Number-theoretic applications of ergodic theory

Author:
S.D. Ramawadh

Supervisor:
Dr. M.F.E. de Jeu

Bachelor thesis
Leiden University, 18 July 2008

Contents

1 Ergodic theory

This thesis is about number-theoretic applications of ergodic theory. In this chapter, we will study the parts of ergodic theory which will be necessary for us when we will consider its applications. This chapter is intended for anyone who has some knowledge of measure theory.

1.1 Measure-preserving transformations

A important class of transformations on probability spaces are the measurable transformations:

Definition 1.1. Let $(X,\mathcal{U},\mu)$ and $(Y,\mathcal{V},\lambda)$ be two arbitrary probability spaces. A transformation $T:(X,\mathcal{U},\mu)\to(Y,\mathcal{V},\lambda)$ is called measurable if $A\in\mathcal{V}\Rightarrow T^{-1}A\in\mathcal{U}$.

The definition of a measure-preserving transformation should not be too surprising:

Definition 1.2. Let $(X,\mathcal{U},\mu)$ and $(Y,\mathcal{V},\lambda)$ be two arbitrary probability spaces. A transformation $T:(X,\mathcal{U},\mu)\to(Y,\mathcal{V},\lambda)$ is called measure-preserving if T is measurable and $\lambda(A)=\mu\left(T^{-1}A\right)$ for all $A\in\mathcal{V}$.

Measure-preserving transformations may be defined in a rather simple way, but measure-preserving transformations on the same probabilty spaces already have a very strong property:

Theorem 1.3 (Poincaré's Recurrence Theorem). Let $T:(X,\mathcal{U},\mu)\to(X,\mathcal{U},\mu)$ be a measure-preserving transformation. Let $A\in\mathcal{U}$ with $\mu(A)>0$, then almost all points of A return infinitely often to A under positive iteration of T.

Proof. Note that the above statement can be formulated as follows: there exists a $B\subset A$ with $\mu(B)=\mu(A)>0$ such that for all $x\in A$ there exists a sequence of natural numbers $n_1<n_2<n_3<\cdots$ with $T^{n_i}(x)\in B$ for all i.

We will first prove the existence of such a set B and a sequence. Let $A\in\mathcal{U}$ with $\mu(A)>0$ and define for $N\geq 0$ the set $A_N=\cup_{n=N}^{\infty}T^{-n}A$. Then $\cap_{n=0}^{\infty}A_n$ is exactly the set of points of X that appear infinitely often in A under positive iteration of T. The set $B=A\cap(\cap_{n=0}^{\infty}A_n)$ is exactly the set of points in A that return infinitely often in A. For each point $x\in B$ we can find a sequence of natural numbers $(n_i)_{i=1}^{\infty}$ such that $T^{n_i}(x)\in A$ for all i (this follows from the way we defined B). However, because $T^{n_j-n_i}\left(T^{n_i}(x)\right)\in A$ for all i,j, we find that $T^{n_i}(x)\in B$ for all i.

Finally, we need to show that $\mu(A)=\mu(B)$. We know that $\mu(A_n)=\mu(A_{n+1})$ for all n, because T is measure-preserving. It follows that $\mu(A_0)=\mu(A_n)$ for all n. Since we also know that $A_0\supset A_1\supset\cdots$, it follows that $\mu\left(\cap_{n=0}^{\infty}A_n\right)=\mu(A_0)$. Note that $A\subset A_0$, from which it now follows that $\mu(B)=\mu\left(A\cap(\cap_{n=0}^{\infty}A_n)\right)=\mu(A\cap A_0)=\mu(A)$. □

Note that, although it doesn't appear that way in the proof, it is important that μ is finite, as the next example demonstrates:

Example 1.4. Consider the set of integers and the measure μ which gives each number a measure of 1 and consider the measure-preserving transformation given by $T(x)=x+1$. If we let $A=\{0\}$, then $\mu(A)=1>0$ and each of the sets A_n (as defined in the proof of Theorem 1.3) has infinite measure. It is obvious that the theorem is incorrect in this case. The proof is also incorrect: since $\cap_{n=0}^{\infty}A_n=\emptyset$, we find $0=\mu\left(\cap_{n=0}^{\infty}A_n\right)\neq\mu(A_0)=\infty$.

1.2 Ergodicity

Let T be a measure-preserving transformation on the probability space $(X, \mathcal{U}, \mu)$. If there is a measurable set A with the property $T^{-1}A = A$, then it is also true that $T^{-1}(X \backslash A) = X \backslash A$. Therefore, we may consider the restriction of T to A (notation: $T|_A$). If either $\mu(A) = 0$ or $\mu(A) = 1$, then the transformation is not simplified in a significant way (neglecting null sets is a common practice in measure theory). Because of this, it may be a good idea to study those measure-preserving transformations that can't be simplified in this way. Such transformations are called ergodic:

Definition 1.5. Let T be a measure-preserving transformation on the probability space $(X, \mathcal{U}, \mu)$. Then T is called ergodic if the only measurable sets with $T^{-1}A = A$ satisfy either $\mu(A) = 0$ or $\mu(A) = 1$.

Ergodic theory had its origins in statistical mechanics. Suppose that a dynamical system describes a path $\gamma(t)$ in the phase-space and that $f(\gamma(t))$ is the value of a certain quantity. With experiments we can only determine the so-called time mean $\frac{1}{n}\sum_{k=0}^{n-1} f(\gamma(t_k))$, while we can only attempt to calculate the space mean $\int_X f \, d\mu$. The following theorem, Birkhoff's Ergodic Theorem, is the most important theorem in ergodic theory and relates the two previously mentioned means:

Theorem 1.6 (Birkhoff's Ergodic Theorem). Let T be an ergodic transformation on the probability space $(X, \mathcal{U}, \mu)$. Then, for all $f \in L^1(X, \mathcal{U}, \mu)$ we have:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} f\left(T^k(x)\right) \quad = \quad \int_X f \, d\mu$$

almost everywhere.

For the proof of this theorem we refer to [5, Theorem 1.14]. Note that ergodic transformations are absolutely not artificial as the following theorem implies:

Theorem 1.7. Let X be a compact metrisable space and let T be a continuous transformation on X. Then there exists a measure μ so that T is ergodic with respect to this measure.

The proof of this theorem can be found in [5, Theorem 6.10].

A problem that arises naturally is the one of identifying ergodic transformations. The next theorem is one of many that we can use to prove ergodicity. Before we get to this theorem, we will first prove a helpful lemma.

Lemma 1.8. Let $(X, \mathcal{U}, \mu)$ be a probability space and T an ergodic transformation on this space. Then all sets $A \in \mathcal{U}$ with $\mu\left(T^{-1}A \triangle A\right) = 0$ satisfy either $\mu(A) = 0$ or $\mu(A) = 1$.

Proof. Let $A \in \mathcal{U}$ and $\mu\left(T^{-1}A \triangle A\right) = 0$. For all $n \geq 0$ we have $\mu\left(T^{-n}A \triangle A\right) = 0$, because $T^{-n}A \triangle A \subset \cup_{i=0}^{n-1} T^{-i}\left(T^{-1}A \triangle A\right)$. Let $A_\infty = \cap_{n=0}^{\infty} \cup_{i=n}^{\infty} T^{-i}A$. The sets $\cup_{i=n}^{\infty} T^{-i}A$ decrease for increasing n and each of them has measure equal to A (because $\mu\left(\cup_{i=n}^{\infty} T^{-i}A \triangle A\right) \leq \mu\left(T^{-n}A \triangle A\right) = 0$), so it follows that $\mu\left(A_\infty \triangle A\right) = 0$, and thus $\mu(A_\infty) = \mu(A)$. Since we have $T^{-1}A_\infty = \cap_{n=0}^{\infty} \cup_{i=n}^{\infty} T^{-(i+1)}A = \cap_{n=0}^{\infty} \cup_{i=n+1}^{\infty} T^{-i}A = A_\infty$ and because T is ergodic, it now follows that $\mu(A_\infty) = 0$ of $\mu(A_\infty) = 1$. Since $\mu(A_\infty) = \mu(A)$, it now follows that either $\mu(A) = 0$ of $\mu(A) = 1$. □

Theorem 1.9. Let $(X, \mathcal{U}, \mu)$ be a probability space and T a measure-preserving transformation on this space. Let $p \geq 1$ be an integer. The following statements are equivalent:

1. T is ergodic.
2. If $f \in L^p(\mu)$ and $(f \circ T)(x) = f(x)$ for almost all $x \in X$, then f is constant almost everywhere.

Proof. The proof is split up in two parts.
$1 \Rightarrow 2$: Let T be ergodic and suppose that f is measurable with $f \circ T = f$ almost everywhere. We may assume that f is real-valued. Define for $k \in \mathbb{Z}$ and $n > 0$ the set $X(k,n) = f^{-1}\left(\left[\frac{k}{2^n}, \frac{k+1}{2^n}\right)\right)$. We have $T^{-1}X(k,n) \triangle X(k,n) \subset \{x \ : \ (f \circ T)(x) \neq f(x)\}$ and thus $\mu\left(T^{-1}X(k,n) \triangle X(k,n)\right) = 0$. It now follows (Lemma 1.8) that $\mu(X(k,n)) = 0$ or $\mu(X(k,n)) = 1$. For each fixed n, $\cup_{k\in\mathbb{Z}} X(k,n) = X$ is a disjoint union and thus there is exactly one k_n with $\mu(X(k_n,n)) = 1$. Let $Y = \cap_{n=1}^{\infty} X(k_n,n)$, then $\mu(Y) = 1$ and f is constant in Y. Therefore, f is constant almost everywhere. Since every member of $L^p(\mu)$ is measurable, the result follows.

$2 \Rightarrow 1$: Suppose $T^{-1}A = A$ with $A \in \mathcal{U}$. The characteristic function χ_A is measurable, so $\chi_A \in L^p(\mu)$ for all $p \geq 1$. We also have $(\chi_A \circ T)(x) = \chi_A(x)$ for all $x \in X$ so χ_A is constant almost everywhere. This means that either $\chi_A = 0$ almost everywhere or $\chi_A = 1$ almost everywhere. It now follows that $\mu(A) = \int_X \chi_A \, d\mu$ is either 0 or 1. Therefore, T is ergodic. □

1.3 Unique ergodicity

Let X be a set and T an ergodic transformation on this set. While this implies that there is an ergodic measure μ, it is not necessarily the only ergodic measure with respect to T.

Example 1.10. Consider the unit circle S^1, viewed as the interval $[0,1)$ with its endpoints joined together. Consider $T : S^1 \to S^1$ given by $T(x) = 2x \pmod 1$. This transformation is better known as the doubling map. One can show that it is ergodic using Theorem 1.9 (see [5, p.30, example (4)] for details). An ergodic measure is the Haar-Lebesgue measure, which measures the lengths of arcs. However, another ergodic measure is the Dirac measure δ_0, which assigns a set A measure 1 only if $0 \in A$ (or else the set has measure zero).

The following definition should hardly be surprising.

Definition 1.11. Let X be a set and $T : X \to X$ a transformation. We call T uniquely ergodic if there is exactly one ergodic measure.

The following theorem reveals an important property of unique ergodicity:

Theorem 1.12. Let T be a continuous transformation on a compact metrisable space X. The following four statements are equivalent:

1. T is uniquely ergodic.
2. There is a probability measure μ such that for all continuous functions f and all $x \in X$:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} f\left(T^k(x)\right) \quad = \quad \int_X f \, d\mu.$$

3. For every continuous function f, $\frac{1}{n}\sum_{k=0}^{n-1} f\left(T^k(x)\right)$ converges uniformly to a constant.

4. For every continuous function f, $\frac{1}{n}\sum_{k=0}^{n-1} f\left(T^k(x)\right)$ converges pointwise to a constant.

The proof of this theorem can be found in [5, Theorem 6.19]. Note the subtle yet important difference when compared to the ergodic case: the average converges for all $x \in X$ instead of almost all $x \in X$. However, we did assume that X is compact and metrisable, but this is the case in many practical situations, and so this assumption will not be a huge problem.

1.4 Mixing and weak-mixing

The following theorem is a corollary of Birkhoff's Ergodic Theorem:

Theorem 1.13. Let $(X,\mathcal{U},\mu)$ be a probability space and let $T : X \to X$ be measure-preserving. Then T is ergodic if and only if for all $A, B \in \mathcal{U}$ we have:

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n-1} \mu\left(T^{-k}A \cap B\right) \quad = \quad \mu(A)\mu(B).$$

Proof. Suppose T is ergodic. If we let $f = \chi_A$ and multiply both sides in the equality of Theorem 1.6 with χ_B (where χ is the characteristic function), then we find:

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n-1} \chi_A\left(T^k(x)\right)\chi_B \quad = \quad \mu(A)\chi_B.$$

The left-hand side cannot exceed 1, so we can use the Dominated Convergence Theorem and find:

$$\begin{aligned}
\int \lim_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n-1} \chi_A\left(T^k(x)\right)\chi_B \, d\mu &= \int \mu(A)\chi_B \, d\mu; \\
\lim_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n-1} \left(\int \chi_A\left(T^k(x)\right)\chi_B \, d\mu\right) &= \mu(A)\int \chi_B \, d\mu; \\
\lim_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n-1} \mu\left(T^{-k}A \cap B\right) &= \mu(A)\mu(B).
\end{aligned}$$

Conversely, suppose that the convergence holds. Let $A \in \mathcal{U}$ with $T^{-1}A = A$, then we find:

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n-1} \mu\left(A\right) \quad = \quad \left(\mu(A)\right)^2.$$

It now follows that $\mu(A) = \left(\mu(A)\right)^2$, so that $\mu(A) = 0$ or $\mu(A) = 1$. Therefore, T must be ergodic. □

By changing the method of convergence in Theorem 1.13 we get the definitions of respectively weak-mixing and mixing. In order to define weak-mixing, we need the following definition:

Definition 1.14. A subset $M \subset \mathbb{Z}_{\geq 0}$ is called a subset of density zero if:

$$\lim_{n\to\infty}\left(\frac{\text{cardinality}(J \cap \{0,1,\ldots,n-1\})}{n}\right) = 0.$$

Now we can define weak-mixing and mixing:

Definition 1.15. If T is a measure-preserving transformation on the probability space $(X,\mathcal{U},\mu)$, then T is called weak-mixing if for all $A, B \in \mathcal{U}$ there exists a subset $J(A,B) \subset \mathbb{Z}_{\geq 0}$ of density zero such that:

$$\lim_{J(A,B)\not\ni n\to\infty} \mu\left(T^{-n}A \cap B\right) \quad = \quad \mu(A)\mu(B).$$

Definition 1.16. If T is a measure-preserving transformation on the probability space $(X,\mathcal{U},\mu)$, then T is called mixing if for all $A, B \in \mathcal{U}$:

$$\lim_{n\to\infty} \mu\left(T^{-n}A \cap B\right) \quad = \quad \mu(A)\mu(B).$$

The difference between ergodicity and (weak-)mixing becomes clear when we look at the previous definitions and theorem from a practical point of view. Suppose we have a large bath filled with water and that we also have an amount of painting powder in your favourite colour. Now suppose that we add the painting powder to the water in a subset B. If we mix the water in the bath, then the way the colour will spread depends on the way we mix the water. Suppose that we mix the water in an ergodic way (i.e., the flow of the water can be described by an ergodic transformation), then we see from Theorem 1.13 by substituting $B = X$ that eventually the colour will be evenly spread. However, the theorem says that this is convergence in mean: the colour will not behave as smoothly as you probably would like to. Now suppose that the water gets stirred in a mixing way. It follows from the definition that the colour spreads nicely and that the intensity converges asymptotically. The weak-mixing case is much like the mixing case, except that the colour density may "misbehave" once in a while.

The following theorem relates the notions of ergodicity and (weak-)mixing:

Theorem 1.17. Let T be a measure-preserving transformation on the probability space $(X,\mathcal{U},\mu)$:

(a) If T is mixing, then T is weak-mixing.

(b) If T is weak-mixing, then T is ergodic.

Proof. We will prove the statements one by one:

(a) This follows trivially: take $J(A,B) = \emptyset$.

(b) Consider the sequence $(a_n)_{n=0}^{\infty}$ given by $a_n = \mu\left(T^{-n}A \cap B\right) - \mu(A)\mu(B)$. We know that $|a_n| \leq 1$ holds for all n. Let $\alpha_{J(A,B)}(n)$ denote the cardinality of $J(A,B) \cap \{0,1,\ldots,n-1\}$. Let $\epsilon > 0$, then there is an N_ϵ such that for all $n \geq N_\epsilon$ and $n \notin J(A,B)$ we have $|a_n| < \epsilon$ and for all $n \geq N_\epsilon$ we have $\left(\alpha_{J(A,B)}(n)/n\right) < \epsilon$. We find:

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n-1}|a_n| \quad &= \quad \lim_{n\to\infty}\left(\frac{1}{n}\sum_{n\in J(A,B)\cap\{0,\ldots,n-1\}}|a_n| + \frac{1}{n}\sum_{n\notin J(A,B)\cap\{0,\ldots,n-1\}}|a_n|\right)\\ &< \quad \frac{1}{n}\alpha_{J(A,B)}(n) + \epsilon \;<\; 2\epsilon.\end{aligned}$$

Now let $\epsilon \to 0$ to find that the average of the sequence $(a_n)_{n=0}^{\infty}$ converges to zero. This implies $\lim_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n-1}\mu\left(T^{-k}A \cap B\right) = \mu(A)\mu(B)$. By Theorem 1.13, T must be ergodic. □

The statements in Theorem 1.17 are all one-way implications. One can show that there exists a weak-mixing transformation which is not mixing [5, p. 40]. However, we will not discuss it here since it is beyond the scope of this thesis. It is much easier to show that there is an ergodic transformation which is not weak-mixing.

Example 1.18. Consider the unit circle S^1, viewed as the interval $[0, 1)$ with its endpoints joined together. Consider the transformation $T : S^1 \to S^1$ given by $T(x) = x + \alpha \pmod 1$ with α irrational. One can prove that this transformation is ergodic. In fact, we will do so in Chapter 2 (Corollary 2.3), but let's assume for now that T is ergodic. Intuitively, it is clear that this transformation is not weak-mixing, because any arc on the unit circle will remain an arc whenever we apply T (T rotates points over an angle α). Therefore, the arc will never spread over the whole unit circle and thus T cannot be weak-mixing.

Sometimes, it may be useful to consider the product of transformations:

Definition 1.19. Suppose that S is a measure-preserving transformation on $(X, \mathcal{U}, \mu)$ and T is a measure-preserving transformation on $(Y, \mathcal{V}, \lambda)$. The direct product of S and T is the measure-preserving transformation on the probability space $(X \times Y, \mathcal{U} \times \mathcal{V}, \mu \times \lambda)$ given by $(S \times T)(x, y) = (S(x), T(y))$.

A natural question that rises is: will $S \times T$ inherit the property of ergodicity or (weak-)mixing if S or T has it (and vice versa)? The following theorem answers this question:

Theorem 1.20. Let S and T be measure-preserving transformations:

(a) If one of S and T is weak-mixing and the other is ergodic, then $S \times T$ is ergodic.

(b) If $S \times T$ is ergodic for each ergodic T, then S is weak-mixing.

(c) $S \times T$ is weak-mixing if and only if both S and T are weak-mixing.

(d) $S \times T$ is mixing if and only if both S and T are mixing.

The proof of statements (a), (b) and (c) can be found in [1, Theorem 4.10.6] and the proof of (c) in [2, Theorems 10.1.2 and 10.1.3].

2 First digits of powers

In this chapter we will encounter the first number-theoretic problem in which we will apply ergodic theory to solve it. After a short introduction to the problem we will look at the ergodic aspect of the problem.

2.1 Introduction

Take your favourite positive integer k and consider its powers k^n with $n \in \mathbb{N}$. With little effort, one can determine the last digits of each power. For example, if k ends in a 5 (i.e. $k \equiv 5 \bmod 10$), then all its powers will also end in a 5. We see that the problem of the last digits of powers can be solved in a straightforward way.

Instead, we will look at the following problem. Let k be any positive integer. Can we determine the first digits of the positive integer powers of k? If so, can we determine the relative frequency with which these digits will appear as first digits of k^n?

Example 2.1. We will illustrate a more generalized question as follows. We let $k = 7$ so that we will consider the powers of 7. How will the first digits be distributed among the numbers $1, \ldots, 9$? How will this distribution be affected when we multiply all powers with a constant $c = 2$? And how will this distribution be affected when we write all powers with respect to base $b = 5$ (i.e. the quinary system)? If we consider the first 50 powers of 7, then the distribution will be the following:

Table 1: The first digit problem for the first 50 powers of 7

	1	2	3	4	5	6	7	8	9
Powers of 7	15	8	7	5	4	3	3	2	3
Powers of 7 multiplied by 2	15	7	8	4	4	4	3	2	3
Powers of 7 in the quinary system	22	13	9	6	–	–	–	–	–

At first sight, it seems that in all cases the lower numbers will appear more often than the higher numbers. However, we only considered the first 50 powers, so we have no reason to expect that this will hold for all powers of 7. Also, note how multiplying all powers by 2 has very little effect in the distribution. Is this true for any combination of powers of k and any constant c?

Another question we will try to answer is the question on the simultaneous distribution. Consider the powers of two different positive integers, say k_1 and k_2. What would the simultaneous distribution of the first digits look like? If the simultaneous distribution turns out to be the product of the respective distributions, then this implies a statistical independence between the first digits of those powers.

We now formulate the question that we will try to answer in this chapter:

> Let $k_1, \ldots, k_n$, $p_1, \ldots, p_n$, $c_1, \ldots, c_n$ and $b_1, \ldots, b_n$ be positive integers with $b_i \geq 2$ for all i. What is the relative frequency with which for all i the numbers p_i will be the (string of) first digits of $c_i \cdot k_i^m$ ($m \in \mathbb{N}$) when both are written with respect to base b_i?

2.2 Translations on $\mathbb{T}^n$

For now, let's forget about the problem we posed in Section 2.1 and focus on a different problem. Its relevance will become clear in Section 2.3.

Consider the n-dimensional torus $\mathbb{T}^n$. It is the n-fold product of unit circles S^1. We can regard these either additively (the interval $[0,1)$ with its endpoints joined together, with addition of numbers modulo 1 as operation) or multiplicatively (the unit circle in $\mathbb{C}$ with multiplication as operation). Note that $\mathbb{T}^n$ is a topological group. Both representations are isomorphic because of the map $\phi : x \to e^{2\pi i x}$, and therefore we can use both representations in a way we like.

Consider the translation T_γ on $\mathbb{T}^n$. It is given by $T_\gamma : (x_1, \ldots, x_n) \to (x_1 + \gamma_1, \ldots, x_n + \gamma_n)$, where $\gamma \in \mathbb{R}^n$ and γ_i denotes the ith coordinate of γ. For which vectors γ is the transformation T_γ ergodic and what is the corresponding ergodic measure? The following theorem gives us the answer.

Theorem 2.2. The translation T_γ on $\mathbb{T}^n$ is uniquely ergodic if and only if $\gamma_1, \ldots, \gamma_n$ are linearly independent over $\mathbb{Q}$.

Proof. Throughout this proof we will consider $\mathbb{T}^n$ multiplicatively. The group homomorphisms $c : \mathbb{T}^n \to S^1$ can be written as $c_{m_1,\ldots,m_n}(x_1, \ldots, x_n) = e^{2\pi i(m_1 x_1 + \ldots + m_n x_n)}$ with $m_i \in \mathbb{Z}$. These maps will be called *characters*. The characters are eigenfunctions of the translation T_γ, because we have: $c_{m_1,\ldots,m_n}\left(T_\gamma(x_1, \ldots, x_n)\right) = e^{2\pi i(m_1(x_1+\gamma_1)+\ldots+m_n(x_n+\gamma_n))} = e^{2\pi i(m_1\gamma_1+\ldots+m_n\gamma_n)} c_{m_1,\ldots,m_n}(x_1, \ldots, x_n)$.

Suppose that the eigenvalue of T_γ is an integer, then this implies that T_γ is periodic. Let A be a sufficiently small set. Because of the periodicity of T_γ, the orbit of A under this transformation is a set which does not all of $\mathbb{T}^n$. Moreover, if A has measure $0 < \epsilon \ll 1$ (positive yet much smaller than 1), then the orbit of A under T_γ will be a set whose measure is also neither 0 or 1. By definition, T_γ cannot be ergodic. Therefore, we want $e^{2\pi i(m_1\gamma_1+\ldots+m_n\gamma_n)}$ not to be an integer. which is equivalent with saying that $\gamma_1, \ldots, \gamma_n$ are linearly independent over $\mathbb{Q}$.

From now on, we assume that the coordinates of γ are linearly independent over $\mathbb{Q}$. We will first consider the case $m_1 = \cdots = m_n = 0$. We then find:

$$\begin{aligned} \left| \frac{1}{p} \sum_{k=0}^{p-1} c_{m_1,\ldots,m_n}\left(T_\gamma^k(x_1, \ldots, x_n)\right) \right| &= \left| \frac{1}{p} \sum_{k=0}^{p-1} e^{2\pi i k(m_1\gamma_1+\ldots+m_n\gamma_n)} \right| \cdot \left| c_{m_1,\ldots,m_n}(x_1, \ldots, x_n) \right| \\ &= \left| \frac{1}{p} \sum_{k=0}^{p-1} 1 \right| = 1, \end{aligned}$$

from which it follows that

$$\frac{1}{p} \sum_{k=0}^{p-1} c_{m_1,\ldots,m_n}\left(T_\gamma^k(x_1, \ldots, x_n)\right) \to 1 = \int_{\mathbb{T}^n} c_{0,\ldots,0}\, d\mu$$

uniformly on $\mathbb{T}^n$.

We now consider all other cases and find:

$$\begin{aligned}\left|\frac{1}{p}\sum_{k=0}^{p-1} c_{m_1,\ldots,m_n}\left(T_\gamma^k(x_1,\ldots,x_n)\right)\right| &= \left|\frac{1}{p}\sum_{k=0}^{p-1} e^{2\pi ik(m_1\gamma_1+\ldots+m_n\gamma_n)}\right|\cdot\left|c_{m_1,\ldots,m_n}(x_1,\ldots,x_n)\right| \\ &= \left|\frac{1-e^{2\pi ip(m_1\gamma_1+\ldots+m_n\gamma_n)}}{p\left(1-e^{2\pi i(m_1\gamma_1+\ldots+m_n\gamma_n)}\right)}\right| \\ &\leq \left|\frac{2}{p\left(1-e^{2\pi i(m_1\gamma_1+\ldots+m_n\gamma_n)}\right)}\right|.\end{aligned}$$

Letting $p\to\infty$ yields us the following result:

$$\frac{1}{p}\sum_{k=0}^{p-1} c_{m_1,\ldots,m_n}\left(T_\gamma^k(x_1,\ldots,x_n)\right) \;\to\; 0 \;=\; \int_{\mathbb{T}^n} c_{m_1,\ldots,m_n}\,d\mu$$

uniformly on $\mathbb{T}^n$. Since characters are trigonometric functions, we have for any trigonometric polynomial ϕ (finite linear combination of characters) the following:

$$\frac{1}{n}\sum_{k=0}^{n-1}\phi\left(T_\gamma^k(x_1,\ldots,x_n)\right) \;\to\; \int_{\mathbb{T}^n}\phi\,d\mu.$$

We now use a famous approximation theorem of Weierstrass, which says that continuous functions are the uniform limit of trigonometric polynomials. Therefore, for all continuous functions f we have:

$$\frac{1}{n}\sum_{k=0}^{n-1} f\left(T_\gamma^k(x_1,\ldots,x_n)\right) \;\to\; \int_{\mathbb{T}^n} f\,d\mu$$

uniformly on $\mathbb{T}^n$. By Theorem 1.12, T_γ is uniquely ergodic. □

If we let $n=1$ in Theorem 2.2, we get the following statements about rotations on the unit circle:

Corollary 2.3. The rotation on the unit circle given by $T(x)=x+\alpha \pmod 1$ is ergodic only if α is irrational. Moreover, if T is ergodic, then it is uniquely ergodic.

The ergodic measure for this transformation is the Haar measure. On $\mathbb{T}^n$, this is equivalent with the n-fold product of the one-dimensional Lebesgue measure. We will be interested in the way the orbit spreads on $\mathbb{T}^n$. Since the characteristic is not a trigonometric function nor continuous, this does not follow directly from Theorem 2.2. However, we can still derive this by another approximation method.

Theorem 2.4. Let $\Delta\subset\mathbb{T}^n$ be measurable. Then we have $\lim_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n-1}\chi_\Delta\left(T_\gamma^k(x)\right)=\mu(\Delta)$, where Δ is the Lebesgue measure.

Proof. Choose two continuous functions $f_1\leq\chi_\Delta\leq f_2$ with $\int(f_2-f_1)\,d\mu<\epsilon$. Denote $B_1=\frac{1}{n}\sum_{k=0}^{n-1}f_1(T_\gamma^k(x))$, $B_2=\frac{1}{n}\sum_{k=0}^{n-1}f_2(T_\gamma^k(x))$ and $B=\frac{1}{n}\sum_{k=0}^{n-1}\chi_\Delta(T_\gamma^k(x))$, then we have:

$$\int\chi_\Delta\,d\mu-\epsilon\leq\int f_1\,d\mu=\lim_{n\to\infty}B_1\leq\liminf_{n\to\infty}B\leq\limsup_{n\to\infty}B\leq\lim_{n\to\infty}B_2=\int f_2\,d\mu\leq\int\chi_\Delta\,d\mu+\epsilon.$$

Letting $\epsilon\to 0$ turns all inequalities into equalities, which gives us:

$$\liminf_{n\to\infty}B=\limsup_{n\to\infty}B=\int\chi_\Delta\,d\mu=\mu(\Delta).$$

□

2.3 Distribution

Let's return to the first digits problem posed in Section 2.1. We saw that the 8 didn't appear that much as the first digit of a power of 7. Therefore, one may think that there is a large string of digits starting with an 8 which will never appear as the string of first digits of a power of 7. For example, is there a power of 7 that starts with 87727964797? The following theorem assures us that the answer is yes.

Theorem 2.5. Let $k_1, \ldots, k_n$, $p_1, \ldots, p_n$, $c_1, \ldots, c_n$ and $b_1, \ldots, b_n$ be positive integers with $b_i \geq 2$ for all i. If, for $m \in \mathbb{N}$, the numbers $\left\{\log_{b_i}(c_i \cdot k_i^m)\right\}$ $(i = 1, \ldots, n)$ are linearly independent over $\mathbb{Q}$, the relative frequency with which for all i the numbers p_i will be the (string of) first digits of $c_i \cdot k_i^m$, when both are written with respect to base b_i, is $\prod_{i=1}^{n} \log_{b_i}\left(\frac{p_i+1}{p_i}\right)$.

Proof. Note that we can rephrase the statement by saying that there are positive integers $l_1, \ldots, l_n$ such that $c_i \cdot k_i^m = b_i^{l_i} p_i + q_i$ and $0 \leq q_i < b_i^{l_i}$ for all i. This is equivalent with saying that for all i, $b_i^{l_i} p_i \leq c_i \cdot k_i^m < b_i^{l_i}(p_i + 1)$ holds. Taking the logarithm with base b_i yields us the following inequality:

$$l_i + \log_{b_i}\left(\frac{p_i}{c_i}\right) \;\leq\; m \log_{b_i} k_i \;<\; l_i + \log_{b_i}\left(\frac{p_i+1}{c_i}\right).$$

Now write $g_i = \left\lfloor \log_{b_i}\left(\frac{p_i}{c_i}\right)\right\rfloor + 1$. It now follows that:

$$0 \leq \log_{b_i}\left(\frac{p_i}{c_i}\right) - (g_i - 1) \leq m \log_{b_i}(k_i) - l_i - (g_i - 1) < \log_{b_i}\left(\frac{p_i+1}{c_i}\right) - (g_i - 1) \leq 1,$$

and therefore:

$$\log_{b_i}\left(\frac{p_i}{c_i \cdot b_i^{g_i-1}}\right) \leq \left\{m \log_{b_i} k_i\right\} \leq \log_{b_i}\left(\frac{p_i+1}{c_i \cdot b_i^{g_i-1}}\right),$$

where $\{\cdot\}$ denotes the fractional part.

Theorem 2.4 now tells us that the relative frequency with which $c_i \cdot k_i^m$ starts with p_i for all i is equal to the Lebesgue measure of the set $\Delta = \prod_{i=1}^{n}\left[\log_{b_i}\left(\frac{p_i}{c_i \cdot b_i^{g_i-1}}\right), \log_{b_i}\left(\frac{p_i+1}{c_i \cdot b_i^{g_i-1}}\right)\right]$. Thus, the relative frequency is:

$$\begin{aligned}
\mu(\Delta) &= \mu\left(\prod_{i=1}^{n}\left[\log_{b_i}\left(\frac{p_i}{c_i \cdot b_i^{g_i-1}}\right), \log_{b_i}\left(\frac{p_i+1}{c_i \cdot b_i^{g_i-1}}\right)\right]\right) \\
&= \prod_{i=1}^{n}\left[\log_{b_i}\left(\frac{p_i+1}{c_i \cdot b_i^{g_i-1}}\right) - \log_{b_i}\left(\frac{p_i}{c_i \cdot b_i^{g_i-1}}\right)\right] \\
&= \prod_{i=1}^{n} \log_{b_i}\left(\frac{p_i+1}{p_i}\right).
\end{aligned}$$

The theorem is now proved. □

Example 2.6. As stated before, there is a power of 7 that begins with 87727964797. For example, 7^{1001} starts with this string of digits. The relative frequency in this case is $\log\left(\frac{87727964798}{87727964797}\right) \approx 4.95 \cdot 10^{-12}$, which means that it will require a lot of work to determine another power of 7 which also starts with 87727964797.

Theorem 2.5 holds a lot of surprising results. For example, larger numbers appear less often as (string of) first digits than smaller numbers. Also, multiplying the powers by some other number has no effect on the distribution whatsoever. The most surprising result, however, is that the distribution is the same for each number whose powers we consider. Note that Theorem 2.5 also tells us that the first digits of powers are statistically independent: the simultaneous distribution is the product of the distributions.

3 Coefficients of continued fractions

This chapter is about the second of the three problems we will discuss. The problem requires a longer introduction, but the way to solve it follows rather straightforwardly.

3.1 Introduction

Numbers can be represented in many ways. In this chapter we consider the continued fractions:

Definition 3.1. A continued fraction is a fraction of the form

$$a_0 + \cfrac{1}{a_1 + \frac{1}{a_2 + \cdots}},$$

where a_0 is an integer and the other numbers a_i are positive integers.

Instead of writing a number using these fractions, it is more common to specify the eventually positive sequence of coefficients $(a_n)_{n=0}^{\infty} = \langle a_0; a_1, a_2, \ldots \rangle$. We will pose the following theorem without any proof. It lists some properties of continued fractions:

Theorem 3.2. Let $x \in \mathbb{R}$ be arbitrary:

(a) The number x has a continued fraction representation.

(b) Every continued fraction converges.

(c) The continued fraction representation of x is finite if and only if x is rational.

(d) The continued fraction representation of x is unique if and only if x is irrational.

For the proof, see [4, Chapter 10].

Given a number $x \in \mathbb{R}$, one may be interested in finding a continued fraction representation. Fortunately, one can determine the coefficients easily. The coefficient a_0 is the integer part of x, so $a_0 = \lfloor x \rfloor$. The other coefficients together are the fractional part of a and can be determined as follows. To determine a_1, we look at $\{x\}$, where $\{\cdot\}$ denotes the fractional part, and consider its reciprocal if $\{x\}$ is not equal to zero (if we have $\{x\} = 0$, then we have determined a continued fraction representation for x). We will call this number x_1, so $x_1 = \frac{1}{\{x\}}$. Then we have $a_1 = \lfloor x_1 \rfloor$. To determine the other coefficients, we use the same process as described above to determine a_1. However, if we want to determine the coefficient a_i ($i \geq 2$), we look at x_{i-1} instead of x. Note that the sequence of continued fractions always converges [4, Chapter 10].

Example 3.3. Let's look at some numbers written in continued fraction representation:

- $7\frac{1}{3} = \langle 7; 3 \rangle$, but also $7\frac{1}{3} = \langle 7; 2, 1 \rangle$. We see that the representation of rationals is not unique in this way, as Theorem 3.3(d) implies.
- $-4\frac{2}{5} = \langle -5; 1, 1, 2 \rangle$ (and also $-4\frac{2}{5} = \langle -5; 1, 1, 1, 1 \rangle$).
- $\sqrt{2} = \langle 1; 2, 2, 2, 2, 2, \ldots \rangle$, where $a_k = 2$ for all $k \neq 0$.

- $e = \langle 2; 1, 2, 1, 1, 4, 1, 1, 6, 1, \ldots \rangle$, where $a_{3k+1} = a_{3k+3} = 1$ and $a_{3k+2} = 2(k+1)$ for all $k \geq 0$.
- $\pi = \langle 3; 3, 7, 15, 1, 292, 1, 1, 1, 2, 1, 3, 1, 14, \ldots \rangle$, where there is no regular behaviour apparent in the coefficients.

Now that we have seen many examples of continued fractions, we are ready to pose the problem:

> Let $a_1, \ldots, a_n$ be arbitrary real numbers and $b_1, \ldots, b_n$ be arbitrary positive integers. What is the relative frequency with which, for all i, b_i appears as the kth coefficient ($k \neq 0$) in the continued fraction representation of a_i?

Having another look at Example 3.4, we can already deduce that, whatever the answer may be, it will be fundamentally different from the answer of the first digits problem. We can deduce this as follows. Suppose that we were able to find a distribution formula with the same conditions as in Section 2.3. Such a formula would definitely hold in the one-dimensional case (i.e. the case where we only look at one number). Now look at the number $\sqrt{2}$. The integer 2 appears infinitely often in its continued fraction, thus we expect 2 to have a high relative frequency. However, 2 does not appear at all in the continued fraction of 0, so its relative frequency should also be zero. We have reached a contradiction here. Therefore, we may expect two different answers.

3.2 The Gauss transformation

Once again, let's forget about the continued fractions for a while. In this section, we will look at the Gauss transformation.

Definition 3.4. The Gauss transformation is the transformation $G : (0, 1] \to [0, 1]$ given by $G(x) = \left\{\frac{1}{x}\right\}$, where $\{\cdot\}$ denotes the fractional part.

One can prove that the Gauss transformation is ergodic, but we can do much better. The ergodic measure is the so-called Gauss measure.

Definition 3.5. The Gauss measure is the measure γ on $[0, 1]$ given by

$$\gamma(A) \quad = \quad \frac{1}{\ln 2} \int_A \frac{1}{1+x}\, dx$$

with $A \subset [0, 1]$. The latter integral is a Lebesgue integral.

The Gauss transformation is one of the many transformations in the class of piecewise monotonic transformations.

Definition 3.6. A transformation T on the interval $(0, 1)$ is called piecewise monotonic if the interval $(0, 1)$ can be split up into a countable number of subintervals $\Delta_1, \Delta_2, \ldots$ such that T is strictly monotonic on each subinterval. The transformation T need not be defined on the endpoints of a subinterval.

One can easily verify that the Gauss transformation G is piecewise monotonic. To see this, take the subintervals $\Delta_i = \left(\frac{1}{1+i}, \frac{1}{i}\right)$. Since $G(x) = 0$ whenever x is 0, 1 or $\frac{1}{p}$ with p an arbitrary positive integer, we see that $G(x) = 0$ holds for the endpoints of any subinterval. Furthermore, $G(x)$ is continuous on the interior of each of the subintervals and the derivative of G is strictly negative, thus G is decreasing on each of the subintervals. Therefore, G is piecewise monotonic.

The following theorem shows that piecewise monotonic transformation are mixing if they satisfy certain conditions:

Theorem 3.7. Let T be a piecewise monotonic transformation on $(0,1)$ and denote the subintervals on which T is monotonic by Δ_i. Suppose that the following conditions hold:

1. For all i: $T(\Delta_i) = (0,1)$ and T is twice continuously differentiable on Δ_i.
2. There exists an $s \in \mathbb{N}$ such that:
$$\inf_{\Delta_i} \inf_{x \in \Delta_i} \left| \frac{dT^s}{dx} \right| \;=\; C_1 \;>\; 1.$$
3. We have:
$$\sup_{\Delta_i} \sup_{x_1, x_2 \in \Delta_i} \frac{\left| \frac{d^2T}{dx^2}(x_1) \right|}{\left(\frac{dT}{dx}(x_2) \right)^2} \;=\; C_2 \;<\; \infty.$$

Then there is a invariant normalized Borel measure μ. Moreover:

(a) The measure μ is equivalent to the Lebesgue measure λ, and there exists a constant $K > 0$ such that $\frac{1}{K} \leq \frac{d\mu}{d\lambda} < K$.

(b) The transformation T is mixing with respect to μ.

The proof can be found in [2, Theorem 10.8.4]. We are not really interested in the theorem itself. However, a corollary of this theorem is very important to us.

Corollary 3.8. The Gauss transformation G is mixing.

Proof. We will show that G satisfies the conditions of Theorem 3.8.

1. As noted earlier, G is discontinuous whenever $x = \frac{1}{p}$ with $p \in \mathbb{N}$ or if $x = 0$. Therefore, G is continuous on each interval Δ_i. We also saw earlier that G is piecewise monotonic. Since $G\left(\frac{1}{i+1}\right) = 0 \equiv 1$ and $G\left(\frac{1}{i}\right) = 0$, and since G is also strictly decreasing on each Δ_i, this implies that $T(\Delta_i) = (0,1)$. Also, G is twice continuously differentiable on all Δ_i, because the function $f(x) = \frac{1}{x}$ is twice continuously differentiable there.
2. Consider the transformation G^2. We know that $\left|\frac{dG}{dx}\right| = \frac{1}{x^2} \geq 1$ for $x \in (0,1)$. Since we have $\left|\frac{dG}{dx}\right| \geq \frac{9}{4}$ for $0 < x \leq \frac{2}{3}$ and $0 < G(x) < \frac{1}{2}$ for $\frac{2}{3} < x < 1$, we find:
$$\left|\frac{dG^2}{dx}(x)\right| = \left|\frac{dG}{dx}(x)\right| \cdot \left|\frac{dG}{dx}(G(x))\right| \geq \frac{9}{4}$$
for all $x \in (0,1)$. Therefore, G satisfies this condition.
3. Since $|G''(x)| = \frac{2}{x^3}$ and $|G'(x)| = \frac{1}{x^2}$ on all Δ_i, we have:
$$\frac{|G''(x_1)|}{|G'(x_2)|^2} \;\leq\; \frac{\left|G''\left(\frac{1}{i+1}\right)\right|}{\left|G'\left(\frac{1}{i}\right)\right|^2} \;=\; \frac{2(i+1)^3}{i^4} \;\leq\; 16.$$
The last inequality is true because $\sup_{i \in \mathbb{N}} \frac{2(i+1)^3}{i^4} = 16$. Therefore, G satisfies this condition.

Since the transformation G satisfies all conditions, Theorem 3.8(b) tells us now that G is mixing. $\square$

Note that Theorem 3.8 does not tell us what the measure μ is. However, it can be shown that the Gauss transformation G is mixing with respect to the Gauss measure γ defined earlier, see [2, p. 174].

3.3 Distribution

Now it's time to solve the problem posed in Section 3.1. Since we are only looking at the coefficients a_i with $i \in \mathbb{N}$, we can set $a_0 = 0$. Note that this means that we only have to consider numbers in $[0, 1]$ (remember that $1 = \langle 0; 1 \rangle$). The Gauss transformation is a helpful tool for determining the coefficents of a continued fraction. We will only consider the irrational numbers. This is because the rational numbers have finite continued fraction representation and thus there is no real need to determine the distribution: one can easily obtain the coefficients by Theorem 3.3. Note that ignoring the rationals is allowed, since the set of rationals in $[0, 1]$ is a null set with respect to the Gauss measure (this follows from Theorem 3.8(a)).

Theorem 3.9. Let $x \in [0, 1]$ be irrational and consider its coefficients in continued fraction representation. Then $a_n = k$ if and only if $\frac{1}{k+1} \leq G^{n-1}(x) \leq \frac{1}{k}$, where G is the Gauss transformation.

Proof. Let $x \in [0, 1]$ be irrational with continued fraction representation $x = \langle a_1, a_2, a_3, \ldots \rangle$. The effect of G formulated in terms of this representation is $\langle a_1, a_2, a_3, \ldots \rangle \to \langle a_2, a_3, \ldots \rangle$, i.e. a shift. It follows directly that, after applying G exactly $n - 1$ times, a_n is the leading coefficient. From this is follows directly that, if $\frac{1}{k+1} < G^{n-1}(x) < \frac{1}{k}$, then $a_n = k$.

Conversely, consider $G^{n-1}(x) = \langle k, a_{n+1}, \ldots \rangle$, where we assume that $a_n = k$. Because of the way G works, we know that $0 < \langle a_{n+1}, a_{n+2}, \ldots \rangle < 1$ (it cannot be equal to either bound, because that would imply that a_{n+2} does not exist). Therefore, we can write $G^{n-1}(x) = \frac{1}{k+\alpha}$ with $\alpha \in (0, 1)$. But this means that $\frac{1}{k+1} < G^{n-1}(x) < \frac{1}{k}$. $\square$

With the theorem proved, it is easy to give an answer to the problem in Section 3.1.

Theorem 3.10. Let $b_1, \ldots, b_n$ be arbitrary positive integers. For almost all real numbers $a_1, \ldots, a_n$, the relative frequency with which b_i appears as the kth coefficient ($k \neq 0$) in the continued fraction representation of a_i for all i is:

$$\prod_{i=1}^{n} \frac{\ln\left(1 + \frac{1}{b_i}\right) - \ln\left(1 + \frac{1}{b_i+1}\right)}{\ln 2}.$$

Proof. Suppose that $n = 1$. In this case, the relative frequency with which b appears as a coefficient in the continued fraction representation of a is the Gauss measure of the interval $\left[\frac{1}{b+1}, \frac{1}{b}\right]$ by Corollary 3.9 and Theorem 3.10. We find:

$$\begin{aligned} \gamma\left(\left[\frac{1}{b+1}, \frac{1}{b}\right]\right) &= \frac{1}{\ln 2} \int_{\frac{1}{b+1}}^{\frac{1}{b}} \frac{1}{1+x}\, dx \\ &= \frac{\ln\left(1 + \frac{1}{b}\right) - \ln\left(1 + \frac{1}{b+1}\right)}{\ln 2}. \end{aligned}$$

By Theorem 1.20(d), the distribution in the case $n > 1$ is exactly n times the distribution in the one-dimensional case. $\square$

Note the subtle difference between the statements of Theorem 2.5 and Theorem 3.12. As opposed to the first theorem, this one holds almost everywhere. This means that there can be a non-empty null set for which the statement is not true. For example, $\sqrt{2}$ and e do not have coefficient 3 in their continued fraction representation, while Theorem 3.12 says that the relative frequency in both cases is approximately 0.0931 for almost all real numbers. Hence exceptions can occur, i.e., Theorem 3.12 is optimal, but ergodic theory provides no further information on the structure of the irrational numbers in the expectional set if it is non-empty.

4 Fractional parts of polynomials

This chapter is about the final number-theoretic problem. While it is easy to explain the problem, solving it is definitely not as straightforward as we have seen in the previous two chapters.

4.1 Introduction

Consider your favourite polynomial with real coefficients in one variable. We will call the polynomial P and the variable n. As the chapter's name implies, we will consider the fractional parts of polynomials. More specifically, we will consider the sequence $(a_k)_{k=1}^{\infty}$ given by $a_k = \{P(k)\}$, where $\{\cdot\}$ once again denotes the fractional part.

To illustrate the problem before formally posing it, we consider three different polynomials, namely $P_1(n) = \pi n^2$, $P_2(n) = n^2\sqrt{3}$ and $P_3(n) = \frac{1}{3}n^2$. For each of these three polynomials, we will look at the first digit after the decimal point for $n = 1, \ldots, 50$. The table belows lists the distribution we find:

Table 2: The first digits of the first 50 fractional parts of the polynomials P_1, P_2 and P_3

	0	1	2	3	4	5	6	7	8	9
$P_1(n) = \pi n^2$	6	6	5	2	6	6	3	5	4	7
$P_2(n) = n^2\sqrt{3}$	2	4	8	4	5	7	6	6	6	2
$P_3(n) = \frac{1}{3}n^2$	16	0	0	34	0	0	0	0	0	0

The distribution of $P_1(n)$ and $P_2(n)$ seem random. The distribution of $P_3(n)$, however, is significantly different from the other two. You may already have an idea on why this case is different from the other two cases and it is very likely that your idea is correct (see Section 4.2).

As promised, we will now formally state the problem. We are interested in the way the fractional parts of polynomials are distributed. Since the first digit after the decimal point can be a 0 (as opposed to, for example, the problem of Section 2), we need to state the problem carefully.

> Let c be a string of digits, where each digit is one of $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$ and let $P(n)$ be a real-valued polynomial in one variable. What is the relative frequency with which c appears as the first digits of the sequence $a_n = \{P(n)\}$ $(n \in \mathbb{N})$?

4.2 Polynomials with rational coefficients

Looking back at Table 2 (Section 4.1), one may believe that the difference between the distribution of $P_3(n)$ and the other two distributions is because $\frac{1}{3}$ is rational and that $\sqrt{3}$ and π are not. While this is indeed the case (as we will see later), we wish to generalize this observation to non-monomial polynomials.

As such, one can think of many possibilities. For example one may think that the leading coefficient should be rational, or that at least one coefficient should be rational, or even that all coefficients should be rational. To put these ideas all to the test, we repeat the experiment seen in Section 4.1 with the following polynomials: $P_1(n) = \frac{1}{6}n^2 + \frac{1}{6}n$, $P_2(n) = n^2\sqrt{6} + \frac{1}{6}n$ and $P_3(n) = \frac{1}{6}n^2 + n\sqrt{6}$.

Table 3: The first digits of the first 50 fractional parts of the polynomials P_1, P_2 and P_3

	0	1	2	3	4	5	6	7	8	9
$P_1(n) = \frac{1}{6}n^2 + \frac{1}{6}n$	33	0	0	17	0	0	0	0	0	0
$P_2(n) = n^2\sqrt{6} + \frac{1}{6}n$	4	10	4	2	4	5	7	7	4	3
$P_3(n) = \frac{1}{6}n^2 + n\sqrt{6}$	4	6	5	5	6	6	5	5	5	3

The result of Table 3 is obvious: periodic behaviour appears whenever the polynomial has rational coefficients. While two examples of this phenomenon is not a proof, we can actually prove that this holds in the general case.

Theorem 4.1. Let $P(n)$ be a real-valued polynomial in one variable. If $P(n)$ has rational coefficients, then the sequence $a_n = \{P(n)\}$ $(n \in \mathbb{N})$ is periodic.

Proof. Consider first the case in which $P(n)$ is one term. Write $P(n) = \frac{a}{b}n^m$ with $a, b \in \mathbb{Z}$ coprime. Suppose that there is a $k \in \mathbb{N}$ such that $P(n+k) \equiv P(n) \pmod 1$ for all $n \in \mathbb{N}$, which means that $\frac{a}{b}(n+k)^m \equiv \frac{a}{b}n^m \pmod 1$. It then follows that $\frac{(n+k)^m - n^m}{b} \equiv 0 \pmod 1$. Now note that, by the binomial formula, that $(n+k)^m - n^m = \sum_{t=0}^{m-1} \binom{m}{t} n^t k^{m-t}$ is divisible by k. Since, in order to guarantee periodic behaviour, we want $(n+k)^m - n^m$ to be divisible by b, we choose $k = b$. We see that such a number $k \in \mathbb{N}$ always exists.

Generalizing this to arbitrary polynomials with rational coefficients in one variable is easy. Let $P(n)$ be such a polynomial. Write $P(n) = \sum_{j=0}^{m} \frac{a_j}{b_j} n^j$ with $a_j, b_j \in \mathbb{Z}$ for all j and $\gcd(a_j, b_j) = 1$ for all j. If we look at one term of this polynomial (i.e., choose a specific j), we obtain a different polynomial of which we already know that the sequence of fractional parts is periodic and that this period is at most b_j. Since we can do this for all j, the behaviour of $P(n)$ is still periodic, but with period at most $\operatorname{lcm}(b_0, \ldots, b_m)$. This period is finite and thus the sequence of fractional parts of any polynomial with rational coefficients in one variable is periodic. □

Not only Theorem 4.1 is useful for our problem, but its proof is very useful as well. For example, we can give an estimate on the period of the sequence $\{P(n)\}_{n=1}^{\infty}$ whenever $P(n)$ has rational coefficients. We can also see directly that the proof fails if even one of the coefficients is irrational: we cannot write this coefficient as a fraction.

Example 4.2. Let's look back at the polynomials we used for the experiments listed in Table 2 and Table 3. The sequence corresponding to the polynomial $\frac{1}{3}n^2$ has period 3, as was deduced correctly in the proof of Theorem 4.1. In fact, the proof of Theorem 4.1 tells us that the sequence $\{P(n)\}_{n=1}^{\infty}$, where $P(n) = \frac{a}{b}n^m$ with $a, b \in \mathbb{Z}$ coprime, has exactly period b.
Something remarkable, however, is going on when considering the polynomial $\frac{1}{6}n^2 + \frac{1}{6}n$. The sequences of $\frac{1}{6}n^2$ and $\frac{1}{6}n$ are both periodic with period 6. However, the sequence corresponding to the polynomial $P(n) = \frac{1}{6}n^2 + \frac{1}{6}n$ is periodic with period 3.

4.3 Other polynomials

Now that we solved the case where all coefficients are rational, we will consider the other case where at least one of the coefficients is irrational.

The problem here is that there is no dynamical system apparent (in contrast to Chapter 3), nor can we easily transform the problem into one where there is an obvious dynamical system (in contrast to Section 2). However, there is a transformation that will be of great value when trying to find the distribution of the sequence of fractional parts of the polynomial $P(n)$, where this polynomial has at least one irrational coefficient.

For now, we will assume that the leading coefficient is irrational and also that it is the only irrational coefficient. Write $P(n) = \alpha n^m + a_{m-1}n^{m-1} + \cdots + a_0$. The relevant transformation is the transformation $T : \mathbb{T}^m \to \mathbb{T}^m$ given by:

$$T\left[(x_1, x_2, \ldots, x_m)\right] \quad = \quad (x_1 + \beta, x_2 + x_1, \ldots, x_m + x_{m-1}) \pmod 1.$$

The following theorem relates T^n to the polynomial $P(n)$:

Theorem 4.3. Let T be the transformation as defined above and let $P(n)$ be a polynomial of degree m with irrational leading coefficient. Then, for a unique choice of $\beta, x_1, \ldots, x_m$, the last coordinate of T^n is equal to $\{P(n)\}$ (with $n \in \mathbb{N}$) for $n \geq m$. Moreover, the corresponding β is irrational.

Proof. We will first find a closed form for the last coordinate of T^n. We claim that the closed form for the j-th coordinate after n iterates is $\sum_{i=0}^{m} \binom{n}{i} x_{j-i}$ (with $1 \leq j \leq m$), where we define $x_0 = \beta$, $x_k = 0$ for $k < 0$ and $\binom{n}{i} = 0$ whenever $i > n$. We will prove by induction (with respect to j and n) that this closed form is correct.

The closed form for $j = 1$ is obvious for all n, so suppose that the closed form is correct for $j = k$ and all n, we then need to prove that it is also correct for $j = k + 1$ and all n.

Suppose first that $n = 1$. The closed form is then equal to $x_j + x_{j-1}$, which is indeed correct. Next, suppose that the closed form is correct for $n = l$, then we need to prove that it is correct for $n = l + 1$. By the induction hypothesis for j, the closed form for the $(k+1)$-th coordinate after $l + 1$ iterates is:

$$\begin{aligned} \sum_{i=0}^{m} \binom{l}{i} x_{k+1-i} \; + \; \sum_{i=0}^{m} \binom{l}{i} x_{k-i} \quad &= \quad \sum_{i=0}^{m} \left[\binom{l}{i} + \binom{l}{i-1} \right] x_{k+1-i} \\ &= \quad \sum_{i=0}^{m} \binom{l+1}{i} x_{k+1-i}. \end{aligned}$$

We have now proven that the closed form is correct.

Finally, we need to prove that $\sum_{i=0}^{m} \binom{n}{i} x_{m-i}$ is equal to $P(n)$ for a suitable choice of the parameters $\beta = x_0, x_1, \ldots, x_m$. Note that the equality $\sum_{i=0}^{m} \binom{n}{i} x_{m-i} = P(n)$ gives a system of linear equations corresponding to the powers of n (if $n \geq m$). We can easily solve the system, since the method of backwards substitution appears naturally: the coefficient of n^i is determined by only the parameters $\beta = x_0, \ldots, x_{m-i}$. The existence of the solution, and thus the theorem, has now been proved. Note also that the leading coefficient is only determined by β, so that β must be irrational. $\square$

Example 4.4. As an illustration of Theorem 4.3 (and especially its proof), let's determine the initial conditions of the transformation such that the closed form for the last coordinate is $P(n) = n^2\sqrt{6}+\frac{1}{6}n$ (as seen in Section 4.2). Since the degree of $P(n)$ is 2, our transformation is $T[(x_1, x_2)] = (x_1 + \beta, x_2 + x_1) \pmod 1$.
To find the values of x_1 and x_2, we need to solve $\sum_{i=0}^{2} \binom{n}{i} x_{2-i} = n^2\sqrt{6} + \frac{1}{6}n$. Expanding this equation gives us: $x_2 + nx_1 + \frac{n(n-1)}{2}\beta = n^2\sqrt{6} + \frac{1}{6}n$. Note that this equation is equivalent with the system

$$\begin{cases} \frac{1}{2}\beta & & & = & \sqrt{6} \\ -\frac{1}{2}\beta & + & x_1 & = & \frac{1}{6} \\ & & x_2 & = & 0 \end{cases}$$

which can easily be solved. The solution is $\beta = 2\sqrt{6}$, $x_1 = \frac{1}{6} + \sqrt{6}$ and $x_2 = 0$. The transformation has now been determined.

4.4 Distribution

In the previous section we have been looking at the transformation $T : \mathbb{T}^m \to \mathbb{T}^m$ given by:

$$T[(x_1, x_2, \ldots, x_m)] \quad = \quad (x_1 + \beta, x_2 + x_1, \ldots, x_m + x_{m-1}) \pmod 1,$$

where β is irrational. In this section we will use this transformation and ergodic theory to find the distribution of the first digits of the sequence $\{P(n)\}_{n=1}^{\infty}$, where the polynomial $P(n)$ has at least one irrational coefficient.

First of all, we will prove that the transformation T is uniquely ergodic. The complete proof is beyond the scope of this thesis, but we will mention all theorems involved in the proof.

Theorem 4.5. The transformation $T : \mathbb{T}^m \to \mathbb{T}^m$ given by
$T[(x_1, x_2, \ldots, x_m)] = (x_1 + \beta, x_2 + x_1, \ldots, x_m + x_{m-1}) \pmod 1$ with β irrational is ergodic.

Proof. Suppose that $f = f \circ T$ for $f \in L^2(\mathbb{T}^m)$. Write $x = (x_1, \ldots, x_m) \in \mathbb{T}^m$ and $n = (n_1, \ldots, n_m)$, both as row vector. The Fourier series of f and $f \circ T$ are:

$$f(x) \;=\; \sum_{n\in\mathbb{Z}^m} c_n e^{2\pi i n\cdot x} \quad \text{and} \quad f(T(x)) \;=\; \sum_{n\in\mathbb{Z}^m} c_n e^{2\pi i n\cdot T(x)}.$$

Note also that we can write:

$$\begin{aligned} n \cdot T(x) &= (n_1, \ldots, n_m) \cdot (x_1 + \beta, \ldots, x_m + x_{m-1}) \\ &= n_1\beta + (n_1 + n_2)\, x_1 + \cdots + (n_{m-1} + n_m)\, x_{m-1} + n_m x_m \\ &= n_1\beta + (nA) \cdot x, \end{aligned}$$

where A is the $(m \times m)$-matrix with $a_{i;i} = a_{i;i-1} = 1$ and $a_{ij} = 0$ in all other cases $(1 \le i, j \le m)$ and nA is the matrix product of the $(1 \times m)$-matrix n with the matrix A. Using this formula and the substitution

$\tilde{n} = nA$, we can rewrite the Fourier series of $f \circ T$ as follows:

$$\begin{aligned} f(T(x)) &= \sum_{n\in\mathbb{Z}^m} c_n e^{2\pi i n\cdot T(x)} \\ &= \sum_{n\in\mathbb{Z}^m} c_n e^{2\pi i(n_1\beta+(nA)\cdot x)} \\ &= \sum_{n\in\mathbb{Z}^m} c_n e^{2\pi i n_1\beta}\cdot e^{2\pi i(nA)\cdot x} \\ &= \sum_{\tilde{n}\in\mathbb{Z}^m} c_{\tilde{n}A^{-1}}\cdot e^{2\pi i\left(\tilde{n}A^{-1}\right)_1\beta}\cdot e^{2\pi i\tilde{n}\cdot x}, \end{aligned}$$

where $\left(\tilde{n}A^{-1}\right)_1$ denotes the first coordinate of the vector $\tilde{n}A^{-1} = n$.

Since $f = f \circ T$, we find when comparing Fourier coefficients: $c_n = c_{nA^{-1}} \cdot e^{2\pi i\left(nA^{-1}\right)_1\beta}$, which is equivalent with $c_{nA} = c_n \cdot e^{2\pi i n_1\beta}$. Therefore, we have $|c_{nA}| = |c_n|$ for all $n \in \mathbb{Z}^m$. Since $\sum |c_n|^2 < \infty$, we have $c_n \neq 0$ only if $nA^r = n$ for some integer $r \geq 1$. However, this implies $n = (n_1, 0, \ldots, 0)$. In this case, we have $c_n = c_n e^{2\pi i n_1\beta}$, and because β is irrational, we have $c_n = 0$ unless $n = 0$. Therefore, we have $f = c_0$ almost everywhere. By Theorem 1.9, T is ergodic. □

Proving the unique ergodicity of T involves the notion of group extensions (in this case the groups are topological groups). Let $(X, \mathcal{U}, \mu)$ be a probability space and T a μ-invariant transformation. Let G be a compact group and $\phi : X \to G$ a continuous transformation. The group extension of the system (X, T) (i.e. the transformation T on the set X) is $(X \times G, T)$ and is defined by $T(x, g) = (T(x), \phi(x)g)$, where $x \in X$ and $g \in G$. If we let ν be the Haar measure on G, then $\mu \times \nu$ will be a T-invariant measure on $X \times G$. One can prove the following theorem regarding group extensions.

Theorem 4.6. Let the system (X, T) be uniquely ergodic and let the group extension $(X \times G, \mathcal{U}_{X\times G}, \mu \times \nu, T)$ be ergodic. Then the system $(X \times G, T)$ is uniquely ergodic.

The proof of this theorem can be found in [3, Proposition 3.10].

Using these two theorems, one can prove:

Theorem 4.7. The transformation $T : \mathbb{T}^m \to \mathbb{T}^m$ given by
$T[(x_1, x_2, \ldots, x_m)] = (x_1 + \beta, x_2 + x_1, \ldots, x_m + x_{m-1}) \pmod 1$ with β irrational is uniquely ergodic.

Proof. We will prove the theorem using induction with respect to m.

Let $m = 1$, so that we get the transformation $T : S^1 \to S^1$ given by $T(x) = x + \beta$, where β is irrational. We know that it is uniquely ergodic (Corollary 2.3).

Suppose that the theorem is true for $m = k$. We can extend the system to a transformation T on the torus $\mathbb{T}^{k+1}$. By applying a group extension we "add" a dimension (i.e. T is no longer a transformation on $\mathbb{T}^k$, but a transformation on $\mathbb{T}^{k+1}$). The compact group is $G = S^1$ and the continuous transformation involved is $\phi : \mathbb{T}^k \to S^1$ given by $\phi[(x_1, \ldots, x_k)] = x_k$. We know that this group extension is ergodic (Theorem 4.5). By Theorem 4.6, T is uniquely ergodic for $m = k + 1$. □

Note that the ergodic measure follows directly from the group extensions. It is the m-fold product of the Haar-measure, which is the Haar-Lebesgue measure on $\mathbb{T}^m$. Using the natural projection on the last coordinate, we have the following:

Corollary 4.8. Let $P(n)$ be a real-valued polynomial with irrational leading coefficient. Then, for $n \in \mathbb{N}$, $\{P(n)\}$ is uniformly distributed on $[0, 1]$.

Proof. The unique ergodicity of T implies the unique ergodicity of the transformation $\mathcal{T} : S^1 \to S^1$ satisfying $\mathcal{T}(x_m) = T[(x_1, \ldots, x_m)]$. We already know that the ergodic measure is the Lebesgue measure. The distribution now follows straightforwardly: a fractional part starts with c if and only if, for certain $n \in \mathbb{N}$, we have $\mathcal{T}^n(x) \in \left[c,\ c + 10^{-(\lfloor \log_{10} c \rfloor + 1)}\right)$ (note that $c + 10^{-(\lfloor \log_{10} c \rfloor + 1)}$ is the smallest number which is greater than c and has the same amount of digits as c). The Lebesgue measure of this interval is exactly $10^{-(\lfloor \log_{10} c \rfloor + 1)}$. It now follows that $\{P(n)\}$ is uniformly distributed on $[0, 1]$. □

Note how the preceding discussion only holds for polynomials with leading coefficient irrational. Of course, we also want to find the distribution in the case that the polynomial $P(n)$ has leading coefficient rational, but still has at least one irrational coefficient. Fortunately, the distribution in this case remains the same as before:

Corollary 4.9. Let $P(n)$ be a real-valued polynomial with at least one irrational coefficient. Then, for $n \in \mathbb{N}$, $\{P(n)\}$ is uniformly distributed on $[0, 1]$.

For the proof, see [2, Theorem 7.2.1] (in particular step 3 of the proof).

5 Summary

The aim of this thesis was to solve some number-theoretic problems using ergodic theory. The first problem was to determine in what way the first digits of powers of natural numbers are distributed. We were also interested in the way this distribution would be affected if we looked at the simultaneous distribution. The second problem dealt with continued fractions. Given an arbitrary number, what can we say about the distribution of the coefficients in the continued fraction representation (i.e., how often can we expect to see a positive integer appear as coefficient?). In this case, we were also interested in the way this distribution would be affected when we would look at the simultaneous distribution. The final problem was about the fractional parts of polynomials. Once again, we were interested in the distribution of the fractional parts. But, before we could tackle these problems, we had to brush up on ergodic theory.

In the first chapter we gave a short introduction to ergodic theory. The chapter is by no means a thorough introduction to ergodic theory, but it discussed those parts that would be useful in later chapters. Important notions we discussed in this chapter are ergodicity, unique ergodicity and (weak-)mixing. We also discussed interpretations and properties of these concepts.
The second chapter was about the problem of the first digits of powers. After a short introduction to the problem we saw that this problem could be formulated in terms of translations on the n-dimensional torus $\mathbb{T}^n$. Theorem 2.5 is the main result of this chapter.
The problem of continued fractions was studied in the third chapter. We needed the Gauss transformation to solve this problem. During the discussion of this transformation, we shortly touched upon the subject of piecewise monotonic transformations. The main result of this chapter is Theorem 3.12.
In the fourth chapter we took on the final problem. If anything, this chapter has shown that not all number-theoretic problems can be solved in a straightforward manner. We considered a certain affine transformation on the n-dimensional torus $\mathbb{T}^n$ and shortly mentioned group extensions. Corollary 4.9 is the main result of this chapter.

References

[1] M. Brin, G. Stuck. *Introduction to dynamical systems*, Cambridge University Press, 2002.

[2] I.P. Cornfeld, S.V. Fomin, Ya. Sinai. *Ergodic Theory*, Springer-Verlag, 1982.

[3] H. Furstenberg. *Recurrence in ergodic theory and combinatorial number theory*, Princeton University Press, 1981.

[4] G.H. Hardy, E.M. Wright. *An introduction to the theory of numbers*, Oxford University Press, 1979.

[5] P. Walters. *An introduction to ergodic theory*, Springer-Verlag, 2002.