Junior Research Seminar: Diophantine Analysis and Approximations [1] [2]

Ramin Takloo-Bighash[3], Steven Miller[4], Harald Helfgott[5], Florin Spinu[6]

February 23, 2003

[1]Homepage: http://www.math.princeton.edu/∼mathlab/
[2]Mondays $6:30-7:30$pm and Tuesdays $8:30-9:30$pm, Fine 314
[3]E-mail: rtakloo@math.princeton.edu
[4]E-mail:sjmiller@math.princeton.edu
[5]E-mail:haraldh@math.princeton.edu
[6]E-mail:fspinu@math.princeton.edu

Contents

Chapter 1

Preliminaries

1.1 Notation

1. $\mathbb{W}$: the set of whole numbers: $\{1, 2, 3, 4, \dots\}$.
2. $\mathbb{N}$: the set of natural numbers: $\{0, 1, 2, 3, \dots\}$.
3. $\mathbb{Z}$: the set of integers: $\{\dots, -2, -1, 0, 1, 2, \dots\}$.
4. $\mathbb{Q}$: the set of rational numbers: $\{x : x = \frac{p}{q}, p, q \in \mathbb{Z}, q \neq 0\}$.
5. $\mathbb{R}$: the set of real numbers.
6. $\mathbb{C}$: the set of complex numbers: $\{z : z = x + iy,\ x, y \in \mathbb{R}\}$.
7. $\mathbb{Z}/n\mathbb{Z}$: the additive group of integers mod n.
8. $(\mathbb{Z}/n\mathbb{Z})^*$: the multiplicative group of invertible elements mod n.
9. $a|b$: a divides b, *i.e.* the remainder after integer division $\frac{b}{a}$ is 0.
10. (a, b) : greatest common divisor (gcd) of a and b, often written $\gcd(a, b)$.
11. $x \equiv y(\mod n)$: there exists an integer a such that $x = y + an$.
12. wlog : without loss of generality.
13. s.t. : such that.
14. $\forall$: for all.

15. $\exists$: there exists.

16. big O notation : $A(x) = O(B(x))$, read "$A(x)$ is of order $B(x)$", means $\exists C > 0$ such that $\forall x, |A(x)| \leq C\, B(x)$.

17. $|S|$ or $\#S$: number of elements in the set S.

18. p : usually a prime number.

19. n : usually an integer.

1.2 Efficient Algorithms

For computational purposes, often having an algorithm to compute a quantity is not enough; we need an algorithm which will compute *quickly*. Below we study three standard problems, and show how to either rearrange the operations more efficiently, or give a more efficient algorithm than the obvious candidate.

1.2.1 Polynomial Evaluation

Let $f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$. The obvious way to evaluate is to calculate x^n and multiply by a_n (n multiplications), calculate x^{n-1} and multiply by a_{n-1} ($n-1$ multiplications) and add, et cetera. There are n additions and $\sum_{k=0}^{n} k$ multiplications, for a total of $n + \frac{n(n+1)}{2}$ operations. Thus, the standard method leads to $O(n^2)$ computations.

Instead, consider the following:

$$\left(\Big((a_n x + a_{n-1})x + a_{n-2}\Big)x + \cdots + a_1\right)x + a_0. \tag{1.1}$$

For example,

$$7x^4 + 4x^3 - 3x^2 - 11x + 2 = \left(\Big((7x+4)x - 3\Big)x - 11\right)x + 2. \tag{1.2}$$

Evaluating the long way takes 14 steps; cleverly rearranging takes 8 steps.

Exercise 1.2.1. *Prove that the second method takes at most $2n$ steps to evaluate $a_n x^n + \cdots a_0$.*

1.2.2 Exponentiation

Consider x^n. The obvious way to evaluate involves $n-1$ multiplications. By writing n in base two, we can evaluate x^n in at most $2\log_2 n$ steps.

Let k be the largest integer such that $2^k \leq n$. Then $\exists a_i \in \{0,1\}$ such that

$$n \;=\; a_k 2^k + a_{k-1}2^{k-1} + \cdots + a_1 2 + a_0. \tag{1.3}$$

It costs k multiplications to evaluate x^{2^i}, $i \leq k$. How? Consider $y_0 = x^{2^0}$, $y_1 = y_0 \cdot y_0 = x^{2^0} \cdot x^{2^0} = x^{2^1}$, $y_2 = y_1 \cdot y_1 = x^{2^2}, \ldots, y_k = y^{k-1} \cdot y^{k-1} = x^{2^k}$.

Then

$$\begin{aligned} x^n &= x^{a_k 2^k + a_{k-1}2^{k-1} + \cdots + a_1 2 + a_0} \\ &= x^{a_k 2^k} \cdot x^{a_{k-1}2^{k-1}} \cdots x^{a_1 2} \cdot x^{a_0} \\ &= \left(x^{2^k}\right)^{a_k} \cdot \left(x^{2^{k-1}}\right)^{a_{k-1}} \cdots \left(x^2\right)^{a_1} \cdot \left(x^1\right)^{a_0} \\ &= y_k^{a_k} \cdot y_{k-1}^{a_{k-1}} \cdots y_1^{a_1} \cdot y_0^{a_0}. \end{aligned} \tag{1.4}$$

As each $a_i \in \{0,1\}$, we have at most $k+1$ multiplications above (if $a_i = 1$ we have the term y_i in the product, if $a_i = 0$ we don't).

Thus, it costs k multiplications to evaluate the x^{2^i} ($i \leq k$), and at most another k multiplications to finish calculating x^n. As $k \leq \log_2 n$, we see that x^n can be determined in at most $2\log_2 n$ steps.

Note, however, that we do need more storage space for this method, as we need to store the values $y_i = x^{2^i}$, $i \leq \log_2 n$.

Exercise 1.2.2. *Instead of expanding n in base two, expand n in base three. How many calculations are needed to evaluate x^n this way? Why is it preferable to expand in base two rather than any other base?*

1.2.3 Euclidean Algorithm

The Euclidean Algorithm is an efficient way to determine the greatest common divisor of x and y, denoted $\gcd(x,y)$ or (x,y). Without loss of generality, assume $1 < x < y$.

The obvious way to determine $\gcd(x,y)$ is to divide x and y by all positive integers up to x. This takes at most $2x$ steps.

Let $[z]$ denote the greatest integer less than or equal to z. We write

$$y = \left[\frac{y}{x}\right] \cdot x + r_1,\ 0 \leq r_1 < x. \tag{1.5}$$

Exercise 1.2.3. *Prove that* $r_1 \in \{0, 1, \ldots, x-1\}$.

Exercise 1.2.4. *Prove* $\gcd(x, y) = \gcd(r_1, x)$. *Hint:* $r_1 = y - \left[\frac{y}{x}\right] \cdot x$.

We proceed in this manner until r_k equals zero or one. As each execution results in $r_i < r_{i-1}$, we proceed at most x times (although later we prove we need to apply these steps at most $2\log_2 x$ times).

$$\begin{aligned}
x &= \left[\frac{x}{r_1}\right] \cdot r_1 + r_2,\ 0 \leq r_2 < r_1 \\
r_1 &= \left[\frac{r_1}{r_2}\right] \cdot r_2 + r_3,\ 0 \leq r_3 < r_2 \\
r_2 &= \left[\frac{r_2}{r_3}\right] \cdot r_3 + r_4,\ 0 \leq r_4 < r_3 \\
&\vdots \\
r_{k-2} &= \left[\frac{r_{k-2}}{r_{k-1}}\right] \cdot r_{k-1} + r_k,\ 0 \leq r_k < r_{k-1}.
\end{aligned} \tag{1.6}$$

Exercise 1.2.5. *Prove that if* $r_k = 0$, *then* $\gcd(x, y) = r_{k-1}$, *while if* $r_k = 1$, *then* $\gcd(x, y) = 1$.

We now analyze how large k can be. The key observation is the following:

Lemma 1.2.6. *Consider three adjacent remainders in the expansion:* r_{i-1}, r_i *and* r_{i+1} *(where* $y = r_{-1}$ *and* $x = r_0$*). Then* $\gcd(r_i, r_{i-1}) = \gcd(r_{i+1}, r_i)$, *and* $r_{i+1} < \frac{r_{i-1}}{2}$.

Proof: We have the following relation:

$$r_{i-1} = \left[\frac{r_{i-1}}{r_i}\right] \cdot r_i + r_{i+1},\ 0 \leq r_{i+1} < r_i. \tag{1.7}$$

If $r_i \leq \frac{r_{i-1}}{2}$, then as $r_{i+1} < r_i$, we immediately conclude that $r_{i+1} < r_i$. If $r_i > \frac{r_{i-1}}{2}$, then we note that

$$r_{i+1} = r_{i-1} - \left[\frac{r_{i-1}}{r_i}\right] \cdot r_i. \tag{1.8}$$

But $\left[\frac{r_{i-1}}{r_i}\right] = 1$ (easy exercise). Thus $r_{i-1} < \frac{r_{i-1}}{2}$. $\square$

We count how often we apply Euclid's Algorithm. Going from $(x, y) = (r_0, r_{-1})$ to (r_1, r_0) costs one application. Every two applications leads to the first entry in the last pair being at most half of the second entry of the first pair.

Thus, if k is the largest integer such that $2^k \leq x$, we see we apply Euclid's Algorithm at most $1 + 2k \leq 1 + 2\log_2 x$ times. Each application requires one integer division, where the remainder is the input for the next step.

We have proven

Lemma 1.2.7. *Euclid's Algorithm requires at most $1 + 2\log_2 x$ divisions to find the greatest common denominator of x and y.*

Let us assume that $r_i = \gcd(x, y)$. Thus, the last equation before Euclid's Algorithm terminated was

$$r_{i-2} = \left[\frac{r_{i-2}}{r_{i-1}}\right] \cdot r_{i-1} + r_i,\ 0 \leq r_i < r_{i-1}. \tag{1.9}$$

Therefore, we can find integers a_{i-1} and b_{i-2} such that

$$r_i = a_{i-1}r_{i-1} + b_{i-2}r_{i-2}. \tag{1.10}$$

Looking at the second to last application of Euclid's algorithm, we find that there are integers a'_{i-2} and b'_{i-3} such that

$$r_{i-1} = a'_{i-2}r_{i-2} + b'_{i-3}r_{i-3}. \tag{1.11}$$

Substituting for $r_{i-1} = r_{i-1}(r_{i-2}, r_{i-3})$ in the expansion of r_i yields that there are integers a_{i-2} and b_{i-3} such that

$$r_i = a_{i-2}r_{i-2} + b_{i-3}r_{i-3}. \tag{1.12}$$

Continuing by induction, and recalling $r_i = \gcd(x, y)$ yields

Lemma 1.2.8. *There exist integers a and b such that $\gcd(x, y) = ax + by$. Moreover, Euclid's Algorithm gives a constructive procedure to find a and b.*

Exercise 1.2.9. *Find a and b such that $a \cdot 244 + b \cdot 313 = \gcd(244, 313)$.*

Exercise 1.2.10. *Add details to complete an alternate proof of the existence of a and b with $ax + by = \gcd(x, y)$:*

1. *Let d be the smallest positive value attained by $ax + by$ as we vary $a, b \in \mathbb{Z}$. Such a d exists: consider $(a, b) = (1, 0)$ or $(0, 1)$. Thus, $d = ax + by$. We now show $d = \gcd(x, y)$.*
2. *$\gcd(x, y)|d$.*
3. *Let $e = Ax + By > 0$. Then $d|e$. Therefore, for any choice of $A, B \in \mathbb{Z}$, $d|(Ax + By)$.*
4. *$d|x$ and $d|y$ (consider clever choices of A and B; one choice gives $d|x$, one gives $d|y$). Therefore $d|\gcd(x, y)$. As we've shown $\gcd(x, y)|d$, this completes the proof.*

Note this is a non-constructive proof. By minimizing $ax + by$, we obtain $\gcd(x, y)$, but we have no idea how many steps is required. Prove that a solution will be found either among pairs (a, b) with $a \in \{1, \ldots, y - 1\}$ and $-b \in \{1, \ldots, x - 1\}$, or $-a \in \{1, \ldots, y - 1\}$ and $b \in \{1, \ldots, x - 1\}$.

1.3 Mean Value Theorem

We recall some notation:

$[a, b] = \{x : a \leq x \leq b\}$. IE, $[a, b]$ is all x between a and b, including a and b. $(a, b) = \{x : a < x < b\}$. IE, (a, b) is all x between a and b, not including the endpoints a and b.

Theorem 1.3.1 (Mean Value Theorem). *Let $h(x)$ be differentiable on $[a, b]$. Then $\exists c \in [a, b]$ such that*

$$h(b) - h(a) = h'(c) \cdot (b - a). \tag{1.13}$$

What is the physical interpretation of the Mean Value Theorem? Let $h(x)$ represent the distance from the starting point at time x. The average speed from a to b is the distance traveled ($h(b) - h(a)$) divided by the elapsed time ($b - a$). As $h'(x)$ represents the speed at time x, the MVT says that there is some intermediate time at which you are traveling at the average speed.

The MVT follows immediately from the Intermediate Value Theorem:

Theorem 1.3.2 (Intermediate Value Theorem). *Let f be a continuous function on $[a, b]$. $\forall C$ between $f(a)$ and $f(b)$, $\exists c \in [a, b]$ such that $f(c) = C$. In other words, all intermediate values of a continuous function are obtained.*

1.3.1 Sketch of Proof of the MVT

The MVT follows from Rolle's Theorem:

Theorem 1.3.3 (Rolle's Theorem). *Let f be differentiable on $[a, b]$, and assume $f(a) = f(b) = 0$. Then there exists a $c \in [a, b]$ such that $f'(c) = 0$.*

Why? Assume Rolle's Theorem. Consider the function

$$h(x) = f(x) - \frac{f(b) - f(a)}{b - a}(x - a) - f(a). \tag{1.14}$$

.

Note $h(a) = f(a) - f(a) = 0$ and $h(b) = f(b) - (f(b) - f(a)) - f(a) = 0$. Thus, the conditions of Rolle's Theorem are satisfied for $h(x)$, and there is some $c \in [a, b]$ such that $h'(c) = 0$. But

$$h'(c) \;=\; f'(c) - \frac{f(b) - f(a)}{b - a}. \tag{1.15}$$

Rewriting yields $f(b) - f(a) = f'(c) \cdot (b - a)$.

Thus, it is sufficient to prove Rolle's Theorem to prove the MVT.

Without loss of generality, assume $f'(a)$ and $f'(b)$ are non-zero. If either were zero, we would be done.

Multiplying $f(x)$ by -1 if needed, we may assume $f'(a) > 0$.

Case 1**:** $f'(b) < 0$**:** As $f'(a) > 0$ and $f'(b) < 0$, the Intermediate Value Theorem, applied to $f'(x)$, asserts that all intermediate values are attained. As

$f'(b) < 0 < f'(a)$, this implies the existence of a $c \in (a, b)$ such that $f'(c) = 0$.

Case 2**:** $f'(b) > 0$**:** $f(a) = f(b) = 0$, and the function f is increasing at a and b. If x is real close to a, then $f(x) > 0$ because $f'(a) > 0$.

This follows from the fact that

$$f'(0) \;=\; \lim_{x \to 0} \frac{f(x) - f(0)}{x}. \tag{1.16}$$

As $f'(0) > 0$, the limit is positive. As the denominator is positive for $x > 0$, the numerator must be positive. Thus, $f(x)$ must be greater than $f(0)$ for small x.

Similarly, $f'(b) > 0$ implies $f(x) < f(b) = 0$ for x near b.

Therefore, the function $f(x)$ is positive for x slightly greater than a and negative for x slightly less than b. If the first derivative were always positive, then $f(x)$ could never be negative as it starts at 0 at a. This can be seen by again using the limit definition of the first derivative to show that if $f'(x) > 0$, then the function is increasing near x. See the next section for more details.

Thus, the first derivative cannot always be positive. Either there must be some point $y \in (a, b)$ such that $f'(y) = 0$ (and we are then done!) or $f'(y) < 0$. By the IVT, as 0 is between $f'(a)$ (which is positive) and $f'(y)$ (which is negative), there is some $c \in (a, y) \subset [a, b]$ such that $f'(c) = 0$.

1.3.2 Sign of the Derivative

As it is such an important concept, let us show that $f'(x) > 0$ implies $f(x)$ is increasing at x. The definition of the derivative gives

$$f'(x) \;=\; \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}. \tag{1.17}$$

If $\Delta x > 0$, the denominator is positive. As the limit is positive, for Δx sufficiently small, the numerator must be positive. Thus, Δx positive and small implies $f(x + \Delta x) > f(x)$.

If $\Delta x < 0$, the denominator is negative. As the limit is positive, for Δx sufficiently small, the numerator must be negative. Thus, Δx negative and small implies $f(x + \Delta x) < f(x)$.

Therefore, if $f'(x)$ is positive, then f is increasing at x. Similarly we can show if $f'(x)$ is negative then f is decreasing at x.

1.3.3 Intermediate Value Theorem

We have reduced all our proofs to the intuitively plausible IVT: if C is between $f(a)$ and $f(b)$ for some continuous function f, then $\exists c \in (a, b)$ such that $f(c) = C$.

Here is a sketch of a proof using the method Divide and Conquer. Without loss of generality, assume $f(a) < C < f(b)$. Let x_1 be the midpoint of $[a, b]$. If $f(x_1) = C$ we are done. If $f(x_1) < C$, we look at the interval $[x_1, b]$. If $f(x_1) > C$ we look at the interval $[a, x_1]$.

In either case, we have a new interval, call it $[a_1, b_1]$, such that $f(a_1) < C < f(b_1)$, and the interval has size half that of $[a, b]$. Continuing in this manner, constantly taking the midpoint and looking at the appropriate half-interval, we see one of two things may happen.

First, we may be lucky and one of the midpoints may satisfy $f(x_n) = C$. In this case, we have found the desired point c.

Second, no midpoint works. Thus, we divide infinitely often, getting a sequence of points x_n. This is where rigorous mathematical analysis is required.

We claim the sequence of points x_n converge to some number $X \in (a, b)$. Clearly it can't be an endpoint. We keep getting smaller and smaller intervals (of half the size of the previous and contained in the previous) where $f(x) < C$ at the left endpoint, and $f(x) > C$ at the right endpoint. By continuity at the point X, eventually $f(x)$ must be close to $f(X)$ for x close to X.

If $f(X) < C$, then eventually the right endpoint cannot be greater than C; if $f(X) > C$, eventually the left endpoint cannot be less than C. Thus, $f(X) = C$.

1.3.4 Taylor Series

Using just the Mean Value Theorem, we prove the n^{th} Taylor Series Approximation. Namely, if f is differentiable at least $n + 1$ times on $[a, b]$, then $\forall x \in [a, b]$, $f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(x - a)^k$ plus an error that is at most $\max_{a \le c \le x} |f^{(n+1)}(c)| \cdot |x - a|^{n+1}$.

Assuming f is differentiable $n+1$ times on $[a, b]$, we apply the MVT multiple times to bound the error between $f(x)$ and its Taylor Approximations.

Let

$$
\begin{aligned}
f_n(x) &= \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(x-a)^k \\
h(x) &= f(x) - f_n(x). \qquad (1.18)
\end{aligned}
$$

$f_n(x)$ is the n^{th} Taylor Series Approximation to $f(x)$. Note $f_n(x)$ is a polynomial of degree n.

We want to bound $|h(x)|$ for $x \in [a,b]$. Without loss of generality (basically, for notational convenience), we may assume $a = 0$ and $f(a) = 0$.

Thus, $h(0) = 0$. Applying the MVT to h yields

$$
\begin{aligned}
h(x) &= h(x) - h(0) \\
&= h'(c_1) \cdot (x-0) \\
&= \Big(f'(c_1) - f_n'(c_1)\Big)x \\
&= \Big(f'(c_1) - \sum_{k=1}^{n} \frac{f^{(k)}(0)}{k!} \cdot k(c_1-0)^{k-1}\Big)x \\
&= \Big(f'(c_1) - \sum_{k=1}^{n} \frac{f^{(k)}(0)}{(k-1)!} c_1^{k-1}\Big)x \\
&= h_1(c_1)x. \qquad (1.19)
\end{aligned}
$$

We now apply the MVT to $h_1(u)$. Note that $h_1(0) = 0$. Therefore

$$
\begin{aligned}
h_1(c_1) &= h_1(c_1) - h_1(0) \\
&= h_1'(c_2) \cdot (c_1 - 0) \\
&= \Big(f''(c_2) - f_n''(c_2)\Big)c_1 \\
&= \Big(f''(c_2) - \sum_{k=2}^{n} \frac{f^{(k)}(0)}{(k-1)!} \cdot (k-1)(c_2-0)^{k-2}\Big)c_1 \\
&= \Big(f''(c_2) - \sum_{k=2}^{n} \frac{f^{(k)}(0)}{(k-2)!} c_2^{k-2}\Big)c_1 \\
&= h_2(c_1)c_1. \qquad (1.20)
\end{aligned}
$$

Therefore,

$$h(x) = f(x) - f_n(x) = h_2(c_2)c_1x, \quad c_2 \in [0, c_1],\ c_1 \in [0, x]. \tag{1.21}$$

Proceeding in this way a total of n times yields

$$h(x) \quad = \quad \Big(f^{(n)}(c_n) - f^{(n)}(0)\Big)c_{n-1}c_{n-2}\cdots c_2c_1x. \tag{1.22}$$

Applying the MVT to $f^{(n)}(c_n) - f^{(n)}(0)$ gives $f^{(n+1)}(c_{n+1}) \cdot (c_n - 0)$. Thus,

$$h(x) = f(x) - f_n(x) = f^{(n+1)}(c_{n+1})c_n\cdots c_1x, \quad c_i \in [0, x]. \tag{1.23}$$

Therefore

$$|h(x)| = |f(x) - f_n(x)| = M_{n+1}|x|^{n+1}, \quad M_{n+1} = \max_{c\in[0,x]} |f^{(n+1)}(c)|. \tag{1.24}$$

Thus, if f is differentiable $n + 1$ times, the n^{th} Taylor Series Approximation to $f(x)$ is correct within a multiple of $|x|^{n+1}$; further, the multiple is bounded by the maximum value of $f^{(n+1)}$ on $[0, x]$.

1.4 Probabilities of Discrete Events

1.4.1 Introduction

Let $\Omega = \{\omega_1, \omega_2, \omega_3, \dots\}$ be an at most countable set of events. We call Ω the **sample (or outcome) space**. We call the elements $\omega \in \Omega$ the **events**. Let $x : \Omega \to \mathbb{R}$. That is, for each event $\omega \in \Omega$, we attach a real number $x(\omega)$. We call x a **random variable**.

Example 1.4.1. *Flip a fair coin* 3 *times. The possible outcomes are*

$$\Omega = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}. \tag{1.25}$$

One possible random variable is $x(\omega)$ *equals the number of heads in* ω*. Thus,* $x(HHT) = 2$ *and* $x(TTT) = 0$*.*

Example 1.4.2. *Let Ω be the space of all flips of a fair coin where all but the last flip are tails, and the last is a head. Thus, $\Omega = \{H, TH, TTH, TTTH, \dots\}$. One possible random variable is $x(\omega)$ is the number of tails; another is $x(\omega)$ equals the number of the flip which is a head.*

We say $p(\omega)$ is a **probability function** on Ω if

1. $0 \leq p(\omega_i) \leq 1$ for all $\omega_i \in \Omega$.
2. $p(\omega) = 0$ if $\omega \notin \Omega$.
3. $\sum_i p(\omega_i) = 1$.

We call $p(\omega)$ the probability of event ω.

Often, we have a random variables where $x(\omega) = \omega$. In a convenient abuse of notation, we write X for Ω and x for $x(\omega)$ and ω. For example, consider two rolls of a fair die. Let X be the result of the first roll, and Y of the second. Then the sample space is $X = Y = \{1, 2, 3, 4, 5, 6\}$.

In general, consider X and Y with x_i occurring with probability $p(x_i)$ and y_j occurring with probability $q(y_j)$. We analyze the **joint probability** $r(x, y)$ of observing x and y.

X and Y are **independent** if $\forall x, y$, $r(x, y) = p(x)q(y)$. In the example of rolling a fair die twice, $r(x, y) = p(x)q(y) = \frac{1}{6} \cdot \frac{1}{6}$ if $x, y \in X = Y$, and 0 otherwise.

Exercise 1.4.3. *Consider again two rolls of a fair die. Now, let X represent the first roll, and Y the sum of the first two rolls. Prove X and Y are not independent.*

Events X_1 through X_N are **independent** if $p(x_1, \dots, x_N) = p_1(x_1) \cdots p(x_N)$.

Exercise 1.4.4. *Construct three events such that any two are independent, but all three are not independent. Hint: roll a fair die twice.*

1.4.2 Means

If $x(\omega) = \omega$, the **mean (or expected value)** of an event x is defined by

$$\bar{x} = \sum_i x_i p(x_i). \tag{1.26}$$

More generally, for a sample space Ω with events ω and a random variable $x(\omega)$, we have

$$\bar{x}(\omega) \;=\; \sum_i x(\omega_i)p(\omega_i). \tag{1.27}$$

For example, the mean of one roll of a fair die is 3.5.

Exercise 1.4.5. *Let X be the number of tosses of a fair coin needed before getting the first head. Thus, $X = \{1, 2, \dots\}$. Calculate $p(x_i)$ and $\bar{x}$. We could let Ω be the space of all tosses of a fair coin where all but the last toss are tails, and the last toss is a head. Then $x(\omega)$ is the number of tosses of ω.*

Instead of writing $\bar{x}$, we often write $E[x]$ or $E[X]$, read as **the expected value of x or X**. More generally, we would have $\bar{x}(\omega)$ and $E[x(\omega)]$.

The k^{th} moment of X is the expected value of x^k:

$$E[x^k] \;=\; \sum_i x_i^k p(x_i) \tag{1.28}$$

or

$$E[x^k(\omega)] \;=\; \sum_i x^k(\omega_i)p(\omega_i). \tag{1.29}$$

Lemma 1.4.6 (Additivity of the Means). *Let X and Y be two independent events with joint probability $r(x, y) = p(x)q(y)$. Let $z = x + y$. Then $E[z] = E[x+y] = E[x] + E[y]$.*

Proof:

$$\begin{aligned}
E[x+y] &= \sum_{(i,j)} (x_i + y_j) r(x_i, y_j) \\
&= \sum_i \sum_j (x_i + y_j) p(x_i) q(y_j) \\
&= \sum_i \sum_j x_i p(x_i) q(y_j) + \sum_i \sum_j y_j p(x_i) q(y_j) \\
&= \sum_i x_i p(x_i) \sum_j q(y_j) + \sum_i p(x_i) \sum_j y_j q(y_j) \\
&= E[x] \cdot 1 + 1 \cdot E[y] = E[x] + E[y].
\end{aligned} \tag{1.30}$$

The astute reader may notice that some care is needed to interchange the order of summations. If $\sum_i \sum_j |x_i y_j| r(x_i, y_j) < \infty$, then Fubini's Theorem is applicable, and we may interchange the summations at will.

We used the two events were independent to go from $\sum_{(i,j)} x_i r(x_i, y_j)$ to $\sum_i x_i p(x_i) \sum_j q(y_j) = E[x]$. Lemma 1.4.6 is true even if the two events are not independent.

If the events are not independent, we encounter sums like $\sum_i \sum_j x_i r(x_i, y_j)$; however, $\sum_j r(x_i, y_j) = p(x_i)$. Why? By summing over all possible y, we are asking what is the probability that $x = x_i$; we do not care what y is. Thus, $\sum_i \sum_j x_i r(x_i, y_j) = \sum_i x_i p(x_i) = E[x]$, and similarly for the other piece.

Exercise 1.4.7. *Write out the proof of the generalization of Lemma 1.4.6, where X and Y are not assumed independent.*

Given an outcome space $X = \{x_1, x_2, \dots\}$ with probabilities $p(x_i)$, let aX be shorthand for the event a times X with outcome space $\{ax_1, ax_2, \dots\}$ and probabilities $p_a(ax_i) = p(x_i)$.

Lemma 1.4.8. *Let X_1 through X_N be a finite collection of independent events. Let a_1 through a_N be real constants. Then*

$$E[a_1 x_1 + \cdots + a_N x_N] = a_1 E[x_1] + \cdots + a_N E[x_N]. \tag{1.31}$$

Lemma 1.4.9. *Let X and Y be independent events. Then $E[xy] = E[x]E[y]$.*

Exercise 1.4.10. *Prove Lemmas 1.4.8 and 1.4.9.*

1.4.3 Variances

The **variance** σ_x^2 (and its square-root, the **standard deviation** σ_x) measure how spread out a probability distribution is. Assume $x(\omega) = \omega$. Given an event X with mean $\bar{x}$, we define the standard deviation σ_x^2 by

$$\sigma_x^2 = \sum_i (x_i - \bar{x})^2 p(x_i). \tag{1.32}$$

More generally, given a sample space Ω, events ω, and a random variable $x : \Omega \to \mathbb{R}$,

$$\sigma_{x(\omega)}^2 = \sum_i \Big(x(\omega_i) - \bar{x}(\omega)\Big) p(\omega_i). \tag{1.33}$$

Exercise 1.4.11. *Let $X = \{0, 25, 50, 75, 100\}$ with probabilities $\{.2, .2, .2, .2, .2\}$. Let Y be the same outcome space, but with probabilities $\{.1, .25, .3, .25, .1\}$. Calculate the means and the variances of X and Y.*

For computing variances, instead of equation 1.32 one often uses

Lemma 1.4.12. $\sigma_x^2 = E[x^2] - E[x]^2$.

Proof: Recall $\bar{x} = E[x]$. Then

$$\begin{aligned}
\sigma_x^2 &= \sum_i \Big(x_i - E[x]\Big)^2 p(x_i) \\
&= \sum_i (x_i^2 - 2x_i E[x] + E[x]^2) p(x_i) \\
&= \sum_i x_i^2 p(x_i) - 2E[x] \sum_i x_i p(x_i) + E[x]^2 \sum_i p(x_i) \\
&= E[x^2] - 2E[x]^2 + E[x]^2 = E[x^2] - E[x]^2. \qquad (1.34)
\end{aligned}$$

The main result on variances is

Lemma 1.4.13 (Variance of a Sum). *Let X and Y be two independent events. Then $\sigma_{x+y}^2 = \sigma_x^2 + \sigma_y^2$.*

Proof: We constantly use the expected value of a sum of independent events is the sum of expected values (Lemma 1.4.6 and Lemma 1.4.8).

$$\begin{aligned}
\sigma_{x+y}^2 &= E[(x+y)^2] - E[(x+y)]^2 \\
&= E[x^2 + 2xy + y^2] - \Big(E[x] + E[y]\Big)^2 \\
&= \Big(E[x^2] + 2E[xy] + E[y^2]\Big) - \Big(E[x]^2 + 2E[x]E[y] + E[y]^2\Big) \\
&= \Big(E[x^2] - E[x]^2\Big) + \Big(E[y^2] - E[y]^2\Big) + 2\Big(E[xy] - E[x]E[y]\Big) \\
&= \sigma_x^2 + \sigma_y^2 + 2\Big(E[xy] - E[x]E[y]\Big). \qquad (1.35)
\end{aligned}$$

By Lemma 1.4.9, $E[xy] = E[x]E[y]$, completing the proof.

Lemma 1.4.14. *Consider n independent copies of the same event (for example, n flips of a coin or n rolls of a die). Then $\sigma_{nx} = \sqrt{n}\sigma_x$.*

Exercise 1.4.15. *Prove Lemma 1.4.14.*

Note that, if the event X has units of meters, then the variance σ_x^2 has units meters-squared, and the standard deviation σ_x and the mean $\bar{x}$ have units meters. Thus, it is the standard deviation that gives a good measure of the deviations of an event around the mean.

There are, of course, alternate measures one can use. For example, one could consider

$$\sum_i (x_i - \bar{x}) p(x_i). \tag{1.36}$$

Unfortunately, this is a signed quantity, and large positive deviations can cancel with large negatives. This leads us to consider

$$\sum_i |x_i - \bar{x}| p(x_i). \tag{1.37}$$

While this has the advantage of avoiding cancellation of errors (as well as having the same units as the events), the absolute value function is not a good function analytically. For example, it is not differentiable. This is primarily why we consider the standard deviation (the square-root of the variance).

Exercise 1.4.16. *Consider the following set of data: for $i \in \{1, \ldots, n\}$, given x_i one observes y_i. Believing that X and Y are linearly related, find the best fit straight line. Namely, determine constants a and b that minimize the error (calculated via the variance)*

$$\sum_{i=1}^{n} \Big(y_i - (a x_i + b) \Big)^2 \;=\; \sum_{i=1}^{n} \Big(\textit{Observed}_i - \textit{Predicted}_i \Big)^2. \tag{1.38}$$

Hint: use Multi-variable Calculus to find linear equations for a and b, and then solve with Linear Algebra.

If instead of measuring total error by the squares of the individual error (for example, using the absolute value), closed form expressions for a and b become significantly harder.

If one requires that $a = 0$, show that the b leading to least error is $b = \bar{y} = \frac{1}{n} \sum_i y_i$.

1.4.4 Random Walks

Consider the classical problem of a drunk staggering home from a lamp post late at night. We flip a fair coin N times. With probability $\frac{1}{2}$ we get heads (tails). For each head (tail) the drunk staggers one unit to the right (left). How far do we expect the drunk to be?

It is very unlikely the drunk will be very far to the left or right.

Exercise 1.4.17. *Let x be $+1$ if we flip a head, -1 for a tail. For a fair coin, prove $E[x] = 0$, $\sigma_x^2 = 1$, $\sigma_x = 1$.*

Exercise 1.4.18. *Let $p_N(y)$ be the probability that after N flips of a fair coin, the drunk is y units to the right of the origin (lamp post).*

1. *Prove $p_N(y) = p_N(-y)$.*
2. *Consider $N = 2M$. Prove $p_{2M}(2k) = \binom{2M}{M+k}\frac{1}{2^{2M}}$, where $\binom{n}{r} = \frac{n!}{r!(n-r)!}$*
3. *Use Stirling's formula ($n! \approx n^n e^{-n}\sqrt{2\pi n} = \sqrt{2\pi}n^{n+\frac{1}{2}}e^{-n}$) to approximate $p_N(y)$.*

Label the coin tosses X_1 through X_N. Let X denote a generic toss of the coin, and Y_N be the distance of the drunkard after N tosses. By Lemma 1.4.8, $E[y_N] = E[x_1 + \cdots + x_N] = E[x_1] + \cdots + E[x_N]$. As each $E[x_i] = E[x] = 0$, $E[y_N] = 0$.

Thus, we expect the drunkard to be at the lamp post. How spread out is his expected position? By Lemma 1.4.14,

$$\sigma_{y_N} = \sigma_{Nx} = \sqrt{N}\sigma_x = \sqrt{N}. \tag{1.39}$$

This means that a *typical* distance from the origin is $\sqrt{N}$. This is called a *diffusion process* and is very common in the real world.

1.4.5 Bernoulli Process

Recall $\binom{N}{r} = \frac{N!}{r!(N-r)!}$ is the number of ways to choose r objects from N objects when order does not matter. Consider n independent repetitions of an event with only two possible outcomes. We typically call one outcome **success** and the other **failure**, the event a **Bernoulli Trial**, and a collection of independent Bernoulli Trials a **Bernoulli Process**.

In each Bernoulli Trial, let there be probability p of success and $q = 1 - p$ of failure. Often, we represent a success with 1 and a failure with 0.

Exercise 1.4.19. *For a Bernoulli Trial, show $\bar{x} = p$, $\sigma_x^2 = pq$, and $\sigma_x = \sqrt{pq}$.*

Let Y_N be the number of successes in N trials. Clearly, the possible values are $Y_N = \{0, 1, \cdots, N\}$. We analyze $p_N(k)$. Rigorously, the sample space Ω is all possible sequences of N trials, and the random variable $y_N : \Omega \to \mathbb{R}$ is given by $y_N(\omega)$ equals the number of successes in ω.

If $k \in Y_N$, we need k successes and $N - k$ failures. We don't care what order we have them (ie, if $k = 4$ and $N = 6$ then $SSFSSF$ and $FSSSSF$ both contribute). Each such string of k successes and $N - k$ failures has probability of $p^k \cdot (1-p)^{N-k}$. There are $\binom{N}{k}$ such strings.

Thus, $p_N(k) = \binom{N}{k} p^k \cdot (1-p)^{N-k}$ if $k \in \{0, 1, \cdots, N\}$ and 0 otherwise.

By clever algebraic manipulations, one can directly evaluate the mean $\overline{y_N}$ and the variance $\sigma_{y_N}^2$; however, Lemmas 1.4.8 and 1.4.14 allow one to calculate both quantities immediately, once one knows the mean and variance for one occurrence.

Lemma 1.4.20. *For a Bernouilli Process with N trials, each having probability p of success, the expected number of successes is $\overline{y_N} = Np$, and the variance is $\sigma_{y_N}^2 = Npq$.*

Exercise 1.4.21. *Prove Lemma 1.4.20.*

Consider the following problem: Let $Z = \{0, 1, 2, \dots\}$ be the number of trials before the first success. What is $\bar{z}$ and σ_z^2?

First, we determine $p(k)$, the probability that the first success occurs after k trials. Clearly this probability is non-zero only for k a positive integer, in which case the string of results must be $k-1$ failures followed by 1 success. Therefore,

$$p(k) \;=\; p \cdot (1-p)^{k-1} \text{ if } k \in \{1, 2, \dots\}, \text{ and } 0 \text{ otherwise.} \tag{1.40}$$

To determine the mean $\bar{z}$ we must evaluate

$$\begin{aligned} \bar{z} &= \sum_{k=1}^{\infty} k \cdot p \cdot (1-p)^{k-1} \\ &= p \sum_{k=1}^{\infty} k q^{k-1}, \; 0 < q = 1 - p < 1. \end{aligned} \tag{1.41}$$

Consider the geometric series

$$f(q) \;=\; \sum_{k=0}^{\infty} q^k \;=\; \frac{1}{1-q}. \tag{1.42}$$

A careful analysis shows we can differentiate term by term if $0 \le q < 1$. Then

$$f'(q) \;=\; \sum_{k=0}^{\infty} k q^{k-1} \;=\; \frac{1}{(1-q)^2}. \tag{1.43}$$

Recalling $q = 1 - p$ and substituting yields

$$\begin{aligned} \bar{z} &= p \sum_{k=1}^{\infty} k q^{k-1} \\ &= \frac{p}{\Big(1-(1-p)\Big)^2} \;=\; \frac{1}{p}. \end{aligned} \tag{1.44}$$

Differentiating under the summation sign is a powerful tool in Probability Theory.

Exercise 1.4.22. *Calculate σ_z^2. Hint: differentiate $f(q)$ twice.*

1.4.6 Poisson Distribution

Divide the unit interval into N equal pieces. Consider N independent Bernoulli Trials, one for each sub-interval. If the probability of a success is $\frac{\lambda}{N}$, then by Lemma 1.4.20 the expected number of successes is $N \cdot \frac{\lambda}{N} = \lambda$.

We consider the limit as $N \to \infty$. Obviously, we still expect λ successes in each interval, but what is the probability of 3λ successes? How long do we expect to wait between successes?

We call this a **Poisson process with parameter** λ. For example, look at the midpoints of the N intervals. At each midpoint we have a Bernoulli Trial with probability of success $\frac{\lambda}{N}$ and failure $1 - \frac{\lambda}{N}$.

We determine the $N \to \infty$ limits. For fixed N, the probability of k successes in a unit interval is

$$\begin{aligned} p_N(k) &= \binom{N}{k}\Big(\frac{\lambda}{N}\Big)^k\Big(1-\frac{\lambda}{N}\Big)^{N-k} \\ &= \frac{N!}{k!(N-k)!}\frac{\lambda^k}{N^k}\Big(1-\frac{\lambda}{N}\Big)^{N-k} \\ &= \frac{N\cdot(N-1)\cdots(N-k+1)}{N\cdot N\cdots N}\frac{\lambda^k}{k!}\Big(1-\frac{\lambda}{N}\Big)^{N}\Big(1-\frac{\lambda}{N}\Big)^{-k} \\ &= 1\cdot\Big(1-\frac{1}{N}\Big)\cdots\Big(1-\frac{k-1}{N}\Big)\frac{\lambda^k}{k!}\Big(1-\frac{\lambda}{N}\Big)^{N}\Big(1-\frac{\lambda}{N}\Big)^{-k} \end{aligned} \tag{1.45}$$

For fixed, finite k, as $N\to\infty$, the first k factors in $p_N(k)$ tend to 1, $\Big(1-\frac{\lambda}{N}\Big)^N\to e^{-\lambda}$, and $\Big(1-\frac{\lambda}{N}\Big)^{-k}\to 1$.

Thus, we are led to the **Poisson Distribution**: Given a parameter λ (interpreted as the expected number of occurrences per unit interval), the probability of k occurrences in a unit interval is $p(k)=\frac{\lambda^k}{k!}e^{-\lambda}$ for $k\in\{0,1,2,\dots\}$.

Exercise 1.4.23. *Check that $p(k)$ given above is a probability distribution. Namely, show $\sum_{k\geq 0}p(k)=1$.*

Exercise 1.4.24. *Show, for the Poisson Distribution, that the mean $\bar{x}=\lambda$ and the variance $\sigma_x^2=\lambda$. Hint: let*

$$f(\lambda) \;=\; \sum_{k=0}^{\infty}\frac{\lambda^k}{k!} \;=\; e^{\lambda}. \tag{1.46}$$

Differentiate once to determine the mean, twice to determine the variance.

1.4.7 Continuous Poisson Distribution

We calculate a very important quantity related to the Poisson Distribution (with parameter λ), namely, how long does one expect to wait between successes?

We've discussed that we expect λ successes per unit interval, and we've calculated the probability of k successes per unit interval.

Start counting at 0, and assume the first success is at x. What is $p_S(x)$? As before, we divide each unit interval into N equal pieces, and consider a Bernoulli Trial at the midpoint of each sub-interval, with probability $\frac{\lambda}{N}$ of success.

We have approximately $\frac{x-0}{1/N} = Nx$ midpoints from 0 to x (with N midpoints per unit interval). Let $\lceil y \rceil$ be the smallest integer greater than or equal to y. Then we have $\lceil Nx \rceil$ midpoints, where the results of the Bernoulli Trials of the first $\lceil Nx \rceil - 1$ midpoints are all failures and the last is a success.

Thus, the probability of the first success occuring in an interval of length $\frac{1}{N}$ containing x (with N divisions per unit interval) is

$$p_{N,S}(x) \quad = \quad \left(1 - \frac{\lambda}{N}\right)^{\lceil Nx \rceil - 1} \cdot \left(\frac{\lambda}{N}\right)^{1}. \tag{1.47}$$

For N large, the above converges to $e^{-\lambda x}\frac{\lambda}{N}$.

We say $p(x)$ is a **continuous probability distribution on** $\mathbb{R}$ if

1. $p(x) \geq 0$ for all $x \in \mathbb{R}$.
2. $\int_{\mathbb{R}} p(x)dx = 1$.
3. Probability$(a \leq x \leq b) = \int_a^b p(x)dx$.

We call $p(x)$ the **probability density function**.

Thus, as $N \to \infty$, we see the probability density function $p_S(x) = \lambda e^{-\lambda x}$. In the special case of $\lambda = 1$, we get the standard exponential decay, e^{-x}.

For instance, let $\pi(M)$ be the number of primes that are at most M. The Prime Number Theorem states $\pi(M) = \frac{M}{\log M}$ plus lower order terms.

Thus, the average spacing between primes around M is about $\log M$. We can model the distribution of primes as a Poisson Process, with parameter $\lambda = \lambda_M = \frac{1}{\log M}$. While possible locations of primes (obviously) is discrete (it must be an integer, and in fact the location of primes aren't independent), a Poisson model often gives very good heuristics.

We can often renormalize so that $\lambda = 1$. This is denoted **unit mean spacing**. For example, one can show the M^{th} prime p_M is about $M \log M$, and spacings between primes around p_M is about $\log M$. Then the normalized primes, $q_M \approx \frac{p_M}{\log M}$ will have unit mean spacing and $\lambda = 1$.

1.4.8 Central Limit Theorem

$X_1, X_2, X_3, \ldots$ are an infinite sequence of random variables such that the X_j are independent identically distributed random variables (abbreviated i.i.d.r.v.) with $E[X_j] = \bar{X}_j = 0$ (can always renormalize by shifting) and variance $E[X_j^2] = 1$. Let $S_N = \sum_{j=1}^{N} X_j$.

Theorem 1.4.25. *Fix $-\infty < a \leq b < \infty$. Then as $N \to \infty$,*

$$\textit{Prob}\Big(\frac{S_N}{\sqrt{N}} \in [a,b]\Big) \to \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{t^2}{2}}\, dt. \tag{1.48}$$

The probability function is called the Gaussian or the Normal distribution. This is the universal curve of probability. Note how robust the Central Limit Theorem is: it doesn't depend on fine properties of the X_j.

1.5 Iteration of Functions

We first consider iterating a linear transformation, for definiteness, a 2×2 map from $\mathbb{R}^2 \to \mathbb{R}^2$. We see here that the limiting behaviour is extremely well controlled by the initial conditions, and in general small changes in initial conditions yield small changes in the limit.

The situation is very different for iterations of non-linear maps. We will discuss several well-known examples (Newton's method, Mandelbrot sets), and discuss the barest beginning of chaotic behaviour.

1.5.1 Linear Functions

Rabbits eat grass. Foxes eat rabbits. You know the story.

Say that, at time n, there are x_n rabbits and y_n foxes. The more rabbits there are, the more rabbits will be born now; the more foxes there are, the more rabbits will die. Conversely, the more rabbits there are, the more foxes will be able to survive; but the more foxes there are beyond a certain point, the more foxes will die of starvation.

Notice that we are assuming that foxes can reproduce only after gorging themselves on rabbit. Leaving that aside, we will make the additional assumption that all dependences are linear. (Here we are considering our own interests, not the

foxes' or the rabbits'.) In other words,

$$x_{n+1} = ax_n - by_n,$$
$$y_{n+1} = cx_n - dy_n.$$

We can write

$$\vec{w}_n = \begin{pmatrix} x_n \\ y_n \end{pmatrix}.$$

Then

$$\vec{w}_{n+1} = A\vec{w}_n,$$

where

$$A = \begin{pmatrix} a & -b \\ c & -d \end{pmatrix}.$$

By induction it follows that

$$\vec{w}_n = A^n \vec{w}_0.$$

This comes in quite handy if we want to get at a closed expression for the number of foxes. How so? Let $\vec{v}_1$, $\vec{v}_2$ be the eigenvectors of A; let λ_1 and λ_2 be the corresponding eigenvalues. Let $\vec{w}_0 = \alpha_1\vec{v}_1 + \alpha_2\vec{v}_2$. Then

$$\vec{w}_n = \alpha_1\lambda_1^n\vec{v}_1 + \alpha_2\lambda_2^n\vec{v}_2.$$

This expression is dominated by whichever of λ_1 and λ_2 has the largest absolute value, if either.

You can see how this generalizes to arbitrarily many variables (species).

Exercise 1.5.1 (One-variable case). *Let $x_{n+1} = ax_n + b$. Give an expression for x_n in terms of x_0, a and b.*

1.5.2 Newton's method

Newton's method is an algorithm to find roots of equations. Let f be a differentiable function on $\mathbb{R}$, and assume we want to find a solution to $f(x) = 0$. Start with x_0 such that $f(x_0)$ is small (we call x_0 the initial guess). Draw the tangent to the graph of f at x_0, which is given by the equation

$$y - f(x_0) = f'(x_0) \cdot (x - x_0). \tag{1.49}$$

Let x_1 be the x-intercept of the tangent line; x_1 is the next guess for the root. Simple algebra gives

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}. \tag{1.50}$$

We now iterate, and apply the above procedure to x_1, obtaining

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}. \tag{1.51}$$

If we let $g(x) = x - \frac{f(x)}{f'(x)}$, we notice we have the sequence

$$x_0, \;\; g(x_0), \;\; g(g(x_0)), \;\; \ldots \tag{1.52}$$

This sequence will, we hope, converge to the root x. That is in fact true for x_0 close enough to x and for f good. How close x_0 has to be is a tricky matter. If there are several roots to f, which root the sequence converges to is something that depends on the initial value x_0 and the function f. This dependence is a very tricky matter. In fact its behaviour is what is known technically as **chaotic**. Informally, we can say that we have **chaos** when tiny changes in the initial value give us very palpable changes in the output.

Exercise 1.5.2. *Let $f(x)$ be a degree n polynomial with complex coefficients. By the Fundamental Theorem of Algebra, there are n (not necessarily distinct) roots. Assume there are m distinct roots. Assign m colors, one to each root. Given a point $x \in C$, we color x with the color of the root that x approaches under Newton's method. Write a computer program to color such sets for some simple polynomials, for example for $x^n - 1 = 0$ for $n = 2, 3$ or 4.*

Exercise 1.5.3. *Assume $f'(x) > 0$. Show that if $g(x) = x$, then $f(x) = 0$. We say such an x is a **fixed point** of g.*

Exercise 1.5.4. *Let $f(x) = x^2 - 3$, and guess $x_0 = 2$. Investigate the sequence of x_n's. What do they seem to converge to? How quickly do they seem to converge?*

For good problems on Newton's method, see [Ru], problems 3.16, 3.18, and 5.25.

1.5.3 The Mandelbrot set

The Mandelbrot set is one of the best-known examples of chaos. Start with any complex number z_0, and define $\{z_n\}$ by

$$z_{n+1} = z_n^2 + z_n. \tag{1.53}$$

We are iterating the quadratic map $f(z) = z^2 + z$. Thus, if we started with z_0, we have the values

$$z_0, \;\; z_0^2 + z_0, \;\; (z_0^2 + z_0)^2 + (z_0^2 + z_0), \;\; \ldots \tag{1.54}$$

Does the sequence $\{z_n\}$ converge? The answer to this can be different for z_0 and $z_0 + \epsilon$, where ϵ is arbitrarily small. The set of complex numbers z_0 for which the sequence convergence is called the **Mandelbrot set**. This is a **fractal**, that is, a set parts of which look much like the set as a whole.

Chapter 2

Algebraic and Transcendental Numbers

2.1 Definitions and Cardinalities of Sets

2.1.1 Definitions

A function $f : A \to B$ is **one-to-one** if $f(x) = f(y)$ implies $x = y$; f is **onto** if given any $b \in B$, $\exists a \in A$ with $f(a) = b$. f is a **bijection** if f is a one-to-one and onto function.

We say two sets A and B **have the same cardinality** (ie, are the same size) if there is a bijection $f : A \to B$. We denote this by $|A| = |B|$. If A has finitely many elements (say n elements), A is **finite** and $|A| = n < \infty$.

Exercise 2.1.1. *Show two finite sets have the same cardinality if and only if they have the same number of elements.*

Exercise 2.1.2. *If f is a bijection from A to B, prove there is a bijection $g = f^{-1}$ from B to A.*

A is **countable** if there is a bijection between A and the integers $\mathbb{Z}$. A is **at most countable** if A is either finite or countable.

Recall a binary relation R is an **equivalence relation** if

1. Reflexive: $R(x, x)$ is true (x is equivalent to x).
2. Symmetric: $R(x, y)$ true implies $R(y, x)$ is true (if x is equivalent to y then y is equivalent to x).

3. Transitive: $R(x, y)$ and $R(y, z)$ true imply $R(x, z)$ is true (if x is equivalent to y and y is equivalent to z, then x is equivalent to z).

We often denote equivalence by $\equiv$ or $=$.

Exercise 2.1.3. *Let $x, y, z \in \mathbb{Z}$, and let $n \in \mathbb{Z}$ be given. Define $R(x, y)$ to be true if $n|(x - y)$ and false otherwise. Prove R is an equivalence relation. We denote it by $x \equiv y$.*

Exercise 2.1.4. *Let x, y, z be subsets of X (for example, $X = \mathbb{Q}, \mathbb{R}, \mathbb{C}, \mathbb{R}^n$, et cetera). Define $R(x, y)$ to be true if $|x| = |y|$ (the two sets have the same cardinality), and false otherwise. Prove R is an equivalence relation.*

2.1.2 Countable Sets

We show several sets are countable. Consider the set of non-negative integers $\mathbb{N}$. Define $f : \mathbb{N} \to \mathbb{Z}$ by $f(2n) = n$, $f(2n + 1) = -n - 1$. By inspection, we see f gives the desired bijection.

Consider $\mathbb{W} = \{1, 2, 3, \dots\}$ (the positive integers). Then $f : \mathbb{W} \to \mathbb{Z}$ defined by $f(2n) = n$, $f(2n + 1) = -n$ gives the desired bijection.

Thus, we have proved

Lemma 2.1.5. *To show a set S is countable, it is sufficient to find a bijection from S to either $\mathbb{Z}$, $\mathbb{N}$ or $\mathbb{W}$.*

We need the intuitively plausible

Lemma 2.1.6. *If $A \subset B$, then $|A| \leq |B|$.*

We can then prove

Lemma 2.1.7. *If $f : A \to C$ is a one-to-one function (not necessarily onto), then $|A| \leq |C|$. Further, if $C \subset A$, then $|A| = |C|$.*

Exercise 2.1.8. *Prove Lemmas 2.1.6 and 2.1.7.*

If A and B are sets, the **cartesian product** $A \times B$ is $\{(a, b) : a \in A, b \in B\}$.

Theorem 2.1.9. *If A and B are countable, so is $A \cup B$ and $A \times B$.*

Proof: we have bijections $f : \mathbb{N} \to A$ and $g : \mathbb{N} \to B$. Thus, we can label the elements of A and B by

$$\begin{aligned} A &= \{a_0, a_1, a_2, a_3, \dots\} \\ B &= \{b_0, b_1, b_2, b_3, \dots\}. \end{aligned} \tag{2.1}$$

Assume $A \cap B$ is empty. Define $h : \mathbb{N} \to A \cup B$ by $h(2n) = a_n$ and $h(2n+1) = b_{n-1}$. We leave to the reader the case when $A \cap B$ is not empty.

To prove the second claim, consider the following function $h : W \to A \times B$:

$$\begin{aligned} &h(1) = (a_0, b_0) \\ &h(2) = (a_1, b_0), h(3) = (a_1, b_1), h(4) = (a_0, b_1) \\ &h(5) = (a_2, b_0), h(6) = (a_2, b_1), h(7) = (a_2, b_2), h(8) = (a_1, b_2), h(9) = (a_0, b_2) \\ &\quad \vdots \\ &h(n^2+1) = (a_n, b_0), h(n^2+2) = (a_n, b_{n-1}), \dots, \\ &\qquad h(n^2+n+1) = (a_n, b_n), h(n^2+n+2) = (a_{n-1}, b_n), \dots, \\ &\qquad h((n+1)^2) = (a_0, b_n) \\ &\quad \vdots \end{aligned} \tag{2.2}$$

Basically, look at all pairs of integers in the first quadrant (including those on the axes). Thus, we have pairs (a_x, b_y). The above function h starts at $(0, 0)$, and then moves through the first quadrant, hitting each pair once and only once, by going up and over. Draw the picture! □

Corollary 2.1.10. *Let A_i be countable $\forall i \in \mathbb{N}$. Then for any n, $A_1 \cup \cdots \cup A_n$ and $A_1 \times \cdots \times A_n$ are countable, where the last set is all n-tuples $(a_1, \dots, a_n)$, $a_i \in A_i$. Further, $\cup_{i=0}^{\infty} A_i$ is countable. If each A_i is at most countable, then $\cup_{i=0}^{\infty} A_i$ is at most countable.*

Exercise 2.1.11. *Prove Corollary 2.1.10. Hint: for $\cup_{i=0}^{\infty} A_i$, mimic the proof used to show $A \times B$ is countable.*

As the natural numbers, integers and rationals are countable, by taking each $A_i = \mathbb{N}$, $\mathbb{Z}$ or $\mathbb{Q}$ we immediately obtain

Corollary 2.1.12. *$\mathbb{N}^n$, $\mathbb{Z}^n$ and $\mathbb{Q}^n$ are countable. Hint: proceed by induction. For example write $\mathbb{Q}^{n+1}$ as $\mathbb{Q}^n \times \mathbb{Q}$.*

Exercise 2.1.13. *Prove there are countably many rationals in the interval $[0, 1]$.*

2.1.3 Algebraic Numbers

Consider a polynomial $f(x) = 0$ with rational coefficients. By multiplying by the least common multiple of the denominators, we can clear the fractions. Thus, without loss of generality it is sufficient to consider polynomials with integer coefficients.

The **algebraic numbers**, $\mathcal{A}$, are the set of all $x \in \mathbb{C}$ such that there is a polynomial of finite degree and integer coefficients (depending on x, of course!) such that $f(x) = 0$. The remaining complex numbers are the **transcendentals**.

The **algebraic numbers of degree** n, $\mathcal{A}_n$, are the set of all $x \in \mathcal{A}$ such that

1. there exists a polynomial with integer coefficients of degree n such that $f(x) = 0$
2. there is no polynomial g with integer coefficients and degree less than n with $g(x) = 0$.

Thus, $\mathcal{A}_n$ is the subset of algebraic numbers x where for each $x \in \mathcal{A}_n$, the degree of the smallest polynomial f with integer coefficients and $f(x) = 0$ is n.

Exercise 2.1.14. *Show the following are algebraic: any rational, the square-root of any rational, the cube-root of any rational,* $r^{\frac{p}{q}}$ *where* $r, p, q \in \mathbb{Q}$, $i = \sqrt{-1}$, $\sqrt{3\sqrt{2} - 5}$.

Theorem 2.1.15. *The Algebraic Numbers are countable.*

Proof: If we show each $\mathcal{A}_n$ is at most countable, then as $\mathcal{A} = \cup_{n=1}^{\infty} \mathcal{A}_n$, by Corollary 2.1.10 $\mathcal{A}$ is at most countable.

Recall the **Fundamental Theorem of Algebra (FTA):** Let $f(x)$ be a polynomial of degree n with complex coefficients. Then $f(x)$ has n (not necessarily distinct) roots. Of course, we will only need a weaker version, namely that the Fundamental Theorem of Algebra holds for polynomials with integer coefficients.

Fix an $n \in \mathbb{N}$. We now show $\mathcal{A}_n$ is at most countable. We can represent every integral polynomial $f(x) = a_n x^n + \cdots + a_0$ by an $(n+1)$-tuple $(a_0, \ldots, a_n)$. By Corollary 2.1.12, the set of all $(n+1)$-tuples with integer coefficients ($\mathbb{Z}^{n+1}$) is countable. Thus, there is a bijection from $\mathbb{N}$ to $\mathbb{Z}^{n+1}$, and we can index each $(n+1)$-tuple $a \in \mathbb{Z}^{n+1}$:

$$\{a : a \in \mathbb{Z}^{n+1}\} = \bigcup_{i=1}^{\infty} \{\alpha_i\}, \tag{2.3}$$

where each $\alpha_i \in \mathbb{Z}^{n+1}$.

For each tuple α_i (or $a \in \mathbb{Z}^{n+1}$), there are n roots. Let R_{α_i} be the roots of the integer polynomial associated to α_i. The roots in R_{α_i} need not be distinct, and the roots may solve an integer polynomial of smaller degree. For example, $f(x) = (x^2-1)^4$ is a degree 8 polynomial. It has two roots, $x = 1$ with multiplicity 4 and $x = -1$ with multiplicity 4, and each root is a root of a degree 1 polynomial.

Let $R_n = \{x \in \mathbb{C} : x \text{ is a root of a degree } n \text{ polynomial}\}$. One can show that

$$R_n \;=\; \bigcup_{i=1}^{\infty} R_{\alpha_i} \;\supset\; \mathcal{A}_n. \tag{2.4}$$

By Lemma 2.1.10, R_n is countable. Thus, by Lemma 2.1.6, as R_n is at most countable, A_n is at most countable.

Therefore, each $\mathcal{A}_n$ is at most countable, so by Corollary 2.1.10 $\mathcal{A}$ is at most countable. As $\mathcal{A}_1 \supset \mathbb{Q}$ (given $\frac{p}{q} \in \mathbb{Q}$, consider $qx - p = 0$), $\mathcal{A}_1$ is at least countable. As we've shown $\mathcal{A}_1$ is at most countable, this implies $\mathcal{A}_1$ is countable. Thus, $\mathcal{A}$ is countable. □

2.1.4 Transcendental Numbers

A set is **uncountable** if there is no bijection between it and the rationals (or the integers, or any countable set).

Theorem 2.1.16. *The set of irrationals in* $[0,1]$ *is uncountable.*

Proof: Let $I = [0,1] - \mathbb{Q} = \{x : 0 \le x \le 1 \text{ and } x \notin \mathbb{Q}\}$. Assume that I is countable (the case where I is finite is even easier).

We can write every number in I in a base two expansion, say $y = .y_1y_2y_3y_4\cdots$, $y_i \in \{0,1\}$, $y = \sum_i y_i 2^{-1}$. Certain numbers can be written two different ways. For example, $0.010011111111111\cdots = .0101$. As we are assuming I is countable, including both representations of these numbers is equivalent to taking the union of two countable sets, which by Theorem 2.1.9 is countable.

Further, we can add back all the rationals in $[0,1]$, as there are countably many rationals in $[0,1]$. Call this set S (the union of the irrationals, the alternate representation of some of the irrationals, and the rationals). As X is contained in the union of three at most countable sets (and two are countable), X is countable by Theorem 2.1.9.

There is therefore a bijection between $\mathbb{N}$ and X. We can enumerate the elements by $\{x_1, x_2, x_3, \dots\}$.

For each x_i, let $.x_{i1}x_{i2}x_{i3}\cdots x_{ii}\cdots$ be its binary expansion. We list the countable members of X:

$$\begin{aligned} x_1 &= x_{11}x_{12}x_{13}x_{14}\cdots \\ x_2 &= x_{21}x_{22}x_{23}x_{24}\cdots \\ x_3 &= x_{31}x_{32}x_{33}x_{34}\cdots \\ &\vdots \\ x_n &= x_{n1}x_{n2}x_{n3}x_{n4}\cdots x_{nn}\cdots \\ &\vdots \end{aligned} \tag{2.5}$$

We construct a real number $x \in [0,1]$ not in X. As this was supposed to be (more than a) complete list of all reals in $[0,1]$, this will contradict the assumption that I is countable.

Consider the number $z = .z_1z_2z_3\cdots z_n\cdots$ defined by $z_n = 1 - x_{nn}$. Can z be one of the numbers in our list? For example, could $z = x_m$?

No, as they differ in the m^{th} digit. Thus, z is not on our list, violating the assumption that we had a complete enumeration. Note we had to be careful and make sure we included all equivalent ways of writing the same number. Thus, while z disagrees with the base two expansion of x_m, it cannot be an equivalent way of representing x_m, as all equivalent ways of representing x_m are in our list. This is merely an annoying technical detail.

Thus, the set of irrationals in $[0,1]$ is not countable. $\Box$.

The above proof is due to Cantor ($1873 - 1874$), and is known as **Cantor's Diagonalization Argument**. Note Cantor's proof shows that *most* numbers are transcendental, though it doesn't tell us *which* numbers are transcendental. We can easily show many numbers (such as $\sqrt{3 + 2^{\frac{3}{5}}\sqrt{7}}$) are algebraic. What of other numbers, such as π and e?

Lambert (1761), Legendre (1794), Hermite (1873) and others proved π irrational; Legendre (1794) also proved π irrational. In 1882 Lindemann proved π transcental.

What about e? Euler (1737) proved that e and e^2 are irrational, Liouville (1844) proved e is not an algebraic number of degree 2, and Hermite (1873) proved e is transcendental.

Liouville (1851) showed transcendental numbers exist; we will discuss his construction later.

2.2 Properties of e

Recall

Definition 2.2.1 (algebraic and transcendental numbers). *A complex number x is algebraic if it satisfies a polynomial equation*

$$f(x) = 0 \tag{2.6}$$

for some non zero polynomial $f(X)$ with integer coefficients. A real (complex) number which is not algebraic is called transcendental.

The algebraic and transcendental numbers are complementary subsets of the complex numbers.

Exercise 2.2.2. *Show that if there is a polynomial f of degree n with rational coefficients such that $f(x) = 0$ then there is a polynomial g of degree n with integer coefficients such that $g(x) = 0$. Thus, it is sufficient to study roots of polynomials with integer coefficients.*

Exercise 2.2.3. *Show that $\sqrt{2} \notin \mathbb{Q}$.*

Examples:

1. $qX - p = 0 \Rightarrow$ "every rational number is algebraic";
2. $qX^n - p = 0 \Rightarrow$ "every root of a rational number is also algebraic";
3. $X^2 + 1 = 0 \Rightarrow i$ is algebraic.

Question 2.2.4. *How large is the subset of algebraic numbers inside the real line (complex plane)?*

Remark 2.2.5. *the set of algebraic numbers is countable, and hence the set of transcendentals is uncountable, so the algebraic number are very "sparse".*

2.2.1 e is Irrational

One of the many ways to define the number e, the base of the natural logarithm, is to write it as the sum of the following infinite series:

$$e = \sum_{n=1}^{\infty} \frac{1}{n!} \tag{2.7}$$

Now, let us denote the partial sums of the above series by

$$s_m = \sum_{n=1}^{m} \frac{1}{n!} \tag{2.8}$$

Hence e is the limit of the convergent sequence s_m. This ideea will be the main tool in analyzing the nature of e.

Theorem 2.2.6 (Euler, 1737). *The number e is irrational.*

Proof. Assume that $e \in \mathbb{Q}$. Then we can write $e = \frac{p}{q}$, where p, q are positive integers.

Now,

$$\begin{aligned} e - s_m &= \sum_{n+m+1}^{\infty} \frac{1}{n!} = \\ &= \frac{1}{(m+1)!}\left\{1 + \frac{1}{m+1} + \frac{1}{(m+1)(m+2)} + ...\right\} \\ &< \frac{1}{(m+1)!}\left\{1 + \frac{1}{m+1} + \frac{1}{(m+1)^2} + \frac{1}{(m+1)^3} + ...\right\} \\ &= \frac{1}{(m+1)!}\frac{1}{1-\frac{1}{m+1}} = \frac{1}{m!m} \end{aligned} \tag{2.9}$$

Hence we obtain

$$0 < e - s_m < \frac{1}{m!m}. \tag{2.10}$$

In particular, taking $m = q$ we get:

$$\begin{aligned} 0 \;<\; & e - s_q \;<\; \frac{1}{q!} \\ 0 \;<\; & q!e - q!s_q \;<\; \frac{1}{q} \end{aligned} \tag{2.11}$$

which is clearly impossible since the left hand side of the last equation, namely $q!e - q!s_q$, would have to be an integer between 0 and 1. This contradicts our assumption that e was rational. □

2.2.2 e is Transcendental

Theorem 2.2.7 (Hermite,1873). *The number e is transcendental.*

Proof. The proof is again by contradiction. Assume that e is algebraic. Then it must satisfy a polynomial equation

$$a_n X^n + ... + a_1 X + a_0 = 0 \tag{2.12}$$

where $a_0, a_1, .., a_n$ are integer numbers, and we can assume without loss of generality that $a_0, a_n \neq 0$.

Now consider a polynomial $f(X)$ of degree r, and associate to it the following linear combination of its derivatives:

$$F(X) = f(X) + f'(X) + ... + f^{(r)}(X) \tag{2.13}$$

Now, the polynomial $F(X)$ has the property that

$$\frac{d}{dx}\left[e^{-x}F(x)\right] = e^{-x}f(x). \tag{2.14}$$

As $F(X)$ is differentiable, applying the Mean Value Theorem to $e^{-x}F(X)$ on the interval $[0, k]$ for k any integer gives

$$e^{-k}F(k) - F(0) = -ke^{-c_k}f(c_k), \quad \text{for some} \quad c_k \in (0, k), \tag{2.15}$$

or, equivalently

$$F(k) - e^k F(0) = -ke^{k-c_k}f(c_k) =: \epsilon_k. \tag{2.16}$$

Now, if we plug in the previous equation the values $k = 0, 1, ..., n$ we get the following system of equations:

$$F(0) - F(0) = 0 =: \epsilon_0$$

$$F(1) - eF(0) = -e^{1-c_1} f(c_1) =: \epsilon_1$$

$$F(2) - e^2 F(0) = -2e^{2-c_2} f(c_2) =: \epsilon_2 \tag{2.17}$$

.....

$$F(n) - e^n F(0) = -ne^{n-c_n} f(c_n) =: \epsilon_n$$

We multiply the first equation by a_0, the second by $a_1, \ldots$, the $(n+1)^{st}$ by a_n. Adding the resulting equations gives

$$\sum_{k=0}^{n} a_k F(k) - \left(\sum_{k=0}^{n} a_k e^k \right) F(0) = \sum_{k=0}^{n} a_k \epsilon_k. \tag{2.18}$$

Notice that on the left hand side we have exactly the polynomial equation that is satisfied by e:

$$\sum_{k=0}^{n} a_k e^k = 0; \tag{2.19}$$

hence Equation 2.18 reduces to

$$\sum_{k=0}^{n} a_k F(k) = \sum_{k=0}^{n} a_k \epsilon_k. \tag{2.20}$$

So far we had complete freedom in our choice of f and its associate F, and the previous equation always hold. In what follows we choose a special polynomial f in order to reach a contradiction.

Take a large prime p, large enough such that $p > |a_0|$ and $p > n$. Let f equal

$$\begin{aligned} f(X) &= \frac{1}{(p-1)!} X^{p-1} (1-X)^p (2-X)^p \cdots (n-X)^p \\ &= \frac{1}{(p-1)!} \Big((n!)^p X^{p-1} + \text{higher order terms} \Big). \end{aligned} \tag{2.21}$$

Though it plays no role in the proof, we note that the degree of f is

$$\deg(f) := r = (n+1)p - 1. \tag{2.22}$$

In what follows we make a number of remarks which will help us finish the proof. By $p\mathbb{Z}$ we mean the set of integer multiples of p.

Remark 2.2.8. *For $i \geq p$, $f^{(i)}(j) \in p\mathbb{Z}, \forall j \in \mathbb{N}$.*

Proof: Differentiate Equation 2.21 $i \geq p$ times. The only terms which survive bring down at least a $p!$. As each term of $f(x)$ is an integer over $(p-1)!$, we see that all surviving terms are multiplied by p.

Remark 2.2.9. *For $0 \leq i < p$, $f^{(i)}(j) = 0, j = 1, 2, .., n$.*

Proof: The multiplicity of a root of a polynomial gives the order of vanishing of the polynomial at that particular root. As $j = 1, 2, \ldots, n$ are roots of multiplicity p, differentiating $f(x)$ less than p times yields a polynomial which still vanishes at these j.

Remark 2.2.10. $F(1), F(2), \ldots, F(n) \in p\mathbb{Z}$.

Proof: Recall that $F(j) = f(j) + f'(j) + .. + f^{(r)}(j)$. By the first remark, $f^{(i)}(j)$ is a multiple of p for $i \geq p$ and any integer j. By the second remark, $f^{(i)}(j) = 0$ for $0 \leq i < p$ and $j = 1, 2, \ldots, n$. Thus, $F(j)$ is a multiple of p for these j.

Remark 2.2.11. *For $0 \leq i \leq p-2$, $f^{(i)}(0) = 0$.*

Proof: Similar to the second remark, we note that $f^{(i)}(0) = 0$ for $0 \leq i < p-2$, because 0 is a root of $f(x)$ of multiplicity $p-1$.

Remark 2.2.12. *$F(0)$ is not a multiple of p.*

Proof: By the first remark, $f^{(i)}(0)$ is a multiple of p for $i \geq p$; by the fourth remark, $f^{(i)}(0) = 0$ for $0 \leq i \leq p-2$. Since

$$F(0) = f(0) + f'(0) + \cdots + f^{(p-2)}(0) + f^{(p-1)}(0) + f^{(p)}(0) + \cdots + f^{(r)}(0), \tag{2.23}$$

to prove $F(0)$ is a not multiple of p it is sufficient to prove $f^{(p-1)}(0)$ is a not multiple of p (as all other terms *are* multiples of p).

However, from the Taylor Series expansion of f in Equation 2.21, we see that

$$f^{(p-1)}(0) \;=\; (n!)^p \;+\; \Big(\text{terms that are multiples of } p\Big). \tag{2.24}$$

Since we chose $p > n$, $n!$ is not divisible by p, proving the remark.

We resume the proof of the transcendence of e.

We also chose p such that a_0 is not divisible by p. This fact plus the above remarks imply first that $\sum_k a_k F(k)$ is an integer, and second that

$$\sum_{k=0}^{n} a_k F(k) \equiv a_0 F(0) \not\equiv 0 \bmod p. \tag{2.25}$$

Thus, $\sum_k a_k F(k)$ is a non-zero integer.

Let us recall equation 2.20:

$$\sum_{k=0}^{n} a_k F(k) = a_1\epsilon_1 + \cdots + a_n\epsilon_n. \tag{2.26}$$

We have already proved that the left hand side is a non-zero integer. We analyze the sum on the right hand side. We have

$$\epsilon_k = -ke^{k-c_k} f(c_k) = \frac{-ke^{k-c_k} c_k^{p-1}(1-c_k)^p \cdots (n-c_k)^p}{(p-1)1}. \tag{2.27}$$

As $0 \le c_k \le k \le n$ we obtain

$$\begin{aligned} |\epsilon_k| &\le \frac{e^k k^p (1\cdot 2\cdots n)^p}{(p-1)!} \\ &\le \frac{e^n (n!n)^p}{(p-1)!} \to 0 \quad \text{as} \quad p \to \infty. \end{aligned} \tag{2.28}$$

Now recall that n is fixed, and so are the constants $a_0, \cdots, a_n$ (they define the polynomial equation supposedly satisfied by e), and the only thing that varies in our argument is the prime number p. Hence, by choosing a prime number p large enough, we can make sure that all ϵ_k's are uniformly small, in particular we can make them small enough such that the following holds:

$$\left| \sum_{k=1}^{n} a_k \epsilon_k \right| < 1 \tag{2.29}$$

To be more precise, we only have to choose p such that $p > n, |a_0|$ and:

$$\frac{e^n (n!n)^p}{(p-1)!} < \frac{1}{\sum_{k=0}^{n} |a_k|} \tag{2.30}$$

In this way we reach a contradiction in the identity 2.20 where the left hand side is a non-zero integer, while the right hand side is a real number of absolute value < 1. □

Exercise 2.2.13. *In the above proof, we assumed $a_0, a_n \neq 0$. Prove that if a number is algebraic, one can always find a polynomial such that the leading term and the constant term are both non-zero.*

Exercise 2.2.14. *For fixed n, prove that as $p \to \infty$, $\frac{(n!n)^p}{(p-1)!} \to 0$. Hint: Let $C = n!n$. Choose $p > 2(2C)^4$. Then $(p-1)! > (p-1)(p-2)\cdots(p-\frac{p}{2}) \approx (\frac{p}{2})^{\frac{p}{2}}$. Substitute and compare.*

Chapter 3

Introduction to Number Theory

3.1 Dirichlet's Box Principle

Definition 3.1.1 (Dirichlet's Box Principle / Pidgeon Hole Principle). *Consider n boxes, and place $n+1$ objects in the n boxes. Then some box contains at least two objects.*

We will use Dirichlet's Box Principle to find good rational approximations to irrational numbers.

3.1.1 Approximation by Rationals

Let $\alpha \in \mathbb{R} - \mathbb{Q}$ be an irrational number. We are looking for a rational number $\frac{p}{q}$ such that $\left|\alpha - \frac{p}{q}\right|$ is small, so that $\frac{p}{q}$ is a good rational approximation to α.

Lemma 3.1.2. *Let $\alpha \in \mathbb{R} - \mathbb{Q}$. Then there exist $p, q \in \mathbb{Z}, q \neq 0$ such that:*

$$\left|\alpha - \frac{p}{q}\right| \leq \frac{1}{q} \tag{3.1}$$

Proof. It is enough to prove this for $\alpha \in (0,1)$. Let $q \geq 1$ and divide the interval $[0,1)$ into q intervals $[\frac{p}{q}, \frac{p+1}{q})$ of length $\frac{1}{q}$. Then α belongs to one of these intervals. For some $0 < p < q$ we then have:

$$\alpha \in \left[\frac{p}{q}, \frac{p+1}{q}\right) \Rightarrow \left|\alpha - \frac{p}{q}\right| \leq \frac{1}{q}. \tag{3.2}$$

To obtain a better approximation, we start with an irrational number $\alpha \in (0,1)$ and an integer parameter $Q > 1$. As before, divide the interval $(0,1)$ into Q equal pieces $(\frac{a}{Q}, \frac{a+1}{Q})$. Consider the $Q+1$ numbers inside the interval $(0,1)$:

$$\{\alpha\}, \{2\alpha\}, ..., \{(Q+1)\alpha\}, \tag{3.3}$$

where $\{x\}$ denotes the fractional part of x. Letting $[x]$ denote the greatest integer less than or equal to x, we have $x = [x] + \{x\}$.

By Dirichlet's Box Principle, at least two of these numbers, say $\{q_1\alpha\}$ and $\{q_2\alpha\}$, belong to a common interval of length $\frac{1}{Q}$. Without loss of generality, we may take $1 \leq q_1 < q_2 \leq Q+1$.

Hence

$$\Big|\{q_2\alpha\} - \{q_1\alpha\}\Big| \leq \frac{1}{Q} \tag{3.4}$$

and

$$|(q_2\alpha - n_2) - (q_1\alpha - n_1)| \leq \frac{1}{Q}, \quad n_i = [q_i\alpha]. \tag{3.5}$$

Now let $q = q_1 - q_2$, $1 \leq q \leq Q$ and $p = n_1 - n_2 \in \mathbb{Z}$. Then

$$\Big|q\alpha - p\Big| \leq \frac{1}{Q} \tag{3.6}$$

and hence

$$\Big|\alpha - \frac{p}{q}\Big| \leq \frac{1}{qQ} \leq \frac{1}{q^2}. \tag{3.7}$$

We have proven

Theorem 3.1.3. *Given $\alpha \in \mathbb{R}$, there exist $p, q \in \mathbb{Z}, q \neq 0$, such that*

$$\Big|\alpha - \frac{p}{q}\Big| < \frac{1}{q^2}. \tag{3.8}$$

3.2 Counting the Number of Primes

3.2.1 Euclid

Lemma 3.2.1 (Euclid). *There are infinitely many primes.*

Proof by contradiction: Assume there are only finitely many primes, say $p_1, p_2, \ldots, p_n$. Consider

$$x = p_1 p_2 .. p_n + 1. \tag{3.9}$$

x cannot be prime, as we are assuming p_1 through p_n is a complete list of primes. Thus, x is composite, and divisible by a prime. However, p_i cannot divide x, as it gives a remainder of 1. Thus, x would have to be divisible by some prime not in our list, again contradicting the assumption that p_1 through p_n is a complete enumeration of the primes. □

Exercise 3.2.2. *Try, using Euclid's argument, to find an explicit lower bound (as weak as you like) to the function:*

$$\pi(X) = \#\{p : p \text{ is prime and } p \leq X\}. \tag{3.10}$$

3.2.2 Dirichlet's Theorem

Theorem 3.2.3 (Primes in Arithmetic Progressions). *Let a and b be relatively prime integers. Then there are infinitely many primes in the progression $an + b$. Further, for a fixed a, to first order all relatively prime b give progressions having the same number of primes.*

Notice that the condition $(a, b) = 1$ is necessary. If $\gcd(a, b) > 1$, $an + b$ can never be prime. Dirichlet's remarkable result is that this condition is also sufficient.

Exercise 3.2.4. *Dirichlet's theorem is not easy to prove, but try to prove it in the particular case $a = 4, b = -1$, i.e. for the arithmetic progression $4n - 1$, using an argument similar to Euclid's. Proving there are infinitely many primes of the form $4n + 1$ is a lot harder.*

3.2.3 Prime Number Theorem

Theorem 3.2.5. *(Prime Number Theorem or PNT) As $X \to \infty$,*

$$\pi(X) \sim \frac{X}{\log X} \tag{3.11}$$

The Prime Number Theorem was proved in 1896 by Jacques Hadamard and Charles Jean Gustave Nicolas Baron de la Vallee Poussin. Of course, we need to quantify what $\pi(X) \sim \frac{X}{\log X}$ means. Basically, there is an error function $E(X)$ such that $|\pi(X) - \frac{X}{\log X}| \leq E(X)$, and $E(X)$ grows slower than $\frac{X}{\log X}$.

A weaker version was proved by Pafnuty Chebyshev (around 1850).

Theorem 3.2.6 (Chebyshev). *There exist explicit positive constants A and B such that, for $n > 30$:*

$$\frac{AX}{\log X} \leq \frac{\pi(X)}{X} \leq \frac{BX}{\log X}. \tag{3.12}$$

Chebyshev showed one can take $A = \log\left(\frac{2^{\frac{1}{2}}3^{\frac{1}{3}}4^{\frac{1}{4}}}{30^{\frac{1}{30}}}\right) \approx .921$ and $B = \frac{6A}{5} \approx 1.105$, which are indeed very close to 1. To highlight the method, we will use cruder arguments and prove the theorem for a smaller A and a larger B.

Chebyshev's argument uses an identity using von Mangoldt's Lambda function $\Lambda(n)$, where $\Lambda(n) = \log p$ if $m = p^k$ for some prime p, and 0 otherwise.

Define the function

$$T(X) = \sum_{1 \leq n \leq X} \Lambda(n) \left[\frac{X}{n}\right] = \sum_{n \geq 1} \Lambda(n) \left[\frac{X}{n}\right]. \tag{3.13}$$

Exercise 3.2.7. *Show that $T(X) = \sum_{n \leq X} \log n$.*

Now, it is easy to see (compare upper and lower sums) that

$$\sum_{n \leq X} \log n = \int_1^X \log \mathrm{t}\ \mathrm{dt} + O(\log X) = X \log X - X + O(\log X), \tag{3.14}$$

giving a good approximation to the function $T(X)$. The trick is to look at

$$T(X) - 2T\Big(\frac{X}{2}\Big) = \sum_n \Lambda(n) \left(\left[\frac{X}{n}\right] - 2\left[\frac{X}{2n}\right] \right) \tag{3.15}$$

By the previous remarks, the LHS $= X \log 2 + O(\log X)$. Also,

$$\text{RHS} \leq \sum_{p \leq X} (\log p) \frac{\log X}{\log p} = \pi(X) \log X. \tag{3.16}$$

Hence we immediately obtain the lower bound:

$$\pi(X) \geq \frac{X \log 2}{\log X} + O(\log X) \tag{3.17}$$

Exercise 3.2.8. *Prove the bound in Equation 3.16.*

To obtain an upper bound for $\pi(X)$, we notice that, since $[2\alpha] \geq 2[\alpha]$, the sum in equation (3.15) has only positive terms. By dropping terms we get a lower bound.

$$\begin{aligned} T(X) - 2T\Big(\frac{X}{2}\Big) &\geq \sum_{X/2<n\leq X} \Lambda(n)\left(\left[\frac{X}{n}\right] - 2\left[\frac{X}{2n}\right]\right) \\ &\geq \sum_{X/2<p\leq X} \log p \\ &\geq \log\Big(\frac{X}{2}\Big) \sum_{X/2<p\leq X} 1 \\ &= \log\Big(\frac{X}{2}\Big)\left(\pi(X) - \pi\Big(\frac{X}{2}\Big)\right) \end{aligned} \tag{3.18}$$

Hence we obtain an upper bound for the number of primes between $\frac{X}{2}$ and X:

$$\pi(X) - \pi(X/2) \leq \frac{X \log 2}{\log(\frac{X}{2})} + O(1) \tag{3.19}$$

Now, if we write inequality (3.19) for $X, \frac{X}{2}, \frac{X}{2^2}, \ldots$ we get

$$\begin{aligned} \pi(X) - \pi(X/2) &\leq 2\frac{X/2}{\log(X/2)} \\ \pi(X/2) - \pi(X/2^2) &\leq 2\frac{X/2^2}{\log(X/2^2)} \\ &\vdots \\ \pi(X/2^{k-1}) - \pi(X/2^k) &\leq 2\frac{X/2^k}{\log(X/2^k)} \end{aligned} \tag{3.20}$$

as long as $\frac{X}{2^k} \geq 1$, i.e. $k \leq [\log_2 X] = k_0$. Summing the above inequalities we get on the left hand side a telescoping sum. All the terms cancel, except for the leading term $\pi(X)$ and $\pi(X/2^{k_0}) = 0$.

Thus

$$\pi(X) \leq 2 \sum_{k=1}^{k_0} \frac{X/2^k}{\log(X/2^k)} \tag{3.21}$$

To evaluate the sum in the above inequality we split it into two parts, k "small" and k "large". More precisely, let $n_0 = \log_2(X^{1/10})$ so that $2^{n_0} = X^{1/10}$ and note that:

$$2 \sum_{k>n_0} \frac{X/2^k}{\log(X/2^k)} \leq 2 \sum_{k>n_0} \frac{X}{2^k} \leq \frac{2X}{2^{n_0}} = \frac{2X}{X^{1/10}} = 2X^{9/10}. \tag{3.22}$$

Hence the contribution from k "large" is very small compared to what we expect (i.e. order of magnitude $\frac{X}{\log X}$), or we can say that the main term comes from the sum over k small.

We now evaluate the contribution from small k.

$$2 \sum_{k=1}^{n_0} \frac{X}{2^k} \frac{1}{\log(X/2^k)} \leq \frac{2X}{\log(X/2^{n_0})} \sum_{k=1}^{n_0} \frac{1}{2^{n_0}} \leq \frac{2X}{\log(X^{9/10})} = \frac{20}{9} \frac{X}{\log X} \tag{3.23}$$

Hence the right hand side of the equation (3.21) is made up of two parts, a main term of size $\frac{BX}{\log X}$ coming from equation (3.23), and a lower order term coming from equation (3.22).

For X sufficiently large,

$$\pi(X) \leq \frac{BX}{\log X} \tag{3.24}$$

where B can be any constant strictly bigger than $\frac{20}{9}$.

To obtain Chebyshev's better constant we would have to work a little harder along these lines, but it is the same method.

Gathering equations (3.17) and (3.24) we see we have proven

$$\frac{AX}{\log X} \leq \pi(X) \leq \frac{BX}{\log X}. \tag{3.25}$$

While this is not an asymptotic for $\pi(X)$, it does give the right order of magnitude for $\pi(X)$, namely $\frac{X}{\log X}$.

Exercise 3.2.9. *Using Chebyshev's Theorem, Prove Bertrand's Postulate: for any integer $n \geq 1$, there is always a prime number between n and $2n$.*

3.3 Partial Summation

Lemma 3.3.1 (Partial Summation: Discrete Version). *Let $A_N = \sum_{n=1}^{N} a_n$. then*

$$\sum_{n=M}^{N} a_n b_n = A_N b_N - A_{M-1} b_M + \sum_{n=M}^{N-1} A_n (b_n - b_{n+1}) \tag{3.26}$$

Proof. Since $A_n - A_{n_1} = a_n$,

$$\begin{aligned} \sum_{n=M}^{N} a_n b_n &= \sum_{n=M}^{N} (A_n - A_{n-1}) b_n \\ &= (A_N - A_{N-1}) b_N + (A_{N-1} - A_{N-2}) b_{N-1} + \cdots + (A_M - A_{M-1}) b_M \\ &= A_N b_N + (-A_{N-1} b_N + A_{N-1} b_{N-1}) + \cdots + (-A_M b_{M+1} + A_M b_M) - a_{M-1} b_M \\ &= A_N b_N - a_{M_1} b_M + \sum_{n=M}^{N-1} A_n (b_n - b_{n+1}). \end{aligned} \tag{3.27}$$

□

Lemma 3.3.2 (Abel's Summation Formula - Integral Version). *Let $h(x)$ be a continuously differentiable function. Let $A(x) = \sum_{n \leq x} a_n$. Then*

$$\sum_{n \leq x} a_n h(n) = A(x) h(x) - \int_1^x A(u) h'(u) du \tag{3.28}$$

See, for example, [Ru], page 70.

Partial Summation allows us to take knowledge of one quantity and convert that to knowledge of another.

For example, suppose we know that

$$\sum_{p\le x}\log p \;=\; x + O(x^{\frac{1}{2}+\epsilon}). \tag{3.29}$$

We use this to glean information about $\sum_{p\le x} 1$.

Define

$$h(n) = \frac{1}{\log n} \quad \text{and} \quad a_n = \begin{cases} \log n & \text{if } n \text{ is prime} \\ 0 & \text{otherwise.} \end{cases} \tag{3.30}$$

Applying partial summation to $\sum_{p\le x} a_n h(n)$ will give us knowledge about $\sum_{p\le x} 1$. Note as long as $h(n) = \frac{1}{\log n}$ for n prime, it doesn't matter how we define $h(n)$ elsewhere; however, to use the integral version of Partial Summation, we need h to be a differentiable function.

Thus

$$\begin{aligned} \sum_{p\le x} 1 &= \sum_{p\le x} a_n h(n) \\ &= \Big(x + O(x^{\frac{1}{2}+\epsilon})\Big)\frac{1}{\log x} - \int_2^x \Big(u + O(u^{\frac{1}{2}+\epsilon})\Big)h'(u)du. \end{aligned} \tag{3.31}$$

The main term ($A(x)h(x)$) equals $\frac{x}{\log x}$ plus a significantly smaller error.

We now calculate the integral, noting $h'(u) = -\frac{1}{u\log^2 u}$. The error piece in the integral gives a constant multiple of

$$\int_2^x \frac{u^{\frac{1}{2}+\epsilon}}{u\log^2 u}du. \tag{3.32}$$

As $\frac{1}{\log^2 u} \le \frac{1}{\log^2 2}$ for $2 \le u \le x$, the integral is bounded by

$$\frac{1}{\log^2 2}\int_2^x u^{-\frac{1}{2}+\epsilon} < \frac{1}{\log^2 2}\frac{1}{\frac{1}{2}+\epsilon}x^{\frac{1}{2}+\epsilon}, \tag{3.33}$$

which is significantly less than $A(x)h(x) = \frac{x}{\log x}$.

We now need to handle the other integral:

$$\int_2^x \frac{u}{u\log^2 u}du \;=\; \int_2^x \frac{1}{\log^2 u}du. \tag{3.34}$$

The obvious approximation to try is $\frac{1}{\log^2 u} \leq \frac{1}{\log^2 2}$. Unfortunately, plugging this in bounds the integral by $\frac{x}{\log^2 2}$. This is larger than the expected main term, $A(x)h(x)$!

As a rule of thumb, whenever you are trying to bound something, try the simplest, most trivial bounds first. Only if they fail should you try to be clever.

Here, we need to be clever, as we are bounding the integral by something larger than the observed terms.

We split the integral into two pieces:

$$\int_2^x = \int_2^{\sqrt{x}} + \int_{\sqrt{x}}^x \tag{3.35}$$

For the first piece, we use the trivial bound for $\frac{1}{\log^2 u}$. Note the interval has length $\sqrt{x} - 2 < \sqrt{x}$. Thus, the first piece contributes at most $\frac{x^{\frac{1}{2}}}{\log^2 2}$, significantly less than $A(x)h(x)$.

The reason trivial bounds failed for the entire integral is the length was too large (of size x); there wasn't enough decay in the function.

The advantage of splitting the integral in two is that in the second piece, even though most of the length of the original interval is here (it is of length $x - \sqrt{x} \approx x$), the function $\frac{1}{\log^2 u}$ is small here. Instead of bounding it by a constant, we now bound it by substituting in the smallest value of u on this interval, $\sqrt{x}$. Thus, the contribution from this integral is at most $\frac{x-\sqrt{x}}{\log^2 \sqrt{x}} < \frac{4x}{\log^2 x}$. Note that this is significantly less than the main term $A(x)h(x) = \frac{x}{\log x}$.

Chapter 4

Fourier Analysis and the Equi-Distribution of $\{n\alpha\}$

4.1 Inner Product of Functions

We define the exponential function by means of the series

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}, \tag{4.1}$$

which converges everywhere. Given the Taylor series expansion of $\sin x$ and $\cos x$, we can verify the identity

$$e^{ix} = \cos x + i \sin x. \tag{4.2}$$

Exercise 4.1.1. *Prove e^x converges for all $x \in \mathbb{R}$ (even better, for all $x \in \mathbb{C}$. Show the series for e^x also equals*

$$\lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n, \tag{4.3}$$

which you may remember from compound interest problems.

Exercise 4.1.2. *Prove, using the series definition, that $e^{x+y} = e^x e^y$. Use this fact to calculate the derivative of e^x. If instead you try to differentiate the series directly, you must justify the derivative of the infinite sum is the infinite sum of the derivatives.*

Remember the definition of **inner or dot product**: for two vectors $\vec{v} = (v_1, \cdots, v_n)$, $\vec{w} = (w_1, \cdots, w_n)$, we take the *inner product* $\vec{v} \cdot \vec{w}$ (also denoted $\langle v, w \rangle$) to mean

$$\vec{v} \cdot \vec{w} \;=\; \langle v, w \rangle = \sum_i v_i \bar{w}_i. \tag{4.4}$$

Further, the length of a vector v is

$$|v| \;=\; \langle v, v \rangle. \tag{4.5}$$

We generalize this for functions. For definiteness, assume f and g are functions from $[0, 1]$ to $\mathbb{C}$. Divide the interval $[0, 1]$ into n equal pieces. Then we can represent the functions by

$$f(x) \;\longleftrightarrow\; \left(f(0), f\Big(\frac{1}{n}\Big), \ldots, f\Big(\frac{n-1}{n}\Big) \right), \tag{4.6}$$

and similarly for g. Call these vectors f_n and g_n. As before, we consider

$$\langle f_n, g_n \rangle \;=\; \sum_{i=0}^{n-1} f\Big(\frac{i}{n}\Big) \cdot \bar{g}\Big(\frac{i}{n}\Big). \tag{4.7}$$

In general, as we continue to divide the interval ($n \to \infty$), the above sum diverges. For example, if f and g are identically 1, the above sum is n.

There is a natural rescaling: we multiply each term in the sum by $\frac{1}{n}$, the size of the sub-interval. Note for the constant function, the sum is now independent of n.

Thus, for good f and g we are led to

$$\langle f, g \rangle \;=\; \lim_{n\to\infty} \sum_{i=0}^{n-1} f\Big(\frac{i}{n}\Big) \cdot \bar{g}\Big(\frac{i}{n}\Big) \frac{1}{n} = \int_0^1 f(x)\overline{g(x)} dx. \tag{4.8}$$

The last result follows by Riemann Integration.

Definition 4.1.3. *We say two continuous functions on $[0, 1]$ are orthogonal (or perpendicular) if their dot product equals zero.*

Exercise 4.1.4. *Prove x^n and x^m are not perpendicular on $[0, 1]$ for $n \neq m$.*

We will see that the exponential function behaves very nicely under the inner product. Define

$$e_n(x) = e^{2\pi inx} \text{ for } n \in \mathbb{Z}. \tag{4.9}$$

Then a straightforward calculation shows

$$\langle e_n(x), e_m(x)\rangle = \begin{cases} 1 & \text{if } n = m \\ 0 & \text{otherwise.} \end{cases} \tag{4.10}$$

Thus $e_0(x), e_1(x), e_2(x), \cdots$ are an **orthogonal set** of functions, which means they are pairwise perpendicular. As each function has length 1, we say the functions $e_n(x)$ are an **orthonormal set** of functions.

Exercise 4.1.5. *Prove $\langle e_n(x), e_m(x)\rangle$ is 1 if $n = m$ and 0 otherwise.*

4.2 Fourier Series and $\{n\alpha\}$

4.2.1 Fourier Series

Let f be continuous and periodic on $\mathbb{R}$ with period one. Define the n**th Fourier coefficient** $\hat{f}(n)$ of f to be

$$\hat{f}(n) = a_n = \langle f(x), e_n(x)\rangle = \int_0^1 f(x)e^{-2\pi inx}dx. \tag{4.11}$$

Returning to the intuition of $\mathbb{R}^m$, we can think of the $e_n(x)$'s as an infinite set of perpendicular directions. The above is simply the projection of f in the direction of $e_n(x)$.

Exercise 4.2.1. *Show*

$$\langle f(x) - \hat{f}(n)e_n(x), e_n(x)\rangle = 0. \tag{4.12}$$

This agrees with our intuition, namely, that if you remove the projection in a certain direction, what is left is perpendicular to that direction.

The N^{th} **partial Fourier series** of f is

$$s_N(x) = \sum_{n=-N}^{N} \hat{f}(n)e_n(x). \tag{4.13}$$

Exercise 4.2.2. *Prove*

1. $\langle f(x) - s_N(x), e_n(x)\rangle = 0$ *if* $|n| \leq N$.
2. $|\hat{f}(n)| \leq \int_0^1 |f(x)|dx$.
3. *If* $\langle f, f\rangle < \infty$, *then* $\sum_{n=-\infty}^{\infty} |\hat{f}(n)|^2 \leq \langle f, f\rangle$.
4. *If* $\langle f, f\rangle < \infty$, *then* $\lim_{|n|\to\infty} \hat{f}(n) = 0$.

As $\langle f(x) - s_N(x), e_n(x)\rangle = 0$ if $|n| \leq N$, we might think that we just have to let N go to infinity to obtain a series s_∞ such that

$$\langle f(x) - s_\infty(x), e_n(x)\rangle = 0. \tag{4.14}$$

Assume that for a periodic function $g(x)$ to be orthogonal to $e_n(x)$ for every n it must be zero for every x. Then $f(x) - s_\infty(x) = 0$, and hence $f = s_\infty$. Voilá – an expression for f as a sum of exponentials! Be careful, however. We have just glossed over the two central issues – completeness and, even worse, convergence. We will now see a way of avoiding some of our problems.

4.2.2 Weighted partial sums

Define

$$\begin{aligned} D_N(x) &= \sum_{n=-N}^{N} e_n(x) = \frac{\sin((2N+1)\pi x)}{\sin \pi x}, \\ F_N(x) &= \frac{\sin^2(N\pi x)}{N \sin^2 \pi x} = \frac{1}{N} \sum_{n=0}^{N-1} D_n(x). \end{aligned} \tag{4.15}$$

Here F stands for Féjer, D for Dirichlet. In general, functions which we are interested in taking their inner product against f are called **kernels**; thus, the Dirichlet kernel, the Féjer kernel, etc.

Note that, no matter what N is, $F_N(x)$ is positive for all x.

We say that a sequence $f_1(x), f_2(x), f_3(x), \ldots$ of functions is an **approximation to the identity** if

1. $f_N(x) \geq 0$ for all x and every N;
2. $\int_0^1 f_N(x)dx = 1$;

3. $\lim_{N\to\infty} \int_{\delta}^{1-\delta} f_N(x)dx = 0$ if $0 < \delta < \frac{1}{2}$.

Theorem 4.2.3. *The Féjer kernels $F_1(x), F_2(x), F_3(x), \ldots$ are an approximation to the identity.*

Proof: The first property is immediate. The second follows from the observation that $F_N(x)$ can be written as

$$F_N(x) = e_0(x) + \frac{N-1}{N}\Big(e_{-1}(x) + e_1(x)\Big) + \cdots, \tag{4.16}$$

and all integrals are zero but the first, which is 1.
To prove the third property, note that $F_N(x) \leq \frac{1}{N\sin^2 \pi\delta}$ for $\delta \leq x \leq 1-\delta$. □

Let f be a continuous, periodic function on $\mathbb{R}$ with period one. Thus, we can consider f as a function on just $[0,1]$, with $f(0) = f(1)$. Define

$$T_N(x) = \int_0^1 f(y)F_N(x-y)dy. \tag{4.17}$$

Recall the following definition and theorem:

Definition 4.2.4 (Uniform Continuity). *A continuous function is uniformly continuous if given an $\epsilon > 0$, there exists a $\delta > 0$ such that $|x-y| < \delta$ implies $|f(x) - f(y)| < \epsilon$. Note that the same δ works for all points.*

Theorem 4.2.5. *Any continuous function on a closed, compact interval is uniformly continuous.*

Exercise 4.2.6. *Show x^n is uniformly continuous on $[a,b]$ for $-\infty < a < b < \infty$.*

Theorem 4.2.7. *Given $\epsilon > 0$, there is an N such that*

$$|f(x) - T_N(x)| \leq \epsilon \tag{4.18}$$

for every $x \in [0,1]$.

Proof. For any positive N,

$$
\begin{aligned}
T_N(x) - f(x) &= \int_0^1 f(x-y)F_N(y)dy - f(x)\cdot 1 \\
&= \int_0^1 f(x-y)F_N(y)dy - \int_0^1 f(x)F_n(y)dy \text{ (property 2 of } F_N) \\
&= \int_0^\delta \Big(f(x-y) - f(x)\Big)F_N(y)dy \\
&+ \int_\delta^{1-\delta} \Big(f(x-y) - f(x)\Big)F_N(y)dy \\
&+ \int_{1-\delta}^1 \Big(f(x-y) - f(x)\Big)F_N(y)dy.
\end{aligned} \tag{4.19}
$$

Let $\delta \in (0, 1/2)$. Then, using the fact that the $F_N(x)$'s are an approximation to the identity, we find

$$
\left|\int_\delta^{1-\delta} \Big(f(x-y) - f(x)\Big)F_N(y)dy\right| \leq 2\max|f(x)| \cdot \int_\delta^{1-\delta} F_N(y)dy. \tag{4.20}
$$

Since

$$
\lim_{N\to\infty} \int_\delta^{1-\delta} F_N(y)dy = 0, \tag{4.21}
$$

we obtain

$$
\lim_{N\to\infty} \int_\delta^{1-\delta} (f(x-y) - f(x))F_N(y)dy = 0. \tag{4.22}
$$

Thus, by choosing N large enough (where large depends on δ), we can insure that this piece is at most $\frac{\epsilon}{3}$.

It remains to estimate what happens near zero. Since f is continuous and $[0, 1]$ is compact, f is uniformly continuous. Thus, we can choose δ small enough that $|f(x-y) - f(x)| < \frac{\epsilon}{3}$ for any x and any positive $y < \delta$. Then

$$
\left|\int_0^\delta \Big(f(x-y) - f(x)\Big)F_N(y)dy\right| \leq \int_0^\delta \frac{\epsilon}{3}F_N(y)dy \leq \frac{\epsilon}{3}\int_0^1 F_N(y)dy \leq \frac{\epsilon}{3}. \tag{4.23}
$$

Similarly

$$\left| \int_{1-\delta}^{1} \Big(f(x-y) - f(x) \Big) F_N(y) dy \right| \leq \frac{\epsilon}{3}. \tag{4.24}$$

Therefore

$$|T_N(x) - f(x)| \leq \epsilon \tag{4.25}$$

for all N sufficiently large. □

Definition 4.2.8 (Trigonometric Polynomials). *Any finite linear combination of the functions $e_n(x)$ is called a trigonometric polynomial.*

From Theorem 4.2.7 we immediately get the Stone-Weierstrass theorem:

Theorem 4.2.9 (Stone-Weierstrass). *Any continuous period function can be uniformly approximated by trigonometric polynomials.*

4.2.3 Equidistribution

We say that a sequence $\{x_n\}$, $x_n \in [0,1]$ is *equidistributed* if

$$\lim_{N\to\infty} \frac{1}{2N+1} \#\{n : |n| \leq N, x_n \in (a,b)\} = b - a \tag{4.26}$$

for all $(a,b) \subset [0,1]$.

Theorem 4.2.10 (Weyl). *Let α be an irrational number in $[0,1]$. Let $x_n = \{n\alpha\}$, where $\{y\}$ denotes the fractional part of y. Then the sequence $\{x_n\}$ is equidistributed.*

Proof. We will estimate $\frac{1}{2N+1} \sum_{-N}^{N} \chi_{(a,b)}(x_n)$ as $N \to \infty$, where $\chi_{(a,b)}$ is the function taking the value 0 outside (a,b) and 1 inside (a,b). We call $\chi_{(a,b)}$ the **characteristic function** of the interval (a,b).

Thus, we must show

$$\lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} \chi_{(a,b)}(x_n) = b - a. \tag{4.27}$$

Consider $e_k(x) = e^{2\pi i k x}$. Since $x_n = \{n\alpha\} = n\alpha - [n\alpha]$ and $e_k(x) = e_k(x+m)$ for every integer m,

$$e_k(x_n) = e^{2\pi i k n \alpha}. \tag{4.28}$$

Hence

$$\begin{aligned}
\frac{1}{2N+1}\sum_{n=-N}^{N} e_k(x_n) &= \frac{1}{2N+1}\sum_{n=-N}^{N} e_k(n\alpha) \\
&= \frac{1}{2N+1}\sum_{n=-N}^{N} (e^{2\pi i k \alpha})^n \\
&= \begin{cases} 1 & \text{if } k = 0 \\ \frac{1}{2N+1}\frac{e_k(-N\alpha)-e_k((N+1)\alpha)}{1-e_k(\alpha)} & \text{if } k > 0. \end{cases}
\end{aligned} \tag{4.29}$$

Now for a fixed irrational α, $|1 - e_k(\alpha)| > 0$. Therefore if $k \neq 0$:

$$\lim_{N\to\infty} \frac{1}{2N+1}\frac{e_k(-N\alpha) - e_k((N+1)\alpha)}{1 - e_k(\alpha)} = 0. \tag{4.30}$$

Let $P(x) = \sum_k a_k e_k(x)$ be a finite sum (ie, $P(x)$ is a trigonometric polynomial). By possibly adding some zero coefficients, we can write $P(x)$ as a sum over a symmetric range: $P(x) = \sum_{k=-K}^{K} a_k a_k(x)$.

Exercise 4.2.11. *Show* $\int_0^1 P(x)dx = a_0$.

By the above arguments, we have shown that for any (finite) trigonometric polynomial $P(x)$:

$$\lim_{N\to\infty} \frac{1}{2N+1}\sum_{n=-N}^{N} P(x_n) \;\to\; a_0 \;=\; \int_0^1 P(x)dx. \tag{4.31}$$

Consider two approximations to the characteristic function $\chi_{(a,b)}$:

1. f_{1m}: $f_{1m}(x) = 1$ if $a + \frac{1}{m} \leq x \leq b - \frac{1}{m}$, drops linearly to 0 at a and b, and is zero elsewhere.

2. f_{2m}: $f_{1m}(x) = 1$ if $a \leq x \leq b$, drops linearly to 0 at $a - \frac{1}{m}$ and $b + \frac{1}{m}$, and is zero elsewhere.

Note there are trivial modifications if $a = 0$ or $b = 1$. Clearly

$$f_{1m}(x) \;\leq\; \chi_{(a,b)}(x) \;\leq\; f_{2m}(x). \tag{4.32}$$

Therefore

$$\frac{1}{2N+1}\sum_{n=-N}^{N} f_{1m}(x_n) \;\leq\; \frac{1}{2N+1}\sum_{n=-N}^{N} \chi_{(a,b)}(x_n) \;\leq\; \frac{1}{2N+1}\sum_{n=-N}^{N} f_{2m}(x_n). \tag{4.33}$$

By Theorem 4.2.7, for each m, given $\epsilon > 0$ we can find trigonometric polynomials $P_{1m}(x)$ and $P_{2m}(x)$ such that $|P_{1m}(x) - f_{1m}(x)| < \epsilon$ and $|P_{2m}(x) - f_{2m}(x)| < \epsilon$.

As f_{1m} and f_{2m} are continuous functions, we can replace

$$\frac{1}{2N+1}\sum_{n=-N}^{N} f_{im}(x_n) \;\text{ with }\; \frac{1}{2N+1}\sum_{n=-N}^{N} P_{im}(x_n) \tag{4.34}$$

at a cost of at most ϵ.

As $N \to \infty$,

$$\frac{1}{2N+1}\sum_{n=-N}^{N} P_{im}(x_n) \;\to\; \int_0^1 P_{im}(x)dx. \tag{4.35}$$

But $\int_0^1 P_{1m}(x)dx = (b-a) - \frac{1}{m}$ and $\int_0^1 P_{2m}(x)dx = (b-a) + \frac{1}{m}$. Therefore, given m and ϵ, we can choose N large enough so that

$$(b-a) - \frac{1}{m} - \epsilon \;\leq\; \frac{1}{2N+1}\sum_{n=-N}^{N} \chi_{(a,b)}(x_n) \;\leq\; (b-a) + \frac{1}{m} + \epsilon. \tag{4.36}$$

Letting m tend to ∞ and ϵ tend to 0, we see $\frac{1}{2N+1}\sum_{n=-N}^{N} \chi_{(a,b)}(x_n) \to b - a$. □

Exercise 4.2.12. *Rigorously do the necessary book-keeping to prove the previous theorem.*

Exercise 4.2.13. *Prove*

1. *If $\alpha \in \mathbb{Q}$, then $\{n\alpha\}$ is periodic.*
2. *If $\alpha \notin \mathbb{Q}$, then no two $\{n\alpha\}$ are equal.*

Chapter 5

Introduction to Continued Fractions

5.1 Introduction

5.1.1 Example

Consider the equation $x^2 + y^2 = 2$. If we try to solve for $x, y \in \mathbb{C}$, we quickly find that, not only are there infinitely many solutions, but given x we can easily determine y. Namely, $y = \sqrt{2 - x^2}$.

Instead of looking for complex solutions, we could restrict x and y to be in $\mathbb{R}$. If $|x| \leq \sqrt{2}$, the same argument works.

If we restrict x and y to be integers, we find there are only four solutions: $(1, 1)$, $(1, -1)$, $(-1, 1)$ and $(-1, -1)$. Once we add restrictions (such as $x, y \in \mathbb{N}$ or $x, y \in \mathbb{Z}$) we have a **Diophantine equation**.

Explicitly, a Diophantine equation is an equation with integer coefficients, where the solutions are restricted to being integers or rationals. These equations are named in honor of the Greek Mathematician Diophantus (approximately 200 to 280 A.D.), who studied equations of this form.

Returning to our example, if instead we allow $x, y \in Q$, how many solutions are there? Can we still parametrize them as easily as when x, y were in $\mathbb{R}$ or $\mathbb{C}$?

We know one rational solution, $(x, y) = (1, 1)$. In turns out that, for quadratic equations like this, knowing one rational solution is enough to find all rational solutions.

The equation $x^2 + y^2 = 2$ is a circle centered at the origin with radius $\sqrt{2}$. Consider the straight line passing through $(1, 1)$ (which is on this circle) with rational slope t.

Exercise 5.1.1. *Prove that the other point of intersection of the line with the circle is also a rational solution. Further, show all rational solutions are obtained in this way.*

5.1.2 Goal of the Course

The main result we shall prove is

Theorem 5.1.2 (Roth's Theorem). *Let α be a real algebraic number (a root of a polynomial equation with integer coefficients). Then, given any $\epsilon > 0$, there are only finitely many relatively prime pairs of integers (p, q) such that*

$$\left|\alpha - \frac{p}{q}\right| < \frac{1}{q^{2+\epsilon}}; \tag{5.1}$$

however, there are infinitely many pairs of relatively prime integers such that

$$\left|\alpha - \frac{p}{q}\right| < \frac{1}{q^2}. \tag{5.2}$$

This should be reminiscent of p-series from calculus. $\sum \frac{1}{n^p}$ converges for $p > 1$ and diverges for $p \leq 1$; ie, there is a sharp boundary where an infinitesimally small change leads to wildly different behaviour. For example, $\sum n^{-1}$ diverges, while $\sum n^{-(1+10^{-1000})}$ converges.

5.2 Continued Fractions

5.2.1 Introduction

Idea: there are various ways of representing numbers. There are decimal expansions, binary expansions, et cetera. If you have something complicated, one way to express it is to write it in terms of something simpler. Continued fractions is an example of this.

Decimal expansion is very simple:

$$\begin{aligned} x &= x_n 10^n + x_{n-1}10^{n-1} + \cdots + x_1 10^1 + x_0 + x_{-1}10^{-1} + x_{-2}10^{-2} + \cdots \\ & \quad x_i \in \{0, 1, \ldots, 9\}. \end{aligned} \tag{5.3}$$

Exercise 5.2.1. *Let x have a periodic decimal expansion. For example, assume $\exists N_0 \in \mathbb{N}$ and $a_1, \ldots, a_n \in \{0, \ldots, 9\}$ such that*

$$\begin{aligned} x &= x_m x_{m-1} \cdots x_1 x_0 . x_{-1} \cdots x_{N_0+1} x_{N_0} a_1 \cdots a_n a_1 \cdots a_n a_1 \cdots a_n \cdots \\ &= x_m x_{m-1} \cdots x_1 x_0 . x_{-1} \cdots x_{N_0+1} x_{N_0} \overline{a_1 \cdots a_n} \end{aligned} \tag{5.4}$$

Prove that x is rational, and bound the size of the denominator.

Continued Fractions is a much more sophisticated machine than decimal expansion. Any finite continued fraction (with integer components) will be a rational number, and vice versa. This is a lot cleaner than something that goes on to infinity and is periodic. A periodic Continued Fraction is actually the solution of a quadratic equation with integer coefficients, which is very different than a periodic decimal expansion.

A lot of very complicated numbers (for example, e), have very simple Continued Fraction expansions.

Using Continued Fractions of numbers, you can get very interesting results on how to approximate numbers by rationals. For example, if you have the decimal expansion of a number, if you truncate the decimal expansion at some point, you get a rational approximation (some integer divided by a power of ten).

You can do this with a continued fraction: you can *cut* it at some point and get a rational number, and use that rational number to approximate the number we started with. We will see that *this is the best approximation you can have*; we will, of course, quantify what we mean by best approximation.

What does this remind us of? Fourier Series or Taylor Series: for a given expansion, the first n terms of a Fourier Series (or Taylor Series) give the best approximation of a certain order to the given function.

A finite continued fraction has this type of power: it is a very sophisticated machine.

5.2.2 Definition

A **Finite Continued Fraction** is a number of the form

$$a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{\ddots + \cfrac{1}{a_n}}}} \tag{5.5}$$

One doesn't want to write something like this every time, so we introduce the following shorthand notations. The first is

$$a_0 + \frac{1}{a_1+}\frac{1}{a_2+}\cdots\frac{1}{a_n} \tag{5.6}$$

A better notation is

$$[a_0, a_1, \ldots, a_n]. \tag{5.7}$$

Exercise 5.2.2. *Show* $[a_0] = a_0$, $[a_0, a_1] = a_0 + \frac{1}{a_1} = \frac{a_0a_1+1}{a_1}$, *and* $[a_0, a_1, a_2] = \frac{a_0(a_1a_2+1)+a_2)}{a_1a_2+1}$.

5.2.3 Elementary Properties of Continued Fractions

Lemma 5.2.3. *Let* $[a_0, \ldots, a_n]$ *be a Continued Fraction. Then*

1. $\Big[a_0, \ldots, a_n\Big] = \Big[a_0, \ldots, a_{n-2}, a_{n-1} + \frac{1}{a_n}\Big]$.
2. $\Big[a_0, \ldots, a_n\Big] = \Big[a_0, \ldots, a_{m-1}, [a_m, \ldots, a_n]\Big]$.

These are the most basic properties of Continued Fractions, and will be used constantly below.

Exercise 5.2.4. *Prove Lemma 5.2.3.*

5.2.4 Convergence to a Continued Fraction

We saw that $[a_0] = a_0$, $[a_0, a_1] = \frac{a_0a_1+1}{a_1}$, and $[a_0, a_1, a_2] = \frac{a_0(a_1a_2+1)+a_2}{a_1a_2+1}$. In general, when we simplify everything (if the continued fraction has finitely many terms), we get the ratio of two numbers. We denote this by $\frac{p_n}{q_n} = \frac{p_n(a_0,\ldots,a_n)}{q_n(a_0,\ldots,a_n)}$, where p_n and q_n are polynomials with integer coefficients of $a_0, a_1, \ldots, a_n$.

Theorem 5.2.5. *For any* $m \in \{2, \ldots, n\}$ *we have*

1. $p_0 = a_0$, $p_1 = a_0a_1 + 1$, *and* $p_m = a_mp_{m-1} + p_{m-2}$.
2. $q_0 = 1$, $q_1 = a_1$, *and* $q_m = a_mq_{m-1} + q_{m-2}$.

We prove this by induction. First, we check a few cases.

By definition, $[a_0] = \frac{a_0}{1}$, which is $\frac{p_0}{q_0}$. $[a_0, a_1] = \frac{a_0 a_1 + 1}{a_1}$, which agrees with $\frac{p_1}{q_1}$.

$[a_0, a_1, a_2]$ should be $\frac{a_2 p_1 + p_0}{a_2 q_1 + q_0}$. As $p_1 = a_0 a_1 + 1$ and $q_1 = a_1$, direct substitution gives $\frac{a_2(a_0 a_1 + 1) + a_0}{a_2 a_1 + 1}$.

We have proved the basis case (and two others just for fun). We now show that if

$$[a_0, \ldots, a_m] = \frac{p_m}{q_m} = \frac{a_m p_{m-1} + p_{m-2}}{a_m q_{m-1} + q_{m-2}} \tag{5.8}$$

then

$$[a_0, \ldots, a_{m+1}] = \frac{p_{m+1}}{q_{m+1}} = \frac{a_{m+1} p_m + p_{m-1}}{a_{m+1} q_m + q_{m-1}}. \tag{5.9}$$

We calculate the continued fraction of $x = [a_0, \ldots, a_m, a_{m+1}]$. By Lemma 5.2.3, this is the same as the continued fraction of $y = [a_0, \ldots, a_{m-1}, a_m + \frac{1}{a_{m+1}}]$. Note, of course, that $x = y$; we use a different letter to emphasize that x has a continued fraction expansion with $m + 2$ terms, and y has a continued fraction expansion with $m + 1$ terms (remember we start counting with a_0).

We consider the Continued Fraction of y; it will have its own expansion with numerator P_m and denominator Q_m. By induction (we are assuming we know the theorem for all continued fractions with m terms), $y = \frac{P_m}{Q_m}$.

Therefore,

$$\frac{P_m}{Q_m} = \frac{\left(a_m + \frac{1}{a_{m+1}}\right) P_{m-1} + P_{m-2}}{\left(a_m + \frac{1}{a_{m+1}}\right) Q_{m-1} + Q_{m-2}}. \tag{5.10}$$

But the first m terms terms of y are the same as those of x. Thus, $P_{m-1} = p_{m-1}$, and similarly for Q_{m-1}, P_{m-2}, and Q_{m-2}.

Substituting gives

$$\frac{P_m}{Q_m} = \frac{\left(a_m + \frac{1}{a_{m+1}}\right) p_{m-1} + p_{m-2}}{\left(a_m + \frac{1}{a_{m+1}}\right) q_{m-1} + q_{m-2}}. \tag{5.11}$$

Standard algebra gives

$$\frac{P_m}{Q_m} = \frac{(a_m a_{m+1} + 1) p_{m-1} + p_{m-2} a_{m+1}}{(a_m a_{m+1} + 1) q_{m-1} + q_{m-2} a_{m+1}}. \tag{5.12}$$

This is the same as

$$\frac{a_{m+1}(a_m p_{m-1} + p_{m-2}) + p_{m-1}}{a_{m+1}(a_m q_{m-1} + q_{m-2}) + q_{m-1}} = \frac{a_{m+1}p_m + p_{m-1}}{a_{m+1}q_m + q_{m-1}}, \tag{5.13}$$

where the last step (substituting in with p_m and q_m) follows from the inductive assumption. This completes the proof. □

A cute example is

$$[1, 1, \ldots, 1] \;=\; 1 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{\ddots_{+\frac{1}{1}}}}} \;=\; \frac{p_n}{q_n}. \tag{5.14}$$

where we have $n+1$ ones. What are the p_i's and the q_i's? $p_0 = 1$, $p_1 = 2$, $p_m = p_{m-1} + p_{m-2}$. Similarly, we get $q_0 = 1$, $q_1 = 1$, and $q_m = q_{m-1} + q_{m-2}$.

Let F_m be the m^{th} Fibonacci number: $F_0 = 1$, $F_1 = 1$, $F_2 = 2$, $F_3 = 5$, and $F_m = F_{m-1} + F_{m-2}$.

Thus, $[1, 1, \cdots, 1] = \frac{F_{n+1}}{F_n}$. As we let the number of ones go to infinity, we can show this will converge to the golden ratio (also called the golden mean), $\frac{1+\sqrt{5}}{2}$.

Notice how beautiful Continued Fractions are. A simple expression like this captures the golden ratio, which has many deep, interesting properties. In base ten, $.111111\ldots$ is just $\frac{1}{9}$.

Exercise 5.2.6. *Let $r_n = \frac{F_{n+1}}{F_n}$. Show that the even terms, r_{2m}, are increasing and the odd terms, r_{2m+1}, are decreasing.*

Exercise 5.2.7. *Investigate $\lim_{n\to\infty}(r_n - r_{n-1})$ for the Fibonacci numbers. Show r_n converges to the golden ratio, $\frac{1+\sqrt{5}}{2}$.*

5.2.5 Observation

We know $\frac{p_n}{q_n} = \frac{a_n p_{n-1} + p_{n-2}}{a_n q_{n-1} + q_{n-2}}$. Consider the difference $p_n q_{n-1} - p_{n-1} q_n$.

Using the recursion relations, this difference also equals

$$(a_n p_{n-1} + p_{n-2}) q_{n-1} - p_{n-1}(a_n q_{n-1} + q_{n-2}). \tag{5.15}$$

This is the same (expand and cancel) as $p_{n-2}q_{n-1} - p_{n-1}q_{n-2}$.

The key observation is as follows: $p_n q_{n-1} - p_{n-1} q_n = -(p_{n-1}q_{n-2} - p_{n-2}q_{n-1})$. The index has reduced by one, and there has been a sign change. Repeat, and we get $p_{n-2}q_{n-3} - p_{n-3}q_{n-2}$. Doing $n-1$ times in total, we get $(-1)^{n-1}(p_1 q_0 - p_0 q_1)$.

Substituting $p_1 = a_0a_1 + 1$, $q_1 = a_1$, $p_0 = a_0$ and $q_0 = 1$ gives

Lemma 5.2.8.

$$p_nq_{n-1} - p_{n-1}q_n = (-1)^{n-1}. \tag{5.16}$$

So, even though a priori this difference should depend on a_0 through a_n, it is in fact just -1 to a power.

Similarly, one can show

Lemma 5.2.9.

$$p_nq_{n-2} - p_{n-2}q_n = (-1)^n a_n. \tag{5.17}$$

Notice that the consecutive convergents to the continued fraction, $\frac{p_n}{q_n}$, $\frac{p_{n-1}}{q_{n-1}}$ and $\frac{p_{n-2}}{q_{n-2}}$, satisfy

Lemma 5.2.10.

$$\frac{p_n}{q_n} - \frac{p_{n-1}}{q_{n-1}} = \frac{(-1)^{n-1}}{q_nq_{n-1}} \tag{5.18}$$

and

$$\frac{p_n}{q_n} - \frac{p_{n-2}}{q_{n-2}} = \frac{(-1)^n a_n}{q_nq_{n-2}} \tag{5.19}$$

To prove this, divide the previous relations by q_nq_{n-1} and q_nq_{n-2}.

5.2.6 Continued Fractions with Positive Terms

Let $x = \frac{p_n}{q_n}$ be the continued fraction of $[a_0, \ldots, a_n]$. We call a_0, $a_1, \ldots, a_n$ the **quotients** of the continued fraction. Let $x_m = \frac{p_m}{q_m}$.

Theorem 5.2.11. *If the quotients a_0 to a_n are positive, then the sequence x_{2m} is an increasing sequence, the sequence x_{2m+1} is a decreasing sequence, and for every m, $x_{2m} < x < x_{2m+1}$ (if $n \neq 2m$ or $2m+1$).*

Proof: x_{2m} increasing means $x_0 < x_2 < x_4 < \ldots$. By Lemma 5.2.10,

$$x_{2(m+1)} - x_{2m} = \frac{(-1)^{2m}a_{2m}}{q_{2m}q_{2(m+1)}}. \tag{5.20}$$

Everything on the right hand side is positive, so $x_{2(m+1)} > x_{2m}$. The result for the odd terms is proved similarly; there we will have $(-1)^{2m+1}$ instead of $(-1)^{2m}$, and we will see the odd terms are decreasing.

We know $x_0 < x_2 < x_4 < \ldots$ and $\cdots < x_5 < x_3 < x_1$. We know x_n, the last guy, is either an x_{2m} or an x_{2m+1} (depending on whether n is odd or even). It must be sandwiched somewhere in the middle. We will verify that $x_{2m+1} - x_{2m}$ is positive. Thus, x_n must be between the two.

We want to see how x_{2m}, x_{2m+1}, x_{2m+2} and x_{2m+3} should be ordered. We claim the ordering should be

$$x_{2m} < x_{2m+2} < x_{2m+3} < x_{2m+1} \tag{5.21}$$

Clearly, as the even terms are increasing and the odd terms are decreasing, $x_{2m} < x_{2m+2}$ and $x_{2m+3} < x_{2m+1}$. Thus, we need only show that x_{2m+3} is greater than x_{2m+2}. This follows immediately from Lemma 5.2.10 (take $n = 2m + 3$ in the lemma).

If n is even, x_n is greater than all the other even terms; if n is odd, x_n is less than all the other odd terms. Collecting the results now yields the theorem.

Chapter 6

Second Lecture

6.1 Another Introduction

Given $x \in \mathbb{R}$, how does one calculate the continued fraction expansion? We first describe the algorithm for determining the decimal expansion, and then we give an algorithm for finding the continued fraction expansion.

6.1.1 Decimal Expansion

Recall $[x]$ is the largest integer less than or equal to x.

Exercise 6.1.1. *Find* $[x]$ *for* $x = -2, 2.9, 3, 3.1, 3.14, \pi, 3.15$ *and* $\frac{29}{5}$. *Does* $[x + y] = [x] + [y]$*? Does* $[xy] = [x] \cdot [y]$*?*

For example, we calculate the decimal expansion of $x = 9.75$. $[x] = [9.75] = 9$. Call this x_1: $x_1 = [x]$.

How do we retrieve the next digit, 7? Look at $x - x_1$. This will be $.75$; if we multiply by 10, we get 7.5, and we note that the greatest integer less than or equal to 7.5 is 7.

Thus, look at $[10(x - x_1)] = 7$, and define $x_2 = 10(x - x_1) = 7.5$. Iterating the above procedure yields the base ten expansion.

Exercise 6.1.2. *Formally write down the procedure to find the base ten expansion of a positive number* x*. Discuss the modifications needed if* x *is negative.*

6.1.2 Continued Fraction Expansion

We expect to get something like

$$x = a_0 + \frac{1}{a_1 + \frac{1}{a_2 + \cdots}}, \tag{6.1}$$

where the a_i are positive integers.
Obviously, $a_0 = [x]$, the greatest integer at most x. Then

$$x - [x] = \frac{1}{a_1 + \frac{1}{a_2 + \cdots}}. \tag{6.2}$$

and the inverse is

$$x_1 = \frac{1}{x - [x]} = a_1 + \frac{1}{a_2 + \frac{1}{a_3 + \cdots}}. \tag{6.3}$$

Therefore, the second digit of the continued fraction expansion is $[x_1] = a_1$.
Let $x_2 = \frac{1}{x_1 - [x_1]}$ and iterate.

Exercise 6.1.3. *Formally write down the procedure to find the continued fraction expansion of a positive number x. Discuss the modifications needed if x is negative.*

Exercise 6.1.4. *Find the first few terms in the continued fraction expansions of $\sqrt{2}$, $\sqrt{3}$, π and e.*

6.1.3 Dynamical Interpretation

We are defining a map

$$f(x) = \frac{1}{x - [x]},\ x > 1. \tag{6.4}$$

If $x > 1$, then $f(x) > 1$ (and is infinite only if $x \in \mathbb{N}$). As $f(x) > 1$, we can apply f to $f(x)$ and get $f(f(x))$. As long as the initial value is greater than 1, we can keep iterating. The results will always be greater than one (and finite for non-integer input).

If we start with $x \in [0, 1)$, then $x - [x] = x$. Thus, for $x \in [0, 1)$, $f(x) = \frac{1}{x}$. If $x > 1$, then $f(x) \neq \frac{1}{x}$: it will be shifted.

Exercise 6.1.5. *Graph $f(x)$, $f(f(x))$, and $f(f(f(x))) = (f \circ f \circ f)(x)$.*

Draw a diagonal map $g(x) = x$. Given x_0, look at $f(x_0)$, and find x such that $g(x) = f(x_0)$. Thus, go from $\Big(x_0, f(x_0)\Big)$ to $\Big(f(x_0), f(x_0)\Big)$. Then, project this point to $\Big(f(x_0), f(f(x_0))\Big)$, and continue the process indefinitely.

Exercise 6.1.6. *Find all points in $[0, 1]$ such that when you iterate infinitely often the above, it converges to a fixed point on the curve. What are the conditions on points in $[0, 1]$ that lead to interesting behaviour? (Extremely hard!)*

See the papers of S. Zakeri at the University of Pennsylvania.

Exercise 6.1.7. *Fact: the continued fraction expansion of a rational number is finite. Prove this implies that if $x \in \mathbb{Q}$, then eventually you must land on a singular point (ie, you are eventually sent to infinity).*

Observation: if you start with a rational number, there are finite numbers taken before the process explodes; if you start with a number x which statisfies a degree 2 equation, the process is periodic.

6.2 Positive, Simple Convergents

Definition 6.2.1 (Positive Continued Fraction). *A continued fraction $[a_0, \dots, a_n]$ is positive if each $a_i > 0$.*

Definition 6.2.2 (Simple Continued Fraction). *A continued fraction is simple if each a_i is a positive integer.*

Definition 6.2.3 (quotients or convergents). *If $x_m = [a_0, \dots, a_m] = \frac{p_m}{q_m}$, then $\frac{p_m}{q_m}$ is the m^{th} quotient or convergent.*

Recall Theorem 5.2.11: *If the quotients are positive, then x_{2n} is an increasing sequence, x_{2n+1} is a decreasing sequence, and for all n, $x_{2n} < x < x_{2n+1}$.*

The proof followed from looking at successive quotients, Lemma 5.2.10.

What is the goal? A decimal expansion of a number converges to the given number, even if the decimal expansion is infinite. We want to prove an analogous property for continued fractions. We described a process which associates a continued fraction to each number. We now show this process is well defined, namely, that the continued fraction does equal the initial number.

Looking at Theorem 5.2.11, we show the even (odd) quotients converge to x from below (above).

Theorem 6.2.4. *Let $[a_0, \ldots, a_n]$ be a positive, simple continued fraction. Then*

1. *$q_n \geq q_{n-1}$ $\forall n \geq 1$, and $q_n > q_{n-1}$ if $n > 1$.*
2. *$q_n \geq n$, with strict inequality if $n > 3$.*

Proof: Recall $q_0 = 1$, $q_1 = a_1 \geq 1$, $q_n = a_n q_{n-1} + q_{n-2}$. Each $a_n > 0$ and is an integer. Thus, $a_n \geq 1$ and $a_n q_{n-1} + q_{n-2} \geq q_{n-1}$, yielding $q_n \geq q_{n-1}$. If $n > 1$, $q_{n-2} > 0$, giving a strict inequality.

We prove the other claim by induction. Suppose $q_{n-1} \geq n - 1$. Then $q_n = a_n q_{n-1} + q_{n-2} \geq q_{n-1} + q_{n-2} \geq (n-1) + 1 = n$. If at one point the inequality is strict, it is strict from that point onward. □

Exercise 6.2.5. *What can one prove about the p_ns?*

Theorem 6.2.6. *Given a continued fraction expansion $[a_0, \ldots, a_n]$ with quotient $\frac{p_n}{q_n}$. Then $\frac{p_n}{q_n}$ is reduced.*

Proof: assume not, and let $d|p_n$ and $d|q_n$. Then $d|(p_n q_{n-1} - q_n p_{n-1})$. By Lemma 5.2.8, $p_n q_{n-1} - q_n p_{n-1} = (-1)^{n-1}$. Thus, $d|(-1)^{n-1}$, which implies $d = \pm 1$ and $\frac{p_n}{q_n}$ is reduced.

6.3 Representation of Numbers by Continued Fractions

Lemma 6.3.1. *Given $x = [a_0, \ldots, a_N]$. If N is odd, there is another continued fraction which also equals x, but with an even number of terms (and vice-versa).*

This is equivalent to the non-uniqueness in decimal expansions. For example, $3.499999999\cdots = 3.50$. We make the **convention** that we throw away any decimal expansion ending with all 9s and replace it with the appropriate expansion ending in 0.

Where does the ambiguity come from? Assume we have two continued fractions such that $[a_0, \ldots, a_N] = [a_0, \ldots, a_N - 1, 1]$. For example,

$$a_1 + \frac{1}{a_2} = a_1 + \frac{1}{(a_2 - 1) + \frac{1}{1}}. \tag{6.5}$$

The only caveat is that we cannot have a zero in a continued fraction expansion. Thus, the above is a correct proof *only if* $a_N \neq 1$; in the example given, we need $a_2 \neq 1$.

If $a_N = 1$, we consider a slight modification. For example, if $a_4 = 1$, we have

$$a_1 + \frac{1}{a_2 + \frac{1}{a_3 + \frac{1}{1}}} = a_1 + \frac{1}{a_2 + \frac{1}{a_3 + 1}}, \tag{6.6}$$

which completes the proof. □

Consider $[a_0, a_1, a_2, \ldots, a_N]$. Define $a'_n = [a_n, \ldots, a_N]$, the tail of the continued fraction. Then $[a_0, \ldots, a_N] = [a_0, \ldots, a_{n-1}, a'_n]$; however, the second continued fraction is positive but not necessarily simple (as a'_n need not be an integer).

Theorem 6.3.2. *Suppose $[a_0, \ldots, a_N]$ is positive and simple. Then $[a'_n] = a_n$ except when both $n = N - 1$ and $a_N = 1$, in which case $a_{N-1} = [a'_{N-1}] + 1$.*

Proof: a'_n is a continued fraction given by

$$a'_n = a_n + \frac{1}{a_{n+1} + \frac{1}{\ddots}}. \tag{6.7}$$

We just need to make sure that

$$\frac{1}{a_{n+1} + \frac{1}{\ddots}} < 1. \tag{6.8}$$

How could this equal 1 or more? The only possibility is if $a_{n+1} = 1$ and the sum of the remaining terms is 0. This happens only if both $n = N-1$ and $a_N = 1$, proving the theorem. □

Uniqueness Assumption (Notation): whenever we write a finite continued fraction, we assume $a_N \neq 1$, where N corresponds to the last term. Again, this is similar to notation from base ten expansion.

Chapter 7

Third Lecture

7.1 Interesting Problem

Question 7.1.1. *Is there a function $f : \mathbb{R}^+ \to \mathbb{R}^+$ such that $\lim_{x\to a} f(x) = 0$?*

We will prove there is no such function.

Question 7.1.2. *Is there an $f : \mathbb{R}^+ \to \mathbb{R}^+$ such that $f(x)f(y) \leq (x-y)^2$ for $x \neq y$?*

If such a function exists, the first problem has an affirmative answer. Fix a. As $f(a) \neq 0$, we have

$$\lim_{x\to a} f(x) = \lim_{x\to a} \frac{(x-a)^2}{f(a)} = 0. \tag{7.1}$$

Consider now

Question 7.1.3. *Is there an $f : \mathbb{Q}^+ \to \mathbb{Q}^+$ such that $f(x)f(y) \leq (x-y)^2$ for $x \neq y$?*

Yes: define $f(\frac{p}{q}) = \frac{1}{q^2}$, $(p,q) = 1$.

We now show there is no function satisfying the conditions of Question 7.1.1.

Proof: for any a, the value is $f(a)$. Form the circle with diameter from $(a,0)$ to $\Big(a, f(a)\Big)$ and center $\Big(a, \frac{f(a)}{2}\Big)$, and do the same at b. If $a \neq b$, we claim that the two circles are disjoint.

If the sum of the radii of two circles is less than the distance between the two centers, clearly the circles are disjoint.

In our case, the radii are $\frac{f(a)}{2}$ and $\frac{f(b)}{2}$. The distance between the centers is $\sqrt{(a-b)^2 + \left(\frac{f(a)-f(b)}{2}\right)^2}$. Thus, it is enough to show that

$$\frac{f(a)}{2} + \frac{f(b)}{2} \leq \sqrt{(a-b)^2 + \left(\frac{f(a)-f(b)}{2}\right)^2}. \tag{7.2}$$

We square both sides, and show the $\leq$ is correct if f is the function

$$\begin{aligned} \left(\frac{f(a)+f(b)}{2}\right)^2 &\leq (a-b)^2 + \left(\frac{f(a)-f(b)}{2}\right)^2 \\ \frac{f(a)^2 - 2f(a)f(b) + f(b)^2}{4} &\leq (a-b)^2 + \frac{f(a)^2 + 2f(a)f(b) + f(b)^2}{4} \\ f(a)f(b) &\leq (a-b)^2 \end{aligned} \tag{7.3}$$

The worst scenario is if the two circles exactly touch (if we have $=$ and not $<$). We use the solution of Question 7.1.3.

Prove there are only countably many such circles that can be placed. Hint: each circle contains a rational tuple. As $\mathbb{Q}^2$ is countable, we can enumerate the circles.

But there are uncountably many circles if we are studying a real-valued function! □

7.2 Uniqueness of Continued Fraction Expansions

Theorem 7.2.1 (Uniqueness of Continued Fraction Expansion). *Let $x = [a_0, \ldots, a_N]$ $= [b_0, \ldots, b_M]$ be continued fractions with $a_N, b_M > 1$. Then $N = M$ and $a_i = b_i$ for $i = 0$ to $N = M$.*

We proceed by induction. $a_0 = [x]$, $b_0 = [x]$. If $[a_0, \ldots, a_N] = [b_0, \ldots, b_N]$, then

$$\Big[[x], [a_1, \ldots, a_N]\Big] = \Big[[x], [b_1, \ldots, b_M]\Big]. \tag{7.4}$$

Then

$$[x] + \frac{1}{[a_1, \ldots, a_N]} = [x] + \frac{1}{[b_1, \ldots, b_M]}. \tag{7.5}$$

Thus, $[a_1, \ldots, a_N] = [b_1, \ldots, b_M]$. We now have one fewer component, and the proof follows by induction. $\square$

Given x, we can associate a continued fraction to x. $a_0 = [x]$,

$$x = a_0 + \frac{1}{a_1'} = a_0 + \frac{1}{a_1 + \frac{1}{a_2'}}, \tag{7.6}$$

and so on. We write a prime over the last component to signify it need not be an integer; ie, it is the real number (greater than or equal to 1) that gives an equality. Note the previous components are integer, and the last is like a remainder.

If $\xi_0 \neq 0$, $\frac{1}{\xi_0} = a_1 + \xi_1$. If $\xi_1 \neq 0$. $\frac{1}{\xi_1} = a_2 + \xi_2$, et cetera, where in general $a_i' = \xi_i^{-1}$.

If at some point $\xi_i = 0$, the process terminates. This means we have something like

$$x = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{\ddots + \frac{1}{a_N}}}}. \tag{7.7}$$

Theorem 7.2.2. *A number is rational if and only if its continued fraction expansion is finite.*

Proof: clearly, if the continued fraction expansion is finite, then the number is rational. The other direction is much harder.

Let $x = \frac{h}{k}$, $(h, k) = 1$, $k > 0$. Then

$$\frac{h}{k} = a_0 + \xi_0, \; h = a_0 k + \xi_0 k, \; 0 \leq \xi_0 < k \rightarrow 0 \leq \xi_o k < k. \tag{7.8}$$

Basically, $\xi_0 k$ is the remainder of the division of h by k. We continue this process.

$k_1 = \xi_0 k$, and

$$a_0' = \frac{1}{\xi_0} = \frac{k}{k\xi_0} = \frac{k}{k_1} = a_1 + \xi_1 \tag{7.9}$$

We now have $k_1 \xi_1 < k_1$, and we define $k_2 = k_1 \xi_1$. We started with k and now have k_1, k_2, et cetera, a decreasing sequence of positive numbers $k > k_1 > k_2 > \cdots$. The sequence must eventually terminate, as each iteration gives us a smaller non-negative number.

We now have

$$\frac{1}{\xi_1} = \frac{k_1}{k_1\xi_1} = \frac{k_1}{k_2} = a_2 + \xi_2, \tag{7.10}$$

where $k_2 > k_2\xi_2 = k_3$.

Exercise 7.2.3. *Let x have a periodic decimal expansion. Prove that x is rational.*

Exercise 7.2.4. *Let x be rational. What can you say about its decimal expansion?*

7.3 Convergence

How well do continued fractions converge to the given number? Recall $x = [a_0, a_1, \ldots, a_n, a'_{n+1}]$. Then

$$x = \frac{a'_{n+1}p_n + p_{n-1}}{a'_{n+1}q_n + q_{n-1}}. \tag{7.11}$$

How large is $\left|x - \frac{p_n}{q_n}\right|$, the difference between x and the n^{th} convergent?

$$\begin{aligned} \left|x - \frac{p_n}{q_n}\right| &= \frac{a'_{n+1}p_n + p_{n-1}}{a'_{n+1}q_n + q_{n-1}} - \frac{p_n}{q_n} \\ &= \frac{p_{n-1}q_n - p_nq_{n-1}}{q_n(a'_{n+1}q_n + q_{n-1})} \\ &= \frac{(-1)^n}{q_nq'_{n+1}} \end{aligned} \tag{7.12}$$

as $q'_1 = a'_1$, $q'_n = a'_nq_{n-1} + q_{n-2}$, and by Lemma 5.2.8, $p_{n-1}q_n - p_nq_{n-1} = (-1)^n$.

How large can q'_{n+1} be? How small?

Note $a_{n+1} < a'_{n+1} < a_{n+1} + 1$. Well, they could be equal, but only if we have a finite continued fraction.

For simplicity, we are **assuming we have an infinite continued fraction, so we don't need to worry about trivial modifications at the last component**. Thus, we are assuming $x \notin \mathbb{Q}$.

Note a'_n is what we need to *truncate* an infinite continued fraction. Thus, we initially have $x = [a_0, \ldots, a_n, a_{n+1}, \ldots] = [a_0, \ldots, a'_n]$.

Thus,

$$q'_{n+1} = a'_{n+1}q_n + q_{n-1} > a_{n+1}q_n + q_{n-1} = q_{n+1} \tag{7.13}$$

and

$$\begin{aligned} q'_{n+1} &< (a_{n+1}+1)q_n + q_{n-1} = a_{n+1}q_n + q_{n-1} + q_n \\ &= q_{n+1} + q_n \le a_{n+2}q_{n+1} + q_n = q_{n+2}, \end{aligned} \tag{7.14}$$

as a_{n+2} is a positive integer.
We have proven

Theorem 7.3.1.

$$\frac{1}{q_n q_{n+2}} < \left| x - \frac{p_n}{q_n} \right| < \frac{1}{q_n q_{n+1}}, \tag{7.15}$$

or

$$\frac{1}{q_{n+2}} < \left| p_n - q_n x \right| < \frac{1}{q_{n+1}}. \tag{7.16}$$

For example, for $x = [1,1,1,1,\dots]$, we get the Fibonacci numbers, and $\frac{F_{n+1}}{F_n} \to \left(\frac{1+\sqrt{5}}{2}\right)^n$.

Thus, even when multiplied by a huge number like q_n, these differences go to zero exponentially fast! *This* is why continued fractions are so useful.

Chapter 8

Fourth Lecture

8.1 Review

Definition 8.1.1. *By* $[a_0, a_1, \dots]$ *we mean*

$$\lim_{n\to\infty} [a_0, a_1, \dots, a_n]. \tag{8.1}$$

Exercise 8.1.2. *Prove*

1. *The even convergents are increasing and bounded.*
2. *The odd convergents are decreasing and bounded.*

Theorem 8.1.3. *The limit*

$$\lim_{n\to\infty} [a_0, a_1, \dots, a_n] \tag{8.2}$$

exists (for a_i *non-negative integers.*

By Exercise 8.1.2, the even convergents $\frac{p_{2n}}{q_{2n}}$ converges (say to A), and the odd convergents $\frac{p_{2n-1}}{q_{2n-1}}$ converge (say to B). We need only show $A = B$, or, equivalently,

$$\left| \frac{p_{2n}}{q_{2n}} - \frac{p_{2n-1}}{q_{2n-1}} \right| \to 0. \tag{8.3}$$

Recall Lemma 5.2.8, which states $p_{2n}q_{2n-1} - p_{2n-1}q_{2n} = (-1)^{2n-1}$
Looking at

$$\left|\frac{p_{2n}}{q_{2n}} - \frac{p_{2n-1}}{q_{2n-1}}\right| = \left|\frac{p_{2n}q_{2n-1} - p_{2n-1}q_{2n}}{q_{2n}q_{2n-1}}\right| = \frac{1}{q_{2n}q_{2n-1}} \leq \frac{1}{2n(2n-1)} \to 0, \quad (8.4)$$

as by Lemma 6.2.4, $q_n \geq n$. Thus, $A = B$. □

Theorem 8.1.4. *[Uniqueness Theorem] Let x have an infinite continued fraction expansion. If $x = [a_0, a_1, \dots] = [b_0, b_1, \dots]$, then $a_i = b_i$.*

As there is no last digit, we do not need to worry about the ambiguity in the last digit. This is markedly different than the slight non-uniqueness in finite continued fraction expansions.

Exercise 8.1.5. *Prove Theorem 8.1.4.*

Remark 8.1.6. *The continued fraction of a number is equal to that number.*

Consider the Taylor Series of a function. For good functions, the Taylor Series equals the function. If, however, we change the function away from the origin, then the Taylor Series will no longer agree further on.

One can write down the continued fraction of a number. There is no reason why this continued fraction should be equal to the given number. One must *prove* that these are the same.

Recall Theorem 7.3.1:

$$\left|x - \frac{p_n}{q_n}\right| < \frac{1}{q_n q_{n+1}} < \frac{1}{q_n^2}. \quad (8.5)$$

8.2 Periodic Continued Fractions

Consider a periodic continue fraction

$$[a_0, a_1, \dots, a_k, \dots, a_{k+m}, a_k, \dots, a_{k+m}, a_k, \dots, a_{k+m}, \dots]. \quad (8.6)$$

For example,

$$[1, 2, 3, 4, 5, 6, 7, 8, 9, 7, 8, 9, 7, 8, 9, 7, 8, 9, \dots]. \quad (8.7)$$

Theorem 8.2.1. *A number x has a peroidic continued fraction if and only if it satisfies a quadratic equation; ie, $\exists A, B, C \in \mathbb{Z}$ such that $Ax^2 + Bx + C = 0$.*

First Direction: If x has a periodic continued fraction, then x satisfies a quadratic equation with integer coefficients.

Let x have a periodic continued fraction:

$$\begin{aligned} x &= [a_0, a_1, \ldots, a_{L-1}, a_L, \ldots, a_{L+k-1}, a_L, \ldots, a_{L+k-1}, a_L, \ldots] \\ &= [a_0, a_1, \ldots, a_{L-1}, a'_L], \end{aligned} \tag{8.8}$$

where

$$\begin{aligned} a'_L &= [a_L, a_{L+1}, a_{L+2}, \ldots] \\ &= [a_L, a_{L+1}, \ldots, a_{L+k-1}, a'_L] \end{aligned} \tag{8.9}$$

As

$$a'_L = \frac{p'a'_L + p''}{q'a'_L + q''}, \tag{8.10}$$

a'_L solves the quadratic equation

$$q'(a'_L)^2 + (q'' - q')a'_L - p'' = 0. \tag{8.11}$$

As

$$x = [a_0, a_1, \ldots, a_{L-1}, a'_L] \tag{8.12}$$

upon solving we obtain

$$x = \frac{p_{L-1}a'_L + p_{L-2}}{q_{L-1}a'_L + q_{L-2}}, \tag{8.13}$$

which gives

$$a'_L = \frac{p_{L-2} - xq_{L-2}}{q_{L-1}x - p_{L-1}}. \tag{8.14}$$

Substituting the above for a'_L in Equation 8.11 yields

$$q'\left(\frac{p_{L-2} - xq_{L-2}}{q_{L-1}x - p_{L-1}}\right)^2 + (q'' - q')\left(\frac{p_{L-2} - xq_{L-2}}{q_{L-1}x - p_{L-1}}\right) - p'' \;=\; 0. \tag{8.15}$$

Multiplying through by $(q_{L-1}x - p_{L-1})^2$, we find that x solves a quadratic equation with non-zero quadratic coefficient.

Second Direction: If x satisfies a quadratic equation with integer coefficients, then x has a periodic continued fraction.

Assume x solves

$$ax^2 + bx + c = 0. \tag{8.16}$$

Further, we may assume the quadratic equation is irreducible over $\mathbb{Z}$ or $\mathbb{Q}$ (if not, then x would satisfy a linear equation). Thus, Equation 8.16 has no rational roots.

We must show x has a periodic continued fraction expansion.

We write $x = [a_0, a_1, \dots]$. We may write $x = [a_0, \dots, a_{n-1}, a'_n]$ and we get

$$x = \frac{p_{n-1}a'_n + p_{n-2}}{q_{n-1}a'_n + q_{n-2}}. \tag{8.17}$$

Substitute $\frac{p_{n-1}a'_n+p_{n-2}}{q_{n-1}a'_n+q_{n-2}}$ for x in Equation 8.16. Clear denominators by multiplying through by $(q_{n-1}a'_n + q_{n-2})^2$. We find a'_n satisfied the following quadratic equation

$$A_n(a'_n)^2 + B_n a'_n + C_n = 0, \tag{8.18}$$

where a messy (but straightforward) calculation gives

$$\begin{aligned} A_n &= ap_{n-1}^2 + bp_{n-1}q_{n-1} + cq_{n-1}^2 \\ B_n &= 2ap_{n-1}q_{n-2} + b(p_{n-1}q_{n-2} + p_{n-2}q_{n-1}) + 2cq_{n-1}q_{n-2} \\ C_n &= ap_{n-2}^2 + bp_{n-2}q_{n-2} + cq_{n-2}^2. \end{aligned} \tag{8.19}$$

Remark 8.2.2. $A_n \neq 0$.

Proof: If $A_n = 0$, then dividing the expression for A_n by q_{n-1}^2 gives $\frac{p_{n-1}}{q_{n-1}}$ satisfies Equation 8.16; however, Equation 8.16 has no rational solutions.

Thus, we have

$$A_n y^2 + B_n y + C_n = 0, \quad y = a'_n \text{ is a solution, } A_n \neq 0. \tag{8.20}$$

The discriminant of the above quadratic is (another messy but straightforward calculation)

$$\Delta = b_n^2 - 4A_nC_n = b^2 - 4ac. \tag{8.21}$$

By Theorem 7.3.1,

$$xq_{n-1} - p_n = \frac{\delta_{n-1}}{q_{n-1}},\ |\delta_{n-1}| < 1. \tag{8.22}$$

Thus,

$$A_n = a\left(xq_{n-1} + \frac{\delta_{n-1}}{q_{n-1}}\right)^2 + bq_{n-1}\left(xq_{n-1} + \frac{\delta_{n-1}}{q_{n-1}}\right) + cq_{n-1}^2. \tag{8.23}$$

Taking absolute values and remembering that $ax^2 + bx + c = 0$ gives

$$\left|(ax^2+bx+c)q_{n-1}^2 + 2ax\delta_{n-1} + a\frac{\delta_{n-1}^2}{q_{n-1}^2} + b\delta_{n-1}\right| \leq 2\cdot|a|\cdot|x| + |b| + |a|. \tag{8.24}$$

As $C_n = A_{n-1}$ we find that

$$\begin{aligned} |C_n| &\leq 2\cdot|a|\cdot|x| + |b| + |a| \\ B_n^2 - 4A_nC_n &= b^2 - 4ac \\ B_n &\leq \sqrt{|4A_nC_n| + |b^2 - 4ac|} < \sqrt{4\Big(2|a|\cdot|x| + |b| + |a|\Big)^2 + |b^2 - 4ac|}. \end{aligned} \tag{8.25}$$

We have shown

Lemma 8.2.3. *There is an M such that, for all n,*

$$|A_n|, |B_n|, |C_n| < M. \tag{8.26}$$

Thus, by Dirichlet's Box Principle, we can find three triples such that

$$(A_{n_1}, B_{n_1}, C_{n_1}) = (A_{n_2}, B_{n_2}, C_{n_2}) = (A_{n_3}, B_{n_3}, C_{n_3}). \tag{8.27}$$

We get three numbers a'_{n_1}, a'_{n_2} and a'_{n_3} which all solve the same quadratic equation (Equation 8.18), and the polynomial *is not* the zero polynomial as $A_n \neq 0$.

As any non-zero polynomial has at most two distinct roots, two of the three a_{n_i}s are equal. Without loss of generality, assume $a'_{n_1} = a'_{n_2}$.

This implies periodicity. Why?

$$\begin{aligned}[a_{n_1}, a_{n_1+1}, \ldots, a_{n_2}, \ldots] &= [a_{n_1}, a_{n_1+1}, \ldots, a'_{n_2}] \\ &= [a_{n_1}, a_{n_1+1}, \ldots, a'_{n_1}]. \end{aligned} \tag{8.28}$$

Notice we have *no idea* where the periodicity begins. The previous statement uses a'_n converts from an infinite continued fraction $[a_0, a_1, \ldots]$ to a finite continued fraction $[a_0, a_1, \ldots, a'_n]$.

Exercise 8.2.4. *Show* $\sqrt{2} = [1, 2, 2, 2, \ldots]$. *Hint:*

$$\begin{aligned}\sqrt{2} &= 1 + (\sqrt{2} - 1) \\ &= 1 + \frac{1}{2 + (\sqrt{2} - 1)} \\ &= 1 + \frac{1}{2 + \frac{1}{2 + (\sqrt{2} - 1)}}. \end{aligned} \tag{8.29}$$

Exercise 8.2.5. *Show* $\sqrt{3} = [1, 1, 2, 1, 2, 1, 2, \ldots]$.

Chapter 9

Approximations to Irrational Numbers

9.1 Convergents Give the Best Approximations

Theorem 9.1.1. *Let $x = [a_0, \dots]$ with convergents $\frac{p_n}{q_n}$. Then for $0 < q \leq q_n$, if $\frac{p}{q} \neq \frac{p_n}{q_n}$, then*

$$\left| x - \frac{p_n}{q_n} \right| \; < \; \left| x - \frac{p}{q} \right|. \tag{9.1}$$

Among numbers with bounded denominator, the continued fraction is the best approximation to the irrational.

The Theorem will follow from

Theorem 9.1.2. *Under the same assumptions,*

$$|p_n - q_n x| \; < \; |p - qx|. \tag{9.2}$$

Proof: Suppose p and q are relatively prime. By Theorem 7.3.1

$$|p_n - q_n x| \; < \; |p_{n-1} - q_{n-1} x|. \tag{9.3}$$

Thus, it is sufficient to investigate $q_{n-1} < q \leq q_n$.

Case 1: $q = q_n$
If $q = q_n$, we must have $p \neq p_n$ (otherwise $\frac{p}{q} = \frac{p_n}{q_n}$).
Therefore

$$\left|\frac{p_n}{q_n} - \frac{p}{q}\right| \;\geq\; \frac{1}{q_n} \tag{9.4}$$

if $p \neq p_n$ (as $q = q_n$, and $|p - p_n| \geq 1$). Thus, $\frac{p_n}{q_n}$ isn't too close to $\frac{p}{q}$.
Again by Theorem 7.3.1,

$$\left|\frac{p_n}{q_n} - x\right| \;\leq\; \frac{1}{q_n q_{n+1}} \;\leq\; \frac{1}{3q_n}. \tag{9.5}$$

We are assuming $n > 1$. This will give $q_{n+1} \geq 3$. Thus, $\frac{p_n}{q_n}$ is close to x.
Consider the ball of radius $\frac{1}{3q_n}$ about $\frac{p_n}{q_n}$. Then x is within this ball; however, $\frac{p}{q}$ is *not* within this ball. $\frac{p}{q}$ is at least $\frac{1}{q_n}$ units from $\frac{p_n}{q_n}$.
Therefore, the closest x can be to $\frac{p}{q}$ is $\frac{2}{3q_n}$, or $|x - \frac{p}{q}| \geq \frac{2}{3q_n}$. But $|x - \frac{p_n}{q_n}| \leq \frac{1}{3q_n}$. Therefore,

$$\left|x - \frac{p_n}{q_n}\right| \;<\; \left|x - \frac{p}{q}\right|, \tag{9.6}$$

as was to be proved.

Case 2: $q_{n-1} < q < q_n$
By our assumptions on q, $\frac{p}{q} \neq \frac{p_n}{q_n}$ or $\frac{p_{n-1}}{q_{n-1}}$.
We will find μ and ν such that

$$\mu p_n + \nu p_{n-1} \;=\; p, \quad \mu q_n + \nu q_{n-1} \;=\; q. \tag{9.7}$$

Assume relations of the above form. Multiplying the first by q_{n-1} and the second by p_{n-1} yields

$$\begin{aligned} \mu(p_n q_{n-1} - p_{n-1} q_n) &= p q_{n-1} - q p_{n-1} \\ \mu &= \pm(p q_{n-1} - q p_{n-1}). \end{aligned} \tag{9.8}$$

Similarly we find that

$$\nu \;=\; \pm(p q_n - q p_{n-1}), \tag{9.9}$$

where we use Lemma 5.2.8 to get $p_n q_{n-1} - p_{n-1} q_n = \pm 1$.
Thus, we can find integer μ and ν such that Equation 9.7 is true.

As $q = \mu q_n + \nu q_{n-1} < q_n$, μ and ν must have opposite signs.

Further, we know $p_n - q_n x$ and $p_{n-1} - q_{n-1} x$ have opposite signs (the even convergents are increasing, the odd convergents are decreasing: see Exercise 8.1.2).

Therefore, $\mu(p_n - q_n x)$ and $\nu(p_{n-1} - q_{n-1} x)$ have the same sign. But

$$p - qx \;=\; \mu(p_n - q_n x) + \nu(p_{n-1} - q_{n-1} x). \tag{9.10}$$

Thus

$$|p - qx| \;>\; |p_{n-1} - q_{n-1} x| \;>\; |p_n - q_n x|. \tag{9.11}$$

The above is the desired inequality. □

Exercise 9.1.3. *Show that* $q_n \geq 3$ *for* $n \geq 2$.

Chapter 10

Measure Theory, Sizes of Well-Approximated Sets, and Height Functions

10.1 Naive measure theory

10.1.1 Reconsidering length and area

What is the length of $\mathbb{Q}$? What is the area of $\mathbb{Q} \times \mathbb{Q}$? The answer is not intuitively clear: on the one hand, both $\mathbb{Q}$ and $\mathbb{Q}\times\mathbb{Q}$ are enumerable, and hence their elements form small subsets of the line and plane; on the other hand, $\mathbb{Q}$ and $\mathbb{Q} \times \mathbb{Q}$ are dense – you cannot have any open set outside it. Perhaps we should rephrase our question: can we extend our concept of area from the sets on which we normally use it to sets such as $\mathbb{Q} \times \mathbb{Q}$?

Of course, this raises two further questions: what is our concept of area? To what kind of sets do we usually apply it? As many of you must know from sad personal experience, applying certain familiar concepts to objects that look perfectly reasonable can result in absurdities and contradictions (for example, Russell's Paradox from set theory). Nevertheless, I think we may presume that we can talk about the length of intervals and the area of triangles and rectangles with a clean conscience. If you were taught geometry properly, you may have derived the area of a polygon from the single postulate that

$$\text{Area}\left(\bigcup_{i=1}^{\infty} S_i\right) = \sum_{i=1}^{\infty} \text{Area}(S_i) \tag{10.1}$$

when $S_1, S_2, \cdots$ are disjoint triangles or rectangles. Thus, starting with the area of triangles and rectangles, we can extend and calculate the area of polygons.

What about the areas of more exotic sets than polygons? We try and generalize the above construction, and see what the largest class of sets is where we can determine the area.

The set – which we shall call the set $\mathcal{R}$ of *measurable sets* from now on – will have to fulfill the following properties for proofs such as the one for the area of the polygon to work:

1. $A, B \in \mathcal{R} \Rightarrow A \cap B \in \mathcal{R}$,
2. $A, B \in \mathcal{R} \Rightarrow A - B \in \mathcal{R}$,
3. $A_1, A_2, \cdots \in \mathcal{R} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{R}$.

We will assume furthermore that $\mathcal{R}$ is a subset of the set of all subsets of a fixed Euclidean space $\mathbb{R}^n$. Thus, for areas, we assume the elements of $\mathcal{R}$ are subsets of the plane.

Our area function will be any function $\sigma : \mathcal{R} \to \mathbb{R}_0^+$ ($\mathbb{R}_0^+$ is the set of non-negative numbers) satisfying the following properties:

1. $\sigma(\emptyset) = 0$, where $\emptyset$ is the empty set.
2. $\sigma\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} A_i$ for $A_1, A_2, \cdots \in \mathcal{R}$ disjoint.

Finally, we require $\mathcal{R}$ to contain all traditional objects (intervals for $n = 1$, rectangles for $n = 2$, boxes for $n = 3$, etc.) and not to be so large that we obtain contradictions. On the intervals, rectangles, etc., we will set σ to have the same value as the standard corresponding notion.

Thus, for $n = 1$, we call σ the *length*, and if I is the interval $[a, b], [a, b), (a, b]$ or (a, b), we define $\sigma(I) = b - a$.

If $n = 2$, we call σ the *area*, and if R is a rectangle, with length l and width w, we define $\sigma(R) = l \cdot w$.

If $n = 3$, we call σ the *volume*, and if B is a box with length l and width w and height h, we define $\sigma(B) = l \cdot w \cdot h$.

And so on for higher dimensions. In general, we talk about the **measure** of a set, where in one-dimension measure is length, in two-dimensions measure is area, and so on.

Definition 10.1.1. *[Lebesgue Measure] We call the measure derived from intervals on the line, rectangles in the plane, boxes in three-space (and so on) the Lebesgue Measure.*

So what can $\mathcal{R}$ be? We refer the interested reader to more advanced texts. We'll just say that one can make it large enough to contain all sets we will actually apply σ to. Don't try this at home, though.

Exercise 10.1.2. *Knowing the area of a rectangle, show one can determine the area of an arbitrary triangle.*

Exercise 10.1.3. *Find the area of a regular n-gon, where each side has length x.*

Exercise 10.1.4 (Russell's Paradox). *A set is not just a collection of all objects satisfying a given property. Consider the potential set $\mathcal{R}$ defined by*

$$\mathcal{R} \ = \ \{x : x \in \mathcal{R} \text{ if and only if } x \notin x\}. \tag{10.2}$$

Most objects are not elements of themselves, thus most objects will be in $\mathcal{R}$. For example, the set of natural number $\mathbb{N}$ is not a natural number, so $\mathbb{N} \in \mathcal{R}$.

If $\mathcal{R}$ were a set, we could ask, "Is $\mathcal{R} \in \mathcal{R}$? Show that both $\mathcal{R} \in \mathcal{R}$ and $\mathcal{R} \notin \mathcal{R}$ contradict the definition of $\mathcal{R}$. Thus, $\mathcal{R}$ is not a set.

10.1.2 Measure of the Rationals

Let

$$\mathbb{Q}_{[0,1]} = [0,1] \cap \mathbb{Q}. \tag{10.3}$$

What is the measure $\sigma(\mathbb{Q}_{[0,1]})$ of $\mathbb{Q}_{[0,1]}$?

It follows from our postulates that, if $A \subset B$, then $\sigma(A) \leq \sigma(B)$. Now let $x_1, x_2, x_3, \ldots$ be the elements of $\mathbb{Q}_{[0,1]}$. (As you know, $\mathbb{Q}$ is ennumerable, and hence so is $\mathbb{Q}_{[0,1]}$.) For $\epsilon > 0$ arbitrary, define

$$U_{i,\epsilon} = \left(x_i - \frac{\epsilon}{2 \cdot 2^i}, x_i + \frac{\epsilon}{2 \cdot 2^i}\right). \tag{10.4}$$

Then

$$\begin{aligned} \mathbb{Q}_{[0,1]} \ &\subset \ \bigcup_{i=1}^{\infty} U_{i,\epsilon} \\ \sigma\left(\bigcup_{i=1}^{\infty} U_{i,\epsilon}\right) \ &\leq \ \sum_{i=1}^{\infty} \frac{\epsilon}{2^i} \ \leq \ \epsilon. \end{aligned} \tag{10.5}$$

We have shown that, given any $\epsilon > 0$, we can find a union of intervals that contains $\mathbb{Q}_{[0,1]}$ and has measure less than ϵ. Thus, $\sigma\left(\mathbb{Q}_{[0,1]}\right) = 0$.

Exercise 10.1.5. *Show that $\mathbb{Q}_{[0,1]} \times \mathbb{Q}_{[0,1]}$ has zero area.*

In general, any countable set will have measure zero. (The argument above works for all such sets.) However, not every set with measure zero will be countable. This allows us to distinguish between the sizes of different uncountable subsets of the real line. You may remember from your distant past (last week) that for any irrational number $x \in [0,1]$ there are infinitely many integers p, q, $0 \leq p \leq q$, $q > 0$, such that

$$\left| x - \frac{p}{q} \right| \leq \frac{1}{q^2} \tag{10.6}$$

We will now see that the set of privileged points that can be approximated a great deal more closely than every number can by (10.6) is actually rather small.

10.2 Measures of Sets with Given Continued Fraction Approximations

10.2.1 $\left|x - \frac{p}{q}\right| \leq \frac{C}{q^{2+\epsilon}}$ Infinitely Often

Theorem 10.2.1. *Let C, ϵ be positive constants. Let S be the set of all points $x \in [0,1]$ such that there are infinitely many integers p, q with*

$$\left| x - \frac{p}{q} \right| \leq \frac{C}{q^{2+\epsilon}}. \tag{10.7}$$

Then $\sigma(S) = 0$.

Proof. Let $N > 0$. Let S_N be the set of all points $x \in [0,1]$ such that there are $p, q \in \mathbb{Z}$, $q > N$, for which

$$\left| x - \frac{p}{q} \right| \leq \frac{C}{q^{2+\epsilon}}. \tag{10.8}$$

If $x \in S$, then $x \in S_N$ for every N. Thus, if we can show that the measure of the sets S_N becomes arbitrarily small as $N \to \infty$, then the measure of S must be zero.

How big can S_N be? For a given q, there are at most q choices for p. Given a pair (p,q), how many x's are there within $\frac{C}{q^{2+\epsilon}}$ of $\frac{p}{q}$? Clearly, the set of such points is the interval

$$I_{p,q} \ = \ \Big(\frac{p}{q} - \frac{C}{q^{2+\epsilon}}, \frac{p}{q} + \frac{C}{q^{2+\epsilon}}\Big). \tag{10.9}$$

Note that the measure of $I_{p,q}$ is $\frac{2C}{q^{2+\epsilon}}$. Let I_q be the set of all x in $[0,1]$ that are within $\frac{C}{q^{2+\epsilon}}$ of a rational number with denominator q. Then

$$I_q \ \subset \ \bigcup_{p=0}^{q} I_{p,q} \tag{10.10}$$

and therefore

$$\begin{aligned} \sigma(I_q) \ &\leq \ \sum_{p=0}^{q} \sigma(I_{p,q}) \\ &= \ (q+1)\cdot\frac{2C}{q^{2+\epsilon}} \\ &= \ \frac{q+1}{q}\frac{2C}{q^{1+\epsilon}} < \frac{4C}{q^{1+\epsilon}}. \end{aligned} \tag{10.11}$$

Then

$$\begin{aligned} \sigma(S_N) \ &\leq \ \sum_{q>N} \sigma(I_q) \\ &= \ \sum_{q>N} \frac{4C}{q^{1+\epsilon}} \\ &< \ \frac{4C}{1+\epsilon} N^{-\epsilon}. \end{aligned} \tag{10.12}$$

Thus, as N goes to infinity, $\sigma(S_N)$ goes to zero. As $S \subset S_N$, $\sigma(S) = 0$. $\square$

Exercise 10.2.2. *Instead of working with* $\left|x - \frac{p}{q}\right| \leq \frac{C}{q^{2+\epsilon}}$, *show the same argument works for* $\left|x - \frac{p}{q}\right| \leq \frac{C}{f(q)}$, *where* $\sum \frac{q}{f(q)} < \infty$.

We can, however, improve on (10.6) somewhat.

10.2.2 $\left|x - \frac{p}{q}\right| \leq \frac{1}{q^2\sqrt{5}}$

Theorem 10.2.3. *Let $x \in [0,1]$ be an irrational number. Then there are infinitely many integers p, q such that*

$$\left|x - \frac{p}{q}\right| \leq \frac{1}{q^2\sqrt{5}}. \tag{10.13}$$

Proof. Note that any finite continued fraction is rational, and any rational number can be represented as a finite continued fraction. Thus, we may assume our number x has an infinite continued fraction expansion.

We will show that, of any three consecutive approximations $\frac{p_{n-1}}{q_{n-1}}$, $\frac{p_n}{q_n}$, $\frac{p_{n+1}}{q_{n+1}}$ to x coming from the continuous fraction expansion of x, at least one satisfies (10.13). Let

$$b_{i+1} = \frac{q_{i-1}}{q_i}. \tag{10.14}$$

Then

$$\begin{aligned}\left|\frac{p_i}{q_i} - x\right| &= \frac{1}{q_i q'_{i+1}} = \frac{1}{q_i(a'_{i+1}q_i + q_{i-1})} \\ &= \frac{1}{q_i(a'_{i+1}q_i + b_{i+1}q_i)} \\ &= \frac{1}{q_i^2}\frac{1}{a'_{i+1} + b_{i+1}}.\end{aligned} \tag{10.15}$$

It is thus enough to prove that

$$a'_i + b_i > \sqrt{5} \tag{10.16}$$

for at least one of any three consecutive values $m-1$, m, $m+1$ of i. Assume

$$\begin{aligned}a'_{n-1} + b_{n-1} &\leq \sqrt{5}, \\ a'_n + b_n &\leq \sqrt{5}.\end{aligned}$$

By definition

$$a'_{n-1} = a_{n-1} + \frac{1}{a'_n} \tag{10.17}$$

and

$$\frac{1}{b_n} = \frac{q_{n-1}}{q_{n-2}} = \frac{a_{n-1}q_{n-2} + q_{n-3}}{q_{n-2}} = a_{n-1} + \frac{q_{n-3}}{q_{n-2}} = a_{n-1} + b_{n-1}. \tag{10.18}$$

Hence

$$\frac{1}{a'_n} + \frac{1}{b_n} = a'_{n-1} + b_{n-1} \leq \sqrt{5}. \tag{10.19}$$

Therefore

$$\begin{aligned} 1 &= a'_n \frac{1}{a'_n} \leq a'_n \left(\sqrt{5} - \frac{1}{b_n}\right) \leq (\sqrt{5} - b_n)\left(\sqrt{5} - \frac{1}{b_n}\right) \\ &= 6 - \sqrt{5}\left(b_n + \frac{1}{b_n}\right). \end{aligned} \tag{10.20}$$

In other words

$$b_n + \frac{1}{b_n} \leq \sqrt{5}. \tag{10.21}$$

Since b_n is rational, the inequality must be strict. Completing the square we obtain

$$b_n > \frac{1}{2}(\sqrt{5} - 1). \tag{10.22}$$

Now suppose

$$\begin{aligned} a'_{m-1} + b_{m-1} &\leq \sqrt{5} \\ a'_m + b_m &\leq \sqrt{5} \\ a'_{m+1} + b_{m+1} &\leq \sqrt{5}. \end{aligned} \tag{10.23}$$

Applying the above reasoning to $n = m$, $n = m + 1$, we obtain

$$\begin{aligned} b_m &> \frac{1}{2}(\sqrt{5} - 1) \\ b_{m+1} &> \frac{1}{2}(\sqrt{5} - 1). \end{aligned} \tag{10.24}$$

By (10.17) with $n = m + 1$ and (10.18) and (10.19) with $n = m$,

$$\begin{aligned} a_m &= \frac{1}{b_{m+1}} - b_m \\ &< \frac{1}{b_{m+1}} - \frac{1}{2}(\sqrt{5}-1) \\ &< \frac{1}{\frac{1}{2}(\sqrt{5}-1)} - \frac{1}{2}(\sqrt{5}-1) \\ &= \frac{1}{2}(\sqrt{5}+1) - \frac{1}{2}(\sqrt{5}-1) \\ &= 1. \end{aligned} \tag{10.25}$$

However, a_m is a positive integer, and there are no positive integers less than 1. Contradiction. □

From the above, we see that the approximation is often better than $\frac{1}{\sqrt{5}q^2}$. For example, if our continued fraction expansion has infinitely many 3s in its expansion, we can do at least as well as $\frac{1}{3q^2}$ infinitely often.

Exercise 10.2.4. *Show that $\frac{1}{\sqrt{5}q^2}$ is the best one can have for all irrationals by studying the golden mean, $\frac{1+\sqrt{5}}{2} = [1, 1, 1, \dots]$.*

Exercise 10.2.5. *Let x be any irrational other than the golden mean. How well can x be approximated? See, for example, [HW].*

10.3 Height Functions and Diophantine Equations

We will now discuss a standard technique for solving or obtaining descriptions of the solutions to a diophantine equation (are there any? are there infinitely many?). The technique consists in defining a function from the set of all solutions to the integers (this is called a *height function*) and examining the properties of the ordering of the solutions according to the values taken by the height function. As this sounds rather abstract, we will examine what was historically the first example.

10.3.1 Fermat's Equation

Lemma 10.3.1. *Any positive integral solution to*

$$x^2 + y^2 = z^2 \tag{10.26}$$

with x, y coprime and x even must satisfy

$$x = 2ab,\ y = a^2 - b^2,\ z = a^2 + b^2 \tag{10.27}$$

for some coprime positive integers a, b, one of them even.

Proof. Since x is even and x and y are coprime, y must be odd. Hence z is odd. Therefore $\frac{1}{2}(z-y)$ and $\frac{1}{2}(z+y)$ are integral. By (10.26),

$$\left(\frac{x}{2}\right)^2 = \left(\frac{z+y}{2}\right)\left(\frac{z-y}{2}\right). \tag{10.28}$$

x and y coprime implies that y and z are coprime as well. Hence $\frac{1}{2}(z-y)$ and $\frac{1}{2}(z+y)$ are coprime. Therefore, as their product is a square, they must both be squares:

$$\frac{z+y}{2} = a^2\ \frac{z-y}{2} = b^2. \tag{10.29}$$

Then a and b satisfy (10.27). Since y is odd, a and b are of opposite parity. □

Exercise 10.3.2. *Assume $x, y, z \in \mathbb{Z}$, $x^2 + y^2 = z^2$ and x, y and z are co-prime (thus, no two share a common factor). Prove exactly one of x and y is even, and exactly one is odd. Hint: clearly x and y are not both even; if both were odd, what is $x^2 + y^2$ congruent to mod 4? Can you find a $z \in \mathbb{Z}$ whose square is congruent to this mod 4?*

Exercise 10.3.3. *Prove that if uv is a square and u, v have no factors in common, then u and v are both squares.*

Theorem 10.3.4 (Fermat). *There are no positive integral solutions to*

$$x^4 + y^4 = z^2 \tag{10.30}$$

Proof. Suppose there are integral solutions to the given equation. Order all solutions (x, y, z) according to the value of z. (In the general framework, this is the same as saying that we define our height function to be $(x, y, z) \mapsto z$.) Now let (x_0, y_0, z_0) be the solution with the least z_0. (If there are several such solutions, choose one among them.) Clearly x_0 and y_0 will have to be coprime, as otherwise we could divide x, y and z by $\gcd(x, y)$ and thereby obtain a smaller solution.

If x_0 and y_0 are both odd, then

$$\begin{aligned} x_0^4 &\equiv 1(\bmod 4) \\ y_0^4 &\equiv 1(\bmod 4) \end{aligned} \tag{10.31}$$

and hence

$$z_0^2 = x_0^4 + y_0^4 \equiv 2(\bmod 4). \tag{10.32}$$

This cannot be, as no square can be congruent to 2 modulo 4.

Hence either x_0 or y_0 is even. We can assume without loss of generality that x_0 is even.

As we can write $x_0^4 + y_0^4 = z^2$ as

$$(x_0^2)^2 + (y_0^2)^2 \ = \ z_0^2, \tag{10.33}$$

by Lemma 10.3.1 we have

$$x_0^2 = 2ab, \ \ y_0^2 = a^2 - b^2, \ \ z_0 = a^2 + b^2, \tag{10.34}$$

with exactly one of a, b odd. If a were even and b odd, then y_0^2 would have to be congruent to -1 modulo 4, and this is impossible. Hence a is odd and b is even. Write $b = 2c$. Then

$$\left(\frac{x_0}{2}\right)^2 = \frac{x_0^2}{4} = \frac{2ab}{4} = \frac{4ac}{4} = ac. \tag{10.35}$$

Since a and b are coprime, a and c are coprime. Hence a and c must both be squares. Write

$$a = d^2, \ \ c = f^2. \tag{10.36}$$

Then (remembering $b = 2c = 2f^2$)

$$y^2 = a^2 - b^2 = d^4 - 4f^4, \tag{10.37}$$

and so

$$(2f^2)^2 + y^2 = (d^2)^2. \tag{10.38}$$

Applying Lemma 10.3.1 again, we obtain

$$2f^2 = 2lm, \ \ d^2 = l^2 + m^2 \tag{10.39}$$

for some coprime positive integers l, m. Since $f^2 = lm$, both l and m are squares:

$$l = r^2, \ m = s^2. \tag{10.40}$$

Therefore

$$d^2 = l^2 + m^2 \tag{10.41}$$

can be written as

$$d^2 = r^4 + s^4. \tag{10.42}$$

But

$$d \leq d^2 = a \leq a^2 < a^2 + b^2 = z_0, \tag{10.43}$$

and thus (10.42) is a solution to (10.30) with a value of z_0 smaller than z. Contradiction. □

We say a solution of $x^4 + y^4 = z^2$ is trivial if x, y or z is zero. The above argument proves that there are no non-trivial integer solutions.

Exercise 10.3.5. *Prove that the equation $x^4 + y^4 = z^2$ has no non-trivial integer solutions implies that $x^4 + y^4 = z^4$ has no non-trivial solutions.*

Exercise 10.3.6 (Fermat's Equation). *Assume $x^m + y^m = z^m$ has no non-trivial integer solutions for some $m \in N$. Prove that for any $a \in \mathbb{N}$, $x^{am} + y^{am} = z^{am}$ has no non-trivial integer solutions. Thus, it is enough to show there are no non-trivial solutions to Fermat's Equation for odd primes and for $m = 4$ (note we must do $m = 4$, as by Pythagoras there* are *solutions when $m = 2$).*

10.3.2 Method of Descent

We sketch an alternate proof that there are no non-trivial solutions of $x^4+y^4 = z^2$.

Definition 10.3.7 (Height of a solution). *Given an integer solution (x, y, z), we define the height of the solution, $h(x, y, z)$, by*

$$h(x, y, z) = \max\Big(|x|, |y|\Big). \tag{10.44}$$

Exercise 10.3.8. *Given any $C > 0$, prove there are only finitely many integer solutions of $x^4+y^4 = z^2$ with $h(x, y, z) \leq C$. More generally, given any polynomial with integer coefficients $p(x, y, z) = 0$, prove there are only finitely many integer solutions with height less than C.*

Let (x_0, y_0, z_0) be a non-trivial solution of $x^4+y^4 = z^2$. An identical argument as before leads to the existence of another non-trivial solution

$$r^4 + s^4 = d^2. \tag{10.45}$$

Exercise 10.3.9. *Show $r \neq 0, s \neq 0$, which implies that (r, s, d) is a non-trivial integer solution.*

Exercise 10.3.10. *Prove $h(r, s, d) < h(x_0, y_0, z_0)$. Hint: first prove*

$$r < \max\Big(|x_0|, |y_0|\Big) \text{ and } s < \Big(|x_0|, |y_0|\Big). \tag{10.46}$$

Thus, given a non-trivial integer solution (x_0, y_0, z_0) we can always find another non-trivial integer solution with smaller height.

We now apply the above construction to the non-trivial integer solution (r, s, d) and obtain another non-trivial integer solution (r_2, s_2, d_2) with *strictly smaller height*. We then apply the same construction to $(r_2, s_,d_2)$ and obtain another non-trivial integer solution (r_3, s_3, d_3) with *strictly smaller height*. And so on.

This is the Method of Descent. Starting with one non-trivial integer solution, we construct an infinite sequence of non-trivial integer solutions, each solution *strictly smaller* than the previous. Here our concept of *smaller* comes from our height function.

But as each solution has strictly smaller height, and the initial height $h(x_0, y_0, z_0)$ was finite, we cannot continue constructing smaller solutions indefinitely; in fact, we can only proceed at most $h(x_0, y_0, z_0) + 1$ times.

Thus, as the Method of Descent gives infinitely many solutions, we reach a contradiction. Our only assumption was that there existed a non-trivial integer solution (x_0, y_0, z_0); therefore, there are no non-trivial integer solutions of $x^4 + y^4 = z^2$.

Chapter 11

Fifth Lecture

11.1 Convergents are the Best Rational Approximations

Theorem 11.1.1. *Let* $x = [a_0, a_1, \dots]$ *with* n^{th} *convergent* $\frac{p_n}{q_n}$. *Suppose* $n > 1$, $0 < q \leq q_n$ *and* $\frac{p}{q} \neq \frac{p_n}{q_n}$. *Then*

$$\left|\frac{p_n}{q_n} - x\right| \; < \; \left|\frac{p}{q} - x\right|. \tag{11.1}$$

Proof. Suppose $(p, q) = 1$; if p and q are not relatively prime, the proof is easier. It is sufficient to show

$$|p_n - q_n x| \; < |p_{n-1} - q_{n-1} x|. \tag{11.2}$$

It will be sufficient to prove the above for $q_{n-1} < q \leq q_n$.

Case 1**:** $q = q_n$ We handled this case last time.

Case 2**:** $q_{n-1} < q < q_n$
Thus,

$$\frac{p}{q} \; \neq \; \frac{p_n}{q_n}, \frac{p_{n-1}}{q_{n-1}}. \tag{11.3}$$

Find μ, ν such that

$$\begin{aligned} \mu p_n + \nu p_{n-1} &= p \\ \mu q_n + \nu q_{n-1} &= q. \end{aligned} \tag{11.4}$$

As

$$\begin{vmatrix} p_n & p_{n-1} \\ q_n & q_{n-1} \end{vmatrix} = (-1)^{n-1}, \tag{11.5}$$

by Cramer's rule we can find such a μ and ν.
Thus,

$$\begin{aligned} \mu &= \pm(pq_{n-1} - qp_{n-1}) \\ \nu &= \pm(pq_n - qp_n). \end{aligned} \tag{11.6}$$

Since

$$q = \mu q_n + \nu q_{n-1} < q_n, \tag{11.7}$$

we find that μ and ν have opposite signs.
On the other hand,

$$p_n - q_n x,\ p_{n-1} - q_{n-1} x \tag{11.8}$$

have opposite signs; therefore,

$$\mu(p_n - q_n x),\ \nu(p_{n-1} - q_{n-1} x) \tag{11.9}$$

have the same signs.
This implies

$$p - qx = \mu(p_n - q_n x) + \nu(p_{n-1} - q_n x), \tag{11.10}$$

and there is no cancellation (as the two terms on the right have the same sign).
As ν is an integer, we find

$$\begin{aligned} \Big|p - qx\Big| &> \Big|\nu(p_{n-1} - q_{n-1})\Big| \\ &\geq \Big|p_{n-1} - q_{n-1} x\Big| \\ &> \Big|p_n - q_n x\Big|. \end{aligned} \tag{11.11}$$

This completes the proof. □

Thus, the convergents provide the best rational approximation to a given number. Now that we know the convergents are the best rational approximations, we now investigate *how well* they approximate.

11.2 Weaker Approximation Properties of Convergents

In our proof that every irrational can be approximated (infinitely often) as well as $\frac{1}{\sqrt{5}q^2}$, we proved

Theorem 11.2.1. *Of any three consecutive convergents to a continued fraction, at least one satisfies*

$$\left|x - \frac{p}{q}\right| \ < \ \frac{1}{\sqrt{5}q^2}. \tag{11.12}$$

One can show

Theorem 11.2.2. *If $\frac{p}{q}$ satisfies $\left|x - \frac{p}{q}\right| < \frac{1}{2q^2}$, then $\frac{p}{q}$ is a convergent of x.*

Exercise 11.2.3. *Prove the above theorem.*

Theorem 11.2.4. *Of any two consecutive convergents, one will satisfy*

$$\left|x - \frac{p}{q}\right| \ < \ \frac{1}{2q^2}. \tag{11.13}$$

Proof. Let x be irrational. We can write

$$\frac{p}{q} - x \ = \ \frac{\theta\epsilon}{q^2}, \ \epsilon = \pm 1, \ 0 < \theta < \frac{1}{2}. \tag{11.14}$$

Extend $\frac{p}{q}$ as a finite continued fraction $[a_0, \ldots, a_n]$. There is non-uniqueness in finite continued fractions (can make it have either even or odd number of terms). Choose $\epsilon = (-1)^{n-1}$. Find w such that

$$x = \frac{wp_n + p_{n-1}}{wq_n + q_{n-1}}, \ \frac{p_n}{q_n} = \frac{p}{q}. \tag{11.15}$$

We will consider $\frac{p_n}{q_n}$ and $\frac{p_{n-1}}{q_{n-1}}$.

Claim 11.2.5. *$\frac{p_{n-1}}{q_{n-1}}$, $\frac{p_n}{q_n}$ are in fact convergents to x.*

Choose w like this:

$$\begin{aligned} wq_nx + q_{n-1}x &= wp_n + p_{n-1} \\ wq_nx - wp_n &= p_{n-1} - q_{n-1} \\ \text{Therefore } w &= \frac{p_{n-1} - q_{n-1}x}{q_nx - p_n}. \end{aligned} \tag{11.16}$$

Lemma 11.2.6. *If $x = \frac{P\zeta+R}{Q\zeta+S}$ with $S > 1$ and P, Q, R, S are integers such that $Q > S > 0$ and $PS - QR = \pm 1$, then $\frac{R}{S}$ and $\frac{P}{Q}$ are two consecutive convergents to x.*

How does Lemma 11.2.6 imply the Theorem? In order to use the lemma, we need to show that $\zeta > 1$, which translates to showing $w > 1$.

Now,

$$\begin{aligned} \frac{\epsilon\theta}{q^2} &= \frac{\epsilon\theta}{q_n^2} \\ &= \frac{p_n}{q_n} - x \\ &= \frac{p_n}{q_n} - \frac{wp_n + p_{n-1}}{wq_n + q_{n-1}} \\ &= \frac{p_nq_{n-1} - p_{n-1}q_n}{q_n(wq_n + q_{n-1})} \\ &= \frac{(-1)^{n-1}}{q_n(wq_n + q_{n-1})}. \end{aligned} \tag{11.17}$$

Therefore,

$$\frac{q_n}{wq_n + q_{n-1}} = \theta \tag{11.18}$$

which gives

$$w = \frac{1}{\theta} - \frac{q_{n-1}}{q_n} > 1, \tag{11.19}$$

which completes the proof of the Theorem.

□

We must now prove Lemma 11.2.6.

Proof. Let

$$\frac{P}{Q} \;=\; [a_0, \ldots, a_n] \;=\; \frac{p_n}{q_n}. \tag{11.20}$$

We must have $P = p_n$ and $Q = q_n$, as these are reduced fractions. Thus, (P, Q) are relatively prime, $Q > 0$.

Choose n such that $PS - QR = \pm 1 = (-1)^{n-1}$. In particular, we have

$$\begin{aligned} p_n S - q_n R &= (-1)^{n-1} \\ &= p_n q_{n-1} - p_{n-1} q_n. \end{aligned} \tag{11.21}$$

Rewriting gives

$$p_n(S - q_{n-1}) \;=\; q_n(R - p_{n-1}). \tag{11.22}$$

As $q_n | p_n(S - q_{n-1})$, this implies $q_n | S - q_{n-1}$.
As $q_n = Q > S > 0$, $q_n > q_{n-1} > 0$, we must have

$$|S - q_{n-1}| < q_n. \tag{11.23}$$

As $q_n | S - q_{n-1}$, this forces $S = q_{n-1}$ and $R = p_{n-1}$.
Hence

$$x = \frac{p_n \zeta + p_{n-1}}{q_n \zeta + q_{n-1}} \tag{11.24}$$

which implies

$$x = [a_0, \ldots, a_{n-1}, \zeta], \tag{11.25}$$

proving the lemma.

□

11.3 Exponent (or Order) of Approximation

Definition 11.3.1 (approximated to order n). *ξ is approximated by rationals to order n (n need not be an integer) if $\exists k = k(\xi)$ such that*

$$\left|\frac{p}{q} - \xi\right| < \frac{k(\xi)}{q^n} \tag{11.26}$$

has only finitely many solutions.

Equivalently,

Definition 11.3.2 (approximation exponent). *ξ has order (or exponent) $\tau(\xi)$ if $\tau(\xi)$ is the smallest number such that $\forall e > \tau(\xi)$, the inequality*

$$\left|\frac{p}{q} - \xi\right| < \frac{}{q^e} \tag{11.27}$$

has only finitely many solutions.

Example 11.3.3. *A rational number has approximation exponent 1 and no more.*

Why? If $\xi = \frac{a}{b}$ and $r = \frac{s}{t} \neq \frac{a}{b}$, then $sb - at \neq 0$. Thus, $|sb - at| \geq 1$ (as it is integral). This implies

$$\begin{aligned} \left|\xi - \frac{s}{t}\right| &= \left|\frac{a}{b} - \frac{s}{t}\right| \\ &= \frac{|sb - at|}{bt} \\ &\geq \frac{1}{bt}. \end{aligned} \tag{11.28}$$

If the rational ξ had approximation exponent $e > 1$ we would find

$$|\xi - r| < \frac{1}{t^e}, \text{ which implies } \frac{1}{t^e} > \frac{1}{bt}. \tag{11.29}$$

Therefore, $t^{e-1} < b$. Since b is fixed, there are only finitely many such t. □

Example 11.3.4. *An irrational number has approximation exponent at least 2. We did this using Dirichlet's Box Principle (among other proofs).*

Theorem 11.3.5 (Liouville). *A real algebraic number of degree n is* not *approximateable to order larger than n. In other words, if α satisfies a polynomial equation with integer coefficients of degree n, then $\tau(\alpha) \leq n$.*

Corollary 11.3.6. *A quadratic irrational number ξ has approximation exponent exactly* 2*.*

Proof. We have previously shown every irrational has $\tau(\xi) \geq 2$; however, by Liouville's Theorem, $\tau(\xi) \leq 2$. Therefore, $\tau(\xi) = 2$. □

We will show later that

Theorem 11.3.7 (Roth 1955**).** *For* any *algebraic number* α*,* $\tau(\alpha) = 2$*.*

Given bounds like $\frac{1}{Mq^e}$, there are two directions we can go. We can try and improve M (Hurwitz), or we can try and improve e (Roth).

The larger e or M, the better our number can be approximated by rationals. As far as approximations go, a rational number is the *simplist* number one can have.

Liouville's Theorem says that algebraic numbers can't be approximated by rationals too quickly; transcendentals are numbers that can be approximated extremely rapidly.

Chapter 12

Liouville's Theorem Constructing Transcendentals

12.1 Review of Approximating by Rationals

Definition 12.1.1 (Approximated by rationals to order n). *A real number x is approximated by rationals to order n if there exist a constant $k(x)$ (possibly depending on x) such that there are infinitely many rational $\frac{p}{q}$ with*

$$\left|x - \frac{p}{q}\right| < \frac{k(x)}{q^n}. \tag{12.1}$$

Recall that Dirichlet's Box Principle gives us:

$$\left|x - \frac{p}{q}\right| < \frac{1}{q^2} \tag{12.2}$$

for infintely many fractions $\frac{p}{q}$. This was proved by choosing a large parameter Q, and considering the $Q+1$ fractionary parts $\{qx\} \in [0,1)$ for $q \in \{0,\ldots,Q\}$. The box principle ensures us that there must be two different q's, say:

$$0 \leq q_1 < q_2 \leq Q \tag{12.3}$$

such that both $\{q_1 x\}$ and $\{q_2 x\}$ belong to the same interval $[\frac{a}{Q}, \frac{a+1}{Q})$, for some $0 \leq a \leq Q-1$. Note that there are exactly Q such intervals partitioning $[0,1)$, and $Q+1$ fractionary parts! Now, the length of such an interval is $\frac{1}{Q}$ so we get

$$|\{q_2 x\} - \{q_1 x\}| < \frac{1}{Q}. \tag{12.4}$$

There exist integers p_1 and p_2 such that

$$\{q_1 x\} = q_1 x - p, \ \{q_2 x\} = q_2 x - p. \tag{12.5}$$

Letting $p = p_2 - p_1$ we find

$$|(q_2 - q_1)x - p| \leq \frac{1}{Q} \tag{12.6}$$

Let $q = q_2 - q_1$, so $1 \leq q \leq Q$, and the previous equation can be rewriten as

$$\left| x - \frac{p}{q} \right| < \frac{1}{qQ} \leq \frac{1}{q^2} \tag{12.7}$$

Now, letting $Q \to \infty$, we get an infinite collection of rational fractions $\frac{p}{q}$ satisfying the above equation. If this collection contains only finitely many distinct fractions, then one of these fractions, say $\frac{p_0}{q_0}$, would occur for infintely many choices Q_k of Q, thus giving us:

$$\left| x - \frac{p_0}{q_0} \right| < \frac{1}{qQ_k} \to 0, \tag{12.8}$$

as $k \to \infty$. This implies that $x = \frac{p_0}{q_0} \in \mathbb{Q}$. So, unles x is a rational number, we can find infinitely many *distinct* rational numbers $\frac{p}{q}$ satisfying Equation 12.7. This means that any real, irrational number can be approximated to order $n = 2$ by rational numbers.

12.2 Liouville's Theorem

Theorem 12.2.1 (Liouville's Theorem). *Let x be a real algebraic number of degree n. Then x is approximated by rationals to order at most n.*

Proof. Let

$$f(X) \ = \ a_n X^n + \cdots a_1 X + a_0 \tag{12.9}$$

be the polynomial with integer coefficients of smallest degree (minimal polynomial) such that x satisfies

$$f(x) \;=\; 0. \tag{12.10}$$

Note that $\deg x = \deg f$ and the condition of minimality implies that $f(X)$ is irreducible over $\mathbb{Z}$. Further, a well known result from algebra states that a polynomial irreducible over $\mathbb{Z}$ is also irreducible over $\mathbb{Q}$.

In particular, as $f(X)$ is irreducible over $\mathbb{Q}$, $f(X)$ does not have any rational roots. If it did, then $f(X)$ would be divisible by a linear polynomial $(X - \frac{a}{b})$. Let $G(X) = \frac{f(X)}{X - \frac{a}{b}}$. Clear denominators (multiply throughout by b), and let $g(X) = bG(X)$. Then $\deg g = \deg f - 1$, and $g(x) = 0$. This contradicts the minimality of f (we choose f to be a polynomial of smallest degree such that $f(x) = 0$). Therefore, f is non-zero at every rational.

Let

$$M \;=\; \sup_{|z-x|<1} |f'(z)|. \tag{12.11}$$

Let now $\frac{p}{q}$ be a rational such that $\left|x - \frac{p}{q}\right| < 1$. The Mean Value Theorem gives us that

$$\left|f\left(\frac{p}{q}\right) - f(x)\right| = \left|f'(c)\left(x - \frac{p}{q}\right)\right| \leq M\left|x - \frac{p}{q}\right| \tag{12.12}$$

where c is some real number between x and $\frac{p}{q}$; $|c - x| < 1$ for $\frac{p}{q}$ moderately close to x.

Now we use the fact that $f(X)$ does not have any rational roots:

$$0 \neq f\left(\frac{p}{q}\right) = a_n\left(\frac{p}{q}\right)^n + \cdots + a_0 = \frac{a_n p^n + \cdots a_1 p^{n-1} q + a_0 q^n}{q^n} \tag{12.13}$$

The numerator of the last term is a nonzero integer, hence it has absolute value at least 1. Since we also know that $f(x) = 0$ it follows that

$$\left|f\left(\frac{p}{q}\right) - f(x)\right| = \left|f\left(\frac{p}{q}\right)\right| = \frac{|a_n p^n + \cdots a_1 p^{n-1} q + a_0 q^n|}{q^n} \geq \frac{1}{q^n}. \tag{12.14}$$

Combining the equations 12.12 and 12.14, we get:

$$\frac{1}{q^n} \leq M\left|x - \frac{p}{q}\right| \;\Rightarrow\; \frac{1}{Mq^n} \leq \left|x - \frac{p}{q}\right| \tag{12.15}$$

whenever $|x - \frac{p}{q}| < 1$. This last equation shows us that x can be approximated by rationals to order at most n. For assume it was otherwise, namely that x can be approximated to order $n + \epsilon$. Then we would have an infinite sequence of distinct rational numbers $\{\frac{p_i}{q_i}\}_{i \geq 1}$ and a constant $k(x)$ depending only on x such that

$$\left| x - \frac{p_i}{q_i} \right| \ < \ \frac{k(x)}{q_i^{n+\epsilon}}. \tag{12.16}$$

Since the numbers $\frac{p_i}{q_i}$ converge to x we can assume that they already are in the interval $(x - 1, x + 1)$. Hence they also satisfy Equation 12.15:

$$\frac{1}{q_i^n} \leq M \left| x - \frac{p_i}{q_i} \right|. \tag{12.17}$$

Combining the last two equations we get

$$\frac{1}{M q_i^n} \leq \left| x - \frac{p_i}{q_i} \right| \ < \ \frac{k(x)}{q_i^{n+\epsilon}}, \tag{12.18}$$

hence

$$q_i^{\epsilon} \ < \ M \tag{12.19}$$

and this is clearly impossible for arbitrarily large q since $\epsilon > 0$ and $q_i \to \infty$. □

Exercise 12.2.2. *Justify the fact that if $\{\frac{p_i}{q_i}\}_{i \geq 1}$ is a rational approximation to order $n \geq 1$ of x, then $q_i \to \infty$.*

Remark 12.2.3. *So far we have seen that the order to which an algebraic number can be approximated by rationals is bounded by its degree. Hence if a real, irrational number $\alpha \notin \mathbb{Q}$ can be approximated by rationals to an arbitrary large order, then α must be transcendental! This provides us with a recipe for constructing transcendental numbers.*

12.3 Constructing Transcendental Numbers

12.3.1 $\sum_m 10^{-m!}$

The following construction of transcendental numbers is due to Liouville.

Theorem 12.3.1. *The number*

$$x \;=\; \sum_{m=1}^{\infty} \frac{1}{10^{m!}} \tag{12.20}$$

is transcendental.

Proof. The series defining x is convergent, since it is dominated by the geometric series $\sum \frac{1}{10^m}$. In fact, the series converges very rapidly and it is this high rate of convergence that will yield x is transcendental.

Fix N large, and let $n > N$. Write

$$\frac{p_n}{q_n} \;=\; \sum_{m=1}^{n} \frac{1}{10^{m!}} \tag{12.21}$$

with $p_n, q_n > 0$ and $(p_n, q_n) = 1$. Then $\{\frac{p_n}{q_n}\}_{n\geq 1}$ is a monotone increasing sequence converging to x. In particular, all these rational numbers are distinct. Not also that q_n must divide $10^{n!}$, which implies

$$q_n \leq 10^{n!}. \tag{12.22}$$

Using this, we get

$$\begin{aligned} 0 < x - \frac{p_n}{q_n} \;&=\; \sum_{m>n} \frac{1}{10^{m!}} = \frac{1}{10^{(n+1)!}} \left(1 + \frac{1}{10^{n+2}} + \frac{1}{10^{(n+2)(n+3)}} + \cdots\right) \\ &<\; \frac{2}{10^{(n+1)!}} = \frac{2}{(10^{n!})^{n+1}} \\ &<\; \frac{2}{q_n^{n+1}} \leq \frac{2}{q_n^N}. \end{aligned} \tag{12.23}$$

This gives an approximation by rationals of order N of x. Since N can be chosen arbitrarily large, this implies that x can be approximated by rationals to arbitrary order. We can conclude, in view of our precious remark 12.2.3 that x is transcendental. □

12.3.2 $[10^{1!}, 10^{2!}, \dots]$

Theorem 12.3.2. *The number*

$$y = [10^{1!}, 10^{2!}, \dots] \tag{12.24}$$

is transcendental.

Proof. Let $\frac{p_n}{q_n}$ be the continued fraction of $[10^{1!} \cdots 10^{n!}]$. Then

$$\begin{aligned} \left| y - \frac{p_n}{q_n} \right| &= \frac{1}{q_n q'_{n+1}} = \frac{1}{q_n(a'_{n+1}q_n + q_{n-1})} \\ &< \frac{1}{a_{n+1}} = \frac{1}{10^{(n+1)!}}. \end{aligned} \tag{12.25}$$

Since $q_k = a_n q_{k-1} + q_{n-2}$, it implies that $q_k > q_{k-1}$ Also, $q_{k+1} = a_{k+1}q_n + q_{k-1}$, so we get

$$\frac{q_{k+1}}{q_k} = a_{k+1} + \frac{q_{k-1}}{q_k} < a_{k+1} + 1. \tag{12.26}$$

Hence writing this inequality for $k = 1, \cdots, n-1$ we obtain

$$\begin{aligned} q_n = q_1 \frac{q_2}{q_1}\frac{q_3}{q_2} \cdots \frac{q_n}{q_{n-1}} &< (a_1+1)(a_2+1)\cdots(a_n+1) \\ &= (1+\frac{1}{a_1})\cdots(1+\frac{1}{a_n})a_1 \cdots a_n \\ &< 2^n a_1 \cdots a_n = 2^n 10^{1!+\cdots+n!} \\ &< 10^{2n!} = a_n^2 \end{aligned} \tag{12.27}$$

Combining equations 12.25 and 12.27 we get:

$$\begin{aligned} \left| y - \frac{p_n}{q_n} \right| &< \frac{1}{a_{n+1}} = \frac{1}{a_n^{n+1}} \\ &< \left(\frac{1}{a_n^2}\right)^{\frac{n}{2}} < \left(\frac{1}{q_n^2}\right)^{\frac{n}{2}} \\ &= \frac{1}{q_n^{n/2}}. \end{aligned} \tag{12.28}$$

In this way we get, just as in the previous theorem, an approximation of y by rationals to arbitrary order. This proves that y is transcendental.

□

12.3.3 Buffon's Needle and π

Consider a collection of infinitely long parallel lines in the plane, where the spacing between any two adjacent lines is d. Let the lines be located at $x = 0, \pm d, \pm 2d, \ldots$. Consider a rod of length l, where for convenience we assume $l < d$.

If we were to *randomly* throw the rod on the plane, what is the probability it hits a line? This question was first asked by Buffon in 1733.

Because of the vertical symmetry, we may assume the center of the rod lies on the line $x = 0$, as shifting the rod (without rotating it) up or down will not alter the number of intersections. By the horizontal symmetry, we may assume $-\frac{d}{2} \le x < \frac{d}{2}$. We posit that all values of x are equally likely. As x is continuous distributed, we may add in $x = \frac{d}{2}$ without changing the probability. The probability density function of x is $\frac{dx}{d}$.

Let θ be the angle the rod makes with the x-axis. As each angle is equally likely, the probability density function of θ is $\frac{d\theta}{2\pi}$.

We assume that x and θ are chosen independently. Thus, the probability density for (x, θ) is $\frac{dxd\theta}{d\cdot 2\pi}$.

The projection of the rod (making an angle of θ with the x-axis) along the x-axis is $l \cdot |\cos\theta|$. If $|x| \le l \cdot |\cos\theta|$, then the rod hits exactly one vertical line exactly once; if $x > l \cdot |\cos\theta|$, the rod does not hit a vertical line. Note that if $l > d$, a rod could hit multiple lines, making the arguments more involved.

Thus, the probability a rod hits a line is

$$\begin{aligned} p &= \int_{\theta=0}^{2\pi} \int_{x=-l\cdot|\cos\theta|}^{l\cdot|\cos\theta|} \frac{dxd\theta}{d\cdot 2\pi} \\ &= \int_{\theta=0}^{2\pi} \frac{l\cdot|\cos\theta|}{d}\frac{d\theta}{2\pi} \\ &= \frac{2l}{\pi d}. \end{aligned} \tag{12.29}$$

Exercise 12.3.3. *Show*

$$\frac{1}{2\pi}\int_0^{2\pi} |\cos\theta| d\theta \;=\; \frac{2}{\pi}. \tag{12.30}$$

Let A be the random variable which is the number of intersections of a rod of length l thrown against parallel vertical lines separated by $d > l$ units. Then

$$A = \begin{cases} 1 & \text{with probability } \frac{2l}{\pi d} \\ 0 & \text{with probability } 1 - \frac{2l}{\pi d} \end{cases}. \tag{12.31}$$

If we were to throw N rods independently, since the expected value of a sum is the sum of the expected values (Lemma 1.4.8), we expect to observe

$$N \cdot \frac{2l}{\pi d} \tag{12.32}$$

intersections.

Turning this around, let us throw N rods, and let I be the number of observed intersections of the rods with the vertical lines. Then

$$I \approx N \cdot \frac{2l}{\pi d} \quad \rightarrow \quad \pi \approx \frac{N}{I} \cdot \frac{2l}{d}. \tag{12.33}$$

The above is an *experimental* formula for π!

Chapter 13

Poissonian Behavior and $\{n^k\alpha\}$

13.1 Equidistribution

We say a sequence of number $x_n \in [0,1)$ is equidistributed if

$$\lim_{N\to\infty} \frac{\#\{n : 1 \le n \le N \text{ and } x_n \in [a,b]\}}{N} \;=\; b-a \tag{13.1}$$

for any subinterval $[a,b]$ of $[0,1]$.

Recall Weyl's Result, Theorem 4.2.10: If $\alpha \notin \mathbb{Q}$, then the fractional parts $\{n\alpha\}$ are equidistributed. Equivalently, $n\alpha$ mod 1 is equidistributed.

Similarly, one can show that for any integer k, $\{n^k\alpha\}$ is equidistributed. See Robert Lipshitz's paper for more details.

13.2 Point Masses and Induced Probability Measures

Recall from physics the concept of a unit point mass located at $x = a$. Such a point mass has no length (or, in higher dimensions, width or height), but finite mass. As mass is the integral of the density over space, a finite mass in zero volume (or zero length on the line) implies an infinite density.

We can make this more precise by the notion of an Approximation to the Identity. See also Theorem 4.2.3.

Definition 13.2.1 (Approximation to the Identity). *A sequence of functions $g_n(x)$ is an approximation to the identity (at the origin) if*

1. $g_n(x) \ge 0$.

2. $\int g_n(x)dx = 1$.

3. *Given $\epsilon, \delta > 0$ there exists $N > 0$ such that for all $n > N$, $\int_{|x|>\delta} g_n(x)dx < \epsilon$.*

We represent the limit of any such family of $g_n(x)$s by $\delta(x)$.

If $f(x)$ is a nice function (say near the origin its Taylor Series converges) then

$$\int f(x)\delta(x)dx \;=\; \lim_{n\to\infty} \int f(x)g_n(x) \;=\; f(0). \tag{13.2}$$

Exercise 13.2.2. *Prove Equation 13.2.*

Thus, in the limit the functions g_n are acting like point masses. We can consider the probability densities $g_n(x)dx$ and $\delta(x)dx$. For $g_n(x)dx$, as $n \to \infty$, almost all the probability is concentrated in a narrower and narrower band about the origin; $\delta(x)dx$ is the limit with all the mass at one point. It is a discrete (as opposed to continuous) probability measure.

Note that $\delta(x-a)$ acts like a point mass; however, instead of having its mass concentrated at the origin, it is now concentrated at a.

Exercise 13.2.3. *Let*

$$g_n(x) \;=\; \begin{cases} n & \text{if } |x| \le \frac{1}{2n} \\ 0 & \text{otherwise} \end{cases} \tag{13.3}$$

Prove $g_n(x)$ is an approximation to the identity at the origin.

Exercise 13.2.4. *Let*

$$g_n(x) \;=\; c\frac{\frac{1}{n}}{\frac{1}{n^2} + x^2}. \tag{13.4}$$

Find c such that the above is an approximation to the identity at the origin.

Given N point masses located at x_1, $x_2, \ldots, x_N$, we can form a probability measure

$$\mu_N(x)dx = \frac{1}{N}\sum_{n=1}^{N} \delta(x - x_n)dx. \tag{13.5}$$

Note $\int \mu_N(x)dx = 1$, and if $f(x)$ is a nice function,

$$\int f(x)\mu_N(x)dx \;=\; \frac{1}{N}\sum_{n=1}^{N} f(x_n). \tag{13.6}$$

Exercise 13.2.5. *Prove Equation 13.6 for nice $f(x)$.*

Note the right hand side of Equation 13.6 looks like a Riemann sum. Or it *would* look like a Riemann sum if the x_ns were equidistributed. In general the x_ns will not be equidistributed, but assume for any interval $[a,b]$ that as $N \to \infty$, the fraction of x_ns ($1 \le n \le N$) in $[a,b]$ goes to $\int_a^b p(x)dx$ for some nice function $p(x)$:

$$\lim_{N\to\infty} \frac{\#\{n : 1 \le n \le N \text{ and } x_n \in [a,b]\}}{N} \to \int_a^b p(x)dx. \tag{13.7}$$

In this case, if $f(x)$ is nice (say twice differentiable, with first derivative uniformly bounded), then

$$\begin{aligned} \int f(x)\mu_N(x)dx &= \frac{1}{N}\sum_{n=1}^{N} f(x_n) \\ &\approx \sum_{k=-\infty}^{\infty} f\Big(\frac{k}{N}\Big)\frac{\#\{n : 1 \le n \le N \text{ and } x_n \in \left[\frac{k}{N}, \frac{k+1}{N}\right]\}}{N} \\ &\to \int f(x)p(x)dx. \end{aligned} \tag{13.8}$$

Definition 13.2.6 (Convergence to $p(x)$). *If the sequence of points x_n satisfies Equation 13.7 for some nice function $p(x)$, we say the probability measures $\mu_N(x)dx$ converge to $p(x)dx$.*

13.3 Neighbor Spacings

We now consider finer questions. Let α_n be a collection of points in $[0,1)$. We order them by size:

$$0 \le \alpha_{\sigma(1)} \le \alpha_{\sigma(2)} \le \cdots \le \alpha_{\sigma(N)}, \tag{13.9}$$

where σ is a permutation of $123\cdots N$. Note the ordering depends crucially on N. Let $\beta_j = \alpha_{\sigma(j)}$.

We consider how the differences $\beta_{j+1} - \beta_j$ are distributed. We will use a slightly different definition of distance, however.

Recall $[0,1)$ is equivalent to the unit circle under the map $x \to e^{2\pi i x}$. Thus, the numbers $.999$ and $.001$ are actually very close; however, if we used the standard definition of distance, then $|.999 - .001| = .998$, which is quite large. Wrapping $[0,1)$ on itself (identifying 0 and 1), we see that $.999$ and $.001$ are separated by $.002$.

Definition 13.3.1 (mod 1 distance). *Let $x, y \in [0,1)$. We define the mod 1 distance from x to y, $||x-y||$, by*

$$||x-y|| \ = \ \min\Big\{|x-y|, \ 1-|x-y|\Big\}. \tag{13.10}$$

Exercise 13.3.2. *Show that the mod 1 distance between any two numbers in $[0,1)$ is at most $\frac{1}{2}$.*

In looking at spacings between the β_js, we have $N-1$ pairs of neighbors:

$$(\beta_2, \beta_1), \ (\beta_3, \beta_2), \ \ldots, \ (\beta_N, \beta_{N-1}). \tag{13.11}$$

These pairs give rise to spacings $\beta_{j+1} - \beta_j \in [0,1)$.

We can also consider the pair (β_1, β_N). This gives rise to the spacing $\beta_1 - \beta_N \in [-1,0)$; however, as we are studying this sequence mod 1, this is equivalent to $\beta_1 - \beta_N + 1 \in [0,1)$.

Henceforth, whenever we perform any arithmetic operation, we always mean mod 1; thus, our answers always live in $[0,1)$

Definition 13.3.3 (Neighbor Spacings). *Given a sequence of numbers α_n in $[0,1)$, fix an N and arrange the numbers α_n ($n \le N$) in increasing order. Label the new sequence β_j; note the ordering will depend on N. Let $\beta_{-j} = \beta_{N-j}$ and $\beta_{N+j} = \beta_j$.*

1. *The nearest neighbor spacings are the numbers $\beta_{j+1} - \beta_j$, $j = 1$ to N.*
2. *The k^{th}-neighbor spacings are the numbers $\beta_{j+k} - \beta_j$, $j = 1$ to N.*

Remember to take the differences $\beta_{j+k} - \beta_j$ mod 1.

Exercise 13.3.4. *Let $\alpha = \sqrt{2}$, and let $\alpha_n = \{n\alpha\}$ or $\{n^2\alpha\}$. Calculate the nearest neighbor and the next-nearest neighbor spacings in each case for $N = 10$.*

Definition 13.3.5 (wrapped unit interval). *We call $[0,1)$, when all arithmetic operations are done mod 1, the wrapped unit interval.*

13.4 Poissonian Behavior

Let $\alpha \notin \mathbb{Q}$. Fix a positive integer k, and let $\alpha_n = \{n^k\alpha\}$. As $N \to \infty$, look at the ordered α_ns, denoted by β_n. How are the nearest neighbor spacings of β_n distributed? How does this depend on k? On α? On N?

Before discussing this problem, we consider a simpler case. Fix N, and consider N independent random variables x_n. Each random variable is chosen from the uniform distribution on $[0, 1)$; thus, the probability that $x_n \in [a, b)$ is $b - a$.

Let y_n be the x_ns arranged in increasing order. How do the neighbor spacings behave?

First, we need to decide what is the correct scale to use for our investigations. As we have N objects on the wrapped unit interval, we have N nearest neighbor spacings. Thus, we expect the average spacing to be $\frac{1}{N}$.

Definition 13.4.1 (Unfolding). *Let $z_n = Ny_n$. The numbers $z_n = Ny_n$ have unit mean spacing. Thus, while we expect the average spacing between adjacent y_ns to be $\frac{1}{N}$ units, we expect the average spacing between adjacent z_ns to be 1 unit.*

So, the probability of observing a spacing as large as $\frac{1}{2}$ between adjacent y_ns becomes negligible as $N \to \infty$. What we should ask is what is the probability of observing a nearest neighbor spacing of adjacent y_ns that is *half* the average spacing. In terms of the z_ns, this will correspond to a spacing between adjacent z_ns of $\frac{1}{2}$ a unit.

13.4.1 Nearest Neighbor Spacings

By symmetry, on the wrapped unit interval the expected nearest neighbor spacing is independent of j. Explicitly, we expect $\beta_{j+1} - \beta_j$ to have the same distribution as $\beta_{i+1} - \beta_i$.

What is the probability that, when we order the x_ns in increasing order, the next x_n after x_1 is located between $\frac{t}{N}$ and $\frac{t+\Delta t}{N}$? Let the x_ns in increasing order be labeled $y_1 \le y_2 \le \cdots \le y_N$, $y_n = x_{\sigma(n)}$.

As we are choosing the x_ns independently, there are $\binom{N-1}{1}$ choices of subscript n such that x_n is nearest to x_1. This can also be seen by symmetry, as each x_n is equally likely to be the first to the *right* of x_1 (where, of course, $.001$ is just a little to the right of $.999$), and we have $N - 1$ choices left for x_n.

The probability that $x_n \in \left[\frac{t}{N}, \frac{t+\Delta t}{N}\right]$ is $\frac{\Delta t}{N}$.

For the remaining $N-2$ of the x_ns, each must be further than $\frac{t+\Delta t}{N}$ from x_n. Thus, they must *all* lie in an interval (or possibly two intervals if we wrap around) of length $1-\frac{t+\Delta t}{N}$. The probability that they all lie in this region is $\left(1-\frac{t+\Delta t}{N}\right)^{N-2}$.

Thus, if $x_1 = y_l$, we want to calculate the probability that $||y_{l+1} - y_l|| \in \left[\frac{t}{N}, \frac{t+\Delta t}{N}\right]$. This is

$$\begin{aligned} \text{Prob}\left(||y_{l+1} - y_l|| \in \left[\frac{t}{N}, \frac{t+\Delta t}{N}\right]\right) &= \binom{N-1}{1} \cdot \frac{\Delta t}{N} \cdot \left(1 - \frac{t+\Delta t}{N}\right)^{N-2} \\ &= \left(1 - \frac{1}{N}\right) \cdot \left(1 - \frac{t+\Delta t}{N}\right)^{N-2} \Delta t. \end{aligned} \tag{13.12}$$

For N enormous and Δt small,

$$\begin{aligned} \left(1 - \frac{1}{N}\right) &\approx 1 \\ \left(1 - \frac{t+\Delta t}{N}\right)^{N-2} &\approx e^{-(t+\Delta t)} \approx e^{-t}. \end{aligned} \tag{13.13}$$

Thus

$$\text{Prob}\left(||y_{l+1} - y_l|| \in \left[\frac{t}{N}, \frac{t+\Delta t}{N}\right]\right) \to e^{-t}\Delta t. \tag{13.14}$$

Remark 13.4.2. *The above argument is infinitesimally wrong. Once we've located y_{l+1}, the remaining x_ns do not need to be more than $\frac{t+\Delta t}{N}$ units to the right of $x_1 = y_l$; they only need to be further to the right than y_{l+1}. As the incremental gain in probabilities for the locations of the remaining x_ns is of order Δt, these contributions will not influence the large N, small Δt limits. Thus, we ignore these effects.*

To rigorously derive the limiting behavior of the nearest neighbor spacings using the above arguments, one would integrate over x_m ranging from $\frac{t}{N}$ to $\frac{t+\Delta t}{N}$, and the remaining events x_n would be in the a segment of length $1 - x_m$. As

$$\left|\left(1 - x_m\right) - \left(1 - \frac{t+\Delta t}{N}\right)\right| \le \frac{\Delta t}{N}, \tag{13.15}$$

this will lead to corrections of higher order in Δt, hence negligible.

We can rigorously avoid this by instead considering the following:

1. Calculate the probability that all the other x_ns are at least $\frac{t}{N}$ units to the right of x_1. This is

$$p_t \;=\; \Big(1-\frac{t}{N}\Big)^{N-1} \;\to\; e^{-t}. \tag{13.16}$$

2. Calculate the probability that all the other x_ns are at least $\frac{t+\Delta t}{N}$ units to the right of x_1. This is

$$p_{t+\Delta t} \;=\; \Big(1-\frac{t+\Delta t}{N}\Big)^{N-1} \;\to\; e^{-(t+\Delta t)}. \tag{13.17}$$

3. The probability that no x_ns are within $\frac{t}{N}$ units to the right of x_1 but at least one x_n is between $\frac{t}{N}$ and $\frac{t+\Delta t}{N}$ units to the right is $p_{t+\Delta t}-p_t$:

$$\begin{aligned} p_t - p_{t+\Delta t} &\to e^{-t}-e^{-(t+\Delta t)} \\ &= e^{-t}\Big(1-e^{-\Delta t}\Big) \\ &= e^{-t}\left(1-1+\Delta t+O\Big((\Delta t)^2\Big)\right. \\ &\to e^{-t}\Delta t. \end{aligned} \tag{13.18}$$

Definition 13.4.3 (Unfolding Spacings). *If $y_{l+1}-y_l \in \left[\frac{t}{N},\frac{t+\Delta t}{N}\right]$, then $N(y_{l+1}-y_l)\in[t,t+\Delta t]$. The new spacings $z_{l+1}-z_l$ have unit mean spacing. Thus, while we expect the average spacing between adjacent y_ns to be $\frac{1}{N}$ units, we expect the average spacing between adjacent z_ns to be 1 unit.*

13.4.2 k^{th} Neighbor Spacings

Similarly, one can easily analyze the distribution of the k^{th} neighbor spacings when each x_n is chosen independently from the uniform distribution on $[0,1)$.

Again, consider $x_1=y_l$. Now we want to calculate the probability that y_{l+k} is between $\frac{t}{N}$ and $\frac{t+\Delta t}{N}$ units to the *right* of y_l.

Therefore, we need exactly $k-1$ of the x_ns to lie between 0 and $\frac{t}{N}$ units to the right of x_1, exactly one x_n (which will be y_{l+k}) to lie between $\frac{t}{N}$ and $\frac{t+\Delta t}{N}$ units to the right of x_1, and the remaining x_ns to lie at least $\frac{t+\Delta t}{N}$ units to the right of y_{l+k}.

Remark 13.4.4. *We face the same problem discussed in Remark 13.4.2; a similar argument will show that ignoring these affects will not alter the limiting behavior. Therefore, we will make these simplifications.*

There are $\binom{N-1}{k-1}$ ways to choose the x_ns that are at most $\frac{t}{N}$ units to the right of x_1; there is then $\binom{(N-1)-(k-1)}{1}$ ways to choose the x_n between $\frac{t}{N}$ and $\frac{t+\Delta t}{N}$ units to the right of x_1.

Thus,

$$\begin{aligned}
&\text{Prob}\left(||y_{l+k} - y_l|| \in \left[\frac{t}{N}, \frac{t+\Delta t}{N}\right]\right) \; = \\
= \quad & \binom{N-1}{k-1}\left(\frac{t}{N}\right)^{k-1} \cdot \binom{(N-1)-(k-1)}{1}\frac{\Delta t}{N} \cdot \left(1 - \frac{t+\Delta t}{N}\right)^{N-(k+1)} \\
= \quad & \frac{(N-1)\cdots(N-1-(k-2))}{N^{k-1}}\frac{(N-1)-(k-1)}{N}\frac{t^{k-1}}{(k-1)!}\left(1 - \frac{t+\Delta t}{N}\right)^{N-(k+1)}\Delta t \\
\to \quad & \frac{t^{k-1}}{(k-1)!}e^{-t}\Delta t. \qquad (13.19)
\end{aligned}$$

Again, one way to avoid the complications is to integrate over x_m ranging from $\frac{t}{N}$ to $\frac{t+\Delta t}{N}$.

Or, similar to before, we can proceed more rigorously as follows:

1. Calculate the probability that exactly $k-1$ of the other x_ns are at most $\frac{t}{N}$ units to the right of x_1, and the remaining $(N-1)-(k-1)$ of the x_ns are at least $\frac{t}{N}$ units to the right of x_1. As there are $\binom{N-1}{k-1}$ ways to choose $k-1$ of the x_ns to be at most $\frac{t}{N}$ units to the right of x_1, this probability is

$$\begin{aligned}
p_t \quad &= \quad \binom{N-1}{k-1}\left(\frac{t}{N}\right)^{k-1}\left(1 - \frac{t}{N}\right)^{(N-1)-(k-1)} \\
&\to \quad \frac{N^{k-1}}{(k-1)!}\frac{t^{k-1}}{N^{k-1}}e^{-t} \\
&\to \quad \frac{t^{k-1}}{k-1!}e^{-t}. \qquad (13.20)
\end{aligned}$$

2. Calculate the probability that exactly $k-1$ of the other x_ns are at most $\frac{t}{N}$ units to the right of x_1, and the remaining $(N-1)-(k-1)$ of the x_ns are at least $\frac{t+\Delta t}{N}$ units to the right of x_1. Similar to the above, this gives

$$\begin{aligned} p_t &= \binom{N-1}{k-1}\Big(\frac{t}{N}\Big)^{k-1}\Big(1-\frac{t+\Delta t}{N}\Big)^{(N-1)-(k-1)} \\ &\to \frac{N^{k-1}}{(k-1)!}\frac{t^{k-1}}{N^{k-1}}e^{-(t+\Delta t)} \\ &\to \frac{t^{k-1}}{(k-1)!}e^{-(t+\Delta t)}. \end{aligned} \tag{13.21}$$

3. The probability that exactly $k-1$ of the x_ns are within $\frac{t}{N}$ units to the right of x_1 and at least one x_n is between $\frac{t}{N}$ and $\frac{t+\Delta t}{N}$ units to the right is $p_{t+\Delta t}-p_t$:

$$p_t - p_{t+\Delta t} \ \to\ \frac{t^{k-1}}{(k-1)!}e^{-t} - \frac{t^{k-1}}{(k-1)!}e^{-(t+\Delta t)} \ \to\ \frac{t^{k-1}}{(k-1)!}e^{-t}\Delta t. \tag{13.22}$$

Note that when $k=1$, we recover the nearest neighbor spacings.

13.5 Induced Probability Measures

We have proven the following:

Theorem 13.5.1. *Consider N independent random variables x_n chosen from the uniform distribution on the wrapped unit interval $[0,1)$. For fixed N, arrange the x_ns in increase order, labeled $y_1 \le y_2 \le \cdots \le y_N$.*

Form the induced probability measure $\mu_{N,1}$ from the nearest neighbor spacings. Then as $N \to \infty$ we have

$$\mu_{N,1}(t)dt \ =\ \frac{1}{N}\sum_{n=1}^{N}\delta\Big(t - N(y_n - y_{n-1})\Big)dt \ \to\ e^{-t}dt. \tag{13.23}$$

Equivalently, using $z_n = Ny_n$:

$$\mu_{N,1}(t)dt \;=\; \frac{1}{N}\sum_{n=1}^{N}\delta\Big(t-(z_n-z_{n-1})\Big)dt \;\to\; e^{-t}dt. \tag{13.24}$$

More generally, form the probability measure from the k^{th} nearest neighbor spacings. Then as $N\to\infty$ we have

$$\mu_{N,k}(t)dt \;=\; \frac{1}{N}\sum_{n=1}^{N}\delta\Big(t-N(y_n-y_{n-k})\Big)dt \;\to\; \frac{t^{k-1}}{(k-1)!}e^{-t}dt. \tag{13.25}$$

Equivalently, using $z_n = Ny_n$:

$$\mu_{N,k}(t)dt \;=\; \frac{1}{N}\sum_{n=1}^{N}\delta\Big(t-(z_n-z_{n-k})\Big)dt \;\to\; \frac{t^{k-1}}{(k-1)!}e^{-t}dt. \tag{13.26}$$

Definition 13.5.2 (Poissonian Behavior). *We say a sequence of points x_n has Poissonian Behavior if in the limit as $N\to\infty$ the induced probability measures $\mu_{N,k}(t)dt$ converge to $\frac{t^{k-1}}{(k-1)!}e^{-t}dt$.*

Exercise 13.5.3. *Let $\alpha\in\mathbb{Q}$, and define $\alpha_n = \{n^m\alpha\}$ for some positive integer m. Show the sequence of points α_n does not have Poissonian Behavior.*

Exercise 13.5.4. *Let $\alpha\notin\mathbb{Q}$, and define $\alpha_n = \{n\alpha\}$. Show the sequence of points α_n does not have Poissonian Behavior. Hint: for each N, show the nearest neighbor spacings take on at most three distinct values (the three values depend on N). As only three values are ever assumed for a fixed N, $\mu_{N,1}(t)dt$ cannot converge to $e^{-t}dt$.*

13.6 Non-Poissonian Behavior

Conjecture 13.6.1. *With probability one (with respect to Lebesgue Measure, see Definition 10.1.1), if $\alpha\notin\mathbb{Q}$, if $\alpha_n = \{n^2\alpha\}$ then the sequence of points α_n is Poissonian.*

There are constructions which show certain irrationals give rise to non-Poissonian behavior.

Theorem 13.6.2. *Let $\alpha \notin \mathbb{Q}$ such that $\left|\alpha - \frac{p_n}{q_n}\right| < \frac{a_n}{q_n^3}$ holds infinitely often, with $a_n \to 0$. Then there exist integers $N_j \to \infty$ such that $\mu_{N_j,1}(t)$ does not converge to $e^{-t}dt$.*

As $a_n \to 0$, eventually $a_n < \frac{1}{10}$ for all n large. Let $N_n = q_n$, where $\frac{p_n}{q_n}$ is a good rational approximation to α:

$$\left|\alpha - \frac{p_n}{q_n}\right| < \frac{a_n}{q_n^3}. \tag{13.27}$$

Remember that all subtractions are performed on the wrapped unit interval. Thus, $||.999 - .001|| = .002$.

We look at $\alpha_k = \{k^2\alpha\}$, $1 \le k \le N_n = q_n$. Let the β_ks be the α_ks arranged in increasing order, and let the γ_ks be the numbers $\{k^2\frac{p_n}{q_n}\}$ arranged in increasing order:

$$\begin{array}{ccccccc} \beta_1 & \le & \beta_2 & \le & \cdots & \le & \beta_N \\ \gamma_1 & \le & \gamma_2 & \le & \cdots & \le & \gamma_N. \end{array} \tag{13.28}$$

13.6.1 Preliminaries

Lemma 13.6.3. *If $\beta_l = \alpha_k = \{k^2\alpha\}$, then $\gamma_l = \{k^2\frac{p_n}{q_n}\}$. Thus, the same permutation orders both the α_ks and the γ_ks.*

Proof. Multiplying both sides of Equation 13.27 by $k^2 \le q_n^2$ yields

$$\left|k^2\alpha - k^2\frac{p_n}{q_n}\right| \;<\; k^2\frac{a_n}{q_n^2} \;\le\; \frac{a_n}{q_n} \;<\; \frac{1}{2q_n}. \tag{13.29}$$

Thus, $k^2\alpha$ and $k^2\frac{p_n}{q_n}$ differ by at most $\frac{1}{2q_n}$. Therefore

$$\left|\left|\left\{k^2\alpha\right\} - \left\{k^2\frac{p_n}{q_n}\right\}\right|\right| \;<\; \frac{1}{2q_n}. \tag{13.30}$$

As the numbers $\{m^2\frac{p_n}{q_n}\}$ all have denominators of size at most $\frac{1}{q_n}$, we see that $\{k^2\frac{p_n}{q_n}\}$ is the closest of the $\{m^2\frac{p_n}{q_n}\}$ to $\{k^2\alpha\}$.

This implies that if $\beta_l = \{k^2\alpha\}$, then $\gamma_l = \{k^2\frac{p_n}{q_n}\}$, completing the proof.

□

Exercise 13.6.4. *Prove the ordering is as claimed. Hint: about each $\beta_l = \{k^2\alpha\}$, the closest number of the form $\{c^2\frac{p_n}{q_n}\}$ is $\{k^2\frac{p_n}{q_n}\}$.*

13.6.2 Proof of Theorem 13.6.2

Exercise 13.6.5. *Assume* $||a-b||, ||c-d|| < \frac{1}{10}$. *Show*

$$||(a-b)-(c-d)|| \ < \ ||a-b|| + ||c-d||. \tag{13.31}$$

Proof of Theorem 13.6.2: We have shown

$$||\beta_l - \gamma_l|| \ < \ \frac{a_n}{q_n}. \tag{13.32}$$

Thus, as $N_n = q_n$:

$$\Big|\Big|N_n(\beta_l - \gamma_l)\Big|\Big| \ < \ a_n, \tag{13.33}$$

and the same result holds with l replaced by $l-1$.
By Exercise 13.6.5,

$$\Big|\Big|N_n(\beta_l - \gamma_l) - N_n(\beta_{l-1} - \gamma_{l-1})\Big|\Big| \ < \ 2a_n. \tag{13.34}$$

Rearranging gives

$$\Big|\Big|N_n(\beta_l - \beta_{l-1}) - N_n(\gamma_l - \gamma_{l-1})\Big|\Big| \ < \ 2a_n. \tag{13.35}$$

As $a_n \rightarrow 0$, this implies the difference between $\Big|\Big|N_n(\beta_l - \beta_{l-1})\Big|\Big|$ and $\Big|\Big|N_n(\gamma_l - \gamma_{l-1})\Big|\Big|$ goes to zero.

The above distance calculations were done mod 1. The actual differences will differ by an integer. Thus,

$$\mu^{\alpha}_{N_n,1}(t)dt \ = \ \frac{1}{N_n}\sum_{l=1}^{N_n}\delta\Big(t - N_n(\beta_l - \beta_{l-1})\Big) \tag{13.36}$$

and

$$\mu^{\frac{p_n}{q_n}}_{N_n,1}(t)dt \ = \ \frac{1}{N_n}\sum_{l=1}^{N_n}\delta\Big(t - N_n(\gamma_l - \gamma_{l-1})\Big) \tag{13.37}$$

are extremely close to one another; each point mass from the difference between adjacent β_ls is either within a_n units of a point mass from the difference between adjacent γ_ls, or is within a_n units of a point mass an integer number of units from a point mass from the difference between adjacent γ_ls. Further, $a_n \rightarrow 0$.

Note, however, that if $\gamma_l = \{k^2 \frac{p_n}{q_n}\}$, then

$$N_n \gamma_l \ = \ q_n \Big\{ k^2 \frac{p_n}{q_n} \Big\} \ \in \ \mathbb{N}. \tag{13.38}$$

Thus, the induced probability measure $\mu^{\frac{p_n}{q_n}}_{N_n,1}(t)dt$ formed from the γ_ls is supported on the integers! Thus, it is impossible for $\mu^{\frac{p_n}{q_n}}_{N_n,1}(t)dt$ to converge to $e^{-t}dt$.

As $\mu^{\alpha}_{N_n,1}(t)dt$, modulo some possible integer shifts, is arbitrarily close to $\mu^{\frac{p_n}{q_n}}_{N_n,1}(t)dt$, the sequence $\{k^2\alpha\}$ is *not* Poissonian along the subsequence of Ns given by N_n, where $N_n = q_n$, q_n is a denominator in a good rational approximation to α. $\square$

13.6.3 Measure of $\alpha \notin \mathbb{Q}$ with Non-Poissonian Behavior along a sequence N_n

What is the (Lebesgue) measure of $\alpha \notin \mathbb{Q}$ such that there are infinitely many n with

$$\Big| \alpha - \frac{p_n}{q_n} \Big| \ < \ \frac{a_n}{q_n^3}, \ a_n \to 0. \tag{13.39}$$

If the above holds, then for any constant $k(\alpha)$, for n large (large depends on both α and $k(\alpha)$) we have

$$\Big| \alpha - \frac{p_n}{q_n} \Big| \ < \ \frac{k(\alpha)}{q_n^{2+\epsilon}}. \tag{13.40}$$

By Theorem 10.2.1, this set has (Lebesgue) measure or size 0. Thus, almost no irrational numbers satisfy the conditions of Theorem 13.6.2, where *almost no* is relative to the (Lebesgue) measure.

Exercise 13.6.6. *In a topological sense, how many algebraic numbers satisfy the conditions of Theorem 13.6.2? How many transcendental numbers satisfy the conditions?*

Exercise 13.6.7. *Let α satisfy the conditions of Theorem 13.6.2. Consider the sequence N_n, where $N_n = q_n$, q_n the denominator of a good approximation to α. We know the induced probability measures $\mu^{\frac{p_n}{q_n}}_{N_n,1}(t)dt$ and $\mu^{\alpha}_{N_n,1}(t)dt$ do not converge to $e^{-t}dt$. Do these measures converge to anything?*

Remark 13.6.8. *In [RSZ] it is shown that for most α satisfying the conditions of Theorem 13.6.2, there* is *a sequence N_j along which $\mu^{\alpha}_{N_n,1}(t)dt$* does *converge to $e^{-t}dt$.*

Chapter 14

Sixth Lecture: (The Start of the) Proof of Roth's Theorem

14.1 Statement of Roth's Theorem

Theorem 14.1.1 (Roth's Theorem). *Let α be a real algebraic number (a root of a polynomial equation with integer coefficients). Then, given any $\epsilon > 0$, there are only finitely many relatively prime pairs of integers (p, q) such that*

$$\left|\alpha - \frac{p}{q}\right| < \frac{1}{q^{2+\epsilon}}; \tag{14.1}$$

however, there are infinitely many pairs of relatively prime integers such that

$$\left|\alpha - \frac{p}{q}\right| < \frac{1}{q^2}. \tag{14.2}$$

Remark 14.1.2. *Note that by replacing ϵ with 2ϵ we can rewrite the above as: there exists a $c(\alpha)$ such that*

$$\left|\alpha - \frac{p}{q}\right| < \frac{c(\alpha)}{q^2} \tag{14.3}$$

for infinitely many $\frac{p}{q}$.

Exercise 14.1.3. *Prove the above remark.*

14.1.1 Application of Roth's Theorem to Solving Diophantine Equations

As an application to investigating solutions of Diophantine equations, we prove

Lemma 14.1.4. *There are only finitely many integer solutions $(x, y) \in \mathbb{Z}^2$ to*

$$x^3 - 2y^3 = a. \tag{14.4}$$

Proof. Let $\rho = e^{2\pi i/3} = (-1)^{1/3} = -\frac{1}{2} + i\frac{\sqrt{3}}{2}$. Then

$$x^3 - 2y^3 = (x - 2^{1/3}y)(x - \rho 2^{1/3}y)(x - \rho^2 2^{1/3}y), \tag{14.5}$$

and therefore

$$\begin{aligned} \left|\frac{a}{y^3}\right| &= \left|\frac{x}{y} - 2^{1/3}\right| \left|\frac{x}{y} - \rho 2^{1/3}\right| \left|\frac{x}{y} - \rho^2 2^{1/3}\right| \\ &\geq \left|\frac{x}{y} - 2^{1/3}\right| \left|\mathrm{Im}(\rho 2^{1/3})\right| \left|\mathrm{Im}(\rho^2 2^{1/3})\right| \\ &= \frac{3}{2^{4/3}} \left|\frac{x}{y} - 2^{1/3}\right|. \end{aligned} \tag{14.6}$$

Hence every solution (x, y) to $x^3 - 2y^3 = a$ is a solution to

$$\left|2^{1/3} - \frac{x}{y}\right| \leq \frac{3 \cdot 2^{-4/3}}{|y|^3}. \tag{14.7}$$

By Roth's Theorem there are only finitely many such solutions. □

Note Liouville's Theorem is *not* strong enough to allow us to conclude there are only finitely many integer solutions. As $2^{1/2}$ is an algebraic number of degree 3, Liouville's Theorem says $2^{1/3}$ is approximable by rationals to at most order 3. Thus, the possibility that $2^{1/3}$ *is* approximable by rationals to order 3 is *not* ruled out by Liouville's Theorem.

14.1.2 abc Conjecture and Roth's Theorem

Conjecture 14.1.5 (abc Conjecture). *Let $\epsilon > 0$. If a, b and c are coprime integers, then*

$$c \leq k_\epsilon \Big(\prod_{p|abc} p \Big)^{1+\epsilon}. \tag{14.8}$$

The abc conjecture implies many great results in mathematics: Fermat's Last Theorem follows in a line, Roth's Theorem in a page, Mordell's conjecture in two pages.

What makes the abc conjecture so difficult to test is that the constant $k = k_\epsilon$; so, if you find a counter-example, just bump up ϵ!

We *will* prove Roth's theorem (which does imply that abc *may* be true).

For more information on the relation of abc to Fermat's Last Theorem and Roth's Theorem, see [GT].

14.2 Review of Liouville's Theorem

Given a real algebraic number α of degree $d > 1$, Liouville showed that, except for finitely many $\frac{p}{q}$,

$$\left|\alpha - \frac{p}{q}\right| \ \geq \ \frac{1}{q^{d+\epsilon}}. \tag{14.9}$$

In other words, a real algebraic number of degree d is approximatable by rationals to at most order d. We quickly recall the proof of Liouville's Theorem; we will, however, slightly change the mechanics of the proof to emphasize concepts which will be crucial for proving Roth's Theorem.

1. We first construct $f(x)$, the minimal irreducible polynomial of α. By construction, $f(\alpha) = 0$.

2. We then prove this polynomial vanishes at $\frac{p}{q}$; ie, $f\left(\frac{p}{q}\right) = 0$.

Exercise 14.2.1. *Show $f\left(\frac{p}{q}\right) = 0$ contradicts the irreducibility of $f(x)$.*

How do we prove f vanishes at $\frac{p}{q}$?
Let

$$f(x) \ = \ a_d x^d + a_{d-1}x^{d-1} + \cdots + a_0, \ a_i \in \mathbb{Z}. \tag{14.10}$$

Substituting gives

$$f\left(\frac{p}{q}\right) \ = \ \frac{N}{q^d}, \ N \in \mathbb{Z}. \tag{14.11}$$

We find an upper bound for $f(\frac{p}{q})$ by using the Taylor Expansion about the point $x = \alpha$. As $f(\alpha) = 0$, there is no constant term in the Taylor Expansions. We may assume $\frac{p}{q}$ satisfied $|\alpha - \frac{p}{q}| < 1$.

$$\begin{aligned} f(x) &= \sum_{i=1}^{d} \frac{1}{i!} \frac{\mathrm{d}^i f}{\mathrm{dx}^i}(\alpha) \cdot (x-\alpha)^i \\ \left| f\left(\frac{p}{q}\right) \right| = \left| \frac{N}{q^d} \right| &\leq \left| \frac{p}{q} - \alpha \right| \cdot \sum_{i=1}^{d} \left| \frac{1}{i!} \frac{\mathrm{d}^i f}{\mathrm{dx}^i}(\alpha) \right| \cdot \left| \frac{p}{q} - \alpha \right|^{i-1} \\ &\leq \left| \frac{p}{q} - \alpha \right| \cdot d \cdot \max_i \left| \frac{1}{i!} \frac{\mathrm{d}^i f}{\mathrm{dx}^i}(\alpha) \cdot 1^{i-1} \right| \\ &\leq \left| \frac{p}{q} - \alpha \right| \cdot A(\alpha), \end{aligned} \tag{14.12}$$

where $A(\alpha) = \max_i \left| \frac{1}{i!} \frac{\mathrm{d}^i f}{\mathrm{dx}^i}(\alpha) \right|$.

If α were approximable by rationals to order $d + \epsilon$, then there would exist a constant $B(\alpha)$ and infinitely many $\frac{p}{q}$ such that

$$\left| \frac{p}{q} - \alpha \right| \leq \frac{B(\alpha)}{q^{d+\epsilon}}. \tag{14.13}$$

Combining yields

$$\left| f\left(\frac{p}{q}\right) \right| \leq \frac{A(\alpha)B(\alpha)}{q^{d+\epsilon}}. \tag{14.14}$$

If q is large, $\frac{A(\alpha)B(\alpha)}{q^{\epsilon}} < 1$.
Therefore

$$|N| \leq \frac{A(\alpha)B(\alpha)}{q^{\epsilon}}. \tag{14.15}$$

As we may take q arbitrarily large (because we are assuming α is approximable by rationals to order $d + \epsilon$), this implies $N = 0$. Thus, $f\left(\frac{p}{q}\right) = 0$. $\square$

14.3 Generalizing Liouville's Construction to get Roth's Theorem

We will need to construct a generalization of the polynomial f in order to prove Roth's Theorem.

Suppose $\beta_i = \frac{r_i}{s_i}$ satisfy the conditions of Roth's Theorem for $i = 1$ to m.

Definition 14.3.1 (height). *Assume p and q are relatively prime. We define the height of the fraction $\frac{p}{q}$ to be*

$$H\Big(\frac{p}{q}\Big) \;=\; \max\Big(|p|, |q|\Big). \tag{14.16}$$

We number the β_is such that

$$H(\beta_1) \le H(\beta_2) \le \cdots \le H(\beta_m). \tag{14.17}$$

We can make $H(\beta_1)$ large, and $H(\beta_i)$ much larger than $H(\beta_{i-1})$, as we are assuming we have infinitely many solutions giving good approximations. Thus, we can find infinitely many rationals $\beta_n = \frac{p_n}{q_n}$ with $q_n \to \infty$. Thus, $H(\beta_n) \to \infty$.

We will construct a polynomial $f(X_1, \ldots, X_m)$ with $f(\alpha, \ldots, \alpha) = 0$ and $f(\beta_1, \ldots, \beta_m) = 0$, and get a contradiction from this.

A polynomial of one variable cannot have infinitely many zeros – this is where we obtained the contradiction in Liouville's Theorem.

Thinks are very different in higher dimensions.

Exercise 14.3.2. *Consider the irreducible polynomial of four variables $xy - zw$. This polynomial is zero infinitely often; moreover, it is zero for infinitely many rationals. Consider, for example, $(0, m, 0, n)$, $m, n \in \mathbb{Q}$.*

14.4 Equivalent Formulation of Roth's Theorem

Definition 14.4.1 ($Ineq_*(\alpha, \beta)$). *Let*

$$\begin{array}{rl} \alpha & \text{be a real algebraic number} \\ \beta & \text{be a rational number with } |\beta - \alpha| \le 1 \\ H(\beta) & \text{be the height of } \beta. \end{array} \tag{14.18}$$

If α and β satisfy

$$|\alpha - \beta| \ \leq \ \frac{1}{H(\beta)^{2+\epsilon}} \tag{14.19}$$

we write $Ineq_*(\alpha, \beta)$.

Remark 14.4.2. *Note that by replacing ϵ with 2ϵ we can rewrite the above as: there exists a $c(\alpha)$ such that*

$$|\alpha - \beta| \ \leq \ \frac{c(\alpha)}{H(\beta)^{2+\epsilon}} \tag{14.20}$$

Exercise 14.4.3. *Prove the above remark.*

Lemma 14.4.4. *Given $\alpha \notin \mathbb{Q}$, $Ineq_*(\alpha, \beta)$ has finitely many solutions if and only if Roth's Theorem is true.*

Proof: Assume Roth's Theorem is true for $\alpha \notin \mathbb{Q}$:

$$\left|\alpha - \frac{p}{q}\right| \quad \leq \quad \frac{1}{q^{2+\epsilon}}. \tag{14.21}$$

Then as

$$|p| \ \leq \ |q| \cdot \Big(|\alpha| + 1\Big), \tag{14.22}$$

we have

$$\begin{aligned} \left|\alpha - \frac{p}{q}\right| \quad &\leq \quad \frac{1}{q^{2+\epsilon}} \\ &\leq \quad \frac{(|\alpha|+1)^{2+\epsilon}}{(|\alpha|+1)^{2+\epsilon} q^{2+\epsilon}} \\ &\leq \quad \frac{(|\alpha|+1)^{2+\epsilon}}{|p|^{2+\epsilon}}. \end{aligned} \tag{14.23}$$

As $\left|\alpha - \frac{p}{q}\right| < \frac{1}{q^{2+\epsilon}}$, combining the last two inequalities gives

$$\begin{aligned}\left|\alpha - \frac{p}{q}\right| &\leq \frac{(|\alpha|+1)^{2+\epsilon}}{\max(|p|,|q|)^{2+\epsilon}} \\ &= \frac{(|\alpha|+1)^{2+\epsilon}}{H\left(\frac{p}{q}\right)}. \end{aligned} \tag{14.24}$$

Thus, if Roth's Theorem holds for α, $Ineq_*(\alpha,\beta)$ holds infinitely often.

Conversely, assume $Ineq_*(\alpha,\beta)$ holds infinitely often. Let $\beta = \frac{p}{q}$. If $q > |p|$, then $H(\beta) = q$, and the result of Roth's Theorem immediately follows for this β.

Therefore, it is enough to consider the case when $Ineq_*(\alpha,\beta)$ holds and $|p| > q$. Thus

$$\left|\alpha - \frac{p}{q}\right| \leq \frac{1}{|p|^{2+\epsilon}}. \tag{14.25}$$

As $|p| > q$, this yields

$$|p| \leq |q| \cdot \Big(|\alpha| + 1\Big), \tag{14.26}$$

which implies

$$\frac{1}{H(\beta)} \geq \frac{1}{(|\alpha|+1)q}. \tag{14.27}$$

Therefore

$$\left|\alpha - \frac{p}{q}\right| \leq \frac{1}{H(\beta)^{2+\epsilon}} \leq \frac{1}{(|\alpha|+1)^{2+\epsilon}} \frac{1}{q^{2+\epsilon}}. \tag{14.28}$$

By Remark 14.1.2 we are done. □

14.5 Algebraic Numbers and Integers

Definition 14.5.1 (algebraic numbers and integers). *An algebraic integer α is a number $\alpha \in \mathbb{C}$ which satisfies*

$$a_d \alpha^d + \cdots + a_0 = 0,\ a_i \in Z, d \in \mathbb{N}. \tag{14.29}$$

An algebraic integer is an algebraic number whose equation has $a_d = 1$.

Exercise 14.5.2. *Find $a \in \mathbb{Z}$ such that $\alpha = \frac{1+\sqrt{5}}{2}$ satisfies $ax^2 - x - 1$. Is α an algebraic integer? Is $\frac{1}{\alpha}$ an algebraic number or integer?*

Lemma 14.5.3 (Reduction Lemma). *If $Ineq_*(\alpha, \beta)$ has finitely many solutions for all algebraic integers α, then it will have finitely many solutions for all algebraic numbers.*

Proof: Suppose α is an algebraic number, and suppose Roth's theorem fails for α. We will find an algebraic integer for which Roth's theorem fails.

Our assumption means there are infinitely many β such that

$$|\beta - \alpha| \ \leq \ \frac{1}{H(\beta)^{2+\epsilon}}. \tag{14.30}$$

We find a D such that $D\alpha$ is an algebraic integer. We have

$$a_d\alpha^d + a_{d-1}\alpha^{d-1} + a_{d-2}\alpha^{d-2} + \cdots + a_0 \ = \ 0. \tag{14.31}$$

Multiply the above equation by a_d^{d-1}. Regrouping we find

$$(a_d\alpha)^2 + a_{d-1}(a_d\alpha)^{d-1} + a_{d-2}a_d(a_d\alpha)^{d-2} + \cdots + a_0a_d^{d-1} \ = \ 0. \tag{14.32}$$

Hence, $a_d\alpha$ satisfies the equation

$$X^d + a_{d-1}X^{d-1} + a_{d-2}a_dX^{d-2} + \cdots + a_0a_d^{d-1} \ = \ 0. \tag{14.33}$$

As all the coefficients are integers, we find $a_d\alpha$ is an algebraic integer. Note $D = a_d \in \mathbb{Z}$.

Choose β such that $H(\beta) > H(D)^{1+\frac{6}{\epsilon}}$. We can do this as we are assuming there are infinitely many β_n giving good rational approximations.

Exercise 14.5.4. *Prove $H(D\beta) \leq H(D)H(\beta)$.*

We are assuming Roth's Theorem fails for the algebraic number α; we now show Roth's Theorem fails for the algebraic integer $D\alpha = a_d\alpha$. We have

$$
\begin{aligned}
|D\alpha - D\beta| &= |D| \cdot |\alpha - \beta| \\
&\leq \frac{|D|}{H(\beta)^{2+\epsilon}} \\
&\leq \frac{H(D)}{H(\beta)^{2+\epsilon}} \\
&= \frac{H(D)}{H(\beta)^{2+\frac{\epsilon}{2}}} \cdot \frac{1}{H(\beta)^{\frac{\epsilon}{2}}} \\
&\leq \frac{H(D)}{\left(\frac{H(D\beta)}{H(D)}\right)^{2+\frac{\epsilon}{2}}} \cdot \frac{1}{\left(H(D)^{1+\frac{6}{\epsilon}}\right)^{\frac{\epsilon}{2}}} \\
&= \frac{1}{H(D\beta)^{2+\frac{\epsilon}{2}}}. \qquad (14.34)
\end{aligned}
$$

We have shown that, if Roth's Theorem fails for an algebraic number, than it also fails for an algebraic integer. □

14.6 Needed Preliminaries

Definition 14.6.1. *$P(X_1, \ldots, X_m) \in k[X_1, \ldots, X_m]$ means P is a polynomial in m variables with coefficients in k. We will often take $k = \mathbb{Z}$ or $\mathbb{R}$.*

Definition 14.6.2. *Let $P(X_1, \ldots, X_m) \in \mathbb{R}[X_1, \ldots, X_m]$. Let $(i_1, \ldots, i_m)$ be an m-tuple of non-negative integers. Define*

$$
\begin{aligned}
|P| &= \textit{maximum absolute value of coefficients of } P \\
\partial_{i_1,\ldots,i_m} P &= \frac{1}{i_1! \cdots i_m!} \frac{\partial^{i_1+\cdots+i_m}}{\partial x_1^{i_1} \cdots \partial x_m^{i_m}} P \qquad (14.35)
\end{aligned}
$$

Lemma 14.6.3. *Let $P \in \mathbb{Z}[X_1, \ldots, X_m]$. Then*

1. *$\partial_{i_1,\ldots,i_m} P \in \mathbb{Z}[X_1, \ldots, X_m]$.*
2. *If $\deg_{X_h} P \leq r_h$ for $1 \leq h \leq m$, then $|\partial_{i_1,\ldots,i_m} P| \leq 2^{r_1+\cdots+r_m}|P|$.*

Sketch of Proof: $\partial_i X^j = \binom{j}{i} X^{j-i}$, and $\binom{j}{i} \in \mathbb{N}$ for $i \leq j$ if $i, j \in \mathbb{N}$. This gives the first statement.

For the second, recall the coefficients in the expansion of $(x+y)^j$ are $\binom{j}{i}$. Thus, taking $x = y = 1$, we have $2^j \geq \binom{j}{i}$. This will give the second statement.

Definition 14.6.4. *Let*

1. $P(X_1, \ldots, X_m) \in k[X_1, \ldots, X_m]$,
2. $(\alpha_1, \ldots, \alpha_m) \in k^m$ *be a point,*
3. $r_1, \ldots, r_m$ *are positive integers.*

Then $Index(P) = Ind(P)$ *with respect to* $(\alpha_1, \ldots, \alpha_m)$ *and* $r_1, \ldots, r_m$ *is the smallest value of* $\frac{i_1}{r_1} + \cdots + \frac{i_m}{r_m}$ *such that* $\partial_{i_1,\ldots,i_m} P(\alpha_1, \ldots, \alpha_m) \neq 0$.

If P *is the zero polynomial, we define* $Ind(P) = \infty$.

Lemma 14.6.5. $Ind(P) \geq 0$, *and equals zero when* $P(\alpha_1, \ldots, \alpha_m) \neq 0$.

Exercise 14.6.6. *Prove the above lemma.*

Lemma 14.6.7. *Let* P, P' *be two polynomials in* $k[X_1, \ldots, X_m]$. *Given positive integers* $r_1, \ldots, r_m$ *and a point* $(\alpha_1, \ldots, \alpha_m) \in k^m$, *then*

1. $Ind(\partial_{i_1,\ldots,i_m} P) \geq Ind(P) - \left(\frac{1_1}{r_1} + \cdots + \frac{i_m}{r_m}\right)$.
2. $Ind(P + P') \geq \min(Ind(P), Ind(P'))$.
3. $Ind(PP') = Ind(P) + Ind(P')$.

Chapter 15

Seventh Lecture: The Proof of Roth's Theorem

15.1 Wronskian

15.1.1 Standard Wronskian

Let $\phi_0(x), \ldots, \phi_{l-1}(x)$ be a collection of polynomials. Consider the determinant of the matrix

$$\left(\frac{1}{\mu!} \frac{\mathrm{d}^\mu}{\mathrm{dx}^\mu} \phi_\nu(x) \right)_{\mu,\nu=0,\ldots,l-1} \tag{15.1}$$

Thus, for $l-1=2$ we have

$$\begin{vmatrix} \phi_0(x) & \phi_1(x) & \phi_2(x) \\ \phi_0'(x) & \phi_1'(x) & \phi_2'(x) \\ \frac{1}{2!}\phi_0''(x) & \frac{1}{2!}\phi_1''(x) & \frac{1}{2!}\phi_2''(x) \end{vmatrix} \tag{15.2}$$

Lemma 15.1.1. *Let $\phi_0(x), \ldots, \phi_{l-1}(x)$ have rational coefficients. Then they are linearly independent if and only if the Wronskian is non-zero.*

Proof: see a course on differential equations.

15.1.2 Definition of Generalized Wronskian

We generalized the Wronskian to polynomials in more than one variable. Consider

$$\Delta \;=\; \frac{1}{i_1!\cdots i_p!}\Big(\frac{\partial}{\partial x_1}\Big)^{i_1}\cdots\Big(\frac{\partial}{\partial x_p}\Big)^{i_p}, \tag{15.3}$$

with $i_1 + \cdots i_p$ the order.

Let $\phi_0(x_1,\ldots,x_p),\ldots,\phi_{l-1}(x_1,\ldots,x_p)$ be l polynomials in p variables. We can have differential operators Δs (such as the Δ above) act on our ϕ_νs.

By Δ_ν we mean some operator with tuple $(i_1,\ldots,i_p)$ such that $i_1+\cdots i_p=\nu$. Thus, the order of $\Delta_\nu \le \nu$. Often we do not care which tuple $(i_1,\ldots,i_p)$ we have; we often only care about the order, which is at most $i_1+\cdots+i_p$. For convenience, we will write Δ_ν for such an operator, although $\Delta_{i_1,\ldots,i_p}$ would be more accurate.

Definition 15.1.2 (Generalized Wronskian). *All quantities as above,*

$$G(x_1,\ldots,x_p) \;=\; \det\Big(\Delta_\mu\phi_\nu(x_1,\ldots,x_p)\Big)_{\mu,\nu=0,\ldots,l-1}. \tag{15.4}$$

Exercise 15.1.3. *Verify that if $\phi_0,\ldots,\phi_{l-1}$ are linearly dependent over $\mathbb{Q}$, then all of their Wronskians are identically zero.*

15.1.3 Properties of the Generalized Wronskian

Lemma 15.1.4. *If $\phi_0,\ldots,\phi_{l-1}$ are linearly independent, then some Wronskian is non-zero.*

Proof. Let $k\in\mathbb{N}$ be larger than all the exponents of the individual x_is, and consider the following l one-variable polynomials

$$P_\nu(t) \;=\; \phi_\nu\Big(t,t^k,t^{k^2},\ldots,t^{k^{p-1}}\Big). \tag{15.5}$$

The polynomials $P_\nu(t)$ are linearly independent. Why?

$$\phi_\nu(x_1,\ldots,x_p) \;=\; \sum_{s_1=0}^{k-1}\cdots\sum_{s_p=0}^{k-1} b_\nu(s_1,\ldots,s_p)x_1^{s_1}\cdots x_p^{s_p}. \tag{15.6}$$

Then if

$$\sum_{\nu=0}^{l-1} c_\nu P_\nu(t) \tag{15.7}$$

is identically zero, we get

$$\sum_{\nu=0}^{l-1} c_\nu \sum_{s_1=0}^{k-1} \cdots \sum_{s_p=0}^{k-1} b_\nu(s_1 \cdots s_p) t^{s_1+ks_2+\cdots k^{p-1}s_p} \tag{15.8}$$

is identically zero.

As the $s_i \in \{0, 1, \ldots, k-1\}$, for any integer m there is only one way to write m as $s_1 + ks_2 + \cdots + k^{p-1}s_p$. Therefore,

$$\sum_{\nu=0}^{k-1} c_\nu \phi_\nu(x_1, \ldots, x_p) = 0, \tag{15.9}$$

a contradiction.

□

Hence, $P_0(t), \ldots, P_{l-1}(t)$ are linearly independent. Hence the *standard Wronskian* will be non-zero. The Wronskian is

$$W(t) \;=\; \det\Big(\frac{1}{\mu!}\frac{\mathrm{d}^\mu}{\mathrm{dt}^\mu}\phi_\nu(t, t^k, \ldots, t^{k^{p-1}})\Big)_{0\le\mu,\nu\le l-1}. \tag{15.10}$$

Now

$$\frac{\mathrm{d}}{\mathrm{dt}}P_\nu \;=\; \frac{\partial}{\partial x_1}\phi_\nu\Big|_{x_1=t} + kt^{k-1}\frac{\partial}{\partial x_2}\phi_\nu\Big|_{x_2=t^k} + \cdots + k^{p-1}t^{k^{p-1}-1}\frac{\partial}{\partial x_p}\phi_\nu\Big|_{x_p=t^{k^{p-1}}}. \tag{15.11}$$

Thus,

$$\frac{\mathrm{d}^\mu}{\mathrm{dt}^\mu} \tag{15.12}$$

is a linear combination of stuff, ie, there will be lots of Δs, and the order of any Δ appearing in $\frac{\mathrm{d}^\mu}{\mathrm{dt}^\mu}$ will *not* exceed μ.

Example:

$$\begin{pmatrix} \phi_0 & \phi_1 \\ \phi_0' & \phi_1' \end{pmatrix} = \begin{pmatrix} \phi_0 & \phi_1 \\ \Delta_1\phi_0 + \Delta_2\phi_0 & \Delta_1\phi_1 + \Delta_2\phi_1 \end{pmatrix} \tag{15.13}$$

Then this equals

$$\begin{pmatrix} \phi_0 & \phi_1 \\ \Delta_1\phi_0 & \Delta_1\phi_1 \end{pmatrix}\Big|_{x_1=t, x_2=t^k} + \begin{pmatrix} \phi_0 & \phi_1 \\ \Delta_2\phi_0 & \Delta_2\phi_1 \end{pmatrix}\Big|_{x_1=t, x_2=t^k} \tag{15.14}$$

15.2 More Properties

Lemma 15.2.1. *Let $R(x_1, \ldots, x_p)$ be a polynomial in $p \geq 2$ variables with integral coefficients, R not identically zero. Let R be of degree at most r_j in the variable x_j, $1 \leq j \leq p$. Then there exists an integer l satisfying $1 \leq l \leq r_p + 1$ and differential operators $\Delta_0, \ldots, \Delta_{l-1}$ on $x_1, \ldots, x_{p-1}$ (with order of Δ_ν at most ν) such that if*

$$F(x_1, \ldots, x_p) \;=\; \det\left(\Delta_\mu \frac{1}{\nu!}\frac{\partial^\nu}{\partial x_p^\nu} R\right)_{0 \leq \mu, \nu \leq l-1}, \tag{15.15}$$

Then

1. *F has integral coefficients and F is not identically zero.*
2. *We have $F(x_1, \ldots, x_p) = u(x_1, \ldots, x_p)v(x_p)$, u and v have integral coefficients with the degree of u at most lr_j ($1 \leq j \leq p-1$) and v i of degree at most lr_p.*

Proof. Consider all representations of R in the form

$$R(x_1, \ldots, x_p) \;=\; \phi_0(x_p)\psi_0(x_1, \ldots, x_{p-1}) + \cdots + \phi_{l-1}(x_p)\psi_{l-1}(x_1, \ldots, x_{p-1}) \tag{15.16}$$

in such a way that each ϕ_i, ψ_i has rational coefficients and each $\phi_n u$ is of degree at most r_p and ψ_ν is of degree at most r_j in each x_j for $1 \leq j \leq p-1$. Such a representation is possible. Collect common powers of x_p and factor out. Consider $\phi_\nu(x_p) = x_p^\nu$. In this case, $l = r_p + 1$.

Choose the smallest l where we have such a representation.

Claim 15.2.2. *$\phi_0(x_p), \ldots, \phi_{l-1}(x_p)$ are linearly independent.*

Assume the last can be written in terms of the others:

$$\phi_{l-1}(x_p) \;=\; d_0\phi_0(x_p) + \cdots + d_{l-2}\phi_{l-2}(x_p). \tag{15.17}$$

Then

$$R \;=\; \phi_0(\psi_0 + d_0\psi_1) + \cdots + \phi_{l-2}(\psi_{l-2} + d_{l-2}\psi_{l-1}). \tag{15.18}$$

Similar arguments yield all are linearly independent.

Let $W(x_p)$ be the Wronskian of $\phi_0(x_p), \ldots, \phi_{l-1}(x_p)$. This is non-zero, and has rational coefficients.

Let $G(x_1, \ldots, x_{p-1})$ be some non-vanishing generalized Wronskian of $\psi_0, \ldots, \psi_{l-1}$. This is

$$\det\Big(\Delta_\mu \psi_n u(x_1, \ldots, x_{p-1})\Big). \tag{15.19}$$

Then

$$\begin{aligned} G(x_1, \ldots, x_{p-1})W(x_p) &= \det\Big(\sum_{\rho=0}^{l-1} \Delta_\mu \frac{1}{\nu!}\frac{\partial^\nu}{\partial x_p^\nu}\phi_\rho(x_p)\psi_\rho(x_1, \ldots, x_{p-1})\Big) \\ &= \det\Big(\Delta_\mu \frac{1}{\nu!}\frac{\partial^\nu}{\partial x_p^\nu} R\Big) \not\equiv 0. \end{aligned} \tag{15.20}$$

Since $G(x_1, \ldots, x_{p-1})W(x_p)$ is integral, $\exists q \in \mathbb{Q}^\times$ such that gG and $g^{-1}W$ are integral – take out common denominators and multiply through. Let one of them be u and the other v.

The assertion about the degrees of u and v follow by direct calculation.

□

Lemma 15.2.3. *R as above, and suppose all the coefficients of R have absolute value bounded by B. Then all the coefficients of F are bounded by*

$$\Big[(r_1+1)\cdots(r_p+1)\Big]^l l! B^l 2^{(r_1+\cdots+r_p)l}, \quad l \leq r+p+1. \tag{15.21}$$

Exercise 15.2.4. *Prove the above lemma.*

Chapter 16

Lang-Trotter Construction for Continued Fraction of α

16.1 Description of When the Method is Applicable

We give a construction to find the continued fraction expansion of many algebraic numbers.

Theorem 16.1.1 (Lang-Trotter). *Consider a positive real algebraic number α of degree d. Let $P_0(x)$ be its minimal polynomial. Assume*

1. *the leading coefficient of $P_0(x)$ is positive;*
2. *$P_0(x)$ has exactly one positive simple root, α, and $\alpha > 1$.*

Then we can construct a continued-fraction expansion of α just by looking at a sequence of polynomials of degree d.

16.2 Proof of Lang-Trotter Method

We construct a sequence of polynomials by induction. Assume we have constructed polynomials $P_0(x)$, $P_1(x), \ldots, P_n(x)$ satisfying the above conditions.

By induction, $P_n(x)$ has positive leading coefficient and only one positive root, which is greater than one. We want to construct $P_{n+1}(x)$ with the same properties.

Let y_n be the sole positive root of $P_n(x)$. Let $a_n = [y_n]$, the greatest integer less than or equal to y_n. Since the leading coefficient of P_n is positive, $P_n(x) \to$

∞ as $x \to \infty$. Now y_n is the only positive root, and, in particular, is a simple root. If $P_n(a_n) > 0$, then $P_n(y + \delta) < 0$ for $\delta > 0$ small. Thus we would get another positive root by the Intermediate Value Theorem. Hence $P_n(a_n) < 0$.

Define

$$\begin{aligned} Q_n(x) &= P_n(x + a_n) \\ P_{n+1}(x) &= -x^d Q_n(x^{-1}). \end{aligned} \tag{16.1}$$

Then $Q_n(x)$ has a root between 0 and 1, and no other positive roots. It follows that $P_{n+1}(x)$ has only one positive root, y_{n+1}, satisfying $y_{n+1} > 1$. The constant term of $Q_n(x)$ is $P_n(a_n) < 0$. Hence $-P_n(a_n)$, the leading coefficient of $P_{n+1}(x)$, is positive.

One can show y_{n+1} is a simple root of $P_{n+1}(x)$; therefore, $P_{n+1}(x)$ satisfies the desired conditions.

We now show that the a_ns give the continued fraction expansion of α.

Recall $a_n = [y_n]$. The root y_{n+1} satisfies $Q_n(y_{n+1}^{-1}) = 0$.

Therefore

$$P_n(y_{n+1}^{-1} + a_n) = 0. \tag{16.2}$$

Which implies

$$\begin{aligned} y_{n+1}^{-1} + a_n &= y_n \\ \implies y_{n+1} &= (y_n - a_n)^{-1} \\ \implies a_{n+1} &= [y_n - a_n]^{-1} = [y_n - [y_n]]^{-1}. \end{aligned} \tag{16.3}$$

Recalling how we construct the continued fraction expansion of a number, we see that the a_ns are just the coefficients of the continued fraction expansion of α.

16.3 Applying the Lang-Trotter Method

First, one must make sure that the given polynomial is irreducible, with exactly one positive root (which is greater than 1).

Let $P_n(y_n) = 0$. In each iteration, we need to find $a_n = \lfloor y_n \rfloor$. Thus, we need a ballpark approximation for y_n.

One method is divide and conquer, looking for sign changes of $P_n(x)$ for $x > 1$.

Another approach is to apply Newton's method to obtain a sequence of guesses $y_{n,k} \to y_n$.

Let $y_{n,0}$ be our first guess to y_n. It is unlikely, however, that $y_{n,0}$ is correct. Looking at the graph of $P_n(x)$, we have the point $\Big(y_{n,0}, P_n(y_{n,0})\Big)$. The slope of the tangent line at $y_{n,0}$ is $P_n'(y_{n,0})$.

The tangent line at $y_{n,0}$ has equation

$$z - P(y_{n,0}) \;=\; P_n'(y_{n,0}) \cdot (y - y_{n,0}), \tag{16.4}$$

where unfortunately y is the horizontal variable and we are using z as the vertical variable.

The horizontal intercept arises from $z = 0$. Label the corresponding y by $y_{n,1}$. Thus,

$$y_{n,1} \;=\; y_{n,0} - \frac{P_n(y_{n,0})}{P_n(y_{n,0})}. \tag{16.5}$$

We continue by induction – given $y_{n,k}$, $y_{n,k+1}$ is given by

$$y_{n,k+1} \;=\; y_{n,k} - \frac{P_n(y_{n,k})}{P_n(y_{n,k})}. \tag{16.6}$$

One can show that $y_{n,k} \to y_n$. For more information see [Ru], exercises 3.16, 3.18 and $6.??$.

Chapter 17

Eighth Lecture: The Proof of Roth's Theorem

17.1 Review of Index

Definition 17.1.1 (Index). *Let $P(x_1, \ldots, x_p)$ be a polynomial in p variables (not identically zero). Let $\alpha_1, \ldots, \alpha_p$ be real numbers and $r_1, \ldots, r_p$ be positive integers. The **index** θ of p at $(\alpha_1, \ldots, \alpha_p)$ relative to $(r_1, \ldots, r_p)$ is the smallest value of*

$$\frac{i_1}{r_1} + \cdots + \frac{i_p}{r_p} \tag{17.1}$$

such that

$$\frac{1}{i_1! \cdots i_p!} \frac{\partial^{i_1+\cdots+i_p}}{\partial x_1^{i_1} \cdots \partial x_p^{i_p}} P(\alpha_1, \ldots, \alpha_p) \neq 0. \tag{17.2}$$

Note $\theta \geq 0$ and $\theta = 0$ implies $P(\alpha_1, \ldots, \alpha_p) = 0$.

Lemma 17.1.2. *Let P and Q be polynomials. Assume*

1. *ind$(P, Q) \geq \min($ind$(P),$ ind$(Q))$.*
2. *ind$(PQ) =$ ind$(P) +$ ind(Q).*
3. *$F = \det($above$) = U(x_1, \ldots, x_{p-1}))V(x_p)$.*

Then

$$\textit{ind}(F)(\alpha_1,\ldots,\alpha_p,r_1,\ldots,r_p) = \textit{ind}(U)(\alpha_1,\ldots,\alpha_{p-1},r_1,\ldots,r_{p-1}) + \textit{ind}(V)(\alpha_p,r_p). \tag{17.3}$$

17.2 Key Equations: Equations 17.6 through 17.14

We just need to prove Roth's Theorem for algebraic integers α with

$$x^n + a_1x^{n-1} + \cdots + a_{n-1}x + a_n = 0, \;\; a_i \in \mathbb{Z}. \tag{17.4}$$

Let

$$A \;=\; \max\Big(1, |a_1|, \ldots, |a_n|\Big). \tag{17.5}$$

Let $m, \delta, q_1, h_1, \ldots, q_m, h_m, r_1, \ldots, r_m$ satisfy

$$0 < \delta < \frac{1}{m}. \tag{17.6}$$

$$10^m\delta^{\frac{m}{2}} + 2(1+3\delta)nm^{\frac{1}{2}} < \frac{m}{2}. \tag{17.7}$$

$$r_m \;>\; \frac{10}{\delta}, \frac{r_{j-1}}{r_j} > \frac{1}{\delta}, \;\; j = 2,\ldots,m. \tag{17.8}$$

$$\log q_1 \;>\; \frac{1}{\delta^2}\Big[2m + 1 + 2m\log(1+A) + 2m\log(1+|\alpha|)\Big]. \tag{17.9}$$

$$r_j \log q_j \geq r_1 \log q_1. \tag{17.10}$$

Define λ, γ, m, B_1 by

$$\lambda = 4(1+3\delta)nm^{\frac{1}{2}}. \tag{17.11}$$

$$\gamma \;=\; \frac{m-\lambda}{2}. \tag{17.12}$$

$$\eta \;=\; 10^m\delta^{\frac{m}{2}}. \tag{17.13}$$

$$B_1 \;=\; \Big[q_1^{\delta r_1}\Big]. \tag{17.14}$$

Note Equation 17.7 is equivalent to $\eta < \gamma$.

Lemma 17.2.1. *[Roth's Lemma] Suppose Equations 17.6 through 17.11 are satisfied. Let $h_1, \ldots, h_m$ be given with $(h_i, q_i) = 1$ for $i = 1, \ldots, m$. Then there is a polynomial $Q(x_1, \ldots, x_m)$ with integer coefficients. It is of degree at most r_j in x_j and it has the following properties:*

1. *ind$(Q)(\alpha, \ldots, \alpha)$ relative to $(r_1, \ldots, r_m)$ is at least $\gamma - \eta$.*
2. *$Q\Big(\frac{n_1}{q_1}, \ldots, \frac{n_m}{q_m}\Big) \neq 0$.*
3. $$\left|\frac{1}{i_1! \cdots i_m!}\Big(\frac{\partial}{\partial x_1}\Big)^{i_1} \cdots \Big(\frac{\partial}{\partial x_m}\Big)^{i_m} Q(\alpha, \ldots, \alpha)\right| \;<\; B_1^{1+3\delta}. \tag{17.15}$$

Remark 17.2.2. *If we have a polynomial of degree 100, say $P(x)$, and the index of P at α relative to $r = 100$ is high. Say the index $\frac{i}{100}$ is a large number. Then i is of size comparable to 100. Thus the derivatives vanish to high order, and it is like $P(x) = (x - \alpha)^i Q(x)$.*

17.3 Proof of Roth (Assuming Lemma 17.2.1)

Let $k > 2$ and assume

$$\left|\alpha - \frac{h}{q}\right| \;<\; \frac{1}{q^k} \tag{17.16}$$

has infinitely many solutions. Since α is irrational, there are infinitely many solutions with $(h, q) = 1$. Choose an m such that $m > 4nm^{\frac{1}{2}}$, $m > 16n^2$.

Also, choose m large so that

$$\frac{2m}{m - 4nm^{\frac{1}{2}}} \;<\; k. \tag{17.17}$$

Choose δ small such that

$$m - 4(1 + 3\delta)nm^{\frac{1}{2}} - 2\eta \;>\; 0. \tag{17.18}$$

Recall Equation 17.13: $\eta = 10^m \delta^{\frac{m}{2}}$. Thus, as $\delta \to 0$, $\eta \to 0$.

We are doing a small perturbation.

After we choose δ small enough to satisfy the above, we further choose δ small enough so that

$$\frac{2m(1+4\delta)}{m - 4(1+3\delta)nm^{\frac{1}{2}} - 2\eta} \; < \; k. \tag{17.19}$$

This is possible as $k < 2$.

This gives

$$\frac{m(1+4\delta)}{\gamma - \eta} \; < \; k. \tag{17.20}$$

Let $(h_1, q_1) = 1$ with q_1 *very* large so that Equation 17.9 is satisfied. Choose $h_2, q_2, \ldots, h_m, q_m$, with $(h_j, q_j) = 1$ for $j = 2, \ldots, m$ such that $\frac{\log q_j}{\log q_{j-1}} > \frac{2}{\delta}$.

Choose

$$r_1 \; > \; \frac{10}{\delta} \frac{\log q_m}{\log q_1}. \tag{17.21}$$

Now choose $r_2, \ldots, r_m$ recursively such that

$$\frac{r_1 \log q_1}{\log q_j} \; \leq \; r_j 1 + \frac{r_1 \log q_1}{q_j}. \tag{17.22}$$

Exercise 17.3.1. *Show that the above choice leads to Equation 17.10 is satisfied.*

Now we take

$$\frac{r_j \log q_j}{r_1 \log q_1} \; < \; 1 + \frac{\log q_j}{r_1 \log q_1} \; \leq \; 1 + \frac{\log q_m}{r_1 \log q_1} \; < \; 1 + \frac{\delta}{10}, \tag{17.23}$$

where the last bit follows from Equation 17.21.

We now have

$$\begin{aligned} r_m \quad &\geq \quad \frac{r_1 \log q_1}{\log q_m} \; \geq \; \frac{10}{\delta} \\ \frac{r_{j-1}}{r_j} \quad &> \quad \frac{\log q_j}{\log q_{j-1}} \left(1 + \frac{\delta}{10}\right)^{-1} \; > \; \frac{1}{\delta}. \end{aligned} \tag{17.24}$$

Roth's Lemma (Lemma 17.2.1) gives

$$Q\Big(\frac{n_1}{q_1},\ldots,\frac{n_m}{q_m}\Big) \neq 0. \tag{17.25}$$

Q has integer coefficients and is of degree at most r_j in x_j. We therefore have

$$\begin{aligned}\left|Q\Big(\frac{n_1}{q_1},\ldots,\frac{n_m}{q_m}\Big)\right| &\geq \frac{1}{q_1^{r_1}}\cdots\frac{1}{q_m^{r_m}} \\ &> \frac{1}{q_1^{mr_1(1+\delta)}},\end{aligned} \tag{17.26}$$

where the last follows from Equation 17.23. What we are saying is

$$r_1+\cdots+r_m < mr_1(1+\delta). \tag{17.27}$$

On the other hand,

$$\begin{aligned}Q\Big(\frac{h_1}{q_1},\ldots,\frac{h_m}{q_m}\Big) &= \sum_{i_1=0}^{r_1}\cdots\sum_{i_m=0}^{r_m}\frac{1}{i_1!\cdots i_m!}\frac{\partial^{i_1+\cdots+i_m}Q}{\partial x_1^{i_1}\cdots\partial x_m^{i_m}}\times \\ &\quad\times\Big(\alpha,\ldots,\alpha\Big)\cdot\Big(\frac{h_1}{q_1}-\alpha\Big)^{i_1}\cdots\Big(\frac{h_m}{q_m}-\alpha\Big)^{i_m}\end{aligned} \tag{17.28}$$

By the first part of Roth's Lemma, we have that if

$$\frac{i_1}{r_1}+\cdots+\frac{i_m}{r_m} < \gamma-\eta, \tag{17.29}$$

then the coefficient $Q_{i_1,\ldots,i_m}$.
In every other term,

$$\begin{aligned}\left|\Big(\alpha,\ldots,\alpha\Big)\cdot\Big(\frac{h_1}{q_1}-\alpha\Big)^{i_1}\cdots\Big(\frac{h_m}{q_m}-\alpha\Big)^{i_m}\right| < \frac{1}{q_1^{ki_1}\cdots q_m^{ki_m}} &= \frac{1}{(q_1^{i_1}\cdots q_m^{i_m})^k} \\ &\leq \frac{1}{q_1^{r_1(\gamma-\eta)k}},\end{aligned} \tag{17.30}$$

where $q_j \geq q_1^{\frac{r_1}{r_j}}$ as $\frac{r_1\log q_1}{\log q_j} \leq r_j$.
Thus

$$\begin{aligned}\frac{1}{q^{mr_1(1+\delta)}} &\leq \left|Q\Big(\frac{h_1}{q_1},\dots,\frac{h_m}{q_m}\Big)\right| \\ &\leq q^{-r_1(\gamma-\eta)k}(1+r_1)\cdots(1+r_m)B_1^{1+3\delta} \\ &\leq B_1^{1+4\delta}q_1^{-r_1(\gamma-\eta)k} \qquad (17.31)\end{aligned}$$

because, as r_1 is the largest,

$$(1+r_1)\cdots(1+r_m) \leq 2^{r_1+\cdots+r_m} < 2^{mr_1} < B_1^{\delta}. \qquad (17.32)$$

Therefore, we find

$$\frac{1}{q^{mr_1(1+\delta)}} < q^{(1+4\delta)\delta r_1 - r_1(\gamma-\eta)k}, \qquad (17.33)$$

with

$$k < \frac{m(1+4\delta)}{\gamma-\eta}. \qquad (17.34)$$

This proves there are only finitely many solutions.

Exercise 17.3.2. *Try to bound what the largest solution to Roth's Inequality is.*

Chapter 18

Ninth Lecture: The Proof of Roth's Theorem

18.1 Preliminaries

Fix $r_1, \ldots, r_m$ and $B \geq 1$. We choose $R(x_1, \ldots, x_m)$ such that

1. R has integral coefficients and is *not* identically zero.
2. R is of degree at most r_j in x_j, $1 \leq j \leq m$.
3. The coefficients of R have absolute value at most B.

Definition 18.1.1 (R_m). *Let*

$$R_m \ = \ R_m(B; r_1, \ldots, r_m) \tag{18.1}$$

be the set of all such polynomials.

Let $q_1, \ldots, q_m$ be positive integers, $h_1, \ldots, h_m$ numbers relatively prime to $q_1, \ldots, q_m$ respectively.

Let $\theta(R)$ denote the index of $R(x_1, \ldots, x_m)$ at the point $\left(\frac{h_1}{r_1}, \ldots, \frac{h_m}{q_m}\right)$ relative to $r_1, \ldots, r_m$.

Definition 18.1.2 (Index). *The index is the supremum of $\frac{j_1}{r_1} + \cdots + \frac{j_m}{r_m}$ which satisfy*

$$\frac{1}{j_1! \cdots j_m!} \Big(\frac{\partial}{\partial x_1}\Big)^{j_1} \cdots \Big(\frac{\partial}{\partial x_m}\Big)^{j_m} R\Big(\frac{h_1}{q_1}, \ldots, \frac{h_m}{q_m}\Big) \ \neq \ 0. \tag{18.2}$$

$\Theta_M(B, q_1, \ldots, q_m; r_1, \ldots, r_m)$ is the supremum of $\theta(R)$, with $R \in R_m$, over all choices of $h_1, \ldots, h_m$.

18.2 Lemmas

Lemma 18.2.1.

$$\Theta_1(B; q_1, r_1) \ \leq \ \frac{\log B}{r_1! \log q_1}. \tag{18.3}$$

Proof. The definition of θ says $R(x_1)$ vanishes at $x_1 = \frac{h_1}{q_1}$ to order $\theta(R)r_1$.

Therefore, as$(h_1, q_1) = 1$,

$$\begin{aligned} R(x_1) &= \Big(x - \frac{h_1}{q_1}\Big)^{\theta \cdot r_1} Q(x_1) \\ &= (q_1 x - h_1)^{\theta \cdot r_1} \cdot \frac{1}{q_1^{\theta r_1}} Q(x_1) \\ &= (q^{\theta r_1} x^{\theta r_1} \cdots) \cdot (\cdots) \\ \big|q_1^{\theta r_1} \cdots\big| &\leq B \ \text{ therefore } \ q_1^{\theta r_1} \ \leq \ B. \end{aligned} \tag{18.4}$$

□

Lemma 18.2.2. *Let $p \geq 2$ be a positive integer. Let $r_1, \ldots, r_p$ be positive integers satisfying*

$$r_p \ > \ \frac{10}{\delta}, \ \ \frac{r_{j-1}}{r_j} \ > \ \frac{1}{\delta}, \ \ j = 2, \ldots, p, \tag{18.5}$$

where $0 < \delta < 1$, $q_1, \ldots, q_p$ are positive integers. Then

$$\Theta_p(G; q_1, \ldots, q_p; r_1, \ldots, r_p) \ \leq \ 2 \max_l \Big(\Phi + \Phi^{\frac{1}{2}} + \delta^{\frac{1}{2}}\Big) \tag{18.6}$$

where $1 \leq l \leq r_p + 1$ and

$$\begin{aligned} \Phi &= \Theta_1(M; q_p; l r_p) + \Theta_{p-1}(M; q_1, \ldots, q_{p-1}; l r_1, \ldots, l r_{p-1}) \\ M &= (r_1 + 1)^{pl} l! B^l 2^{p l r_1}. \end{aligned} \tag{18.7}$$

Idea: why should something like this hold? Start with a polynomial satisfying certain conditions. Some kind of mixed Wronskian. Calculate determinant, get product of two polynomials, first polynomial of the first $p-1$ variables, last polynomial in just the last (p^{th}) variable.

Since starting polynomials have bounds on size of coefficients, so too will the coefficients of these polynomials be bounded. What we get is the index of the big determinant polynomial (relative to the strings of numbers) will be the index of the polynomial of the first plus the index of the second (because they are independent variables). This is the addition formula for indices.

Lemma 18.2.3 (Lemma 7). *Let m be a positive integer, $0 < \delta < m^{-1}$, $r_1, \ldots, r_m$ integers with $r_m > \frac{10}{\delta}$, $\frac{r_{j-1}}{r_j} > \frac{1}{\delta}$ for $j = 2$ to m, $\log q_1 > \frac{m(2m+1)}{\delta}$ and $r_j \log q_j > r_1 \log q_1$ for $2 \leq j \leq m$ then*

$$\Theta_m(\delta_1^{\delta r_1}; q_1, \ldots, q_m; r_1, \ldots, r_m) \ < \ 10^m \delta^{(\frac{1}{2})^m}. \tag{18.8}$$

Lemma 18.2.4 (Lemma 8 – The Main Lemma). *Let $r_1, \ldots, r_m$ be any positive integers, $\lambda > 0$. The the number of*

$$\begin{array}{rcl} 0 \ \leq & j_1 & \leq r_1 \\ 0 \ \leq & j_2 & \leq r_2 \\ & \vdots & \\ 0 \ \leq & j_m & \leq r_m \end{array} \tag{18.9}$$

satisfying

$$\frac{j_1}{r_1} + \cdots + \frac{j_m}{r_m} \ \leq \ \frac{m-\lambda}{2} \tag{18.10}$$

does not *exceed*

$$2\frac{2\sqrt{m}}{\lambda}(r_1+1)\cdots(r_m+1). \tag{18.11}$$

18.3 Sketch of Proof

Recall the setup: α is an algebraic integer, $f(x) = x^n + a_1 x^{n-1} + \cdots + a_n$, $f(\alpha) = 0$, $A = \max(1, |a_1|, \ldots, |a_n|)$, $m, \delta, q_1, h_1, \ldots, q_m, h_m$ satisfy Equations 17.6 to 17.14.

We have

Lemma 18.3.1 (Lemma 9). *$h_1, \ldots, h_m$ integers relatively prime to $q_1, \ldots, q_m$ respectively, $Q = Q(x_1, \ldots, x_m)$ has integer coefficients, $\deg_{x_j}(Q) \leq r_j$. Then*

1. *The index of Q at $(\alpha, \ldots, \alpha)$ is at least $\gamma - \eta$.*
2. *$Q\left(\frac{h_1}{q_1}, \ldots, \frac{h_m}{q_m}\right) \neq 0$.*
3. *$|Q_{i_1,\ldots,i_m}(\alpha, \ldots, \alpha)| < B_1^{1+3\delta}$.*

Proof. Let

$$\begin{aligned} W(x_1, \ldots, x_m) &= \sum_{s_1=0}^{r_1} \cdots \sum_{s_m=0}^{r_m} c(s_1, \ldots, s_m) x_1^{s_1} \cdots x_m^{s_m} \\ 0 &\leq c(s_1, \ldots, s_m) \leq B_1. \end{aligned} \tag{18.12}$$

The number of such equals $(1 + B_1)^r$, $r = (r_1 + 1) \cdots (r_m + 1)$.
Consider the derivatives

$$W_{j_1,\ldots,j_m}(x_1, \ldots, x_m) = \frac{1}{\cdots} \tag{18.13}$$

Consider all of these derivatives where

$$0 \leq j_i \leq r_i, \quad \frac{j_1}{r_1} + \cdots + \frac{j_m}{r_m} \leq \gamma. \tag{18.14}$$

The number of all such, by Lemma 8, is $D \leq \frac{2\sqrt{m}}{\lambda} r$.
We consider $W_{j_1,\ldots,j_m}(x, \ldots, x)$. Divide by $f(x)$, and let

$$T_{j_1,\ldots,j_m}(W; x) \tag{18.15}$$

be the remainder.

The coefficients of W are bounded by B_1, the coefficients of $W_{j_1,\ldots,j_m}(x_1, \ldots, x_m)$ are bounded by $2^{r_1+\cdots r_m} B_1 \leq 2^{mr_1} B_1$, which is at most $B_1^{1+\delta}$. This follows from Equation 17.9 (and possibly other equations.

Thus, the coefficients of $W_{j_1,\ldots,j_m}$ is at most $B_1^{1+\delta}$.

When we put $x_1 = x_2 = \cdots = x_m = x$, we have a bound of $r B_1^{1+\delta} < B_1^{1+2\delta}$ for the coefficients of $W_{j_1,\ldots,j_m}(x, \ldots, x)$.

Next, consider

$$W_{j_1,\dots,j_m}(x,\dots,x) \;=\; w_s x^s + w_{s-1}x^{s-1} + \cdots + w_0 \tag{18.16}$$

and divide by $f(x)$. Remember $f(x) = x^n + \cdots$, and assume $s \geq n$ (otherwise nothing to do).

We get, on performing the division,

$$W_{j_1,\dots,j_m}(x,\dots,x) w_s x^{s-n} f(x) \tag{18.17}$$

plus stuff of the form either $w_\nu - a_{s-\nu}w_s$ or w_ν. Thus, the absolute values is at most $B_1^{1+2\delta}(1+A)$.

We proceed by induction. Finally, we get the coefficients of $T_{j_1,\dots,j_m}(W;x)$ are bounded by

$$(1+A)^{s-n+1}B_1^{1+2\delta}. \tag{18.18}$$

By our conditions, any such number is at most $B_1^{1+3\delta}$.

The coefficients of $T_{j_1,\dots,j_m}(W;x)$ are bounded by $B_1^{1+3\delta}$. Thus, each coefficient has at most $1 + 2B_1^{1+3\delta}$ possibilities. We have a polynomial of degree n, so we have n choices to make, giving $\left(1 + 2B_1^{1+3\delta}\right)^n$. Further, remember we have D choices for $j_1,\dots,j_m$, giving $\left(1 + 2B_1^{1+3\delta}\right)^{nD}$. This is bounded by $(1+B_1)^r$, which is the number of all possible Ws. We get

$$\left(1 + 2B_1^{1+3\delta}\right)^{nD} \;<\; (1+B_1)^r, \tag{18.19}$$

where the left hand side is the number of all possible D derivatives, and again the right hand side is the number of possible Ws.

By the Pidgeon-Hole Principle, there are two Ws such that, after division by $f(x)$, all of their derivatives are equal.

Let these two polynomials be W_1 and W_2. Let $W^* = W_1 - W_2$. Then

$$f(x) | W^*_{j_1,\dots,j_m}(x,\dots,x), \;\; 0 \leq j_i \leq r_i, \;\; \sum_{i=1}^m \frac{j_i}{r_i} \leq \gamma. \tag{18.20}$$

Observation: $W^* \in R_m(q_1^{\delta r_1}; r_1,\dots,r_m)$.

By Lemma 7, we have a bound for the indices of everything in R_m. We get Θ_m at any point $(\frac{h_1}{q_1},\dots,\frac{h_m}{q_m})$ is at most η. We choose a polynomial Q such that

$$Q(x_1, \ldots, x_m) \;=\; \frac{1}{k_1! \cdots k_m!} \Big(\frac{\partial}{\partial x_1}\Big)^{k_1} \cdots \Big(\frac{\partial}{\partial x_m}\Big)^{k_m} W \tag{18.21}$$

with

$$\frac{k_1}{r_1} + \cdots + \frac{k_m}{r_m} \;<\; \eta \tag{18.22}$$

such that

$$Q\Big(\frac{h_1}{q_1}, \ldots, \frac{h_m}{q_m}\Big) \;\neq\; 0. \tag{18.23}$$

We check and show that Q satisfies the necessary conditions.

By construction, the index of W^* at the point $(\alpha, \ldots, \alpha)$, then the polynomial vanishes, with $\sum \frac{j_i}{r_i} < \gamma$. By shifting the index at most η, and the index is at most γ, then the index is at least $\gamma - \eta$.

□

Chapter 19

Kuzmin's Theorem

19.1 Introduction

Given $\alpha \in \mathbb{R}$, we calculate its continued fraction expansion. Without loss of generality, we may assume $\alpha \in (0,1)$, as this shift changes only the zeroth digit.

Thus,

$$\alpha \ = [0, a_1, a_2, a_3, a_4, \dots] \tag{19.1}$$

Given any sequence of positive integers a_i, we can construct a number α with these as its digits. However, for a generic α chosen randomly in $(0,1)$, how often do we expect to observe digits in the continued fraction expansion equal to 1? To 2? To 3? And so on.

For a given C, what is the measure of the set of $\alpha \in (0,1)$ such that

$$\left|\alpha - \frac{p}{q}\right| \ < \ \frac{C}{q^2} \tag{19.2}$$

holds only finitely often? More generally, instead of $\frac{C}{q^2}$ we could have $\frac{1}{q^2 \log q}$ or any such expression.

19.2 Distribution of $a_1(\alpha) = k$

What is the measure of $\alpha \in (0,1)$ such that $a_1(\alpha) = 7$?

$$\alpha \;=\; \cfrac{1}{7+\cfrac{1}{a_2+\cfrac{1}{\ddots}}} \tag{19.3}$$

Clearly, if $\alpha \leq \frac{1}{8}$, $a_1 \geq 8$. A little thought shows that if $\frac{1}{8} < \alpha \leq \frac{1}{7}$, then $a_1(\alpha) = 7$, because $a_1 = \lfloor \frac{1}{\alpha} \rfloor$. Note $\lfloor x \rfloor$ is the greatest integer less than or equal to x.

So, the measure of $\alpha \in (0,1)$ such that $a_1 = 7$ is $\frac{1}{7} - \frac{1}{8}$. This is $\frac{1}{7\cdot 8}$, which is approximately $\frac{1}{7^2}$.

More generally, we find that the measure of $\alpha \in (0,1)$ such that $a_1(\alpha) = k$ is $\frac{1}{k(k+1)} \approx \frac{1}{k^2}$.

19.3 Distribution of $a_n(\alpha) = k$

Suppose one already has digits $a_1, \ldots, a_{n-1}$. The $\alpha \in (0,1)$ whose first $n-1$ digits are these numbers is a segment of $(0,1)$.

We want the sub-interval where $a_n = k$. Thus, we want to find

$$\frac{|\{\alpha \in (0,1) : \text{if } i \leq n-1, a_i(\alpha) = a_i; a_n(\alpha) = k\}|}{|\{\alpha \in (0,1) : \text{if } i \leq n-1, a_i(\alpha) = a_i\}|} \tag{19.4}$$

Lemma 19.3.1. *The above is at least $\frac{1}{3k^2}$ and at most $\frac{2}{k^2}$.*

The calculation hinges on the fact that this interval is $\left[\frac{p_n k + p_{n-1}}{q_n k + q_{n-1}}, \frac{p_n(k+1)+p_{n-1}}{q_n(k+1)+q_{n-1}}\right]$ $|p_{n-1}q_n - p_n q_{n-1}| = 1$.

For $\alpha \in (0,1)$, by Lemma 19.3.1 we obtain

$$\begin{aligned} \frac{1}{3k^2}\,|\{\alpha : i \leq n-1, a_i(\alpha) = a_i\}| \;&\leq\; |\{\alpha : i \leq n-1, a_i(\alpha) = a_i; a_n(\alpha) = k\}| \\ &\leq\; \frac{2}{k^2}\,|\{\alpha : i \leq n-1, a_i(\alpha) = a_i\}| \end{aligned} \tag{19.5}$$

and thus, summing over all possible a_i, $i \leq n-1$,

$$\frac{1}{3k^2} \;\leq\; |\{\alpha : a_n(\alpha) = k\}| \leq \frac{2}{k^2}. \tag{19.6}$$

Corollary 19.3.2. *There exist constants $0 < C_1 < C_2 < \infty$ such that*

$$\frac{C_1}{k} \ < \ |\{\alpha \in (0,1) : a_n(\alpha) \geq k\}| \ < \ \frac{C_2}{k}. \tag{19.7}$$

Proof. The sum $\frac{1}{k^2} + \frac{1}{(k+1)^2} + \cdots \approx \frac{1}{k}$. □

19.4 Measure of α with Bounded Digits in their Continued Fraction Expansion

Theorem 19.4.1. *Consider all $\alpha \in (0,1)$ such that for all n, $a_n(\alpha) \leq K$ for some fixed constant K. The set of such α has measure 0.*

Proof. We look and see what percent of the sub-intervals we keep losing. Let $\beta = 1 - \frac{C_1}{K}$, where C_1 is as in Corollary 19.3.2. We can show that the probability that the first n digits are all at most K is β^n, as the requirement that each a_i be at most K causes us to keep at most β of the sub-intervals we still have. As $n \to \infty$, β^n tends to 0. □

If α has $a_i(\alpha) > N$, then $\left|\alpha - \frac{p_n}{q_n}\right| \leq \frac{1}{Nq_n^2}$. Letting $\epsilon = \frac{1}{N}$, we see we can approximate to within $\frac{\epsilon}{q_n^2}$. As almost all α have infinitely many i with $a_i(\alpha) > N$, for almost all α we can find infinitely many $\frac{p_n}{q_n}$ such that the approximation is as good as $\frac{\epsilon}{q_n^2}$.

19.5 Measure of α with Digits in their Continued Fraction Expansion Growing

Theorem 19.5.1. *If*

$$\sum_{n=1}^{\infty} \frac{1}{k_n}$$

converges, the set $\{\alpha \in (0,1) : a_i(\alpha) \leq k_i\}$ has positive measure.

Remark 19.5.2. *Let $k_i \geq 2$. An infinite product*

$$\left(1 - \frac{1}{k_1}\right) \cdot \left(1 - \frac{1}{k_2}\right) \cdot \left(1 - \frac{1}{k_3}\right) \cdots \tag{19.8}$$

converges to zero if and only if

$$\sum \frac{1}{k_i} \tag{19.9}$$

diverges.

For example,

$$\prod_n \left(1 - \frac{1}{n}\right) \ \to \ \frac{1}{2}\frac{2}{3}\frac{3}{4}\cdots \to 0. \tag{19.10}$$

Consider α with convergents $\frac{p_n}{q_n}$. For almost all $\alpha \in (0,1)$, the q_ns will grow exponentially. Explicitly, there exist constants B_1 and B_2 such that

$$e^{B_1 n} \ \leq \ q_n \ \leq \ e^{B_2 n}. \tag{19.11}$$

Using $q_{n+1} = a_n q_n + q_{n-1}$, we show

$$a_1 \cdots a_n \ \leq \ q_n \ \leq \ 2a_1 \cdots a_n. \tag{19.12}$$

Thus, it is sufficient to bound $a_1 \cdots a_n$.

Say $x_1 \cdots x_n = c$; call this an *n-dimensional hyperbola*. Assume further that $c_1 \leq x_1 \cdots x_n \leq c_2$. Using that sizes of certain sub-intervals with $a_n(\alpha) = k$ is of size $\frac{1}{k^2}$, we are led to integrating

$$\int_V \frac{1}{x_1^2 + \cdots + x_n^2} \tag{19.13}$$

over a region of space trapped between two n-dimensional hyperbolas. This is similar to the following: if we want to estimate the number of integer solutions to $500 \leq x^2 + y^2 \leq 1000$, the main term arises from an annulus with radii $\sqrt{500}$ and $\sqrt{1000}$.

We have a function $f(n)$ with $f(n) \leq \frac{c}{n}$. What is the measure of the set of all $\alpha \in (0,1)$ such that

$$\left|\alpha - \frac{p_n}{q_n}\right| \ < \ \frac{f(q_n)}{q_n} \tag{19.14}$$

for infinitely many n? If $f(n) = \frac{1}{n \log n}$ (or, in general, if $\sum f(n)$ diverges), the measure is one. What if, instead, $f(n) = \frac{1}{n \log^{1+\epsilon}(n)}$, for some small but positive ϵ? The measure is then zero!

The first of the two assertions is basically Theorem 19.5.1 strengthened by information coming from (19.11). The second one can be proven as follows. Suppose $\sum f(n)$ converges. Let $E_n = \{\alpha \in (0,1) : |\alpha - \frac{p}{n}| < f(n)/n\}$. The measure of E_n is approximately $f(n)$. If the set of all $\alpha \in (0,1)$ contained in infinitely many E_{q_n}'s had positive measure, $\sum_n \mu(E_{q_n})$ couldn't converge. Then $\sum f(n)$ couldn't converge. Contraditction. Therefore the set of all $\alpha \in (0,1)$ contained in infinitely many E_{q_n}'s has zero measure. This is the same as the set of all $\alpha \in (0,1)$ for which

$$\left|\alpha - \frac{p_n}{q_n}\right| \ < \ \frac{f(q_n)}{q_n} \tag{19.15}$$

for infinitely many n.

19.6 Needed Technical Results

Let $z_n(\alpha)$ be the continued fraction from the n^{th} digit onward:

$$\begin{aligned} \alpha &= \frac{1}{a_1 + \frac{1}{a_2 + \frac{1}{a_3 + \cdots}}} \\ z_n(\alpha) &= \frac{1}{a_{n+1} + \frac{1}{a_{n+2} + \frac{1}{a_{n+3} + \cdots}}} \end{aligned} \tag{19.16}$$

As always, we assume $\alpha \in (0,1)$.
Define, for $x \in (0,1)$:

$$m_n(x) \ = \ |\{\alpha \in (0,1) : z_n(\alpha) < x\}|. \tag{19.17}$$

We have

$$m_{n+1} \ = \ \sum_{k=1}^{\infty} m_n\Big(\frac{1}{k}\Big) - m_n\Big(\frac{1}{k+x}\Big). \tag{19.18}$$

For example, if we want $z_2(\alpha) < .3$, we have lots of different sets that work

$$\left(\frac{1}{1+.3}, \frac{1}{1}\right), \ \left(\frac{1}{2+.3} + \frac{1}{2}\right), \ \left(\frac{1}{3+.3}, \frac{1}{3}\right), \ldots \tag{19.19}$$

We find

$$\begin{aligned} m_1(x) &= x \\ m_2(x) &= \sum_{k=1}^{\infty} \frac{1}{k} - \frac{1}{k+x} \\ &= \sum_{k=1}^{\infty} m_1\left(\frac{1}{k}\right) - m_1\left(\frac{1}{k+x}\right). \end{aligned} \tag{19.20}$$

We expect it to converge to a function $\phi(x)$ such that

$$\phi(x) = \phi\left(\frac{1}{k}\right) - \phi\left(\frac{1}{k+x}\right). \tag{19.21}$$

A formula of type $\phi(x) = C\ln(1+x)$ will satisfy the above. It turns out that $\phi(x)$ *is* of this form, with $C = \frac{1}{\ln 2}$.

Kuzmin showed that if a sequence of functions $\{f_n\}$ satisfies

$$f_{n+1}(x) = \sum_{k=1}^{\infty} \frac{1}{(m+x)^2} f\left(\frac{1}{k+x}\right) \tag{19.22}$$

and we have some bounds on $f_1(x)$, say $0 < f_1(x) < B$, $f_1'(x) < B$, we find

$$f_n(x) = \frac{a}{1+x} + C_1 e^{-C_2\sqrt{n}}, \quad a = \frac{1}{\ln 2}\int_0^1 f_1(x)dx. \tag{19.23}$$

If $\{m_n\}$ satisfies (19.18), then $\{f_n\} = \{m_n'(x)\}$ satisfies (19.22). Setting $f_1 = m_1'(x)$, we obtain

$$m_n(x) = \log_2(1+x) + C_3 e^{-C_2\sqrt{n}}. \tag{19.24}$$

19.7 Kuzmin's Theorem

What is the probability that $a_n = 37$? We know, as $n \to \infty$, that the probability is at least $\frac{1}{3\cdot 37^2}$ and at most $\frac{2}{37^2}$. We can do better now. The probability is

$$m_n\left(\frac{1}{37}\right) - m_n\left(\frac{1}{37+1}\right) \quad \to \quad \log_2\left(1+\frac{1}{37}\right) - \log_2\left(\frac{1}{37+1}\right) \tag{19.25}$$

We can sharpen the above and incorporate an error bound.

Theorem 19.7.1. *Let α be chosen uniformly in $(0,1)$. Then*

$$\textit{Prob}(a_n = k) \;=\; \log_2\left(1+\frac{1}{k(k+2)}\right) \;+\; \frac{C_3}{k(k+1)}e^{-C_2\sqrt{n-1}}. \tag{19.26}$$

Proof: the probability we want is simply

$$m_{n-1}\left(\frac{1}{k}\right) \;-\; m_{n-1}\left(\frac{1}{k+1}\right) \;=\; \int_{\frac{1}{k+1}}^{\frac{1}{k}} m'_{n-1}(x)dx. \tag{19.27}$$

We note $f_n(x) = m'_(x)$ satisfied Equations 19.22 and 19.23, and the result follows by integration.

Levy (see [Le]) proved the error term may be taken as $Ae^{-\lambda n}$.

19.8 Strengthened Versions of Kuzmin's Theorem

We can strengthen (19.24) further. Let $f : \mathbb{Z}^+ \to \mathbb{R}$ be such that $f(n) < Cn^{\frac{1}{2}-\epsilon}$. For almost all $\alpha \in (0,1)$ (ie, all α except for a set of measure zero),

$$\lim_{N\to\infty}\sum_{n=1}^{\infty} f\left(a_n(\alpha)\right) \;=\; \sum_{n=1}^{\infty} f(n)\log_2\left(1+\frac{1}{n(n+2)}\right). \tag{19.28}$$

We can take

$$f(n) = \begin{cases} 0 & \text{if } n \neq k \\ 1 & \text{if } n = k \end{cases} \tag{19.29}$$

and regain Theorem 19.7.1. If we take

$$f(n) \;=\; \ln n \tag{19.30}$$

then

$$\frac{1}{N}\sum_{n=1}^{N} \ln a_n \;=\; \log\, \sqrt[N]{a_1\cdots a_N}. \tag{19.31}$$

Chapter 20

Kuzmin Experiments

20.1 Statement of Problem

Let $p(a_n(x) = k)$ denote the probability that the n^{th} digit of the continued fraction expansion of x is k. Kuzmin's theorem states that, for almost any number $x \in (0, 1)$,

$$\left| p(a_n(x) = k) \; - \; \log_2\left(1 + \frac{1}{k(k+2)}\right) \right| \; \leq \; \frac{A}{k(k+1)} \cdot e^{-B\sqrt{n-1}}, \qquad (20.1)$$

for some constants A and B. As an aside: Levy proves a better bound, showing the error is at most Ce^{-Dn}.

Thus, for n large, the probability that the n^{th} digit is k is approximately independent of n.

Fix a range of digits, say n runs from M to $N - 1$. What is the expected number of digits equal to k? Is it approximately $(N - M)\log_2(1 + \frac{1}{k(k+2)})$?

20.2 Direct Solution

We give a difficult, but direct, solution. We have a range of digits that we are investigating. For definiteness, let us say we are looking at digits 1 to 1000. Let S_{1000} be the set of all strings of 1000 numbers, where each digit is an integer.

Given a number $x \in (0, 1)$, there is a unique $s(x) \in S_{1000}$ such that the first 1000 digits in the continued fraction expansion of x are the same (and in the same order) as those from $s(x)$.

Given x, let $\text{num}_{1000}(x) = \text{num}_{1000;1}(x)$ be the number of ones in the first 1000 digits of its continued fraction expansion. By an abuse of notation, we will write $\text{num}_{1000}(s(x))$ for the number of ones in the first 1000 digits of $s(x)$.

We say $x \equiv s(x)$ if the first 1000 digits of the continued fraction of x is $s(x)$.

Thus, the expected number of ones in the first 1000 digits of continued fraction expansions of numbers in $(0, 1)$ is

$$\sum_{s \in S_{1000}} \text{num}_{1000}(s(x)) \cdot \text{Prob}(x : x \equiv s(x)). \tag{20.2}$$

Thus, to solve the original question using this method, it is necessary to know what is the measure of the sets $\{x : x \equiv s(x)\}$ for each $s \in S_{1000}$.

This is a very complicated question. It can be computed by brute force, but it is quite involved, and Kuzmin's Theorem is not applicable. Obviously, there is nothing special about the first 1000 digits and looking at the number of occurrences of 1; we could investigate the number of occurrences of k from digits M to $N - 1$.

We now show a better way.

20.3 Solution via Linearity of Expected Values

Recall $a_n(x)$ is the n^{th} digit of the continued fraction expansion of x. We define the following indicator variables:

$$A_{n,k}(x) \ = \ \begin{cases} 1 & \text{if } a_n(x) = k \\ 0 & \text{if } a_n(x) \neq k \end{cases} \tag{20.3}$$

Thus,

$$\text{num}_{M,N;k}(x) \ = \ \sum_{n=M}^{N-1} A_{n,k}(x) \tag{20.4}$$

is the number of digits n of x which equal k with $M \leq n < N$.

Let $q_k = \log_2\left(1 + \frac{1}{k(k+2)}\right)$. Kuzmin's Theorem states that the probability $a_n(x) = k$ is, up to a small error (for n large), q_k. Thus, each $A_{n,k}(x) = 1$ with probability approximately q_k and 0 with probability approximately $1 - q_k$.

Hence, the expected value of $A_{n,k}(x)$ is

$$E[A_{n,k}(x)] \;=\; 1 \cdot q_k + 0 \cdot (1 - q_k) \;=\; q_k. \tag{20.5}$$

The power of Kuzmin's theorem is that the above expectation is, up to a very small error, independent of n. In other words, the main term of these expectations is q_k.

We now use the linearity of the expected value: the expected value of a sum is the sum of the expected values. This simple observation allows us to avoid having to calculate the measure of sets $\{x : x \equiv_{M,N-1} s(x)\}$ for each $s \in S_{M,N-1}$, where the notation is an obvious generalization of before.

We have

$$\begin{aligned} E\left[\text{num}_{M,N;k}(x)\right] &= E\left[\sum_{n=M}^{N-1} A_{n,k}(x)\right] \\ &= \sum_{n=M}^{N-1} E\left[A_{n,k}(x)\right] \\ &= \sum_{n=M}^{N-1} (q_k + \text{small error}) \\ &\approx (N-M)q_k + (\text{small error}) \cdot (N-M). \end{aligned} \tag{20.6}$$

In words, we have shown that if we look at $N - M$ digits, we expect to see $(N - M)q_k$ occurrences of the digit k. While the proof above is for *consecutive* digits, a similar proof works for *any* set of $N - M$ digits.

Very concretely, if we were to look at digits $50,001$ to digits $100,000$, the number of ones we expect to see would be

$$50,000 \cdot \log_2\left(1 + \frac{1}{1(1+2)}\right) \;\approx\; 50,000 \cdot .415. \tag{20.7}$$

20.4 Generalization

Fix k_1, k_2, M and $N - 1$. What is the expected number of occurrences of the pair (k_1, k_2) among digits M to $N - 1$?

For example, let $(k_1, k_2) = (1, 1)$. If we had a string of digits

$$1, 1, 1, 2, 3, 453, 1, 17, 5, 4, 1, 2, 10, 2, 1, 1, 19 \tag{20.8}$$

in the continued fraction expansion of x, this would contribute three pairs of $(1,1)$. Note that $1,1,1$ counts as *two* pairs.

Using Kuzmin's Theorem, one can show (for n large) that the probability that $a_n(x) = k_1$ and $a_{n+1}(x) = k_2$ is q_{k_1,k_2} plus significantly smaller corrections. Note, similar to Kuzmin (which looks at just one digit) that we have approximate n-independence in the probabilities q_{k_1,k_2}. Note (do the calculation!) that $q_{k_1,k_2} \neq q_{k_1}q_{k_2}$.

We again define indicator variables:

$$B_{n,k_1,k_2}(x) \;=\; \begin{cases} 1 & \text{if } a_n(x) = k_1 \text{ and } a_{n+1}(x) = k_2 \\ 0 & \text{otherwise} \end{cases} \tag{20.9}$$

Similar to before, we have the expected value of $B_{n,k_1,k_2}(x)$ is q_{k_1,k_2}.

Let $\text{num}_{M,N,k_1,k_2}(x)$ equal the number of $n \in [M, N-1]$ such that $a_n(x) = k_1$ and $a_{n+1}(x) = k_2$. We have

$$\begin{aligned} E\left[\text{num}_{M,N,k_1,k_2}(x)\right] &= E\left[\sum_{n=M}^{N-1} B_{n,k_1,k_2}(x)\right] \\ &= \sum_{n=M}^{N-1} E\left[B_{n,k_1,k_2}(x)\right] \\ &= \sum_{n=M}^{N-1} (q_{k_1,k_2} + \text{small error}) \\ &= (N-M)q_{k_1,k_2} + (\text{small error}) \cdot (N-M) \\ &\approx (N-M)q_{k_1,k_2}. \end{aligned} \tag{20.10}$$

20.5 General Comments

Using binary indicator variables is an extremely useful and standard trick. Note that we do *not* have the digits in a continued fraction expansion are independent of each other. Fortunately, such a fact is not needed to calculate the expected number of occurrences of $a_n(x) = k$ or $a_n(x) = k_1$ and $a_{n+1}(x) = k_2$ for $n \in [M, N-1]$.

We can't avoid having to do some averaging over all $x \in (0,1)$, but we don't want to have to deal with sets such as *the set of* $x \in (0,1)$ *such that there are exactly* 237 *ones in the first* 1000 *digits of the continued fraction expansion*. It

would be very difficult and time consuming to determine the measure of each such set (for all the sets we would need).

Fortunately, using binary indicator variables, we can avoid such a subdivision. Instead, we study significantly simpler sets such as *the set of* $x \in (0,1)$ *such that* $a_n(x) = k$. This is much easier, and in fact the measure of such sets (for n large) is Kuzmin's theorem.

Appendix A

Robert Lipshitz's Junior Project: Numerical results concerning the distribution of $\{n^2\alpha\}$

The following is Robert Lipshitz's paper from the $2000-2001$ Junior Research Seminar / Undergraduate Mathematics Laboratory at Princeton. For a copy of his programs, please go to

$http://www.math.princeton.edu/\sim mathlab/projects/n2alpha/main.html$

A.1 Introduction

This paper presents numerical evidence concerning the following problem. Fix an irrational number α and an integer $N > 0$. By $\{x\}$ we denote the fractional part of x. Consider $\{n^2\alpha\}$ for n from 1 to N. Write these numbers in increasing order and let β_j be the j^{th} of them. The problem is concerned with the distribution of the differences between consecutive β_j. Specifically, for almost all α the consecutive spacings are expected to behave as one would expect for numbers chosen at random in $[0, 1)$. We state the conjectures precisely below.

The question is of mathematical interest, but also of physical interest. It arises in the study of certain quantum mechanical systems where one is concerned with the onset of chaotic behavior. For example, the question arises in the study of the "quantum kicked rotator" [CGI].

A.2 Known Results

This section provides some background information about the problem. First we prove that $\{n^2\alpha\}$ is equidistributed, following the treatment in [Ca]. Then we state and prove a few easy results about random numbers, for comparison with our case, and finally describe a few of the results proved in [RSZ].

We begin with some preliminary definitions and then set up general machinery which will make proving the equidistribution of $\{n^2\alpha\}$ relatively easy.

Definition A.2.1. *We say that a sequence $\{\alpha_k\}_{k=1}^{\infty}$ in $[0,1]$ is **equidistributed** if for any interval $[a,b] \subset [0,1]$,*

$$\lim_{n\to\infty} \frac{1}{n} \left(\#\left\{\alpha_k \in [a,b], 1 \le k \le n\right\}\right)$$

exists and is equal to $b-a$.

Notice that taking $[a,b] \subset (0,1)$ rather than $[a,b] \subset [0,1]$ in the above would give an equivalent definition.

Definition A.2.2. *We say a sequence of real numbers $\{\alpha_k\}_{k=1}^{\infty}$ is **equidistributed mod 1** if the sequences of fractional parts of α_k is equidistributed.*

Theorem A.2.3. *For $0 \le \alpha_n < 1$, $\{\alpha_n\}$ is equidistributed if and only if for any $\mathcal{C}^\infty$ function $f : [0,1] \to \mathbb{C}$ it is true that $\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} f(\alpha_k) = \int_0^1 f(x)dx$.*

Proof: Suppose $\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} f(\alpha_k) = \int_0^1 f(x)dx$ for any $\mathcal{C}^\infty$ function $f : [0,1] \to \mathbb{C}$. Fix $[a,b] \subset (0,1)$ and fix ϵ with $0 < \epsilon < b-a$. It is not hard to construct a $\mathcal{C}^\infty$ $f : [0,1] \to \mathbb{R}$ such that $0 \le f(x) \le 1$ for all $x \in [0,1]$ and:

$$f(x) = \begin{cases} 0 & \text{if } x < a-\epsilon \text{ or } x > b+\epsilon \\ 1 & \text{if } a < x < b \end{cases}$$

Then,

$$b-a \le \int_0^1 f(x)dx \le b-a+2\epsilon$$

Also, $\sum_{k=1}^{n} f(\alpha_k) \ge \#\left\{\alpha_k \in [a,b], 1 \le k \le n\right\}$ so

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1} f(\alpha_k) \ge \limsup_{n\to\infty} \frac{1}{n} \left(\#\left\{\alpha_k \in [a,b], 1 \le k \le n\right\}\right)$$

Hence, since we assumed that $\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} f(\alpha_k) = \int_0^1 f(x)dx$ for any f, we have $b-a+2\epsilon \geq \limsup_{n\to\infty} \frac{1}{n}\left(\#\left\{\alpha_k \in [a,b], 1\leq k\leq n\right\}\right)$ and letting ϵ tend to zero we obtain

$$b-a \geq \limsup_{n\to\infty} \frac{1}{n}\left(\#\left\{\alpha_k \in [a,b], 1\leq k\leq n\right\}\right).$$

Applying the symmetric argument to a function $g(x)$ with $0 \leq g(x) \leq 1$ for all $x \in [0,1]$ and

$$g(x) = \begin{cases} 0 & \text{if } x < a \text{ or } x > b \\ 1 & \text{if } a+\epsilon < x < b-\epsilon \end{cases}$$

and letting $\epsilon \to 0$ yields the inequality

$$b-a \leq \liminf_{n\to\infty} \frac{1}{n}\left(\#\left\{\alpha_k \in [a,b], 1\leq k\leq n\right\}\right)$$

so in fact $\lim_{n\to\infty} \frac{1}{n}\left(\#\left\{\alpha_k \in [a,b], 1\leq k\leq n\right\}\right)$ exists and is equal to $b-a$.

Conversely, suppose that for any interval $[a,b] \subset [0,1]$,

$$\lim_{n\to\infty} \frac{1}{n}\left(\#\left\{\alpha_k \in [a,b], 1\leq k\leq n\right\}\right) = b-a.$$

Fix $\epsilon > 0$ and $f : [0,1] \to \mathbb{C}$. Since f is continuous on a compact set and hence uniformly continuous, there is an $N \in \mathbb{N}$ such that $|x-y| \leq \frac{1}{N} \Rightarrow |f(x)-f(y)| < \epsilon$. Consider the partition of $[0,1]$ into $0 < 1/N < 2/N < 3/N < ... < N/N = 1$. There is an $M \in \mathbb{N}$ such that for $m > M$ and any a between 1 and $N-1$,

$$\left|\frac{\#\{\alpha_i \in [a/N,(a+1)/N], 1\leq i\leq m\}}{m} - \frac{1}{N}\right| < \frac{\epsilon}{N}$$

Fix m. Let $A_j = \#\{\alpha_i \in [j/N, (j+1)/N] \text{ with } 1 \le i \le m\}$. Now,

$$\begin{aligned}
\left|\frac{1}{m}\sum_{i=1}^{m} f(\alpha_i) - \int_0^1 f(x)dx\right| &\le \left|\frac{1}{m}\sum_{i=1}^{m} f(\alpha_i) - \frac{1}{N}\sum_{j=0}^{N-1} f(j/N)\right| \\
&\qquad + \left|\frac{1}{N}\sum_{j=0}^{N-1} f(j/N) - \int_0^1 f(x)dx\right| \\
&< \left|\sum_{j=0}^{N-1}\left[\left(\sum_{\alpha_i\in[j/N,j+1/N]} \frac{f(\alpha_i)}{m}\right) - \frac{f(j/N)}{N}\right]\right| + \frac{N\epsilon}{N} \\
&\le \sum_{j=1}^{N-1}\left(A_j\frac{\epsilon}{m} + \left|\frac{A_j}{m} - \frac{1}{N}\right|\right) + \epsilon \\
&\le \epsilon + \sum_{j=0}^{N-1}\frac{\epsilon}{N} + \epsilon = 3\epsilon
\end{aligned}$$

Thus, since ϵ was arbitrary, this implies that

$$\left|\frac{1}{m}\sum_{i=1}^{m} f(\alpha_i) - \int_0^1 f(x)dx\right| \to 0$$

as $m \to \infty$, which completes the proof. □

Theorem A.2.4. *A sequence $\{\alpha_n\}_{n=1}^{\infty}$ is equidistributed mod 1 if and only if for all $t \in \mathbb{Z}$, $t \neq 0$, $\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} e^{2\pi i t\alpha_k} = 0$.*

Proof: Observe that by periodicity of $e^{2\pi iz}$ we may assume that all α_n are between 0 and 1 and replace "equidistributed mod 1" by simply "equidistributed."

Suppose that for all $t \neq 0$, $\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} e^{2\pi i t\alpha_k} = 0$. Fix a $\mathcal{C}^\infty$ function $f : [0,1] \to \mathbb{C}$. From elementary Fourier analysis, there is a sequence $\{a_n\}_{n=-\infty}^{\infty}$ with $\lim_{n\to\infty} n^2 a_n = 0$ such that

$$f(x) = \sum_{n=-\infty}^{\infty} a_n e^{2\pi inx}.$$

(Since $\lim_{n\to\infty} n^2 a_n = 0$, the series converges uniformly by the Weirstrass M

test.) $\int_0^1 f(x)dx = a_0$. Finally,

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n} f(\alpha_k) &= \lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}\sum_{t=-\infty}^{\infty} a_t e^{2\pi i t\alpha_k}\\ &= \sum_{t=-\infty}^{\infty}\lim_{n\to\infty}\frac{a_t}{n}\sum_{k=1}^{n} e^{2\pi i t\alpha_k}\\ &= a_0,\end{aligned}$$

where the change in the order of summation is justified by uniform convergence. Thus, $\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n} f(\alpha_k) = \int_0^1 f(x)dx$ so by the previous theorem, $\{\alpha_n\}_{n=1}^{\infty}$ is equidistributed.

The converse is an immediate corollary of the previous theorem. □

The following technical lemma is taken verbatim from [Ca] page 71.

Lemma A.2.5. *Let $u_1, u_2, ..., u_Q$ be in $\mathbb{C}$, and let $1 \le H \le Q$. Then,*

$$\begin{aligned}H^2\left|\sum_{1\le q\le Q} u_q\right|^2 \le{}& H(H+Q-1)\sum_{1\le q\le Q}|u_q|^2\\ &+ 2(H+Q-1)\sum_{0<h<H}(H-h)\sum_{1\le q\le Q-h}\bar{u}_q u_{q+h}.\end{aligned}$$

Proof: For convenience, set $u_q = 0$ for $q \le 0$ or $q > Q$.

$$H\sum_{1\le q\le Q} u_q = \sum_{0\le p\le H+Q}\sum_{0\le r\le H} u_{p-r}$$

so by the Schwartz inequality,

$$\begin{aligned}H^2\left|\sum_{1\le q\le Q} u_q\right|^2 &= \left|\sum_{0<p<H+Q}\sum_{0\le r<H} u_{p-r}\right|^2\\ &\le \sum_{0<p<H+Q}\left|\sum_{0\le r<H} u_{p-r}\right|^2\\ &= (H+1-1)\sum_{\substack{0<p<H+Q\\ 0\le r,s<H}} u_{p-r}\bar{u}_{p-s}\end{aligned}$$

Now, $|u_q|^2$ occurs for $r = s = p - q$ if $p - q \geq 0$. $p - q$ can be anything less than H, so the term occurs H times. $u_q \bar{u}_{q+h}$ or $\bar{u_q} u_{q+h}$ occurs if $0 < h < H$, $H - h$ times, and the result follows immediately. □

Corollary A.2.6. *For a sequence $\{z_q\}_{q=1}^{\infty}$, suppose that for each $h > 0$*

$$\frac{1}{Q} \sum_{1 \leq q \leq Q} e^{2\pi i (z_{q+h} - z_q)} \to 0$$

as $Q \to \infty$. Then $\frac{1}{Q} \sum_{1 \leq q \leq Q} e^{2\pi i z_q} \to 0$ as $Q \to \infty$.

Proof: In the above, taking $u_q = e^{2\pi i z_q}$,

$$\left| \sum_{1 \leq q \leq Q} \frac{u_q}{Q} \right|^2 \leq \frac{H + Q - 1}{Q^2 H} \sum_{1 \leq q \leq Q} 1 + 2 \frac{H + Q - 1}{QH} \sum_{0 < h \leq H} \frac{H - h}{QH} \sum_{1 \leq q \leq Q - h} e^{2\pi i (z_{q+h} - z_q)}.$$

The first term approaches $1/H$ as $Q \to \infty$ while the second goes to zero. Letting H tend to infinity then yields the result. □

Theorem A.2.7. *Suppose $z_{q+h} - z_q$ is equidistributed mod 1 for all $h \in \mathbb{Z}$, $h > 0$. Then so is z_q.*

Proof: By assumption and theorem A.2.4, for any $t \in \mathbb{Z}$, $t \neq 0$,

$$\frac{1}{Q} \sum_{1 \leq q \leq Q} e^{2\pi i t (z_{q+h} - z_q)} \to 0$$

as $Q \to \infty$. Hence, by the preceding corollary, $\frac{1}{Q} \sum_{1 \leq q \leq Q} e^{2\pi i t z_q} \to 0$ as $Q \to \infty$. Hence, by theorem A.2.4 again, $\{z_q\}_{q=1}^{\infty}$ is equidistributed mod 1. □ Now the result we want follows quite easily.

Theorem A.2.8. *For α irrational, $\{n\alpha\}_{n=0}^{\infty}$ is equidistributed mod 1.*

Proof: Consider $\frac{1}{N} \sum_{k=0}^{N} e^{2\pi i k t \alpha}$. The series is geometric, and equals $\frac{1}{N} \frac{1 - e^{2\pi i (N+1) t \alpha}}{1 - e^{2\pi i t \alpha}}$, which goes to 0 as N goes to infinity. □

Theorem A.2.9. *For α irrational, $\{n^2\alpha\}_{n=1}^{\infty}$ is equidistributed mod 1.*

Proof: For any fixed integer $h > 0$, $(n+h)^2\alpha - n^2\alpha = 2h\alpha + h^2$ which is equidistributed by the previous theorem. Thus, by Theorem A.2.7, $\{n^2\alpha\}$ is equidistributed mod 1. □

It is actually true that for any polynomial f with at least one coefficient (other than the constant term) irrational, $\{f(n)\}_{n=1}^{\infty}$ is equidistributed. This can be proved in a manner analogous to the previous theorem, by induction on the degree of the polynomial. See [Ca], page 73.

One way of interpreting the preceding result is that the sequence $\{n^2\alpha\}$ behaves, in this respect, like a uniformly distributed sequence of independent random numbers. The questions that this paper attempts to address numerically relate to how far this analogy can be pushed. We now state and prove a few easy results about such sequences of random numbers, to be used for comparison later. The main issue we are interested in is consecutive spacing of numbers in the sequences. For a more detailed (and more elegant) treatment, the reader is advised to consult the first chapter of [Fe].

With probability 1, a sequence $X_1, ..., X_n$ of uniformly distributed independent random variables in $[0,1]$ partitions $[0,1]$ into $n+1$ subintervals. We are interested in the lengths of these subintervals.

Lemma A.2.10. *The probability that the length of a given subinterval is at least t is $(1-t)^n$.*

Proof: Let L_i denote the length of the i^{th} interval. Observe that the distribution of L_i is independent of i. To see this, imagine choosing $n+1$ random points on a circle instead of n random points on a line. In this case, it is clear that the lengths of the intervals all have the same distribution. Cutting the circle at the $(n+1)^{st}$ point yields n points chosen at random in the line, and thus in this case as well the distribution of L_i is independent of i.

Consider the left-most interval. It will have length at least t if none of the X_i is less than t. This happens with probability $(1-t)^n$. □

Corollary A.2.11. *Let $X_1, X_2, ...$ be a sequence of independent random variables in $[0,1]$. For fixed n, $X_1, ..., X_n$ partition $[0,1]$ into n+1 subintervals. Let L_n denote the length of the first interval in this partition. Then, for fixed t, the probability that $nL_n > t$ approaches e^{-t} as $n \to \infty$.*

Obviously there is nothing special about choosing the first interval in the above corollary. This result leads us to the following definition. Given a sequence $\{\alpha_n\}_{n=1}^{\infty}$ in $[0,1]$, for fixed N let $\beta_{1,N}, \beta_{2,N}, ..., \beta_{N,N}$ denote the sequence

$\alpha_1, \alpha_2, ..., \alpha_N$ reordered in increasing order. Define the **first consecutive spacing measure** to be

$$\mu(N) = \frac{1}{N-1} \sum_{j=1}^{N-1} \delta_{N(\beta_{j+1}-\beta_j)}$$

where δ is the Dirac $delta$-function. Note that this is normalized to be a probability measure. From the above corollary, we expect that if the α_n "behave like" random numbers then $\mu(N) \to e^{-x}dx$ as $N \to \infty$.

Slightly more generally, one can define the $\mathbf{k^{th}}$ **consecutive spacing measure** by

$$\mu_k(N) = \frac{1}{N-k} \sum_{j=1}^{N-k} \delta_{N(\beta_{j+k}-\beta_j)}.$$

With a little more work one sees that for random numbers one expects that $\mu_k(N) \to \frac{x^k}{k!}e^{-x}dx$ as $N \to \infty$.

We now restrict our attention to the sequences $\{n^2\alpha\}$. For such a sequence, let $\mu_k(N, \alpha)$ denote the k^{th} consecutive spacing measure. It is conjectured in [RSZ] that for almost all α (in the sense of Lebesgue), $\mu_k(N, \alpha) \to \frac{x^k}{k!}e^{-x}dx$. (In fact, the authors make a slightly stronger claim, expressed in terms of m-level correlations. For computational reasons this claim is more difficult to test, and we shall not discuss it further in this paper.)

We note that it is not true that for any irrational α, $\mu_k(N, \alpha) \to \frac{x^k}{k!}e^{-x}dx$. Problems arise for α with very good rational approximations. For example:

Proposition A.2.12. *Let α be an irrational number such that there is a sequence of rational numbers p_n/q_n with $|\alpha - p_n/q_n| < a_n/q_n^3$, where $a_n \to 0$ as $n \to \infty$. Then, there is a sequence of integers $N_j \to \infty$ such that $\mu_1(N_j, \alpha)$ does not converge to $e^{-x}dx$.*

Proof: What we shall actually show is that either $\mu_1(N_j, \alpha)$ does not converge or it converges to a measure supported on the integers.

Fix n with $a_n < 1/2$. We let $N_n = q_n$ and consider $\mu(\alpha, N_n)$. For $k < N_n$, $\left|\{k^2\alpha\} - \{k^2\frac{p_n}{q_n}\}\right| \leq \frac{a_n}{q_n}$. Let $\beta_1, ..., \beta_N$ be $\{1^2\alpha\}, \{2^2\alpha\}, ..., \{N^2\alpha\}$ written in increasing order, and let $\gamma_1, ..., \gamma_N$ be $\{1^2\frac{p_n}{q_n}\}, \{2^2\frac{p_n}{q_n}\}, ..., \{N^2\frac{p_n}{q_n}\}$ written in increasing order. Then, for $j < N_n$, since $a_n < 1/2$, if $\beta_j = \alpha_l$ then $\gamma_j = \{l^2\frac{p_n}{q_n}\}$. Thus, $|N_n(\beta_{j+1} - \beta_j) - N_n(\gamma_{j+1} - \gamma_j)| \leq 2a_n$. Since the $a_n \to 0$, the result now follows immediately. □

We note that the set of irrationals which satisfy the conditions of the proposition has measure zero. In fact, for any $\epsilon > 0$, the set of irrationals with infinitely many rational approximations p_n/q_n with $|\alpha - p_n/q_n| < 1/q_n^{2+\epsilon}$ has measure zero; see [Ki].

(Unfortunately, in a topological sense, "almost all" irrationals have approximations better than $1/q^k$ for any k. That is, the set of irrationals that do not is a countable union of nowhere dense sets. It is not hard to write down irrationals with these properties; an example is $\sum_{n=1}^{\infty} 10^{-n!}$. In numerical tests, such numbers behave just like rational numbers, so we ignore them below.)

Ironically, it is proved in [RSZ] that for most irrationals with approximations as good as in the above proposition, there is *some* sequence of integers N_j along which $\mu_k(N_j, \alpha) \to \frac{x^k}{k!}e^{-x}dx$.

It is interesting to compare the $\{n^2\alpha\}$, where behavior seems to be quite random, to the $\{n\alpha\}$ case, where it is not. As we saw, the sequence $\{n\alpha\}_{n=1}^{\infty}$ is equidistributed. However, in this case, for any α, for fixed N, $\beta_{i+1,N} - \beta_{i,N}$ takes on at most three different values. The following table shows the three values of $\beta_{i+1,N} - \beta_{i,N}$ for $\alpha = \sqrt{2}$ and a few different N:

N	values		
5	.171573	.242641	
10	.171573	.071068	.100505
15	.100505	.071068	.029437
20	.071068	.041631	.029437
30	.041631	.029437	.012193

The reason is that the consecutive spacings are determined by the "best approximations of the second kind" (in the language of [Ki]). (The rational number p/q is a best approximation of the second kind to α if $|q\alpha - p| < |q'\alpha - p'|$ for any p', q' with $q' < q$.) Writing down the proof that $\beta_{i+1,N} - \beta_{i,N}$ takes on only three values is somewhat cumbersome, though not difficult. The values $\beta_{i+1,N} - \beta_{i,N}$ can take on are either of the form $|q\alpha - p|$, where p/q is a best approximation to α or $|\alpha - p + q'\alpha - p'|$ where p/q and p'/q' are successive best approximations to α. All best approximations come from the continued fraction convergents to α. The reader should compare the following table, of convergents to $\sqrt{2}$ with the

previous one:

n	p_n/q_n	$\lvert q_n\alpha - p_n\rvert$	$\lvert q_n\alpha - p_n + q_{n-1}\alpha - p_{n-1}\rvert$
1	1	.414213	
2	$3/2$	.171573	.242641
3	$7/5$	.071068	.100505
4	$17/12$	.029437	.041631
5	$41/29$	.012193	.017244

A.3 Computations

The new content of this paper consists of several computational tests of the conjectures in [RSZ]. Although some computational tests were performed in [CGI], the emphasis in that paper was physical rather than mathematical, and tests were not particularly extensive. The "idea" behind most of the computations that I performed to test the conjectures is simply to compare cumulative distribution function of $\mu_k(\alpha, N)$ with the expected result. (Very often I only considered the case $k = 1$.) I worked primarily with the irrationals $\sqrt{2}$, e, and π, as well as an irrational, which we will call η, constructed in [RSZ] which is known to be poorly behaved (not Poissonian in the stronger sense that they consider) but which seems, somewhat surprisingly, to be well behaved in the contexts we consider.

Let us begin to get a sense of convergence by looking at graphs of expected and actual cumulative distribution functions for $\mu_1(\alpha, N)$ for various α and N. Firstly, for $\sqrt{2}$, we notice relatively rapid convergence:

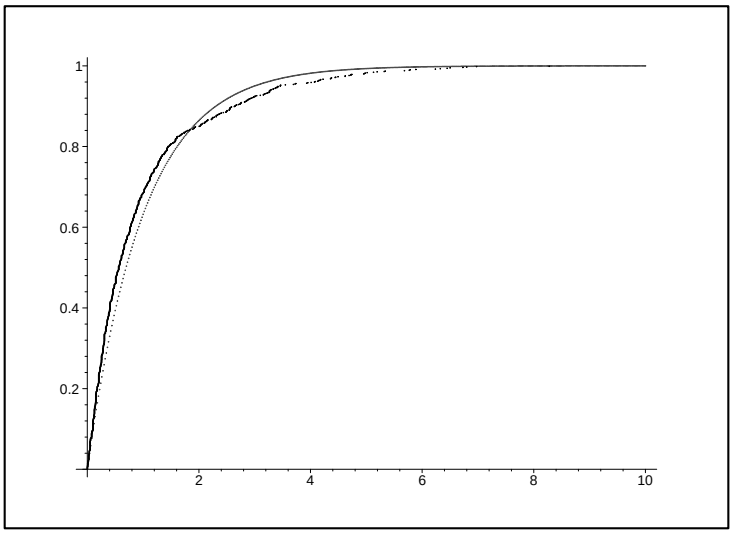

Figure 1: $\alpha = \sqrt{2}$, $N = 1,000$

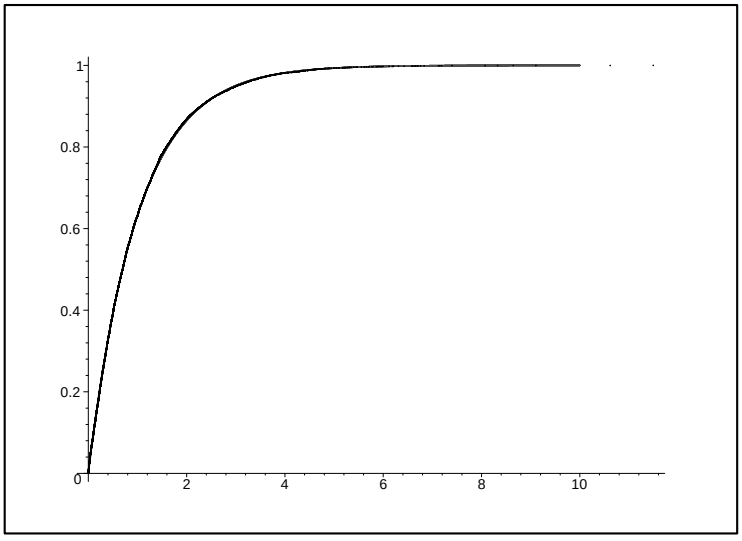

Figure 2: $\alpha = \sqrt{2}$, $N = 10,000$

Notice that by 10,000 points the two graphs are essentially indistinguishable. For π, convergence seems to be even faster:

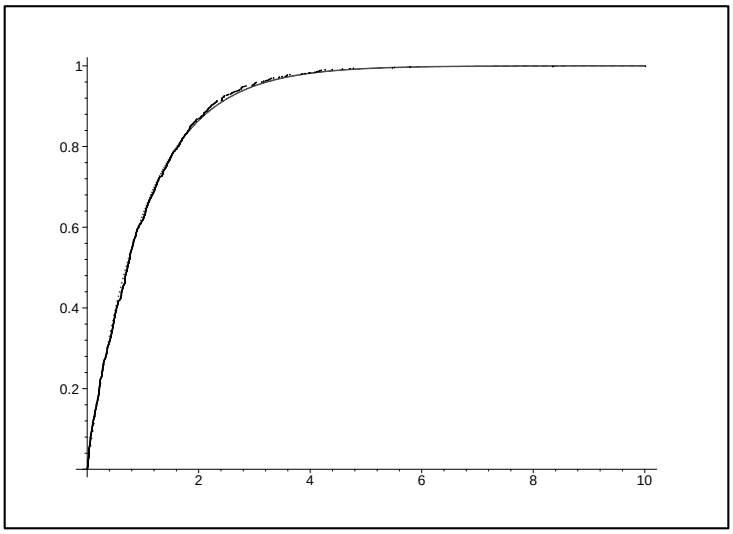

Figure 3: $\alpha = \pi$, $N = 1,000$

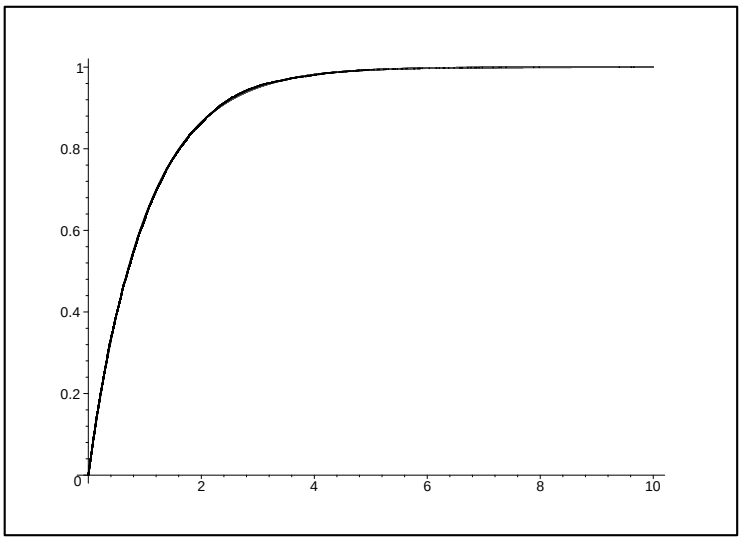

Figure 4: $\alpha = \pi$, $N = 10,000$

For η, the first few pictures make the issue a lot less clear. It is not until roughly $N = 50,000$ that it becomes relatively clear that η is behaving as it ought.

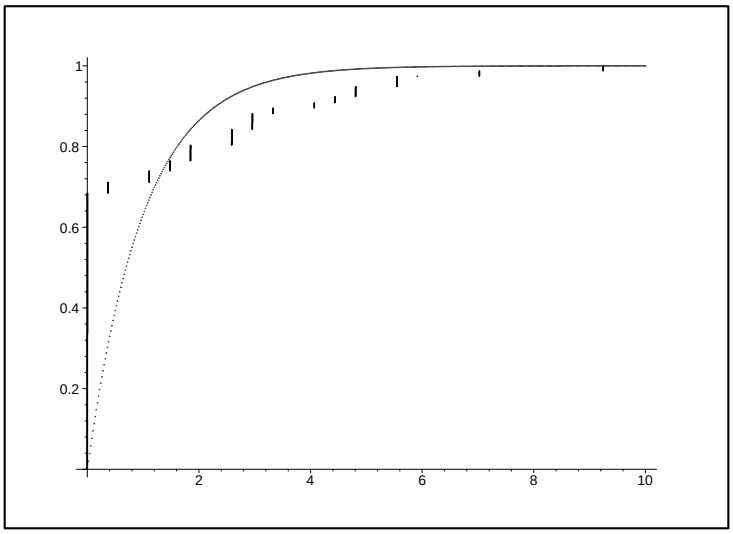

Figure 5: $\alpha = \eta$, $N = 1,000$

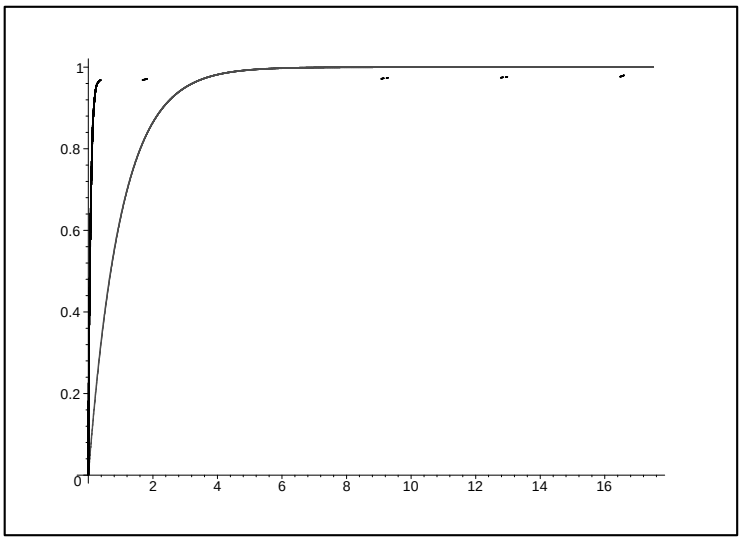

Figure 6: $\alpha = \eta$, $N = 10,000$

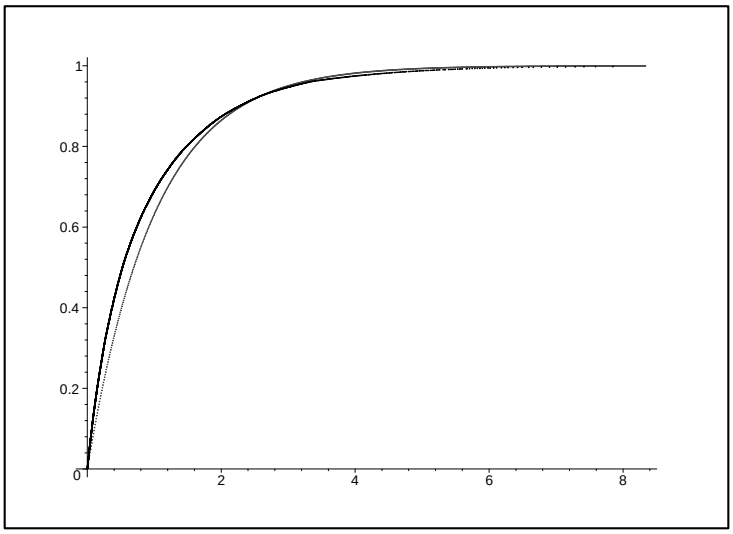

Figure 7: $\alpha = \eta$, $N = 100,000$

Of course, while looking at pictures is pleasant, one can only extract a certain amount of information from them. In order to make more precise comparisons, and to be able to take larger N, we define a numerical measure of discrepancy, which we shall denote $disc(\alpha, N)$. The idea is that $disc$ should measure the difference between the cumulative distribution functions for $\mu_1(\alpha, N)$ and $e^{-x}dx$. For computational purposes, we define $disc$ to be the maximum difference between the two cumulative distribution functions that occurs at one of the points $N(\beta_{j+1} - \beta_j)$. With this definition, $disc$ can thus be calculated in $N \log N$ time, but still captures the difference between the two cumulative distribution functions to within at most $1/N$.

Following is a table of $disc$ for various α and N

	$\frac{35}{17}$	$\sqrt{2}$	e	π	η
$1,000$	.991	.077813	.0392078	.029832	.68223
$10,000$	.9991	.007698	.0119735	.007069	.74749
$50,000$	.99982	.012865	.0068190	.007297	.093145
$75,000$	.99988	.012728	.0078838	.005619	.041912
$100,000$	.99991	.010729	.0070657	.004573	.021142

From the table, one notices several things. Firstly, in all cases except 35/17 (which is just included for comparison), the discrepancy seems to be going to zero. For η it seems to be doing so more slowly than for the other three. It might be worthwhile increasing N even further to make sure that in fact $disc$ is going to zero.

(Doing so is not feasible with my current computing resources.) Convergence seems to be slightly faster for e and π than for $\sqrt{2}$; that there is a difference is perhaps not surprising, given that e and π are very different kinds of numbers from $\sqrt{2}$, but the author does not feel competent to speculate on the precise reason.

Secondly, one notices that the decrease is not monotone. One idea proposed to explain this was that the jumps might be related to the continued fraction expansions. That is, for N near denominators of convergents, perhaps one finds that $disc$ is unexpectedly large. I tested for such patterns with $\sqrt{2}$. The first few convergents of $\sqrt{2}$ are 1, 3/2, 7/5, 17/12, 41/29, 99/70, 239/169, 577/408, 1393/985, 3363/2378, 8119/5741, 19601/13860, 47321/33461, and 114243/80782. Here is a data table for $\sqrt{2}$ with denominators of convergents in bold.

N	$disc$
408	.1091138
552	.1298027
697	.1240015
841	.1018524
985	.0796401
1333	.0516779
1682	.0338395
2030	.0199573
2378	.0169253
3219	.0118520
4060	.0154372
4900	.0160554
5741	.0124257
7771	.0094704
9801	.0095135
11830	.0083271
13860	.0084538
18760	.0110632
23661	.0078211
28561	.0080084
33461	.0101271
45291	.0120665
57122	.0123425
68952	.0123332
80782	.0121800

I find no particular relation between closeness to the denominator of a convergent and $disc$, suggesting that other factors, probably complicated and perhaps figments of the definitions, are at work. Still, this might be worth looking at further.

Another question of interest is whether the μ_k are independent for different k. For example, one might consider whether the distribution of pairs $(N(\beta_{k+1} - \beta_k), N(\beta_{k+2} - \beta_k))$ is, in the limit, the product of the two one dimensional distributions. (Equivalently, one might ask about the distribution of $(N(\beta_{k+1} - \beta_k), N(\beta_{k+2} - \beta_{k+1}))$; this is actually what I investigated.) The same question can be asked about correlations between more than two terms. The expectation is that for almost all irrationals, the distribution should approach the product of the one dimensional distributions, and my data appears to support this hypothesis in all of the cases that I tested.

In higher dimensions I was, unfortunately, unable to be quite as clever with defining $disc$. The best way that I could think of to calculate $disc$ for dimensions greater than 2 took on the order of N^2 computations, and the $N \log N$ algorithm that I thought of for 2 dimensions was sufficiently complicated that it didn't seem worth implementing. Speed, thus, became an issue, and I used a combination of Maple and C code. This worked satisfactorily.

Here is a data table of $disc$ for the two dimensional case, for $\alpha = \sqrt{2}$ and $\alpha = \eta$.

	$\sqrt{2}$	η
$1,000$	$.091572$	$.650591$
$10,000$	$.009135$	$.860561$
$15,000$	$.012914$	$.762592$
$20,000$	$.011899$	$.590495$
$30,000$	$.008189$	$.307924$
$50,000$	$.012719$	$.138787$

The data is, unfortunately, not quite as convincing as before, and more experiments would be in order.

Here is another data table, for the five dimensional case.

	$\sqrt{2}$	η
1000	.072419	0.097487
5000	.015055	0.426626
10000	.010035	0.542570
15000	.011936	0.576099
20000	.008507	0.594172
25000	.010824	0.582399
30000	.010747	0.516590
35000	.008606	0.415605
40000	.010173	0.338313
45000	.009174	0.284612

This data actually looks a bit more convincing of convergence than the other, but not a lot. It looks entirely possible that $disc$ for $\sqrt{2}$ hovers around .01. More tests would definitely be in order.

A.4 For Those Who Come After

The idea of the "undergraduate mathematics laboratory" is for students to do experiments that other students will continue. This section of the paper, instead of presenting results, discusses briefly the methods I used and difficulties I encountered, to help those who come after, and briefly mentions possible areas for further study. More specific information about what programs I used and how they work can be found in the electronic "readme" and the comments in the programs.

The first issue is always how many digits to keep. I kept 30. If one goes only up to $N = 100,000$ (10^5), $N^3 \leq 10^{15}$, and thus 30 should certainly be enough. When working with C instead of Maple, my variables were double precision floating points. This might cause concern about digits, but I only used C after already calculating the $\beta_{n+1} - \beta_n$.

The rate limiting step in the calculations is calculating $disc$. This was not a big problem for the one dimensional case, where running time with $N = 100,000$ was a few minutes. (It was much longer until I realized that I should convert α into a floating point number before doing the calculations in Maple.) For higher dimensional cases, however, Maple became woefully inadequate, taking several minutes for $N = 5,000$ and forever for much higher N. I thus switched to C, which was much, much faster. The largest cases I did ($N = 45,000$) took a

few minutes in C. The other limit is memory; for $N = 100,000$ I was using 250 megabytes of memory to store the arrays. Going much further, thus, is not feasible at the moment.

One thing the next person to work on this problem should do, of course, is to fill in the data tables I presented a bit more. This should not be hard; the programs are already in place. It would be nice to have the several variable discrepancy for more α and for higher N. In particular, it might be nice to compare several algebraic numbers of various degrees, considering e and π throughout, and maybe looking at several different possible choices of η.

At the level at which I worked, not a lot of complicated programming was required; this is nice. Unfortunately, the next thing to look at, I think, is what are referred to in [RSZ] as "M-level correlations." Computing these by brute force requires $O(N^M)$ computations, which is rather unfriendly. I am not sure if this can be reduced. Probably to have a chance of the computations finishing one should use C instead of Maple. Even if it is only possible to take M a few thousand, it would still be worthwhile, I think.

I wish my successors the best of luck, and hope that this paper has been of some use.

Bibliography

[CGI] G. Casati, I. Guarneri, and F. M. Izrailev, *Statistical Properties of the Quasi-Energy Spectrum of a Simple Integrable System*, Phys. Lett. A 124 (1987), $263 - 266$.

[Ca] J. W. S. Cassels, *An Introduction to Diophantine Approximation*, Cambridge University Press, London 1957.

[Da] H. Davenport, *Multiplicative Number Theory,* 2*nd edition*, Graduate Texts in Mathematics **74**, Springer-Verlag, New York, 1980, revised by H. Montgomery.

[Fe] W. Feller, *An Introduction to Probability Theory and its Applications*, Vol. II. Second edition. John Wiley & Sons, Inc., New York-London-Sydney 1971.

[GT] A. Granville and T. Tucker, *It's as easy as* abc, Notices of the AMS, volume 49, number 10 (November 2002).

[HS] M. Hindry and J. Silverman, *Diophantine geometry: An introduction*, Graduate Texts in Mathematics, vol. 201, Springer, New York, 2000.

[HW] G. Hardy and E. Wright, *An Introduction to the Theory of Numbers*, fifth edition, Oxford Science Publications, Clarendon Press, Oxford, 1995.

[Ki] A. Y. Khinchin, *Continued Fractions*, Third Edition, The University of Chicago Press, Chicago 1964.

[La] S. Lang, *Introduction to Diophantine Approximations*, Addison-Wesley, Reading, 1966.

[LT] S. Lang and H. Trotter, *Continued fractions for some algebraic numbers*, J. Reine Angew. Math. **255**, 1972, $112 - 134$.

[Le] P. L$\acute{e}$vy, *Sur les lois de probabilit$\acute{e}$ dont dependent les quotients complets et incomplets d'une fraction continue*, Bull. Soc. Math., **57**, 1929, $178-194$.

[Ro] Roth, *Rational approximations to algebraic numbers*, Mathematika 2, 1955, $1-20$.

[Ru] W. Rudin, *Principles of Mathematical Analysis*, third edition, International Series in Pure and Applied Mathematics, McGraw-Hill Inc., New York, 1976.

[RSZ] Z. Rudnick, P. Sarnak, and A. Zaharescu, *The Distribution of Spacings Between the Fractional Parts of* $n^2\alpha$, Invent. Math. 145 (2001), no. 1, $37-57$.