \underbrace{\binom{n+0}{0}}_{\substack{=1 \\ \text{(by (119))}}} \underbrace{v_{n-0}}_{=v_n} + \sum_{r=1}^{n} \binom{n+r}{r} v_{n-r} = v_n + \sum_{r=1}^{n} \binom{n+r}{r} v_{n-r}.$$

Thus,

$$v_n^2 - v_n = \sum_{r=1}^{n} \binom{n+r}{r} v_{n-r}. \tag{757}$$

The same argument (applied to $(w_0, w_1, w_2, \ldots)$ instead of $(v_0, v_1, v_2, \ldots)$) yields

$$w_n^2 - w_n = \sum_{r=1}^{n} \binom{n+r}{r} w_{n-r}. \tag{758}$$

However, our induction hypothesis says that (756) holds for $j < n$. In other words, $v_j = w_j$ holds for each $j < n$. In other words, $v_j = w_j$ holds for each $j \in \{0, 1, \ldots, n-1\}$. Substituting $n - r$ for j in this result, we conclude the following:

$$v_{n-r} = w_{n-r} \qquad \text{holds for each } r \in \{1, 2, \ldots, n\}.$$

Hence, the right hand sides of the equalities (757) and (758) are equal. Thus, their left hand sides are equal as well. In other words, $v_n^2 - v_n = w_n^2 - w_n$. Moreover, v_n is a positive integer (since $(v_0, v_1, v_2, \ldots)$ is a sequence of positive integers), and w_n is a positive integer (similarly). Thus, Lemma A.8.6 (applied to $x = v_n$ and $y = w_n$) yields that $v_n = w_n$. This is precisely what we needed to prove. Thus, the induction step is complete. Hence, (756) is proved.]

Now, we have proved (756); in other words, we have shown that $(v_0, v_1, v_2, \ldots) = (w_0, w_1, w_2, \ldots)$.

Forget that we fixed $(v_0, v_1, v_2, \ldots)$ and $(w_0, w_1, w_2, \ldots)$. We thus have shown that if $(v_0, v_1, v_2, \ldots)$ and $(w_0, w_1, w_2, \ldots)$ are two sequences $(u_0, u_1, u_2, \ldots)$ of positive integers satisfying (755), then $(v_0, v_1, v_2, \ldots) = (w_0, w_1, w_2, \ldots)$. In other words, there is **at most one** such sequence $(u_0, u_1, u_2, \ldots)$. Thus, Claim 2 is proved.]

Having proved Claim 1 and Claim 2, we have thus solved Exercise 7.7.10. □

□

A.9. Homework set #8 discussion

The following are discussions of the problems on homework set #8 (Section 8.3).

A.9.1. Discussion of Exercise 8.3.1

Discussion of Exercise 8.3.1. Exercise 8.3.1 appears in [Engel98, Exercise 8.33] and in [19s, Exercise 2.10.19]. We shall outline two solutions for it: a relatively elegant one and a relatively straightforward one.

First solution to Exercise 8.3.1 (sketched). (See [19s, solution to Exercise 2.10.19] for the details of this solution.) We shall show that $\sum_{(x,y)\in Z} \frac{1}{xy} = 1$.

Let W be the set of all pairs $(x,y) \in [n]^2$ satisfying $x \perp y$ and $x+y \leq n$. Any pair $(x,y) \in [n]^2$ satisfying $x \perp y$ must satisfy either $x+y \leq n$ or $x+y > n$ (but not both at the same time), and thus must belong to exactly one of the two sets W and Z. Hence, we can split the sum $\sum_{\substack{(x,y)\in[n]^2; \\ x\perp y}} \frac{1}{xy}$ as follows:

$$\sum_{\substack{(x,y)\in[n]^2; \\ x\perp y}} \frac{1}{xy} = \sum_{(x,y)\in W} \frac{1}{xy} + \sum_{(x,y)\in Z} \frac{1}{xy}. \tag{759}$$

On the other hand, let A be the set of all pairs $(x,y) \in [n]^2$ satisfying $x \perp y$ and $x < y$. Let B be the set of all pairs $(x,y) \in [n]^2$ satisfying $x \perp y$ and $x = y$. Let C be the set of all pairs $(x,y) \in [n]^2$ satisfying $x \perp y$ and $x > y$.

Now, any pair $(x,y) \in [n]^2$ satisfying $x \perp y$ must satisfy exactly one of the three relations $x < y$ and $x = y$ and $x > y$, and thus must belong to exactly one of the three sets A, B and C. Thus, we can split the sum $\sum_{\substack{(x,y)\in[n]^2; \\ x\perp y}} \frac{1}{xy}$ as follows:

$$\sum_{\substack{(x,y)\in[n]^2; \\ x\perp y}} \frac{1}{xy} = \sum_{(x,y)\in A} \frac{1}{xy} + \sum_{(x,y)\in B} \frac{1}{xy} + \sum_{(x,y)\in C} \frac{1}{xy}. \tag{760}$$

We shall now study the three sums on the right hand side of this equality. We begin with the first one: We claim that

$$\sum_{(x,y)\in A} \frac{1}{xy} = \sum_{(x,y)\in W} \frac{1}{x(x+y)}. \tag{761}$$

[*Proof of (761):* It is fairly straightforward to see that the two maps

$$\begin{aligned} f : W &\to A, \\ (u,v) &\mapsto (u, u+v) \end{aligned}$$

and

$$\begin{aligned} g : A &\to W, \\ (u, v) &\mapsto (u, v - u) \end{aligned}$$

are well-defined[440] and mutually inverse[441]. Hence, these two maps are bijections. In particular, this means that f is a bijection. Hence, we can substitute $(u, u + v)$ for (x, y) in the sum $\sum\limits_{(x,y)\in A} \dfrac{1}{xy}$. We thus obtain

$$\sum_{(x,y)\in A} \frac{1}{xy} = \sum_{(u,v)\in W} \frac{1}{u(u+v)} = \sum_{(x,y)\in W} \frac{1}{x(x+y)}$$

(here, we have renamed the summation index (u, v) as (x, y)). This proves (761).]

Next, we claim that

$$\sum_{(x,y)\in C} \frac{1}{xy} = \sum_{(x,y)\in W} \frac{1}{y(x+y)}. \tag{762}$$

[*Proof of (762):* This is similar to the proof of (761). It is fairly straightforward to see that the two maps

$$\begin{aligned} f : W &\to C, \\ (u, v) &\mapsto (u + v, v) \end{aligned}$$

and

$$\begin{aligned} g : C &\to W, \\ (u, v) &\mapsto (u - v, v) \end{aligned}$$

are well-defined[442] and mutually inverse[443]. Hence, these two maps are bijections. In particular, this means that f is a bijection. Hence, we can substitute $(u + v, v)$ for (x, y) in the sum $\sum\limits_{(x,y)\in C} \dfrac{1}{xy}$. We thus obtain

$$\sum_{(x,y)\in C} \frac{1}{xy} = \sum_{(u,v)\in W} \underbrace{\frac{1}{(u+v)v}}_{=\frac{1}{v(u+v)}} = \sum_{(u,v)\in W} \frac{1}{v(u+v)} = \sum_{(x,y)\in W} \frac{1}{y(x+y)}$$

[440] Proving this is fairly easy: The only part that is not completely trivial is showing that $u \perp v$ implies $u \perp u + v$ and $u \perp v - u$. But this can easily be derived from Proposition 3.4.4 (indeed, Proposition 3.4.4 **(b)** entails $\gcd(u, u + v) = \gcd(u, v)$ and $\gcd(u, v - u) = \gcd(u, v)$).

[441] This is completely straightforward to check.

[442] Proving this is fairly easy: The only part that is not completely trivial is showing that $u \perp v$ implies $u + v \perp v$ and $u - v \perp v$. But this can easily be derived from Proposition 3.4.4 (indeed, Proposition 3.4.4 **(b)** entails $\gcd(v, u + v) = \gcd(v, u)$ and $\gcd(v, u - v) = \gcd(v, u)$).

[443] This is completely straightforward to check.

(here, we have renamed the summation index (u,v) as (x,y)). This proves (762).]

Next, we claim that

$$\sum_{(x,y)\in B} \frac{1}{xy} = 1. \tag{763}$$

[*Proof of (763):* The set B consists of all pairs $(x,y) \in [n]^2$ satisfying $x \perp y$ and $x = y$. In other words, B consists of all pairs $(x,x) \in [n]^2$ satisfying $x \perp x$ (since the condition $x = y$ allows us to rewrite the pair (x,y) as (x,x)). However, the only such pair is $(1,1)$ (indeed, a positive integer x always satisfies $\gcd(x,x) = x$, and therefore will satisfy $x \perp x$ if and only if $x = 1$). Hence, $B = \{(1,1)\}$. Consequently,

$$\sum_{(x,y)\in B} \frac{1}{xy} = \sum_{(x,y)\in\{(1,1)\}} \frac{1}{xy} = \frac{1}{1\cdot 1} = 1.$$

This proves (763).]

Now, (760) becomes

$$\begin{aligned}
\sum_{\substack{(x,y)\in[n]^2;\\ x\perp y}} \frac{1}{xy} &= \underbrace{\sum_{(x,y)\in A} \frac{1}{xy}}_{\substack{= \sum_{(x,y)\in W} \frac{1}{x(x+y)} \\ \text{(by (761))}}} + \underbrace{\sum_{(x,y)\in B} \frac{1}{xy}}_{\substack{=1 \\ \text{(by (763))}}} + \underbrace{\sum_{(x,y)\in C} \frac{1}{xy}}_{\substack{= \sum_{(x,y)\in W} \frac{1}{y(x+y)} \\ \text{(by (762))}}} \\
&= \sum_{(x,y)\in W} \frac{1}{x(x+y)} + 1 + \sum_{(x,y)\in W} \frac{1}{y(x+y)} \\
&= \underbrace{\sum_{(x,y)\in W} \frac{1}{x(x+y)} + \sum_{(x,y)\in W} \frac{1}{y(x+y)}}_{= \sum_{(x,y)\in W} \left(\frac{1}{x(x+y)} + \frac{1}{y(x+y)} \right)} + 1 \\
&= \sum_{(x,y)\in W} \underbrace{\left(\frac{1}{x(x+y)} + \frac{1}{y(x+y)} \right)}_{= \frac{y+x}{xy(x+y)} = \frac{1}{xy}} + 1 = \sum_{(x,y)\in W} \frac{1}{xy} + 1.
\end{aligned}$$

Comparing this with (759), we obtain

$$\sum_{(x,y)\in W} \frac{1}{xy} + \sum_{(x,y)\in Z} \frac{1}{xy} = \sum_{(x,y)\in W} \frac{1}{xy} + 1$$

Subtracting $\sum_{(x,y)\in W} \frac{1}{xy}$ from both sides of this equality, we obtain $\sum_{(x,y)\in Z} \frac{1}{xy} = 1$. This solves Exercise 8.3.1. $\square$

Second solution to Exercise 8.3.1 (sketched). Let us rename the set Z as Z_n, to stress its dependence on n. Thus, we must compute $\sum\limits_{(x,y)\in Z_n} \dfrac{1}{xy}$. We claim that

$$\sum_{(x,y)\in Z_n} \frac{1}{xy} = 1.$$

We shall prove this by induction on n:

Induction base: It is easy to see that $\sum\limits_{(x,y)\in Z_n} \dfrac{1}{xy} = 1$ for $n = 1$ (since the only element of Z_1 is the pair $(1,1)$).

Induction step: Let $m > 1$ be an integer. Assume (as the induction hypothesis) that $\sum\limits_{(x,y)\in Z_{m-1}} \dfrac{1}{xy} = 1$. We must prove that $\sum\limits_{(x,y)\in Z_m} \dfrac{1}{xy} = 1$.

We shall achieve this by analyzing how Z_m differs from Z_{m-1}. Recall the definitions of these two sets:

- The set Z_m consists of all pairs $(x,y) \in [m]^2$ satisfying $x \perp y$ and $x + y > m$.
- The set Z_{m-1} consists of all pairs $(x,y) \in [m-1]^2$ satisfying $x \perp y$ and $x + y > m - 1$.

Thus, roughly speaking, the sets Z_m and Z_{m-1} agree in "most" of their elements. To make this more precise, let us see which elements of Z_m fail to belong to Z_{m-1}:

- An element $(x,y) \in Z_m$ will always belong to Z_{m-1}, unless one of x and y equals m (because $x + y > m$ always implies $x + y > m - 1$, whereas $(x,y) \in [m]^2$ implies $(x,y) \in [m-1]^2$ unless one of x and y equals m). Hence, the sum $\sum\limits_{(x,y)\in Z_m} \dfrac{1}{xy}$ can be split as follows:

$$\sum_{(x,y)\in Z_m} \frac{1}{xy} = \sum_{\substack{(x,y)\in Z_m; \\ (x,y)\in Z_{m-1}}} \frac{1}{xy} + \sum_{\substack{(x,y)\in Z_m; \\ \text{one of } x \text{ and } y \text{ equals } m}} \frac{1}{xy}. \tag{764}$$

We can likewise analyze which elements of Z_{m-1} fail to belong to Z_m:

- An element $(x,y) \in Z_{m-1}$ will always belong to Z_m, unless $x + y = m$ (because $x + y > m - 1$ implies $x + y > m$ unless $x + y = m$, whereas $(x,y) \in [m-1]^2$ always implies $(x,y) \in [m]^2$). Hence, the sum $\sum\limits_{(x,y)\in Z_{m-1}} \dfrac{1}{xy}$ can be split as follows:

$$\sum_{(x,y)\in Z_{m-1}} \frac{1}{xy} = \sum_{\substack{(x,y)\in Z_{m-1}; \\ (x,y)\in Z_m}} \frac{1}{xy} + \sum_{\substack{(x,y)\in Z_{m-1}; \\ x+y=m}} \frac{1}{xy}. \tag{765}$$

Now, subtracting (765) from (764), we obtain

$$\begin{aligned}
&\sum_{(x,y)\in Z_m} \frac{1}{xy} - \sum_{(x,y)\in Z_{m-1}} \frac{1}{xy} \\
&= \left(\sum_{\substack{(x,y)\in Z_m; \\ (x,y)\in Z_{m-1}}} \frac{1}{xy} + \sum_{\substack{(x,y)\in Z_m; \\ \text{one of } x \text{ and } y \text{ equals } m}} \frac{1}{xy} \right) - \left(\sum_{\substack{(x,y)\in Z_{m-1}; \\ (x,y)\in Z_m}} \frac{1}{xy} + \sum_{\substack{(x,y)\in Z_{m-1}; \\ x+y=m}} \frac{1}{xy} \right) \\
&= \underbrace{\sum_{\substack{(x,y)\in Z_m; \\ (x,y)\in Z_{m-1}}} \frac{1}{xy} - \sum_{\substack{(x,y)\in Z_{m-1}; \\ (x,y)\in Z_m}} \frac{1}{xy}}_{\substack{=0 \\ \text{(since these two sums are equal)}}} + \underbrace{\sum_{\substack{(x,y)\in Z_m; \\ \text{one of } x \text{ and } y \text{ equals } m}} \frac{1}{xy}}_{= \sum\limits_{\substack{(x,y)\in [m]^2; \\ x\perp y; \\ (x=m \text{ or } y=m)}}} - \underbrace{\sum_{\substack{(x,y)\in Z_{m-1}; \\ x+y=m}} \frac{1}{xy}}_{= \sum\limits_{\substack{(x,y)\in [m-1]^2; \\ x\perp y; \\ x+y=m}}} \\
&= \sum_{\substack{(x,y)\in [m]^2; \\ x\perp y; \\ (x=m \text{ or } y=m)}} \frac{1}{xy} - \sum_{\substack{(x,y)\in [m-1]^2; \\ x\perp y; \\ x+y=m}} \frac{1}{xy}. \qquad (766)
\end{aligned}$$

Now, let us simplify the first sum on the right hand side. A pair $(x, y) \in [m]^2$ satisfying $x \perp y$ cannot satisfy $x = m$ and $y = m$ simultaneously (because this would mean that $m \perp m$, but this is impossible because of $m > 1$). Hence, we can

split the sum $\displaystyle\sum_{\substack{(x,y)\in[m]^2;\\ x\perp y;\\ (x=m \text{ or } y=m)}} \frac{1}{xy}$ as follows:

$$\sum_{\substack{(x,y)\in[m]^2;\\ x\perp y;\\ (x=m \text{ or } y=m)}} \frac{1}{xy} = \underbrace{\sum_{\substack{(x,y)\in[m]^2;\\ x\perp y;\\ x=m \text{ but not } y=m}} \frac{1}{xy}}_{\substack{= \sum\limits_{\substack{y\in[m];\\ m\perp y;\\ \text{not } y=m}} \frac{1}{my} \\ = \sum\limits_{\substack{y\in[m-1];\\ m\perp y}} \frac{1}{my}}} + \underbrace{\sum_{\substack{(x,y)\in[m]^2;\\ x\perp y;\\ y=m \text{ but not } x=m}} \frac{1}{xy}}_{\substack{= \sum\limits_{\substack{x\in[m];\\ x\perp m;\\ \text{not } x=m}} \frac{1}{xm} \\ = \sum\limits_{\substack{x\in[m-1];\\ x\perp m}} \frac{1}{xm}}}$$

$$= \underbrace{\sum_{\substack{y\in[m-1];\\ m\perp y}} \frac{1}{my}}_{\substack{= \sum\limits_{\substack{x\in[m-1];\\ m\perp m-x}} \frac{1}{m(m-x)} \\ \text{(here, we have substituted } m-x \text{ for } y \text{ in the sum)}}} + \sum_{\substack{x\in[m-1];\\ x\perp m}} \frac{1}{xm}$$

$$= \underbrace{\sum_{\substack{x\in[m-1];\\ m\perp m-x}}}_{\substack{= \sum\limits_{\substack{x\in[m-1];\\ x\perp m}} \\ \text{(since it is not hard to see that } m\perp m-x \text{ is equivalent to } x\perp m)}} \frac{1}{m(m-x)} + \sum_{\substack{x\in[m-1];\\ x\perp m}} \frac{1}{xm}$$

$$= \sum_{\substack{x\in[m-1];\\ x\perp m}} \frac{1}{m(m-x)} + \sum_{\substack{x\in[m-1];\\ x\perp m}} \frac{1}{xm} = \sum_{\substack{x\in[m-1];\\ x\perp m}} \underbrace{\left(\frac{1}{m(m-x)} + \frac{1}{xm}\right)}_{=\frac{1}{x(m-x)}}$$

$$= \sum_{\substack{x\in[m-1];\\ x\perp m}} \frac{1}{x(m-x)}. \tag{767}$$

Meanwhile, the second sum on the right hand side of (766) rewrites as follows:

$$\sum_{\substack{(x,y)\in[m-1]^2;\\ x\perp y;\\ x+y=m}} \frac{1}{xy} = \underbrace{\sum_{\substack{x\in[m-1];\\ x\perp m-x}}}_{\substack{= \sum\limits_{\substack{x\in[m-1];\\ x\perp m}} \\ \text{(since it is not hard} \\ \text{to see that } x\perp m-x \text{ is} \\ \text{equivalent to } x\perp m)}} \frac{1}{x(m-x)}$$

$$\left(\begin{array}{c}\text{here, we have substituted } (x, m-x) \\ \text{for } (x,y) \text{ in the sum, since the} \\ \text{condition } x+y=m \text{ allows rewriting } y \text{ as } m-x\end{array}\right)$$

$$= \sum_{\substack{x\in[m-1];\\ x\perp m}} \frac{1}{x(m-x)}. \tag{768}$$

In view of (767) and (768), we can rewrite (766) as

$$\sum_{(x,y)\in Z_m} \frac{1}{xy} - \sum_{(x,y)\in Z_{m-1}} \frac{1}{xy} = \sum_{\substack{x\in[m-1];\\ x\perp m}} \frac{1}{x(m-x)} - \sum_{\substack{x\in[m-1];\\ x\perp m}} \frac{1}{x(m-x)} = 0.$$

Hence,

$$\sum_{(x,y)\in Z_m} \frac{1}{xy} = \sum_{(x,y)\in Z_{m-1}} \frac{1}{xy} = 1 \qquad \text{(by the induction hypothesis)}.$$

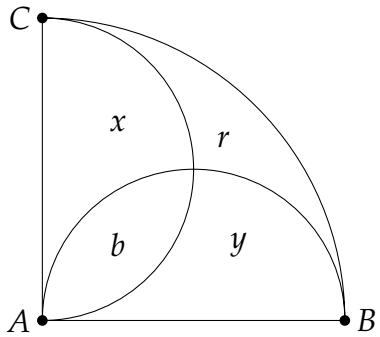
This completes the induction step. Thus, Exercise 8.3.1 is solved again. □

□

A.9.2. Discussion of Exercise 8.3.2

Discussion of Exercise 8.3.2. First and foremost: The two parts of Exercise 8.3.2 come from different subjects, but agree in the rough idea of their solution, which is to compute the area of a shape (in part **(a)**) or the size of a set (in part **(b)**) by first computing the area of a larger shape or the size of a larger set, and then subtracting the part that was overcounted. In the case of the size of a finite set (which is what we have in part **(b)**), this is a trick we have used many times before (always relying on the difference rule – i.e., on Theorem 7.1.8). In the case of the area of a shape, the conceptual grounding is more sophisticated (areas require measure theory to rigorously define, whereas the size of a finite set is one of the most basic notions in mathematics); but the idea is the same.

Solution to Exercise 8.3.2 (sketched). **(a)** The two half-circles subdivide the quarter-circle into four regions. We denote the areas of these four regions by b, r, x, y in the following way:



.

[444] Thus, the area of the red region is r, whereas the area of the blue region is b. We claim that $r = b$.

Indeed, let $a = |AB| = |AC|$. Thus, the quarter-circle has radius a; hence, its area is $\frac{1}{4}\pi a^2$. In other words,

$$b + r + x + y = \frac{1}{4}\pi a^2 \tag{769}$$

(since the quarter-circle has been subdivided into four regions with areas b, r, x, y, and thus its area is $b + r + x + y$).

However, the lower semicircle has diameter AB and thus radius $\frac{1}{2}\underbrace{|AB|}_{=a} = \frac{1}{2}a$. Hence, its area is $\frac{1}{2}\pi\left(\frac{1}{2}a\right)^2 = \frac{1}{8}\pi a^2$. In other words,

$$b + y = \frac{1}{8}\pi a^2 \tag{770}$$

(since the lower semicircle has been subdivided into two regions with areas b and y, and thus its area is $b + y$).

A similar argument (using the upper semicircle instead of the lower one) shows that $b + x = \frac{1}{8}\pi a^2$. Subtracting this equality from (769), we obtain $(b + r + x + y) -$

[444] Here is how to define b, r, x, y without referencing the picture: We shall refer to the semicircle with diameter AB as the *lower semicircle*. We shall refer to the semicircle with diameter AC as the *upper semicircle*. Let $\mathbb{A}(U)$ denote the area of any shape U. Now, we set

$$\begin{aligned}
b &= \mathbb{A}\left(\left(\text{lower semicircle}\right) \cap \left(\text{upper semicircle}\right)\right); \\
r &= \mathbb{A}\left(\left(\text{quarter-circle}\right) \setminus \left(\left(\text{lower semicircle}\right) \cup \left(\text{upper semicircle}\right)\right)\right); \\
x &= \mathbb{A}\left(\left(\text{upper semicircle}\right) \setminus \left(\text{lower semicircle}\right)\right); \\
y &= \mathbb{A}\left(\left(\text{lower semicircle}\right) \setminus \left(\text{upper semicircle}\right)\right).
\end{aligned}$$

$(b+x) = \frac{1}{4}\pi a^2 - \frac{1}{8}\pi a^2$. This simplifies to $r+y = \frac{1}{8}\pi a^2$. Comparing this with (770), we obtain $r+y = b+y$. Hence, $r = b$. This solves Exercise 8.3.2 **(a)**.

(b) Exercise 8.3.2 **(b)** is [17f-hw8s, Lemma 0.3], where I give three proofs. The shortest of them is the second proof, which I will now reproduce:

The map σ is a permutation of $[n]$, thus a bijection $[n] \to [n]$. Hence, it has an inverse map σ^{-1}. Now, the map

$$\begin{aligned} \{i \in [n] \mid \sigma(i) \geq j\} &\to \{i \in [n] \mid i \geq j\}, \\ i &\mapsto \sigma(i) \end{aligned}$$

is clearly well-defined. Furthermore, this map is a bijection (indeed, its inverse sends each i to $\sigma^{-1}(i)$). Hence, the bijection principle yields

$$|\{i \in [n] \mid \sigma(i) \geq j\}| = |\{i \in [n] \mid i \geq j\}|.$$

In other words,

$$(\text{\# of all } i \in [n] \text{ satisfying } \sigma(i) \geq j) = (\text{\# of all } i \in [n] \text{ satisfying } i \geq j).$$

We have

$$\begin{aligned} &\left(\text{\# of all } i \in [n] \text{ satisfying } \underbrace{i \geq j > \sigma(i)}_{\iff (i \geq j \text{ but } \textbf{not } \sigma(i) \geq j)}\right) \\ &= (\text{\# of all } i \in [n] \text{ satisfying } i \geq j \text{ but } \textbf{not } \sigma(i) \geq j) \\ &= (\text{\# of all } i \in [n] \text{ satisfying } i \geq j) \\ &\qquad - (\text{\# of all } i \in [n] \text{ satisfying } i \geq j \text{ and } \sigma(i) \geq j) \end{aligned}$$

(by the difference rule) and

$$\begin{aligned} &\left(\text{\# of all } i \in [n] \text{ satisfying } \underbrace{\sigma(i) \geq j > i}_{\iff (\sigma(i) \geq j \text{ but } \textbf{not } i \geq j)}\right) \\ &= (\text{\# of all } i \in [n] \text{ satisfying } \sigma(i) \geq j \text{ but } \textbf{not } i \geq j) \\ &= \underbrace{(\text{\# of all } i \in [n] \text{ satisfying } \sigma(i) \geq j)}_{\substack{=(\text{\# of all } i \in [n] \text{ satisfying } i \geq j) \\ \text{(this has been shown above)}}} \\ &\qquad - \left(\text{\# of all } i \in [n] \text{ satisfying } \underbrace{\sigma(i) \geq j \text{ and } i \geq j}_{\iff (i \geq j \text{ and } \sigma(i) \geq j)}\right) \\ &\qquad\qquad \text{(by the difference rule)} \\ &= (\text{\# of all } i \in [n] \text{ satisfying } i \geq j) \\ &\qquad - (\text{\# of all } i \in [n] \text{ satisfying } i \geq j \text{ and } \sigma(i) \geq j). \end{aligned}$$

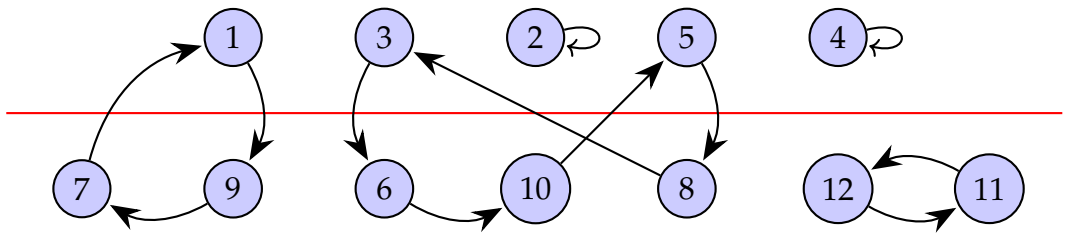
Comparing these two equalities, we obtain

$$(\#\text{ of all } i \in [n] \text{ satisfying } i \geq j > \sigma(i)) = (\#\text{ of all } i \in [n] \text{ satisfying } \sigma(i) \geq j > i).$$

This solves Exercise 8.3.2 **(b)**. □

Let us illustrate the claim of Exercise 8.3.2 **(b)** on an example:

Example A.9.1. Let $n = 12$ and let $j = 6$. Let σ be the permutation of $[n]$ that sends $1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12$ to $9, 2, 6, 4, 8, 10, 1, 3, 7, 5, 12, 11$, respectively. As in Example 6.2.4 and Example 6.2.7, we can represent σ by a diagram:



.

Here, we have drawn the $j-1$ nodes $1, 2, \ldots, j-1$ above the red line, and drawn the remaining $n-j+1$ nodes $j, j+1, \ldots, n$ below the red line. Now,

$$(\#\text{ of all } i \in [n] \text{ satisfying } i \geq j > \sigma(i))$$

is the # of all arrows that cross the red line from bottom to top, whereas

$$(\#\text{ of all } i \in [n] \text{ satisfying } \sigma(i) \geq j > i)$$

is the # of all arrows that cross the red line from top to bottom. (We assume that no arrow crosses the red line more than once; this can be achieved, e.g., by making all arrows straight.) Hence, Exercise 8.3.2 **(b)** says that the # of arrows crossing the red line from bottom to top equals the # of arrows crossing the red line from top to bottom. Stated in this way, this feels rather intuitive – particularly if one remembers that the arrows form several disjoint cycles, and each cycle must enter the top region as often as it leaves it. It is essentially a manifestation of the "what goes in must come out" principle. This thinking can actually be made rigorous, leading to a different solution to Exercise 8.3.2 **(b)**. (See [17f-hw8s, Third proof of Lemma 0.3] for this solution.)

□

A.10. Homework set #9 discussion

The following are discussions of the problems on homework set #9 (Section 8.4).

A.10.1. Discussion of Exercise 8.4.1 (TODO: add details!)

Discussion of Exercise 8.4.1. Exercise 8.4.1 is [18s-mt2s, Exercise 1]. For the sake of completeness, let us copy the solution from [18s-mt2s]. We will first need an auxiliary concept:

Definition A.10.1. Let $n \in \mathbb{N}$ and $d \in \mathbb{N}$. Let $h \in [d]$. An n-tuple $(x_1, x_2, \ldots, x_n) \in [d]^n$ will be called *h-even* if the number h occurs in it an even number of times (i.e., the number of $i \in [n]$ satisfying $x_i = h$ is even). (For example, the 3-tuple $(1, 4, 4)$ is 4-even and 3-even but not 1-even.)

This definition generalizes the concept of "1-even" used in Exercise 7.8.1. Let us now state the following generalization of Exercise 7.8.1:

Proposition A.10.2. Let $n \in \mathbb{N}$, and let d be a positive integer. Let $h \in [d]$. Then, the # of h-even n-tuples in $[d]^n$ is $\frac{1}{2}\left(d^n + (d-2)^n\right)$.

Indeed, Exercise 7.8.1 is the particular case of Proposition A.10.2 for $h = 1$; but conversely, Proposition A.10.2 can be derived from Exercise 7.8.1 by "renaming 1 as h".

Here is a rigorous way to make this "renaming" argument:

Proof of Proposition A.10.2 (sketched). There is clearly some permutation σ of $[d]$ such that $\sigma(1) = h$. (For example, we can let σ be the transposition swapping 1 with h when $h \neq 1$, and otherwise we can just set $\sigma = \mathrm{id}$.) Fix such a σ. It is easy to see that the map

$$\begin{aligned} [d]^n &\to [d]^n, \\ (x_1, x_2, \ldots, x_n) &\mapsto (\sigma(x_1), \sigma(x_2), \ldots, \sigma(x_n)) \end{aligned}$$

(that is, the map that applies σ to each entry of an n-tuple in $[d]^n$) is a bijection (since σ is a bijection). Furthermore, an n-tuple $(x_1, x_2, \ldots, x_n) \in [d]^n$ is 1-even if and only if its image $(\sigma(x_1), \sigma(x_2), \ldots, \sigma(x_n))$ under this bijection is h-even (because the number 1 occurs in the n-tuple $(x_1, x_2, \ldots, x_n)$ at the same positions at which the number $\sigma(1) = h$ occurs in the n-tuple $(\sigma(x_1), \sigma(x_2), \ldots, \sigma(x_n))$). Thus, there is a bijection

$$\begin{aligned} \left\{\text{1-even } n\text{-tuples in } [d]^n\right\} &\to \left\{h\text{-even } n\text{-tuples in } [d]^n\right\}, \\ (x_1, x_2, \ldots, x_n) &\mapsto (\sigma(x_1), \sigma(x_2), \ldots, \sigma(x_n)). \end{aligned}$$

Hence, we can use the bijection principle to see that

$$\begin{aligned} \left|\left\{h\text{-even } n\text{-tuples in } [d]^n\right\}\right| &= \left|\left\{\text{1-even } n\text{-tuples in } [d]^n\right\}\right| \\ &= \left(\#\text{ of 1-even } n\text{-tuples in } [d]^n\right) \\ &= \frac{1}{2}\left(d^n + (d-2)^n\right) \qquad \text{(by our solution to Exercise 7.8.1)}. \end{aligned}$$

In other words, the # of h-even n-tuples in $[d]^n$ is $\frac{1}{2}\left(d^n + (d-2)^n\right)$. This proves Proposition A.10.2.

(Alternatively, of course, Proposition A.10.2 can be proved in the same way as we solved Exercise 7.8.1, just with all "1"s in the appropriate places replaced by "h"s.) □

Solution to Exercise 8.4.1 (sketched). We first make the following claim:

> *Observation 1:* Let $h \in [d]$. Then, the # of $(n-1)$-tuples in $[d]^{n-1}$ that are **not** h-even is $\frac{1}{2}\left(d^{n-1} - (d-2)^{n-1}\right)$.

[*Proof of Observation 1:* Proposition A.10.2 (applied to $n-1$ instead of n) shows that the # of h-even $(n-1)$-tuples in $[d]^{n-1}$ is $\frac{1}{2}\left(d^{n-1} + (d-2)^{n-1}\right)$. In other words,

$$\left(\#\text{ of } h\text{-even } (n-1)\text{-tuples in } [d]^{n-1}\right) = \frac{1}{2}\left(d^{n-1} + (d-2)^{n-1}\right).$$

Now, the difference rule yields

$$\begin{aligned}
&\left(\#\text{ of } (n-1)\text{-tuples in } [d]^{n-1} \text{ that are } \textbf{not } h\text{-even}\right) \\
&= \underbrace{\left(\#\text{ of all } (n-1)\text{-tuples in } [d]^{n-1}\right)}_{=d^{n-1}} - \underbrace{\left(\#\text{ of } h\text{-even } (n-1)\text{-tuples in } [d]^{n-1}\right)}_{=\frac{1}{2}\left(d^{n-1}+(d-2)^{n-1}\right)} \\
&= d^{n-1} - \frac{1}{2}\left(d^{n-1} + (d-2)^{n-1}\right) = \frac{1}{2}\left(d^{n-1} - (d-2)^{n-1}\right).
\end{aligned}$$

In other words, the # of $(n-1)$-tuples in $[d]^{n-1}$ that are **not** h-even is $\frac{1}{2}\left(d^{n-1} - (d-2)^{n-1}\right)$. This proves Observation 1.]

We can construct each first-even n-tuple $(x_1, x_2, \ldots, x_n)$ in $[d]^n$ as follows:

- First, we choose the value of x_1. We denote this value by h. There are d options for this decision (since this value must belong to $[d]$).
- Next, we choose the $(n-1)$-tuple $(x_2, x_3, \ldots, x_n)$. Note that the entry $h = x_1$ must occur an **odd** number of times in this $(n-1)$-tuple $(x_2, x_3, \ldots, x_n)$ (because we want the n-tuple $(x_1, x_2, \ldots, x_n)$ to be first-even, so that x_1 must occur an **even** number of times in this n-tuple; but the $(n-1)$-tuple $(x_2, x_3, \ldots, x_n)$ is missing its very first occurrence, and thus must contain it an **odd** number of times). In other words, the $(n-1)$-tuple $(x_2, x_3, \ldots, x_{n-1})$ must **not** be h-even. Thus, there are $\frac{1}{2}\left(d^{n-1} - (d-2)^{n-1}\right)$ options for this decision (since Observation 1 yields that the number of $(n-1)$-tuples in $[d]^{n-1}$ that are **not** h-even is $\frac{1}{2}\left(d^{n-1} - (d-2)^{n-1}\right)$).

Hence, the dependent product rule shows that the total # of first-even n-tuples $(x_1, x_2, \ldots, x_n)$ in $[d]^n$ is

$$d \cdot \frac{1}{2}\left(d^{n-1} - (d-2)^{n-1}\right) = \frac{1}{2}d\left(d^{n-1} - (d-2)^{n-1}\right).$$

This solves Exercise 8.4.1. □

□

A.10.2. Discussion of Exercise 8.4.2 (TODO: add details!)

Discussion of Exercise 8.4.2. Exercise 8.4.2 is a generalization of `https://math.stackexchange.com/questions/897948` (which is the particular case for $p = 1$). Let me sketch a combinatorial and two algebraic solutions:

First solution to Exercise 8.4.2 (sketched). Let S be the $2n$-element set $\{-1, -2, \ldots, -n\} \cup \{1, 2, \ldots, n\}$. Let U be the set of all $(n+p)$-element subsets of S. For each $i \in [n]$, we let A_i be the set of all $J \in U$ that contain neither i nor $-i$ (that is, that satisfy $i \notin J$ and $-i \notin J$). The Principle of Inclusion and Exclusion (specifically, Theorem 7.8.6) shows that

$$\begin{aligned}&(\#\text{ of } x \in U \text{ satisfying } x \notin A_i \text{ for all } i \in [n])\\ &= \sum_{I \subseteq [n]} (-1)^{|I|} (\#\text{ of } x \in U \text{ satisfying } x \in A_i \text{ for all } i \in I)\,. \qquad (771)\end{aligned}$$

Now, let us compute both the left and the right hand side of this equality, and see what it then turns into.

The left hand side is an easy counting problem: We have

$$\begin{aligned}&(\#\text{ of } x \in U \text{ satisfying } x \notin A_i \text{ for all } i \in [n])\\ &= (\#\text{ of } J \in U \text{ satisfying } J \notin A_i \text{ for all } i \in [n])\\ &\qquad\qquad (\text{here, we have renamed the index } x \text{ as } J)\\ &= (\#\text{ of } J \in U \text{ such that each } i \in [n] \text{ satisfies either } i \in J \text{ or } -i \in J)\\ &\qquad\qquad (\text{by the definition of } A_i)\\ &= \binom{n}{p} \cdot 2^{n-p}. \qquad (772)\end{aligned}$$

Indeed, the last equality sign here follows from the dependent product rule, since the following decision procedure can be used to construct a set $J \in U$ such that each $i \in [n]$ satisfies either $i \in J$ or $-i \in J$:

- We choose the p elements $i \in [n]$ for which J will contain **both** i and $-i$. (Make sure you understand why there should be exactly p such elements!) There are $\binom{n}{p}$ options for this decision.
- For each of the remaining $n - p$ elements $i \in [n]$, we decide whether J will contain i or $-i$. (Indeed, J will have to contain exactly one of i and $-i$ for each of these i.) These are $n - p$ decisions, and each of them allows for 2 options; thus, we have 2^{n-p} options here in total.

The right hand side of (771) is even easier to compute: If I is any subset of $[n]$,

then

$$\begin{aligned}
&(\#\text{ of } x \in U \text{ satisfying } x \in A_i \text{ for all } i \in I) \\
&= (\#\text{ of } J \in U \text{ satisfying } J \in A_i \text{ for all } i \in I) \\
&\qquad\qquad (\text{here, we have renamed the index } x \text{ as } J) \\
&= (\#\text{ of } J \in U \text{ such that each } i \in [n] \text{ satisfies neither } i \in J \text{ nor } -i \in J) \\
&\qquad\qquad (\text{by the definition of } A_i) \\
&= (\#\text{ of } (n+p)\text{-element subsets } J \text{ of } S \setminus (\pm I)) \\
&\qquad\qquad (\text{where } \pm I \text{ denotes the set } I \cup \{-i \mid i \in I\}) \\
&= \binom{2n - 2|I|}{n+p} \qquad (773)
\end{aligned}$$

(since $|S \setminus (\pm I)| = \underbrace{|S|}_{=2n} - \underbrace{|\pm I|}_{=2|I|} = 2n - 2|I|$). Hence,

$$\begin{aligned}
&\sum_{I \subseteq [n]} (-1)^{|I|} \underbrace{(\#\text{ of } x \in U \text{ satisfying } x \in A_i \text{ for all } i \in I)}_{\substack{= \binom{2n-2|I|}{n+p} \\ (\text{by } (773))}} \\
&= \underbrace{\sum_{I \subseteq [n]}}_{= \sum_{k=0}^{n} \sum_{\substack{I \subseteq [n]; \\ |I| = k}}} (-1)^{|I|} \binom{2n-2|I|}{n+p} = \sum_{k=0}^{n} \underbrace{\sum_{\substack{I \subseteq [n]; \\ |I| = k}} (-1)^{|I|} \binom{2n-2|I|}{n+p}}_{= (-1)^k \binom{n}{k} \binom{2n-2k}{n+p}} \\
&= \sum_{k=0}^{n} (-1)^k \binom{n}{k} \binom{2n-2k}{n+p}. \qquad (774)
\end{aligned}$$

Hence, (771) becomes

$$\begin{aligned}
&(\#\text{ of } x \in U \text{ satisfying } x \notin A_i \text{ for all } i \in [n]) \\
&= \sum_{I \subseteq [n]} (-1)^{|I|} (\#\text{ of } x \in U \text{ satisfying } x \in A_i \text{ for all } i \in I) \\
&= \sum_{k=0}^{n} (-1)^k \binom{n}{k} \binom{2n-2k}{n+p} \qquad (\text{by } (774)).
\end{aligned}$$

Comparing this with (772), we obtain

$$\sum_{k=0}^{n} (-1)^k \binom{n}{k} \binom{2n-2k}{n+p} = \binom{n}{p} \cdot 2^{n-p} = 2^{n-p} \binom{n}{p}.$$

This solves Exercise 8.4.2. □

Our second solution to Exercise 8.4.2 is algebraic; it uses induction on n. The main ingredient is the following simple lemma (which is generally quite useful for recursively simplifying sums of the form $\sum_{k=0}^{n}(-1)^k\binom{n}{k}a_k$):

Lemma A.10.3. Let n be a positive integer. Let $a_0, a_1, \ldots, a_n$ be any $n+1$ numbers. Then,

$$\sum_{k=0}^{n}(-1)^k\binom{n}{k}a_k = \sum_{k=0}^{n-1}(-1)^k\binom{n-1}{k}(a_k - a_{k+1}).$$

Proof of Lemma A.10.3. We have

$$\begin{aligned}
&\sum_{k=0}^{n}(-1)^k \underbrace{\binom{n}{k}}_{\substack{=\binom{n-1}{k-1}+\binom{n-1}{k} \\ \text{(by Theorem 4.3.7)}}} a_k \\
&= \sum_{k=0}^{n}\underbrace{(-1)^k\left(\binom{n-1}{k-1}+\binom{n-1}{k}\right)a_k}_{=(-1)^k\binom{n-1}{k-1}a_k+(-1)^k\binom{n-1}{k}a_k} \\
&= \sum_{k=0}^{n}\left((-1)^k\binom{n-1}{k-1}a_k + (-1)^k\binom{n-1}{k}a_k\right) \\
&= \sum_{k=0}^{n}(-1)^k\binom{n-1}{k-1}a_k + \sum_{k=0}^{n}(-1)^k\binom{n-1}{k}a_k. \qquad (775)
\end{aligned}$$

Now, $0-1=-1\notin\mathbb{N}$ and thus $\binom{n-1}{0-1}=0$ (by an application of (118)). Now,

$$\begin{aligned}
&\sum_{k=0}^{n}(-1)^k\binom{n-1}{k-1}a_k \\
&=(-1)^0\underbrace{\binom{n-1}{0-1}}_{=0}a_0+\sum_{k=1}^{n}(-1)^k\binom{n-1}{k-1}a_k \\
&\qquad\text{(here, we have split off the addend for } k=0 \text{ from the sum)} \\
&=\underbrace{(-1)^0\,0a_0}_{=0}+\sum_{k=1}^{n}(-1)^k\binom{n-1}{k-1}a_k \\
&=\sum_{k=1}^{n}(-1)^k\binom{n-1}{k-1}a_k=\sum_{k=0}^{n-1}\underbrace{(-1)^{k+1}}_{=-(-1)^k}\underbrace{\binom{n-1}{(k+1)-1}}_{=\binom{n-1}{k}}a_{k+1} \\
&\qquad\text{(here, we have substituted } k+1 \text{ for } k \text{ in the sum)} \\
&=\sum_{k=0}^{n-1}\left(-(-1)^k\right)\binom{n-1}{k}a_{k+1}=-\sum_{k=0}^{n-1}(-1)^k\binom{n-1}{k}a_{k+1}. \qquad (776)
\end{aligned}$$

Also, $n\geq 1$ (since n is a positive integer), so that $n-1\in\mathbb{N}$. Hence, Proposition 4.3.4 (applied to $n-1$ and n instead of n and k) yields $\binom{n-1}{n}=0$ (since $n>n-1$). Now,

$$\begin{aligned}
&\sum_{k=0}^{n}(-1)^k\binom{n-1}{k}a_k \\
&=(-1)^n\underbrace{\binom{n-1}{n}}_{=0}a_n+\sum_{k=0}^{n-1}(-1)^k\binom{n-1}{k}a_k \\
&\qquad\text{(here, we have split off the addend for } k=n \text{ from the sum)} \\
&=\underbrace{(-1)^n\,0a_n}_{=0}+\sum_{k=0}^{n-1}(-1)^k\binom{n-1}{k}a_k \\
&=\sum_{k=0}^{n-1}(-1)^k\binom{n-1}{k}a_k. \qquad (777)
\end{aligned}$$

Now, (775) becomes

$$\begin{aligned}
\sum_{k=0}^{n}(-1)^k\binom{n}{k}a_k &= \underbrace{\sum_{k=0}^{n}(-1)^k\binom{n-1}{k-1}a_k}_{\substack{=-\sum_{k=0}^{n-1}(-1)^k\binom{n-1}{k}a_{k+1} \\ \text{(by (776))}}} + \underbrace{\sum_{k=0}^{n}(-1)^k\binom{n-1}{k}a_k}_{\substack{=\sum_{k=0}^{n-1}(-1)^k\binom{n-1}{k}a_k \\ \text{(by (777))}}} \\
&= -\sum_{k=0}^{n-1}(-1)^k\binom{n-1}{k}a_{k+1} + \sum_{k=0}^{n-1}(-1)^k\binom{n-1}{k}a_k \\
&= \sum_{k=0}^{n-1}(-1)^k\binom{n-1}{k}a_k - \sum_{k=0}^{n-1}(-1)^k\binom{n-1}{k}a_{k+1} \\
&= \sum_{k=0}^{n-1}\underbrace{\left((-1)^k\binom{n-1}{k}a_k - (-1)^k\binom{n-1}{k}a_{k+1}\right)}_{=(-1)^k\binom{n-1}{k}(a_k-a_{k+1})} \\
&= \sum_{k=0}^{n-1}(-1)^k\binom{n-1}{k}(a_k-a_{k+1}).
\end{aligned}$$

This proves Lemma A.10.3. □

Second solution to Exercise 8.4.2. Forget that we fixed n and p. We proceed by induction on n:

Induction base: It is straightforward to see that each $p \in \mathbb{Z}$ satisfies

$$\binom{0}{p} = 2^{0-p}\binom{0}{p}. \tag{778}$$

[445] Hence, it is easy to see that Exercise 8.4.2 holds for $n = 0$ [446].

Induction step: Let m be a positive integer. Assume (as the induction hypothesis) that Exercise 8.4.2 holds for $n = m - 1$. We must prove that Exercise 8.4.2 holds for $n = m$.

We have assumed that Exercise 8.4.2 holds for $n = m - 1$. In other words, each $p \in \mathbb{Z}$ satisfies

$$\sum_{k=0}^{m-1} (-1)^k \binom{m-1}{k} \binom{2(m-1)-2k}{m-1+p} = 2^{m-1-p} \binom{m-1}{p}. \tag{779}$$

Now, let $p \in \mathbb{Z}$. Then, Lemma A.10.3 (applied to $n = m$ and $a_k = \binom{2m-2k}{m+p}$)

[445] *Proof of (778):* Let $p \in \mathbb{Z}$. We are in one of the following three cases:

Case 1: We have $p = 0$.

Case 2: We have $p > 0$.

Case 3: We have $p < 0$.

Let us first consider Case 1. In this case, we have $p = 0$. Thus, $2^{0-p} = 2^{0-0} = 2^0 = 1$, so that $\underbrace{2^{0-p}}_{=1} \binom{0}{p} = \binom{0}{p}$. Thus, (778) holds. Hence, we have proved (778) in Case 1.

Let us now consider Case 2. In this case, we have $p > 0$. Thus, Proposition 4.3.4 (applied to 0 and p instead of n and k) yields that $\binom{0}{p} = 0$. Comparing this with $2^{0-p} \underbrace{\binom{0}{p}}_{=0} = 0$, we obtain $\binom{0}{p} = 2^{0-p} \binom{0}{p}$. Hence, we have proved (778) in Case 2.

Finally, let us consider Case 3. In this case, we have $p < 0$. Thus, $p \notin \mathbb{N}$, so that $\binom{0}{p} = 0$ (by an application of (118)). Comparing this with $2^{0-p} \underbrace{\binom{0}{p}}_{=0} = 0$, we obtain $\binom{0}{p} = 2^{0-p} \binom{0}{p}$. Hence, we have proved (778) in Case 3.

We have now proved (778) in all three Cases 1, 2 and 3. Thus, (778) always holds.

[446] *Proof.* For each $p \in \mathbb{Z}$, we have

$$\begin{aligned}
\sum_{k=0}^{0} (-1)^k \binom{0}{k} \binom{2\cdot 0-2k}{0+p} &= \underbrace{(-1)^0}_{=1} \underbrace{\binom{0}{0}}_{=1} \underbrace{\binom{2\cdot 0-2\cdot 0}{0+p}}_{\substack{=\binom{0}{p} \\ \text{(since } 2\cdot 0-2\cdot 0=0 \text{ and } 0+p=p)}} \\
&= \binom{0}{p} = 2^{0-p} \binom{0}{p} \qquad \text{(by (778))}.
\end{aligned}$$

In other words, Exercise 8.4.2 holds for $n = 0$.

yields

$$\begin{aligned}&\sum_{k=0}^{m}(-1)^k\binom{m}{k}\binom{2m-2k}{m+p}\\&=\sum_{k=0}^{m-1}(-1)^k\binom{m-1}{k}\left(\binom{2m-2k}{m+p}-\binom{2m-2(k+1)}{m+p}\right).\end{aligned}\tag{780}$$

We shall now simplify the difference $\binom{2m-2k}{m+p}-\binom{2m-2(k+1)}{m+p}$ on the right side of this equality.

Indeed, it is easy to see that every $n\in\mathbb{R}$ and $i\in\mathbb{R}$ satisfy

$$\binom{n}{i}-\binom{n-2}{i}=\binom{n-2}{i-2}+2\binom{n-2}{i-1}\tag{781}$$

[447].

[447] *Proof of (781):* Let $n\in\mathbb{R}$ and $i\in\mathbb{R}$. Then, Theorem 4.3.7 (applied to $k=i$) yields

$$\begin{aligned}\binom{n}{i}&=\underbrace{\binom{n-1}{i-1}}_{\substack{=\binom{n-1-1}{i-1-1}+\binom{n-1-1}{i-1}\\ \text{(by Theorem 4.3.7,}\\ \text{applied to } n-1 \text{ and } i-1\\ \text{instead of } n \text{ and } k)}}+\underbrace{\binom{n-1}{i}}_{\substack{=\binom{n-1-1}{i-1}+\binom{n-1-1}{i}\\ \text{(by Theorem 4.3.7,}\\ \text{applied to } n-1 \text{ and } i\\ \text{instead of } n \text{ and } k)}}\\&=\binom{n-1-1}{i-1-1}+\binom{n-1-1}{i-1}+\binom{n-1-1}{i-1}+\binom{n-1-1}{i}\\&=\binom{n-1-1}{i-1-1}+2\binom{n-1-1}{i-1}+\binom{n-1-1}{i}\\&=\binom{n-2}{i-2}+2\binom{n-2}{i-1}+\binom{n-2}{i}\end{aligned}$$

(since $n-1-1=n-2$ and $i-1-1=i-2$). Subtracting $\binom{n-2}{i}$ from both sides of this equality, we obtain

$$\binom{n}{i}-\binom{n-2}{i}=\binom{n-2}{i-2}+2\binom{n-2}{i-1}.$$

Thus, (781) is proven.

Now, for each $k \in \mathbb{Z}$, we have

$$\begin{aligned}
&\binom{2m-2k}{m+p} - \binom{2m-2(k+1)}{m+p} \\
&= \binom{2m-2k}{m+p} - \binom{2m-2k-2}{m+p} \qquad (\text{since } 2m-2(k+1) = 2m-2k-2) \\
&= \binom{2m-2k-2}{m+p-2} + 2\binom{2m-2k-2}{m+p-1} \\
&\qquad (\text{by (781) (applied to } n = 2m-2k \text{ and } i = m+p)) \\
&= \binom{2(m-1)-2k}{(m-1)+(p-1)} + 2\binom{2(m-2)-2k}{(m-1)+p}. \qquad (782)
\end{aligned}$$

(since $m+p-2 = (m-1)+(p-1)$ and $m+p-1 = (m-1)+p$ and $2m-2k-$

$2 = 2(m-1) - 2k$). Hence, (780) becomes

$$\begin{aligned}
&\sum_{k=0}^{m} (-1)^k \binom{m}{k} \binom{2m-2k}{m+p} \\
&= \sum_{k=0}^{m-1} (-1)^k \binom{m-1}{k} \underbrace{\left(\binom{2m-2k}{m+p} - \binom{2m-2(k+1)}{m+p} \right)}_{\substack{= \binom{2(m-1)-2k}{(m-1)+(p-1)} + 2\binom{2(m-2)-2k}{(m-1)+p} \\ \text{(by (782))}}} \\
&= \sum_{k=0}^{m-1} \underbrace{(-1)^k \binom{m-1}{k} \left(\binom{2(m-1)-2k}{(m-1)+(p-1)} + 2\binom{2(m-2)-2k}{(m-1)+p} \right)}_{= (-1)^k \binom{m-1}{k} \binom{2(m-1)-2k}{(m-1)+(p-1)} + 2\cdot(-1)^k \binom{m-1}{k} \binom{2(m-2)-2k}{(m-1)+p}} \\
&= \sum_{k=0}^{m-1} \left((-1)^k \binom{m-1}{k} \binom{2(m-1)-2k}{(m-1)+(p-1)} + 2\cdot(-1)^k \binom{m-1}{k} \binom{2(m-2)-2k}{(m-1)+p} \right) \\
&= \underbrace{\sum_{k=0}^{m-1} (-1)^k \binom{m-1}{k} \binom{2(m-1)-2k}{(m-1)+(p-1)}}_{\substack{= 2^{m-1-(p-1)} \binom{m-1}{p-1} \\ \text{(by (779), applied to } p-1 \text{ instead of } p)}} \\
&\qquad + 2 \cdot \underbrace{\sum_{k=0}^{m-1} (-1)^k \binom{m-1}{k} \binom{2(m-2)-2k}{(m-1)+p}}_{\substack{= 2^{m-1-p} \binom{m-1}{p} \\ \text{(by (779))}}} \\
&= \underbrace{2^{m-1-(p-1)}}_{\substack{= 2^{m-p} \\ \text{(since } m-1-(p-1) = m-p)}} \binom{m-1}{p-1} + \underbrace{2 \cdot 2^{m-1-p}}_{\substack{= 2^{(m-1-p)+1} = 2^{m-p} \\ \text{(since } (m-1-p)+1 = m-p)}} \binom{m-1}{p} \\
&= 2^{m-p} \binom{m-1}{p-1} + 2^{m-p} \binom{m-1}{p} = 2^{m-p} \left(\binom{m-1}{p-1} + \binom{m-1}{p} \right).
\end{aligned}$$

Comparing this with

$$2^{m-p} \underbrace{\binom{m}{p}}_{\substack{= \binom{m-1}{p-1} + \binom{m-1}{p} \\ \text{(by Theorem 4.3.7,} \\ \text{applied to } n=m \text{ and } k=p)}} = 2^{m-p} \left(\binom{m-1}{p-1} + \binom{m-1}{p} \right),$$

we obtain

$$\sum_{k=0}^{m} (-1)^k \binom{m}{k} \binom{2m-2k}{m+p} = 2^{m-p} \binom{m}{p}.$$

Forget that we fixed p. We thus have shown that every $p \in \mathbb{Z}$ satisfies

$$\sum_{k=0}^{m} (-1)^k \binom{m}{k} \binom{2m-2k}{m+p} = 2^{m-p} \binom{m}{p}.$$

In other words, Exercise 8.4.2 holds for $n = m$. This completes the induction step. Thus, Exercise 8.4.2 is solved again. □

Our third solution to Exercise 8.4.2 is algebraic as well, but unlike our rather down-to-earth second solution above, it relies on several binomial identities (while avoiding the use of induction). The first one is the following identity ([19fco-hw3s, §6.6, Theorem 6.13]):

Lemma A.10.4. Let $p \in \mathbb{N}$ and $q \in \mathbb{R}$. Then,

$$\sum_{i=0}^{p} (-1)^i \binom{p}{i} \binom{x-i}{q} = \binom{x-p}{q-p} \qquad \text{for all } x \in \mathbb{R}.$$

References to the proof of Lemma A.10.4. Lemma A.10.4 is proved in [19fco-hw3s, §6.6, Theorem 6.13]. (The proof proceeds by first reducing it to the case $q \in \mathbb{N}$, which constitutes [19fco-hw3s, Exercise 6], and then solving the latter exercise in four different ways. The fourth solution is particularly of note, as it involves finite differences of integer sequences, which are a variation on the finite differences of polynomials we have seen in Example 5.4.2. It can easily be rewritten in terms of the latter.) □

Our next binomial identity is a mashup of Corollary 4.3.17 with the trinomial revision formula:

Lemma A.10.5. Let $n \in \mathbb{N}$ and $p \in \mathbb{Z}$. Then,

$$\sum_{k=0}^{n} \binom{n}{k} \binom{k}{p} = 2^{n-p} \binom{n}{p}.$$

Proof of Lemma A.10.5. We are in one of the following three cases:

Case 1: We have $p < 0$.

Case 2: We have $p > n$.

Case 3: We have neither $p < 0$ nor $p > n$.

Let us first consider Case 1. In this case, we have $p < 0$. Thus, $p \notin \mathbb{N}$. Hence, (118) (applied to $k = p$) yields $\binom{n}{p} = 0$. Hence, $2^{n-p} \underbrace{\binom{n}{p}}_{=0} = 0$. Comparing this with

$$\sum_{k=0}^{n} \binom{n}{k} \underbrace{\binom{k}{p}}_{\substack{=0 \\ \text{(by (118),} \\ \text{since } p \notin \mathbb{N})}} = \sum_{k=0}^{n} \binom{n}{k} 0 = 0,$$

we obtain $\sum_{k=0}^{n} \binom{n}{k} \binom{k}{p} = 2^{n-p} \binom{n}{p}$. Thus, Lemma A.10.5 is proved in Case 1.

Let us next consider Case 2. In this case, we have $p > n$. Hence, $\binom{n}{p} = 0$ (by Proposition 4.3.4, applied to $k = p$).

Let $k \in \{0, 1, \ldots, n\}$. Then, $k \leq n < p$ (since $p > n$), thus $p > k$. Hence, Proposition 4.3.4 (applied to k and p instead of n and k) yields $\binom{k}{p} = 0$. Hence, $\binom{n}{k} \underbrace{\binom{k}{p}}_{=0} = 0$.

Forget that we fixed k. We thus have proved the equality $\binom{n}{k} \binom{k}{p} = 0$ for each $k \in \{0, 1, \ldots, n\}$. Summing these equalities over all $k \in \{0, 1, \ldots, n\}$, we obtain $\sum_{k=0}^{n} \binom{n}{k} \binom{k}{p} = \sum_{k=0}^{n} 0 = 0$. Comparing this with $2^{n-p} \underbrace{\binom{n}{p}}_{=0} = 0$, we obtain $\sum_{k=0}^{n} \binom{n}{k} \binom{k}{p} = 2^{n-p} \binom{n}{p}$. Thus, Lemma A.10.5 is proved in Case 2.

Let us finally consider Case 3. In this case, we have neither $p < 0$ nor $p > n$. Hence, we have $p \geq 0$ (since we don't have $p < 0$) and $p \leq n$ (since we don't have $p > n$). From $p \leq n$, we obtain $n \geq p$, and thus $n - p \in \mathbb{N}$. Also, from $p \geq 0$, we obtain $-p \leq 0 \leq n - p$ (since $n - p \in \mathbb{N}$) and $p \in \mathbb{N}$. Now,

$$\begin{aligned} \sum_{k=0}^{n} \underbrace{\binom{n}{k} \binom{k}{p}}_{\substack{= \binom{n}{p} \binom{n-p}{k-p} \\ \text{(by Proposition 7.5.2} \\ \text{(applied to } a=k \text{ and } b=p))}} &= \sum_{k=0}^{n} \binom{n}{p} \binom{n-p}{k-p} \\ &= \binom{n}{p} \sum_{k=0}^{n} \binom{n-p}{k-p}. \end{aligned} \tag{783}$$

However, each $k \in \{-p, -p+1, \ldots, -1\}$ satisfies $k \notin \mathbb{N}$ (since $k \leq -1 < 0$) and thus

$$\binom{n-p}{k} = 0 \tag{784}$$

(by (118)). Now,

$$\begin{aligned}
\sum_{k=0}^{n} \binom{n-p}{k-p} &= \sum_{k=-p}^{n-p} \binom{n-p}{k} \qquad \text{(here, we have substituted } k \text{ for } k-p \text{ in the sum)} \\
&= \sum_{k=-p}^{-1} \underbrace{\binom{n-p}{k}}_{\substack{=0 \\ \text{(by (118)} \\ \text{(since } k \notin \mathbb{N}))}} + \underbrace{\sum_{k=0}^{n-p} \binom{n-p}{k}}_{\substack{=2^{n-p} \\ \text{(by Corollary 4.3.17} \\ \text{(applied to } n-p \text{ instead of } n))}} \\
&\qquad \text{(here, we have split the sum at } k=0 \text{ (since } -p \leq 0 \leq n-p)) \\
&= \underbrace{\sum_{k=-p}^{-1} 0}_{=0} + 2^{n-p} = 2^{n-p}.
\end{aligned}$$

Thus, (783) becomes

$$\sum_{k=0}^{n} \binom{n}{k}\binom{k}{p} = \binom{n}{p} \underbrace{\sum_{k=0}^{n} \binom{n-p}{k-p}}_{=2^{n-p}} = \binom{n}{p} 2^{n-p} = 2^{n-p}\binom{n}{p}.$$

Thus, Lemma A.10.5 is proved in Case 3.

We have now proved Lemma A.10.5 in all three Cases 1, 2 and 3. Thus, Lemma A.10.5 always holds. □

We need yet another simple binomial identity (a "symmetric" form of the trinomial revision formula for nonnegative integer arguments):

Lemma A.10.6. Let $n, i, k \in \mathbb{N}$. Then,

$$\binom{n}{i}\binom{n-i}{k} = \binom{n}{k}\binom{n-k}{i}.$$

Proof of Lemma A.10.6. Theorem 7.5.4 (applied to $a = n-i$ and $b = k$) yields

$$\binom{n}{n-i}\binom{n-i}{k} = \binom{n}{k}\binom{n-k}{(n-i)-k}. \tag{785}$$

However, Theorem 4.3.10 (applied to i instead of k) yields

$$\binom{n}{i} = \binom{n}{n-i}.$$

Multiplying both sides of this equality by $\binom{n-i}{k}$, we obtain

$$\begin{aligned}\binom{n}{i}\binom{n-i}{k} &= \binom{n}{n-i}\binom{n-i}{k} \\ &= \binom{n}{k}\binom{n-k}{(n-i)-k} \qquad \text{(by (785))}. \end{aligned} \tag{786}$$

Now, we are in one of the following two cases:
Case 1: We have $k \leq n$.
Case 2: We have $k > n$.

Let us first consider Case 1. In this case, we have $k \leq n$. Thus, $n-k \in \mathbb{N}$. Hence, Theorem 4.3.10 (applied to $n-k$ and i instead of n and k) yields

$$\binom{n-k}{i} = \binom{n-k}{(n-k)-i} = \binom{n-k}{(n-i)-k} \tag{787}$$

(since $(n-k)-i = (n-i)-k$). Now, (786) becomes

$$\binom{n}{i}\binom{n-i}{k} = \binom{n}{k}\underbrace{\binom{n-k}{(n-i)-k}}_{\substack{=\binom{n-k}{i} \\ \text{(by (787))}}} = \binom{n}{k}\binom{n-k}{i}.$$

Thus, Lemma A.10.6 is proved in Case 1.

Let us now consider Case 2. In this case, we have $k > n$. Thus, $\binom{n}{k} = 0$ (by Proposition 4.3.4). Now, (786) becomes

$$\binom{n}{i}\binom{n-i}{k} = \underbrace{\binom{n}{k}}_{=0}\binom{n-k}{(n-i)-k} = 0.$$

Comparing this with $\underbrace{\binom{n}{k}}_{=0}\binom{n-k}{i} = 0$, we obtain

$$\binom{n}{i}\binom{n-i}{k} = \binom{n}{k}\binom{n-k}{i}.$$

Thus, Lemma A.10.6 is proved in Case 2.

We have thus proved Lemma A.10.6 in both Cases 1 and 2. Thus, Lemma A.10.6 always holds. □

Third solution to Exercise 8.4.2 (sketched). It is easy to see that Exercise 8.4.2 holds if $n+p<0$ [448]. Thus, we WLOG assume that $n+p \geq 0$. Hence, $n+p \in \mathbb{N}$ (since $n+p$ is an integer).

[448] *Proof.* Assume that $n+p<0$. Then, $n+p \notin \mathbb{N}$. Hence, each $k \in \{0,1,\ldots,n\}$ satisfies $\binom{2n-2k}{n+p}=0$ (by (118)). Thus,

$$\sum_{k=0}^{n}(-1)^k\binom{n}{k}\underbrace{\binom{2n-2k}{n+p}}_{=0}=\sum_{k=0}^{n}(-1)^k\binom{n}{k}0=0. \tag{788}$$

On the other hand, $p=\underbrace{(n+p)}_{<0}-\underbrace{n}_{>0}<0$ and thus $p \notin \mathbb{N}$. Hence, $\binom{n}{p}=0$ (by (118)). Thus,

$$2^{n-p}\underbrace{\binom{n}{p}}_{=0}=0.$$

Comparing this with (788), we obtain $\sum_{k=0}^{n}(-1)^k\binom{n}{k}\binom{2n-2k}{n+p}=2^{n-p}\binom{n}{p}$. Hence, Exercise 8.4.2 is solved under the assumption that $n+p<0$.

Now, we have

$$\sum_{k=0}^{n}(-1)^k\binom{n}{k}\binom{2n-2k}{n+p}$$

$$=\sum_{i=0}^{n}(-1)^i\binom{n}{i}\underbrace{\binom{2n-2i}{n+p}}_{\substack{=\binom{(n-i)+(n-i)}{n+p}\\ =\sum_{k=0}^{n+p}\binom{n-i}{k}\binom{n-i}{n+p-k}\\ \text{(by Proposition 7.5.1,}\\ \text{applied to } n-i,\ n-i \text{ and } n+p \text{ instead of } x,\ y \text{ and } n)}}$$

$$=\sum_{i=0}^{n}(-1)^i\binom{n}{i}\underbrace{\sum_{k=0}^{n+p}\binom{n-i}{k}\binom{n-i}{n+p-k}}_{\substack{=\sum_{k=0}^{n}\binom{n-i}{k}\binom{n-i}{n+p-k}\\ \text{(here, we have changed the upper range of the summation;}\\ \text{indeed, it is not hard to see that all addends with } k>n\\ \text{are zero, as are all addends with } k>n+p\text{, so that the value}\\ \text{of the sum is not affected by our change)}}}$$

$$=\sum_{i=0}^{n}(-1)^i\binom{n}{i}\sum_{k=0}^{n}\binom{n-i}{k}\binom{n-i}{n+p-k}$$

$$=\sum_{k=0}^{n}\sum_{i=0}^{n}(-1)^i\underbrace{\binom{n}{i}\binom{n-i}{k}}_{\substack{=\binom{n}{k}\binom{n-k}{i}\\ \text{(by Lemma A.10.6)}}}\binom{n-i}{n+p-k}$$

$$=\sum_{k=0}^{n}\sum_{i=0}^{n}(-1)^i\binom{n}{k}\binom{n-k}{i}\binom{n-i}{n+p-k}$$

$$=\sum_{k=0}^{n}\binom{n}{k}\underbrace{\sum_{i=0}^{n}(-1)^i\binom{n-k}{i}\binom{n-i}{n+p-k}}_{\substack{=\sum_{i=0}^{n-k}(-1)^i\binom{n-k}{i}\binom{n-i}{n+p-k}\\ \text{(here, we removed all addends with } i>n-k\text{,}\\ \text{since these addends were 0 anyway)}}}$$

$$=\sum_{k=0}^{n}\binom{n}{k}\underbrace{\sum_{i=0}^{n-k}(-1)^i\binom{n-k}{i}\binom{n-i}{n+p-k}}_{\substack{=\binom{n-(n-k)}{(n+p-k)-(n-k)}\\ \text{(by Lemma A.10.4,}\\ \text{applied to } p=n-k \text{ and } q=n+p-k \text{ and } x=n)}}$$

$$= \sum_{k=0}^{n} \binom{n}{k} \underbrace{\binom{n-(n-k)}{(n+p-k)-(n-k)}}_{=\binom{k}{p}} = \sum_{k=0}^{n} \binom{n}{k}\binom{k}{p} = 2^{n-p}\binom{n}{p}$$

(by Lemma A.10.5). This solves Exercise 8.4.2 again. □

□

A.10.3. Discussion of Exercise 8.4.3 (TODO: add details!)

Discussion of Exercise 8.4.3. Exercise 8.4.3 is problem 11 from the IMO Shortlist 1991 (except that we have replaced the number 1991 by n). The key to the solution is the following variant (and corollary) of Proposition 4.9.18:

Proposition A.10.7. Let a and b be two numbers such that $a \neq 0$. Let $(x_0, x_1, x_2, \ldots)$ be an (a,b)-recurrent sequence with $x_0 = 0$ and $x_1 = 1$. Let $n \in \{1,2,3,\ldots\}$. Then,

$$\frac{1}{n}\left(x_{n+1} + bx_{n-1}\right) = \sum_{k=0}^{n-1} \frac{1}{n-k}\binom{n-k}{k} a^{n-2k} b^k.$$

Proof of Proposition A.10.7. We have $n > 0$ (since $n \in \{1,2,3,\ldots\}$). Thus, $\binom{0}{n} = 0$ (by Proposition 4.3.4, applied to 0 and n instead of n and k).

We have $n \in \{1,2,3,\ldots\} \subseteq \{-1,0,1,\ldots\}$. Thus, Proposition 4.9.18 yields

$$\begin{aligned}
x_{n+1} &= \sum_{k=0}^{n} \binom{n-k}{k} a^{n-2k} b^k \\
&= \sum_{k=0}^{n-1} \binom{n-k}{k} a^{n-2k} b^k + \underbrace{\binom{n-n}{n}}_{=\binom{0}{n}=0} a^{n-2n} b^n \\
&\qquad \text{(here, we have split off the addend for } k = n \text{ from the sum)} \\
&= \sum_{k=0}^{n-1} \binom{n-k}{k} a^{n-2k} b^k + \underbrace{0 a^{n-2n} b^n}_{=0} \\
&= \sum_{k=0}^{n-1} \binom{n-k}{k} a^{n-2k} b^k. \qquad (789)
\end{aligned}$$

From $n \in \{1,2,3,\ldots\}$, we obtain $n-2 \in \{-1,0,1,\ldots\}$. Hence, Proposition 4.9.18 (applied to $n-2$ instead of n) yields

$$\begin{aligned}
x_{(n-2)+1} &= \sum_{k=0}^{n-2} \binom{n-2-k}{k} a^{(n-2)-2k} b^k \\
&= \sum_{k=1}^{n-1} \underbrace{\binom{n-2-(k-1)}{k-1}}_{\substack{=\binom{n-k-1}{k-1} \\ \text{(since } n-2-(k-1)=n-k-1)}} \underbrace{a^{(n-2)-2(k-1)}}_{\substack{=a^{n-2k} \\ \text{(since } (n-2)-2(k-1)=n-2k)}} b^{k-1} \\
&\qquad \text{(here, we have substituted } k-1 \text{ for } k \text{ in the sum)} \\
&= \sum_{k=1}^{n-1} \binom{n-k-1}{k-1} a^{n-2k} b^{k-1}.
\end{aligned}$$

Multiplying both sides of this equality by b, we find

$$\begin{aligned}
bx_{(n-2)+1} &= b \sum_{k=1}^{n-1} \binom{n-k-1}{k-1} a^{n-2k} b^{k-1} = \sum_{k=1}^{n-1} \binom{n-k-1}{k-1} a^{n-2k} \underbrace{bb^{k-1}}_{=b^k} \\
&= \sum_{k=1}^{n-1} \binom{n-k-1}{k-1} a^{n-2k} b^k.
\end{aligned}$$

Comparing this with

$$\begin{aligned}
\sum_{k=0}^{n-1} \binom{n-k-1}{k-1} a^{n-2k} b^k &= \underbrace{\binom{n-0-1}{0-1}}_{\substack{=0 \\ \text{(by (118)} \\ \text{(since } 0-1 \notin \mathbb{N}))}} a^{n-2\cdot 0} b^0 + \sum_{k=1}^{n-1} \binom{n-k-1}{k-1} a^{n-2k} b^k \\
&\qquad \text{(here, we have split off the addend for } k=0 \text{ from the sum)} \\
&= \underbrace{0 a^{n-2\cdot 0} b^0}_{=0} + \sum_{k=1}^{n-1} \binom{n-k-1}{k-1} a^{n-2k} b^k \\
&= \sum_{k=1}^{n-1} \binom{n-k-1}{k-1} a^{n-2k} b^k,
\end{aligned}$$

we obtain

$$\sum_{k=0}^{n-1} \binom{n-k-1}{k-1} a^{n-2k} b^k = bx_{(n-2)+1} = bx_{n-1} \tag{790}$$

(since $(n-2)+1 = n-1$).

However, every $k \in \{0,1,\ldots,n-1\}$ satisfies

$$\frac{1}{n-k}\binom{n-k}{k} = \frac{1}{n}\left(\binom{n-k}{k} + \binom{n-k-1}{k-1}\right). \tag{791}$$

[*Proof of (791):* Let $k \in \{0, 1, \ldots, n-1\}$. Then, $k \leq n-1 < n$, so that $k \neq n$ and thus $n-k \neq 0$. Thus, we can divide by $n-k$. Also, we can divide by n (since $n > k \geq 0$ and thus $n \neq 0$).

Now, Lemma A.4.9 (applied to $n-k$ instead of n) yields

$$(n-k)\binom{n-k-1}{k-1} = k\binom{n-k}{k}. \tag{792}$$

Now,

$$\begin{aligned}
\frac{1}{n-k}\binom{n-k}{k} - \frac{1}{n}\binom{n-k}{k} &= \underbrace{\left(\frac{1}{n-k} - \frac{1}{n}\right)}_{=\frac{k}{n(n-k)}}\binom{n-k}{k} = \frac{k}{n(n-k)}\binom{n-k}{k} \\
&= \frac{1}{n(n-k)} \underbrace{k\binom{n-k}{k}}_{\substack{=(n-k)\binom{n-k-1}{k-1} \\ \text{(by (792))}}} \\
&= \frac{1}{n(n-k)}(n-k)\binom{n-k-1}{k-1} = \frac{1}{n}\binom{n-k-1}{k-1}.
\end{aligned}$$

Adding $\frac{1}{n}\binom{n-k}{k}$ to both sides of this equality, we find

$$\frac{1}{n-k}\binom{n-k}{k} = \frac{1}{n}\binom{n-k}{k} + \frac{1}{n}\binom{n-k-1}{k-1} = \frac{1}{n}\left(\binom{n-k}{k} + \binom{n-k-1}{k-1}\right).$$

This proves (791).]

Now,

$$\begin{aligned}
&\sum_{k=0}^{n-1} \underbrace{\frac{1}{n-k}\binom{n-k}{k}}_{\substack{=\frac{1}{n}\left(\binom{n-k}{k}+\binom{n-k-1}{k-1}\right) \\ \text{(by (791))}}} a^{n-2k}b^k \\
&= \sum_{k=0}^{n-1} \frac{1}{n}\left(\binom{n-k}{k}+\binom{n-k-1}{k-1}\right) a^{n-2k}b^k \\
&= \frac{1}{n}\sum_{k=0}^{n-1}\left(\binom{n-k}{k}+\binom{n-k-1}{k-1}\right) a^{n-2k}b^k \\
&= \frac{1}{n}\underbrace{\sum_{k=0}^{n-1}\binom{n-k}{k}a^{n-2k}b^k}_{\substack{=x_{n+1} \\ \text{(by (789))}}} + \frac{1}{n}\underbrace{\sum_{k=0}^{n-1}\binom{n-k-1}{k-1}a^{n-2k}b^k}_{\substack{=bx_{n-1} \\ \text{(by (790))}}} \\
&= \frac{1}{n}x_{n+1} + \frac{1}{n}bx_{n-1} = \frac{1}{n}\left(x_{n+1}+bx_{n-1}\right).
\end{aligned}$$

This proves Proposition A.10.7. □

Now, solving Exercise 8.4.3 is a matter of finding the right (a,b)-recurrent sequence:

Solution to Exercise 8.4.3 (sketched). Set $a=1$ and $b=-1$. Thus, $a=1\neq 0$.

Consider the sequence $(g_0, g_1, g_2, \ldots)$ defined in Example 4.7.7. This sequence $(g_0, g_1, g_2, \ldots)$ is $(1,-1)$-recurrent. In other words, it is (a,b)-recurrent (since $a=1$ and $b=-1$). Moreover, it satisfies $g_0=0$ and $g_1=1$. Thus, Proposition A.10.7 (applied to $x_i = g_i$) yields

$$\begin{aligned}
\frac{1}{n}\left(g_{n+1}+bg_{n-1}\right) &= \sum_{k=0}^{n-1}\frac{1}{n-k}\binom{n-k}{k}\underbrace{a^{n-2k}}_{\substack{=1^{n-2k} \\ \text{(since } a=1)}}\underbrace{b^k}_{\substack{=(-1)^k \\ \text{(since } b=-1)}} \\
&= \sum_{k=0}^{n-1}\frac{1}{n-k}\binom{n-k}{k}\underbrace{1^{n-2k}}_{=1}(-1)^k = \sum_{k=0}^{n-1}\frac{(-1)^k}{n-k}\binom{n-k}{k} \\
&= \sum_{i=0}^{n-1}\frac{(-1)^i}{n-i}\binom{n-i}{i}.
\end{aligned}$$

Hence,

$$\sum_{i=0}^{n-1}\frac{(-1)^i}{n-i}\binom{n-i}{i} = \frac{1}{n}\left(g_{n+1}+\underbrace{b}_{=-1}g_{n-1}\right) = \frac{1}{n}\left(g_{n+1}-g_{n-1}\right).$$

Thus, in order to solve Exercise 8.4.3, it suffices to prove the equality

$$g_{n+1} - g_{n-1} = (-1)^{(n+1)//3} (1 + [3 \mid n]) . \tag{793}$$

But this is entirely straightforward: Both sides of this equality depend only on the remainder $n\%6$ (not on n itself), because the sequence $(g_0, g_1, g_2, \ldots)$ is 6-periodic (as we have seen in Example 4.7.7). Thus, in order to prove it for all positive integers n, it suffices to prove it for all $n \in \{1, 2, 3, 4, 5, 6\}$ only (since these 6 values of n represent all possible remainders that $n\%6$ can take). But this can be done mechanically.[449] This solves Exercise 8.4.3. □

□

A.10.4. Discussion of Exercise 8.4.4 (TODO: add details!)

Discussion of Exercise 8.4.4. Exercise 8.4.4 is a quickie. The answers to both parts **(a)** and **(b)** are "$mn - 1$ moves". No matter how you are breaking up the chocolate, you will always have it broken down into 1×1-pieces after $mn - 1$ moves (no fewer, no more).

Indeed, any time you make a move, the # of pieces of chocolate increases by 1. Thus, it will take precisely $mn - 1$ moves to get this # from its initial value 1 to its final value mn (which it has to take if we have broken up the entire bar into 1×1-squares; after all, the total area of the chocolate is invariant). □

[449] Here are a few more details: Let us forget that we fixed n. Thus, we must prove (793) for all positive integers n. In other words, we must prove that the two sequences

$$(g_{n+1} - g_{n-1})_{n \in \{1,2,3,\ldots\}} = (g_2 - g_0, g_3 - g_1, g_4 - g_2, g_5 - g_3, g_6 - g_4, \ldots)$$

and

$$\begin{aligned} &\left((-1)^{(n+1)//3} (1 + [3 \mid n])\right)_{n \in \{1,2,3,\ldots\}} \\ &= \left((-1)^{2//3} (1 + [3 \mid 1]), (-1)^{3//3} (1 + [3 \mid 2]), (-1)^{4//3} (1 + [3 \mid 3]), \ldots\right) \end{aligned}$$

are identical. But both of these two sequences are 6-periodic (indeed, the sequence $(g_{n+1} - g_{n-1})_{n \in \{1,2,3,\ldots\}}$ is 6-periodic because the sequence $(g_0, g_1, g_2, \ldots)$ is 6-periodic; meanwhile, the sequence $\left((-1)^{(n+1)//3} (1 + [3 \mid n])\right)_{n \in \{1,2,3,\ldots\}}$ is 6-periodic because it is straightforward to see that

$$(-1)^{(n+1)//3} = \begin{cases} 1, & \text{if } n\%6 \in \{0, 1, 5\}; \\ -1, & \text{if } n\%6 \in \{2, 3, 4\} \end{cases} \qquad \text{and}$$

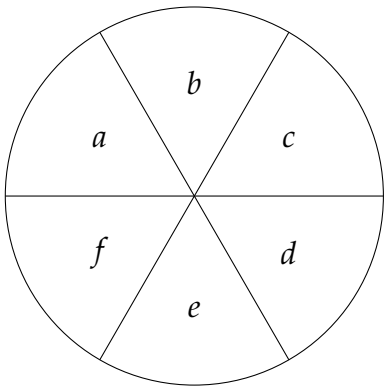
$$1 + [3 \mid n] = \begin{cases} 1, & \text{if } n\%6 \in \{1, 2, 4, 5\}; \\ 2, & \text{if } n\%6 \in \{0, 3\} \end{cases}$$

for each $n \in \mathbb{Z}$). Hence, in order to prove that these two sequences are identical, it suffices to show that they agree in their first 6 entries. But this is mechanical.

A.10.5. Discussion of Exercise 8.4.5 (TODO: add details!)

Discussion of Exercise 8.4.5. Exercise 8.4.5 is a simple application of invariants.

Indeed, if a state has the numbers a, b, c, d, e, f written into the 6 sectors as follows:



,

then we define the *fingerprint* of this state as the number $a - b + c - d + e - f$.

It is easy to see that each move leaves the fingerprint of the state unchanged (since, for example,

$$(a+1) - (b+1) + c - d + e - f = a - b + c - d + e - f$$

for any numbers a, b, c, d, e, f). The fingerprint of the initial state (in which the numbers are $1, 0, 1, 0, 0, 0$) is $1 - 0 + 1 - 0 + 0 - 0 = 2$; thus, it remains 2 throughout the process. Hence, it can never become 0. This entails that the numbers can never all become equal (because if all numbers are equal, then the fingerprint of the state is $a - a + a - a + a - a = 0$). In other words, you cannot ensure that all six sectors have the same number written in them. This solves Exercise 8.4.5.

Exercise 8.4.5 is [Engel98, Chapter 1, Example E3]. □

A.10.6. Discussion of Exercise 8.4.6 (TODO: add details!)

Discussion of Exercise 8.4.6. The word "maximum" is a red herring. In truth, no matter what moves you make, the number of cents you will have gained by the end of the procedure (i.e., by the time you are left with n heaps) is always $\binom{n}{2}$. Here are two ways to prove this:

First proof: Apply strong induction on n. Let's say that in the first move, you split the original heap (with n chips) into one heap with a chips and one heap with b chips; thus you gain ab cents. Note that $a + b = n$ and $a < n$ and $b < n$. Then you keep making moves, until you have disassembled both heaps into many one-chip heaps. By the induction hypothesis, we know that you gain $\binom{a}{2}$ cents from disassembling the a-chip heap (since $a < n$) and that you gain $\binom{b}{2}$ cents from disassembling the b-chip heap (since $b < n$). Hence, in total, you gain

$ab + \binom{a}{2} + \binom{b}{2}$ cents. But since $ab + \binom{a}{2} + \binom{b}{2} = \binom{n}{2}$ (check this!), this shows that you gain $\binom{n}{2}$ cents. This completes the induction step.

Second proof: Let me define a *dispersed couple* to mean a set $\{u, v\}$ of two chips that lie in different heaps. Of course, this concept depends on the state. In the initial state, there are no dispersed couples, since all chips lie in the same heap. In the end state (when you are left with n heaps, each containing exactly one chip), there are precisely $\binom{n}{2}$ dispersed couples (since any two distinct chips form a dispersed couple). Thus, any set $\{u, v\}$ of two chips becomes a dispersed couple in one of your moves (and then remains a dispersed couple forever, since there are no moves that would bring chips from different heaps back together). Furthermore, the number of cents you gain from a move is precisely the number of dispersed couples that are created by this move (i.e., the number of sets $\{u, v\}$ of two chips that were not a dispersed couple before the move, but are a dispersed couple after the move). Thus, the total number of cents you gain during the game is the total number of dispersed couples created by the end of the game. But this latter number is $\binom{n}{2}$ (since the initial state has no dispersed couples, whereas the end state has $\binom{n}{2}$ of them). This, again, proves our claim. □

A.10.7. Discussion of Exercise 8.4.7 (TODO: add details!)

Discussion of Exercise 8.4.7. Exercise 8.4.7 is Exercise 1 on the USAJMO 2019. For now, I refer there (or to `https://artofproblemsolving.com/wiki/index.php/2019_USAJMO_Problems/Problem_1` or to `https://www.maa.org/sites/default/files/pdf/AMC/usamo/2019/2019-USAJMO-Solutions.pdf`) for the solution. □

A.10.8. Discussion of Exercise 8.4.8 (TODO: add details!)

Discussion of Exercise 8.4.8. Exercise 8.4.8 is not just superficially similar to Exercise 8.2.7; it can also be solved by the same kind of reasoning. I shall outline the solution briefly, mostly just focussing on what is different:

Solution to Exercise 8.4.8 (sketched). Let V be the set of all violinists. We identify any state with the finite subset

$$\{(i, v) \mid i \in \mathbb{Z}, \text{ and } v \text{ is a violinist staying in room } i\}$$

of $\mathbb{Z} \times V$. Thus, states are finite subsets of $\mathbb{Z} \times V$. For example, the state shown in (400) is the subset

$$\{(3, a), (3, b), (4, c), (5, d), (5, e)\}$$

of $\mathbb{Z} \times V$ (assuming that the cells shown in (400) correspond to the rooms $1, 2, \ldots, 11$). For another example, the subset $\{(1, \alpha), (1, \beta), (4, \gamma)\}$ of $\mathbb{Z} \times V$ (with α, β, γ being three distinct violinists) is the state in which the two violinists α and β are staying in room 1 and the violinist γ is staying in room 4 (and there are no other violinists in the hotel).

Thus, any move removes two pairs of the form (i, α) and (i, β) from the state and inserts the two pairs $(i-1, \alpha)$ and $(i+1, \beta)$ into the state. (Here, of course, α is the violinist that moves to room $i-1$, and β is the violinist that moves to room $i+1$.)

If $S \subseteq \mathbb{Z} \times V$ is any state, then $\operatorname{Occ} S$ shall denote the set of all occupied rooms in S. Thus,

$$\operatorname{Occ} S = \{i \mid (i, v) \in S\}.$$

This is a finite subset of $\mathbb{Z}$.

We define the *entropy* of a state S to be the sum $\sum_{(i,v) \in S} 2^i$ (recalling that S is a finite subset of $\mathbb{Z} \times V$). This is a rational number (since the sum is finite).

It is easy to see that this entropy increases each night. In other words:

> *Claim 1:* Whenever two violinists move apart, the entropy of the state increases.

[*Proof of Claim 1:* Similar to the proof of Claim 1 in the solution to Exercise 8.2.7.]

Once again (as in the solution to Exercise 8.2.7), this is useful but not sufficient to solve the exercise. Again, we need to show that only finitely many possible states can be reached.

Let us do this. We let N be the # of violinists in the hotel. (Clearly, this # does not change during the process.) We shall show that the violinists never move "too far" away from their original rooms; more precisely, they will always stay within $2N$ rooms of the interval between the leftmost and the rightmost room occupied in the initial state. We shall now make this precise.

Let S_0 be the initial state. Consider a sequence of g (successive) moves, starting from state S_0 and leading to states $S_1, S_2, S_3, \ldots, S_g$ in this order (i.e., the first move transforms state S_0 into state S_1; the next move transforms state S_1 into state S_2; and so on). As in the solution to Exercise 8.2.7, we have

$$N = |S_0| = |S_1| = \cdots = |S_g|. \tag{794}$$

We WLOG assume that $N > 0$. Thus, the sets $S_0, S_1, \ldots, S_g$ are nonempty finite sets (because (794) shows that each of them has N elements); therefore, so are the sets $\operatorname{Occ}(S_0), \operatorname{Occ}(S_1), \ldots, \operatorname{Occ}(S_g)$. Hence, the minima and maxima of these latter sets $\operatorname{Occ}(S_0), \operatorname{Occ}(S_1), \ldots, \operatorname{Occ}(S_g)$ are well-defined. Let $\alpha = \min(\operatorname{Occ}(S_0))$ and $\omega = \max(\operatorname{Occ}(S_0))$. (Note that the sets $\operatorname{Occ}(S_0), \operatorname{Occ}(S_1), \ldots, \operatorname{Occ}(S_g)$ are going to play some of the roles that the sets $S_0, S_1, \ldots, S_g$ used to play in our solution to Exercise 8.2.7. In the latter solution, violinists could not be roommates, which is why we didn't have to introduce $\operatorname{Occ} S$.)

Now, we shall see that the violinists don't spread "too fast" through the hotel. To be more specific, we claim the following:

Claim 2: We have $\min(\operatorname{Occ}(S_{i+1})) \geq \min(\operatorname{Occ}(S_i)) - 1$ for each $i \in \{0, 1, \ldots, g-1\}$.

[*Proof of Claim 2:* This is similar to proving Claim 2 in the solution of Exercise 8.2.7, but easier, since no violinist moves by more than 1 room in a single night.]

We next introduce another notation. If S is any finite subset of $\mathbb{Z}$, then we define the subset S^{0+} of $\mathbb{Z}$ by

$$S^{0+} = S \cup \{s+1 \mid s \in S\} = \{i \in \mathbb{Z} \mid i \in S \text{ or } i-1 \in S\}.$$

In other words, S^{0+} is the set of all integers i that belong to S themselves or have their left neighbor $i-1$ belong to S. Clearly, for any finite subset S of $\mathbb{Z}$, we have

$$\left|S^{0+}\right| \leq 2 \cdot |S| \tag{795}$$

(this follows easily from (277)) and

$$S \subseteq S^{0+}. \tag{796}$$

Our next claim shows that if at least one of two consecutive rooms is occupied at some point, then this property will remain valid in all future states (even though it will not always be the same room that is occupied, or even the same # of rooms):

Claim 3: We have $(\operatorname{Occ}(S_i))^{0+} \subseteq (\operatorname{Occ}(S_{i+1}))^{0+}$ for each $i \in \{0, 1, \ldots, g-1\}$.

[*Proof of Claim 3:* This is similar to proving Claim 3 in the solution of Exercise 8.2.7, but easier, since we only need to take care of two (not three) adjacent rooms.]

Let us draw some conclusions from Claim 3. It is **not** true that $\operatorname{Occ}(S_0) \subseteq \operatorname{Occ}(S_1) \subseteq \cdots \subseteq \operatorname{Occ}(S_g)$, since an occupied room can become unoccupied after a move. However, Claim 3 shows that we have

$$(\operatorname{Occ}(S_0))^{0+} \subseteq (\operatorname{Occ}(S_1))^{0+} \subseteq \cdots \subseteq \left(\operatorname{Occ}(S_g)\right)^{0+}. \tag{797}$$

In words, this is saying that if a room or the next room to its right is occupied at some time, then it will always remain the case that this room or the next room is occupied.

In the rest of this solution, we shall use the notation $[p, q]$ for the integer interval $\{p, p+1, \ldots, q\}$, just as we did in the solution of Exercise 8.2.7.

Our next claim will show that the leftmost occupied room will always remain to the right of the room $\alpha - 2N$ (so it cannot "wander off" to the left too far):

Claim 4: Let $m \in \{0, 1, \ldots, g\}$. Then, $\min(\operatorname{Occ}(S_m)) > \alpha - 2N$.

[*Proof of Claim 4:* This is similar to proving Claim 4 in the solution of Exercise 8.2.7. This time, we need to use $|\operatorname{Occ}(S_m)| \leq |S_m| = N$.]

Claim 5: Let $m \in \{0, 1, \ldots, g\}$. Then, $\max(\operatorname{Occ}(S_m)) < \omega + 2N$.

[*Proof of Claim 5:* This is similar to proving Claim 5 in the solution of Exercise 8.2.7.]

Now, Claim 4 and Claim 5 can be combined to the following:

Claim 6: Let $m \in \{0, 1, \ldots, g\}$. Then, $S_m \subseteq [\alpha - 2N + 1, \omega + 2N - 1] \times V$.

[*Proof of Claim 6:* Using Claim 4 and Claim 5, it is easy to see that $\operatorname{Occ}(S_m) \subseteq [\alpha - 2N + 1, \omega + 2N - 1]$ (indeed, this can be proved just as we showed Claim 6 in the solution of Exercise 8.2.7). However, it is easy to see that $S \subseteq (\operatorname{Occ} S) \times V$ for any state $S \subseteq \mathbb{Z} \times V$ (indeed, this follows from the definition of $\operatorname{Occ} S$). Applying this to $S = S_m$, we obtain

$$S_m \subseteq \underbrace{\operatorname{Occ}(S_m)}_{\subseteq [\alpha - 2N + 1, \omega + 2N - 1]} \times V \subseteq [\alpha - 2N + 1, \omega + 2N - 1] \times V.$$

This proves Claim 6.]

We note that $\omega = \max(\operatorname{Occ}(S_0)) \geq \min(\operatorname{Occ}(S_0)) = \alpha$ and thus $\underbrace{\omega}_{\geq \alpha} + \underbrace{2N - 1}_{\substack{> -2N+1 \\ (\text{since } N > 0)}} > \alpha - 2N + 1$. Hence, the interval $[\alpha - 2N + 1, \omega + 2N - 1]$ is nonempty, and its size is

$$|[\alpha - 2N + 1, \omega + 2N - 1]| = (\omega + 2N - 1) - (\alpha - 2N + 1) + 1 = \omega - \alpha + 4N - 1.$$

Now, we can use (e.g.) the pigeonhole principle: Claim 1 entails that the entropy of the state increases with every move. Thus,

$$(\text{the entropy of } S_0) < (\text{the entropy of } S_1) < \cdots < \left(\text{the entropy of } S_g\right).$$

Hence, the entropies of the $g + 1$ states $S_0, S_1, \ldots, S_g$ are distinct. Therefore, these $g + 1$ states $S_0, S_1, \ldots, S_g$ must themselves be distinct. However, all these $g + 1$ states $S_0, S_1, \ldots, S_g$ are subsets of the set $[\alpha - 2N + 1, \omega + 2N - 1] \times V$ (by Claim 6). Hence, we have found $g + 1$ distinct subsets of the set $[\alpha - 2N + 1, \omega + 2N - 1] \times V$. By the pigeonhole principle, this entails that

$$\begin{aligned} g + 1 &\leq (\text{\# of all subsets of } [\alpha - 2N + 1, \omega + 2N - 1] \times V) \\ &= 2^{|[\alpha - 2N + 1, \omega + 2N - 1] \times V|} = 2^{|[\alpha - 2N + 1, \omega + 2N - 1]| \cdot |V|} = 2^{(\omega - \alpha + 4N - 1) \cdot |V|} \\ &\qquad (\text{since } |[\alpha - 2N + 1, \omega + 2N - 1]| = \omega - \alpha + 4N - 1). \end{aligned}$$

In other words, $g \leq 2^{(\omega - \alpha + 4N - 1) \cdot |V|} - 1$.

Now, forget that we fixed g and $S_1, S_2, \ldots, S_g$. We thus have shown that any sequence of g (successive) moves, starting from state S_0, must satisfy $g \leq 2^{(\omega - \alpha + 4N - 1) \cdot |V|} - 1$. In other words, there is no sequence of (successive) moves, starting from state S_0, that has more than $2^{(\omega - \alpha + 4N - 1) \cdot |V|} - 1$ moves. In other words, the moves cannot go on for more than $2^{(\omega - \alpha + 4N - 1) \cdot |V|}$ nights (if the initial state is S_0). Hence, the moving will stop after a finite number of days. This solves Exercise 8.4.8. $\square$

Let me mention a major difference between Exercise 8.4.8 and Exercise 8.2.7: In Exercise 8.4.8, if we treat the violinists as indistinguishable, then the final state (i.e., the set of rooms that are occupied after the moving has stopped) is independent of the moves! That is, it does not matter which pairs of violinists choose to move apart first; in all possible scenarios, the rooms that end up occupied at the end will be the same. This result (which differs dramatically from the behavior of final states in Exercise 8.2.7) appears, e.g., in [Klivan18, Proposition 5.2.1] (in much greater generality); it is part of the theory of the chip-firing game (also known as the abelian sandpile model) on graphs. There are now two books ([CorPer18], [Klivan18]) out on this theory; in particular, [Klivan18, §5.3] explicitly describes the final state in Exercise 8.4.8 in the case when all violinists start out in the same room. □

A.10.9. Discussion of Exercise 8.4.9 (TODO: add details!)

Discussion of Exercise 8.4.9. Exercise 8.4.9 is a particular case of the following more general exercise (which, in turn, can be regarded as a particular case of [AlCuHu16, Theorem 3.10, "extreme polynomial"]):

Exercise A.10.1. Fix a positive integer $k \geq 2$ and two integers $u \in \mathbb{N}$ and $v \in \mathbb{N}$. Define a sequence $(a_0, a_1, a_2, \ldots)$ of positive rational numbers recursively by setting

$$a_n = 1 \qquad \text{for each } n < k \tag{798}$$

and

$$a_n = \frac{a_{n-1}a_{n-k+1} + u\left(a_{n-1} + a_{n-2} + \cdots + a_{n-k+1}\right) + v}{a_{n-k}} \qquad \text{for each } n \geq k. \tag{799}$$

Prove that a_n is a positive integer for each integer $n \geq 0$.

Solution to Exercise A.10.1 (sketched). This is similar to the above solution to Exercise 8.1.8, so we restrict ourselves to an outline.

Again, we begin by playing around with the recursive equation.

Let n be an integer satisfying $n \geq k+1$. Thus, $n \geq n-1 \geq k$ (since $n \geq k+1$); hence, (799) holds. Multiplying both sides of (799) by a_{n-k}, we obtain

$$a_n a_{n-k} = a_{n-1}a_{n-k+1} + u\left(a_{n-1} + a_{n-2} + \cdots + a_{n-k+1}\right) + v. \tag{800}$$

The same argument (applied to $n-1$ instead of n) yields

$$a_{n-1}a_{n-k-1} = a_{n-2}a_{n-k} + u\left(a_{n-2} + a_{n-3} + \cdots + a_{n-k}\right) + v.$$

Subtracting this equality from (800), we find

$$\begin{aligned} a_n a_{n-k} - a_{n-1} a_{n-k-1} &= \left(a_{n-1} a_{n-k+1} + u\left(a_{n-1} + a_{n-2} + \cdots + a_{n-k+1}\right) + v\right) \\ &\quad - \left(a_{n-2} a_{n-k} + u\left(a_{n-2} + a_{n-3} + \cdots + a_{n-k}\right) + v\right) \\ &= a_{n-1} a_{n-k+1} - a_{n-2} a_{n-k} + u a_{n-1} - u a_{n-k}. \end{aligned}$$

Adding $a_{n-2} a_{n-k} + u a_{n-k} + a_{n-1} a_{n-k-1}$ to both sides of this equality, we obtain

$$a_n a_{n-k} + a_{n-2} a_{n-k} + u a_{n-k} = a_{n-1} a_{n-k-1} + a_{n-1} a_{n-k+1} + u a_{n-1}.$$

In other words,

$$a_{n-k}\left(a_n + a_{n-2} + u\right) = a_{n-1}\left(a_{n-k+1} + a_{n-k-1} + u\right).$$

Dividing both sides of this equality by $a_{n-1} a_{n-k}$, we obtain

$$\frac{a_n + a_{n-2} + u}{a_{n-1}} = \frac{a_{n-k+1} + a_{n-k-1} + u}{a_{n-k}}. \tag{801}$$

Now, forget that we fixed n. We thus have proved the equality (801) for each integer $n \geq k+1$.

Now, let us define a number

$$b_n = \frac{a_n + a_{n-2} + u}{a_{n-1}} \tag{802}$$

for each integer $n \geq 2$. Thus, we have defined a sequence $(b_2, b_3, b_4, \ldots)$ of rational numbers. It is easy to see that $b_2, b_3, \ldots, b_k$ are integers (since $a_{k-1}, a_{k-2}, \ldots, a_0$ are all equal to 1, and thus all denominators involved in computing $b_2, b_3, \ldots, b_k$ are 1s).

Now, the equality (801) entails that

$$b_n = b_{n-(k-1)}$$

for each integer $n \geq k+1$ (because its left hand side is b_n, while its right hand side is $b_{n-(k-1)}$). In other words, $b_{m+(k-1)} = b_m$ for each integer $m \geq 2$. In other words, the sequence $(b_2, b_3, b_4, \ldots)$ is $(k-1)$-periodic. Thus, each entry of this sequence equals one of the first $k-1$ entries $b_2, b_3, \ldots, b_k$ of this sequence. Since the first $k-1$ entries $b_2, b_3, \ldots, b_k$ of this sequence are integers (as we have seen above), we thus conclude that each entry of this sequence is an integer. In other words, $b_2, b_3, b_4, \ldots$ are integers.

From here on, we argue similarly to our above solution to Exercise 8.1.8: Solving the equality (802) for a_n, we obtain

$$a_n = b_n a_{n-1} - a_{n-2} - u \qquad \text{for each integer } n \geq 2.$$

Using this equality (and the fact that $b_2, b_3, b_4, \ldots$ are integers), we can prove (by a straightforward strong induction on n) that a_n is a positive integer for each $n \in \mathbb{N}$ (because the induction hypothesis yields that a_{n-1} and a_{n-2} are integers, so that $b_n a_{n-1} - a_{n-2} - u$ is an integer, and the positivity follows from the statement of the exercise). Thus, Exercise A.10.1 is solved. $\square$

Exercise 8.4.9 is a particular case of Exercise A.10.1. (Namely, if we set $k = 4$ and $u = 0$ and $v = 1$ in Exercise A.10.1, then the sequence $(a_0, a_1, a_2, \ldots)$ becomes the sequence $(t_0, t_1, t_2, \ldots)$ from Exercise 8.4.9.) $\square$

A.10.10. Discussion of Exercise 8.4.10 (TODO: add details!)

Discussion of Exercise 8.4.10. This is similar to the above solution to Exercise A.10.1, so we restrict ourselves to an outline.

Let n be an integer satisfying $n \geq 4$. The definition of the sequence $(t_0, t_1, t_2, \ldots)$ yields

$$t_n = \frac{t_{n-1}^2 + qt_{n-1}t_{n-2} + t_{n-2}^2}{t_{n-3}},$$

so that

$$t_n t_{n-3} = t_{n-1}^2 + qt_{n-1}t_{n-2} + t_{n-2}^2.$$

The same argument (applied to $n-1$ instead of n) yields

$$t_{n-1}t_{n-4} = t_{n-2}^2 + qt_{n-2}t_{n-3} + t_{n-3}^2.$$

Subtracting this equality from the previous one, we find

$$\begin{aligned} t_n t_{n-3} - t_{n-1}t_{n-4} &= \left(t_{n-1}^2 + qt_{n-1}t_{n-2} + t_{n-2}^2\right) - \left(t_{n-2}^2 + qt_{n-2}t_{n-3} + t_{n-3}^2\right) \\ &= t_{n-1}^2 + qt_{n-1}t_{n-2} - qt_{n-2}t_{n-3} - t_{n-3}^2. \end{aligned}$$

Adding $qt_{n-2}t_{n-3} + t_{n-3}^2 + t_{n-1}t_{n-4} + qt_{n-1}t_{n-3}$ to both sides of this equality, we obtain

$$\begin{aligned} & t_n t_{n-3} + qt_{n-2}t_{n-3} + t_{n-3}^2 + qt_{n-1}t_{n-3} \\ &= t_{n-1}^2 + qt_{n-1}t_{n-2} + t_{n-1}t_{n-4} + qt_{n-1}t_{n-3}. \end{aligned}$$

In other words,

$$t_{n-3}\left(t_n + qt_{n-1} + qt_{n-2} + t_{n-3}\right) = t_{n-1}\left(t_{n-1} + qt_{n-2} + qt_{n-3} + t_{n-4}\right).$$

Dividing both sides of this equality by $t_{n-1}t_{n-2}t_{n-3}$, we obtain

$$\frac{t_n + qt_{n-1} + qt_{n-2} + t_{n-3}}{t_{n-1}t_{n-2}} = \frac{t_{n-1} + qt_{n-2} + qt_{n-3} + t_{n-4}}{t_{n-2}t_{n-3}}. \tag{803}$$

Now, forget that we fixed n. We thus have proved the equality (803) for each integer $n \geq 4$.

Now, let us define a number

$$b_n = \frac{t_n + qt_{n-1} + qt_{n-2} + t_{n-3}}{t_{n-1}t_{n-2}}$$

for each integer $n \geq 3$. Thus, we have defined a sequence $(b_3, b_4, b_5, \ldots)$ of rational numbers. It is easy to see that b_3 is an integer (since $t_2 = 1$ and $t_1 = 1$ and $t_3 \in \mathbb{Z}$).

Now, the equality (803) entails that

$$b_n = b_{n-1}$$

for each integer $n \geq 4$ (because its left hand side is b_n, while its right hand side is b_{n-1}). In other words, $b_3 = b_4 = b_5 = \cdots$. Hence, all entries $b_3, b_4, b_5, \ldots$ of the sequence $(b_3, b_4, b_5, \ldots)$ are integers (since b_3 is an integer).

From here on, we can conclude our solution to Exercise 8.4.10 similarly to our above solution to Exercise A.10.1.

Having solved Exercise 8.4.10, let me point out two generalizations (although I only know how to solve one).

Exercise 8.4.10 is a particular case (for $k = 3$) of the following more general exercise:

Exercise A.10.2. Fix a positive integer $k \geq 2$ and an integer $q \in \mathbb{N}$. Define a sequence $(a_0, a_1, a_2, \ldots)$ of positive rational numbers recursively by setting

$$a_n = 1 \qquad \text{for each } n < k \tag{804}$$

and

$$a_n = \frac{\left(a_{n-1}^2 + a_{n-2}^2 + \cdots + a_{n-k+1}^2\right) + q \sum\limits_{1 \leq i < j \leq k-1} a_{n-i} a_{n-j}}{a_{n-k}} \qquad \text{for each } n \geq k. \tag{805}$$

Prove that a_n is a positive integer for each integer $n \geq 0$.

Also, Exercise 8.4.10 is a particular case (for $k = 3$) of the following more general exercise:

Exercise A.10.3. Fix some $k \in \{2, 3, 4\}$ and an integer $q \in \mathbb{N}$. Define a sequence $(a_0, a_1, a_2, \ldots)$ of positive rational numbers recursively by setting

$$a_n = 1 \qquad \text{for each } n < k \tag{806}$$

and

$$a_n = \frac{\left(a_{n-1}^2 + a_{n-2}^2 + \cdots + a_{n-k+1}^2\right) + q a_{n-1} a_{n-k+1}}{a_{n-k}} \qquad \text{for each } n \geq k. \tag{807}$$

Prove that a_n is a positive integer for each integer $n \geq 0$.

Exercise A.10.2 and Exercise A.10.3 can be solved along the same lines as our above solution to Exercise 8.4.10. (Alas, Exercise A.10.3 becomes false for $k = 5$ (indeed, for $q = 1$ and $k = 5$, we have $a_{11} \notin \mathbb{Z}$); this spoils the pattern somewhat.) $\square$

A.11. Homework set #10A discussion

The following are discussions of the problems on homework set #10A (Subsection 9.1.5).

A.11.1. Discussion of Exercise 9.1.4

Discussion of Exercise 9.1.4. The key to Exercise 9.1.4 is to recall the identity $u^2 - v^2 = (u - v)(u + v)$. So the exercise is about writing n as a product of two integers of the form $u - v$ and $u + v$. When can two integers be written in the form $u - v$ and $u + v$? The answer to this turns out to be "exactly when they are congruent modulo 2 (that is, have the same parity)". Let us state this as a lemma[450]:

Lemma A.11.1. Let a and b be two integers. Then:

(a) If $a \equiv b \bmod 2$, then there exist two integers $u, v \in \mathbb{Z}$ such that $u + v = a$ and $u - v = b$.

(b) If there exist two integers $u, v \in \mathbb{Z}$ such that $u + v = a$ and $u - v = b$, then $a \equiv b \bmod 2$.

Proof of Lemma A.11.1. **(a)** Assume that $a \equiv b \bmod 2$. Let $x = \dfrac{a+b}{2}$ and $y = \dfrac{a-b}{2}$. Thus, $x + y = \dfrac{a+b}{2} + \dfrac{a-b}{2} = a$ and $x - y = \dfrac{a+b}{2} - \dfrac{a-b}{2} = b$.

From $a \equiv b \bmod 2$, we obtain $2 \mid a - b$, and thus $\dfrac{a-b}{2} \in \mathbb{Z}$. In other words, $y \in \mathbb{Z}$ (since $y = \dfrac{a-b}{2}$). Also, from $x + y = a$, we obtain $x = a - y \in \mathbb{Z}$ (since $a \in \mathbb{Z}$ and $y \in \mathbb{Z}$). We thus know that x and y are integers (since $x \in \mathbb{Z}$ and $y \in \mathbb{Z}$) and satisfy $x + y = a$ and $x - y = b$. Hence, there exist two integers $u, v \in \mathbb{Z}$ such that $u + v = a$ and $u - v = b$ (namely, $u = x$ and $v = y$). This proves Lemma A.11.1 **(a)**.

(b) Assume that there exist two integers $u, v \in \mathbb{Z}$ such that $u + v = a$ and $u - v = b$. Consider these u and v. From $u + v = a$, we obtain

$$a = u + v = u - v + \underbrace{2}_{\equiv 0 \bmod 2} v \equiv u - v + 0v = u - v = b \bmod 2.$$

This proves Lemma A.11.1 **(b)**. $\square$

[450] Note the similarity to Lemma A.4.5 (which we will not, however, use here).

With Lemma A.11.1 in hand, we can quickly reduce Exercise 9.1.4 to showing that n can be written as a product of two integers that have the same parity if and only if $n \not\equiv 2 \bmod 4$. But this is easy to check:

- If n is divisible by 4, then $n = 2 \cdot \dfrac{n}{2}$ is a way to write n as a product of two integers that have the same parity.
- If n is odd, then $n = 1 \cdot n$ is a way to write n as a product of two integers that have the same parity.
- On the other hand, if n is neither divisible by 4 nor odd, then we cannot write n as a product ab of two integers a and b that have the same parity, because such a product must either be divisible by 4 (if a and b are both even) or be odd (if a and b are both odd).

Now, it remains to observe that the statement "n is either divisible by 4 or odd" is equivalent to "$n \not\equiv 2 \bmod 4$".

Just for the sake of completeness, here is a detailed version of the solution we just outlined:

Solution to Exercise 9.1.4. We are in one of the following three cases:

Case 1: We have $4 \mid n$.

Case 2: We have $2 \nmid n$.

Case 3: We have neither $4 \mid n$ nor $2 \nmid n$.

Let us first consider Case 1. In this case, we have $4 \mid n$. Thus, $n\%4 = 0$ [451]. Hence, $n \not\equiv 2 \bmod 4$ [452].

From $4 \mid n$, we obtain $\dfrac{n}{4} \in \mathbb{Z}$. Thus, $2 \cdot \dfrac{n}{4}$ is an even integer. In other words, $\dfrac{n}{2}$ is an even integer (since $2 \cdot \dfrac{n}{4} = \dfrac{n}{2}$). Therefore, $\dfrac{n}{2} \equiv 0 \equiv 2 \bmod 2$. Hence, Lemma A.11.1 **(a)** (applied to $a = \dfrac{n}{2}$ and $b = 2$) yields that there exist two integers $u, v \in \mathbb{Z}$ such that $u + v = \dfrac{n}{2}$ and $u - v = 2$ (since $\dfrac{n}{2}$ and 2 are integers). These two integers u and v must necessarily satisfy $n = u^2 - v^2$ (since $u^2 - v^2 = \underbrace{(u-v)}_{=2}\underbrace{(u+v)}_{=\frac{n}{2}} = 2 \cdot \dfrac{n}{2} = n$). Hence, n can be represented in the form $n = u^2 - v^2$ for some $u, v \in \mathbb{Z}$.

Thus, we have shown that both statements "n can be represented in the form $n = u^2 - v^2$ for some $u, v \in \mathbb{Z}$" and "$n \not\equiv 2 \bmod 4$" are true (since we have shown

[451] *Proof.* Proposition 3.3.2 **(b)** (applied to 4 and n instead of n and u) shows that we have $4 \mid n$ if and only if $n\%4 = 0$. Thus, we have $n\%4 = 0$ (since $4 \mid n$).

[452] *Proof.* Assume the contrary. Thus, $n \equiv 2 \bmod 4$, so that $2 \equiv n \bmod 4$. Hence, Proposition 3.3.2 **(c)** (applied to 4, n and 2 instead of n, u and c) yields $2 = n\%4$ (since $2 \in \{0, 1, \ldots, 4-1\}$). Hence, $n\%4 = 2 \neq 0$. This contradicts $n\%4 = 0$. This contradiction shows that our assumption was false; qed.

that $n \not\equiv 2 \bmod 4$). Therefore, these two statements are equivalent. In other words, n can be represented in the form $n = u^2 - v^2$ for some $u, v \in \mathbb{Z}$ if and only if $n \not\equiv 2 \bmod 4$. Thus, Exercise 9.1.4 is solved in Case 1.

Let us next consider Case 2. In this case, we have $2 \nmid n$. In other words, n is odd. Hence, $n \not\equiv 2 \bmod 4$ [453].

On the other hand, $n \equiv 1 \bmod 2$ (since n is odd). Therefore, Lemma A.11.1 **(a)** (applied to $a = n$ and $b = 1$) yields that there exist two integers $u, v \in \mathbb{Z}$ such that $u + v = n$ and $u - v = 1$. These two integers u and v must necessarily satisfy $n = u^2 - v^2$ (since $u^2 - v^2 = \underbrace{(u-v)}_{=1}\underbrace{(u+v)}_{=n} = 1 \cdot n = n$). Hence, n can be represented in the form $n = u^2 - v^2$ for some $u, v \in \mathbb{Z}$.

Thus, we have shown that both statements "n can be represented in the form $n = u^2 - v^2$ for some $u, v \in \mathbb{Z}$" and "$n \not\equiv 2 \bmod 4$" are true (since we have shown that $n \not\equiv 2 \bmod 4$). Therefore, these two statements are equivalent. In other words, n can be represented in the form $n = u^2 - v^2$ for some $u, v \in \mathbb{Z}$ if and only if $n \not\equiv 2 \bmod 4$. Thus, Exercise 9.1.4 is solved in Case 2.

Finally, let us consider Case 3. In this case, we have neither $4 \mid n$ nor $2 \nmid n$. Thus, $4 \nmid n$ (since we don't have $4 \mid n$) and $2 \mid n$ (since we don't have $2 \nmid n$). From $2 \mid n$, we obtain $\dfrac{n}{2} \in \mathbb{Z}$. In other words, $\dfrac{n}{2}$ is an integer. This integer $\dfrac{n}{2}$ cannot be even[454], and therefore must be odd. Hence, we have $n \equiv 2 \bmod 4$ [455]. Thus, we don't have $n \not\equiv 2 \bmod 4$. Also, we have $n \equiv 0 \bmod 2$ (since $2 \mid n$).

Now, let $u, v \in \mathbb{Z}$ be integers satisfying $n = u^2 - v^2$. We shall derive a contradiction. Indeed, set $a = u + v$ and $b = u - v$. Thus, Lemma A.11.1 **(b)** yields $a \equiv b \bmod 2$ (since $u + v = a$ and $u - v = b$). Also, $n = u^2 - v^2 = \underbrace{(u+v)}_{=a}\underbrace{(u-v)}_{=b} = ab$. If the integer b was odd, then we would have $b \equiv 1 \bmod 2$ and therefore $n = \underbrace{a}_{\equiv b \equiv 1 \bmod 2}\underbrace{b}_{\equiv 1 \bmod 2} \equiv 1 \cdot 1 = 1 \not\equiv 0 \bmod 2$, which would contradict $n \equiv 0 \bmod 2$. Hence, the integer b cannot be odd. Thus, b must be even. In other words, $b = 2q$ for some integer q. Consider this q. Also, $a \equiv b \equiv 0 \bmod 2$ (since b is even); hence, the integer a is even. In other words, $a = 2p$ for some integer p. Consider this p.

[453] *Proof.* Assume the contrary. Thus, $n \equiv 2 \bmod 4$. Hence, Proposition 3.2.6 **(e)** (applied to n, 2, 4 and 2 instead of a, b, n and m) yields $n \equiv 2 \bmod 2$ (since $2 \mid 4$). Hence, $n \equiv 2 \equiv 0 \bmod 2$; in other words, n is even. That is, $2 \mid n$. This contradicts $2 \nmid n$. This contradiction shows that our assumption was false; qed.

[454] *Proof.* Assume the contrary. Thus, the integer $\dfrac{n}{2}$ is even. In other words, $2 \mid \dfrac{n}{2}$; therefore, $\dfrac{\left(\frac{n}{2}\right)}{2} \in \mathbb{Z}$. In other words, $\dfrac{n}{4} \in \mathbb{Z}$ (since $\dfrac{\left(\frac{n}{2}\right)}{2} = \dfrac{n}{4}$). Hence, $4 \mid n$. But this contradicts $4 \nmid n$. This contradiction shows that our assumption was false; qed.

[455] *Proof.* The integer $\dfrac{n}{2}$ is odd. In other words, $\dfrac{n}{2} = 2u + 1$ for some $u \in \mathbb{Z}$. Consider this u. Multiplying the equality $\dfrac{n}{2} = 2u + 1$ by 2, we obtain $n = 2(2u+1) = \underbrace{4}_{\equiv 0 \bmod 4} u + 2 \equiv 0u + 2 = 2 \bmod 4$.

Now, $n = \underbrace{a}_{=2p} \underbrace{b}_{=2q} = (2p)(2q) = 4pq$ shows that $4 \mid n$; but this contradicts $4 \nmid n$.

Forget that we fixed u, v. We thus have obtained a contradiction for any two integers $u, v \in \mathbb{Z}$ satisfying $n = u^2 - v^2$. Hence, no two such integers u, v exist. In other words, n cannot be represented in the form $n = u^2 - v^2$ for some $u, v \in \mathbb{Z}$.

Thus, we have shown that both statements "n can be represented in the form $n = u^2 - v^2$ for some $u, v \in \mathbb{Z}$" and "$n \not\equiv 2 \bmod 4$" are false (since we have shown that we don't have $n \not\equiv 2 \bmod 4$). Therefore, these two statements are equivalent. In other words, n can be represented in the form $n = u^2 - v^2$ for some $u, v \in \mathbb{Z}$ if and only if $n \not\equiv 2 \bmod 4$. Thus, Exercise 9.1.4 is solved in Case 3.

We have now solved Exercise 9.1.4 in all three Cases 1, 2 and 3. Thus, our solution to Exercise 9.1.4 is complete. □

□

A.11.2. Discussion of Exercise 9.1.5 (TODO: add details!)

Discussion of Exercise 9.1.5. I don't remember where I found Exercise 9.1.5. It is, in some sense, a (noticeably) harder version of Exercise 9.1.4. I will only sketch the solution, after first outhousing part of it into the following lemma:

Lemma A.11.2. Let n be a positive integer that is neither a prime nor a power of 2. Then, there exist two integers $a \geq 3$ and $b \geq 3$ satisfying $2n = ab$ and $a \not\equiv b \bmod 2$.

Proof of Lemma A.11.2 (sketched). First of all, we recall that n is not a power of 2. Hence, the number n has at least one odd prime divisor d (since otherwise, the prime factorization of n would only consist of 2's, but this would mean that n is a power of 2). Consider such a d. Thus, $n = de$ for some positive integer e (since d is a positive divisor of n). Moreover, $d \geq 3$ (since d is an odd prime). Note that $n \neq d$ (since n is not a prime, but d is a prime), and thus $e \neq 1$ (since otherwise, we would have $e = 1$ and thus $n = d \underbrace{e}_{=1} = d$, which would contradict $n \neq d$). Hence, $e \geq 2$ (since e is a positive integer) and thus $2e \geq 2 \cdot 2 = 4 \geq 3$.

Now, from $n = de$, we obtain $2n = 2de = d(2e)$. Moreover, the two integers d and $2e$ satisfy $d \geq 3$ and $2e \geq 3$ and $d \not\equiv 2e \bmod 2$ (since d is odd while $2e$ is even). Hence, there exist two integers $a \geq 3$ and $b \geq 3$ satisfying $2n = ab$ and $a \not\equiv b \bmod 2$ (namely, $a = d$ and $b = 2e$). This proves Lemma A.11.2. □

We can now solve Exercise 9.1.5:

Solution to Exercise 9.1.5 (sketched). We must prove the following two claims:

> *Claim 1:* If n can be represented in the form $n = uv - \binom{u}{2}$ for some $u, v \in \mathbb{Z}$ satisfying $v \geq u \geq 3$, then n is neither a prime nor a power of 2.

Claim 2: If n is neither a prime nor a power of 2, then n can be represented in the form $n = uv - \binom{u}{2}$ for some $u, v \in \mathbb{Z}$ satisfying $v \geq u \geq 3$.

[*Proof of Claim 1:* Assume that n can be represented in the form $n = uv - \binom{u}{2}$ for some $u, v \in \mathbb{Z}$ satisfying $v \geq u \geq 3$. Consider these u, v. Thus,

$$n = uv - \underbrace{\binom{u}{2}}_{=\frac{u(u-1)}{2}} = uv - \frac{u(u-1)}{2} = u\left(v - \frac{u-1}{2}\right).$$

Multiplying this equality by 2, we get

$$2n = 2u\left(v - \frac{u-1}{2}\right) = u(2v - u + 1).$$

Set $w = 2v - u + 1$. Then, $2n = u\underbrace{(2v - u + 1)}_{=w} = uw$. Hence, u and w are divisors of $2n$. Both of these divisors are ≥ 3 (since $u \geq 3$ and $w = 2\underbrace{v}_{\geq u} - u + \underbrace{1}_{\geq 0} \geq 2u - u + 0 = u \geq 3$), and furthermore have different parity (since their difference $u - \underbrace{w}_{=2v-u+1} = u - (2v - u + 1) = 2(u - v) - 1$ is odd). Thus, one of these two divisors must be odd. Hence, $2n$ has an odd divisor that is ≥ 3 (since both of these divisors are ≥ 3). This shows that $2n$ is not a power of 2 (since a power of 2 has no odd divisor that is ≥ 3 [456]). Hence, n is not a power of 2 either.

It remains to show that n is not a prime. Indeed, assume the contrary. Thus, n is a prime. Hence, from $n \mid 2n = uw$, we conclude that we have $n \mid u$ or $n \mid w$ (by Theorem 9.1.8, applied to $p = n$, $a = u$ and $b = w$). If $n \mid u$, then $u \geq n$ (since $u \geq 3$ is positive) and therefore $2n = \underbrace{u}_{\geq n}\underbrace{w}_{\geq 3} \geq n \cdot 3 = 3n > 2n$, which is clearly absurd. The same argument (with the roles of u and w interchanged) yields an absurd conclusion if $n \mid w$. Thus, neither $n \mid u$ nor $n \mid w$ can hold. This contradicts the fact that we have $n \mid u$ or $n \mid w$. This contradiction completes our proof of Claim 1.]

[*Proof of Claim 2:* Assume that n is neither a prime nor a power of 2. Thus, Lemma A.11.2 shows that there exist two integers $a \geq 3$ and $b \geq 3$ satisfying $2n = ab$ and $a \not\equiv b \bmod 2$. Consider these a and b.

We WLOG assume that $b \geq a$ (since we can otherwise achieve this by swapping a with b).

We have $a \not\equiv b \bmod 2$. In other words, $2 \nmid a - b$. Thus, $a - b$ is odd, so that we have $a - b \equiv 1 \bmod 2$. However, $a + b \equiv a - b \bmod 2$ (since $(a + b) - (a - b) = 2b$

[456] This is easiest to see using prime factorization.

is clearly even). Therefore, $a+b \equiv a-b \equiv 1 \bmod 2$. In other words, $a+b$ is odd. Hence, $a+b-1$ is even, so that $\dfrac{a+b-1}{2} \in \mathbb{Z}$.

Set $u = a$ and $v = \dfrac{a+b-1}{2}$. Then, both u and v are integers (since $u = a \in \mathbb{Z}$ and $v = \dfrac{a+b-1}{2} \in \mathbb{Z}$). Moreover, the definitions of u and v yield

$$u\left(2v-u+1\right) = a\left(2 \cdot \frac{a+b-1}{2} - a + 1\right) = ab = 2n.$$

Solving this equality for n, we find

$$n = \frac{u\left(2v-u+1\right)}{2} = uv - \underbrace{\frac{u\left(u-1\right)}{2}}_{=\binom{u}{2}} = uv - \binom{u}{2}.$$

Furthermore, $u = a \geq 3$. Moreover,

$$\begin{aligned} v &= \frac{a+b-1}{2} \geq \frac{a+a-1}{2} \qquad \left(\text{since } b \geq a\right) \\ &= \underbrace{a}_{=u} - \underbrace{\frac{1}{2}}_{<1} > u-1 \end{aligned}$$

and thus $v \geq u$ (since v and u are integers). Hence, $v \geq u \geq 3$. Thus, n can be represented in the form $n = uv - \binom{u}{2}$ for some $u, v \in \mathbb{Z}$ satisfying $v \geq u \geq 3$ (indeed, we have just found such u, v). This proves Claim 2.]

With Claim 1 and 2 proved, we have solved Exercise 9.1.5. □

□

A.11.3. Discussion of Exercise 9.1.6

Discussion of Exercise 9.1.6. Both parts of Exercise 9.1.6 are important results in number theory (see, e.g., [Grinbe19c, Corollary 5.6] and [Grinbe19c, Proposition 3.1]). Let us give two solutions:

First solution to Exercise 9.1.6. **(a)** Let $k \in \{1, 2, \ldots, p-1\}$. Thus, $k \in \{1, 2, \ldots, p-1\} \subseteq \{0, 1, \ldots, p-1\}$ and $p \equiv 0 \bmod p$. Hence, Exercise 9.1.2 (applied to $a = p$ and $b = 0$) yields

$$\binom{p}{k} \equiv \binom{0}{k} \bmod p. \tag{808}$$

However, from $k \in \{1, 2, \ldots, p-1\}$, we obtain $k \geq 1 > 0$. Hence, Proposition 4.3.4 (applied to $n = 0$) yields $\binom{0}{k} = 0$. Thus, (808) rewrites as $\binom{p}{k} \equiv 0 \bmod p$. In other words, $p \mid \binom{p}{k}$. This solves Exercise 9.1.6 **(a)**.

(b) Let $k \in \{0, 1, \ldots, p-1\}$. It is clear that $\underbrace{p}_{\equiv 0 \bmod p} - 1 \equiv 0 - 1 = -1 \bmod p$. Hence, Exercise 9.1.2 (applied to $a = p - 1$ and $b = -1$) yields

$$\binom{p-1}{k} \equiv \binom{-1}{k} = (-1)^k \bmod p \qquad \text{(by (122))}.$$

This solves Exercise 9.1.6 **(b)**. □

Second solution to Exercise 9.1.6. **(a)** Let $k \in \{1, 2, \ldots, p-1\}$. Then, Exercise 4.5.4 **(b)** (applied to $n = p$ and $m = k$) yields

$$\frac{\gcd(p,k)}{p}\binom{p}{k} \in \mathbb{Z}. \tag{809}$$

However, k is coprime to p (by Proposition 9.1.5, applied to $i = k$). In other words, $\gcd(k,p) = 1$. Now, Proposition 3.4.4 **(b)** (applied to $a = p$ and $b = k$) yields $\gcd(p,k) = \gcd(k,p) = 1$. Hence, $\frac{\gcd(p,k)}{p}\binom{p}{k} = \frac{1}{p}\binom{p}{k} = \frac{\binom{p}{k}}{p}$. Thus, (809) rewrites as follows:

$$\frac{\binom{p}{k}}{p} \in \mathbb{Z}.$$

In other words, $p \mid \binom{p}{k}$. This solves Exercise 9.1.6 **(a)** again.

(b) We proceed by induction on k:

Induction base: We have $\binom{p-1}{0} = 1$ (by (119), applied to $n = p - 1$) and $(-1)^0 = 1$. Thus, $\binom{p-1}{0} = 1 \equiv 1 = (-1)^0 \bmod p$. In other words, Exercise 9.1.6 **(b)** holds for $k = 0$.

Induction step: Let $m \in \{1, 2, \ldots, p-1\}$. Assume (as the induction hypothesis) that Exercise 9.1.6 **(a)** holds for $k = m - 1$. We must show that Exercise 9.1.6 **(b)** holds for $k = m$.

We have assumed that Exercise 9.1.6 **(b)** holds for $k = m - 1$. In other words, $\binom{p-1}{m-1} \equiv (-1)^{m-1} \bmod p$.

But Exercise 9.1.6 **(a)** (applied to $k = m$) yields $p \mid \binom{p}{m}$. In other words, $\binom{p}{m} \equiv 0 \bmod p$. In view of

$$\binom{p}{m} = \binom{p-1}{m-1} + \binom{p-1}{m} \qquad \text{(by Theorem 4.3.7, applied to } n = p \text{ and } k = m\text{)},$$

this rewrites as $\binom{p-1}{m-1} + \binom{p-1}{m} \equiv 0 \bmod p$. In other words,

$$\binom{p-1}{m} \equiv -\underbrace{\binom{p-1}{m-1}}_{\equiv (-1)^{m-1} \bmod p} \equiv -(-1)^{m-1} = (-1)^m \bmod p.$$

In other words, Exercise 9.1.6 **(b)** holds for $k = m$. This completes the induction step. Thus, Exercise 9.1.6 **(b)** is solved (again). □

□

A.11.4. Discussion of Exercise 9.1.7

Discussion of Exercise 9.1.7. Exercise 9.1.7 is a classical and important fact (in particular, it is the main tool in the proof of the Chevalley–Warning theorem). The following solution, which strategically uses our Corollary 7.8.8, is probably the shortest:[457]

Solution to Exercise 9.1.7. From $k \in \{0, 1, \ldots, p-2\}$, we obtain $k \leq p-2 < p-1$. Hence, Corollary 7.8.8 **(a)** (applied to $n = p-1$ and $m = k$) yields

$$\sum_{i=0}^{p-1} (-1)^{p-1-i} \binom{p-1}{i} i^k = 0.$$

However,

$$\sum_{i=0}^{p-1} (-1)^{p-1-i} \underbrace{\binom{p-1}{i}}_{\substack{\equiv (-1)^i \bmod p \\ \text{(by Exercise 9.1.6 } \textbf{(b)} \\ \text{(applied to } i \text{ instead of } k))}} i^k \equiv \sum_{i=0}^{p-1} \underbrace{(-1)^{p-1-i} (-1)^i}_{\substack{= (-1)^{(p-1-i)+i} = (-1)^{p-1} \\ \text{(since } (p-1-i)+i = p-1)}} i^k$$

$$= \sum_{i=0}^{p-1} (-1)^{p-1} i^k = (-1)^{p-1} \sum_{i=0}^{p-1} i^k,$$

so that

$$(-1)^{p-1} \sum_{i=0}^{p-1} i^k \equiv \sum_{i=0}^{p-1} (-1)^{p-1-i} \binom{p-1}{i} i^k = 0 \bmod p.$$

[457] See [Grinbe19a, proof of Lemma 3.8] for a different solution.

Multiplying both sides of this congruence by $(-1)^{p-1}$, we find

$$(-1)^{p-1}(-1)^{p-1}\sum_{i=0}^{p-1} i^k \equiv (-1)^{p-1}\,0 = 0 \bmod p.$$

In view of

$$\underbrace{(-1)^{p-1}(-1)^{p-1}}_{\substack{=\left((-1)^{p-1}\right)^2=(-1)^{2(p-1)}=1 \\ \text{(since } 2(p-1) \text{ is even)}}} \sum_{i=0}^{p-1} i^k = \sum_{i=0}^{p-1} i^k,$$

this rewrites as $\sum_{i=0}^{p-1} i^k \equiv 0 \bmod p$. This solves Exercise 9.1.7. □

□

A.11.5. Discussion of Exercise 9.1.8

Discussion of Exercise 9.1.8. Exercise 9.1.8 is Problem #10501 **(b)** from the American Mathematical Monthly (suggested by Roger B. Eggleton; see [EggWes98] for the solution); it also appears (in a fairly representative particular case) as problem #2175 in Crux Mathematicorum (proposed by Christopher J. Bradley). See OEIS Sequence A063647 for more references.

Solution to Exercise 9.1.8. Let U denote the set of all pairs (j,k) of positive integers satisfying $\frac{1}{j} - \frac{1}{k} = \frac{1}{n}$. Then, $|U|$ is the # of such pairs. Thus, $|U| = u$ (since u was defined to be the # of such pairs).

Let V be the set of all integers $i \in [n-1]$ satisfying $i \mid n^2$.

For each $(j,k) \in U$, we have $n-j \in V$ [458]. Hence, we can define a map

$$\begin{aligned} f : U &\to V, \\ (j,k) &\mapsto n-j. \end{aligned}$$

Consider this map f.

We shall now construct an inverse to this map f. Indeed, if $x \in V$, then

[458] *Proof.* Let $(j,k) \in U$. Thus, (j,k) is a pair of positive integers satisfying $\frac{1}{j} - \frac{1}{k} = \frac{1}{n}$ (by the definition of U). From $\frac{1}{j} - \frac{1}{k} = \frac{1}{n}$, we obtain $\frac{1}{j} - \frac{1}{k} - \frac{1}{n} = 0$. Now,

$$(n-j)(n+k) - n^2 = kn - jn - jk = njk \underbrace{\left(\frac{1}{j} - \frac{1}{k} - \frac{1}{n}\right)}_{=0} = 0.$$

Hence, $(n-j)(n+k) = n^2$. Thus, $n-j \mid n^2$ (since $n-j$ and $n+k$ are integers). Moreover, $j \geq 1$ (since j is a positive integer), so that $n - \underbrace{j}_{\geq 1} \leq n-1$. On the other hand, $n+k$ is a positive integer (since n and k are positive integers). Hence, $n+k \neq 0$. Thus, from $(n-j)(n+k) = n^2$, we obtain $n-j = \frac{n^2}{n+k}$. Thus, $n-j$ is positive (since n and $n+k$ are positive). Combining this with $n-j \leq n-1$, we obtain $0 < n-j \leq n-1$. Thus, $n-j \in \{1,2,\ldots,n-1\}$ (since $n-j$ is an integer). In other words, $n-j \in [n-1]$ (since $[n-1] = \{1,2,\ldots,n-1\}$).

We now know that $n-j$ is an integer and satisfies $n-j \in [n-1]$ and $n-j \mid n^2$. In other words, $n-j$ is an integer $i \in [n-1]$ satisfying $i \mid n^2$. In other words, $n-j \in V$ (since V is the set of all integers $i \in [n-1]$ satisfying $i \mid n^2$). Qed.

$\left(n-x, \dfrac{n^2}{x}-n\right) \in U$ [459]. Hence, we can define a map

$$\begin{aligned} g : V &\to U, \\ x &\mapsto \left(n-x, \frac{n^2}{x}-n\right). \end{aligned}$$

Consider this map g.

Now, $f \circ g = \mathrm{id}$ [460] and $g \circ f = \mathrm{id}$ [461]. Thus, the two maps f and g are

[459] *Proof.* Let $x \in V$. Thus, x is an integer $i \in [n-1]$ satisfying $i \mid n^2$ (since V was defined to be the set of all such integers i). In other words, x is an integer satisfying $x \in [n-1]$ and $x \mid n^2$.

From $x \in [n-1] = \{1, 2, \ldots, n-1\}$, we obtain $1 \leq x \leq n-1$. Hence, $x \geq 1 > 0$. Thus, $x \neq 0$. Hence, from $x \mid n^2$, we obtain $\dfrac{n^2}{x} \in \mathbb{Z}$. Hence, $\dfrac{n^2}{x} - n \in \mathbb{Z}$ as well (since $n \in \mathbb{Z}$). In other words, $\dfrac{n^2}{x} - n$ is an integer.

Moreover, $x \leq n-1 < n$, so that $\dfrac{n^2}{x} > \dfrac{n^2}{n}$ (since $n^2 > 0$ (because n is positive)). Hence, $\dfrac{n^2}{x} > \dfrac{n^2}{n} = n$, so that $\dfrac{n^2}{x} - n > 0$. Hence, $\dfrac{n^2}{x} - n$ is a positive integer (since we already know that $\dfrac{n^2}{x} - n$ is an integer). Also, $n - x > 0$ (since $x < n$), so that $n - x$ is a positive integer (since $n - x$ is an integer).

Now, we know that $\left(n-x, \dfrac{n^2}{x}-n\right)$ is a pair of positive integers (since $n-x$ and $\dfrac{n^2}{x}-n$ are positive integers) satisfying $\dfrac{1}{n-x} - \dfrac{1}{\dfrac{n^2}{x}-n} = \dfrac{1}{n}$ (this equality can be checked by a straightforward computation). In other words, $\left(n-x, \dfrac{n^2}{x}-n\right)$ is a pair (j,k) of positive integers satisfying $\dfrac{1}{j} - \dfrac{1}{k} = \dfrac{1}{n}$. In other words, $\left(n-x, \dfrac{n^2}{x}-n\right) \in U$ (since U is the set of all such pairs (j,k)). Qed.

[460] *Proof.* Let $x \in V$. Then, the definition of g yields $g(x) = \left(n-x, \dfrac{n^2}{x}-n\right)$. Applying the map f to both sides of this equality, we find

$$f(g(x)) = f\left(\left(n-x, \frac{n^2}{x}-n\right)\right) = n-(n-x) = x = \mathrm{id}(x).$$

Hence, $(f \circ g)(x) = f(g(x)) = \mathrm{id}(x)$.

Forget that we fixed x. We thus have shown that $(f \circ g)(x) = \mathrm{id}(x)$ for each $x \in V$. In other words, $f \circ g = \mathrm{id}$.

[461] *Proof.* Let $y \in U$. Then, y is a pair (j,k) of positive integers satisfying $\dfrac{1}{j} - \dfrac{1}{k} = \dfrac{1}{n}$ (since U is the set of all such pair (j,k)). Consider this (j,k). Thus, $y = (j,k)$.

Solving the equality $\dfrac{1}{j} - \dfrac{1}{k} = \dfrac{1}{n}$ for k, we obtain

$$k = \frac{1}{\dfrac{1}{j} - \dfrac{1}{n}} = \frac{jn}{n-j}. \tag{810}$$

mutually inverse. Hence, the map f is invertible, i.e., is bijective. In other words, f is a bijection from U to V. Hence, the bijection principle yields $|U| = |V|$.

Now, recall that $|U| = u$. Hence,

$$u = |U| = |V| = \left(\text{the \# of all integers } i \in [n-1] \text{ satisfying } i \mid n^2\right)$$

(since V is the set of all integers $i \in [n-1]$ satisfying $i \mid n^2$). This solves Exercise 9.1.8. □

As an aside, it is easy to solve Exercise 9.1.3 again using Exercise 9.1.8. (We leave the details to the mythical interested reader.) □

Applying the map $g \circ f$ to both sides of the equality $y = (j,k)$, we obtain

$$\begin{aligned}
(g \circ f)(y) = (g \circ f)((j,k)) &= g\left(\underbrace{f(j,k)}_{\substack{=n-j \\ \text{(by the definition of } f)}}\right) = g(n-j) \\
&= \left(\underbrace{n-(n-j)}_{=j}, \underbrace{\frac{n^2}{n-j} - n}_{\substack{=\frac{jn}{n-j}=k \\ \text{(by (810))}}}\right) \qquad \text{(by the definition of } g) \\
&= (j,k) = y = \operatorname{id}(y).
\end{aligned}$$

Forget that we fixed y. We thus have shown that $(g \circ f)(y) = \operatorname{id}(y)$ for each $y \in U$. In other words, $g \circ f = \operatorname{id}$.

References

[17f-hw2s] Darij Grinberg, *UMN Fall 2017 Math 4707 & Math 4990 homework set #2 with solutions*, http://www.cip.ifi.lmu.de/~grinberg/t/17f/hw2s.pdf

[17f-hw3s] Darij Grinberg, *UMN Fall 2017 Math 4990 homework set #3 with solutions*, http://www.cip.ifi.lmu.de/~grinberg/t/17f/hw3os.pdf

[17f-hw4s] Darij Grinberg, *UMN Fall 2017 Math 4990 homework set #4 with solutions*, http://www.cip.ifi.lmu.de/~grinberg/t/17f/hw4os.pdf

[17f-hw6s] Darij Grinberg, *UMN Fall 2017 Math 4990 homework set #6 with solutions*, http://www.cip.ifi.lmu.de/~grinberg/t/17f/hw6os.pdf

[17f-hw7s] Darij Grinberg, *UMN Fall 2017 Math 4990 homework set #7 with solutions*, http://www.cip.ifi.lmu.de/~grinberg/t/17f/hw7os.pdf

[17f-hw8s] Darij Grinberg, *UMN Fall 2017 Math 4990 homework set #8 with solutions*, http://www.cip.ifi.lmu.de/~grinberg/t/17f/hw8os.pdf

[18f-hw2s] Darij Grinberg, *UMN Fall 2018 Math 5705 homework set #2 with solutions*, http://www.cip.ifi.lmu.de/~grinberg/t/18f/hw2s.pdf

[18f-hw3s] Darij Grinberg, *UMN Fall 2018 Math 5705 homework set #3 with solutions*, http://www.cip.ifi.lmu.de/~grinberg/t/18f/hw3s.pdf

[18f-hw4s] Darij Grinberg, *UMN Fall 2018 Math 5705 homework set #4 with solutions*, http://www.cip.ifi.lmu.de/~grinberg/t/18f/hw4s.pdf

[18f-mt1s] Darij Grinberg, *UMN Fall 2018 Math 5705 midterm #1 with solutions*, http://www.cip.ifi.lmu.de/~grinberg/t/18f/mt1s.pdf

[18s-hw2s] Darij Grinberg, *UMN Spring 2018 Math 4707 homework set #2 with solutions*, http://www.cip.ifi.lmu.de/~grinberg/t/18s/hw2s.pdf

[18s-hw3s] Darij Grinberg, *UMN Spring 2018 Math 4707 homework set #3 with solutions*, http://www.cip.ifi.lmu.de/~grinberg/t/18s/hw3s.pdf

[18s-mt1s] Darij Grinberg, *UMN Spring 2018 Math 4707 midterm #1 with solutions*, http://www.cip.ifi.lmu.de/~grinberg/t/18s/mt1s.pdf

[18s-mt2s] Darij Grinberg, *UMN Spring 2018 Math 4707 midterm #2 with solutions*, http://www.cip.ifi.lmu.de/~grinberg/t/18s/mt2s.pdf

[19s] Darij Grinberg, *Introduction to Modern Algebra (UMN Spring 2019 Math 4281 notes)*, 1 October 2020. http://www.cip.ifi.lmu.de/~grinberg/t/19s/notes.pdf

[19s-hw0s] Darij Grinberg, *UMN Spring 2019 Math 4281 homework set #0 with solutions*, `http://www.cip.ifi.lmu.de/~grinberg/t/19s/hw0s.pdf`

[19s-mt3s] Darij Grinberg, *UMN Spring 2019 Math 4281 midterm #3 with solutions*, `http://www.cip.ifi.lmu.de/~grinberg/t/19s/mt3s.pdf`

[19fco] Darij Grinberg, *Enumerative Combinatorics (Drexel Fall 2019 Math 222 notes)*, 18 September 2020. `http://www.cip.ifi.lmu.de/~grinberg/t/19fco/n/n.pdf`

[19fco-hw0s] Darij Grinberg, *Drexel Fall 2019 Math 222 homework set #0 with solutions*, `http://www.cip.ifi.lmu.de/~grinberg/t/19fco/hw0s.pdf`

[19fco-hw2s] Darij Grinberg, *Drexel Fall 2019 Math 222 homework set #2 with solutions*, `http://www.cip.ifi.lmu.de/~grinberg/t/19fco/hw2s.pdf`

[19fco-hw3s] Darij Grinberg, *Drexel Fall 2019 Math 222 homework set #3 with solutions*, `http://www.cip.ifi.lmu.de/~grinberg/t/19fco/hw3s.pdf`

[19fco-mt2s] Darij Grinberg, *Drexel Fall 2019 Math 222 midterm #2 with solutions*, `http://www.cip.ifi.lmu.de/~grinberg/t/19fco/mt2s.pdf`

[Aigner07] Martin Aigner, *A Course in Enumeration*, Graduate Texts in Mathematics #238, Springer 2007.

[AigZie14] Martin Aigner, Günter M. Ziegler, *Proofs from the Book*, 6th edition, Springer 2018.

[AlCuHu16] Joshua Alman, Cesar Cuenca, Jiaoyang Huang, *Laurent phenomenon sequences*, Journal of Algebraic Combinatorics **43** (2016), pp. 589–633.

[AleKur91] P. Alekseev, L. Kurlyandchik, *Summa minimumov i minimum summy*, Kvant 1991/3, pp. 49–55. `http://kvant.mccme.ru/1991/03/summa_minimumov_i_minimum_summ.htm`

[AndAnd03] Titu Andreescu, Dorin Andrica, *360 Problems for Mathematical Contests*, Gil 2003.

[AndAnd14] Titu Andreescu, Dorin Andrica, *Equations with Solution in Terms of Fibonacci and Lucas Sequences*, An. St. Univ. Ovidius Constanta **22** (3), 2014, pp. 5–12.

[AndCri17] Titu Andreescu, Vlad Crisan, *Mathematical Induction: A powerful and elegant method of proof*, XYZ Press 2017.

[AndDos10] Titu Andreescu, Gabriel Dospinescu, *Problems from the Book*, 2nd edition, XYZ Press 2010.

[AndDos12] Titu Andreescu, Gabriel Dospinescu, *Straight from the Book*, XYZ Press 2012.

[AndEne11] Titu Andreescu, Bogdan Enescu, *Mathematical Olympiad Treasures*, 2nd edition, Birkhäuser 2011.

[AndEri04] George E. Andrews, Kimmo Eriksson, *Integer Partitions*, Cambridge University Press 2004.

[AndFen04] Titu Andreescu, Zuming Feng, *A Path to Combinatorics for Undergraduates: Counting Strategies*, Springer 2004.

[AndTet18] Titu Andreescu, Marian Tetiva, *Sums and Products*, XYZ Press 2018.

[AnDoMu17] Titu Andreescu, Gabriel Dospinescu, Oleg Mushkarov, *Number Theory: Concepts and Problems*, XYZ Press 2017.

[AnFeLe01] Titu Andreescu, Zuming Feng, George Lee, Jr. (eds.), *Mathematical Olympiads 2000–2001: Problems and Solutions From Around the World*, MAA 2002.

[Armstro19] Drew Armstrong, *MTH 309: Discrete Mathematics, Summer 2019*, 9 August 2019.
`https://www.math.miami.edu/~armstrong/309sum19.php`

[Bartle15] Padraic Bartlett, *various courses*, 2015.

[BecGeo10] Matthias Beck, Ross Geoghegan, *The Art of Proof: Basic Training for Deeper Mathematics*, Springer 2010.
`http://math.sfsu.edu/beck/aop.html`

[BeDeQu11] Arthur T. Benjamin, Halcyon Derks, Jennifer J. Quinn, *The Combinatorialization of Linear Recurrences*, The Electronic Journal of Combinatorics **18(2)** (2011), #P12.

[BenPlo08] Arthur T. Benjamin and Sean S. Plott, *A combinatorial approach to Fibonomial coefficients*, Fibonacci Quarterly **46/47**, issue 1, pp. 7–9.

[BenQui03] Arthur T. Benjamin and Jennifer J. Quinn, *Proofs that Really Count: The Art of Combinatorial Proof*, The Mathematical Association of America, 2003.

[BenQui08] Arthur T. Benjamin and Jennifer J. Quinn, *An Alternate Approach to Alternating Sums: A Method to DIE for*, The College Mathematics Journal, Volume 39, Number 3, May 2008, pp. 191-202(12).

[BiFoFo95] Wang Zheng Bing, Robert Fokkink, Wan Fokkink, *A Relation Between Partitions and the Number of Divisors*, The American Mathematical Monthly **102**, No. 4 (Apr., 1995), pp. 345–347.
`https://www.cs.vu.nl/~wanf/pubs/partition.pdf`

[BMPrWe09] Mireille Bousquet-Mélou, James Propp, Julian West, *Perfect matchings for the three-term Gale-Robinson sequences*, The Electronic Journal of Combinatorics **16** (2009), #R125.

[Bogart17] Kenneth P. Bogart, *Combinatorics Through Guided Discovery*, 2017.
`http://bogart.openmathbooks.org/`
Solutions available at `https://math.dartmouth.edu/news-resources/electronic/kpbogart/FDL-bogart/kpbogart-FDL.tar.gz`

[Bogomo19] Alexander Bogomolny, *Cut the Knot*, 2019.
`https://www.cut-the-knot.org/`
Use the Wayback Machine for some links that have disappeared.

[Bollob10] Béla Bollobás, *The Art of Mathematics: Coffee Time in Memphis*, Cambridge University Press, 3rd printing 2010.

[Bona17] Miklós Bóna, *A walk through combinatorics : an introduction to enumeration and graph theory*, World Scientific 2017.

[BorMea16] Peter Borg, Karen Meagher, *The Katona cycle proof of the Erdös–Ko–Rado theorem and its possibilities*, Journal of Algebraic Combinatorics **43**(4), 915–939.

[Bossae20] Jens Bossaert, *Curiosa Mathematica*, 27 March 2020.
`https://users.ugent.be/~jebossae/docs/curiosa_2.pdf`

[BoyDip12] William E. Boyce, Richard C. DiPrima, *Elementary Differential Equations*, 10th edition, Wiley 2012.

[BraTho05] David M. Bradley, Hugh Thomas, *Peg Jumping for Fun and Profit*, arXiv:math/0411275v2.
`https://arxiv.org/abs/math/0411275v2`

[Brown20] K. S. Brown, *A Quasi-Periodic Sequence*, 10 May 2020, MathPages.
`https://www.mathpages.com/home/kmath162/kmath162.htm`

[Brualdi09] Richard A. Brualdi, *Introductory Combinatorics*, 5th edition, Pearson 2009.

[Burton11] David M. Burton, *Elementary Number Theory*, 7th edition, McGraw-Hill 2011.

[Caminh18] Antonio Caminha Muniz Neto, *An Excursion through Elementary Mathematics, Volume III: Discrete Mathematics and Polynomial Algebra*, Springer 2018.

[Carl17] Merlin Carl, *Wie kommt man drauf? Einführung in das mathematische Problemlösen*, Springer 2017.
See the Springer book page for solutions and additional exercises.

[CarSpe04] Gabriel D. Carroll, David Speyer, *The Cube Recurrence*, The Electronic Journal of Combinatorics **11** (2014), #R73.

[ChaSed97] Gengzhe Chang, Thomas W. Sederberg, *Over and Over Again*, Anneli Lax New Mathematical Library **39**, The Mathematical Association of America 1997.

[Chen08] Kuo-Jye Chen, *Andrews's argument proves a theorem of Zeckendorf*, Mathematical Medley **34** (2008), no. 1, pp. 13–15.

[Chen20] Evan Chen, *various notes and problem collections on mathematical olympiads*, `https://web.evanchen.cc/` . See, in particular, the OTIS Excerpts (`https://web.evanchen.cc/excerpts.html`).

[Cirtoa06] Vasile Cirtoaje, *Algebraic inequalities: old and new methods*, Gil, Zalău 2006.

[ClaHef04] W. Edwin Clark, Jim Hefferon, *Elementary Number Theory*, 5 January 2004.
`http://joshua.smcvt.edu/numbertheory/book.pdf`

[Cohen78] Daniel I. A. Cohen, *Basic techniques of combinatorial theory*, John Wiley & Sons 1978.

[CoLiOs15] David A. Cox, John Little, Donal O'Shea, *Ideals, Varieties, and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra*, 4th edition, Springer 2015.

[Comtet74] Louis Comtet, *Advanced Combinatorics: The Art of Finite and Infinite Expansions*, D. Reidel Publishing Company, 1974.

[ConradE] Keith Conrad, *Euler's theorem*.
`https://kconrad.math.uconn.edu/blurbs/ugradnumthy/eulerthm.pdf`

[CorPer18] Scott Corry, David Perkinson, *Divisors and Sandpiles: An Introduction to Chip-Firing*, AMS 2018.
See `https://web.archive.org/web/20180428212224/http://people.reed.edu/~davidp/divisors_and_sandpiles/draft_01.28.2018.pdf` for a preprint.

[CraPom05] Richard Crandall, Carl Pomerance, *Prime Numbers: A Computational Perspective*, 2nd edition, Springer 2005.

[Day16] Martin V. Day, *An Introduction to Proofs and the Mathematical Vernacular*, 7 December 2016.
https://web.archive.org/web/20180712152432/https://www.math.vt.edu/people/day/ProofsBook/IPaMV.pdf .

[Dijkst80] Edsger W. Dijkstra, *On a theorem by Lambek and Moser*, Report EWD753, 1980.
https://www.cs.utexas.edu/users/EWD/ewd07xx/EWD753.PDF

[DJMP11] Dusan Djukic, Vladimir Jankovic, Ivan Matic, Nikola Petrovic, *The IMO Compendium: A Collection of Problems Suggested for The International Mathematical Olympiads: 1959-2009*, 2nd edition, Springer 2011.
See https://www.imomath.com/index.php?options=350&lmm=0 for errata.

[DLPS16] Birgit van Dalen, Julian Lyczak, Quintijn Puite, Merlijn Staps, *Finaletraining Wiskunde Olympiade*, 21 November 2016.
https://www.wiskundeolympiade.nl/phocadownload/finaletraining/finaletraining_binnenwerk_uitwerkingen.pdf

[DuaOli13] Rui Duarte, António Guedes de Oliveira, *A short proof of a famous combinatorial identity*, arXiv:1307.6693v1.
https://arxiv.org/abs/1307.6693v1

[Dudley12] Underwood Dudley, *Elementary Number Theory*, 2nd edition, Courier 2012.

[DynUsp06] E. B. Dynkin, V. A. Uspenskii, *Mathematical Conversations: Multicolor Problems, Problems in the Theory of Numbers, and Random Walks*, Dover 2006.

[Edward05] Harold M. Edwards, *Essays in Constructive Mathematics*, Springer 2005.
See https://www.math.nyu.edu/faculty/edwardsh/eserrata.pdf for errata.

[EggWes98] Roger B. Eggleton, Western Maryland College Problems Group, *AMM problem #10501 with solution*, The American Mathematical Monthly, **105** (4), pp. 372.

[Elman20] Richard Elman, *Lectures on Abstract Algebra*, 28 September 2020.
https://www.math.ucla.edu/~rse/algebra_book.pdf

[Emerso69] Edgar I. Emerson, *Recurrent sequences in the equation $DQ^2 = R^2 + N$*, Fibonacci Quarterly **7** (1969), issue 3, pp. 231–242.

[Engel97] Konrad Engel, *Sperner theory*, Cambridge University Press 1997.

[Engel98] Arthur Engel, *Problem-Solving Strategies*, Springer 1998.

[ErdSze78] P. Erdös, G. Szekeres, *Some Number Theoretic Problems on Binomial Coefficients*, Australian Math. Soc. Gazette **5** (1978), pp. 97–99. `http://bsmath.hu/~p_erdos/1978-46.pdf`

[Euler190] Leonhard Euler, *Consideratio quarumdam serierum, quae singularibus proprietatibus sunt praeditae*, Novi Commentarii academiae scientiarum Petropolitanae **3**, pp. 86–108, Euler archive #E190.

[Excali97] Anon, *A Proof of The Lambek and Moser Theorem*, Mathematical Excalibur **4** (1), 1997–1998, p. 2.

[FinRos16] Benjamin Fine, Gerhard Rosenberger, *Number theory: An introduction via the distribution of primes*, 2nd edition, Birkhäuser 2016.

[FomZel02] Sergey Fomin, Andrei Zelevinsky, *The Laurent Phenomenon*, Advances in Applied Mathematics **28** (2002), pp. 119–144.

[Foucar12] Simon Foucart, *Notes on Problem Solving*, 2012. `https://www.math.tamu.edu/~foucart/teaching/notes/PbSolving.pdf`

[Frankl20] Peter Frankl, *Old and new applications of Katona's circle*, arXiv:2011.14252v1.

[Gale98] David Gale, *Tracking the automatic ant and other mathematical explorations*, Springer 1998.

[GalQua20] Jean Gallier and Jocelyn Quaintance, *Algebra, Topology, Differential Calculus, and Optimization Theory For Computer Science and Engineering*, 22 March 2020. `https://www.cis.upenn.edu/~jean/gbooks/geomath.html`

[Galvin17] David Galvin, *Basic discrete mathematics*, 13 December 2017. `http://www-users.math.umn.edu/~dgrinber/comb/60610lectures2017-Galvin.pdf` (The URL might change, and the text may get updated. In order to reliably obtain the version of 13 December 2017, you can use the archive.org Wayback Machine: `https://web.archive.org/web/20180205122609/http://www-users.math.umn.edu/~dgrinber/comb/60610lectures2017-Galvin.pdf` .)

[Galvin20] David Galvin, *Notre Dame Math 43900: Problem Solving in Math*, lecture notes (work in progress), `https://www3.nd.edu/~dgalvin1/43900/43900_F20/index.html` (follow the overleaf link, once there allow javascript and note the "Download PDF" button).

[Gardne91] Martin Gardner, *Fractal Music, Hypercards and More...*, W. H. Freeman & Co. 1991.

[GelAnd17] Răzvan Gelca, Titu Andreescu, *Putnam and Beyond*, 2nd edition, Springer 2017.

[Golomb76] Solomon W. Golomb, *The "Sales Tax" Theorem*, Mathematics Magazine **49** (1976), no. 4, pp. 187–189.

[GolTch83] E. Goles, M. Tchuente, *Iterative behaviour of generalized majority functions*, Mathematical Social Sciences **4** (1983), pp. 197–204.

[Goodma15] Frederick M. Goodman, *Algebra: Abstract and Concrete*, edition 2.6, 1 May 2015.
`http://homepage.math.uiowa.edu/~goodman/algebrabook.dir/book.2.6.pdf` .

[Gould72] H. W. Gould, *A new greatest common divisor property of the binomial coefficients*, Fibonacci Quarterly **10** (1972), issue 6, pp. 579–584 and 628, errata in **11** (1973), issue 2, p. 168.

[Granvi05] Andrew Granville, *Binomial coefficients modulo prime powers*, preprint.
`https://web.archive.org/web/20181024055320/http://ebooks.bharathuniv.ac.in/gdlc1/gdlc1/EngineeringMergedLibraryv3.0/AndrewGranville/BinomialCoefficientsModuloPrimePowers(5579)/BinomialCoefficientsModuloPrimePowers-AndrewGranville.pdf`

[Grinbe07] Darij Grinberg, *The Vornicu-Schur inequality and its variations*, 13 August 2007.
`http://www.cip.ifi.lmu.de/~grinberg/VornicuS.pdf`

[Grinbe08] Natalia Grinberg, *Lösungsstrategien: Mathematik für Nachdenker*, Harri Deutsch 2008.

[Grinbe15] Darij Grinberg, *Notes on the combinatorial fundamentals of algebra*, 10 January 2019.
`http://www.cip.ifi.lmu.de/~grinberg/primes2015/sols.pdf`
The numbering of theorems and formulas in this link might shift when the project gets updated; for a "frozen" version whose numbering is guaranteed to match that in the citations above, see `https://github.com/darijgr/detnotes/releases/tag/2019-01-10` .

[Grinbe16] Darij Grinberg, *18.781 (Spring 2016): Floor and arithmetic functions*, 13 April 2019.

[Grinbe18] Darij Grinberg, *Zeckendorf family identities generalized*, arXiv:1103.4507v2, 10 June 2018.

[Grinbe19a] Darij Grinberg, *The Lucas and Babbage congruences*, 10 January 2019.

[Grinbe19b] Darij Grinberg, *math.stackexchange answer #2786815 (answer to "The selection of a linear recurrent sequence over the field is LRS too")*, Mathematics Stack Exchange.
`https://math.stackexchange.com/q/2786815`

[Grinbe19c] Darij Grinberg, *Fleck's binomial congruence using circulant matrices*, 14 July 2019.

[GriSne07] Charles M. Grinstead, J. Laurie Snell, *Introduction to Probability*, 2nd edition, 4 July 2006.

[GrKnPa94] Ronald L. Graham, Donald E. Knuth, Oren Patashnik, *Concrete Mathematics, Second Edition*, Addison-Wesley 1994.
See `https://www-cs-faculty.stanford.edu/~knuth/gkp.html` for errata.

[Guicha20] David Guichard, *An Introduction to Combinatorics and Graph Theory*, 30 January 2020.
`https://www.whitman.edu/mathematics/cgt_online/book/`

[Gunder10] David S. Gunderson, *Handbook of Mathematical Induction: Theory and Applications*, CRC Press 2010.

[Hadama09] Jacques Hadamard, *Lessons in Geometry: I. Plane Geometry*, AMS 2008.
`https://bookstore.ams.org/mbk-57`

[Hammac15] Richard Hammack, *Book of Proof*, 3.2rd edition, 24 July 2020.
`https://www.people.vcu.edu/~rhammack/BookOfProof/`

[Harju11] Tero Harju, *Lecture Notes in Combinatorial Enumeration*, 12 June 2016.
`https://users.utu.fi/harju/combinenum/combinenum11.htm`

[Heffer20] Jim Hefferon, *Linear Algebra*, 4th edition 2020.
`http://joshua.smcvt.edu/linearalgebra`

[Hender16] Nik Henderson, *What is Zeckendorf's Theorem?*, 23 July 2016.
`https://math.osu.edu/sites/math.osu.edu/files/henderson_zeckendorf.pdf`

[HilHog72] A. P. Hillman and V. E. Hoggatt, Jr., *A Proof of Gould's Pascal Hexagon Conjecture*, Fibonacci Quarterly **10** (1972), issue 6, pp. 579–584 and 598.

[Hoggat72] V. E. Hoggatt, Jr., *Generalized Zeckendorf theorem*, Fibonacci Quarterly **10** (1972), issue 1, pp. 89–93.

[Honsbe12] Ross Honsberger, *Episodes in Nineteenth and Twentieth Century Euclidean Geometry*, MAA 2012.

[Honsbe70] Ross Honsberger, *Ingenuity in Mathematics*, MAA 1970.

[Hung07] Pham Kim Hung, *Secrets in Inequalities, volume 1*, GIL 2007.

[Ivanov08] Nikolai V. Ivanov, *Linear Recurrences*, 17 January 2008. `https://nikolaivivanov.files.wordpress.com/2014/02/ivanov2008arecurrence.pdf`

[Ivanov18] O. A. Ivanov, *Portal through Mathematics*, Anneli Lax New Mathematical Library **52**, MAA Press 2018.

[Jarvin20] Olli Järviniemi, *Ollin Opas Olympiatason Ongelmanratkaisuun*, 9 November 2020. `https://matematiikkakilpailut.fi/kirjallisuus/OOOO.pdf`

[Jones75] James P. Jones, *Diophantine Representation of the Fibonacci numbers*, Fibonacci Quarterly **13** (1975), issue 1, pp. 84–88.

[Joyce17] David Joyce, *Introduction to Modern Algebra*, 5 December 2017. `https://mathcs.clarku.edu/~djoyce/ma225/algebra.pdf`

[Joyner08] W. D. Joyner, *Mathematics of the Rubik's cube*, 19 August 2008. `https://web.archive.org/web/20160304122348/http://www.permutationpuzzles.org/rubik/webnotes/` (link to the PDF at the bottom).

[Jukna11] Stasys Jukna, *Extremal Combinatorics with Applications in Computer Science*, 2nd edition, Springer 2011. See `http://www.thi.informatik.uni-frankfurt.de/~jukna/EC_Book_2nd/index.html` for a draft and errata.

[Kamp15] Peter H. van der Kamp, *Somos-4 and Somos-5 are arithmetic divisibility sequences*, Journal of Difference Equations and Applications **22** (2016), issue 4. See arXiv:1505.00194v1 for a preprint.

[Kedlay20] Kiran Kedlaya et al., *The Putnam archive*, 2020. `https://kskedlaya.org/putnam-archive/`

[Klivan18] Caroline J. Klivans, *The Mathematics of Chip-Firing*, CRC Press 2018.

[Knapp16] Anthony W. Knapp, *Basic Algebra*, digital second edition 2016. `http://www.math.stonybrook.edu/~aknapp/download.html`

[LamMos54] Joachim Lambek, Leo Moser, *Inverse and Complementary Sequences of Natural Numbers*, The American Mathematical Monthly, **61** (7), Aug-Sep 1954, pp. 454–458.

[Larson83] Loren C. Larson, *Problem-Solving Through Problems*, Springer 1983.

[Lee07] Hojoo Lee, *Topics in Inequalities: Theorems and Techniques*, 8 August 2007.
`https://www.isinj.com/mt-usamo/TopicsinInequalities1stedition-HojooLee(2007).pdf`

[LeeShi19] Jonghyeon Lee, Heesung Shin, *The chromatic polynomial for cycle graphs*, Korean J. Math. **27**, No 2 (2019) pp. 525–534, arXiv:1907.04320v1.

[LeLeMe16] Eric Lehman, F. Thomson Leighton, Albert R. Meyer, *Mathematics for Computer Science*, revised Tuesday 6th June 2018,
`https://courses.csail.mit.edu/6.042/spring18/mcs.pdf` .

[LeLoPo08] Hojoo Lee, Tom Lovering, and Cosmin Pohoata, *Infinity*, 19 October 2008.
`http://prac.im.pwr.wroc.pl/~kwasnicki/pl/o/infinity.pdf`

[Lengye06] Tamás Lengyel, *A counting based proof of the generalized Zeckendorf's theorem*, Fibonacci Quarterly **44** (2006), issue 4, pp. 324–325.

[Liu68] C. L. Liu, *Introduction to Combinatorial Mathematics*, McGraw–Hill 1968.

[Liu85] C. L. Liu, *Elements of discrete mathematics*, 2nd edition, McGraw–Hill 1985.

[Loehr11] Nicholas A. Loehr, *Bijective Combinatorics*, Chapman & Hall/CRC 2011.

[Loehr19] Nicholas A. Loehr, *An Introduction to Mathematical Proofs*, CRC Press 2020.
See `https://personal.math.vt.edu/nloehr/prfbook.html` for errata and solutions.

[Macgil17] Gary MacGillivray, *Discrete Math Study Materials*, 2017.
`https://www.math.uvic.ca/faculty/gmacgill/guide/index.html`

[Malouf92] Janice L. Malouf, *An integer sequence from a rational recursion*, Discrete Mathematics **110** (1992), pp. 257–261.

[Markus83] Aleksei Ivanovich Markushevich, *Recursion sequences*, Mir Publishers, Moscow, 2nd printing 1983.

[Mazur10] David R. Mazur, *Combinatorics: A Guided Tour*, MAA 2010.

[Mcclea17] John McCleary, *Exercises in (Mathematical) Style: Stories of Binomial Coefficients*, 2nd printing, MAA Press 2018.

[Melian01] María Victoria Melián, *Linear recurrence relations with constant coefficients*, 9 April 2001.
`http://matematicas.uam.es/~mavi.melian/CURSO_15_16/web_Discreta/recurrence.pdf`

[MerJoh19] Mircea Merca, Warren P. Johnson, *American Mathematical Monthly problem #12022 with solution*, The American Mathematical Monthly, 126:8 (2019), pp. 755–763.

[Mildor06] Thomas J. Mildorf, *Olympiad Inequalities*, 4 August 2006.
`https://www.isinj.com/mt-usamo/OlympiadInequalities-ThomasMildorf(2006).pdf`

[Narkie00] Władysław Narkiewicz, *The Development of Prime Number Theory: From Euclid to Hardy and Littlewood*, Springer 2000.

[Neale17] Vicky Neale, *Closing the Gap: The Quest to Understand Prime Numbers*, Oxford University Press 2017.

[Negut05] Andrei Negut, *Problems for the Mathematical Olympiads*, Gil 2005.

[Newste20] Clive Newstead, *An Infinite Descent into Pure Mathematics*, version 0.5 preview, 26 September 2020.
`https://infinitedescent.xyz`

[NiZuMo91] Ivan Niven, Herbert S. Zuckerman, Hugh L. Montgomery, *An Introduction to the Theory of Numbers*, 5th edition, Wiley 1991.
See `http://www-personal.umich.edu/~hlm/nzm/index.html` for errata and supplements.

[Pak10] Igor Pak, *Lectures on Discrete and Polyhedral Geometry*, 20 April 2010.
`https://www.math.ucla.edu/~pak/geompol8.pdf`

[Parvar11] Amir Hossein Parvardi, *Lifting The Exponent Lemma (LTE)*, version 6, 7 April 2011.
`https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.221.5543`

[PeWiZe97] Marko Petkovšek, Herbert S. Wilf, Doron Zeilberger, *A=B*, 27 April 1997, A.K. Peters 1996.
`https://www.math.upenn.edu/~wilf/AeqB.html`

[PolSze78] George Pólya, Gabor Szegő, *Problems and Theorems in Analysis I*, Springer 1978 (reprinted 1998).

[Polya73] George Polya, *How to Solve It: A New Aspect of Mathematical Method*, Princeton University Press, 2nd edition 1973.

[Polya81] George Polya, *Mathematical Discovery*, combined edition, John Wiley & Sons 1981.

[Poorte04] Alfred J. van der Poorten, *Elliptic curves and continued fractions*, arXiv:math/0403225v3.

[Riorda68] John Riordan, *Combinatorial Identities*, Robert E. Krieger Publishing Company, Huntington, NY 1979.

[RitHee13] Benoît Rittaud, Albrecht Heeffer, *The Pigeonhole Principle, Two Centuries Before Dirichlet*, The Mathematical Intelligencer, 36(2), pp. 27–29.

[Rivlin90] Theodore J. Rivlin, *Chebyshev Polynomials: From Approximation Theory to Algebra and Number Theory*, 2nd edition, Wiley 1990.

[Russel16] Matthew Christopher Russell, *Using experimental mathematics to conjecture and prove theorems in the theory of partitions and commutative and non-commutative recurrences*, PhD thesis at Rutgers University, 2016.

[Schole05] John Scholes aka kalva, *Maths problems*, ca. 2005.
The original site is no longer available, but parts of it have been archived in several places:
`https://webee.technion.ac.il/people/aditya/www.kalva.demon.co.uk/`
`https://prase.cz/kalva/putnam.html`
`https://web.archive.org/web/20070930165736/http://www.kalva.demon.co.uk/`
`https://web.archive.org/web/20020927115327/http://www.kalva.demon.co.uk/`

[Shapir76] L. W. Shapiro, *A Catalan triangle*, Discrete Mathematics **14** (1976), pp. 83–90.

[ShChYa62] David Oskarovich Shklarsky, Nikolai Nikolaevich Chentzov, Isaak Moiseevich Yaglom, *The USSR Olympiad Problem Book*, Freeman 1962 (reprinted by Dover 1993).

[Smid09] Vita Smid, *Inclusion-Exclusion Principle: Proof by Mathematical Induction*, 2 December 2009.
`https://faculty.math.illinois.edu/~nirobles/files453/iep_proof.pdf`

[SpeStu09] David Speyer, Bernd Sturmfels, *Tropical Mathematics*, Mathematics Magazine **82** (2009), pp. 163–173.

[Spivey19] Michael Z. Spivey, *The Art of Proving Binomial Identities*, CRC Press 2019.

[SpQuBa20] Eckard Specht, Erhard Quaisser, Patrick Bauermann, *50 Jahre Bundeswettbewerb Mathematik: Die schönsten Aufgaben*, 2te Auflage, Springer 2020.

[Stanle01] Richard P. Stanley, *Enumerative Combinatorics, volume 2*, Cambridge University Press 2001.
See `http://math.mit.edu/~rstan/ec/` for errata.

[Stanle11] Richard P. Stanley, *Enumerative Combinatorics, volume 1*, Second edition, version of 15 July 2011. Available at `http://math.mit.edu/~rstan/ec/` .
See `http://math.mit.edu/~rstan/ec/` for errata.

[Stanle15] Richard P. Stanley, *Catalan Numbers*, 1st edition, Cambridge University Press 2015.

[Steele04] J. Michael Steele, *The Cauchy–Schwarz Master Class: An Introduction to the Art of Mathematical Inequalities*, Cambridge University Press 2004.
See `http://www-stat.wharton.upenn.edu/~steele/Publications/Books/CSMC/CSMC_Errata.pdf` for errata.

[Stein09] William Stein, *Elementary Number Theory: Primes, Congruences, and Secrets*, Springer 2009.

[Steinb06] Mark Steinberger, *Algebra*, 31 August 2006.
`https://web.archive.org/web/20180821125315/https://www.albany.edu/~mark/algebra.pdf`

[Strasz65] S. Straszewicz, *Mathematical Puzzles and Problems from the Polish Mathematical Olympiads*, Pergamon Press 1965.

[Strick20] Neil Strickland, *Linear mathematics for applications*, 11 February 2020.
`https://neilstrickland.github.io/linear_maths/notes/linear_maths.pdf`

[Suksom20] Warut Suksompong, *Graph Cycles and Olympiad Problems*, Mathematics Magazine **93** (2020), pp. 284–287.

[Sved84] Marta Sved, *Counting and Recounting: The Aftermath*, The Mathematical Intelligencer **6** (1984), no. 2, pp. 44–45.

[Swanso20] Irene Swanson, *Introduction with Analysis with Complex Numbers*, 2020.
`https://web.archive.org/web/20201012174324/https://people.reed.edu/~iswanson/analysisconstructR.pdf`

[TAoCP3] Donald A. Knuth, *The Art of Computer Programming, volume 3: Sorting and Searching*, 2nd edition, Addison–Wesley 1998.
See `https://www-cs-faculty.stanford.edu/~knuth/taocp.html` for errata.

[TAoCP4A] Donald A. Knuth, *The Art of Computer Programming, volume 4A: Combinatorial Algorithms, Part 1*, Addison–Wesley 2011.
See `https://www-cs-faculty.stanford.edu/~knuth/taocp.html` for errata.

[TatWin05] Leung Tat-Wing, *The Method of Infinite Descent*, Mathematical Excalibur **10** (2005), nr. 4.

[Tomesc85] Ioan Tomescu, *Problems in Combinatorics and Graph Theory*, translated by Robert A. Melter, John Wiley & Sons 1985.

[UspHea39] J. V. Uspensky, M. A. Heaslet, *Elementary Number Theory*, 1st edition, McGraw-Hill 1939.
See `https://projecteuclid.org/euclid.bams/1183502492` for some corrections.

[VelWag20] Daniel J. Velleman, Stan Wagon, *Bicycle or Unicycle? A Collection of Intriguing Mathematical Puzzles*, AMS/MAA Problem Books **36**, MAA Press 2020.

[Vorobi02] Nicolai N. Vorobiev, *Fibonacci Numbers*, Translated from the Russian by Mircea Martin, Springer 2002 (translation of the 6th Russian edition).

[Wagner05] Carl G. Wagner, *Basic Combinatorics*, 14 February 2005.
`http://www.math.utk.edu/~wagner/papers/comb.pdf`

[Wagner08] David G. Wagner, *C&O 330: Introduction to Combinatorial Enumeration*, 9 August 2008.
`https://web.archive.org/web/20110124001354/http://www.math.uwaterloo.ca/~dgwagner/MATH249/co330.pdf`

[Walker87] Elbert A. Walker, *Introduction to Abstract Algebra*, Random House-/Birkhauser, New York, 1987.

[Wallis85] John Wallis, *A Treatise of Algebra*, London 1685.

[Wemyss14] Michael Wemyss, *A very elementary proof that the Somos 5 sequence is integer valued*, 28 April 2014.
`https://www.maths.gla.ac.uk/~mwemyss/Somos5proof.pdf`

[White10a] Dennis White, *Math 4707: Chromatic Polynomials*, 3 November 2010.
`http://www-users.math.umn.edu/~reiner/Classes/Chromatic.pdf`

[White10b] Dennis White, *Math 4707: Inclusion-Exclusion and Derangements*, 18 October 2010.
`http://www-users.math.umn.edu/~reiner/Classes/Derangements.pdf`

[YagYag64] Akiva M. Yaglom, Isaac M. Yaglom, *Challenging Mathematical Problems with Elementary Solutions, volume I: Combinatorial Analysis and Probability Theory*, Holden-Day 1964 (reprinted by Dover 1987).

[YagYag67] Akiva M. Yaglom, Isaac M. Yaglom, *Challenging Mathematical Problems with Elementary Solutions, volume II: Problems from Various Branches of Mathematics*, Holden-Day 1967 (reprinted by Dover 1987).

[Yiu12] Paul Yiu, *Problem Solving and Recreational Mathematics*, 2 August 2012. `http://math.fau.edu/Yiu/Oldwebsites/PSRM2012/00PSRM2012Chapters1to44.pdf`

[Zagier96] Don Zagier, *Problems posed at the St Andrews Colloquium, 1996*. `http://www-groups.mcs.st-andrews.ac.uk/~john/Zagier/Problems.html`

[ZawHit09] Alexander Zawaira, Gavin Hitchcock, *A Primer for Mathematics Competitions*, Oxford University Press 2009.

[Zeitz17] Paul Zeitz, *The Art and Craft of Problem Solving*, 3rd edition, Wiley 2017.

[Zhao20] Yufei Zhao, *Math Olympiad training handouts*, `https://yufeizhao.com/olympiad/` .

[Zill17] Dennis G. Zill, *A First Course in Differential Equations with Modeling Applications*, Cengage 2017.