PreCalculus: Concrete Abstractions

Gudfit

Contents

Chapter 1

Ideas & Motivations

Welcome to Concrete Abstractions (with some theory) by me (Gudfit). The point of these notes is to cover everything I think is important as I build up to my current knowledge, while keeping it free and accessible for everyone from kids to adults.

I aim for each set of notes to be max 100 pages, as rigorous as possible, and far-reaching too.

That means I'll cover the axioms and proofs of the most interesting stuff, plus I'll pull in other subjects we've already touched on to show how math builds on itself like funky Lego.

Building on my existing Logic and Set Theory notes, this text balances intuition with formality.

Disclaimer: These notes are unabashedly adapted from Concrete Mathematics by Donald Knuth (and co-authors Graham and Patashnik). Unironically, you might as well read his book. When I first made the original version in 2020, I essentially took his work and translated it into something I could understand personally. I take no credit for the brilliance here, just the re-phrasing.

Chapter 2

Recurrence Relations

With the foundations of set theory, functions, and the principle of mathematical induction established, we turn our attention to the analysis of algorithms and dynamic processes. Many mathematical problems are best described not by a static equation, but by a rule that defines the state of a system in terms of its previous states. Such rules are called recurrence relations.

Our approach to solving these problems follows a three-stage methodology:

1. **Abstraction**: Construct a mathematical model of the problem.
2. **Recursion**: Derive a recurrence relation that describes the problem in terms of smaller instances of itself.
3. **Closed Form**: Determine a non-recursive formula for the n-th term and prove its equivalence to the recursive definition.

2.1 The Tower of Hanoi

We begin with the most basic problem, the Tower of Hanoi, a puzzle attributed to the French mathematician Édouard Lucas in 1883.

Definition 2.1.1. ***The Setup***. The puzzle consists of three vertical pegs and n disks of distinct sizes. Initially, the disks are stacked on the first peg in order of decreasing size, with the largest at the bottom and the smallest at the top.

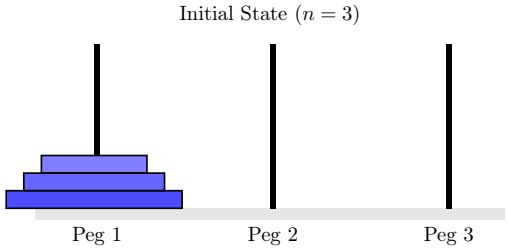


Figure 2.1: The Tower of Hanoi with $n = 3$ disks.

The objective is to transfer the entire stack to a different peg, adhering to the following constraints (L-rules):

1. Only one disk may be moved at a time.
2. A disk may only be placed on top of a larger disk or an empty peg.

We seek the minimum number of moves, denoted T_n, required to transfer n disks.

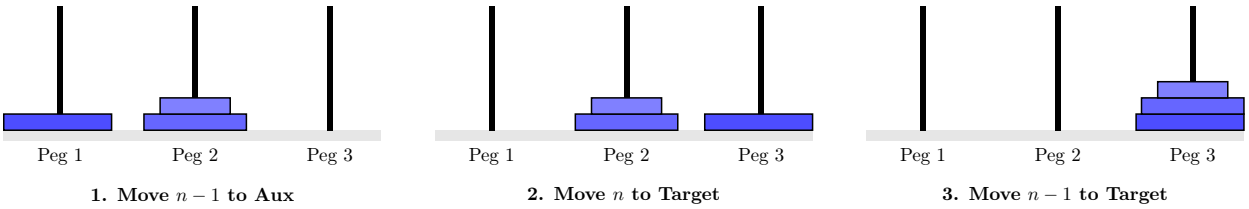
Developing the Recurrence

We analyse the problem for small values of n to identify a pattern.

- **Case** $n = 0$: No disks require zero moves. $T_0 = 0$.
- **Case** $n = 1$: Simply move the single disk to the target peg. $T_1 = 1$.
- **Case** $n = 2$: We must move the top (small) disk to an auxiliary peg (1 move), move the bottom (large) disk to the target peg (1 move), and place the small disk on top of the large one (1 move). Thus, $T_2 = 3$.

Consider the general case for n disks. To move the largest disk (disk n) from the source peg to the target peg, the $n-1$ disks sitting above it must first be moved to the third (auxiliary) peg. The largest disk is then free to move. Finally, the stack of $n-1$ disks must be moved from the auxiliary peg onto the target peg.

This recursive strategy is illustrated in Figure 2.2.



1. Move $n-1$ to Aux **2. Move n to Target** **3. Move $n-1$ to Target**

Figure 2.2: The recursive strategy

The strategy outlines a sufficient procedure:

1. Move $n-1$ disks to the auxiliary peg: T_{n-1} moves.
2. Move the largest disk to the target peg: 1 move.
3. Move $n-1$ disks from the auxiliary peg to the target peg: T_{n-1} moves.

The total moves required by this strategy is $T_{n-1} + 1 + T_{n-1}$. Therefore, the minimum number of moves satisfies the inequality:

$$T_n \leq 2T_{n-1} + 1$$

Conversely, to move disk n, all disks above it *must* be removed. Moving the stack of $n-1$ disks requires at least T_{n-1} moves. Moving disk n requires at least 1 move. Reassembling the stack requires another T_{n-1} moves. Thus, any valid sequence of moves must satisfy:

$$T_n \geq 2T_{n-1} + 1$$

Combining these inequalities yields the recurrence relation.

Definition 2.1.2. ***Recurrence for Tower of Hanoi***. The minimum number of moves T_n required to solve the Tower of Hanoi puzzle with n disks is given by:

$$T_n = \begin{cases} 0 & \text{if } n = 0 \\ 2T_{n-1} + 1 & \text{if } n > 0 \end{cases}$$

2.2 Closed-Form Solution

While the recurrence relation allows us to compute T_n, the computational effort increases linearly with n. We seek a closed-form formula $f(n)$ that allows direct calculation.

Calculating the first few terms:

n	0	1	2	3	4	5	6
T_n	0	1	3	7	15	31	63

The sequence resembles powers of 2. Specifically, $T_n = 2^n - 1$. We formally prove this equivalence using mathematical induction.

***Theorem* 2.2.1. Closed Form of Hanoi**. For all $n \in \mathbb{N}$, the recurrence $T_n = 2T_{n-1} + 1$ with $T_0 = 0$ is equivalent to:

$$T_n = 2^n - 1$$

Proof. **Base Case** ($n = 0$): From the closed form, $2^0 - 1 = 1 - 1 = 0$. This matches the definition $T_0 = 0$.

Inductive Step: Assume the hypothesis holds for some $k \in \mathbb{N}$, i.e., $T_k = 2^k - 1$. We examine T_{k+1}:

$$\begin{aligned} T_{k+1} &= 2T_k + 1 && \text{by the recurrence relation} \\ &= 2(2^k - 1) + 1 && \text{by the inductive hypothesis} \\ &= 2^{k+1} - 2 + 1 \\ &= 2^{k+1} - 1 \end{aligned}$$

The formula holds for $k + 1$. By the Principle of Mathematical Induction, $T_n = 2^n - 1$ for all $n \in \mathbb{N}$. ■

Alternative Derivation: Linearisation

We can also derive the closed form without guessing the pattern by transforming the recurrence. Consider the relation $T_n = 2T_{n-1} + 1$. Adding 1 to both sides yields:

$$T_n + 1 = 2T_{n-1} + 2 = 2(T_{n-1} + 1)$$

Let $U_n = T_n + 1$. The recurrence becomes:

$$U_n = 2U_{n-1}$$

with the initial condition $U_0 = T_0 + 1 = 1$. This describes a geometric progression with ratio 2.

$$U_n = 2^n \cdot U_0 = 2^n$$

Returning to our original variable:

$$T_n = U_n - 1 = 2^n - 1$$

This method confirms our previous result through direct manipulation of the recurrence structure.

2.3 Lines in the Plane

Following the analysis of the Tower of Hanoi, we turn to a geometric problem posed by the Swiss mathematician Jacob Steiner in 1826. The fundamental question is one of maximisation: what is the maximum number of regions, denoted L_n, into which a plane can be divided by n straight lines?

Developing the Recurrence

We begin by establishing the base cases through geometric intuition.

- **Case** $n = 0$: With no lines, the plane is a single, undivided region. Thus, $L_0 = 1$.
- **Case** $n = 1$: A single line divides the plane into two half-planes. Thus, $L_1 = 2$.
- **Case** $n = 2$: Two distinct lines can either be parallel (creating 3 regions) or intersecting. If they intersect, they divide the plane into four regions. As we seek the maximum, $L_2 = 4$.

Observing the sequence $1, 2, 4$, one might conjecture that $L_n = 2^n$. However, introducing a third line reveals the flaw in this assumption. To maximise regions, the third line must intersect both existing lines, but it

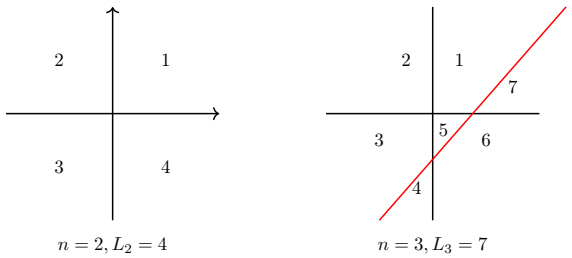


Figure 2.3: Partitioning the plane with 2 and 3 lines. Note that for $n = 3$, the third line (red) intersects the previous two at distinct points, adding 3 new regions.

cannot intersect them at the same point (which would reduce the number of new regions formed). Following this construction, shown in Figure 2.3, we find $L_3 = 7$, not 8.

To determine the general recurrence, consider the state where $n-1$ lines are already drawn, defining L_{n-1} regions. We introduce the n-th line. To maximise the number of new regions:

1. The new line must not be parallel to any of the existing $n-1$ lines.
2. The new line must not pass through any existing intersection points.

Under these conditions, the n-th line intersects the existing $n-1$ lines at $n-1$ distinct points. These intersection points divide the new line into n distinct segments (including the two infinite rays at either end).

Each of these n segments divides an existing region into two parts. Consequently, the addition of the n-th line adds exactly n new regions to the plane.

Definition 2.3.1. ***Recurrence for Lines in the Plane***. The maximum number of regions L_n defined by n lines in the plane is given by:

$$L_n = \begin{cases} 1 & \text{if } n = 0 \\ L_{n-1} + n & \text{if } n > 0 \end{cases}$$

Closed-Form Solution

We define the function $L : \mathbb{N} \to \mathbb{N}$ by the recurrence above. Unfolding the recurrence yields a clear pattern:

$$\begin{aligned} L_n &= L_{n-1} + n \\ &= L_{n-2} + (n-1) + n \\ &= \vdots \\ &= L_0 + 1 + 2 + \cdots + (n-1) + n \\ &= 1 + \sum_{i=1}^{n} i \end{aligned}$$

We recall the formula for the sum of the first n integers: $\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$. Substituting this into our expression yields the closed form.

***Theorem* 2.3.1. Closed Form of Steiner's Problem**. For all $n \in \mathbb{N}$, the maximum number of regions defined by n lines is:

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

Proof. **Base Case ($n = 0$)**: The formula yields $\frac{0(1)}{2} + 1 = 1$, which matches $L_0 = 1$.

Inductive Step: Assume the hypothesis holds for some $k \in \mathbb{N}$, so $L_k = \frac{k(k+1)}{2} + 1$. We calculate L_{k+1}:

$$\begin{aligned} L_{k+1} &= L_k + (k+1) && \text{by the recurrence definition} \\ &= \left(\frac{k(k+1)}{2} + 1\right) + (k+1) && \text{by the inductive hypothesis} \\ &= \frac{k(k+1) + 2(k+1)}{2} + 1 \\ &= \frac{(k+1)(k+2)}{2} + 1 \end{aligned}$$

This result matches the closed-form formula for $n = k+1$. Thus, the equivalence holds for all $n \in \mathbb{N}$. ■

Variation: The Zig-Zag Problem

We now extend this reasoning to "bent" lines. A bent line is defined as two rays meeting at a vertex; essentially a line with a single angle. We denote the maximum number of regions defined by n bent lines as Z_n.

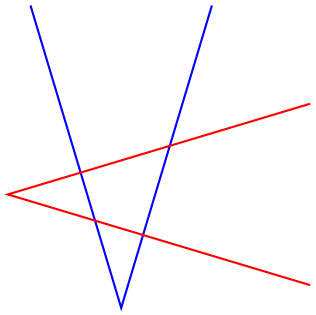


Figure 2.4: Two bent lines can create up to $Z_2 = 7$ regions.

Intuitively, a bent line is topologically similar to two intersecting straight lines, but with the regions "behind" the vertex removed. Specifically, if we extended the two rays of a bent line into full straight lines, the vertex would be their intersection point. The "cut" eliminates the two unbounded regions formed by the extensions.

Therefore, placing n bent lines is analogous to placing $2n$ straight lines, minus the regions lost at the n vertices. Each vertex reduces the region count by 2 compared to the full straight-line case.

***Theorem* 2.3.2. Closed Form for Bent Lines**. The maximum number of regions Z_n defined by n bent lines is given by:

$$Z_n = L_{2n} - 2n = 2n^2 - n + 1$$

Proof. We substitute the closed form for L_{2n} into the relation $Z_n = L_{2n} - 2n$:

$$\begin{aligned} Z_n &= \left(\frac{2n(2n+1)}{2} + 1\right) - 2n \\ &= n(2n+1) + 1 - 2n \\ &= 2n^2 + n + 1 - 2n \\ &= 2n^2 - n + 1 \end{aligned}$$

■

Comparing the two variations, for large n, $L_n \approx \frac{1}{2}n^2$ while $Z_n \approx 2n^2$. Consequently, using bent lines allows us to define approximately four times as many regions as straight lines for the same number of objects.

2.4 The Josephus Problem

The Josephus Problem is a classic problem in combinatorics and computer science, named after the Jewish-Roman historian Flavius Josephus. The problem is derived from an account of the Siege of Yodfat in 67 CE, where Josephus and 40 rebels were trapped in a cave by Roman soldiers. Preferring death to capture, they decided to form a circle and proceed to kill every k-th person until only one remained. Josephus, wishing to survive, calculated the position where he should stand to be the last survivor.

Mathematically, we consider n people arranged in a circle, indexed from 1 to n. Beginning with the second person, we eliminate every k-th person. We seek the position denoted $J(n)$ of the survivor. For the scope of this chapter, we restrict our initial analysis to the case where $k = 2$.

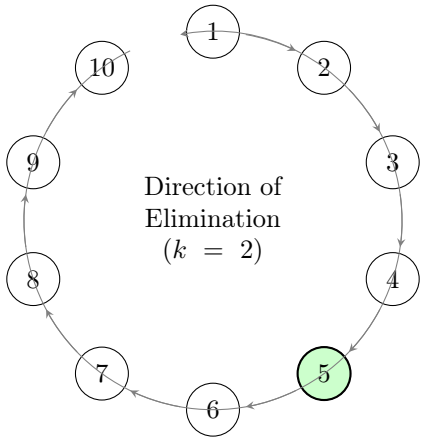
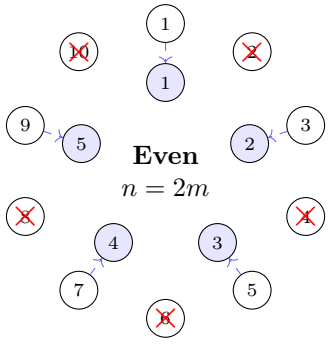


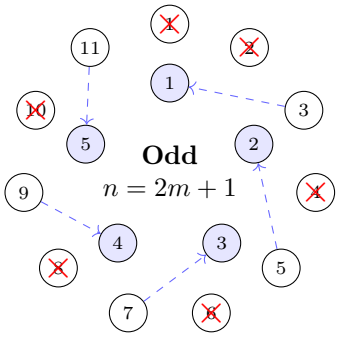
Figure 2.5: The Josephus circle for $n = 10$. Elimination starts with 2. The green node (5) is the survivor.

Developing the Recurrence

Let $J(n)$ denote the survivor's position in a circle of size n with a step of $k = 2$. To derive the recurrence, we analyse how the circle reduces after a single pass of eliminations.



Even Case ($n = 10$)
Elimination of evens leaves 5 odd positions. $1 \to 1$, $3 \to 2$, etc.



Odd Case ($n = 11$)
Elimination of evens *and* 1 leaves 5 odd positions. $3 \to 1$, $5 \to 2$, etc.

Figure 2.6: Visualisation of the reduction step. The outer ring represents the original circle. Red crosses mark eliminations in the first pass. The inner ring represents the re-indexed sub-problem $J(m)$. The dashed arrows show the mapping from the original position to the new recursive position.

- **Base Case**: With $n = 1$, the survivor is trivially the only person. Thus, $J(1) = 1$.

- **Even Case ($n = 2m$)**: In the first pass around the circle, people at positions $2, 4, 6, \ldots, 2m$ are eliminated. This leaves the m people at positions $1, 3, 5, \ldots, 2m-1$. The person originally at position $2i-1$ is now the i-th person in the remaining circle (see Figure 2.6, left). If the survivor is the x-th person in this reduced circle, their original position was $2x-1$. Thus:

$$J(2m) = 2J(m) - 1$$

- **Odd Case ($n = 2m+1$)**: In the first pass, people at positions $2, 4, \ldots, 2m$ are eliminated. The next person to be eliminated is at position 1 (wrapping around). This leaves m people at positions $3, 5, \ldots, 2m+1$. The person originally at position $2i+1$ corresponds to the i-th person in the new circle (see Figure 2.6, right). Recovering the original index gives:

$$J(2m+1) = 2J(m) + 1$$

This yields a complete system of recurrence relations for the Josephus problem with $k = 2$:

$$\begin{aligned} J(1) &= 1 \\ J(2n) &= 2J(n) - 1 \quad \text{for } n \geq 1 \\ J(2n+1) &= 2J(n) + 1 \quad \text{for } n \geq 1 \end{aligned} \tag{2.1}$$

Closed-Form Solution

We compute the first few values of $J(n)$ to identify a pattern.

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$J(n)$	1	1	3	1	3	5	7	1	3	5	7	9	11	13	15	1

The values of $J(n)$ reset to 1 whenever n is a power of 2 and increase by 2 thereafter. Let us express n as $n = 2^m + l$, where 2^m is the largest power of 2 such that $2^m \leq n$, and $0 \leq l < 2^m$ is the remainder. The table suggests the closed form:

$$J(2^m + l) = 2l + 1$$

***Theorem* 2.4.1. Closed Form of the Josephus Problem**. For any $n \in \mathbb{Z}^+$, let $n = 2^m + l$ where $0 \leq l < 2^m$. Then:

$$J(n) = 2l + 1$$

Proof. We proceed by induction on m.

Base Case ($m = 0$): Here $n = 1$ and $l = 0$. The formula gives $2(0) + 1 = 1$, which matches $J(1) = 1$.

Inductive Step: Let $m > 0$. Assume the formula holds for all $k < 2^m$. We verify it for n such that $2^m \leq n < 2^{m+1}$. Let $n = 2^m + l$.

1. **If n is even**: Let $n = 2k$. Then l must be even. We write $k = 2^{m-1} + l/2$.

$$\begin{aligned} J(2^m + l) &= 2J(2^{m-1} + l/2) - 1 && \text{by (2.1)} \\ &= 2(2(l/2) + 1) - 1 && \text{by hypothesis} \\ &= 2l + 2 - 1 = 2l + 1 \end{aligned}$$

2. **If n is odd**: Let $n = 2k + 1$. Then l is odd. We write $k = 2^{m-1} + (l-1)/2$.

$$\begin{aligned} J(2^m + l) &= 2J(2^{m-1} + (l-1)/2) + 1 && \text{by (2.1)} \\ &= 2(2((l-1)/2) + 1) + 1 && \text{by hypothesis} \\ &= 2(l-1) + 2 + 1 = 2l + 1 \end{aligned}$$

The hypothesis holds for all n. ■

Radix Interpretation

To gain deeper insight into the solution $J(n) = 2l + 1$, let us examine the structure of n and l using binary representation. Recall that we decomposed n as $n = 2^m + l$, where 2^m is the largest power of 2 not exceeding n, and $0 \leq l < 2^m$.

In binary notation, the term 2^m corresponds to a 1 followed by m zeros. Since $l < 2^m$, the binary representation of l requires at most m bits. Therefore, when we write $n = 2^m + l$, the 2^m term corresponds precisely to the most significant bit (the leading 1), and l corresponds to the remaining bits.

Let the binary expansion of n be given by:

$$n = (b_m b_{m-1} \dots b_1 b_0)_2$$

By definition of binary notation:

$$n = 1 \cdot 2^m + b_{m-1} \cdot 2^{m-1} + \cdots + b_1 \cdot 2^1 + b_0 \cdot 2^0$$

Comparing this to $n = 2^m + l$, we identify:

- The leading bit $b_m = 1$.
- The remainder l is represented by the remaining bits: $l = (0 b_{m-1} \dots b_1 b_0)_2$.

Now, let us construct the survivor's position $J(n) = 2l + 1$ using these bits.

1. **Calculate** $2l$: Multiplying a binary number by 2 is equivalent to a left shift (appending a 0).
$$2l = 2 \cdot (b_{m-1} \dots b_0)_2 = (b_{m-1} \dots b_0 0)_2$$
2. **Calculate** $2l + 1$: Adding 1 changes the least significant bit from 0 to 1.
$$2l + 1 = (b_{m-1} \dots b_0 1)_2$$

Combining these observations, we can express the transformation from n to $J(n)$ directly in terms of their bit patterns:

$$n = (1 b_{m-1} \dots b_0)_2 \quad \xrightarrow{J} \quad J(n) = (b_{m-1} \dots b_0 1)_2$$

The operation performed is a cyclic left shift on the binary representation of n. The most significant bit (which is always 1) is removed from the front and appended to the end.

Example 2.4.1. Let $n = 19$. 19 in binary $= 10011_2$ and $\rightarrow 00111_2$ after cyclic shift. The result is 00111_2 or 7 in decimal. To confirm we check against the formula. Using the formula: $19 = 16+3$, so $J(19) = 2(3)+1 = 7$.

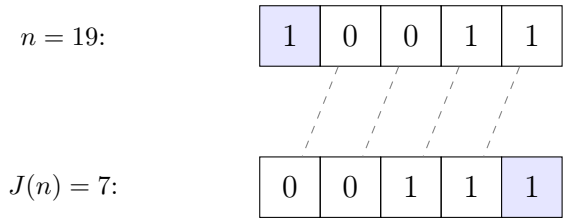


Figure 2.7: Calculating $J(19)$ via cyclic bit shift.

2.5 Generalised Josephus Recurrence

We now extend the Josephus problem to a broader class of recurrence relations. Let us define a function $f : \mathbb{Z}^+ \to \mathbb{R}$ subject to the following recursive definitions involving constants $\alpha, \beta, \gamma \in \mathbb{R}$:

$$\begin{aligned} f(1) &= \alpha \\ f(2n) &= 2f(n) + \beta \quad \text{for } n \geq 1 \\ f(2n+1) &= 2f(n) + \gamma \quad \text{for } n \geq 1 \end{aligned} \tag{2.2}$$

Remark. The original Josephus problem $J(n)$ corresponds to the specific case where $(\alpha, \beta, \gamma) = (1, -1, 1)$.

To determine the closed-form solution for $f(n)$, we employ the **Repertoire Method**. This approach relies on the observation that the recurrence is linear in the parameters α, β, and γ. Consequently, the solution must take the form of a linear combination:

$$f(n) = \alpha A(n) + \beta B(n) + \gamma C(n) \tag{2.3}$$

Our goal is to determine the coefficient functions $A(n)$, $B(n)$, and $C(n)$. We do this by testing specific functions $f(n)$ (our "repertoire") for which we know the behavior, substituting them into the recurrence to find relations between the parameters, and solving the resulting system of equations.

2.5.1 The Repertoire Method

We begin by tabulating the values of $f(n)$ for small n to establish a hypothesis for the functions A, B, C. Let $n = 2^m + l$, where $0 \leq l < 2^m$.

n	Recursive Expansion	In terms of α, β, γ		
1	$f(1)$	1α		
2	$2f(1) + \beta$	2α	$+1\beta$	
3	$2f(1) + \gamma$	2α		$+1\gamma$
4	$2f(2) + \beta$	4α	$+3\beta$	
5	$2f(2) + \gamma$	4α	$+2\beta$	$+1\gamma$
6	$2f(3) + \beta$	4α	$+1\beta$	$+2\gamma$
7	$2f(3) + \gamma$	4α		$+3\gamma$
8	$2f(4) + \beta$	8α	$+7\beta$	
9	$2f(4) + \gamma$	8α	$+6\beta$	$+1\gamma$

From this table, we can form initial conjectures:

- The coefficient of α appears to be 2^m. Thus, we guess $A(n) = 2^m$.
- The coefficient of γ, $C(n)$, appears to be l (the remainder).
- The coefficient of β, $B(n)$, appears to be $2^m - 1 - l$.

We now prove these conjectures formally using the Repertoire Method.

Step 1: Determine $A(n)$

Consider the case where $\alpha = 1$ and $\beta = \gamma = 0$. The recurrence becomes:

$$f(1) = 1, \quad f(2n) = 2f(n), \quad f(2n+1) = 2f(n)$$

Substituting these parameters into the general form (2.3):

$$f(n) = 1 \cdot A(n) + 0 \cdot B(n) + 0 \cdot C(n) \implies f(n) = A(n)$$

We prove that $A(n) = 2^m$, where 2^m is the highest power of 2 dividing n, by induction on m.

Proof. **Base Case** $(m = 0)$: $n = 1$. $f(1) = 1$, and $2^0 = 1$. Holds.

Inductive Step: Assume $A(k) = 2^{m-1}$ for all k such that $2^{m-1} \leq k < 2^m$. Let n be in the range $2^m \leq n < 2^{m+1}$.

- If n is even, $n = 2k$. Then $A(2k) = 2A(k) = 2(2^{m-1}) = 2^m$.
- If n is odd, $n = 2k + 1$. Then $A(2k + 1) = 2A(k) = 2(2^{m-1}) = 2^m$.

This establishes our first equation:

$$A(n) = 2^m \tag{2.4}$$

■

Step 2: The Constant Function

Let $f(n) = 1$ for all n. We substitute this into the recurrence (2.2) to find the required parameters.

- Base: $1 = \alpha \implies \alpha = 1$.
- Even: $f(2n) = 2f(n) + \beta \implies 1 = 2(1) + \beta \implies \beta = -1$.
- Odd: $f(2n + 1) = 2f(n) + \gamma \implies 1 = 2(1) + \gamma \implies \gamma = -1$.

Substituting these values ($f(n) = 1, \alpha = 1, \beta = -1, \gamma = -1$) into the ansatz (2.3) yields:

$$1 = 1 \cdot A(n) - 1 \cdot B(n) - 1 \cdot C(n)$$

This provides our second equation:

$$A(n) - B(n) - C(n) = 1 \tag{2.5}$$

Step 3: The Linear Function

Let $f(n) = n$ for all n. We substitute this into the recurrence.

- Base: $1 = \alpha \implies \alpha = 1$.
- Even: $2n = 2(n) + \beta \implies \beta = 0$.
- Odd: $2n + 1 = 2(n) + \gamma \implies \gamma = 1$.

Substituting these values ($f(n) = n, \alpha = 1, \beta = 0, \gamma = 1$) into the ansatz yields:

$$n = 1 \cdot A(n) + 0 \cdot B(n) + 1 \cdot C(n)$$

This provides our third equation:

$$A(n) + C(n) = n \tag{2.6}$$

Step 4: Solving the System

We now have a system of three equations:

1. $A(n) = 2^m$
2. $A(n) - B(n) - C(n) = 1$
3. $A(n) + C(n) = n$

We solve for $B(n)$ and $C(n)$ in terms of n and $A(n)$. From (3):

$$C(n) = n - A(n) = n - 2^m = l$$

From (2):

$$B(n) = A(n) - 1 - C(n) = 2^m - 1 - l$$

***Theorem* 2.5.1. Closed Form of Generalised Josephus**. For any constants α, β, γ, the solution to the recurrence (2.2) for $n = 2^m + l$ is:

$$f(n) = 2^m\alpha + (2^m - 1 - l)\beta + l\gamma$$

2.5.2 Radix Interpretation of the Generalised Form

The closed-form solution derived via the Repertoire Method can be interpreted elegantly using a modified binary representation. Recall the recurrence $f(2n+j) = 2f(n) + \beta_j$ where $\beta_0 = \beta$ and $\beta_1 = \gamma$. Let $n = (b_m b_{m-1} \dots b_0)_2$. Unfolding the recurrence yields:

$$\begin{aligned} f((b_m \dots b_0)_2) &= 2f((b_m \dots b_1)_2) + \beta_{b_0} \\ &= 2\left(2f((b_m \dots b_2)_2) + \beta_{b_1}\right) + \beta_{b_0} \\ &\vdots \\ &= 2^m f(b_m) + 2^{m-1}\beta_{b_{m-1}} + \dots + \beta_{b_0} \end{aligned}$$

Since $b_m = 1$, $f(b_m) = f(1) = \alpha$. Thus:

$$f(n) = 2^m \alpha + \sum_{i=0}^{m-1} 2^i \beta_{b_i}$$

We define the Relaxed Radix Representation $(c_m c_{m-1} \dots c_0)_2$ as $\sum c_i 2^i$. Then the solution is simply:

$$f(n) = (\alpha \beta_{b_{m-1}} \dots \beta_{b_0})_2$$

where $\beta_{b_i} = \beta$ if $b_i = 0$ and γ if $b_i = 1$.

The Block Transformation Property

Why does this relaxed representation equate to the cyclic shift found earlier for the original Josephus Problem? For the original problem, $\alpha = 1, \beta = -1, \gamma = 1$. The solution in relaxed radix is $f(n) = (1\beta_{b_{m-1}} \dots \beta_{b_0})_2$. In our relaxed notation, zeros in the input n become -1s.

$$f((1, 0, \dots, 0)_2) = (1, -1, \dots, -1)_2$$

We assert that a block of the form $(1, -1, \dots, -1)_2$ is equivalent to $(0, \dots, 0, 1)_2$.

Proof. Consider the value of a block starting with 1 followed by k entries of -1.

$$\begin{aligned} (1, \underbrace{-1, \dots, -1}_{k})_2 &= 1 \cdot 2^k - \sum_{i=0}^{k-1} 2^i \\ &= 2^k - (2^k - 1) \\ &= 1 \end{aligned}$$

Thus, any block $(1, 0, \dots, 0)_2$ in the original binary string becomes $(1, -1, \dots, -1)_2$ in the relaxed string, which evaluates to 1. This is equivalent to shifting the leading 1 of that block to the least significant position of that block. Applying this logic recursively explains the cyclic shift. ■

The Radix Change Theorem

We can generalise this result further by changing the radix of the domain and codomain. Instead of doubling the input $(2n)$ and the output $(2f(n))$, we consider dividing the input by d and multiplying the output by c.

***Theorem* 2.5.2. Radix Change Theorem**. Let $f : \mathbb{Z}^+ \to \mathbb{R}$ be defined by the recurrence:

$$\begin{aligned} f(j) &= \alpha_j && \text{for } 1 \le j < d \\ f(dn + j) &= cf(n) + \beta_j && \text{for } 0 \le j < d \text{ and } n \ge 1 \end{aligned} \tag{2.7}$$

Let the input n have the base-d representation $(b_m b_{m-1} \dots b_0)_d$. Then the value of $f(n)$ is obtained by interpreting the corresponding coefficients as a number in base c:

$$f((b_m b_{m-1} \dots b_0)_d) = (\alpha_{b_m} \beta_{b_{m-1}} \dots \beta_{b_0})_c$$

Proof. The proof follows by induction on m, the number of digits. **Base Case**: If $n < d$, then $n = (b_0)_d$. By the recurrence definition, $f(n) = \alpha_{b_0}$, which matches the formula $(\alpha_{b_0})_c$.

Inductive Step: Let n have $m+1$ digits. We can write $n = dn' + b_0$, where $n' = (b_m \dots b_1)_d$. By the recurrence:

$$f(n) = cf(n') + \beta_{b_0}$$

By the inductive hypothesis, $f(n') = (\alpha_{b_m}\beta_{b_{m-1}} \dots \beta_{b_1})_c$. Substituting this back:

$$\begin{aligned} f(n) &= c\left[\sum_{i=1}^{m-1} \beta_{b_i} c^{i-1} + \alpha_{b_m} c^m\right] + \beta_{b_0} \\ &= \sum_{i=1}^{m-1} \beta_{b_i} c^i + \alpha_{b_m} c^{m+1} + \beta_{b_0} c^0 \\ &= (\alpha_{b_m}\beta_{b_{m-1}} \dots \beta_{b_0})_c \end{aligned}$$

Thus, the theorem holds. ■

Example 2.5.1. Consider the recurrence defined by:

$$\begin{aligned} f(1) = 34, \quad f(2) &= 5 \\ f(3n) &= 10f(n) + 76 \\ f(3n+1) &= 10f(n) - 2 \\ f(3n+2) &= 10f(n) + 8 \end{aligned}$$

Here, the input radix is $d = 3$ and the output multiplier is $c = 10$. The recurrence parameters are:

- Base cases: $\alpha_1 = 34$, $\alpha_2 = 5$.
- Recursive steps ($j \in \{0, 1, 2\}$): $\beta_0 = 76$, $\beta_1 = -2$, $\beta_2 = 8$.

We evaluate $f(19)$. First, we expand 19 in base 3:

$$19 = 2 \cdot 3^2 + 0 \cdot 3^1 + 1 \cdot 3^0 = (201)_3$$

Using the Radix Change Theorem, we map the digits $(2, 0, 1)$ to coefficients $(\alpha_2, \beta_0, \beta_1)$ and interpret them in base 10:

$$\begin{aligned} f(19) &= (\alpha_2\beta_0\beta_1)_{10} \\ &= 5 \cdot 10^2 + 76 \cdot 10^1 + (-2) \cdot 10^0 \\ &= 500 + 760 - 2 \\ &= 1258 \end{aligned}$$

2.6 A Five-Parameter Generalisation

We conclude our investigation into Josephus-style recurrences by solving a general five-parameter recurrence. This problem unifies the concepts of the Repertoire Method and the Relaxed Radix Representation.

Let $h : \mathbb{Z}^+ \to \mathbb{R}$ be defined by the following recurrence relation:

$$\begin{aligned} h(1) &= \alpha \\ h(2n) &= 4h(n) + \gamma_0 n + \beta_0 \quad \text{for } n \geq 1 \\ h(2n+1) &= 4h(n) + \gamma_1 n + \beta_1 \quad \text{for } n \geq 1 \end{aligned} \tag{2.8}$$

Note that the multiplicative factor is now $k = 4$. Based on the linearity of the recurrence, we posit a closed-form solution as a linear combination of the five parameters:

$$h(n) = \alpha A(n) + \gamma_0 B(n) + \gamma_1 C(n) + \beta_0 D(n) + \beta_1 E(n) \tag{2.9}$$

We must determine the five coefficient functions $A(n)$ through $E(n)$.

Applying the Radix Change Theorem

Consider the subset of cases where $\gamma_0 = \gamma_1 = 0$. The recurrence simplifies to:

$$h(2n + j) = 4h(n) + \beta_j \quad \text{for } j \in \{0, 1\}$$

This is a radix-change recurrence with input radix $d = 2$ and output multiplier $c = 4$. By the Radix Change Theorem, for $n = (1b_{m-1} \dots b_0)_2$, the solution is the relaxed 4-radix representation:

$$h(n) = (\alpha\beta_{b_{m-1}} \dots \beta_{b_0})_4 \tag{2.10}$$

In this restricted case, the ansatz (2.9) reduces to $h(n) = \alpha A(n) + \beta_0 D(n) + \beta_1 E(n)$. Comparing this with (2.10), we can identify A, D, and E.

The Coefficient $A(n)$ If we set $\beta_0 = \beta_1 = 0$, then $h(n) = (\alpha 0 \dots 0)_4 = \alpha \cdot 4^m$. Thus:

$$A(n) = 4^m \quad \text{where } 2^m \le n < 2^{m+1}$$

The Coefficients $D(n)$ and $E(n)$ These functions represent the contribution of the β terms in the base-4 expansion.

- $D(n)$ corresponds to $\beta_0 = 1, \alpha = 0, \beta_1 = 0$. It effectively sums powers of 4 for every bit in n that is 0 (excluding the leading bit).
- $E(n)$ corresponds to $\beta_1 = 1, \alpha = 0, \beta_0 = 0$. It sums powers of 4 for every bit in n that is 1 (excluding the leading bit).

Applying the Repertoire Method

To find $B(n)$ and $C(n)$, we require additional equations. We select repertoire functions to generate a linear system.

Repertoire 1: $h(n) = 1$ We require the recurrence to hold for all n. Recall that if a linear equation $Cn + D = 0$ holds for all integers n, then the coefficients must be identically zero ($C = 0$ and $D = 0$). Applying this principle:

- $1 = \alpha \implies \alpha = 1$.
- $1 = 4(1) + \gamma_0 n + \beta_0 \implies \gamma_0 n + (3 + \beta_0) = 0$. Thus $\gamma_0 = 0$ and $\beta_0 = -3$.
- $1 = 4(1) + \gamma_1 n + \beta_1 \implies \gamma_1 n + (3 + \beta_1) = 0$. Thus $\gamma_1 = 0$ and $\beta_1 = -3$.

Substituting these parameters into the ansatz yields:

$$A(n) - 3D(n) - 3E(n) = 1 \tag{2.11}$$

Repertoire 2: $h(n) = n$ Substituting $h(n) = n$:

- Base: $1 = \alpha \implies \alpha = 1$.
- Even: $2n = 4n + \gamma_0 n + \beta_0 \implies 0 = (2 + \gamma_0)n + \beta_0 \implies \gamma_0 = -2, \beta_0 = 0$.
- Odd: $2n + 1 = 4n + \gamma_1 n + \beta_1 \implies 0 = (2 + \gamma_1)n + (\beta_1 - 1) \implies \gamma_1 = -2, \beta_1 = 1$.

This yields the equation:

$$A(n) - 2B(n) - 2C(n) + E(n) = n \tag{2.12}$$

Repertoire 3: $h(n) = n^2$ Substituting $h(n) = n^2$:

- Base: $1 = \alpha \implies \alpha = 1$.
- Even: $(2n)^2 = 4n^2 + \gamma_0 n + \beta_0 \implies 0 = \gamma_0 n + \beta_0 \implies \gamma_0 = 0, \beta_0 = 0$.
- Odd: $(2n+1)^2 = 4n^2 + \gamma_1 n + \beta_1 \implies 4n^2 + 4n + 1 = 4n^2 + \gamma_1 n + \beta_1$. Equating terms gives $\gamma_1 = 4, \beta_1 = 1$.

This yields the equation:

$$A(n) + 4C(n) + E(n) = n^2 \tag{2.13}$$

System Solution

We have constructed a system of five relations.

(i) $A(n) = 4^m$
(ii) $\alpha A(n) + \beta_0 D(n) + \beta_1 E(n) = (\alpha\beta_{b_{m-1}} \dots \beta_{b_0})_4$
(iii) $A(n) - 3D(n) - 3E(n) = 1$
(iv) $A(n) - 2B(n) - 2C(n) + E(n) = n$
(v) $A(n) + 4C(n) + E(n) = n^2$

We solve for the unknown functions $C(n)$ and $B(n)$ algebraically. From (2.13), we isolate $C(n)$:

$$C(n) = \frac{n^2 - A(n) - E(n)}{4}$$

From (2.12), we isolate $B(n)$:

$$B(n) = \frac{A(n) - 2C(n) + E(n) - n}{2}$$

Substituting the expression for $C(n)$ into the equation for $B(n)$:

$$\begin{aligned} B(n) &= \frac{A(n) - \frac{2(n^2 - A(n) - E(n))}{4} + E(n) - n}{2} \\ &= \frac{2A(n) - (n^2 - A(n) - E(n)) + 2E(n) - 2n}{4} \\ &= \frac{3A(n) + 3E(n) - n^2 - 2n}{4} \end{aligned}$$

Thus, the closed-form solution is fully determined by the parameters α, β, γ and the radix-derived functions A, D, and E.

2.7 Exercises

Part I: The Geometry of Recursion

1. **The Constrained Tower of Hanoi.** Consider the Tower of Hanoi problem with n disks and three pegs: A, B, and C. Suppose that moving a disk directly between peg A and peg B is forbidden. Every move must involve peg C (e.g., A $\to$ C, C $\to$ B, etc.).

   (a) Let T_n denote the minimum number of moves required to transfer n disks from A to B. Show that $T_n = 3T_{n-1} + 2$ for $n \geq 1$.

   (b) Prove that $T_n = 3^n - 1$.

   (c) What is the minimum number of moves required to move the stack from A to C?

   Remark. Note that the symmetry between destinations is broken by the constraint.

2. **Bounded Regions.** In the *Lines in the Plane* problem, we established that n lines define $L_n = \frac{n(n+1)}{2} + 1$ regions. Some of these regions are unbounded (extending to infinity), while others are bounded (finite area).

   (a) By observing the regions "on the edge" of the arrangement, determine the number of unbounded regions created by n lines in general position.

   (b) Deduce a formula for B_n, the maximum number of bounded regions defined by n lines.

   (c) Express B_n using binomial coefficients.

3. **Planes in Space.** Let P_n be the maximum number of regions defined by n planes in three-dimensional space.

   (a) Explain why P_n satisfies the recurrence $P_n = P_{n-1} + L_{n-1}$, where L_{n-1} is the number of regions defined by $n-1$ lines in a plane.

   Remark. Consider what happens on the surface of the n-th plane.

   (b) Derive the closed-form solution for P_n.

4. **The Zig-Zag Construction.** We proved that n bent lines can define at most $Z_n = 2n^2 - n + 1$ regions. However, we did not prove that this maximum is achievable. Show that the following construction achieves Z_n regions: The j-th bent line (for $1 \leq j \leq n$) has its vertex at $(n^{2j}, 0)$ and passes through the points $(n^{2j} - n^j, 1)$ and $(n^{2j} - n^j - n^{-n}, 1)$.

   Remark. You do not need to calculate the exact intersections; rather, argue why the steepness of the lines ensures that every segment of the new bent line intersects all previous lines.

5. **Angle Constraints.** Is it possible to obtain the maximum Z_n regions with n bent lines if we require the angle at every vertex to be $30°$? Justify your answer.

6. **Map Colouring.** Prove that the regions formed by any number of lines in the plane can be coloured with just two colours such that no two regions sharing a boundary segment have the same colour.

   Remark. Use induction. When adding the n-th line, what happens to the colouring of the regions on one side of that line?

Part II: Algebraic Recurrences and the Repertoire Method

7. **The Linear Recurrence.** Solve the recurrence relation defined by:

$$Q_0 = \alpha; \quad nQ_n = (n + \beta)Q_{n-1} \text{ for } n > 0.$$

   Express your answer in terms of binomial coefficients.

8. **Non-linear Transformation.** Solve the following recurrence for Q_n:

$$Q_0 \neq 0; \quad Q_n(1 + Q_{n-1}) = Q_{n-1} \text{ for } n \geq 1.$$

   Remark. This recurrence is non-linear. Try computing the first few terms, or consider the sequence $x_n = 1/Q_n$.

9. **The Flawed Induction.** Let $H(n) = J(n+1) - J(n)$. From the relation $J(2n) = 2J(n) - 1$ and $J(2n+1) = 2J(n) + 1$, we can derive:

$$H(2n) = J(2n+1) - J(2n) = 2$$

$$H(2n+1) = J(2n+2) - J(2n+1) = (2J(n+1) - 1) - (2J(n) + 1) = 2H(n) - 2$$

   It appears possible to prove that $H(n) = 2$ for all n by induction.

   (a) Check the base cases. Does $H(1) = 2$? Does $H(2) = 2$?

(b) Calculate $H(n)$ for $n = 1, \ldots, 10$.
(c) Locate exactly where the inductive argument breaks down.

10. The Repertoire Method (Four Parameters). Use the Repertoire Method to solve the general four-parameter recurrence $g(n)$:

$$\begin{aligned} g(1) &= \alpha \\ g(2n+j) &= 3g(n) + \gamma n + \beta_j \quad \text{for } j \in \{0,1\} \text{ and } n \geq 1. \end{aligned}$$

Remark. Follow the method from the text. Use $g(n) = 1$ and $g(n) = n$ as your repertoire functions. Note that the multiplicative factor is 3, not 2, which will affect the powers in your coefficient functions.

11. The Five-Parameter Solution. Complete the solution for the five-parameter recurrence $h(n)$ introduced in the text:

$$\begin{aligned} h(1) &= \alpha \\ h(2n) &= 4h(n) + \gamma_0 n + \beta_0 \\ h(2n+1) &= 4h(n) + \gamma_1 n + \beta_1 \end{aligned}$$

Specifically, solve the system of equations derived in the text to find explicit formulas for $B(n)$ and $C(n)$. Verify your solution by calculating $h(5)$ using both the recursive definition and your closed form.

12. Cyclic Shift Generalisation. Let $n = (b_m b_{m-1} \ldots b_0)_2$. We saw that for the standard Josephus problem, $J(n) = (b_{m-1} \ldots b_0 b_m)_2$ (a 1-bit cyclic left shift). Consider a variant where $k = 2$, but the survivor is determined by a different elimination rule such that the closed form is a *2-bit* cyclic left shift:

$$f((b_m b_{m-1} \ldots b_0)_2) = (b_{m-2} \ldots b_0 b_m b_{m-1})_2$$

Determine the recurrence relations (for $f(2n)$, $f(4n+1)$, etc.) that produce this behaviour.

Part III: Advanced Theory

13. ⋆ The Josephus Permutation. Let us define the Josephus permutation P_n on the set $\{1, \ldots, n\}$ not just by the survivor, but by the *order* of elimination. For $n = 7$ and $k = 2$, the elimination order is $2, 4, 6, 1, 5, 3, 7$. Thus the survivor is 7.

(a) What is the elimination order for $n = 8$?
(b) Let $\pi_n(i)$ denote the position of the i-th person in the elimination sequence (e.g., for $n = 7$, $\pi_7(2) = 1$ because person 2 is eliminated first). Derive a recurrence relation for $\pi_n(x)$.

14. ⋆ Compact Josephus. Let $D = (b_m \ldots b_0)_2$. Prove that the general Josephus recurrence $f(n)$ with parameters α, β, γ can be computed via the following algorithm: Start with $v = \alpha$. For i from $m-1$ down to 0, set:

$$v \leftarrow \begin{cases} 2v + \beta & \text{if } b_i = 0 \\ 2v + \gamma & \text{if } b_i = 1 \end{cases}$$

Show that this algorithm is equivalent to the relaxed radix representation derived in the text.

Chapter 3

Sums

Having explored recurrence relations, we observed that solutions often take the form of summations. For instance, the solution to the Lines in the Plane problem was $L_n = 1 + \sum_{i=1}^{n} i$. To handle more complex recursive structures, we require a rigorous framework for manipulating sums. We begin by formalising the concept of a sequence.

3.1 Sequences

Intuitively, a sequence is an ordered list of objects. Mathematically, we define it strictly as a function.

Definition 3.1.1. ***Sequence***. A sequence of elements from a set A is a function $f : \mathbb{N} \to A$. The value $f(n)$ is denoted by a_n and is called the n-th term of the sequence. We denote the sequence variously as:

$$f = \{a_n\}_{n\in\mathbb{N}}, \text{ or } (a_n), \text{ or simply } \{a_n\}$$

Remark. We distinguish between the sequence, denoted by (a_n), and the set of its values (range), denoted by $\{a_n\}$. For example, if $a_n = (-1)^n$, the sequence $(a_n) = (-1, 1, -1, \dots)$ is infinite and oscillating, whereas the set of values $\{a_n\} = \{-1, 1\}$ is finite.

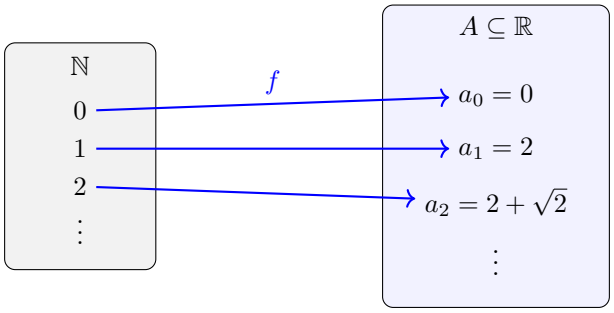


Figure 3.1: A sequence $f(n) = n + \sqrt{n}$ viewed as a mapping from $\mathbb{N}$ to $\mathbb{R}$.

Example 3.1.1. Real Sequence. Let $f : \mathbb{N} \to \mathbb{R}$ be defined by the formula $a_n = n + \sqrt{n}$. This generates the sequence:

$$0, \quad 2, \quad 2 + \sqrt{2}, \quad 3 + \sqrt{3}, \quad \dots$$

Cardinality and Indexing

A fundamental property of sequences is their domain. Since the domain is $\mathbb{N}$, a sequence is inherently countably infinite.

- **Infinite Nature**: Removing a finite subset K from $\mathbb{N}$ leaves a set $\mathbb{N} \setminus K$ that has the same cardinality as $\mathbb{N}$. That is, $|\mathbb{N}| = |\mathbb{N} \setminus K|$.
- **Re-indexing**: Because of this bijection, we can enumerate the elements of a sequence using any countably infinite subset of $\mathbb{N}$ as the index set.

Consequently, we are not bound to start indexing at 0. We may define a sequence on $\mathbb{N}^+ = \mathbb{N} \setminus \{0\}$, denoted $\{a_n\}_{n\geq 1}$, or generally $f : T \to A$ where T is any countably infinite set.

Finite Sequences

While a general sequence is infinite, practical computation often deals with finite segments.

Definition 3.1.2. ***Finite Sequence***. A finite sequence is a function $f : K \to A$ where K is a finite subset of $\mathbb{N}$ (or $\mathbb{Z}$). Typically $K = \{1, 2, \ldots, n\}$. We denote this as $\{a_k\}_{k=1}^n$, where n is the length of the sequence.

Remark. If $n = 0$, the set K is empty. This defines the *empty sequence*, denoted e.

Example 3.1.2. Restricted Domain. Let $a_n = \frac{n}{(n-2)(n-5)}$. The function is undefined at $n = 2$ and $n = 5$. A finite sequence can be constructed by restricting the domain to a set K that excludes these poles, for example, $f : \mathbb{N} \setminus \{2, 5\} \to \mathbb{R}$.

Example 3.1.3. Integer Domain. Let $T = \{-1, -2, 3, 4\}$. A function $f : T \to \mathbb{R}$ defined by a_n for $n \in T$ constitutes a finite sequence with domain T, despite the indices not being strictly natural numbers. This flexibility allows us to handle indices from the set of integers $\mathbb{Z}$ as long as the set K remains finite.

3.2 Finite Sums

We define a finite sum as the addition of consecutive terms from a sequence $\{a_n\}$ over a specified range. Let $f : \mathbb{Z}^+ \to \mathbb{Q}$ be a sequence of rational numbers.

Notation 3.2.1. Sigma Notation.

$$\sum_{k=1}^{n} a_k = a_1 + a_2 + \cdots + a_n$$

Here, k is the *index of summation*, 1 is the *lower bound*, and n is the *upper bound*. We can express the index set $K = \{1, \ldots, n\}$ explicitly in the notation:

$$\sum_{k=1}^{n} a_k = \sum_{1\leq k\leq n} a_k = \sum_{k\in K} a_k$$

This generalised form allows us to sum over arbitrary finite sets K.

Sums over Sub-sequences

We frequently require the sum of terms satisfying a specific property $P(k)$. This restricts the summation to a sub-sequence.

$$\sum_{P(k)} a_k = \sum_{k\in K} a_k \quad \text{where } K = \{n \in \mathbb{N} : P(n)\}$$

For the sum to be well-defined in this context, we assume:

1. The statement $P(n)$ is decidable (strictly True or False).
2. The set K is finite.

Example 3.2.1. Sum of Even Indices. Consider summing the even terms of $\{a_n\}$ up to $2n$. This can be expressed in two equivalent ways:

- **Property-based**: $P(k)$ is " $1 \leq k \leq 2n$ AND k is even".

$$\sum_{\substack{1\leq k\leq 2n \\ k\in \text{Even}}} a_k = a_2 + a_4 + \cdots + a_{2n}$$

- **Transformation-based**: Let $k = 2j$.

$$\sum_{j=1}^{n} a_{2j} = a_2 + a_4 + \cdots + a_{2n}$$

Example 3.2.2. Odd Numbers under 100. Let $P(n)$ be the property "$1 \leq n < 100$ and n is ODD".

$$K = \{n \in \mathbb{N} : P(n)\} = \{1, 3, 5, \ldots, 99\}$$

Using the substitution $k \to 2n + 1$, the set can be described by $0 \leq n \leq 49$.

$$\sum_{P(n)} a_n = \sum_{n=0}^{49} a_{2n+1} = a_1 + a_3 + \cdots + a_{99}$$

Example 3.2.3. Squared Terms. Let $P(n)$ be the property $1 \leq n < 100$. The sum of the terms $(2n + 1)^2$ satisfying this property is

$$\sum_{P(n)} (2n+1)^2 = \sum_{1\leq n<100} (2n+1)^2 = \sum_{n=1}^{99} (2n+1)^2 = 3^2 + 5^2 + \cdots + (2(99)+1)^2$$

3.3 The Iverson Bracket

Mathematical analysis often requires combining arithmetic with logic. To manipulate sums with complex boundary conditions algebraically, we introduce the *Iverson bracket*, derived from Kenneth Iverson's programming language APL.

Definition 3.3.1. ***Iverson Bracket***. Let $P(x)$ be a statement that is either true or false. The characteristic function $[P(x)]$ is defined as:

$$[P(x)] = \begin{cases} 1 & \text{if } P(x) \text{ is true} \\ 0 & \text{if } P(x) \text{ is false} \end{cases}$$

This function acts as a filter for the set of values where $P(x)$ holds. It allows us to rewrite a sum over a restricted set K as a sum over all integers (or naturals), with the bracket eliminating unwanted terms.

$$\sum_{k\in K} a_k = \sum_{k} a_k [P(k)] \tag{3.1}$$

where $K = \{k : P(k) \text{ is true}\}$.

Example 3.3.1. Primes. Consider the sum of the reciprocals of prime numbers $p \leq n$. The condition $P(x)$ is the intersection of two predicates: $P_1(x)$ ("x is prime") and $P_2(x)$ ("$x \leq n$").

$$\sum_{p} \frac{1}{p} [p \text{ prime}][p \leq n]$$

Here, the term is non-zero only when p satisfies both conditions simultaneously.

Note. Be careful with notation: Standard convention uses capital letters (e.g., N) for sets and lowercase (e.g., n) for variables. Some texts loosely write $n \leq N$, which implies a number is less than a set—a category error. The correct form is $n \in N$ or comparing variables $p \leq n$.

Manipulating Sums with Iverson Brackets

The Iverson bracket provides an elegant way to prove arithmetic identities involving logic.

***Theorem* 3.3.1. Absolute Value Identity**. For any $x \in \mathbb{R}$:

$$x \cdot ([x > 0] - [x < 0]) = |x|$$

Proof. We evaluate the expression based on the sign of x:

1. **Case** $x > 0$: $[x > 0] = 1$ and $[x < 0] = 0$. The expression becomes $x(1 - 0) = x$.
2. **Case** $x < 0$: $[x > 0] = 0$ and $[x < 0] = 1$. The expression becomes $x(0 - 1) = -x$.
3. **Case** $x = 0$: $[x > 0] = 0$ and $[x < 0] = 0$. The expression becomes $0(0) = 0$.

In all cases, the result matches the definition of $|x|$. ■

Empty and Vacuous Sums

The definition of summation requires careful handling of bounds. Consider the sum:

$$S = \sum_{k=4}^{0} q_k$$

The index set is $K = \{k \in \mathbb{Z} : 4 \leq k \leq 0\}$. This set is empty ($K = \emptyset$) because there are no integers satisfying this condition. By convention, a sum over an empty set is the additive identity:

$$\sum_{k \in \emptyset} a_k = 0$$

Applications of Index Sets

Understanding the index set is crucial when the summation involves composite functions.

Example 3.3.2. Sum over Squares. Evaluate $\sum_{0 \leq k^2 \leq 5} a_{k^2}$. We first determine the index set K. We require integers k such that $0 \leq k^2 \leq 5$.

$$k^2 \in \{0, 1, 4\} \implies k \in \{0, 1, -1, 2, -2\}$$

Thus, $K = \{-2, -1, 0, 1, 2\}$. The sum expands to:

$$\sum_{k \in K} a_{k^2} = a_{(-2)^2} + a_{(-1)^2} + a_{0^2} + a_{1^2} + a_{2^2}$$

Simplifying the indices:

$$= a_4 + a_1 + a_0 + a_1 + a_4 = a_0 + 2a_1 + 2a_4$$

Remark. distinct values of k map to the same term index k^2, causing terms to appear multiple times.

3.4 Sums and Recurrences

We have established the definition of sequences and finite sums. We now turn our attention to the evaluation of these sums. A fundamental observation is that any sum can be expressed as a recurrence relation. Let S_n denote the sum of the first n terms of a sequence $\{a_k\}$.

$$S_n = \sum_{k=0}^{n} a_k \tag{3.2}$$

Separating the last term yield the recurrence:

$$S_n = S_{n-1} + a_n \quad \text{for } n > 0 \tag{3.3}$$

with the boundary condition $S_0 = a_0$. (Recall that $\sum_{k=0}^{-1} a_k = 0$). Conversely, many recurrence relations can be reduced to sums. If we can solve the recurrence for S_n, we obtain a closed-form expression for the summation.

The Repertoire Method for Sums

Consider the problem of finding a closed formula for the sum of an arithmetic progression:

$$S_n = \sum_{k=0}^{n} (a + bk)$$

Expressing this as a recurrence, we define $R_n = S_n$. Then:

$$R_0 = \alpha, \quad R_n = R_{n-1} + \beta + \gamma n \tag{3.4}$$

where the term added at each step is linear in n. For the specific arithmetic sum above, $R_0 = a$ (so $\alpha = a$), and the term $a_n = a + bn$ corresponds to $\beta = a$ and $\gamma = b$.

To solve the general recurrence (3.4), we employ the Repertoire Method. Since the recurrence is linear in its parameters α, β, γ, the solution must be a linear combination of these parameters:

$$R_n = A(n)\alpha + B(n)\beta + C(n)\gamma \tag{3.5}$$

We determine the coefficient functions $A(n), B(n)$, and $C(n)$ by testing specific "repertoire" cases where the answer is known.

- **The Constant Case ($R_n = 1$).** Let $R_n = 1$ for all n. Substituting into the recurrence (3.4):

  $$1 = 1 + \beta + \gamma n \implies 0 = \beta + \gamma n$$

  For this to hold for all n, we must have $\beta = 0$ and $\gamma = 0$. The initial condition gives $R_0 = \alpha \implies \alpha = 1$. Substituting these values ($\alpha = 1, \beta = 0, \gamma = 0$) into the ansatz (3.5):

  $$1 = A(n)(1) + B(n)(0) + C(n)(0) \implies A(n) = 1$$

- **The Linear Case ($R_n = n$).** Let $R_n = n$ for all n. Substituting into the recurrence:

  $$n = (n-1) + \beta + \gamma n \implies 1 = \beta + \gamma n$$

  This implies $\beta = 1$ and $\gamma = 0$. The initial condition is $R_0 = 0 \implies \alpha = 0$. Substituting ($\alpha = 0, \beta = 1, \gamma = 0$) into the ansatz:

  $$n = A(n)(0) + B(n)(1) + C(n)(0) \implies B(n) = n$$

- **The Quadratic Case ($R_n = n^2$).** Let $R_n = n^2$. Substituting into the recurrence:

  $$\begin{aligned} n^2 &= (n-1)^2 + \beta + \gamma n \\ n^2 &= n^2 - 2n + 1 + \beta + \gamma n \\ 0 &= (1 + \beta) + n(\gamma - 2) \end{aligned}$$

  Equating coefficients yields $\beta = -1$ and $\gamma = 2$. The initial condition is $R_0 = 0^2 = 0 \implies \alpha = 0$. Substituting ($\alpha = 0, \beta = -1, \gamma = 2$) into the ansatz:

  $$n^2 = A(n)(0) + B(n)(-1) + C(n)(2) \implies 2C(n) - B(n) = n^2$$

  Using the known value $B(n) = n$, we solve for $C(n)$:

  $$2C(n) - n = n^2 \implies C(n) = \frac{n^2 + n}{2} = \frac{n(n+1)}{2}$$

***Theorem* 3.4.1. General Solution for Linear Sum Recurrences**. The solution to the recurrence $R_n = R_{n-1} + \beta + \gamma n$ with $R_0 = \alpha$ is:

$$R_n = \alpha + n\beta + \frac{n(n+1)}{2}\gamma$$

Applying this to our original arithmetic sum $S_n = \sum_{k=0}^{n}(a + bk)$, where $\alpha = a, \beta = a, \gamma = b$:

$$S_n = a + na + \frac{n(n+1)}{2}b = (n+1)a + \frac{n(n+1)}{2}b$$

This confirms the standard result derived via simple summation laws: $\sum a + b \sum k$.

Summation Laws

To manipulate sums efficiently without expanding them into recurrences, we rely on three fundamental algebraic laws. Let K be a finite index set.

- **Distributive Law**: Constants can be factored out of the summation.

$$\sum_{k \in K} ca_k = c \sum_{k \in K} a_k$$

- **Associative Law**: A sum of terms can be split into separate sums.

$$\sum_{k \in K} (a_k + b_k) = \sum_{k \in K} a_k + \sum_{k \in K} b_k$$

- **Commutative Law**: The order of summation does not affect the result. If $Q : K \to K$ is a permutation (a bijection of K onto itself), then:

$$\sum_{k \in K} a_k = \sum_{k \in K} a_{Q(k)}$$

Manipulating Domains

The associative law applies to splitting the summand. We also require rules for splitting the *index set* (the domain). For any two sets of predicates defining domains K and K', the summation over the union can be decomposed using the Inclusion-Exclusion Principle.

$$\sum_{k \in K} a_k + \sum_{k \in K'} a_k = \sum_{k \in K \cap K'} a_k + \sum_{k \in K \cup K'} a_k \tag{3.6}$$

Proof. Let $P(k)$ and $Q(k)$ be the predicates defining sets K and K' respectively. Using Iverson brackets, the identity relies on the property:

$$[P(k)] + [Q(k)] = [P(k) \vee Q(k)] + [P(k) \wedge Q(k)]$$

Multiplying by a_k and summing over all k yields the result. This follows directly from the fact that $|A|+|B| = |A \cup B| + |A \cap B|$ for finite sets. ■

Geometric Sums

A sequence $\{a_n\}$ is geometric if the ratio of consecutive terms is constant, i.e., $a_{n+1}/a_n = q$ for all n. The general term is $a_n = a_0 q^n$. The sum of a geometric sequence is given by:

$$S_n = \sum_{k=0}^{n} a_0 q^k$$

***Theorem* 3.4.2. Geometric Sum Formula**. For $q \neq 1$:

$$\sum_{k=0}^{n} a_0 q^k = a_0 \left(\frac{1 - q^{n+1}}{1 - q} \right)$$

Proof. Multiply the sum S_n by q:

$$\begin{aligned} S_n &= a_0 + a_0 q + \cdots + a_0 q^n \\ qS_n &= \quad a_0 q + \cdots + a_0 q^n + a_0 q^{n+1} \end{aligned}$$

Subtracting the second equation from the first eliminates the intermediate terms (a telescoping effect):

$$S_n(1 - q) = a_0 - a_0 q^{n+1}$$

Dividing by $(1 - q)$ yields the formula. ■

Example 3.4.1. Powers of 2. Evaluate $S_n = \sum_{k=0}^{n} 2^{-k}$. Here $a_0 = 1$ and $q = 1/2$.

$$S_n = \frac{1 - (1/2)^{n+1}}{1 - 1/2} = \frac{1 - (1/2)^{n+1}}{1/2} = 2 \left(1 - \frac{1}{2^{n+1}} \right) = 2 - \frac{1}{2^n}$$

The Summation Factor Method

We can often reduce a linear recurrence relation to a summation by introducing a summation factor. This technique is particularly useful when the recurrence is of the form $a_n T_n = b_n T_{n-1} + c_n$.

Consider the Tower of Hanoi recurrence derived in Chapter 1:

$$T_0 = 0, \quad T_n = 2T_{n-1} + 1$$

We seek to eliminate the coefficient 2. We divide the recurrence by 2^n:

$$\frac{T_n}{2^n} = \frac{2T_{n-1}}{2^n} + \frac{1}{2^n} = \frac{T_{n-1}}{2^{n-1}} + \frac{1}{2^n}$$

Let $S_n = \frac{T_n}{2^n}$. The recurrence simplifies to a sum:

$$S_0 = 0, \quad S_n = S_{n-1} + 2^{-n}$$

Unfolding the recurrence:

$$S_n = \sum_{k=1}^{n} 2^{-k}$$

Using the geometric sum formula (shifted to start at $k = 1$):

$$S_n = \left(2 - \frac{1}{2^n} \right) - 1 = 1 - \frac{1}{2^n}$$

Finally, we recover T_n by inverting the substitution $T_n = 2^n S_n$:

$$T_n = 2^n \left(1 - \frac{1}{2^n} \right) = 2^n - 1$$

This recovers the closed form solution $T_n = 2^n - 1$, demonstrating the power of reducing recurrences to sums.

3.5 The Summation Factor Method Explained

In our analysis of the Tower of Hanoi, we simplified the recurrence $T_n = 2T_{n-1} + 1$ by dividing by 2^n. This was not an arbitrary trick; it is an application of a general technique for solving linear first-order recurrences of the form:

$$a_n T_n = b_n T_{n-1} + c_n \quad \text{for } n \geq 1 \tag{3.7}$$

Here, $\{a_n\}, \{b_n\}$, and $\{c_n\}$ are arbitrary sequences, and T_0 is given as an initial condition. Our goal is to reduce this recurrence to a simple summation by introducing a summation factor s_n.

Deriving the Technique

We multiply the original recurrence (3.7) by a factor s_n (where $s_n \neq 0$):

$$s_n a_n T_n = s_n b_n T_{n-1} + s_n c_n$$

We wish to choose s_n such that the term involving T_{n-1} matches the form of the term on the left-hand side, but indexed at $n-1$. Specifically, we impose the condition:

$$s_n b_n = s_{n-1} a_{n-1} \tag{3.8}$$

Under this condition, the recurrence becomes:

$$s_n a_n T_n = s_{n-1} a_{n-1} T_{n-1} + s_n c_n$$

Let us define a new variable $S_n = s_n a_n T_n$. The relation simplifies to:

$$S_n = S_{n-1} + s_n c_n$$

This is a standard summation recurrence. Unfolding it yields:

$$\begin{aligned} S_n &= S_0 + \sum_{k=1}^{n} s_k c_k \\ &= s_0 a_0 T_0 + \sum_{k=1}^{n} s_k c_k \end{aligned}$$

Using the defining property $s_0 a_0 = s_1 b_1$, we can rewrite S_0 as $s_1 b_1 T_0$. Thus:

$$S_n = s_1 b_1 T_0 + \sum_{k=1}^{n} s_k c_k$$

To recover T_n, we invert the substitution $S_n = s_n a_n T_n$:

$$T_n = \frac{1}{s_n a_n} \left(s_1 b_1 T_0 + \sum_{k=1}^{n} s_k c_k \right) \tag{3.9}$$

Determining the Summation Factor

We must now solve the auxiliary recurrence $s_n = s_{n-1} \frac{a_{n-1}}{b_n}$ to find an explicit formula for s_n. Calculating the first few terms:

$$\begin{aligned} s_2 &= s_1 \frac{a_1}{b_2} \\ s_3 &= s_2 \frac{a_2}{b_3} = s_1 \frac{a_1 a_2}{b_2 b_3} \\ s_4 &= s_3 \frac{a_3}{b_4} = s_1 \frac{a_1 a_2 a_3}{b_2 b_3 b_4} \end{aligned}$$

By induction, the general form is:

$$s_n = s_1 \frac{\prod_{k=1}^{n-1} a_k}{\prod_{k=2}^{n} b_k} \tag{3.10}$$

The constant s_1 can be chosen arbitrarily (e.g., $s_1 = 1$) provided it is non-zero.

***Theorem* 3.5.1. General Solution for Linear First-Order Recurrences**. Any recurrence of the form $a_n T_n = b_n T_{n-1} + c_n$ has the closed-form solution:

$$T_n = \frac{1}{s_n a_n} \left(s_1 b_1 T_0 + \sum_{k=1}^{n} s_k c_k \right)$$

where the summation factor is given by $s_n = \frac{a_{n-1}}{b_n} s_{n-1}$.

Application: Tower of Hanoi

Let us revisit $T_n = 2T_{n-1} + 1$ with $T_0 = 0$. Comparing this to the general form $a_n T_n = b_n T_{n-1} + c_n$, we identify:

$$a_n = 1, \quad b_n = 2, \quad c_n = 1$$

We compute the summation factor s_n. Using the recursive definition $s_n = s_{n-1} \frac{a_{n-1}}{b_n}$:

$$s_n = s_{n-1} \frac{1}{2}$$

This is a geometric progression. If we choose $s_1 = \frac{1}{2}$ (consistent with the pattern $s_n = 2^{-n}$ starting from $n = 1$), then:

$$s_n = \frac{1}{2^n}$$

Substituting these values into the general solution formula (3.9):

$$\begin{aligned} T_n &= \frac{1}{(1)(2^{-n})} \left(s_1 (2)(0) + \sum_{k=1}^{n} \frac{1}{2^k} (1) \right) \\ &= 2^n \sum_{k=1}^{n} \left(\frac{1}{2} \right)^k \end{aligned}$$

The sum is a standard geometric series $\sum_{k=1}^{n} q^k$ with $q = 1/2$.

$$\sum_{k=1}^{n} \left(\frac{1}{2} \right)^k = \frac{1/2(1 - (1/2)^n)}{1 - 1/2} = 1 - \frac{1}{2^n}$$

Therefore:

$$T_n = 2^n \left(1 - \frac{1}{2^n} \right) = 2^n - 1$$

This confirms our previous result using a systematic method applicable to any linear recurrence.

3.6 Exercises

1. **Iverson Brackets and Bounds.** Evaluate the sum

$$\sum_{k} [1 \le j \le k \le n]$$

as a function of j and n.

Remark. Treat the Iverson bracket as a filter for the index set of k. What are the effective lower and upper limits imposed by the inequality?

2. **Repertoire Method for Alternating Sums.** Use the Repertoire Method to find a closed form for the alternating sum of squares:
$$S_n = \sum_{k=0}^{n} (-1)^k k^2$$
   (a) Express S_n as a recurrence relation involving S_{n-1}.
   (b) Assume a solution of the form $S_n = (-1)^n(An^2 + Bn + C)$. Substitute this into the recurrence to determine the constants A, B, and C.

3. **Double Summation.** Evaluate the sum
$$\sum_{k=1}^{n} k2^k$$
   by rewriting it as a double sum:
$$\sum_{1 \le j \le k \le n} 2^k$$
   Remark. Interchange the order of summation. Sum over k first (geometric series), then sum the result over j.

4. **The Summation Factor.** Use a summation factor to solve the recurrence:
$$\begin{aligned} T_0 &= 5 \\ 2T_n &= nT_{n-1} + 3n! \quad \text{for } n > 0 \end{aligned}$$
   Remark. Rewrite the recurrence in the form $a_n T_n = b_n T_{n-1} + c_n$ and calculate $s_n = s_{n-1} \frac{a_{n-1}}{b_n}$.

5. **The Perturbation Method.** The perturbation method exploits the identity $S_{n+1} = S_n + a_{n+1} = a_0 + \sum_{k=0}^{n} a_{k+1}$. Use this method to evaluate the following sums in order (assume $n \ge 0$):
   (a) $S_n = \sum_{k=0}^{n} (-1)^{n-k}$
   (b) $T_n = \sum_{k=0}^{n} (-1)^{n-k} k$
   (c) $U_n = \sum_{k=0}^{n} (-1)^{n-k} k^2$

6. **Sum of Cubes via Squares.** We wish to evaluate $\square_n = \sum_{k=1}^{n} k^3$.
   (a) First, prove the identity:
$$\left(\sum_{k=1}^{n} k \right)^2 + \sum_{k=1}^{n} k^2 = 2 \sum_{1 \le j \le k \le n} jk$$
   (b) Apply the standard formula for $\sum_{j=1}^{k} j$ to the right-hand side.
   (c) Deduce the closed form for $\sum_{k=1}^{n} k^3$.

7. ⋆ **Backward Induction.** Sometimes it is possible to prove a statement $P(n)$ by showing that it holds for an infinite sequence of integers (like powers of 2) and that $P(n) \implies P(n-1)$. Consider the Arithmetic Mean-Geometric Mean (AM-GM) inequality:
$$P(n): \quad x_1 \dots x_n \le \left(\frac{x_1 + \cdots + x_n}{n} \right)^n \quad \text{for } x_i \ge 0.$$
   (a) Prove $P(2)$ directly.
   (b) Prove that $P(n)$ and $P(2)$ imply $P(2n)$. This establishes $P(n)$ for all $n = 2^k$.
   (c) Prove that $P(n)$ implies $P(n-1)$.

   Remark. Set x_n to be the arithmetic mean of the first $n-1$ terms.

   (d) Explain why these steps are sufficient to prove $P(n)$ for all $n \in \mathbb{Z}^+$.

Chapter 4

Multiple Sums

Just as sequences can be summed to form single sums, arrays and matrices of values lead naturally to double sums. A double sum involves adding terms indexed by two variables, typically i and j.

4.1 Definition and General Properties

Consider a finite collection of terms $a_i b_j$ where $1 \leq i, j \leq 3$. The sum of all such terms can be written as:

$$\sum_{1 \leq i,j \leq 3} a_i b_j = a_1 b_1 + a_1 b_2 + a_1 b_3 + a_2 b_1 + \cdots + a_3 b_3$$

This expression represents the sum of all elements in a 3×3 matrix where the (i, j)-th entry is $a_i b_j$.

Definition 4.1.1. ***Double Sum***. For index sets I and J, the double sum over $I \times J$ is defined as the iterated sum:

$$\sum_{i \in I, j \in J} a_{i,j} = \sum_{i \in I} \left(\sum_{j \in J} a_{i,j} \right) = \sum_{j \in J} \left(\sum_{i \in I} a_{i,j} \right) \tag{4.1}$$

We often write $\sum_{i,j}$ as shorthand for $\sum_{i \in I, j \in J}$.

The equality of the two iterated forms (summing rows then columns vs. columns then rows) is a consequence of the commutativity and associativity of addition for finite sums.

The General Distributive Law

A powerful property of double sums arises when the summand is separable, meaning it can be written as a product of a term depending only on i and a term depending only on j.

Example 4.1.1. Factorising a 3×3 Sum. Evaluate $S = \sum_{1 \leq i,j \leq 3} a_i b_j$.

$$\begin{aligned} S &= a_1 b_1 + a_1 b_2 + a_1 b_3 \\ &\quad + a_2 b_1 + a_2 b_2 + a_2 b_3 \\ &\quad + a_3 b_1 + a_3 b_2 + a_3 b_3 \\ &= a_1(b_1 + b_2 + b_3) + a_2(b_1 + b_2 + b_3) + a_3(b_1 + b_2 + b_3) \\ &= (a_1 + a_2 + a_3)(b_1 + b_2 + b_3) \end{aligned}$$

Thus, the double sum factors into the product of two single sums.

We generalise this observation to arbitrary index sets defined by predicates.

***Theorem* 4.1.1. General Distributive Law**. Let $P(i)$ and $Q(j)$ be predicates defining finite index sets $I = \{i : P(i)\}$ and $J = \{j : Q(j)\}$. Then:

$$\sum_{i \in I, j \in J} a_i b_j = \left(\sum_{i \in I} a_i \right) \left(\sum_{j \in J} b_j \right) \tag{4.2}$$

Proof. Let the domain of summation be defined by the joint predicate $R(i,j) = P(i) \wedge Q(j)$. The characteristic function of a conjunction satisfies $[P(i) \wedge Q(j)] = [P(i)][Q(j)]$. We rewrite the sum using Iverson brackets:

$$\begin{aligned}
\sum_{i,j} a_i b_j [P(i)][Q(j)] &= \sum_i \left(\sum_j a_i b_j [P(i)][Q(j)] \right) \\
&= \sum_i \left(a_i [P(i)] \sum_j b_j [Q(j)] \right) \quad \text{since } a_i[P(i)] \text{ is constant w.r.t } j \\
&= \left(\sum_j b_j [Q(j)] \right) \left(\sum_i a_i [P(i)] \right) \quad \text{factoring out the constant sum}
\end{aligned}$$

This equates to $(\sum_{i \in I} a_i)(\sum_{j \in J} b_j)$. ■

Symmetry in Double Sums

Often, we wish to sum elements over a restricted region, such as the upper triangular part of a matrix. Consider a square array of terms $a_i a_j$ for $1 \leq i, j \leq n$:

$$A = \begin{pmatrix} a_1 a_1 & a_1 a_2 & \dots & a_1 a_n \\ a_2 a_1 & a_2 a_2 & \dots & a_2 a_n \\ \vdots & \vdots & \ddots & \vdots \\ a_n a_1 & a_n a_2 & \dots & a_n a_n \end{pmatrix}$$

We define the sum of the upper triangle (including the diagonal) as:

$$S_\nabla = \sum_{1 \leq i \leq j \leq n} a_i a_j$$

And the sum of the lower triangle (including the diagonal) as:

$$S_\Delta = \sum_{1 \leq j \leq i \leq n} a_i a_j$$

***Lemma* 4.1.1.** *Symmetry of Triangular Sums.* For any sequence $\{a_k\}$, the sum over the upper triangle equals the sum over the lower triangle:

$$\sum_{1 \leq i \leq j \leq n} a_i a_j = \sum_{1 \leq j \leq i \leq n} a_i a_j$$

Proof. Let $P(i,j)$ be the predicate $1 \leq i, j \leq n$. The upper sum is defined by the condition $P(i,j) \wedge (i \leq j)$.

$$S_\nabla = \sum_{P(i,j)} a_i a_j [i \leq j]$$

We perform a change of variables: swap the names of indices i and j.

$$S_\nabla = \sum_{P(j,i)} a_j a_i [j \le i]$$

Since $P(i,j)$ is symmetric (i.e., $P(j,i) = P(i,j)$) and multiplication is commutative ($a_j a_i = a_i a_j$):

$$S_\nabla = \sum_{P(i,j)} a_i a_j [j \le i] = S_\Delta$$

Thus, the two sums are identical. ■

This symmetry allows us to compute sums over triangular regions by relating them to the sum over the entire square, a technique we will explore in the following sections.

4.2 Manipulation of Double Sums

We have established the definitions and distributive properties of double sums. We now apply these tools to derive non-trivial identities and inequalities. A recurring theme is the exploitation of symmetry in the index set to simplify summation.

Symmetry and Simplification

Let us return to the problem of summing over a triangular domain. We previously defined:

$$S_5 = \sum_{1 \le i \le j \le n} a_i a_j$$
$$S_4 = \sum_{1 \le j \le i \le n} a_i a_j$$

We proved by symmetry that $S_5 = S_4$. We now seek an explicit formula for S_5 in terms of single sums.

Consider the sum over the entire square domain $1 \le i, j \le n$. Using the general distributive law:

$$S_\square = \sum_{1 \le i,j \le n} a_i a_j = \left(\sum_{i=1}^{n} a_i \right) \left(\sum_{j=1}^{n} a_j \right) = \left(\sum_{k=1}^{n} a_k \right)^2$$

We can partition the square domain into three disjoint sets: the upper triangle (strictly above diagonal), the lower triangle (strictly below), and the diagonal itself ($i = j$). However, it is often more convenient to work with the non-strict inequalities ($i \le j$ and $j \le i$).

$$2S_5 = S_5 + S_4 = \sum_{1 \le i \le j \le n} a_i a_j + \sum_{1 \le j \le i \le n} a_i a_j$$

We apply the property for combining summation domains (which follows from the properties of Iverson brackets or set cardinality):

$$\sum_{k \in Q} x_k + \sum_{k \in R} x_k = \sum_{k \in Q \cup R} x_k + \sum_{k \in Q \cap R} x_k$$

Let Q be the condition $i \le j$ and R be $j \le i$ (within the bounds $1 \le i, j \le n$).

- Intersection ($Q \cap R$): $i \le j$ and $j \le i$ implies $i = j$.
- Union ($Q \cup R$): $i \le j$ or $j \le i$ is always true for any pair of real numbers. Thus, the union covers the entire square $1 \le i, j \le n$.

Substituting these back into the equation:

$$\begin{aligned} 2S_5 &= \sum_{1\le i,j\le n} a_i a_j + \sum_{1\le i=j\le n} a_i a_j \\ &= \left(\sum_{k=1}^{n} a_k\right)^2 + \sum_{k=1}^{n} a_k^2 \end{aligned}$$

Solving for S_5:

$$\sum_{1\le i\le j\le n} a_i a_j = \frac{1}{2}\left[\left(\sum_{k=1}^{n} a_k\right)^2 + \sum_{k=1}^{n} a_k^2\right] \tag{4.3}$$

This identity relates the sum of products in the upper triangle to the square of the sum and the sum of squares.

Deriving Chebyshev's Inequalities

Example 4.2.1. Lagrange's Identity. We now tackle a more complex problem involving two sequences $\{a_n\}$ and $\{b_n\}$. Consider the sum:

$$S = \sum_{1\le j<k\le n} (a_k - a_j)(b_k - b_j) \tag{4.4}$$

Our goal is to simplify this expression and derive an inequality.

Symmetrisation Technique

Let $P(j,k)$ denote the bounds $1 \le j,k \le n$.

$$S = \sum_{P(j,k),j<k} (a_k - a_j)(b_k - b_j)$$

We swap the indices j and k. Since $P(j,k)$ is symmetric ($P(k,j) = P(j,k)$), the sum becomes:

$$S = \sum_{P(j,k),k<j} (a_j - a_k)(b_j - b_k)$$

Notice that $(a_j - a_k)(b_j - b_k) = (-(a_k - a_j))(-(b_k - b_j)) = (a_k - a_j)(b_k - b_j)$. The term is invariant under the swap. Thus, we can write:

$$2S = \sum_{j<k} (a_k - a_j)(b_k - b_j) + \sum_{k<j} (a_k - a_j)(b_k - b_j)$$

We can relax the strict inequalities to $\le$ because when $j = k$, the term is $(a_k - a_k)(b_k - b_k) = 0$.

$$2S = \sum_{j\le k} \text{term} + \sum_{k\le j} \text{term}$$

Using the domain combination formula again ($Q \cup R$ is the whole square, $Q \cap R$ is the diagonal where terms are 0):

$$2S = \sum_{1\le j,k\le n} (a_k - a_j)(b_k - b_j)$$

Expansion and Simplification

We expand the product in the summand:

$$(a_k - a_j)(b_k - b_j) = a_k b_k - a_j b_k - a_k b_j + a_j b_j$$

Consequently, the double sum splits into four distinct components:

$$2S = \sum_{1\le j,k\le n} a_k b_k - \sum_{1\le j,k\le n} a_j b_k - \sum_{1\le j,k\le n} a_k b_j + \sum_{1\le j,k\le n} a_j b_j$$

We evaluate each component independently:

1. $\displaystyle\sum_{j,k} a_k b_k = \sum_{k=1}^{n}\left(\sum_{j=1}^{n} 1\right) a_k b_k = n\sum_{k=1}^{n} a_k b_k$

2. $\displaystyle\sum_{j,k} a_j b_j = \sum_{j=1}^{n}\left(\sum_{k=1}^{n} 1\right) a_j b_j = n\sum_{j=1}^{n} a_j b_j$

3. $\displaystyle\sum_{j,k} a_j b_k = \left(\sum_{j=1}^{n} a_j\right)\left(\sum_{k=1}^{n} b_k\right)$ (by General Distributive Law)

4. $\displaystyle\sum_{j,k} a_k b_j = \left(\sum_{k=1}^{n} a_k\right)\left(\sum_{j=1}^{n} b_j\right)$

Substituting these results back into the expression for $2S$:

$$2S = 2n\sum_{k=1}^{n} a_k b_k - 2\left(\sum_{k=1}^{n} a_k\right)\left(\sum_{k=1}^{n} b_k\right)$$

Dividing by 2 yields Lagrange's Identity:

$$\sum_{1\le j<k\le n} (a_k - a_j)(b_k - b_j) = n\sum_{k=1}^{n} a_k b_k - \left(\sum_{k=1}^{n} a_k\right)\left(\sum_{k=1}^{n} b_k\right) \tag{4.5}$$

Chebyshev's Sum Inequalities

The identity derived above allows us to determine the relationship between the sum of products and the product of sums based on the monotonicity of the sequences.

Case 1: Similarly Sorted Sequences Assume $a_1 \le a_2 \le \cdots \le a_n$ and $b_1 \le b_2 \le \cdots \le b_n$. For any $j < k$, we have $a_j \le a_k$ (so $a_k - a_j \ge 0$) and $b_j \le b_k$ (so $b_k - b_j \ge 0$). The product $(a_k - a_j)(b_k - b_j) \ge 0$. Since every term in the sum S is non-negative, $S \ge 0$.

$$n\sum a_k b_k - \left(\sum a_k\right)\left(\sum b_k\right) \ge 0$$

***Theorem* 4.2.1. Chebyshev's Inequality (Sorted)**. If $\{a_k\}$ and $\{b_k\}$ are both non-decreasing (or both non-increasing), then:

$$\left(\sum_{k=1}^{n} a_k\right)\left(\sum_{k=1}^{n} b_k\right) \le n\sum_{k=1}^{n} a_k b_k$$

Proof. Assume both sequences are non-decreasing:

$$a_1 \le a_2 \le \cdots \le a_n, \quad b_1 \le b_2 \le \cdots \le b_n$$

For any pair of indices $1 \leq j < k \leq n$, we have $a_k - a_j \geq 0$ and $b_k - b_j \geq 0$. Consequently, their product is non-negative:

$$(a_k - a_j)(b_k - b_j) \geq 0$$

Since every term in the sum $S = \sum_{1 \leq j < k \leq n} (a_k - a_j)(b_k - b_j)$ is non-negative, we must have $S \geq 0$. Substituting this inequality into Lagrange's Identity (4.4):

$$0 \leq n \sum_{k=1}^{n} a_k b_k - \left(\sum_{k=1}^{n} a_k \right) \left(\sum_{k=1}^{n} b_k \right)$$

Rearranging the terms yields the claim. ■

Note. Equality holds if and only if $(a_k - a_j)(b_k - b_j) = 0$ for all $j < k$. This occurs if one of the sequences is constant.

Case 2: Oppositely Sorted Sequences Assume $a_1 \leq a_2 \leq \cdots \leq a_n$ and $b_1 \geq b_2 \geq \cdots \geq b_n$. For any $j < k$, $a_k - a_j \geq 0$ but $b_k - b_j \leq 0$. The product term is ≤ 0, so $S \leq 0$.

***Theorem* 4.2.2. Chebyshev's Inequality (Opposed).** If $\{a_k\}$ is non-decreasing and $\{b_k\}$ is non-increasing, then:

$$\left(\sum_{k=1}^{n} a_k \right) \left(\sum_{k=1}^{n} b_k \right) \geq n \sum_{k=1}^{n} a_k b_k$$

Proof. Assume $a_1 \leq \cdots \leq a_n$ and $b_1 \geq \cdots \geq b_n$. For any $j < k$, we have $a_k - a_j \geq 0$ but $b_k - b_j \leq 0$. Thus:

$$(a_k - a_j)(b_k - b_j) \leq 0$$

This implies the sum $S \leq 0$. Substituting into Lagrange's Identity:

$$0 \geq n \sum_{k=1}^{n} a_k b_k - \left(\sum_{k=1}^{n} a_k \right) \left(\sum_{k=1}^{n} b_k \right)$$

Rearranging gives the inequality. ■

4.3 Harmonic Sums and Bounds

We now turn our attention to a classic problem involving harmonic numbers, demonstrating how different manipulations of the summation domain can lead to varying forms of the solution, ultimately revealing a closed-form identity.

Example 4.3.1. Reciprocal Differences. Evaluate the sum:

$$S_n = \sum_{1 \leq j < k \leq n} \frac{1}{k - j} \tag{4.6}$$

Approach 1: Fixing the Upper Bound

We first express the double sum as an iterated sum. The condition $1 \leq j < k \leq n$ implies that k ranges from 2 to n, and for each k, j ranges from 1 to $k - 1$.

$$S_n = \sum_{k=2}^{n} \sum_{j=1}^{k-1} \frac{1}{k - j}$$

Let us substitute the index in the inner sum. Let $m = k - j$. As j goes from 1 to $k - 1$:

- When $j = 1, m = k - 1$.
- When $j = k - 1, m = 1$.

Thus, the inner sum becomes $\sum_{m=1}^{k-1} \frac{1}{m}$. This is precisely the definition of the harmonic number H_{k-1}.

$$S_n = \sum_{k=2}^{n} H_{k-1} \tag{4.7}$$

Shifting the index by letting $i = k - 1$ (so i ranges from 1 to $n - 1$):

$$S_n = \sum_{i=1}^{n-1} H_i \tag{4.8}$$

This gives us S_n as a sum of harmonic numbers. While useful, it is not a closed form in terms of n and H_n.

Approach 2: Fixing the Lower Bound

Alternatively, we can iterate over j first. The condition $1 \leq j < k \leq n$ implies j ranges from 1 to $n - 1$, and for each j, k ranges from $j + 1$ to n.

$$S_n = \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} \frac{1}{k - j}$$

Let $m = k - j$.

- When $k = j + 1, m = 1$.
- When $k = n, m = n - j$.

The inner sum becomes $\sum_{m=1}^{n-j} \frac{1}{m} = H_{n-j}$.

$$S_n = \sum_{j=1}^{n-1} H_{n-j}$$

Reversing the order of summation by letting $i = n - j$ (as j goes $1 \to n - 1$, i goes $n - 1 \to 1$):

$$S_n = \sum_{i=1}^{n-1} H_i$$

This confirms the result from Approach 1 but does not provide a new form.

Approach 3: Change of Variables for Closed Form

To find a closed form solution $S_n = f(n, H_n)$, we perform a change of variables on the original double sum indices. Let $k' = k - j$ and $j' = j$. Then $k = k' + j'$. The condition $1 \leq j < k \leq n$ transforms as follows:

$$1 \leq j' < k' + j' \leq n \implies 1 \leq j' \leq n - k'$$

Also, since $k > j$, $k' \geq 1$. The maximum value for k' occurs when $j' = 1$ and $k = n$, so $k' \leq n - 1$. Rewriting the sum in terms of j' and k':

$$S_n = \sum_{k'=1}^{n-1} \sum_{j'=1}^{n-k'} \frac{1}{k'}$$

Notice that the term $\frac{1}{k'}$ is independent of j'. Thus, the inner sum is simply adding a constant value $(n-k')$ times.

$$S_n = \sum_{k'=1}^{n-1} \frac{1}{k'}(n-k')$$

Splitting the sum:

$$\begin{aligned} S_n &= \sum_{k'=1}^{n-1} \left(\frac{n}{k'} - 1\right) \\ &= n\sum_{k'=1}^{n-1} \frac{1}{k'} - \sum_{k'=1}^{n-1} 1 \\ &= nH_{n-1} - (n-1) \end{aligned}$$

Using the identity $H_{n-1} = H_n - \frac{1}{n}$:

$$\begin{aligned} S_n &= n\left(H_n - \frac{1}{n}\right) - n + 1 \\ &= nH_n - 1 - n + 1 \\ &= nH_n - n \end{aligned}$$

***Theorem* 4.3.1. Sum of Harmonic Numbers**. Combining the results from Approach 1 and Approach 3 yields a fundamental identity for the sum of the first n harmonic numbers:

$$\sum_{k=1}^{n} H_k = (n+1)H_n - n \tag{4.9}$$

Proof. From Approach 1, $S_{n+1} = \sum_{k=1}^{n} H_k$. From Approach 3, replacing n with $n+1$:

$$S_{n+1} = (n+1)H_{n+1} - (n+1)$$

Using $H_{n+1} = H_n + \frac{1}{n+1}$:

$$\begin{aligned} S_{n+1} &= (n+1)\left(H_n + \frac{1}{n+1}\right) - (n+1) \\ &= (n+1)H_n + 1 - n - 1 \\ &= (n+1)H_n - n \end{aligned}$$

■

4.4 Sum of Squares: Five Methods

The summation of squares, denoted $\square_n = \sum_{k=0}^{n} k^2$, is a classical problem that can be solved using a variety of techniques. We present five distinct methods to derive the closed-form solution. Each method illustrates a different strategy applicable to summation problems in general.

Method 1: Mathematical Induction

The most direct method is to guess the formula and verify it. **Hypothesis**: $\square_n = \frac{n(n+1)(2n+1)}{6}$. We rewrite the formula for algebraic convenience as:

$$\square_n = \frac{n(n+1/2)(n+1)}{3}$$

Proof. **Base Case** ($n = 0$): $\sum_{k=0}^{0} k^2 = 0$. The formula gives $\frac{0(1/2)(1)}{3} = 0$.

Inductive Step: Assume the formula holds for $n - 1$.

$$\square_{n-1} = \frac{(n-1)(n-1/2)n}{3}$$

We evaluate $\square_n = \square_{n-1} + n^2$:

$$\begin{aligned}
\square_n &= \frac{n(n-1)(n-1/2)}{3} + n^2 \\
&= n\left[\frac{(n-1)(n-1/2)}{3} + n\right] \\
&= n\left[\frac{n^2 - \frac{3}{2}n + \frac{1}{2} + 3n}{3}\right] \\
&= n\left[\frac{n^2 + \frac{3}{2}n + \frac{1}{2}}{3}\right] \\
&= \frac{n(n+1/2)(n+1)}{3}
\end{aligned}$$

This matches the hypothesis. Thus, the formula is valid for all $n \geq 0$. ■

Method 2: The Perturbation Method

Often we do not have a candidate formula to verify. The Perturbation Method constructs a solution algebraically. Let $S_n = \sum_{k=0}^{n} a_k$. We express S_{n+1} in two ways: by separating the first term and by separating the last term.

$$S_n + a_{n+1} = a_0 + \sum_{k=0}^{n} a_{k+1}$$

We attempt to express the sum $\sum a_{k+1}$ in terms of S_n.

First Attempt: Perturbing $\sum k^2$ Let $S_n = \square_n = \sum_{k=0}^{n} k^2$.

$$\begin{aligned}
S_n + (n+1)^2 &= 0^2 + \sum_{k=0}^{n} (k+1)^2 \\
&= \sum_{k=0}^{n} (k^2 + 2k + 1) \\
&= S_n + 2\sum_{k=0}^{n} k + \sum_{k=0}^{n} 1
\end{aligned}$$

The S_n terms on both sides cancel out. This fails to solve for S_n, but it serendipitously yields the formula for the sum of the first n integers:

$$(n+1)^2 = 2\sum_{k=0}^{n} k + (n+1) \implies \sum_{k=0}^{n} k = \frac{n(n+1)}{2}$$

Second Attempt: Perturbing $\sum k^3$ To find the sum of squares, we must perturb the sum of cubes. Let $C_n = \sum_{k=0}^{n} k^3$.

$$\begin{aligned} C_n + (n+1)^3 &= \sum_{k=0}^{n} (k+1)^3 \\ &= \sum_{k=0}^{n} (k^3 + 3k^2 + 3k + 1) \\ &= C_n + 3\square_n + 3\sum_{k=0}^{n} k + \sum_{k=0}^{n} 1 \end{aligned}$$

The C_n terms cancel, leaving an equation we can solve for $\square_n$:

$$(n+1)^3 = 3\square_n + 3\frac{n(n+1)}{2} + (n+1)$$

Solving for $\square_n$:

$$\begin{aligned} 3\square_n &= (n+1)^3 - \frac{3n(n+1)}{2} - (n+1) \\ &= (n+1)\left[(n+1)^2 - \frac{3}{2}n - 1\right] \\ &= (n+1)\left[n^2 + 2n + 1 - \frac{3}{2}n - 1\right] \\ &= (n+1)\left(n^2 + \frac{1}{2}n\right) = n(n+1)(n+1/2) \end{aligned}$$

Dividing by 3 recovers the standard formula.

Method 3: The Repertoire Method

We generalize the recurrence relation used for arithmetic sums:

$$R_0 = \alpha, \quad R_n = R_{n-1} + \beta + \gamma n + \delta n^2$$

The solution is a linear combination of basis functions:

$$R_n = A(n)\alpha + B(n)\beta + C(n)\gamma + D(n)\delta$$

From our previous work (where $\delta = 0$), we know:

$$A(n) = 1, \quad B(n) = n, \quad C(n) = \frac{n^2 + n}{2}$$

We need to determine $D(n)$. We choose the repertoire function $R_n = n^3$. Substituting $R_n = n^3$ into the recurrence:

$$\begin{aligned} n^3 &= (n-1)^3 + \beta + \gamma n + \delta n^2 \\ n^3 &= n^3 - 3n^2 + 3n - 1 + \beta + \gamma n + \delta n^2 \\ 0 &= n^2(\delta - 3) + n(\gamma + 3) + (\beta - 1) \end{aligned}$$

For this to hold for all n, we must have $\delta = 3, \gamma = -3, \beta = 1$ (and $\alpha = 0$ since $R_0 = 0$). Substituting these parameters into the general solution form:

$$n^3 = 0 + n(1) + \frac{n^2 + n}{2}(-3) + D(n)(3)$$

Solving for $D(n)$:

$$3D(n) = n^3 - n + \frac{3}{2}(n^2 + n) = n(n + 1/2)(n + 1)$$

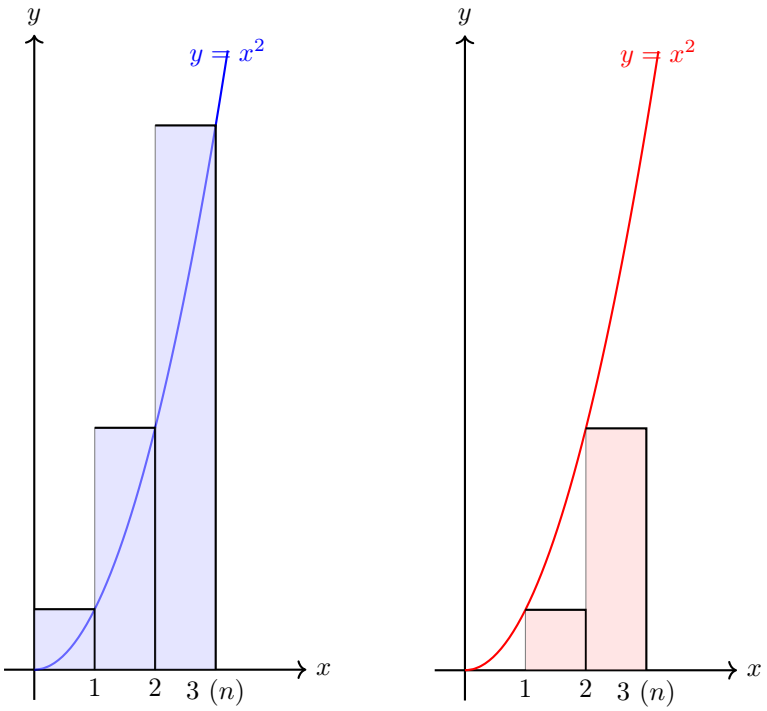
Thus, $D(n) = \frac{n(n+1/2)(n+1)}{3}$. Our target sum $\square_n$ corresponds to the case $\alpha = \beta = \gamma = 0$ and $\delta = 1$.

$$\square_n = D(n)(1) = \frac{n(n+1)(2n+1)}{6}$$

Method 4: The Method of Exhaustion

We now turn to a geometric approach that determines the sum $\square_n = \sum_{k=1}^{n} k^2$ by relating it to the area under a parabola. This method, known as the method of exhaustion, relies solely on algebraic inequalities and geometric intuition.

Consider the region in the Cartesian plane bounded by the curve $y = x^2$, the vertical line $x = n$, and the x-axis. Let $A(n)$ denote the exact area of this region. While we do not yet know the formula for $A(n)$, we can approximate it using rectangles of unit width.



Upper Sum U_n Lower Sum L_n

Figure 4.1: Geometric representation of the upper and lower bounds for the area under $y = x^2$. U_n overestimates the area (blue), while L_n underestimates it (red).

We construct two sets of rectangles to bound the area from above and below:

1. **Upper Approximation (U_n)**: We place rectangles of width 1 such that the top-right corner of the k-th rectangle touches the curve at (k, k^2). The total area of these n rectangles is the sum of their heights:
$$U_n = 1 \cdot 1^2 + 1 \cdot 2^2 + \cdots + 1 \cdot n^2 = \sum_{k=1}^{n} k^2 = \square_n$$
Since these rectangles completely cover the parabolic region, $A(n) < U_n$.

2. **Lower Approximation (L_n)**: We place rectangles of width 1 such that the top-left corner of the k-th rectangle touches the curve at $(k-1, (k-1)^2)$. The total area is:
$$L_n = 1 \cdot 0^2 + 1 \cdot 1^2 + \cdots + 1 \cdot (n-1)^2 = \sum_{k=0}^{n-1} k^2 = \square_n - n^2$$
Since these rectangles fit entirely inside the region, $L_n < A(n)$.

Combining these observations yields the fundamental inequality:

$$\square_n - n^2 < A(n) < \square_n \tag{4.10}$$

Determining the Area Algebraically We suspect that the area $A(n)$ is related to n^3, given the dimensions (base n, height n^2). To pinpoint the coefficient, we appeal to the following algebraic lemma.

***Lemma* 4.4.1.** *Cubic Bounds for Sum of Squares.* For any positive integer m:

$$\sum_{k=1}^{m-1} k^2 < \frac{m^3}{3} < \sum_{k=1}^{m} k^2$$

Proof. Consider the identity $(k+1)^3 - k^3 = 3k^2 + 3k + 1$. Summing this from $k = 1$ to $m-1$, the LHS telescopes to $m^3 - 1$.

$$m^3 - 1 = 3\sum_{k=1}^{m-1} k^2 + 3\sum_{k=1}^{m-1} k + (m-1)$$

Since the terms $3k$ and 1 are positive, we immediately see that $3\sum k^2 < m^3 - 1 < m^3$, which implies $\sum_{k=1}^{m-1} k^2 < m^3/3$. A similar manipulation with $(k-1)^3$ establishes the upper bound. ■

This lemma strongly suggests that the exact area is $A(n) = \frac{n^3}{3}$.

Evaluating the Error Term Let us define the "error" or discrepancy between our discrete sum $\square_n$ and the continuous area $n^3/3$ as:

$$E_n = \square_n - \frac{n^3}{3}$$

We can determine E_n precisely by finding a recurrence relation for it. Recall that $\square_n = \square_{n-1} + n^2$. Substituting the definition of E_n:

$$E_n + \frac{n^3}{3} = \left(E_{n-1} + \frac{(n-1)^3}{3}\right) + n^2$$

Rearranging to isolate $E_n - E_{n-1}$:

$$\begin{aligned} E_n - E_{n-1} &= n^2 + \frac{(n-1)^3 - n^3}{3} \\ &= n^2 + \frac{n^3 - 3n^2 + 3n - 1 - n^3}{3} \\ &= n^2 - n^2 + n - \frac{1}{3} \\ &= n - \frac{1}{3} \end{aligned}$$

This gives us a very simple recurrence for the error term:

$$E_0 = 0, \quad E_n = E_{n-1} + n - \frac{1}{3}$$

Solving this recurrence is straightforward, as it is merely a sum:

$$\begin{aligned} E_n &= \sum_{k=1}^{n} \left(k - \frac{1}{3}\right) \\ &= \sum_{k=1}^{n} k - \sum_{k=1}^{n} \frac{1}{3} \\ &= \frac{n(n+1)}{2} - \frac{n}{3} \\ E_n &= \frac{3n^2 + n}{6} \end{aligned}$$

The Final Formula We can now reconstruct the exact formula for the sum of squares by adding the error term back to the area approximation:

$$\begin{aligned}\square_n &= \frac{n^3}{3} + E_n \\ &= \frac{2n^3}{6} + \frac{3n^2+n}{6} \\ &= \frac{n(n+1)(2n+1)}{6}\end{aligned}$$

Thus, the method of exhaustion allows us to derive the exact discrete sum by correcting the continuous area estimate.

Remark. (Preview of integration). Later, when we introduce the definite integral, we will be able to write the area under $y = x^2$ from 0 to n as $\int_0^n x^2\,dx = n^3/3$. Method 4 is exactly this idea carried out by the classical method of exhaustion, using only finite sums and inequalities instead of limits. In that sense, the integral method is just a faster, more systematic version of Method 4.

Method 5: Expand and Contract

This sophisticated method replaces a single sum with a double sum ("expand"), rearranges the summation order, and simplifies ("contract"). We begin by expressing k^2 as a sum of k terms:

$$\square_n = \sum_{k=1}^{n} k^2 = \sum_{k=1}^{n}\sum_{j=1}^{k} k$$

The domain of summation is $1 \le j \le k \le n$. We swap the summation order to iterate over j first. The bounds become $1 \le j \le n$ and $j \le k \le n$.

$$\square_n = \sum_{j=1}^{n}\sum_{k=j}^{n} k$$

The inner sum is an arithmetic series starting at j and ending at n.

$$\begin{aligned}\sum_{k=j}^{n} k &= \frac{(j+n)(n-j+1)}{2} \\ &= \frac{n(n+1)+j-j^2}{2}\end{aligned}$$

Substituting this back into the expression for $\square_n$:

$$\square_n = \frac{1}{2}\sum_{j=1}^{n}\left[n(n+1)+j-j^2\right]$$

Notice that the term $\sum j^2$ is exactly our original unknown, $\square_n$.

$$\begin{aligned}\square_n &= \frac{1}{2}\left[\sum_{j=1}^{n} n(n+1) + \sum_{j=1}^{n} j - \sum_{j=1}^{n} j^2\right] \\ \square_n &= \frac{1}{2}\left[n^2(n+1) + \frac{n(n+1)}{2} - \square_n\right] \\ 2\square_n &= n^2(n+1) + \frac{n(n+1)}{2} - \square_n \\ 3\square_n &= n(n+1)\left(n+\frac{1}{2}\right) \\ 3\square_n &= n(n+1)\left(\frac{2n+1}{2}\right) \\ \square_n &= \frac{n(n+1)(2n+1)}{6}\end{aligned}$$

A Double Sum Identity

Using Method 5, we can prove a useful identity linking sums of powers to products of indices.

$$\sum_{k=1}^{n} k^3 + \sum_{k=1}^{n} k^2 = 2 \sum_{1 \le j \le k \le n} j \cdot k \tag{4.11}$$

Proof. Evaluate the RHS:

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

This splits directly into the sum of cubes plus the sum of squares. ■

4.5 Exercises

1. **A Fallacious Derivation.** Identify the specific step where the following derivation fails, and explain why the manipulation is invalid.

$$\begin{aligned}\left(\sum_{j=1}^{n} a_j\right)\left(\sum_{j=1}^{n} \frac{1}{a_j}\right) &= \sum_{j=1}^{n}\sum_{k=1}^{n} \frac{a_j}{a_k} \\ &= \sum_{j=1}^{n}\sum_{k=1}^{n} \frac{a_k}{a_k} \quad \text{(by renaming } j \text{ to } k\text{)} \\ &= \sum_{j=1}^{n}\sum_{k=1}^{n} 1 \\ &= n^2\end{aligned}$$

2. **Chebyshev's Inequality.** Let $a_k = k$ and $b_k = H_k$ (the harmonic numbers).
   (a) Are these sequences similarly sorted or oppositely sorted?
   (b) Apply Chebyshev's Sum Inequality to derive an upper bound for $\sum_{k=1}^{n} kH_k$.
   (c) Compare this bound with the exact value derived in the previous chapter using summation by parts.

3. **Telescoping with Parity.** Evaluate the alternating sum:

$$S_n = \sum_{k=1}^{n} \frac{(-1)^k k}{4k^2 - 1}$$

4. **The Minimum Matrix.** Evaluate the double sum $\sum_{1 \le i,j \le n} \min(i, j)$.

   Remark. Decompose the square domain into the diagonal ($i = j$) and the two strict triangular regions ($i < j$ and $j < i$). Use symmetry.

5. ⋆ **Square of Harmonics.** Use the symmetry of double sums to prove:

$$\sum_{k=1}^{n} \frac{H_k}{k} = \frac{1}{2}\left(H_n^2 + \sum_{k=1}^{n} \frac{1}{k^2}\right)$$

   Remark. Write H_k as a sum, converting the LHS into a double sum over $1 \le j \le k \le n$. Relate this to the square of the harmonic sum $(\sum \frac{1}{k})^2$.

Chapter 5

Finite Calculus

Up to this point, our primary instruments for investigating functions on $\mathbb{Z}$ or $\mathbb{N}$ have been recurrence relations (such as $T_n = 2T_{n-1} + 1$) and finite sums (such as $S_n = \sum_{k=0}^{n} a_k$). While effective, these methods often require ad-hoc algebraic manipulations. In this chapter, we introduce the *difference operator*. This operator systematically describes how a discrete function changes from one integer to the next, providing a unified framework for working with sequences and sums.

5.1 The Difference Operator

Let $f : \mathbb{Z} \to \mathbb{R}$ be a function mapping integers to real numbers. We seek a way to quantify how f changes from one step to the next. Since the domain is discrete, the smallest possible step we can take is 1.

Definition 5.1.1. ***Difference Operator***. The *forward difference operator*, denoted by Δ, is defined by the rule:

$$(\Delta f)(n) = f(n+1) - f(n).$$

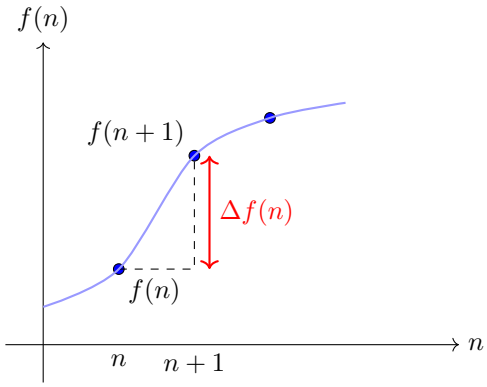


Figure 5.1: Geometric interpretation of $\Delta f(n)$ as the vertical step between consecutive terms.

From an operator-theoretic perspective, Δ is a mapping from the space of functions to itself. Formally, if $\mathcal{F} = \{f : \mathbb{Z} \to \mathbb{R}\}$, then $\Delta : \mathcal{F} \to \mathcal{F}$.

Example 5.1.1. Elementary Differences.

1. **Constant Function**: If $f(n) = c$, then $(\Delta f)(n) = c - c = 0$.
2. **Identity Function**: If $f(n) = n$, then $(\Delta f)(n) = (n+1) - n = 1$.
3. **Quadratic Function**: If $f(n) = n^2$, then:

$$(\Delta f)(n) = (n+1)^2 - n^2 = (n^2 + 2n + 1) - n^2 = 2n + 1.$$

In particular, differences send constants to 0, linear functions to constants, and quadratics to linear functions; in general, Δ lowers the degree of a polynomial by one.

Difference Tables

For a function defined by a sequence of values, we can visualise the action of Δ using a difference table. Let $f(n) = n^2$.

n	0	1	2	3	...
$f(n)$	0	1	4	9	...
$\Delta f(n)$	1	3	5	...	

Table 5.1: Difference table for $f(n) = n^2$. The entry $\Delta f(n)$ is placed between n and $n+1$ to indicate it represents the interval change.

Linearity

The difference operator shares the fundamental algebraic property of the derivative: it is linear.

***Proposition* 5.1.1.** *Linearity of* Δ. Let $f, g : \mathbb{Z} \to \mathbb{R}$ and let $a, b \in \mathbb{R}$ be constants. Then:

$$\Delta(af + bg) = a\Delta f + b\Delta g.$$

Proof. We evaluate the operator at an arbitrary integer n:

$$\begin{aligned}\Delta(af + bg)(n) &= (af + bg)(n+1) - (af + bg)(n) \\ &= af(n+1) + bg(n+1) - af(n) - bg(n) \\ &= a[f(n+1) - f(n)] + b[g(n+1) - g(n)] \\ &= a(\Delta f)(n) + b(\Delta g)(n)\end{aligned}$$

This holds for all n, proving the proposition. ■

Remark. (Relation to Recurrences). The difference operator provides an alternative notation for recurrence relations. The statement $g(n) = \Delta f(n)$ is logically equivalent to the recurrence $f(n+1) = f(n) + g(n)$. Thus, finding the difference is equivalent to determining the step-by-step evolution of the sequence.

The Anti-Difference

Just as the derivative operation is not injective (many functions can have the same derivative), the difference operator is not invertible.

Consider two functions $f_1(n) = c_1$ and $f_2(n) = c_2$ where $c_1 \neq c_2$.

$$\Delta f_1(n) = 0 \quad \text{and} \quad \Delta f_2(n) = 0.$$

Since $\Delta f_1 = \Delta f_2$ despite $f_1 \neq f_2$, we cannot define a unique inverse function Δ^{-1}. However, we can define the concept of an *anti-difference* (or indefinite sum) up to an additive constant.

If $\Delta F(n) = f(n)$, we say that $F(n)$ is the anti-difference of $f(n)$, denoted formally as:

$$\sum f(n)\delta n = F(n) + C,$$

where C is any function $p(n)$ such that $\Delta p(n) = 0$. Over the integers, C is simply a constant.

5.2 The Fundamental Theorem of Finite Calculus

The utility of the difference operator culminates in its application to summation. A recurring theme in mathematics is that operations often have "inverse" procedures that are computationally easier than the direct method. Here, summing *differences* is trivial due to the telescoping nature of the operation.

Consider the sum of $\Delta f(k)$ over a finite interval $a \leq k \leq b-1$:

$$\begin{aligned}\sum_{k=a}^{b-1} \Delta f(k) &= \sum_{k=a}^{b-1} (f(k+1) - f(k)) \\ &= (f(a+1) - f(a)) \\ &+ (f(a+2) - f(a+1)) \\ &+ \dots \\ &+ (f(b) - f(b-1))\end{aligned}$$

Adding these rows, the term $f(a+1)$ cancels with $-f(a+1)$, $f(a+2)$ cancels with $-f(a+2)$, and so forth. The only terms surviving the cancellation are $-f(a)$ from the first row and $f(b)$ from the final row.

***Theorem* 5.2.1. Telescoping Sum (Fundamental Theorem).** Let f be a function defined on the integers. Let $g(k) = \Delta f(k)$. Then for any integers $a < b$:

$$\sum_{k=a}^{b-1} g(k) = f(b) - f(a).$$

Proof. Let $S_n = \sum_{k=a}^{n-1} \Delta f(k)$. We proceed by induction on n, with $n \geq a$. **Base Case ($n = a$)**: The sum is empty ($\sum_{k=a}^{a-1}$), so $S_a = 0$. The formula gives $f(a) - f(a) = 0$. **Inductive Step**: Assume $\sum_{k=a}^{n-1} \Delta f(k) = f(n) - f(a)$. Then:

$$\begin{aligned}\sum_{k=a}^{n} \Delta f(k) &= \left(\sum_{k=a}^{n-1} \Delta f(k) \right) + \Delta f(n) \\ &= (f(n) - f(a)) + (f(n+1) - f(n)) \\ &= f(n+1) - f(a)\end{aligned}$$

Thus the theorem holds for all $b > a$. ■

Note. This theorem establishes that summing the incremental changes of a function over an interval simply yields the net change of the function across that interval. Note that the upper limit of the summation is $b-1$, not b. This is because $\Delta f(b-1) = f(b) - f(b-1)$ is the specific step that brings the sequence to the value $f(b)$.

Remark. This theorem parallels the Fundamental Theorem of Calculus, which relates the integral of a derivative to the values of the function at the boundaries. Here, the sum of the differences (discrete derivatives) collapses to the difference of the function values at the boundaries. The discrete sum stops at $b-1$ precisely because $\Delta f(b-1)$ is the step that bridges $f(b-1)$ to $f(b)$.

5.3 The Basis Problem and Factorial Powers

We have established that the difference operator Δ is linear, allowing us to compute the difference of any polynomial by summing the differences of its terms. However, a practical difficulty arises when applying Δ to standard powers n^k.

Consider the function $f(n) = n^3$. Applying the difference operator yields:

$$\begin{aligned}\Delta(n^3) &= (n+1)^3 - n^3 \\ &= (n^3 + 3n^2 + 3n + 1) - n^3 \\ &= 3n^2 + 3n + 1\end{aligned}$$

While the leading term $3n^2$ is simple, the operation generates lower-order "debris" $(3n+1)$. In general, Δn^k results in a polynomial of degree $k-1$, but the coefficients are complicated combinations involving binomial coefficients. This suggests that the standard basis $\{1, n, n^2, \dots\}$ is not the natural basis for the difference operator. We require a new family of functions that transforms cleanly under Δ.

Definition 5.3.1. ***Falling Factorial Power***. For $m \in \mathbb{N}$ and $x \in \mathbb{R}$, the *falling factorial power*, denoted $x^{\underline{m}}$, is the product of m terms starting at x and decreasing by 1:

$$x^{\underline{m}} = x(x-1)(x-2)\cdots(x-m+1) \quad \text{for } m \geq 1.$$

We define the base case $x^{\underline{0}} = 1$.

Example 5.3.1. Falling Factorials.

- $x^{\underline{1}} = x$
- $x^{\underline{2}} = x(x-1) = x^2 - x$
- $x^{\underline{3}} = x(x-1)(x-2) = x^3 - 3x^2 + 2x$

Note that $x^{\underline{m}}$ is a monic polynomial in x of degree m. Consequently, any polynomial of degree m can be expressed as a linear combination of falling factorials $x^{\underline{0}}, \dots, x^{\underline{m}}$.

For completeness, we also define the complementary concept, though it is less central to the operator Δ.

Definition 5.3.2. ***Rising Factorial Power***. For $m \in \mathbb{N}$ and $x \in \mathbb{R}$, the *rising factorial power* $x^{\overline{m}}$ is defined by:

$$x^{\overline{m}} = x(x+1)\cdots(x+m-1) \quad \text{for } m \geq 1,$$

with $x^{\overline{0}} = 1$.

The Discrete Power Rule

The definition of the falling factorial suggests a structural similarity to standard powers. We now demonstrate that these functions possess a difference property remarkably similar to the power rule found in differential calculus, but adapted for discrete steps.

***Theorem* 5.3.1. Finite Power Rule**. For all integers $m \geq 1$ and all real x:

$$\Delta\big(x^{\underline{m}}\big) = m\, x^{\underline{m-1}}.$$

Proof. We expand the term $(x+1)^{\underline{m}}$ and $x^{\underline{m}}$ to identify common factors.

$$\begin{aligned}(x+1)^{\underline{m}} &= (x+1)(x)(x-1)\cdots(x-m+2) \\ x^{\underline{m}} &= \quad\quad (x)(x-1)\cdots(x-m+2)(x-m+1)\end{aligned}$$

The product $(x)(x-1)\cdots(x-m+2)$ is precisely $x^{\underline{m-1}}$. Factoring this out:

$$\begin{aligned}\Delta\big(x^{\underline{m}}\big) &= (x+1)^{\underline{m}} - x^{\underline{m}} \\ &= x^{\underline{m-1}}\left[(x+1) - (x-m+1)\right] \\ &= x^{\underline{m-1}}\left[x + 1 - x + m - 1\right] \\ &= m\, x^{\underline{m-1}}\end{aligned}$$

■

This theorem confirms that falling factorials are the "eigen-basis" for the difference operator: applying Δ simply reduces the exponent by one and multiplies by the original exponent.

Remark. (Finite Integration). Because $\Delta x^{\underline{m}} = m x^{\underline{m-1}}$, we can immediately deduce the inverse operation. If we wish to find a function $F(n)$ such that $\Delta F(n) = n^{\underline{k}}$, we can reverse the rule:

$$\Delta \left(\frac{n^{\underline{k+1}}}{k+1} \right) = n^{\underline{k}}.$$

This process of finding an anti-difference (sometimes called *finite integration*) allows us to sum falling factorials directly using the Fundamental Theorem of Finite Calculus derived in the previous section.

5.4 Discrete Primitives and Indefinite Sums

Given a function f, the difference Δf quantifies the change of f from one step to the next. It is natural to inquire about the inverse operation: given a sequence of changes, can we reconstruct the original function? This motivates the concept of a *discrete primitive*.

Definition 5.4.1. ***Discrete Primitive***. Let $g : \mathbb{Z} \to \mathbb{R}$. A function $F : \mathbb{Z} \to \mathbb{R}$ is called a *discrete primitive* (or anti-difference) of g if:

$$\Delta F(n) = F(n+1) - F(n) = g(n) \quad \text{for all } n.$$

We denote the collection of all such primitives by the indefinite sum notation:

$$F = \sum g.$$

Example 5.4.1. Simple Primitive. Let $g(n) = 1$ for all n. We seek a function $F(n)$ such that $F(n+1) - F(n) = 1$. The function $F(n) = n$ satisfies this condition:

$$\Delta F(n) = (n+1) - n = 1 = g(n).$$

Thus, $F(n) = n$ is a primitive of the constant function 1.

Uniqueness of Primitives

Primitives are not unique. If $F(n)$ is a primitive, then $F(n) + C$ (where C is a constant) is also a primitive, since $\Delta(F + C) = \Delta F + \Delta C = g(n) + 0 = g(n)$. We now formalise this observation.

***Theorem* 5.4.1. Discrete Fundamental Theorem (Constant Case)**. Let F_1 and F_2 be two discrete primitives of the same function g. Then their difference is a constant:

$$F_1(n) - F_2(n) = C \quad \text{for all } n \in \mathbb{Z},$$

for some constant $C \in \mathbb{R}$.

Proof. Let $H(n) = F_1(n) - F_2(n)$. Applying the difference operator:

$$\begin{aligned} \Delta H(n) &= \Delta(F_1 - F_2)(n) \\ &= \Delta F_1(n) - \Delta F_2(n) \\ &= g(n) - g(n) \\ &= 0 \end{aligned}$$

If $\Delta H(n) = 0$, then $H(n+1) - H(n) = 0$, implying $H(n+1) = H(n)$ for all n. By induction, $H(n)$ must be constant for all $n \in \mathbb{Z}$. ■

Remark. (Periodic Generalisation). If we extend the domain from $\mathbb{Z}$ to $\mathbb{R}$, the condition $\Delta H(x) = 0$ implies $H(x+1) = H(x)$. This defines a *1-periodic function*, not necessarily a constant. For example, $C(x) = \sin(2\pi x)$ satisfies $\Delta C(x) = 0$. However, when restricted to integer inputs, any 1-periodic function behaves exactly like a constant.

Indefinite Sum Notation

We introduce a specific notation to represent the general solution to the difference equation $\Delta F = g$.

Definition 5.4.2. ***Indefinite Sum***. Let $g : \mathbb{Z} \to \mathbb{R}$. An *indefinite sum* of g, denoted by $\sum g(x)\,\delta x$, is defined as:

$$\sum g(x)\,\delta x = F(x) + C,$$

where $\Delta F(x) = g(x)$ and C is an arbitrary constant.

Note. There is no genuine inverse operator Δ^{-1} because Δ is not injective. For example, if $f_1(n) = c_1$ and $f_2(n) = c_2$ are two different constant functions, then $\Delta f_1 = 0$ and $\Delta f_2 = 0$. Thus, knowing only that the difference is 0 does not allow us to distinguish between f_1 and f_2. This ambiguity is captured by the additive constant C.

5.5 Definite Sums

We now turn the idea of a discrete primitive into a definite summation operator that evaluates the total change of a function between two integer bounds.

Definition 5.5.1. ***Definite Sum***. Let $g : \mathbb{Z} \to \mathbb{R}$. Suppose F is a discrete primitive of g, such that $\Delta F = g$. The *definite sum* of g from a to b is defined by:

$$\sum_a^b g(x)\,\delta x := F(x)\Big|_a^b = F(b) - F(a).$$

Note. This definition is well-posed. If G is another primitive of g, then $G(n) = F(n) + C$. The constant cancels out in the difference:

$$G(b) - G(a) = (F(b) + C) - (F(a) + C) = F(b) - F(a).$$

The Connection to Ordinary Sums

To understand the utility of this notation, we must establish the relationship between the definite sum $\sum_a^b \ldots \delta x$ and the standard summation notation $\sum_{k=a}^n$. Recall that $g(x) = F(x+1) - F(x)$. We evaluate the definite sum for small intervals:

- **Case** $b = a$: $\sum_a^a g(x)\delta x = F(a) - F(a) = 0$.
- **Case** $b = a + 1$: $\sum_a^{a+1} g(x)\delta x = F(a+1) - F(a) = \Delta F(a) = g(a)$.
- **Case** $b = a + 2$:

$$\begin{aligned}\sum_a^{a+2} g(x)\delta x &= F(a+2) - F(a) \\ &= (F(a+2) - F(a+1)) + (F(a+1) - F(a)) \\ &= g(a+1) + g(a)\end{aligned}$$

By extending this pattern inductively, we arrive at a fundamental identity.

***Theorem* 5.5.1. Definite Sum as Ordinary Sum**. For integers $b \geq a$:

$$\sum_a^b g(x)\,\delta x = \sum_{k=a}^{b-1} g(k).$$

Proof. We proceed by induction on b. The base cases $b = a$ and $b = a + 1$ are verified above. Assume the relation holds for $b = a + k$.

$$\begin{aligned}\sum_{a}^{a+k+1} g(x)\,\delta x &= F(a+k+1) - F(a) \\ &= (F(a+k+1) - F(a+k)) + (F(a+k) - F(a)) \\ &= g(a+k) + \sum_{a}^{a+k} g(x)\,\delta x \\ &= g(a+k) + \sum_{j=a}^{a+k-1} g(j) \quad \text{(by hypothesis)} \\ &= \sum_{j=a}^{a+k} g(j)\end{aligned}$$

Thus, the identity holds for all $b \geq a$. ■

***Corollary* 5.5.1.** *Evaluation of Finite Sums.* To evaluate a standard sum $\sum_{k=a}^{n} g(k)$, we express it as a definite sum with upper limit $n + 1$:

$$\sum_{k=a}^{n} g(k) = \sum_{a}^{n+1} g(x)\,\delta x = F(n+1) - F(a),$$

where $\Delta F = g$.

Finite Power Rule for Indefinite Sums

In the previous section, we established the Finite Power Rule: $\Delta(x^{\underline{m}}) = m x^{\underline{m-1}}$. By inverting this rule, we obtain a method for taking anti-differences of falling factorials.

***Theorem* 5.5.2. Finite Power Rule for Indefinite Sums**. For any integer $m \geq 0$:

$$\sum x^{\underline{m}}\,\delta x = \frac{x^{\underline{m+1}}}{m+1} + C,$$

where C is an arbitrary constant.

Proof. Let $F(x) = \frac{x^{\underline{m+1}}}{m+1}$. We verify that $\Delta F(x) = x^{\underline{m}}$:

$$\Delta F(x) = \frac{1}{m+1}\Delta\left(x^{\underline{m+1}}\right) = \frac{1}{m+1}(m+1)x^{\underline{m}} = x^{\underline{m}}.$$

Thus, $F(x)$ is the primitive of $x^{\underline{m}}$. ■

Combining this with the corollary above provides a closed-form solution for sums of falling factorials.

***Corollary* 5.5.2.** *Sum of Falling Factorial Powers.* For integers $m \geq 0$ and $n \geq 0$:

$$\sum_{k=0}^{n} k^{\underline{m}} = \frac{(n+1)^{\underline{m+1}}}{m+1}.$$

Proof. We convert the ordinary sum to a definite sum with limit $n + 1$:

$$\sum_{k=0}^{n} k^{\underline{m}} = \sum_{0}^{n+1} x^{\underline{m}}\delta x = \left[\frac{x^{\underline{m+1}}}{m+1}\right]_{0}^{n+1}.$$

Evaluating at the bounds:

$$\frac{(n+1)^{\underline{m+1}}}{m+1} - \frac{0^{\underline{m+1}}}{m+1}.$$

For $m \geq 0$, the term $m+1 \geq 1$, so $0^{\underline{m+1}} = 0 \cdot (-1) \cdots = 0$. The result follows. ■

Example 5.5.1. Elementary Sums. Using this corollary, we can derive standard summation formulas instantly:

- **Sum of Ones ($m = 0$):**

$$\sum_{k=0}^{n} 1 = \sum_{k=0}^{n} k^{\underline{0}} = \frac{(n+1)^{\underline{1}}}{1} = n+1.$$

- **Sum of Integers ($m = 1$):**

$$\sum_{k=0}^{n} k = \sum_{k=0}^{n} k^{\underline{1}} = \frac{(n+1)^{\underline{2}}}{2} = \frac{(n+1)n}{2}.$$

Application: Sum of Squares

We now demonstrate the power of the operator method by deriving the formula for $\sum_{k=0}^{n} k^2$ without resorting to geometric intuition or induction guesses. The procedure follows three logical steps.

(i) **Basis Transformation.** We must express the summand k^2 in terms of the falling factorial basis $\{k^{\underline{0}}, k^{\underline{1}}, k^{\underline{2}}, \dots\}$. Recall that $k^{\underline{2}} = k(k-1) = k^2 - k$. Rearranging this, we get $k^2 = k^{\underline{2}} + k$. Since $k = k^{\underline{1}}$, we have:

$$k^2 = k^{\underline{2}} + k^{\underline{1}}.$$

(ii) **Finite Integration.** We seek the primitive $F(k)$ such that $\Delta F(k) = k^2$. By linearity, we integrate term by term using the Finite Power Rule:

$$\sum k^2 \delta k = \sum (k^{\underline{2}} + k^{\underline{1}}) \delta k = \frac{k^{\underline{3}}}{3} + \frac{k^{\underline{2}}}{2} + C$$

Let $F(k) = \frac{k^{\underline{3}}}{3} + \frac{k^{\underline{2}}}{2}$.

(iii) **Evaluation.** We evaluate the definite sum from 0 to $n+1$:

$$\sum_{k=0}^{n} k^2 = F(n+1) - F(0)$$

First, note that $F(0) = 0$. We compute $F(n+1)$:

$$\begin{aligned} F(n+1) &= \frac{(n+1)^{\underline{3}}}{3} + \frac{(n+1)^{\underline{2}}}{2} \\ &= \frac{(n+1)n(n-1)}{3} + \frac{(n+1)n}{2} \\ &= (n+1)n\left[\frac{n-1}{3} + \frac{1}{2}\right] \\ &= (n+1)n\left[\frac{2(n-1)+3}{6}\right] \\ &= (n+1)n\left[\frac{2n+1}{6}.\right] \end{aligned}$$

Thus, we recover the classic formula:

$$\sum_{k=0}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}.$$

5.6 Definite Sums and Power Sums

We now develop general tools for manipulating definite sums and apply them to evaluate sums of falling factorials and ordinary powers. Recall that if F is a discrete primitive of g (i.e., $\Delta F = g$), the definite sum over the integer interval $[a, b)$ is defined as:

$$\sum_a^b g(x)\,\delta x := F(x)\Big|_a^b = F(b) - F(a).$$

Algebraic Properties

The definite sum shares the fundamental algebraic properties of the definite integral, specifically regarding the manipulation of limits.

***Proposition* 5.6.1.** *Properties of Limits.* For any integers a, b, c and any function g:

1. **Reversal of Limits:** Swapping the bounds negates the sum.
$$\sum_b^a g(x)\,\delta x = -\sum_a^b g(x)\,\delta x.$$
2. **Chasles' Relation (Additivity):** Intermediate bounds can be inserted or removed.
$$\sum_a^b g(x)\,\delta x + \sum_b^c g(x)\,\delta x = \sum_a^c g(x)\,\delta x.$$

Proof. Let F be a primitive of g.

1. By definition, $\sum_b^a g(x)\,\delta x = F(a) - F(b) = -(F(b) - F(a)) = -\sum_a^b g(x)\,\delta x$.
2. Expanding the sums yields:
$$(F(b) - F(a)) + (F(c) - F(b)) = F(c) - F(a) = \sum_a^c g(x)\,\delta x.$$

■

Power Sums of Falling Factorials

In the previous section, we established the Finite Power Rule for indefinite sums:

$$\sum x^{\underline{m}}\,\delta x = \frac{x^{\underline{m+1}}}{m+1} + C$$

We can now apply the Fundamental Theorem of Finite Calculus to evaluate definite sums of falling factorials.

$$\sum_a^b x^{\underline{m}}\,\delta x = \frac{x^{\underline{m+1}}}{m+1}\Big|_a^b = \frac{b^{\underline{m+1}} - a^{\underline{m+1}}}{m+1}$$

By setting the lower bound to 0, we obtain a particularly simple formula:

$$\sum_0^n x^{\underline{m}}\,\delta x = \frac{n^{\underline{m+1}}}{m+1}.$$

***Theorem* 5.6.1. Sum of Falling Powers**. For integers $m \geq 0$ and $n \geq 0$:

$$\sum_{k=0}^{n-1} k^{\underline{m}} = \sum_{0}^{n} x^{\underline{m}} \, \delta x = \frac{n^{\underline{m+1}}}{m+1}$$

Proof. Using the relationship $\sum_a^b g(x)\,\delta x = \sum_{k=a}^{b-1} g(k)$, we set $a = 0$ and $b = n$.

$$\sum_{k=0}^{n-1} k^{\underline{m}} = \frac{n^{\underline{m+1}}}{m+1} - \frac{0^{\underline{m+1}}}{m+1}$$

Since $m \geq 0$, $m+1 \geq 1$, so $0^{\underline{m+1}} = 0$. The result follows. ■

Example 5.6.1. Sum of the First n Integers. Let $m = 1$. Since $k^{\underline{1}} = k$:

$$\sum_{k=0}^{n-1} k = \frac{n^{\underline{2}}}{2} = \frac{n(n-1)}{2}$$

To sum up to n (i.e., $\sum_{k=0}^{n}$), we simply substitute $n+1$ for n in the result, yielding $\frac{(n+1)n}{2}$.

Converting Ordinary Powers

While falling factorials integrate cleanly, mathematical problems often present ordinary powers (k^2, k^3). To compute sums like $\sum k^2$, we transform the ordinary powers into a linear combination of falling factorials.

***Lemma* 5.6.1.** *Basis Transformation for k^2.* $k^2 = k^{\underline{2}} + k^{\underline{1}}$.

Proof. Expanding the falling factorials gives $k^{\underline{2}} + k^{\underline{1}} = k(k-1) + k = k^2 - k + k = k^2$. ■

Example 5.6.2. Sum of Squares. We evaluate $S = \sum_{k=0}^{n-1} k^2$ by substituting the basis transformation.

$$\begin{aligned}\sum_{k=0}^{n-1} k^2 &= \sum_{k=0}^{n-1} \left(k^{\underline{2}} + k^{\underline{1}}\right) \\ &= \sum_{k=0}^{n-1} k^{\underline{2}} + \sum_{k=0}^{n-1} k^{\underline{1}} \quad \text{(by linearity)}\end{aligned}$$

Applying the Sum of Falling Powers theorem to each term:

$$\begin{aligned}S &= \frac{n^{\underline{3}}}{3} + \frac{n^{\underline{2}}}{2} \\ &= \frac{n(n-1)(n-2)}{3} + \frac{n(n-1)}{2} \\ &= n(n-1)\left(\frac{n-2}{3} + \frac{1}{2}\right) \\ &= \frac{n(n-1)(2n-1)}{6}\end{aligned}$$

This provides the sum for $0 \leq k < n$. To find the standard sum $\sum_{k=0}^{n} k^2$, we replace n with $n+1$, recovering the familiar $\frac{n(n+1)(2n+1)}{6}$.

For higher powers, the decomposition requires more terms.

***Lemma* 5.6.2.** *Basis Transformation for k^3.*

$$k^3 = k^{\underline{3}} + 3k^{\underline{2}} + k^{\underline{1}}.$$

Proof.

$$\begin{aligned} k^{\underline{3}} + 3k^{\underline{2}} + k^{\underline{1}} &= k(k-1)(k-2) + 3k(k-1) + k \\ &= k(k-1)\left[(k-2)+3\right] + k \\ &= k(k-1)(k+1) + k \\ &= k(k^2-1) + k \\ &= k^3 - k + k = k^3. \end{aligned}$$

■

Example 5.6.3. Sum of Cubes. We evaluate $\sum_{k=0}^{n-1} k^3$ by integrating the transformed polynomial term by term.

$$\begin{aligned} \sum_{k=0}^{n-1} k^3 &= \sum_{0}^{n} \left(x^{\underline{3}} + 3x^{\underline{2}} + x^{\underline{1}}\right) \delta x \\ &= \left[\frac{x^{\underline{4}}}{4} + 3\frac{x^{\underline{3}}}{3} + \frac{x^{\underline{2}}}{2}\right]_0^n \\ &= \frac{n^{\underline{4}}}{4} + n^{\underline{3}} + \frac{n^{\underline{2}}}{2} \end{aligned}$$

Substituting the definitions $n^{\underline{4}} = n(n-1)(n-2)(n-3)$, etc., and performing the algebraic simplification (left as an exercise) yields:

$$\sum_{k=0}^{n-1} k^3 = \left(\frac{n(n-1)}{2}\right)^2$$

Again, replacing n with $n+1$ gives the standard result for $\sum_{k=1}^{n} k^3$, which is $\left(\frac{n(n+1)}{2}\right)^2$.

5.7 Negative Falling Powers and Harmonic Numbers

Thus far, our definition of the falling factorial $x^{\underline{m}}$ has been restricted to non-negative integers m. We now extend this concept to negative indices, preserving the essential algebraic properties of the operator.

Definition 5.7.1. ***Negative Falling Powers***. For $m > 0$, the negative falling factorial $x^{\underline{-m}}$ is defined by:

$$x^{\underline{-m}} = \frac{1}{(x+1)(x+2)\cdots(x+m)}$$

Specifically:

$$x^{\underline{-1}} = \frac{1}{x+1}, \quad x^{\underline{-2}} = \frac{1}{(x+1)(x+2)}$$

This definition ensures that the fundamental shift identity $x^{\underline{m+n}} = x^{\underline{m}}(x-m)^{\underline{n}}$ remains valid for all integers m, n.

***Theorem* 5.7.1. Extended Finite Power Rule**. For any integer m (positive, negative, or zero):

$$\Delta\left(x^{\underline{m}}\right) = m\, x^{\underline{m-1}}$$

Proof. We verify this for the negative case. Let $m = -k$ where $k > 0$.

$$\begin{aligned} \Delta\left(x^{\underline{-k}}\right) &= (x+1)^{\underline{-k}} - x^{\underline{-k}} \\ &= \frac{1}{(x+2)(x+3)\cdots(x+k+1)} - \frac{1}{(x+1)(x+2)\cdots(x+k)} \end{aligned}$$

To combine these fractions, we find a common denominator $D = (x+1)(x+2)\cdots(x+k+1)$.

$$\begin{aligned}\Delta\big(x^{\underline{-k}}\big) &= \frac{(x+1)-(x+k+1)}{D} \\ &= \frac{-k}{(x+1)(x+2)\cdots(x+k+1)} \\ &= -k\, x^{\underline{-(k+1)}} \\ &= m\, x^{\underline{m-1}}\end{aligned}$$

Thus, the power rule holds universally. ■

Consequently, for any $m \neq -1$, the indefinite sum is given by:

$$\sum x^{\underline{m}}\,\delta x = \frac{x^{\underline{m+1}}}{m+1} + C$$

The Special Case $m = -1$

The case $m = -1$ yields a division by zero in the power rule formula ($m + 1 = 0$). This is analogous to the integral $\int x^{-1}dx = \ln x$ in calculus. In the discrete realm, the logarithmic role is played by the *harmonic numbers*.

Definition 5.7.2. ***Harmonic Numbers***. For an integer $x \geq 1$, the x-th harmonic number H_x is the sum of the reciprocals of the first x integers:

$$H_x = \sum_{k=1}^{x} \frac{1}{k}$$

We define $H_0 = 0$.

***Proposition* 5.7.1.** *Difference of Harmonic Numbers.* The harmonic numbers satisfy the difference equation:

$$\Delta H_x = \frac{1}{x+1} = x^{\underline{-1}}$$

Proof.

$$\Delta H_x = H_{x+1} - H_x = \left(\sum_{k=1}^{x+1} \frac{1}{k}\right) - \left(\sum_{k=1}^{x} \frac{1}{k}\right) = \frac{1}{x+1}$$

■

This identifies H_x as an anti-difference of $x^{\underline{-1}}$.

***Theorem* 5.7.2. Unified Falling Powers Theorem**. For any integers $a \leq b$ and $m \in \mathbb{Z}$:

$$\sum_{a}^{b} x^{\underline{m}}\,\delta x = \begin{cases} \dfrac{x^{\underline{m+1}}}{m+1}\Big|_a^b & \text{if } m \neq -1, \\ H_x\Big|_a^b & \text{if } m = -1. \end{cases}$$

Note. H_x is only defined for non-negative integers. If the domain extends to negative integers, we must use the generalised harmonic definition, but for this course, we restrict $x \geq 0$.

We present two distinct proofs.

Method 1: The Operator Viewpoint. We proceed by identifying the discrete primitive for each case.

Case 1: $m \neq -1$**.** From the Extended Finite Power Rule, we established that $\Delta\big(x^{\underline{m+1}}\big) = (m+1)x^{\underline{m}}$. Since the difference operator is linear, we can divide by the constant $m+1$:

$$\Delta\left(\frac{x^{\underline{m+1}}}{m+1}\right) = \frac{1}{m+1}\Delta\big(x^{\underline{m+1}}\big) = x^{\underline{m}}$$

Thus, $F(x) = \frac{x^{\underline{m+1}}}{m+1}$ is a primitive of $x^{\underline{m}}$. By the definition of the definite sum:

$$\sum_a^b x^{\underline{m}}\,\delta x = F(b) - F(a) = \frac{x^{\underline{m+1}}}{m+1}\Big|_a^b$$

Case 2: $m = -1$**.** Recall the property of harmonic numbers derived previously:

$$\Delta H_x = \frac{1}{x+1} = x^{\underline{-1}}$$

This identifies H_x as the primitive of $x^{\underline{-1}}$. Consequently:

$$\sum_a^b x^{\underline{-1}}\,\delta x = H_x\Big|_a^b$$

■

Method 2: The Classical Sigma Viewpoint. We rely on the equivalence $\sum_a^b g(x)\delta x = \sum_{k=a}^{b-1} g(k)$ and telescoping sums.

Case 1: $m \neq -1$**.** Consider the difference of the term $(k+1)^{\underline{m+1}} - k^{\underline{m+1}}$. By the discrete power rule, this expands to $(m+1)k^{\underline{m}}$. We can rearrange this to isolate the term $k^{\underline{m}}$:

$$k^{\underline{m}} = \frac{1}{m+1}\left((k+1)^{\underline{m+1}} - k^{\underline{m+1}}\right)$$

We sum this expression from $k = a$ to $b-1$:

$$\sum_{k=a}^{b-1} k^{\underline{m}} = \frac{1}{m+1}\sum_{k=a}^{b-1}\left((k+1)^{\underline{m+1}} - k^{\underline{m+1}}\right)$$

The right-hand side is a telescoping sum of the form $\sum(\phi_{k+1} - \phi_k)$, which collapses to $\phi_b - \phi_a$.

$$\sum_{k=a}^{b-1} k^{\underline{m}} = \frac{1}{m+1}\left(b^{\underline{m+1}} - a^{\underline{m+1}}\right)$$

This matches the operator result.

Case 2: $m = -1$**.** In sigma notation, the sum is:

$$\sum_{k=a}^{b-1} k^{\underline{-1}} = \sum_{k=a}^{b-1}\frac{1}{k+1}$$

We perform an index shift. Let $j = k+1$. As k ranges from a to $b-1$, j ranges from $a+1$ to b.

$$\sum_{j=a+1}^{b}\frac{1}{j}$$

Using the definition of harmonic numbers $H_n = \sum_{i=1}^{n}\frac{1}{i}$, this sum represents the difference between the sum up to b and the sum up to a:

$$\sum_{j=a+1}^{b}\frac{1}{j} = \left(\sum_{j=1}^{b}\frac{1}{j}\right) - \left(\sum_{j=1}^{a}\frac{1}{j}\right) = H_b - H_a$$

■

5.8 Exponentials and Geometric Progressions

We now investigate functions that are proportional to their own differences, analogous to e^x in calculus (if you know calculus). Consider $f(x) = c^x$ for some constant c.

***Proposition* 5.8.1.** *Difference of Exponentials.* For any constant $c \in \mathbb{R}$ then $\Delta(c^x) = (c-1)c^x$

Proof. $\Delta(c^x) = c^{x+1} - c^x = c \cdot c^x - c^x = (c-1)c^x$. ■

To find the primitive of c^x, we simply divide by the constant factor $(c-1)$, provided $c \neq 1$.

$$\sum c^x \, \delta x = \frac{c^x}{c-1} + C$$

***Theorem* 5.8.1. Geometric Progression Formula**. For $c \neq 1$ and integers $a \leq b$:

$$\sum_a^b c^x \, \delta x = \frac{c^x}{c-1}\Big|_a^b = \frac{c^b - c^a}{c-1}$$

Proof. We previously established that the difference of an exponential function is proportional to itself $\Delta(c^x) = (c-1)c^x$. For $c \neq 1$, we can invert this relation to find the primitive.

$$\Delta\left(\frac{c^x}{c-1}\right) = \frac{1}{c-1}\Delta(c^x) = \frac{1}{c-1}(c-1)c^x = c^x$$

Thus, $F(x) = \frac{c^x}{c-1}$ is the primitive. By definition:

$$\sum_a^b c^x \, \delta x = F(b) - F(a) = \frac{c^b - c^a}{c-1}$$

■

Remark. Converting to standard sigma notation ($\sum_{k=a}^{b-1} c^k$), this recovers the standard formula for the sum of a geometric series.

Example 5.8.1. Powers of 2. Let $c = 2$. Then $\Delta(2^x) = (2-1)2^x = 2^x$. The function 2^x is its own difference, just as e^x is its own derivative.

$$\sum_{k=0}^{n} 2^k = \sum_0^{n+1} 2^x \, \delta x = \frac{2^{n+1} - 2^0}{2-1} = 2^{n+1} - 1$$

5.9 Summation by Parts

One of the most useful tools in finite calculus is *summation by parts*, a discrete identity that comes from the product rule for differences. It lets us systematically handle sums involving products of functions, such as $k2^k$ or kH_k, by transforming them into simpler sums.

The Discrete Product Rule

Let $u(x)$ and $v(x)$ be functions on $\mathbb{Z}$. We compute the difference of their product:

$$\begin{aligned}
\Delta(u(x)v(x)) &= u(x+1)v(x+1) - u(x)v(x) \\
&= u(x+1)v(x+1) - u(x)v(x+1) + u(x)v(x+1) - u(x)v(x) \\
&= v(x+1)\big[u(x+1) - u(x)\big] + u(x)\big[v(x+1) - v(x)\big] \\
&= v(x+1)\Delta u(x) + u(x)\Delta v(x)
\end{aligned}$$

In this discrete setting, one of the factors is evaluated at the shifted point $x+1$. It is convenient to introduce the *shift operator* E, defined by $Ev(x) = v(x+1)$.

***Proposition* 5.9.1.** *Discrete Product Rule.* $\Delta(uv) = u\Delta v + Ev\Delta u$

Summation by Parts Formula

By applying the Fundamental Theorem of Finite Calculus to the product rule, we derive the summation formula. Summing both sides from a to b:

$$\sum_a^b \Delta(uv)\,\delta x = \sum_a^b u\Delta v\,\delta x + \sum_a^b Ev\Delta u\,\delta x$$

The left side telescopes to $u(x)v(x)\Big|_a^b$. Rearranging terms yields the formula.

***Theorem* 5.9.1. Summation by Parts**. For functions u, v and integers $a \le b$:

$$\sum_a^b u(x)\Delta v(x)\,\delta x = u(x)v(x)\Big|_a^b - \sum_a^b Ev(x)\Delta u(x)\,\delta x$$

In standard sigma notation:

$$\sum_{k=a}^{b-1} u_k\Delta v_k = [u_k v_k]_a^b - \sum_{k=a}^{b-1} v_{k+1}\Delta u_k$$

This formula transforms a difficult sum $\sum u\Delta v$ into a potentially simpler sum $\sum Ev\Delta u$. The strategy is to choose u such that Δu is simpler than u (e.g., a polynomial of lower degree) and v such that Δv is a manageable part of the original summand.

Example: Summing $k2^k$

Evaluate $S_n = \sum_{k=0}^n k2^k$. We identify the parts: let $u(x) = x$ and $\Delta v(x) = 2^x$. We compute the necessary components:

- $\Delta u(x) = 1$ (degree reduced from 1 to 0).
- $v(x) = 2^x$ (since $\Delta 2^x = 2^x$).
- $Ev(x) = 2^{x+1}$.

Applying summation by parts on the interval $[0, n+1)$:

$$\begin{aligned}
\sum_0^{n+1} x2^x\,\delta x &= [x2^x]_0^{n+1} - \sum_0^{n+1} 2^{x+1}\cdot 1\,\delta x \\
&= \big((n+1)2^{n+1} - 0\cdot 2^0\big) - 2\sum_0^{n+1} 2^x\,\delta x \\
&= (n+1)2^{n+1} - 2[2^x]_0^{n+1} \\
&= (n+1)2^{n+1} - 2(2^{n+1}-1) \\
&= (n+1)2^{n+1} - 2^{n+2} + 2 \\
&= 2^{n+1}(n+1-2) + 2 \\
&= (n-1)2^{n+1} + 2
\end{aligned}$$

Example: Summing kH_k

Evaluate $\sum_{k=0}^{n-1} kH_k$. Let $u(x) = H_x$ and $\Delta v(x) = x = x^{\underline{1}}$. Components:

- $\Delta u(x) = x^{\underline{-1}} = \frac{1}{x+1}$.
- $v(x) = \frac{x^{\underline{2}}}{2}$.
- $Ev(x) = \frac{(x+1)^{\underline{2}}}{2} = \frac{(x+1)x}{2}$.

Applying summation by parts on $[0, n)$:

$$\begin{aligned}\sum_0^n H_x x\,\delta x &= \left[H_x \frac{x^{\underline{2}}}{2}\right]_0^n - \sum_0^n \frac{(x+1)x}{2}\cdot\frac{1}{x+1}\,\delta x \\ &= \left(H_n \frac{n(n-1)}{2} - 0\right) - \frac{1}{2}\sum_0^n x\,\delta x \\ &= \frac{n(n-1)}{2}H_n - \frac{1}{2}\left[\frac{x^{\underline{2}}}{2}\right]_0^n \\ &= \frac{n(n-1)}{2}H_n - \frac{n(n-1)}{4} \\ &= \frac{n(n-1)}{2}\left(H_n - \frac{1}{2}\right)\end{aligned}$$

This yields a compact formula for the sum of harmonic numbers weighted by their index.

5.10 Exercises

Part I: Basics and Definitions

1. **Zero Falling Power.** Determine the value of $0^{\underline{m}}$ for any integer m (positive, negative, or zero).

   Remark. Recall the definition $x^{\underline{m}} = x(x-1)\dots(x-m+1)$ for $m > 0$, $x^{\underline{0}} = 1$, and $x^{\underline{-m}} = 1/((x+1)\dots(x+m))$.

2. **Evaluating Sums via Basis Transformation.** Evaluate the sum of a quadratic polynomial:

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

3. **Negative Falling Powers.** While the text demonstrated the theory of negative falling factorials, their utility shines in simplifying partial fraction sums. Evaluate the following sum using finite calculus:

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

   Remark. Express the summand as a falling factorial $k^{\underline{m}}$ with a negative exponent. Apply the Unified Falling Powers Theorem to evaluate the definite sum. Verify your answer for $n = 1$ and $n = 2$.

Part II: Techniques and Applications

4. **Algebraic Proof of Summation by Parts.** The general rule for summation by parts was given as:

$$\sum_a^b u(x)\Delta v(x)\,\delta x = u(x)v(x)\Big|_a^b - \sum_a^b Ev(x)\Delta u(x)\,\delta x.$$

In standard sigma notation, this is equivalent to:

$$\sum_{0\le k<n} (a_{k+1} - a_k)b_k = a_n b_n - a_0 b_0 - \sum_{0\le k<n} a_{k+1}(b_{k+1} - b_k).$$

Prove this sigma-notation formula directly using the distributive, associative, and commutative laws, without invoking the difference operator Δ.

5. **Summation by Parts Practice.** Evaluate the following definite sums using summation by parts:
   (a) $\sum_0^n x3^x \delta x$
   (b) $\sum_0^n H_x \delta x$

   Remark. Let $u(x) = H_x$ and $\Delta v(x) = 1$.

6. **Perturbation on Harmonic Sums.** Attempt to evaluate the sum $\sum_{k=1}^n kH_k$ by the perturbation method. You will find that the term you are trying to solve for cancels out. Instead of the value of the sum, what identity regarding harmonic numbers does this failed attempt produce?

7. **Iterated Summation by Parts.** Evaluate the sum $\sum_{k=0}^n k^2 2^k$.

   Remark. You will need to apply summation by parts twice.

8. ⋆ **Discrete Integration I.** Compute $\Delta(c^x)$ and use it to deduce the value of the sum:

$$\sum_{k=1}^n \frac{k}{c^k} \quad \text{and} \quad \sum_{k=1}^n \frac{(-2)^{\underline{k}}}{k}.$$

   Remark. Consider $\Delta(xc^{-x})$. You may need to apply summation by parts.

9. ⋆⋆ **Discrete Integration II.** Deduce the value of the sum:

$$\sum_{k=1}^n \binom{k}{2} / c^{k-1}.$$

Part III: Theory and Structure

10. **Symmetry of the Difference Rule.** The text derives the discrete product rule:

$$\Delta(uv) = u\Delta v + Ev\Delta u$$

    (a) At first glance, the left-hand side $\Delta(uv)$ appears symmetric with respect to u and v, while the right-hand side $u\Delta v + Ev\Delta u$ contains a shift operator E only on v, breaking the symmetry. Is this formula correct? If so, show that the RHS is algebraically equivalent to $v\Delta u + Eu\Delta v$.
    (b) Derive a symmetric version of the rule that treats u and v identically, perhaps involving terms like $u(x+1)$ and $v(x+1)$.

11. ⋆ **Generalised Product Rule.** The product rule for two functions is $\Delta(uv) = u\Delta v + Ev\Delta u$.
    (a) Derive the difference rule for a product of three functions, $\Delta(u(x)v(x)w(x))$, in terms of $\Delta u, \Delta v, \Delta w$ and the shift operator E.
    (b) Generalise this result. If $\Delta(f_1 \dots f_n)$ is expanded, what is the coefficient of the term containing exactly k difference operators?

12. **Laws of Exponents for Rising Factorials.** Just as falling factorials satisfy $x^{\underline{m+n}} = x^{\underline{m}}(x-m)^{\underline{n}}$, rising factorials obey a similar exponent law.
    (a) State and prove the law of exponents for $x^{\overline{m+n}}$.
    (b) Use this law to define $x^{\overline{-m}}$ for $m > 0$.

13. Factorial Power Identities. Prove the following identities relating ordinary powers, falling factorials, and rising factorials. Assume all denominators are non-zero.

(a) $x^{\underline{n}} = (-1)^n(-x)^{\overline{n}}$ and $x^{\overline{n}} = (-1)^n(-x)^{\underline{n}}$

(b) $x^{\underline{m}}(x-m)^{\underline{n}} = x^{\underline{m+n}}$ and $x^{\underline{m}}(x-m)^{\underline{n}} = x^{\underline{n}}(x-n)^{\underline{m}}$

14. Rising Factorials. The text defined the rising factorial as $x^{\overline{m}} = x(x+1)\cdots(x+m-1)$.

(a) Prove that the difference of a rising factorial is given by:

$$\Delta(x^{\overline{m}}) = m(x+1)^{\overline{m-1}}.$$

Remark. Note the shift in the base compared to the falling factorial rule.

(b) Using the relationship between rising and falling factorials, or by direct summation, prove that:

$$\sum_0^n x^{\overline{m}}\delta x = \frac{n^{\overline{m+1}}}{m+1}.$$

15. Alternative Factorial Reflection. Prove the reflection formula:

$$x^{\underline{m}} = (-1)^m(-x)^{\overline{m}} = \frac{1}{(x+1)^{\overline{-m}}}.$$

Explain the condition under which the last equality holds.

16. ⋆ Higher Order Differences. We define the higher order differences recursively by $\Delta^n f(x) = \Delta(\Delta^{n-1}f(x))$.

(a) Recall that $\Delta = E - I$. Use the Binomial Theorem on the operator algebra to prove the explicit formula:

$$\Delta^n f(x) = \sum_{k=0}^{n}(-1)^{n-k}\binom{n}{k}f(x+k).$$

(b) Use this formula to calculate $\Delta^3 f(0)$ for the function $f(x) = x^3$. Check your answer by constructing a small difference table.

17. ⋆⋆ Newton's Series (Discrete Taylor Formula). We aim to prove that any polynomial $P(x)$ of degree d can be uniquely expressed as a sum of falling factorials:

$$P(x) = \sum \_k = 0^d c\_k x^{\underline{k}},$$

where the coefficients are given by $c_k = \frac{\Delta^k P(0)}{k!}$.

(a) **The Action on the Basis.** Recall that $\Delta x^{\underline{k}} = kx^{\underline{k-1}}$. Show that applying the difference operator j times to the basis element $x^{\underline{k}}$ yields:

$$\Delta^j(x^{\underline{k}})\Big|_{x=0} = \begin{cases} k! & \text{if } j = k \\ 0 & \text{if } j \neq k \end{cases}.$$

Remark. Consider the cases $j > k$ (derivative kills the term) and $j < k$ (term contains a factor of x, which becomes 0).

(b) **Isolating the Coefficient.** Apply the operator Δ^j to the entire summation $P(x) = \sum_{k=0}^d c_k x^{\underline{k}}$ and evaluate at $x = 0$ to prove the formula for c_k.

Chapter 6

Integer Functions

In our exploration of sums and recurrences, we frequently encounter expressions that require mapping real numbers to integers. To treat these rigorous discrete structures analytically, we introduce the floor and ceiling functions. These functions allow us to discretise continuous variables, a process fundamental to computer science and number theory.

6.1 Floors and Ceilings

We begin by defining the functions formally using the order properties of the integers.

Definition 6.1.1. ***Floor and Ceiling***. For any real number $x \in \mathbb{R}$:

1. The floor of x, denoted $\lfloor x \rfloor$, is the greatest integer less than or equal to x.

$$\lfloor x \rfloor := \max\{a \in \mathbb{Z} : a \leq x\}$$

2. The ceiling of x, denoted $\lceil x \rceil$, is the least integer greater than or equal to x.

$$\lceil x \rceil := \min\{a \in \mathbb{Z} : a \geq x\}$$

The existence and uniqueness of these values are guaranteed by the well-ordering principle and the Archimedean property of the real numbers.

Fundamental Properties

The utility of these functions arises from their algebraic properties and their interaction with inequalities.

***Proposition* 6.1.1.** *Basic Inequalities.* For all $x \in \mathbb{R}$ and $n \in \mathbb{Z}$:

(i) $\lfloor x \rfloor = x \iff x \in \mathbb{Z}$.
(ii) $\lceil x \rceil = x \iff x \in \mathbb{Z}$.
(iii) $x - 1 < \lfloor x \rfloor \leq x \leq \lceil x \rceil < x + 1$.
(iv) $\lfloor -x \rfloor = -\lceil x \rceil$.
(v) $\lceil -x \rceil = -\lfloor x \rfloor$.

The relationship between the floor and ceiling of a non-integer is strictly defined. We can express the difference $\lceil x \rceil - \lfloor x \rfloor$ using the Iverson bracket as $[x \notin \mathbb{Z}]$. Specifically:

$$\lceil x \rceil - \lfloor x \rfloor = \begin{cases} 0 & \text{if } x \in \mathbb{Z} \\ 1 & \text{if } x \notin \mathbb{Z} \end{cases} \tag{6.1}$$

Crucially, we can convert inequalities involving floors and ceilings into inequalities involving only x and integers. This allows us to "remove" the floor or ceiling brackets during derivations.

***Theorem* 6.1.1. Integer Bounds**. Let $x \in \mathbb{R}$ and $n \in \mathbb{Z}$.

(a) $\lfloor x \rfloor = n \iff n \leq x < n+1$.
(b) $\lceil x \rceil = n \iff n-1 < x \leq n$.
(c) $x < n \iff \lfloor x \rfloor < n$.
(d) $n \leq x \iff n \leq \lfloor x \rfloor$.
(e) $x \leq n \iff \lceil x \rceil \leq n$.
(f) $n < x \iff n < \lceil x \rceil$.

Proof. We prove (c) as an illustrative example. Let $x < n$. Since $\lfloor x \rfloor \leq x$, it follows immediately that $\lfloor x \rfloor < n$. Conversely, let $\lfloor x \rfloor < n$. From property (iii) of the Basic Inequalities, we know $x < \lfloor x \rfloor + 1$. Since $\lfloor x \rfloor$ and n are integers, $\lfloor x \rfloor < n$ implies $\lfloor x \rfloor \leq n-1$. Therefore, $x < (n-1)+1 = n$. The other properties follow similarly from the definitions. ■

Translation Invariance

Integers can be moved in and out of floor and ceiling brackets freely. However, this property does not hold for multiplication.

***Lemma* 6.1.1.** *Integer Translation.* For any $x \in \mathbb{R}$ and $n \in \mathbb{Z}$:

$$\lfloor x+n \rfloor = \lfloor x \rfloor + n \quad \text{and} \quad \lceil x+n \rceil = \lceil x \rceil + n$$

Proof. We prove the floor case. Let $m = \lfloor x \rfloor$. By definition, $m \leq x < m+1$. Adding n to all parts of the inequality yields:

$$m+n \leq x+n < m+n+1$$

By Theorem 6.1.1(a), this inequality implies that $\lfloor x+n \rfloor = m+n$. Substituting $m = \lfloor x \rfloor$ gives the result. ■

Note. This linearity applies only to additive integers. It is generally false that $\lfloor nx \rfloor = n\lfloor x \rfloor$. For example, if $n = 2$ and $x = 1/2$, then $\lfloor 2 \cdot 0.5 \rfloor = 1$ but $2\lfloor 0.5 \rfloor = 0$.

The Fractional Part

It is often useful to decompose a real number into its integer and non-integer components.

Definition 6.1.2. ***Fractional Part***. The fractional part of x is defined as:

$$\{x\} = x - \lfloor x \rfloor$$

It follows that $0 \leq \{x\} < 1$ and $x = \lfloor x \rfloor + \{x\}$.

If we write $x = n + \theta$ where $n \in \mathbb{Z}$ and $0 \leq \theta < 1$, then by the uniqueness of the floor function, it must be that $n = \lfloor x \rfloor$ and $\theta = \{x\}$.

Using this decomposition, we can analyse the behaviour of the floor function over a sum of real numbers. While $\lfloor x+y \rfloor$ is not generally $\lfloor x \rfloor + \lfloor y \rfloor$, the difference is bounded.

***Proposition* 6.1.2.** *Sum of Floors.* For any $x, y \in \mathbb{R}$:

$$\lfloor x+y \rfloor = \lfloor x \rfloor + \lfloor y \rfloor + \lfloor \{x\} + \{y\} \rfloor$$

Or equivalently:

$$\lfloor x+y \rfloor = \begin{cases} \lfloor x \rfloor + \lfloor y \rfloor & \text{if } \{x\} + \{y\} < 1 \\ \lfloor x \rfloor + \lfloor y \rfloor + 1 & \text{if } \{x\} + \{y\} \geq 1 \end{cases}$$

Proof. Let $x = \lfloor x \rfloor + \{x\}$ and $y = \lfloor y \rfloor + \{y\}$. Then:

$$\lfloor x + y \rfloor = \lfloor \lfloor x \rfloor + \lfloor y \rfloor + \{x\} + \{y\} \rfloor$$

Since $\lfloor x \rfloor + \lfloor y \rfloor$ is an integer, we can extract it using 6.1.1:

$$\lfloor x + y \rfloor = \lfloor x \rfloor + \lfloor y \rfloor + \lfloor \{x\} + \{y\} \rfloor$$

Since $0 \leq \{x\} < 1$ and $0 \leq \{y\} < 1$, the sum of fractional parts satisfies $0 \leq \{x\} + \{y\} < 2$. Consequently, $\lfloor \{x\} + \{y\} \rfloor$ can only be 0 or 1. ■

Binary Representations and Logarithms

The floor and ceiling functions appear naturally when analysing the binary representation of integers. Consider the number of bits required to represent a positive integer n.

Example 6.1.1. Consider $n = 35$. Since $2^5 = 32$ and $2^6 = 64$, we have $2^5 < 35 \leq 2^6$. Taking the base-2 logarithm, $5 < \log_2 35 \leq 6$. Using the property that $\lceil x \rceil = k \iff k - 1 < x \leq k$, we find $\lceil \log_2 35 \rceil = 6$. In binary, $35 = (100011)_2$, which has 6 bits.

Does the number of bits always equal $\lceil \log_2 n \rceil$? Consider $n = 32 = (100000)_2$. This has 6 bits. However, $\log_2 32 = 5$, so $\lceil \log_2 32 \rceil = 5$. The formula fails for powers of 2. The correct relation involves the floor function.

***Proposition* 6.1.3.** *Bit Length.* Let $n \in \mathbb{Z}^+$ and let m be the number of bits in the binary representation of n. Then:

$$m = \lfloor \log_2 n \rfloor + 1$$

Proof. If n has m bits, its leading bit represents 2^{m-1}. Thus, the value of n is bounded by:

$$2^{m-1} \leq n < 2^m$$

Taking logarithms base 2:

$$m - 1 \leq \log_2 n < m$$

By Theorem 6.1.1(a), this is equivalent to $\lfloor \log_2 n \rfloor = m - 1$. Thus $m = \lfloor \log_2 n \rfloor + 1$. ■

6.2 Nested Floors and Ceilings

We often encounter expressions where floor or ceiling functions are nested inside continuous functions. Under specific conditions, the inner floor or ceiling can be removed or simplified.

***Theorem* 6.2.1. Nested Square Roots.** For any non-negative real number $x \geq 0$:

$$\lfloor \sqrt{\lfloor x \rfloor} \rfloor = \lfloor \sqrt{x} \rfloor$$

Proof. Let $m = \lfloor \sqrt{\lfloor x \rfloor} \rfloor$. By definition:

$$m \leq \sqrt{\lfloor x \rfloor} < m + 1$$

Squaring all terms (since they are non-negative):

$$m^2 \leq \lfloor x \rfloor < (m + 1)^2$$

We now apply Theorem 6.1.1 to the inequalities.

- From $m^2 \leq \lfloor x \rfloor$, applying property (d), we get $m^2 \leq x$.
- From $\lfloor x \rfloor < (m + 1)^2$, applying property (c), we get $x < (m + 1)^2$.

Combining these, we have $m^2 \leq x < (m+1)^2$. Taking the square root:

$$m \leq \sqrt{x} < m+1$$

This is the definition of $m = \lfloor \sqrt{x} \rfloor$. Thus, the identity holds. ■

The analogous result $\lceil \sqrt{\lceil x \rceil} \rceil = \lceil \sqrt{x} \rceil$ also holds. We can generalise this property to a broader class of functions.

***Theorem* 6.2.2. Generalised Nesting**. Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous, monotonically increasing function with the property that:

$$f(x) \in \mathbb{Z} \implies x \in \mathbb{Z}$$

Then for all x in the domain:

$$\lfloor f(\lfloor x \rfloor) \rfloor = \lfloor f(x) \rfloor \quad \text{and} \quad \lceil f(\lceil x \rceil) \rceil = \lceil f(x) \rceil$$

Proof. We prove the ceiling case: $\lceil f(\lceil x \rceil) \rceil = \lceil f(x) \rceil$.If $x \in \mathbb{Z}$, then $\lceil x \rceil = x$ and the equality is trivial.Assume $x \notin \mathbb{Z}$. Then $x < \lceil x \rceil$. Since f is monotonically increasing, $f(x) < f(\lceil x \rceil)$. Since the ceiling function is non-decreasing, we have:

$$\lceil f(x) \rceil \leq \lceil f(\lceil x \rceil) \rceil$$

To prove equality, we assume for the sake of contradiction that $\lceil f(x) \rceil < \lceil f(\lceil x \rceil) \rceil$.

Remark. Remark: We assume the Intermediate Value Theorem from analysis (we prove this theorem in the analysis notes that comes after this notes).

Since f is continuous, there must exist a y such that $x < y < \lceil x \rceil$ and $f(y) = \lceil f(x) \rceil$.However, $\lceil f(x) \rceil$ is an integer, so $f(y) \in \mathbb{Z}$. By the hypothesis of the theorem, this implies $y \in \mathbb{Z}$.But there are no integers strictly between x and $\lceil x \rceil$. This contradiction implies that our assumption was false. Therefore, $\lceil f(x) \rceil = \lceil f(\lceil x \rceil) \rceil$. ■

Example 6.2.1. Consider the function $f(x) = \frac{x+m}{n}$ for integers n, m. This function is continuous and increasing. Property Check: If $\frac{x+m}{n} = k \in \mathbb{Z}$, then $x = nk - m$, which is an integer. Thus, the condition holds. We can therefore apply the Generalised Nesting theorem to derive the useful identity:

$$\left\lfloor \frac{\lfloor x \rfloor + m}{n} \right\rfloor = \left\lfloor \frac{x+m}{n} \right\rfloor$$

6.3 Integers in Intervals

A fundamental application of the floor and ceiling functions is the enumeration of integers within specific ranges. To facilitate this, we adopt a notation that explicitly distinguishes intervals of real numbers from discrete sets of integers.

Notation 6.3.1. Interval Notation Let $\alpha, \beta \in \mathbb{R}$. We denote intervals of real numbers using standard bracket notation:

$$[\alpha, \beta] = \{x \in \mathbb{R} : \alpha \leq x \leq \beta\}$$
$$[\alpha, \beta) = \{x \in \mathbb{R} : \alpha \leq x < \beta\}$$
$$(\alpha, \beta] = \{x \in \mathbb{R} : \alpha < x \leq \beta\}$$
$$(\alpha, \beta) = \{x \in \mathbb{R} : \alpha < x < \beta\}$$

When we wish to refer specifically to the set of *integers* within an interval, we employ the notation $[\alpha \dots \beta]$. For example:

$$[\alpha \dots \beta] = [\alpha, \beta] \cap \mathbb{Z} = \{n \in \mathbb{Z} : \alpha \leq n \leq \beta\}$$

Counting Integers

We seek a rigorous method to calculate the cardinality of these integer sets. Using the Iverson bracket and the properties of floors and ceilings derived in the previous chapter, we can establish precise counting formulas.

Consider the half-open interval $[\alpha \dots \beta)$. The number of integers in this set is given by the sum:

$$\sum_n [n \in [\alpha \dots \beta)] = \sum_n [\alpha \le n < \beta]$$

Recall the integer bounds properties:

- $\alpha \le n \iff \lceil \alpha \rceil \le n$
- $n < \beta \iff n < \lceil \beta \rceil$

Thus, the condition $\alpha \le n < \beta$ is equivalent to $\lceil \alpha \rceil \le n < \lceil \beta \rceil$. The sum becomes:

$$\sum_{n=\lceil \alpha \rceil}^{\lceil \beta \rceil - 1} 1$$

This is a sum of ones, and the number of terms is simply the upper bound minus the lower bound.

***Theorem* 6.3.1. Integers in Intervals**. Let $\alpha, \beta \in \mathbb{R}$ with $\alpha \le \beta$. The number of integers in the standard intervals is given by:

(i) $|[\alpha \dots \beta)| = \lceil \beta \rceil - \lceil \alpha \rceil$
(ii) $|(\alpha \dots \beta]| = \lfloor \beta \rfloor - \lfloor \alpha \rfloor$
(iii) $|[\alpha \dots \beta]| = \lfloor \beta \rfloor - \lceil \alpha \rceil + 1$
(iv) $|(\alpha \dots \beta)| = \lceil \beta \rceil - \lfloor \alpha \rfloor - 1$ (assuming $\alpha < \beta$)

Proof. We have derived (i) above. For (ii), observe that $n \in (\alpha, \beta] \iff \alpha < n \le \beta$. Using integer bound properties, this is equivalent to $\lfloor \alpha \rfloor < n \le \lfloor \beta \rfloor$, or $\lfloor \alpha \rfloor + 1 \le n \le \lfloor \beta \rfloor$. The count is $\lfloor \beta \rfloor - (\lfloor \alpha \rfloor + 1) + 1 = \lfloor \beta \rfloor - \lfloor \alpha \rfloor$. Formula (iii) follows from $\lceil \alpha \rceil \le n \le \lfloor \beta \rfloor$. Formula (iv) follows from $\lfloor \alpha \rfloor < n < \lceil \beta \rceil$, which implies $\lfloor \alpha \rfloor + 1 \le n \le \lceil \beta \rceil - 1$. ■

6.4 The Casino Problem

We now apply our knowledge of summation, integer functions, and intervals to solve a more complex problem involving number theory and combinatorics.

Definition 6.4.1. *The Setup*. Consider a roulette wheel with 1,000 slots numbered $1, \dots, 1000$. A number n is deemed a **winner** if it is divisible by the floor of its cube root. That is:

$$\lfloor \sqrt[3]{n} \rfloor \mid n$$

We assume a "Strong Rule" where each number comes up exactly once in 1000 plays. If the Casino pays \$5 for a win and collects \$1 for a loss, does the player have an advantage?

To determine the player's advantage, we must calculate the total number of winning numbers, denoted W.

$$W = \sum_{n=1}^{1000} [\text{n is a winner}] = \sum_{n=1}^{1000} [\lfloor \sqrt[3]{n} \rfloor \mid n]$$

If W is sufficiently large (specifically, if $5W - (1000 - W) > 0 \implies 6W > 1000 \implies W > 166$), the player wins in the long run.

Derivation of the Solution

- **Change of Variables** The term $\lfloor \sqrt[3]{n} \rfloor$ changes value much more slowly than n, so it is convenient to introduce a new variable
$$k = \lfloor \sqrt[3]{n} \rfloor.$$
The condition $k = \lfloor \sqrt[3]{n} \rfloor$ is equivalent to
$$k \le \sqrt[3]{n} < k+1 \quad \Longleftrightarrow \quad k^3 \le n < (k+1)^3.$$
We rewrite the sum using the Generalised Distributive Law for sums, explicitly iterating over both k and n:
$$W = \sum_k \sum_n [k = \lfloor \sqrt[3]{n} \rfloor] \, [k \mid n] \, [1 \le n \le 1000]. \tag{6.2}$$
Substituting the inequality form for the floor function,
$$W = \sum_k \sum_n [k^3 \le n < (k+1)^3] \, [k \mid n] \, [1 \le n \le 1000].$$

- **Handling the Bounds** Since $1 \le n \le 1000$, the value
$$k = \lfloor \sqrt[3]{n} \rfloor$$
ranges from $\lfloor \sqrt[3]{1} \rfloor = 1$ up to $\lfloor \sqrt[3]{1000} \rfloor = 10$. For a given integer k, the corresponding n lie in the interval
$$[k^3, (k+1)^3).$$
Concretely:
  - For $k = 1$, the interval is $[1, 8)$.
  - For $k = 9$, the interval is $[729, 1000)$.
  - For $k = 10$, the interval is $[1000, 1331)$.

The condition $1 \le n \le 1000$ interacts only with the top interval:
  - For $k < 10$, we have $(k+1)^3 \le 10^3 = 1000$, so the full interval $[k^3, (k+1)^3)$ is contained in $[1, 1000]$.
  - For $k = 10$, the interval is $[1000, 1331)$, so the only possible n in our range is $n = 1000$.

We therefore split off the boundary case $k = 10$ and then sum over $1 \le k < 10$.

The Boundary Case ($k = 10$). The only candidate is $n = 1000$. We check whether it is a winner:
$$\lfloor \sqrt[3]{1000} \rfloor \mid 1000 \quad \Longleftrightarrow \quad 10 \mid 1000,$$
which is true. Thus this contributes exactly 1 to W.

The Main Sum ($1 \le k < 10$). For $1 \le k < 10$, the constraint $n \le 1000$ is redundant, since $(k+1)^3 \le 10^3 = 1000$. Thus, any integer n satisfying $k^3 \le n < (k+1)^3$ automatically satisfies $n \le 1000$. Hence the summation simplifies to:
$$W = 1 + \sum_{k=1}^{9} \sum_n [k^3 \le n < (k+1)^3] \, [k \mid n].$$

- **Summing Multiples** The condition $[k \mid n]$ means n must be a multiple of k. Write $n = km$ for some integer m. Substituting into the inequalities,
$$k^3 \le km < (k+1)^3 \quad \Longleftrightarrow \quad k^2 \le m < \frac{(k+1)^3}{k}.$$
Thus, for a fixed k, the admissible integers m are precisely those in the half-open interval
$$\left[k^2, \ \frac{(k+1)^3}{k} \right).$$

Hence

$$W = 1 + \sum_{k=1}^{9} \sum_{m} \left[m \in \left[k^2,\ \frac{(k+1)^3}{k} \right) \right].$$

By Theorem 6.3.1(i), the number of integers in an interval $[a, b)$ is

$$\#\{n \in \mathbb{Z} : a \le n < b\} = \lceil b \rceil - \lceil a \rceil.$$

Applying this with $a = k^2$ and $b = \dfrac{(k+1)^3}{k}$, we obtain

$$\text{Count}_k = \left\lceil \frac{(k+1)^3}{k} \right\rceil - \lceil k^2 \rceil.$$

Since k is an integer, $\lceil k^2 \rceil = k^2$. We now simplify the upper endpoint:

$$\frac{(k+1)^3}{k} = \frac{k^3 + 3k^2 + 3k + 1}{k} = k^2 + 3k + 3 + \frac{1}{k}.$$

Using the translation property $\lceil x + n \rceil = \lceil x \rceil + n$ (for integer n),

$$\left\lceil k^2 + 3k + 3 + \frac{1}{k} \right\rceil = k^2 + 3k + 3 + \left\lceil \frac{1}{k} \right\rceil.$$

For $k \ge 1$ we have $0 < \dfrac{1}{k} \le 1$, so $\lceil 1/k \rceil = 1$. Therefore

$$\left\lceil \frac{(k+1)^3}{k} \right\rceil = k^2 + 3k + 4,$$

and the number of valid m for this k is

$$\text{Count}_k = (k^2 + 3k + 4) - k^2 = 3k + 4.$$

- **Final Calculation** Substituting back into the expression for W,

$$W = 1 + \sum_{k=1}^{9} (3k + 4).$$

This is a straightforward arithmetic series:

$$\begin{aligned} W &= 1 + 3 \sum_{k=1}^{9} k + \sum_{k=1}^{9} 4 \\ &= 1 + 3 \cdot \frac{9 \cdot 10}{2} + 4 \cdot 9 \\ &= 1 + 135 + 36 \\ &= 172. \end{aligned}$$

Thus there are $W = 172$ winning numbers between 1 and 1000. The player's net expected gain over 1000 spins under the Strong Rule is

$$6W - 1000 = 6 \cdot 172 - 1000 = 1032 - 1000 = 32 > 0.$$

Equivalently, the expected profit per spin is

$$\frac{32}{1000} = 0.032 \text{ dollars} = 3.2 \text{ cents}.$$

Since $W = 172 \ge 167$, the player indeed has a statistical advantage.

6.5 Spectrum Partitions

An intriguing application of the floor function is the partitioning of the set of integers based on irrational multipliers. This concept was formalised by H.J.S. Smith in 1876 and is known as the theory of *spectra*.

Definition 6.5.1. ***Spectrum***. For any real number α, the spectrum of α, denoted $\text{Spec}(\alpha)$, is the sequence of integers:

$$\text{Spec}(\alpha) = \{\lfloor \alpha \rfloor, \lfloor 2\alpha \rfloor, \lfloor 3\alpha \rfloor, \dots\}$$

Formally, as a set: $\text{Spec}(\alpha) = \{\lfloor n\alpha \rfloor : n \in \mathbb{Z}^+\}$.

Note that if $\alpha < 1$, the sequence is not strictly increasing and $\text{Spec}(\alpha)$ forms a multiset. For example, $\text{Spec}(1/2) = \{0, 1, 1, 2, 2, \dots\}$. If $\alpha \geq 1$, the elements are distinct.

Example 6.5.1. Spectra of $\sqrt{2}$ and $2 + \sqrt{2}$. Let us compute the first few terms of the spectra for $\alpha = \sqrt{2} \approx 1.414$ and $\beta = 2 + \sqrt{2} \approx 3.414$.

- $\text{Spec}(\sqrt{2}) = \{1, 2, 4, 5, 7, 8, 9, 11, 12, 14, 15, 16, \dots\}$
- $\text{Spec}(2 + \sqrt{2}) = \{3, 6, 10, 13, 17, 20, \dots\}$

Comparing these two sets reveals a remarkable pattern:

- They appear disjoint: no number seems to belong to both.
- Their union seems to cover all positive integers: $1, 2, (3), 4, 5, (6), 7, 8, 9, (10), \dots$

This observation suggests that these two spectra partition the set of positive integers $\mathbb{Z}^+$.

The Partition Theorem

The phenomenon observed above is not a coincidence. It is a specific instance of a general theorem concerning irrational numbers.

***Theorem* 6.5.1. Beatty's Theorem (Spectrum Partition).** Let $\alpha, \beta \in \mathbb{R} \setminus \mathbb{Q}$ be positive irrational numbers such that:

$$\frac{1}{\alpha} + \frac{1}{\beta} = 1$$

Then the sets $A = \text{Spec}(\alpha)$ and $B = \text{Spec}(\beta)$ form a partition of $\mathbb{Z}^+$. That is:

1. $A \cap B = \emptyset$ (Disjointness)
2. $A \cup B = \mathbb{Z}^+$ (Covering)

Proof. **Part 1: Disjointness.** We proceed by contradiction. Assume there exists an integer $k \in A \cap B$. By definition, there exist integers $n, m \in \mathbb{Z}^+$ such that:

$$k = \lfloor n\alpha \rfloor \quad \text{and} \quad k = \lfloor m\beta \rfloor$$

Using the inequality property of the floor function ($\lfloor x \rfloor = k \iff k \leq x < k + 1$):

$$k \leq n\alpha < k + 1 \tag{6.3}$$

$$k \leq m\beta < k + 1 \tag{6.4}$$

Since α and β are irrational, $n\alpha$ and $m\beta$ cannot be integers. Therefore, the strict inequalities hold:

$$k < n\alpha < k + 1 \quad \text{and} \quad k < m\beta < k + 1$$

Dividing (6.3) by α and (6.4) by β:

$$\frac{k}{\alpha} < n < \frac{k+1}{\alpha} \quad \text{and} \quad \frac{k}{\beta} < m < \frac{k+1}{\beta}$$

Adding these two inequalities together:

$$k\left(\frac{1}{\alpha}+\frac{1}{\beta}\right) < n+m < (k+1)\left(\frac{1}{\alpha}+\frac{1}{\beta}\right)$$

Substitute the hypothesis $1/\alpha + 1/\beta = 1$:

$$k < n+m < k+1$$

This implies that the integer $n+m$ lies strictly between the consecutive integers k and $k+1$, which is impossible. Thus, A and B must be disjoint.

Part 2: Covering. The idea is to count how many elements of A and B lie in the range $\{1,\ldots,N\}$ for any arbitrary N. Let $A_N = A \cap [1,N]$ and $B_N = B \cap [1,N]$. Let $a(N) = |A_N|$ and $b(N) = |B_N|$. We will show that $a(N)+b(N) = N$. Since A and B are disjoint subsets of $\mathbb{Z}^+$, if their cardinalities sum to N within the range $[1,N]$, they must partition the range.

Observe that $\lfloor n\alpha \rfloor \leq N \iff n\alpha < N+1 \iff n < \frac{N+1}{\alpha}$. Since $n \in \mathbb{Z}^+$, the number of such values is exactly:

$$a(N) = \left\lfloor \frac{N+1}{\alpha} \right\rfloor$$

Similarly, for β:

$$b(N) = \left\lfloor \frac{N+1}{\beta} \right\rfloor$$

We compute the sum:

$$a(N)+b(N) = \left\lfloor \frac{N+1}{\alpha} \right\rfloor + \left\lfloor \frac{N+1}{\beta} \right\rfloor$$

Let $x = \frac{N+1}{\alpha}$ and $y = \frac{N+1}{\beta}$. By hypothesis, $x+y = (N+1)(1/\alpha+1/\beta) = N+1$. Since α, β are irrational, x and y are irrational. We can write $x = \lfloor x \rfloor + \{x\}$ and $y = \lfloor y \rfloor + \{y\}$ with $0 < \{x\}, \{y\} < 1$. Summing them:

$$x+y = \lfloor x \rfloor + \lfloor y \rfloor + \{x\} + \{y\} = N+1$$

Thus $\{x\} + \{y\} = N+1-(\lfloor x \rfloor + \lfloor y \rfloor)$, which is an integer. Since $0 < \{x\}+\{y\} < 2$, the only possible integer value is 1.

$$\{x\}+\{y\} = 1 \implies \lfloor x \rfloor + \lfloor y \rfloor = (N+1)-1 = N$$

Therefore, $a(N)+b(N) = N$. Since this holds for all N, every integer is included in exactly one of the sets A or B. ■

This theorem confirms our initial observation for $\alpha = \sqrt{2}$ and $\beta = 2+\sqrt{2}$, since:

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

Counting Elements in Spectra

Before discussing finite partitions, let us derive a formula for the number of elements in a spectrum up to a certain value n. Let $N(\alpha, n) = |\{m \in \mathrm{Spec}(\alpha) : m \leq n\}|$.

$$N(\alpha,n) = \sum_{k\geq 1} [\lfloor k\alpha \rfloor \leq n]$$

Using the integer bound property $\lfloor x \rfloor \leq n \iff x < n+1$:

$$N(\alpha,n) = \sum_{k\geq 1} [k\alpha < n+1] = \sum_{k\geq 1} \left[k < \frac{n+1}{\alpha}\right]$$

This sum counts the number of positive integers strictly less than $(n+1)/\alpha$.

$$N(\alpha, n) = \left\lceil \frac{n+1}{\alpha} \right\rceil - 1 \tag{6.5}$$

Applying this to our partition problem, the total number of elements $\leq n$ in both sets is:

$$|A_n| + |B_n| = \left(\left\lceil \frac{n+1}{\alpha} \right\rceil - 1 \right) + \left(\left\lceil \frac{n+1}{\beta} \right\rceil - 1 \right)$$

Using the identity $\lceil x \rceil + \lceil y \rceil = x + y$ proved in the context of the theorem (since $1/\alpha + 1/\beta = 1$ implies $(n+1)/\alpha + (n+1)/\beta = n+1 \in \mathbb{Z}$), we find $|A_n| + |B_n| = n$. This confirms that for every finite n, the two sets account for exactly n integers, consistent with the partition property.

6.6 Floor and Ceiling Sums

We conclude with a classic summation involving the floor of a square root:

$$S_n = \sum_{k=0}^{n-1} \lfloor \sqrt{k} \rfloor$$

We begin by reindexing the sum using a new variable $m = \lfloor \sqrt{k} \rfloor$. The condition $m = \lfloor \sqrt{k} \rfloor$ is equivalent to the inequality $m \leq \sqrt{k} < m+1$, which upon squaring becomes:

$$m^2 \leq k < (m+1)^2$$

Using Iverson brackets to express the constraint, we rewrite the sum as:

$$\begin{aligned} S_n &= \sum_{k=0}^{n-1} \sum_{m \geq 0} m \, [m = \lfloor \sqrt{k} \rfloor] \\ &= \sum_{m \geq 0} m \sum_{k=0}^{n-1} [m^2 \leq k < (m+1)^2] \end{aligned}$$

The inner sum counts the number of integers $k \in [0, n-1]$ that satisfy $m^2 \leq k < (m+1)^2$. This is equivalent to finding the cardinality of the intersection of intervals:

$$|[m^2, (m+1)^2) \cap [0, n)|$$

The Case of a Perfect Square

Assume first that $n = a^2$ is a perfect square. The variable k ranges over $[0, a^2)$. For any m in the range $0 \leq m \leq a-1$, the entire interval $[m^2, (m+1)^2)$ lies strictly within $[0, a^2)$. Conversely, for $m \geq a$, the interval is disjoint from $[0, a^2)$ and contributes nothing to the sum.

For each valid m, the number of integers k is simply the length of the interval:

$$(m+1)^2 - m^2 = 2m + 1$$

Thus, the sum simplifies to:

$$S_{a^2} = \sum_{m=0}^{a-1} m(2m+1) = \sum_{m=0}^{a-1} (2m^2 + m)$$

We evaluate this using standard summation formulas for consecutive integers and squares:

$$\sum_{m=0}^{a-1} m = \frac{(a-1)a}{2}, \quad \sum_{m=0}^{a-1} m^2 = \frac{(a-1)a(2a-1)}{6}$$

Substituting these into the expression for S_{a^2}:

$$\begin{aligned} S_{a^2} &= 2\left(\frac{(a-1)a(2a-1)}{6}\right) + \frac{(a-1)a}{2} \\ &= \frac{a(a-1)}{6}\left(2(2a-1)+3\right) \\ &= \frac{a(a-1)(4a+1)}{6} \end{aligned}$$

The General Case

Let n be an arbitrary positive integer. We define $a = \lfloor\sqrt{n}\rfloor$. It follows that $a^2 \leq n < (a+1)^2$. The sum can be partitioned into two segments: the terms up to the largest perfect square less than n, and the remaining terms.

$$S_n = \sum_{k=0}^{a^2-1} \lfloor\sqrt{k}\rfloor + \sum_{k=a^2}^{n-1} \lfloor\sqrt{k}\rfloor$$

The first part corresponds exactly to S_{a^2}, which we computed above.

$$\sum_{k=0}^{a^2-1} \lfloor\sqrt{k}\rfloor = S_{a^2} = \frac{a(a-1)(4a+1)}{6}$$

For the second part, the index k ranges from a^2 to $n-1$. In this range, we have:

$$a^2 \leq k \leq n-1 < n < (a+1)^2$$

Taking the square root implies $a \leq \sqrt{k} < a+1$, so $\lfloor\sqrt{k}\rfloor = a$ for all terms in this summation. The number of terms is $(n-1) - a^2 + 1 = n - a^2$. Therefore:

$$\sum_{k=a^2}^{n-1} \lfloor\sqrt{k}\rfloor = a(n-a^2)$$

Combining these results yields the general closed-form solution.

***Theorem* 6.6.1. Sum of Integer Roots**. For any integer $n \geq 1$, let $a = \lfloor\sqrt{n}\rfloor$. Then:

$$\sum_{k=0}^{n-1} \lfloor\sqrt{k}\rfloor = \frac{a(a-1)(4a+1)}{6} + a(n-a^2)$$

Proof. Let

$$S_n = \sum_{k=0}^{n-1} \lfloor\sqrt{k}\rfloor \quad \text{and} \quad a = \lfloor\sqrt{n}\rfloor.$$

By definition of the floor function, for any integer $m \geq 0$ we have

$$\lfloor\sqrt{k}\rfloor = m \iff m \leq \sqrt{k} < m+1 \iff m^2 \leq k < (m+1)^2.$$

So the values of k with $\lfloor\sqrt{k}\rfloor = m$ are exactly the integers in the interval

$$[m^2, (m+1)^2).$$

Step 1: Determine which m occur. Since $0 \leq k \leq n-1$, we are only interested in values of m for which the interval

$$[m^2, (m+1)^2)$$

intersects $[0, n)$.

By the definition of $a = \lfloor \sqrt{n} \rfloor$ we have

$$a^2 \le n < (a+1)^2.$$

Thus:

- For $0 \le m \le a-1$, we have $(m+1)^2 \le a^2 \le n$, so the whole interval $[m^2, (m+1)^2)$ lies inside $[0, n)$.
- For $m = a$, the interval is $[a^2, (a+1)^2)$. Its intersection with $[0, n)$ is $[a^2, n)$, which is non-empty.
- For $m \ge a+1$, we have $m^2 \ge (a+1)^2 > n$, so $[m^2, (m+1)^2)$ does not intersect $[0, n)$ at all.

Therefore, only $m = 0, 1, \ldots, a$ contribute to the sum.

Step 2: Count how many k give each m. For $0 \le m \le a-1$, the integers k with $\lfloor \sqrt{k} \rfloor = m$ are exactly

$$k = m^2, m^2+1, \ldots, (m+1)^2 - 1.$$

The number of such k is the length of this interval:

$$(m+1)^2 - m^2 = 2m + 1.$$

For $m = a$, the integers k lie in the intersection $[a^2, (a+1)^2) \cap [0, n) = [a^2, n)$, so they are

$$k = a^2, a^2+1, \ldots, n-1,$$

and the number of such k is

$$(n-1) - a^2 + 1 = n - a^2.$$

Step 3: Write S_n as a sum over m. We now sum m times the count of k for which $\lfloor \sqrt{k} \rfloor = m$:

$$S_n = \sum_{k=0}^{n-1} \lfloor \sqrt{k} \rfloor = \sum_{m=0}^{a-1} m(2m+1) \; + \; a(n - a^2).$$

Step 4: Evaluate the polynomial sum. We simplify

$$\sum_{m=0}^{a-1} m(2m+1) = \sum_{m=0}^{a-1} (2m^2 + m) = 2 \sum_{m=0}^{a-1} m^2 + \sum_{m=0}^{a-1} m.$$

Using the standard formulae

$$\sum_{m=0}^{a-1} m = \frac{(a-1)a}{2}, \qquad \sum_{m=0}^{a-1} m^2 = \frac{(a-1)a(2a-1)}{6},$$

we obtain

$$\begin{aligned} \sum_{m=0}^{a-1} m(2m+1) &= 2 \cdot \frac{(a-1)a(2a-1)}{6} + \frac{(a-1)a}{2} \\ &= \frac{(a-1)a(2a-1)}{3} + \frac{(a-1)a}{2} \\ &= \frac{a(a-1)}{6}\big(2(2a-1) + 3\big) \\ &= \frac{a(a-1)(4a+1)}{6}. \end{aligned}$$

Step 5: Combine everything. Substituting back into the expression for S_n gives

$$S_n = \frac{a(a-1)(4a+1)}{6} \; + \; a(n - a^2),$$

where $a = \lfloor \sqrt{n} \rfloor$. ■

6.7 Exercises

1. **Floor and Ceiling Symmetry.** Show that the expression
$$\left\lfloor \frac{2x+1}{2} \right\rfloor - \left\lfloor \frac{2x+1}{4} \right\rfloor + \left\lfloor \frac{2x+1}{4} \right\rfloor$$
simplifies to either $\lfloor x \rfloor$ or $\lceil x \rceil$. Which one is it, and why?

2. **Ceiling-Floor Conversion.** Prove that for all integers n and all positive integers m:
$$\left\lceil \frac{n}{m} \right\rceil = \left\lfloor \frac{n+m-1}{m} \right\rfloor$$
This identity provides an algebraic method to convert ceilings to floors without using the reflection law involving negative arguments.

3. **Integers in Open Intervals.** Prove that the open interval (α, β), where $\alpha < \beta$, contains exactly $\lceil \beta \rceil - \lfloor \alpha \rfloor - 1$ integers. Explain why the case $\alpha = \beta$ must be excluded from this formula.

4. **Logarithmic Floors.** Find a necessary and sufficient condition on the real number $b > 1$ such that
$$\lfloor \log_b \lfloor x \rfloor \rfloor = \lfloor \log_b x \rfloor$$
for all real $x \geq 1$. Formally, determine a predicate $P(b)$ such that the equality holds for all $x \geq 1$ if and only if $P(b)$ is true.

5. **Repetitive Sequence.** Consider the sequence where the integer k appears k times:
$$1, 2, 2, 3, 3, 3, 4, 4, 4, 4, 5, 5, 5, 5, 5, \ldots$$
Show that the n-th term of this sequence is given by:
$$a_n = \left\lfloor \sqrt{2n} + \frac{1}{2} \right\rfloor$$
Remark. Determine the range of indices n for which the value is m.

6. **Floor Inequality.** Prove or disprove the following inequality for all real numbers x, y:
$$\lfloor x \rfloor + \lfloor y \rfloor + \lfloor x + y \rfloor \leq \lfloor 2x \rfloor + \lfloor 2y \rfloor$$

7. ⋆ **Complex Logarithmic Sum.** Assuming n is a non-negative integer, find a closed form for the sum:
$$\sum_{1 \leq k < 2^n} \left\lfloor \frac{1}{\left\lfloor \frac{2^{\lfloor \log_2 k \rfloor}}{k} \right\rfloor 4^{\lfloor \log_2 k \rfloor}} \right\rfloor$$
Remark. This expression looks daunting, but notice that for a fixed $m = \lfloor \log_2 k \rfloor$, the term 2^m is constant. Group the terms by the value of $\lfloor \log_2 k \rfloor$.

8. ⋆ **Summing Floors.** Evaluate the sum
$$\sum_{0 \leq k < m} \left\lfloor \frac{x+k}{m} \right\rfloor$$
in the case where $x \geq 0$. Use the substitution $\lfloor y \rfloor = \sum_j [1 \leq j \leq y]$ to rewrite the term $\lfloor (x+k)/m \rfloor$, then swap the order of summation to sum over k first. Does your result agree with the identity $\lfloor nx \rfloor = \sum_{0 \leq k < n} \lfloor x + k/n \rfloor$?

9. ⋆⋆ **Square Root Recurrence.** Solve the recurrence relation defined by:
$$a_0 = 1; \quad a_n = a_{n-1} + \lfloor \sqrt{a_{n-1}} \rfloor \quad \text{for } n > 0$$
Find an explicit formula for a_n or describe the sequence's behaviour.

Chapter 7

Number Theory

Note. This Chapter should be revision most things have already been proven in my previous notes so i wont over bloat the chapter.

We restrict our universe of discourse to the integers $\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\}$ and the natural numbers $\mathbb{N} = \{0, 1, 2, \dots\}$.

7.1 Divisibility and The Division Algorithm

The primary structural characteristic of the integers is multiplicative. We begin by formalising the concept of one number being a "part" of another.

***Axiom* 7.1.1. The Well-Ordering Principle**. Every non-empty subset of $\mathbb{N}$ contains a least element.

Definition 7.1.1. ***Divisibility***. Let $a, b \in \mathbb{Z}$. We say that a divides b, denoted $a \mid b$, if there exists an integer c such that:

$$b = a \cdot c$$

In this context, a is a *divisor* of b, and b is a *multiple* of a. If no such integer c exists, we write $a \nmid b$.

Note. It follows immediately that 1 and -1 divide every integer, and every non-zero integer divides 0.

Divisors naturally occur in pairs: if $a \mid b$, then $(-a) \mid b$, as $b = (-a)(-k)$. We call $(-a, -k)$ the *associated divisors* to (a, k). Usually, we restrict our attention to positive divisors to avoid redundancy.

Fundamental Properties of Divisibility

The divisibility relation is transitive and linear.

***Proposition* 7.1.1.** *Basic Facts of Divisibility.* For integers m, n, k:

1. **Antisymmetry**: If $m \mid n$ and $n \mid m$, then $m = \pm n$.
2. **Linearity**: If $m \mid n_1$ and $m \mid n_2$, then m divides any linear combination of them:

$$m \mid (un_1 + vn_2) \quad \text{for all } u, v \in \mathbb{Z}$$

3. **Transitivity**: If $m \mid n$ and $n \mid k$, then $m \mid k$.
4. **Bounds**: If $m \mid n$ and $n \neq 0$, then $|m| \leq |n|$.

A crucial observation for primality testing and factorisation algorithms is the distribution of divisors relative to the square root.

***Lemma* 7.1.1.** *The Square Root Bound.* If $n > 0$ is composite, it possesses a divisor m such that $1 < m \leq \sqrt{n}$. Equivalently, if (m, k) is a divisor pair of n, then $\min(|m|, |k|) \leq \sqrt{n}$.

Proof. Assume for contradiction that both $|m| > \sqrt{n}$ and $|k| > \sqrt{n}$. Then $|mk| > \sqrt{n} \cdot \sqrt{n} = n$, which contradicts $n = mk$. ■

Example 7.1.1. Divisors of 60. To find all divisors of $n = 60$, we need only test integers $m \leq \sqrt{60} \approx 7.7$. Checking $m \in \{1, 2, 3, 4, 5, 6, 7\}$, we find the pairs:

$$(1, 60), (2, 30), (3, 20), (4, 15), (5, 12), (6, 10)$$

Since 7 does not divide 60, we stop.

The Division Algorithm

Often, an integer a does not divide b perfectly. In such cases, we wish to quantify the "error" or residue left over. This is formalised by the Division Theorem, attributed to Euclid (circa 300 BCE).

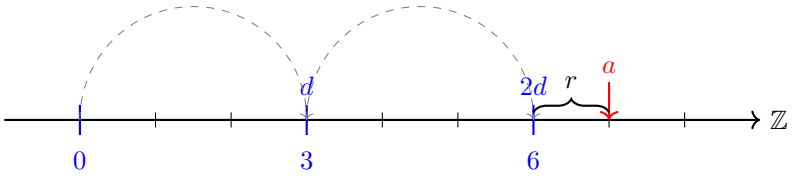


Figure 7.1: Visualisation of $a = dq + r$. Here $a = 7$ and $d = 3$. The quotient $q = 2$ represents the number of full steps of size d, and $r = 1$ is the remainder.

***Theorem* 7.1.1. Euclid's Division Theorem**. For any integers a and $d \neq 0$, there exist unique integers q (the quotient) and r (the remainder) such that:

$$a = dq + r \quad \text{and} \quad 0 \leq r < |d|$$

Proof. **Existence**: Consider the set S of all non-negative integers that can be represented as $a - dt$ for some integer t:

$$S = \{a - dt \mid t \in \mathbb{Z}\} \cap \mathbb{N}$$

First, we show S is non-empty.

- If $a \geq 0$, choose $t = 0$. Then $a - d(0) = a \in S$.
- If $a < 0$, choose $t = -|a|d$. Since $d \neq 0$, $|d| \geq 1$, and the term $a - d(-|a|d) = a + d^2|a| \geq a + |a| \geq 0$. Thus the value is in S.

By the Well-Ordering Principle, every non-empty set of non-negative integers contains a least element. Let r be the minimum element of S. By definition, there exists some q such that $r = a - dq$, which implies $a = dq + r$. Since $r \in S$, $r \geq 0$.

We claim $r < |d|$. Suppose for the sake of contradiction that $r \geq |d|$. Consider $r' = r - |d|$.

$$r' = (a - dq) - |d| = a - d(q \pm 1),$$

Where

- If $d > 0$, then $|d| = d$ and $r' = (a - dq) - d = a - d(q + 1)$.
- If $d < 0$, then $|d| = -d$ and $r' = (a - dq) - (-d) = a - d(q - 1)$.

In both cases, r' can be written in the form $a - dt$ for some integer t (namely $t = q + 1$ if $d > 0$ and $t = q - 1$ if $d < 0$). Thus $r' \in S$. But $0 \leq r' < r$, which contradicts the minimality of r in S. Hence our assumption $r \geq |d|$ is false, and we must have $0 \leq r < |d|$.

Uniqueness: Suppose there exist two pairs (q, r) and (q', r') satisfying the conditions.

$$a = dq + r = dq' + r'$$

Without loss of generality, assume $r \geq r'$.

$$d(q - q') = r' - r$$

Taking absolute values:

$$|d| \cdot |q - q'| = |r' - r|$$

Since $0 \leq r, r' < |d|$, the distance $|r' - r|$ must be strictly less than $|d|$. The only multiple of $|d|$ strictly less than $|d|$ is 0. Thus, $|r' - r| = 0 \implies r = r'$. It follows immediately that $d(q - q') = 0$, and since $d \neq 0$, $q = q'$. ■

We can formalise the remainder operation using the floor function introduced in the previous chapter.

Definition 7.1.2. ***The Modulo Operator***. For $a \in \mathbb{Z}$ and $b \neq 0$, we define the binary operation mod as:

$$a \bmod b = a - b \left\lfloor \frac{a}{b} \right\rfloor$$

This yields the unique remainder r defined in the Division Theorem.

Residue Classes: Squares Modulo 4

The remainder classification allows us to partition the integers. For example, any integer n is either even ($n = 2k$) or odd ($n = 2k + 1$). This partition has arithmetic consequences.

***Proposition* 7.1.2.** *Squares Modulo 4.* The square of any integer n is either divisible by 4 or leaves a remainder of 1 when divided by 4.

Proof.

- **Case 1**: n is even ($n = 2k$). Then $n^2 = (2k)^2 = 4k^2$. The remainder is 0.
- **Case 2**: n is odd ($n = 2k + 1$). Then $n^2 = 4k^2 + 4k + 1 = 4(k^2 + k) + 1$. The remainder is 1.

Consequently, no integer of the form $4k + 2$ or $4k + 3$ can be a perfect square. ■

7.2 Radix Representation

A fundamental application of the Division Algorithm is the representation of numbers in different bases. Given a base $b \in \mathbb{N}, b > 1$, any positive integer n can be expressed uniquely as:

$$n = a_k b^k + a_{k-1} b^{k-1} + \cdots + a_1 b^1 + a_0$$

where the coefficients (digits) satisfy $0 \leq a_i < b$. We write $n = (a_k a_{k-1} \ldots a_0)_b$.

Conversion Algorithms

There are two primary methods for determining the digits a_i.

Method 1: Iterative Division (Upward Evaluation) Observe that $n = b(\ldots) + a_0$. Thus, a_0 is the remainder of n divided by b.

$$n = q_0 b + a_0$$

The next digit, a_1, is the remainder of the quotient q_0 divided by b:

$$q_0 = q_1 b + a_1$$

We repeat this process until the quotient is zero.

Example 7.2.1. Base 7 Conversion. Represent 1749 in base 7.

$$\begin{aligned} 1749 &= 249 \cdot 7 + \mathbf{6} \\ 249 &= 35 \cdot 7 + \mathbf{4} \\ 35 &= 5 \cdot 7 + \mathbf{0} \\ 5 &= 0 \cdot 7 + \mathbf{5} \end{aligned}$$

Reading the remainders from last to first: $1749 = (5046)_7$.

Method 2: Highest Power Subtraction (Downward Evaluation) Alternatively, one may find the largest power $b^k \leq n$, determine $a_k = \lfloor n/b^k \rfloor$, let $r_k = n \bmod b^k$, and repeat for b^{k-1}.

Note. Existence and uniqueness of this representation follow directly from Euclid's Division Theorem. Briefly, we apply the division algorithm to n with divisor b to obtain

$$n = q_0 b + a_0, \quad 0 \leq a_0 < b.$$

If $q_0 = 0$, we are done with $k = 0$. Otherwise, we repeat with q_0:

$$q_0 = q_1 b + a_1, \quad 0 \leq a_1 < b,$$

and so on. Since the quotients q_i form a strictly decreasing sequence of non-negative integers, this process terminates after finitely many steps, yielding

$$n = a_k b^k + a_{k-1} b^{k-1} + \cdots + a_1 b + a_0$$

for some k and digits $0 \leq a_i < b$. Uniqueness also follows from the division algorithm: if

$$n = a_k b^k + \cdots + a_0 = a'_k b^k + \cdots + a'_0$$

with $0 \leq a_i, a'_i < b$, then comparing both sides modulo b forces $a_0 = a'_0$. Subtracting these and dividing by b, we obtain a shorter equality of the same form. Inducting on k shows $a_i = a'_i$ for all i.

7.3 Greatest Common Divisors

We often analyse the shared structural properties of two integers.

Definition 7.3.1. ***Greatest Common Divisor***. An integer d is a greatest common divisor (gcd) of a and b if:

1. $d \mid a$ and $d \mid b$ (it is a common divisor).
2. For any integer c, if $c \mid a$ and $c \mid b$, then $c \mid d$.

Note that this definition relies on divisibility, not magnitude. In the ring of integers, if d satisfies this, then $-d$ also satisfies it. We denote the unique *positive* gcd as $\gcd(a, b)$. If $\gcd(a, b) = 1$, we say a and b are relatively prime (denoted $a \perp b$).

Euclid's Algorithm

The computation of the gcd is based on a simple reduction property.

***Lemma* 7.3.1.** *GCD Reduction.* For any integers m, n, q:

$$\gcd(m, n - qm) = \gcd(m, n)$$

Proof. Let $d = \gcd(m, n)$. Since $d \mid m$ and $d \mid n$, d must divide any linear combination $n - qm$. Thus d is a common divisor of m and $n - qm$. Conversely, let c be any common divisor of m and $n - qm$. Then c divides m and c divides $(n - qm) + qm = n$. Thus c is a common divisor of m and n. Since the sets of common divisors are identical, their greatest elements are identical. ■

Setting $q = \lfloor n/m \rfloor$, we have $n - qm = n \bmod m$. This yields the recursive step for Euclid's Algorithm:

$$\gcd(n, m) = \gcd(m, n \bmod m)$$

By repeatedly applying the modulo operator, we generate a strictly decreasing sequence of remainders.

Let $r_0 = a$ and $r_1 = b$ with $a \geq b > 0$. We define the sequence:

$$\begin{aligned} r_0 &= q_1 r_1 + r_2 && (0 < r_2 < r_1) \\ r_1 &= q_2 r_2 + r_3 && (0 < r_3 < r_2) \\ &\vdots \\ r_{n-2} &= q_n r_{n-1} + r_n && (0 < r_n < r_{n-1}) \\ r_{n-1} &= q_{n+1} r_n + 0 \end{aligned}$$

The algorithm terminates when the remainder is 0. The last non-zero remainder r_n is the gcd.

***Theorem* 7.3.1. Correctness of Euclid's Algorithm**. The last non-zero remainder r_n produced by the Euclidean algorithm is exactly $\gcd(a, b)$.

Proof. We must show that r_n is a common divisor and that it is the greatest.

Step 1: r_n is a common divisor. From the last line, $r_{n-1} = q_{n+1} r_n$, so $r_n \mid r_{n-1}$. Moving up one step: $r_{n-2} = q_n r_{n-1} + r_n$. Since r_n divides both terms on the RHS, $r_n \mid r_{n-2}$. By induction, r_n divides all previous remainders, including $r_1 = b$ and $r_0 = a$. Thus r_n is a common divisor.

Step 2: r_n is the greatest. Let c be any common divisor of a and b. From the first line, $r_2 = a - q_1 b$. Since $c \mid a$ and $c \mid b$, $c \mid r_2$. From the second line, $r_3 = b - q_2 r_2$. Since $c \mid b$ and $c \mid r_2$, $c \mid r_3$. By induction, c divides every remainder r_k. Specifically, $c \mid r_n$. Since any common divisor divides r_n, r_n must be the maximal such integer. ■

Remark. Kronecker (1823-1891) observed that the algorithm can be accelerated by choosing the *least absolute remainder*. Instead of $0 \leq r < b$, we choose r such that $|r| \leq b/2$, allowing negative remainders.

We can express the algorithm recursively using the modern modulo definition:

$$f(a, b) = \begin{cases} a & \text{if } b = 0 \\ f(b, a \bmod b) & \text{if } b > 0 \end{cases} \tag{7.1}$$

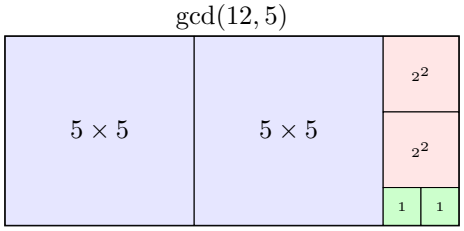


Figure 7.2: Geometric interpretation of Euclid's Algorithm for $\gcd(12, 5)$.

Ideals and Linear Combinations

A profound insight of number theory is that the greatest common divisor of a and b can naturally be generated by linear combinations of a and b. To prove this rigorously, we introduce the concept of an *ideal*.

Definition 7.3.2. ***Ideal***. For $a, b \in \mathbb{Z}$, the ideal generated by a and b, denoted (a, b), is the set of all integer linear combinations of a and b:

$$(a, b) = \{ua + vb \mid u, v \in \mathbb{Z}\}$$

Similarly, the principal ideal generated by a single integer d is $(d) = \{ud \mid u \in \mathbb{Z}\}$.

We now show that the complex structure of (a, b) collapses into a simple principal ideal.

***Lemma* 7.3.2.** *Principal Ideal Lemma.* For any $a, b \in \mathbb{Z}$, there exists a $d \in \mathbb{Z}$ such that $(a, b) = (d)$. Furthermore, $d = \gcd(a, b)$.

Proof. If $a = b = 0$, then $(0, 0) = \{0\} = (0)$, so $d = 0$. Assume at least one of a, b is non-zero. The set (a, b) contains non-zero integers (e.g., $a^2 + b^2 > 0$). Let $S = (a, b) \cap \mathbb{Z}^+$. By the Well-Ordering Principle, S contains a smallest positive element; call it d.

We first show $(d) \subseteq (a, b)$. Since $d \in (a, b)$, any multiple kd is a linear combination of elements in (a, b), and thus is in (a, b).

We next show $(a, b) \subseteq (d)$. Let $c \in (a, b)$. By the Division Theorem, we can write $c = qd + r$ with $0 \leq r < d$. Rearranging, $r = c - qd$. Since $c \in (a, b)$ and $d \in (a, b)$, their linear combination r must be in (a, b). However, d is the *smallest* positive element of (a, b), and $r < d$. The only non-negative integer in (a, b) strictly smaller than d is 0. Thus $r = 0$, implying $c = qd$. Hence $c \in (d)$.

We conclude $(a, b) = (d)$. Finally, we verify $d = \gcd(a, b)$.

1. Since $a \in (a, b) = (d)$, $d \mid a$. Similarly $d \mid b$.
2. Let k be any common divisor of a and b. Since $d \in (a, b)$, $d = ua + vb$ for some u, v. Because k divides a and b, k divides $ua + vb$, so $k \mid d$.

■

***Theorem* 7.3.2. Bézout's Identity**. For any integers a and b, there exist integers u and v such that:

$$\gcd(a, b) = ua + vb$$

This follows directly from the fact that $d \in (a, b)$. This identity is computationally accessible via the Extended Euclidean Algorithm.

Example 7.3.1. Linear Combination. Consider $a = 26, b = 18$.

$$\begin{aligned} 26 &= 1 \cdot 18 + 8 \\ 18 &= 2 \cdot 8 + 2 \\ 8 &= 4 \cdot 2 + 0 \end{aligned}$$

Thus $\gcd(26, 18) = 2$. To find the combination, we back-substitute:

$$\begin{aligned} 2 &= 18 - 2(8) \\ 2 &= 18 - 2(26 - 1 \cdot 18) \\ 2 &= 3 \cdot 18 - 2 \cdot 26 \end{aligned}$$

Here $u = -2$ and $v = 3$.

Properties of the GCD

The Euclidean algorithm allows us to derive fundamental theorems about relative primality and linear scaling.

***Lemma* 7.3.3.** *Euclid's Lemma (Division Lemma).* If $a \mid bc$ and $a \perp b$, then $a \mid c$.

Proof. Since $a \perp b$, $\gcd(a, b) = 1$. By Bézout's Identity (which follows from the extended Euclidean algorithm), there exist integers u, v such that $ua + vb = 1$. Multiplying by c:

$$uac + vbc = c$$

Clearly $a \mid uac$. By hypothesis, $a \mid bc$, so $a \mid vbc$. By linearity, a divides the sum, so $a \mid c$. ■

This lemma allows us to extend the concept of relative primality from pairs to products.

***Theorem* 7.3.3. Coprimality with Products**. If an integer a is relatively prime to each of the numbers $b_1, b_2, \ldots, b_k$, then a is relatively prime to their product $B = b_1 b_2 \ldots b_k$.

Proof. We prove the base case $k = 2$ by contradiction; the general result follows by mathematical induction.

Base Case: Assume $a \perp b$ and $a \perp c$, but $a \not\perp bc$. By definition, $\gcd(a, bc) = d > 1$. Thus, there exists a common divisor d such that $d \mid a$ and $d \mid bc$.

- Since $a \perp b$, and $d \mid a$, any common divisor of d and b would also divide a and b. The only such divisor is 1, so $d \perp b$.
- Similarly, since $a \perp c$, and $d \mid a$, we must have $d \perp c$.

We now have $d \mid bc$ and $d \perp b$. By Euclid's Lemma (Division Lemma), this implies $d \mid c$. However, we established that $d \perp c$ (i.e., $\gcd(d, c) = 1$). The only positive integer that divides c and is relatively prime to c is 1. Thus $d = 1$. This contradicts the assumption that $d > 1$. Therefore, $a \perp bc$. ■

We next examine how the gcd scales with linear multiplication.

***Theorem* 7.3.4. Distributivity of GCD**. For any positive integer k:

$$\gcd(ka, kb) = k \cdot \gcd(a, b)$$

Proof. Apply Euclid's algorithm to the pair (ka, kb). Every equation $R_{i-2} = q_i R_{i-1} + R_i$ in the expansion of $\gcd(a, b)$ is simply multiplied by k. The quotients q_i remain unchanged, but the remainders become kr_i. The last non-zero remainder is $kr_n = k \cdot \gcd(a, b)$. ■

This property is essential for reducing fractions. If we write a fraction a/b and factor out $d = \gcd(a, b)$ such that $a = da'$ and $b = db'$, then:

$$\gcd(a, b) = d \cdot \gcd(a', b') \implies d = d \cdot 1$$

Thus $\gcd(a', b') = 1$, proving that any fraction can be reduced to relatively prime terms.

This scaling property implies that dividing out the gcd leaves the resulting quotients relatively prime.

***Theorem* 7.3.5. Reduction to Coprime Factors**. Let $d = \gcd(a, b)$. If we write $a = a_1 d$ and $b = b_1 d$, then:

$$a_1 \perp b_1$$

Proof. Using the Distributivity Theorem:

$$\begin{aligned} \gcd(a, b) &= \gcd(a_1 d, b_1 d) \\ d &= d \cdot \gcd(a_1, b_1) \end{aligned}$$

Since $d \neq 0$, we can divide both sides by d to obtain $1 = \gcd(a_1, b_1)$. Thus $a_1 \perp b_1$. ■

Remark. This theorem is the theoretical basis for simplifying fractions. If we interpret a fraction a/b as a pair of integers (a, b) with $b \neq 0$, this result shows that we can always divide out the greatest common divisor to obtain a pair (a_1, b_1) where the components are relatively prime.

7.4 Least Common Multiples

We define the dual concept to the divisor.

Definition 7.4.1. ***Least Common Multiple***. The least common multiple of a and b, denoted $\text{lcm}(a, b)$, is the smallest positive integer m such that $a \mid m$ and $b \mid m$.

The set of common multiples corresponds to the intersection of the ideals (a) and (b). A fundamental property of the lcm is that it "contains" all other common multiples.

***Theorem* 7.4.1. Universal Divisibility of LCM**. Any common multiple of a and b is divisible by $\text{lcm}(a, b)$.

Proof. Let $L = \text{lcm}(a, b)$ and let M be any common multiple of a and b. We divide M by L using the Division Algorithm:

$$M = qL + r \quad \text{where } 0 \leq r < L$$

We must show that $r = 0$. Rearranging for r:

$$r = M - qL$$

Since $a \mid M$ and $a \mid L$, it follows by linearity that $a \mid (M - qL)$, so $a \mid r$. Similarly, $b \mid M$ and $b \mid L$ implies $b \mid r$. Thus, r is a common multiple of a and b. However, $r < L$, and L is defined as the *least positive* common multiple. The only non-negative common multiple strictly smaller than L is 0. Therefore, $r = 0$, which implies $M = qL$. Thus $L \mid M$. ■

The set of common multiples is simply the intersection of the ideal (a) and (b). The structure of the integers implies a tight relationship between the gcd and lcm.

***Theorem* 7.4.2. The Product Theorem**. For any positive integers a and b:

$$\gcd(a, b) \cdot \text{lcm}(a, b) = ab$$

Proof. Let $d = \gcd(a, b)$. We can write $a = da'$ and $b = db'$ where $a' \perp b'$. The product $ab = d^2a'b'$. Consider the integer $m = da'b'$.

- $m = ab'$, so $a \mid m$.
- $m = ba'$, so $b \mid m$.

Thus m is a common multiple. Any common multiple M must be divisible by a, so $M = ka = kda'$. Since $b \mid M$, $db' \mid kda'$, which implies $b' \mid ka'$. Since $b' \perp a'$, we must have $b' \mid k$, so $k = jb'$. Thus $M = jb'da' = jm$. This shows m is the *least* common multiple. Multiplying m by d:

$$d \cdot \text{lcm}(a, b) = d(da'b') = (da')(db') = ab$$

■

***Corollary* 7.4.1.** *Relativity.* $\text{lcm}(a, b) = ab$ if and only if $a \perp b$.

Famous Conjectures

While the definitions above are elementary, they lead to questions of immense difficulty.

- **Twin Prime Conjecture**: Are there infinitely many primes p such that $p + 2$ is also prime? The pairs $(3, 5), (5, 7), (11, 13)$ suggest a pattern that has yet to be proven infinite.
- **Fermat's Last Theorem**: Do there exist positive integers a, b, c such that $a^n + b^n = c^n$ for $n \geq 3$? Pierre de Fermat claimed a proof in the margins of a book in 1637, but it was not until 1994 that Andrew Wiles provided a rigorous proof using elliptic curves.
- **Catalan's Conjecture**: The equation $x^a - y^b = 1$ has only one solution in natural numbers $x, a, y, b > 1$: $3^2 - 2^3 = 1$. This was proven by Preda Mihăilescu in 2002.

7.5 Prime Numbers

Having established the linear structure of the integers through the division algorithm and the greatest common divisor, we now turn to their multiplicative building blocks.

Definition 7.5.1. ***Prime and Composite***. An integer $p > 1$ is called a prime if its only positive divisors are 1 and p. An integer $n > 1$ that is not prime is called composite.

Note. The number 1 is neither prime nor composite; it is a unit. Excluding 1 from the set of primes is a necessary convention to ensure the uniqueness of factorisation.

The Fundamental Theorem of Arithmetic

The central pillar of elementary number theory is the assertion that every integer can be decomposed uniquely into prime factors. This theorem justifies the chemical analogy of primes as the "atoms" of the integers.

***Theorem* 7.5.1. The Fundamental Theorem of Arithmetic**. Every positive integer $n > 1$ can be written as a product of primes. Furthermore, this representation is unique, up to the ordering of the factors.

$$n = p_1 p_2 \dots p_k$$

The proof consists of two distinct parts: existence and uniqueness.

Existence

We proceed by contradiction. Suppose there exists an integer $n > 1$ that cannot be written as a product of primes. By the Well-Ordering Principle, there must be a *smallest* such integer; let us call it m.

- If m is prime, then $m = m$ is a valid product of primes (a single factor). This contradicts the assumption that m has no representation.
- If m is composite, then by definition $m = ab$ for integers $1 < a, b < m$.

Since a and b are strictly smaller than m, and m is the *smallest* number without a prime factorisation, both a and b must have prime factorisations.

$$a = p_1 \dots p_r, \quad b = q_1 \dots q_s$$

Consequently, $m = ab = p_1 \dots p_r q_1 \dots q_s$. This is a prime factorisation of m, yielding a contradiction. Thus, no such integer exists.

Uniqueness

The proof of uniqueness is more subtle and relies on Euclid's Lemma, which connects divisibility by primes to the algebraic structure of the integers.

***Lemma* 7.5.1.** *Euclid's Lemma.* If p is a prime and $p \mid ab$, then $p \mid a$ or $p \mid b$.

Proof. Suppose $p \mid ab$ and $p \nmid a$. We must show $p \mid b$. Since p is prime and does not divide a, the greatest common divisor of a and p must be 1. Thus, $a \perp p$. By Bézout's Identity, there exist integers u, v such that:

$$ua + vp = 1$$

Multiplying both sides by b:

$$uab + vpb = b$$

Since $p \mid ab$, there exists an integer k such that $ab = kp$. Substituting this:

$$u(kp) + vpb = b \implies p(uk + vb) = b$$

Thus, b is a multiple of p, so $p \mid b$. ■

We now prove the uniqueness of the factorisation. Suppose an integer n has two factorisations:

$$n = p_1 p_2 \dots p_k = q_1 q_2 \dots q_m$$

where p_i and q_j are primes. Consider p_1. Since p_1 divides n, it divides the product $q_1 q_2 \dots q_m$. By Euclid's Lemma, p_1 must divide at least one factor q_j. Since q_j is prime, its only divisors are 1 and itself. As $p_1 > 1$, we must have $p_1 = q_j$. We divide both sides by p_1 (and q_j) and repeat the argument. The process must terminate with all factors matched, implying $k = m$ and the sets of primes are identical.

7.6 Canonical Representation

By grouping identical primes, we can express any positive integer n in standard form:

$$n = \prod_p p^{n_p} \tag{7.2}$$

where the product runs over all prime numbers, and the exponent n_p is a non-negative integer. For any specific n, only finitely many exponents n_p are non-zero.

This representation transforms multiplicative problems into additive ones. Mapping an integer to its sequence of exponents creates an isomorphism between the positive integers and the space of integer sequences.

$$n \longleftrightarrow \langle n_2, n_3, n_5, n_7, \dots \rangle$$

GCD and LCM via Exponents

Let integers a and b be represented by the exponent sequences $\langle a_p \rangle$ and $\langle b_p \rangle$ respectively.

$$a = \prod_p p^{a_p}, \quad b = \prod_p p^{b_p}$$

The divisibility relation $a \mid b$ holds if and only if $a_p \leq b_p$ for all primes p. It follows immediately that the greatest common divisor and least common multiple correspond to the component-wise minimum and maximum of these exponents.

***Theorem* 7.6.1. GCD and LCM Representation**.

$$\gcd(a, b) = \prod_p p^{\min(a_p, b_p)}$$
$$\operatorname{lcm}(a, b) = \prod_p p^{\max(a_p, b_p)}$$

This provides an elegant algebraic proof of the Product Theorem derived in the previous section.

Proof. For any real numbers x, y, we have the identity $x + y = \min(x, y) + \max(x, y)$. Multiplying the

expressions for gcd and lcm:

$$\begin{aligned}\gcd(a,b)\cdot\operatorname{lcm}(a,b) &= \prod_p p^{\min(a_p,b_p)}\cdot\prod_p p^{\max(a_p,b_p)}\\ &= \prod_p p^{\min(a_p,b_p)+\max(a_p,b_p)}\\ &= \prod_p p^{a_p+b_p}\\ &= \left(\prod_p p^{a_p}\right)\left(\prod_p p^{b_p}\right) = ab\end{aligned}$$

■

Application: Factorial Factors

A classic application of number theory involves determining the prime factorisation of a factorial $n!$. Since $n!$ grows extremely rapidly, direct factorisation is impossible for large n. However, we can determine the exponent of a prime p in $n!$ using the properties of the floor function.

Let $\epsilon_p(m)$ denote the exponent of the highest power of p that divides m. We seek $E_p(n!) = \epsilon_p(1\cdot 2\cdot\cdots\cdot n)$. By the properties of logarithms (or exponents), $E_p(n!) = \sum_{k=1}^{n}\epsilon_p(k)$.

Consider the contribution of multiples of p.

- Every multiple of p ($p, 2p, 3p\ldots$) contributes at least one factor of p. The number of such multiples in $\{1,\ldots,n\}$ is $\lfloor n/p\rfloor$.
- Every multiple of p^2 ($p^2, 2p^2\ldots$) contributes an *additional* factor of p. The number of such multiples is $\lfloor n/p^2\rfloor$.
- Generally, every multiple of p^k contributes an additional factor.

***Theorem* 7.6.2. Legendre's Formula**. The exponent of a prime p in the prime factorisation of $n!$ is:

$$E_p(n!) = \sum_{k=1}^{\infty}\left\lfloor\frac{n}{p^k}\right\rfloor$$

Note. The summation is finite, as the terms become zero when $p^k > n$.

Example 7.6.1. Zeros of 100!. To find the number of trailing zeros in 100!, we calculate the exponent of 10 in its factorisation. Since $10 = 2\cdot 5$, the number of zeros is determined by the number of pairs of prime factors 2 and 5. As 5 is larger, it is the limiting factor.

$$E_5(100!) = \left\lfloor\frac{100}{5}\right\rfloor + \left\lfloor\frac{100}{25}\right\rfloor + \left\lfloor\frac{100}{125}\right\rfloor + \ldots$$

$$E_5(100!) = 20 + 4 + 0 = 24$$

Thus, 100! ends in 24 zeros.

The Failure of Unique Factorisation

To appreciate the Fundamental Theorem of Arithmetic, one must realise that it is not a trivial property of all mathematical structures. In more general algebraic systems (rings), unique factorisation may fail.

Consider the set of numbers $\mathbb{Z}[\sqrt{-5}] = \{a + b\sqrt{-5} \mid a, b \in \mathbb{Z}\}$. This set is closed under addition and multiplication. We define a norm function $N(z) = a^2 + 5b^2$ which has the property $N(xy) = N(x)N(y)$.

In this system, the number 6 can be factored in two distinct ways:

$$6 = 2 \cdot 3$$

$$6 = (1 + \sqrt{-5})(1 - \sqrt{-5})$$

One can prove that $2, 3, 1 + \sqrt{-5}$, and $1 - \sqrt{-5}$ are all "irreducible" (they cannot be broken down further into non-unit factors) in this system. Furthermore, 2 is not associated with $1 \pm \sqrt{-5}$. Consequently, $\mathbb{Z}[\sqrt{-5}]$ is not a Unique Factorisation Domain (UFD). This highlights that the property of primes in $\mathbb{Z}$ (that if p divides a product, it must divide a factor), is a special geometric property of the integers not shared by all number systems.

In the standard integers $\mathbb{Z}$, the concepts of irreducible (has no factors) and prime (divides a product implies divides a factor) coincide. In $\mathbb{Z}[\sqrt{-5}]$, the number 2 is irreducible (cannot be split) but *not* prime (it divides $(1 + \sqrt{-5})(1 - \sqrt{-5}) = 6$, but divides neither factor). The Fundamental Theorem of Arithmetic holds in $\mathbb{Z}$ precisely because irreducibles are primes.

7.7 The Distribution of Primes

Having established the fundamental nature of prime numbers as the atomic elements of the integers, we turn our attention to their distribution. While the location of any specific prime appears random, the overall behaviour of the prime numbers follows precise laws when viewed at scale.

The Infinitude of Primes

The most basic question regarding the set of prime numbers $\mathbb{P}$ is whether it is finite. This question was settled by Euclid around 300 BCE with one of the most elegant proofs in mathematics.

***Theorem* 7.7.1. Infinitude of Primes**. The set of prime numbers is infinite.

Proof. We proceed by contradiction. Assume that there are finitely many primes, and let the complete set be $\mathbb{P} = \{p_1, p_2, \ldots, p_m\}$. Construct the integer Q defined as the product of all primes plus one:

$$Q = \left(\prod_{i=1}^{m} p_i \right) + 1$$

By the Fundamental Theorem of Arithmetic, Q must have a prime divisor, say q. If q were in our set $\mathbb{P}$, then $q = p_k$ for some k. Consequently, q divides the product $\prod p_i$. Since $q \mid Q$ and $q \mid \prod p_i$, by linearity, q must divide their difference:

$$q \mid \left(Q - \prod_{i=1}^{m} p_i \right) \implies q \mid 1$$

This is impossible, as no prime divides 1. Therefore, q cannot be in the set $\{p_1, \ldots, p_m\}$. This contradicts the assumption that $\mathbb{P}$ contains all primes. Thus, the set of primes must be infinite. ■

Prime Gaps

While there are infinitely many primes, they become less frequent as numbers grow larger. A natural question is whether the gap between consecutive primes is bounded. That is, does there exist a constant C such that for any prime p_n, $p_{n+1} - p_n < C$? We prove constructively that we can find intervals of arbitrary length containing no primes.

***Theorem* 7.7.2. Arbitrary Prime Gaps**. For any integer $k \geq 1$, there exists a sequence of k consecutive composite integers.

Proof. Consider the sequence of k integers starting at $(k+1)!+2$.

$$S = \{(k+1)!+2, \quad (k+1)!+3, \quad \ldots, \quad (k+1)!+(k+1)\}$$

Let $x_j = (k+1)!+j$ for $2 \leq j \leq k+1$. By definition of the factorial, $j \mid (k+1)!$. Clearly, $j \mid j$. By linearity of divisibility, $j \mid ((k+1)!+j)$, so $j \mid x_j$. Since $x_j > j > 1$, each number in the sequence has a non-trivial divisor and is therefore composite. ■

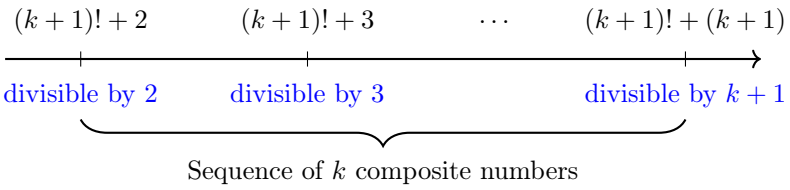


Figure 7.3: Visualisation of the prime gap construction.

Example 7.7.1. Gap of size 5. To find 5 consecutive composites ($k = 5$), we compute $(5+1)!+2 = 720+2 = 722$. The sequence $722, 723, 724, 725, 726, 727$ contains composites divisible by $2, 3, 4, 5, 6$ respectively. (Note: 727 is actually prime, marking the end of the gap).

The Prime Number Theorem

Although the exact distribution of primes is erratic, their average density behaves predictably. To quantify this, we introduce the prime counting function.

Definition 7.7.1. ***Prime Counting Function***. The function $\pi : \mathbb{R} \to \mathbb{N}$ counts the number of primes less than or equal to x:

$$\pi(x) = \sum_{p \leq x} 1 \quad \text{where } p \in \mathbb{P}$$

In 1896, Hadamard and de la Vallée Poussin independently proved that $\pi(x)$ approximates to $x/\ln x$ as x approaches infinity.

***Theorem* 7.7.3. The Prime Number Theorem.**

$$\lim_{x \to \infty} \frac{\pi(x)}{x/\ln x} = 1$$

This theorem implies that the probability of a randomly selected integer n being prime is approximately $1/\ln n$. This probabilistic intuition underpins modern cryptography.

Remark. I know i haven't introduced limits yet but all this means is "as x gets bigger and bigger close to infinity, both $\pi(x)$ and $x/\ln x$ also gets bigger and bigger to infinity, but they grow at essentially the same rate so their ratio gets closer and closer to 1."

Primality Testing

In many applications, particularly RSA encryption, we must determine whether a large integer n is prime. The most elementary method is trial division.

***Lemma* 7.7.1.** *Trial Division Bound.* If n is a composite integer, then n has a prime divisor p such that $p \leq \sqrt{n}$.

Proof. Since n is composite, we can write $n = ab$ with $1 < a \leq b < n$. If $a > \sqrt{n}$, then $b \geq a > \sqrt{n}$, which implies $ab > \sqrt{n} \cdot \sqrt{n} = n$, a contradiction. Thus $a \leq \sqrt{n}$. By the Fundamental Theorem of Arithmetic, a has a prime factor $p \leq a \leq \sqrt{n}$. By transitivity, $p \mid n$. ■

This lemma optimises trial division: to test n, we need only check divisors up to $\lfloor\sqrt{n}\rfloor$. However, for numbers with hundreds of digits, $\sqrt{n}$ is still astronomically large.

Computational Complexity

Primality testing is a central problem in complexity theory.

1. **Probabilistic Tests**: Algorithms such as the Miller-Rabin test can determine if a number is composite with high certainty. If the test returns "composite", the number is definitely composite. If it returns "probably prime", the probability of error is bounded by a parameter ϵ (e.g., 2^{-100}).
2. **Deterministic Tests**: In 2002, Agrawal, Kayal, and Saxena (AKS) proved that determining primality is in the complexity class **P**. This means there exists a deterministic algorithm that runs in polynomial time relative to the number of digits (bits) of n.

Open Problems

Despite centuries of study, simple questions about primes remain unanswered.

- **Goldbach's Conjecture**: Every even integer greater than 2 is the sum of two primes ($4 = 2 + 2, 6 = 3 + 3, 8 = 3 + 5$).
- **Twin Prime Conjecture**: There are infinitely many pairs of primes $(p, p + 2)$. Examples include $(3, 5), (11, 13)$, and $(101, 103)$.

7.8 Congruences and Modular Arithmetic

With the foundations of divisibility and prime factorisation established, we now introduce modular arithmetic. This framework, developed by Gauss, simplifies number theory by focusing on the cyclical properties of integers — treating numbers as equivalent if they leave the same remainder upon division.

Modular Congruence

We begin by defining the core relation.

Definition 7.8.1. ***Congruence***. Let $m \in \mathbb{Z}^+$ be a fixed integer called the *modulus*. Two integers a and b are said to be congruent modulo m, denoted $a \equiv b \pmod{m}$, if m divides their difference:

$$m \mid (a - b)$$

Equivalently, a and b leave the same remainder when divided by m.

***Proposition* 7.8.1.** *Equivalence Relation.* The relation $\equiv \pmod{m}$ is an equivalence relation on $\mathbb{Z}$. For all $a, b, c \in \mathbb{Z}$:

- **Reflexivity**: $a \equiv a \pmod{m}$.
- **Symmetry**: If $a \equiv b \pmod{m}$, then $b \equiv a \pmod{m}$.
- **Transitivity**: If $a \equiv b \pmod{m}$ and $b \equiv c \pmod{m}$, then $a \equiv c \pmod{m}$.

This relation partitions the set of integers $\mathbb{Z}$ into m disjoint equivalence classes, called residue classes. We denote the set of these classes as $\mathbb{Z}_m = \{[0]_m, [1]_m, \ldots, [m-1]_m\}$. Typically, we work with the standard representative of each class, the remainder in the range $[0, m - 1]$.

Arithmetic Properties

A crucial feature of congruences is that they respect the operations of addition and multiplication.

***Lemma* 7.8.1.** *Modular Arithmetic Operations.* If $a \equiv b \pmod{m}$ and $c \equiv d \pmod{m}$, then:

1. $a + c \equiv b + d \pmod{m}$
2. $a \cdot c \equiv b \cdot d \pmod{m}$
3. $a^k \equiv b^k \pmod{m}$ for any $k \in \mathbb{N}$

Proof. We prove the multiplication property. By definition, $a = b + km$ and $c = d + lm$ for integers k, l.

$$\begin{aligned} ac &= (b + km)(d + lm) \\ &= bd + blm + dkm + klm^2 \\ &= bd + m(bl + dk + klm) \end{aligned}$$

Since m divides the second term, $ac - bd$ is divisible by m. Thus $ac \equiv bd \pmod{m}$. ■

***Corollary* 7.8.1.** *Polynomial Congruence.* Let $f(x)$ be a polynomial with integer coefficients. If $a \equiv b \pmod{m}$, then $f(a) \equiv f(b) \pmod{m}$.

This property allows us to replace large numbers with their small remainders at any stage of a calculation without altering the final modular result.

Example 7.8.1. Power Calculation. Compute the last digit of 7^{100}. We work modulo 10.

$$7^2 = 49 \equiv -1 \pmod{10}$$

$$7^{100} = (7^2)^{50} \equiv (-1)^{50} \equiv 1 \pmod{10}$$

The last digit is 1.

Modular Inverses

While addition, subtraction, and multiplication behave intuitively modulo m, division is more complex. The congruence $ax \equiv b \pmod{m}$ does not always have a solution for x. To "divide" by a, we need to find its multiplicative inverse.

Definition 7.8.2. ***Multiplicative Inverse***. An integer a is invertible modulo m if there exists an integer x such that:

$$ax \equiv 1 \pmod{m}$$

If such an x exists, it is unique modulo m and is denoted a^{-1}.

***Theorem* 7.8.1. Existence of Inverses**. The inverse $a^{-1} \pmod{m}$ exists if and only if $\gcd(a, m) = 1$.

Proof. **Necessity**: If $ax \equiv 1 \pmod{m}$, then $ax - 1 = km$ for some integer k. Rearranging gives $ax - km = 1$. Any common divisor of a and m must divide the linear combination $ax - km = 1$. Thus $\gcd(a, m)$ must be 1.

Sufficiency: If $\gcd(a, m) = 1$, then by Bézout's Identity, there exist integers u, v such that $ua + vm = 1$. Taking this equation modulo m:

$$ua + vm \equiv 1 \pmod{m} \implies ua \equiv 1 \pmod{m}$$

Thus u is the inverse of a. ■

The Extended Euclidean Algorithm can be used to compute a^{-1} efficiently.

The Chinese Remainder Theorem

We often encounter systems of simultaneous congruences. The Chinese Remainder Theorem (CRT) provides a condition for the existence and uniqueness of solutions to such systems.

***Theorem* 7.8.2. The Chinese Remainder Theorem**. Let $m_1, m_2, \ldots, m_k$ be pairwise relatively prime integers (i.e., $\gcd(m_i, m_j) = 1$ for $i \neq j$). Let $M = m_1 m_2 \ldots m_k$. For any integers $a_1, a_2, \ldots, a_k$, the system of congruences:

$$\begin{aligned} x &\equiv a_1 \pmod{m_1} \\ x &\equiv a_2 \pmod{m_2} \\ &\vdots \\ x &\equiv a_k \pmod{m_k} \end{aligned}$$

has a unique solution modulo M.

Proof. **Existence**: Let $M_i = M/m_i$. Since $m_j \mid M_i$ for all $j \neq i$, and $\gcd(m_i, M_i) = 1$. Since $\gcd(m_i, M_i) = 1$, there exists an inverse y_i such that $M_i y_i \equiv 1 \pmod{m_i}$. Construct the solution:

$$x = \sum_{i=1}^{k} a_i M_i y_i$$

Let us verify this modulo m_j. For any term $i \neq j$, $m_j \mid M_i$, so $a_i M_i y_i \equiv 0 \pmod{m_j}$. The only non-zero term is $i = j$:

$$x \equiv a_j M_j y_j \equiv a_j(1) \equiv a_j \pmod{m_j}$$

Thus x satisfies all congruences.

Uniqueness: Suppose x and x' are two solutions. Then $x \equiv x' \pmod{m_i}$ for all i. This implies $m_i \mid (x - x')$. Since the moduli m_i are pairwise coprime, their product M must divide $(x - x')$. Thus $x \equiv x' \pmod{M}$. ■

Example 7.8.2. Sun Tzu's Puzzle. Find a number x that leaves a remainder of 2 when divided by 3, 3 when divided by 5, and 2 when divided by 7.

$$x \equiv 2 \pmod{3}, \quad x \equiv 3 \pmod{5}, \quad x \equiv 2 \pmod{7}$$

Here $M = 3 \cdot 5 \cdot 7 = 105$.

- $M_1 = 35$. $35 \equiv 2 \pmod{3}$. Inverse of 2 mod 3 is 2. Term: $2 \cdot 35 \cdot 2 = 140$.
- $M_2 = 21$. $21 \equiv 1 \pmod{5}$. Inverse of 1 mod 5 is 1. Term: $3 \cdot 21 \cdot 1 = 63$.
- $M_3 = 15$. $15 \equiv 1 \pmod{7}$. Inverse of 1 mod 7 is 1. Term: $2 \cdot 15 \cdot 1 = 30$.

$x = 140 + 63 + 30 = 233$. $233 \pmod{105} = 23$. Solution: $x \equiv 23 \pmod{105}$.

7.9 Fermat's Little Theorem and Euler's Theorem

We now explore properties related to powers in modular arithmetic. These theorems are fundamental to primality testing and public-key cryptography.

***Theorem* 7.9.1. Fermat's Little Theorem**. If p is a prime number, then for any integer a:

$$a^p \equiv a \pmod{p}$$

If $p \nmid a$, we can divide by a to get:

$$a^{p-1} \equiv 1 \pmod{p}$$

Proof. Assume $p \nmid a$. Consider the set of multiples $S = \{a, 2a, 3a, \ldots, (p-1)a\}$. We claim that no two elements in S are congruent modulo p. If $ia \equiv ja \pmod{p}$ for $1 \leq i, j < p$, then $p \mid a(i-j)$. Since $p \nmid a$, $p \mid (i-j)$. Given the range of i, j, this implies $i = j$. Thus, the elements of S are simply the integers $\{1, 2, \ldots, p-1\}$ in some permuted order modulo p. Multiplying all elements in S:

$$a \cdot 2a \cdot \cdots \cdot (p-1)a \equiv 1 \cdot 2 \cdot \cdots \cdot (p-1) \pmod{p}$$

$$a^{p-1}(p-1)! \equiv (p-1)! \pmod{p}$$

Since p is prime, $(p-1)!$ is coprime to p and can be cancelled.

$$a^{p-1} \equiv 1 \pmod{p}$$

■

Euler's Generalisation

Euler generalised this result to composite moduli using the Euler phi function $\phi(n)$, defined as the count of positive integers less than n that are relatively prime to n.

***Theorem* 7.9.2. Euler's Theorem**. If $\gcd(a, n) = 1$, then:

$$a^{\phi(n)} \equiv 1 \pmod{n}$$

This theorem reduces to Fermat's Little Theorem when $n = p$, as $\phi(p) = p - 1$.

Application: Public-Key Cryptography

The theoretical results of modular arithmetic underpin the Diffie-Hellman key exchange and the RSA cryptosystem, which secure modern internet communications.

The Discrete Logarithm Problem

Consider the equation $g^x \equiv y \pmod{p}$.

- **Exponentiation**: Given g, x, p, computing y is easy (using modular exponentiation).
- **Discrete Logarithm**: Given g, y, p, finding x is computationally infeasible for large p.

This asymmetry defines a *one-way function*.

Diffie-Hellman Key Exchange

This protocol allows two parties (Alice and Bob) to establish a shared secret over an insecure channel.

1. **Setup**: A large prime p and a generator g are public.
2. **Alice**: Chooses secret a, sends $A = g^a \pmod{p}$.
3. **Bob**: Chooses secret b, sends $B = g^b \pmod{p}$.
4. **Shared Secret**:
   - Alice computes $S = B^a \pmod{p} = (g^b)^a = g^{ab}$.
   - Bob computes $S = A^b \pmod{p} = (g^a)^b = g^{ab}$.

An eavesdropper sees only g, p, A, B. To find $S = g^{ab}$, they would need to solve the Discrete Logarithm Problem to find a or b, which is assumed to be hard.

7.10 Exercises

Part I: Computational and Elementary Properties

1. **Radix Conversion.**
   (a) Express the decimal number 2024 in base 7.
   (b) Express the binary number $(1011011)_2$ in base 16 (hexadecimal) without converting to decimal first.

2. **The Euclidean Algorithm.** For the pair $(a, b) = (89, 55)$:
   (a) Use the Euclidean algorithm to compute $\gcd(a, b)$. Observe the sequence of remainders. Do you recognise these numbers?
   (b) Find integers u and v such that $89u + 55v = \gcd(89, 55)$.

3. **Linearity of GCD and LCM.** Prove or disprove the following identities for positive integers k, m, n:
   (a) $\gcd(km, kn) = k \cdot \gcd(m, n)$.
   (b) $\operatorname{lcm}(km, kn) = k \cdot \operatorname{lcm}(m, n)$.

   Remark. Use the canonical prime factorisation representation $n = \prod p_i^{e_i}$.

4. **Recursive LCM.** Using the identity $\gcd(m, n) \cdot \operatorname{lcm}(m, n) = mn$ and the result from the previous exercise, prove that if $n \geq m > 0$:

$$\operatorname{lcm}(n, m) = \frac{n \cdot \operatorname{lcm}(m, n \bmod m)}{n \bmod m}$$

   provided $n \bmod m \neq 0$.

5. **Divisibility Rules.** Using modular arithmetic, prove the following standard divisibility rules for an integer n given in decimal representation $n = (a_k a_{k-1} \dots a_0)_{10}$.
   (a) $n \equiv \sum_{i=0}^{k} a_i \pmod 9$. Consequently, $9 \mid n$ if and only if the sum of its digits is divisible by 9.
   (b) $n \equiv \sum_{i=0}^{k} (-1)^i a_i \pmod{11}$. Consequently, $11 \mid n$ if and only if the alternating sum of its digits is divisible by 11.

Part II: Modular Arithmetic and Structure

6. **Nested Moduli.** Let x, y, n be positive integers. Prove that $(x \bmod ny) \bmod y = x \bmod y$

   Remark. Write $x = q(ny) + r$ where $0 \leq r < ny$. Then express r in terms of y.

7. **The Definition of Modulo 0.** In our definition of $a \equiv b \pmod m$, we typically assume $m \geq 1$.
   (a) Based strictly on the definition $m \mid (a - b)$, interpret the meaning of $a \equiv b \pmod 0$.
   (b) What does the "set of integers modulo 0", $\mathbb{Z}_0$, look like?

8. **Linear Diophantine Equations.** A linear Diophantine equation is an equation of the form $ax+by = c$ where we seek integer solutions for x and y.
   (a) Prove that this equation has a solution if and only if $d \mid c$, where $d = \gcd(a, b)$.
   (b) If (x_0, y_0) is a specific solution, prove that all other solutions are of the form:

$$x = x_0 + \frac{b}{d}t, \quad y = y_0 - \frac{a}{d}t \quad \text{for } t \in \mathbb{Z}$$

9. **Primes of the form $4k + 3$.** Euclid's proof of the infinitude of primes can be adapted to specific arithmetic progressions.
   (a) Show that any odd number is of the form $4k + 1$ or $4k + 3$.
   (b) Show that the product of two numbers of the form $4k + 1$ is also of the form $4k + 1$.

(c) Suppose there are finitely many primes of the form $4k+3$, say $\{p_1, \ldots, p_m\}$. Consider $N = 4p_1 \ldots p_m - 1$. Show that N must have a prime factor of the form $4k+3$ that is not in the list, yielding a contradiction.

10. Euclid Numbers. Let p_n denote the n-th prime number. We define the n-th Euclid number as $E_n = p_n\# + 1$, where $p_n\# = \prod_{i=1}^{n} p_i$ is the primorial.

(a) Calculate E_1, E_2, E_3, and E_4. Determine their prime factorisations.
(b) Prove that for any n, $E_n \equiv 1 \pmod{p_i}$ for all $1 \leq i \leq n$.
(c) **Disprove:** "Every prime number p appears as a factor of some Euclid number E_n."

Remark. Consider the prime $p = 5$. Use the fact that E_n is odd for $n \geq 1$, and consider the modulo 10 or modulo 5 behaviour of the primorial.

Part III: Advanced Topics

11. Rationality and the Fundamental Theorem.

(a) Let n be a positive integer. Prove that $\sqrt{n}$ is rational if and only if n is a perfect square.

Remark. Consider the canonical prime factorisation $n = \prod p_i^{e_i}$. What must be true of the exponents e_i for $\sqrt{n}$ to be an integer? What if $\sqrt{n} = a/b$?

(b) Generalise this to show that $\sqrt[k]{n}$ is rational if and only if n is a perfect k-th power.

12. ⋆ Legendre's Formula and Base Representation. Recall Legendre's Formula $E_p(n!) = \sum_{k=1}^{\infty} \lfloor n/p^k \rfloor$. Let $S_p(n)$ be the sum of the digits of n when written in base p. Prove the alternate form of Legendre's Formula:

$$E_p(n!) = \frac{n - S_p(n)}{p-1}$$

Remark. Write $n = \sum a_i p^i$. Evaluate $\lfloor n/p^k \rfloor$ in terms of the coefficients a_i and sum the series.

13. ⋆ Wilson's Theorem.

(a) For a prime p, consider the polynomial $f(x) = (x-1)(x-2)\cdots(x-(p-1)) - (x^{p-1}-1)$ in $\mathbb{Z}_p[x]$.
(b) Determine the degree of this polynomial.
(c) How many roots does the polynomial $x^{p-1} - 1$ have in $\mathbb{Z}_p$?
(d) Use this to prove Wilson's Theorem: $(p-1)! \equiv -1 \pmod{p}$.

14. ⋆ Complex Roots of Unity and Modular Patterns. We can extend modular arithmetic concepts to complex numbers to find closed forms for modular reductions. Let $\omega = \frac{-1+i\sqrt{3}}{2} = e^{i2\pi/3}$. Note that $\omega^3 = 1$ and $1 + \omega + \omega^2 = 0$.

(a) Verify the identity for the parity bit:

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

(Here we treat the result 0 or 1 as an integer, not a class).
(b) Find constants $a, b, c \in \mathbb{C}$ such that for any integer $n \geq 0$:

$$n \bmod 3 = a + b\omega^n + c\omega^{2n}$$

Remark. Set up a system of equations for $n = 0, 1, 2$. Note that the pattern $n \bmod 3$ is periodic with period 3, matching the powers of ω.

(c) Generalise this method. How could one express $n \bmod k$ using k-th roots of unity?

15. ⋆⋆ The RSA Cryptosystem. Let $p = 61$ and $q = 53$.

(a) Compute the modulus $n = pq$ and the totient $\phi(n) = (p-1)(q-1)$.
(b) Choose a public exponent $e = 17$. Verify that $\gcd(e, \phi(n)) = 1$.
(c) Find the private exponent d such that $de \equiv 1 \pmod{\phi(n)}$.
(d) To encrypt the message $M = 65$, compute $C \equiv M^e \pmod{n}$.
(e) Verify the decryption by computing $C^d \pmod{n}$.

Chapter 8

The Archimedean Principle and Completeness

In the preceding chapters of this volume, we rigorously explored discrete mathematical structures: recurrence relations, sums, and the number-theoretic properties of the integers $\mathbb{Z}$. To progress to analysis, we must extend our domain to the rational numbers $\mathbb{Q}$. For the formal set-theoretic construction of $\mathbb{Q}$ as the field of quotients of $\mathbb{Z}$, the reader is referred to our companion notes on *Set Theory*.

While $\mathbb{Q}$ forms an ordered field, it is insufficient for continuous analysis due to inherent "gaps" in the number line — sequences that appear to converge fail to resolve to a value within $\mathbb{Q}$. To bridge the discrete and the continuous, we now formally define the completeness property of the real numbers $\mathbb{R}$. This chapter establishes the topological foundations required for calculus, culminating in the Archimedean Principle, the existence of roots, the density of rationals and therefore being the perfect springboard for the next set of notes.

8.1 Metric Properties: The Absolute Value

Before establishing the structure of $\mathbb{R}$, we must formalise the notion of magnitude. The absolute value function underpins the definitions of limits, neighbourhoods, and convergence.

Definition 8.1.1. ***Absolute Value***. For $x \in \mathbb{R}$, the absolute value $|x|$ is defined by:

$$|x| := \sqrt{x^2} = \begin{cases} x & \text{if } x \geq 0 \\ -x & \text{if } x < 0 \end{cases}$$

***Theorem* 8.1.1. Fundamental Properties**. Let $a, b \in \mathbb{R}$.

1. **Non-negativity:** $|a| \geq 0$, with equality if and only if $a = 0$.
2. **Multiplicativity:** $|ab| = |a||b|$.
3. **Triangle Inequality:** $|a + b| \leq |a| + |b|$.
4. **Reverse Triangle Inequality:** $||a| - |b|| \leq |a - b|$.

Proof. We prove (iii). Squaring both sides (permissible as both are non-negative) preserves the inequality order:

$$|a + b|^2 = (a + b)^2 = a^2 + 2ab + b^2 = |a|^2 + 2ab + |b|^2$$

Since $ab \leq |ab| = |a||b|$, it follows that:

$$|a|^2 + 2ab + |b|^2 \leq |a|^2 + 2|a||b| + |b|^2 = (|a| + |b|)^2$$

Taking square roots yields $|a + b| \leq |a| + |b|$. The generalisation to n terms, $|\sum a_i| \leq \sum |a_i|$, follows by induction. ■

8.2 The Completeness of Real Numbers

The fundamental distinction between $\mathbb{Q}$ and $\mathbb{R}$ lies in the existence of least upper bounds. We begin by formalising bounds.

Definition 8.2.1. ***Supremum and Infimum***. Let $S \subseteq \mathbb{R}$ be a non-empty set.

1. S is bounded above if there exists $u \in \mathbb{R}$ such that $\forall s \in S, s \leq u$.
2. The supremum (least upper bound) of S, denoted $\sup S$, is a value $\alpha \in \mathbb{R}$ such that:
   - α is an upper bound of S.
   - If $\gamma < \alpha$, then γ is not an upper bound of S.
3. Analogously, the infimum (greatest lower bound), denoted $\inf S$, is the greatest lower bound.

Remark. Distinguish $\sup S$ from $\max S$. The maximum must belong to the set S, whereas the supremum need not. For the open interval $S = (0, 1)$, $\sup S = 1$ but $\max S$ does not exist.

***Axiom* 8.2.1. The Completeness Axiom**. Every non-empty subset of $\mathbb{R}$ that is bounded above has a supremum in $\mathbb{R}$.

We demonstrate the necessity of this axiom by proving the incompleteness of the rationals.

***Proposition* 8.2.1.** *Incompleteness of* $\mathbb{Q}$. The set $S = \{q \in \mathbb{Q} \mid q^2 < 2\}$ is bounded above but has no supremum in $\mathbb{Q}$.

Proof. The set is bounded above by 2. Suppose $b = \sup S \in \mathbb{Q}$. By trichotomy, $b^2 < 2$, $b^2 = 2$, or $b^2 > 2$. We establish later that no rational squares to 2.

- If $b^2 < 2$: We construct a rational $b + \epsilon \in S$ with $\epsilon > 0$. Let $\epsilon = \min(1, \frac{2-b^2}{2b+1})$. Then $(b + \epsilon)^2 < 2$, contradicting that b is an upper bound.
- If $b^2 > 2$: We construct an upper bound $b - \epsilon < b$. Let $\epsilon = \frac{b^2-2}{2b}$. Then $(b - \epsilon)^2 > 2$, contradicting that b is the *least* upper bound.

Thus, $\sup S \notin \mathbb{Q}$. ■

8.3 The Archimedean Principle

The Completeness Axiom implies that the set of integers is not bounded within the reals. This property, known as the Archimedean Principle, ensures that no real number is infinitely large, nor is any positive real number infinitely small.

***Theorem* 8.3.1. Archimedean Property**. The set of natural numbers $\mathbb{N}$ is not bounded above in $\mathbb{R}$.

Proof. Assume for contradiction that $\mathbb{N}$ is bounded above. By the Completeness Axiom, let $\alpha = \sup \mathbb{N}$. Since α is the least upper bound, $\alpha - 1$ is not an upper bound. Thus, there exists $n \in \mathbb{N}$ such that $n > \alpha - 1$. This implies $n + 1 > \alpha$. Since $n + 1 \in \mathbb{N}$, this contradicts the assumption that α is an upper bound for all natural numbers. ■

This theorem yields critical corollaries for analysis.

***Corollary* 8.3.1.** *Consequences of Archimedes.* Let $x, y \in \mathbb{R}$ with $x > 0$.

1. There exists $n \in \mathbb{N}$ such that $nx > y$.
2. For any $\epsilon > 0$, there exists $n \in \mathbb{N}$ such that $0 < \frac{1}{n} < \epsilon$.

The Integer Part (Floor Function)

Using the Archimedean property, we connect our previous discrete analysis to the continuous domain by proving the existence of the floor function.

***Theorem* 8.3.2. Existence of Integer Part**. For any $x \in \mathbb{R}$, there exists a unique integer n such that $n \leq x < n+1$. We denote this $n = \lfloor x \rfloor$.

Proof. **Existence:** Consider the set $S = \{k \in \mathbb{Z} \mid k \leq x\}$. By the Archimedean property, $\mathbb{Z}$ is unbounded below, so S is non-empty. S is bounded above by x. By the well-ordering principle (adapted for integers bounded above), S contains a maximal element. Let $n = \max S$. Since $n \in S$, $n \leq x$. Since n is the maximum, $n+1 \notin S$, implying $n+1 > x$. **Uniqueness:** Suppose m and n satisfy the condition with $m < n$. Then $m+1 \leq n \leq x < m+1$, a contradiction. ■

Density of Rational Numbers

The Archimedean Principle allows us to prove that rational numbers are "dense" in the real line—meaning they appear in every open interval, regardless of size.

***Theorem* 8.3.3. Density of $\mathbb{Q}$.** Any open interval $(a, b) \subset \mathbb{R}$ with $a < b$ contains at least one rational number.

Proof. Since $b - a > 0$, by the Archimedean Principle, there exists $n \in \mathbb{N}$ such that $n(b-a) > 1$, or $nb - na > 1$. This implies the interval (na, nb) has length greater than 1. We select an integer strictly between na and nb. Let $m = \lfloor na \rfloor + 1$. From the definition of the floor function, $na < m \leq na + 1$. Since $nb - na > 1$, we have $na + 1 < nb$. Combining these yields $na < m < nb$. Dividing by n yields $a < \frac{m}{n} < b$. Thus, $q = m/n \in \mathbb{Q}$ lies in (a, b). ■

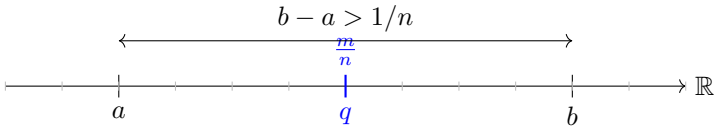


Figure 8.1: Constructing a rational q between a and b by choosing a step size $1/n$ smaller than the interval width.

***Theorem* 8.3.4. Density of Irrationals**. The set of irrational numbers $\mathbb{R} \setminus \mathbb{Q}$ is dense in $\mathbb{R}$.

Proof. Let $x < y$. Consider the interval $(x - \sqrt{2}, y - \sqrt{2})$. By the density of $\mathbb{Q}$, there exists $q \in \mathbb{Q}$ such that $x - \sqrt{2} < q < y - \sqrt{2}$. Rearranging, we find $x < q + \sqrt{2} < y$. Let $\xi = q + \sqrt{2}$. Since the sum of a rational and an irrational is irrational (closure properties), $\xi \in \mathbb{R} \setminus \mathbb{Q}$ and $\xi \in (x, y)$. ■

Existence of Roots

While $\mathbb{Q}$ is dense, it is not complete. We verify the structure of $\mathbb{R}$ by proving the existence of roots. We begin by explicitly constructing the square root of 2, utilizing a precise epsilon-style argument to rule out trichotomy cases.

***Theorem* 8.3.5. Existence of $\sqrt{2}$ in $\mathbb{R}$.** There exists a real number $\alpha > 0$ such that $\alpha^2 = 2$.

Proof. Define the set $A = \{x \in \mathbb{R} \mid x > 0 \text{ and } x^2 \leq 2\}$. First, $1 \in A$, so A is non-empty. Second, 2 is an upper bound for A (since $3^2 > 2$), so A is bounded above. By the Completeness Axiom, let $\alpha = \sup A$. Since $1 \in A$, we have $\alpha \geq 1$. We must show $\alpha^2 = 2$. We do so by ruling out $\alpha^2 < 2$ and $\alpha^2 > 2$.

Case 1: Assume $\alpha^2 < 2$. We construct a number larger than α that is still in A. Let $h = \frac{2-\alpha^2}{4\alpha}$. Since $\alpha \geq 1$ and $\alpha^2 < 2$, we have $0 < h < 1$. Consider $\alpha_1 = \alpha + h$. Then $(\alpha + h)^2 = \alpha^2 + 2\alpha h + h^2 = \alpha^2 + h(2\alpha + h)$. Since $h < 1$ and $1 \leq \alpha \leq 2$ then $(\alpha + h)^2 < \alpha + h(2\alpha + 2) \leq \alpha^2 + h(4\alpha)$. Substituting $h = \frac{2-\alpha^2}{4\alpha}$:

$$(\alpha + h)^2 < \alpha^2 + \left(\frac{2 - \alpha^2}{4\alpha}\right) 4\alpha = \alpha^2 + 2 - \alpha^2 = 2$$

Thus, $\alpha_1 \in A$. Since $\alpha_1 > \alpha$, this contradicts that α is an upper bound.

Case 2: Assume $\alpha^2 > 2$. We construct an upper bound smaller than α. Let $h = \frac{\alpha^2-2}{4\alpha}$. Since $\alpha > 1$, $h > 0$. Also $h < \frac{\alpha^2}{4\alpha} = \frac{\alpha}{4} < \alpha$, so $\alpha - h > 0$. Consider $\alpha_0 = \alpha - h$. Then $(\alpha - h)^2 = \alpha^2 - 2\alpha h + h^2 > \alpha^2 - 2\alpha h$. Substituting $h = \frac{\alpha^2-2}{4\alpha}$:

$$(\alpha - h)^2 > \alpha^2 - 2\alpha\left(\frac{\alpha^2 - 2}{4\alpha}\right) = \alpha^2 - \frac{\alpha^2 - 2}{2} = \frac{\alpha^2}{2} + 1$$

Since $\alpha^2 > 2$, $\frac{\alpha^2}{2} + 1 > 2$. Thus $(\alpha - h)^2 > 2$.This implies α_0 is an upper bound for A. Since $\alpha_0 < \alpha$, this contradicts that α is the *least* upper bound. Since neither inequality holds, $\alpha^2 = 2$. ■

We can generalize this result to any n-th root using the binomial expansion.

***Theorem* 8.3.6. Existence of n-th Roots**. Fix $n \in \mathbb{N}, n \geq 2$. For any $a > 0$, there exists a unique positive real number r such that $r^n = a$. We write $r = \sqrt[n]{a}$.

Proof. Let $S = \{s \in \mathbb{R} \mid s \geq 0 \wedge s^n \leq a\}$. The set is non-empty ($0 \in S$) and bounded above (by $\max(1, a)$). By the Completeness Axiom, let $r = \sup S$. We utilise the identity: $x^n - y^n = (x - y)\sum_{k=0}^{n-1} x^{n-1-k}y^k$.

Case 1: $r^n < a$. Let $\delta = a - r^n$. For $\epsilon \in (0, 1)$:

$$(r + \epsilon)^n - r^n = \epsilon \sum_{k=0}^{n-1}(r + \epsilon)^{n-1-k}r^k < \epsilon \sum_{k=0}^{n-1}(r + 1)^{n-1-k}(r + 1)^k = \epsilon \cdot n(r + 1)^{n-1}$$

Choosing $\epsilon < \frac{\delta}{n(r+1)^{n-1}}$ yields $(r + \epsilon)^n < r^n + \delta = a$, placing $r + \epsilon$ in S, a contradiction.

Case 2: $r^n > a$. Let $\delta = r^n - a$. For $\epsilon \in (0, r)$:

$$r^n - (r - \epsilon)^n = \epsilon \sum_{k=0}^{n-1} r^{n-1-k}(r - \epsilon)^k < \epsilon \cdot nr^{n-1}$$

Choosing $\epsilon < \frac{\delta}{nr^{n-1}}$ yields $(r-\epsilon)^n > r^n-\delta = a$, making $r-\epsilon$ an upper bound smaller than r, a contradiction. Therefore, $r^n = a$. ■

Having established the existence of $\sqrt{2}$ in $\mathbb{R}$, we conclude by determining its nature with respect to $\mathbb{Q}$.

***Theorem* 8.3.7. Irrationality of $\sqrt{2}$.** The number $\sqrt{2}$ is irrational.

Proof. Assume $\sqrt{2} \in \mathbb{Q}$. Then $\sqrt{2} = m/n$ for $m, n \in \mathbb{N}$ with $\gcd(m, n) = 1$. Squaring gives $2 = m^2/n^2 \implies 2n^2 = m^2$. Since $2 \mid m^2$, by Euclid's Lemma, $2 \mid m$. Let $m = 2k$. Then $2n^2 = (2k)^2 = 4k^2 \implies n^2 = 2k^2$. Thus $2 \mid n^2 \implies 2 \mid n$. We have $2 \mid m$ and $2 \mid n$, contradicting $\gcd(m, n) = 1$. Thus $\sqrt{2} \notin \mathbb{Q}$. ■

8.4 Exercises

1. Prove that for any $x, y, z \in \mathbb{R}$, $x^2 + y^2 + z^2 \geq xy + yz + zx$.

Remark. Start from the fact that $a^2 \geq 0$ for any a. Consider expressions of the form $(x - y)^2$.

2. Arithmetic-Geometric Mean Inequality.

(a) For any non-negative a, b, prove that $\sqrt{ab} \leq \frac{a+b}{2}$. Prove that equality holds if and only if $a = b$.
(b) If $0 < a < b$, prove the chain of inequalities: $a < \sqrt{ab} < \frac{a+b}{2} < b$.
(c) If $0 < a < b$ and $n \geq 2$ is an integer, prove that $\sqrt[n]{a} < \sqrt[n]{b}$.

3. Use the Triangle Inequality to prove the **Reverse Triangle Inequality**: $||a| - |b|| \leq |a - b|$.

Note. Hint: Write $a = (a-b)+b$ and apply the standard Triangle Inequality. Repeat for $b = (b-a)+a$.

4. ⋆ Let $x > 0$. Show that $x > 1$ if and only if $1/x < 1$. Subsequently, prove that if $y > x \geq 1$, then $x + \frac{1}{x} < y + \frac{1}{y}$.

Remark. Consider the behaviour of the function $f(z) = z + 1/z$.

8. For each of the following sets, find the supremum, infimum, maximum, and minimum, if they exist.

(a) $S_1 = \{x \in \mathbb{R} \mid x^2 < 5\}$
(b) $S_2 = \{(-1)^n + 1/n \mid n \in \mathbb{N}\}$
(c) $S_3 = \{a/(a+1) \mid a > 0\}$
(d) $S_4 = \{r \in \mathbb{Q} \mid r^3 < 8\}$

9. Prove that if a non-negative real number x satisfies $0 \leq x \leq \epsilon$ for all $\epsilon > 0$, then $x = 0$.

10. Let $S \subseteq \mathbb{R}$ be a non-empty set. State and prove the characterisation of $\beta = \inf S$ that is analogous to the Approximation Property of the Supremum.

11. Let $S \subseteq \mathbb{R}$ be a non-empty set of real numbers and let $c \in \mathbb{R}$. Define the set $cS = \{cs \mid s \in S\}$.

(a) If $c > 0$ and S is bounded above, show that cS is bounded above and $\sup(cS) = c \sup S$.
(b) If $c < 0$ and S is bounded above, show that cS is bounded below and $\inf(cS) = c \sup S$.

12. A set $S \subseteq \mathbb{R}$ is bounded if and only if there exists a number $C > 0$ such that $|x| \leq C$ for all $x \in S$. Prove this equivalence.

13. ⋆ Prove that $\log_2(3)$ is irrational.

Remark. Assume $\log_2(3) = a/b$. Rewrite in exponential form and derive a contradiction via the Fundamental Theorem of Arithmetic.

14. ⋆ Let A and B be two non-empty sets of real numbers such that for every $x \in A$ and every $y \in B$, we have $x \leq y$.

(a) Prove that $\sup A$ and $\inf B$ both exist.
(b) Prove that $\sup A \leq \inf B$.
(c) Provide an example where $\sup A = \inf B$.

16. Let $a < b$. Show that $a < \frac{1}{2}(a + b) < b$. Conclude that every open interval (a, b) is non-empty.

17. Use the Archimedean property to prove that for any real number $y > 0$, there exists an $n \in \mathbb{N}$ such that $n - 1 \leq y < n$.

18. Find a rational number and an irrational number between 1.7320508 and 1.7320509.

19. Is the set of rational numbers with denominators equal to a power of 3 (i.e., numbers of the form $m/3^n$ for $m \in \mathbb{Z}, n \in \mathbb{N}$) dense in $\mathbb{R}$? Provide a rigorous proof.

20. ⋆ Let x be an irrational number. Prove that the set $S = \{m + nx \mid m, n \in \mathbb{Z}\}$ is dense in $\mathbb{R}$.