Chapter 1

Introduction

Dynamics is a discipline with roots that go back at least to the time of Newton and questions about the motion of heavenly bodies. More recently, dynamics has grown into a broad discipline that encompasses differential equations, population biology models and group actions on homogeneous spaces. In the late 1800's, Poincaré initiated a qualitative approach to dynamical problems and, following this approach, various fields have emerged where the phase space of the dynamical system (the space that represents all possible states of a system in motion) is an abstract mathematical space. A dynamical system, in this more abstract approach, consists of a set X and a function or transformation T defined on X and with values in X. In ergodic theory, also known as measurable dynamics, X and T are given a measurable structure. We will occasionally deal with questions from topological dynamics, where X is a topological space, usually a compact metric space, and T a continuous map or a homeomorphism.

Ergodic theory uses tools of measure theory to study the long-term behavior of abstract dynamical systems. Its origins lie in statistical mechanics with, for example, the study of hard-sphere gases. While this is outside the scope of this book, we describe briefly how a dynamical system is obtained from this setting. To study the time evolution of n particles in space, each particle can be represented by a position vector and a momentum vector, so with six numbers for

each particle, a point in $\mathbb{R}^{6n}$ can be thought of as representing the state of the system at a moment in time. The physical laws that govern the motion can be modeled by differential equations and, using a classical theorem of Liouville, it can be shown that there are regions of the phase space, which sometimes are compact sets, where there is a measure that respects the motion (a concept that is called an invariant measure and that is studied in Chapter 3). In this way one obtains a law of motion, in this case a flow, defined on a measure space.

In the system that we just discussed, as in most systems arising in nature, time is continuous. Before describing a reduction to the discrete-time case we introduce some notation. Let x be a point in the phase space and let $T_t(x)$ represent the state of the system after t units of time have elapsed from the moment the system is in state x. We observe two properties of T_t. First, when $t = 0$ the map T_0 should be the identity transformation, and next if $T_{t_1}(x)$ is the state of the system after t_1 units of time have elapsed, the state after an additional t_2 units of time should be the same as the state after $t_1 + t_2$ units of time have elapsed from the initial configuration. So the family of maps $\{T_t\}$ should satisfy the following equations:

$$T_0(x) = x,$$
$$T_{t_1+t_2}(x) = T_{t_1}(T_{t_2}(x)).$$

A family of maps T_t satisfying this "time invariance" property is called a flow; we think of it as modeling the evolution of continuous time in a system. While continuous-time systems are the first systems to arise in a natural way, discrete-time dynamical systems had already been suggested by Poincaré as good approximations of continuous systems, and they are important as they show the qualitative range of behavior (such as ergodicity and mixing) that is present in continuous-time systems. For example, a motion picture is a discrete approximation of the reality it is filming. A discrete-time dynamical system may be obtained from any continuous system by fixing an initial time and considering its integer multiples. Thus, to approximate the dynamics of a continuous system one can fix a "small" time t_0 and analyze the motion of the systems at times that are integer multiples of t_0. More precisely, define a self-map or transformation T of the measure space

by $T(x) = T_{t_0}(x)$. So, given an initial state x, $T(x)$ represents the state of the system after t_0 units of time. Then $T^2(x) = T(T(x)) = T_{t_0+t_0}(x)$ would represent the state of the system at time $t_0+t_0 = 2t_0$ units of time, and $T^n(x)$ is similarly defined to represent the state of the system after nt_0 units of time when starting at x. Thus, one arrives at discrete-time dynamical systems defined on measure spaces, the subject on which this text focuses.

The next important simplification in ergodic theory reduces the phase space X to an interval in $\mathbb{R}$, or more generally, a Legesgue space. It can be shown that measure-theoretically (i.e., up to a measurable isomorphism) all nonpathological spaces (e.g., complete separable metric spaces with a Borel measure) are isomorphic to a finite or infinite interval in $\mathbb{R}$ (with possibly a countable number of isolated points of positive mass called atoms). Therefore, our first study is that of transformations T on a space X that is usually a finite, or in some cases infinite, interval in $\mathbb{R}$. Chapter 2 develops the theory of Lebesgue measure on the real line, and then the basic concepts of measure theory that are needed for the study of the dynamical properties of recurrence and ergodicity. Some examples of dynamical systems are more naturally described on other measure spaces such as symbolic spaces or subsets of $\mathbb{R}^d$, so we briefly discuss how measure theory on the line generalizes to these spaces.

An important development, that we do not cover in this book, in ergodic theory was the introduction by Kolmogorov of the notion of entropy. Using entropy as developed by Kolmogorov and Sinai, Ornstein proved in 1970 a remarkable theorem classifying Bernoulli automorphisms up to isomorphism. Our development of ergodic theory stops short of the notion of entropy and we refer to the Bibliographical Notes for further references.

While ergodic theory originated in statistical mechanics and initially dealt only with measure-preserving transformations on finite measure spaces, already in the 1930's mathematicians were interested in infinite measure spaces, and later in relaxing the measure-preserving condition and considering what we call nonsingular transformations. We emphasize finite measure-preserving transformations

but also consider notions such as recurrence and ergodicity in the context of infinite measure-preserving transformation in Chapter 3.

Equally important as the examples that have come from physics are examples and questions that originated in other areas of mathematics. One of the early results in ergodic theory is the beautiful theorem of Weyl on the equidistribution of numbers, which is studied in Chapter 5. A more recent example, that is only mentioned without proof in this book, is Furstenberg's ergodic theoretic proof, published in 1977, of the celebrated theorem of Szemerédi on arithmetic progressions. This theorem states that any set of integers of positive upper (Banach) density has infinitely many arithmetic progressions of a given length. To prove this theorem Furstenberg associated to each set of integers of positive density a finite measure space and a transformation defined on it. He then proved a remarkable theorem known as the Multiple Recurrence Theorem which, under this correspondence, implies the Szemerédi theorem. The methods in Furstenberg's proof have proven useful in solving other problems. In particular they have played a role in Green and Tao's recent solution of a 300-year old problem on the existence of arithmetic progressions in the primes.