Rigorous Elementary Mathematics

Volume 1: Algebra

Samer Seraj

Existsforall Academy

Copyright

ISBN 978-1-7389501-0-2

Acknowledgements

> "At the age of eleven, I began Euclid, with my brother as tutor. This was one of the great events of my life, as dazzling as first love. I had not imagined there was anything so delicious in the world. From that moment until I was thirty-eight, mathematics was my chief interest and my chief source of happiness."
>
> – *Bertrand Russell, Autobiography*

> "Mathematicians, like Proust and everyone else, are at their best when writing about their first love."
>
> – *Gian-Carlo Rota, Discrete Thoughts*

I express my gratitude to:

- The Almighty Creator, for providing me with this blessed and privileged life.
- My parents, for financing my mathematical education, and for supporting me during the time that this book series was written.
- My friends, for their companionship and for listening to me talk about mathematics.
- Euclid, for writing the *Elements*, which showed the world the meaning of eternal rigour.

Special thanks is extended to Warren Bei for reading the manuscript and offering numerous suggestions, many of which were implemented. Any remaining mathematical errors or mistakes in the typesetting are my responsibility alone.

Contents

Preface

> "In studying a philosopher, the right attitude is neither reverence nor contempt, but first a kind of hypothetical sympathy, until it is possible to know what it feels like to believe in his theories, and only then a revival of the critical attitude, which should resemble, as far as possible, the state of mind of a person abandoning opinions which he has hitherto held. Contempt interferes with the first part of this process, and reverence with the second. Two things are to be remembered: that a man whose opinions and theories are worth studying may be presumed to have had some intelligence, but that no man is likely to have arrived at complete and final truth on any subject whatever."
>
> – *Bertrand Russell, A History of Western Philosophy*

Mathematics is the study of ultimate regularity. Regularity entails order or predictability. Its antithesis is chaos. When there is regularity, there are discernible objects at play. In other words, there is structure. Wherever there is structure, there is symmetry. Symmetry means that, while one aspect of the object changes, another remains unchanged. The present trilogy is an effort to rigorously systematize and provide an exposition of those aspects of elementary mathematics that appeal to the author. In the course of writing, it became evident that there are three recurring themes among the proof techniques used, all of which are forms of symmetry:

1. The discrete Fubini's principle instructs us to write the same thing in two different ways. For example, we have applied this principle in several ways:
   - There is a theorem dedicated to the ways in which one may switch from iterating through the rows of a matrix to columns or vice versa (Theorem 3.19).
   - It is used to prove the Girard-Newton sums (Theorem 10.53).
   - It is used in the proof of Chebyshev's inequality (Corollary 11.19).
   - The presented proof of Hölder's inequality (Theorem 11.25) utilizes it.
2. Antisymmetry in a partial order is a powerful method of proof that lets us break down the strong notion of equality into the conjunction of two individually weaker statements. The two examples from elementary algebra that we have used in this book are:
   - Set equality can be broken into two subset relations, specifically $A = B$ if and only if both $A \subseteq B$ and $B \subseteq A$ (see Theorem 1.3).
   - Real number equality can be split into two inequality relations, meaning $x = y$ if and only if both $x \leq y$ and $y \leq x$ (see Definition 4.38).

3. Modding out by an equivalence relation allows us to focus on the essential properties of objects which are preserved under the relation. Most of this volume focuses on the usual equality, such as numbers being equal to each other if they are exactly the same number, or functions being equal to each other if they have the same domain and map the same inputs to the same outputs (Definition 1.23), or formal polynomials being equal if they represent the same sequence of coefficients (Definition 10.4). The other volumes present more interesting notions of equivalence.

It is our hope that the reader will keep these proof techniques in mind while reading the book, and that the impression of the importance of symmetry will grow as the reader encounters the methods time and again.

The intended audience consists of students of math contests, competitions, and olympiads who want want to take a rigorous second look at the results that they might be accustomed to taking for granted, and teachers, coaches, and trainers who want to reinforce their own understanding of what they teach.

Suggestions, comments, and error submissions would be greatly appreciated. These may include suggestions for strengthening or generalizing theorems, and additional material. Messages may be sent to

academy@existsforall.com

Samer Seraj
Mississauga, Ontario, Canada
March 27, 2023

Chapter 1

Sets and Maps

> "From the paradise, that Cantor created for us, no-one shall be able to expel us."
>
> – *David Hilbert, Über das Unendliche*

> "The axiom of choice is obviously true, the well-ordering principle obviously false, and who can tell about Zorn's lemma?"
>
> – *Jerry Bona*

Sets are collections of elements. One way of formulating all objects in the mathematical universe is to boil them down to sets, and, ultimately, combinations of the empty set. We will look at some of the axioms that sets must satisfy and standard methods of constructing sets. Then we will study functions. A function is a machine that takes inputs from a set and, for each input, provides exactly one output in another set. There are numerous general properties of functions that we will see here in preparation for upcoming material. We will end with a proof method that sets allow us to describe: proof by induction and its variants, which are the well-ordering principle and complete induction.

1.1 Sets

The notation of standard logical operators will be used here. Brief definitions may be found in the list of symbols at the end of the book, under the heading of "Logic."

Definition 1.1. A **set** is a collection of elements, where the elements are not ordered (meaning there is no "first" or "second" or otherwise ranked element) and there are no repeated elements. If physically possible on paper, the elements of a set can be written out by enclosing its elements in any order from left to right within curly brackets like $\{\ldots\}$. We denote that x is an **element of** the set S by writing $x \in S$; if x is not an element of S, we denote it as $x \notin S$. The notions of "element" and "element of" are undefined, but we have to start somewhere. The set with no elements is called the **empty set** and is denoted by $\varnothing$. A set that is not the empty set is called **non-empty**. Sets S and T are said to be **equal** if

$$\forall x, (x \in S \iff x \in T).$$

Equality of two sets S and T is written using an **equals sign** $=$ like $S = T$. Equal sets may be substituted in for each other in math. A **singleton** is a set of a single element, that is a set of the form $\{x\}$, where x is the only element. An **unordered pair** is a set with exactly two elements like $\{x, y\}$.

Example. In the modern mathematical universe, all objects (such as numbers, spaces, and functions) can be expressed in terms of sets, though most of these constructions are too advanced for us, and often too complicated for practical usage. The set of integers is denoted by

$$\mathbb{Z} = \{\ldots, -2, -1, -0, 1, 2, \ldots\},$$

the set of positive integers by

$$\mathbb{Z}_+ = \{1, 2, 3, \ldots\},$$

and the set of non-negative integers by

$$\mathbb{Z}_{\geq 0} = \{0, 1, 2, 3, \ldots\}.$$

In some contexts, the notation $\mathbb{N}$ and the term "natural numbers" is used but we have intentionally avoided this notation and terminology because there is no universal agreement on whether the naturals should include 0. We will study integers in some detail in Chapter 2 and in great detail in Volume 3.

Definition 1.2. A set X is said to be a **subset** of a set Y if

$$\forall x, (x \in X \implies x \in Y),$$

and this is denoted by $X \subseteq Y$. A **proper subset** X of a set Y is a subset other than Y itself. In other words, there exists an element (not necessarily a unique element) that is inside Y but outside X. The assertion that X is a proper subset of Y is denoted by $X \subsetneq Y$. If $X \subseteq Y$, then Y is called a **superset** of X. If we wish to emphasize the superset relation instead of the subset relation, then is acceptable to reverse the notation and write $Y \supseteq X$ instead of $X \subseteq Y$, and $Y \supsetneq X$ instead of $X \subsetneq Y$. We will avoid the notation $\subset$ and $\supset$ to prevent ambiguity; $\subseteq$ and $\supseteq$ denote that equality is possible (but not necessary), whereas $\subsetneq$ and $\supsetneq$ denote that equality is impossible. To denote that X is not a subset of Y we write $X \not\subseteq Y$, and $Y \not\supseteq X$ denotes that Y is not a superset of X.

Example. For every set S, it is vacuously true that $\varnothing \subseteq S$. This includes the fact that $\varnothing$ is a subset of itself. In fact, $\varnothing$ is a proper subset of every non-empty set, yet, in contrast, the only subset of $\varnothing$ is $\varnothing$, so the empty set has no proper subsets.

Theorem 1.3. The subset relation on the subsets of a particular set produces a partial order, which we will define more generally in Definition 4.44. What this means is that for all subsets X, Y, Z of a particular set, the following properties hold:

1. Reflexivity: $X \subseteq X$
2. Antisymmetry: if $X \subseteq Y$ and $Y \subseteq X$, then $X = Y$. Reflexivity gives the converse, thanks to the substitution property of equality.
3. Transitivity: if $X \subseteq Y$ and $Y \subseteq Z$, then $X \subseteq Z$

All of these properties follow from the definition of the subset relation. We leave the verification of the details to the reader.

Definition 1.4. One way of constructing sets is to use **set builder notation**. This allows us to denote and define sets like

$$\{x \in S : p(x)\} = \{x : x \in S \wedge p(x)\},$$

where $p(x)$ is a predicate (meaning, a property) in the variable x. The colon is read as "such that" and could be replaced by a vertical bar $|$. So we are defining $\{x \in S : p(x)\}$ to be the subset of S such that those are all the elements $x \in S$ that make $p(x)$ true. The variable x could be replaced by a list of distinct variables.

One way in which we might wish extend set builder notation is to say that "this is the set of all possible sets that have a certain property." So there is a certain condition and we want the power to summon into a set *every* set that satisfies this predicate, not just a subset of some fixed set. Such a strong ability quickly leads to trouble, as shown by Russell's paradox below.

Theorem 1.5 (Russell's paradox)**.** Suppose it is possible to summon all sets x in the mathematical universe that satisfy a certain predicate $p(x)$ in the variable x. Then we can use this power to deduce a contradiction. Due to Russell's paradox, we choose to restrict the ability to build sets according to satisfying a certain predicate (in set builder notation) by only allowing the ability to carve out a subset of a set, and not carve out a part of the entire mathematical universe. That is, $\{x \in S : p(x)\}$ is allowed but $\{x : p(x)\}$ is not allowed; the latter is called "unrestricted comprehension." But the issue of Russell's paradox remains if there is a set whose elements are all possible sets, so we must also stipulate that there is no "set of all sets."

Proof. Suppose, for contradiction, that unrestricted comprehension is acceptable. This allows us to define the set

$$S = \{x : x \notin x\}.$$

Then there are two possibilities: $S \in S$ or $S \notin S$.

- If $S \in S$, then $S \notin S$ by the definition of S. This contradicts the premise of the case in which we are working.
- If $S \notin S$, then $\neg(S \notin S)$ must be true by the definition of S, which is equivalent to $S \in S$. Again, we have contradicted the premise of the case in which we are working.

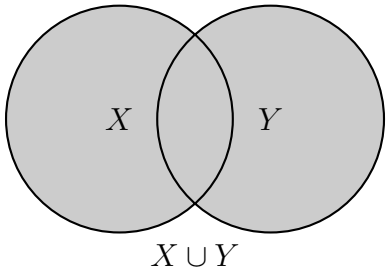
In either case, we have a contradiction. Thus, we cannot allow unrestricted comprehension, and we choose the safe-seeming option of summoning only subsets of known sets.
Now suppose there exists a set V that contains all sets. Then we can still define

$$T = \{x \in V : x \notin x\}.$$

Once again, $T \in T$ leads to $T \notin T$ and $T \notin T$ leads to $T \in T$. Since Russell's paradox takes hold again, there can be no such object as the "set of all sets." ■

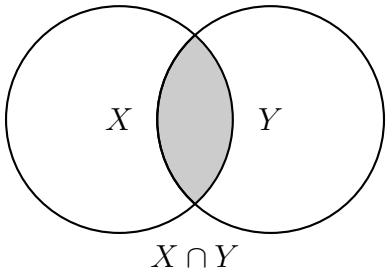
Definition 1.6. Several common constructions of sets using set builder notation are as follows, given sets X and Y:

- **Union**: $X \cup Y = \{x : x \in X \vee x \in Y\}$



$X \cup Y$

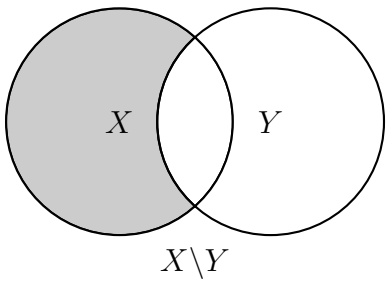
- **Intersection**: $X \cap Y = \{x : x \in X \wedge x \in Y\}$



$X \cap Y$

- **Difference**: also called the **relative complement** of X excluding Y, this is

$$X \backslash Y = X - Y = \{x : x \in X \wedge x \notin Y\}$$

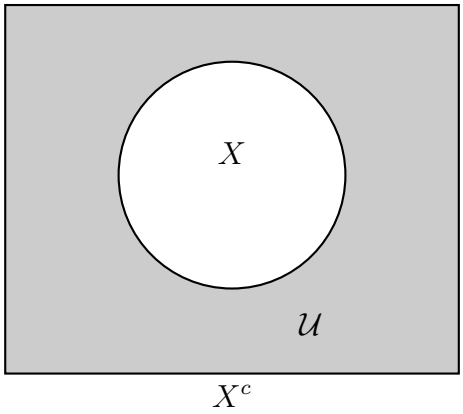


$X \backslash Y$

- **Complement**: if all relevant sets are subsets of some "universal" set $\mathcal{U}$, then we define

$$\overline{X} = X^c = \{x : x \in \mathcal{U} \wedge x \notin X\} = \mathcal{U} \backslash X.$$

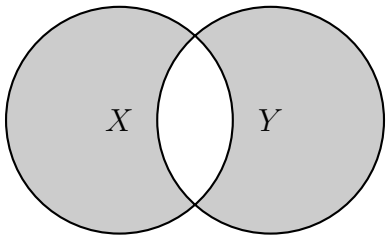
Any mention of a complements implies that we are working in an overarching set $\mathcal{U}$. A very useful application of the complement is to write the difference as $X \backslash Y = X \cap (Y^c)$.



X^c

- **Symmetric difference**: using difference twice,

$$X \oplus Y = (X \backslash Y) \cup (Y \backslash X) = (X \cup Y) \backslash (X \cap Y)$$



$X \oplus Y$

There is no universally accepted order of operations on sets, but it is generally accepted that complement takes precedence over everything other than brackets (this is similar to exponents for numbers). So we can write $X \cap Y^c$ instead of $X \cap (Y^c)$.

Lemma 1.7. If X, Y, Z are sets, then:

1. $X \subseteq Z$ and $Y \subseteq Z$ if and only if $X \cup Y \subseteq Z$.

2. $Z \subseteq X$ and $Z \subseteq Y$ if and only if $Z \subseteq X \cap Y$.

These should be clear from the definition of the subset relation.

Theorem 1.8. It is easy to see that for any set X, taking the complement twice yields $(X^c)^c$. Also, $\varnothing^c = \mathcal{U}$ and $\mathcal{U}^c = \varnothing$. The following table provides more such equalities. These identities can be useful in simplifying expressions involving sets. Let X, Y, Z be sets, $\varnothing$ denote the empty set as usual, and, if relevant, let $\mathcal{U}$ denote a universal set in which all work is being done. Then:

Identity	Union	Intersection
Complement	$X \cup X^c = \mathcal{U}$	$X \cap X^c = \varnothing$
Annihilation	$X \cup \mathcal{U} = \mathcal{U}$	$X \cap \varnothing = \varnothing$
Identity	$X \cup \varnothing = X$	$X \cap \mathcal{U} = X$
Idempotent	$X \cup X = X$	$X \cap X = X$
Absorption	$X \cup (X \cap Y) = X$	$X \cap (X \cup Y) = X$
Commutative	$X \cup Y = Y \cup X$	$X \cap Y = Y \cap X$
Associative	$X \cup (Y \cup Z) = (X \cup Y) \cup Z$	$X \cap (Y \cap Z) = (X \cap Y) \cap Z$
Distributive	$X \cup (Y \cap Z) = (X \cup Y) \cap (X \cup Z)$	$X \cap (Y \cup Z) = (X \cap Y) \cup (X \cap Z)$
De Morgan's	$(X \cup Y)^c = X^c \cap Y^c$	$(X \cap Y)^c = X^c \cup Y^c$

The associative laws allow us to write a sequence of unions or a sequence of intersections without parentheses, but not if unions and intersections are mixed. The commutative laws allow us to write a sequence of unions in any order, and similarly for intersections, but again not if they are mixed. When unions and intersections both appear, other equivalences like the absorption laws and distributive laws may be helpful. Unlike set union and intersection, set difference is neither commutative not associative, meaning it is not necessarily true that $X \backslash Y = Y \backslash X$. However, the symmetric difference of sets is both commutative and associative.

Set identities can be proven by using other identities, logical definitions and logical equivalences of the predicates involved, or antisymmetry of the subset relation. We will not cover any of these proofs because they are too cumbersome and space-consuming.

Corollary 1.9. By repeated applications of de Morgan's laws, these two laws may be extended to the union of many sets or intersection of many sets:

$$(p_1 \cup p_1 \cup \cdots \cup p_n)^c \equiv p_1^c \cap p_2^c \cap \cdots \cap p_n^c,$$
$$(p_1 \cap p_1 \cap \cdots \cap p_n)^c \equiv p_1^c \cup p_2^c \cup \cdots \cup p_n^c.$$

Thanks to de Morgan's laws, it is often more convenient to work with complements than to take a difference with a universal set.

Problem 1.10. Let A and B be sets such that their union is a universal set $\mathcal{U}$ and their intersection is the empty set $\varnothing$. Prove that $A^c = B$.

Theorem 1.11. Without getting into the names or technical statements of the axioms involved, here are a few properties satisfied by sets:

- No set can be an element of itself.
- For any two sets X, Y, there is a "pairing" set $\{X, Y\}$ containing the two. This can be used repeatedly to produce a set of n sets. Or we can take $X = Y$ to get that there is a singleton set $\{X\}$ (which is different from X).

- If S is a set of sets, then there is a set that is the union of the elements of S. For example,
$$\bigcup\{\{1,2,3\},\{2,3,4\},\{6,7\}\} = \{1,2,3,4,6,7\}.$$

Definition 1.12. Two sets A and B are said to be **disjoint** if they have no shared elements, meaning $A \cap B = \varnothing$. A non-empty set $\mathcal{A}$ of sets is said to be **pairwise disjoint** if each pair of distinct sets $A, B \in \mathcal{A}$ is disjoint; by distinct, we mean that $A \neq B$. We will not classify the empty set as pairwise disjoint so that there will never be a need to clarify that a pairwise disjoint set is non-empty; one could easily define otherwise if desired. The union of a pairwise disjoint set of sets is is called a **disjoint union**. If we want to emphasize the fact that a union is specifically a disjoint union, we use the rigid $\sqcup$ symbol instead of the rounded $\cup$ symbol.

Example. Since $\varnothing \cap \varnothing = \varnothing$, our definition says that $\varnothing$ and $\varnothing$ are disjoint. All sets whose only element is one set are pairwise disjoint. This means that, although we have defined the empty set $\varnothing$ to not be pairwise disjoint, the set containing the empty set $\{\varnothing\}$ is pairwise disjoint.

Definition 1.13. An **ordered pair** is similar to a set of two elements, except in two regards: order matters, meaning there is a "firs2 t" element and a "second" element, and the two elements are allowed to be the same. If the first element is a and the second element is b, then the ordered pair is written as (a, b). According to Kuratowski, it is possible to define this ordered pair using sets as
$$\{\{a\},\{a,b\}\},$$
but that would not serve any purpose at this level. Two ordered pairs (a, b) and (c, d) are said to be **equal** if they are equal component-wise, meaning $a = c$ and $b = d$.

Definition 1.14. Given an ordered pair of sets (A, B), we denote and define its **Cartesian product** to be the set
$$A \times B = \{(a, b) : a \in A, b \in B\}.$$
So it is the set of all ordered pairs such that the first entry is from A and the second entry is from B. This definition will be extended to a Cartesian product of n ordered sets in Definition 3.4 after we define lists in Definition 3.1.

Definition 1.15. Let X and Y be sets. A **binary relation** on the ordered pair of sets (X, Y) is a subset of $X \times Y$. If the name of the relation is R, the fact that $(x, y) \in R$ may be denoted by xRy, which we utter as "x is related to y". In a sense, a binary relation assigns a "true" or "false" value to each element of $X \times Y$, with the true ones being the elements of R.

Definition 1.16. If S is a non-empty set, an **equivalence relation** on S is a binary relation $\sim$ on $S \times S$ that satisfies the following three properties for all $a, b, c \in S$:

1. Reflexive: $a \sim a$
2. Symmetric: $a \sim b$ implies $b \sim a$
3. Transitive: $a \sim b$ and $b \sim c$ together imply $a \sim c$

Definition 1.17. Given an equivalence relation $\sim$ on S, and given an element $a \in S$, the **equivalence class** of a is denoted by and defined as

$$[a] = \{b \in S : a \sim b\},$$

which is the set of all $b \in S$ to which a is related. If a is an element of S, and C is an equivalence class of S under $\sim$ such that $a \in C$, then a is said to be a **representative** of C.

Theorem 1.18. An equivalence relation on S splits S into non-empty disjoint equivalence classes whose union is S. Conversely, a partition of a set induces an equivalence relation such that two elements are equivalent to each other if and only if they are in the same component. The following are some useful consequences of this partition idea:

1. Every element of S is in some equivalence class, so the classes cover all of S. On the other hand, no element is in more than one class.

2. All elements in an equivalence class are pairwise equivalent to each other under this equivalence relation.

3. No two elements from different equivalence classes are equivalent to each other.

4. If $a \in S$ is an element of an equivalence class C, then the class of a is equal to C, meaning $[a] = C$. As a result, if $b \in [a]$, then $[a] = [b]$.

5. If $a \sim b$ then $[a] = [b]$. Similarly, any two equivalence classes that share an element are in fact equal. In a sense, a single common element causes the classes to collapse into each other.

6. If $a \nsim b$ then $[a] \cap [b] = \varnothing$. Similarly, any two different equivalence classes are disjoint.

The proofs of these facts are cumbersome so we will not cover them here. The fact the the set of equivalences classes forms a partition of the set will be proven when we study group theory in Volume 2.

Example. Equivalence relations are among the most important instances of binary relations, though we will see other examples of binary relations, like orders in Section 4.3. The benefit of an equivalence relation is that, even though it does not necessarily mean exact equality, it allows us to zero in on some similar quality of objects or some property that is preserved between two equivalence classes no matter which of their representatives are chosen. Examples of equivalence relations include the congruence or similarity of triangles and polygons, equipollence of directed line segments in the study of vectors (angles between vectors are preserved no matter which representative directed line segments are chosen), equipotence of sets with regards to cardinality, and congruence of integers in modular arithmetic. As a basic example, Definition 1.19 asserts that equality relations, such as those between sets or numbers or polynomials, are equivalence relations.

Definition 1.19. Equality of objects, such as numbers, in a set can be interpreted as the special "finest" equivalence relation on the set, meaning the equivalence classes are precisely the singletons of the set. So equality holds between two objects if and only if they are the

same object. This relation is denoted by the equals sign = like $A = B$, which is called an **equation**. Its negation is denoted by $A \neq B$, which means A and B are different objects. As an equivalence relation, equality satisfies reflexivity, symmetry and transitivity.

Theorem 1.20 (Substitution property)**.** If a and b are equal objects meaning $a = b$, then as equal objects, we can replace a with b or b with a in an equation or any other mathematical statement without altering the truth value of the statement.

The substitution property should make sense intuitively.

Definition 1.21. There are several definitions and applications of the term **well-defined** of which we are aware, all of them being about avoiding ambiguity in interpretation.

- A representation of an object is called well-defined if there is no other such representation. For example, for each integer $b \geq 2$, integers have exactly one terminating representation in base-b.

- A function is well-defined if it produces the same output for a given input, regardless of which of the multiple representations of the input are used. For example, $0.999\ldots$ and 1.0 represent the same rational number, but a function should be independent of the the symbolic representation. An anti-example is the "function" that takes rational numbers in fraction form and outputs the numerator. This is not a function at all because there are (infinitely) many fractional representations of each rational number.

- A function that is defined on a set of equivalence classes is well-defined if, even if the function is defined in terms of elements of the equivalence classes, the function does not produce different outputs for different elements of the same class. For example, in the arithmetic of $\mathbb{Z}_n$ that we will mention in Volume 3, the addition

  $$[a] + [b] = [a + b]$$

  is well-defined as it is independent of the choice of representatives a and b of their respective classes.

- Notation is well-defined if we take advantage of some convenient property to relax the notation without introducing ambiguity. For example, associativity of multiplication allows us to write $a \cdot b \cdot c$ without parentheses, as opposed to $a \cdot (b \cdot c)$ or $(a \cdot b) \cdot c$.

If a representation, function or piece of notation is not well-defined, then we call it **ill-defined**.

1.2 Functions

Definition 1.22. A **variable** is an unknown or unspecified mathematical object, and may be represented using a symbol such as x. A variable derives its name from the fact that it is allowed to *vary* over a certain domain (which is often the real numbers).

Definition 1.23. Let X and Y be non-empty sets. A binary relation R on (X, Y) is called **functional** if xRy_1 and xRy_2 together imply that $y_1 = y_2$; so each $x \in X$ is related to *at most* one element of Y. A relation R is called **serial** if

$$\forall x \in X, \exists y \in Y, xRy,$$

so each $x \in X$ is related to *at least* one element of Y. A **function** from X to Y is a binary relation on (X, Y) that is both functional and serial. Less formally, a function $f : X \to Y$ takes inputs from X, and for each $x \in X$, produces exactly one output $y \in Y$, where the output (otherwise known as the **image** of x under f) is denoted by $f(x)$. We usually refer to the function itself as f, though the function can also be denoted by $f(x)$ if x is a general variable that does not represent any particular element of X. In the equation $y = f(x)$, there is exactly one y value assigned to each x value, so x is called the **independent variable** and y is called the **dependent variable**. While notation such as $f(x) = x + 1$ is usually used to define how outputs are found using inputs, there is also the "maps to" notation $x \mapsto x + 1$. There are some sets associated with a function:

- The **domain** X of f is the set of all inputs and is denoted by $\mathrm{Dom}(f)$.
- The **range** of f is the set of all outputs that actually occur and is denoted by $\mathrm{Rng}(f)$. It is accepted in set theory that the range of a function is always a set. So if f is a function with domain X, then the range set denoted by and defined as

  $$\{f(x) : x \in X\} = \{y : \exists x, (x \in X \wedge y = f(x))\}$$

  exists. We have defined set-builder notation in Definition 1.4, and it can be modified in the way described here. If the domain is X, the range of f may be denoted by $f(X)$, which is a special case of the more general image notation in Definition 1.25.
- It is not always practical to write down the exact range, so we often prefer to deal with a **codomain** Y instead, which is a set that contains the range. When the notation $f : X \to Y$ is used instead of just f, it means the codomain Y is specified. If a codomain Y is specified, we can say that f is **Y-valued**.

Two functions are said to be **equal** if they have the same domain and they produce the same output for the same element of the domain. In particular, it is possible for two functions to be equal without having the same codomain. We denote that f and g are equal functions by $f = g$.

Example. The most basic functions are **constant functions**, where $f(x) = c$ for a fixed value c for all x in the domain. After constant functions, the next most basic function with domain X is $\mathrm{Id}_X : X \to X$, which is defined as $\mathrm{Id}_X(x) = x$ for all $x \in X$. This is called the **identity function** on X.

Definition 1.24. If the domain of a function f is **restricted** to a subset I of the domain, then we denote the new function with the restricted domain by $f|_I$.

Definition 1.25. For a set of values I in the domain of f, the image of I under f is denoted by and defined as

$$f(I) = \{f(x) : x \in I\}.$$

This is called the **image** of I under f.

Definition 1.26. If there are functions $f : X \to Y$ and $g : Y \to Z$, the **composite function** or **composition** $g \circ f : X \to Z$ is defined as

$$(g \circ f)(x) = g(f(x))$$

for all $x \in X$. In this case, it is said that g is **left-composed** with f and f is **right-composed** with g. Composition can occur several times or not at all:

- If a function f is applied n times, then we denote the composition

$$\underbrace{(f \circ f \circ \cdots \circ f)}_{n \text{ copies of } f}$$

  by f^n. This is different from exponents and does *not* denote $(f(x))^n$. Note that in order to compose a function with itself, its range has to be a subset of its domain.

- By definition, f^1 is the same as just f.

- For any function f, the notation f^0 means f is applied zero times, so $f^0(x) = x$ for all $x \in X$. In other words, f^0 is the identity function on the domain of f.

Theorem 1.27. Function composition is associative. This means that for functions $f : X \to Y, g : Y \to Z, h : Z \to W$, it holds that

$$(h \circ g) \circ f = h \circ (g \circ f).$$

So it is not ambiguous to simply write it as $h \circ g \circ f$ without any parentheses, and similarly for any other finite sequence of compositions. However, function composition is not commutative in general.

So far, we know what f^n means for $n \geq 0$, from which arises the question of how f^n should be interpreted for negative integers n. If we had a definition of f^{-1}, we could define f^{-n} as n applications of f^{-1}, but we still need a reasonable definition of f^{-1}. For non-negative integers n, it is true that $f \circ f^n = f^{n+1} = f^n \circ f$. If we want this property to hold for $n = -1$, then we would require something along the lines of (but not necessarily exactly)

$$f \circ f^{-1} = \mathrm{Id} = f^{-1} \circ f,$$

where Id is an identity function on some set. So, intuitively, f^{-1} should undo or invert f, and vice versa. As we will see, a function has an inverse if and only if certain circumstances hold. Moreover, we will have to be careful about specifying the domain and codomain of f^{-1}.

Definition 1.28. Let $f : X \to Y$ be a function. Then:

1. A function $g : Y \to X$ is called a **left-inverse** of f if $g \circ f = \mathrm{Id}_X$.

2. A function $g : Y \to X$ is called a **right-inverse** of f if $f \circ g = \mathrm{Id}_Y$.

3. A function $g : Y \to X$ is called an **inverse** of f if g is both a left-inverse and a right-inverse of f. Then g is denoted as f^{-1}. A function f is called **invertible** if it has an inverse g.

Theorem 1.29. Suppose $g : Y \to X$ is a left-inverse of $f : X \to Y$ and $h : Y \to X$ is a right-inverse of f. Then $g = h$, and this common function is an inverse of f.

Proof. Suppose f, g, h are as stated. Then

$$g = g \circ \mathrm{Id}_Y = g \circ (f \circ h) = (g \circ f) \circ h = \mathrm{Id}_X \circ h = h.$$

So $g = h$ is both a left-inverse and a right-inverse of f, making it an inverse of f. ■

Problem 1.30. Show that if an inverse of a function exists, it is unique. In other words, if g and h are inverses of f then $g = h$. This allows us to say "the" inverse instead of "an" inverse. Moreover, find a counterexample to the assertion that when a left-inverse exists for a function it is unique, and similarly for right-inverses. As a hint, use small finite sets for the counterexamples.

Theorem 1.31. Suppose $f : X \to Y$ and $g : Y \to X$ are functions. Then:

1. g is a left-inverse of f if and only if f is a right-inverse of g.

2. f is invertible with g being its inverse if and only if g is invertible with f being its inverse. This allows us to say that f and g are "inverses of each other," instead of specifying which one is the inverse of the other.

3. If f has an inverse, then its inverse is invertible, and the inverse of the inverse of f is f.

Proof. We will prove these one after another as each will help to prove the next one:

1. By the definition of left and right-inverses, the proposition $g \circ f = \mathrm{Id}_X$ is simultaneously equivalent to saying that g is a left-inverse of f and that f is a right-inverse of g.

2. By the definition of an inverse, f is invertible with g being its inverse if and only if g is a left-inverse and a right-inverse of f. By the previous part, this is equivalent to f being a right-inverse and a left-inverse of g, which is equivalent to g being invertible with f being its inverse.

3. Suppose f has an inverse f^{-1}. By the previous part, this implies that f^{-1} is invertible with f being its inverse. Thus, $(f^{-1})^{-1} = f$.

■

Theorem 1.32. Suppose $f : X \to Y$ and $g : Y \to Z$ are invertible functions. Then $g \circ f$ is invertible and

$$(g \circ f)^{-1} = f^{-1} \circ g^{-1}.$$

More generally, induction (see Section 1.3) can be used to prove that

$$(f_n \circ \cdots \circ f_2 \circ f_1)^{-1} = f_1^{-1} \circ f_2^{-1} \circ \cdots \circ f_n^{-1}.$$

Proof. Let the inverse of f be $f^{-1} : Y \to X$ and let the inverse of g be $g^{-1} : Z \to Y$. Then $f^{-1} \circ g^{-1}$ is a function that maps Z to X. Then

$$(f^{-1} \circ g^{-1}) \circ (g \circ f) = f^{-1} \circ (g^{-1} \circ g) \circ f = f^{-1} \circ \mathrm{Id}_Y \circ f = f^{-1} \circ f = \mathrm{Id}_X,$$
$$(g \circ f) \circ (f^{-1} \circ g^{-1}) = g \circ (f \circ f^{-1}) \circ g^{-1} = g \circ \mathrm{Id}_Y \circ g^{-1} = g \circ g^{-1} = \mathrm{Id}_Z.$$

This proves that $g \circ f$ is invertible with inverse $f^{-1} \circ g^{-1}$. ■

Definition 1.33. To find convenient criteria that allow us to establish the existence of a left-inverse, right-inverse or inverse, we will define the following conditions for a function $f : X \to Y$:

1. f is **injective** or **one-to-one** if for all x_1 and x_2 in X,

   $$f(x_1) = f(x_2) \implies x_1 = x_2.$$

   This means that for each output in the range of f, there is exactly one input in the domain of f that leads to that output. Alternatively formulated, each codomain element has *at most* one domain element mapping to it. By contrapositive, we can also say that different domain elements map to different codomain elements.

2. f is **surjective** or **onto** if for all $y \in Y$, there exists *at least* one $x \in X$ such that $f(x) = y$. In other words, the range of f is the specified codomain of f.

3. f is **bijective** if it is both injective and surjective.

Example. Though we will not prove it, $f(x) = x^n$ is bijective from $\mathbb{R}$ to $\mathbb{R}$ for any fixed odd integer n. On the other hand, bijectivity of x^n does not hold if n is an even integer, but we regain bijectivity if we restrict the domain and codomain to $\mathbb{R}_{\geq 0}$.

Definition 1.34. Let $f : X \to Y$ be a function such that $Y \subseteq X$, so that it is possible to compose f with itself. If there exists a positive integer n such that $f^n(x) = x$ for all $x \in X$, then we say that f is **cyclic**. In this case, the smallest positive integer m satisfying $f^m(x) = x$ for all x is called the **order** of f. If the order of a cyclic function f is 2, then f is called an **involution**.

Problem 1.35. Prove that every cyclic function is injective.

Problem 1.36. Show that the function $f : \mathbb{R}\backslash\{0, 1\} \to \mathbb{R}\backslash\{0, 1\}$, defined by $f(x) = \dfrac{x - 1}{x}$ is a cyclic function of order 3.

Definition 1.37. If, in set builder notation, we wish to carve out the subset of the image of a set X under a function f such that we take only those elements $x \in X$ that satisfy a predicate $p(x)$, then we get the set

$$\{f(x) : x \in X \wedge p(x)\}.$$

If f is injective, then we may instead use the notation

$$\{f(x) \in f(X) : p(x)\}.$$

We have imposed injectivity here because, as a result of there being a unique input x corresponding to the output $f(x)$, we can test $p(x)$ on the unique x. Otherwise, if there are multiple x's corresponding to $f(x)$, then $p(x)$ might be true for some and false for others, which makes us question whether $f(x)$ should be in the set.

We will need the following definition to complete our discussion of inverses.

Definition 1.38. Let $f : X \to Y$ be a function. For each $y \in Y$, the **preimage** of y under f is denoted and defined as

$$f^{-1}(y) = \{x \in X : f(x) = y\}.$$

Moreover, for any subset Z of Y, the **preimage** of Z under f is denoted by and defined as

$$f^{-1}(Z) = \bigcup_{z \in Z} f^{-1}(z) = \bigcup \left\{ f^{-1}(z) : z \in Z \right\}.$$

Preimages are not mere abstractions for the sake of abstraction. For example, they are of fundamental importance in combinatorics, as we will see in Volume 2. In the theory of real functions, they are related to root-finding, as the following definition shows.

Definition 1.39. If $f : X \to Y$ is a function where Y is a subset of the real (or complex) numbers, then the preimage $f^{-1}(0)$ is called the **zero set** of f and its elements are called the **roots** or **zeros** of f. The elements of $f^{-1}(0)$ are also called the **solutions** of the equation $f(x) = 0$.

The determination of the roots of functions is a fundamental problem in algebra. Part of what makes root-finding an even more frequent activity than one might imagine is that, for any fixed $y \in Y$, the determination of the preimage $f^{-1}(y)$ is the same as finding the roots of $f(x) - y$ or, equivalently, the solutions of $f(x) = y$. We will develop methods of determining the roots of quadratic functions in Theorem 9.5 and Theorem 9.11 and, more generally, for polynomials in Section 10.1.

Definition 1.40. A **choice function** f with domain X, where each element of X is a non-empty set, is a function such that

$$\forall S \in X, f(S) \in S.$$

So a choice function maps each set $S \in X$ to an element of S. The **axiom of choice** states that every set of non-empty sets has a choice function defined on X.

In essence, the axiom of choice says that, given a collection of non-empty boxes, we have the power of choose exactly one item from each box. This makes sense for a finite number of boxes, but the result becomes more controversial as the number of boxes climb to infinity and beyond. There are pathological consequences such as the Banach-Tarski paradox and the well-ordering of the real continuum.

Theorem 1.41. Let $f : X \to Y$ be a function. Then we can determine when f has what kind of inverse according to whether it is injective, surjective or bijective.

1. f is injective if and only if f has a left-inverse.
2. f is surjective if and only if f has a right-inverse.
3. f is bijective if and only if f has an inverse.

An implication of the last property is that the inverse of a bijective function is also bijective.

Proof. We will need the axiom of choice from set theory for the second part of the proof.

1. Suppose f is injective. Define $g : Y \to X$ so that if $y \in f(X)$ then $g(y) = x$ is the unique element of $f^{-1}(y)$ where the uniqueness is an implication of injectivity, and if $y \notin f(X)$ then $g(y) = z$ for some constant $z \in X$ since we are assuming that X is non-empty. Then $g \circ f = \mathrm{Id}_X$ and so g is a left-inverse of f.

   In the other direction, suppose $g : Y \to X$ is a left-inverse of f. If $f(x_1) = f(x_2)$ for some $x_1, x_2 \in X$ then applying g to both sides yields

   $$x_1 = g(f(x_1)) = g(f(x_2)) = x_2,$$

   which proves injectivity.

2. Suppose f is surjective. Then for each $y \in Y$, the preimage $f^{-1}(y)$ is nonempty. Using the axiom of choice, for each $y \in Y$, we can arbitrarily choose $x \in f^{-1}(y)$ and define $g : Y \to X$ by $g(y) = x$. Then $f \circ g = \mathrm{Id}_Y$ and so g is a right-inverse of f.

   In the other direction, suppose $g : Y \to X$ is a right-inverse of f. To show that f is surjective, we need to show that for each $y \in Y$, there exists an $x \in X$ such that $f(x) = y$. Given $y \in Y$, we let $x = g(y)$. Indeed, applying f to both sides yields

   $$f(x) = f(g(y)) = y.$$

3. An inverse of f exists if and only if it has both a left-inverse and a right-inverse. By the previous two arguments, this is true if and only if f is both injective and surjective, which is equivalent to f being bijective. However, if we wish to bypass the controversial axiom of choice, there exists a proof of this assertion that does not rely on it.

   In one direction, suppose f is bijective. Injectivity is equivalent to the proposition that, for each $y \in Y$, the preimage $f^{-1}(y)$ has at most one element. Surjectivity is equivalent to the proposition that, for each $y \in Y$, the preimage $f^{-1}(y)$ has at least one element. So bijectivity is equivalent to saying that the preimage $f^{-1}(y)$ of each $y \in Y$

is a singleton. So we construct $g : Y \to X$ by mapping each $y \in Y$ to the unique x such that $f(x) = y$. As desired, it holds that $g \circ f = \mathrm{Id}_X$ and $f \circ g = \mathrm{Id}_Y$.

In the other direction, suppose f has an inverse g. Then g is a left-inverse and a right-inverse of f. Then we can simply repeat the second direction of each of the previous arguments to show that the existence of a left-inverse implies injectivity and that the existence of a right-inverse implies surjectivity. Neither argument uses the axiom of choice. Thus, f is bijective.

Suppose f is a bijective function whose inverse is g. By Theorem 1.31, f is invertible with g being its inverse if and only if g is invertible with f being its inverse. Thus, by the last property in the above list, g is also bijective. ■

Example 1.42. Describe a method of finding the inverse of a bijective function $f : X \to Y$ where X and Y are subsets of $\mathbb{R}$, and where we have an algebraic expression for the definition of f. Apply this method to finding the inverse of the function $f : \mathbb{R} \to \mathbb{R}$ that is defined as $f(x) = (2x + 1)^3$ for all real x.

Solution. Assume that $f : X \to Y$ has an inverse $g : Y \to X$. By the definition of an inverse, $g(f(x)) = x$ for all $x \in X$ and $f(g(y)) = y$ for all $y \in Y$. The former is not too useful in practice. However, if we have an algebraic expression for f, then $f(g(y)) = y$ can be used to get an equation in terms of $g(y)$ that we can solve for $g(y)$. It then remains to be shown that this function g is both a left-inverse and a right-inverse, which can be done by checking that g indeed maps each element of X to some element of Y, and algebraically verifying that $g \circ f = \mathrm{Id}_X$ and $f \circ g = \mathrm{Id}_Y$.
Let us apply this method now. Let $f(x) = (2x + 1)^3$ have domain and codomain $\mathbb{R}$. Assume that f has an inverse $g : \mathbb{R} \to \mathbb{R}$. Then

$$y = f(g(y)) = (2g(y) + 1)^3,$$

which we can solve to get

$$g(y) = \frac{\sqrt[3]{y} - 1}{2}.$$

It can easily be verified that g maps each real number to some real number, and that

$$f \circ g = \mathrm{Id}_{\mathbb{R}} = g \circ f,$$

so g is indeed the inverse of f.
Instead of going through the formal steps that we have stated, an informal method is as follows: We can start with the equation $f(x) = y$ which has y written in terms of x, and manipulate the equation until x is isolated in terms of y. Then the expression in terms of y needs to be shown to be both a left-inverse and a right-inverse of f as in the formal method. Some people also like to switch all x's with y's and vice versa. ■

Problem 1.43. Let X and Y be non-empty sets. Show that there exists an injection $f : X \to Y$ if and only if there exists a surjection $g : Y \to X$. You may use the axiom of choice or results that depend on it.

Problem 1.44. Find functions f, g, h, all with domains and codomains $\mathbb{R}$, such that $g \circ f = h \circ f$ and $g \neq h$. What about satisfying $f \circ g = f \circ h$ and $g \neq h$?

Problem 1.45. Find a function $f : \mathbb{R} \to \mathbb{R}$ such that f is injective and not surjective. Separately, find a function $g : \mathbb{R} \to \mathbb{R}$ that is surjective and not injective.

Problem 1.46. Let $f : A \to B$ and $g : B \to C$ be functions. Prove that:

1. If $g \circ f$ is injective, then f is injective. Also, find an example of $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ such that $g \circ f$ is injective, but g is not injective.

2. If $g \circ f$ is surjective, then g is surjective. Also, find an example of $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ such that $g \circ f$ is surjective, but f is not surjective.

Problem 1.47. Find functions $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ such that $g \circ f$ is bijective and $f \circ g$ is not bijective.

The following is a powerful theorem. Due to the difficulty of the proof, we will skip over providing it. However, we will come back to it in Volume 2 to prove a finite version.

Theorem 1.48 (Schröder-Bernstein theorem)**.** If X and Y are non-empty sets and there exists an injection $f : X \to Y$ and an injection $g : Y \to X$, then there exists a bijection $h : X \to Y$. By Problem 1.43 (the solution to which depends on the axiom of choice), each of the following conditions are also individually sufficient for proving the existence of a bijection $h : X \to Y$:

1. There exists a surjection $f : X \to Y$ and a surjection $g : Y \to X$.

2. There exists a surjection $f : X \to Y$ and an injection $h : X \to Y$.

Problem 1.49. Two sets A and B are said to have the "same cardinality" if there exists a bijection $f : A \to B$. Prove that the following sets, strangely enough, all have the same cardinality:

$$[0, 1], (0, 1), \mathbb{R}, (0, \infty), [0, \infty).$$

As a hint, Theorem 1.48 will be useful for some of the cases. In Volume 3, we will prove that $\mathbb{R}$ is "uncountable" in the sense that a bijection cannot be set up between $\mathbb{R}$ and $\mathbb{Z}_+$, thereby proving that all of the sets here are also uncountable.

Definition 1.50. The **power set** of a set S is the set of all subsets of S, including the empty set $\varnothing$ and S itself. The power set of S is denoted by $\mathcal{P}(S)$. It is an axiom that for each set, its power set exists.

Example. The power set of $\{1, 2, 3\}$ is

$$\{\varnothing, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}.$$

As an edge case, $\mathcal{P}(\varnothing) = \{\varnothing\}$.

Theorem 1.51 (Cantor's theorem)**.** There is no surjective function from a non-empty set to its power set, but for each set there exists an injective function from the set to its power set.

Proof. The proof will be reminiscent of the proof of Russell's paradox (Theorem 1.5). Let X be a non-empty set. Suppose, for contradiction, that there exists a surjection $f : X \to \mathcal{P}(X)$. Then we define the Cantor diagonal set of f to be the set

$$D = \{x \in X : x \notin f(x)\}.$$

Note that D is a subset of X, so it is an element of $\mathcal{P}(X)$. The assumed surjectivity of f implies that

$$\exists y \in X, f(y) = D.$$

Now there are two possibilities: $y \in D$ or $y \notin D$.

- If $y \in D$, then $y \notin f(y) = D$. This contradicts the premise of the case.
- If $y \notin D$, then $\neg(y \notin f(y))$, which means $y \in D$. This again contradicts the premise of this case.

Since we get a contradiction either way, f cannot be surjective. However, $g : X \to \mathcal{P}(X)$, defined by $x \mapsto \{x\}$ is an injection. ■

Corollary 1.52. What Cantor's theorem implies is that we can repeatedly take power sets of some infinite set X to get sets

$$X, \mathcal{P}(X), \mathcal{P}(\mathcal{P}(X)), \mathcal{P}(\mathcal{P}(\mathcal{P}(X))) \dots$$

that are "bigger and bigger infinities" in the sense that there can be no surjection from an earlier set to a later set in this sequence of successive power sets, yet a "copy" of each set exists inside each later set via an injection.

Proof. Let the sequence of power sets be $p_0, p_1, p_2, p_3, \dots$ so that p_k is the set where the power set has been applied k times to the original set X. Suppose, for contradiction, that there is an surjection $f : p_m \to p_n$ for some non-negative integers $m < n$. We know that there are injections $p_m \to p_{m+1}$ and $p_{m+1} \to p_{m+2}$ and so on until $p_{n-2} \to p_{n-1}$, so composing them yields an injection $p_m \to p_{n-1}$. By Problem 1.43, there is a surjection $g : p_{n-1} \to p_m$. By composing $f \circ g$, we get a surjection $p_{n-1} \to p_n$, which contradicts Cantor's theorem. Thus, no lower set can be mapped to a higher set via a surjection, thus creating an infinite strict hierarchy of infinite sets. ■

Corollary 1.53 (Cantor's paradox)**.** As a consequence of Cantor's theorem, we get a new proof that there is no set of all sets.

Proof. Suppose, for contradiction, that there exists a set V of all sets. Then the power set $\mathcal{P}(V)$ exists. The idea is to construct a surjection from V to $\mathcal{P}(V)$, which will contradict Cantor's theorem. Since every element of $\mathcal{P}(V)$ is a set, every element of $\mathcal{P}(V)$ is an element of V. So we can construct a surjection $f : V \to \mathcal{P}(V)$ by mapping the elements of $\mathcal{P}(V)$ (which are all elements of V) from V to $\mathcal{P}(V)$ by the identity function, and mapping any other elements of V to $\mathcal{P}(V)$ arbitrarily. For example, we can define

$$f(x) = \begin{cases} x & \text{if } x \in \mathcal{P}(V) \\ \varnothing & \text{if } x \notin \mathcal{P}(V) \end{cases}.$$

This is a surjection because each element of $\mathcal{P}(V)$ gets hit at least once. ■

Problem 1.54. When I was a student in high school, I used to play with a Rubik's cube at school, while riding on public transit, and everywhere else. On several such occasions, people came up to me and mentioned that they had previously managed to "solve" five of the six faces (meaning for each of those five faces, its nine squares had stickers of the same colour), but that they had been unable to figure out how to solve the sixth face. There is no method of solving the Rubik's cube of which I am aware that proceeds by solving faces. This is because the cube is made up of pieces, including edges that have two stickers that are always adjacent to each other and corners that have three stickers that are always together. Methods for solving the cube are more about placing the edges and corners into the correct slots. Even without this knowledge, how did I conclude that they were lying?

1.3 Induction

There exist several standard methods of proof, including direct proofs, proof by contrapositive, proof by exhaustion or casework, and proof by contradiction. There is one more powerful technique called induction that we can bring into play, thanks to our knowledge of sets.

Definition 1.55. For each positive integer n, the notation $[n]$ refers to the set $\{1, 2, \ldots, n\}$, which is called a **section** of the positive integers $\mathbb{Z}_+ = \{1, 2, 3, \ldots\}$. Similarly, for each non-negative integer n, the notation $[n]^*$ refers to the section $\{0, 1, 2, \ldots, n\}$ of the non-negative integers $\mathbb{Z}_{\geq 0} = \{0, 1, 2, 3, \ldots\}$.

Theorem 1.56. The following three assertions are all true and equivalent to each other.

1. **Well-ordering principle**: Let X be a non-empty subset of the integers $\mathbb{Z}$ such that X has a lower bound, meaning

$$\exists b \in \mathbb{Z}, \forall x \in X, x \geq b.$$

   Then X has a minimal or least element, meaning

$$\exists m \in X, \forall x \in X, x \geq m.$$

   Thanks to the antisymmetry of real inequalities, if there are two minimal elements m_1, m_2, they must be equal (due to $m_1 \geq m_2$ and $m_2 \geq m_1$), so the minimal element is

2. **Principle of mathematical induction**: Let X be a subset of the positive integers $\mathbb{Z}_+$ such that $1 \in X$ and

$$\forall n \in \mathbb{Z}_+, (n \in X \implies n + 1 \in X).$$

   Then $X = \mathbb{Z}_+$. One can think of this as a domino effect.

3. **Principle of complete induction**: Let X be a subset of the positive integers $\mathbb{Z}_+$ such that $1 \in X$ and

$$\forall n \in \mathbb{Z}_+, ([n] \subseteq X \implies n+1 \in X).$$

Then $X = \mathbb{Z}_+$. This is a version of induction where we can assume more in the hypothesis of the second condition, which gives us more power.

The proof of this tri-fold equivalence is a bit too technical for our purposes. It is often given in introductory courses in computer science.

True understanding of the well-ordering principle and the two forms of induction listed will have to await the various examples that will occur in practice. The well-ordering principle is quite useful in number theory, as we will see in the proofs of Euclidean division, Bézout's lemma, and the uniqueness of prime factorization in Volume 3.

Corollary 1.57. Let Y be a subset of $\mathbb{Z}$ such that Y has an upper bound c, meaning

$$\exists c \in \mathbb{Z}, \forall y \in Y, y \leq c.$$

Then Y has a maximum element n, meaning

$$\exists n \in Y, \forall y \in Y, y \leq n.$$

Proof. The idea is to use the well-ordering principle on another set where the upper bound is converted into a lower bound. Suppose Y is a subset of $\mathbb{Z}$ with upper bound c. Then $y \leq c$ for every $y \in Y$. The idea is to use a "reflection" to rewrite this as $1 \leq 1+c-y$. Then the set

$$1 + c - Y = \{1 + c - y : y \in Y\}$$

is a subset of $\mathbb{Z}_+$ and so has a lower bound of 1. By the well-ordering principle, $1 + c - Y$ has a least element x_0. So

$$\forall y \in Y, 1 + c - y \geq x_0,$$

which we can rewrite as $\forall y \in Y, y \leq 1 + c - x_0$. So $1 + c - x_0$ is an upper bound for Y and it suffices to show that $1 + c - x_0 \in Y$. Indeed, since $x_0 \in 1 + c - Y$, there exists a $y_0 \in Y$ such that $x_0 = 1 + c - y_0$, which we can rewrite as $1 + c - x_0 = y_0 \in Y$, as desired. ■

Problem 1.58. A weaker version of the well-ordering principle states that, if X is a nonempty subset of $\mathbb{Z}_+$, then X has a least element. Show that this version is not really weaker by proving that it implies the version stated in Theorem 1.56.

The well-ordering principle is a highly useful method by itself that does not usually need any modification. We will, however, combine the well-ordering principle with proof by contradiction when we study proof by infinite descent in Diophantine analysis in Volume 3.

Theorem 1.59 (Proof by induction)**.** Let $P(n)$ be a predicate in the variable n which ranges over the positive integers. Then:

1. If $P(1)$ is true and, for every positive integer n, the implication $P(n) \implies P(n+1)$ is true, then $P(n)$ is true for all positive integers n. This is called (ordinary) **induction**.

2. If $P(1)$ is true and, for every positive integer n, the implication

$$(P(1) \wedge P(2) \wedge \cdots \wedge P(n)) \implies P(n+1)$$

is true, then $P(n)$ is true for all positive integers. This is called **strong** or **complete induction**.

Here, $P(1)$ is called the **base case** and either of the two implications is called the **inductive step**, with the premise of the inductive step being called the **induction hypothesis**.

Proof. Using set builder notation, let

$$S = \{n \in \mathbb{Z}_+ : P(n) \text{ is true}\}.$$

1. Suppose $P(1)$ is true and that $P(n) \implies P(n+1)$ is true for every positive integer n. This means $1 \in S$ and

$$\forall n \in \mathbb{Z}_+, (n \in S \implies n+1 \in S).$$

By the principle of mathematical induction, $S = \mathbb{Z}_+$, so $\forall n \in \mathbb{Z}_+$, $P(n)$ is true.

2. Suppose $P(1)$ is true and that

$$(P(1) \wedge P(2) \wedge \cdots \wedge P(n)) \implies P(n+1)$$

is true for every positive integer n. This means $1 \in S$ and

$$\forall n \in \mathbb{Z}_+, ([n] \subseteq S \implies n+1 \in S).$$

By the principle of complete induction, $S = \mathbb{Z}_+$ again.

■

There is another form of induction called structural induction which applies to recursively defined sets, but we will not touch it.

Example 1.60. For each positive integer n, it holds that

$$1 + 2 + \cdots + n = \frac{n(n+1)}{2}.$$

Solution. We will use induction. The base case is true because

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

Suppose the sum holds for some positive integer n. So we are assuming that

$$1 + 2 + \cdots + n = \frac{n(n+1)}{2}.$$

Adding $n+1$ to both sides yields

$$\begin{aligned} 1+2+\cdots+n+(n+1) &= \frac{n(n+1)}{2}+(n+1) \\ &= \frac{n(n+1)+2(n+1)}{2} \\ &= \frac{(n+1)(n+2)}{2}, \end{aligned}$$

which finishes the inductive step and so completes the proof by induction. This solution was rather mechanical. We will provide a more illuminating proof when we study arithmetic series in Theorem 6.3. ■

Here are some tips with regards to proof by induction:

1. Mathematical induction is different from scientific induction. The former is a rigorous tool pertaining to the mathematical universe, whereas the latter is an tool for extrapolating evidence into general assumptions about the physical world.

2. It might be necessary to manually verify several base cases before an inductive step can be applied conveniently. That is, verify $P(1), P(2), \ldots, P(m)$ manually for some positive integer m, and then show that $P(n) \implies P(n+1)$ for all integers $n \geq m$. In fact a strong induction argument might require that the previous m cases are known to hold, such as in the theory of linear recursions which we will see when we study combinatorics in Volume 2.

3. Instead of $\mathbb{Z}_+$, it is fine to iterate over a different "countable" set like the even integers because the definition of countable is that it is a set that is in bijection with $\mathbb{Z}_+$.

4. Regardless of which variant of induction is used, what is of utmost importance is that the argument actually induces a domino effect. For example, always manually verify that that induction step works when $P(1)$ is the induction hypothesis. That is, we need to be able to go from the base case to the first non-base case. This is the most precarious step and the source of fallacies.

5. On the whole, think of induction as allowing us to know the truth of some statements that should have taken an infinite lifetime to verify, even though we live finite lives. This is the power of mathematics. The idea is similar to the universal generalization rule of inference for quantifiers, where we fix an arbitrary element and prove something about it, and all of a sudden the property holds for all such elements.

Problem 1.61. Prove that, for each positive integer n,

$$\frac{1}{1^2}+\frac{1}{2^2}+\frac{1}{3^2}+\cdots+\frac{1}{n^2} \leq 2-\frac{1}{n}.$$

This proves that the infinite sum

$$\frac{1}{1^2}+\frac{1}{2^2}+\frac{1}{3^2}+\cdots$$

evaluates to some number strictly less than 2.

Problem 1.62. Prove that, for every positive integer n, the following double inequality holds:

$$\sqrt{n} \leq \frac{1}{\sqrt{1}} + \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{3}} + \cdots + \frac{1}{\sqrt{n}} \leq 2\sqrt{n}.$$

Problem 1.63. For positive integers n, find a closed formula for the sum

$$\frac{1}{2!} + \frac{2}{3!} + \frac{3}{4!} + \cdots + \frac{n}{(n+1)!}.$$

As a hint, evaluate the expression for $n = 1, 2, 3$ and find a pattern which can be proven by induction. Here, $n!$ refers to the product $1 \cdot 2 \cdot 3 \cdots (n-1) \cdot n$.

Chapter 2

Arithmetic

"God made the integers, all else is the work of man."

– *Leopold Kronecker*

"We are justified in calling numbers a free creation of the human mind."

– *Richard Dedekind, Essays on the Theory of Numbers*

The study of algebra begins with the study of numbers and their associated operations. We will study abstract operations first, and then move into the standard addition, multiplication, subtraction and division of real numbers. Afterwards, we will complete our study of arithmetic by looking at exponents, radicals and logarithms.

2.1 Operations

We use binary operations on a daily basis, but without thinking of them as functions. Binary operations are about putting together two objects of the same type to produce a third object of the same type. Let us study operations abstractly as functions.

Definition 2.1. Given a set S, a **binary operation** on S is a function $f : S \times S \to S$. Instead of denoting it as $f(x, y)$, a symbol like $\star$ or $\circ$ (unrelated to function composition) denoting f is placed in between the two ordered inputs like $x \star y$. This is called **infix notation**, as opposed to the usual **prefix notation** of functions.

Example. Addition and multiplication are examples of binary operations on $\mathbb{R}$. Of course, we denote addition as $a + b$ instead of $+(a, b)$ and multiplication as $a \times b$ instead of $\times(a, b)$.

Theorem 2.2. If $a = b$ is an equation, then we may apply a function f that has a, b in its domain to get $f(a) = f(b)$. For example, this works with binary operations where one entry is fixed. That is, if $\star$ is a binary operation on a set S and $a, b, c \in S$, then

$$a = b \implies c \star a = c \star b,$$
$$a = b \implies a \star c = b \star c.$$

However, "doing the same thing to both sides" in a way that what we are applying is not a function does not necessarily work; for example, taking the numerators of the true equation $\frac{1}{2} = \frac{2}{4}$ yields the false equation $1 = 2$.

Proof. The result does not need much justification. Simply note that the same function being applied to equal inputs (so the inputs are the same objects) yields the same output for both by the functional property of a function (see Definition 1.23). In the special case, $\star(\cdot,\cdot)$ is a function from $S \times S$ to S, so fixing the first or second entry produces a function from S to S. ■

Corollary 2.3. Let $\star$ be a binary operation on a set S and $a, b, c, d \in S$. If $a = b$ and $c = d$, then

$$a \star c = b \star d.$$

This is a very useful result in arithmetic because it allows us to "smash" equations together.

Proof. This is proven by the sequence

$$\begin{cases} a = b & \implies a \star c = b \star c \\ c = d & \implies b \star c = b \star d \end{cases} \implies a \star c = b \star d.$$

We have used transitivity of equality here.

■

Binary operations might have one or more of the following convenient properties.

Definition 2.4. Let $\circ$ and $\star$ be binary operations on a set S.

- It is said that $\circ$ is **commutative** if for all $a, b \in S$,

$$a \circ b = b \circ a.$$

- It is said that $\circ$ is **associative** if for all $a, b, c \in S$,

$$a \circ (b \circ c) = (a \circ b) \circ c.$$

- It is said that $\star$ is **left-distributive** over $\circ$ if for all $a, b, c \in S$,

$$c \star (a \circ b) = (c \star a) \circ (c \star b)$$

  and $\star$ is **right-distributive** over $\circ$ if for all $a, b, c \in S$,

$$(a \circ b) \star c = (a \star c) \circ (b \star c).$$

  If $\star$ is both left-distributive and right-distributive over $\circ$ then we say that $\star$ is **distributive** over $\circ$.

Example. Addition and multiplication of real numbers are both commutative and associative, and multiplication is distributive over addition. Subtraction and division are neither commutative nor associative. Those who are familiar with matrix multiplication will recognize that it is associative but not commutative. An example of a commutative but not associative operation is the midpoint operation $(a, b) \mapsto \dfrac{a+b}{2}$ on real numbers.

Problem 2.5. Write the definition of commutativity, associativity, and each kind of distributivity in function notation, like $\circ(a, b)$. This exercise should give the reader an appreciation for why the usual notation is much easier for the human mind to grasp in practice.

Theorem 2.6. Let $\star$ and $\circ$ be binary operations on S. If $\star$ is commutative, then $\star$ is left-distributive over $\circ$ if and only if $\star$ is right-distributive over $\circ$.

Proof. Suppose $\star$ is commutative. Then $\star$ is left-distributive over $\circ$ if and only if for all $a, b, c \in S$,

$$c \star (a \circ b) = (c \star a) \circ (c \star b).$$

By applying commutativity of $\star$ to each of the three applications of $\star$, this is equivalent to

$$(a \circ b) \star c = (a \star c) \circ (b \star c),$$

which is the definition of right-distributivity. We can go in the other direction as well. ■

Definition 2.7. Let $\circ$ be a binary operation on S. An element $e \in S$ is an **identity** for $\circ$ if for all $a \in S$,

$$e \circ a = a \circ e = a.$$

Example. For the addition of real numbers, 0 is an identity. For the multiplication of real numbers, 1 is an identity. An arbitrary binary operation does not necessarily have an identity.

Theorem 2.8. Let $\circ$ be a binary operation on S. If $\circ$ has an identity, then the identity is unique.

Proof. Suppose $\circ$ has two identities $e_1, e_2 \in S$. Since e_1 is an identity and e_2 is an element of S, $e_1 \circ e_2 = e_2$. Since e_2 is an identity and e_1 is an element of S, $e_1 \circ e_2 = e_1$. Thus,

$$e_1 = e_1 \circ e_2 = e_2,$$

meaning e_1 and e_2 are the same element. ■

Definition 2.9. Let $\circ$ be a binary operation on S with an identity e. Then b is said to be a **left-inverse** of a if $b \circ a = e$, and b is said to be a **right-inverse** of a if $a \circ b = e$. If b is both a left-inverse and a right-inverse of a, then b is said to be an **inverse** of a.

Example. For the addition of real numbers, $-a$ is an inverse of a. For the multiplication of non-zero real numbers, $a^{-1} = \dfrac{1}{a}$ is a multiplicative inverse of a. For an arbitrary binary operation that has an identity, an arbitrary element of the underlying set does not necessarily have an inverse.

Theorem 2.10. Let $\circ$ be an associative binary operation on S with an identity. If an element a has an inverse, then the inverse of a is unique.

Proof. Let the identity of $\circ$ be e. Suppose a has two inverses b and c. Then

$$b = b \circ e = b \circ (a \circ c) = (b \circ a) \circ c = e \circ c = c,$$

so b and c are the same element. This proof is easier to comprehend if one begins with the central equality and then moves outward on the left and right separately. ■

Definition 2.11. If $\star$ is a binary operation on a set S and R is a subset of S, then R is said to be **closed** under $\star$ if, for all $a, b, \in R$, it holds that $a \star b \in R$.

There are three abstract algebraic structures that will occasionally or rarely crop up in our studies. We have thought it to be prudent to define them early on for the reader's convenience. A full course in university-level abstract algebra would be necessary to truly appreciate the definitions.

Definition 2.12. A **group** is an ordered pair $(G, \star)$ of a set G and a binary operation $\star$ on G such that $\star$ is associative, there is an identity element, and every element has an inverse. If $\star$ is also commutative, then the group is called **abelian**. Often, G itself is referred to as the group even though it is just the set component of the ordered pair that is the group.

We will use groups to study counting under symmetry in Volume 2.

Definition 2.13. A **ring** is an ordered triple $(R, \circ, \star)$ of a set R and binary operations $\circ$ and $\star$ such that $(R, \circ)$ is an abelian group, $\star$ is associative, and $\star$ is distributive over $\circ$. If $\star$ is also commutative, then the ring is called **commutative**. It is not necessary for $\star$ to have an identity the way $\circ$ does, but if $\star$ is known to have an identity, then it is good to mention it.

Definition 2.14. A **field** is a commutative ring such that the second operation has an identity and every element that is not the identity of the first operation has an inverse under the second operation. Alternatively formulated, a field is an ordered triple $(F, \circ, \star)$ of a set F and binary operations $\circ$ and $\star$ such that $(F, \circ)$ is an abelian group, and if x is the identity for $\circ$, then $(F \backslash \{x\}, \star)$ is also an abelian group, and $\star$ is distributive over $\circ$.

With the abstractions out of the way, let us focus on real numbers and their special subsets, the integers and the rationals.

Definition 2.15. We will not formally construct the **real numbers** out of sets, but indeed they can be boiled down to sets. The reader should already be familiar with how real numbers behave since they fulfil our natural sense of quantification. There is one semi-mysterious property of the real numbers called Dedekind completeness (see Definition 4.19), but it does not matter much until one begins to learn calculus. The basic operations of real number arithmetic are **addition** and **multiplication**, which then lead to **subtraction** and **division**. The set of real numbers is denoted by $\mathbb{R}$, positive real numbers by $\mathbb{R}_+$, and non-negative real numbers by $\mathbb{R}_{\geq 0}$.

Students ordinarily take an analysis course at some point where they are told that real numbers can be defined as equivalence classes of Cauchy sequences of rationals. As this definition is beyond the scope of our writing, we will instead focus on learning about the rules that real numbers satisfy.

Theorem 2.16. For real numbers a, b, c :

- $a = b$ implies $a + c = b + c$ and $c + a = c + b$
- $a = b$ implies $ac = bc$ and $ca = cb$

The two rules above allow us to deduce two more rules for real numbers a, b, c, d :

- $a = b$ and $c = d$ together imply $a + c = b + d$
- $a = b$ and $c = d$ together imply $ac = bd$

Four rules analogous to the above four rules hold for subtraction and for division by non-zero numbers, which will become evident when we define subtraction in terms of addition (Definition 2.22) and division in terms of multiplication (Definition 2.37).

Proof. This is a direct application of Theorem 2.2. ■

Definition 2.17. Since we have numerous operations, we have to set down rules that govern how we interpret expressions that include operations. The mnemonic to remember for the order in which operations are interpreted is **PEMDAS**, which stands for: Parentheses, Exponents, Multiplication, Division, Addition, Subtraction. Some people use **BEDMAS** where the B stands for Brackets.

1. Firstly, if there are parentheses, we evaluate what is inside them first. If there are nested parentheses, meaning sets of parentheses within sets of parentheses, we evaluate them from the innermost parentheses outward. For example, in $(1 + 2) \cdot 3$, the $1 + 2$ is evaluated first.
2. We will describe exponents and how to work with them in Theorem 2.50.
3. Then come multiplication and division. For example, in $1 + 2 \cdot 3$, the $2 \cdot 3$ is evaluated first. We will see that division boils down to multiplication, so they are of equal rank.
4. Of lowest rank are addition and subtraction. We will see that subtraction boils down to addition, so they are of equal rank.

Example. There are some "puzzles" that have garnered attention in recent times such as the following: compute $3 \div 3 \cdot 3$. There are two possible interpretations:

$$(3 \div 3) \cdot 3 = 1 \cdot 3 = 3,$$
$$3 \div (3 \cdot 3) = 3 \div 9 = \frac{1}{3}.$$

An issue with such problems is the ambiguity stemming from the usage of the division symbol instead of expressing the division as a fraction. Secondly, in PEMDAS and BEDMAS, it is not clarified whether multiplication or division has priority, nor it is necessary. Simply use parentheses to avoid confusion if needed. The burden of clarity is on the writer.

Theorem 2.18. Let a, b, c be any real numbers. Then the following rules hold:

1. Addition is commutative: $a + b = b + a$
2. Addition is associative: $a + (b + c) = (a + b) + c$
3. Multiplication is commutative: $ab = ba$

4. Multiplication is associative: $a(bc) = (ab)c$
5. Zero is an additive identity: $a + 0 = 0 + a = a$
6. One is a multiplicative identity: $1 \cdot a = a \cdot 1 = a$
7. Multiplication distributes over addition: $a(b + c) = ab + ac$

Depending on the context, the word **trivial** might refer to the 0 or 1 if we are talking about a trivial number.

We will take these properties for granted, as a subset of the axioms for the real numbers. However, they should be intuitively clear by interpreting $a + b$ as starting from a on the number line and moving away by b units, and $a \cdot b$ as meaning "a copies of b" in the case of positive integers.

Definition 2.19. When we use the distributive property to go from $a(b + c)$ to $ab + ac$, it is called **expanding**. The reverse process is called **factoring**.

Theorem 2.20. For every real number a, there is a unique real number $-a$, such that $a + (-a) = 0$. We call $-a$ the **additive inverse** or **negation** of a.

It should be intuitively clear that $-a$ is the mirror image of a across 0 on the number line.

Definition 2.21. The essence of the **integers** is that repeatedly adding 1 (which represents singleness) to itself produces all of the positive integers, that there is a 0 integer that represents "nothing," and the negative integers are the additive inverses of the positive integers. The set of integers is denoted by $\mathbb{Z}$, the **positive integers** by $\mathbb{Z}_+$, and the set of **non-negative integers** by $\mathbb{Z}_{\geq 0}$, where non-negative means the positives along with zero. The integers form a group under addition, and a commutative ring under addition and multiplication together.

Definition 2.22. We define **subtraction** by b as the addition of the negation of b. In other words, $a - b$ is defined as $a + (-b)$. Note that subtraction is not associative because

$$(a - b) - c \neq a - (b - c)$$

unless $c = 0$. Subtraction is also not commutative. To be clear, we interpret the expression $a - b - c$ as applying operations from left to right, so that it is $a + (-b) + (-c)$.

Lemma 2.23 (Additive cancellation law)**.** If a, b, c are real numbers such that

$$a + c = b + c,$$

then $a = b$.

Proof. We can add $-c = -c$ to $a + c = b + c$ to get $a = b$. This is allowed because we established in Theorem 2.16 that equations may be added to each other. ■

Theorem 2.24. For any real number a, it holds that $a \cdot 0 = 0$.

Proof. Since 0 is the additive identity, $0 + 0 = 0$. Multiplying both sides by a and using the distributive law yields

$$a \cdot 0 = a \cdot (0 + 0) = a \cdot 0 + a \cdot 0.$$

Now we can subtract $a \cdot 0 = a \cdot 0$ from both sides to get $a \cdot 0 = 0$, as desired. ■

Theorem 2.25. For any real number a, it holds that $-a = (-1) \cdot a$. So negating is equivalent to multiplying by the negation of the multiplicative identity.

Proof. By Theorem 2.20, $-a$ is defined as the number that satisfies the equation $a+(-a) = 0$. So -1 is the number that satisfies the equation $1 + (-1) = 0$. Multiplying both sides by a and using the distributive law yields

$$a + (-1) \cdot a = 0.$$

Equating this with $a + (-a) = 0$ yields

$$a + (-a) = a + (-1) \cdot a$$

and subtracting a from both sides gives $-a = (-1) \cdot a$. ■

Corollary 2.26. For any real numbers a and b, it holds that $(-a)b = -(ab)$. Consequently, $a(-b)$ represents the same number. So the expression $-ab$ is well-defined as it can refer to any of $(-a)b$ or $-(ab)$ or $a(-b)$.

Proof. Since $-c = (-1) \cdot c$ holds for any real number c, we find by associativity that

$$(-a)b = ((-1)a)b = (-1)(ab) = -(ab).$$

For the consequence,

$$(-a)b = -(ab) = -(ba) = (-b)a = a(-b).$$

■

Corollary 2.27. For any real numbers a and b, it holds that $(-a)(-b) = ab$. As the well-known saying goes, "double negatives makes a positive."

Proof. Since $a + (-a) = 0$ and $b + (-b) = 0$, multiplying the equations yields the following sequence of deductions:

$$\begin{aligned}
(a + (-a))(b + (-b)) &= 0 \cdot 0 \\
ab + a(-b) + (-a)b + (-a)(-b) &= 0 \\
ab + (-ab) + (-ab) + (-a)(-b) &= 0 \\
-ab + (-a)(-b) &= 0 \\
(-a)(-b) &= ab.
\end{aligned}$$

In the last step, we added ab to both sides. ■

Problem 2.28. For a real number a, what is the negative of the negative of a? Rigorously prove your answer.

Definition 2.29. The **absolute value** of a real number is its distance from 0 on the real number line. More precisely, we define it as

$$|x| = \begin{cases} x & \text{if } x \text{ is non-negative} \\ -x & \text{if } x \text{ is negative} \end{cases}.$$

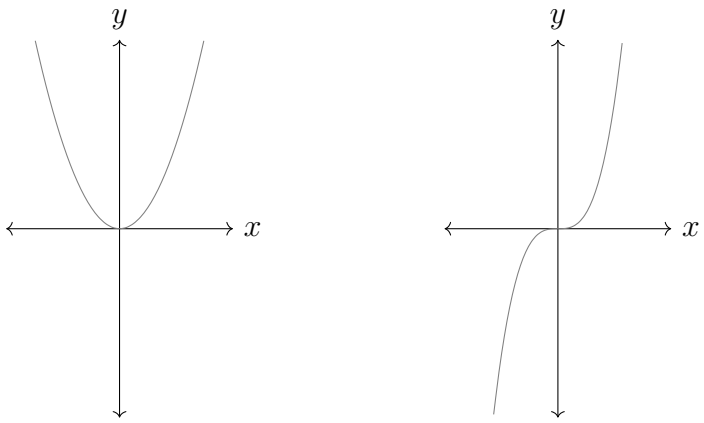
Note that, if x is negative, then tacking on an extra negative sign as shown makes the output positive, by the solution to Problem 2.28. Absolute value is a very useful function that we will study in detail in Section 5.1.

Definition 2.30. There are two basic kinds of symmetry for functions f whose domain and range are subsets of $\mathbb{R}$ such that $x \in \text{Dom}(f) \iff -x \in \text{Dom}(f)$:

- The function f is **even** if $f(x) = f(-x)$ for all $x \in \text{Dom}(f)$.
- The function f is **odd** if $f(x) = -f(-x)$ for all $x \in \text{Dom}(f)$.

It becomes evident why we refer to these definitions as instances of symmetry if one graphs even or odd functions on the Cartesian plane.

Example. Even and odd functions are defined as such because the function $f(x) = x^n$ with domain $\mathbb{R}$ is even for even integers n and odd for odd integers n. We will explain exponents in Section 2.2 for those looking for a definition. The graphs of x^2 and x^3 are given below on the left and right, respectively.



Even and odd functions have interesting properties with regards to their sums, differences, products, quotients and compositions. These properties are too numerous to exhaustively list here, but the reader is encouraged to look into them elsewhere. We have mentioned a couple of the more interesting properties in Example 2.31 and Problem 2.32.

Example 2.31. Suppose f is a function whose domain and range are subsets of $\mathbb{R}$ such that $x \in \text{Dom}(f) \iff -x \in \text{Dom}(f)$. Show that there exists a unique even function f_e and a unique odd function f_o, both with domains equal to $\text{Dom}(f)$, such that

$$f(x) = f_e(x) + f_o(x)$$

for all $x \in \text{Dom}(f)$.

Solution. We will first prove uniqueness of f_e and f_o. If f_e and f_o exist, then

$$\begin{aligned} f(x) &= f_e(x) + f_o(x) \\ f(-x) &= f_e(-x) + f_o(-x) \\ &= f_e(x) - f_o(x). \end{aligned}$$

Solving this system for f_e and f_o yields

$$\begin{aligned} f_e(x) &= \frac{f(x) + f(-x)}{2}, \\ f_o(x) &= \frac{f(x) - f(-x)}{2} \end{aligned}$$

for all $x \in \mathrm{Dom}(f)$. This proves uniqueness, since f_e and f_o must be two specific functions in terms of f. Moreover, it can be readily verified that f_e is even and f_o is odd, and that $f = f_e + f_o$, which proves existence. ■

Problem 2.32. Find all functions $f : \mathbb{R} \to \mathbb{R}$ that are both even and odd.

Theorem 2.33. For every non-zero real number a, there is a unique real number a^{-1}, such that $a \cdot a^{-1} = 1$. We call a^{-1} the **multiplicative inverse** or **reciprocal** of a.

The intuitive idea behind reciprocals is that a segments of length a^{-1} on the number line come together to make a segment of length 1. In other words, if a segment of 1 were to be divided into a segments, then each small segment would have length a^{-1}.

Definition 2.34. The **rationals** can be produced by taking the multiplicative inverse of each non-zero integer, and then adding each such inverse to itself arbitrarily many finite times. Since repeated addition is simply multiplication, set builder notation gives us

$$\mathbb{Q} = \{ab^{-1} : a \in \mathbb{Z}_{\geq 0}, b \in \mathbb{Z} \backslash \{0\}\}.$$

The rationals are denoted by $\mathbb{Q}$, the **positive rationals** by $\mathbb{Q}_+$, and the **non-negative rationals** by $\mathbb{Q}_{\geq 0}$. The rationals form a field, just like the reals. When we get to complex numbers, we will see that they form a field too in Theorem 8.5.

Lemma 2.35 (Multiplicative cancellation law)**.** If a, b, c are real numbers such that $ac = bc$ and c is non-zero, then $a = b$.

Proof. We can multiply $ac = bc$ by $c^{-1} = c^{-1}$ to get $a = b$. This is allowed because we established in Theorem 2.16 that equations may be multiplied together. ■

Theorem 2.36. If a and b are real numbers such that $ab = 0$ then $a = 0$ or $b = 0$. For this reason, we like to take every term in an equation to one side and factor it so that we have a factored expression equal to 0. This allows us to deduce that at least one of the factors is equal to 0.

Proof. Assuming $ab = 0$, it suffices to show that $b \neq 0$ implies $a = 0$. Since b is non-zero, b has a multiplicative inverse b^{-1}. By multiplying $ab = 0$ by $b^{-1} = b^{-1}$, we get

$$a = (ab)b^{-1} = 0 \cdot b^{-1} = 0.$$

So $b = 0$ or $a = 0$. Here, we have used Theorem 2.24. ■

Definition 2.37. We define **division** by a non-zero real number b as the multiplication by multiplicative inverse of b. In other words, $\frac{a}{b}$ is defined as $a \cdot b^{-1}$. Note that this means

$$a^{-1} = 1 \cdot a^{-1} = \frac{1}{a},$$

which is called the reciprocal of a. The division expression $\frac{a}{b}$ is called a **fraction**. The fraction bar, called a **vinculum**, that separates the **numerator** a and **denominator** b implicitly says that we are putting parentheses around the numerator and around the denominator. For example, $\frac{a+b}{c+d}$ means $\frac{(a+b)}{(c+d)}$. Unlike subtraction, an expression such as $a/b/c$ without parentheses has no commonly accepted definition and so it is ill-defined; it could represent $\frac{a}{\left(\frac{b}{c}\right)}$ or $\frac{\left(\frac{a}{b}\right)}{c}$.

Problem 2.38. Compute $\frac{-1}{-1}$.

Lemma 2.39. For any non-zero real numbers a and b, it holds that

$$(ab)^{-1} = a^{-1}b^{-1}.$$

Proof. Indeed, we find that

$$(ab)(a^{-1}b^{-1}) = (ab)(b^{-1}a^{-1}) = a(bb^{-1})a^{-1} = aa^{-1} = 1$$

so the multiplicative inverse of ab is $a^{-1}b^{-1}$. This is reminiscent of Theorem 1.32. ■

Problem 2.40. Similar to negation (see Problem 2.28), applying the multiplicative inverse twice to a non-zero real number returns the original number. This means $(a^{-1})^{-1} = a$ for $a \neq 0$.

Theorem 2.41. Fractions have several properties that are useful in computations and simplifications. For real numbers a, b, c, d such that c and d are non-zero, the following identities hold with any additional restrictions stated:

1. Addition and subtraction: $\frac{a}{c} \pm \frac{b}{c} = \frac{a \pm b}{c}$, where both $\pm$ signs are the same sign
2. Multiplication: $\frac{a}{c} \cdot \frac{b}{d} = \frac{ab}{cd}$
3. Simplification: $\frac{ad}{cd} = \frac{a}{c}$
4. Reciprocal: $\left(\frac{a}{c}\right)^{-1} = \frac{c}{a}$ if $a \neq 0$

5. Division: $\frac{\left(\frac{a}{c}\right)}{\left(\frac{b}{d}\right)} = \frac{ad}{bc}$ if $b \neq 0$. The following special cases are useful in our experience:

$$\frac{\left(\frac{1}{a}\right)}{b} = \frac{1}{ab}, \frac{a}{\left(\frac{1}{b}\right)} = ab, \frac{1}{\left(\frac{a}{b}\right)} = \frac{b}{a}, \frac{1}{\left(\frac{1}{a}\right)} = a, \frac{\left(\frac{a}{b}\right)}{\left(\frac{c}{b}\right)} = \frac{a}{c}, \frac{\left(\frac{a}{b}\right)}{\left(\frac{a}{c}\right)} = \frac{c}{b}.$$

It is not true in general that $\frac{a+b}{c+d} = \frac{a}{c} + \frac{b}{d}$.

Proof. The identities can be deduced from the definitions of inverses and fractions, along with Lemma 2.39, Problem 2.40, and the distributive property. At times, we will also use the earlier identities in this list to prove later identities in this list.

1. $\frac{a}{c} \pm \frac{b}{c} = ac^{-1} \pm bc^{-1} = (a \pm b)c^{-1} = \frac{a \pm b}{c}$
2. $\frac{a}{c} \cdot \frac{b}{d} = (ac^{-1})(bd^{-1}) = (ab)(c^{-1}d^{-1}) = (ab)(cd)^{-1} = \frac{ab}{cd}$
3. $\frac{ad}{cd} = (ad) \cdot (cd)^{-1} = adc^{-1}d^{-1} = ac^{-1}dd^{-1} = ac^{-1} = \frac{a}{c}$
4. $\frac{a}{c} \cdot \frac{c}{a} = \frac{ac}{ca} = \frac{a}{a} = 1$, so $\frac{c}{a}$ is the multiplicative of inverse of $\frac{a}{c}$
5. $\frac{\left(\frac{a}{c}\right)}{\left(\frac{b}{d}\right)} = \frac{a}{c} \cdot \left(\frac{b}{d}\right)^{-1} = \frac{a}{c} \cdot \frac{d}{b} = \frac{ad}{bc}$

By choosing $a = b = c = d = 1$, we find that $\frac{1+1}{1+1} = 1 \neq 2 = \frac{1}{1} + \frac{1}{1}$. ■

Definition 2.42. The act of turning the sum $\frac{a}{b} + \frac{c}{d}$ into the single equivalent fraction $\frac{ad+bc}{bd}$ is called **finding a common denominator** and bd is called the **common denominator** in this case. Any other denominator that allows for turning the sum of multiple fractions into one fraction also fulfils this definition.

Definition 2.43. Given several fractions with integer denominators, the **lowest common denominator** is the lowest common multiple of the integer denominators, and the number may be used as a common denominator to add the fractions together.

Problem 2.44. Prove the following identities for real numbers a and $b \neq 0$:

- $\frac{-a}{-b} = \frac{a}{b}$
- $\frac{-a}{b} = -\left(\frac{a}{b}\right) = \frac{a}{-b}$

As such, in addition to Definition 2.34, we may equivalently define the rationals as

$$\mathbb{Q} = \left\{ \frac{a}{b} : a \in \mathbb{Z}, b \in \mathbb{Z} \backslash \{0\} \right\}.$$

Definition 2.45. If there are one or more fractions in an equation, then the act of multiplying both sides of the equation by the product of the denominators, in order to produce an equation with no fractions, is called **clearing the denominators**. If the denominators are all integers, we can get away with multiplying by only their lowest common multiple.

Theorem 2.46. If a, b, c, d are real numbers such that $\frac{a}{b} = \frac{c}{d}$ (so b and d must be non-zero), then

$$\frac{a}{b} = \frac{c}{d} = \frac{a+c}{b+d}.$$

Moreover, if $b \neq d$, then

$$\frac{a}{b} = \frac{c}{d} = \frac{a-c}{b-d}.$$

Proof. We will work backwards using reversible steps:

$$\begin{aligned} \frac{a}{b} &= \frac{a+c}{b+d} \\ a(b+d) &= b(a+c) \\ ab + ad &= ba + bc \\ ad &= bc \\ \frac{a}{b} &= \frac{c}{d}, \end{aligned}$$

which we have assumed to be true. For the second result, we can apply the first result as follows:

$$\begin{aligned} &\frac{a}{b} = \frac{c}{d} = \frac{-c}{-d} \\ \implies &\frac{a-c}{b-d} = \frac{a+(-c)}{b+(-d)} = \frac{a}{b} = \frac{-c}{-d} = \frac{c}{d}. \end{aligned}$$

■

The reader is expected to be familiar with how to apply arithmetic operations to integers and rationals using standard computational algorithms.

Problem 2.47. Note that $\frac{a}{b} = \frac{c}{d}$ if and only if $ad = bc$, assuming the denominators are never 0. This leads us to defining the **ratio binary relation** on $(\mathbb{R} \times \mathbb{R}) \backslash \{(0,0)\}$, which is defined by

$$(a, b) \sim (c, d) \iff ad = bc.$$

Prove that this is an equivalence relation. The equivalence class with (a, b) in it can be represented by the **ratio** $a : b$.

Problem 2.48. Reviewing Definition 1.34 if needed, show that the following functions are involutions.

1. $f : \mathbb{R} \to \mathbb{R}$ defined by $c - x$ where c is any real constant
2. $g : \mathbb{R} \backslash \{0\} \to \mathbb{R} \backslash \{0\}$ defined by $\frac{c}{x}$ where c is any non-zero real constant

2.2 Exponents, Radicals, and Logarithms

Definition 2.49. The **exponential function** $f_b : D \to \mathbb{R}$ with a fixed real base b is a function that is denoted with input x by $f_b(x) = b^x$. For us, the domain is defined as

$$D = \begin{cases} \mathbb{R} & \text{if } b > 0 \\ \mathbb{R}_+ & \text{if } b = 0 \\ \mathbb{Z} \cup \left\{ \frac{m}{n} : m, n \in \mathbb{Z}, n \text{ is odd}, \gcd(m, n) = 1 \right\} & \text{if } b < 0 \end{cases}.$$

In the expression b^x, the whole expression b^x is called a **power**, b is called the **base**, and x is called the **exponent**. We will not formally define the exponential function, but we have noted numerous properties of powers in Theorem 2.50 that are convenient for computation. Going back to PEMDAS or BEDMAS, the E stands for "exponent." This means that exponents have a higher rank than the other four arithmetic operations.

Example. In $a^b + c$, the a^b is computed first and then c is added in. Another, more important, example is that $-a^n$ is not necessarily equal to $(-a)^n$, but rather it is the negation of a^n. This is because $-a^n$ means $(-1) \cdot a^n$, and multiplication comes after exponentiation.

Theorem 2.50. The exponential function and powers satisfy the following arithmetic properties:

1. The exponential function is bijective onto $\mathbb{R}_+$ when b is positive but not equal to 1. The exponential function is identically 1 for $b = 1$, and identically 0 for $b = 0$.

2. If n is a positive integer and b is real, then

$$b^n = \underbrace{b \cdot b \cdots b}_{n \text{ factors of } b}.$$

3. If $b \neq 0$ is real, then powers with zero exponent can be computed as $b^0 = 1$. If $b = 0$, then $b^0 = 0^0$ is undefined due to the following dichotomy:

$$\begin{aligned} x \neq 0 &\implies 0^x = 0, \\ b \neq 0 &\implies b^0 = 1. \end{aligned}$$

   So it is not clear what 0^0 "should" be. Depending on the context, we might choose a temporary convention as appropriate and convenient.

4. If b is a positive real, then powers with negative exponents are defined in terms of powers with positive exponents as

$$b^{-x} = (b^x)^{-1} = \frac{1}{b^x}.$$

   This is well-defined for positive b because $b^x = 0$ if and only if $b = 0$.

5. Powers with the same positive base b can be multiplied as

$$b^x \cdot b^y = b^{x+y}.$$

As a consequence, if $b \neq 0$, then we can divide powers with the same base as

$$\frac{b^x}{b^y} = b^x \cdot (b^y)^{-1} = b^x \cdot b^{-y} = b^{x-y}.$$

6. If a, b are positive reals, then powers with the same exponent can be multiplied as

$$a^x \cdot b^x = (ab)^x.$$

As a consequence, if $b \neq 0$, then we can divide powers with the same exponent as

$$\frac{a^x}{b^x} = a^x \cdot (b^x)^{-1} = a^x \cdot b^{-x} = a^x \cdot (b^{-1})^x = \left(\frac{a}{b}\right)^x.$$

7. The power of a power with a positive base b can be computed as

$$(b^x)^y = (b^y)^x = b^{xy}.$$

This includes the rational exponent case

$$b^{\frac{p}{q}} = (b^{\frac{1}{q}})^p = (b^p)^{\frac{1}{q}}$$

for integers p and $q \neq 0$. We show how to interpret and compute powers with rational exponents using radicals in Theorem 2.55. Note that there is a difference between $b^{(x^y)}$ and $(b^x)^y$. If no parentheses are used, then it is convention to interpret b^{x^y} as $b^{(x^y)}$.

If b is negative, there are analogues of most of these rules that work for negative bases as long as the exponents are from the stated domain.

Example 2.51. Find a pattern in the sequence of powers of -1:

$$(-1)^1, (-1)^2, (-1)^3, (-1)^4 \ldots.$$

Solution. We can prove by induction on $n \geq 1$ that

$$(-1)^n = \begin{cases} 1 & \text{if } n \text{ is even} \\ -1 & \text{if } n \text{ is odd} \end{cases}.$$

This pattern of alternating between 1 and -1 is frequently used as a tool in summation notation. By multiplying the term of index k by $(-1)^k$, we can indicate that the terms have alternating signs. For example,

$$\sum_{k=1}^{n} \frac{(-1)^k}{k} = -\frac{1}{1} + \frac{1}{2} - \frac{1}{3} + \cdots + \frac{(-1)^n}{n}.$$

Finite and infinite series are defined in Definition 3.10. ■

Example 2.52. Let n be an odd positive integer. Show that the functions $f(x) = x^n$ and $g(x) = x^{\frac{1}{n}}$ are inverse functions of each other. To be clear, these are well-defined functions whose domains are $\mathbb{R}$.

Solution. Letting $m = \dfrac{1}{n}$, it holds that $mn = 1$. So we find that

$$f(g(x)) = (x^m)^n = x^{mn} = x^1 = x,$$
$$g(f(x)) = (x^n)^m = x^{nm} = x^1 = x.$$

Since both compositions yield the identity function, f and g are inverse functions. ■

Theorem 2.53. Here are a couple of tricks of the trade that are worth knowing:

1. Adding b^x to itself n times, where n is a positive integer, yields nb^x. If $b = n$, then we get
$$\underbrace{b^x + b^x + \cdots + b^x}_{n \text{ copies of } b^x} = b \cdot b^x = b^{x+1}.$$
2. By distributivity,
$$b^x \pm b^y = b^y(b^{x-y} \pm 1).$$
This is usually useful for only $x \geq y$.

Definition 2.54. If n is a positive integer, then an **n^{th} root** of a real number b is defined as any real (or complex) number r such that $r^n = b$. Complex numbers provides a complete theory of roots (see Theorem 8.30). Real numbers are messier, but it is thankfully true that every positive real b has a unique *positive* n^{th} root; we denote this root in **radical** notation as $\sqrt[n]{b}$ or, for $n = 2$, just $\sqrt{b}$. There are two more kinds of roots:

- For a positive real number b, the negative of $\sqrt[n]{b}$ satisfies
$$(-\sqrt[n]{b})^n = (-1)^n b.$$
This equals b for even positive integers n. Thus, $-\sqrt[n]{b}$ is called the **negative n^{th} root** of b if b is a positive real number and n is an even positive integer. The uniqueness of the positive root implies the uniqueness of the negative root (since they are negatives of each other), assuming the latter exists.
- For a negative real number b and odd positive integer n, there exists a unique real number r such that $r^n = b$. For example, $(-2)^3 = -8$, so we can write $\sqrt[3]{-8} = -2$.

Theorem 2.55. For a positive real b and a positive integer n,
$$\sqrt[n]{b} = b^{\frac{1}{n}}.$$

If we add a positive integer m to the mix, we get that
$$\sqrt[n]{b^m} = b^{\frac{m}{n}}.$$

We can also bring a negative into the exponent by applying a reciprocal, like
$$\frac{1}{\sqrt[n]{b^m}} = \frac{1}{b^{\frac{m}{n}}} = b^{-\frac{m}{n}}.$$

Proof. Suppose b is a positive real number and n is a positive integer. Let $r = \sqrt[n]{b}$. Then

$$b = r^n \implies b^{\frac{1}{n}} = (r^n)^{\frac{1}{n}} = r^{n \cdot \frac{1}{n}} = r^1 = r = \sqrt[n]{b}.$$

For the second result, we can apply the first result and an exponent rule to get

$$\sqrt[n]{b^m} = (b^m)^{\frac{1}{n}} = b^{\frac{m}{n}}.$$

The third rule is evident from its statement. ■

Theorem 2.56. If a and b are positive real numbers and n is a positive integer, then

$$\sqrt[n]{a} \cdot \sqrt[n]{b} = \sqrt[n]{ab},$$
$$\frac{\sqrt[n]{a}}{\sqrt[n]{b}} = \sqrt[n]{\frac{a}{b}} \text{ for } b > 0.$$

As a corollary, simplifications such as

$$\sqrt[n]{a^n b} = a\sqrt[n]{b}$$

are possible. There are analogous rules for negative a, b and mixed signs among a, b, when n is an odd positive integer.

Proof. Given positive real a, b and a positive integer n,

$$\sqrt[n]{a} \cdot \sqrt[n]{b} = a^{\frac{1}{n}} \cdot b^{\frac{1}{n}} = (ab)^{\frac{1}{n}} = \sqrt[n]{ab},$$
$$\frac{\sqrt[n]{a}}{\sqrt[n]{b}} = \frac{a^{\frac{1}{n}}}{b^{\frac{1}{n}}} = \left(\frac{a}{b}\right)^{\frac{1}{n}} = \sqrt[n]{\frac{a}{b}} \text{ for } b > 0.$$

We can apply the first rule to get that $\sqrt[n]{a^n b} = \sqrt[n]{a^n} \cdot \sqrt[n]{b} = a\sqrt[n]{b}$. In practice, extracting the largest n^{th} power a^n from under the hood of an n^{th} root involves finding a prime factorization when the integer is large. ■

In general, use your head when applying exponent rules. For example, a negative base can be taken to a rational exponent whose denominator is odd, such as

$$(-8)^{\frac{2}{3}} = ((-8)^2)^{\frac{1}{3}} = 64^{\frac{1}{3}} = 4$$
$$= ((-8)^{\frac{1}{3}})^2 = (-2)^2 = 4.$$

However, if the denominator is even, issues like

$$1 = 1^{\frac{1}{2}} = ((-1)(-1))^{\frac{1}{2}} = (-1)^{\frac{1}{2}}(-1)^{\frac{1}{2}}$$

can occur where the right side is not even defined without complex numbers. Even with complex numbers,

$$(-1)^{\frac{1}{2}}(-1)^{\frac{1}{2}} = i^2 = -1,$$

which is a contradiction. See Definition 8.1 for the definition of this mysterious letter i. There is also the fallacy that

$$\sqrt{(-1)^2} = (-1)^{\frac{2}{2}} = (-1)^1 = -1,$$

whereas the fact is that $\sqrt{(-1)^2} = \sqrt{1} = 1$. In conclusion, exponent rules should be not be blindly applied when a base is negative.

Problem 2.57. By **rationalizing the denominator** of a fraction, we refer to a process of finding the same rational number in fraction form but with a rational denominator. For example,

$$\frac{a}{\sqrt{b}} = \frac{a\sqrt{b}}{\sqrt{b}\cdot\sqrt{b}} = \frac{a\sqrt{b}}{b}.$$

Completing such a process can result in more convenient arithmetic manipulations, such as finding common denominators. See if you can rationalize the denominators of

$$\frac{1}{\sqrt{2}+\sqrt{3}} \text{ and } \frac{1}{\sqrt{2}-\sqrt{3}+\sqrt{5}}$$

using the observation that

$$(\sqrt{a}-\sqrt{b})(\sqrt{a}+\sqrt{b}) = a-b.$$

Definition 2.58. Since an exponential function $f(x) = b^x$ is bijective onto $\mathbb{R}_+$ for $0 < b \neq 1$, we can define its inverse $g : \mathbb{R}_+ \to \mathbb{R}$, which is written as $g(y) = \log_b y$. This is called the **logarithm** with base b.

Similar to exponents, and often proven using exponent laws, there exist logarithmic laws.

Theorem 2.59. By the definition of inverses, we get our first two logarithmic identities for $0 < b \neq 1$:

1. For all $x \in \mathbb{R}$, it holds that $\log_b(b^x) = g(f(x)) = x$.
2. For all $y \in \mathbb{R}_+$, it holds that $b^{\log_b y} = f(g(y)) = y$.

These two rules combined tell us that $b^x = y$ if and only if $\log_b y = x$, with one direction of this biconditional statement being proven by each composition. Try showing it. Switching between exponential notation and logarithmic notation is an important technique.

Theorem 2.60. Let $b > 0$ and $c > 0$ be bases not equal to 1, $x > 0$ and $y > 0$ be arguments, and a be any real number. Then the following rules hold:

1. $\log_b 1 = 0$
2. $\log_b b = 1$
3. Product rule: $\log_b(xy) = \log_b x + \log_b y$
4. Quotient rule: $\log_b \dfrac{x}{y} = \log_b x - \log_b y$
5. Power Rule: $\log_b(x^a) = a \log_b x$
6. Change-of-base formula: $\log_b x = \dfrac{\log_c x}{\log_c b}$
7. Reciprocal rule: $\log_b c = \dfrac{1}{\log_c b}$

8. $\log_{(b^a)}(x^a) = \log_b x$
9. $\log_{\frac{1}{b}} x = -\log_b x$

Proof. The proofs largely involve a combination of turning logarithmic equations into exponential equations, letting expressions equal to variables, and using exponent rules. We will also work backwards from the desired logarithmic identities while making sure that our steps are reversible, utilize the fact that exponential functions are bijective, and make repeated use of the identity $b^{\log_b y} = y$ that we derived in Theorem 2.59. We encourage the reader to try proving these identities independently before reading the brief proofs provided.

1. Converting to exponential form, $\log_b 1 = 0$ if and only if $b^0 = 1$, which is true.
2. Converting to exponential form, $\log_b b = 1$ if and only if $b^1 = b$, which is true.
3. Converting to exponential form, $\log_b(xy) = \log_b x + \log_b y$ if and only if
$$xy = b^{\log_b x + \log_b y} = b^{\log_b x} \cdot b^{\log_b y},$$
which is true.
4. Converting to exponential form, $\log_b \dfrac{x}{y} = \log_b x - \log_b y$ if and only if
$$\frac{x}{y} = b^{\log_b x - \log_b y} = \frac{b^{\log_b x}}{b^{\log_b y}},$$
which is true.
5. Converting to exponential form, $\log_b(x^a) = a \log_b x$ if and only if
$$x^a = b^{a \log_b x} = \left(b^{\log_b x}\right)^a,$$
which is true.
6. Clearing the denominator turns it into
$$\log_c b \log_b x = \log_c x.$$
Converting to exponential form yields
$$c^{\log_c b \log_b x} = x.$$
This is true because the left side reduces as
$$c^{\log_c b \log_b x} = \left(c^{\log_c b}\right)^{\log_b x} = b^{\log_b x} = x.$$
7. By the change-of-base formula,
$$\frac{1}{\log_c b} = \frac{\log_c c}{\log_c b} = \log_b c.$$

8. By the change-of-base formula and power rule,

$$\log_{b^a}(x^a) = \frac{\log_b x^a}{\log_b b^a} = \frac{a \log_b x}{a \log_b b} = \frac{a \log_b x}{a} = \log_b x.$$

9. By the change-of-base-formula,

$$\log_{\frac{1}{b}} x = \frac{\log_b x}{\log_b \left(\frac{1}{b}\right)} = \frac{\log_b x}{-1} = -\log_b x.$$

■

Corollary 2.61. The chain rule

$$\log_b x = \frac{\log_c x}{\log_c b}$$

can equivalently be written as

$$(\log_c b)(\log_b x) = \log_c x,$$

which makes it seem like the two b's cancel out. The generalization to a product

$$(\log_a b_1)(\log_{b_1} b_2) \cdots (\log_{b_n} x) = \log_a x$$

is called the **chain rule for logarithms**. As always, all of the bases must be positive (and not 1), and all of the inputs, otherwise known as arguments, must be positive.

Chapter 3

Indexed Objects

> "Fundamental progress has to do with the reinterpretation of basic ideas."
>
> – *Alfred North Whitehead*

> "It is in this gesture of "going beyond," to be something in oneself rather than the pawn of a consensus, the refusal to stay within a rigid circle that others have drawn around one - it is in this solitary act that one finds true creativity. All others things follow as a matter of course."
>
> – *Alexander Grothendieck*

The intuition behind indexing is that if we have some distinguishable index cards, then we can write a (not necessarily unique) symbol on each card. In particular, if the index cards are stacked in an order, then we can speak of the symbol at a particular point or "index" in the order. The indexed objects that we will study are lists or tuples, sequences, series and products, some of which have variants with infinite entries or terms. We will end by studying a powerful method of reinterpreting nested sums and products, which we will call the "discrete Fubini's principle" [4].

3.1 Sequences

Definition 3.1. If I and S are non-empty sets and $f : I \to S$ is a function, then we define the following terms:

- f is an **indexing function** and may be be called an **I-indexed family**
- I is the **indexing set** and its elements are called **indices**
- $\text{Rng}(f)$ is the **indexed set** and its elements are called **entries**

In a context where the function f is treated as a tool for indexing, f_i denotes $f(i)$ and is called the **entry at index** i, and f is denoted by $(f_i)_{i\in I}$. Two indexing functions are said to be **equal** if they are equal as functions, meaning the indexing sets are the same and corresponding indices lead to the same output. There are two common specific types of indexing functions:

- If $I = [n]$ for some positive integer n, then f is called a **list** or **tuple** and may be denoted as $(f_i)_{i=1}^n$ or $(f_1, f_2, \ldots, f_n)$. More generally, the indexing set in a list may be taken to be any non-empty finite set, which is any set that is equipotent (meaning, in bijection) with $[n]$ for some positive integer n.

- If $I = \mathbb{Z}_+$, then f is called a **sequence** and may be denoted as $(f_i)_{i=1}^{\infty}$ or $(f_1, f_2, f_3, \ldots)$. More generally, the indexing set in a sequence may be taken to be any countably infinite set, which is any set that is equipotent with $\mathbb{Z}_+$.

The more general definitions of lists and sequences can be useful, such as with Fibonacci and Catalan numbers because the first index of each is 0 instead of 1; these two sequences will be explored in Volume 2.

We have defined lists and sequences rather formally. The reader would be well-advised to not worry about the formal definition in most cases, and instead simply think of them as finite or countably infinite sets written down in a particular order from left to right.

Definition 3.2. Though it is not a requirement for an indexing function, the indexing set might have a total order (and so the associated strict total order) on it (see Definition 4.38). For example, there is the standard order on $\mathbb{R}$ that is inherited by $[n]$ and $\mathbb{Z}_+$. This can be beneficial because it places the outputs in that order. Suppose $f : I \to S$ is an indexing function. If the I and S have total orders on them (and so the associated strict total orders), then f is said to be **strictly increasing**, **strictly decreasing**, **strictly monotone**, **non-decreasing**, **non-increasing**, or **monotone** according to whether f satisfies this property as a function (see Definition 4.24).

Strictly monotone lists will appear in the proof of the Erdős-Szekeres theorem when we study the pigeonhole principle in Volume 2, and non-decreasing and strictly increasing tuples will be a part of the rearrangement inequality (Theorem 11.18) and Chebyshev's inequality (Corollary 11.19) when we study multivariable inequalities.

Definition 3.3. If $A : I \to S$ is an indexing function where S is a set of sets, then we can denote the union of all sets that are elements of the indexed set by

$$\bigcup_{i \in I} A_i = \{a : \exists i \in I, a \in A_i\}.$$

We defined the Cartesian product of an ordered pair of sets in Definition 1.14. Equipped with the definition of a list of sets, we can generalize this concept.

Definition 3.4. Given a list of sets $(A_1, A_2, \ldots, A_n)$ for a positive integer n, we denote and define its **Cartesian product** to be the set

$$A_1 \times A_2 \times \cdots \times A_n = \{(a_1, a_2, \ldots, a_n) : a_1 \in A_1, a_2 \in A_2, \ldots, a_n \in A_n\}.$$

In words, it is the set of all n-tuples such that, for each index $i \in [n]$, the i^{th} entry is from A_i. If all of the A_i are the same set A, then we can use the notation

$$A^n = \underbrace{A \times A \times \cdots \times A}_{n \text{ copies of } A}.$$

Example. If one or more of the component sets of a Cartesian product are $\varnothing$, then the whole Cartesian product is $\varnothing$ because no tuples can be created. This also makes sense by contrapositive because, if a Cartesian produce is non-empty, then each of the component sets must contain at least one element.

An idea that we will now explore is that of counting the number of entries in a list without writing out every entry. That is, we will be calculating the length of a list, given some description of the list; this is an important skill in practice. Inevitably, there will need to be some pattern in the list in order for this task to be feasible, and we will usually need to know the first and last entries.

Example 3.5. The pages (not individual sheets) of a book are numbered from 1 to 100. How many pages are there in between 10 and 40 inclusive? Inclusive means we include the pages 10 and 40 at the ends.

Solution. It is tempting to say that the answer is $40 - 10 = 30$ but this is incorrect. The list of pages is

$$(10, 11, 12, \ldots, 40).$$

We subtract 9 from each entry to turn it into the list $(1, 2, 3, \ldots, 31)$. This means there are 31 pages from page 10 to page 40 inclusive. ■

Problem 3.6. The pages of a book are numbered from 1 to 100. How many pages are there in between 10 and 40 exclusive? Exclusive means we do not include the pages 10 and 40 at the ends.

In Example 3.5 and Problem 3.6, we saw that it is easy to make an "off-by-one error." More importantly, we found a method for determining the length of a list. The method generalizes nicely: assuming the given list has distinct entries at each index, we can apply injections to the list (so that the entries at all indices remain distinct at each step) until the list is of the form $(1, 2, 3, \ldots, n)$. This will mean the number of entries is n. Examples of useful injections to keep in our repertoire are division or multiplication by a non-zero constant, and addition and subtraction (we will call these "list transformations"). Addition or subtraction can be thought of as translating the list, and multiplication or division by a non-zero constant can be interpreted as scaling the list.

Problem 3.7. Find the number of multiples of 3 that are strictly greater than 110 and strictly less than 1000.

Problem 3.8. Though we will define arithmetic and geometric sequences later in Section 6.1, this is an appropriate time for the reader to prove the following statements.

1. Let a and $d \neq 0$ be real numbers. Then the number of entries in the list
$$(a + nd, a + (n + 1)d, \ldots, a + md)$$
is $m - n + 1$.
2. Let $b \neq 0$ and $r \neq 0, \pm 1$ be real numbers. Then the number of entries in the list
$$(br^n, br^{n+1}, \ldots, br^m)$$
is $m - n + 1$.

Note that the reason for the restrictions on the constants is to ensure that the terms of the sequences are all distinct.

Problem 3.9. Find the number of perfect cubes (these are cubes of integers) that are strictly greater than -1001 and strictly less than 1001.

3.2 Sums and Products

Definition 3.10. Finite and infinite series are defined as follows.

1. If n is a positive integer and $a : [n] \to \mathbb{R}$ is a list, then the corresponding **finite series** is denoted by and defined as

$$\sum_{i=1}^{n} a_i = a_1 + a_2 + \cdots + a_n.$$

If the indexing set is $\{p, p+1, p+2, \ldots, q\}$ for integers $p \leq q$, then the corresponding finite series can be denoted by

$$\sum_{i=p}^{q} a_i = a_p + a_{p+1} + \cdots + a_q.$$

If the upper bound on the index in a sum is less than the lower bound, then this "empty sum" is defined to be 0; this convention can help with writing general formulas where there are edge cases.

More generally, for any finite indexing set I, if $a : I \to \mathbb{R}$ is a list then the corresponding finite series is denoted by $\sum_{i \in I} a_i$, which denotes the sum of the outputs $a(i)$. This notation is acceptable because, thanks to the commutativity of addition, the order in which the a_i are added does not matter.

2. If $a : \mathbb{Z}_+ \to \mathbb{R}$ is a sequence, then the corresponding **infinite series** is denoted and defined as

$$\sum_{i=1}^{\infty} a_i = a_1 + a_2 + a_3 + \cdots = \lim_{n \to \infty} \sum_{i=1}^{n} a_i.$$

For each positive integer n, the term $\sum_{i=1}^{n} a_i$ is called the n^{th} **partial sum** of the series, and the value of the infinite series is the value that the partial sums approach as n gets arbitrarily large. This idea of **convergence** can be made formal using limits from calculus. If the indexing set is $\{p, p+1, p+2, \ldots\}$ for some integer p, then the corresponding infinite series is

$$\sum_{i=p}^{\infty} a_i = a_p + a_{p+1} + a_{p+2} + \cdots = \lim_{n \to \infty} \sum_{i=p}^{n} a_i,$$

and for each positive integer n, its n^{th} partial sum is $\sum_{i=p}^{p+n-1} a_i$. Unlike a finite series, if J is a countably infinite set and $a : J \to \mathbb{R}$ is a sequence, then the expression $\sum_{j \in J} a_j$ is not necessarily well-defined. This is because rearranging the terms of an infinite

series does not necessarily lead to the same final value. Readers interested in analysis should look up absolute convergence, conditional convergence, and the Riemann series theorem.

Both finite series and infinite series are called **series** or a **sum** for short. In a series, each **term** a_i is called a **summand**. The symbol $\sum$ in the context of series is called **sigma notation**.

Example. It is important to note that not all sequences lead to a convergent series. Here are some examples:

- Of course,
$$1 + 2 + 3 + 4 + \cdots$$
does not converge and instead goes to infinity because the partial sums $1 + 2 + \cdots + n$ get arbitrarily large.

- Even though the individual terms of the **harmonic series**
$$1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots$$
get arbitrarily close to 0, the partial sums (surprisingly!) get arbitrarily large.

- The partial sums of **Grandi's series**
$$1 + (-1) + 1 + (-1) + \cdots$$
oscillate between 1 ad 0, so it does not converge.

Example 3.11. Show that partial sums of the harmonic series
$$1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots$$
approach infinity as $n \to \infty$.

Solution. We will aim to get a lower bound on the m^{th} partial sum of the harmonic series. And that lower bound needs to go to infinity as n goes to infinity. The crucial observation is that we should look at the harmonic series in the chunks that exist in between terms whose indices are powers of 2. For example,
$$\begin{aligned} 1 + \frac{1}{2} + \left(\frac{1}{3} + \frac{1}{4}\right) + \left(\frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8}\right) &> \frac{1}{2} + \frac{1}{2} + \left(\frac{1}{4} + \frac{1}{4}\right) + \left(\frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8}\right) \\ &= \frac{1}{2} + \frac{1}{2} + 2 \cdot \frac{1}{4} + 4 \cdot \frac{1}{8} \\ &= \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2}. \end{aligned}$$

In general, for each positive integer $n \geq 2$,

$$\sum_{k=2^{n-1}+1}^{2^n} \frac{1}{k} > \sum_{k=2^{n-1}+1}^{2^n} \frac{1}{2^n} = \frac{2^n}{2} \cdot \frac{1}{2^n} = \frac{1}{2}.$$

Thus, for each positive integer n,

$$\sum_{k=1}^{2^n} \frac{1}{k} = 1 + \sum_{i=0}^{n-1} \sum_{k=2^i+1}^{2^{i+1}} \frac{1}{k} > \frac{1}{2} + n \cdot \frac{1}{2} = \frac{n+1}{2}.$$

Since $\frac{n+1}{2}$ grows unboundedly as $n \to \infty$ and since the partial sums of the harmonic series are strictly increasing, the partial sums go to infinity as $n \to \infty$. ■

Problem 3.12. The sum $\sum_{i=s}^{t} a_i$ starts with $i = s$ and then i goes up by increments of 1 up to and including $i = t$. Find a way of expressing (in summation notation) the sum of the same a_i but in the other direction, from $i = t$ to $i = s$. Remember, the counter or indexing variable can only go up by 1's, not down and not by any other increment.

Definition 3.13. The multiplicative analogue of sigma notation is **pi notation**. The symbol $\sum$ for finite and infinite series is replaced by $\prod$ for finite and infinite **products**. For positive integers n, finite sets I, and integers $p \leq q$, the meanings of

$$\prod_{i=1}^{n} a_i, \prod_{i=p}^{q} a_i, \prod_{i \in I} a_i, \prod_{i=1}^{\infty} a_i, \prod_{i=p}^{\infty} a_i$$

are analogous to the definitions of $\sum$ notation, where addition is replaced by multiplication. The multiplicative analogue of a summand is a **multiplicand**. For each positive integer n, just like how the n^{th} partial sum of an infinite series is the sum of the first n summands, the n^{th} **partial product** of an infinite product is the product of the first n multiplicands.

Definition 3.14. The empty sum is defined as the additive identity $\sum_{i \in \varnothing} a_i = 0$ and the empty product is defined as the multiplicative identity $\prod_{i \in \varnothing} a_i = 1$. This convention can be useful at times.

Definition 3.15. There are many scenarios in which there are **nested** sums or products (possibly both), which means the summands or multiplicands themselves contain sums or products; this nesting can have any finite depth. Sums and products in one nesting should never share symbols for indices. When writing or interpreting a nesting, one should go from left to right and, at each step into the nesting, consider each appearance of the symbols for indices from previous levels to be a constant. For example,

$$\prod_{i \in I} \sum_{j \in J} a_i b_j = \prod_{i \in I} \left(a_i \sum_{j \in J} b_j \right) = \left(\prod_{i \in I} a_i \right) \cdot \left(\sum_{j \in J} b_j \right)^{|I|},$$

where $|I|$ denotes the number of elements in I. We were able to factor a_i out of $\sum_{j \in J} a_i b_j$ in the first step using the distributive law because i is a constant at that depth. For those with experience with basic programming, this iterative mechanism of sigma and pi notation is akin to loops, with nested sums and products being like nested loops. To avoid ambiguity, it is possible to use parentheses like

$$\prod_{i \in I} \sum_{j \in J} a_i b_j = \prod_{i \in I} \left(\sum_{j \in J} (a_i b_j) \right)$$

to clarify what lies in the summands or multiplicands, but this is usually not necessary.

Definition 3.16. If the indexing set is

$$I = \{(i, j) \in [n] \times [n] : i < j\}$$

for some positive integer n, then two other ways of writing the sum $\sum_{(i,j) \in I} a_{i,j}$ are

$$\sum_{1 \leq i < j \leq n} a_{i,j} = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} a_{i,j}.$$

Both deserve some commentary:

- The left side is an example of summation notation where it is not clear which symbols represent the indices, and even if we knew them, it has not been made clear to what set they belong. In such cases, the convention is to determine which symbols have a fixed value (such as n), assume that the other symbols are all indices (such as i and j), and iterate over all integer values of these presumed indices that satisfy all other restrictions (such as the fact that i and j are integers that fall in $[1, n]$ and satisfy $i < j$). We will see much of this when studying Vieta's formulas (Theorem 10.43) and the principle of inclusion-exclusion in Volume 2.
- The right side is an example of writing the sum in a way that gives explicit instructions on how to perform the iteration, like a computer program. It is not always feasible, nor is it always necessary. Moreover, it is an example of a nested summation where the range of an inner index depends on an outer index, since j begins at $i + 1$.

When there are multiple indices like i and j in the subscript, $a_{i,j}$ may be written without the comma as a_{ij} if there is no chance of mistaking ij as a multiplication or concatenation.

Theorem 3.17. There are a few ways in which sums and products are commonly manipulated. The basic variants are stated below, and we expect that the reader can adapt them to other circumstances, such as the index starting at an integer other than 1, as needed. If I is a finite indexing set, then:

1. $\sum_{i \in I} (a_i + b_i) = \sum_{i \in I} a_i + \sum_{i \in I} b_i$ and $\prod_{i \in I} a_i b_i = \left(\prod_{i \in I} a_i \right) \cdot \left(\prod_{i \in I} b_i \right)$

2. $\displaystyle\sum_{i\in I} ca_i = c\cdot\sum_{i\in I} a_i$ and $\displaystyle\prod_{i\in I} ca_i = c^{|I|}\cdot\prod_{i\in I} a_i$

3. $\displaystyle\sum_{i\in I}\sum_{j\in J} a_i b_j = \sum_{i\in I}\left(a_i\sum_{j\in J} b_j\right) = \left(\sum_{i\in I} a_i\right)\cdot\left(\sum_{j\in J} b_j\right)$

We state these without proof as they are straightforward consequences of the rules of arithmetic. There are many other ways in which summations and products may be manipulated. We advise the reader to not memorize these rules, but rather to understand the meaning of the notation and apply the rules of arithmetic as seen fit.

Definition 3.18. Informally, if m and n are positive integers, then an $m\times n$ **matrix** is an array or table of numbers with m rows and n columns. More formally, it is an indexing function whose indexing set is $[m]\times[n]$.

Theorem 3.19 (Discrete Fubini's principle)**.** The essence of this principle is that, if there are nested summations, then it may be possible to alter the order in which the nesting occurs. This cannot be done blindly because there are numerous ways in which the indexing variables might depend on each other. While we are not aware of one general result that captures all possibilities, the results below are some of the most common instances of this idea, and the method used to prove them can be modified to suit other scenarios. Let m and n be positive integers and let a_{ij} be a real number for every $(i,j)\in[m]\times[n]$. Then

$$\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij} = \sum_{j=1}^{n}\sum_{i=1}^{m} a_{ij}.$$

If $m=n$ then

$$\sum_{i=1}^{n}\sum_{j=i}^{n} a_{ij} = \sum_{j=1}^{n}\sum_{i=1}^{j} a_{ij},$$

$$\sum_{i=1}^{n}\sum_{j=1}^{i} a_{ij} = \sum_{j=1}^{n}\sum_{i=j}^{n} a_{ij},$$

$$\sum_{i=1}^{n}\sum_{j=1}^{n+1-i} a_{ij} = \sum_{j=1}^{n}\sum_{i=1}^{n+1-j} a_{ij},$$

$$\sum_{i=1}^{n}\sum_{j=n+1-i}^{n} a_{ij} = \sum_{j=1}^{n}\sum_{i=n+1-j}^{n} a_{ij}.$$

The same results hold if we replace all of the summations $\sum$ with products $\prod$. There is no need to memorize these identities because they follow immediately from the proof technique.

Proof. We are tempted to provide the following "proof without words."

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mn} \end{bmatrix}$$

As an explanation, the idea is to place the a_{ij} from the sum in an $m \times n$ matrix that has m rows and n columns, where a_{ij} occupies the entry at the intersection of row i and column j. Then one way to find the sum of the a_{ij} is to add the sum of all rows and another way to find the sum of the a_{ij} is to add the sum of the all columns. The two methods yield the same final sum, so the two expressions are equal. The same method works for the other four sums, as shown below.

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ & a_{22} & a_{23} & \cdots & a_{2n} \\ & & a_{33} & \cdots & a_{3n} \\ & & & \ddots & \vdots \\ & & & & a_{nn} \end{bmatrix}, \begin{bmatrix} a_{11} & & & & \\ a_{21} & a_{22} & & & \\ a_{31} & a_{32} & a_{33} & & \\ \vdots & \vdots & \vdots & \ddots & \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn} \end{bmatrix},$$

$$\begin{bmatrix} a_{11} & \cdots & a_{1(n-2)} & a_{1(n-1)} & a_{1n} \\ a_{21} & \cdots & a_{2(n-2)} & a_{2(n-1)} & \\ a_{31} & \cdots & a_{3(n-2)} & & \\ \vdots & \iddots & & & \\ a_{n1} & & & & \end{bmatrix}, \begin{bmatrix} & & & & a_{1n} \\ & & & a_{2(n-1)} & a_{2n} \\ & & a_{3(n-2)} & a_{3(n-1)} & a_{3n} \\ & \iddots & \vdots & \vdots & \vdots \\ a_{n1} & \cdots & a_{n(n-2)} & a_{n(n-1)} & a_{nn} \end{bmatrix}$$

■

Problem 3.20. Let n be a positive integer and let a_{ij} be a real number for every $(i, j) \in [n] \times [n]$ such that $i < j$. Prove that

$$\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} a_{ij} = \sum_{j=2}^{n} \sum_{i=1}^{j-1} a_{ij}.$$

The discrete Fubini's principle is commonly applied in combinatorics, especially for proving more complicated combinatorial identities.

Chapter 4

Equality and Order

> "An equation has no meaning for me unless it expresses a thought of God."
>
> – *Srinivasa Ramanujan*

> "For an analyst, an equation is just two inequalities."
>
> – *Anonymous*

Equations and inequalities are the two most effective ways of relating real numbers to each other. These relations not only take on the form of static propositions, but also dynamic predicates where we want to find all the values of variables that make the equation or inequality true. This act, known as finding the "solutions" to the predicate, is a core aspect of mathematical activity. We will look at some common methods of solving equations and inequalities, while observing that the technique underlying all of them is expansion-and-contraction: find a possibly larger set of possibilities and eliminate anything that does not work. Finally, we will take a look at abstract orders, including total orders, partial orders, and well-orders, with a nod towards the fact that antisymmetry in a partial order is one of the most powerful tools lurking in the background of mathematics.

4.1 Equations

Definition 4.1. We previously defined an equation as the assertion that two objects are the same (Definition 1.19). An **equation** can also refer to a predicate like

$$A(x_1, x_2 \ldots, x_n) = B(x_1, x_2, \ldots, x_n),$$

where A and B are expressions of some sort, usually an amalgam of functions and operations acting on the variables x_i. Not all assignments of the x_i necessarily result in a true equation. If a tuple $(x_1, x_2 \ldots, x_n)$ of fixed values cause this equation to be true, then we say that the tuple **satisfies** the equation. Finding all such tuples is called **solving** the equation. If all possible tuples within a certain domain satisfy the equation, then the equation is said to be an **identity** on that domain.

Example. An example of an ordinary equation with specified objects is $1 = 1$. An example of a predicate equation is $x + 1 = 2x$. Its only solution is the number 1, so we say that $x = 1$ is the only solution. An example of an identity that holds for all $(x, y) \in \mathbb{R} \times \mathbb{R}$ (and in fact, for complex numbers too) is the **difference of squares** factorization

$$x^2 - y^2 = (x - y)(x + y).$$

Definition 4.2. The simplest equations are **linear equations** which have sums of variables times constants, called **coefficients**, equal to a constant. These equations look like

$$a_1x_1 + a_2x_2 + \cdots + a_nx_n = b,$$

where the x_i are the variables, the a_i are the coefficients, and b is the constant.

Example 4.3 (One-variable linear equation)**.** Find all real numbers x such that $ax + b = 0$, where b is a real constant and a is a non-zero real constant.

Solution. Suppose x is a real number that is a solution to the equation. Then $ax + b = 0$. First we subtract b from both sides (this means applying the function $f(t) = +(t, -b)$ to both sides) to get

$$ax = -b.$$

Then we divide both sides by the non-zero real a (this means applying the function $g(t) = \times(t, a^{-1})$ to both sides) to get

$$x = -\frac{b}{a}.$$

So the only possible solution is this number. However, we need to verify that it actually works. If $x = -\dfrac{b}{a}$, then

$$ax + b = a \cdot \left(-\frac{b}{a}\right) + b = -b + b = 0,$$

so it does work. We can either substitute in all solutions found like this to check which ones work, or we can check that all steps were reversible in the "only if" direction. This check is needed because the "only if" direction merely finds a *superset* of the set of solutions, but that does not mean that all of those values are solutions. Technically, what we have done here is use the fact that $P(x)$ and $Q(x)$ are predicates such that

$$\forall x \in \mathbb{R}, (P(x) \implies Q(x))$$

and $P(c)$ is true for some real constant c, then $Q(c)$ is true. This produces the subset relation

$$S = \{x \in \mathbb{R} : P(x)\} \subseteq \{x \in \mathbb{R} : Q(x)\} = T,$$

since

$$c \in S \iff P(c) \implies Q(c) \iff c \in T.$$

This can be conceptualized as expanding to a larger set of potential solutions and then contracting back to the smaller set of actual solutions. Ironically, the larger set might be easier to express. In physical situations and mathematical scenarios, there might be elements of the larger set that are not elements of the smaller set, say due to an extra restriction. In such cases, these elements are called **extraneous solutions**. ■

Problem 4.4. As we saw in the solution to Example 4.3, a standard technique used to solve an equation is to apply the same function to both sides of the equation. Often, we apply a binary operation with the first entry as a side of the equation and the second entry as some specific number, which makes it equivalent to apply a univariate function to both sides. Although equalities are preserved under such an application of a function, show that an inequation like $a \neq b$ is not necessarily preserved. A specific example will suffice.

Definition 4.5. A **system of equations** is a finite collection of equations. A **solution** to the system is a tuple of values for the variables such that the tuple of values satisfies every equation; we typically wish to find *all* solutions, though this is not always feasible and we might count ourselves lucky to find any solutions at all.

Barring the theory of linear algebra, there are three main elementary techniques for finding all solutions to a system of linear equations: substitution, elimination, and "smashing." All of the techniques rely on the expansion-contraction method: find a superset of the solutions in the "only if" direction and then remove any that do not actually work in the "if" direction.

- **Substitution**: Isolate a variable in one equation by writing it as an expression in terms of the other variables in the equation. Then remove this variable from the other equations by substituting the expression into all of the other equations in which the variable appears.
- **Elimination**: Given two equations $a = b$ and $c = d$, a third equation that is true is $ta + sc = tb + sd$ for any real s, t. By adding a suitable multiple of one equation to a suitable multiple of another equation, we might be able to simplify the equation, say by removing a variable. This is a modification of Theorem 2.16.
- **Smashing**: Add or multiply together all or most of the given equations. This works because if $a = b$ and $c = d$, then $a + c = b + d$ and $ac = bd$. Again, the engine driving this is Theorem 2.16.

We will see another technique for solving equations that utilizes symmetry to induce an order when we study inequalities (see Theorem 4.37).

Example 4.6 (Substitution)**.** Solve the following system of equations:

$$\begin{aligned} x^2 + y^2 &= 1, \\ x &= y. \end{aligned}$$

Solution. Suppose (x, y) is a solution to both equations. Thanks to the second equation, we can substitute x in for y in the first equation. This yields

$$2x^2 = 1 \iff x^2 = \frac{1}{2} \iff x = \pm\frac{1}{\sqrt{2}},$$

and y has the same value as x. Since the steps are reversible, the only two solutions are

$$\begin{aligned} x = y &= \frac{1}{\sqrt{2}}, \\ x = y &= -\frac{1}{\sqrt{2}}. \end{aligned}$$

By "the only," we mean that they are both solutions and there are no other solutions. ■

Example 4.7 (Elimination)**.** Solve the following system of equations:

$$4a + 3b = 90,$$
$$3a + 4b = 85.$$

Solution. We will use Theorem 2.16 several times in order to add and subtract equations together. Adding the two equations yields

$$7a + 7b = 175 \implies a + b = 25.$$

Subtracting the two equations yields

$$a - b = 5.$$

Adding the two new equations yields

$$2a = 30 \implies a = 15.$$

Subtracting the two new equations yields

$$2b = 20 \implies b = 10.$$

Substituting $(a, b) = (15, 10)$ into the original two equations shows that both equations are satisfied, so $(15, 10)$ is the only solution. ■

Problem 4.8. Let p, q, r be real constants. Suppose x, y, z are non-zero real numbers that solve the following system of equations:

$$p = x + \frac{1}{y},$$
$$q = y + \frac{1}{z},$$
$$r = z + \frac{1}{x}.$$

Express $xyz + \dfrac{1}{xyz}$ in terms of p, q, r using arithmetic operations.

Problem 4.9. Find all real numbers a such that a is its own additive inverse. Similarly, find all real numbers b such that b is its own multiplicative inverse.

Problem 4.10. It is important to ensure that each step of manipulating an equation makes sense and does not accidentally make an illegal move. Observe the following "proof" that starts with any two equal numbers and ends with $1 = 2$:

$$\begin{aligned} a &= b \\ a^2 &= ab \\ a^2 - b^2 &= ab - b^2 \\ (a - b)(a + b) &= (a - b)b \\ a + b &= b \\ 2a &= a \\ 2 &= 1. \end{aligned}$$

Can you spot the error?

Example 4.11. Find all solutions to the following system of equations:

$$\begin{aligned} 2x + 3y &= 4, \\ 6x + 9y &= 13. \end{aligned}$$

Solution. Suppose there exists a solution (x, y). Then the second equation implies that

$$2x + 3y = \frac{13}{3}.$$

The first equation says that

$$2x + 3y = 4.$$

Then $4 = \frac{13}{3}$, which contradicts the fact that these two numbers are unequal. Thus, there are no solutions. ■

Example 4.12. Find all solutions of the following system of equations:

$$\begin{aligned} 2x + 3y &= 4, \\ 4x + 6y &= 8. \end{aligned}$$

Solution. Suppose (x, y) is a solution to the system. The two equations imply each other because the second is twice the first and the first is half the second. We need only consider the first equation because the second equation provides no new information. We can isolate y in $2x + 3y = 4$ to get

$$y = \frac{4 - 2x}{3}.$$

So every solution is of the form $\left(t, \frac{4 - 2t}{3}\right)$ for some real t. The question is whether all elements of the set

$$\left\{\left(t, \frac{4 - 2t}{3}\right) : t \in \mathbb{R}\right\}$$

are solutions. We can check that this is the case by plugging these coordinates into $2x + 3y$, which yields

$$\begin{aligned} 2x + 3y &= 2t + 3\left(\frac{4 - 2t}{3}\right) \\ &= 2t + 4 - 2t \\ &= 4. \end{aligned}$$

So this is indeed the complete set of solutions. This way of expressing the set of all solutions in terms of one or more variables like t that have a high degree of freedom is called a **parametrization**, with t being called a **parameter**. ■

4.2 Inequalities

The reader should be familiar with the notion of order on the real numbers. This allows us to produce inequalities like $2 \geq 1$ and $0 < 1$ that compare where the numbers lie in relation to each other on the number line. Let us review the properties of real number inequalities. First of all, how do we compare two real numbers in decimal representation?

Definition 4.13. Two non-negative integers may be compared according to how many 1's are required to be added up to be equal to each. For example,

$$3 = 1 + 1 + 1 > 1 + 1 = 2.$$

Now let a and b be positive real numbers whose decimal representations are

$$a = a_n a_{n-1} \dots a_1 a_0 . a_{-1} a_{-2} a_{-3} \dots,$$
$$b = b_m b_{m-1} \dots b_1 b_0 . b_{-1} b_{-2} b_{-3} \dots,$$

where we choose decimal representations such that it is not true that all digits of sufficiently small indices are 9 (it will be discussed in Volume 3 that there always exists a unique representation that avoids this property). If $n > m$ then $a > b$, and if $m > n$ then $b > a$. Now suppose $n = m$. Then we go through the digits from left to right and compare the digits of a, b that have corresponding indices. If there are no conflicting digits, then $a = b$. Otherwise, suppose $a_k \neq b_k$ are the first, meaning leftmost, digits at which a conflict occurs. If $a_k > b_k$ then $a > b$, and if $a_k < b_k$ then $a < b$, regardless of what happens at digits further to the right. This property of base representations makes comparing decimal representations easy. As for what happens if one or both of a, b are non-positive:

1. All negatives are less than all positives
2. Zero is less than all positives and greater than all negatives
3. If both a, b are negative then we can use the fact that $a > b$ if and only if $-b > -a$ to reduce it to the problem of comparing two positive reals

The method described of comparing two positive reals in decimal representation can be proven using the base-10 expansion and a geometric series, but we have not covered the prerequisite material yet.

Definition 4.14. There is a difference between **strict** inequalities $<$ in which equality is not an option, and **non-strict** inequalities $\leq$ in which equality is possible. This is because the non-strict version $a \leq b$ is equivalent to '$a < b$ or $a = b$," and the strict version is equivalent to "$a \leq b$ and $a \neq b$." This is due to the trichotomy law (Theorem 4.16).

Example. The difference between strict and non-strict can have a giant effect effect in an argument, in terms of strength or validity, so one should pay attention to which one applies. A special case is $x \geq 0$ versus $x > 0$, where non-negative versus positive can make a world of difference.

Definition 4.15. An inequality I is **stronger** or **sharper** than an inequality J (and J is called **weaker**) if I is stronger as a conclusion, meaning I implies J.

Example. The strict inequality $a > b$ implies its non-strict counterpart $a \geq b$, so $a > b$ is stronger than $a \geq b$ (though we would still have to check that $a > b$ is in fact true). Note that if $a \geq b$ holds, it is not necessarily true that $a > b$ because $a = b$ might hold instead. On that note, $a = b$ implies $a \geq b$, but the converse does not necessarily hold because $a > b$ might be true.

Let us now observe some of the properties of inequalities to which we have become accustomed through practice.

Theorem 4.16. Inequalities of real numbers satisfy many properties, the most important of which are as follows for all $a, b, c \in \mathbb{R}$:

1. $a \leq a$ is true but $a < a$ is false
2. Trichotomy law: exactly one of $a < b, a = b, b < a$ is true. As a result, $a \leq b$ and $a > b$ are negations, and $a \geq b$ and $a < b$ are negations.
3. Connexity: at least one of $a \leq b$ and $b \leq a$ is true
4. Antisymmetry: if both $a \leq b$ and $b \leq a$ are true, then $a = b$
5. Transitivity: If $a \leq b$ and $b \leq c$, then $a \leq c$; if one or both of $a \leq b$ or $b \leq c$ is replaced by its strict counterpart in the hypothesis, then $a \leq c$ can be replaced by its strict counterpart in the conclusion.
6. The compatibility rules:
   - Additive compatibility: If $a > b$, then $a + c > b + c$. It follows that if $a \geq b$, then $a + c \geq b + c$.
   - Multiplicative compatibility: If $a > b$ and c is positive, then $ac > bc$. It follows that if $a \geq b$ and c is non-negative, then $ac \geq bc$. If c is negative, then applying these results leads to

$$\begin{aligned} a > b &\implies a(-c) > b(-c) \implies bc > ac, \\ a \geq b &\implies a(-c) \geq b(-c) \implies bc \geq ac. \end{aligned}$$

   Since subtraction is addition in disguise and division by a non-zero real is multiplication in disguise, there are analogues of the compatibility rules for them too. We leave it to the reader to write them out.

Problem 4.17. Let a, b, c, d be real. As an analogue of Theorem 2.16 for inequalities, prove that we can "add" and "multiply" inequalities as follows:

- If $a > b$ and $c > d$, then $a + c > b + d$.
- If $a \geq b$ and $c > d$, then $a + c > b + d$.

- If $a \geq b$ and $c \geq d$, then $a + c \geq b + d$.
- If $a > b > 0$ and $c > d > 0$, then $ac > bd$.

Multiplication results involving non-strict inequalities or non-positive numbers are left to the reader to discover in the same manner.

Theorem 4.18. If a and b are integers, then $a > b$ if and only if $a - 1 \geq b$. An equivalent form of the second proposition is $a \geq b + 1$. This result allows us to switch between a strict inequality and a non-strict inequality when we are dealing with integers.

Proof. Let a and b be integers such that $a > b$. Then $a - b > 0$. This means that $a - b$ is a positive integer, so $a - b \geq 1$, which is equivalent to $a - 1 \geq b$ or $a \geq b + 1$. Conversely, suppose $a - 1 \geq b$. Then $a > a - 1 \geq b$, so $a > b$ by transitivity. ■

There are several concepts related to bounds of which one should be aware.

Definition 4.19. Let S be a non-empty set of real numbers.

1. An **upper bound** on S is a real number a such that for all $x \in S$, it holds that $x \leq a$. A **lower bound** on S is a real number b such that for all $x \in S$, it holds that $x \geq b$. Upper or lower bounds do not depend on each other and it is possible that one or both do not exist for a given set.

2. A **maximum** of S is an upper bound on S that is also an element of S. A **minimum** is S is a lower bound on S that is also an element of S. Maxima and minima do not always exist, even if upper and lower bounds exist, but they always exist for finite sets. However, when a maximum or minimum does exist, it is unique (try proving this!). The maximum is denoted by $\max(S)$ and the minimum is denoted by $\min(S)$. These functions are easily extended to non-set collections such as lists and multisets (these will feature prominently in Volume 2)

3. A **supremum** of S is a minimum element of the set of upper bounds on S. An **infimum** of S is a maximum element of the set of lower bounds on S. There is a semi-secret but extremely important property of the real numbers called **Dedekind completeness** which says that if S has an upper bound then S has a supremum, and if S has a lower bound then S has an infimum. Note that $\sup(S)$ and $\inf(S)$, as they are denoted, are not necessarily elements of S.

There are many properties of the $\max, \min, \sup, \inf$ functions that are too numerous to list here. For example, for any positive α,

$$\max\{\alpha x : x \in S\} = \alpha \cdot \max(S)$$

and for any negative β,

$$\max\{\beta x : x \in S\} = \beta \cdot \min(S).$$

Another example is that, if we define max and min as binary operations on ordered pairs, then they are commutative and associative; this will be helpful when analyzing the gcd and lcm functions in Volume 3. We will simply allow such properties to come up in practice.

Theorem 4.20 (Archimedean property)**.** For every real number r, there exists an integer n such that $n > r$.

Proof. Suppose, for contradiction, the negation of the Archimedean property. This means that there exists a real number r such that, for all integers n,

$$n \leq r.$$

So the integers are bounded above by r. By the Dedekind completeness of $\mathbb{R}$ there exists a least upper bound ℓ of $\mathbb{Z}$. The trick is to note that $\ell - 1$ is not an upper bound on $\mathbb{Z}$ because it is strictly less than the supremum ℓ. So there exists an integer m such that $\ell - 1 < m$. But then $\ell < m + 1$. Since $m + 1$ is an integer, this contradicts the fact that ℓ is an upper bound on $\mathbb{Z}$. Thus, our initial assumption was wrong. Instead, the Archimedean property holds, making the integers "unbounded." ■

Problem 4.21. Find an infinite sequence of positive rationals $q_1, q_2, q_3, \ldots$ that gets closer and closer to 0 but never reaches 0. More precisely, the sequence should satisfy the following condition:

$$\forall \epsilon > 0, \exists N \in \mathbb{Z}_+, \forall n \in \mathbb{Z}_+ : n \geq N \implies q_n < \epsilon.$$

Definition 4.22. To effectively describe sets of real numbers such as domains, ranges, codomains, we introduce **interval notation**. Let a and b be real numbers. There are nine types of intervals:

- The **closed** interval $[a, b]$ is the set $\{x \in \mathbb{R} : a \leq x \leq b\}$.
- The **open** interval (a, b) is the set $\{x \in \mathbb{R} : a < x < b\}$.
- The interval $[a, b)$ is the set $\{x \in \mathbb{R} : a \leq x < b\}$.
- The interval $(a, b]$ is the set $\{x \in \mathbb{R} : a < x \leq b\}$.
- The interval $[a, \infty)$ is the set $\{x \in \mathbb{R} : a \leq x\}$.
- The interval (a, ∞) is the set $\{x \in \mathbb{R} : a < x\}$.
- The interval $(-\infty, a]$ is the set $\{x \in \mathbb{R} : x \leq a\}$.
- The interval $(-\infty, a)$ is the set $\{x \in \mathbb{R} : x < a\}$.
- The interval $(-\infty, \infty)$ is the set $\mathbb{R}$ of all real numbers.

Intervals are often combined in various ways using set operations, such as union, intersection, complement and difference.

Problem 4.23. Find a set $S \subseteq \mathbb{R}$ such that S has an upper bound and a lower bound, but S does not have a maximum or minimum.

Definition 4.24. Let X be a non-empty subset of $\mathbb{R}$. Then the function $f : X \to \mathbb{R}$ is said to be:

- **Strictly increasing** if for all $x_1, x_2 \in X$, it holds that $x_1 < x_2 \implies f(x_1) < f(x_2)$.
- **Strictly decreasing** if for all $x_1, x_2 \in X$, it holds that $x_1 < x_2 \implies f(x_1) > f(x_2)$.
- **Non-decreasing** if for all $x_1, x_2 \in X$, it holds that $x_1 < x_2 \implies f(x_1) \leq f(x_2)$.
- **Non-increasing** if for all $x_1, x_2 \in X$, it holds that $x_1 < x_2 \implies f(x_1) \geq f(x_2)$.

Moreover, a function that is strictly increasing or strictly decreasing is called **strictly monotone**, and a function that is non-decreasing or non-increasing is called **monotone**.

Composing (strictly) monotonic functions produces more such functions in ways that we leave it to the reader to discover. For example, composing two strictly decreasing functions produces a strictly increasing function, like a double negative turning into a positive.

Theorem 4.25. Let X be a subset of $\mathbb{R}$ and let $f : X \to \mathbb{R}$ be function. Let x_1 and x_2 be any elements of X. Then:

1. If f is strictly increasing or non-decreasing then $f(x_1) < f(x_2) \implies x_1 < x_2$.
2. If f is strictly decreasing or non-increasing then $f(x_1) < f(x_2) \implies x_1 > x_2$.

Proof. The two proofs are very similar.

1. Suppose f is strictly increasing or non-decreasing and that $f(x_1) < f(x_2)$. Suppose for contradiction that $x_1 \geq x_2$. Then $f(x_1) \geq f(x_2)$, which is a contradiction. So $x_1 < x_2$.
2. Suppose f is strictly decreasing or non-increasing and that $f(x_1) < f(x_2)$. Suppose for contradiction that $x_1 \leq x_2$. Then $f(x_1) \geq f(x_2)$, which is a contradiction. So $x_1 > x_2$.

■

Lemma 4.26. Strictly monotone functions, whether strictly increasing or strictly decreasing, are injective.

Proof. We will show the proof for strictly increasing functions. Suppose f is strictly increasing. To prove that f is injective, it suffices to show that

$$x_1 \neq x_2 \implies f(x_1) \neq f(x_2).$$

There are two cases, $x_1 > x_2$ or $x_1 < x_2$. These are handled as follows:

$$\begin{aligned} x_1 > x_2 &\implies f(x_1) > f(x_2) \implies f(x_1) \neq f(x_2), \\ x_1 < x_2 &\implies f(x_1) < f(x_2) \implies f(x_1) \neq f(x_2). \end{aligned}$$

We leave the analogous proof for strictly decreasing functions to the reader as an exercise. ■

Theorem 4.27. Let $f : X \to Y$ be a strictly monotone function where X is a subset of $\mathbb{R}$ and Y is the range of f. Then f is invertible and:

1. If f is strictly increasing, then f^{-1} inherits this property.

2. If f is strictly decreasing, then f^{-1} inherits this property.

Proof. By Lemma 4.26 and the fact that Y is the range of f, we get that f is bijective onto Y. This is equivalent to f having an inverse $f^{-1} : Y \to X$.
For the inheritance property, we will tackle each case separately. Suppose y_1 and y_2 are elements of Y such that $y_1 < y_2$ Let $f^{-1}(y_1) = x_1$ and $f^{-1}(y_2) = x_2$.

1. If f is strictly increasing, then

$$f(x_1) = y_1 < y_2 = f(x_2) \implies f^{-1}(y_1) = x_1 < x_2 = f^{-1}(y_2),$$

   where we have used the fact that $f(x_1) < f(x_2) \implies x_1 < x_2$ from Theorem 4.25. Thus, f^{-1} is also strictly increasing.

2. If f is strictly decreasing, then

$$f(x_1) = y_1 < y_2 = f(x_2) \implies f^{-1}(y_1) = x_1 > x_2 = f^{-1}(y_2),$$

   where we have used the fact that $f(x_1) < f(x_2) \implies x_1 > x_2$. Thus, f^{-1} is also strictly decreasing.

This inheritance property does not extend to non-decreasing and non-increasing functions because they are not necessarily injective, and so such a function does not even necessarily have an inverse. ■

Theorem 4.28. The following are examples of functions with properties of strict monotonicity:

1. The reciprocal function $f(x) = \dfrac{1}{x}$ is strictly decreasing on $(-\infty, 0)$ and it is also strictly decreasing on $(0, \infty)$, but the results are separate. The reciprocal of any positive real is greater than the reciprocal of any negative real.

2. For any odd positive integer n, $f(x) = x^n$ is strictly increasing on all of $\mathbb{R}$. As a consequence, by Theorem 4.27, its inverse function $g(x) = \sqrt[n]{x} = x^{\frac{1}{n}}$ is strictly increasing on all of $\mathbb{R}$.

3. For even positive integers n, $f(x) = x^n$ is strictly decreasing on $(-\infty, 0]$ and strictly increasing on $[0, \infty)$. As a consequence, the inverse function $g(x) = \sqrt[n]{x} = x^{\frac{1}{n}}$ is strictly increasing on $[0, \infty)$, which is the domain of g.

4. For real $b > 1$, $f(x) = b^x$ is strictly increasing on $\mathbb{R}$. As such, its inverse $g(x) = \log_b x$ is also strictly increasing on $\mathbb{R}_+$, which the domain of g.

5. For real $0 < b > 1$, $f(x) = b^x$ is strictly decreasing on $\mathbb{R}$. As such, its inverse $g(x) = \log_b x$ is also strictly decreasing on $\mathbb{R}_+$, which is the domain of g.

Note that, since the exponential function b^x (for positive $b \neq 1$) is strictly monotonic, it is injective. Thus, we can "equate exponents" in $b^x = b^y$ to get $x = y$.

Proof. The first property is a result of multiplicative compatibility as we just have to manipulate the inequality $x_1 < x_2$ into $\dfrac{1}{x_2} < \dfrac{1}{x_1}$. Rigorous proofs of the remaining properties follow from calculus, which is beyond the scope of the text. ■

Note that the compatibility rules for inequalities (Theorem 4.16) follow from applying the monotone functions $f(t) = +(t, c)$ and $g(t) = \times(t, c)$ to inequalities.

Lemma 4.29 (Equality lemmas)**.** In inequalities, we often want to know if and when equality holds. There is a pair of lemmas that come in handy for this purpose:

1. If
$$a_1 \geq a_2 \geq \cdots \geq a_n,$$
then $a_1 = a_n$ if and only if
$$a_1 = a_2 = \cdots = a_n.$$

2. If
$$a_1 \geq b_1, a_2 \geq b_2, \ldots, a_n \geq b_n,$$
then
$$a_1 + a_2 + \cdots + a_n \geq b_1 + b_2 + \cdots + b_n,$$
where equality holds if and only if
$$a_1 = b_1, a_2 = b_2, \ldots, a_n = b_n.$$

Proof. Both of these lemmas are useful for determining the equality cases of multivariable inequalities, such as those studied in Chapter 11.

1. If all of the a_k are equal, then clearly $a_1 = a_n$. Conversely, suppose $a_1 = a_n$. We know that, for each $k \in [n]$,
$$a_1 \geq a_k \geq a_n \implies a_1 = a_k = a_n,$$
where we have used antisymmetry. So all of the a_k are equal.

2. The inequality
$$a_1 + a_2 + \cdots + a_n \geq b_1 + b_2 + \cdots + b_n$$
holds by adding together all of the inequalities in the hypothesis. The tricky part is the equality criterion. If equality holds in all of the $a_k \geq b_k$, then adding up the equations yields
$$a_1 + a_2 + \cdots + a_n = b_1 + b_2 + \cdots + b_n.$$
Conversely, suppose this last equation holds. We can rewrite it as
$$(a_1 - b_1) + (a_2 - b_2) + \cdots + (a_n - b_n) = 0.$$
Each summand $a_k - b_k$ is non-negative by assumption. If any of the $a_k \geq b_k$ is strict then $a_k - b_k$ is positive, leading to the contradiction
$$(a_1 - b_1) + (a_2 - b_2) + \cdots + (a_n - b_n) > 0.$$
Thus, $a_k = b_k$ for all $k \in [n]$.

■

Problem 4.30. Let $(a_k)_{k=1}^n$ and $(b_k)_{k=1}^n$ be n-tuples of positive reals such that

$$a_1 \geq b_1, a_2 \geq b_2, \ldots, a_n \geq b_n.$$

Prove that

$$a_1 a_2 \cdots a_n \geq b_1 b_2 \cdots b_n,$$

where equality holds if and only if

$$a_1 = b_1, a_2 = b_2, \ldots, a_n = b_n.$$

Definition 4.31. Just like equations, **inequalities** can take the form of predicates, where the **solution** is the set of values of the variables that satisfy the inequality. When we study multivariable inequalities momentarily and again in Chapter 11, we will find inequalities that are **identities** in the sense that they hold for all real values of the variables (or some other large domain like the positive reals).

Example 4.32. Solve $-2x + 3 > 0$ for all real x that satisfy it.

Solution. Suppose x is a real number that satisfies the inequality. Just like with equations, we apply the same operation to both sides until we isolate x:

$$-2x + 3 > 0 \iff -2x > -3 \iff x < \frac{3}{2}.$$

The steps are reversible, so the set of solutions is $\left(-\infty, \frac{3}{2}\right)$. Note that dividing both sides by -2 flipped the inequality sign because $f(t) = \times(t, (-2)^{-1})$ is a decreasing function due to $(-2)^{-1}$ being negative. In general, one has to be careful when multiplying or dividing by a negative real number, or a variable that might store a negative value. ■

Example 4.33 (Sign analysis)**.** Find all real x such that $\dfrac{x+2}{x-3} < 4$.

Solution. It would be ideal if we could clear the denominator, but we do not know its sign and so we do not know whether the inequality symbol would be flipped. Instead, we take the constant over to the other side and combine everything:

$$0 > \frac{x+2}{x-3} - 4 = \frac{x + 2 - 4(x-3)}{x-3} = \frac{-3x + 14}{x-3}.$$

Thus, we want to know the real numbers x for which $\dfrac{-3x+14}{x-3}$ is negative. We can figure this out using a technique called "sign analysis" where we look at the signs of the factors $-3x + 14$ and $x - 3$, since the sign of $\dfrac{-3x+14}{x-3}$ is the sign of $-3x + 14$ times the sign of $x - 3$. Neither factor can have a 0 sign because $-3x + 14$ would make the quotient 0 and $x - 3 = 0$ would cause division by 0.

Note that, if $(+)$ represents a positive sign and $(-)$ represents a negative sign, then

$$(+)\cdot(+) = (-)\cdot(-) = (+),$$
$$(+)\cdot(-) = (-)\cdot(+) = (-).$$

Since we want an overall negative sign, we only care for the cases where

$$-3x + 14 < 0 \text{ and } x - 3 > 0,$$
$$-3x + 14 > 0 \text{ and } x - 3 < 0.$$

This yields the two intervals

$$\{x \in \mathbb{R} : -3x + 14 < 0\} \cap \{x \in \mathbb{R} : x - 3 > 0\} = \left(\frac{14}{3}, \infty\right) \cap (3, \infty) = \left(\frac{14}{3}, \infty\right),$$
$$\{x \in \mathbb{R} : -3x + 14 > 0\} \cap \{x \in \mathbb{R} : x - 3 < 0\} = \left(-\infty, \frac{14}{3}\right) \cap (-\infty, 3) = (-\infty, 3).$$

This yields a final solution of

$$(-\infty, 3) \cup \left(\frac{14}{3}, \infty\right)$$

Sign analysis becomes increasing difficult as the number of factors in the numerator and denominator increases due to increased casework. As such, we will develop a more efficient alternative called "interval analysis" in Example 10.40. ■

The following is a result that is easy to overlook, especially because of its name, but it has powerful implications, some of which we will see when we study multivariable inequalities, such as the Cauchy-Schwarz inequality in Theorem 11.9.

Theorem 4.34 (Trivial inequality)**.** The inequality $x^2 \geq 0$ is an identity on the real numbers, with equality holding if and only if $x = 0$.

Proof. This is easy to see by doing casework on the sign of x. If $x = 0$, then $x^2 = 0$. Regardless of whether x is positive or negative, $x^2 = x \cdot x$ is positive, moreover, equality does not hold in these cases. ■

Problem 4.35. Prove that

$$\frac{x^2 + y^2}{2} \geq xy$$

for all reals x, y, with equality holding if and only if $x = y$.

Theorem 4.36 (Bernoulli's inequality)**.** If m is a non-negative integer and x is a real number such that $x \geq -1$, then

$$(1 + x)^m \geq 1 + mx.$$

If $x = -1$ and $m = 0$, then we use the convention that $0^0 = 1$. Moreover, equality holds if and only if $m = 0$ or $m = 1$ or $x = 0$.

Though it might seem like a passing curiosity, this is not an inequality that lives in a vacuum. For example, it implies that for any positive real number r, positive integer t, and integer $n \geq 2$,

$$\left(1 + \frac{r}{n}\right)^{nt} > 1 + nt \cdot \frac{r}{n} = 1 + rt.$$

For those familiar with the basics of investing, this shows that compound interest at a rate of r is a strictly better deal than simple interest at a rate of r, as long as the number of times n that the compound interest is applied in a single time period is at least 2.

Proof. First we will prove that the stated equality cases are indeed sufficient for equality. If $m = 0$, then

$$(1 + x)^m = (1 + x)^0 = 1 = 1 + 0 \cdot x = 1 + mx.$$

If $m = 1$, then

$$(1 + x)^m = (1 + x)^1 = 1 + x = 1 + 1 \cdot x = 1 + mx.$$

If $x = 0$, then

$$(1 + x)^m = (1 + 0)^m = 1 = 1 + m \cdot 0 = 1 + mx.$$

Now suppose $x \neq 0$. We will prove, by induction on $m \geq 2$, that the strict inequality

$$(1 + x)^m > 1 + mx$$

holds. In the base case $m = 2$, we want to prove that

$$(1 + x)^2 > 1 + 2x.$$

Upon expanding the left side and cancelling terms common to both sides, this becomes equivalent to $x^2 > 0$, which is true by the trivial inequality because $x \neq 0$. Now suppose there exists an integer $m \geq 2$ such that

$$(1 + x)^m > 1 + mx.$$

Then multiplying both sides of the induction hypothesis by $1 + x$ (recall that this is positive) yields

$$\begin{aligned}(1 + x)^{m+1} &\geq (1 + mx)(1 + x) \\ &= 1 + x + mx + mx^2 \\ &= 1 + (m + 1)x + mx^2 \\ &> 1 + (m + 1)x.\end{aligned}$$

Thus, we have proven Bernoulli's inequality by induction.

The corollary about interest is true because the formula for compound interest is $p\left(1 + \frac{r}{n}\right)^{nt}$ and the formula for simple interest is $p(1 + rt)$. Here, p is the principal (your original amount of money), r is the interest rate as a decimal like 0.05 (not percentage), t is the number of time periods elapsed (this is usually the number of years), and n is the number of times that the compound interest is applied per time period (if one time period is a year, then examples of n could be $4, 12, 365$). See Example 11.5 for why compounding more times leads to a better deal for the receiver than compounding fewer times. ■

Theorem 4.37. A technique involving orders that will occasionally be useful in solving inequalities or equations or proving identities is symmetry.

1. For solving inequalities or equations, if, for each pair of variables out of some set of variables, pairwise swapping them does not alter the truth of the statement, then we can assume that that set of variables fall in a certain order. After finding solutions under this order, all other solutions can be found by finding all permutations of these solutions.

2. For proving multivariable inequalities that are symmetric in some variables, the story is simpler. It simply suffices to assume a certain order of the symmetric variables and prove that the inequality holds in that case, and this proof will translate over too all other orders. For example, we will see Schur's inequality (Theorem 11.14) which says that for all non-negative real numbers x, y, z and all positive real numbers t, it holds that
$$x^t(x-y)(x-z) + y^t(y-z)(y-x) + z^t(z-x)(z-y) \geq 0.$$
Since x, y, z are symmetric here, we can and will assume without loss of generality that $x \geq y \geq z$ in the proof.

If two symmetric variables are known to represent unequal numbers, we can even assume a strict inequality between them.

4.3 Abstract Orders

The inequality relation on real numbers can be generalized. Given a set, we might want to place its elements in a particular order. This idea of ordering elements of a set is formalized as follows.

Definition 4.38. A **total order** on a set is a binary relation $\leq$ on the set that behaves very similarly to non-strict inequalities on real numbers. To be precise, $\leq$ is a total order on the set S if for all $a, b, c \in S$ the following properties hold:

- Connexity: $a \leq b$ or $b \leq a$
- Antisymmetry: if $a \leq b$ and $b \leq a$, then $a = b$
- Transitivity: if $a \leq b$ and $b \leq c$, then $a \leq c$

The associated **strict total order** $<$ is a binary relation on S that is defined as: $a < b$ holds if it is not true that $b \leq a$. Equivalently, $a < b$ holds if $a \leq b$ and $a \neq b$. The strict total order satisfies the following properties for all $a, b, c \in S$:

- Irreflexivity: $a < a$ is false
- Asymmetry (different from antisymmetry): if $a < b$ is true, then $b < a$ is false
- Transitivity: if $a < b$ and $b < c$, then $a < c$

- Trichotomy law: exactly one of $a < b, a = b, b < a$ is true

Note that the trichotomy law implies irreflexivity and asymmetry, so we do not really need to point out the first two properties listed, other than for pedagogical purposes.

Example. The $\leq$ relation on the real numbers is a total order, and it induces the strict total order $<$ on the real numbers. Visually, it allows us to place the real numbers on a line in such a way that two numbers a, b satisfy $a < b$ if and only if a is to the left of b.

Problem 4.39. Let us go in the reverse direction by using a strict total order to induce a total order. Prove that if $<$ is a transitive trichotomous binary relation, then the binary relation $\leq$ is a total order, where $\leq$ is defined as: $a \leq b$ if and only if $a < b$ or $a = b$.

In the context of manual combinatorics (i.e. counting elements of a finite set by listing them out), there is a particular total order that can be helpful, which we will define momentarily as the lexicographical order.

Definition 4.40. A collection of distinct symbols or letters is called an **alphabet**. A finite list of symbols of the alphabet is called a **word** or **string**; the symbols are ordered from right to left. The **length** of a string is the number of symbols in it.

Example. The **binary alphabet** is the set $\{0, 1\}$ and its strings are finite lists of 0's and 1's.

Definition 4.41. Suppose $\mathcal{A}$ is an alphabet with a total order $\geq$ on it with the corresponding strict total order $>$. We can define a total order, called **lexicographical order**, on the set of strings made of n symbols from $\mathcal{A}$, for a fixed positive integer n, as follows. Suppose there are strings

$$a = a_1 a_2 \cdots a_{n-1} a_n,$$
$$b = b_1 b_1 \cdots b_{n-1} b_n$$

such that $a \neq b$. Then there exists an index $1 \leq i \leq n$ where $a_i \neq b_i$. Let j be the smallest such index at which a and b differ. If $a_j > b_j$ then $a > b$, and if $a_j < b_j$ then $a < b$.

Example. Lexicographical order is used in dictionaries and encyclopedias so that there is a standard way to find entries according to their alphanumeric label. The way that we compared decimal representations of real numbers in Definition 4.13 is a variation of lexicographical order where the strings have countably infinite letters.

Definition 4.42. As we have defined it, lexicographical order tells us how to compare only strings that have the same number of symbols. Dictionaries extend this idea by defining a new "blank" letter that is smaller than every other letter, and adding exactly enough blanks to the right end of the smaller word so that the two words have the same number of letters. Then it is possible to compare them using lexicographical order. There is a different extension, called **shortlex order**, that is sometimes more convenient from an organizational perspective. It says that, given two strings of different length, the string with fewer letters is automatically smaller than the string with more letters. This is not so useful in a dictionary setting when we might not know the exact spelling of a word (and so the length of the word might be unknown, making it difficult to locate).

Example. If we follow the example of dictionaries, aaa comes before ab. In shortlex order, ab comes before aaa.

Problem 4.43. In shortlex order, list all binary alphabet strings that consist of one, two, or three symbols. The total order on the binary alphabet is given by $0 < 1$.

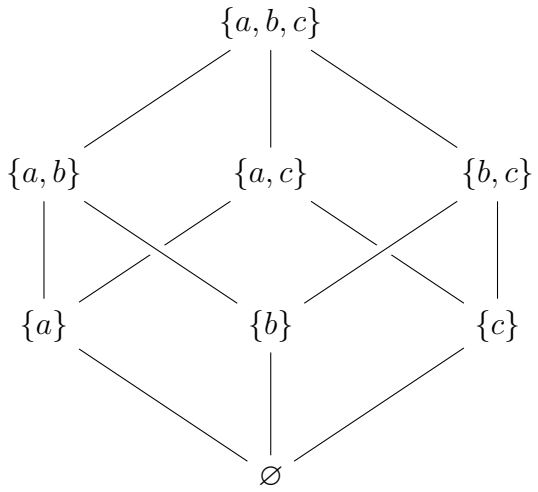
What if there are notions of ordering where antisymmetry and transitivity hold, but connexity does not? That is, maybe it is possible to preserve the essence of inequalities without it being possible to compare every pair of elements. This motivates the next definition.

Definition 4.44. A **partial order** on a set S is a binary relation $\leq$ on S (recall from Definition 1.15 that this is a subset of $S \times S$) such that for all $a, b, c \in S$ the following properties hold:

1. Reflexivity: $a \leq a$
2. Antisymmetry: if $a \leq b$ and $b \leq a$, then $a = b$
3. Transitivity: if $a \leq b$ and $b \leq c$, then $a \leq c$

Note that connexity implies reflexivity, so every total order is a partial order.

Example. Given a set S, the subset relation $\subseteq$ on its power set $\mathcal{P}(S)$ is a partial order. Below is what is known as *Hasse diagram* of the partial order structure of the subsets of $\{a, b, c\}$.



In Volume 3, we will define that a positive integer a divides a positive integer b if there exists a (necessarily positive) integer c such that $ac = b$. This relation is denoted by $a \mid b$, and it is an important partial order on the positive integers.

We cannot stress enough the incredible power of antisymmetry. It allows us to use two instances of the weaker notion of order to achieve an equality (equality is stronger than order because reflexivity says that if $a = b$ then $a \leq b$, but the other direction is not true in general). Antisymmetry will come up over and over again in our studies, especially with respect to numbers, subsets, and divisibility. For example, the "sandwiching"

$$a \leq b \leq a \implies a = b$$

will come up in probability with $b = 0$ and $b = 1$. In number theory, the antisymmetry of divisibility is an indispensable tool. In combinatorics and set theory, we have a variation of antisymmetry called the Schröder-Bernstein theorem.

Definition 4.45. One more kind of ordering, which we will not study abstractly, are well-orders. A **well-order** on a set S is a total order on S such that every non-empty subset of S has a least element under this ordering. More precisely, if the well-order is $\leq$ and T is a non-empty subset of S, then

$$\exists b \in T, \forall a \in T, b \leq a.$$

Example. The only example of a well-order that we will need is the ordinary order on $\mathbb{Z}_+$, or more generally, any subset of $\mathbb{Z}$ that has a lower bound. This is due to the well-ordering principle, which is described in Theorem 1.56. Note that $\mathbb{Z}$ itself is not well-ordered because its elements descend downwards infinitely. Under the ordinary order, $\mathbb{R}$ is not well-ordered either, but it is one of the pathological consequences of the axiom of choice that a well-order can be defined on any set, including the reals.

We conclude by noting that partial orders include total orders, and total orders include well-orders, thus forming a nice hierarchy of orders.

Chapter 5

Special Functions

> "For fifteen days I struggled to prove that no functions analogous to those I have since called Fuchsian functions could exist; I was then very ignorant. Every day I sat down at my work table where I spent an hour or two; I tried a great number of combinations and arrived at no result. One evening, contrary to my custom, I took black coffee; I could not go to sleep; ideas swarmed up in clouds; I sensed them clashing until, to put it so, a pair would hook together to form a stable combination. By morning I had established the existence of a class of Fuchsian functions, those derived from the hypergeometric series. I had only to write up the results which took me a few hours."
>
> – *Henri Poincaré, Science et Methode*

> "A mathematician is a device for turning coffee into theorems."
>
> – *Alfréd Rényi*

Having seen general properties of functions in Section 1.2, we will now see specific examples of functions that are common in mathematics. This includes absolute value, the signum function, floor function, and ceiling function. Absolute value naturally introduces us to the real triangle inequality, which has geometric and complex variants. The floor and ceiling functions are essential for rounding in a precise manner, and they have many properties, the most useful of which we will study.

5.1 Absolute Value

Definition 5.1. A particular definition of a function is said to be **piecewise** if the definition depends on different cases. To be clear, being piecewise is a property of the way a function is presented, not a property of a function itself. Nevertheless it is commonplace to speak of "piecewise functions" if there is a relevant definition of the function that is piecewise.

Example. A function might have different definitions on the intervals $(-\infty, 3]$, $(3, 17)$, and $[17, \infty)$. We denote such presentations by placing curly bracket on the left side of the cases as follows:

$$f(x) = \begin{cases} 2x^2 & \text{if } x \in (-\infty, 3] \\ 3x - 1 & \text{if } x \in (3, 17) \\ 5x^3 + 4x & \text{if } x \in [17, \infty) \end{cases}.$$

Some examples of piecewise functions that we will see are the absolute value and signum functions, and the floor and ceiling functions.

Recall, from Definition 2.29, that the absolute value of a real number is its distance from 0 on the real number line. So the absolute value function $|\cdot| : \mathbb{R} \to \mathbb{R}$ is defined by

$$|x| = \begin{cases} x & \text{if } x \geq 0 \\ -x & \text{if } x < 0 \end{cases}.$$

Theorem 5.2. There are several properties of the absolute value function that often come in handy. We state them here and leave the proofs to the reader because they are straightforward. Let x and y be real numbers and n be an integer. Then:

1. Instead of the given piecewise definition, an alternate unified definition of absolute value is $|x| = \sqrt{x^2}$, where we choose the non-negative square root as usual.
2. $|x| = 0$ if and only if $x = 0$, and otherwise $|x| > 0$. So $|x| \geq 0$ always.
3. It is true that $|x| \geq x$, with equality holding if and only if $x \geq 0$.
4. $|xy| = |x| \cdot |y|$
5. For $x \neq 0$, $\left|\frac{1}{x}\right| = \frac{1}{|x|}$, and so $\left|\frac{y}{x}\right| = \frac{|y|}{|x|}$
6. $|x^n| = |x|^n$, and if n is even then $|x^n| = |x|^n = x^n$.
7. $|-x| = |x|$, and so $|x - y| = |y - x|$
8. Applying absolute value twice (or more times) is the same as applying it once, meaning $||x|| = |x|$, which is called an idempotent property.

Theorem 5.3. We often have to "open up" absolute values in equations or inequalities and do casework. There are three basic scenarios, for a real variable x and a non-negative constant c :

1. $|x| = c$ if and only if $x = \pm c$
2. $|x| > c$ if and only if $x > c$ or $x < -c$
3. $|x| < c$ if and only if $-c < x < -c$

We leave the proofs to the reader. The analogous statements for negative c are trivial or vacuous, and we encourage the reader to explore them.

Theorem 5.4 (Real triangle inequality)**.** For any real numbers x and y,

$$|x| + |y| \geq |x + y|.$$

Equality holds if and only if both of x and y are non-negative or both are non-positive. Another way of stating the equality condition that is sometimes useful is that $x = 0$ or $y = 0$ or $x = cy$ for some real $c > 0$.

Proof. We will work backwards from the given inequality using reversible steps. Since both sides of $|x|+|y| \geq |x+y|$ are non-negative, it is a reversible step to square both sides, which yields

$$\begin{aligned}(|x|+|y|)^2 \geq |x+y|^2 &\iff |x|^2 + 2|x|\cdot|y| + |y|^2 \geq |x+y|^2 \\ &\iff x^2 + 2|xy| + y^2 \geq (x+y)^2 \\ &\iff x^2 + 2|xy| + y^2 \geq x^2 + 2xy + y^2 \\ &\iff |xy| \geq xy,\end{aligned}$$

which is true. Equality holds if and only if $|xy| = xy$, which is true if and only if xy is non-negative. And xy is non-negative if and only if x and y are both non-negative or both non-positive. For the other way of stating the equality condition, if $x = 0$ or $y = 0$ then equality clearly holds. If neither x nor y is 0 then $\operatorname{sgn}(x) = \operatorname{sgn}(y)$, where the signum function is defined in Definition 5.7. Then we can simply define $c = \dfrac{x}{y}$ which must be positive. ■

Example 5.5 (Reverse triangle inequality)**.** Prove that, for all real x and y,

$$||x| - |y|| \leq |x - y|.$$

Solution. We could square the inequality as in the proof of the Triangle Inequality, and, in fact, this would make it easier to derive the equality criterion. However, since we are not asked to find the equality condition, let us practice using the triangle inequality instead. We get the two inequalities

$$\begin{aligned}|x| &\leq |x-y| + |y|, \\ |y| &\leq |y-x| + |x|.\end{aligned}$$

These are equivalent to

$$\begin{aligned}|x| - |y| &\leq |x-y|, \\ |y| - |x| &\leq |x-y|.\end{aligned}$$

Since $||x| - |y|| = \max(|x| - |y|, |y| - |x|)$, the desired inequality follows. ■

Problem 5.6. Prove that, for any real numbers $x_1, x_2, \ldots, x_n$,

$$|x_1| + |x_2| + \cdots + |x_n| \geq |x_1 + x_2 + \cdots + x_n|.$$

Also prove that equality holds if and only if all the x_k are non-negative or all of them are non-positive.

Definition 5.7. The **signum function** gives the sign of a real number, meaning

$$\operatorname{sgn}(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases}.$$

Theorem 5.8. The following properties hold for the signum function for real x and y :

1. $\operatorname{sgn}(-x) = -\operatorname{sgn}(x)$
2. $\operatorname{sgn}(x) = \dfrac{x}{|x|} = \dfrac{|x|}{x}$ for $x \neq 0$
3. $\operatorname{sgn}(xy) = \operatorname{sgn}(x) \cdot \operatorname{sgn}(y)$
4. $\operatorname{sgn}\left(\dfrac{y}{x}\right) = \operatorname{sgn}(xy)$ if $x \neq 0$

The proofs are straightforward so we leave them to the reader. The second property is proven by casework, and the third and fourth properties follow from it.

5.2 Rounding

When working with a non-integer real number, we sometimes want to work with a nearby integer instead. This leads to several notions.

Definition 5.9. The **floor** function, **ceiling** function, and **fractional part** of a real number are defined respectively as follows:

$$\begin{aligned} \lfloor x \rfloor &= \max\{m \in \mathbb{Z} : m \leq x\}, \\ \lceil x \rceil &= \min\{n \in \mathbb{Z} : n \geq x\}, \\ \{x\} &= x - \lfloor x \rfloor. \end{aligned}$$

Example. Let's compute the floor, ceiling, and fractional part of -2.6. It is clear that $\lfloor -2.6 \rfloor = -3$ and $\lceil -2.6 \rceil = -2$. Perhaps counter to intuition,

$$\{-2.6\} = -2.6 - \lfloor -2.6 \rfloor = -2.6 + 3 = 0.4,$$

which is neither -0.6 nor 0.6.

Problem 5.10. Prove that, for any integer n,

$$\left\lfloor \frac{n}{2} \right\rfloor \cdot \left\lceil \frac{n}{2} \right\rceil = \left\lfloor \frac{n^2}{4} \right\rfloor.$$

Lemma 5.11. For each real number x, there exists exactly one integer in each of the intervals $[x, x+1)$ and $(x-1, x]$.

Proof. We will show the proof for $[x, x+1)$ and leave the proof for $(x-1, x]$ to the reader as the latter is analogous to the former. Suppose for contradiction that there is no integer in the interval $[x, x+1)$. Let n be the greatest integer that is less than x. Then $n < x$ and, since there is no integer in $[x, x+1)$, the next integer is $n+1 \geq x+1$. But this means $n \geq x$,

which contradicts $n < x$. So there must exist an integer in $[x, x+1)$. Now suppose there are two integers n and m in $[x, x+1)$. Then we have the inequalities

$$x \leq n < x+1,$$
$$x \leq m < x+1.$$

The negation of the latter is $-x-1 < -m \leq -x$ and adding this to the first inequality yields

$$-1 < n-m < 1.$$

But $n-m$ is an integer and the only integer that is strictly between -1 and 1 is 0. Thus $n = m$ and so there is a unique integer in $[x, x+1)$. ■

Theorem 5.12. Let m and n be integers and x be a real number. Then the following two sequences of equivalences hold:

$$\lfloor x \rfloor = m \iff x-1 < m \leq x \iff m \leq x < m+1,$$
$$\lceil x \rceil = n \iff x \leq n < x+1 \iff n-1 < x \leq n.$$

Proof. Combining the original definitions of the floor and ceiling functions with Lemma 5.11 allows us to redefine $\lfloor x \rfloor$ as the unique integer in the interval $(x-1, x]$ and $\lceil x \rceil$ as the unique integer in the interval $[x, x+1)$. The first equivalence of each sequence follows from this, and the second equivalence of each sequence is a simple matter of rewriting the inequalities. ■

Corollary 5.13. Using Theorem 5.12, we can present the floor and ceiling functions in a piecewise manner via infinitely many cases:

$$\lfloor x \rfloor = \begin{cases} \vdots & \\ -2 & \text{if } -2 \leq x < -1 \\ -1 & \text{if } -1 \leq x < 0 \\ 0 & \text{if } 0 \leq x < 1 \\ 1 & \text{if } 1 \leq x < 2 \\ 2 & \text{if } 2 \leq x < 3 \\ \vdots & \end{cases}, \quad \lceil x \rceil = \begin{cases} \vdots & \\ -2 & \text{if } -3 < x \leq -2 \\ -1 & \text{if } -2 < x \leq -1 \\ 0 & \text{if } -1 < x \leq 0 \\ 1 & \text{if } 0 < x \leq 1 \\ 2 & \text{if } 1 < x \leq 2 \\ \vdots & \end{cases}.$$

Problem 5.14. Prove that

$$\lfloor x \rfloor = -\lceil -x \rceil$$

for all real x. In terms of graphing on the Cartesian plane, this means that applying a reflection across the x-axis and across the y-axis allows us to switch between the floor function and ceiling function.

Corollary 5.15. For any real number x, it holds that

$$0 \leq \{x\} < 1.$$

This follows from combining the definition $\{x\} = x - \lfloor x \rfloor$ with the fact that, if $\lfloor x \rfloor = m$, then $m \leq x < m+1$ or $0 \leq x - m < 1$ (Theorem 5.12). For those familiar with modular arithmetic (this will be the theme of Volume 3), it should be evident that taking the fractional part of a real number is an analogous "reduction modulo 1" function for real numbers.

Theorem 5.16. The following equivalences among inequalities hold. They are interesting and applicable, as they allow us to switch a variable with its ceiling or floor when being compared to an integer. If x is a real number and p is an integer, then:

$$\begin{aligned} x < p &\iff \lfloor x \rfloor < p, \\ p \leq x &\iff p \leq \lfloor x \rfloor, \\ p < x &\iff p < \lceil x \rceil, \\ x \leq p &\iff \lceil x \rceil \leq p. \end{aligned}$$

Proof. One direction of each biconditional statement is trivial and the other direction makes the inequality sharper. We will prove the first two equivalences and we leave the remaining two as exercises to the reader, as they require arguments similar to those used to prove the first two.
Since $\lfloor x \rfloor \leq x$, it follows immediately that $x < p$ implies $\lfloor x \rfloor < p$. Conversely, suppose $\lfloor x \rfloor < p$. Since $\lfloor x \rfloor$ and p are both integers, Theorem 4.18 tells us that $\lfloor x \rfloor \leq p - 1$ or $\lfloor x \rfloor + 1 \leq p$. If $\lfloor x \rfloor = m$ then we know that $m \leq x < m + 1 \leq p$. Thus, $x < p$.
For the second equivalence, it is immediately true that $p \leq \lfloor x \rfloor$ implies $p \leq x$. Now suppose $p \leq x$. Then p is an element of the set $\{m \in \mathbb{Z} : m \leq x\}$ Since $\lfloor x \rfloor$ is the maximal element of this set, it must be true that $p \leq \lfloor x \rfloor$, as desired. ■

Problem 5.17. Show that the following identities hold for integers $n > 0$ and m :

$$\left\lfloor \frac{m}{n} \right\rfloor = \left\lceil \frac{m-n+1}{n} \right\rceil \text{ and } \left\lceil \frac{m}{n} \right\rceil = \left\lfloor \frac{m+n-1}{n} \right\rfloor.$$

They are interesting as they allow us to change between floor and ceiling functions for rational inputs.

Theorem 5.18. There is a useful technique of extracting integers out of the floor and ceiling functions and fractional part. If x is a real number and p is an integer, then:

$$\begin{aligned} \lfloor x + p \rfloor &= \lfloor x \rfloor + p, \\ \lceil x + p \rceil &= \lceil x \rceil + p, \\ \{x + p\} &= \{x\}. \end{aligned}$$

Proof. We will not prove the identity for the ceiling function since it is analogous to the proof for the floor function. For the floor function, let $\lfloor x+p \rfloor = m$. Then $m \leq x+p < m+1$ or

$$m - p \leq x < m - p + 1.$$

This means $\lfloor x \rfloor = m - p$, which is equivalent to the desired identity. Using this identity allows us to show that

$$\begin{aligned} \{x+p\} &= x + p - \lfloor x + p \rfloor \\ &= x + p - \lfloor x \rfloor - p \\ &= x - \lfloor x \rfloor \\ &= \{x\}. \end{aligned}$$

■

Example 5.19. The usual rounding function takes 6.2 to 6, and 6.7 to 7, and 6.5 to 7. Show that the rounding function may be encapsulated by the definition

$$\operatorname{round}(x) = \left\lfloor x + \frac{1}{2} \right\rfloor.$$

Solution. Indeed, we find that

$$\begin{aligned} \left\lfloor x + \frac{1}{2} \right\rfloor &= \left\lfloor \lfloor x \rfloor + \{x\} + \frac{1}{2} \right\rfloor \\ &= \lfloor x \rfloor + \left\lfloor \{x\} + \frac{1}{2} \right\rfloor \\ &= \begin{cases} \lfloor x \rfloor & \text{if } 0 \le \{x\} < 0.5 \\ \lfloor x \rfloor + 1 & \text{if } 0.5 \le \{x\} < 1 \end{cases}, \end{aligned}$$

which matches our intuition for rounding. ■

Theorem 5.20 (Hermite's identity)**.** For all real x and positive integers m,

$$\lfloor x \rfloor + \left\lfloor x + \frac{1}{m} \right\rfloor + \left\lfloor x + \frac{2}{m} \right\rfloor + \cdots + \left\lfloor x + \frac{m-1}{m} \right\rfloor = \lfloor mx \rfloor.$$

In particular, we can take $x = \dfrac{n}{m}$ for an integer n to produce an interesting identity.

Proof. Fix a positive integer m and let

$$f(x) = \lfloor x \rfloor + \left\lfloor x + \frac{1}{m} \right\rfloor + \cdots + \left\lfloor x + \frac{m-2}{m} \right\rfloor + \left\lfloor x + \frac{m-1}{m} \right\rfloor - \lfloor mx \rfloor.$$

We want to show that f is identically 0. Miraculously, we find that f is periodic (periodicity is defined in Definition 7.1) because

$$\begin{aligned} f\left(x + \frac{1}{m}\right) &= \left\lfloor x + \frac{1}{m} \right\rfloor + \left\lfloor x + \frac{2}{m} \right\rfloor + \cdots + \left\lfloor x + \frac{m-1}{m} \right\rfloor + \lfloor x + 1 \rfloor - \lfloor mx + 1 \rfloor \\ &= \left\lfloor x + \frac{1}{m} \right\rfloor + \left\lfloor x + \frac{2}{m} \right\rfloor + \cdots + \left\lfloor x + \frac{m-1}{m} \right\rfloor + \lfloor x \rfloor - \lfloor mx \rfloor \\ &= f(x). \end{aligned}$$

So it suffices to show that f is identically 0 on the interval $\left[0, \frac{1}{m}\right)$. For x in this interval, it turns out that every term in the expression for f vanishes because of the following reasoning. It holds that $0 \leq x < \frac{1}{m}$ which implies

$$0 \leq \frac{k}{m} \leq x + \frac{k}{m} < \frac{k+1}{m} \leq 1.$$

This means $0 \leq x + \frac{k}{m} < 1$ or $\left\lfloor x + \frac{k}{m} \right\rfloor = 0$ for $k = 0, 1, \ldots, m-1$. Moreover, if $1 \leq x < \frac{1}{m}$ then $0 \leq mx < 1$ which tells us that $\lfloor mx \rfloor = 0$. This completes the proof.
Note that the identity

$$\lceil x \rceil + \left\lceil x - \frac{1}{m} \right\rceil + \left\lceil x - \frac{2}{m} \right\rceil + \cdots + \left\lceil x - \frac{m-1}{m} \right\rceil = \lceil mx \rceil$$

for ceiling functions can be derived by applying Hermite's Identity to $-x$ and using the fact that $\lceil x \rceil = -\lfloor -x \rfloor$ (Problem 5.14). We did not include it in the statement of the theorem because it is neither widely cited nor particularly applicable. ■

Theorem 5.21. For any real x and y, the following inequalities hold:

$$\lfloor x \rfloor + \lfloor y \rfloor \leq \lfloor x + y \rfloor \leq \lfloor x \rfloor + \lfloor y \rfloor + 1,$$
$$\lceil x \rceil + \lceil y \rceil - 1 \leq \lceil x + y \rceil \leq \lceil x \rceil + \lceil y \rceil.$$

Proof. We will show that the floor function inequalities hold. The ceiling function inequalities follow from applying them to $-x$ and $-y$ because

$$\lceil x \rceil = -\lfloor -x \rfloor,$$

by Problem 5.14. Note that

$$\lfloor x + y \rfloor = \lfloor \lfloor x \rfloor + \{x\} + \lfloor y \rfloor + \{y\} \rfloor = \lfloor x \rfloor + \lfloor y \rfloor + \lfloor \{x\} + \{y\} \rfloor.$$

So it suffices to prove

$$0 \leq \lfloor \{x\} + \{y\} \rfloor \leq 1.$$

Since $\lfloor \{x\} + \{y\} \rfloor$ is an integer, it is equivalent to show that $\lfloor \{x\} + \{y\} \rfloor$ is either 0 or 1. Since $0 \leq \{x\} < 1$ and $0 \leq \{y\} < 1$, adding them yields $0 \leq \{x\} + \{y\} < 2$. Indeed, the possibilities then are

$$\lfloor \{x\} + \{y\} \rfloor = \begin{cases} 0 & \text{if } 0 \leq \{x\} + \{y\} < 1, \\ 1 & \text{if } 1 \leq \{x\} + \{y\} < 2. \end{cases}$$

■

Corollary 5.22. The floor and ceiling functions are non-decreasing, meaning if x and y are real numbers such that $x < y$ then

$$\lfloor x \rfloor \leq \lfloor y \rfloor \text{ and } \lceil x \rceil \leq \lceil y \rceil.$$

Proof. Let $k > 0$ be defined as $y = x + k$. Since $k > 0$ means $\lfloor k \rfloor \geq 0$, the floor function inequality in Theorem 5.21 tells us that

$$\lfloor x \rfloor \leq \lfloor x \rfloor + \lfloor k \rfloor \leq \lfloor x + k \rfloor = \lfloor y \rfloor.$$

Similarly, since $k > 0$ means $\lceil k \rceil \geq 1$, the ceiling function inequality in Theorem 5.21 tells us that

$$\lceil x \rceil \leq \lceil x \rceil + \lceil k \rceil - 1 \leq \lceil x + k \rceil = \lceil y \rceil.$$

■

We end the section with a general theorem. It has several ramifications, as we will see in Corollary 5.24. Notice that the floor and ceiling functions are idempotent, meaning

$$\lfloor \lfloor x \rfloor \rfloor = \lfloor x \rfloor \text{ and } \lceil \lceil x \rceil \rceil = \lceil x \rceil.$$

We can ask whether we can insert a function in between the two applications of the floor or ceiling function and still retain equality.

Theorem 5.23 (McEliece's theorem)**.** Let f be a continuous and monotonically increasing function, defined on some interval $I \subseteq \mathbb{R}$, with the property that

$$f(x) \in \mathbb{Z} \implies x \in \mathbb{Z}.$$

Then, for all $x \in I$,

$$\lfloor f(x) \rfloor = \lfloor f(\lfloor x \rfloor) \rfloor \text{ and } \lceil f(x) \rceil = \lceil f(\lceil x \rceil) \rceil .$$

Of course, each identity holds only if the expression on either side of it is defined, as x being in I does not necessarily mean that $\lfloor x \rfloor$ and $\lceil x \rceil$ are in I.

Proof. We will prove the theorem for floor functions and leave the analogous argument for ceiling functions as an exercise to the reader. The reader may also find the proof for ceiling functions in [2].
So we want to prove that $\lfloor f(x) \rfloor = \lfloor f(\lfloor x \rfloor) \rfloor$. Recall that $\lfloor x \rfloor \leq x$. If $\lfloor x \rfloor = x$ then the identity immediately holds. Now suppose $\lfloor x \rfloor < x$, implying x is not an integer. By the fact that f is monotonically increasing,

$$f(\lfloor x \rfloor) < f(x).$$

Since the floor function is non-decreasing, we get

$$\lfloor f(\lfloor x \rfloor) \rfloor \leq \lfloor f(x) \rfloor.$$

If equality holds in this inequality, then we are done. So suppose for contradiction that

$$\lfloor f(\lfloor x \rfloor) \rfloor < \lfloor f(x) \rfloor.$$

Then $\lfloor f(\lfloor x \rfloor) \rfloor < \lfloor f(x) \rfloor \leq f(x)$. Now there are two possibilities, both of which will lead to the same conclusion. One possibility is that $\lfloor f(x) \rfloor = f(x)$. The other possibility is

that $\lfloor f(\lfloor x \rfloor) \rfloor < \lfloor f(x) \rfloor < f(x)$. In the latter case, since $\lfloor f(x) \rfloor$ is an integer, we can use Theorem 5.16 to take off the floor function on the left side and get

$$f(\lfloor x \rfloor) < \lfloor f(x) \rfloor < f(x).$$

In this case, we can invoke continuity of f to use the intermediate value theorem, which tells us that there exists $y \in (\lfloor x \rfloor, x)$ such that $f(y) = \lfloor f(x) \rfloor$. In either case, $\lfloor f(x) \rfloor = f(z)$ for some z. Since $f(z) = \lfloor f(x) \rfloor$ is an integer, z must be an integer by the special property of f. But $\lfloor x \rfloor < z \leq x$, so z cannot be an integer since x is assumed to not be an integer and $\lfloor x \rfloor$ is the maximum integer that is less than or equal to x. So $\lfloor f(\lfloor x \rfloor) \rfloor < \lfloor f(x) \rfloor$ leads to a contradiction, forcing it to be the case that $\lfloor f(\lfloor x \rfloor) \rfloor = \lfloor f(x) \rfloor$. ■

Corollary 5.24. There are several immediate examples of continuous and monotonically increasing functions f that are defined on an interval $I \subseteq \mathbb{R}$ and satisfy the property

$$f(x) \in \mathbb{Z} \implies x \in \mathbb{Z}.$$

They are:

- $f(x) = \sqrt[n]{x}$ for integers $n \geq 1$, where the domain is $[0, \infty)$ for even n and $\mathbb{R}$ for odd n
- $f(x) = \log_b x$ for integer bases $b \geq 2$, where the domain is $[1, \infty)$
- $f(x) = \dfrac{x+m}{n}$ for integers $n > 0$ and m, where the domain is $\mathbb{R}$

  In particular, if we choose $m = 0$ and a sequence of positive integers $n_1, n_2, \ldots, n_k$ and replace x with $\dfrac{x}{n_1}$, then repeated application of McEliece's theorem yields the nested division formulas

$$\lfloor \ldots \lfloor \lfloor x/n_1 \rfloor / n_2 \rfloor \ldots / n_k \rfloor = \left\lfloor \frac{x}{n_1 n_2 \cdots n_k} \right\rfloor,$$
$$\lceil \ldots \lceil \lceil x/n_1 \rceil / n_2 \rceil \ldots / n_k \rceil = \left\lceil \frac{x}{n_1 n_2 \cdots n_k} \right\rceil.$$

Chapter 6

Closed Forms

> "There are actually formulas in the literature ("nameless here for evermore") for certain counting functions $f(n)$ whose evaluation requires listing all (or almost all) of the $f(n)$ objects being counted! Such a "formula" is completely worthless."
>
> – *Richard Stanley, Enumerative Combinatorics I*

Having studied sequences, sums, and products in general in Chapter 3, we will now take a look at particular instances of them that are amenable to being expressed in closed form. First we will look at arithmetic and geometric sequences and series. Then we will move on to studying a curious technique called telescoping that is surprisingly applicable.

6.1 Arithmetic and Geometric

Given a finite series or product, it is desirable to find a "closed formula" for it, and similarly for the partial sums or products of an infinite series or product. How a closed formula can be found, if at all possible, depends on the definition of the indexed object. We will find formulas for two of the most common sequences and series, namely arithmetic and geometric sequence and series. There is no universally accepted definition of a closed formula, but we can encapsulate the sentiment informally as follows.

Definition 6.1. A formula is said to be **closed** or be in **closed form** if the number of operations in it is constant or has a constant upper bound across all values of all variables.

Example. The formula $\underbrace{p+p+\cdots+p}_{n \text{ copies of } p}$ is not closed because there are $n-1$ operations, which grows without an upper bound as n goes to infinity. On the other hand, pn is considered to be a closed expression for the same quantity. What exactly is meant by "the number of operations" is ambiguous in Definition 6.1. For example, multiplying two very large integers takes a computer algorithm more steps than multiplying two small integers, but both computations could be considered to be one multiplication operation. Moreover, many "closed" formulas in higher mathematics involve integrals or use functions, such as the Gamma function Γ or the Riemann zeta function ζ, even though it might be difficult to evaluate the outputs of these functions at given inputs.

Definition 6.2. An **arithmetic sequence** with **initial term** a and **common difference** d is

$$(a+(n-1)d)_{n=1}^{\infty} = (a, a+d, a+2d, \ldots).$$

For each positive integer n, the sum of the first n terms of an arithmetic sequence is called an **arithmetic series**.

Theorem 6.3. If the initial term of an arithmetic sequence $(a_n)_{n=1}^{\infty}$ is a and common difference is d, then the resulting arithmetic series with n terms evaluates to

$$a_1 + a_2 + \cdots + a_n = \left(\frac{a + a_n}{2}\right) \cdot n = \frac{(2a + (n-1)d)n}{2}$$

for each positive integer n. The first formula is more memorable because the second lacks symmetry.

Proof. The sum can be written as

$$\sum_{k=1}^{n} (a + (k-1)d) = an + d \cdot \sum_{k=1}^{n} (k-1).$$

So it is a matter of finding the sum

$$1 + 2 + \cdots + (n-1).$$

The first idea is one of which the reader should take careful note because it can be useful elsewhere: set the desired quantity equal to a variable. The second idea is to "reflect" the sum by writing it in reverse order. Let

$$S = \sum_{k=1}^{n-1} k = \sum_{k=1}^{n-1} (n-k).$$

Adding the two yields

$$2S = \sum_{k=1}^{n-1} k + \sum_{k=1}^{n-1} (n-k) = \sum_{k=1}^{n-1} n = (n-1)n.$$

So the sum of the first $n-1$ positive integers is $S = \dfrac{(n-1)n}{2}$. This allows to conclude that

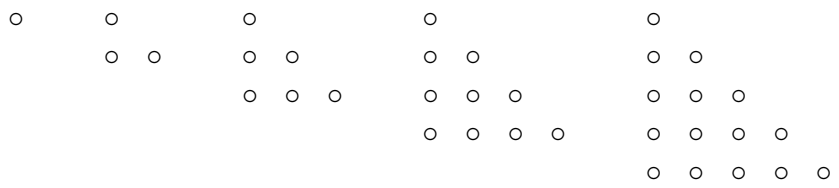
$$\sum_{k=1}^{n} (a + (k-1)d) = an + d \cdot \frac{(n-1)n}{2} = \frac{(2a + (n-1)d)n}{2} = \frac{(a + a_n)n}{2}.$$

According to legend, Gauss was told by his teacher to evaluate the sum $1+2+3+\cdots+100$, and the teacher was astonished when he quickly produced the answer by using this reflection method. For this reason, the method is sometimes called **Gauss's trick**. ■

Definition 6.4. For each positive integer n, the n^{th} **triangular number** is

$$1 + 2 + \cdots + n = \frac{n(n+1)}{2}.$$

The name is derived from the fact that $1 + 2 + \cdots + n$ dots can be stacked in a triangular array as shown below.



Problem 6.5. For each positive integer n, find the sum of the first n odd positive integers

$$1 + 3 + 5 + \cdots + (2n - 1).$$

Definition 6.6. A **geometric sequence** with **initial term** b and **common ratio** r is

$$(br^{n-1})_{n=1}^{\infty} = (b, br, br^2, \ldots).$$

For each positive integer n, the sum of the first n terms of a geometric sequence is called a **geometric series**. An **infinite geometric series** is the sum of all terms in a geometric sequence. This does not always converge to a number.

Theorem 6.7. If the initial term of a geometric sequence $(b_n)_{n=1}^{\infty}$ is b and common ratio is r, then the resulting geometric series with n terms evaluates to

$$b_1 + b_2 + \cdots + b_n = \begin{cases} bn & \text{if } r = 1 \\ \dfrac{b(r^n - 1)}{(r - 1)} & \text{if } r \neq 1 \end{cases}$$

for each positive integer n. If $b = 0$ then the infinite geometric series is 0, so suppose $b \neq 0$. Then the infinite geometric series does not converge to a number if $|r| \geq 1$. However, if $|r| < 1$, then this infinite sum converges to

$$\sum_{k=0}^{\infty} br^k = \frac{b}{1 - r}$$

Proof. First we tackle the finite case. The formula for $r = 1$ is obviously true, so we assume that $r \neq 1$. As with an arithmetic series, we set the desired quantity equal to a variable. We let

$$S = b + br + br^2 + \cdots + br^{n-1}.$$

Then we create another sum that has many of the same terms, so that we can take the difference of the two. Multiplying the first equation by r yields

$$rS = br + br^2 + br^3 + \cdots + br^n.$$

Subtracting the first equation from the second gives

$$\begin{aligned} (r - 1)S &= rS - S \\ &= (br + br^2 + br^3 + \cdots + br^n) - (b + br + br^2 + \cdots + br^{n-1}) \\ &= br^n - b = b(r^n - 1). \end{aligned}$$

By isolating S, we get the desired formula

$$S = \frac{b(r^n - 1)}{r - 1}.$$

Now we will work on the infinite case with $b \neq 0$. The following are the cases of $|r| \geq 1$.

- If $r = 1$, then the series has partial sums bn. If $b > 0$, then bn grows without an upper bound as n goes to infinity; if $b < 0$, then bn goes to $-\infty$ as n goes to infinity.
- If $r = -1$, then the partial sums oscillate between b and $-b$, which also does not converge.
- If $r > 1$, then the partial sum $\dfrac{b(r^n - 1)}{r - 1}$ has a constant denominator and a numerator that grows arbitrarily large as n goes to infinity.
- If $r < -1$, then the partial sum $\dfrac{b(r^n - 1)}{r - 1}$ again has a constant denominator, and a numerator that grows arbitrarily large for even integers n as n goes to infinity.

So there is no convergence for $|r| \geq 1$. Finally, we consider the "Goldilocks zone" $|r| < 1$, which is equivalent to $-1 < r < 1$. Here, we rewrite the partial sum as

$$\frac{b(r^n - 1)}{r - 1} = \frac{br^n}{r - 1} + \frac{b}{1 - r}.$$

As n goes to infinity, the first term goes to 0 because its numerator gets arbitrarily close to 0 while the denominator is constant; the second term is constant. This means the partial sums converge to $\dfrac{b}{1 - r}$.
Note that making these arguments precise would require defining convergence and divergence, which would take us to the realm of limits from calculus. ■

Problem 6.8. Let p and $r \neq 1$ be real constants and n be a positive integer. Evaluate

$$\underbrace{(\cdots((pr + p)r + p)r \cdots + p)r + p}_{\text{number of multiplications by } r \text{ is } n-1}.$$

Example 6.9. Prove the following facts about arithmetic and geometric sequences.

1. Let $(a_n)_{n=1}^{\infty}$ be an arithmetic sequence. For any distinct indices i and j, the common difference is $\dfrac{a_i - a_j}{i - j}$. Also, for any index $n > 1$, a_n is the arithmetic mean of its neighbours, meaning
$$a_n = \frac{a_{n-1} + a_{n+1}}{2}.$$
2. Let $(b_n)_{n=1}^{\infty}$ be a geometric sequence. If the initial term and common ratio r are non-zero, then, for any distinct indices i and j, $\dfrac{b_i}{b_j} = r^{i-j}$; in some cases, such as if r is

known to be positive or $i - j$ is an odd integer, it is possible to isolate for r. Also, for any index $n > 1$, if b_n is known to be positive (for example, if the initial term and common ratio are both positive), then b_n is the geometric mean of its neighbours, meaning

$$b_n = \sqrt{b_{n-1}b_{n+1}}.$$

Solution. These are all a matter of straightforward computation by unwrapping definitions.

1. Let the initial term of the arithmetic sequence be a and the common difference be d. Then
$$\frac{a_i - a_j}{i - j} = \frac{(a + (i-1)d) - (a + (j-1)d)}{i - j} = d.$$
For the second result,
$$\frac{a_{n-1} + a_{n+1}}{2} = \frac{(a + (n-1-1)d) + (a + (n+1-1)d)}{2} = a + (n-1)d = a_n.$$

2. Let the initial term of the geometric sequence be $b \neq 0$ and the common ratio be $r \neq 0$. Then
$$\frac{b_i}{b_j} = \frac{br^{i-1}}{br^{j-1}} = r^{(i-1)-(j-1)} = r^{i-j}.$$
For the second result,
$$b_{n-1}b_{n+1} = (br^{n-1-1})(br^{n+1-1}) = (br^{n-1})^2 = b_n^2.$$
If we know that b_n is positive, then we can take the positive square root of both sides to get
$$b_n = \sqrt{b_{n-1}b_{n+1}}.$$

■

Definition 6.10. An **arithmetico-geometric sequence** results from multiplying each term of an arithmetic sequence with the corresponding term of a geometric sequence. So if the arithmetic sequence has initial term a and common difference d, and if the geometric sequence has initial term b and common ratio r, then the arithmetico-geometric sequence is

$$((a + (n-1)d) \cdot br^{n-1})_{n=1}^{\infty}.$$

An **arithmetico-geometric series** is defined as one would expect.

Theorem 6.11. Using the same technique that is used to find the sum of a geometric series, an arithmetico-geometric series can be evaluated as

$$\sum_{k=1}^{n} (a + (k-1)d) \cdot br^{k-1} = \frac{ab - (a + nd)br^n}{1 - r} + \frac{dbr(1 - r^n)}{(1 - r)^2},$$

assuming $r \neq 1$. If $r = 1$, then this is simply an arithmetic sequence.

Proof. Suppose $r \neq 1$. We want to find a closed form for

$$S = ab + (a+d)br + (a+2d)br^2 + \cdots + [a+(n-1)d] \cdot br^{n-1}.$$

Multiplying both sides of this equation by the common ratio r of the associated geometric sequence yields

$$Sr = abr + (a+d)br^2 + \cdots + [a+(n-2)d] \cdot br^{n-1} + [a+(n-1)d] \cdot br^n.$$

Subtracting the latter equation from the former yields

$$\begin{aligned} S(1-r) &= S - Sr \\ &= ab + (dbr + dbr^2 + \cdots + dbr^{n-1}) - [a+(n-1)d] \cdot br^n \\ &= ab + (dbr + dbr^2 + \cdots + dbr^{n-1} + dbr^n) - (abr^n + ndbr^n) \\ &= ab - (a+nd)br^n + \frac{dbr(1-r^n)}{1-r}, \end{aligned}$$

where we used the formula for a geometric series in the last step. Dividing both sides by $1-r$ gives the final formula. It is not necessary to memorize this complicated formula, considering the relatively low frequency with which arithmetico-geometric series appear in problems. We have derived it only to showcase the technique, which can be replicated easily. ■

Theorem 6.12. If $|r| < 1$, then

$$\sum_{k=1}^{\infty} (a+(k-1)d) \cdot br^{k-1} = \frac{ab}{1-r} + \frac{dbr}{(1-r)^2}.$$

We could classify all cases of convergence like with geometric series but the casework is not worth the effort in this context.

Proof. The infinite sum is

$$\begin{aligned} \sum_{k=1}^{\infty} (a+(k-1)d) \cdot br^{k-1} &= \lim_{n\to\infty} \sum_{k=1}^{n} (a+(k-1)d) \cdot br^{k-1} \\ &= \lim_{n\to\infty} \left[\frac{ab - (a+nd)br^n}{1-r} + \frac{dbr(1-r^n)}{(1-r)^2} \right] \\ &= \frac{ab}{1-r} + \frac{dbr}{(1-r)^2} - \lim_{n\to\infty} \frac{(a+nd)br^n}{1-r} - \lim_{n\to\infty} \frac{dbr^{n+1}}{(1-r)^2}. \end{aligned}$$

In the second limit, the denominator is constant and the numerator goes to 0 as n goes to infinity. In the first limit, the denominator is again 0, but the numerator is a bit trickier. Without getting into the formal definition of limits, the crux of the matter is that, while $a+nd$ goes to infinity linearly as n goes to infinity, br^n goes to 0 at an exponential rate as n goes to infinity. So the former is overtaken by the latter and the whole term disappears. Thus, we are left with only the two terms outside the limits. ■

6.2 Telescoping

Definition 6.13. A series is said to **telescope** if it can be written in a way that many terms cancel each other out. Just like sums, it is possible for products to telescope. The name is derived from how it is possible to collapse some telescopes so that intermediate components disappear from external view.

Example. The product

$$\frac{1}{2} \cdot \frac{2}{3} \cdot \frac{3}{4} \cdots \frac{n-1}{n} \cdot \frac{n}{n+1} = \frac{1}{n+1}$$

telescopes because each denominator cancels with the subsequent numerator, leaving us with only the leftmost numerator and rightmost denominator in the end.

Problem 6.14. Prove that, for any positive integer n,

$$1 \cdot 1! + 2 \cdot 2! + 3 \cdot 3! + \cdots + n \cdot n! = (n+1)! - 1.$$

A telescoping series sometimes goes hand-in-hand with a technique called partial fraction decomposition.

Definition 6.15. Without defining **partial fraction decomposition** in great detail, the general idea is that, if we have an expression of the form $\dfrac{f(x)}{g_1(x)g_2(x)\cdots g_n(x)}$ for functions f and $g_1, g_2, \ldots, g_n$, then it could be useful to decompose it as

$$\frac{f(x)}{g_1(x)g_2(x)\cdots g_n(x)} = \frac{f_1(x)}{g_1(x)} + \frac{f_2(x)}{g_2(x)} + \cdots + \frac{f_n(x)}{g_n(x)}$$

for suitable functions $f_1, f_2, \ldots, f_n$.

Of course, the question is how the f_i can be found. There exists a general theorem for rational functions (see Definition 10.35) that is useful in the integration of such functions in calculus. For our purposes, it will suffice to deal only with distinct linear functions $g_i(x) = a_i x + b_i$ and the constant function $f(x) = 1$. In such cases, we will commit three mathematical sins to find the f_i :

- We will assume that the f_i exist.
- We will assume that the f_i are equal to constants c_i.
- We will clear the denominators and solve for the f_i by substituting in "forbidden" values such as $-\dfrac{b_i}{a_i}$ for x that would have originally caused

  $$g_1(x)g_2(x)\cdots g_n(x) = 0,$$

  which might as well be fraternizing with division by 0.

Why is all this acceptable? Technically, it has to do with the fundamental theorem of algebra (Theorem 10.28), for which we are not yet ready. However, as long as we find f_i that work in the end, the process can be swept under the rug. Let us see a classic example of a series that can be made to telescope using partial fraction decomposition.

Example 6.16. If n is a positive integer, then ind a closed expression for

$$\frac{1}{1\cdot 2}+\frac{1}{2\cdot 3}+\frac{1}{3\cdot 4}+\cdots+\frac{1}{n(n+1)}$$

in terms of n.

Solution. Each term is of the form $\dfrac{1}{k(k+1)}$. We start by assuming that there exist constants A and B such that

$$\frac{A}{k}+\frac{B}{k+1}=\frac{1}{k(k+1)}.$$

Clearing the denominators yields

$$A(k+1)+Bk=1.$$

If we want A to disappear, we can substitute $k=-1$ and get $B=-1$. If we want B to disappear, we can substitute $k=0$ and get $A=1$. Our hope that it is in fact true that

$$\frac{1}{k}-\frac{1}{k+1}=\frac{1}{k(k+1)},$$

which we can verify by adding the fractions on the left side. This means the series in the problem becomes

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

Many consecutive terms cancel each other out, leaving only the terms on the far left and far right. This yields an answer of $1-\dfrac{1}{n+1}=\dfrac{n}{n+1}$. ■

Problem 6.17. Find the sum of the infinite series

$$\frac{1}{1\cdot 2\cdot 3}+\frac{1}{2\cdot 3\cdot 4}+\frac{1}{3\cdot 4\cdot 5}+\cdots$$

by finding the partial sum of the first n terms and observing what happens as n goes to infinity.

Next we (re)turn to the problem of finding a closed form for the sum of the p^{th} powers of the first n positive integers:

$$1^p+2^p+\cdots+n^p.$$

Using Gauss's trick, we have already evaluated the sum for $p=1$ as $\dfrac{n(n+1)}{2}$ (see Theorem 6.3). Now we consider $p=2$ and $p=3$. For $p=2$, we can use telescoping as follows. This method can be generalized to recursively find formulas for all p using the binomial theorem. We will do so in Volume 2.

Theorem 6.18. For each positive integer n,

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

Proof. The motivation is to build a formula for this expression of squares through the lower instance of Gauss's formula. This could be possible by adding the instances of

$$(k+1)^3 - k^3 = 3k^2 + 3k + 1.$$

We evaluate the sum of this quantity from $k = 0$ to $k = n$, in two different ways. On the left side, we get a telescoping series

$$\sum_{k=0}^{n} [(k+1)^3 - k^3] = (n+1)^3.$$

Using the right side,

$$\begin{aligned}\sum_{k=0}^{n} (3k^2 + 3k + 1) &= 3 \cdot \sum_{k=0}^{n} k^2 + 3 \cdot \sum_{k=0}^{n} k + \sum_{k=0}^{n} 1 \\ &= 3 \cdot \sum_{k=1}^{n} k^2 + \frac{3n(n+1)}{2} + (n+1).\end{aligned}$$

Equating the two sums allows us to perform the isolation

$$\sum_{k=1}^{n} k^2 = \frac{1}{3} \cdot \left[(n+1)^3 - \frac{3n(n+1)}{2} - (n+1)\right] = \frac{n(n+1)(2n+1)}{6}.$$

■

Problem 6.19. Miraculously, it holds that

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

We invite the reader to prove this by induction and to take note that it is often possible to prove closed formulas by induction if one knows the destination formula. The reader might be interested in knowing that the author of this book proved the following proposition by an induction argument in a paper with Dr. Edward Barbeau [5]:

If $(a_1, a_2, \ldots, a_n)$ is a strictly increasing tuple of n positive integers, then

$$a_1^3 + a_2^2 + \cdots + a_n^3 \geq (a_1 + a_2 + \cdots + a_n)^2,$$

with equality holding if and only if $(a_1, a_2, \ldots, a_n) = (1, 2, \ldots, n)$.

Example 6.20. Show that, if x is a real number such that $|x| < 1$ then

$$\prod_{k=0}^{\infty} (1 + x^{2^k}) = \sum_{k=0}^{\infty} x^k.$$

Solution. By applications of the difference of squares factorization, we can prove by induction on positive integers n that the n^{th} partial product is equal to

$$\begin{aligned}\prod_{k=0}^{n}(1+x^{2^k}) &= (1+x)(1+x^2)(1+x^4)\cdots(1+x^{2^n}) \\ &= \frac{(1-x)(1+x)(1+x^2)(1+x^4)\cdots(1+x^{2^n})}{1-x} \\ &= \frac{1-x^{2^{n+1}}}{1-x} \\ &= \frac{1}{1-x} - \frac{x^{2^{n+1}}}{1-x}.\end{aligned}$$

As $n \to \infty$, the second term goes to 0 because the numerator $x^{2^{n+1}}$ gets arbitrarily close to 0, while the denominator $1-x$ remains constant. So

$$\prod_{k=0}^{\infty}(1+x^{2^k}) = \lim_{n\to\infty}\prod_{k=0}^{n}(1+x^{2^k}) = \frac{1}{1-x}.$$

Since $|x| < 1$, this is the formula for an infinite geometric whose initial term is 1 and common ratio is x. This was not an example of a telescoping product per se, but the way the terms collapsed into each other in the numerator of the partial product is reminiscent of telescoping. ■

Problem 6.21. Evaluate the infinite product $\displaystyle\prod_{k=2}^{\infty}\frac{k^3-1}{k^3+1}$. You should evaluate the finite version up to n multiplicands first and then take the limit as $n \to \infty$.

Chapter 7

Trigonometric Functions

"Young man, in mathematics you don't understand things. You just get used to them."

– *John von Neumann*

We will look at the six trigonometric functions here in terms of unit circle geometry. The trigonometric functions are ubiquitous in mathematics. We will start with a general study of periodic functions, and then specifically see properties of the the main trigonometric functions. Afterwards, we will take an exhaustive look at trigonometric identities.

7.1 Periodic Functions

We will begin with some general facts about periodic functions, since it will turn out that trigonometric functions are all periodic.

Definition 7.1. A non-constant function f is said to be **periodic** if its domain is a subset of $\mathbb{R}$ and there exists a positive real constant p such that

$$x \in \text{Dom}(f) \iff x + p \in \text{Dom}(f)$$

for all real x, and

$$f(x + p) = f(x)$$

for all $x \in \text{Dom}(f)$. Such a number p is called a **period** of f. If there is a smallest positive real constant p with this property, then it is called the **minimal period** of f. However, we often simply refer to the minimal period as "the" period, such as in the case of trigonometric functions.

Example. Not every periodic function has a minimal period. For example, the **Dirichlet function** $\delta : \mathbb{R} \to \mathbb{R}$, which is also called the **indicator function** of the rationals, is defined by

$$\delta(x) = \begin{cases} 1 & \text{if } x \in \mathbb{Q} \\ 0 & \text{if } x \notin \mathbb{Q} \end{cases}.$$

For every $r \in \mathbb{Q}$, $x + r$ is rational if x is rational and $x + r$ is irrational if x is irrational. Thus, for every rational r, $\delta(x + r) = \delta(x)$. This means that, for every period, a smaller period exists, since there exists a sequence of positive rationals $\frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots$ that get arbitrarily close to 0.

On the other hand, it is true that every continuous periodic function with domain $\mathbb{R}$ has a minimal period. We will not prove this here since we have not provided a rigorous definition of continuity.

Theorem 7.2. If f is a periodic function with a period p, then $x + pn \in \text{Dom}(f)$ and $f(x + pn) = f(x)$ for all $x \in \text{Dom}(f)$ and all integers n.

Proof. We will prove this by induction on $n \geq 0$ first, and then we will take care of $n \leq 0$ by a separate induction. The result is true for $n = 0$ by the definition of a function. Now suppose there exists an $n \geq 0$ such that $x + pn \in \text{Dom}(f)$ and $f(x + pn) = f(x)$ for all $x \in \text{Dom}(f)$. By the definition of a periodic function, $x + pn + p = x + p(n + 1)$ is in $\text{Dom}(f)$ and

$$f(x + p(n + 1)) = f((x + pn) + p) = f(x + pn) = f(x)$$

for all $x \in \text{Dom}(f)$. This completes the first induction.
Now we will handle $n \leq 0$. Again, the base case is trivial. Now suppose there exists an $n \leq 0$ such that $x + pn \in \text{Dom}(f)$ and $f(x + pn) = f(x)$ for all $x \in \text{Dom}(f)$. By the definition of a periodic function, since $x + pn = x + p(n - 1) + p$ is in $\text{Dom}(f)$, so is $x + p(n - 1)$, and

$$f(x) = f(x + pn) = f(x + p(n - 1) + p) = f(x + p(n - 1))$$

for all $x \in \text{Dom}(f)$. This completes the second induction. ■

Theorem 7.3. Suppose $f : X \to \mathbb{R}$ is periodic with minimal period p, where X is a subset of $\mathbb{R}$. If $ax + b$ and $cx + d$ are non-constant linear functions with domains $\mathbb{R}$, then

$$k(x) = af(cx + d) + b$$

is periodic with minimal period $\dfrac{p}{|c|}$.

Proof. First we show that $\dfrac{p}{|c|}$ is a period of k. We will start by proving that

$$x \in \text{Dom}(k) \iff x + \frac{p}{|c|} \in \text{Dom}(k).$$

Since $ax + b$ and $cx + d$ are bijective from $\mathbb{R}$ to $\mathbb{R}$, it can be calculated that $\text{Dom}(k)$ is the preimage of X under $cx + d$. So $x \in \text{Dom}(k)$ if and only if $cx + d \in X$. Since f is periodic, $cx + d \in X$ if and only if $cx \pm p + d \in X$, where the $\pm$ sign is taken to be positive if c is positive and negative if c is negative. This is equivalent to

$$c\left(x + \frac{p}{|c|}\right) + d \in X.$$

Again, using the fact that $\text{Dom}(k)$ is the preimage of X under $cx + d$, the last condition is true if and only if $x + \dfrac{p}{|c|} \in \text{Dom}(k)$, as desired. Moreover, $\dfrac{p}{|c|}$ is positive and

$$\begin{aligned} k\left(x + \frac{p}{|c|}\right) &= af\left(c\left(x + \frac{p}{|c|}\right) + d\right) + b \\ &= af(cx \pm p + d) + b \\ &= af(cx + d) + b \\ &= k(x), \end{aligned}$$

where, as before, the $\pm$ sign is positive if c is positive and negative if c is negative.
Now suppose for contradiction that there exists a period q of f that is strictly smaller than $\frac{p}{|c|}$. If $x \in \text{Dom}(k)$ then $x+q$ and $x-q$ are also in $\text{Dom}(k)$, and all three lead to the same value in the image of k. Now we have the sequence of implications for all $x \in \text{Dom}(k)$:

$$\begin{aligned} k(x \pm q) &= k(x) \\ af(c(x \pm q) + d) + b &= af(cx+d) + b \\ f(cx + d \pm cq) &= f(cx+d) \\ f(cx + d + |c|q) &= f(cx+d) \end{aligned}$$

For every real y, there exists a real $x = \dfrac{y-d}{c}$ such that $y = cx + d$. So for every real y,

$$f(y + |c|q) = f(y).$$

But $|c|q < p$, which contradicts the minimality of p as a period of f. ■

Although trigonometric functions have purely algebraic definitions in terms of infinite series, it will be more illuminating for us to define them geometrically.

Definition 7.4. Recall that the ratio of the circumference divided by the diameter is called π, which has the same value for all circles. The circumference of a unit circle is

$$\text{circumference} = \pi \cdot \text{diameter} = 2\pi.$$

For this reason, instead of using 360° to measure a full rotation around a circle, we can use 2π **radians** to measure a full rotation, and fractions of 2π radians to measure other sizes of rotation that are less than or greater than a full rotation. With radians, we don't mention the units by convention, unlike the degrees symbol. Also, if necessary, we can multiply by the conversion factor

$$\frac{360^\circ}{2\pi} = 1$$

or its reciprocal to convert between degrees and radians.

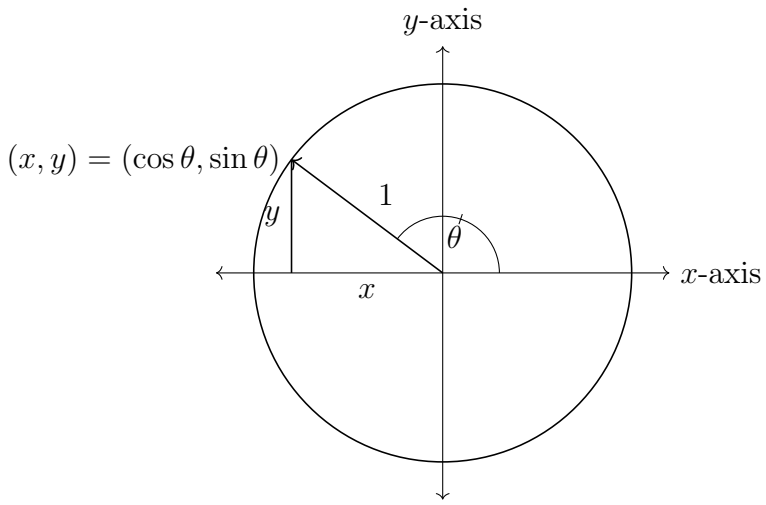
We are now ready to define the six trigonometric functions, whose arguments (inputs) will be in radians unless otherwise specified.

Definition 7.5. Let θ be a real number, which we interpret as an angle in radians. Starting at $(1,0)$ on the unit circle, we rotate around the origin by angle θ counterclockwise while staying on the unit circle and land on the point (x,y) on the unit circle. Note that if θ is negative, this is a clockwise rotation. Then we can define the following **trigonometric functions**:

1. Sine: $\sin\theta = y$
2. Cosine: $\cos\theta = x$
3. Tangent: $\tan\theta = \dfrac{y}{x} = \dfrac{\sin\theta}{\cos\theta}$

4. Secant: $\sec\theta = \dfrac{1}{x} = \dfrac{1}{\cos\theta}$
5. Cosecant: $\csc\theta = \dfrac{1}{y} = \dfrac{1}{\sin\theta}$
6. Cotangent: $\cot\theta = \dfrac{x}{y} = \dfrac{\cos\theta}{\sin\theta}$

If we draw the line segment from (x, y) to the origin, and draw the perpendicular distance from (x, y) to the x-axis, then the trigonometric functions can be interpreted as ratios between the legs and hypotenuse of the resulting triangle. The caveat to this approach is that we have to allow for negative lengths in the triangles since x or y can be negative, depending on the quadrant in which it lies. We also have to allow for degenerate triangles since (x, y) could lie on the boundary between quadrants, as in the x-axis or y-axis.



Definition 7.6. If we restrict ourselves to angles in only the first quadrant where the angles are acute and x and y are positive, then we can interpret the trigonometric functions in an equivalent way that has geometric significance in relation to right triangles. If the angle is θ then there exist right triangles with angles $\theta, \dfrac{\pi}{2} - \theta, \dfrac{\pi}{2}$. Let the side opposite to θ have length o, the adjacent side have length a, and the hypotenuse have length h. This triangle is similar to a triangle with the same angles but with a hypotenuse of length 1 that can be placed inside the first quadrant. The ratios of sides allows us to conclude that:

$$\sin\theta = \frac{o}{h},\ \cos\theta = \frac{a}{h},\ \tan\theta = \frac{o}{a}$$

Similar definitions hold for the other three (reciprocal) trigonometric functions.

Theorem 7.7. The reader should be familiar with the fact that $30° - 60° - 90°$ right triangles have sides in the ratio $1 : \sqrt{3} : 2$, and that $45° - 45° - 90°$ right triangles have sides in the ratio $1 : 1 : \sqrt{2}$. These triangles allow us to compute the following common values of the three main trigonometric functions:

Function	$\theta = \frac{\pi}{6}$	$\theta = \frac{\pi}{4}$	$\theta = \frac{\pi}{3}$	$\theta = \frac{\pi}{2}$
sin	$\frac{1}{2}$	$\frac{\sqrt{2}}{2}$	$\frac{\sqrt{3}}{2}$	1
cos	$\frac{\sqrt{3}}{2}$	$\frac{\sqrt{2}}{2}$	$\frac{1}{2}$	0
tan	$\frac{\sqrt{3}}{3}$	1	$\sqrt{3}$	undefined

The trigonometric functions are periodic. This makes sense since angles that are 2π apart lead to the same coordinates x and y. However, the period of not every trigonometric function is 2π, as some periods are smaller. Note that when we discuss trigonometric functions, "the" period refers to the minimal period.

Theorem 7.8. The domains, ranges, periods, and sets of roots of the six trigonometric functions are summarized as follows:

Function	Domain	Range	Period	Roots
sin	$\mathbb{R}$	$[-1, 1]$	2π	$\{\pi n : n \in \mathbb{Z}\}$
cos	$\mathbb{R}$	$[-1, 1]$	2π	$\left\{\frac{\pi}{2} + \pi n : n \in \mathbb{Z}\right\}$
tan	$\mathbb{R}\backslash\left\{\frac{\pi}{2} + \pi n : n \in \mathbb{Z}\right\}$	$\mathbb{R}$	π	$\{\pi n : n \in \mathbb{Z}\}$
sec	$\mathbb{R}\backslash\left\{\frac{\pi}{2} + \pi n : n \in \mathbb{Z}\right\}$	$\mathbb{R}\backslash(-1, 1)$	2π	None
csc	$\mathbb{R}\backslash\{\pi n : n \in \mathbb{Z}\}$	$\mathbb{R}\backslash(-1, 1)$	2π	None
cot	$\mathbb{R}\backslash\{\pi n : n \in \mathbb{Z}\}$	$\mathbb{R}$	π	$\left\{\frac{\pi}{2} + \pi n : n \in \mathbb{Z}\right\}$

Problem 7.9. Use the fact that we can uniquely represent each real as some $\tan\theta$ within one period of the tangent function to prove that all finite open intervals (a, b) for $a < b$ have the same cardinality as all of $\mathbb{R}$. In other words, find a bijection from (a, b) to $\mathbb{R}$.

Definition 7.10. A **sinusoidal function** is a function that is a result of left-composing or right-composing a non-constant linear function with sine to produce a function of the form

$$f(x) = a\sin(cx + d) + b$$

or similarly with cosine. The **amplitude** of f is half the difference between its largest output and smallest output. Since such a function will inherit the periodicity of sine or cosine by Theorem 7.3, its **period** is defined in the standard way.

Theorem 7.11. The amplitude of the sinusoidal function

$$f(x) = a\sin(cx + d) + b$$

is $|a|$, and its period is $\dfrac{2\pi}{|c|}$. The same result holds for cosine.

Proof. Since the range of $\sin(cx + d)$ is $[-1, 1]$,

$$\text{Rng}(f) = \begin{cases} [-a + b, a + b] & \text{if } a > 0 \\ [a + b, -a + b] & \text{if } a < 0 \end{cases},$$

both of which can be proven using inequalities. So if $a > 0$ then the amplitude is half of $(a + b) - (-a + b) = 2a$, and if $a < 0$ then the amplitude is half of $(-a + b) - (a + b) = -2a$. Either way, the amplitude is $2|a|$. The period follows from the fact that the period of sine is 2π and Theorem 7.3. The logic is identical for cosine. ■

Thus far, we have defined the trigonometric functions, computed their values at some common inputs, observed their domains, ranges, and periods, and studied the composition of sine functions with linear functions. The next step is to consider whether we can find inverse trigonometric functions. As a consequence of the periodicity of each trigonometric function f, the preimage of each value in the range of f has infinitely many elements in the domain of f. So true inverses do not exist. However, we can rectify this situation.

Definition 7.12. Let f be a trigonometric function. Suppose we can restrict f to a domain I such that the $\text{Rng}(f|_I) = \text{Rng}(f)$ and $f|_I$ is injective. Then we can define the function g so that, for each $y \in \text{Rng}(f)$, we choose $g(y)$ to be the unique element x of the preimage $f^{-1}(y)$ that lies in I. We call x the **principal value** of the preimage $f^{-1}(y)$. The set of principal values I of f is the range of g, and by convention I is chosen to an interval close to 0. **Inverse trigonometric functions** g are defined as follows:

Function	Definition	Domain	Range (Principal Values)
$\arcsin y = x$	$\sin x = y$	$[-1, 1]$	$\left[-\dfrac{\pi}{2}, \dfrac{\pi}{2}\right]$
$\arccos y = x$	$\cos x = y$	$[-1, 1]$	$[0, \pi]$
$\arctan y = x$	$\tan x = y$	$\mathbb{R}$	$\left(-\dfrac{\pi}{2}, \dfrac{\pi}{2}\right)$

Inverse functions $\text{arccot}, \text{arcsec}, \text{arccsc}$ can also be defined. As they are less common and as there is some disagreement among authors about the choice of principal values for arcsec and arccsc, we have omitted them from the discussion.

7.2 Trigonometric Identities

There are many identities that we can develop involving trigonometric functions. We begin with single-variable identities. In general, it should be assumed that we do not allow for division by 0, so the domains of the identities should be restricted as necessary. It is too tedious to always explicitly state the restrictions.

Theorem 7.13 (Pythagorean identity)**.** For each trigonometric function f, let the expression $f^2(\theta)$ refer to the squared expression $(f(\theta))^2$, not the composition $f(f(\theta))$. This is the convention for trigonometric functions. Then

$$\begin{aligned}\sin^2\theta + \cos^2\theta &= 1,\\ \tan^2\theta + 1 &= \sec^2\theta,\\ 1 + \cot^2\theta &= \csc^2\theta.\end{aligned}$$

We will usually not list identities about $\sec, \csc, \cot$ since they readily reduce to results about $\sin, \cos, \tan$, but we have made an exception in this case due to the frequency with which the Pythagorean identities appear in practice. The functions $\sec, \csc, \cot$ also sometimes make formulas more confusing than the "simplification" is worth.

Proof. Given an angle θ, let (x, y) be the point $(\cos\theta, \sin\theta)$ on the unit circle, as prescribed by the definition of the trigonometric functions. Then we draw line segment between (x, y) and the origin, and the perpendicular distance from (x, y) to the x-axis. This results in a right triangle with hypotenuse 1 and legs of length $|x|$ and $|y|$. By the Pythagorean theorem,

$$\sin^2\theta + \cos^2\theta = x^2 + y^2 = |x|^2 + |y|^2 = 1.$$

The other two identities follow from dividing the first Pythagorean identity by $\cos^2\theta$ or $\sin^2\theta$. ■

Theorem 7.14. The reflection identities are:

	Even-Odd	Complementary Angle	Supplementary Angle
Reflection Across	$y = 0$	$x = y$	$x = 0$
Sine	$\sin(-\theta) = -\sin\theta$	$\sin\left(\frac{\pi}{2} - \theta\right) = \cos\theta$	$\sin(\pi - \theta) = \sin\theta$
Cosine	$\cos(-\theta) = \cos\theta$	$\cos\left(\frac{\pi}{2} - \theta\right) = \sin\theta$	$\cos(\pi - \theta) = -\cos\theta$
Tangent	$\tan(-\theta) = -\tan\theta$	$\tan\left(\frac{\pi}{2} - \theta\right) = \cot\theta$	$\tan(\pi - \theta) = -\tan\theta$

Notice that right-composing a linear function with cosine produces sine and vice versa.

We do not prove these identities. Although they readily follow from making geometric observations about right triangles in the unit circle, there can be extensive casework on quadrants that is cumbersome enough to not be conducive to our purposes.

Theorem 7.15. This set of identities involve horizontal translations of trigonometric functions by some standard values. The phase shift identities are:

	Shift by $\frac{\pi}{2}$	Shift by π	Shift by 2π
Sine	$\sin\left(\theta \pm \frac{\pi}{2}\right) = \pm\cos\theta$	$\sin(\theta \pm \pi) = -\sin\theta$	$\sin(\theta \pm 2\pi) = \sin\theta$
Cosine	$\cos\left(\theta \pm \frac{\pi}{2}\right) = \mp\sin\theta$	$\cos(\theta \pm \pi) = -\cos\theta$	$\cos(\theta \pm 2\pi) = \cos\theta$
Tangent	$\tan\left(\theta \pm \frac{\pi}{2}\right) = -\cot\theta$	$\tan(\theta \pm \pi) = \tan\theta$	$\tan(\theta \pm 2\pi) = \tan\theta$

Proof. The phase shift identities with shifts by 2π in the right-most column are easy to prove since we know that the periods of sin and cos are 2π, and the period of tan is π. The tan identities in the bottom row follow from finding the quotient of the analogous identities for sine and cosine. The remaining identities follow from applying the reflections identities. For illustrative purposes, here is an example:

$$\sin\left(\theta + \frac{\pi}{2}\right) = \sin\left(\frac{\pi}{2} - (-\theta)\right) = \cos(-\theta) = \cos\theta.$$

The reader should verify the others independently. ■

The next step is to jump from single-variable identities to identities in two variables.

Theorem 7.16. It may be desirable to decompose the trigonometric function of a sum or difference of angles into an expression involving the trigonometric functions of the individual angles. This idea is encapsulated in the angle sum and angle difference identities:

	Angle Sum	Angle Difference
Sine	$\sin(\alpha+\beta) = \sin\alpha\cos\beta + \cos\alpha\sin\beta$	$\sin(\alpha-\beta) = \sin\alpha\cos\beta - \cos\alpha\sin\beta$
Cosine	$\cos(\alpha+\beta) = \cos\alpha\cos\beta - \sin\alpha\sin\beta$	$\cos(\alpha-\beta) = \cos\alpha\cos\beta + \sin\alpha\sin\beta$
Tangent	$\tan(\alpha+\beta) = \frac{\tan\alpha + \tan\beta}{1 - \tan\alpha\tan\beta}$	$\tan(\alpha-\beta) = \frac{\tan\alpha - \tan\beta}{1 + \tan\alpha\tan\beta}$

Proof. Taking the first identity $\sin(\alpha+\beta) = \sin\alpha\cos\beta + \cos\alpha\sin\beta$ for granted allows us to derive the rest using it and the reflection identities. ■

Corollary 7.17. The double angle and half angle identities are:

	Double Angle	Half Angle
Sine	$\sin(2\theta) = 2\sin\theta\cos\theta$	$\sin^2\frac{\theta}{2} = \frac{1-\cos\theta}{2}$
Cosine	$\cos(2\theta) = \cos^2\theta - \sin^2\theta$ $= 2\cos^2\theta - 1$ $= 1 - 2\sin^2\theta$	$\cos^2\frac{\theta}{2} = \frac{1+\cos\theta}{2}$
Tangent	$\tan(2\theta) = \frac{2\tan\theta}{1-\tan^2\theta}$	$\tan\frac{\theta}{2} = \frac{\sin\theta}{1+\cos\theta} = \frac{1-\cos\theta}{\sin\theta}$

Note that taking the plus-minus square root in the half angle formula for sine or cosine leads to two possible isolations. The sign of $\sin\frac{\theta}{2}$ or $\cos\frac{\theta}{2}$ depends on the quadrant in which $\frac{\theta}{2}$ lies.

Proof. The double angle identities follow from substituting in $\theta = \alpha = \beta$ into the identities for $\sin(\alpha+\beta), \cos(\alpha+\beta), \tan(\alpha+\beta)$. To produce the second and third double angle identities for cosine, we apply the Pythagorean identity to the first one. The half angle identities for sine and cosine follow from the double angle identities for cosine using the natural substitution. The half angle identities for tangent are more troublesome: taking the quotient of the half angle identities for sine and cosines yields

$$\tan^2\frac{\theta}{2} = \frac{\sin^2\frac{\theta}{2}}{\cos^2\frac{\theta}{2}} = \frac{1-\cos\theta}{1+\cos\theta}.$$

Multiplying the numerator and denominator by $1-\cos\theta$ and applying the Pythagorean identity yields

$$\tan^2\frac{\theta}{2} = \frac{(1-\cos\theta)^2}{1-\cos^2\theta} = \left(\frac{1-\cos\theta}{\sin\theta}\right)^2.$$

Then we notice that $1-\cos\theta$ is always positive, and $\tan\frac{\theta}{2}$ always has the same sign as $\sin\theta$ (this is left as an exercise to the reader). Thus, we can remove the squares. The second identity can be proven by replacing the step where we multiplied the numerator and denominator by $1-\cos\theta$ instead with multiplying the numerator and denominator by $1+\cos\theta$. ■

Problem 7.18. Prove the triple angle identities:

$$\sin(3\theta) = 3\sin\theta - 4\sin^3\theta,$$
$$\cos(3\theta) = 4\cos^3\theta - 3\cos\theta.$$

Can you use these to develop a formula for $\tan(3\theta)$ in terms of $\tan\theta$?

Theorem 7.19. The final set of identities allow us to convert between a sum or difference and a product.

Product-to-Sum	Sum-to-Product
$\sin(\alpha+\beta)+\sin(\alpha-\beta)=2\sin\alpha\cos\beta$	$\sin\alpha+\sin\beta=2\sin\left(\frac{\alpha+\beta}{2}\right)\cos\left(\frac{\alpha-\beta}{2}\right)$
$\sin(\alpha+\beta)-\sin(\alpha-\beta)=2\cos\alpha\sin\beta$	$\sin\alpha-\sin\beta=2\sin\left(\frac{\alpha-\beta}{2}\right)\cos\left(\frac{\alpha+\beta}{2}\right)$
$\cos(\alpha-\beta)+\cos(\alpha+\beta)=2\cos\alpha\cos\beta$	$\cos\alpha+\cos\beta=2\cos\left(\frac{\alpha+\beta}{2}\right)\cos\left(\frac{\alpha-\beta}{2}\right)$
$\cos(\alpha-\beta)-\cos(\alpha+\beta)=2\sin\alpha\sin\beta$	$\cos\alpha-\cos\beta=-2\sin\left(\frac{\alpha+\beta}{2}\right)\sin\left(\frac{\alpha-\beta}{2}\right)$

Proof. The product-to-sum identities follow from expanding the two terms on the left side of each identity using the angle sum and angle difference identities. The sum-to-product identities follow from the product-to-sum identities. For illustrative purposes, we provide one of the proofs, and the rest is left to the reader: Suppose α and β are angles. Let $x=\dfrac{\alpha+\beta}{2}$ and $y=\dfrac{\alpha-\beta}{2}$. Then $x+y=\alpha$ and $x-y=\beta$. Therefore,

$$\begin{aligned}\sin\alpha+\sin\beta&=\sin(x+y)+\sin(x-y)\\&=2\sin x\cos y\\&=2\sin\left(\frac{\alpha+\beta}{2}\right)\cos\left(\frac{\alpha-\beta}{2}\right).\end{aligned}$$

■

Overall, we have discussed over fifty trigonometric identities, if we count all the variations in signs. It is not feasible to memorize all of them. We recommend initially memorizing the reflection identities and the expansion of $\sin(\alpha+\beta)$, and use them to derive other identities as needed. In the process of using identities in problems, the more frequently occurring ones will fall into memory in due time.

Chapter 8

Complex Numbers

> "Between two truths of the real domain, the easiest and shortest path quite often passes through the complex domain."
>
> – *Paul Painlevé*

The complex numbers are a two-dimensional extension of real numbers, and surprisingly form a field. They have significant algebraic ramifications, such as in the theory of polynomials. We will start off by studying the rectangular form of complex numbers and follow it up with the polar form. We will then observe that complex numbers are essentially Cartesian coordinates with an extra multiplicative structure, which we will see has a geometric interpretation involving rotation.

8.1 Rectangular Form

Definition 8.1. A **complex number** is an expression in the **rectangular form** $a + bi$, where a and b are real numbers, and i is a solution to the equation

$$x^2 + 1 = 0.$$

Note that since the squares of all real numbers are non-negative, whereas $i^2 = -1$, it must be true that i is distinct from all real numbers. The set of complex numbers is denoted by the symbol $\mathbb{C}$. The complex number $a+bi$ corresponds with the Cartesian coordinates (a, b). So $i = 0 + 1i$ corresponds with $(0, 1)$.

Definition 8.2. For a complex number $a + bi$, the **real part** is $\mathrm{Re}(a + bi) = a$ and the **imaginary part** is $\mathrm{Im}(a + bi) = b$ (note that the imaginary part is the real number b, not bi). The real part and the imaginary parts are called the **components** of the complex number. Two complex numbers are equal if and only if their real parts are equal and their imaginary parts are equal.

Definition 8.3. For a complex number $z = a + bi$, if $b = 0$ then z maybe be interpreted as the real number a. As such, $\mathbb{R}$ is a subset of $\mathbb{C}$. If instead $a = 0$, then we call the complex number $z = bi$ **pure imaginary**.

Definition 8.4. We define **addition** and **multiplication** on complex numbers in way that, if the two complex numbers in question are real, the definitions of these operations on $\mathbb{C}$ boil down to the definitions of the operations on $\mathbb{R}$:

- Addition is component-wise, so that

$$(a + bi) + (c + di) = (a + c) + (b + d)i.$$

- Multiplication respects the distributive law, so that

$$\begin{aligned}(a + bi) \cdot (c + di) &= ac + adi + bci + bdi^2 \\ &= ac + adi + bci - bd \\ &= (ac - bd) + (ad + bc)i.\end{aligned}$$

Theorem 8.5. The complex numbers with addition and multiplication form a field (see Definition 2.14). To recap in an expanded fashion, this means:

1. Addition and multiplication are commutative
2. Addition and multiplication are associative
3. Multiplicative is distributive over addition
4. There exists an additive identity and a multiplicative identity
5. Every element has an additive inverse
6. Even element that is not the additive identity has a multiplicative inverse

By Theorem 2.10 and Theorem 2.8, identities and inverses in a field are unique.

Proof. We provide the proofs and ideas for some of the properties. The rest are left as exercises to the reader.

4. It can be verified that $0 = 0 + 0i$ is an additive identity, and that $1 = 1 + 0i$ is a multiplicative identity.
5. It can be verified that the negative $-1 \cdot (a + bi) = (-a) + (-b)i$ is an additive inverse of $a + bi$. We denote the additive inverse of $a + bi$ by $-(a + bi)$.
6. Finding a multiplicative inverse of $a+bi \neq 0$ requires slightly more effort. Suppose $c+di$ exists such that $(c+di)(a+bi) = 1$. The critical observation is that $(a+bi)(a+(-b)i)$ expands to the real number $a^2 + b^2$. So multiplying both sides of the equation by $a + (-b)i$ yields $(c + di)(a^2 + b^2) = a + (-b)i$ or

$$c + di = \frac{a}{a^2 + b^2} + \frac{-b}{a^2 + b^2}i.$$

It can then be manually verified that this is indeed a multiplicative inverse. We denote that multiplicative inverse of $a + bi$ by $(a + bi)^{-1}$.

■

Definition 8.6. Due to the existence of additive inverses for all complex numbers and multiplicative inverses for all non-zero complex numbers, we can define the inverse operations of **subtraction** and **division** on $\mathbb{C}$:

- Subtraction by a complex number $c+di$ is the addition of its additive inverse:
$$(a+bi)-(c+di)=(a-c)+(b-d)i.$$
- Division by a non-zero complex number $c+di$ is multiplication by its multiplicative inverse:
$$\frac{a+bi}{c+di}=(a+bi)(c+di)^{-1}.$$

Definition 8.7. We can define **powers** of a complex number z, as long as we only allow integer exponents n for the time being:

- If $n=0$ and $z\neq 0$, then $z^0=1$.
- If n is a positive integer, then $z^n=\underbrace{z\cdot z\cdots z}_{n \text{ copies of } z}$.
- If n is a negative integer then $z^n=(z^{-1})^{-n}=\underbrace{z^{-1}\cdot z^{-1}\cdots z^{-1}}_{(-n) \text{ copies of } z^{-1}}$.

Theorem 8.8. The usual exponents rules hold for complex numbers z and w, and integers exponents n and m, assuming we never take a base of 0 to the power of a negative exponent or divide by 0:

1. Negative exponent means reciprocal: $z^{-n}=\dfrac{1}{z^n}=(z^n)^{-1}$
2. Multiplying powers with the same base: $z^n\cdot z^m=z^{n+m}$
3. Dividing powers with the same base: $\dfrac{z^n}{z^m}=z^{n-m}$
4. Powers of powers: $(z^n)^m=z^{nm}$
5. Multiplying powers with the same exponent: $z^n\cdot w^n=(zw)^n$
6. Dividing powers with the same exponent: $\dfrac{z^n}{w^n}=\left(\dfrac{z}{w}\right)^n$

Problem 8.9. By Example 2.51, the powers $(-1)^n$ oscillate between 1 and -1 for positive integers n. Find a pattern in the powers of i.

As mentioned earlier, each complex number $a+bi$ can be interpreted as a point (a,b) on the Cartesian plane, and vice versa. This allows us to see complex numbers visually, similar to how we can see real number on the number line. Addition of complex numbers is analogous to addition of coordinates, as they are both component-wise. As we will see, the true power of this new interpretation is the fact that complex numbers can be multiplied, which is a definition that we did not have for vanilla coordinates.

Definition 8.10. When finding the inverse of $z = a + bi$, we found the complex number number $a - bi$ to be useful. We call

$$\overline{z} = a - bi$$

the **complex conjugate** of $a + bi$. Geometrically, the function $z \mapsto \overline{z}$ is a reflection across the x-axis. The function that is the reflection across the y-axis is given by $z \mapsto -\overline{z}$.

Theorem 8.11. The complex conjugate has several useful and interesting properties. If z and w are complex numbers, a is a real number, and n is an integer then:

1. Conjugate of a real number: $\overline{z} = z$ if and only if $z = a$ for some real number a
2. Conjugate of a pure imaginary number: $\overline{z} = -z$ if and only if $z = ai$ for some real number a
3. Real and imaginary parts: $\mathrm{Re}(z) = \dfrac{z + \overline{z}}{2}$ and $\mathrm{Im}(z) = \dfrac{z - \overline{z}}{2i}$
4. Conjugation is an involution: $\overline{(\overline{z})} = z$
5. Conjugation distributes over addition: $\overline{z + w} = \overline{z} + \overline{w}$
6. Conjugation distributes over multiplication: $\overline{z \cdot w} = \overline{z} \cdot \overline{w}$
7. Conjugation commutes with negation: $\overline{(-z)} = -\overline{z}$, and so $\overline{z - w} = \overline{z} - \overline{w}$
8. Conjugation commutes with taking inverse: $(\overline{z})^{-1} = \overline{(z^{-1})}$, and so $\overline{\left(\dfrac{w}{z}\right)} = \dfrac{\overline{w}}{\overline{z}}$ for $z \neq 0$
9. Conjugation commutes with taking powers: $(\overline{z})^n = \overline{z^n}$

These are easy to verify, so their proofs are left to the reader.

Definition 8.12. During the process of finding the inverse of $z = a + bi$, the expression $a^2 + b^2$ also came up. We call the real number

$$|a + bi| = \sqrt{a^2 + b^2}$$

the **modulus** or **magnitude** or **absolute value** of $a + bi$. The reason that we take the square root of it is that the final expression then has an easy geometric interpretation. If we draw the line segment between $a + bi$ and 0, along with the perpendicular distance from $a + bi$ to the x-axis, then we have produced a right triangle with lengths of length $|a|$ and $|b|$ (these are real absolute values). The hypotenuse is the distance between $a + bi$ and the origin, and the Pythagorean theorem tells us that it equals $\sqrt{a^2 + b^2} = |a + bi|$.

Theorem 8.13. The complex modulus has some convenient properties. If z and w are complex numbers, and n is an integer then:

1. When applied to complex numbers with 0 imaginary part, the complex modulus restricts to becoming the real absolute value.
2. Zero modulus: $|z| = 0$ if and only if $z = 0$

3. Modulus and conjugation: $z \cdot \overline{z} = |z|^2$; in particular, if z lies on the complex unit circle then $|z| = 1$ and so $\overline{z} = \dfrac{1}{z}$
4. Modulus is preserved under reflections: $|z| = |-z| = |\overline{z}| = \left|\overline{(-z)}\right|$
5. Modulus distributes over multiplication: $|z \cdot w| = |z| \cdot |w|$
6. Modulus commutes with taking inverse: $\left|z^{-1}\right| = |z|^{-1}$, and so $\left|\dfrac{w}{z}\right| = \dfrac{|w|}{|z|}$ for $z \neq 0$
7. Modulus commutes with taking powers: $|z^n| = |z|^n$
8. Real and imaginary parts: $|z| \geq |\mathrm{Re}(z)|$ and $|z| \geq |\mathrm{Im}(z)|$
9. **Complex triangle inequality**: $|z| + |w| \geq |z + w|$ where equality holds if and only if $z = 0$ or $w = 0$ or $z = \alpha w$ for some real $\alpha > 0$. This means z and w fall on the same ray emanating from the origin.

Proof. We leave all but the triangle inequality as exercises to the reader, as the rest are straightforward to verify. In order to prove this inequality, we will use the substitutions $z = a+bi$ and $w = c+di$, and work backwards through repeated squaring and simplification:

$$\begin{aligned}
&|z| + |w| \geq |z + w| \\
\iff &|a + bi| + |c + di| \geq |a + bi + c + di| \\
\iff &\sqrt{a^2 + b^2} + \sqrt{c^2 + c^2} \geq \sqrt{(a + c)^2 + (b + d)^2} \\
\iff &a^2 + b^2 + c^2 + d^2 + 2\sqrt{(a^2 + b^2)(c^2 + d^2)} \geq a^2 + c^2 + b^2 + d^2 + 2ac + 2bd \\
\iff &\sqrt{(a^2 + b^2)(c^2 + d^2)} \geq ac + bd.
\end{aligned}$$

This actually follows from the two-dimensional special case of the Cauchy-Schwarz inequality (see the proof in Theorem 11.9), but for now we will prove it by expansion.

$$\begin{aligned}
\sqrt{(a^2 + b^2)(c^2 + d^2)} \geq ac + bd \iff &\sqrt{(a^2 + b^2)(c^2 + d^2)} \geq |ac + bd| \\
\iff &(a^2 + b^2)(c^2 + d^2) \geq (ac + bd)^2 \\
\iff &a^2c^2 + a^2d^2 + b^2c^2 + b^2d^2 \geq a^2c^2 + b^2d^2 + 2abcd \\
\iff &a^2d^2 + b^2c^2 \geq 2abcd \\
\iff &(ad - bc)^2 \geq 0,
\end{aligned}$$

which follows from the trivial inequality.
Equality holds if and only if $ad = bc$ and $|ac+bd| = ac+bd$. It is easy to verify that if $z = 0$ or $w = 0$ or $z = \alpha w$ for some positive real α, then the above equality condition is implied. So we must prove the other direction. Suppose $ad = bc$ and $|ac + bd| = ac + bd$, and that $z \neq 0$ and $w \neq 0$. By some basic linear algebra, $ad = bc$ means that the position vectors with arrowheads (a, b) and (c, d) are linearly dependent, which means they must lie on the same line through the origin. By an equivalent criterion for linear dependence, since $z \neq 0$,

there exists a real α such that $z = \alpha w$. All we need to do now is prove that they point in the same direction or, equivalently, lie on the same ray through the origin. To do this, we will prove that $\alpha > 0$. Indeed, by $|ac + bd| = ac + bd$, we get

$$|ac + bd| = |\alpha c \cdot d + \alpha d \cdot d| = |\alpha| \cdot (c^2 + d^2),$$
$$ac + bd = \alpha c \cdot d + \alpha d \cdot d = \alpha \cdot (c^2 + d^2),$$

so $|\alpha| = \alpha$, making α non-negative. Also, $\alpha \neq 0$ because $\alpha = 0$ would imply the contradiction $z = 0$. Therefore, α is strictly positive. ■

Theorem 8.14 (Complex geometric series)**.** For any non-negative integer t and complex number $z \neq 1$,

$$1 + z + z^2 + \cdots + z^t = \frac{z^{t+1} - 1}{z - 1}.$$

Proof. The proof is the same as that for summing a geometric series with real summands. Let

$$S = 1 + z + z^2 + \cdots + z^t.$$

By multiplying both sides by z, we get

$$zS = z + z^2 + z^3 + \cdots + z^{t+1}.$$

Subtracting the latter equation from the former yields $(1 - z)S. = 1 - z^{t+1}$. Therefore,

$$S = \frac{1 - z^{t+1}}{1 - z}.$$

■

Problem 8.15. Prove the following variant of the formula for a finite geometric series. Let $z \neq 1$ be a complex number. Let α and β be positive integers and γ be a non-negative integer such that $\gamma = \alpha - q\beta$ for some non-negative integer q. Then

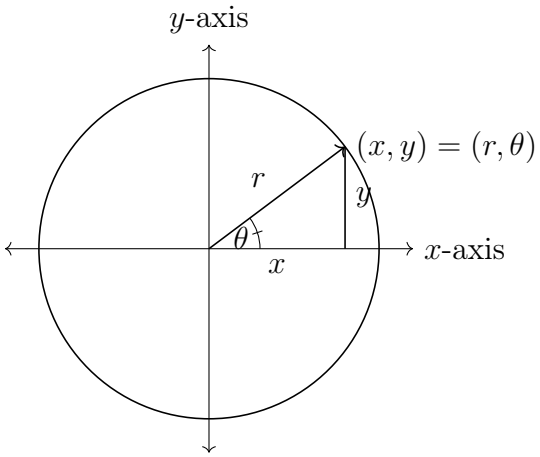
$$z^\alpha + z^{\alpha-\beta} + z^{\alpha-2\beta} + \cdots + z^\gamma = \frac{z^{\alpha+\beta} - z^\gamma}{z^\beta - 1}.$$

8.2 Polar Form

So far, we have been working with complex numbers in the rectangular form $a + bi$. There is another way of representing complex numbers which illuminates further properties.

Definition 8.16. Instead of locating points $z \neq 0$ on the Cartesian plane using rectangular coordinates (x, y), we can use **polar coordinates**. Polar coordinates also have two components (r, θ). The first component $r \in (0, \infty)$ represents the circle centered at the origin on which z lies; we compute r as the length of the radius from z to the origin, which is $|z|$. The second component $\theta \in [0, 2\pi)$ is the angle by which the positive x-axis has to be rotated counterclockwise in order to coincide with the aforementioned radius. These coordinates $(r, \theta) \in (0, \infty) \times [0, 2\pi)$ are unique representations of points on the plane, and so

are in bijection with rectangular coordinates $(x, y) \in \mathbb{R} \times \mathbb{R} \backslash \{(0, 0)\}$. If we interpret z as a complex number, then we define the **argument** of z to be $\arg z = \theta$. Similar to rectangular coordinates, two complex numbers are equal if and only if their moduli are equal and their arguments are equal.



Now we will tackle the question of how to convert between rectangular and polar coordinates.

Theorem 8.17. Suppose $z \neq 0$ is a point on the Cartesian plane. If the polar coordinates for z are (r, θ), then the corresponding rectangular coordinates (x, y) are

$$\begin{aligned} x &= r \cdot \cos\theta, \\ y &= r \cdot \sin\theta. \end{aligned}$$

Proof. Let $x' = \cos\theta$ and $y' = \sin\theta$. Then (x', y') is the point on the unit circle that makes an angle of θ with the positive x-axis, by the definition of sine and cosine. The radius from the origin to (x', y') has length 1. By dilating the unit circle from the origin by a factor of r, (x', y') is mapped to (rx', ry'), which coincides with z. Thus,

$$z = (rx', ry') = (r \cdot \cos\theta, r \cdot \sin\theta).$$

This use of dilation can be made formal using similar triangles, which we leave to the reader. ■

Theorem 8.18. Suppose $z \neq 0$ is a point on the Cartesian plane. If the rectangular coordinates for x are (x, y), then the corresponding polar coordinates (r, θ) are

$$\begin{aligned} r &= \sqrt{x^2 + y^2}, \\ \theta &= \begin{cases} \arccos\dfrac{x}{r} & \text{if } y \geq 0 \\ 2\pi - \arccos\dfrac{x}{r} & \text{if } y < 0 \end{cases}. \end{aligned}$$

Authors often present a formula for θ using arctan but we have preferred to express θ using arccos because it involves fewer cases.

Proof. The fact that $r = \sqrt{x^2 + y^2}$ follows from the fact that r is the distance from (x, y) to 0. To compute θ, we will tackle the case $y \geq 0$ followed by $y < 0$. In both cases, we will use the fact that Definition 7.12 says that the range of arccos is $[0, \pi]$.

- Suppose $y \geq 0$. By observing the unit circle, this means $\theta \in [0, \pi]$. We know that $x = r \cdot \cos\theta$. Since $\theta \in \text{Rng}(\arccos)$, we can isolate θ to get
$$\theta = \arccos \frac{x}{r}.$$
- Now suppose $y < 0$. By observing the unit circle, this means $0 < \theta < 2\pi$ which is equivalent to $0 < 2\pi - \theta < \pi$. Using the period of cosine and a reflection identity,
$$\cos(2\pi - \theta) = \cos(-\theta) = \cos\theta.$$
As before, $\cos\theta = \dfrac{x}{r}$, so $\cos(2\pi - \theta) = \dfrac{x}{r}$. Since $(2\pi - \theta) \in \text{Rng}(\arccos)$, we can isolate θ to get
$$\theta = 2\pi - \arccos \frac{x}{r}.$$

■

Definition 8.19. We defined $\arg(z)$ to lie in $[0, 2\pi)$, but it often suffices to use any value of the set
$$\text{Arg}(z) = \{\arg(z) + 2\pi k : k \in \mathbb{Z}\}$$
because, if combined with a fixed modulus, they all lead to the same point in the plane. We say that angles α and β are **congruent modulo** 2π if $\alpha - \beta = 2\pi k$ for some integer k, and we denote this relationship by $\alpha \equiv \beta$.

Lemma 8.20. If $\alpha = \beta$ then of course $\alpha \equiv \beta$. The converse is not necessarily true, but we can establish equality using congruence if the angles lie within the same revolution. To be precise, if we know that $\alpha \equiv \beta$ then $\alpha = \beta$ if either of the following criteria hold:

1. Both α and β lie in the same interval $[x, y)$ or $(x, y]$ or (x, y) where $0 < y - x \leq 2\pi$
2. Both α and β lie in the same interval $[x, y]$ where $0 \leq y - x < 2\pi$

The closed and open nature of an endpoint of the interval is important, as is the strictness or non-strictness of the upper bound of the inequality.

Proof. We provide a proof for the interval $[x, y)$ in the first case and leave the rest as an exercise in mimicry to the reader. As both α and β lie in $[x, y)$, we have the inequalities
$$\begin{aligned} &x \leq \alpha < y, \\ &x \leq \beta < y \implies -y < -\beta \leq -x. \end{aligned}$$
Their sum is $x - y < \alpha - \beta < y - x$ or
$$0 \leq |\alpha - \beta| < y - x \leq 2\pi.$$
Since $\alpha \equiv \beta$, we know that $\alpha - \beta$ is an integer multiple of 2π. The only multiply of 2π in $[0, 2\pi)$ is 0, which completes the proof. ■

Definition 8.21. Let $z = x + iy$ be a non-zero element of the complex plane. By the existence of polar coordinates, we know there exists an ordered pair $(r, \theta) \in (0, \infty) \times [0, 2\pi)$ such that $x = r \cdot \cos\theta$ and $y = r \cdot \sin\theta$. This means

$$z = x + iy = r(\cos\theta + i\sin\theta).$$

This is called the **trigonometric form** of a complex number. Note that we may replace θ with any angle congruent to θ modulo 2π without changing the underlying point in the plane. So, even though we call it "the" trigonometric form, this form is not unique. As a short hand, we define

$$re^{i\theta} = r(\cos\theta + i\sin\theta),$$

where e is a real number called Euler's constant.

It can be shown that is definition is consistent with the infinite series expressions for $e^{i\theta}, \cos\theta$, and $\sin\theta$ but we will not delve into this. Instead, we will demonstrate that this exponential notation is appropriate by proving several analogues of exponent laws for expressions of the form $e^{i\theta}$.

Theorem 8.22. The follow exponent laws hold for any real α and β, positive reals r and s, and any integer n :

1. $re^{i\alpha} \cdot se^{i\beta} = rse^{i(\alpha+\beta)}$
2. $(re^{i\alpha})^{-1} = \dfrac{1}{r}e^{i(-\alpha)}$ and so $\overline{re^{i\theta}} = re^{i(-\theta)}$
3. $\dfrac{re^{i\alpha}}{se^{i\beta}} = \dfrac{r}{s}e^{i(\alpha-\beta)}$
4. $(re^{i\alpha})^n = r^n e^{i(n\alpha)}$ (de Moivre's formula)

Proof. We prove these identities one by one.

1. By the sine and cosine angle sum identities (Theorem 7.16),

$$\begin{aligned} re^{i\alpha} \cdot se^{i\beta} &= r(\cos\alpha + i\sin\alpha) \cdot s(\cos\beta + i\sin\beta) \\ &= rs\left((\cos\alpha\cos\beta - \sin\alpha\sin\beta) + i(\sin\alpha\cos\beta + \cos\alpha\sin\beta)\right) \\ &= rs\left(\cos(\alpha+\beta) + i\sin(\alpha+\beta)\right) \\ &= rse^{i(\alpha+\beta)}. \end{aligned}$$

2. Using the fact that $(a+bi)^{-1} = \dfrac{a}{a^2+b^2} + i\dfrac{-b}{a^2+b^2}$ and the Pythagorean identity and

reflection identities,

$$\begin{aligned}(re^{i\alpha})^{-1} &= (r\cos\alpha + ir\sin\alpha)^{-1} \\ &= \frac{r\cos\alpha}{r^2(\cos^2\alpha+\sin^2\alpha)} + i\frac{-r\sin\alpha}{r^2(\cos^2\alpha+\sin^2\alpha)} \\ &= \frac{1}{r}(\cos\alpha - i\sin\alpha) \\ &= \frac{1}{r}(\cos(-\alpha)+i\sin(-\alpha)) \\ &= \frac{1}{r}e^{i(-\alpha)}.\end{aligned}$$

As a consequence, since $(re^{i\theta})\cdot\overline{(re^{i\theta})} = r^2$ (due to $z\cdot\overline{z} = |z|^2$, mentioned in Theorem 8.13)

$$\overline{(re^{i\theta})} = r^2\cdot(re^{i\theta})^{-1} = re^{i(-\theta)}.$$

3. Using the previous two parts,

$$\frac{re^{i\alpha}}{se^{i\beta}} = (re^{i\alpha})\cdot(se^{i\beta})^{-1} = re^{i\alpha}\cdot\frac{1}{s}e^{i(-\beta)} = \frac{r}{s}e^{i(\alpha-\beta)}.$$

4. To prove de Moivre's formula, we will split it into the cases $n=0, n\geq 1$, and $n\leq -1$. For $n=0$,

$$(re^{i\theta})^0 = 1 = r^0(\cos 0 + i\sin 0) = r^0e^{i0} = r^0e^{i(0\theta)}.$$

For $n\geq 1$, this is an induction exercise using

$$re^{i\alpha}\cdot se^{i\beta} = rse^{i(\alpha+\beta)}.$$

If $n\leq -1$, let $m=-n\geq 1$. By previous parts,

$$(re^{i\theta})^n = (re^{i\theta})^{(-1)\cdot m} = \left((re^{i\theta})^m\right)^{-1} = (r^me^{i(m\theta)})^{-1} = r^{-m}e^{i(-m\theta)} = r^ne^{i(n\theta)}.$$

■

Problem 8.23. Prove that

$$\sin\theta = \frac{e^{i\theta}-e^{i(-\theta)}}{2i} \text{ and } \cos\theta = \frac{e^{i\theta}+e^{i(-\theta)}}{2}.$$

Lemma 8.24. We now establish some results about trigonometric forms.

1. For all real θ, $e^{i\theta} = \cos\theta + i\sin\theta = 1$ if and only if $\theta\equiv 0$.
2. For all real α and β, $e^{i\alpha} = e^{i\beta}$ if and only if $\alpha\equiv\beta$.
3. For all real θ, $\arg(e^{i\theta})\equiv\theta$.
4. For all real α and β, $\arg(e^{i\alpha}) = \arg(e^{i\beta})$ if and only if $\alpha\equiv\beta$.

Proof. We will prove each part in succession, using earlier parts to prove later parts.

1. Suppose $e^{i\theta} = \cos\theta + i\sin\theta = 1$. Then we can equate real and imaginary parts to get $\cos\theta = 1$ and $\sin\theta = 0$, which simultaneously hold only at integer multiples of 2π. So $\theta \equiv 0$.

2. Subsequently, we have the following equivalences:

$$e^{i\alpha} = e^{i\beta} \iff \frac{e^{i\alpha}}{e^{i\beta}} = 1 \iff e^{i(\alpha-\beta)} = 1 \iff \alpha - \beta \equiv 0 \iff \alpha \equiv \beta.$$

3. Let $e^{i\theta} = x + iy$ and $\arg(x+iy) = \phi$. We know that $x = \cos\phi$ and $y = \sin\phi$. Then

$$e^{i\theta} = x + iy = \cos\phi + i\sin\phi = e^{i\phi}.$$

   By the previous part, this means $\arg(e^{i\theta}) = \phi \equiv \theta$.

4. This follows from the previous two parts. If $\arg(e^{i\alpha}) = \arg(e^{i\beta})$ then using the fact that $\arg(e^{i\alpha}) \equiv \alpha$ and $\arg(e^{i\beta}) \equiv \beta$, we get $\alpha \equiv \beta$. Conversely, if $\alpha \equiv \beta$ then we know that $e^{i\alpha} = e^{i\beta}$. Taking the argument of each sides yields $\arg(e^{i\alpha}) = \arg(e^{i\beta})$.

■

Corollary 8.25. We can redefine $\arg z$ for a non-zero complex number z to be the unique real number $\theta \in [0, 2\pi)$ such that $z = |z| \cdot e^{i\theta}$.

Proof. Suppose z is a non-zero complex number with modulus $r = |z|$ and $\arg z = \theta$. We know that $z = re^{i\theta}$ is a trigonometric form of z. Suppose z has another trigonometric form $se^{i\phi}$ for some $s > 0$ and $\phi \in [0, 2\pi)$. Then $r = |z| = |se^{i\phi}| = s$ and $e^{i\phi} = e^{i\theta}$. By Lemma 8.24, this means $\phi \equiv \theta$. But they both lie in $[0, 2\pi)$ which forces $\phi = \theta$. Therefore, $\arg z$ is the unique real number $\theta \in [0, 2\pi)$ such that $z = |z| \cdot e^{i\theta}$. ■

Corollary 8.26. Let z and w be non-zero complex numbers and n be an integer. Then the following analogues of logarithmic rules (Theorem 2.60) hold for arg modulo 2π :

1. $\arg(zw) \equiv \arg z + \arg w$
2. $\arg(z^{-1}) \equiv -\arg z$
3. $\arg\left(\frac{w}{z}\right) \equiv \arg w - \arg z$
4. $\arg(z^n) \equiv n\arg(z)$
5. $\arg z = 0$ if and only if z is a positive real number, and $\arg z = \pi$ if and only if z is a negative real number

Proof. We show how to prove the first one and leave the rest as an exercise to the reader. We will use the fact that $\arg(rz) = \arg(z)$ for any positive real r, which can be shown using similar triangles. We will also use the various analogues of exponent laws that we have proven for complex numbers in trigonometric form (Lemma 8.24), along with the fact that $\arg(e^{i\theta}) \equiv \theta$. Let $z = re^{i\alpha}$ and $w = se^{i\beta}$ where $\alpha = \arg z$ and $\beta = \arg w$. Then

$$\arg(zw) = \arg(re^{i\alpha} \cdot se^{i\beta}) = \arg(e^{i\alpha} \cdot e^{i\beta}) = \arg(e^{i(\alpha+\beta)}) \equiv \alpha + \beta = \arg z + \arg w.$$

■

The following corollary establishes a link between the polar form and geometry, specifically by interpreting multiplication by a complex number as a combination of dilation and rotation.

Corollary 8.27. The function $s_w : \mathbb{C} \to \mathbb{C}$, defined by $s_w(z) = zw$ has a geometric interpretation on the complex plane. It means we apply a positive dilation to z from the origin by a factor of $|w|$ and rotate it counterclockwise around the origin by an angle measuring $\arg w$. This is called a spiral similarity. In a specific case, if w lies on the complex unit circle, meaning $|w| = 1$, then this is function causes purely a rotation.

Proof. The geometric interpretation follows from the following two facts that we have previously stated:

$$|zw| = |z| \cdot |w|,$$
$$\arg(zw) \equiv \arg z + \arg w.$$

The special case is easy to see as well. ■

Problem 8.28. Determine a complex number w such that multiplying any complex number z by w results in z being rotated counter-clockwise by $\dfrac{\pi}{2} = 90°$.

Problem 8.29. Let z be a non-zero complex number. Then prove the following rules for arg :

1. $\arg(-z) \equiv \pi + \arg z$
2. $\arg(\overline{z}) \equiv -\arg z$

As a precursor to the fundamental theorem of algebra (Theorem 10.28), which tells us how many complex roots a polynomial with complex coefficients has, we will now compute all the n^{th} roots of a non-zero complex number in trigonometric form.

Theorem 8.30. A non-zero complex number z has exactly n distinct n^{th} roots. Since we have not actually defined n^{th} roots of complex numbers, this is made more precise in the language of equations by saying that there are exactly n distinct solutions $x = w$ of the equation $x^n - z$, where x is a variable and z is a fixed non-zero complex number. Moreover, if z is given in trigonometric form, we can compute each root w in trigonometric form.

Proof. Let $z = re^{i\theta}$. Then $w = se^{i\phi}$ is an n^{th} root of z if and only if

$$re^{i\theta} = (se^{i\phi})^n = s^n e^{i(n\phi)}.$$

This holds if and only if $s = \sqrt[n]{r}$ and $\theta \equiv n\phi$. The former condition is an explicit statement about the modulus of w, so will now focus on the latter condition. The condition $\theta \equiv n\phi$ is equivalent to there existing an integer k such that $\theta + 2\pi k = n\phi$, which we can write as

$$\phi = \frac{\theta + 2\pi k}{n}.$$

This would seem to imply that there is a value of ϕ for each integer k, but it possible that many such angles lead to the same point in the plane. We need to classify such overlaps. The possibility of $se^{\frac{\theta+2\pi k_1}{n}} = se^{\frac{\theta+2\pi k_2}{n}}$ is equivalent to

$$\frac{\theta + 2\pi k_2}{n} \equiv \frac{\theta + 2\pi k_2}{n},$$

which is equivalent to the existence of an integer j such that

$$\frac{\theta + 2\pi k_2}{n} + 2\pi j = \frac{\theta + 2\pi k_2}{n}.$$

Upon simplification, this is equivalent to $k_1 \equiv k_2 \pmod{n}$. As there are n distinct classes modulo n (to learn about modular arithmetic, see Volume 3), corresponding to

$$k = 0, 1, 2, \ldots, n-1,$$

each such k leads to a distinct root

$$w_k = \sqrt[n]{r}e^{\frac{\theta+2\pi k}{n}}.$$

As a geometric interpretation, notice that

$$\frac{w_{k+1}}{w_k} = e^{i\frac{2\pi}{n}},$$

where we define $w_n = w_0$. This means, each root can be found by rotating the preceding root around the origin by $\dfrac{2\pi}{n}$ radians. Subsequently, the n roots form a regular n-gon centered at the origin. ■

Students who want to learn more about the usage of complex numbers in geometry, especially in olympiad problems, should refer to Evan Chen's acclaimed EGMO book [6].

Definition 8.31. The n^{th} **roots of unity** are the n complex n^{th} roots of 1. They are given by

$$e^{i\frac{2\pi k}{n}}, \text{ for } k = 0, 1, 2, \ldots, n-1.$$

We will study these in Volume 3 in the context of number theory, specifically cyclotomic polynomials.

Chapter 9

Quadratics

> "If I have seen further it is by standing on the shoulders of giants."
>
> – *Isaac Newton, letter to Robert Hooke*

> "[Taniyama] was not a very careful person as a mathematician. He made a lot of mistakes. But he made mistakes in a good direction. I tried to imitate him. But I've realized that it's very difficult to make good mistakes."
>
> – *Goro Shimura*

We will begin studying polynomials by looking at quadratic functions. First we will see techniques for determining the roots of quadratics, including completing the square which produces the quadratic formula and factoring when it is feasible. We will also look at the shape of the graph of a quadratic function, which can be studied by completing the square and applying the trivial inequality.

9.1 Algebra

After linear functions in one variable, the functions in one variable that are one degree higher in complexity are quadratic functions.

Definition 9.1. A **quadratic function** in one variable is a function $f : \mathbb{C} \to \mathbb{C}$ of the form

$$f(x) = ax^2 + bx + c,$$

where $a \neq 0$ (this prevents it from being linear). The **coefficients** a, b, c are (for now) real constants called the **leading** or **quadratic coefficient**, **linear coefficient**, and **constant**, respectively. Similarly, the **terms** ax^2, bx, c are respectively called the **leading** or **quadratic term**, **linear term**, and **constant term**. By Definition 1.39, a **root** of a quadratic function f is a complex number z such that $f(z) = 0$. A **quadratic equation** is formed by setting a quadratic function equal to 0, which produces

$$ax^2 + bx + c = 0.$$

Solving for the roots of a quadratic function amounts to setting it equal to 0 as shown and determining all complex values of the variable x that satisfy this equation.

Lemma 9.2. If z is a complex number, then we say that w is a complex **square root** of z if $z = w^2$. Suppose a is a real number.

1. If $a = 0$, then the only complex square root of a is 0.
2. If $a > 0$, then the only two complex square roots of a are the real numbers $\pm\sqrt{a}$.
3. If $a < 0$, then the only two complex square roots of a are the pure imaginary numbers $\pm i\sqrt{-a}$ (note that $-a$ is positive).

Proof. Suppose $p + iq$ is a complex square root of the real number a. Then

$$a = (p + iq)^2 = (p^2 - q^2) + i(2pq).$$

Equating real and imaginary parts yields $a = p^2 - q^2$ and $0 = 2pq$, where the latter implies that at least one of p or q is 0. Note that $(-p - iq)^2 = a$ as well, so the negative of a square root of a is also a square root of a. Since one is the negative of the other and we are interested in finding all square roots, we may assume without loss of generality in our hunt that $p \geq 0$. Now we will treat a depending on its sign:

1. Suppose $a = 0$. Due to $pq = 0$, at least one of p or q is 0. Since $a = 0$, we get $|p| = |q|$ from $0 = p^2 - q^2$, meaning both are 0. Thus, 0 is the only square root of 0.
2. Suppose $a > 0$. Both p and q cannot be 0 because then $a = p^2 - q^2$ would be 0. Moreover, p cannot be 0, otherwise it would be true that $a = -q^2 < 0$. Since at least one of p or q is 0, we have $q = 0$. Thus, $p = \sqrt{a}$ and there are exactly two square roots $\pm\sqrt{a}$ of $a > 0$.
3. Suppose $a < 0$. Again, p and q cannot both be 0. Moreover, q cannot be 0, otherwise it would be true that $a = p^2 > 0$. Since at least one of p or q is 0, we have $p = 0$. Thus, $q = \sqrt{-a}$ and there are exactly two square roots $\pm i\sqrt{-a}$ of $a < 0$.

■

Definition 9.3. For convenience of notation, if $a < 0$ then we may choose to denote $\pm i\sqrt{-a}$ by $\pm\sqrt{a}$.

Example. We recommend being very careful with expressions of the form $\sqrt{a}$ for negative a as normal exponent laws do not necessarily hold for them. One famous example of a mistake is

$$-1 = i \cdot i = \sqrt{-1}\sqrt{-1} = \sqrt{(-1) \cdot (-1)} = 1.$$

Problem 9.4. Suppose $f(x) = ax^2 + bx + c$ is a quadratic function.

1. If $c = 0$, then find all complex roots of f in terms of a and b.
2. If $b = 0$, then find all complex roots of f in terms of a and c.

Now we will show a general method of determining the roots of a quadratic, called the quadratic formula, which may be found using a process called "completing the square."

Theorem 9.5 (Quadratic formula)**.** Let $f(x) = ax^2 + bx + c$ be a quadratic function. Then f has exactly two roots, which are

$$z = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

Proof. The standard derivation of the quadratic formula involves a step that requires casework on $\pm$ signs. We present a slightly modified method that circumvents the need to address these technicalities. A complex number z is a root of f if and only if

$$\begin{aligned} az^2 + bz + c &= 0 \\ 4a^2z^2 + 4abz + 4ac &= 0 \\ 4a^2z^2 + 4abz + b^2 &= b^2 - 4ac \\ (2az + b)^2 &= b^2 - 4ac \\ 2az + b &= \pm\sqrt{b^2 - 4ac} \\ z &= \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}. \end{aligned}$$

Each step was reversible, so there are exactly two (possibly equal) roots. ■

Note that, even though the quadratic formula asserts the existence of exactly two roots, one or both of the roots might be extraneous in word problems. This is because there can be additional constraints in the conditions stated in a word problem, such as negative values not being allowed in the x-axis, which might be interpreted as time.

Definition 9.6. The **discriminant** of a quadratic $f(x) = ax^2 + bx + c$ is the real number $b^2 - 4ac$, which appears under the square root in the quadratic formula.

Corollary 9.7. We can classify solution types of a quadratic function $f(x) = ax^2 + bx + c$ according to the sign of its discriminant $D = b^2 - 4ac$:

1. $D = 0$ if and only if the two roots are real and equal. In this case,

$$f(x) = a\left(x + \frac{b}{2a}\right)^2.$$

2. $D > 0$ if and only if the two roots are real and distinct.

3. $D < 0$ if and only if the two roots are non-real and distinct.

Proof. Since the discriminant of $f(x) = ax^2 + bx + c$ is the expression under the square root in the quadratic formula

$$\frac{-b \pm \sqrt{b^2 - 4ac}}{2a},$$

the classification of the solution types based on the sign of D follows immediately. In the $D = 0$ case, we can write the condition as $c = \dfrac{b^2}{4a}$, which leads to

$$\begin{aligned} ax^2 + bx + c &= ax^2 + bx + \frac{b^2}{4a} \\ &= a\left(x + \frac{b}{2a}\right)^2. \end{aligned}$$

■

Problem 9.8. If the coefficients of the quadratic function

$$f(x) = ax^2 + bx + c$$

are all rational, prove that the two roots are rational if and only if the discriminant D is the square of a rational number.

Corollary 9.9. Let $f(x) = ax^2 + bx + c$ be a quadratic function. If its roots are z_1 and z_2, then

$$f(x) = a(x - z_1)(x - z_2).$$

Proof. By the quadratic formula and the difference of squares factorization,

$$\begin{aligned} a(x - z_1)(x - z_2) &= a\left(x - \frac{-b + \sqrt{b^2 - 4ac}}{2a}\right)\left(x - \frac{-b - \sqrt{b^2 - 4ac}}{2a}\right) \\ &= a\left(\left(x + \frac{b}{2a}\right)^2 - \left(\frac{\sqrt{b^2 - 4ac}}{2a}\right)^2\right) \\ &= a\left(x^2 + \frac{b}{a}x + \frac{b^2}{4a^2} - \frac{b^2 - 4ac}{4a^2}\right) \\ &= a\left(x^2 + \frac{b}{a}x + \frac{c}{a}\right) \\ &= ax^2 + bx + c. \end{aligned}$$

■

Corollary 9.10 (Vieta's formulas for quadratics)**.** Two complex numbers z_1 and z_2 are the roots of a quadratic function

$$f(x) = ax^2 + bx + c$$

if and only if the following two equations hold:

$$\begin{aligned} z_1 + z_2 &= -\frac{b}{a}, \\ z_1 z_2 &= \frac{c}{a}. \end{aligned}$$

Proof. Suppose z_1 and z_2 are the roots of f. By the Corollary 9.9, it holds for all real x that

$$\begin{aligned} ax^2 + bx + c &= f(x) \\ &= a(x - z_1)(x - z_2) \\ &= a(x^2 - (z_1 + z_2)x + z_1 z_2) \\ &= ax^2 - a(z_1 + z_2)x + az_1 z_2. \end{aligned}$$

Putting all of the terms on one side, we get

$$(b + a(z_1 + z_2))x + (c - az_1 z_2) = 0$$

for all real x. Substituting in $x = 0$ yields $z_1 z_2 = c$ and then substituting in $x = 1$ yields $z_1 + z_2 = -\dfrac{b}{a}$.

Conversely, suppose z_1 and z_2 are two complex numbers that satisfy $z_1 + z_2 = -\dfrac{b}{a}$ and $z_1 z_2 = \dfrac{c}{a}$. For all real numbers x,

$$\begin{aligned} a(x - z_1)(x - z_2) &= ax^2 - a(z_1 + z_2)x + az_1 z_2 \\ &= ax^2 + bx + c \\ &= f(x). \end{aligned}$$

Substituting z_1 and z_2 into both sides yields $0 = f(z_1)$ and $0 = f(z_2)$, so z_1 and z_2 are the roots of f. ■

Readers are warned that it is a common mistake to forget to divide by a in Vieta's formulas. This might occur when one becomes accustomed to the monic case of $a = 1$. It is also easy to forget the negative sign in the formula for the sum of the roots.

Theorem 9.11. If the coefficients a, b, c of a quadratic function $f(x) = ax^2 + bx + c$ are integers, then there are some practical steps that can be taken to find the roots of f without using the quadratic formula.

1. If $a = 1$, then we can iterate through the pairs of integers (r, s) such that $rs = c$. If we happen to strike upon a pair (r, s) such that $r + s = -b$, then r and s are the roots of f. Subsequently,
$$f(x) = (x - r)(x - s).$$

2. More generally, even if $a \neq 1$, we can iterate through the pairs of integers (p, q) such that $pq = a$ and pairs of integers (r, s) such that $rs = c$. If we happen to strike upon pairs (p, q) and (r, s) such that $ps + qr = -b$, then $\dfrac{r}{p}$ and $\dfrac{s}{q}$ are the roots of f. Subsequently,
$$f(x) = (px - r)(qx - s).$$

This method of writing the quadratic function as a product of two linear functions is called **factorization** and the two linear functions are called **factors**. The technique can be extended to the case where some or all of the coefficients a, b, c are non-integer rationals, as we can set $ax^2 + bx + c$ equal to 0 and clear the denominators. Note that it is possible for the roots to not be rational even if the coefficients are rational, so it is not always possible to factor a quadratic into two linear factors that each have integer coefficients.

Proof. Let $f(x) = ax^2 + bx + c$ be a quadratic function with integer coefficients a, b, c.

1. Suppose $a = 1$. If r, s satisfy $rs = c$ and $r + s = -b$, then Vieta's formulas tell us that r and s are the roots of f. This leads to

$$f(x) = (x - r)(x - s).$$

2. If p, q and r, s satisfy

$$\begin{aligned} pq &= a, \\ ps + qr &= -b, \\ rs &= c, \end{aligned}$$

then dividing the second and third equations by a and using the first equation yields

$$\begin{aligned} \frac{s}{q} + \frac{r}{p} &= -\frac{b}{pq} = -\frac{b}{a}, \\ \frac{r}{p} \cdot \frac{s}{q} &= -\frac{c}{pq} = -\frac{c}{a}. \end{aligned}$$

By Vieta's formulas, $\dfrac{r}{p}$ and $\dfrac{s}{q}$ are the roots of f. This leads to

$$f(x) = a\left(x - \frac{r}{p}\right)\left(x - \frac{s}{q}\right) = (px - r)(qx - s).$$

■

Becoming efficient at factoring boils down to practice. There are tricks available as well, such as eliminating cases using signs, divisibility, and symmetry. The reader will find no dearth resources available elsewhere for engaging in this training.

Theorem 9.12. Let $z = x + iy$ be a non-real complex number. Then z has exactly two complex square roots, which are

$$\sqrt{\frac{\sqrt{x^2 + y^2} + x}{2}} + i \cdot \operatorname{sgn}(y)\sqrt{\frac{\sqrt{x^2 + y^2} - x}{2}}$$

and its negation.

Proof. Suppose a and b are real numbers such that $x+iy=(a+ib)^2$. Then it would also be true that $x+iy=(-a-ib)^2$. So finding one root automatically leads to a second, where the two are negatives of each other. This allows us to assume without loss of generality in our search that, if such a complex number $a+ib$ exists, then $a \geq 0$. Note that a cannot be 0 because then it would be true that $x+iy=(a+ib)^2=-b^2$, which is is real, contradicting the assumption that $z=x+iy$ is non-real. So $a>0$.
Now we expand

$$x+iy=(a+ib)^2=(a^2-b^2)+2abi$$

and equate real and imaginary parts to get

$$\begin{aligned} x &= a^2-b^2, \\ y &= 2ab. \end{aligned}$$

Since $a \neq 0$, we can divide by $2a$ to isolate b as $b=\dfrac{y}{2a}$ and substitute it into the first equation to get $x=a^2-\dfrac{y^2}{4a^2}$ or

$$4a^4-4xa^2-y^2=0.$$

By applying the quadratic formula to the "variable" a^2,

$$a^2=\frac{x \pm \sqrt{x^2+y^2}}{2}.$$

Recall from Theorem 8.13 that $|z| \geq |\mathrm{Re}(z)|$, so $\sqrt{x^2+y^2} \geq |x| \geq x$. This forces us to choose the positive sign in our expression from the quadratic formula, as otherwise a^2 would be non-positive. Then, since $a>0$, we can take the positive square root of a^2 to get

$$a=\sqrt{\frac{\sqrt{x^2+y^2}+x}{2}}.$$

Finally, by rationalizing the denominator below,

$$b=\frac{y}{2a}=\frac{y}{\sqrt{2}\sqrt{\sqrt{x^2+y^2}+x}}=\frac{y}{\sqrt{y^2}}\cdot\sqrt{\frac{\sqrt{x^2+y^2}-x}{2}}.$$

By Theorem 5.8,

$$\frac{y}{\sqrt{y^2}}=\frac{y}{|y|}=\mathrm{sgn}(y)$$

and we do not have to consider the possibility of $y=0$ because $z=x+iy$ is non-real. Therefore,

$$a+ib=\sqrt{\frac{\sqrt{x^2+y^2}+x}{2}}+i\cdot\mathrm{sgn}(y)\sqrt{\frac{\sqrt{x^2+y^2}-x}{2}}.$$

The reader should algebraically verify that $(a+ib)^2=x+iy$. ■

Definition 9.13. Originally, we defined the coefficients of a quadratic function to be real constants. We can extend this to allow them to be complex constants. So we redefine a **quadratic function** to be a function $f : \mathbb{C} \to \mathbb{C}$ defined by $f(x) = ax^2 + bx + c$ for all complex numbers x, where a, b, c are complex constants. We are redefining the scope of quadratic functions now because Theorem 9.12 shows that every complex number has exactly two complex square roots; consequently, the quadratic formula and Vieta's formulas hold for quadratic functions with complex coefficients as well. However, the classification of solution types based on the sign of the discriminant depends on the coefficients being real.

Problem 9.14. Suppose a, b, c are positive real numbers such that $a^2 - b^2c$ is non-negative. Show that the square root decomposition

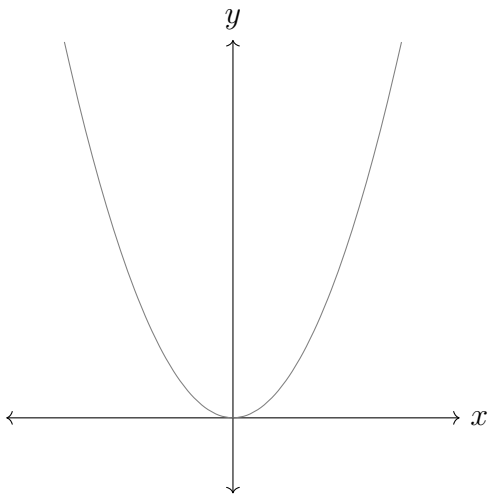
$$\sqrt{a \pm b\sqrt{c}} = \sqrt{\frac{a + \sqrt{a^2 - b^2c}}{2}} \pm \sqrt{\frac{a - \sqrt{a^2 - b^2c}}{2}},$$

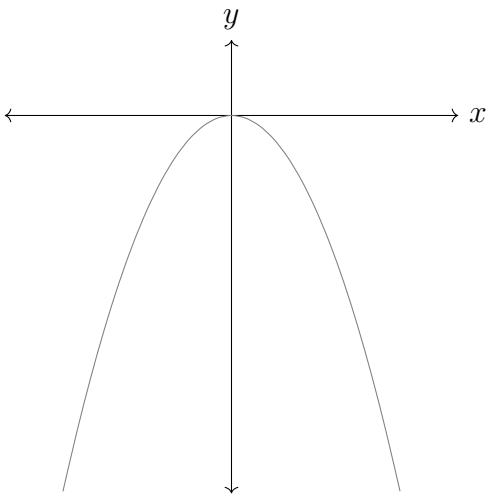
holds, where the $\pm$ signs on either side of the equation are both positive or both negative. In particular, if a, b, c are rational and $a^2 - b^2c$ is the square of a rational number, then we call this decomposition a "denesting" of square roots.

9.2 Graphing

Now, we will be assuming that the coefficients of the quadratic functions involved are real and that their domains are the real numbers. This will allow us to speak of the maximum or minimum possible output of a quadratic function, along with the location in the domain where this extreme value is attained.

Theorem 9.15. The quadratic function $f : \mathbb{R} \to \mathbb{R}$, defined by $f(x) = ax^2 + bx + c$ attains a global extremum of $-\dfrac{b^2 - 4ac}{4a}$ at $x = -\dfrac{-b}{2a}$. By extremum, we mean that $f\left(-\dfrac{-b}{2a}\right)$ is a minimum or maximum of $f(x)$ over all real x. This extremum is a minimum if $a > 0$ and a maximum if $a < 0$. When there is a global minimum we say that the quadratic **opens up**, and when there is is a global maximum we say that the quadratic **opens down**.





Proof. Following the proof of Theorem 9.5, we complete the square as follows:

$$\begin{aligned} f(x) &= ax^2 + bx + c \\ &= \frac{1}{4a}(4a^2x^2 + 4abx) + c \\ &= \frac{1}{4a}(4a^2x^2 + 4abx + b^2) - \frac{b^2}{4a} + c \\ &= \frac{1}{4a}(2ax + b)^2 - \frac{b^2 - 4ac}{4a} \\ &= a\left(x + \frac{b}{2a}\right)^2 - \frac{b^2 - 4ac}{4a}. \end{aligned}$$

Now we can use the trivial inequality. If $a > 0$, then

$$f(x) = a\left(x + \frac{b}{2a}\right)^2 - \frac{b^2 - 4ac}{4a} \geq -\frac{b^2 - 4ac}{4a},$$

where equality holds if and only if $x = -\dfrac{b}{2a}$. If $a < 0$, then

$$f(x) = a\left(x + \frac{b}{2a}\right)^2 - \frac{b^2 - 4ac}{4a} \leq -\frac{b^2 - 4ac}{4a},$$

where equality holds if and only if $x = -\dfrac{b}{2a}$ again. ∎

As a side note, the curves formed by quadratic functions, called a parabola, turn out to be a part of a family of curves called conics, formed by taking cross sections of cones.

Definition 9.16. Since the quadratic $f(x) = ax^2 + bx + c$ attains a global extremum of $-\dfrac{b^2 - 4ac}{4a}$ at $x = -\dfrac{-b}{2a}$, we call the point

$$\left(-\frac{b}{2a}, -\frac{b^2 - 4ac}{4a}\right)$$

on the quadratic's graph its **vertex**. It is for this reason that we call

$$a\left(x + \frac{b}{2a}\right)^2 - \frac{b^2 - 4ac}{4a}$$

the **vertex form** of f. The ordinary $ax^2 + bx + c$ form is called the **standard form**.

Corollary 9.17. Every quadratic function is a result of left-composing and right-composing the function $p(x) = x^2$ with linear functions.

Proof. The result is evident from the vertex form. To be formal, let $f(x) = ax^2 + bx + c$ be a quadratic function. We define the linear functions

$$\begin{aligned} g(x) &= ax - \frac{b^2 - 4ac}{4a}, \\ h(x) &= x + \frac{b}{2a}. \end{aligned}$$

Then $g \circ p \circ h = f$, thanks to the vertex form of f. As a result, the graph of f may be obtained by transforming the graph of the simple function $p(x) = x^2$ via linear deformations. ■

Corollary 9.18. The graph of the quadratic function $f(x) = ax^2 + bx + c$ is symmetric across the vertical line $x = -\dfrac{b}{2a}$ that runs through its vertex. This is the called the **line of symmetry** of the curve.

Proof. Writing f in vertex form and substituting in $x = -\dfrac{b}{2a} \pm r$ for any real r shows that

$$f\left(-\frac{b}{2a} + r\right) = ar^2 - \frac{b^2 - 4ac}{4a} = f\left(-\frac{b}{2a} - r\right).$$

This helps us to graph quadratics because the left or right half of the graph can be used to produce the other half. ■

We remark that non-linear equations can sometimes be coaxed into being amenable to techniques pertaining to linear or quadratic equations, thereby allowing us to use the methods developed for solving linear or quadratic equations. Common techniques include:

1. Substitute variables for expressions, such as expressions involving square roots, reciprocals, powers, or other oft-repeating expressions. After finding values of the new variables, we can set the values equal to the expressions for the new variables in terms of the old variables.

2. If there are n^{th} roots, it can help to take powers of the equation, possibly after some rearrangement of the terms in the equation.

Moreover, just like systems of linear equations, there can be systems of equations involving quadratics functions. The usual methods of elimination and substitution remain effective.

Chapter 10

Polynomials

> "A mathematician who is not also something of a poet will never be a complete mathematician."
>
> – *Karl Weierstrass*

> "Good, [the student who dropped out to study poetry] did not have enough imagination to become a mathematician."
>
> – *David Hilbert*

> "Pure mathematics is, in its way, the poetry of logical ideas."
>
> – *Albert Einstein*

Having seen constant, linear, and quadratic functions, we will now generalize them into polynomials. We will study the main parts of a polynomial, which are its degree, coefficients, and roots. Then we will take a look at the division of polynomials, which is the polynomial analogue of integer division with remainder. Thirdly, we will study rational functions, which are essentially one polynomial divided by another without remainder, but with domain restrictions that are necessary to avoid division by 0. Afterwards, we will look at a general version of Vieta's formulas, and a theorem of Girard and Newton. We will end the chapter by looking at factorizations of multivariable polynomials, including several well-known remarkable ones.

10.1 Degree, Coefficients, and Roots

One way to motivate the definition of polynomials is to consider the functions or expressions that are achieved by multiplying and adding together several linear functions or expressions in the same variable. The general form that will be shown is how those functions would be presented if all factors were multiplied out and like terms were collected.

Definition 10.1. A **term** is an expression of the form ax^d, where a is a complex constant called its **coefficient**, x is the **variable** or **indeterminate**, and d is a non-negative integer called its **degree**. A univariate **polynomial** is a sum of a finite number of such terms, where terms of the same degree (called "like terms") are typically collected using the distributive law so that each term in the polynomial has a different degree. A polynomial looks like

$$a_n x^n + a_{n-1} x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0.$$

Although complex numbers can be substituted in for the variable in a polynomial and the expression can then be evaluated using arithmetic operations, we caution the reader to not

automatically interpret polynomials as functions. The definition of a "formal" polynomial is an infinite sequence of coefficients indexed by $\mathbb{Z}_+$,

$$(a_k)_{k=0}^{\infty} = (a_0, a_1, a_2, a_3, \ldots)$$

such that all elements of sufficiently high index are 0. We will make this idea more formal in the context of general generating functions when we study combinatorics in Volume 2.

Definition 10.2. In a polynomial

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0,$$

the term $a_n x^n$ with the highest degree is not allowed to have $a_n = 0$, and the degree n of this term is called the **degree** of the polynomial. It is denoted by $\deg f = n$. If all the coefficients of a polynomial are 0, then its degree is said to be $-\infty$, in order to be consistent with some relations involving degrees of polynomials that we will explore in Theorem 10.6, and we call that polynomial the **zero polynomial**. For ease of notation, if $\deg f = n$, then we define coefficients corresponding to terms with degree higher than n to be 0. That is, $a_k = 0$ for $k > \deg f$.

Definition 10.3. In a polynomial

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0,$$

we call $a_n \neq 0$ the **leading coefficient** and a_0 the **constant term**.

Example. A non-zero constant a is a polynomial of degree 0, a linear expression $ax + b$ is a polynomial of degree 1, and a quadratic expression $ax^2 + bx + c$ is a polynomial of degree 2.

Definition 10.4. Two polynomials are said to be **equal** if they have the same degree and the coefficients of the two k^{th} degree terms are equal for all k, assuming all like terms have been collected. If we have two different expressions for the same polynomial (this often happens in algebra-based combinatorics), then we can set the coefficients of the corresponding terms equal to each other; this technique is called **comparing coefficients** and we will see more of it in Volumes 2 and 3. A separate, but related, concept is that of being **equal everywhere**, which means the two polynomials take on the same value for every complex input (see Definition 10.8 for polynomials as functions). So being equal is in reference to formal polynomials, whereas being equal everywhere is in reference to polynomials as functions.

Definition 10.5. We define some specific binary and unary operations on formal polynomials in the following way so that the defining equations would be satisfied if any complex number were substituted in for x. Let there be two polynomials

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0,$$
$$g(x) = b_m x^m + b_{m-1} x^{m-1} + \cdots + b_2 x^2 + b_1 x + b_0.$$

1. Addition of polynomials is done term-by-term, so that
$$(f+g)(x) = \sum_{k=0}^{\max(n,m)} (a_k + b_k)x^k.$$

2. Negation of a polynomial simply takes the negative of each coefficient so that
$$(-f)(x) = (-a_n)x^n + (-a_{n-1})x^{n-1} + \cdots + (-a_2)x^2 + (-a_1)x + (-a_0).$$
As such, we can define subtraction of polynomials as
$$f - g = f + (-g).$$
Note that 0 is an additive identity of polynomials and $-f$ is an additive inverse of f.

3. Multiplication of polynomials is defined to be compatible with the arithmetic distributive law, so that
$$(f \cdot g)(x) = \sum_{k=0}^{n+m} c_k x^k,$$
where
$$c_k = a_0 b_k + a_1 b_{k-1} + \cdots + a_{k-1} b_1 + a_k b_0.$$

4. Composition of polynomials is done by applying the distributive law and using the rules for adding and multiplying polynomials. It results in a polynomial, though we are unaware of a neat formula for the coefficients; interested readers may search for Faà di Bruno's formula. We use the ordinary composition notation $f \circ g$ to denote the composition of polynomials.

Theorem 10.6. For each type of operation on polynomials that we have described, we can comment on the degree of the resulting polynomials. If f and g are polynomials, then:

1. $\deg(-f) = \deg(f)$
2. $\deg(f \pm g) \leq \max(\deg f, \deg g)$, and, due to potential cancellation of terms, it is possible to achieve any of the values that can be degrees, which are $-\infty$ and non-negative integers. If $\deg g < \deg f$, we can be more precise and say $\deg(f \pm g) = \deg f$.
3. $\deg(f \cdot g) = \deg f + \deg g$
4. $\deg(f \circ g) = (\deg f) \cdot (\deg g)$ if f and g are non-constant polynomials

Proof. It is not difficult to see that the first result holds in general, the next two results hold for non-zero polynomials, and that the fourth result holds for non-constant polynomials. What we will discuss is how the second and third results force the convention that the degree of the 0 polynomial is $-\infty$. Let f be a non-zero polynomial and g be the 0 polynomial. For the product identity to be satisfied, it must be true that
$$\deg 0 = \deg(f \cdot g) = \deg f + \deg g = \deg f + \deg 0.$$

Since $\deg f$ could be any non-negative integer, there is no real value of $\deg 0$ that can satisfy this identity. However, if we make reasonable extensions of the rules of arithmetic to include $\pm\infty$ as "numbers," then these are candidates for the value of $\deg 0$. Now, in order for the addition identity to hold, we must have

$$\deg 0 = \deg(f - f) \leq \max(\deg f, \deg f) = \deg f.$$

This excludes the possibility of $\deg 0$ being $+\infty$, so the only candidate is $-\infty$. Indeed, it produces the following reasonable results for all polynomials f, including $f = 0$, under extended interpretations of arithmetic:

- $f = f + 0$ and $\deg f \leq \max(\deg f, -\infty) = \deg f$
- $0 = f - f$ and $-\infty \leq \max(\deg f, \deg(-f)) = \deg f$
- $0 = 0 \cdot f$ and $-\infty = -\infty + \deg f$

■

Problem 10.7. Prove that, if f is a non-constant polynomial then left-composing and/or right-composing f with non-constant linear polynomials like $ax + b$ leaves its degree unchanged.

Definition 10.8. A polynomial

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0$$

may be interpreted as a function $f : \mathbb{C} \to \mathbb{C}$ where we replace x with complex numbers z and evaluate

$$f(z) = a_n z^n + a_{n-1} z^{n-1} + \cdots + a_2 z^2 + a_1 z + a_0$$

using the rules of complex arithmetic. In compliance with the definition of a root of a general real- or complex-valued function (Definition 1.39), if z is a complex number such that $f(z) = 0$, then we call z a **root** or **solution** of f.

Problem 10.9. Show that, if all the coefficients of a polynomial are real and positive, then any real root must be negative.

Problem 10.10. Show that a polynomial has 0 as a root if and only if its constant term is 0.

Example 10.11. Let $f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0$ be a polynomial where $a_0 \neq 0$. Show that z is a root of the polynomial with the reversed sequence of coefficients,

$$g(x) = a_0 x^n + a_1 x^{n-1} + \cdots + a_{n-2} x^2 + a_{n-1} x + a_n,$$

if and only if $\dfrac{1}{z}$ is a root of f.

Solution. The fact that f is of degree n lets us know that $a_n \neq 0$. Since $a_0 \neq 0$ and $a_n \neq 0$, the roots of f and g do not include 0, so we can take the reciprocal of any root of either polynomial. Suppose z is a root of g. Then

$$a_0 z^n + a_1 z^{n-1} + \cdots + a_{n-2} z^2 + a_{n-1} z + a_n = 0.$$

Dividing by z^n yields

$$a_0 + a_1 \left(\frac{1}{z}\right) + \cdots + a_{n-2} \left(\frac{1}{z}\right)^{n-2} + a_{n-1} \left(\frac{1}{z}\right)^{n-1} + a_n \left(\frac{1}{z}\right)^n = 0,$$

which shows that $\dfrac{1}{z}$ is a root of f. Since this step is reversible by multiplying by z^n, the other direction holds as well.
An interesting application of this result is to polynomials

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0$$

that are palindromic in the sense that $a_k = a_{n-k}$ for each index $0 \leq k \leq n$. Then z is a root of f if and only if $\dfrac{1}{z}$ is a root of f, which gives us a way of finding a new root from an existing root in this special case. We will see some other methods of finding new roots from old roots in Theorem 10.16 and Problem 10.18. As another side note, it may be readily verified that a polynomial f is palindromic if and only if it satisfies

$$f(x) = x^n \cdot f\left(\frac{1}{x}\right).$$

We will use this fact when studying cyclotomic polynomials in Volume 3. ■

Theorem 10.12 (Rational root theorem)**.** If $\dfrac{p}{q}$ is a reduced fraction that is a rational root of a polynomial

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0$$

with all integer coefficients such that $a_0 \neq 0$, then p divides a_0 and q divides a_n. Since there are finitely many divisors of a_0 and a_n, we can iterate through all such rational numbers $\dfrac{p}{q}$ and test whether $f\left(\dfrac{p}{q}\right) = 0$ to find all the rational roots. (Note that p and q may be negative divisors of a_0 and a_n, respectively.)

Proof. Non-zero constant polynomials have no roots, so we can assume that $\deg f \geq 1$ so that $n \neq 0$. Suppose $\dfrac{p}{q}$ is a rational root of f. Then substituting $\dfrac{p}{q}$ into f yields

$$a_n \left(\frac{p}{q}\right)^n + a_{n-1} \left(\frac{p}{q}\right)^{n-1} + \cdots + a_2 \left(\frac{p}{q}\right)^2 + a_1 \left(\frac{p}{q}\right) + a_0 = 0.$$

We can clear all the denominators by multiplying both sides by q^n, which yields

$$a_n p^n + a_{n-1} p^{n-1} q + \cdots + a_2 p^2 q^{n-2} + a_1 p q^{n-1} + a_0 q^n = 0.$$

Since p divides the leftmost n terms on the left side as well as the right side 0, p must also divide a_0q^n. But we assumed that p is relatively prime to q, so p must divide a_0 by Gauss's divisibility lemma (see Volume 3). Similarly, q divides the rightmost n terms on the left side as well as the right side 0, so q must also divide a_np^n. As q is relatively prime to p, it must be true that q divides a_n by Gauss's divisibility lemma. ■

Corollary 10.13. Suppose

$$f(x) = a_nx^n + a_{n-1}x^{n-1} + \cdots + a_2x^2 + a_1x + a_0$$

is a non-zero polynomial. From the fact that f is non-zero, we know that there exists a non-negative integer k such that $a_k \neq 0$ with $a_0 = a_1 = \cdots = a_{k-1} = 0$ (if $k = 0$ then the list $a_0, a_0, \ldots, a_{k-1}$ is empty). Let

$$g(x) = a_nx^{n-k} + a_{n-1}x^{n-1-k} + \cdots + a_{k+2}x^2 + a_{k+1}x + a_k$$

be the polynomial such that $f(x) = x^k \cdot g(x)$. Then the non-zero roots of f are precisely the roots of g.

Proof. Let g be defined as $f(x) = x^k \cdot g(x)$ where x^k is the maximum power of x that can be factored out of f. If z is a non-zero root of f then

$$0 = f(z) = z^kg(z) \implies g(z) = 0.$$

And if z is a root of g then

$$f(z) = z^kg(z) = z^k \cdot 0 = 0.$$

As a consequence, even if the constant term of f is 0, we can still apply the rational root theorem to g to find all the rational roots of f. ■

Definition 10.14. A **monic** polynomial is a polynomial whose leading coefficient is 1.

There are two more important implications of the rational root theorem, both of which we label under the heading of the integer root theorem.

Corollary 10.15 (Integer root theorem)**.** Let $f(x) = a_nx^n + a_{n-1}x^{n-1} + \cdots + a_2x^2 + a_1x + a_0$ be a polynomial with integer coefficients and $a_n \neq 0$. Then:

1. Any non-zero integer root r of f divides a_k, where k is the smallest index i such that a_i is non-zero. As a consequence, r always divides a_0.

2. If f is monic, then all of its rational roots are integers.

Combining these two results, the only candidates for non-zero rational roots of a monic polynomial with integer coefficients are the factors of a_k, which is a finite list.

Proof. Let $f(x) = a_nx^n + a_{n-1}x^{n-1} + \cdots + a_2x^2 + a_1x + a_0$ be a polynomial with integer coefficients. Let the polynomial g be defined by $f(x) = x^k \cdot g(x)$, where x^k is the maximum power of x that can be factored out of f. This means that $a_k \neq 0$ and $a_0 = a_1 = \cdots = a_{k-1} = 0$. Then:

1. Suppose r is a non-zero integer root of f. We know that r is a root of

$$g(x) = a_n x^{n-k} + a_{n-1}x^{n-1-k} + \cdots + a_{k+2}x^2 + a_{k+1}x + a_k.$$

Since $a_k \neq 0$, we can apply the rational root theorem to the root $\frac{r}{1}$ and say that 1 divides the leading coefficient a_n (which is always true) and r divides a_k, as desired.

If $a_0 = 0$ then r automatically divides a_0. If $a_0 \neq 0$ then $k = 0$ and r still divides $a_0 = a_k$.

2. Suppose f is monic so that $a_n = 1$. Let $r = \frac{p}{q}$ be a reduced rational root of f. If $r = 0$ then r is already an integer, so we can assume that r is a non-zero root. Then $\frac{p}{q}$ is a root of the polynomial g define above, and the rational root theorem says that p divides a_k and q divides 1. So $q = \pm 1$, which means $\frac{p}{q} = \pm p$ is an integer.

■

Notice that the quadratic formula tells us that, if a quadratic has two non-real roots, then the two roots are complex conjugates. This result generalizes.

Theorem 10.16. If f is a polynomial with real coefficients, and z is a complex root of f then $\overline{z}$ is also a root of f.

Proof. This results follows from the fact that if z and w are complex numbers then $\overline{z+w} = \overline{z} + \overline{w}$ and $\overline{z \cdot w} = \overline{z} \cdot \overline{w}$, and that if a is real then $\overline{a} = a$. If

$$f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0$$

is a polynomial with real coefficients and z is a complex root of f, then

$$\begin{aligned} f(\overline{z}) &= a_n(\overline{z})^n + a_{n-1}(\overline{z})^{n-1} + \cdots + a_2(\overline{z})^2 + a_1(\overline{z}) + a_0 \\ &= a_n\overline{z^n} + a_{n-1}\overline{z^{n-1}} + \cdots + a_2\overline{z^2} + a_1\overline{z} + a_0 \\ &= \overline{a_n z^n} + \overline{a_{n-1}z^{n-1}} + \cdots + \overline{a_2 z^2} + \overline{a_1 z} + \overline{a_0} \\ &= \overline{a_n z^n + a_{n-1}z^{n-1} + \cdots + a_2 z^2 + a_1 z + a_0} \\ &= \overline{f(z)} \\ &= \overline{0} = 0. \end{aligned}$$

So $\overline{z}$ is also a root of f. ■

Definition 10.17. Let a and b be rational numbers and c be a positive rational number that is not the square of a rational number. Then we say that the **radical conjugate** of $a + b\sqrt{c}$ is $\overline{a + b\sqrt{c}} = a - b\sqrt{c}$. We leave it to the reader to show that this definition makes sense because the representation $a + b\sqrt{c}$ is unique (meaning if $a_1 + b_1\sqrt{c} = a_2 + b_2\sqrt{c}$, then $a_1 = a_2$ and $b_1 = b_2$.) This is not to be confused with the complex conjugate.

Problem 10.18. Let $r = a + b\sqrt{c}$ where a and b are rational numbers and c is a positive rational number that is not the square of a rational number. If f is a polynomial with rational coefficients and r is a root of f, then show that the radical conjugate $\underline{r}$ of r is also a root of f.

Theorem 10.19 (Equality of polynomials as functions)**.** Let

$$f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0$$

be a polynomial with complex coefficients. Then the following are equivalent conditions on f.

1. All coefficients of f are 0.
2. For all complex z, we get $f(z) = 0$.
3. For $1 + \deg f$ many distinct complex z, we have $f(z) = 0$.

Consequently, the following conditions are equivalent on complex polynomials f and g:

a. The degrees of f and g are equal, and the coefficients of corresponding terms match.

b. For all complex z, we get $f(z) = g(z)$, meaning they are equal everywhere.

c. For $1 + \max(\deg f, \deg g)$ many distinct complex z, we have $f(z) = g(z)$.

Proof. For the first collection of equivalence statements, is easy to see that (1) $\implies$ (2) $\implies$ (3), so it remains to be shown that (3) $\implies$ (1). Let $\deg f = n$, and let $z_0, z_1, z_2, \ldots, z_n$ be $n+1$ distinct complex numbers at which f evaluates to 0. This system of equations may be written in matrix form as

$$\begin{bmatrix} 1 & z_0 & z_0^2 & \cdots & z_0^{n-1} & z_0^n \\ 1 & z_1 & z_1^2 & \cdots & z_1^{n-1} & z_1^n \\ 1 & z_2 & z_2^2 & \cdots & z_2^{n-1} & z_2^n \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & z_{n-1} & z_{n-1}^2 & \cdots & z_{n-1}^{n-1} & z_{n-1}^n \\ 1 & z_n & z_n^2 & \cdots & z_n^{n-1} & z_n^n \end{bmatrix} \cdot \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_{n-1} \\ a_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}$$

The square matrix on the far left is a well-known matrix, called the Vandermonde matrix, that has a non-zero determinant which can be beautifully computed. We will analyze it when studying special cases of the theory of linear recurrences in Volume 2. For now, what we need to know is that the non-zero determinant makes the matrix invertible, and multiplying the left sides of both sides of the equation by this inverse yields

$$(a_0, a_1, a_2, \ldots, a_{n-1}, a_n) = (0, 0, 0 \ldots, 0, 0).$$

For a separate proof that (3) $\implies$ (2), see Theorem 10.34.
For the second collection of equivalent statements, it is again easy to see that $(a) \implies (b) \implies (c)$. For $(c) \implies (a)$, we let $h = f - g$ so that $\deg h \leq \max(\deg f, \deg g)$ and $h(z) = 0$ for $1 + \max(\deg f, \deg g)$ many distinct complex z. By the earlier implication (3) $\implies$ (1), we get that all coefficients of $h = f - g$ are zero. This implies that $\deg f = \deg g$ and that the coefficients of f all match those of g. ∎

Example 10.20. Find an example of a ring or field in which there exists a polynomial that has some non-zero coefficients, but it is still equal to the additive identity, that is the "zero" field or ring element, for all inputs.

Solution. Let p be a prime. By Fermat's little theorem from number theory (see Volume 3),

$$x^p - x \equiv 0 \pmod{p}$$

for every integer x. Therefore, it is possible for a polynomial with coefficients and inputs from $\mathbb{Z}_p$ to have non-zero coefficients while still being the identically "zero" function. ■

10.2 Division and Factoring

Definition 10.21. So far, we have found results for univariate polynomials that have integer, rational, real, and complex coefficients. The sets of such formal polynomials are denoted by $\mathbb{Z}[x], \mathbb{Q}[x], \mathbb{R}[x]$, and $\mathbb{C}[x]$, respectively, and they are called **polynomial rings** over $\mathbb{S}$ where $\mathbb{S}$ is the set from which the coefficients are taken.

While we can define subtraction of polynomials easily, the polynomials rings $\mathbb{Z}[x]$, $\mathbb{Q}[x]$, $\mathbb{R}[x]$, and $\mathbb{C}[x]$ do not form fields because most polynomials do not have a multiplicative inverse that is a polynomial. As such, there is no way of always dividing a polynomial into another polynomial to get a polynomial. However, in some cases a polynomial does divide another polynomial, and in other cases we can do what is called Euclidean division, which is division with a remainder. In Section 10.3, we will look at functions produced by dividing a polynomial function by a polynomial function.

Definition 10.22. Let $\mathbb{S}$ denote one of the sets $\mathbb{Z}, \mathbb{Q}, \mathbb{R}$, or $\mathbb{C}$. If f and g are polynomials in $\mathbb{S}[x]$ then we say that f is **divisible** by g or that g is a **factor** of f or that g **divides** f if there exists a polynomial $h \in \mathbb{S}[x]$ such that $f = g \cdot h$.

Theorem 10.23 (Polynomial Euclidean division)**.** Let $\mathbb{F}$ be a field ($\mathbb{Q}, \mathbb{R}$, and $\mathbb{C}$ are examples of fields, but $\mathbb{Z}$ is not). For any $f \in \mathbb{F}[x]$ and any non-zero $g \in \mathbb{F}[x]$, there exists a $q \in \mathbb{F}[x]$ and an $r \in \mathbb{F}[x]$ such that

$$f = gq + r, \deg r < \deg g.$$

In this notation, if $\deg f \geq \deg g$ then $\deg q = \deg f - \deg g$. Moreover, given this f and non-zero g, the q and r are unique. We call f the **dividend**, g the **divisor**, q the **quotient**, and r the **remainder**. (By the same proof, the same result holds with every instance of $\mathbb{F}$ replaced with $\mathbb{Z}$ if g is monic. For a precise statement pertaining to the coefficients living in rings like $\mathbb{Z}$ instead of fields, read the proof of Theorem 10.54.)

Proof. Let $\mathbb{F}$ be a field, $f \in \mathbb{F}[x]$ and g be a non-zero element of $\mathbb{F}[x]$. First we will show that q and r exist, and then we will prove that they are unique. To prove existence, we will consider two cases:

1. If $\deg f < \deg g$, including the possibility that $f = 0$, then we can choose $q = 0$ and $r = f$ to get
$$f = g \cdot 0 + f = gq + r, \deg r < \deg g.$$

2. The more involved case is where $\deg f \geq \deg g$. We will proceed by strong induction on $\deg f$. Since $\deg g \geq 0$, the base case is $\deg f = 0$, meaning f is a non-zero constant a and g is a non-zero constant b. Since we must have $\deg r < \deg g$, the only choice for r is 0. Then $q = \dfrac{a}{b}$ and we get

$$f = a = b \cdot \frac{a}{b} + 0 = gq + r, \deg r < \deg g.$$

The strong induction hypothesis is that the assertion holds whenever $\deg f \leq n-1$ for some $n - 1 \geq 0$. Now suppose $\deg f = n$. Let

$$f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0,$$

and let $\deg g = m \leq n$ so that

$$g(x) = b_m x^m + b_{m-1}x^{m-1} + \cdots + b_2 x^2 + b_1 x + b_0.$$

If we could find a polynomial $h(x)$ such that $\deg(f - hg) < \deg f$, then either $f = hg + 0$ and we are done, or we can apply the induction hypothesis to $f - hg$ to get a quotient q_0 and a remainder r such that $f - hg = gq_0 + r$ with $\deg r < \deg g$. This would give us

$$f = g(h + q_0) + r,$$

so we could take $q = h + q_0$. To find h, we note that the leading term of hg must equal the leading term of f, so we let $h = \dfrac{a_n}{b_m}x^{n-m}$, which works. Notice that it is precisely in the division by b_m that we require the existence of multiplicative inverses in fields, which does not necessarily hold in arbitrary rings.

At this point, we compute can $\deg q$, now that we know that q and r exist. Since $\deg r < \deg g \leq \deg f$, we find that

$$\deg g + \deg q = \deg(g \cdot q) = \deg(f - r) = \deg f,$$

so $\deg q = \deg f - \deg g$.

Now we prove the uniqueness of q and r. Suppose $f = gq_1 + r_1$ with $\deg r_1 < \deg g$, and $f = gq_2 + r_2$ with $\deg r_2 < \deg g$. Then we get $gq_1 + r_1 = gq_2 + r_2$ or

$$g \cdot (q_1 - q_2) = r_2 - r_1.$$

For the sake of contradiction, suppose $q_1 - q_2 \neq 0$ so that $\deg(q_1 - q_2) \geq 0$. Then

$$\deg(r_2 - r_1) = \deg(g \cdot (q_1 - q_2)) = \deg g + \deg(q_1 - q_2) \geq \deg g.$$

However, we also know that

$$\deg(r_2 - r_1) \leq \max(\deg r_2, \deg r_1) < \deg g.$$

This is a contradiction so we must have $q_1 = q_2$, which implies that $r_1 = r_2$. ■

The Euclidean division theorem for polynomials implicitly tells us that a quotient and a remainder exist, but it does not tell us how to find them. Sometimes we want to find them explicitly, which is where polynomial **long division** comes into play. Assuming that the reader is familiar with long division for integers, it will suffice to show an example of how polynomial long division works out. The idea is the same: if we are dividing f by g, then slice away at the terms of f from highest degree to lowest degree using multiples of g. Let us see an example.

Example 10.24. Find the quotient and remainder when $2x^3 + 11x + 6$ is divided by $x - 8$.

Solution. First we proceed by long division.

$$\begin{array}{rr|lllllll} & & 2x^2 & - & 16x & + & 117 & & \\ \cline{3-8} x+8 & | & 2x^3 & + & 0x^2 & - & 11x & + & 6 \\ & & 2x^3 & + & 16x^2 & & & & \\ \cline{3-5} & & & & -16x^2 & - & 11x & & \\ & & & & -16x^2 & - & 128x & & \\ \cline{5-8} & & & & & & 117x & + & 6 \\ & & & & & & 117x & + & 936 \\ \cline{7-8} & & & & & & & & -930 \end{array}$$

The trouble with long division is that, well, it can take a long time to write out. Thankfully, there is a compressed way of writing down the same process when the divisor is a monic linear polynomial. It is called **synthetic division** we write it out as follows.

$$\begin{array}{r|llll} -8 & 2 & 0 & -11 & 6 \\ \cline{2-4} & & & & \end{array} \quad \longrightarrow \quad \begin{array}{r|llll} -8 & 2 & 0 & -11 & 6 \\ & & -16 & 128 & -936 \\ \cline{2-5} & 2 & -16 & 117 & -930 \end{array}$$

So the quotient is $2x^2 + 16x + 139$ and the remainder is 1118. The general process is:

1. Place the negation of the constant term of the divisor at the top-left corner.

2. Place the coefficients of the dividend in the top row, including any zero coefficients, sequenced from left to right in descending order of the degrees of the corresponding terms.

3. Each number in the second row is found by multiplying the number to its bottom-left by the number from the first step at the top-left corner.

4. Each number in the bottom row is the sum of the two numbers directly above it in the same column.

For example, the sequence in which we filled the entries above is: 2, the -16 in the second row, the -16 in the bottom row, $128, 117, -936, -930$.
We would like to caution the reader in several ways:

- As described, the method only works when the divisor is a monic linear polynomial.
- If the divisor is $x - c$ then we place c in the top-left corner of the synthetic division box, not $-c$.
- If there are "missing" coefficients in the dividend, we still have to place them in the first line of the synthetic division box as 0.

As a bonus, synthetic division can be made to work for division by non-monic linear polynomials as well. If we are dividing f by $ax - b$ then first we use synthetic division on the problem of dividing f by $x - \frac{b}{a}$. Then

$$f(x) = \left(x - \frac{b}{a}\right) q(x) + r(x),$$

where q is the quotient and r is the remainder. This implies that

$$f(x) = (ax - b)\frac{q(x)}{a} + r(x),$$

so the true quotient is $\frac{q(x)}{a}$ and r is still the remainder. This follows from the fact that quotients and remainders are unique in the Euclidean division of polynomials. ■

Problem 10.25. Find the quotient and remainder when $3x^4+10x+5$ is divided by $2x^2+x+4$.

From here on, let $\mathbb{F}$ denote an arbitrary field, unless otherwise specified.

Theorem 10.26 (Polynomial remainder theorem)**.** If $c \in \mathbb{F}$, then dividing $f \in \mathbb{F}[x]$ by $x - c$ yields a constant remainder equal to $f(c)$.

Proof. By Euclidean division,

$$f(x) = (x - c)q(x) + r(x),$$

with $\deg r < \deg(x - c) = 1$. So r must be a constant. Substituting $x = c$ into the equation yields

$$f(c) = r(c) = r.$$

Note that we can generalize the argument to division by a general linear factor $ax - b$. Then the remainder is $f\left(\frac{b}{a}\right)$, though the polynomial remainder theorem is usually stated only for division by monic linear polynomials. ■

Theorem 10.27 (Polynomial factor theorem)**.** If $c \in \mathbb{F}$, then $x - c$ divides $f \in \mathbb{F}[x]$ with no remainder if and only if c is a root of f.

Proof. By the polynomial remainder theorem (Theorem 10.26), the remainder of f upon division by $x - c$ is $f(c)$. Thus, the polynomial factor theorem is equivalent to saying that $f(c) = 0$ if and only if c is a root of f, which follows from the definition of a root. ■

Just like with integers in number theory, there a notion of a greatest common divisor of polynomials, along with related results such as an analogue of polynomials of Bézout's lemma. While these are interesting, we will not discuss them. The reader who learns abstract algebra will surely come across these ideas in a more general setting.

Theorem 10.28 (Fundamental theorem of algebra)**.** Every non-constant polynomial in $\mathbb{C}[x]$ has at least one complex root.

Proof. The proof of this theorem is beyond the scope of our material, as every known proof requires some form of higher mathematics. ■

Definition 10.29. The polynomial factor theorem (Theorem 10.27) tells us that $c \in \mathbb{F}$ is a root of $f \in \mathbb{F}[x]$ if and only if $x - c$ is a factor of f. We define the **multiplicity** of a root c to be the largest integer k such that $(x - c)^k$ is a factor of f.

Theorem 10.30. The following three conditions are equivalent for fields $\mathbb{F}$. For every non-constant polynomial $f \in \mathbb{F}[x]$:

1. f is the product of a constant times $\deg f$ monic linear polynomials from $\mathbb{F}[x]$.
2. f has exactly $\deg f$ roots in $\mathbb{F}$, if the number of times that each root is counted is equal to its multiplicity. This is called counting "with multiplicity."
3. f has at least one root in $\mathbb{F}$.

Note that, while the fundamental theorem of algebra implies that they are all true for $\mathbb{F} = \mathbb{C}$, they might all fail for some other fields.

Proof. In a cyclic fashion, we will show that $1 \implies 2 \implies 3 \implies 1$.

- If $c \in \mathbb{F}$ such that $x - c$ is a factor of f then the polynomial factor theorem tells us that c is a root of f (Theorem 10.27). Since we are assuming that f has $\deg f$ monic linear factors $x - c$, each c is a root. Counting roots with multiplicity, this means there are at *least* $\deg f$ roots of f in $\mathbb{F}$.

  Now we will show by induction on $\deg f$ that there are at *most* $\deg f$ roots of f in $\mathbb{F}$. We need to prove the result for $\deg f \geq 1$, but for the sake of convenience, we will start with the base case $\deg f = 0$ where f is a non-zero constant. If f is a non-zero constant, then it has no roots, and so it is true that f has at most $\deg f$ roots.

  For the induction hypothesis, suppose that the result holds for all non-zero polynomials of degree less than or equal to n for some integer $n \geq 0$, and let f be a polynomial of degree $n + 1$. If f has no roots, then it is true that f has at most $\deg f$ roots. If f has a root c then let k be the multiplicity of c so that

$$f(x) = (x - c)^k q(x)$$

for some $q \in \mathbb{F}[x]$. The polynomial factor theorem tells us that $k \geq 1$, so

$$\begin{aligned} n+1 &= \deg f \\ &= \deg((x-c)^k) + \deg q \\ &= k + \deg q \\ &\geq 1 + \deg q. \end{aligned}$$

Then $\deg q \leq n+1-1 = n$. By the induction hypothesis, q has at most $n+1-k$ roots counted with multiplicity, none of which are c. Moreover, every root of f is either c, or if it is $d \neq c$ then d is a root of q because

$$0 = f(d) = (d-c)^k q(d) \implies q(d) = 0.$$

Therefore, f has at most

$$k + (n+1-k) = n+1 = \deg f$$

roots counted with multiplicity, as desired.

Since we have shown that f has both at least and at most $\deg f$ roots with multiplicity, f has exactly $\deg f$ roots with multiplicity.

- If f has exactly $\deg f \geq 1$ roots in $\mathbb{F}$, then it must have at least one root in $\mathbb{F}$.
- Suppose f has at least one root in $\mathbb{F}$. We will prove by induction on $\deg f \geq 1$ that f is the product of a constant times $\deg f$ monic polynomials that are all from $\mathbb{F}[x]$. In the base case, $\deg f = 1$ and so must have the form $ax + b = a\left(x + \frac{b}{a}\right)$. So the result holds for linear polynomials.

  For the induction hypothesis, we will assume that the result holds for polynomials with degree equal to n for some $n \geq 1$. Suppose f has degree $n+1$. Since we know that f has a root c, the polynomial factor theorem tells us that $f(x) = (x-c)q(x)$ for some $q \in \mathbb{F}[x]$. We know that

  $$\deg f = \deg(x-c) + \deg q = 1 + \deg q,$$

  so $\deg q = \deg f - 1 < \deg f$, which allows us to apply the induction hypothesis on q. This tells us that q is a product of a constant times $\deg q = (\deg f) - 1$ monic linear factors. The rest follows because we just tack on an extra factor $x - c$, according to $f(x) = (x-c)q(x)$.

As a side note, since f has exactly $\deg f$ roots with multiplicity, f has at most $\deg f$ distinct roots. ■

Problem 10.31. Prove that, if p and q are polynomials in $\mathbb{C}[x]$ such that pq is the zero polynomial, then at least one of p or q is the zero polynomial. A consequence of this is that if f, g, h are polynomials such that $fg = fh$ and f is not the zero polynomial, then $f \cdot (g-h) = 0$ and we can "cancel" f from both sides to get $g = h$.

Corollary 10.32. If f is a non-constant polynomial in $\mathbb{C}[x]$ of degree n, then it has exactly n complex roots $z_1, z_2, \ldots, z_n$, including multiplicity, and f factors as

$$a(x - z_1)(x - z_2) \cdots (x - z_n),$$

where a is its leading coefficient, and the factorization is unique up to reordering the factors $x - z_j$. As a result, if two polynomials $f, g \in \mathbb{C}[x]$ have the same set of roots (and so their degrees are equal) and their leading coefficients are equal, then $f = g$.

Proof. We know from Theorem 10.30 that f has exactly n roots $z_1, z_2, \ldots, z_n$ with multiplicity and we know that one way of factoring f is

$$f(x) = a(x - z_1)(x - z_2) \cdots (x - z_n),$$

where a is its leading coefficient. So we just need to prove uniqueness. Suppose there is another factorization

$$f(x) = b(x - w_1)(x - w_2) \cdots (x - w_m).$$

Then we can equate them to get

$$a(x - z_1)(x - z_2) \cdots (x - z_n) = b(x - w_1)(x - w_2) \cdots (x - w_m).$$

The leading terms must be equal, meaning $ax^n = bx^m$, which implies $n = m$ and subsequently $a = b$. Substituting w_1 into both sides yields

$$(w_1 - z_1)(w_1 - z_2) \cdots (w_1 - z_n) = 0,$$

which means $w_1 = z_i$ for some $i \in [n]$. Reordering the z_j if necessary, we may assume without loss of generality that $i = 1$. Via polynomial division, we can remove $x - z_1 = x - w_1$ from both sides to get

$$(x - z_2) \cdots (x - z_n) = (x - w_2) \cdots (x - w_n).$$

Continuing in this way n times, we find that the multisets $\langle z_1, z_2, \ldots, z_n \rangle$ and $\langle z_1, z_2, \ldots, z_m \rangle$ are identical, which shows that the factorization is unique up to reordering the factors.
For the consequence, if f and g have the same roots $z_1, z_2, \ldots, z_n$ and the same leading coefficient a, then

$$f(x) = a(x - z_1)(x - z_2) \cdots (x - z_n) = g(x).$$

■

Theorem 10.33. The list of complex roots of a non-constant polynomial $f \in \mathbb{R}[x]$ is

$$r_1, r_2, \ldots, r_i, z_1, \overline{z_1}, z_2, \overline{z_2}, \ldots, z_j, \overline{z_j}$$

for non-negative integers i and j, where the r_k are real and the z_k and $\overline{z_k}$ are non-real complex numbers that come in such conjugate pairs, and $i + 2j = \deg f$. As a consequence, the number of non-real complex roots of f is even.

Proof. Let f be a non-constant polynomial of degree n with real coefficients, and let its list of roots be $w_1, w_2, \ldots, w_n$. Then

$$f(x) = a(x - w_1)(x - w_2) \cdots (x - w_n),$$

where a is its leading coefficient. If all the w_k are real then there nothing left to prove. Otherwise, suppose a non-real complex real root exists. Since the w_k may be reordered without changing the factored expression for f, we can assume without loss of generality that w_1 is a non-real complex root. We know from Theorem 10.16 that $\overline{w_1}$ is also a root, so we also assume without loss of generality that $w_2 = \overline{w_1}$. Then the product of the first two factors is

$$\begin{aligned}(x - w_1)(x - w_2) &= (x - w_1)(x - \overline{w_1}) \\ &= x^2 - (w_1 + \overline{w_1})x + w_1\overline{w_1} \\ &= x^2 - 2\mathrm{Re}(w_1) + |w_1|^2,\end{aligned}$$

which is a polynomial with real coefficients. The Euclidean division theorem then tells us that the quotient $(x - w_3) \cdots (x - w_n)$ after dividing f by $(x - w_1)(x - w_2)$ also has real coefficients. This allows us to continue repeating the process by taking away pairs of conjugate non-real complex roots until we run out of non-real complex roots. ■

Theorem 10.34 (Polynomial identity theorem)**.** The fundamental theorem of algebra and its corollaries help us to derive the more difficult parts of Theorem 10.19 in a new way, as follows.

1. If $f \in \mathbb{C}[x]$ and f has more than $\deg f$ distinct complex roots then f is zero everywhere.
2. Suppose f and g are polynomials in $\mathbb{C}[x]$ such that both have degree less than or equal to n. If $f(z) = g(z)$ for more than n distinct points $z \in \mathbb{C}$ then f and g are equal everywhere.

These results help us to identify polynomials using a bounded number of points.

Proof. The latter follows from the former, so we will prove them in sequence:

1. One consequence (Theorem 10.30) of the fundamental theorem of algebra is that if $\deg f \geq 1$ then f has at most $\deg f$ distinct roots. By contrapositive, if f has more than $\deg f$ distinct roots, then f is a constant polynomial. If f is a non-zero constant, then it has no roots, which contradicts the requirement that it must have strictly more than $\deg f = 0$ roots. So, instaed, f must be zero everywhere.
2. If $f(z) = g(z)$ for more than n complex numbers z, then $(f - g)(z) = 0$ at each of those points. But we know that

   $$\deg(f - g) \leq \max(\deg f, \deg g) \leq n,$$

   so the fundamental theorem of algebra tells us that $f - g$ is zero everywhere as in the preceding part. This implies that f and g are equal everywhere.

■

10.3 Rational Functions

Definition 10.35. A **rational function** is a function that can be expressed in the form $f(x) = \dfrac{p(x)}{q(x)}$ where p and q are polynomial functions, and q is not the zero polynomial. For our purposes, they will only be polynomials with real coefficients. The domain of f is restricted to those real numbers that are not real roots of q, in order to avoid division by 0.

Theorem 10.36. If a rational function f can be written in the form $\dfrac{p(x)}{q(x)}$ where p and q are polynomials and q is not the zero polynomial, then the real roots of f are the real roots of p, excluding any real roots of q.

Proof. Suppose $f(r) = 0$ for some real r at which f is defined. Then $q(r) \neq 0$, otherwise f would not be defined at r. And $f(r) = \dfrac{p(r)}{q(r)} = 0$ implies $p(r) = 0$, so r is a real root of p, Conversely, if r is a real root of p but not a real root of q, then f is defined at r and

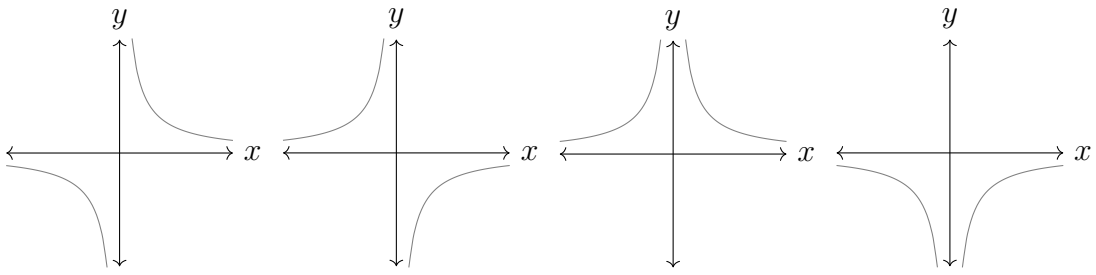
$$f(r) = \frac{p(r)}{q(r)} = \frac{0}{q(r)} = 0.$$

So r is also a real root of f. ■

Definition 10.37. An **asymptote** of a function f is a line to which the f gets arbitrarily close as the inputs of the function approach a certain real value or $+\infty$ or $-\infty$. For our purposes, it is acceptable for f to intersect an asymptote, even infinitely many times. We can classify the notion of asymptote into three types of lines: vertical, horizontal, or oblique, the last of which is also called slant.

1. A **vertical asymptote** is a line of the form $x = c$ for a real constant c. This line may be an asymptote regardless of whether f is defined at c, or what value $f(c)$ takes on if f does happen to be defined at c. The asymptotic behavior of f near c comes in four flavours:

   - $f(x)$ approaches $-\infty$ from the left of $x = c$ but $+\infty$ from the right of $x = c$
   - $f(x)$ approaches $+\infty$ from the left of $x = c$ but $-\infty$ from the right of $x = c$
   - $f(x)$ approaches $+\infty$ from both the left and right of $x = c$
   - $f(x)$ approaches $-\infty$ from both the left and right of $x = c$



2. A **horizontal asymptote** is a line of the form $y = c$ that $f(x)$ approaches as $x \to \infty$ or $x \to -\infty$. For an arbitrary function, these can be two different lines.

3. An **oblique or slant asymptote** is a line $g(x) = ax + b$ with $a \neq 0$ such that $f(x) - g(x)$ goes to 0 as $x \to \infty$ or $x \to -\infty$. For an arbitrarily function, these can be two different lines.

Theorem 10.38. We can classify when asymptotes occur for rational functions and develop methods of finding them. Let $f(x) = \dfrac{p(x)}{q(x)}$ be a rational function where p and q are polynomials. Then:

1. f has a vertical asymptote at $x = c$ if and only if the multiplicity of $x - c$ in q is greater than the multiplicity of $x - c$ in p.

2. f has a horizontal asymptote as $x \to \infty$ if and only if it has a horizontal asymptote as $x \to -\infty$. In this case, the two lines are equal, and we call this line the horizontal asymptote of f. Moreover, f has a horizontal asymptote if and only if $\deg p \leq \deg q$. If $\deg p < \deg q$ then the horizontal asymptote is $y = 0$. If $\deg p = \deg q$ then the horizontal asymptote is $y = \dfrac{a}{b}$, where a is the leading coefficient of p and b is the leading coefficient of q.

3. As with horizontal asymptotes, f has an oblique asymptote as $x \to \infty$ if and only if it has an oblique asymptote as $x \to -\infty$. In this case, the two lines are equal, and we call this line the oblique asymptote of f. Moreover, f has an oblique asymptote if and only if $\deg p = 1 + \deg q$. In this case, the oblique asymptote is the quotient found upon doing Euclidean division of p by q.

A rational function f can have an arbitrarily large finite number of vertical asymptotes, but it is restricted to at most one horizontal or oblique asymptote. That is, the existence of a horizontal asymptote excludes the existence of an oblique asymptote, and vice versa.

Proof. Since asymptotes involve the notion of "approaching," formal proofs of these assertions would be done in the language of limits from calculus. As this is beyond the scope of our exposition, we will provide only intuitive arguments for the "if" direction of the classification of when each type of asymptote occurs, along with informal justification for the technique used to compute the asymptote. Let f be as defined in the problem.

1. If $q(c) = 0$ and $p(c) \neq 0$, then we are essentially "dividing by 0" in $\dfrac{p(x)}{q(x)}$, so the behaviour of f around c is to shoot up to ∞ or down to ∞.

2. Suppose $m = \deg p \leq \deg q = n$. Let

$$p(x) = a_m x^m + a_{m-1} x^{m-1} + \cdots a_2 x^2 + a_1 x + a_0,$$
$$q(x) = b_n x^n + b_{n-1} x^{n-1} + \cdots b_2 x^2 + b_1 x + b_0.$$

The general idea is to divide the numerator and denominator of $\frac{p(x)}{q(x)}$ by x^n and observe what happens to each term of p and q as $x \to \infty$. This gives

$$\frac{p(x)}{q(x)} = \frac{\frac{a_m}{x^{n-m}} + \frac{a_{m-1}}{x^{n-m+1}} + \cdots + \frac{a_2}{x^{n-2}} + \frac{a_1}{x^{n-1}} + \frac{a_0}{x^n}}{b_n + \frac{b_{n-1}}{x} + \cdots + \frac{b_2}{x^{n-2}} + \frac{b_1}{x^{n-1}} + \frac{b_0}{x^n}}.$$

If $m < n$ then every term in the numerator and denominator, except b_n in the denominator, goes to 0 as $x \to \infty$. So $\frac{p(x)}{q(x)} \to \frac{0}{b_n}$ meaning $f(x) \to 0$ as $x \to \infty$. If $m = n$, then the numerator goes to a_m and the denominator goes to b_n as $x \to \infty$, so $f(x) \to \frac{a_m}{b_n}$.

3. Suppose $\deg p = 1 + \deg q$. By the Euclidean division theorem for polynomials,

$$p = qs + r$$

for a quotient s and a denominator r, where $\deg r < \deg q$ and $\deg s = \deg q - \deg p = 1$. Then

$$\frac{p}{q} - s = \frac{r}{q}$$

so the behaviour of $f - s$ the same as the behaviour of $\frac{r}{q}$ as $x \to \infty$. But $\deg r < \deg q$ so by the same argument as for horizontal asymptotes, $\frac{r}{q} \to 0$ as $x \to \infty$, which means s is an oblique asymptote of f.

■

Now that we have explored roots and asymptotes of rational functions, we are ready to tackle the question of when a rational function takes on what sign. There is one lemma that we will need as a prerequisite, which is as follows.

Lemma 10.39. Every polynomial $f \in \mathbb{R}[x]$ can be written as a product of a constant, monic linear polynomials $x + a$ where $a \in \mathbb{R}$, and monic quadratic polynomials $x^2 + bx + c$ where $b, c \in \mathbb{R}$ such that $b^2 - 4c < 0$. The linear or quadratics components might be empty.

Proof. If f is a constant, the result is immediate. If f is non-constant, Theorem 10.33 says that its list of roots is

$$r_1, r_2, \ldots, r_i, z_1, \overline{z_1}, z_2, \overline{z_2}, \ldots, z_j, \overline{z_j},$$

where the r_k are real and the z_k and $\overline{z_k}$ are non-real complex numbers, and $i + 2j = \deg f$. Then

$$f(x) = (x - r_1) \cdots (x - r_i)(x - z_1)(x - \overline{z_1}) \cdots (x - z_j)(x - \overline{z_j}).$$

We can choose the monic linear factors to be $x - r_k$ and the monic quadratic factors to be $(x - z_k)(x - \overline{z_k})$. This works because

$$(x - z_k)(x - \overline{z_k}) = x^2 - 2\mathrm{Re}(z_k) + |z_k|^2,$$

where the coefficients are real and

$$|z_k|^2 = (\text{Re}(z_k))^2 + (\text{Im}(z_k))^2 > (\text{Re}(z_k))^2,$$

so the discriminant is

$$(2\text{Re}(z_k))^2 - 4 \cdot 1 \cdot |z_k|^2 < (2\text{Re}(z_k))^2 - 4 \cdot (\text{Re}(z_k))^2 = 0.$$

■

Let us now see the process of solving a rational inequality that contains every possible obstacle. The method will work for polynomials too, as the denominator is then simply 1.

Example 10.40 (Interval analysis)**.** Find all real x such that

$$f(x) = \frac{-6(x-2)(x+1)^2(x+3)^3(x^2+x+2)}{7(x-2)(x-4)(x+1)(x+5)^4} \geq 0.$$

Solution. First we split this into two problems: finding the real roots of f and finding when the strict inequality $f(x) > 0$ holds. The real roots of f are the real roots of the numerator, excluding the real roots of the denominator. This gives the set

$$\{2, -1, -3\} \backslash \{2, 4, -1, -5\} = \{-3\}.$$

Now we will find when the inequality holds strictly and pop -3 back into the solution in the end. We take f over to the other side to remove the pesky negative sign, and then we remove the leading coefficients to get the equivalent inequality

$$\frac{(x-2)(x+1)^2(x+3)^3(x^2+x+2)}{(x-2)(x-4)(x+1)(x+5)^4} < 0, x \neq 2, 4, -1, -5.$$

Then we cancel out pairs of common linear factors between the numerator and denominator, while noting that the domain of the simplified expression cannot include the roots of those cancelled factors. This gives the equivalent inequality

$$\frac{(x+1)(x+3)^3(x^2+x+2)}{(x-4)(x+5)^4} < 0, x \neq 2, 4, -1, -5.$$

Since the quadratic factor's discriminant is negative and its leading coefficient is positive, it always takes on positive values, so we can divide out by it to get the equivalent inequality

$$\frac{(x+1)(x+3)^3}{(x-4)(x+5)^4} < 0, x \neq 2, 4, -1, -5.$$

The squares of linear factors are always non-negative, so $(x+3)^2 \geq 0$ and $(x+5)^4 \geq 0$. Moreover, it is not possible that $x = -3$ since that would make the whole expression 0, contradicting the fact that we are working on a strict inequality, and it is not possible that $x = -5$ since that is not in the domain of the expression. So $(x+3)^2 > 0$ and $(x+5)^4 > 0$ and we can divide out by them to get the equivalent inequality

$$\frac{(x+1)(x+3)}{x-4} < 0, x \neq 2, 4, -1, -5.$$

Now the expression is simple enough for us to perform **interval analysis**. The list of roots is $-1, -3$ and the list of singularities (i.e. where vertical asymptotes occur) is 4. We will call the list $-3, -1, 4$ the list of "change points," though it is not standard terminology. It should intuitively make sense that it is only at roots and singularities that the sign might change. So the function that the expression $\dfrac{(x+1)(x+3)}{x-4}$ represents always has the same sign when x is strictly in between two consecutive change points. This idea allows us to partition the real line into intervals and check the sign of each factor in each interval.

	$(-\infty, -3)$	$(-3, -1)$	$(-1, 4)$	$(4, \infty)$
$x+1$	$-$	$-$	$+$	$+$
$x+3$	$-$	$+$	$+$	$+$
$x-4$	$-$	$-$	$-$	$+$
$f(x)$	$-$	$+$	$-$	$+$

So our preliminary solution is

$$(-\infty, -3) \cup (-1, 4).$$

We then have to remove the points $\{2, 4, -1, -5\}$ because they are not in the domain of f, which refines the solution to

$$(-\infty, -5) \cup (-5, -3) \cup (-1, 2) \cup (2, 4).$$

Finally, we cannot forget to include the root -3, which gives the final solution

$$(-\infty, -5) \cup (-5, -3] \cup (-1, 2) \cup (2, 4).$$

Those who are interested in knowing why interval analysis works should look at the intermediate value theorem from calculus. ■

A general algorithm for solving rational inequalities should be evident from the example provided. It is essentially a matter of going through a sequence of simplifications, while ensuring that the steps are reversible, and then analyzing the signs of the remaining factors on each open interval defined by consecutive change points. The only step that we did not show is that of factoring the numerator and denominator into linear and irreducible quadratic factors; the fact that this is possible is proven by Lemma 10.39, and we have developed several techniques to make it happen if the circumstances are sufficiently convenient: Theorem 10.12, Corollary 10.15, Theorem 10.16, and Problem 10.18. Even Example 10.11 is helpful in special cases. Another trick is that z is a root if and only if $-z$ is a root, if all terms of odd degree have 0 as the coefficient, since this leaves only even degree terms.

10.4 Symmetry

Definition 10.41. A **term** ax^α in n variables $x_1, x_2, \ldots, x_n$ is an expression of the form

$$ax_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n},$$

where the **coefficient** a is an element of a fixed field $\mathbb{F}$, such as $\mathbb{C}$, and the α_i are non-negative integers. The **multidegree** of this term is the multi-index

$$\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$$

and its **degree** is

$$|\alpha| = \alpha_1 + \alpha_2 + \cdots + \alpha_n.$$

A **polynomial** in the n variables $x_1, x_2, \ldots, x_n$ is a sum of finitely many terms in these variables. The **degree** of a polynomial is the highest degree among the degrees of its non-zero terms. The set of polynomials in the n distinct variables $x_1, x_2, \ldots, x_n$ with coefficients in the field $\mathbb{F}$ is denoted by

$$\mathbb{F}[x_1, x_2, \ldots, x_n].$$

Problem 10.42. Vieta's formulas for quadratics (Corollary 9.10) tells us that if the roots of $ax^2 + bx + c$ are r_1 and r_2 then

$$\begin{aligned} r_1 + r_2 &= -\frac{b}{a} \\ r_2 r_2 &= \frac{c}{a}. \end{aligned}$$

Find analogous formulas for cubic equations, which are univariate polynomials of degree 3.

Inspired by these results for quadratic and cubic polynomials, we seek to generalize them to polynomials of degree $n \geq 1$.

Theorem 10.43 (Vieta's formulas)**.** Suppose

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0$$

is a polynomial of degree $n \geq 1$ with complex coefficients a_i. Then the complex $r_1, r_2, \ldots, r_n$, are the n roots of f if and only if the following n equations hold:

$$\begin{aligned} r_1 + r_2 + \cdots + r_n &= (-1)^1 \cdot \frac{a_{n-1}}{a_n} \\ (r_1 r_2 + \cdots + r_1 r_n) + (r_2 r_3 + \cdots + r_2 r_n) + \cdots + (r_{n-1} r_n) &= (-1)^2 \cdot \frac{a_{n-2}}{a_n} \\ &\vdots \\ \sum_{\substack{J \subseteq [n] \\ |J| = k}} \prod_{j \in J} r_j = \sum_{1 \leq i_1 < i_2 < \cdots < i_k \leq n} r_{i_1} r_{i_2} \cdots r_{i_k} &= (-1)^k \cdot \frac{a_{n-k}}{a_n} \\ &\vdots \\ r_1 r_2 \cdots r_n &= (-1)^n \cdot \frac{a_0}{a_n}. \end{aligned}$$

Proof. For the main direction, supposes $r_1, r_2, \ldots, r_n$, are the n roots of f. We could perform a dry proof by induction on n, but it will be more instructive to consider the expansion of

$$\begin{aligned} f(x) &= a_n x^n + a_{n-1}x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0 \\ &= a_n(x - r_1)(x - r_2)\cdots(x - r_n). \end{aligned}$$

Dividing both sides by a_n yields

$$x^n + \frac{a_{n-1}}{a_n}x^{n-1} + \cdots + \frac{a_2}{a_n}x^2 + \frac{a_1}{a_n}x + \frac{a_0}{a_n} = (x - r_1)(x - r_2)\cdots(x - r_n).$$

Now we will discover the coefficients of the right side upon expansion. In the expansion, before like terms are collected, there are 2^n terms, each having been created by choosing x from $n - k$ of the factors $x - r_i$ and $-r_i$ from the remaining k factors. If we want the coefficient of x^{n-k} on the right side, we will choose x from $n - k$ of the factors and $-r_i$ from the remaining k factors. This yields exactly all of the terms of the form

$$(-r_{i_1})(-r_{i_2})\cdots(-r_{i_k})x^{n-k} = (-1)^k \cdot r_{i_1} r_{i_2} \cdots r_{i_k} x^{n-k}$$

for all integer multi-indices $(i_1, i_2, \ldots, i_k)$ such that $1 \leq i_1 < i_2 < \cdots < i_k \leq n$. But the coefficient of x^{n-k} on the left side is $\dfrac{a_{n-k}}{a_n}$, so

$$\sum_{\substack{J \subseteq [n] \\ |J| = k}} \prod_{j \in J} r_j = \sum_{1 \leq i_1 < i_2 < \cdots < i_k \leq n} r_{i_1} r_{i_2} \cdots r_{i_k} = (-1)^k \cdot \frac{a_{n-k}}{a_n},$$

for each integer k such that $1 \leq k \leq n$, as expected.
Conversely, suppose $r_1, r_2, \ldots, r_n$, are n complex numbers such that the n stated equations hold. According to our described expansion,

$$\begin{aligned} a_n(x - r_1)(x - r_2)\cdots(x - r_n) &= a_n x^n + \sum_{k=1}^{n} (-1)^k a_n \left(\sum_{\substack{J \subseteq [n] \\ |J| = k}} \prod_{j \in J} r_j \right) x^{n-k} \\ &= a_n x^n + \sum_{k=1}^{n} a_{n-k} x^{n-k} \\ &= f(x). \end{aligned}$$

According to Corollary 10.32, f has a unique factorization into linear factors, so $r_1, r_2, \ldots, r_n$ are the n roots of f. ∎

Definition 10.44. Let n, k be positive integers such that $k \neq n$. Given n complex numbers $r_1, r_2, \ldots, r_n$, their k^{th} **symmetric sum** is defined as

$$\sigma_k = \sum_{\substack{J \subseteq [n] \\ |J| = k}} \prod_{j \in J} r_j = \sum_{1 \leq i_1 < i_2 < \cdots < i_k \leq n} r_{i_1} r_{i_2} \cdots r_{i_k}.$$

Corollary 10.45. Since factoring and expanding are two sides of the same coin that is the distributive law, we can reverse-engineer the roots in Vieta's formulas. That is, suppose we are given the n symmetric sums $\sigma_1, \sigma_2, \ldots, \sigma_n$ of n unknown complex numbers $r_1, r_2, \ldots, r_n$. Then a monic polynomial of degree n in $\mathbb{C}[x]$ with $r_1, r_2, \ldots, r_n$ as its roots is

$$(x - r_1)(x - r_2)\cdots(x - r_n) = x^n + \sum_{k=1}^{n} (-1)^k \sigma_k x^{n-k}.$$

The right side is known to us because we know the value of each σ_k, and subsequently we can often obtain information about the r_i using various techniques available to us, such as the rational root theorem (Theorem 10.12), integer root theorem (Corollary 10.15), and the fact that non-real complex roots come in conjugate pairs (Theorem 10.16).

Definition 10.46. A polynomial $f \in \mathbb{F}[x_1, x_2, \ldots, x_n]$ is said to be **symmetric** if swapping any two variables x_i and x_j in f results in the same polynomial.

Example. The polynomial

$$f(x, y) = x^3 + xy^2 + x^2y + y^3$$

is symmetric because

$$f(y, x) = y^3 + yx^2 + y^2x + x^3 = f(x, y),$$

thanks to commutativity of addition and multiplication. We swapped every instance of x with y and vice versa, and checked whether this produces an equivalent polynomial.

Definition 10.47. Let n, k be positive integers such that $k \neq n$. Given n variables $x_1, x_2, \ldots, x_n$, the k^{th} **elementary symmetric polynomial** in $\mathbb{F}[x_1, x_2, \ldots, x_n]$ is defined as

$$e_k(x_1, x_2, \ldots, x_n) = \sum_{1 \leq i_1 < i_2 < \cdots < i_k \leq n} x_{i_1} x_{i_2} \cdots x_{i_k}.$$

Lemma 10.48. Let n be a positive integer. Every polynomial in $\mathbb{F}[x_1, x_2, \ldots, x_n, x_{n+1}]$ may be interpreted as a univariate polynomial in the variable x_{n+1} whose coefficients are polynomials in $\mathbb{F}[x_1, x_2, \ldots, x_n]$. As such, we can say that

$$\mathbb{F}[x_1, x_2, \ldots, x_n, x_{n+1}] = (\mathbb{F}[x_1, x_2, \ldots, x_n,])\,[x_{n+1}].$$

In the language of abstract algebra, although $\mathbb{F}[x_1, x_2, \ldots, x_n]$ is not a field, it is a ring (like $\mathbb{Z}$) and we can still define polynomials whose coefficients are its elements.

Proof. We will show the method of turning a polynomial in $\mathbb{F}[x_1, x_2, \ldots, x_n, x_{n+1}]$ into a polynomial in $(\mathbb{F}[x_1, x_2, \ldots, x_n,])\,[x_{n+1}]$. This is best done by example. If we begin with

$$f(x, y) = 2x^3y^2 + 4x^2y + 5y + xy^2 \in \mathbb{R}[x, y]$$

and wish to write it as an element of $(\mathbb{R}[x])[y]$, we order the terms first by the exponents of y and then the exponents of x. This yields

$$f(x, y) = 2y^2x^3 + y^2x + 4yx^2 + 5y.$$

Finally, we consider x to be a constant and collect like terms in the variable y, which yields

$$f(x, y) = (2x^3 + x)y^2 + (4x^2 + 5)y.$$

This process generalizes. ■

Theorem 10.49. The elementary symmetric polynomials in $\mathbb{F}[x_1, x_2, \ldots, x_n]$ are symmetric polynomials (thank goodness!).

Proof. Let the n variables be $x_1, x_2, \ldots, x_n$ and let x be an extra variable. Then we define

$$f \in \mathbb{F}[x_1, x_2, \ldots, x_n, x] = (\mathbb{F}[x_1, x_2, \ldots, x_n])[x]$$
$$f(x_1, x_2, \ldots, x_n, x) = (x - x_1)(x - x_2) \cdots (x - x_n).$$

Let $e_1, e_2, \ldots, e_n$ be the elementary symmetric polynomials in $x_1, x_2, \ldots, x_n$. By the same argument that we used when proving Vieta's formulas (Theorem 10.43), this expands as

$$x^n - e_1 x^{n-1} + \cdots + (-1)^k e_k x^{n-k} + \cdots + (-1)^{n-1} e_{n-1} x + (-1)^n e_n.$$

But swapping x_i with x_j in $(x - x_1)(x - x_2) \cdots (x - x_n)$ leaves the expression unchanged, which means all of the coefficients of

$$x^n - e_1 x^{n-1} + \cdots + (-1)^k e_k x^{n-k} + \cdots + (-1)^{n-1} e_{n-1} x + (-1)^n e_n.$$

remain unchanged. By comparing coefficients, this simultaneously proves that all the e_k are symmetric polynomials. ■

Theorem 10.50 (Fundamental theorem of symmetric polynomials)**.** Every symmetric polynomial in $\mathbb{F}[x_1, x_2, \ldots, x_n]$ can be uniquely written as a polynomial in terms of the elementary symmetric polynomials $e_1, e_2, \ldots, e_k$.

Proof. There exists an elementary proof of this result using lexicographical order (Definition 4.41) that is due to Gauss, but we will not delve into it as it is cumbersome and would not help with our purposes. See [1]. ■

Corollary 10.51. If we are given a polynomial

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0$$

in $\mathbb{C}[x]$ with unknown roots $r_1, r_2, \ldots, r_n$ but known coefficients a_k, then we can compute any symmetric polynomial $g(r_1, r_2, \ldots, r_n)$ in terms of the coefficients a_k.

Problem 10.52. Let $f(x, y, z) = 2x^3 + 3x^2 + 5x + 7$. Let its complex roots be r, s, t. Compute the value of the symmetric expression $g(r, s, t) = r^2 + s^2 + t^2$.

Theorem 10.53 (Girard-Newton sums)**.** Let n be a positive integer and

$$p(x) = s_0 x^n + s_1 x^{n-1} + s_{n-1} x + s_n$$

be a polynomial in $\mathbb{C}[x]$ with the complex roots $r_1, r_2, \ldots, r_n$, where each root is included in this list as many times as its multiplicity in $p(x)$. For each positive integer k, let

$$p_k = r_1^k + r_2^k + \cdots + r_n^k,$$

with the convention

$$p_0 = \underbrace{1 + 1 + \cdots + 1}_{n \text{ copies of } 1} = n.$$

Then the following identities hold:

$$0 = \begin{cases} ks_k + \sum_{i=0}^{k-1} s_i p_{k-i} & \text{if } 1 \leq k < n \\ \sum_{i=0}^{n} s_i p_{k-i} & \text{if } 1 \leq n \leq k \end{cases}.$$

The coefficients in $p(x)$ have been strategically chosen to be indexed in reverse order so that these identities are easier to remember.

Proof. The identity for $k \geq n$ is easier to establish, and we will start with that. For each $j \in [n]$, r_j is a root of of $p(x)$, so

$$0 = p(r_j) = s_0 r_j^n + s_1 r_j^{n-1} + \cdots + s_{n-2} r_j^2 + s_{n-1} r_j + s_n.$$

Multiplying through by r_j^{k-n} (and using the temporary convention $0^0 = 1$, if needed), yields

$$0 = p(r_j) \cdot r_j^{k-n} = s_0 r_j^k + s_1 r_j^{k-1} + \cdots + s_{n-2} r_j^{k-n+2} + s_{n-1} r_j^{k-n+1} + s_n r_j^{k-n}.$$

By summing this equation over all $j \in [n]$, we get

$$0 = \sum_{j=1}^{n} p(r_j) \cdot r_j^{k-n} = \sum_{j=1}^{n} \sum_{i=0}^{n} s_i r_j^{k-i}.$$

By the discrete Fubini's principle, this is equivalent to

$$0 = \sum_{i=0}^{n} \sum_{j=1}^{n} s_i r_j^{k-i} = \sum_{i=0}^{n} s_i (r_1^{k-i} + r_2^{k-i} + \cdots + r_n^{k-i}) = \sum_{i=0}^{n} s_i p_{k-i},$$

which is the desired identity.
All proofs of the $k < n$ identity of which are aware are more complicated than the first case in one way or another. The cleanest one, in our opinion, is one that uses a tiny amount of calculus. We will show it here, as the derivative is an extremely useful tool to illustrate. The idea is to compute the derivative of $p(x)$ is two different ways and to equate corresponding coefficients. One way is standard, that is

$$p'(x) = ns_0 x^{n-1} + (n-1)s_1 x^{n-2} + \cdots + 2s_{n-2} x + s_{n-1} = \sum_{k=0}^{n-1} (n-k) s_k x^{n-k-1}.$$

The other way is to differentiate the factorization

$$p(x) = s_0 (x - r_1)(x - r_2) \cdots (x - r_n).$$

By a generalizing the product rule from calculus via induction on $n \geq 2$,

$$p'(x) = \sum_{j=1}^{n} \frac{p(x)}{x - r_j}.$$

The summands do not actually have a discontinuity at $x = r_j$, but rather each $\dfrac{p(x)}{x - r_j}$ denotes the quotient upon $p(x)$ being perfectly divided by its factor $x - r_j$. Leaving the details to the reader (it is beneficial to write them out!), the quotient of $p(x)$ upon synthetic division by the linear factor $x - r_j$ is

$$\frac{p(x)}{x - r_j} = \sum_{k=0}^{n-1} x^{n-k-1}(s_0 r_j^k + s_1 r_j^{k-1} + \cdots + s_{k-1} r_j + s_k).$$

Summing this equation over all $j \in [n]$, we get

$$p'(x) = \sum_{j=1}^{n} \frac{p(x)}{x - r_j} = \sum_{j=1}^{n} \sum_{k=0}^{n-1} x^{n-k-1}(s_0 r_j^k + s_1 r_j^{k-1} + \cdots + s_{k-1} r_j + s_k).$$

By two applications of the discrete Fubini's principle, this is equivalent to

$$\begin{aligned} p'(x) &= \sum_{j=1}^{n} \sum_{k=0}^{n-1} x^{n-k-1} \sum_{i=0}^{k} s_i r_j^{k-i} = \sum_{k=0}^{n-1} \sum_{j=1}^{n} x^{n-k-1} \sum_{i=0}^{k} s_i r_j^{k-i} \\ &= \sum_{k=0}^{n-1} x^{n-k-1} \sum_{j=1}^{n} \sum_{i=0}^{k} s_i r_j^{k-i} = \sum_{k=0}^{n-1} x^{n-k-1} \sum_{i=0}^{k} \sum_{j=1}^{n} s_i r_j^{k-i} \\ &= \sum_{k=0}^{n-1} x^{n-k-1} \sum_{i=0}^{k} s_i \sum_{j=1}^{n} r_j^{k-i} = \sum_{k=0}^{n-1} x^{n-k-1} \sum_{i=0}^{k} s_i p_{k-i}. \end{aligned}$$

By comparing coefficients with

$$p'(x) = \sum_{k=0}^{n-1} (n-k) s_k x^{n-k-1},$$

we get for each $k \in [n-1]$ that

$$(n-k)s_k = \sum_{i=0}^{k} s_i p_{k-i} = \sum_{i=0}^{k-1} s_i p_{k-i} + s_k p_0 = \sum_{i=0}^{k-1} s_i p_{k-i} + s_k n.$$

Taking everything to one side and slightly simplifying yields

$$0 = k s_k + \sum_{i=0}^{k-1} s_i p_{k-i}.$$

This is what we wanted to prove. Note that we left out the $k = 0$ identity because it just trivially says that $k = 0$. ■

10.5 Multivariable Factoring

Now we will see the analogue of the polynomial factor theorem (Theorem 10.27) for multivariable polynomials.

Theorem 10.54 (Multivariable factor theorem)**.** Let $n \geq 2$ be an integer. Suppose polynomials $f \in \mathbb{F}[x_1, x_2, \ldots, x_n]$ and $g \in \mathbb{F}[x_1, x_2, \ldots, x_{n-1}]$ exist such that

$$f(x_1, x_2, \ldots, x_{n-1}, g(x_1, x_2, \ldots, x_{n-1})) = 0.$$

Then $x_n - g$ is a factor of f in the sense that there exists a polynomial $h \in \mathbb{F}[x_1, x_2, \ldots, x_n]$ such that

$$f = (x_n - g)h.$$

The same result holds for any x_k instead of x_n.

Proof. The proof is most naturally done in the language of abstract algebra using commutative rings, so we will only briefly sketch it. There is a version of the Euclidean division theorem for polynomials (Theorem 10.23) that works on polynomials with coefficients in commutative rings instead of fields, with one caveat: the divisor has to be monic. We will only require the special case of a linear divisor, which is stated as follows:
If R is a commutative ring, then for all $f \in R[x]$ and $g \in R$, there exists an $h \in R[x]$ and an $r \in R$ such that

$$f(x) = (x - g)h(x) + r.$$

The proof of it follows exactly the line of logic that we used in proving the Euclidean division theorem for polynomials, so we do not repeat it here. We merely comment that it will work, despite the fact that the elements of R do not necessarily have multiplicative inverses, and it will work precisely because $x - g$ is monic. We encourage the reader to prove it, first for $f = 0$ and then by induction on $\deg f \geq 0$.
As a consequence, an analogue of the polynomial factor theorem is that $x - g$ divides f with no remainder (meaning $r = 0$) if and only if $f(g) = 0$.
Now we take R to be the polynomial ring $\mathbb{F}[x_1, x_2, \ldots, x_{n-1}]$, which is a commutative ring, and note that

$$\mathbb{F}[x_1, x_2, \ldots, x_{n-1}, x_n] = (\mathbb{F}[x_1, x_2, \ldots, x_{n-1}])[x_n] = R[x_n]$$

Suppose $f \in R[x_n]$ and $g \in R$ such that $f(g) = 0$. By our analogue of the polynomial factor theorem, there exists an $h \in R[x_n]$ such that $f = (x_n - g)h$, and we are done. ■

Example 10.55. Factor $(x - y)^3 + (y - z)^3 + (z - x)^3$ into a product of a constant times linear factors.

Solution. Let $f(x, y, z) = (x - y)^3 + (y - z)^3 + (z - x)^3$. Note that each of the scenarios $x = y, y = z, z = x$ individually leads to $f(x, y, z) = 0$. This means each of the linear polynomials $x - y, y - z, z - x$ is a factor of f. So we can guess that there is a polynomial g such that

$$f(x, y, z) = (x - y)(y - z)(z - x)g(x, y, z).$$

But $\deg f = 3$, so g must be a constant c. Now we can substitute constants into f such as $(x, y, z) = (3, 2, 1)$ to get

$$c = \frac{(x-y)^3 + (y-z)^3 + (z-x)^3}{(x-y)(y-z)(z-x)} = \frac{1^3 + 1^3 + (-2)^3}{1 \cdot 1 \cdot (-2)} = 3.$$

Indeed, expanding allows us to check that

$$(x-y)^3 + (y-z)^3 + (z-x)^3 = 3(x-y)(y-z)(z-x).$$

Note that there was guesswork involved, since we have not proven a result that says that the quotient upon dividing $f(x, y, z)$ by $x - y$ still has $y - z$ and $z - x$ as factors. ■

Problem 10.56. Factor $(x+y+z)^3 - x^3 - y^3 - z^3$ into a product of a constant times linear factors.

We will end the section by discussing some common multivariable factorizations and listing many others. In general, we like expansions of factored expressions where the simplified expression, after like terms have been collected, has very few terms.

Theorem 10.57 (Difference or sum of powers)**.** Recall the difference of squares factorization

$$x^2 - y^2 = (x-y)(x+y).$$

This generalizes as follows to all integers $n \geq 1$:

1. $x^{n+1} - y^{n+1} = (x-y)(x^n + x^{n-1}y + x^{n-2}y^2 + \cdots + x^2y^{n-2} + xy^{n-1} + y^n)$
2. $x^{2n+1} + y^{2n+1} = (x+y)(x^{2n} - x^{2n-1}y + x^{2n-2}y^2 - \cdots + x^2y^{2n-2} - xy^{2n-1} + y^{2n})$

Proof. We will prove the difference of powers using geometric series and then prove the sum of odd powers using the difference of powers.

1. Although a direct expansion with telescoping would work, let's try something more interesting. The formula for a geometric series (Theorem 8.14) says that, for $z \neq 1$.

   $$1 + z + z^2 + \cdots + z^n = \frac{z^{n+1} - 1}{z - 1}.$$

   This means

   $$z^{n+1} - 1 = (z-1)(z^n + z^{n-1} + \cdots + z + 1).$$

   We can check separately that this new equation holds for $z = 1$ as well. Letting $z = \dfrac{x}{y}$ and multiplying both sides by y^{n+1} yields the desired formula. So this difference of powers formula holds for all complex numbers x and y.

2. Since $2n + 1$ is odd,

   $$x^{2n+1} + y^{2n+1} = x^{2n+1} - (-y)^{2n+1}.$$

   Then we simply apply the difference of powers formula from the first part to the right side.

■

Problem 10.58. There are multiple ways of applying the sum or difference of powers factorizations to composite exponents, as follows:

$$x^{mn} \pm y^{mn} = (x^m)^n \pm (y^m)^n = (x^n)^m \pm (y^n)^m.$$

Factor $x^6 - y^6$ into irreducible factors with real coefficients.

Problem 10.59. Find a quadratic with real coefficients that has a non-trivial third root of unity as a root. By non-trivial, we mean that it is not equal to 1.

Problem 10.60. The following is an exceedingly common lemma. Suppose P is a polynomial with integer coefficients and a, b are integers such that $a \neq b$. Show that the integer $a - b$ divides the integer $P(a) - P(b)$. Deduce that if b is a root of P, then $a - b$ divides $P(a)$.

Example 10.61 (Sophie Germain identity)**.** Factor $x^4 + 4y^4$ as the product of two polynomials, each of which is a sum of two squares.

Solution. We can rewrite the expression as follows using the difference of squares:

$$\begin{aligned} x^4 + 4y^4 &= (x^2)^2 + (2y^2)^2 + 4x^2y^2 - 4x^2y^2 \\ &= (x^2 + 2y^2)^2 - (2xy)^2 \\ &= (x^2 - 2xy + 2y^2)(x^2 + 2xy + 2y^2) \\ &= ((x-y)^2 + y^2)((x+y)^2 + y^2). \end{aligned}$$

■

Theorem 10.62. As a reference, below is a master list of additional well-known factorizations in two or three variables. The reader is encouraged to verify them independently.

$$\begin{aligned} x^4 - y^4 &= (x-y)(x+y)(x^2+y^2) \\ x^4 + 4y^4 &= ((x-y)^2 + y^2)((x+y)^2 + y^2) \\ x^4 + x^2y^2 + y^4 &= (x^2 + xy + y^2)(x^2 - xy + y^2) \\ x^4 + y^4 + (x \pm y)^4 &= 2(x^2 \pm xy + y^2)^2 \end{aligned}$$

$$\begin{aligned} (x+y)^2 - x^2 - y^2 &= 2xy \\ (x+y)^3 - x^3 - y^3 &= 3xy(x+y) \\ (x+y)^5 - x^5 - y^5 &= 5xy(x+y)(x^2+xy+y^2) \\ (x+y)^7 - x^7 - y^7 &= 7xy(x+y)(x^2+xy+y^2)^2 \end{aligned}$$

$$\begin{aligned} (x+y+z)^2 - x^2 - y^2 - z^2 &= 2(xy+yz+zx) \\ (x+y+z)^3 - x^3 - y^3 - z^3 &= 3(x+y)(y+z)(z+x) \\ (x+y+z)^5 - x^5 - y^5 - z^5 &= 5(x+y)(y+z)(z+x)(x^2+y^2+z^2+xy+yz+zx) \\ \frac{(x+y+z)^7 - x^7 - y^7 - z^7}{(x+y)(y+z)(z+x)} &= 7((x^2+y^2+z^2+xy+yz+zx)^2 + xyz(x+y+z)) \end{aligned}$$

$$(x+y+z)(xy+yz+zx)-xyz=(x+y)(y+z)(z+x)$$
$$x^3+y^3+z^3-3xyz=(x+y+z)(x^2+y^2+z^2-xy-yz-zx)$$
$$(x+y+z)^3-24xyz=(-x+y+z)^3+(x-y+z)^3+(x+y-z)^3$$
$$x^3+y^3+z^3+(x+y)^3+(y+z)^3+(z+x)^3=3(x+y+z)(x^2+y^2+z^2)$$

$$(x-y)^3+(y-z)^3+(z-x)^3=3(x-y)(y-z)(z-x)$$
$$(xy^2+yz^2+zx^2)-(x^2y+y^2z+z^2x)=(x-y)(y-z)(z-x)$$
$$(xy^3+yz^3+zx^3)-(x^3y+y^3z+z^3x)=(x+y+z)(x-y)(y-z)(z-x)$$
$$(x^2y^3+y^2z^3+z^2x^3)-(x^3y^2+y^3z^2+z^3x^2)=(xy+yz+zx)(x-y)(y-z)(z-x)$$

$$(xy^2+yz^2+zx^2)(x^2y+y^2z+z^2x)-x^2y^2z^2=(x^2+yz)(y^2+zx)(z^2+xy)$$
$$(xy^2+yz^2+zx^2-xyz)^2+(x^2y+y^2z+z^2x-xyz)^2=(x^2+y^2)(y^2+z^2)(z^2+x^2)$$

Problem 10.63. Suppose a, b, c are complex numbers such that $a+b+c=0$. For each positive integer n, let

$$f(n)=\frac{a^n+b^n+c^n}{n}.$$

Prove that the following two identities hold:

$$f(2)\cdot f(3)=f(5),$$
$$f(2)\cdot f(5)=f(7).$$

Chapter 11

Multivariable Inequalities

> "Mathematics, rightly viewed, possesses not only truth, but supreme beauty - a beauty cold and austere, like that of sculpture, without appeal to any part of our weaker nature, without the gorgeous trappings of painting or music, yet sublimely pure, and capable of a stern perfection such as only the greatest art can show."
>
> – *Bertrand Russell, The Study of Mathematics*

Multivariable inequalities is an area of elementary mathematics that saw tremendous growth the first two decades of the twenty-first century. Entire books could be (and have been) written about them, especially by certain Vietnamese and Romanian practitioners. The advent of advanced techniques that allow for proofs of large classes of inequalities has resulted in problems about inequalities appearing less frequently on competitions and olympiads, and it has also caused the production of unnatural and less beautiful inequalities in the problems literature. We will focus only on classical results with which we feel the reader should be familiar.

11.1 AM-GM and Cauchy-Schwarz

An optimization result generally has two parts: a bound and statement about where equality is achieved in the bound, if at all. If we are lucky, it is possible to make a biconditional statement (recall that this means "if and only if") about when equality occurs. In some cases, there is no known concise presentation of the equality cases, so only sufficient or necessary conditions might be stated. If the inequality is ugly enough, there might be no mention of equality criteria at all.

Problem 11.1. Let a, b, c be real numbers. Then prove that the following classic inequalities hold and find where equality occurs in each case:

$$a^2 + b^2 \geq 2ab$$
$$a^2 + b^2 + c^2 \geq ab + bc + ca.$$

Definition 11.2. For any positive integer n and non-negative real numbers $a_1, a_2, \ldots, a_n$, their **arithmetic mean** (AM for short) is

$$\frac{a_1 + a_2 + \cdots + a_n}{n}$$

and their **geometric mean** (GM for short) is

$$\sqrt[n]{a_1 a_2 \cdots a_n}.$$

Theorem 11.3 (AM-GM inequality)**.** Let n be a positive integer and $a_1, a_2, \ldots, a_n$ be non-negative real numbers. Then

$$\frac{a_1 + a_2 + \cdots + a_n}{n} \geq \sqrt[n]{a_1 a_2 \cdots a_n},$$

with equality holding if and only if all of the a_i are equal.

Proof. We will use a special form of induction called Cauchy induction or forward-backward induction. Let $S(n)$ denote the assertion that

$$\frac{a_1 + a_2 + \cdots + a_n}{n} \geq \sqrt[n]{a_1 a_2 \cdots a_n}$$

for all non-negative a_i, with equality holding if and only if all the a_i are equal. It will suffice to prove the following in sequence:

1. $S(2)$, which establishes the base case ($S(1)$ is trivial can be proven alone)
2. $S(n) \implies S(2n)$, which establishes $S(2^k)$ for all positive integers k when repeatedly applied, starting with $S(2)$
3. $S(n) \implies S(n-1)$, which establishes $S(n)$ for all remaining n since we can work backwards from each $S(2^k)$

Now we will prove that these three hold:

1. We already know that $S(2)$ is true, due to the trivial inequality. See Problem 11.1.
2. Suppose $S(n)$ holds. Then we use $S(2)$ and $S(n)$ to get

$$\begin{aligned}\frac{a_1 + a_2 + \cdots + a_{2n}}{2n} &= \frac{1}{2}\left(\frac{a_1 + a_2 + \cdots + a_n}{n} + \frac{a_{n+1} + a_{n+2} + \cdots + a_{2n}}{n}\right) \\ &\geq \frac{1}{2}\left(\sqrt[n]{a_1 a_2 \cdots a_n} + \sqrt[n]{a_{n+1} a_{n+1} \cdots a_{2n}}\right) \\ &\geq \sqrt[2n]{a_1 a_2 \cdots a_{2n}}.\end{aligned}$$

   We can derive that equality holds if and only if all the a_i are equal using the equality conditions for $S(2)$ and $S(n)$.

3. Now suppose $S(n)$ holds. Miraculously, we find that we can rewrite the arithmetic mean in $S(n-1)$ as follows and apply $S(n)$ to get

$$\begin{aligned}\frac{a_1 + a_2 + \cdots + a_{n-1}}{n-1} &= \frac{1}{n}\left(a_1 + a_2 + \cdots + a_{n-1} + \frac{a_1 + a_2 + \cdots + a_{n-1}}{n-1}\right) \\ &\geq \sqrt[n]{a_1 a_2 \cdots a_{n-1} \cdot \frac{a_1 + a_2 + \cdots + a_{n-1}}{n-1}}.\end{aligned}$$

Upon taking the n^{th} power of each side, cancelling a copy of $\dfrac{a_1+a_2+\cdots+a_{n-1}}{n-1}$ from both sides, and finally taking the $(n-1)^{\text{th}}$ root of each side, this becomes the inequality for $S(n-1)$. The equality condition for $S(n-1)$ follows from the equality condition for $S(n)$.

■

Problem 11.4. In the third step of the proof of Theorem 11.3, we used the substitution

$$a_n = \frac{a_1+a_2+\cdots+a_{n-1}}{n-1}$$

to show that $S(n) \implies S(n-1)$. Use the substitution

$$a_n = \sqrt[n-1]{a_1a_2\cdots a_{n-1}}$$

to provide an alternative proof of this implication.

Example 11.5. Prove that, for all real numbers $r \geq -1$ and positive integers $m > n$, it holds that

$$\left(1+\frac{r}{m}\right)^m \geq \left(1+\frac{r}{n}\right)^n,$$

with equality holding if and only if $r = 0$. This goes hand in hand with Bernoulli's inequality (Theorem 4.36) because, where Bernoulli shows that compound interest is better than simple interest for the receiver, this shows that more compounding at the same rate means the deal keeps getting better for the receiver. What do you think happens as $n \to \infty$?

Solution. Let $r \geq -1$ be a real number and n be a positive integer. We will show that

$$\left(1+\frac{r}{n+1}\right)^{n+1} \geq \left(1+\frac{r}{n}\right)^n,$$

with equality holding if and only if $r = 0$. It suffices to prove this instead because we can repeatedly apply this result to compare the expressions for m and n instead.
First note that

$$r \geq -1 \implies n \geq 1 \geq -r \implies \frac{r}{n} \geq -1 \implies 1+\frac{r}{n} \geq 0.$$

So we can apply the AM-GM inequality to one copy of 1 and n copies of $1+\dfrac{r}{n}$ to get

$$\frac{1+n\cdot\left(1+\frac{r}{n}\right)}{n+1} \geq \sqrt[n+1]{1\cdot\left(1+\frac{r}{n}\right)^n}.$$

This implies the comparison between the expressions for $n+1$ and n that we wanted to see. By the equality criterion for the AM-GM inequality, equality holds if and only if

$$1 = 1+\frac{r}{n} \iff r = 0,$$

as hypothesized.
Interestingly, the benefits of more compounding taper off. It is known that

$$\lim_{n\to\infty}\left(1+\frac{r}{n}\right)^n = e^r,$$

where e is Euler's constant and r is any real number. ■

Problem 11.6 (GM-HM inequality). Let n be a positive integer and $a_1, a_2, \ldots, a_n$ be positive real numbers. Show that

$$\sqrt[n]{a_1 a_2 \cdots a_n} \geq \frac{n}{\frac{1}{a_1} + \frac{1}{a_2} + \cdots + \frac{1}{a_n}},$$

with equality holding if and only if all of the a_i are equal. The quantity on the right side is called the **harmonic mean** (HM for short) of the numbers, hence the name GM-HM inequality.

Problem 11.7. Solve the following two optimization problems.

1. If p is a positive real number, then determine the dimensions of a rectangle with maximal area among all rectangles with perimeter p. The dimensions should be written in terms of p.

2. If S is a positive real number, then determine the dimensions of a box with maximal volume among all boxes with surface area S. The dimensions should be written in terms of S.

Problem 11.8. The AM-GM inequality states that

$$\frac{x^3 + y^3 + z^3}{3} \geq xyz$$

for all non-negative real numbers x, y, z. Determine all triples (x, y, z) of *real* numbers (not just non-negative ones) for which this inequality holds and state a biconditional equality condition.

Theorem 11.9 (Cauchy-Schwarz inequality)**.** Suppose n is a positive integer. For all real numbers $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$, it holds that

$$(a_1^2 + a_2^2 + \cdots + a_n^2)(b_1^2 + b_2^2 + \cdots + b_n^2) \geq (a_1 b_1 + a_2 b_2 + \cdots + a_n b_n)^2.$$

Equality holds if and only if either $a_i = 0$ for all $i \in [n]$ or there exists a real constant r such that $a_i r = b_i$ for all $i \in [n]$.

Proof. If we rearrange the inequality and multiply through by 4, we get the equivalent inequality

$$0 \geq (2a_1 b_1 + 2a_2 b_2 + \cdots + 2a_n b_n)^2 - 4(a_1^2 + a_2^2 + \cdots + a_n^2)(b_1^2 + b_2^2 + \cdots + b_n^2),$$

where the right side looks like a discriminant. This inspires to construct a quadratic function $f : \mathbb{R} \to \mathbb{R}$ defined by

$$f(x) = (a_1^2 + a_2^2 + \cdots + a_n^2)x^2 - (2a_1 b_1 + 2a_2 b_2 + \cdots + 2a_n b_n)x + (b_1^2 + b_2^2 + \cdots + b_n^2),$$

which has the desired discriminant. At this point, we should note that this is not a quadratic function if all the a_i are equal to 0 because then the quadratic term would disappear. However, it is easy to see that the Cauchy-Schwarz inequality holds with both sides being equal in this case. So now we can assume that at least one of the a_i is non-zero.

By rearranging the expression for f so that terms with the same indices are grouped together, we see that it can be rewritten as

$$\begin{aligned} f(x) &= (a_1^2x^2 - 2a_1b_1x + b_1^2) + (a_2^2x^2 - 2a_2b_2x + b_2^2) + \cdots + (a_n^2x^2 - 2a_nb_nx + b_n^2) \\ &= (a_1x - b_1)^2 + (a_2x - b_2)^2 + \cdots + (a_nx - b_n)^2 \geq 0. \end{aligned}$$

The fact that f is always non-negative means that there is either one real solution or no real solutions. This is equivalent to the discriminant being non-positive, which proves the desired inequality.
So equality holds if and only if all of the a_i are equal to 0, or the discriminant is 0. In the latter case, this is equivalent to there existing a real root r of f (two distinct real roots is not possible for f, so the existence of a real root is equivalent to the existence of exactly one real root). But we found that

$$f(x) = (a_1x - b_1)^2 + (a_2x - b_2)^2 + \cdots + (a_nx - b_n)^2,$$

so there exists a real root r if and only if

$$(a_1r - b_1)^2 + (a_2r - b_2)^2 + \cdots + (a_nr - b_n)^2 = 0.$$

By the equality condition of the trivial inequality, this is true if and only if there exists a real number r such that $a_ir = b_i$ for all i. ■

Corollary 11.10. Let n be a positive integer. The following are special cases of the Cauchy-Schwarz inequality that are useful in practice.

1. If $a_1, a_2, \ldots, a_n$ are positive real numbers, then

$$(a_1 + a_2 + \cdots + a_n)\left(\frac{1}{a_1} + \frac{1}{a_2} + \cdots + \frac{1}{a_n}\right) \geq n^2,$$

   with equality holding if and only if all of the a_i are equal.

2. If $a_1, a_2, \ldots, a_n$ are non-negative real numbers and $b_1, b_2, \ldots, b_n$ are positive real numbers, then

$$(a_1b_1 + a_2b_2 + \cdots + a_nb_n)\left(\frac{a_1}{b_1} + \frac{a_2}{b_2} + \cdots + \frac{a_n}{b_n}\right) \geq (a_1 + a_2 + \cdots + a_n)^2.$$

   If the a_i are all positive, then equality holds if and only if all the b_i are equal.

3. If $a_1, a_2, \ldots, a_n$ are positive real numbers and $b_1, b_2, \ldots, b_n$ are real numbers, then

$$(a_1 + a_2 + \cdots + a_n)\left(\frac{b_1^2}{a_1} + \frac{b_2^2}{a_2} + \cdots + \frac{b_n^2}{a_n}\right) \geq (b_1 + b_2 + \cdots + b_n)^2,$$

   with equality holding if and only if there exists a real number r such that $a_ir = b_i$ for all i. This inequality has several names, one of which is **Engel's form**.

Proof. These are all applications of the Cauchy-Schwarz inequality.

1. Let $a_1, a_2, \ldots, a_n$ be positive real numbers. Applying Cauchy-Schwarz to the tuples
$$(\sqrt{a_1}, \sqrt{a_2}, \ldots, \sqrt{a_n}) \text{ and } \left(\frac{1}{\sqrt{a_1}}, \frac{1}{\sqrt{a_2}}, \ldots, \frac{1}{\sqrt{a_n}}\right)$$
yields the desired inequality. Since the a_i are positive, we do not have to worry about the equality case where all of the a_i are zero. Equality holds if and only if there exists a real number r such that $\sqrt{a_i} \cdot r = \dfrac{1}{\sqrt{a_i}}$ for all i, which is equivalent to all of the a_i being equal to each other and $\dfrac{1}{r}$.

2. Let $a_1, a_2, \ldots, a_n$ be non-negative real numbers and $b_1, b_2, \ldots, b_n$ be positive real numbers. Applying Cauchy-Schwarz to the tuples
$$\left(\sqrt{a_1 b_1}, \sqrt{a_2 b_2}, \ldots, \sqrt{a_n b_n}\right) \text{ and } \left(\sqrt{\frac{a_1}{b_1}}, \sqrt{\frac{a_2}{b_2}}, \ldots, \sqrt{\frac{a_n}{b_n}}\right)$$
yields the desired identity. If all of the a_i are positive, then the $\sqrt{a_i b_i}$ are all positive and so equality holds if and only if there exists a real number r such that $\sqrt{a_i b_i} \cdot r = \sqrt{\dfrac{a_i}{b_i}}$ for all i. This equality condition holds if and only if all of the b_i are equal to each other and $\dfrac{1}{r}$.

3. Let $a_1, a_2, \ldots, a_n$ be positive real numbers and $b_1, b_2, \ldots, b_n$ be real numbers. Applying Cauchy-Schwarz to the tuples
$$(\sqrt{a_1}, \sqrt{a_2}, \ldots, \sqrt{a_n}) \text{ if } \left(\frac{b_1}{\sqrt{a_1}}, \frac{b_2}{\sqrt{a_2}}, \ldots, \frac{b_n}{\sqrt{a_n}}\right)$$
yields the desired identity. Since the a_i are all positive, equality holds if and only if there exists a real number r such that $\sqrt{a_i} \cdot r = \dfrac{b_i}{\sqrt{a_i}}$ for all i. This equality condition is true if and only if there exists a real number r such that $a_i r = b_i$ for all i.

■

Problem 11.11. If $\sigma : [n] \to [n]$ is a bijection and $a_1, a_2, \ldots, a_n$ are real numbers, then show that
$$a_1^2 + a_2^2 + \cdots + a_n^2 \geq |a_1 a_{\sigma(1)} + a_2 a_{\sigma(2)} + \cdots + a_n a_{\sigma(n)}|.$$

Problem 11.12 (RMS-AM inequality). Let n be a positive integer and $a_1, a_2, \ldots, a_n$ be real numbers. Show that
$$n \cdot (a_1^2 + a_2^2 + \cdots + a_n^2) \geq (a_1 + a_2 + \cdots + a_n)^2,$$
and as a result,
$$\sqrt{\frac{a_1^2 + a_2^2 + \cdots + a_n^2}{n}} \geq \frac{a_1 + a_2 + \cdots + a_n}{n},$$
with equality holding if and only if all of the a_i are non-negative and equal. The quantity on the left side is called the **root mean square** (RMS for short) of the numbers, hence the name RMS-AM inequality.

Problem 11.13 (Nesbitt's inequality). Prove that, for all positive real numbers a, b, c,

$$\frac{a}{b+c} + \frac{b}{c+a} + \frac{c}{a+b} \geq \frac{3}{2},$$

and show that equality holds if and only if $a = b = c$.

11.2 Schur, Rearrangement, and Chebyshev

Theorem 11.14 (Schur's inequality)**.** For all non-negative real numbers x, y, z and all positive real numbers t, it holds that

$$x^t(x-y)(x-z) + y^t(y-z)(y-x) + z^t(z-x)(z-y) \geq 0.$$

Equality holds if and only if $x = y = z$, or two of x, y, z are equal to each other with the third equal to 0. If t is a non-negative even integer (with the convention that $0^0 = 1$), then the inequality holds for all real x, y, z and the equality condition is the same.

Proof. Let t be a positive real number. Define the three-variable polynomial

$$P(x, y, z) = x^t(x-y)(x-z) + y^t(y-z)(y-x) + z^t(z-x)(z-y).$$

Since P is symmetric in the variables x, y, z, proving that the inequality holds when $x \geq y \geq z$ will prove it for all other orderings, and so all real triples x, y, z. So we may assume without loss of generality that $x \geq y \geq z$. We may rewrite

$$P(x, y, z) = (x-y)(x^t(x-z) - y^t(y-z)) + z^t(x-z)(y-z).$$

Using the ordering $x \geq y \geq z$, the following four inequalities hold for all real x, y, z:

$$\begin{aligned} x - y &\geq 0, \\ x - z &\geq y - z, \\ x - z &\geq 0, \\ y - z &\geq 0. \end{aligned}$$

If x, y, z are non-negative, then $x^t \geq y^t$ as well for all positive real t, thus proving that the above expression for $P(x, y, z)$ must be non-negative.
Alternatively, suppose x, y, z are any real numbers and t is an even positive integer (the $t = 0$ case is easy to verify separately). We will show that $P(x, y, z)$ is non-negative by doing casework on the sign of y. The four inequalities above still hold and will be useful.

- If $y = 0$, then we can show that

  $$P(x, y, z) = (x-z)(x^{t+1} - z^{t+1}),$$

  which must be non-negative because $x \geq y \geq z$ implies that $x \geq 0$ and $0 \geq z$, and $t+1$ is an odd positive integer.

- If $y > 0$, then $x > 0$ too. Then

$$P(x, y, z) = (x - y)(x^t(x - z) - y^t(y - z)) + z^t(x - z)(y - z)$$

  is non-negative because $x^t \geq y^t$, and $z^t \geq 0$ as well since t is an even positive integer.

- If $y < 0$, then $z < 0$ too. Then

$$P(x, y, z) = x^t(x - y)(x - z) + (y - z)(z^t(x - z) - y^t(x - y))$$

  is non-negative because $z^t \geq y^t$, and $x^t \geq 0$ as well since t is an even positive integer.

Thus, Schur's inequality holds for all real x, y, z if t is an even positive integer.
Now we have to prove the biconditional equality condition, which can be stated in logical notation as

$$(x = y \wedge y = z \wedge z = x) \vee (x = y \wedge z = 0) \vee (y = z \wedge x = 0) \vee (z = x \wedge y = 0).$$

It is easy to show that if any one of the four components of this logical disjunction holds, then $P(x, y, z) = 0$. Now suppose the negation of the equality condition holds:

$$(x \neq y \vee y \neq z \vee z \neq x) \wedge (x \neq y \vee z \neq 0) \wedge (y \neq z \vee x \neq 0) \wedge (z \neq x \vee y \neq 0).$$

First we break this up into two disjoint cases: x, y, z are all distinct or at least two of x, y, z are equal. We can further break the latter case into three cases: $x = y$ or $y = z$ or $z = x$, which are three disjoint cases because any two of the equalities holding would cause $x = y = z$, contradicting the condition $(x \neq y) \vee (y \neq z) \vee (z \neq x)$ in the negation. In the $x = y$ case, the negated condition becomes

$$(y \neq z \vee z \neq x) \wedge (z \neq 0) \wedge (y \neq z \vee x \neq 0) \wedge (z \neq x \vee y \neq 0).$$

It must be true that both $y \neq z$ and $z \neq x$, otherwise $x = y = z$ is true which we have already stated to be contradictory. So the $x = y$ case of the negation is equivalent to

$$(x = y) \wedge (x \neq z) \wedge (y \neq z) \wedge (z \neq 0).$$

Similarly, the $y = z$ case is equivalent to

$$(y = z) \wedge (y \neq x) \wedge (z \neq x) \wedge (x \neq 0)$$

and the $z = x$ case is equivalent to

$$(z = x) \wedge (z \neq y) \wedge (x \neq y) \wedge (y \neq 0).$$

We will treat each of these four cases separately by showing that they all lead to $P(x, y, z) > 0$. The reader should check that the arguments below simultaneously hold in the case where x, y, z are non-negative real numbers with t a positive real number, and the case where x, y, z are any real numbers with t an even positive integer.

- In the case that x, y, z are all distinct, we will give an argument for the strict ordering $x > y > z$ which may be adapted to any other strict ordering since P is symmetric in x, y, z. If $x > y > z$, then

$$\begin{aligned} P(x,y,z) &= (x-y)(x^t(x-z) - y^t(y-z)) + z^t(x-z)(y-z) \\ &\geq (x-y)(x^t(x-z) - y^t(y-z)) \\ &> (x-y)(x^t(y-z) - y^t(y-z)) \\ &= (x-y)(y-z)(x^t - y^t) \\ &> 0. \end{aligned}$$

- In the $x = y$ case,

$$P(x,y,z) = z^t(z-x)(z-y) = z^t(z-x)^2,$$

which must be positive because $z \neq 0$ and $x \neq z$. The $y = z$ and $z = x$ cases follow symmetrically.

Thus, equality cannot hold in any of the four cases, so the stated equality condition is biconditional. ■

Corollary 11.15. The $t = 0, 1, 2, 3$ cases of Schur's inequality can appear in various disguised ways using special factorizations of multivariable polynomials. For $t = 0$, we get the classic

$$x^2 + y^2 + z^2 \geq xy + yz + zx.$$

For $t = 1$, a factored version is

$$xyz \geq (-x + y + z)(x - y + z)(x + y - z)$$

and an equivalent form that is stronger than Nesbitt's inequality (Problem 11.13) is

$$\frac{x}{y+z} + \frac{y}{z+x} + \frac{z}{x+y} + \frac{4xyz}{(x+y)(y+z)(z+x)} \geq 2.$$

A factored form of $t = 2$ is

$$(x+y+z)(x^3+y^3+z^3+3xyz) \geq 2(x^2+y^2+z^2)(xy+yz+zx).$$

For $t = 3$, we find that

$$x^2 + y^2 + z^2 + \frac{6xyz}{x+y+z} + \frac{(x+y+z)xyz}{x^2+y^2+z^2} \geq 2(xy+yz+zx).$$

Problem 11.16 (Euler's inequality). Prove that, for every triangle, if R is its circumradius and r is its inradius, then $R \geq 2r$, and show that equality holds if and only if the triangle is equilateral. It will be helpful to note that

$$\begin{aligned} R &= \frac{abc}{\sqrt{(a+b+c)(-a+b+c)(a-b+c)(a+b-c)}}, \\ r &= \frac{1}{2} \cdot \sqrt{\frac{(-a+b+c)(a-b+c)(a+b-c)}{a+b+c}}, \end{aligned}$$

where a, b, c are the side lengths of the triangle. See [9] for a different short proof that uses excircles, published by the author of this book.

Example 11.17 (Gerretsen's inequalities)**.** Prove that, for every triangle, if R is its circumradius, r is its inradius and s is its semiperimeter, then

$$16Rr - 5r^2 \leq s^2 \leq 4R^2 + 4Rr + 3r^2,$$

with equality holding in either inequality if and only if the triangle is equilateral. Relevant formulas for R and r were stated in Problem 11.16 in terms of the side lengths a, b, c of the triangle; the semiperimeter is defined as $\dfrac{a+b+c}{2}$.

Solution. There is a technique called Ravi substitution, which defines

$$\begin{aligned} x &= s - a, \\ y &= s - b, \\ z &= s - c. \end{aligned}$$

Thanks to the triangle inequality, x, y, z are all positive real numbers. We can solve for a, b, c to get

$$\begin{aligned} a &= y + z, \\ b &= z + x, \\ c &= x + y. \end{aligned}$$

Now we can rewrite the formulas for R, r, s in terms of x, y, z instead of a, b, c:

$$\begin{aligned} R &= \frac{(x+y)(y+z)(z+x)}{4\sqrt{(x+y+z)xyz}}, \\ r &= \sqrt{\frac{xyz}{x+y+z}}, \\ s &= x + y + z. \end{aligned}$$

Substituting these into $16Rr - 5r^2 \leq s^2$ and clearing the denominators yields

$$4(x+y)(y+z)(z+x) - 5xyz \leq (x+y+z)^3,$$

which turns out to be another way of writing Schur inequality for $t = 1$,

$$x(x-y)(x-z) + y(y-z)(y-x) + z(z-x)(z-y) \geq 0.$$

Since none of x, y, z can be 0, equality holds if and only if $x = y = z$, which is true if and only if $a = b = c$.

The other side of this double inequality is messier. Without showing the details, we can substitute the formulas for R, r, s from Problem 11.16 into $s^2 \leq 4R^2 + 4Rr + 3r^2$, clear the denominators, expand every term, and rearrange them to produce the equivalent inequality

$$\begin{aligned} &x^4(y-z)^2 + y^4(z-x)^2 + z^4(x-y)^2 \\ &+ 2(u(u-v)(u-w) + v(v-w)(v-u) + w(w-u)(w-v)) \geq 0, \end{aligned}$$

where

$$u = xy, v = yz, w = zx.$$

The first three terms are non-negative by the trivial inequality and the last three terms together are non-negative by Schur's inequality for $t = 1$. Again, since none of x, y, z can be 0, equality holds if and only if $x = y = z$, which is true if and only if the triangle is equilateral. The reader should work out the expansion manually as performing such computations is not uncommon, and so it is a good skill to practice. ■

Problem 11.16 and Example 11.17 not just curiosities. Euler's inequality and Gerretsen's inequalities are the backbone of many proofs that use the powerful *Rrs* technique of proving geometric inequalities about triangles. This is a method that expresses all terms in a geometric inequality purely in terms of R, r, s (as opposed to other quantities, like the side lengths) and then utilizes Euler or Gerretsen, which are usually strong enough. In sharper cases, there is a monster available called Blundon's inequalities.

Theorem 11.18 (Rearrangement inequality)**.** Suppose n is a positive integer and suppose

$$(x_1, x_2, \ldots, x_n) \text{ and } (y_1, y_2, \ldots, y_n)$$

are non-decreasing n-tuples of real numbers. By non-decreasing, we mean that

$$x_1 \leq x_2 \leq \cdots \leq x_n,$$
$$y_1 \leq y_2 \leq \cdots \leq y_n.$$

Then, for every bijection $\sigma : [n] \to [n]$, it holds that

$$\sum_{i=1}^{n} x_i y_{n-i+1} \leq \sum_{i=1}^{n} x_i y_{\sigma(i)} \leq \sum_{i=1}^{n} x_i y_i.$$

We have not seen an equality condition for the rearrangement inequality that holds biconditionally, but the following sufficient criteria hold. If $(x_1, x_2, \ldots, x_n)$ and $(y_1, y_2, \ldots, y_n)$ are strictly increasing instead of merely non-decreasing, meaning

$$x_1 < x_2 < \cdots < x_n,$$
$$y_1 < y_2 < \cdots < y_n,$$

then equality holds in the right inequality if and only if $\sigma(i) = i$ for all i, and equality holds in the left inequality if and only if $\sigma(i) = n - i + 1$ for all i. Regardless of whether the strict inequalities hold, another sufficient equality criterion is that at least one of the n-tuples $(x_1, x_2, \ldots, x_n)$ and $(y_1, y_2, \ldots, y_n)$ consists of all equal entries.

Proof. We will first prove the following assertion by induction on positive integers n: For all n-tuples of real numbers $(x_1, x_2, \ldots, x_n)$ and $(y_1, y_2, \ldots, y_n)$ and all bijections $\sigma : [n] \to [n]$, it holds that

$$\sum_{i=1}^{n} x_i y_{\sigma(i)} \leq \sum_{i=1}^{n} x_i y_i,$$

with equality holding in the case of $(x_1, x_2, \ldots, x_n)$ and $(y_1, y_2, \ldots, y_n)$ being strictly increasing if and only if σ is the identity map on $[n]$.
In the base case $n = 1$, the only bijection $\sigma : [1] \to [1]$ is the identity map $\sigma(1) = 1$. The result clearly holds in this case. Now suppose the result holds for some positive integer n. Let

$$(x_1, x_2, \ldots, x_n, x_{n+1}) \text{ and } (y_1, y_2, \ldots, y_n, y_{n+1})$$

be non-decreasing $(n+1)$-tuples of real numbers, and let $\sigma : [n+1] \to [n+1]$ be a bijection. We split the argument into two cases: $\sigma(n+1) = n+1$ and $\sigma(n+1) \neq n+1$.
If $\sigma(n+1) = n+1$, then we can cancel $x_{n+1}y_{\sigma(n+1)} = x_{n+1}y_{n+1}$ from both sides of the desired inequality

$$\sum_{i=1}^{n+1} x_i y_{\sigma(i)} \leq \sum_{i=1}^{n+1} x_i y_i$$

so that it is equivalent to prove that

$$\sum_{i=1}^{n} x_i y_{\sigma(i)} \leq \sum_{i=1}^{n} x_i y_i.$$

This follows from invoking the induction hypothesis because the restriction $\sigma|_{[n]}$ is a bijection from $[n]$ to $[n]$ in this case, and the non-decreasing property is preserved. For the sufficient equality condition, suppose

$$(x_1, x_2, \ldots, x_n, x_{n+1}) \text{ and } (y_1, y_2, \ldots, y_n, y_{n+1})$$

are strictly increasing. By the induction hypothesis, equality holds in

$$\sum_{i=1}^{n} x_i y_{\sigma(i)} \leq \sum_{i=1}^{n} x_i y_i$$

if and only if $\sigma|_{[n]}$ is the identity map on $[n]$. Since we have assumed that $\sigma(n+1) = n+1$, equality holds in

$$\sum_{i=1}^{n+1} x_i y_{\sigma(i)} \leq \sum_{i=1}^{n+1} x_i y_i$$

if and only if σ is the identity map on $[n+1]$.
On the other hand, suppose $\sigma(k) = n+1$ for some $k \in [n]$. We define a bijection τ based on σ by swapping the outputs corresponding to the inputs k and $n+1$. That is, we define $\tau : [n] \to [n]$ by

$$\tau(i) = \begin{cases} \sigma(i) & \text{if } i \neq k, n+1 \\ \sigma(n+1) & \text{if } i = k \\ \sigma(k) = n+1 & \text{if } i = n+1 \end{cases}.$$

By working backwards, we can show that it is true that

$$\sum_{i=1}^{n+1} x_i y_{\sigma(i)} \leq \sum_{i=1}^{n+1} x_i y_{\tau(i)}$$

because cancelling equal terms from both sides shows that it is equivalent to

$$\begin{aligned} x_k y_{\sigma(k)} + x_{n+1} y_{\sigma(n+1)} &\leq x_k y_{\tau(k)} + x_{n+1} y_{\tau(n+1)} \\ 0 &\leq x_k (y_{\tau(k)} - y_{\sigma(k)}) + x_{n+1}(y_{\tau(n+1)} - y_{\sigma(n+1)}) \\ &= x_k (y_{\sigma(n+1)} - y_{n+1}) + x_{n+1}(y_{n+1} - y_{\sigma(n+1)}) \\ &= (x_{n+1} - x_k)(y_{n+1} - y_{\sigma(n+1)}). \end{aligned}$$

This is true because $x_{n+1} \geq x_k$ and $y_{n+1} \geq y_{\sigma(n+1)}$. Then

$$\begin{aligned} \sum_{i=1}^{n+1} x_i y_{\sigma(i)} &\leq \sum_{i=1}^{n+1} x_i y_{\tau(i)} \\ &= \sum_{i=1}^{n} x_i y_{\tau(i)} + x_{n+1} y_{\tau(n+1)} = \sum_{i=1}^{n} x_i y_{\tau(i)} + x_{n+1} y_{n+1} \\ &\leq \sum_{i=1}^{n} x_i y_i + x_{n+1} y_{n+1} = \sum_{i=1}^{n+1} x_i y_i, \end{aligned}$$

where we used the induction hypothesis in the penultimate step. Thus, the rearrangement inequality holds in the case of $\sigma(n+1) \neq n+1$. For the sufficient equality condition, suppose $(x_1, x_2, \ldots, x_n, x_{n+1})$ and $(y_1, y_2, \ldots, y_n, y_{n+1})$ are strictly increasing. Then the strict inequality

$$(x_{n+1} - x_k)(y_{n+1} - y_{\sigma(n+1)}) > 0$$

holds because $x_{n+1} > x_k$ due to $k < n+1$ and $y_{n+1} > y_{\sigma(n+1)}$ due to $\sigma(n+1) \neq n+1$. So equality cannot hold for the overarching inequality because equality is impossible in this intermediate step. This completes the induction.
Finally, we will work on the left inequality

$$\sum_{i=1}^{n} x_i y_{n-i+1} \leq \sum_{i=1}^{n} x_i y_{\sigma(i)}.$$

Since $y_1 \leq y_2 \leq \cdots \leq y_n$, multiplying through by -1 yields $-y_n \leq -y_{n-1} \leq \cdots \leq -y_1$. By the right inequality that we painstakingly proved by induction, it holds for any bijection $\phi : [n] \to [n]$ that

$$\sum_{i=1}^{n} x_i (-y_{\phi(n-i+1)}) \leq \sum_{i=1}^{n} x_i (-y_{n-i+1}),$$

which is equivalent to

$$\sum_{i=1}^{n} x_i y_{n-i+1} \leq \sum_{i=1}^{n} x_i y_{\phi(n-i+1)}.$$

For any bijection $\sigma : [n] \to [n]$, define the bijection $\phi(i) = \sigma(n-i+1)$ for all i. Then $\sigma(i) = \phi(n-i+1)$; the intuition behind this is that ϕ and σ are reflections of each other. For this ϕ, the preceding inequality is

$$\sum_{i=1}^{n} x_i y_{n-i+1} \leq \sum_{i=1}^{n} x_i y_{\sigma(i)},$$

which is what we wanted to prove. Again, for the sufficient equality condition, suppose

$$(x_1, x_2, \ldots, x_n) \text{ and } (y_1, y_2, \ldots, y_n)$$

are strictly increasing. Then $(x_1, x_2, \ldots, x_n)$ and the reflected n-tuple $(-y_n, -y_{n-1}, \ldots, -y_1)$ are strictly increasing. Equality holds for the left side of the rearrangement inequality if and only if

$$\begin{aligned}
\sum_{i=1}^{n} x_i y_{n-i+1} &= \sum_{i=1}^{n} x_i y_{\sigma(i)}, \\
\sum_{i=1}^{n} x_i (-y_{\sigma(i)}) &= \sum_{i=1}^{n} x_i (-y_{n-i+1}), \\
\sum_{i=1}^{n} x_i (-y_{\phi(n-i+1)}) &= \sum_{i=1}^{n} x_i (-y_{n-i+1}).
\end{aligned}$$

By the equality condition for the right side of the rearrangement inequality, equality holds in the last inequality above if and only if $\phi(n-i+1) = n-i+1$ for all $i \in [n]$, which is equivalent to it being true that $\sigma(i) = n-i+1$ for all i.
The other sufficient equality criterion, where at least one of the tuples

$$(x_1, x_2, \ldots, x_n) \text{ or } (y_1, y_2, \ldots, y_n)$$

is constant, is obvious in general, and we leave its two-line verification to the reader. ■

Corollary 11.19 (Chebyshev's inequality)**.** Suppose n is a positive integer and

$$\begin{aligned}
&(x_1, x_2, \ldots, x_n), \\
&(y_1, y_2, \ldots, y_n)
\end{aligned}$$

are each non-decreasing n-tuples of real numbers. Then

$$\frac{1}{n} \cdot \sum_{i=1}^{n} x_i y_i \geq \left(\frac{1}{n} \cdot \sum_{i=1}^{n} x_i \right) \left(\frac{1}{n} \cdot \sum_{j=1}^{n} y_j \right) \geq \frac{1}{n} \cdot \sum_{i=1}^{n} x_i y_{n-i+1}.$$

A case in which equality holds is when at least one of the n-tuples consists of all equal numbers.

Proof. By applying the rearrangement inequality to the n cyclically defined bijections

$$\begin{aligned}
&\sigma_1 : (1, 2, 3, \ldots, n-2, n-1, n) \mapsto (1, 2, 3, \ldots, n-2, n-1, n), \\
&\sigma_2 : (1, 2, 3, \ldots, n-2, n-1, n) \mapsto (2, 3, 4, \ldots, n-1, n, 1), \\
&\sigma_3 : (1, 2, 3, \ldots, n-2, n-1, n) \mapsto (3, 4, 5, \ldots, n, 1, 2), \\
&\qquad\qquad\qquad\qquad \vdots \\
&\sigma_n : (1, 2, 3, \ldots, n-2, n-1, n) \mapsto (n, 1, 2, \ldots, n-3, n-2, n-1),
\end{aligned}$$

and adding up the rearrangement inequalities, we get

$$n \cdot \sum_{i=1}^{n} x_i y_i \geq \sum_{j=1}^{n} \sum_{i=1}^{n} x_i y_{\sigma_j(i)} \geq n \cdot \sum_{i=1}^{n} x_i y_{n-i+1}.$$

Using the discrete Fubini's principle, the middle term can be rewritten as

$$\begin{aligned} \sum_{j=1}^{n} \sum_{i=1}^{n} x_i y_{\sigma_j(i)} = \sum_{i=1}^{n} \sum_{j=1}^{n} x_i y_{\sigma_j(i)} &= \sum_{i=1}^{n} \left(x_i \sum_{j=1}^{n} y_{\sigma_j(i)} \right) \\ &= \sum_{i=1}^{n} \left(x_i \sum_{j=1}^{n} y_j \right) = \left(\sum_{i=1}^{n} x_i \right) \left(\sum_{j=1}^{n} y_j \right). \end{aligned}$$

If $x_1 = x_2 = \cdots = x_n$ then equality clearly holds, and similarly if $y_1 = y_2 = \cdots = y_n$. ■

11.3 Convexity

Definition 11.20. Let I be an interval (not necessarily bounded) in $\mathbb{R}$ and $f : I \to \mathbb{R}$ be a function. Then f is a **convex function** if, for all $x, y \in I$ such that $x \neq y$ and for all $\lambda \in (0, 1)$,

$$f((1-\lambda)x + \lambda y) \leq (1-\lambda)f(x) + \lambda f(y).$$

Note that this inequality automatically holds for $\lambda = 0$ and $\lambda = 1$ (try it out!), which is why it would be redundant to include them in the definition. If the inequality holds strictly for all $x \neq y$ in I and all $\lambda \in (0, 1)$, then f is said to be a **strictly convex function** (in this case, it would not make sense to speak of $\lambda = 0$ or $\lambda = 1$). If the direction of the inequality is reversed, then we use the corresponding terms **concave** or **strictly concave**. In terms of graphs, convexity means that the line segment joining $(x, f(x))$ to $(y, f(y))$ does not dip below the graph of f from x to y, whereas strict convexity means the segment, excluding the endpoints, lies strictly above the graph of f.

Example. The second derivative of a function (meaning, the derivative of the derivative) is useful in determining its convexity, but we will not use this technique from calculus. Instead, we provide a list of functions that are well-known to be convex:

- x^n on $[0, \infty)$ for odd positive integers n
- x^n on $\mathbb{R}$ for even positive integers n
- b^x on $\mathbb{R}$ for real b such that $0 < b \neq 1$
- $\dfrac{1}{x^n}$ on $(0, \infty)$ for all positive integers n

Theorem 11.21 (Jensen's inequality)**.** Let I be a real interval, $f : I \to \mathbb{R}$ be a function, n be a positive integer, $x_1, x_2, \ldots, x_n \in I$, $\lambda_1, \lambda_2, \ldots, \lambda_n \in [0, 1]$ such that

$$\lambda_1 + \lambda_2 + \cdots + \lambda_n = 1.$$

Then

$$f(\lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_n x_n) \leq \lambda_1 f(x_1) + \lambda_2 f(x_2) + \cdots + \lambda_n f(x_n).$$

If f is strictly convex, equality holds if and only if all x_i, which have non-zero λ_i, are equal. A special case of the inequality occurs at

$$\lambda_1 = \lambda_2 = \cdots = \lambda_n = \frac{1}{n},$$

where the inequality becomes

$$f\left(\frac{x_1 + x_2 + \cdots + x_n}{n}\right) \leq \frac{f(x_1) + f(x_2) + \cdots + f(x_n)}{n}.$$

Proof. We proceed by induction on n. For $n = 1$, $\lambda_1 = 1$ and Jensen says

$$f(\lambda_1 x_1) \leq \lambda_1 f(x_1),$$

which is true with equality. For $n = 2$, Jensen says

$$f(\lambda_1 x_1 + \lambda_2 x_2) \leq \lambda_1 f(x_1) + \lambda_2 f(x_2).$$

This is true by the convexity of f because $\lambda_1 = 1 - \lambda_2$.
Now suppose the result holds for some $n - 1 \in \mathbb{Z}_+$. As the inductive step, suppose we are working in the successive case of n numbers. Our strategy will be to separate $\lambda_n x_n$ from the rest of the terms so that the convexity of f can be used, followed by the induction hypothesis on the remaining $n - 1$ terms. If $\lambda_n = 1$, then all other λ_i are 0, which reduces the scenario to the $n = 1$ case. So suppose $\lambda_n \neq 1$. First we perform the manipulation

$$f(\lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_n x_n) \leq f\left[(1 - \lambda_n)\left(\frac{\lambda_1}{1 - \lambda_n} \cdot x_1 + \cdots + \frac{\lambda_{n-1}}{1 - \lambda_n} \cdot x_{n-1}\right) + \lambda_n x_n\right].$$

Since $(1 - \lambda_n) + \lambda_n = 1$, the convexity of f says that an upper bound on the above expression is

$$(1 - \lambda_n) f\left(\frac{\lambda_1}{1 - \lambda_n} \cdot x_1 + \frac{\lambda_2}{1 - \lambda_n} \cdot x_2 + \cdots + \frac{\lambda_{n-1}}{1 - \lambda_n} \cdot x_{n-1}\right) + \lambda_n f(x_n).$$

Now we wish to use the induction hypothesis on the first term. Note that

$$\sum_{i=1}^{n-1} \frac{\lambda_i}{1 - \lambda_n} = \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_{n-1}}{1 - \lambda_n} = \frac{1 - \lambda_n}{1 - \lambda_n} = 1.$$

Rearranging the indices if needed, we may assume that x_1 is the smallest among the x_i and that x_n is the largest among the x_i. Then

$$x_1 \leq x_1 \cdot \sum_{i=1}^{n-1} \frac{\lambda_i}{1 - \lambda_n} \leq \sum_{i=1}^{n-1} \left(\frac{\lambda_i}{1 - \lambda_n} \cdot x_i\right) \leq x_n \cdot \sum_{i=1}^{n-1} \frac{\lambda_i}{1 - \lambda_n} = x_n,$$

so the central sum actually lies in the original interval I. By the induction hypothesis,

$$\begin{aligned}(1 - \lambda_n) f\left[\sum_{i=1}^{n-1} \left(\frac{\lambda_i}{1 - \lambda_n} \cdot x_i\right)\right] + \lambda_n f(x_n) &\leq (1 - \lambda_n) \sum_{i=1}^{n-1} \left(\frac{\lambda_i}{1 - \lambda_n} f(x_i)\right) + \lambda_n f(x_n) \\ &= \lambda_1 f(x_1) + \lambda_2 f(x_2) + \cdots + \lambda_n f(x_n).\end{aligned}$$

This proves Jensen's inequality. Now we have to tackle the equality condition. Supposing f is strictly convex, we will make some reductions:

- If some λ_k is 1, then all other λ_i are 0. So the condition that "all x_i, for which λ_i is non-zero, are equal" becomes trivial. So we will suppose in the below cases that no λ_k is 1.
- If we can prove the equality criterion for all $\lambda_1, \lambda_2, \ldots, \lambda_n \in (0,1)$, it will imply the case where some λ_k are 0 because those $\lambda_k x_k$ and $\lambda_k f(x_k)$ terms disappear from the inequality, leading to a case where no λ_i is 0. So suppose no λ_k is 0 in the final case below.
- As stated at the ends of the first and second cases, we are assuming that all λ_i are in $(0,1)$. We break this case into several steps:
  - In one direction, if all the x_i are equal, then it is not difficult to verify that equality holds in Jensen. Below, we tackle the other direction.
  - If all the x_i are distinct, then we proceed by induction to show that Jensen is strict. In the base case of $n = 1$, since f is strictly convex, the definition of strict convexity requires that, if $x_1 \neq x_2$ and λ_1, λ_2 are non-zero, then the inequality is strict. Assuming the induction hypothesis for $x_1 < x_2 < \cdots < x_n$ for some $n \geq 2$, we prove it for $x_1 < x_2 < \cdots < x_n < x_{n+1}$ by noting that the application of the inductive hypothesis in the proof of Jensen makes the inequality strict.
  - Now suppose the x_i are neither all distinct nor all equal for some $n > 2$ (n must be strictly greater than 2 because two numbers are either the same as each other or different). Then the terms of Jensen, which says

$$f\left(\sum_{i=1}^{n} \lambda_i x_i\right) \leq \sum_{i=1}^{n} \lambda_i f(x_i),$$

    can be regrouped by collecting like terms as

$$f\left(\sum_{i=1}^{m} \ell_i y_i\right) \leq \sum_{i=1}^{m} \ell_i f(y_i),$$

    where the y_i are distinct and $m < n$.

■

Problem 11.22 (Power means inequality ladder). Let $n \in \mathbb{Z}_+$, $(x_i)_{i=1}^n$ be a list of positive reals, and $(\lambda_i)_{i=1}^n$ be a list of positive reals such that

$$\lambda_1 + \lambda_2 + \cdots + \lambda_n = 1.$$

Prove that, if $r > s$ are non-zero real numbers, then

$$(\lambda_1 x_1^r + \lambda_2 x_2^r + \cdots + \lambda_n x_n^r)^{\frac{1}{r}} \geq (\lambda_1 x_1^s + \lambda_2 x_2^s + \cdots + \lambda_n x_n^s)^{\frac{1}{s}}.$$

Also prove that equality holds if and only if $x_1 = x_2 = \cdots = x_n$.

Theorem 11.23 (Weighted AM-GM inequality)**.** Let $n \in \mathbb{Z}_+$, $(x_i)_{i=1}^n$ be a list of non-negative reals, and $(\lambda_i)_{i=1}^n$ be a list of non-negative reals such that $\lambda_1 + \lambda_2 + \cdots + \lambda_n = 1$. Then

$$\lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_n x_n \geq x_1^{\lambda_1} x_2^{\lambda_2} \cdots x_n^{\lambda_n},$$

where we use the temporary convention that $0^0 = 1$. Equality holds if and only if all x_i, such that $\lambda_i \neq 0$, are equal. We can lift the criterion that the λ_i add to 1, and instead set $\lambda_1 + \lambda_2 + \cdots + \lambda_n = \lambda$ to get the corollary

$$\frac{\lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_n x_n}{\lambda} \geq (x_1^{\lambda_1} x_2^{\lambda_2} \cdots x_n^{\lambda_n})^{\frac{1}{\lambda}}.$$

Setting all $\lambda_i = 1$ in the corollary yields the ordinary AM-GM inequality.

Proof. We assume, without loss of generality, that all x_i are positive, as any x_i being 0 reduces the inequality to a case with fewer variables and where all remaining x_i are positive. The key is to use Jensen's inequality with the convex function $f(x) = e^x$. It is helpful to observe that

$$e^{t \cdot \ln x} = x^t$$

holds for all positive real x and all real t. By Jensen,

$$\begin{aligned} x_1^{\lambda_1} x_2^{\lambda_2} \cdots x_n^{\lambda_n} &= e^{\lambda_1 \ln x_1} e^{\lambda_2 \ln x_2} \cdots e^{\lambda_n \ln x_n} \\ &= e^{\lambda_1 \ln x_1 + \lambda_2 \ln x_2 + \cdots + \lambda_n \ln x_n} \\ &\leq \lambda_1 e^{\ln x_1} + \lambda_2 e^{\ln x_2} + \cdots + \lambda_n e^{\ln x_n} \\ &= \lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_n x_n. \end{aligned}$$

The equality condition for Jensen's inequality says that equality holds if and only if all of the $\ln x_i$, for which $\lambda_i \neq 0$, are equal. By inverting the logarithm, this condition is true if and only if all of the x_i, for which $\lambda_i \neq 0$, are equal. The corollaries are immediately true. ■

Problem 11.24 (Generalized Young's inequality). Let $n \in \mathbb{Z}_+$, $(x_i)_{i=1}^n$ be a list of non-negative reals, and $(\mu_i)_{i=1}^n$ be a list of positive reals, each greater than 1, such that

$$\frac{1}{\mu_1} + \frac{1}{\mu_2} + \cdots + \frac{1}{\mu_n} = 1.$$

Prove that

$$x_1 x_2 \cdots x_n \leq \frac{x_1^{\mu_1}}{\mu_1} + \frac{x_2^{\mu_2}}{\mu_2} + \cdots + \frac{x_n^{\mu_n}}{\mu_n},$$

with equality holding if and only if

$$x_1^{\mu_1} = x_2^{\mu_2} = \cdots = x_n^{\mu_n}.$$

Theorem 11.25 (Hölder's inequality)**.** Let $n, m \in \mathbb{Z}_+$, and $(x_{1,i})_{i=1}^n, (x_{2,i})_{i=1}^n, \ldots, (x_{m,i})_{i=1}^n$ be m lists, each of n non-negative real numbers. Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be n non-negative real numbers that satisfy $\lambda_1 + \lambda_2 + \cdots + \lambda_n = 1$. Then

$$\prod_{j=1}^{m} \left(\sum_{i=1}^{n} x_{j,i} \right)^{\lambda_j} \geq \sum_{i=1}^{n} \left(\prod_{j=1}^{m} x_{j,i}^{\lambda_j} \right),$$

with the convention that $0^0 = 1$. This is written in more leisurely notation as

$$\begin{aligned}&(x_{1,1} + x_{1,2} + \cdots + x_{1,n})^{\lambda_1} \cdot (x_{2,1} + x_{2,2} + \cdots + x_{2,n})^{\lambda_2} \cdots (x_{m,1} + x_{m,2} + \cdots + x_{m,n})^{\lambda_m} \\ &\geq (x_{1,1}^{\lambda_1} x_{2,1}^{\lambda_2} \cdots x_{m,1}^{\lambda_m}) + (x_{1,2}^{\lambda_1} x_{2,2}^{\lambda_2} \cdots x_{m,2}^{\lambda_m}) + \cdots + (x_{1,n}^{\lambda_1} x_{2,n}^{\lambda_2} \cdots x_{m,n}^{\lambda_m}).\end{aligned}$$

Visually, we can place the $x_{j,i}$ in an $m \times n$ matrix with each list occupying a row, so that the inequality's left side has the sum of a row in each multiplicand, and the inequality's right side has the product of a column in each summand. Equality holds if and only if, for all $j \in [m]$ such that $\lambda_j \neq 0$,

$$x_{j,1} : x_{j,2} : \cdots : x_{j,m}$$

forms the same ratio, or one of the lists consists of all 0's.

Proof. If some $x_{i,j}$ is 0, the desired inequality is weaker than (and so it is implied by) Hölder with the entire i^{th} column of the matrix being 0; many terms disappear in the latter to produce a Hölder's inequality with $(n-1)$-length lists. This can be repeated for any other zero terms. If any list consists of all 0's, then the inequality is trivially true as an equality. Thus, we may assume, without loss of generality, that all $x_{j,i}$ are positive. Dividing both sides of the inequality by the left side yields

$$1 \geq \sum_{i=1}^{n} \prod_{j=1}^{m} \left(\frac{x_{j,i}}{\sum_{k=1}^{n} x_{j,k}} \right)^{\lambda_j}.$$

For each $j \in [m]$ and $i \in [n]$, let

$$y_{j,i} = \frac{x_{j,i}}{\sum_{k=1}^{m} x_{j,k}},$$

so that we obtain the condition

$$y_{j,1} + y_{j,2} + \cdots + y_{j,n} = 1$$

for each $j \in [m]$. We wish to prove that

$$\sum_{i=1}^{n} \prod_{j=1}^{m} y_{j,i}^{\lambda_j} \leq 1.$$

By applying the weighted AM-GM inequality n times and the discrete Fubini's principle,

$$\begin{aligned}\sum_{i=1}^{n} \prod_{j=1}^{m} y_{j,i}^{\lambda_j} &\leq \sum_{i=1}^{n} \sum_{j=1}^{m} \lambda_j y_{j,i} = \sum_{j=1}^{m} \sum_{i=1}^{n} \lambda_j y_{j,i} \\ &= \sum_{j=1}^{m} \left(\lambda_j \sum_{i=1}^{n} y_{j,i} \right) \\ &= \sum_{j=1}^{m} \lambda_j = 1,\end{aligned}$$

which establishes the inequality.

Now we need to work on the equality condition. By the equality condition for the weighted AM-GM inequality, equality holds if and only if, for each fixed $i \in [n]$ and for all $j \in [m]$ such that $\lambda_j \neq 0$, the $y_{j,i}$ are equal. Taking all λ_j to be non-zero for ease of notation, this means there exists a positive real t_i such that

$$t_i = y_{1,i} = y_{2,i} = \cdots = y_{m,i}.$$

For each $j \in [m]$, let

$$s_j = x_{j,1} + x_{j,2} + \cdots + x_{j,n} \neq 0.$$

Then, for each $i \in [n]$ and $j \in [m]$,

$$t_i = y_{j,i} = \frac{x_{j,i}}{s_j} \implies x_{j,i} = s_j t_i.$$

Fixing $j \in [m]$, we get

$$x_{j,1} : x_{j,2} : \cdots : x_{j,n} = s_j t_1 : s_j t_2 : \cdots : s_j t_n = t_1 : t_2 : \cdots : t_n,$$

which is the same for every $j \in [m]$. The converse, that the criterion implies equality, is easy to establish algebraically. If one of the lists is all 0's, equality automatically holds; we addressed this possibility at the beginning of the proof. ■

Problem 11.26. Let $n \in \mathbb{Z}_+$, let $(x_k)_{k=1}^n$ and $(y_k)_{k=1}^n$ each be a list of n non-negative real numbers, and let p and q be real numbers greater than 1 such that

$$\frac{1}{p} + \frac{1}{q} = 1.$$

As a special case of Hölder's inequality, prove that

$$\sum_{k=1}^{n} x_k y_k \leq \left(\sum_{k=1}^{n} x_k^p \right)^{\frac{1}{p}} \cdot \left(\sum_{k=1}^{n} x_k^q \right)^{\frac{1}{q}}.$$

Also, prove that equality holds if and only if the list of x_i's is all 0's, or the list of y_i's is all 0's, or there exists a positive real c such that, for all $k \in [n]$, $y_k^q = c x_k^p$.

Theorem 11.27 (Minkowski's inequality)**.** Let $n \in \mathbb{Z}_+$ and $p \geq 1$ be a real number. The **p-norm** of an n-tuple $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ of real numbers is denoted by and defined as

$$\|\mathbf{x}\|_p = (|x_1|^p + |x_2|^p + \cdots + |x_n|^p)^{\frac{1}{p}}.$$

Minkowski says that, for any two n-tuples $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and any real $p \geq 1$,

$$\|\mathbf{x} + \mathbf{y}\|_p \leq \|\mathbf{x}\|_p + \|\mathbf{y}\|_p.$$

This can be considered to be the triangle inequality on the p-norm. Equality holds if and only if all the entries of $\mathbf{x}$ are 0 or all the entries of $\mathbf{y}$ are 0, or there exists a positive real c such that $\mathbf{y} = c\mathbf{x}$.

Proof. The $p = 1$ case follows from the usual triangle inequality, so we will assume $p > 1$. Then $\dfrac{1}{p} < 1$, and we let $q > 1$ be the real number such that

$$\frac{1}{p} + \frac{1}{q} = 1.$$

By the real triangle inequality,

$$\begin{aligned}
\|\mathbf{x} + \mathbf{y}\|_p^p &= \sum_{k=1}^{n} |x_k + y_k|^p \\
&= \sum_{k=1}^{n} |x_k + y_k|^{p-1} \cdot |x_k + y_k| \\
&\leq \sum_{k=1}^{n} |x_k + y_k|^{p-1} \cdot (|x_k| + |y_k|) \\
&= \sum_{k=1}^{n} |x_k + y_k|^{p-1} \cdot |x_k| + \sum_{k=1}^{n} |x_k + y_k|^{p-1} \cdot |y_k| l.
\end{aligned}$$

So far, by the criterion for equality in the real triangle inequality, equality holds if and only if, for each $k \in [n]$, x_k of y_k are not on opposite sides of 0 on the real number line, meaning there exists a real constant $c_k \geq 0$ such that $y_k = c_k x_k$. Moving forward with proving the inequality, we will use the equivalences

$$\frac{1}{p} + \frac{1}{q} = 1 \iff p + q = pq \iff p = (p-1)q \iff 1 = p - \frac{p}{q}.$$

By Hölder's inequality, we continue with

$$\begin{aligned}
\|\mathbf{x} + \mathbf{y}\|_p^p &\leq \sum_{k=1}^{n} |x_k + y_k|^{p-1} \cdot |x_k| + \sum_{k=1}^{n} |x_k + y_k|^{p-1} \cdot |y_k| \\
&\leq \left(\sum_{k=1}^{n} |x_k + y_k|^{(p-1)q} \right)^{\frac{1}{q}} \left(\sum_{k=1}^{n} |x_k|^p \right)^{\frac{1}{p}} + \left(\sum_{k=1}^{n} |x_k + y_k|^{(p-1)q} \right)^{\frac{1}{q}} \left(\sum_{k=1}^{n} |y_k|^p \right)^{\frac{1}{p}} \\
&= \left(\sum_{k=1}^{n} |x_k + y_k|^p \right)^{\frac{1}{q}} \left(\sum_{k=1}^{n} |x_k|^p \right)^{\frac{1}{p}} + \left(\sum_{k=1}^{n} |x_k + y_k|^p \right)^{\frac{1}{q}} \left(\sum_{k=1}^{n} |y_k|^p \right)^{\frac{1}{p}} \\
&= \|\mathbf{x} + \mathbf{y}\|_p^{\frac{p}{q}} \cdot (\|\mathbf{x}\|_p + \|\mathbf{y}\|_p).
\end{aligned}$$

As a result, the inequality is established as

$$\|\mathbf{x}\|_p + \|\mathbf{y}\|_p \geq \|\mathbf{x} + \mathbf{y}\|_p^{p - \frac{p}{q}} = \|\mathbf{x} + \mathbf{y}\|_p.$$

Now we complete the proof of the equality criteria. If $\mathbf{x} = 0$ or $\mathbf{y} = 0$, equality can be verified to hold, so suppose neither list consists of all 0's. Equality holds in the second part of the

proof if and only if equality holds in the application of Hölder's inequality. By the equality condition on Hölder, equality holds if and only if there exist non-negative real constants d_1, d_2 such that, for all $k \in [n]$,

$$d_1|x_k|^p = |x_k + y_k|^{(p-1)q} = |x_k + y_k|^p,$$
$$d_2|y_k|^p = |x_k + y_k|^{(p-1)q} = |x_k + y_k|^p.$$

Then $d_1|x_k|^p = d_2|y_k|^p$ for all $k \in [n]$. Since $\mathbf{x}, \mathbf{y}$ are non-zero lists, $d_1 \neq 0$ and $d_2 \neq 0$. So, letting $c = \left(\dfrac{d_1}{d_2}\right)^{\frac{1}{p}} > 0$, we get $|y_k| = c|x_k|$ for all $k \in [n]$. The equality condition from the first half of the proof requires that x_k, y_k are not of opposite signs, so we can remove the absolute values to get $\mathbf{x} = c\mathbf{y}$. Conversely, the fact that equality holds under the stated condition is easy to establish algebraically. ■

Theorem 11.28 (Popoviciu's inequality)**.** Let I be a real interval, and $f : I \to \mathbb{R}$ be a convex function. Then, for any $a, b, c \in I$,

$$\frac{2}{3} \cdot \left[f\left(\frac{a+b}{2}\right) + f\left(\frac{b+c}{2}\right) + f\left(\frac{c+a}{2}\right)\right] \leq \frac{f(a) + f(b) + f(c)}{3} + f\left(\frac{a+b+c}{3}\right).$$

Proof. Due to a, b, c occupying symmetric positions, we may assume, without loss of generality, that $a \leq b \leq c$. We will split the proof into cases according to how the middle variable b compares to the average $\dfrac{a+b+c}{3}$. There are two cases:

1. Suppose $b \leq \dfrac{a+b+c}{3}$. It may be verified that

$$\frac{a+b+c}{3} \leq \frac{a+c}{2} \leq c,$$
$$\frac{a+b+c}{3} \leq \frac{b+c}{2} \leq c.$$

Then there exist $\lambda_a, \lambda_b \in [0, 1]$ such that

$$\frac{a+c}{2} = (1-\lambda_a) \cdot \frac{a+b+c}{3} + \lambda_a c,$$
$$\frac{b+c}{2} = (1-\lambda_b) \cdot \frac{a+b+c}{3} + \lambda_b c.$$

The sum of these equations is

$$\begin{aligned}\frac{a+b+2c}{2} &= (2-\lambda_a-\lambda_b) \cdot \frac{a+b+c}{3} + (\lambda_a+\lambda_b)c \\ &= (2-\lambda_a-\lambda_b) \cdot \frac{a+b-2c}{3} + 2c,\end{aligned}$$

which can be simplified to

$$\frac{a+b-2c}{2} = (2-\lambda_a-\lambda_b) \cdot \frac{a+b-2c}{3}.$$

If $a+b-2c=0$, then the fact that $a \leq c$ and $b \leq c$ leads to $a=b=c$, in which case Popoviciu holds. Supposing $a+b \neq 2c$, we can cancel it from both sides to get

$$\frac{1}{2} = \frac{2-\lambda_a-\lambda_b}{3} \implies \lambda_a+\lambda_b = \frac{1}{2}.$$

Now we work on developing upper bounds on the left side of the desired inequality. Due to the convexity of f, Jensen says that

$$\begin{aligned} f\left(\frac{a+b}{2}\right) &\leq \frac{f(a)+f(b)}{2}, \\ f\left(\frac{b+c}{2}\right) &\leq (1-\lambda_a) f\left(\frac{a+b+c}{3}\right) + \lambda_a f(c), \\ f\left(\frac{c+a}{2}\right) &\leq (1-\lambda_b) f\left(\frac{a+b+c}{3}\right) + \lambda_b f(c). \end{aligned}$$

Adding the three and dividing both sides by $\frac{3}{2}$ yields Popoviciu, since $\lambda_a+\lambda_b = \frac{1}{2}$.

2. Suppose $b \geq \frac{a+b+c}{3}$. Then we proceed as in the first case, but with the inequalities

$$\begin{aligned} a \leq \frac{a+c}{2} &\leq \frac{a+b+c}{3}, \\ a \leq \frac{a+b}{2} &\leq \frac{a+b+c}{3}. \end{aligned}$$

The reader is strongly encouraged to write the proof of this case by mimicking the technicalities of the first case.

■

A lesser-known inequality involving convexity is that of Karamata. A proof is easily found in various sources, including by an online search.

11.4 Newton and Maclaurin

Definition 11.29. Let n be a positive integer, k be an integer such that $1 \leq k \leq n$ and let $r_1, r_2, \ldots, r_n$ be a list of real numbers. Recall from Definition 10.44 that the k^{th} symmetric sum of this list is

$$\sigma_k = \sum_{\substack{J \subseteq [n] \\ |J|=k}} \prod_{j \in J} r_j = \sum_{1 \leq i_1 < i_2 < \cdots < i_k \leq n} r_{i_1} r_{i_2} \cdots r_{i_k}.$$

Now we define the **k^{th} symmetric mean** as the average of the terms of the k^{th} symmetric sum. Since the number of k-tuples $(i_1, i_2, \ldots, i_n)$ such that

$$1 \leq i_1 < i_2 < \cdots < i_k \leq n$$

is $\binom{n}{k}$ by a bijective counting argument (see Volume 2), it means the k^{th} symmetric mean of the list $r_1, r_2, \ldots, r_n$ is

$$\mu_k = \frac{\sigma_k}{\binom{n}{k}}.$$

We also extend the definitions to $k = 0$ to get the empty sum $\sigma_0 = 1$ and $\mu_0 = \frac{\sigma_0}{\binom{n}{0}} = 1$.

Theorem 11.30 (Newton's inequalities)**.** If $n \geq 2$ is an integer, $r_1, r_2, \ldots, r_n$ are non-negative real numbers, and k is an integer such that $0 < k < n$, then

$$\mu_k^2 \geq \mu_{k-1}\mu_{k+1}.$$

If all r_i are positive, then equality holds for a specific k if and only if all the r_i are equal to each other. As a result, equality holds for a specific k if and only if it holds for all k.

Proof. We will prove this by induction on $n \geq 2$, handling all k at each step after fixing n. In the base case of $n = 2$, the only possibility for k is $k = 1$, and the inequality says

$$\left(\frac{r_1 + r_2}{2}\right)^2 \geq 1 \cdot (r_1 r_2).$$

This is the most basic case of AM-GM, with equality holding if and only if $r_1 = r_2$.
Now suppose the result holds for any $n - 1$ numbers for some integer n such that $n - 1 \geq 2$. To prove Newton's inequality for n numbers $r_1, r_2, \ldots, r_n$, we will first set $(\nu_k)_{k=1}^{n-1}$ to be the symmetric means of $r_1, r_2, \ldots, r_{n-1}$ (specifically, r_n is excluded). Then, for each integer k such that $0 < k < n$, we split μ_k into the terms that do not contain r_n and the terms that do contain r_n. We then utilize the combinatorial identities (see Volume 2)

$$\begin{aligned}
\binom{n}{k} &= \frac{n}{k} \cdot \binom{n-1}{k-1}, \\
\binom{n}{k} &= \binom{n}{n-k} = \frac{n}{n-k} \cdot \binom{n-1}{n-k-1} = \frac{n}{n-k} \cdot \binom{n-1}{k}
\end{aligned}$$

to get

$$\mu_k = \frac{n-k}{n} \cdot \nu_k + \frac{k}{n} \cdot \nu_{k-1} r_n.$$

Convince yourself that this is true. The preceding recursion is the only lemma that we will need.
If we wish to use the above recursion on μ_{k-1} in $\mu_{k-1}\mu_{k+1}$, it will result in a term with ν_{k-2}, which is sensible only for $k \geq 2$. So we will handle $k = 1$ separately at the outset now. If $k = 1$, then Newton says

$$\begin{aligned}
\left(\frac{r_1 + r_2 + \cdots + r_n}{n}\right)^2 &\geq 1 \cdot \left(\sum_{1 \leq i < j \leq n} r_i r_j\right) \cdot \frac{1}{\binom{n}{2}} \\
&= \frac{2}{n(n-1)} \cdot \sum_{1 \leq i < j \leq n} r_i r_j.
\end{aligned}$$

Further backwards manipulations yield

$$\begin{aligned}(n-1)\left[\sum_{k=1}^{n} r_k^2 + \sum_{1\le i<j\le n} r_i r_j\right] &\ge 2n \sum_{1\le i<j\le n} r_i r_j \\ (n-1)\sum_{k=1}^{n} r_k^2 &\ge 2 \sum_{1\le i<j\le n} r_i r_j \\ \sum_{1\le i<j\le n} (r_i^2 + r_j^2) &\ge \sum_{1\le i<j\le n} 2 r_i r_j \\ \sum_{1\le i<j\le n} (r_i - r_j)^2 &\ge 0,\end{aligned}$$

which is true by the trivial inequality. Equality holds if and only if

$$r_1 = r_2 = \cdots = r_n.$$

Now we may assume that $k \ge 2$. By the recursive lemma,

$$\begin{aligned}\mu_{k-1}\mu_{k+1} &= \left(\frac{n-k+1}{n}\cdot \nu_{k-1} + \frac{k-1}{n}\cdot \nu_{k-2} r_n\right)\cdot\left(\frac{n-k-1}{n}\cdot \nu_{k+1} + \frac{k+1}{n}\cdot \nu_k r_n\right) \\ &= \frac{(n-k+1)(n-k-1)}{n^2}\cdot \nu_{k-1}\nu_{k+1} + \frac{(k-1)(n-k-1)}{n^2}\cdot \nu_{k-2}\nu_{k+1} r_n \\ &+ \frac{(n-k+1)(k+1)}{n^2}\cdot \nu_{k-1}\nu_k r_n + \frac{(k-1)(k+1)}{n^2}\cdot \nu_{k-2}\nu_k r_n^2.\end{aligned}$$

We use the induction hypothesis to get

$$\begin{aligned}\nu_{k-1}\nu_{k+1} &\le \nu_k^2, \\ \nu_{k-2}\nu_k &\le \nu_{k-1}^2,\end{aligned}$$

and multiplying them together yields

$$\nu_{k-2}\nu_{k-1}\nu_k\nu_{k+1} \le \nu_{k-1}^2\nu_k^2.$$

We wish to cancel $\nu_{k-1}\nu_k$ from both sides, but we need to handle the possibilities of $\nu_k = 0$ or $\nu_{k-1} = 0$ first.

- Suppose $\nu_k = 0$. Then the fact that $\nu_{k-1}\nu_{k+1} \le \nu_k^2$ says that $\nu_{k-1} = 0$ or $\nu_{k+1} = 0$.
  - Suppose $\nu_{k-1} = 0$. Then the recursion says that $\mu_k = 0$, so the product of any collection of k of the r_i multiply to 0. Then the product of any collection of $k-1$ of the r_i also multiply to 0, giving $\mu_{k-1} = 0$. This makes Newton's inequality true.
  - Suppose $\nu_{k+1} = 0$. By the above expanded upper bound, $\mu_{k-1}\mu_{k+1} \le 0$, so $\mu_{k-1} = 0$ or $\mu_{k+1} = 0$, either of which makes Newton true by the trivial inequality.
- Suppose $\nu_{k-1} = 0$. Then the fact that $\nu_{k-2}\nu_k \le \nu_{k-1}^2$ says that $\nu_{k-2} = 0$ or $\nu_k = 0$.

- Suppose $\nu_{k-2} = 0$. Again, by the expanded upper bound above, $\mu_{k-1}\mu_{k+1} \leq 0$, which we have previously shown to imply Newton.
- Suppose $\nu_k = 0$. We have shown above that the combination of $\nu_k = 0$ and $\nu_{k-1} = 0$ implies Newton.

So now we may assume that $\nu_{k-1}\nu_k \neq 0$ and cancel it from both sides of

$$\nu_{k-2}\nu_{k-1}\nu_k\nu_{k+1} \leq \nu_{k-1}^2\nu_k^2$$

to get

$$\nu_{k-2}\nu_{k+1} \leq \nu_{k-1}\nu_k.$$

Continuing with the upper bounding of $\mu_{k-1}\mu_{k+1}$, we get

$$\begin{aligned}\mu_{k-1}\mu_{k+1} &\leq \frac{(n-k)^2-1}{n^2} \cdot \nu_k^2 + \frac{(k-1)(n-k-1)}{n^2} \cdot \nu_{k-1}\nu_k r_n \\ &\quad + \frac{(n-k+1)(k+1)}{n^2} \cdot \nu_{k-1}\nu_k r_n + \frac{k^2-1}{n^2} \cdot \nu_{k-1}^2 r_n^2.\end{aligned}$$

To collect the middle two like terms, expansion and simplification yield

$$\frac{(k-1)(n-k-1)}{n^2} + \frac{(n-k+1)(k+1)}{n^2} = \frac{2k(n-k)}{n^2} + \frac{2}{n^2}.$$

Finally, we reach the desired upper bound:

$$\begin{aligned}\mu_{k-1}\mu_{k+1} &\leq \frac{(n-k)^2}{k^2} \cdot \nu_k^2 + \frac{2k(n-k)}{n^2} \cdot \nu_k\nu_{k-1}r_n + \frac{k^2}{n^2} \cdot \nu_{k-1}^2 r_n^2 \\ &\quad - \frac{\nu_k^2}{n^2} + \frac{2}{n} \cdot \nu_{k-1}r_n - \frac{\nu_{k-1}^2 r_n^2}{n^2} \\ &= \left(\frac{n-k}{n} \cdot \nu_k + \frac{k}{n} \cdot \nu_{k-1}r_n\right)^2 - \left(\frac{\nu_k}{n} - \frac{\nu_{k-1}r_n}{n}\right)^2 \\ &\leq \mu_k^2 - 0^2 = \mu_k^2.\end{aligned}$$

For the equality condition, suppose all of the r_i are positive. If equality holds, it must have held in both of

$$\begin{aligned}\nu_{k-1}\nu_{k+1} &\leq \nu_k^2, \\ \left(\frac{\nu_k}{n} - \frac{\nu_{k-1}r_n}{n}\right)^2 &\leq 0.\end{aligned}$$

By the equality criterion in the induction hypothesis, if $\nu_{k-1}\nu_{k+1} = \nu_k^2$, then there exists a positive real r such that

$$r = r_1 = r_2 = \cdots = r_{n-1}.$$

By the second condition, which states $\nu_k = \nu_{k-1}r_n$, we can substitute $r > 0$ for each r_i in ν_k and ν_{k-1} to get

$$r^k = r^{k-1}r_n \implies r = r_n.$$

Conversely, it is not difficult to verify algebraically that, if all r_i are positive and equal, then equality holds in Newton's inequality.

■

Problem 11.31. Let n be a positive integer. Prove that, for any $a_1, a_2, \ldots, a_n \in \mathbb{R}$, if

$$\forall k \in [n-1] : \frac{a_{k-1} + a_{k+1}}{2} \geq a_k,$$

then

$$\forall k \in [n-1] : \frac{a_{k+1} - a_0}{k+1} \geq \frac{a_k - a_0}{k}.$$

Theorem 11.32 (Maclaurin's inequalities)**.** Let $n \geq 2$ be an integer, and $r_2, r_2, \ldots, r_n$ be positive real numbers. For each $k \in [n-1]$,

$$(\mu_k)^{\frac{1}{k}} \geq (\mu_{k+1})^{\frac{1}{k+1}},$$

where we have used the notation of symmetric means (Definition 11.29). Equality holds for a specific k if and only if $r_1 = r_2 = \cdots = r_n$. So equality holds for a specific k if and only if it holds for all k. Altogether, the inequalities state that

$$\mu_1 \geq \sqrt{\mu_2} \geq \sqrt[3]{\mu_3} \geq \cdots \geq \sqrt[n]{\mu_n},$$

where d_1 is the arithmetic mean of the r_i and $\sqrt[n]{\mu_n}$ is the geometric mean. So Maclaurin's inequalities point out gradations or refinements in between the AM-GM inequality.

Proof. By taking the natural logarithm of Newton's inequalities (which is acceptable since all of the r_i are positive, so their symmetric sums are also positive, allow them to be fed into logarithms),

$$\begin{aligned} \mu_k &\geq (\mu_{k-1}\mu_{k+1})^{\frac{1}{2}} \\ \ln \mu_k &\geq \frac{\ln \mu_{k-1} + \ln \mu_{k+1}}{2} \\ \frac{-\ln \mu_{k-1} - \ln \mu_{k+1}}{2} &\geq -\ln \mu_k. \end{aligned}$$

Since this works for all $k \in [n-1]$, we have satisfied the hypotheses of Problem 11.31 for $a_i = \ln \mu_i$. So we can use $\mu_0 = 1$ to conclude that

$$\begin{aligned} \frac{-\ln \mu_k}{k} &= \frac{-\ln \mu_k + \ln \mu_0}{k} = \frac{a_k - a_0}{k} \\ &\leq \frac{a_{k+1} - a_0}{k+1} = \frac{-\ln \mu_{k+1} + \ln \mu_0}{k+1} = \frac{-\ln \mu_{k+1}}{k+1}. \end{aligned}$$

This is equivalent to

$$\begin{aligned} \ln\left[(\mu_{k+1})^{\frac{1}{k+1}}\right] &\leq \ln\left[(\mu_k)^{\frac{1}{k}}\right] \\ (\mu_{k+1})^{\frac{1}{k+1}} &\leq (\mu_k)^{\frac{1}{k}}, \end{aligned}$$

which is what we wanted. If equality holds, then it held in the application of Newton's inequality

$$\mu_k \geq (\mu_{k-1}\mu_{k+1})^{\frac{1}{2}}.$$

Since all of the r_i are positive, equality in Newton here requires $r_1 = r_2 = \cdots = r_n$. Conversely, if all of the r_i are equal, it is clear through some algebra that equality holds in Maclaurin. ■

Another inequality involving symmetric expressions is that of Muirhead. We have not included it because of the complexity of the proof. For a complete proof, see [3]. There are many further techniques that can be studied. The interested reader will find no deficit of resources from which such further results and problem-solving techniques can be learned for proving inequalities. For example, see the famous books [8] and [7], as well as [10].

Appendices

Appendix A

Solutions

> "Yes, mathematics has two faces; it is the rigorous science of Euclid but it is also something else. Mathematics presented in the Euclidean way appears as a systematic, deductive science; but mathematics in the making appears as an experimental, inductive science. Both aspects are as old as the science of mathematics itself."
>
> – *George Pólya, How to Solve It*

Solution 1.10. We will use a sequence of set identities as follows:

$$\begin{aligned} A^c &= \mathcal{U}\backslash A = \mathcal{U} \cap A^c = (A \cup B) \cap A^c = (A \cap A^c) \cup (B \cap A^c) = \varnothing \cup (B \cap A^c) \\ &= (B \cap A^c) \cup \varnothing = (B \cap A^c) \cup (B \cap A) = B \cap (A^c \cup A) = B \cap \mathcal{U} \\ &= B. \end{aligned}$$

Other proofs using other methods are possible, but this was the cleanest one that we could develop.

Solution 1.30. Suppose $f : X \to Y$ is a function, and that $g : Y \to X$ and $h : Y \to X$ are inverses of f. Then g is a left-inverse of f and h is a right-inverse of f. By Theorem 1.29, $g = h$.
Now we will show that left-inverses and right-inverses are not necessarily unique.

- Let $f : \{0, 1\} \to \{0, 1, 2\}$ be defined by $0 \mapsto 0$ and $1 \mapsto 1$. Then there are the two distinct left-inverses $g : \{0, 1, 2\} \to \{0, 1\}$ and $h : \{0, 1, 2\} \to \{0, 1\}$, where both map $0 \mapsto 0$ and $1 \mapsto 1$, but g maps $2 \mapsto 0$ and h maps $2 \mapsto 1$.
- Let $f : \{0, 1\} \to \{0\}$ be defined by $0 \mapsto 0$ and $1 \mapsto 0$. Then there are the two distinct right-inverses $g : \{0\} \to \{0, 1\}$ and $h : \{0\} \to \{0, 1\}$, where g maps $0 \mapsto 0$ and h maps $0 \mapsto 1$.

Solution 1.35. Suppose f is a cyclic function of order m. If $f(x) = f(y)$, then applying f by $(m - 1)$ times to each side of the equation gives

$$x = f^m(x) = f^m(y) = y.$$

Thus, cyclicity implies injectivity.

Solution 1.36. Composing f with itself gives

$$f(f(x)) = \frac{\frac{x-1}{x} - 1}{\left(\frac{x-1}{x}\right)} = \frac{1}{1-x}.$$

Composing f over this again yields

$$f^3(x) = \frac{\frac{1}{1-x} - 1}{\left(\frac{1}{1-x}\right)} = x,$$

as desired. In the problem statement, we reverse-engineered the exclusion of both 0 and 1 from both the domain and codomain so that there is no division by 0 at any point of the compositions and so that the codomain is a subset of the domain.

Solution 1.43. A function $f : X \to Y$ is an injection if and only if f has a left-inverse $g : Y \to X$. Next, g is a left-inverse of f if and only if f is a right-inverse of g. Finally, g has a right-inverse if and only if g is surjective. As all of the steps were reversible, we are done.

Solution 1.44. We need g and h to disagree somewhere, but they need to agree on the range of f. Choosing $f(x) = x^2$ so that its range is $\mathbb{R}_{\geq 0}$, we find that we just need g, h to agree on the non-negative reals and disagree on some subset of the negative reals. So we choose $g(x) = x$ and $h(x) = |x|$. Incidentally, this simultaneously provides an example of functions f, g, h with domains and codomains $\mathbb{R}$ such that $f \circ g = f \circ h$ but $g \neq h$.

Solution 1.45. Injectivity without surjectivity is not difficult to achieve, say using an exponential function, such as $f(x) = 2^x$. For surjectivity without injectivity, we can choose a cubic function with two "bends," such as $g(x) = x(x-1)(x+1) = x^3 - x$.

Solution 1.46. Let f and g have the stated domains and codomains.

1. Suppose $f(x) = f(y)$ for some $x, y \in A$. Then applying g to both sides of the equation yields $g(f(x)) = g(f(y))$ or $(g \circ f)(x) = (g \circ f)(y)$. By the injectivity of $g \circ f$, we find that $x = y$, thereby proving the injectivity of f. For the counterexample to the injectivity of g, we first choose $g(x) = x^2$ as a classic non-injective function. Then we recognize that, in order for $g \circ f$ to be injective, it is necessary that there are not negatives in the range of f. Choosing $f(x) = 2^x$ closes the deal.

2. Since $g \circ f$ is surjective, we find that $(g \circ f)(A) = C$ or $g(f(A)) = C$. As a result,

$$f(A) \subseteq B \implies C = g(f(A)) \subseteq g(B) \subseteq C.$$

   So $C \subseteq g(B)$ and $g(B) \subseteq C$, and antisymmetry gives $g(B) = C$, proving that g is surjective. For the counterexample to the surjectivity of f, we realize that it is necessary for f to map to only a subset of $\mathbb{R}$, yet g must be able to take this subset and map it to the entirety of $\mathbb{R}$. To that end, we choose $f(x) = 2^x$ so that its range is the positive reals, and we define $g(x) = \tan x$ on the domain of $\tan x$ and $g(x) = 0$ (this is an arbitrary value) wherever $\tan x$ has a vertical asymptote. Since $\tan x$ takes, for example, $\left(\frac{\pi}{2}, \frac{3\pi}{2}\right)$ and maps it to the entirety of $\mathbb{R}$, this does the job.

Solution 1.47. An initial clue is that neither f nor g can be a bijection, for otherwise the bijection f^{-1} or g^{-1} could be right-composed or left-composed with $g \circ f$ to get that g or f is bijective, respectively; this is impossible because then we could compose $f \circ g$ in either case to get that $f \circ g$ is a bijection, thwarting our efforts. Moreover, by Problem 1.46, if $g \circ f$ is bijective, its injectivity implies the injectivity of f, and its surjectivity implies the surjectivity of g. Putting it all together, we try $f(x) = 2^x$ as it is injective but not surjective. Choosing $g(x) = \log_2 x$ would give the bijection $(g \circ f)(x) = x$, but we still need to define g on the non-positive reals and ensure that $f \circ g$ is not bijective. To simultaneously satisfy these goals, we define $g(x) = 0$ for $x \leq 0$. This does the trick.

Solution 1.49. We will develop bijections, explicitly or implicitly, with domains and codomains $[0,1] \to (0,1)$, $(0,1) \to \mathbb{R}$, $\mathbb{R} \to (0,\infty)$, and $(0,\infty) \to [0,\infty)$.

1. Making use of Schröder-Bernstein (Theorem 1.48), we will produce an injection from $[0,1]$ to $(0,1)$ and vice versa. These are $x \mapsto \dfrac{x}{2} + \dfrac{1}{4}$ and simply $x \mapsto x$.

2. A bijection from $(0,1)$ to $\mathbb{R}$ is given by $x \mapsto \tan\left[\pi\left(x - \dfrac{1}{2}\right)\right]$.

3. A bijection from $\mathbb{R}$ to $(0,\infty)$ is given by $x \mapsto 2^x$.

4. Again, we will make use of Schröder-Bernstein by producing an injection from $(0,\infty)$ to $[0,\infty)$ and vice versa. These maps are $x \mapsto x$ and $x \mapsto x+1$.

By transitivity of the "same cardinality" property (which is proven by taking compositions of bijections), we are done.

Solution 1.54. We label each sticker so that it has a unique identifier (while retaining its colour). Although not every rearrangement of the stickers is achievable by legal moves, every achievable rearrangement can be considered to be a bijection from the possible locations on the six faces to the set of labelled stickers. If five of the faces are solved, meaning there is a bijection from yellow inputs to yellow outputs, and similarly for four more colours, then that leaves only 9 stickers of the same colour for the final face. Since every achievable rearrangement is a bijection as described, the final face is automatically solved.

Solution 1.58. Let X be a non-empty subset of $\mathbb{Z}$ with a lower bound b. Since $x \geq b$ for all $x \in X$, it holds that $x - b + 1 \geq 1$ and so all elements of $X - b + 1 = \{x - b + 1 : x \in X\}$ are in $\mathbb{Z}_+$. By the assumed weak version of the well-ordering principle, $X - b + 1$ has a least element m. So $x - b + 1 \geq m$ for all $x \in X$. Thus,

$$\forall x \in X, x \geq m + b - 1.$$

All we have to do is prove that $m + b - 1 \in X$. Indeed, since $m \in X - b + 1$, there exists an $x_0 \in X$ such that $m = x_0 - b + 1$. Therefore,

$$m + b - 1 = x_0 \in X.$$

Solution 1.61. The base case $n = 1$ is easy to verify. Suppose

$$\frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \cdots + \frac{1}{n^2} \leq 2 - \frac{1}{n}$$

holds for some positive integer n. Adding $\dfrac{1}{(n+1)^2}$ to both sides yields

$$\frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \cdots + \frac{1}{n^2} + \frac{1}{(n+1)^2} \leq 2 - \frac{1}{n} + \frac{1}{(n+1)^2}.$$

It suffices to prove that

$$2 - \frac{1}{n} + \frac{1}{(n+1)^2} \leq 2 - \frac{1}{n+1}.$$

Working backwards by clearing the denominators and expanding reduces it to $0 \leq 1$, which is true. (It is actually known that the infinite sum is exactly $\dfrac{\pi^2}{6}$)

Solution 1.62. The base case for $n = 1$ may be directly verified. Suppose there exists a positive integer n such that

$$\sqrt{n} \leq \frac{1}{\sqrt{1}} + \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{3}} + \cdots + \frac{1}{\sqrt{n}} \leq 2\sqrt{n}.$$

Adding $\dfrac{1}{\sqrt{n+1}}$ to all three components yields

$$\sqrt{n} + \frac{1}{\sqrt{n+1}} \leq \frac{1}{\sqrt{1}} + \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{3}} + \cdots + \frac{1}{\sqrt{n}} + \frac{1}{\sqrt{n+1}} \leq 2\sqrt{n} + \frac{1}{\sqrt{n+1}}.$$

By transitivity of inequalities, it suffices to weaken the inequalities by proving that

$$\sqrt{n+1} \leq \sqrt{n} + \frac{1}{\sqrt{n+1}},$$
$$2\sqrt{n} + \frac{1}{\sqrt{n+1}} \leq 2\sqrt{n+1}.$$

Working backwards by clearing denominators, cancelling or collecting like terms, and squaring both sides as needed, we find that the former is equivalent to $0 \leq n$ and the latter is equivalent to $0 \leq 1$, both of which are true. Thus, the induction is complete. Strangely, we had to weaken our result in the inductive step by widening the gaps between the parts of the inequality in order to have the effect of bringing the induction dominoes closer together and cause a domino chain reaction.

Solution 1.63. By computing the first few cases, we suspect that the formula is

$$\frac{1}{2!} + \frac{2}{3!} + \frac{3}{4!} + \cdots + \frac{n}{(n+1)!} = 1 - \frac{1}{(n+1)!}.$$

The base case $n = 1$ is easy enough to verify. Assuming that the formula holds for some positive integer n, we add $\dfrac{n+1}{(n+2)!}$ to both sides to get

$$\begin{aligned}\frac{1}{2!} + \frac{2}{3!} + \frac{3}{4!} + \cdots + \frac{n}{(n+1)!} + \frac{n+1}{(n+2)!} &= 1 - \frac{1}{(n+1)!} + \frac{n+1}{(n+2)!} \\ &= 1 - \frac{(n+2) - (n+1)}{(n+2)!} \\ &= 1 - \frac{1}{(n+2)!},\end{aligned}$$

as desired. This completes the induction.

Solution 2.5. The following are statements that apply to all elements a, b, c of the underlying set:

- Commutativity of $\circ$: $\circ(a, b) = \circ(b, a)$
- Associativity of $\circ$: $\circ(a, \circ(b, c)) = \circ(\circ(a, b), c)$
- Left-distributivity of $\star$ over $\circ$: $\star(c, \circ(a, b)) = \circ(\star(c, a), \star(c, b))$
- Right-distributivity of $\star$ over $\circ$: $\star(\circ(a, b), c) = \circ(\star(a, c), \star(b, c))$

This should give the reader a sense of how function notation is unwieldy in the case of binary operations and is discarded as such.

Solution 2.28. We find that

$$-(-a) = (-1)((-1)a) = ((-1)(-1))a = (1 \cdot 1)a = 1 \cdot a = a.$$

We have used Corollary 2.27 to compute $(-1)(-1) = 1 \cdot 1 = 1$. Alternatively, $-a$ is defined as the number such that $a + (-a) = 0$. But then a is the number such that $(-a) + a = 0$, so the negation of $-a$ is a itself.

Solution 2.32. Suppose $f : \mathbb{R} \to \mathbb{R}$ is a function that is both even and odd. Then for all real x,

$$-f(x) = f(-x) = f(x).$$

Thus, the only possibility is that $f(x) = 0$ for all real x, Indeed, this function is both even and odd.

Solution 2.38. We know that $(-1)^{-1} = -1$ because $(-1)(-1) = 1 \cdot 1 = 1$, so

$$\frac{-1}{-1} = (-1) \cdot (-1)^{-1} = (-1)(-1) = 1.$$

Solution 2.40. Since a^{-1} is defined as the number such that $a \cdot a^{-1} = 1$, we find that a is the number such that $a^{-1} \cdot a = 1$. So the inverse of a^{-1} is a.

Solution 2.44. The identities are derived from previous results as follows:

- $\dfrac{-a}{-b} = \dfrac{(-1)a}{(-1)b} = \dfrac{-1}{-1} \cdot \dfrac{a}{b} = \dfrac{a}{b}$
- $\dfrac{-a}{b} = \dfrac{(-1)a}{(1)b} = \dfrac{-1}{1} \cdot \dfrac{a}{b} = (-1) \cdot \dfrac{a}{b} = -\dfrac{a}{b}$ and the second equality follows similarly

Solution 2.47. We have to verify reflexivity, symmetry and transitivity. Before we get into the proof, we will prove a lemma, which states that if $(a, b) \sim (c, 0)$ then $b = 0$, and if $(a, b) \sim (0, d)$ then $a = 0$. If $(a, b) \sim (c, 0)$ then $0 = a \cdot 0 = bc$. Since $(0, 0)$ is not a part of the set, $c \neq 0$ and so $b = 0$. For the second assertion, if $(a, b) \sim (0, d)$ then $ad = b \cdot 0 = 0$. Again since $(0, 0)$ is not a part of the set, $d \neq 0$ and so $a = 0$.

1. Reflexivity: $(a, b) \sim (a, b)$ is true because it is means $ab = ba$, which holds by commutativity of multiplication

2. Symmetry: Suppose $(a, b) \sim (c, d)$. Then $ad = bc$. This is equivalent to $cb = da$, which implies $(c, d) \sim (a, b)$.

3. Transitivity: Suppose $(a, b) \sim (c, d)$ and $(c, d) \sim (e, f)$. Then $ad = bc$ and $cf = de$. We want to show that $af = be$, as that would give us $(a, b) \sim (e, f)$. We look at three cases:

   - Suppose $c \neq 0$ and $d \neq 0$. Multiplying $ad = bc$ and $cf = de$ yields $adcf = bcde$. We can divide both sides by cd to get $af = be$.
   - If $c = 0$ then $a = e = 0$ by the lemma, and so $af = 0 = be$.
   - If $d = 0$ then $b = f = 0$ by the lemma, and so $af = 0 = be$ again.

Therefore, the ratio binary binary relation is an equivalence relation.

Solution 2.48. Neither function is the identity function, so neither function is of order 1. So it suffices to show that $f \circ f$ and $g \circ g$ are both the identity function. Indeed,

$$f(f(x)) = c - (c - x) = c - c + x = x,$$
$$g(g(x)) = \frac{c}{\left(\frac{c}{x}\right)} = c \cdot \frac{x}{c} = x.$$

Solution 2.57. Our key tool will be the difference of squares factorization in the form stated in the problem. For the first one,

$$\frac{1}{\sqrt{2} + \sqrt{3}} = \frac{\sqrt{2} - \sqrt{3}}{(\sqrt{2} + \sqrt{3})(\sqrt{2} - \sqrt{3})} = \frac{\sqrt{2} - \sqrt{3}}{2 - 3} = \sqrt{3} - \sqrt{2}.$$

For the second fraction, we will apply the difference of squares twice:

$$\begin{aligned}\frac{1}{\sqrt{2}-\sqrt{3}+\sqrt{5}} &= \frac{\sqrt{2}+\sqrt{5}+\sqrt{3}}{(\sqrt{2}+\sqrt{5}-\sqrt{3})(\sqrt{2}+\sqrt{5}+\sqrt{3})}\\ &= \frac{\sqrt{2}+\sqrt{3}+\sqrt{5}}{2(2+\sqrt{10})}\\ &= \frac{(\sqrt{2}+\sqrt{3}+\sqrt{5})(2-\sqrt{10})}{2(2+\sqrt{10})(2-\sqrt{10})}\\ &= \frac{(\sqrt{2}+\sqrt{3}+\sqrt{5})(2-\sqrt{10})}{2(4-10)}\\ &= \frac{(\sqrt{2}+\sqrt{3}+\sqrt{5})(\sqrt{10}-2)}{12}.\end{aligned}$$

Solution 3.6. Again, the answer is not $40-10=30$. The list of pages is

$$(11, 12, 13, \ldots, 39).$$

We subtract 10 from each entry to turn it into the list $(1, 2, 3, \ldots, 29)$. So there are 29 pages from page 10 to page 40 exclusive.

Solution 3.7. We begin by finding the first and last multiples on 3 in the given interval. By using long division, we find that $110 = 3 \cdot 36 + 2$ and $1000 = 3 \cdot 333 + 1$. So the list of relevant multiples of 3 is

$$(3 \cdot 37, 3 \cdot 38, 3 \cdot 39, \ldots, 3 \cdot 333).$$

We divide every entry by 3 to produce the list $(37, 38, 39, \ldots, 333)$, and then we subtract 36 from each of these entries to get $(1, 2, 3, \ldots, 297)$. Thus, the answer is 297.

Solution 3.8. The two solutions use transformations of lists as usual.

1. We subtract a from each term and then divide each resulting term by d to get the list $(n, n+1, \ldots, m)$. Then we subtract $n-1$ from each term to get $(1, 2, \ldots, m-n+1)$.
2. We divide each term by br^{n-1} to get the list $(r^1, r^2, \ldots, r^{m-n+1})$. Then we apply the $\log_r x$ function to each term to get the list $(1, 2, \ldots, m-n+1)$.

In both cases, the number of indices is $m-n+1$. There is no need to memorize this result, as knowledge of the proof will allow the reader to easily replicate the process.

Solution 3.9. Conveniently, $(-10)^3 = -1000$ and $10^3 = 1000$, and the next cubes are too large in magnitude in either direction. So the list of relevant cubes of integers, ordered from least to greatest, is

$$((-10)^3, (-9)^3, (-8)^3, \ldots, 8^3, 9^3, 10^3).$$

By taking the cube root of each entry, we get the list $(-10, -9, -8, \ldots, 8, 9, 10)$, which has $10+1+9=10=21$ entries in it. We did the final computation by noticing that there are 10 negative entries, one 0 entry, and 10 positive entries.

Solution 3.12. The idea is to make the counter what we are subtracting from t at each step. That is,

$$\begin{aligned}\sum_{i=s}^{t} a_j &= a_s + a_{s+1} + \cdots + a_{t-1} + a_t \\ &= a_t + a_{t-1} + \cdots + a_{s+1} + a_s \\ &= a_t + a_{t-1} + \cdots + a_{t-(t-s-1)} + a_{t-(t-s)} = \sum_{j=0}^{t-s} a_{t-j}.\end{aligned}$$

This is known as a *reflection*, since we are reflecting the indices from right to left.

Solution 3.20. As in the proof of Theorem 3.19, the terms can be placed in a matrix:

$$\begin{bmatrix} a_{12} & a_{13} & \cdots & a_{1n} \\ & a_{23} & \cdots & a_{2n} \\ & & \ddots & \vdots \\ & & & a_{n-1,n} \end{bmatrix}$$

The left side of the identity finds the sum of the matrix by adding the sums of rows, and the right side is the sums of columns.

Solution 4.4. We know that $-2 \neq 2$, but squaring both sides gives $4 \neq 4$, which is untrue since we know that $4 = 4$. Technically, we applied the function $f(t) = t^2$ to both sides.

Solution 4.8. Multiplying the three equations together yields

$$\begin{aligned}pqr &= \left(x + \frac{1}{y}\right)\left(y + \frac{1}{z}\right)\left(z + \frac{1}{x}\right), \\ &= xyz + x + y + z + \frac{1}{x} + \frac{1}{y} + \frac{1}{z} + \frac{1}{xyz}.\end{aligned}$$

Adding the three equations together yields

$$p + q + r = x + y + z + \frac{1}{x} + \frac{1}{y} + \frac{1}{z}.$$

Therefore, we can subtract the added equation from the multiplied equation to get

$$pqr - (p + q + r) = xyz + \frac{1}{xyz}.$$

Solution 4.9. Let a be a real number that is its own additive inverse. Then

$$a + a = 0.$$

By the distributive property,

$$2a = a \cdot (1 + 1) = 0.$$

Dividing both sides by 2, we get that $a = 0$, which indeed works. So 0 is the only solution. Let b be a real number that is its own multiplicative inverse. Then

$$b^2 = b \cdot b = 1.$$

We will use the difference of squares factorization, which, using the distributive property, says

$$(x - y)(x + y) = x \cdot x + x \cdot y - y \cdot x + y \cdot y = x^2 - y^2.$$

Then $b^2 = 1$ can be rewritten as

$$0 = b^2 - 1^2 = (b - 1)(b + 1).$$

Thus either $b - 1 = 0$ or $b + 1 = 0$ by Theorem 2.36. This means the potential solutions are 1 and -1. They are indeed their own inverses, as we can verify by computation:

$$(-1) \cdot (-1) = 1 = 1 \cdot 1.$$

Solution 4.10. The errors occurs when we divides both sides of

$$(a - b)(a + b) = (a - b)b$$

by $a - b$. Since we started with the assumption that $a = b$, this means dividing by $a - b$ is the same as dividing by 0, which has no definition.

Solution 4.17. The idea is to change one variable at a time. By the compatibility rules (Theorem 4.16), we find that:

$$\begin{aligned} a + c &> b + c > b + d, \\ a + c &\geq b + c > b + d, \\ a + c &\geq b + c \geq b + d, \\ ac &> bc > bd. \end{aligned}$$

For further results, as long as we go step by step and use one compatibility rule at a time, we cannot go wrong.

Solution 4.21. We choose the sequence that takes the reciprocals of the integers:

$$\frac{1}{1}, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots.$$

Suppose $\epsilon > 0$ is fixed. We want a positive integer N such that

$$\frac{1}{N}, \frac{1}{N+1}, \frac{1}{N+2}, \frac{1}{N+3}, \cdots$$

are all less than ϵ. An equivalent condition is

$$\frac{1}{\epsilon} < N, N+1, N+2, \ldots.$$

By the Archimedean property, there exists an integer N such that $\dfrac{1}{\epsilon} < N$. It follows that all subsequent integers are also greater than $\dfrac{1}{\epsilon}$.

Solution 4.23. Any open interval such as $(0, 1)$ works. An upper bound is 1 and a lower bound is 0, but neither lives within the set. Since the upper bound 1 is the supremum and the lower bound 0 is the infimum, there is no maximum or minimum.

Solution 4.30. By multiplying together all of the inequalities

$$a_1 \geq b_1, a_2 \geq b_2, \ldots, a_n \geq b_n,$$

we find that

$$a_1 a_2 \cdots a_n \geq b_1 b_2 \cdots b_n$$

is true.
For the equality criterion, the idea is to use the fact that $\log_{10} x$ is a strictly increasing function with the domain being the positive reals and the range being the reals (in fact, the increasing function $\log_b x$, for any base $b > 1$, will work). This will allow us to reduce this problem to its additive variant, which is one of the parts of Lemma 4.29. If

$$a_1 = b_1, a_2 = b_2, \ldots, a_n = b_n,$$

then we can multiply them together to find that equality holds in our inequality. Conversely, suppose

$$a_1 a_2 \cdots a_n = b_1 b_2 \cdots b_n.$$

By taking logs, this can be rewritten as

$$\log_{10} a_1 + \cdots + \log_{10} a_n \geq \log_{10} b_1 + \cdots + \log_{10} b_n.$$

Since a logarithm to a fixed base $b > 1$ is a strictly increasing function,

$$a_k \geq b_k \implies \log_{10} a_k \geq \log_{10} b_k$$

for each $k \in [n]$. By the additive variant, $\log_{10} a_k = \log_{10} b_k$ for each $k \in [n]$, and so $a_k = b_k$ as well.

Solution 4.35. The inequality can be manipulated using reversible steps as follows:

$$\begin{aligned} \frac{x^2 + y^2}{2} &\geq xy \\ x^2 + y^2 &\geq 2xy \\ x^2 - 2xy + y^2 &\geq 0 \\ (x - y)^2 &\geq 0, \end{aligned}$$

which is true by the trivial inequality. Equality holds if and only if $x = y$.

Solution 4.39. We need to prove connexity, antisymmetry, and transitivity of the $\leq$ relation defined.

1. Connexity: We wish to show that $a \leq b$ or $b \leq a$. Logically, it is equivalent to show that if $a \leq b$ is false, then $b \leq a$. Supposing $a \not\leq b$, the trichotomy law implies that the only remaining option is $b < a$. This implies $b \leq a$.

2. Antisymmetry: Suppose $a \leq b$ and $b \leq a$. By definition, this means that $a < b$ or $a = b$, and $b < a$ or $b = a$. Suppose, for contradiction, that $a \neq b$. Then we are forced into saying that $a < b$ and $b < a$ at the same time. This is impossible by the asymmetry of a strict total order. Thus, $a = b$.

3. Transitivity: Suppose $a \leq b$ and $b \leq c$. By definition, this means that $a < b$ or $a = b$, and $b < c$ or $b = c$. This produces four cases:

   - If $a < b$ and $b < c$, then $a < c$.
   - If $a < b$ and $b = c$, then $a < c$.
   - If $a = b$ and $b < c$, then $a < c$.
   - If $a = b$ and $b = c$, then $a = c$.

   In all four cases, it is true that $a \leq c$.

Solution 4.43. Since we are using shortlex order, we can group the strings according to their length:

$$0, 1,$$
$$00, 01, 10, 11,$$
$$000, 001, 010, 011, 100, 101, 110, 111.$$

Note that, among strings with the same number of symbols, we have of course followed lexicographical order. We know that we have found all the strings because a combinatorial multiplication argument implies that there should be 2^n strings of length n, which matches our results for $n = 1, 2, 3$.

Solution 5.6. It is possible to do this by induction, but we will simply square both sides as before. The reason is that we want to display a nice technique for deriving the equality condition when there are numerous terms. Squaring both sides is reversible since both sides are non-negative, and doing it yields

$$|x_1| + |x_2| + \cdots + |x_n| \geq |x_1 + x_2 + \cdots + x_n|$$
$$\sum_{k=1}^{n} |x_k|^2 + \sum_{1 \leq i < j \leq n} 2|x_i| \cdot |x_j| \geq \sum_{k=1}^{n} x_k^2 + \sum_{1 \leq i < j \leq n} 2x_i x_j$$
$$\sum_{1 \leq i < j \leq n} (|x_i x_j| - x_i x_j) \geq 0.$$

Now we will use Lemma 4.29. Notice that each term $(|x_i x_j| - x_i x_j)$ is non-negative. In order for the sum of all such terms to equal 0, each term must be 0. The reason is that if a single such term is positive instead of 0, the whole sum will be positive instead of 0. So we must have

$$|x_i x_j| = x_i x_j$$

for each pair of indices $1 \leq i < j \leq n$. This means x_i and x_j are both non-negative or both non-positive. Applying this rule to the $n-1$ pairs of numbers $(x_1, x_2), (x_2, x_3), \ldots, (x_{n-1}, x_n)$, we find that all of the x_k are non-negative or all of them are non-positive.

Solution 5.10. This identity can be proven by casework on the parity of n. We leave the computations for even $n = 2m$ and odd $n = 2m + 1$ to the reader.

Solution 5.14. Let $m = \lfloor x \rfloor$ and $n = \lceil -x \rceil$. We know that

$$\begin{aligned} x - 1 < m \leq x, \\ -x \leq n < -x + 1 \iff x - 1 < -n \leq x. \end{aligned}$$

This means m and $-n$ are both integers in the interval $(x - 1, x]$ in which we know there exists exactly one integer. Thus, $m = -n$.

Solution 5.17. Let $p = \left\lfloor \dfrac{m}{n} \right\rfloor$. This equation is equivalent to

$$\begin{aligned} p \leq \frac{m}{n} < p + 1 &\iff pn \leq m < (p+1)n \\ &\iff pn \leq m \leq (p+1)n - 1, \end{aligned}$$

where we have used the fact that $n > 0$ to clear the denominator. Similarly, let $q = \left\lceil \dfrac{m}{n} \right\rceil$. Then this equation is equivalent to

$$\begin{aligned} q - 1 < \frac{m}{n} \leq q &\iff (q-1)n < m \leq qn \\ &\iff (q-1)n + 1 \leq m \leq qn. \end{aligned}$$

Finally, note that we can convert the former kind of inequalities into the latter type and vice versa as follows:

$$\begin{aligned} pn \leq m \leq (p+1)n - 1 &\iff (p-1)n + 1 \leq m - n + 1 \leq pm, \\ (q-1)n + 1 \leq m \leq qn &\iff qn \leq m + n - 1 \leq (q+1)n - 1. \end{aligned}$$

This completes the proof because all of our work involved biconditional statements.

Solution 6.5. This is an arithmetic series with initial term 1 and last term $2n - 1$ with a total of n terms, so the answer is

$$\left(\frac{1 + (2n-1)}{2} \right) \cdot n = n^2.$$

Remarkably, the sum of the first n odd positive integers is the n^{th} positive square.

Solution 6.8. By expanding, we notice that the expression

$$\underbrace{(\ldots((pr+p)r+p)r\ldots+p)r+p}_{\text{number of multiplications by } r \text{ is } n-1}$$

is equal to the geometric series

$$pr^{n-1}+pr^{n-2}+\cdots+pr^2+pr+p.$$

This pattern may be proven by induction. Thus, the expression evaluates to

$$p\cdot\frac{r^n-1}{r-1}.$$

The original expression is the natural starting point of the so-called future value formula for annuities in finance, except r is replaced by $r=s+1$ so that the formula can be written as

$$p\cdot\frac{(s+1)^n-1}{s}.$$

Solution 6.14. Using the key fact $(n+1)!=(n+1)\cdot n!$, we find that

$$\begin{aligned}
&1\cdot 1!+2\cdot 2!+3\cdot 3!+\cdots+n\cdot n!\\
&=(2-1)\cdot 1!+(3-1)\cdot 2!+(4-1)\cdot 3!+\cdots+[(n+1)-1]\cdot n!\\
&=(2!-1!)+(3!-2!)+(4!-3!)+\cdots+[(n+1)!-n!]\\
&=(n+1)!-1,
\end{aligned}$$

where we had telescoping in the final step. We will ask the reader for a combinatorial interpretation of this identity in a problem in Volume 2.

Solution 6.17. Each term is of the form

$$\frac{1}{k(k+1)(k+2)}=\frac{1}{2}\left(\frac{1}{k}-\frac{2}{k+1}+\frac{1}{k+2}\right),$$

which we can find through partial fraction decomposition. There is a way to telescope the series by stacking the terms on top of each other and noticing that terms on diagonals cancel out because

$$\frac{1}{k}-\frac{2}{k}+\frac{1}{k}=0.$$

Drawing it out and filling in the details are left as an exercise to the reader. We will instead show a cleverer method that uses an incomplete partial fraction decomposition:

$$\frac{1}{k(k+1)(k+2)}=\frac{1}{2}\left(\frac{1}{k(k+1)}-\frac{1}{(k+1)(k+2)}\right)=f(k)-f(k+1),$$

where $f(k) = \dfrac{1}{2} \cdot \dfrac{1}{k(k+1)}$. Then the n^{th} partial sum is

$$\begin{aligned}\sum_{k=1}^{n} \frac{1}{k(k+1)(k+2)} &= \sum_{k=1}^{n} (f(k) - f(k+1)) \\ &= f(1) - f(n+1) \\ &= \frac{1}{2} \cdot \left(\frac{1}{1 \cdot 2} - \frac{1}{n(n+1)} \right) \\ &= \frac{1}{4} - \frac{1}{2(n+1)(n+2)},\end{aligned}$$

which goes to $\dfrac{1}{4}$ as n goes to infinity because the second term goes to 0.

Solution 6.19. It is easily seen that the base case $n = 1$ holds. Now suppose that

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

for some $n \geq 1$. Then adding $(n+1)^3$ to both sides yields

$$\begin{aligned}1^3 + 2^3 + \cdots + n^3 + (n+1)^3 &= \frac{n^2(n+1)^2}{4} + (n+1)^3 = \frac{n^2(n+1)^2 + 4(n+1)^3}{4} \\ &= \frac{(n+1)^2[n^2 + 4(n+1)]}{4} = \frac{(n+1)^2(n+2)^2}{4} \\ &= \left[\frac{(n+1)(n+2)}{2} \right]^2,\end{aligned}$$

as desired.

Solution 6.21. Using the difference of cubes and sum of cubes formulas

$$\begin{aligned}k^3 - 1 &= (k-1)(k^2 + k + 1), \\ k^3 + 1 &= (k+1)(k^2 - k + 1),\end{aligned}$$

the partial product up to index $n \geq 2$ is equal to

$$\prod_{k=2}^{n} \frac{k^3 - 1}{k^3 + 1} = \prod_{k=2}^{n} \frac{(k-1)(k^2 + k + 1)}{(k+1)(k^2 - k + 1)} = \left(\prod_{k=2}^{n} \frac{k-1}{k+1} \right) \left(\prod_{k=2}^{n} \frac{k^2 + k + 1}{k^2 - k + 1} \right).$$

Now we can evaluate each product separately and then bring them back together. The first product easily telescopes as

$$\prod_{k=2}^{n} \frac{k-1}{k+1} = \frac{1}{3} \cdot \frac{2}{4} \cdot \frac{3}{5} \cdots \frac{n-1}{n+1} = \frac{1 \cdot 2}{n(n+1)}.$$

The second product also telescopes, but due to the unexpected identity

$$k^2 + k + 1 = (k+1)^2 - (k+1) + 1.$$

Letting $f(k) = k^2 - k + 1$, the second product is

$$\prod_{k=2}^{n} \frac{k^2+k+1}{k^2-k+1} = \prod_{k=2}^{n} \frac{f(k+1)}{f(k)} = \frac{f(3)}{f(2)} \cdot \frac{f(4)}{f(3)} \cdots \frac{f(n+1)}{f(n)} = \frac{f(n+1)}{f(2)}.$$

Therefore, the infinite product is

$$\begin{aligned}
\prod_{k=2}^{\infty} \frac{k^3-1}{k^3+1} &= \lim_{n\to\infty} \prod_{k=2}^{n} \frac{k^3-1}{k^3+1} \\
&= \lim_{n\to\infty} \left[\frac{1 \cdot 2}{n(n+1)} \cdot \frac{f(n+1)}{f(2)} \right] \\
&= \lim_{n\to\infty} \frac{2((n+1)^2 - (n+1) + 1)}{n(n+1)(2^2 - 2 + 1)} \\
&= \frac{2}{3} \cdot \lim_{n\to\infty} \frac{n^2+n+1}{n^2+n} \\
&= \frac{2}{3} \cdot \lim_{n\to\infty} \left(1 + \frac{1}{n^2+n} \right).
\end{aligned}$$

As n goes to infinity, $\dfrac{1}{n^2+n}$ goes to 0, so the answer is $\dfrac{2}{3}$.

Solution 7.9. We will use the fact that $\tan : \left(-\frac{\pi}{2}, \frac{\pi}{2}\right) \to \mathbb{R}$ is a bijection. It suffices to find a bijection between every pair of open intervals, because the composition of bijections is a bijection. Let (a, b) and (c, d) be open intervals. The idea is to shift (a, b) over to $(0, b-a)$, scale that to $(0, d-c)$ and then shift that over to (c, d). Altogether, this produces the function

$$f(x) = (x - a) \cdot \frac{d-c}{b-a} + c.$$

This is is a linear function, so it is bijective on $\mathbb{R}$. We can verify that

$$a < x < b \iff c < f(x) < d,$$

so restricting f to the domain (a, b) makes the range (c, d). This completes the proof because restrictions of bijections are also bijections onto the new range.

Solution 7.18. By the addition identities and double angle identities for sine and cosine,

$$\begin{aligned}
\sin(3\theta) &= \sin(\theta + 2\theta) \\
&= \sin\theta\cos(2\theta) + \cos\theta\sin(2\theta) \\
&= \sin\theta(1 - 2\sin^2\theta) + \cos\theta(2\sin\theta\cos\theta) \\
&= \sin\theta - 2\sin^3\theta + 2\sin\theta\cos^2\theta \\
&= \sin\theta - 2\sin^3\theta + 2\sin\theta(1 - \sin^2\theta) \\
&= \sin\theta - 2\sin^3\theta + 2\sin\theta - \sin^3\theta \\
&= 3\sin\theta - 4\sin^3\theta
\end{aligned}$$

and similarly,

$$\begin{aligned}\cos(\theta) &= \cos(\theta + 2\theta)\\ &= \cos\theta\cos(2\theta) - \sin\theta\sin(2\theta)\\ &= \cos\theta(2\cos^2\theta - 1) - \sin\theta(2\sin\theta\cos\theta)\\ &= 2\cos^3\theta - \cos\theta - 2\sin^2\theta\cos\theta\\ &= 2\cos^3\theta - \cos\theta - 2(1-\cos^2\theta)\cos\theta\\ &= 2\cos^3\theta - \cos\theta - 2\cos\theta + 2\cos^3\theta\\ &= 4\cos^3\theta - 3\cos\theta.\end{aligned}$$

As a result, using $\sec^2\theta = 1 + \tan^2\theta$,

$$\begin{aligned}\tan(3\theta) &= \frac{\sin(3\theta)}{\cos(3\theta)}\\ &= \frac{3\sin\theta - 4\sin^3\theta}{4\cos^3\theta - 3\cos\theta}\\ &= \frac{\frac{3\sin\theta}{\cos^3\theta} - 4\frac{\sin^3\theta}{\cos^3\theta}}{4 - \frac{3}{\cos^2\theta}}\\ &= \frac{3\tan\theta\sec^2\theta - 4\tan^3\theta}{4 - 3\sec^2\theta}\\ &= \frac{3\tan\theta(1+\tan^2\theta) - 4\tan^3\theta}{4 - 3(1+\tan^2\theta)}\\ &= \frac{3\tan\theta - \tan^3\theta}{1 - 3\tan^2\theta}.\end{aligned}$$

Solution 8.9. The first few powers of i are: $i^1 = i, i^2 = -1, i^3 = -i, i^4 = 1$, and it can be shown by induction that this pattern cycles. We can also perform an induction on non-positive integer exponents that descend towards $-\infty$ to prove that the periodic pattern extends in the other direction: $i^0 = 1, i^{-1} = -i, i^{-2} = -1, i^{-3} = i$, and so on.

Solution 8.15. By the formula for a complex geometric series (Theorem 8.14), we can evaluate this series as

$$\begin{aligned}z^\alpha + z^{\alpha-\beta} + z^{\alpha-2\beta} + \cdots + z^\gamma &= z^\alpha \cdot \left(1 + \frac{1}{z^\beta} + \frac{1}{z^{2\beta}} + \cdots + \frac{1}{z^{q\beta}}\right)\\ &= z^\alpha \cdot \frac{1 - \left(\frac{1}{z^\beta}\right)^{q+1}}{1 - \frac{1}{z^\beta}} = z^\alpha \cdot \frac{z^{q\beta}\cdot z^\beta - 1}{z^{q\beta}(z^\beta - 1)}\\ &= z^\alpha \cdot \frac{z^{\alpha-\gamma}\cdot z^\beta - 1}{z^{\alpha-\gamma}(z^\beta - 1)} = \frac{z^{\alpha+\beta} - z^\gamma}{z^\beta - 1}.\end{aligned}$$

Solution 8.23. The trigonometric form of a complex number tells us that

$$\begin{aligned}e^{i\theta} &= \cos\theta + i\sin\theta,\\ e^{i(-\theta)} &= \cos(-\theta) + i\sin(-\theta)\\ &= \cos\theta - i\sin\theta.\end{aligned}$$

From there, it is a matter of using the two equations to isolate $\cos\theta$ and $\sin\theta$ like a system of equations.

Solution 8.28. The complex number that we seek is

$$e^{i\frac{\pi}{2}} = \cos\frac{\pi}{2} + i\sin\frac{\pi}{2}0 + i\cdot 1 = i.$$

The reason this works is that

$$\arg(iz) \equiv \arg i + \arg z = \frac{\pi}{2} + \arg z.$$

Solution 8.29. Let $z = re^{i\theta}$. Then the proofs are straightforward using phase shift and reflection identities.

1. $\arg(-z) = \arg(-\cos\theta - i\sin\theta) = \arg(\cos(\pi+\theta) + i\sin(\pi+\theta)) \equiv \pi + \theta \equiv \pi + \arg z$
2. $\arg(\overline{z}) = \arg(\cos\theta - i\sin\theta) = \arg(\cos(-\theta) + i\sin(-\theta)) \equiv -\theta \equiv -\arg z$

Solution 9.4. Let $f(x) = ax^2 + bx + c$ be a quadratic function.

1. Suppose $c = 0$. Then
$$f(x) = ax^2 + bx = x(ax+b).$$
This is equal to 0 if and only if one of x or $ax+b$ is equal to 0. Thus, the roots are 0 and $-\dfrac{b}{a}$.
2. Suppose $b = 0$. Then
$$f(x) = ax^2 + c.$$
Setting this equal to 0, we can isolate $x = \pm\sqrt{-\dfrac{c}{a}}$.

Solution 9.8. Let the coefficients a, b, c be rational. For one direction, suppose

$$\frac{-b+\sqrt{b^2-4ac}}{2a} = r_1,$$
$$\frac{-b-\sqrt{b^2-4ac}}{2a} = r_2$$

for rational numbers r_1 and r_2. Then

$$D = b^2 - 4ac = (a(r_1 - r_2))^2,$$

which is the square of a rational number. Conversely, suppose $D = b^2 - 4ac$ is the square of a rational number. Since we assumed that a, b, c are rational, $\dfrac{-b \pm \sqrt{b^2-4ac}}{2a}$ are also rational for either choice of the $\pm$ sign.

Solution 9.14. It is easy to verify that the identity holds by squaring both sides, which is a reversible step since both sides are non-negative. We leave this as an exercise to the reader. Instead, we will show how we came to the decomposition in the first place.
Suppose there exist non-negative real numbers x and y such that

$$\sqrt{a \pm b\sqrt{c}} = \sqrt{x} \pm \sqrt{y},$$

where the two $\pm$ symbols represent the same sign. Squaring yields

$$a \pm b\sqrt{c} = (x + y) \pm 2\sqrt{xy}.$$

Inspired by how it is possible to equate real parts and equal imaginary parts when dealing with complex numbers, we make a non-rigorous guess that

$$\begin{aligned} a &= x + y, \\ b\sqrt{c} &= 2\sqrt{xy} \end{aligned}$$

will work. Since both sides are positive in the latter equation, the equation is equivalent to its squared counterpart $b^2c = 4xy$. So $x + y = a$ and $xy = \dfrac{b^2c}{4}$. By Vieta's formulas, x and y must be roots of the quadratic $f(z) = z^2 - az + \dfrac{b^2c}{4}$. By the quadratic formula, these roots are

$$x, y = \frac{a \pm \sqrt{a^2 - b^2c}}{2}.$$

As it turns out, these x and y indeed work out to form a decomposition $\sqrt{x} \pm \sqrt{y}$ that can be algebraically verified, as we mentioned at the beginning.

Solution 10.7. Non-constant linear polynomials have degree 1. According to Theorem 10.6, $\deg(f \circ g) = (\deg f) \cdot (\deg g)$ if f and g are non-constant polynomials. The result follows, since multiplication by 1 leaves a number unchanged.

Solution 10.9. Let f be a polynomial with positive real coefficients. Suppose for contradiction that r is a non-negative real root of f. Then $f(r)$ computes to a positive real number, so r cannot be a root of f. Thus, if f has a real root, it must be negative.

Solution 10.10. Let $f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_2x^2 + a_1x + a_0$. If $f(0) = 0$ then all but the constant term disappears and so $a_0 = 0$. Conversely, if $a_0 = 0$ then $f(0) = 0$ because all the non-constant terms have a factor of x.

Solution 10.18. Inspired by how we showed that the non-real complex roots of polynomials with real coefficients come in complex conjugate pairs in Theorem 10.16, we proceed as follows. Let f be a polynomial with rational coefficients, let c be a positive rational number that is not the square of a rational number, and let $r = a + b\sqrt{c}$ be a root of f, where a and b are rational numbers. Now let

$$\begin{aligned} r_1 &= a_1 + b_1\sqrt{c}, \\ r_2 &= a_2 + b_2\sqrt{c}, \end{aligned}$$

where a_1, a_2 and b_1, b_2 are rational numbers. We leave it to the reader to show that

$$\underline{r_1 \cdot r_2} = \underline{r_1} \cdot \underline{r_2},$$
$$\underline{r_1 + r_2} = \underline{r_1} + \underline{r_2}.$$

Moreover, for any rational number t, we can show that $\underline{tr} = t\underline{r}$. Now the proof is exactly the same as the proof for showing that the non-real complex roots of polynomials with real coefficients come in complex conjugate pairs.

Solution 10.25. We perform long division of polynomials as follows:

$$\begin{array}{r|llllllllll}
 & \frac{3}{2}x^2 & - & \frac{3}{4}x & - & \frac{21}{8} & & & & \\
\hline
2x^2+x+4 & 3x^4 & + & 0x^3 & + & 0x^2 & + & 10x & + & 5 \\
 & 3x^4 & + & \frac{3}{2}x^3 & + & 6x^2 & & & & \\
\hline
 & & & -\frac{3}{2}x^3 & - & -6x^2 & + & 10x & & \\
 & & & -\frac{3}{2}x^3 & - & \frac{3}{4}x^2 & - & 3x & & \\
\hline
 & & & & & -\frac{21}{4}x^2 & + & 13x & + & 5 \\
 & & & & & -\frac{21}{4}x^2 & - & \frac{21}{8}x & - & \frac{21}{2} \\
\hline
 & & & & & & & \frac{125}{8}x & + & \frac{31}{2}
\end{array}$$

Therefore the quotient is $\frac{3}{2}x^2 - \frac{3}{4}x - \frac{21}{8}$ and the remainder is $\frac{125}{8}x - \frac{31}{2}$.

Solution 10.31. Suppose, by contrapositive, that neither of p or q is the 0 polynomial. If pq has a root z, then $(pq)(z) = 0$. Then $p(z)q(z) = 0$ and z is a root of at least one of p or q. By Theorem 10.30, p has at most $\deg p$ distinct roots, and q has at most most $\deg q$ distinct roots. This means there are at most $\deg p + \deg q$ values of z, which means pq takes on the value 0 for only finitely many complex numbers. By Theorem 10.19, pq is not 0 everywhere. By the contrapositive of what we just proved, if pq is 0 everywhere, then at least one of p or q is the zero polynomial.

Solution 10.42. Let $ax^3 + bx^2 + cx + d$ be a cubic equation and let its roots be r_1, r_2, r_3. Then expanding the factored form of the equation yields

$$\begin{aligned} ax^3 + bx^2 + cx + d &= a(x - r_1)(x - r_2)(x - r_3) \\ &= ax^3 - a(r_1 + r_2 + r_3)x^2 + a(r_1r_2 + r_2r_3 + r_1r_3)x - ar_1r_2r_3. \end{aligned}$$

Comparing non-leading coefficients yields the formulas:

$$\begin{aligned} r_1 + r_2 + r_3 &= -\frac{b}{a} \\ r_1r_2 + r_2r_3 + r_1r_3 &= \frac{c}{a} \\ r_1r_2r_3 &= -\frac{d}{a}. \end{aligned}$$

Solution 10.52. We can rewrite the expression

$$\begin{aligned} r^2 + s^2 + t^2 &= (r + s + t)^2 - 2(rs + st + rt) \\ &= \sigma_1^2 - 2\sigma_2. \end{aligned}$$

But we know that $\sigma_k = (-1)^k \frac{a_{n-k}}{a_n}$, so $\sigma_1 = -\frac{a_2}{a_3}$ and $\sigma_2 = \frac{a_1}{a_3}$. The answer is

$$\sigma_1^2 - 2\sigma_2 = \frac{a_2^2}{a_3^2} - \frac{2a_1}{a_3} = \frac{a_2^2 - 2a_1a_3}{a_3^2} = -\frac{11}{4}.$$

Solution 10.56. Let $f(x, y, z) = (x+y+z)^3 - x^3 - y^3 - z^3$. Note that each of the scenarios $x = -y, y = -z, z = -x$ individually leads to $f(x, y, z) = 0$. This means each of the linear polynomials $x + y, y + z, z + x$ is a factor of f. So we can guess that there is a polynomial g such that

$$f(x, y, z) = (x + y)(y + z)(z + x)g(x, y, z).$$

But $\deg f = 3$, so g must be a constant c. Now we can substitute constants into f such as $(x, y, z) = (1, 1, 1)$ to get

$$c = \frac{(x + y + z)^3 - x^3 - y^3 - z^3}{(x + y)(y + z)(z + x)} = \frac{3^3 - 1^3 - 1^3 - 1^3}{2 \cdot 2 \cdot 2} = 3.$$

Indeed, expanding allows us to check that

$$(x + y + z)^3 - x^3 - y^3 - z^3 = 3(x + y)(y + z)(z + x).$$

Solution 10.58. We use the difference of squares, and difference and sum of cubes factorizations to get

$$\begin{aligned} x^6 - y^6 &= (x^3)^2 - (y^3)^2 \\ &= (x^3 - y^3)(x^3 + y^3) \\ &= (x - y)(x^2 + xy + y^2)(x + y)(x^2 - xy + y^2), \end{aligned}$$

where the quadratic factors are irreducible as long as we are seeking real coefficients because splitting them into linear factors using the quadratic formula yields complex coefficients.

Solution 10.59. Let ω be a non-trivial third root of unity. Then $\omega^3 = 1$, which tells us that

$$0 = \omega^3 - 1 = (\omega - 1)(\omega^2 + \omega + 1).$$

Since $\omega \neq 1$, we find that

$$\omega^2 + \omega + 1,$$

so a solution is $x^2 + x + 1$.
In general, this technique can be used to find a polynomial with degree less than n that is satisfied by an n^{th} root of unity ω about which we have some additional information. It is always true that $\omega^n - 1$, but sometimes we are capable of producing a sharper polynomial (i.e. a factor of $\omega^n - 1$) and that might be necessary. For related material, see the section on cyclotomic polynomials in Volume 3.

Solution 10.60. Let $P(x) = \sum_{k=0}^{n} a_k x^k$. Then

$$P(a) - P(b) = \sum_{k=0}^{n} a_k (a^k - b^k).$$

By the difference of powers factorizations, $a - b$ divides $a_k - b_k$ for any non-negative integer k. Thus, $a - b$ divides $P(a) - P(b)$. In the case that b is a root of P, it means that $P(b) = 0$, so the result asserts that $a - b$ divides $P(a)$.

Solution 10.63. The idea is to replace each instance of c with $-(a + b)$ and to use the factorizations of $(a+b)^n - a^n - b^n$ for $n = 3, 5, 7$ (Theorem 10.62). For the first identity, this yields

$$\begin{aligned}
\frac{f(5)}{f(3)} &= \frac{3}{5} \cdot \frac{a^5 + b^5 - (a+b)^5}{a^3 + b^3 - (a+b)^3} \\
&= \frac{3}{5} \cdot \frac{5ab(a+b)(a^2 + ab + b^2)}{3ab(a+b)} \\
&= a^2 + ab + b^2 \\
&= \frac{a^2 + b^2 + (a+b)^2}{2} = f(2).
\end{aligned}$$

For the second identity,

$$\begin{aligned}
\frac{f(7)}{f(5)} &= \frac{5}{7} \cdot \frac{a^7 + b^7 - (a+b)^7}{a^5 + b^5 - (a+b)^5} \\
&= \frac{5}{7} \cdot \frac{7ab(a+b)(a^2 + ab + b^2)^2}{5ab(a+b)(a^2 + ab + b^2)} \\
&= a^2 + ab + b^2 \\
&= \frac{a^2 + b^2 + (a+b)^2}{2} = f(2).
\end{aligned}$$

Technically, we should not have divided by any quantities or cancelled any common factors between a numerator and denominator without being sure that the expressions do not assume a null value. However, the proof is cleaner this way and it is clear that a proof without division or cancellation could be produced in the same way by using the same factorizations.

Solution 11.1. By the trivial inequality,

$$a^2 + b^2 \geq 2ab \iff (a-b)^2 \geq 0.$$

Equality holds if and only of $a = b$. Adding the analogous inequalities $b^2 + c^2 \geq 2bc$ and $c^2 + a^2 \geq 2ca$ to it and dividing by 2 yields

$$a^2 + b^2 + c^2 \geq ab + bc + ca.$$

Equality holds if and only if $a = b = c$.

Solution 11.4. Using $S(n)$ and the stated substitution,

$$\begin{aligned}
\frac{1}{n}\left(a_1 + a_2 + \cdots + a_{n-1} + \sqrt[n-1]{a_1 a_2 \cdots a_{n-1}}\right) &\geq \sqrt[n]{a_1 a_2 \cdots a_{n-1} \cdot \sqrt[n-1]{a_1 a_2 \cdots a_{n-1}}} \\
&= \left((a_1 a_2 \cdots a_{n-1})^{1+\frac{1}{n-1}}\right)^{\frac{1}{n}} \\
&= (a_1 a_2 \cdots a_{n-1})^{\frac{1}{n-1}} \\
a_1 + a_2 + \cdots + a_{n-1} + \sqrt[n-1]{a_1 a_2 \cdots a_{n-1}} &= n(a_1 a_2 \cdots a_{n-1})^{\frac{1}{n-1}} \\
a_1 + a_2 + \cdots + a_{n-1} &= (n-1)(a_1 a_2 \cdots a_{n-1})^{\frac{1}{n-1}} \\
\frac{a_1 + a_2 + \cdots + a_{n-1}}{n-1} &= \sqrt[n-1]{a_1 a_2 \cdots a_{n-1}}.
\end{aligned}$$

As in the proof the theorem, the equality condition for $S(n-1)$ follows from that of $S(n)$.

Solution 11.6. The GM-HM inequality is equivalent to

$$\frac{\frac{1}{a_1} + \frac{1}{a_2} + \cdots + \frac{1}{a_n}}{n} \geq \sqrt[n]{\frac{1}{a_1} \cdot \frac{1}{a_2} \cdots \frac{1}{a_n}},$$

which follows from the AM-GM inequality. Using the equality condition for the AM-GM inequality, equality holds if and only if all the $\dfrac{1}{a_i}$ are equal, which is equivalent to all the a_i being equal. As with most proofs of multivariable inequalities, we have worked backwards here. This is fine as long as the steps are reversible. Working backwards is an indispensable technique that only the most rigid pedant would avoid.

Solution 11.7. Both solutions will use the AM-GM inequality.

1. Let a and b be the dimensions of a rectangle with perimeter p and let its area be A. By the AM-GM inequality for two variables,

$$A = ab \leq \left(\frac{a+b}{2}\right)^2 = \frac{p^2}{16}.$$

This upper bound for the area is reached if and only if $a = b = \dfrac{p}{4}$.

2. Let a, b, c be the dimensions of a box with surface area S and let its volume be V. By the AM-GM inequality for three variables,

$$\frac{S}{6} = \frac{ab + bc + ca}{3} \geq (abc)^{\frac{2}{3}} = V^{\frac{2}{3}}.$$

So $V \leq \left(\frac{S}{6}\right)^{\frac{3}{2}}$. This upper bound for the volume is reached if and only if $ab = bc = ca = \frac{S}{6}$, which is true if and only if $a = b = c = \sqrt{\frac{S}{6}}$.

Solution 11.8. The trick is to use the factorization (see the list in Theorem 10.62)

$$\begin{aligned} x^3 + y^3 + z^3 - 3xyz &= (x + y + z)(x^2 + y^2 + z^2 - xy - yz - zx) \\ &= \frac{1}{2}(x + y + z)((x - y)^2 + (y - z)^2 + (z - x)^2). \end{aligned}$$

By the trivial inequality, this is non-negative if and only if $x + y + z \geq 0$. Equality holds if and only if $x = y = z$ or $x + y + z = 0$.

Solution 11.11. By the Cauchy-Schwarz inequality,

$$\begin{aligned} (a_1^2 + a_2^2 + \cdots + a_n^2)^2 &= (a_1^2 + a_2^2 + \cdots + a_n^2)(a_{\sigma(1)}^2 + a_{\sigma(2)}^2 + \cdots + a_{\sigma(n)}^2) \\ &\geq (a_1 a_{\sigma(1)} + a_2 a_{\sigma(2)} + \cdots + a_n a_{\sigma(n)})^2. \end{aligned}$$

By taking the square root of both sides, we get

$$a_1^2 + a_2^2 + \cdots + a_n^2 \geq |a_1 a_{\sigma(1)} + a_2 a_{\sigma(2)} + \cdots + a_n a_{\sigma(n)}|.$$

Solution 11.12. A technique to keep in mind is that all of the numbers in one of the two n-tuples in Cauchy-Schwarz can be filled with (possibly unequal) constants. In this case, we can fill one of the n-tuples with all 1's to get

$$\begin{aligned} n(a_1^2 + a_2^2 + \cdots + a_n^2) &= \underbrace{(1^2 + 1^2 + \cdots + 1^2)}_{n \text{ copies of } 1^2}(a_1^2 + a_2^2 + \cdots + a_n^2) \\ &\geq (a_1 + a_2 + \cdots + a_n)^2. \end{aligned}$$

Since $1 \neq 0$, equality holds if and only if there exists a real number r such that $1 \cdot r = a_i$ for all i, which is equivalent to all the a_i being equal. This inequality leads to

$$\begin{aligned} \sqrt{\frac{a_1^2 + a_2^2 + \cdots + a_n^2}{n}} &\geq \left|\frac{a_1 + a_2 + \cdots + a_n}{n}\right| \\ &\geq \frac{a_1 + a_2 + \cdots + a_n}{n}. \end{aligned}$$

Equality holds in the left inequality if and only if all of the a_i are equal and equality holds in the right inequality if and only if $a_1 + a_2 + \cdots + a_n$ is non-negative. Thus, equality holds in the RMS-AM inequality if and only if all of the a_i are non-negative and equal.

Solution 11.13. By Engel's form of the Cauchy-Schwarz inequality (Corollary 11.10),

$$\begin{aligned}\frac{a}{b+c} + \frac{b}{c+a} + \frac{c}{a+b} &= \frac{a^2}{ab+ca} + \frac{b^2}{bc+ab} + \frac{c^2}{ca+bc} \\ &\geq \frac{(a+b+c)^2}{(ab+ca)+(bc+ab)+(ca+bc)} \\ &= \frac{a^2+b^2+c^2}{2(ab+bc+ca)} + 1.\end{aligned}$$

By Problem 11.1, we know that

$$a^2 + b^2 + c^2 \geq ab + bc + ca,$$

so it holds that

$$\frac{a^2+b^2+c^2}{2(ab+bc+ca)} + 1 \geq \frac{3}{2}.$$

We know that equality holds in the last step if and only if $a = b = c$, so this is a necessary criterion for equality in Nesbitt's inequality. Substituting $a = b = c$ into Nesbitt's inequality shows that this condition is also sufficient.

Solution 11.16. By substituting the stated formulas into Euler's inequality and clearing the denominators, we find that we want to prove that

$$abc \geq (-a+b+c)(a-b+c)(a+b-c).$$

This is another way of writing Schur's inequality for $t = 1$. We leave it to the reader to watch the solution unfurl by expanding the right side of this equality and expanding Schur for $t = 1$. Since a, b, c are side lengths of a triangle, none of them can be 0, so the only case of equality is $a = b = c$.

Solution 11.22. We will use Jensen's inequality. For the interval $(0, \infty)$ on which the following functions in the variable x are defined, we observe the convexity of x^t with a fixed real $t > 1$, the concavity of x^t with $0 < t < 1$, and the convexity of x^t for $t < 0$. Now we break the argument into three cases, according to where r and s lie relative to 0:

1. Suppose $r > s > 0$. Then $\dfrac{r}{s} > 1$, so $f(x) = x^{\frac{r}{s}}$ is convex. By Jensen,

$$(\lambda_1 x_1^s + \lambda_2 x_2^s + \cdots \lambda_n x_n^s)^{\frac{r}{s}} \leq \lambda_1 x_1^r + \lambda_2 x_2^r + \cdots \lambda_n x_n^r.$$

   Taking both sides to the power of $\dfrac{1}{r}$ preserves the direction of the inequality because $g(x) = x^{\frac{1}{r}}$ is an increasing function for $r > 0$.

2. Suppose $r > 0 > s$. Then $\dfrac{r}{s} < 0$, so $f(x) = x^{\frac{r}{s}}$ is convex. The same inequality holds and taking both sides to the power of $\dfrac{1}{r}$ still preserves the inequality because $g(x) = x^{\frac{1}{r}}$ is increasing again due to $r > 0$.

3. Suppose $0 > r > s$. Then $0 < \dfrac{r}{s} < 1$, so $x^{\frac{r}{s}}$ is concave this time. Then the reverse of the inequality from the previous two cases holds, but this is reversed again by taking both sides to the power of $\dfrac{1}{r}$ because $g(x) = x^{\frac{1}{r}}$ is decreasing for $r < 0$.

By Jensen, equality holds if and only if $x_1^s = x_2^s = \cdots = x_n^s$, since all of the λ_i are positive; in particular, none of the λ_i are 0. Since $s \neq 0$, we can invert it to state that equality holds and only if all of the x_i are equal.

Solution 11.24. We will use the weighted AM-GM inequality. To this end, let $\lambda_i = \dfrac{1}{\mu_i}$ for each index $i \in [n]$ so that $\lambda_i \in (0, 1)$, and

$$\lambda_1 + \lambda_2 + \cdots + \lambda_n = 1$$

By the weighted AM-GM inequality,

$$\begin{aligned} x_1 x_2 \cdots x_n &= (x_1^{\mu_1})^{\lambda_1} (x_2^{\mu_2})^{\lambda_2} \cdots (x_n^{\mu_n})^{\lambda_n} \\ &\leq \lambda_1 x_1^{\mu_1} + \lambda_2 x_2^{\mu_2} + \cdots + \lambda_n x_n^{\mu_n} \\ &= \frac{x_1^{\mu_1}}{\mu_1} + \frac{x_2^{\mu_2}}{\mu_2} + \cdots + \frac{x_n^{\mu_n}}{\mu_n}. \end{aligned}$$

None of the $\dfrac{1}{\mu_i}$ can be 0, so the equality criterion of the weighted AM-GM inequality says that equality holds if and only if

$$x_1^{\mu_1} = x_2^{\mu_2} = \cdots = x_n^{\mu_n}.$$

Solution 11.26. In Hölder's inequality, we take $m = 2$ and the two lists to be $(x_k^p)_{k=1}^n$ and $(y_k^q)_{k=1}^n$. From there, the result easily follows because, for any real t, $(t^p)^{\frac{1}{p}} = t$ and $(t^q)^{\frac{1}{q}} = t$. By the equality criterion of Hölder's inequality, equality holds if and only if one of the lists, that is $(x_k^p)_{k=1}^n$ or $(y_k^q)_{k=1}^n$, consists of all 0's or

$$x_1^p : x_2^p : \cdots : x_k^p = y_1^q : y_2^q : \cdots : y_k^q.$$

So each list is a scaled version of the other, which establishes the existence of a scale factor c, which must be positive because all of the x_k and y_k are positive. To be precise about what we mean by a scale factor, it means $y_k^q = c x_k^p$ for all indices $k \in [n]$.

Solution 11.31. Fixing any positive integer n and taking the assumption for granted, we will prove the conclusion by induction on $k \in [n-1]$. In the case base, where $k = 1$, we

prove the following by taking reversible steps:

$$\begin{aligned}
\frac{a_2 - a_0}{2} &\geq \frac{a_1 - a_0}{1} \\
a_2 - a_0 &\geq 2a_1 - 2a_0 \\
\frac{a_2 + a_0}{2} &\geq a_1,
\end{aligned}$$

which is true by assumption. Suppose the conclusion is true for some integer k such that $1 \leq k < n - 1$. To get to $k + 1$, we first manipulate the induction hypothesis to get

$$\begin{aligned}
\frac{a_{k+1} - a_0}{k+1} &\geq \frac{a_k - a_0}{k} \\
k(a_{k+1} - a_0) &\geq (k+1)(a_k - a_0) \\
ka_{k+1} + a_0 &\geq (k+1)a_k.
\end{aligned}$$

Then, by assumption and the above,

$$\begin{aligned}
2a_{k+1} &\leq a_k + a_{k+2} \\
&\leq \frac{k}{k+1} \cdot a_{k+1} + \frac{1}{k+1} \cdot a_0 + a_{k+2} \\
2(k+1)a_{k+1} &\leq ka_{k+1} + a_0 + (k+1)a_{k+2} \\
(k+2)a_{k+1} &\leq (k+1)a_{k+2} + a_0 \\
(k+2)a_{k+1} - (k+2)a_0 &\leq (k+1)a_{k+2} - (k+1)a_0 \\
\frac{a_{k+1} - a_0}{k+1} &\leq \frac{a_{k+2} - a_0}{k+2},
\end{aligned}$$

as desired.

List of Symbols

Arithmetic

$\mathbb{Z}$ integers

$\mathbb{Z}_+$ positive integers

$\mathbb{Z}_{\geq 0}$ non-negative integers

$\mathbb{Q}$ rational numbers

$\mathbb{Q}_+$ positive rationals

$\mathbb{Q}_{\geq 0}$ non-negative rationals

$\mathbb{R}$ real numbers

$\mathbb{R}_+$ positive reals

$\mathbb{R}_{\geq 0}$ non-negative reals

$\mathbb{C}$ complex numbers

$\mathbb{F}$ field

$\pm$ plus or minus

$<, >$ strict inequality

$\leq, \geq$ non-strict inequality

Constants

$\zeta_k = e^{\frac{2k\pi}{m}i}$ m^{th} root of unity

π pi

e Euler's constant

ϕ the golden ratio

i the square root of -1

Functions

$\text{Dom}(f)$ domain

$\text{Rng}(f)$ range

$\lfloor \cdot \rfloor$ floor function

$\lceil \cdot \rceil$ ceiling function

sgn signum function

max maximum function

min minimum function

det determinant

Id_S identity function on S

$f \circ g$ function composition

$n!$ factorial

$\overline{z}$ complex conjugate

$\underline{r}$ radical conjugate

σ permutation

csc cosecant

sin sine

cos cosine

sec secant

cot cotangent

tan tangent

$\bullet$ dot product of vectors

$\times$ cross product of vectors

Logic

$\neg$ negation

$\vee$ disjunction, or

$\wedge$ conjunction, and

$\oplus$ exclusive or, XOR

$\Longrightarrow$ implication

$\iff$ biconditional

$\equiv$ logical equivalence

Miscellaneous

$\exists$ existential quantifier

$\forall$ universal quantifier

$(a_i)_{i\in I}$ sequence indexed by I

$\sum$ summation notation

$\prod$ product notation

$a \sim b$ equivalence relation

Sets

$\varnothing$ empty set

$\in$ element of

$\notin$ not element of

$[n]$ $\{1, 2, \ldots, n\}$ for $n \in \mathbb{Z}_+$

$[n]^*$ $\{0, 1, 2, \ldots, n\}$ for $n \in \mathbb{Z}_{\geq 0}$

S^c set complement

$\cup$ set union

$\cap$ set intersection

$A \backslash B$ set difference

$A \times B$ Cartesian product of sets

A^n $\underbrace{A \times A \times \cdots \times A}_{n \text{ copies of } A}$

$\mathcal{P}(A)$ power set

$A \oplus B$ set symmetric difference

$\subseteq$ subset

$\subsetneq$ proper subset

$\supseteq$ superset

$\mathcal{U}$ universal set

Bibliography

"When we can't think for ourselves, we can always quote."
– *Ludwig Wittgenstein*

[1] David A. Cox at al. *Ideals, Varieties, and Algorithms, fourth edition*. Springer International Publishing, 2015, pp. 347–348.

[2] Donald E. Knuth at al. *Concrete Mathematics, second edition*. Addison-Wesley, 1994, pp. 71–72.

[3] Radmila Manfrino et al. *Inequalities: A Mathematical Olympiad Approach*. Birkhauser Verlag AG, 2009, pp. 43–47.

[4] Titu Andreescu and Zuming Feng. *A Path to Combinatorics for Undergraduates: Counting Strategies*. Birkhauser, 2004, p. 164.

[5] Edward Barbeau and Samer Seraj. "Sum of cubes is square of sum". In: *Notes on Number Theory and Discrete Mathematics* 19(1) (2013), pp. 1–13.

[6] Evan Chen. *Euclidean Geometry in Mathematical Olympiads*. MAA Press, 2016, pp. 95–118.

[7] Pham Kim Hung. *Secrets in Inequalities: Advanced Inequalities*. GIL Publishing House, 2008.

[8] Pham Kim Hung. *Secrets in Inequalities: Basic Inequalities*. GIL Publishing House, 2007.

[9] Samer Seraj. "A Short Proof of Euler's Inequality". In: *Resonance - Journal of Science Education* 20(1) (2015), p. 75.

[10] J. Michael Steele. *The Cauchy-Schwarz Masterclass: An Introduction to the Art of Mathematical Inequalities*. Cambridge University Press, 2004.

Index

"We raise to degrees (of wisdom) whom We please: but over all endued with knowledge is one, the All-Knowing."
– *Qur'an 12:76*

About the Author

> "Why is it the words we write for ourselves are always so much better than the words we write for others?... You write your first draft with your heart. You rewrite with your head. The first key to writing is to write, not to think."
>
> – *Sean Connery, Finding Forrester*

> "If you would be a real seeker after truth, it is necessary that at least once in your life you doubt, as far as possible, all things."
>
> – *René Descartes*

Samer Seraj is the owner of Existsforall Academy Inc., which is a Canadian company that specializes in mathematical education. During his school years, his participation in math contests culminated in his qualification for the Canadian Mathematical Olympiad and the Asian Pacific Mathematics Olympiad in his senior year of high school. He then spent four years learning higher mathematics and earned his undergraduate degree in mathematics from Trinity College at the University of Toronto. At the time, he won two prestigious research grants, presented papers at several conferences, and was elected as President of the student body's Mathematics Union. After graduation, he worked for four years in a mix of roles as a mathematics instructor, curriculum developer, and personnel manager of a team of over five hundred educators at a company based in San Diego, California. More recently, he founded Existsforall Academy, where he enjoys teaching his students. His recent contributions to the Canadian mathematical community have included being a guest editor of the Canadian Mathematical Society's problem-solving journal, Crux Mathematicorum, sitting on the University of Waterloo CEMC's committee for the Problem of the Month, teaching courses at the University of Toronto's math outreach program, Math+, and serving as a trainer of Team Canada for the International Mathematical Olympiad.

https://existsforall.com/

ISBN 978-1-7389501-0-2