Some topics in ergodic theory.
WORKING DRAFT

Yuri Bakhtin

CHAPTER 1

Introduction

I am going to adopt the view that the main purpose of ergodic theory is to study statistical patterns in deterministic or random dynamical systems and, specifically, how these patterns depend on the initial state of the system. Let me first try to explain this without going to technical details.

Suppose that we have a system X evolving in a phase space $\mathbb{X}$. This means that at each time t the state of the system denoted by X_t belongs to $\mathbb{X}$. The system may be rather complex, e.g., X can be a turbulent flow of particles in a fluid or gas and at any given time the description of the state of the flow should involve positions and velocities of all these particles. All these details of the flow at any given time have to be encoded as an element of $\mathbb{X}$.

Studying such a system experimentally may be a difficult task because if the behavior of the system is rich, then tracking all the details of the evolution may be hard or impossible. In many cases predicting the trajectory of such a system may be problematic as well, sometimes due to analytic or computational difficulties, sometimes due to intrinsic randomness or lack of stability in the system and due to its sensitivity to the initial data.

However, one may try a statistical approach to describing the properties of the system. Although concrete realizations $(X_t)_{t\geq 0}$ of the system may be erratic, it is still possible that these realizations follow steady statistical patterns.

A usual approach to experimental analysis of the system is to collect statistical information about the system by making repeated observations or measurements. From the mathematical point of view, a measurement is a way to assign to any state $x \in \mathbb{X}$ a number $f(x)$. In other words, it is a function $f : \mathbb{X} \to \mathbb{R}$. In particular, a measurement f of system X at time t, produces a number $f(X_t)$. One such measurement hardly tells much about the system, so a natural way to proceed is to conduct the same measurement at different times and find the average of the obtained values. If the measurements are made at times $1, 2, \ldots$, then the result of this procedure is

$$\frac{f(X_1) + \ldots + f(X_n)}{n}.$$

The hope is that these time-averaged measurements can serve as approximations to a "true" value one is trying to estimate. Moreover, one hopes that the more observations we average, the closer the resulting estimate is to the

ideal true value. In other words, collecting more and more statistics is useful only if a version of the law of large numbers holds, namely, that there is a number $\bar{f}$ such that $\frac{f(X_1)+\ldots+f(X_n)}{n} \to \bar{f}$ as $n \to \infty$. Moreover, that infinite time horizon average should be independent of the specific initial state of the system or at least be stable with respect to a class of perturbations of the initial state. Otherwise it is hard to assign a meaning to this limiting value. So, some questions that naturally arise are: Does the system possess any statistical regularity, i.e., does it make sense to collect statistics about the system? How do the statistical properties of the system depend on the initial condition? Does the system tend to remember the initial condition or does it forget it in the long run? What is remembered and what is forgotten in the long run?

One way to look at these issues is the *stability* point of view. I would like to illustrate that with a couple of standard mathematical examples.

The first very simple example is a deterministic linear dynamical system with one stable fixed point. A discrete dynamical system is given by a transformation θ of a phase space $\mathbb{X}$. For our example, we take the phase space $\mathbb{X}$ to be the real line $\mathbb{R}$ and define the transformation θ by $\theta(x) = ax$, $x \in \mathbb{R}$, where a is a real number between 0 and 1. To any point $x \in \mathbb{X}$ one can associate its forward orbit $(X_n)_{n=0}^{\infty}$, a sequence of points obtained from $X_0 = x$ by iterations of the map θ, i.e., $X_n = \theta(X_{n-1})$ for all $n \in \mathbb{N}$:

$$
\begin{aligned}
&X_0 = x = \theta^0(x)\\
&X_1 = \theta(X_0) = \theta(x) = \theta^1(x),\\
&X_2 = \theta(X_1) = \theta \circ \theta(x) = \theta^2(x),\\
&X_3 = \theta(X_2) = \theta \circ \theta \circ \theta(x) = \theta^3(x),\\
&\ldots.
\end{aligned}
$$

We are interested in the behavior of the forward orbit $(X_n)_{n=0}^{\infty}$ of x as $n \to \infty$, where n plays the role of time. In this simple example, the analysis is straightforward. Namely, zero is a unique fixed point of the transformation: $\theta(0) = 0$, and since $X_n = a^n x$, $n \in \mathbb{N}$ and $a \in (0, 1)$ we conclude that as $n \to \infty$, X_n converges to that fixed point exponentially fast. Therefore, 0 is a stable fixed point, or a one-point global attractor for the dynamical system defined by θ, i.e., its domain of attraction coincides with $\mathbb{R}$. So, due to the contraction and intrinsic stability that is present in the map θ, there is a fast loss of memory in the system, and no matter what the initial condition is, it gets forgotten in the long run and the points $X_n = \theta^n(x)$ approach the stable fixed point 0 as $n \to \infty$.

A completely different example is the following: let $\mathbb{X} = [0, 1)$ and let θ be defined by $\theta(x) = \{2x\}$. This system exhibits no stability at all. To see that, it is convenient to look at this system using binary representations of numbers in $[0, 1)$: for each $x \in [0, 1)$ there is a sequence $(x_i)_{i=1}^{\infty}$ of numbers

$x_i \in \{0,1\}$ such that $x = \overline{0.x_1x_2x_3\ldots}$ in binary notation, i.e.,

$$x = \sum_{i=1}^{\infty} \frac{x_i}{2^i}.$$

This representation is unique for all x except dyadic numbers. These are numbers with only finitely many 1's in their binary representation. Each dyadic number x allows for one more representation:

$$\overline{0.x_1x_2x_3\ldots x_{m-1}10000\ldots} = \overline{0.x_1x_2x_3\ldots x_{m-1}01111\ldots},$$

so to preserve the uniqueness of binary representations, let us agree that representations ending with $\overline{1111\ldots}$ are not allowed.

The transformation θ acts on binary representations as a shift:

$$\theta(\overline{0.x_1x_2x_3\ldots}) = \overline{0.x_2x_3x_4\ldots}.$$

Therefore,

$$\theta^n(\overline{0.x_1x_2x_3\ldots}) = \overline{0.x_{1+n}x_{2+n}x_{3+n}\ldots}.$$

To see that there is no stability in this system, let us start with any point $x = \overline{0.x_1x_2x_3\ldots} \in [0,1)$. Let us now take another point $y = \overline{0.y_1y_2y_3\ldots} \in [0,1)$. If $y_n \neq x_n$ for infinitely many values of n, then for those values of n, the iterates of $\theta^n(x)$ and $\theta^n(y)$ will not be close, one belonging to $[0,1/2)$ and another to $[1/2,1)$.

We see that stability at the level of trajectories does not hold in the usual sense. However, a statistical version of stability holds true. Namely, one can prove that if f is an integrable Borel function on $[0,1)$, then for almost every $x \in [0,1)$ with respect to the Lebesgue measure on $[0,1)$,

$$\lim_{n\to\infty} \frac{f(x) + f(\theta(x)) + f(\theta^2(x)) + \ldots + f(\theta^{n-1}(x))}{n} = \int_{[0,1)} f(x)dx. \tag{1.1}$$

Since the right-hand side represents the average of f with respect to the Lebesgue measure on $[0,1)$, one can say that, in the long run, statistical properties of this dynamical system are described by the Lebesgue measure or uniform distribution on $[0,1)$. In fact, one interpretation of identity (1.1) is that for Lebesgue almost every initial condition x the *empirical distribution* or *empirical measure* generated by θ and assigning mass n^{-1} to each point $x, \theta(x), \ldots, \theta^{n-1}(x)$ converges to the Lebesgue measure. Since the limiting measure is the same for a broad class of initial conditions x, we can speak of statistical stability: in the long run the initial value gets forgotten in the statistical properties of the system. We also can interpret the convergence in (1.1) as a law of large numbers.

We see that the Lebesgue measure plays a special role for this system since it describes the limits in (1.1). The reason for this is, as we will see is (a) that the Lebesgue measure is *invariant* under θ and (b) it is *ergodic*, i.e., it is not decomposable into two nontrivial invariant measures.

Similar questions are natural to ask if one considers random maps instead of deterministic ones. One natural way the random maps emerge is via random perturbations of deterministic dynamics. Let us describe one example of this kind. It is a 'noisy' modification of our first example above. Recall that in that example we worked with the map $\theta : x \mapsto \theta x$, and 0 was a stable equilibrium point. Let us perturb this dynamical system with noise, i.e., a random perturbation that kick the system out of equilibrium. Suppose we have a sequence $(\xi_n)_{n\in\mathbb{Z}}$ of independent Gaussian random variables with mean 0 and variance σ^2 defined on some probability space $(\Omega, \mathcal{F}, \mathsf{P})$. For every $n \in \mathbb{Z}$ we will now define a random map $\theta_{n,\omega} : \mathbb{R} \to \mathbb{R}$ by

$$\theta_{n,\omega}(x) = ax + \xi_n(\omega).$$

This model is known as an autoregressive-moving-average (ARMA) model of order 1.

A natural analogue of a forward orbit from our first example would be a stochastic process $(X_n)_{n\geq 0}$ emitted from a point $x_0 \in \mathbb{R}$, i.e., satisfying $X_0 = x$ and, for all $n \in \mathbb{N}$,

$$X_n = aX_{n-1} + \xi_n. \tag{1.2}$$

However, the stability issue it is not as straightforward here as in the deterministic case. It is clear that there is no fixed equilibrium point that would serve all maps $\theta_{n,\omega}$ at the same time. The solution of the equation $\theta_{n,\omega}(y) = y$ for some n may be irrelevant for all other values of n. Still this system allows for an ergodic result similar to (1.1). Let μ be the Gaussian measure with mean 0 and variance $\sigma^2/(1 - a^2)$. Then, for any Borel function f integrable with respect to μ, we have that for almost every $(x_0, \omega) \in \mathbb{R} \times \Omega$ with respect to $\mu \times \mathsf{P}$,

$$\lim_{n\to\infty} \frac{f(X_0) + f(X_1) + f(X_2) + \ldots + f(X_{n-1})}{n} = \int_{\mathbb{R}} f(x)\mu(dx).$$

The underlying reason for this result is that μ is a unique invariant distribution for the Markov semigroup associated with the process. This example demonstrates a situation where there is no stability in the straightforward deterministic sense, but there is statistical stability.

Roughly speaking the ergodic approach to stochastic stability is to base the study of the system on the description of the set of invariant distributions and their properties, and so this is the main theme of these notes in the context of random dynamics. Nevertheless, we will begin with the basics of ergodic theory of deterministic transformations. The full tentative plan is to study

(1) Generalities on ergodic theory of deterministic transformations.
(2) Generalities on ergodic theory of random transformations and Markov processes.
(3) Finite and countable state space. Markov chains.
(4) Random dynamics in Euclidean spaces

(5) Continuous time. SDEs.
(6) SPDEs: Stochastic Navier-Stokes, Burgers equations.
(7) Lyapunov exponents and multiplicative ergodic theory.

In the end of this introduction let us discuss the roots of ergodic theory and the origin of the word *ergodic*. In the second half of the nineteenth century, L. Boltzmann worked on foundations of statistical mechanics that was supposed to connect the kinetic theory of gases with thermodynamics. The main idea was that gases in equilibrium obey a distribution and all macroscopic quantities of interest can be derived from that distribution. Boltzmann introduced the notion of *ensemble* (that corresponds to Kolmogorov's notion of probability space) and, among other things, was concerned with the connection with the frequentist notion of probability since computing frequencies is a form of time averaging.

The systems of gas particles Boltzmann was considering preserve the total energy. He called an energy-preserving system an *Ergode* if there were no other conserved quantities, and he introduced the term *Ergodentheorie* for the study of such systems. The greek root *erg* in both *energy* and *ergode* relates to work or activity (*ergon* means *work* in ancient Greek). It takes no external work to move between states with the same energy, and when studying such systems Boltzmann was led to what may be called a weak form of the *ergodic hypothesis*: that the path that the system follows explores all available configurations of particles and their velocities with the same total energy. (The Greek word for road or way is *odos*)

The stronger version of the ergodic hypothesis is usually formulated as "time averages equal ensemble averages", and can be understood as a form of law of large numbers. So, if $X_1, X_2, \ldots$ are configurations that the system evolves through and f is any observable, then

$$\frac{f(X_1) + \ldots + f(X_n)}{n} \to \int_{\mathbb{X}} f(x)\mathsf{P}(dx),$$

where P is the equilibrium distribution, and the right-hand side plays the role of the ensemble average. Of course, no rigorous measure and integration theory or probability theory existed at that time, so the mathematical treatment of these issues had to wait until the 20th century.

CHAPTER 2

Measure-preserving transformations and other basic notions of ergodic theory for deterministic systems

This chapter is devoted to a very sketchy introduction to the ergodic theory of deterministic measure-preserving transformations. Many more topics are studied in introductory texts [**Wal82**], [**Sin76**]. A comprehensive modern course on dynamical systems is [**KH95**].

1. Example: circle rotations

Let us start with an example of a simple dynamical system. Consider a circle $\mathbf{S}^1$. It can be understood as the real line modulo 1: $\mathbf{S}^1 = \mathbb{R}^1/\mathbb{Z}^1$, i.e., it is the result of identification of points on the the real line $\mathbb{R}^1$ with common fractional part. It can also be parametrized by the segment $[0, 1]$ with identified endpoints.

A rotation of the circle is a transformation $\theta : \mathbf{S}^1 \to \mathbf{S}^1$ given by

$$\theta(\omega) = \omega + \alpha \pmod 1, \quad \omega \in \mathbf{S}^1, \tag{2.1}$$

where α is a real number. Let us look at what happens if we keep watching the dynamics emerging under iterations of the map θ, i.e., the compositions $\theta^1 = \theta$, $\theta^2 = \theta \circ \theta$, $\theta^3 = \theta \circ \theta \circ \theta$, etc. The number of iterations plays the role of time, and we are interested in statistical patterns in these iterations over long time intervals.

If the rotation angle α is rational and can be represented as $\alpha = m/n$ for some coprime numbers $m, n \in \mathbb{N}$, then for all $\omega \in \mathbf{S}$, all the points $\omega, \theta\omega, \theta^2\omega, \dots, \theta^{n-1}\omega$ are distinct, but $\theta^n\omega = \omega$ i.e., after n iterations of θ the circle returns to its initial position and then the next n iterations $\theta^n\omega, \dots, \theta^{2n-1}\omega$ coincide with $\omega, \theta\omega \dots, \theta^{n-1}\omega$, and then this repeats for $\theta^{2n}\omega, \dots, \theta^{3n-1}\omega$, and so on. In other words all points ω are *periodic* with period n.

A more interesting situation occurs if α is irrational. Then there are no periodic points and one can show that the *orbit* $\{\omega, \theta\omega, \theta^2\omega \dots\}$ is dense in $\mathbf{S}^1$, and, moreover the following equidistribution theorem established almost simultaneously by P. Bohl [**Boh09**], H. Weyl [**Wey10**], and W. Sierpinski [**Sie10**] (although I have not looked into these papers) holds:

THEOREM 2.1. *Let α be irrational. Then for any point $\omega \in \mathbf{S}^1$, and any interval $I = (a, b) \subset \mathbf{S}^1$,*

$$\lim_{n\to\infty} \frac{\sum_{j=0}^{n-1} \mathbf{1}_{\theta^n\omega\in I}}{n} = b - a. \tag{2.2}$$

The sum in the left-hand side computes the total number of visits of the sequence $(\theta^j\omega)_{j=0}^{\infty}$ to I before time n, and the ratio computes the frequency of those visits relative to the total time n. Notice that the limiting value on the right-hand side equals to $\mathrm{Leb}(I)$, the Lebesgue measure of I. Also notice that the Lebesgue measure plays a special role for this dynamical system, namely it is invariant under rotation θ. It should be intuitively clear what this means. First, imagine mass uniformly spread over the circle and then rotate the circle. The distribution of mass in the circle after the rotation is still uniform, so it coincides with the initial mass distribution.

We will see soon that the fact that the statistics of visits to intervals is described in the long run by an invariant distribution is a general phenomenon explained by ergodic theorems. Here we note that it is easy to extend this result to the following:

THEOREM 2.2. *Let α be irrational. Then for any point $\omega \in \mathbf{S}^1$ and any continuous function $f : \mathbf{S}^1 \to \mathbb{R}$,*

$$\lim_{n\to\infty} \frac{\sum_{j=0}^{n-1} f(\theta^j\omega)}{n} = \int_{\mathbf{S}^1} f(x)dx. \tag{2.3}$$

DERIVATION OF THEOREM 2.2 FROM THEOREM 2.1: Identity (2.2) is a specific case of identity (2.3) for $f = \mathbf{1}_I$. Given (2.2), we can use linearity of sums and integrals to derive (2.3) for functions f representable as

$$f(\omega) = \sum_{k=1}^{m} c_k \mathbf{1}_{I_k}(\omega),$$

for some system of intervals $I_k \subset \mathbf{S}^1$, $k = 1, \dots, m$. Suppose now f is continuous. Therefore, it is uniformly continuous and for every $\varepsilon > 0$, we can find a step function f_ε such that $|f(\omega) - f_\varepsilon(\omega)| < \varepsilon$ for all $\omega \in \mathbf{S}^1$. Then

$$\left|\frac{\sum_{j=0}^{n-1} f(\theta^j\omega)}{n} - \int_{\mathbf{S}^1} f(x)dx\right| \le \left|\frac{\sum_{j=0}^{n-1} f_\varepsilon(\theta^j\omega)}{n} - \int_{\mathbf{S}^1} f_\varepsilon(x)dx\right| + 2\varepsilon, \quad n \in \mathbb{N},$$

so $\limsup_{n\to\infty}$ of the left-hand side does not exceed 2ε. Since $\varepsilon > 0$ is arbitrary, we conclude that the limit of the left-hand side equals zero. □

Theorem 2.2 can be interpreted so that for every observable f from a broad class of functions, the time averages approximate the space averages.

Of course, this property fails for rational values of $\alpha = m/n$ despite the fact that the Lebesgue measure is still invariant. The reason is that in that case, for any $\omega \in \mathbf{S}^1$, the orbit $(\theta^j\omega)_{j\ge0}$ explores only a very small part of $\mathbf{S}^1$ — a finite set, in fact, — whereas the integral in the right-hand side of (2.3) depends on the behavior of f on the entire $\mathbf{S}^1$. This also can be

explained by saying that $\mathbf{S}^1$ is decomposable: one can split $\mathbf{S}^1$ into smaller massive sets composed of periodic orbits such that the dynamical system can be confined to any of these sets and studied independently on each of them.

Moreover, due to this decomposability of the dynamics, the Lebesgue measure is not a unique invariant measure. For example, we can choose one periodic orbit $(\omega, \theta\omega, \dots, \theta^{n-1}\omega)$ and obtain an invariant measure by assigning mass n^{-1} to each of its points:

$$\mu = \frac{1}{n}\sum_{j=0}^{n-1} \delta_{\theta^j\omega},$$

where δ_x denotes the Dirac measure concentrated at x, i.e., for any set A,

$$\delta_x(A) = \mathbf{1}_{x\in A}. \tag{2.4}$$

Notice that this uniform distribution along the orbit of ω is not possible for nonperiodic orbits.

We will see later that the Lebesgue measure can be decomposed into a combination of these discrete invariant measures, i.e., can be represented as their (continual) mixture.

Let us now introduce some rigorous definitions in a more abstract framework.

2. Measure-preserving transformations. Formal definitions

Throughout these notes book we will use concepts from measure-theory based probability theory introduced by A.Kolmogorov in 1930's, see [**Kol50**]. For a comprehensive introduction to probability, we recommend the book by A.Shiryaev [**Shi96**], although this is largely a question of habit and taste. There are several good books that can serve this purpose equally well. For example, a more recent textbook [**KS07**] contains a concise introduction to the mathematical theory of probability and stochastic processes.

We will often work with measurable spaces $(\Omega, \mathcal{F})$, where Ω is any set and $\mathcal{F}$ is a σ-algebra on Ω. If P is a probability measure on $(\Omega, \mathcal{F})$, the triplet $(\Omega, \mathcal{F}, \mathsf{P})$ is called a probability space.

The first two definitions below require only the measurable structure and do not depend on measures.

Definition 2.1. Let $(\Omega, \mathcal{F})$ and $(\tilde\Omega, \tilde{\mathcal{F}})$ be two measurable spaces. A map $X : \Omega \to \tilde\Omega$ is called *measurable* with respect to $(\mathcal{F}, \tilde{\mathcal{F}})$ if for every $A \in \tilde{\mathcal{F}}$, we have $X^{-1}A = \{\omega : X(\omega) \in A\} \in \mathcal{F}$. Often, such a map is also called $\tilde\Omega$-valued random variable. In the case where $(\tilde\Omega, \tilde{\mathcal{F}}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, where $\mathcal{B}(\mathbb{R})$ denotes the Borel σ-algebra on $\mathbb{R}$, X is simply called a random variable.

A specific case of this definition is the following:

DEFINITION 2.2. Let $(\Omega, \mathcal{F})$ be a measurable space. A map or transformation $\theta : \Omega \to \Omega$ is called *measurable* if, for every $A \in \mathcal{F}$, we have $\theta^{-1}A = \{\omega : \theta\omega \in A\} \in \mathcal{F}$.

Here and often in these notes we follow the tradition to denote the value assigned by the transformation θ to argument ω by $\theta\omega$.

A measurable transformation θ defines a *semigroup* of measurable transformations $(\theta^n)_{n\in\mathbb{Z}_+}$. Here θ^n is inductively defined by $\theta^0 = \mathrm{Id}$ (identical transformation, i.e., $\mathrm{Id}(\omega) \equiv \omega$) and for $n \geq 1$, $\theta^n = \theta \circ \theta^{n-1}$. If a transformation is invertible, i.e., θ is a bijection of Ω onto itself, then θ^{-1} is well-defined and in a similar way θ generates a *group* of transformations $(\theta^n)_{n\in\mathbb{Z}}$. Identity $\theta^m \circ \theta^n = \theta^{m+n}$ is valid for all $m, n \in \mathbb{Z}_+$ in the case of the semigroup and for all $m, n \in \mathbb{Z}$ in the case of the group.

Often in these notes we will omit the composition sign "$\circ$" for brevity. For example, this convention allows to rewrite the above group identity written as $\theta^m\theta^n = \theta^{m+n}$.

DEFINITION 2.3. Let $(\Omega, \mathcal{F}, \mathsf{P})$ be a probability space and $(\tilde{\Omega}, \tilde{\mathcal{F}})$ be a measurable space. Let a map $X : \Omega \to \tilde{\Omega}$ be $(\mathcal{F}, \tilde{\mathcal{F}})$-measurable. The *pushforward* of P under X is a measure on $(\Omega, \tilde{\mathcal{F}})$ denoted by $\mathsf{P}\theta^{-1}$ and defined by

$$\mathsf{P}X^{-1}(A) = \mathsf{P}(X^{-1}A) = \mathsf{P}\{\omega : X(\omega) \in A\}, \quad A \in \mathcal{F}.$$

DEFINITION 2.4. Let $(\Omega, \mathcal{F}, \mathsf{P})$ be a probability space and suppose transformation $\theta : \Omega \to \Omega$ be measurable. If the pushforward $\mathsf{P}\theta^{-1}$ coincides with P, i.e.,

$$\mathsf{P}(\theta^{-1}A) = \mathsf{P}(A), \quad A \in \mathcal{F}, \tag{2.5}$$

then we say that P is *invariant* under θ, or that θ *preserves* P, or that θ is P*-preserving.*

It is important to notice that the definition of invariance is based on pre-images under θ, and not forward images. The reason for that is that the pushforwards of measures are defined in terms of pre-images and there is no natural notion of pullbacks of measures.

PROBLEM 2.1. *Suppose transformation θ preserves a measure P. Prove that for any $n \in \mathbb{Z}_+$, θ^n also preserves P. Suppose additionally that θ is invertible and θ^{-1} is measurable. Prove that then a stronger statement holds: for any $n \in \mathbb{Z}$, θ^n also preserves P (in particular, θ^{-1} is P-preserving).*

Let us consider some basic examples of probability spaces with measure-preserving transformations.

Let $\Omega = \{0, 1, \ldots, n-1\}$, $\mathcal{F} = 2^\Omega$ (σ-algebra of all subsets of Ω), and let P be the uniform distribution on Ω. Then a transformation θ is P-preserving if and only if it is a permutation, i.e., a one-to-one map. For instance, the cyclic map $\theta\omega = \omega + 1 \pmod{n}$ is P-preserving.

Continuous analogues of this cyclic transformation are circle rotations preserving the Lebesgue measure on $\mathbf{S}^1$ equipped with the Borel σ-algebra, and, more generally, shifts on a d-dimensional torus $\mathbb{T}^d = \mathbb{R}^d/\mathbb{Z}^d$ given by

$$\theta(\omega_1, \dots, \omega_d) = (\omega_1 + \alpha_1 \pmod 1, \dots, \omega_d + \alpha_d \pmod 1), \tag{2.6}$$

where $\alpha = (\alpha_1, \dots, \alpha_d)$ is a fixed vector in $\mathbb{R}^d$.

One more class of examples transformations preserving the Lebesgue measure is provided by interval exchange transformations. Suppose the probability space is $([0,1), \mathcal{B}([0,1)), \mathrm{Leb})$. Let us take $m \in \mathbb{N}$ and let intervals $I_k = [a_k, b_k)$, $k = 1, \dots, m$, form a partitions of $[0,1)$, i.e., these intervals are mutually disjoint and their union is $[0,1)$. Let $I'_k = [a'_k, b'_k)$, $k = 1, \dots, m$, form another partition such that $b'_k - a'_k = b_k - a_k$, $k = 1, \dots, m$. Then one can define a transformation θ of $[0,1)$ such that for each I_k it coincides with a translation of I_k onto I'_k. It is easy to see that thus defined transformation preserves the Lebesgue measure.

It is often useful to check invariance in terms of test functions. From this point on we will often denote expectation (integral) of a random variable X defined on $(\Omega, \mathcal{F}, \mathsf{P})$ by $\mathsf{E}X$ or $\mathsf{E}X(\omega)$.

LEMMA 2.1. *A measurable map θ on a probability space $(\Omega, \mathcal{F}, \mathsf{P})$ preserves P if and only if for any bounded measurable function $X : \Omega \to \mathbb{R}$,*

$$\mathsf{E}X(\omega) = \mathsf{E}X(\theta\omega). \tag{2.7}$$

PROOF: Definition 2.4 is equivalent to (2.7) for indicator functions $X = \mathbf{1}_A$, $A \in \mathcal{F}$. So sufficiency of (2.7) is obvious. On the other hand, if (2.7) holds for indicators, then it holds for simple functions (finite linear combinations of indicators). Also, any bounded measurable function X can be approximated by a simple one: for any $\varepsilon > 0$ there is a simple function X_ε such that $|X(\omega) - X_\varepsilon(\omega)| \le \varepsilon$. Since (2.7) holds for X_ε, we obtain $|\mathsf{E}X(\omega) - \mathsf{E}X(\theta\omega)| \le 2\varepsilon$. Since ε is arbitrary, (2.7) follows. □

DEFINITION 2.5. *A probability space $(\Omega, \mathcal{F}, \mathsf{P})$ equipped with a measure-preserving transformation θ is often called a metric dynamical system or a measure-preserving dynamical system. Quadruplet notation $(\Omega, \mathcal{F}, \mathsf{P}, \theta)$ is often used.*

3. Poincaré's recurrence theorem

The first truly contentful result that we are going to prove is the following recurrence theorem proved by H. Poincaré [**Poi90**], [**Poi99**].

THEOREM 2.3. *Suppose $(\Omega, \mathcal{F}, \mathsf{P})$ is a probability space, and θ is a P-preserving transformation of Ω. Then for any set $A \in \mathcal{F}$ we have $\mathsf{P}(T(A)) = 0$, where $T(A)$ is set of all points in A that never return to A under iterations of θ, i.e.,*

$$T(A) = \{\omega \in A : \ \theta^n\omega \notin A, \ n \in \mathbb{N}\}.$$

PROOF: Let us prove first that the sets $\theta^{-n}T(A)$, $n \in \mathbb{N}$ are mutually disjoint. In fact, if $\omega \in \theta^{-n}T(A) \cap \theta^{-m}T(A)$ and $m < n$, then we denote $\omega' = \theta^m\omega$ and notice that $\omega' \in T(A)$, and at the same time $\theta^{n-m}\omega' = \theta^{n-m}\theta^m\omega = \omega \in T(A)$ which is a contradiction with the definition of $T(A)$.

Due to θ-invariance of P, $\mathsf{P}(\theta^{-n}T(A)) = \mathsf{P}(T(A))$ for all n, so

$$\sum_{n\in\mathbb{N}} \mathsf{P}(T(A)) = \sum_{n\in\mathbb{N}} \mathsf{P}(\theta^{-n}T(A)) = \mathsf{P}\left(\bigcup_{n\in\mathbb{N}} \theta^{-n}T(A)\right) \le \mathsf{P}(\Omega),$$

which can hold true only if $\mathsf{P}(T(A)) = 0$. □

It is possible to give a quantitative strengthening of this theorem. The following is an adaptation from [**Tao08**].

THEOREM 2.4. *Under the conditions of Theorem 2.3,*

$$\limsup_{n\to\infty} \mathsf{P}(\theta^{-n}A \cap A) \ge \mathsf{P}^2(A), \quad A \in \mathcal{F}.$$

PROOF: The invariance of P implies that for any $N \in \mathbb{N}$,

$$\mathsf{E}\left[\sum_{n=1}^{N} \mathbf{1}_{\theta^{-n}A}\right] = N\mathsf{P}(A).$$

Using the Cauchy–Schwartz inequality or Lyapunov inequality, we obtain

$$\mathsf{E}\left[\sum_{n=1}^{N} \mathbf{1}_{\theta^{-n}A}\right]^2 \ge \left(\mathsf{E}\left[\sum_{n=1}^{N} \mathbf{1}_{\theta^{-n}A}\right]\right)^2 = N^2\mathsf{P}(A)^2. \tag{2.8}$$

The left-hand side of this inequality can be rewritten as

$$\mathsf{E}\left[\sum_{n=1}^{N} \mathbf{1}_{\theta^{-n}A}\right]^2 = \sum_{n,m=1}^{N} \mathsf{P}(\theta^{-n}A \cap \theta^{-m}A) = \sum_{n,m=1}^{N} \mathsf{P}(A \cap \theta^{-|m-n|}A). \tag{2.9}$$

Introducing $L = \limsup_{n\to\infty} \mathsf{P}(A \cap \theta^{-n}A)$, for any $\varepsilon > 0$, we can find $n_0 = n_0(\varepsilon)$ such that

$$\mathsf{P}(A \cap \theta^{-n}A) < L + \varepsilon, \quad n > n_0.$$

Let us now split the sum on the right-hand side of (2.9) into two: over n, m satisfying $|n-m| \ge n_0$ and over n, m satisfying $|n-m| < n_0$. Noticing that there are less than $2n_0N$ terms in the second sum and they all are bounded by 1, we obtain

$$\frac{1}{N^2} \sum_{n,m=1}^{N} \mathsf{P}(A \cap \theta^{-|m-n|}A) < L + \varepsilon + \frac{2n_0(\varepsilon)N}{N^2}.$$

Combining this with (2.8) and (2.9), we obtain

$$\mathsf{P}(A)^2 \le L + \varepsilon + \frac{2n_0(\varepsilon)}{N}.$$

Taking $N \to \infty$ and then $\varepsilon \to 0$, we obtain the desired result. □

PROBLEM 2.2. Using an approriate Bernoulli shift, prove that the lower bound in Theorem 2.4 is sharp, i.e., in general $\mathsf{P}^2(A)$ cannot be replaced by a larger value.

at this point the Bernoulli shift has not been introduced yet

Theorem 2.4 implies that for any set A of positive measure (no matter how small) there are infinitely many times n such that $P(\theta^{-n}A \cap A) > 0$. In particular, Theorem 2.3 can be viewed as a corollary of Theorem 2.4.

DERIVATION OF THEOREM 2.3 FROM THEOREM 2.4: Let us assume that $\mathsf{P}(T(A)) > 0$. Then Theorem 2.4 implies that $\mathsf{P}(\theta^{-n}T(A) \cap T(A))$ is positive for infinitely many $n \in \mathbb{N}$. For those n, there is $\omega \in \theta^{-n}T(A) \cap T(A) \subset \theta^{-n}A \cap A$. So, $\omega \in T(A)$, but $\theta^n\omega \in A$. This contradicts the definition of $T(A)$. □

There is an interesting Zermelo's paradox related to Poincaré's recurrence theorem. Consider a gas, i.e., a collection of particles in a box. The evolution of this system of particles can be described by a Hamiltonian dynamical system preserving the Lebesgue measure in the phase space encoding positions and momenta of all particles. If in the beginning all particles are in one half of the box, then as the system evolves it is natural to expect that these particles eventually will be more or less uniformly distributed between this half and the other one. Poincaré's recurrence theorem allows to conclude though that sooner or later there will be a moment when these particles will all gather within the same half of the box. This seems to contradict both intuition and experience. The explanation is that the time required for such an event to occur in a realistic setting is astronomically large, much larger than the length of any conceivable experiment, and so these events are never observed in practice.

4. Stochastic processes: basic notions

There is a useful interpretation of metric dynamical systems in terms of stationary processes, for example for any random variable $f : \mathbb{X} \to \mathbb{R}$, the process $(X_n)_{n\in\mathbb{N}}$ defined by $X_n(\omega) = f(\theta^n\omega)$ is a stationary process. First, let us recall some basic facts concerning stochastic processes in general.

DEFINITION 2.6. Let $(\mathbb{X}, \mathcal{X})$ be a measurable space and $\mathbb{T}$ be any set. Any collection of $\mathbb{X}$-valued random variables $(X_t)_{t\in\mathbb{T}}$ is called a stochastic process.

The set $\mathbb{T}$ in this definition plays the role of time axis, and if $\mathbb{T} \subset \mathbb{R}$, then for fixed $\omega \in \Omega$, the *realization* $(X_t(\omega))_{t\in\mathbb{T}}$ is also often called the *trajectory* or *sample path* corresponding to ω.

The space of these sample paths is denoted by

$$\mathbb{X}^{\mathbb{T}} = \{x : \mathbb{T} \to \mathbb{X}\}.$$

If $\mathbb{T} = \mathbb{N}$, then $\mathbb{X}^{\mathbb{T}}$ consists of one-sided sequences $(x_1, x_2, \ldots)$, and if $\mathbb{T} = \mathbb{Z}$, then $\mathbb{X}^{\mathbb{T}}$ consists of two-sided sequences $(\ldots, x_{-2}, x_{-1}, x_0, x_1, x_2, \ldots)$.

It is often useful to consider an $(\mathbb{X}, \mathcal{X})$-valued process X as one $\mathbb{X}^{\mathbb{T}}$-valued random variable, and so we need to introduce an appropriate σ-algebra on $\mathbb{X}^{\mathbb{T}}$.

DEFINITION 2.7. Suppose $\mathbb{T} \subset \mathbb{R}$. For $m \in \mathbb{N}$, $n_1 < \ldots < n_m$, and $A_1, \ldots, A_m \in \mathcal{X}$, we denote

$$C_{n_1,\ldots,n_m}(A_1, \ldots, A_m) = \{x \in \mathbb{X}^{\mathbb{T}} : \ x_{n_1} \in A_0, \ldots, x_{n_m} \in A_m\}. \tag{2.10}$$

Sets of this form are called *elementary cylinders*.

Let us stress that all cylinders have non-unique representations of the form (2.10). For example,

$$C_{n_1,\ldots,n_m,n_m+1}(A_1, \ldots, A_m, \mathbb{X}) = C_{n_1,\ldots,n_m}(A_1, \ldots, A_m).$$

DEFINITION 2.8. The *cylindric σ-algebra* $\mathcal{X}^{\mathbb{T}}$ is generated by all elementary cylinders.

For any time $n \in \mathbb{T}$, the projection $\pi_n : \mathbb{X}^{\mathbb{T}} \to \mathbb{X}$ is defined by $\pi_n(x) = x_n$.

THEOREM 2.5. *The cylindric σ-algebra $\mathcal{X}^{\mathbb{T}}$ is generated by maps π_n.*

The family of all elementary cylinders is a π-system, i.e., it is closed under intersection.

THEOREM 2.6 (see [**Shi96**, Lemma II.2.3]). *If the restrictions of two probability measures on a π-system $\mathcal{E}$ coincide, then the measures coincide on $\sigma(\mathcal{E})$.*

The following corollary says that a probability measure on $(\mathbb{X}^{\mathbb{T}}, \mathcal{X}^{\mathbb{T}})$ is uniquely defined by its values on elementary cylinders.

THEOREM 2.7. *Let P and Q be two measures on $(\mathbb{X}^{\mathbb{T}}, \mathcal{X}^{\mathbb{T}})$ such that for any elementary cylinder C, $\mathsf{P}(C) = \mathsf{Q}(C)$. Then $\mathsf{P} = \mathsf{Q}$.*

To formulate an existence result, we need to introduce general cylinders.

DEFINITION 2.9. Let $m \in \mathbb{N}$ and $n_1, \ldots, n_m \subset \mathbb{T}$ satisfy $n_1 < \ldots < n_m$, we denote by

$$\mathcal{X}^{n_1,\ldots,n_m} = \sigma(C_{n_1,\ldots,n_m}(A_1, \ldots, A_m), \quad A_1, \ldots, A_m \in \mathcal{X}).$$

Elements of $\mathcal{X}^{n_1,\ldots,n_m}$ are called *cylinders*.

THEOREM 2.8 (Kolmogorov–Daniell extension theorem). *If $(\mathbb{X}, \mathcal{X})$ is a Borel space and P is a nonnegative function on all cylinders such that for any $m \in \mathbb{N}$ and any $n_1, \ldots, n_m \subset \mathbb{T}$ satisfying $n_1 < \ldots < n_m$, the restriction of P to $\mathcal{X}^{n_1,\ldots,n_m}$ is a probability measure, then there is a unique measure on $\mathcal{X}^{\mathbb{T}}$ such that its restriction on cylinders coincides with P.*

Often in these notes we will work with Borel spaces, as in this theorem. Let us introduce the corresponding definitions.

DEFINITION 2.10. We say that measurable spaces $(\mathbb{X}, \mathcal{X})$ and $(\mathbb{Y}, \mathcal{Y})$ are measurably isomorphic if there is a bijection $\varphi : \mathbb{X} \to \mathbb{Y}$ such that φ is $(\mathcal{X}, \mathcal{Y})$-measurable and φ^{-1} is $(\mathcal{Y}, \mathcal{X})$-measurable.

DEFINITION 2.11. A measurable space $(\mathbb{Y}, \mathcal{Y})$ is called a Borel space if it is measurably isomorphic to some space $(\mathbb{X}, \mathcal{X})$, where $\mathbb{X}$ is a Borel subset of $[0, 1]$, and $\mathcal{X}$ is the σ-algebra induced by $\mathcal{B}([0, 1])$ on $\mathcal{X}$.

REMARK 2.1. The family of Borel spaces $(\mathbb{X}, \mathcal{X})$ is extremely rich and includes, for example, Borel subsets of Polish spaces, see [**Kur66**]. We recall that a metric set $(\mathbb{X}, \rho)$ is called Polish if it is complete and separable.

5. Interpretation of measure-preserving maps via stationary processes

In this section we describe the connection between stationary processes and measure-preserving transformations.

DEFINITION 2.12. Suppose $\mathbb{T} = \mathbb{N}$, $\mathbb{Z}_+$, or $\mathbb{Z}$ and let $(X_t)_{t\in\mathbb{T}}$ be a stochastic process with values in $\mathbb{X}$. This process is called *stationary* if for any $m \in \mathbb{N}$, any $i_1, \dots, i_m \in \mathbb{T}$, and any $r \in \mathbb{N}$, the distributions of $(X_{i_1}, \dots, X_{i_m})$ and $(X_{i_1+r}, \dots, X_{i_m+r})$ coincide, i.e., for any $A_1, \dots, A_m \in \mathcal{G}$,

$$\mathsf{P}\{X_{i_1} \in A_1, \dots, X_{i_m} \in A_m\} = \mathsf{P}\{X_{i_1+r} \in A_1, \dots, X_{i_m+r} \in A_m\}. \tag{2.11}$$

It is sufficient to check this definition for $r = 1$ since the statement for the general r follows by induction.

The simplest example of a stationary process is a sequence of independent and identically distributed (i.i.d.) random variables $(X_n)_{n\in\mathbb{N}}$. By definition, this means that there is a probability P on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that for any m and any $i_1, \dots, i_m \in \mathbb{N}$,

$$\mathsf{P}\{X_{i_1} \in A_1, \dots, X_{i_m} \in A_m\} = P(A_1)P(A_2)\dots P(A_m),$$

so identity (2.11) easily follows.

Intro to Gaussian processes — ?

LEMMA 2.2. *Suppose X is an $(\mathbb{X}, \mathcal{X})$-valued random variable on a probability space $(\Omega, \mathcal{F}, \mathsf{P})$, and θ is P-preserving transformation on Ω. Then the stochastic process $(X_n)_{n\in\mathbb{Z}_+}$ defined by*

$$X_n(\omega) = X(\theta^n\omega) \tag{2.12}$$

is stationary. If θ is measurably invertible, then the process $(X_n)_{n\in\mathbb{Z}}$ defined by the same formula is stationary.

PROOF: Let us check the definition of stationarity for shift $r = 1$:

$$\begin{aligned}
&\mathsf{P}\{\omega : X_{i_1+1}(\omega) \in A_1, \dots, X_{i_m+1}(\omega) \in A_m\} \\
=&\mathsf{P}\{\omega : X(\theta^{i_1+1}\omega) \in A_1, \dots, X(\theta^{i_m+1}\omega) \in A_m\} \\
=&\mathsf{P}\{\omega : X(\theta^{i_1}\theta\omega) \in A_1, \dots, X(\theta^{i_m}\theta\omega) \in A_m\} \\
=&\mathsf{P}\{\omega : X(\theta^{i_1}\omega) \in A_1, \dots, X(\theta^{i_m}\omega) \in A_m\} \\
=&\mathsf{P}\{\omega : X_{i_1}(\omega) \in A_1, \dots, X_{i_m}(\omega) \in A_m\},
\end{aligned}$$

where the third identity follows from the θ-invariance of P. □

One can also make sense of a converse statement: every stationary process can be represented in the form described in Lemma 2.2. Suppose $(X_n)_{n\in\mathbb{Z}_+}$ is an $(\mathbb{X}, \mathcal{X})$-valued stationary process on a probability space $(\Omega, \mathcal{F}, \mathsf{P})$.

One way to look at the process X is to view it as a map from Ω to the space of trajectories $\tilde{\Omega} = \mathbb{X}^{\mathbb{Z}_+}$, i.e., the space of all functions $x : \mathbb{Z}_+ \to \mathbb{X}$, i.e., sequences $x = (x_0, x_1, \dots)$, equipped with cylindric σ-algebra $\tilde{\mathcal{F}} = \mathcal{X}^{\mathbb{Z}_+}$. In fact, the process X is $(\mathcal{F}, \tilde{\mathcal{F}})$-measurable (to see this, it is sufficient to check that X-pre-images of all cylinders are in $\mathcal{F}$). Therefore, we can consider the pushforward $\tilde{\mathsf{P}} = \mathsf{P}X^{-1}$ of P to $(\tilde{\Omega}, \tilde{\mathcal{F}})$ under this map. The result is a probability space $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathsf{P}})$. The process $(\tilde{X}_n)_{n\in\mathbb{Z}_+}$ defined by $\tilde{X}_n(x) = \pi_n x = x_n$, $n \in \mathbb{Z}_+$, is a stationary process under $\tilde{\mathsf{P}}$, because its distribution is the same as that of the original process X.

Let us prove that $\tilde{\mathsf{P}}$ is invariant under the shift map $\theta : \tilde{\Omega} \to \tilde{\Omega}$ defined by $\theta(x_0, x_1, \dots) = (x_1, x_2, \dots)$. To check this, let us recall that, due to the Kolmogorov–Daniell extension theorem, measures on $(\tilde{\Omega}, \tilde{\mathcal{F}}) = (\mathbb{X}^{\mathbb{Z}_+}, \mathcal{X}^{\mathbb{Z}_+})$ are uniquely determined by their values on cylinders

$$C_n(A_0, \dots, A_{n-1}) = \{x \in \mathbb{X}^{\mathbb{Z}_+} : \ x_0 \in A_0, \dots, x_{n-1} \in A_{n-1}\},$$

where $n \in \mathbb{N}$, $A_0, \dots, A_{n-1} \in \mathcal{X}$. So our claim follows from

$$\begin{aligned}
\tilde{\mathsf{P}}(\theta^{-1}C_n(A_0, \dots, A_{n-1})) &= \tilde{\mathsf{P}}(C_{n+1}(\mathbb{X}, A_0, \dots, A_{n-1})) \\
&= \mathsf{P}\{X_1 \in A_0, \dots, X_n \in A_{n-1}\} \\
&= \mathsf{P}\{X_0 \in A_0, \dots, X_{n-1} \in A_{n-1}\} \\
&= \tilde{\mathsf{P}}(C_n(A_0, \dots, A_{n-1})).
\end{aligned}$$

Now we can take the random variable $\pi_0 : \tilde{\Omega} \to \mathbb{X}$ defined by $\pi_0(x) = x_0$ and notice that the process $\tilde{X}_n$ on $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathsf{P}})$ can be represented as $\tilde{X}_n(x) = \pi_0(\theta^n x)$. This is what we wanted, because this representation is of form (2.12). However, to obtain this representation we had to switch from the original probability space to the canonical space of trajectories. The same procedure applies to processes indexed by $\mathbb{N}$ and $\mathbb{Z}$.

It is often convenient to work with stationary processes directly on their trajectory spaces. For example, when working with i.i.d. $(\mathbb{X}, \mathcal{X})$-valued random variables, it is convenient to work with product measure on $(\mathbb{X}^{\mathbb{Z}_+}, \mathcal{X}^{\mathbb{Z}_+})$.

If $\mathbb{X}$ is finite then $(\mathbb{X}^{\mathbb{Z}_+}, \mathcal{X}^{\mathbb{Z}_+})$ equipped with a product measure and coordinate shift θ is called a *Bernoulli shift*.

6. Ergodicity

In this section we work with a probability space $(\Omega, \mathcal{F}, \mathsf{P})$ equipped with a P-preserving transformation θ. The goal of this section is to define and study the notion of ergodicity of $(\Omega, \mathcal{F}, \mathsf{P}, \theta)$. This notion has to deal with the natural question of whether one can decompose Ω into smaller sets and study the dynamics restricted to those sets separately.

Definition 2.13. A set A is called *(backward) invariant* if $\theta^{-1}A = A$, *forward invariant* or *positive invariant* if $\theta A \subset A$, and *almost invariant* or *invariant* mod 0 if $\mathsf{P}(\theta^{-1}A \triangle A) = 0$.

Lemma 2.3. (a) *If $A \in \mathcal{F}$ is invariant, then A is forward invariant.*
(b) *If $A \in \mathcal{F}$ is forward invariant then it is almost invariant.*
(c) *If $A \in \mathcal{F}$ is almost invariant, then there is an invariant set B such that $\mathsf{P}(A \triangle B) = 0$.*
(d) *If $A \in \mathcal{F}$ is forward invariant, then there is an invariant set B such that $\mathsf{P}(A \triangle B) = 0$.*

Proof: To prove part (a), it suffices to notice that invariance of A means that $\omega \in A$ iff $\theta\omega \in A$.

To prove part (b), we first claim that for any forward invariant set A, $A \subset \theta^{-1}A$. This is implied by

$$A \subset \theta^{-1}\theta A \subset \theta^{-1}A,$$

where the first inclusion holds for any set A, and the second one follows from $\theta A \subset A$. Since P is preserved by θ, we have $\mathsf{P}(A) = \mathsf{P}(\theta^{-1}A)$, so $\mathsf{P}(\theta^{-1}A \triangle A) = \mathsf{P}(\theta^{-1}A \setminus A) = 0$.

Part (d) follows from (b) and (c), so it remains to prove (c). Let A be almost invariant. Let

$$B = \liminf_{n\to\infty} \theta^{-n}A = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} \theta^{-k}A.$$

Then $\theta^{-1}B = \liminf_{n\to\infty} \theta^{-n-1}A = B$, i.e., B is invariant. Let us prove $\mathsf{P}(A \triangle B) = 0$. The almost invariance of A and the measure-preserving property of θ imply that for all $n \in \mathbb{N}$,

$$\mathsf{P}(\theta^{-(n+1)}A \triangle \theta^{-n}A) = \mathsf{P}(\theta^{-n}(\theta^{-1}A \triangle A)) = \mathsf{P}(\theta^{-1}A \triangle A) = 0.$$

Therefore,

$$\mathsf{P}(\theta^{-n}A \triangle A) = 0, \quad n \in \mathbb{N}. \tag{2.13}$$

$$A \setminus B = A \cap \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} (\theta^{-k}A)^c \subset A \cap \bigcup_{k=1}^{\infty} (\theta^{-k}A)^c = \bigcup_{k=1}^{\infty} (A \setminus \theta^{-k}A). \tag{2.14}$$

$$B \setminus A = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} \theta^{-k} A \setminus A \subset \bigcup_{n=1}^{\infty} (\theta^{-n} A \setminus A). \tag{2.15}$$

Now our claim follows from (2.13), (2.14), and (2.15). □

It is easy to check that the families $\mathcal{I}_\theta$ and $\mathcal{I}_\theta^*$ of invariant and almost invariant sets are σ-algebras. However, the forward invariant sets do not form a σ-algebra in general and for this reason they are not convenient to work with in the context of measure theory. This is tightly related to the fact that taking pre-image is a more well-behaved operation than taking images, and thus it is easy to define pushforwards of σ-algebras and measures by taking pre-images, and there is no natural way to define pullbacks of these objects via forward images.

PROBLEM 2.3. Give an example of a metric dynamical system such that the collection of forward-invariant sets does not form a σ-algebra. Hint: complements.

DEFINITION 2.14. A transformation is called *ergodic* if every invariant set has measure 0 or 1.

This definition means that one cannot split our metric dynamical system into two.

LEMMA 2.4. *A transformation θ is ergodic if and only if every almost invariant set has measure* 0 *or* 1.

PROOF: The "if" part is obvious since $\mathcal{I} \subset \mathcal{I}^*$. The converse is a direct consequence of part (c) of Lemma 2.3. □

DEFINITION 2.15. A random variable X is called *invariant* if $X(\omega) = X(\theta\omega)$ for all $\omega \in \Omega$.

DEFINITION 2.16. A random variable X is called *almost invariant* or *invariant* mod 0 if $X(\omega) = X(\theta\omega)$ for P-almost all $\omega \in \Omega$.

The following obvious statement gives one more reason to define invariant and almost invariant sets as in Definition 2.13.

LEMMA 2.5. *A set A is invariant if and only if $\mathbf{1}_A$ is invariant. A set A is almost invariant if and only if $\mathbf{1}_A$ is almost invariant.*

DEFINITION 2.17. We say that a (real-valued) random variable X is a.s.-constant if there is a number $c \in \mathbb{R}$ such that $\mathsf{P}\{X = c\} = 1$.

THEOREM 2.9. *The following three conditions are equivalent if θ is a P-preserving transformation:*

(a) *θ is ergodic;*
(b) *every almost invariant random variable is a.s.-constant;*
(c) *every invariant random variable is a.s.-constant;*

PROOF: Condition (b) trivially implies (c), and the latter implies (a) as a specific case for indicator random variables. It remains to derive (b) from (a). So we suppose that X is a random variable almost invariant under an ergodic transformation θ. For every real number t, the set $A(t) = \{X \le t\}$ is almost invariant. Ergodicity of θ and Lemma 2.4 imply $\mathsf{P}\{X \le t\} = 0$ or 1 for any $t \in \mathbb{R}$. Notice that the function $F : \mathbb{R} \to \mathbb{R}$ defined by $F(t) = \mathsf{P}\{X \le t\}$ is the distribution function of X. Thus, $\lim_{t\to-\infty} F(t) = 0$, $\lim_{t\to+\infty} F(t) = 1$, F is nondecreasing and left-continuous. Therefore, there is some point $c \in \mathbb{R}$ such that $F(t) = 0$ for all $t \le c$ and $F(t) = 1$ for all $t > c$. Then $\mathsf{P}\{X = c\} = 1$, and the lemma is proved. □

REMARK 2.2. It is sufficient to verify conditions (b) and (c) of Theorem 2.9, for bounded random variables. In fact, for a general random variable X these conditions can be verified by checking them first for its truncations $X_N = X\mathbf{1}_{|X|\le N}$ and then letting $N \to \infty$.

REMARK 2.3. Conditions (b) and (c) of Theorem 2.9 remain necessary and sufficient conditions for ergodicity if stated for complex-valued random variables instead of real-valued ones. To see that, consider the real and imaginary parts of complex valued random variables.

Let us now check ergodicity for some systems.

LEMMA 2.6. *Let $(\Omega, \mathcal{F}, \mathsf{P})$ be the unit circle $\mathbf{S}^1$ equipped with Borel σ-algebra and Lebesgue measure. If α is irrational, then the circle rotation θ by angle α defined in* (2.1) *is ergodic.*

PROOF: Let us take an almost invariant bounded random variable f and prove that it is a.s.-constant. Every bounded measurable function on $\mathbf{S}$ can be decomposed into a Fourier series convergent in $L^2(\mathbf{S}^1)$:

$$f(\omega) = \sum_{k\in\mathbb{Z}} c_k e^{2\pi i k\omega}, \tag{2.16}$$

where

$$c_k = \int_{\mathbf{S}^1} e^{-2\pi i k\omega} f(\omega) d\omega, \quad k \in \mathbb{Z}.$$

Since f is almost invariant, $f(\omega)$ and $f(\theta\omega)$ coincide almost everywhere. So the Fourier series for $f(\theta\omega)$ coincides with (2.16). Therefore, we can compute c_k as Fourier coefficients for $f(\theta\omega)$:

$$c_k = \int_{\mathbf{S}^1} e^{-2\pi i k\omega} f(\omega+\alpha) d\omega = e^{2\pi i\alpha} \int_{\mathbf{S}^1} e^{-2\pi i k\omega} f(\omega) d\omega = e^{2\pi i\alpha k} c_k, \quad k \in \mathbb{Z}.$$

If α is irrational, then $e^{2\pi i\alpha k} \ne 1$ for all $k \ne 0$, so c_k may be nonzero only for $k = 0$, which means that f is a.s.-constant. □

If $\alpha = m/n$ is rational, then θ is not ergodic since $f(\omega) = e^{2\pi i n\omega}$ is a nonconstant invariant function.

THEOREM 2.10. *Let θ be a shift on the torus $\mathbb{T}^d$ defined by (2.6). It is ergodic if and only if the identity $r_1\alpha_1 + \ldots + r_d\alpha_d = m$ is impossible for any $r_1, \ldots, r_d, m \in \mathbb{Z}$ such that not all of them are equal to 0.*

PROBLEM 2.4. Prove Theorem 2.10.

THEOREM 2.11 (A variation on Kolmogorov's 0-1 Law, REFERENCE????). *Let $(\mathbb{X}, \mathcal{X}, P)$ be a probability space and $(\Omega, \mathcal{F}, \mathsf{P}) = (\mathbb{X}^{\mathbb{N}}, \mathcal{X}^{\mathbb{N}}(\mathbb{X}), P^{\mathbb{N}})$, where $\mathcal{X}^{\mathbb{N}}$ is the cylindric σ-algebra and $P^{\mathbb{N}}$ is the product measure on it with marginal distribution P. Then the shift transformation θ defined by $\theta(\omega_1, \omega_2, \ldots) = (\omega_2, \omega_3, \ldots)$ is measure preserving and ergodic.*

PROOF: The invariance of P under θ follows from the discussion right after Lemma 2.2.

To prove ergodicity, consider an almost invariant set A. It is sufficient to prove that $\mathsf{P}(A) = \mathsf{P}^2(A)$ which can be rewritten as $\mathsf{P}(A \cap A) = \mathsf{P}(A)\mathsf{P}(A)$ or, using the almost invariance of A, as $\mathsf{P}(A \cap \theta^{-n}A) = \mathsf{P}(A)\mathsf{P}(\theta^{-n}A)$ for all $n \in \mathbb{N}$. This means that A is independent of itself or its pre-images under θ^n. To prove this, we will approximate A and $\theta^{-n}A$ (for large n) by cylinders that depend on disjoint coordinate sets and thus are independent.

Since $A \in \mathcal{X}^{\mathbb{N}}$, it can be approximated by cylinders, i.e., for any $\varepsilon > 0$ there is a set $A_\varepsilon \in \mathcal{X}^{\mathbb{N}}$, a number $k(\varepsilon) \in \mathbb{N}$ and a set $B_\varepsilon \in \mathcal{X}^{k(\varepsilon)}$ such that $A_\varepsilon = \{\omega : \ (\omega_1, \ldots, \omega_{k(\varepsilon)}) \in B_\varepsilon\}$, and $\mathsf{P}(A_\varepsilon \triangle A) < \varepsilon$.

Let us take any $n \in \mathbb{N}$ satisfying $n \geq k(\varepsilon)$. Sets A_ε and $\theta^{-n}A_\varepsilon = \{\omega : (\omega_{n+1}, \ldots, \omega_{n+k(\varepsilon)}) \in B_\varepsilon\}$ are independent by the definition of the product measure, so

$$\mathsf{P}(A_\varepsilon \cap \theta^{-n}A_\varepsilon) = \mathsf{P}(A_\varepsilon)\mathsf{P}(\theta^{-n}A_\varepsilon) = \mathsf{P}^2(A_\varepsilon). \tag{2.17}$$

To deal with the left-hand side, we notice that A_ε is very close to an almost invariant set:

$$\begin{aligned}
|\mathsf{P}(A_\varepsilon \cap \theta^{-n}A_\varepsilon) - \mathsf{P}(A)| =& |\mathsf{P}(A_\varepsilon \cap \theta^{-n}A_\varepsilon) - \mathsf{P}(A \cap \theta^{-n}A)| \\
\leq& |\mathsf{P}(A_\varepsilon \cap \theta^{-n}A_\varepsilon) - \mathsf{P}(A_\varepsilon \cap \theta^{-n}A)| \\
&+ |\mathsf{P}(A_\varepsilon \cap \theta^{-n}A) - \mathsf{P}(A \cap \theta^{-n}A)| \\
\leq& |\mathsf{P}(\theta^{-n}(A_\varepsilon \triangle A)| + |\mathsf{P}(A_\varepsilon \triangle A)| \leq 2\varepsilon.
\end{aligned}$$

For the right-hand side of (2.17), we have

$$|\mathsf{P}^2(A_\varepsilon) - \mathsf{P}^2(A)| = (\mathsf{P}(A_\varepsilon) + \mathsf{P}(A)) \cdot |\mathsf{P}(A_\varepsilon) - \mathsf{P}(A)| \leq 2|\mathsf{P}(A_\varepsilon \triangle A)| \leq 2\varepsilon,$$

so $|\mathsf{P}^2(A) - \mathsf{P}(A)| \leq 4\varepsilon$. Since $\varepsilon > 0$ is arbitrary, we have $\mathsf{P}^2(A) = \mathsf{P}(A)$, so $\mathsf{P}(A) = 0$ or 1. □

CHAPTER 3

Ergodic theorems for measure-preserving transformations

1. Von Neumann's ergodic theorem in L^2.

Based on [**Hal60**, pp.13–17].

Suppose we have a metric dynamical system $(\Omega, \mathcal{F}, \mathsf{P}, \theta)$. We are interested in the convergence of time averages

$$A_n f(\omega) = \frac{1}{n} \sum_{k=0}^{n-1} f(\theta^k \omega) \tag{3.1}$$

as $n \to \infty$ for measurable functions $f : \Omega \to \mathbb{R}$. There is an elegant proof of this convergence in the sense of L^2 due to Riesz.

Let us first introduce an L^2 framework. We recall that the space of all measurable functions $f : \mathcal{F} \to \mathbb{C}$ such that $\mathsf{E}|f|^2 < \infty$ with the usual identification of functions that coincide P-almost everywhere, is a complex Hilbert space denoted by $L^2(\Omega, \mathcal{F}, \mathsf{P})$ or L^2 for brevity, with inner product $\langle f, g \rangle = \mathsf{E} f \bar{g}$ and norm $\|f\| = \langle f, f \rangle^{1/2} = (\mathsf{E}|f|^2)^{1/2}$.

For any measurable f, we will denote by Uf the function defined by $Uf(\omega) = f(\theta\omega)$, $\omega \in \Omega$. The following observation was first made by Koopman in [**Koo31**] in the context of Hamiltonian systems:

LEMMA 3.1. *The operator U is an isometry in L^2, i.e., if $f \in L^2$, then $\|Uf\| = \|f\|$.*

PROOF: For any $A \in \mathcal{F}$, we have $U\mathbf{1}_A = \mathbf{1}_{\theta^{-1}A}$. Therefore, the invariance of P under θ implies

$$\langle U\mathbf{1}_A, U\mathbf{1}_A \rangle = \mathsf{E}|\mathbf{1}_{\theta^{-1}A}|^2 = \mathsf{P}(\theta^{-1}A) = \mathsf{P}(A).$$

Also, for any disjoint measurable sets A and B,

$$\langle U\mathbf{1}_A, U\mathbf{1}_B \rangle = \mathsf{E}\mathbf{1}_{\theta^{-1}A}\mathbf{1}_{\theta^{-1}A} = 0.$$

Now we can use linearity of U to derive the result for any simple function $f = \sum_{j=1}^m c_j \mathbf{1}_{A_j}$, where $m \in \mathbb{N}$, $(A_j)_{j=1}^m$ is a collection of mutually disjoint

measurable sets, and $c_j \in \mathbb{C}$ for $j = 1, \ldots, m$:

$$\|Uf\|^2 = \sum_{i,j=1}^{m} c_i \bar{c}_j \langle U\mathbf{1}_{A_i}, U\mathbf{1}_{A_j}\rangle$$
$$= \sum_{i,j=1}^{m} c_i \bar{c}_j \langle \mathbf{1}_{\theta^{-1}A_i}, \mathbf{1}_{\theta^{-1}A_j}\rangle = \sum_{i=1} |c_i|^2 \mathsf{P}(A_i) = \|f\|^2.$$

For a general f, we can find a sequence of simple functions f_n such that $f_n^2(\omega)$ is increasing to $f^2(\omega)$ as $n \to \infty$ for all $\omega \in \mathbb{X}$. Then $(Uf_n)^2$ also increases to $(Uf)^2$ pointwise, and we can use the above identity along with the monotone convergence theorem to write:

$$\|Uf\|^2 = \lim_{n\to\infty} \|Uf_n\|^2 = \lim_{n\to\infty} \|f_n\|^2 = \|f\|^2.$$

□

PROBLEM 3.1. This problem is a generalization of Lemma 3.1. Let $(\Omega, \mathcal{F}, \mathsf{P}, \theta)$ be a metric dynamical system. Let $p \in [1, \infty]$. Prove that if X is a random variable belonging to $L^p(\Omega, \mathcal{F}, \mathsf{P})$, then $UX \in L^p(\Omega, \mathcal{F}, \mathsf{P})$, and $\|UX\|_{L^p(\Omega,\mathcal{F},\mathsf{P})} = \|X\|_{L^p(\Omega,\mathcal{F},\mathsf{P})}$.

Lemma 3.1 shows that studying the convergence of ergodic averages in L^2 is possible via the asymptotic analysis of operators $\frac{1}{n}\sum_{k=0}^{n-1} U^k$, where U is an isometry of a complex Hilbert space.

The first version of the following L^2 ergodic theorem appeared in [**vN32**].

THEOREM 3.1. *Let U be an isometry of a complex Hilbert space $\mathbb{H}$. Let π_I be the orthogonal projection onto the space $I = \{f \in \mathbb{H} : Uf = f\}$. Then, for every $f \in \mathbb{H}$,*

$$A_n f = \frac{1}{n}\sum_{k=0}^{n-1} U^k f \tag{3.2}$$

converges to $\pi_I f$ as $n \to \infty$.

PROOF: Let us denote $G = \{g - Ug : g \in \mathbb{H}\}$ and prove that I and the closure of G are orthogonal complements of each other. If $\langle f, g - Ug\rangle = 0$ for some f and all g, then $\langle f, g\rangle = \langle f, Ug\rangle$ for all g. In particular, $\langle f, f\rangle = \langle f, Uf\rangle$. Using this property and the isometry property, we obtain

$$\langle Uf - f, Uf - f\rangle = \langle Uf, Uf\rangle - \langle Uf, f\rangle - \langle f, Uf\rangle - \langle f, f\rangle$$
$$= \|f\|^2 - \|f\|^2 - \|f\|^2 + \|f\|^2 = 0,$$

so $Uf = f$. Thus $\bar{G}^\perp \subset I$. Since the isometry property can be rewritten as $\langle Uf, Ug\rangle = \langle f, g\rangle$ for all $f, g \in \mathbb{H}$, we can take any $f \in I$, $g \in \mathbb{H}$ and write

$$\langle f, g - Ug\rangle = \langle f, g\rangle - \langle f, Ug\rangle = \langle f, g\rangle - \langle Uf, Ug\rangle = \langle f, g\rangle - \langle f, g\rangle = 0.$$

Therefore, $I \subset \bar{G}^\perp$, and thus $I = \bar{G}^\perp$. So, now we can decompose any $f \in \mathbb{H}$ as $f = f_I + f_G$, where $f_I \in I$ and $f_G \in \bar{G}$ are orthogonal projections of f

onto I and $\bar{G}$, and it is sufficient to see how the ergodic averaging acts on each of the spaces I and $\bar{G}$.

If $f \in I$, then $A_n f = f$.

If $f \in G$, i.e., if $f = g - Ug$ for some $g \in \mathbb{H}$, then

$$\|A_n f\| = \left\| \frac{1}{n} \sum_{k=0}^{n-1} U^k (g - Ug) \right\| = \left\| \frac{1}{n} (g - U^n g) \right\| \le \frac{2}{n} \|g\|$$

Let now $f \in \bar{G}$. Then for any $\varepsilon > 0$ there is $f_\varepsilon = g_\varepsilon - U g_\varepsilon \in G$ such that $\|f - f_\varepsilon\| \le \varepsilon$. Therefore,

$$\|A_n f\| \le \|A_n (f - f_\varepsilon)\| + \|A_n f_\varepsilon\| \le \varepsilon + 2\|g_\varepsilon\|/n,$$

where we used that A_n is a contraction, i.e., $\|Ah\| \le \|h\|$ for all $h \in \mathbb{H}$. So, $\limsup_{n\to\infty} \|A_n f\| \le \varepsilon$, and since the choice of ε is arbitrary, we conclude that $\limsup_{n\to\infty} \|A_n f\| = 0$, and the proof is complete. □

In the context of measure-preserving transformations, the space I always contains all constant functions.

PROBLEM 3.2. Prove that if there are no other almost invariant functions, i.e., the transformation is ergodic, then the spaces I and $I^\perp$ are, respectively, the space of constants and the space of functions $f \in L^2$ with $\mathsf{E} f = 0$. Derive that in this case, the limit in the theorem is deterministic and equal to $\mathsf{E} f$.

In general, one can introduce the invariant σ-algebra $\mathcal{F}_I$ generated by all functions from I. Then $\pi_I f$ can be understood as the conditional expectation $\mathsf{E}(f|\mathcal{F}_I)$.

The simplest cases where this theorem applies are rotations of Euclidean spaces $\mathbb{R}^2$ and $\mathbb{R}^3$ (or their complexifications).

Let us also briefly describe the spectral approach to this theorem based on von Neumann's original idea. He proved that the group $(U^n)_{n\in\mathbb{Z}}$ generated by a unitary operator U admits the following "spectral" representation:

$$U^n = \int_{[-\pi,\pi)} e^{in\varphi} P(d\varphi), \quad n \in \mathbb{Z},$$

for some projector-valued measure P on $[-\pi, \pi)$. So, splitting the domain of integration into a one point set $\{0\}$ and its complement, we obtain

$$A_n = \int_{[-\pi,\pi)} \frac{1}{n} \sum_{k=0}^{n-1} e^{ik\varphi} P(d\varphi) = P(\{0\}) + \int_{[-\pi,\pi)\setminus\{0\}} \frac{e^{in\varphi} - 1}{n(e^{i\varphi} - 1)} P(d\varphi),$$

The integrand in the right-hand side is bounded by 1 and converges to 0, so A_n converges to $P(\{0\})$ which is exactly the projection onto the space of eigenvectors of U associated to eigenvalue $e^{i\cdot 1\cdot 0} = 1$, i.e., the set I of invariant vectors.

Another instance of this general approach is the law of large numbers for L^2-stationary $\mathbb{C}$-valued processes. If X is such a process then there is a

number a and an orthogonal random measure Z on $[-\pi,\pi)$ such that

$$X_n = a + \int_{[-\pi,\pi)} e^{in\varphi} Z(d\varphi), \quad n \in \mathbb{Z}.$$

Therefore, we can write

$$\frac{1}{n}(X_0 + \ldots + X_{n-1}) = Z(\{0\}) + \int_{[-\pi,\pi)\setminus\{0\}} \frac{e^{in\varphi}-1}{n(e^{i\varphi}-1)} Z(d\varphi), \quad n \in \mathbb{N},$$

and conclude that

$$\frac{1}{n}(X_0 + \ldots + X_{n-1}) \stackrel{L^2}{\to} Z(\{0\}), \quad n \to \infty.$$

Therefore, the limit is nonrandom and equals zero if the spectral measure has no atom at zero, i.e., $\mathsf{E}|Z(\{0\})|^2 = 0$.

2. Birkhoff's pointwise ergodic theorem

The following theorem was first proved by G.Birkhoff in [**Bir31**].

THEOREM 3.2. *Let $(\Omega, \mathcal{F}, \mathsf{P}, \theta)$ be a metric dynamical system, $\mathcal{I}$ the σ-algebra of θ-invariant sets, and $f \in L^1(\Omega, \mathcal{F}, \mathsf{P})$. Then, with probability 1,*

$$A_n f \to \mathsf{E}[f|\mathcal{I}], \quad n \to \infty,$$

where the ergodic averages A_n are defined by (3.1) *or, equivalently, by* (3.2) *and $Uf(\omega) = f(\theta\omega)$.*

If in addition θ is ergodic then with probability 1,

$$A_n f \to \mathsf{E}f, \quad n \to \infty. \tag{3.3}$$

There exist various proofs of the pointwise ergodic theorem. They all do not seem as transparent as the proof of the L^2 version. Here, we give one of the simplest and most conceptual proofs constructed by Y. Katznelson and B. Weiss [**KW82**] and based on the idea of the nonstandard analysis proof of T. Kamae [**Kam82**].

PROOF: Due to linearity of time averages and conditional expectations, it is sufficient to prove the convergence for nonnegative function f. Let us introduce

$$\bar{f}(\omega) = \limsup_{n\to\infty} A_n f(\omega), \quad \underline{f}(\omega) = \liminf_{n\to\infty} A_n f(\omega), \quad \omega \in \Omega.$$

A priori we do not know if $\bar{f}$ and $\underline{f}$ are finite, but they are almost invariant. To see that, we write

$$|A_n f(\omega) - A_n(\theta\omega)| \le \frac{1}{n} f(\omega) + \frac{1}{n} f(\theta^n \omega).$$

The first term on the right-hand side clearly converges to 0 as $n \to \infty$ for all ω. The second term converges to 0 almost surely. The latter follows from

the finiteness of $\mathsf{E}f$: for every $\varepsilon > 0$, we have

$$\sum_{n=0}^{\infty} \mathsf{P}\{f(\theta^n\omega) > \varepsilon n\} < \infty,$$

so the Borel–Cantelli Lemma implies that $f(\theta^n\omega) > \varepsilon n$ at most for finitely many values of n.

We only need to show that

$$\mathsf{E}\bar{f}\mathbf{1}_B \le \mathsf{E}f\mathbf{1}_B \le \mathsf{E}\underline{f}\mathbf{1}_B, \quad B \in \mathcal{I}, \tag{3.4}$$

because then we will have $\mathsf{P}\{\bar{f} = \underline{f}\} = 1$, so $\mathsf{P}\{A_n f \to \bar{f}\} = 1$ and $\mathsf{E}\bar{f}\mathbf{1}_B = \mathsf{E}f\mathbf{1}_B$. Also $\bar{f}$ is almost invariant and thus $\mathcal{I}^*$-measurable.

Let us prove the first inequality in (3.4). To that end we will need to use several truncations. Let us take a large number $M > 0$ and define

$$\bar{f}_M(\omega) = \bar{f}(\omega) \wedge M, \quad \omega \in \Omega.$$

For all ω, $\bar{f}_M(\omega)$ is finite and does not exceed $\bar{f}(\omega)$. Let us fix $\varepsilon > 0$, define

$$n(\omega) = \inf\left\{n : \bar{f}_M(\omega) \le A_n f(\omega) + \varepsilon\right\}, \quad \omega \in \Omega,$$

and notice that $n(\omega) < \infty$ for all ω. Almost-invariance of $\bar{f}$ under θ implies almost-invariance of $\bar{f}_M$, so there is a full-measure set Ω' such that $\bar{f}_M(\theta^j\omega) = \bar{f}_M(\omega)$ for all $\omega \in \Omega'$ and $j \in \mathbb{N}$, and we get

$$\sum_{j=0}^{n(\omega)-1} \bar{f}_M(\theta^j\omega) = n(\omega)\bar{f}_M(\omega) \le \sum_{j=0}^{n(\omega)-1} f(\theta^j\omega) + n(\omega)\varepsilon, \quad \omega \in \Omega', \tag{3.5}$$

where in the inequality we used the definition of $n(\omega)$. Note that (3.5) is a lower bound for time averages of f by time averages of $\bar{f}_M$ (an approximation to $\bar{f}$), and to prove the first inequality in (3.4) we need to convert (3.5) into an integral inequality. If only the upper limit of summation were not random, we would have integrated both sides immediately, but n is random. Nevertheless we shall try to obtain a version of (3.5) with a deterministic number L replacing $n(\omega)$. The idea is to split the summation from 0 to $L-1$ into random intervals such that on each of them we can apply (3.5). To achieve that, it is convenient though to work with an auxiliary version of (3.5) where random variable n is replaced by its bounded truncation.

Let us find $N > 0$ such that $\mathsf{P}(C) > 1 - \varepsilon/M$, where $C = \{n(\omega) < N\}$, and introduce $f_{M,\varepsilon} = f\mathbf{1}_C + (f \vee M)\mathbf{1}_{C^c}$ and $\tilde{n} = n\mathbf{1}_C + \mathbf{1}_{C^c}$. Then

$$\sum_{j=0}^{\tilde{n}(\omega)-1} \bar{f}_M(\theta^j\omega) = \tilde{n}(\omega)\bar{f}_M(\omega) \le \sum_{j=0}^{\tilde{n}(\omega)-1} f_{M,\varepsilon}(\theta^j\omega) + \tilde{n}(\omega)\varepsilon, \quad \omega \in \Omega', \tag{3.6}$$

because on C this coincides with (3.5), and on C^c we have $\bar{f}_M \le M \le f_{M,\varepsilon}$.

Let us now choose $L > 0$ large enough to guarantee $NM/L < \varepsilon$ and for all $\omega \in \Omega$ define $n_0(\omega) = 0$ and, inductively,

$$n_k(\omega) = n_{k-1}(\omega) + \tilde{n}(\theta^{n_{k-1}(\omega)}\omega), \quad k \in \mathbb{N}.$$

The purpose of this is to split the sum from 0 to $L-1$ into intervals such that on each of them we can apply (3.6): we define $k(\omega) = \max\{k : n_k(\omega) \le L-1\}$ and write

$$\sum_{j=0}^{L-1} \bar{f}_M(\theta^j\omega) = \sum_{k=1}^{k(\omega)} \sum_{j=n_{k-1}(\omega)}^{n_k(\omega)-1} \bar{f}_M(\theta^j\omega) + \sum_{j=n_{k(\omega)}(\omega)}^{L-1} \bar{f}_M(\theta^j\omega).$$

Applying (3.6) to each of the $k(\omega)$ terms and estimating the last special term by NM (here we use the fact that $\tilde{n}(\omega)$ is uniformly bounded by N, and so is $L - n_{k(\omega)}(\omega)$), we obtain

$$\sum_{j=0}^{L-1} \bar{f}_M(\theta^j\omega) \le \sum_{j=0}^{L-1} f_{M,\varepsilon}(\theta^j\omega) + L\varepsilon + NM.$$

Integrating both sides over B and dividing by L gives

$$\mathsf{E}\bar{f}_M\mathbf{1}_B \le \frac{1}{L}\sum_{j=0}^{L-1} \mathsf{E}f_{M,\varepsilon}(\theta^j\omega)\mathbf{1}_B + \varepsilon + \frac{NM}{L} \le \mathsf{E}f_{M,\varepsilon}\mathbf{1}_B + 2\varepsilon, \tag{3.7}$$

where we used invariance of B and measure-preserving property of θ:

$$\mathsf{E}f_{M,\varepsilon}(\theta^j\omega)\mathbf{1}_{\{\omega\in B\}} = \mathsf{E}f_{M,\varepsilon}(\omega)\mathbf{1}_{\{\omega\in\theta^{-j}B\}} = \mathsf{E}f_{M,\varepsilon}(\omega)\mathbf{1}_{\{\omega\in B\}}.$$

Since $f_{M,\varepsilon} \le f + M\mathbf{1}_{C^c}$, we can estimate the right-hand side of(3.7) with

$$\mathsf{E}f_{M,\varepsilon}\mathbf{1}_B \le \mathsf{E}f\mathbf{1}_B + M\mathsf{P}(C^c) \le \mathsf{E}f\mathbf{1}_B + \varepsilon,$$

we obtain

$$\mathsf{E}\bar{f}_M\mathbf{1}_B \le \mathsf{E}f\mathbf{1}_B + 3\varepsilon. \tag{3.8}$$

Letting $\varepsilon \to 0$ and then $M \to \infty$, we obtain the first inequality in (3.4).

To prove the second inequality in (3.4), we do not even need to introduce a a truncation analogous to $\bar{f}_M$. We simply fix $\varepsilon > 0$, define

$$n(\omega) = \inf\left\{n : \underline{f}(\omega) \ge A_n f(\omega) - \varepsilon\right\}, \quad \omega \in \Omega,$$

introduce $C = \{n(\omega) > N\}$ where N is chosen so that $\mathsf{E}f\mathbf{1}_{C^c} < \varepsilon$, define $f_\varepsilon = f\mathbf{1}_C$ and $\tilde{n} = n\mathbf{1}_C + \mathbf{1}_{C^c}$, and proceed in a similar way to the proof of the first inequality in (3.4).

To prove (3.3), it is sufficient to notice that in the ergodic case, $\mathcal{I}^*$ contains only sets of probability 0 and 1, so $\mathsf{E}[f|\mathcal{I}^*] \overset{\text{a.s.}}{=} \mathsf{E}f$.

□

Problem 3.3. Check the details of the proof of the second inequality in (3.4).

Problem 3.4. Let $(\Omega, \mathcal{F}, \mathsf{P})$ be a probability space and suppose that σ-algebras $\mathcal{G}_0, \mathcal{G}_1 \subset \mathcal{F}$ satisfy the following condition: for every $i \in \{0,1\}$, every set $A \in \mathcal{G}_i$ there is a set $B \in \mathcal{G}_{1-i}$ such that $\mathsf{P}(A \triangle B) = 0$. Then for every random variable $X \in L^1(\Omega, \mathcal{F}, \mathsf{P})$, $\mathsf{E}[X|\mathcal{G}_1] \overset{\text{a.s.}}{=} \mathsf{E}[X|\mathcal{G}_2]$. In particular,

if $\mathcal{G}$ is a σ-subalgebra of $\mathcal{F}$ such that for each $A \in \mathcal{G}$, $\mathsf{P}(A) = 0$ or 1, then $\mathsf{E}[X|\mathcal{G}] \stackrel{\text{a.s.}}{=} \mathsf{E}X$ for any random variable $X \in L^1(\Omega, \mathcal{F}, \mathsf{P})$.

THEOREM 3.3 (L^1 ergodic theorem). *Under the conditions of theorem 3.2,*

$$\lim_{n\to\infty} \mathsf{E}|A_n f - \mathsf{E}[f|\mathcal{I}]| = 0.$$

PROOF: For $\varepsilon > 0$ we can find a bounded function f_ε such that $\mathsf{E}|f_\varepsilon - f| < \varepsilon$. The pointwise ergodic theorem applies to f_ε and since the sequence $A_n f_\varepsilon$ is uniformly bounded, we obtain by bounded convergence theorem that

$$\lim_{n\to\infty} \mathsf{E}|A_n f_\varepsilon - \mathsf{E}[f_\varepsilon|\mathcal{I}]| = 0,$$

i.e.,

$$\mathsf{E}|A_n f_\varepsilon - \mathsf{E}[f_\varepsilon|\mathcal{I}]| < \varepsilon$$

for sufficiently large n. On the other hand, due to the measure preserving property,

$$\mathsf{E}|A_n f_\varepsilon - A_n f| \le \frac{1}{n}\sum_{k=0}^{n-1} \mathsf{E}|f_\varepsilon(\theta^k\omega) - f(\theta^k\omega)| \le \frac{1}{n}\sum_{k=0}^{n-1} \mathsf{E}|f_\varepsilon(\omega) - f(\omega)| < \varepsilon,$$

and due to the Jensen inequality for conditional expectations,

$$\mathsf{E}\,|\mathsf{E}[f_\varepsilon|\mathcal{I}] - \mathsf{E}[f|\mathcal{I}]| \le \mathsf{E}\mathsf{E}[|f_\varepsilon - f||\mathcal{I}] = \mathsf{E}|f_\varepsilon - f| < \varepsilon.$$

Since ε is arbitrary, the theorem follows from the last three inequalities. □

We can now look at the previous examples from the point of view of the ergodic theorem.

THEOREM 3.4 (Kolmogorov's strong law of large numbers, reference????). *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent identically distributed random variables such that X_1 has a finite expectation a. Then*

$$\frac{1}{n}(X_1 + \ldots + X_n) \stackrel{\text{a.s.}}{\to} a, \quad n \to \infty.$$

PROOF: This theorem is a direct consequence of Birkhoff's ergodic theorem and ergodicity of the standard shift i.i.d. sequences. □

Thus the pointwise ergodic theorem can be viewed as a generalization of the strong law of large numbers.

In fact, the ergodic theorem provides the following description of an invariant measure in the ergodic case: the measure of a set A equals the average occupation time for A, i.e., the average fraction of time spent by the trajectory in A:

$$\lim_{n\to\infty} \frac{\sum_{j=0}^{n-1} \mathbf{1}_{\theta^n\omega\in A}}{n} = \mathsf{P}(A). \tag{3.9}$$

Theorems 2.1 and 2.2 can also be understood as a corollaries of ergodicity of circle rotations (Theorem 2.10) and the pointwise ergodic theorem. However, the latter provides convergence only for almost every ω whereas

the statements of those results, in fact, hold for all $\omega \in \mathbf{S}^1$ without exception. This gap can be closed easily.

PROBLEM 3.5. Follow these lines to construct a complete proof of Theorems 2.1 and 2.2.

One can use the ergodic theorem to give equivalent definition of ergodicity:

THEOREM 3.5. *Let θ be a measure preserving tranformation on $(\Omega, \mathcal{F}, \mathsf{P})$. The following statements are equivalent:*

(a) *θ is ergodic;*
(b) *For any $f \in L^1(\Omega, \mathcal{F}, \mathsf{P})$,* (3.3) *holds;*
(c) *For any $A \in \mathcal{F}$,* (3.9) *holds.*

PROOF: Condition (a) implies condition (b) by the ergodic theorem, (c) is a specific case of (b). To derive (a) from (c), we take any invariant set A, use the invariance of A and of A^c to write $A_n \mathbf{1}_A = \mathbf{1}_A$, and since $A_n \mathbf{1}_A$ converges to $\mathsf{P}(A)$ a.s., we conclude that $\mathbf{1}_A \overset{\text{a.s.}}{=} \mathsf{P}(A)$, and this can happen only if $\mathsf{P}(A) = 0$ or 1. □

3. Kingman's subadditive ergodic theorem

THEOREM 3.6 ([**Kin73**]). *Let θ be a measure-preserving and ergodic transformation of a probability space $(\Omega, \mathcal{F}, \mathsf{P})$. Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of L^1 random variables that satisfy*

$$X_{n+m}(\omega) \le X_n(\omega) + X_m(\theta^n\omega), \quad n, m, \in \mathbb{N}. \tag{3.10}$$

Let $a = \inf_{n\in\mathbb{N}}(\mathsf{E}X_n/n)$. Then

$$\frac{X_n}{n} \xrightarrow{\text{a.s.}} a, \quad n \to \infty. \tag{3.11}$$

REMARK 3.1. This is an extension of Birkhoff's ergodic theorem. Other generalizations are possible. We prove only ergodic case.

Examples: Product of independent matrices. Optimal paths. Partition functions in random environment.

PROOF: Our proof follows closely notes [**Lal10**]. Applying property (3.10) inductively, we obtain that for any $m \in \mathbb{N}$ and any finite sequence $(n_k)_{k=1}^m$ of natural numbers,

$$X_{n_1+\ldots+n_m}(\omega) \le X_{n_1}(\omega) + X_{n_2}(\theta^{n_1}\omega) + \ldots + X_{n_m}(\theta^{n_1+\ldots+n_{m-1}}\omega). \tag{3.12}$$

Defining $X_0 \equiv 0$, we see that this inequality remains true also if $(n_k)_{k=1}^m$ is a sequence of nonnegative integers.

Let us now observe that it is sufficient to consider nonpositive random variables. In fact, for a general sequence (X_n) satisfying the conditions of the theorem, the sequence of random variables $(Y_n)_{n\in\mathbb{N}}$ defined by

$$Y_n(\omega) = X_n(\omega) - \sum_{k=0}^{n-1} X_1(\theta^k\omega),$$

is also subadditive in the sense of (3.10), and all these variables are a.s.-nonpositive due to (3.12). If the theorem holds true for $Y_n(\omega)$ then it also holds for X_n since we can apply Birkhoff's ergodic theorem to the second term in the definition of Y_n. So from now on we may assume that $X_n \le 0$.

There are two cases: $a > -\infty$ and $a = -\infty$. If $a > -\infty$, we need to establish the following two inequalities:

$$\liminf_{n\to\infty} \frac{X_n(\omega)}{n} \overset{\text{a.s.}}{\ge} a, \tag{3.13}$$

$$\limsup_{n\to\infty} \frac{X_n(\omega)}{n} \overset{\text{a.s.}}{\le} a. \tag{3.14}$$

If $a = -\infty$, it is sufficient to prove only (3.14) .

Let us start with (3.13). We denote its left-hand side by $b(\omega) \in [-\infty, 0]$ and claim that it is a.s.-constant. In fact, since (3.10) implies $X_{n+1}(\omega) \le X_1(\omega)+X_n(\theta\omega)$, by dividing both sides by n and taking the lim inf we obtain $b(\omega) \le b(\theta\omega)$ for all ω.

PROBLEM 3.6. *Use the invariance of P under θ to prove that $b(\omega) \overset{\text{a.s.}}{=} b(\theta\omega)$.*

So, b is θ-invariant and, due to ergodicity, has to be a.s.-constant with respect to P.

Constants a and b are both nonpositive. Let us prove that if $a > -\infty$, then $b \geq a$. Suppose that $b < a - \varepsilon$ for some $\varepsilon > 0$. If

$$B_m = \left\{\omega : \min_{1\leq n\leq m} X_n(\omega)/n \leq a - \varepsilon\right\}, \quad m \in \mathbb{N},$$

then, for any $\delta > 0$, we can find $m \in \mathbb{N}$ such that $\mathsf{P}(B_m) > 1 - \delta$. Let us fix ω and take an arbitrary $n \in \mathbb{N}$. Let $R = \{j \in \{1, \ldots, nm\} : \theta^j\omega \in B_m\}$. By definition of R, for every $j \in R$ there is $k = k(j) \in \{1, \ldots, m\}$ such that $X_k(\theta^j\omega) \leq k(a - \varepsilon)$.

Assuming $R \neq \emptyset$, let us now define two finite sequences $j_1, \ldots, j_r$ and $k_1, \ldots, k_r$ depending on ω. First, we let j_1 be the smallest number in R and $k_1 = k(j_1)$. Then we inductively set j_i to be the smallest number in R exceeding $j_{i-1} + k_{i-1} - 1$ and $k_i = k(j_i)$ until we get to a number $j_r \leq nm$ such that $k_r = k(j_r)$ satisfies $j_r + k_r > \max R$.

We have

$$R \subset \bigcup_{i=1}^{r} \{j_i, \ldots, j_i + k_i - 1\}, \tag{3.15}$$

where the intervals in the union on the right-hand side are mutually disjoint.

Applying (3.12) to the sequence

$$j_1, k_1, j_2 - (j_1 + k_1), k_2, j_3 - (j_2 + k_2), \ldots, k_r, nm + m - (j_r + k_r),$$

and throwing out nonpositive terms corresponding to $j_1, j_2-(j_1+k_1), \ldots, nm+m-(j_r+k_r)$, we obtain

$$X_{nm+m}(\omega) \leq X_{k_1}(\theta^{j_1}\omega) + X_{k_2}(\theta^{j_2}\omega) + \ldots + X_{k_r}(\theta^{j_r}\omega).$$

We have $X_{k_i}(\theta^{j_i}\omega) < k_i(a - \varepsilon)$ by definition of j_i, k_i, $i = 1, \ldots, r$, so

$$X_{nm+m}(\omega) \leq (k_1 + \ldots + k_r)(a - \varepsilon) \leq (a - \varepsilon)\sum_{i=1}^{nm} \mathbf{1}_{B_m}(\theta^i\omega),$$

where the last inequality follows from (3.15) and $a - \varepsilon < 0$. Note that

$$X_{nm+m}(\omega) \leq (a - \varepsilon)\sum_{i=1}^{nm} \mathbf{1}_{B_m}(\theta^i\omega), \tag{3.16}$$

also trivially holds true if $R = \emptyset$. Let us now divide both sides of (3.16) by $nm + m$ and take expectations:

$$\mathsf{E}\frac{X_{nm+m}}{nm+m} \leq (a - \varepsilon)\frac{nm}{nm+m}\mathsf{P}(B_m).$$

Recalling that $\mathsf{P}(B_m) > 1 - \delta$, letting $n \to \infty$ and using the definition of a, we see that

$$a \leq (a - \varepsilon)(1 - \delta),$$

and we obtain a contradiction if we choose δ sufficiently small.

Let us prove (3.14). If all the maps θ^m, $m \in \mathbb{N}$, are ergodic, then we can use (3.12) to write

$$X_{nm+k}(\omega) \le \sum_{j=0}^{n-1} X_m(\theta^{jm}\omega) + X_k(\theta^{nm}\omega)$$

for $n, m \in \mathbb{N}$ and $k \in \{1, \dots, m\}$, divide both sides by nm, use $X_k \le 0$, let $n \to \infty$, and apply Birkhoff's ergodic theorem to the right-hand side to see that

$$\limsup_{n\to\infty} \frac{X_n(\omega)}{n} \le \frac{\mathsf{E}X_m}{m}, \tag{3.17}$$

and (3.14) follows since m is arbitrary. However, θ^m does not have to be ergodic.

Problem 3.7. Give an example of an ergodic transformation θ such that θ^2 is not ergodic.

So we need to introduce additional averaging to tackle this difficulty. We use (3.12) to write the following m inequalities:

$$\begin{aligned}
X_{nm+k}(\omega) &\le \sum_{j=0}^{n-2} X_m(\theta^{jm}\omega) + X_{m+k}(\theta^{(n-1)m}\omega) \\
X_{nm+k}(\omega) &\le X_1(\omega) + \sum_{j=0}^{n-2} X_m(\theta^{jm+1}\omega) + X_{m+k-1}(\theta^{(n-1)m+1}\omega), \\
X_{nm+k}(\omega) &\le X_2(\omega) + \sum_{j=0}^{n-2} X_m(\theta^{jm+2}\omega) + X_{m+k-2}(\theta^{(n-1)m+2}\omega), \\
&\dots \\
X_{nm+k}(\omega) &\le X_{m-1}(\omega) + \sum_{j=0}^{n-2} X_m(\theta^{jm+m-1}\omega) + X_{k+1}(\theta^{(n-1)m+k-1}\omega).
\end{aligned}$$

Let us take the average of these inequalities and use the nonpositivity of all the random variables involved:

$$X_{nm+k}(\omega) \le \frac{1}{m} \sum_{j=0}^{(n-1)m-1} X_m(\theta^j\omega),$$

Birkhoff's ergodic theorem now applies to the sum on the right-hand side, so dividing both sides by nm and taking $n \to \infty$, we obtain (3.17), which completes the proof. □

CHAPTER 4

Invariant measures

1. Existence of invariant measures

In the last several sections we derived basic dynamical properties of measure-preserving transformations. In particular, we obtained that in the basic framework we have worked with, the statistical properties of dynamics expressed in the form of time averages over long trajectories are described in terms of the invariant measure and the σ-algebra of almost invariant sets.

In this chapter we change the point of view and notice that a typical situation that often leads to interesting and hard problems, is that *a priori* we are given only a measurable space $(\Omega, \mathcal{F})$ and a measurable transformation θ on it. In this case, if we want to study statistical properties of the dynamics with the help of the theory developed above, we first need to find invariant measures, and among them find ergodic ones and only then we may be able to deduce statistical information about trajectories of the system. In general, our conclusions may be different for different initial conditions since there may be several distinct ergodic measures and a function may produce different averages with respect to all these measures.

So we see that it is fundamental to describe the family of all invariant distributions and specifically all ergodic distributions. Notice that here we have made a terminological switch from ergodic transformations to ergodic measures.

First of all we notice that the pushforward of a measure is a linear operator: if μ_1 and μ_2 are finite signed measures (not necessarily probability measures) and a_1, a_2 are two numbers, then $a_1\mu_1 + a_2\mu_2$ is also a finite signed measure, and its pushforward under θ is easily seen to coincide with $a_1\mu_1\theta^{-1} + a_2\mu_2\theta^{-1}$. Therefore, we can speak about a linear operator in the space of finite measures and see straightforwardly that an invariant distribution plays the role of an eigenvector of this operator corresponding to eigenvalue 1. So, we may be interested in studying the structure of the set of all such eigenvectors.

In general, there is a variety of situations that differ from each other in nature.

Even existence of an invariant measure is sometimes hard to establish. However, a compactness argument due to Bogolyubov and Krylov (reference) may often be employed. We give a concrete result here, but we will see developments of this approach throughout these notes.

THEOREM 4.1. *Let Ω be a metric space equipped with Borel σ-algebra $\mathcal{B}(\Omega)$ and a measurable transformation θ on $(\Omega, \mathcal{B}(\Omega))$. Suppose there is a compact set K that is forward invariant under θ, and θ is continuous on K. Then there is an invariant probability measure supported on K.*

PROOF: First let us notice that it is sufficient to consider the case where $K = \Omega$. Let now P be any probability measure on Ω. Let $P_k = P\theta^{-k}$, $k \in \mathbb{Z}_+$ and $\bar{P}_n = \frac{1}{n}\sum_{k=0}^{n-1} P_k$. We would like to extract a weakly convergent subsequence from $(\bar{P}_n)_{n\in\mathbb{Z}_+}$. Let us recall that weak convergence $\mu_n \Rightarrow \mu$ of probability measures $(\mu_n)_{n\in\mathbb{Z}_+}$ to a probability measure μ means that for every bounded continuous function $f : \Omega \to \mathbb{R}$

$$\int_\Omega f(\omega)\mu_n(\omega) \to \int_\Omega f(\omega)\mu(\omega), \quad n \to \infty.$$

The Prokhorov theorem (see [**Bil68**] for a classical introduction to the theory of weak convergence of measures) states that if for any $\varepsilon > 0$ there is a compact set C such that $\mu_n(C) > 1 - \varepsilon$ for all n, then one can extract a weakly convergent subsequence $\mu_{n'}$ from μ_n. In our case, for any $\varepsilon > 0$ we can choose $C = \Omega$, so any sequence of measures on Ω contains a convergent subsequence. So let us choose a sequence $n_i \to \infty$ such that $\bar{P}_{n_i} \Rightarrow \mathsf{P}$ for some probability P, $\quad i \to \infty$. Let us prove that P is θ-invariant. It is sufficient to check

$$\int_\Omega f(\omega)\mathsf{P}(d\omega) = \int_\Omega f(\theta\omega)\mathsf{P}(d\omega),$$

for all continuous bounded functions f since that integrals of functions from that set determines a measure uniquely. For such a function f,

$$\begin{aligned}
&\int_\Omega f(\omega)\mathsf{P}(d\omega) = \lim_{i\to\infty}\int_\Omega f(\omega)\bar{P}_{n_i}(d\omega)\\
&= \lim_{i\to\infty}\frac{1}{n_i}\sum_{k=0}^{n_i-1}\int_\Omega f(\omega)P_k(d\omega)\\
&= \lim_{i\to\infty}\left[\frac{1}{n_i}\sum_{k=1}^{n_i}\int_\Omega f(\omega)P_k(d\omega) + \frac{1}{n_i}\int_\Omega f(\omega)P(d\omega) - \frac{1}{n_i}\int_\Omega f(\omega)P_{n_i}(d\omega)\right]\\
&= \lim_{i\to\infty}\frac{1}{n_i}\sum_{k=0}^{n_i-1}\int_\Omega f(\theta\omega)P_k(d\omega) = \int_\Omega f(\theta\omega)\mathsf{P}(d\omega),
\end{aligned}$$

which completes the proof. □

2. Structure of the set of invariant measures.

THEOREM 4.2. *Let θ be a measurable transformation on $(\Omega, \mathcal{F})$. Suppose P_1 and P_2 are two distinct θ-invariant and ergodic measures on $(\Omega, \mathcal{F})$. Then they are singular to each other.*

PROOF: Let us take a bounded random variable X such that $\mathsf{E}_{\mathsf{P}_1} X \neq \mathsf{E}_{\mathsf{P}_2} X$. By Birkhoff's ergodic theorem, there are sets $A_1, A_2 \in \mathcal{F}$ such that $\mathsf{P}_1(A_1) = \mathsf{P}_2(A_2) = 1$, and that if $i \in \{1, 2\}$, then

$$\frac{1}{n} \sum_{k=0}^{n-1} X(\theta^k \omega) \to \mathsf{E}_{\mathsf{P}_i} X, \quad \omega \in A_i.$$

Therefore, $A_1 \cap A_2 = 0$, and P_1 and P_2 are mutually singular. □

For a measurable space with a measurable transformation $(\Omega, \mathcal{F}, \theta)$, the set $I(\Omega, \mathcal{F}, \theta)$ of invariant probability measures is a convex set, i.e., for every $\mathsf{Q}_1, \mathsf{Q}_2 \in I(\Omega, \mathcal{F}, \theta)$ and every $\alpha \in [0, 1]$, the measure $\mathsf{P} = \alpha \mathsf{Q}_1 + (1 - \alpha)\mathsf{Q}_2$ is also θ-invariant, as a straightforward computation shows. Let us study the extreme points of $I(\Omega, \mathcal{F}, \theta)$. A point $\mu \in I(\Omega, \mathcal{F}, \theta)$ is called an extreme point if conditions $\mathsf{Q}_1, \mathsf{Q}_2 \in I(\Omega, \mathcal{F}, \theta)$, $\alpha \in (0, 1)$, and

$$\mathsf{P} = \alpha \mathsf{Q}_1 + (1 - \alpha)\mathsf{Q}_2, \tag{4.1}$$

imply $\mathsf{Q}_1 = \mathsf{Q}_2 = \mathsf{P}$.

THEOREM 4.3. *A measure P is an extreme point of $I(\Omega, \mathcal{F}, \theta)$ iff $(\Omega, \mathcal{F}, \mathsf{P}, \theta)$ is ergodic.*

PROOF: Let $(\Omega, \mathcal{F}, \mathsf{P}, \theta)$ be not ergodic. Then there is an invariant set B with $\mathsf{P}(B) \in (0, 1)$. Then B^c is also invariant. Let us define measures Q_1 and Q_2 by $\mathsf{Q}_1(A) = \mathsf{P}(A|B)$ and $\mathsf{Q}_2(A) = \mathsf{P}(A|B^c)$, and set $\alpha = \mathsf{P}(B)$. Then (4.1) is satisfied, and both measures $\mathsf{Q}_1, \mathsf{Q}_2$ are invariant: for any set A,

$$\begin{aligned} \mathsf{Q}_1(\theta^{-1}A) = \mathsf{P}(\theta^{-1}A|B) = \frac{\mathsf{P}(\theta^{-1}A \cap B)}{\mathsf{P}(B)} &= \frac{\mathsf{P}(\theta^{-1}A \cap \theta^{-1}B)}{\mathsf{P}(B)} \\ &= \frac{\mathsf{P}(\theta^{-1}(A \cap B))}{\mathsf{P}(B)} = \frac{\mathsf{P}(A \cap B)}{\mathsf{P}(B)} = \mathsf{Q}_1(A), \end{aligned}$$

and a similar computation applies to Q_2. Also, $\mathsf{Q}_1 \neq \mathsf{P}$ since $\mathsf{Q}_1(B) = 1 \neq \mathsf{P}(B)$. Therefore, P is not an extreme point.

Next, let P be ergodic and let decomposition (4.1) hold with $\alpha \in (0, 1)$ and invariant measures $\mathsf{Q}_1, \mathsf{Q}_2$. Let us take any bounded measurable function f. Ergodic theorem implies that for a set A such that $\mathsf{P}(A) = 1$ and

$$\lim_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} f(\theta^k \omega) = \int_\Omega f(\omega) \mathsf{P}(d\omega), \quad \omega \in A. \tag{4.2}$$

Due to (4.1), measures Q_1 and Q_2 are absolutely continuous with respect to P. Since $\mathsf{P}(A^c) = 0$, we have $\mathsf{Q}_1(A^c) = \mathsf{Q}_2(A^c) = 0$, so we conclude that convergence in (4.2) holds almost surely with respect to Q_1 and Q_2, so by ergodic theorem,

$$\int_\Omega f(\omega) \mathsf{P}(d\omega) = \mathsf{E}_{\mathsf{Q}_1}[f|\mathcal{I}], \quad \mathsf{Q}_1\text{-a.s.}$$

Integrating this identity with respect to Q_1 we obtain

$$\int_\Omega f(\omega)\mathsf{P}(d\omega) = \int_\Omega f(\omega)\mathsf{Q}_1(d\omega),$$

and since f was an arbitrary bounded measurable function, we conclude that $\mathsf{P} = \mathsf{Q}_1$. Similarly, we obtain $\mathsf{P} = \mathsf{Q}_2$, so P is an extreme point. □

In general, every θ-invariant probability measure can be represented as a mixture of ergodic θ-invariant measures:

$$\mathsf{P} = \int_{\mathcal{E}} P\mu(dP), \tag{4.3}$$

where μ is a measure on $\mathcal{E}$, the set of all ergodic measures. In the remainder of this section we will make sense of this statement, but let us first consider the following example. Let θ be a rotation of the cylinder $\mathbf{S}^1 \times \mathbb{R}^1$ by an irrational angle α:

$$\theta(x, y) = (\{x + \alpha\}, y).$$

Since for every $y \in \mathbb{R}$, the circle $\mathbf{S}^1 \times \{y\}$ is invariant, every ergodic measure has to be concentrated on one of the circles. Restricted to any circle, θ act as a circle rotation, so for every $y \in \mathbb{R}$, the measure $\mathrm{Leb} \times \delta_y$ is a unique invariant measure on $\mathbf{S}^1 \times \{y\}$ and it is ergodic under θ. We see that all ergodic measures are indexed by $y \in \mathbb{R}$, and representation (4.3) may2 be interpreted as

$$\mathsf{P}(A) = \int_{y\in\mathbb{R}} \mu(dy)\,\mathrm{Leb}(A_y),$$

where $A_y = \{x \in \mathbf{S}^1 : \ (x, y) \in A\}$ for every $y \in \mathbb{R}$. In other words, an invariant measure distributes the mass over multiple fibers (ergodic components), but within individual fibers, the mass is distributed according to an ergodic measure.

One could use an abstract Choquet theorem that says that every point of a compact convex subset of a locally convex topological vector space can be represented as a mixture (integral convex combination) of its extreme points. We will take another, more general and constructive approach. We follow [**Sar09**] and [**EW11**].

Definition 4.1. *A σ-algebra $\mathcal{G}$ is countably generated if $\mathcal{G} = \sigma(E_k, k \in \mathbb{N})$ for some countable family of sets $E_k \in \mathcal{F}, k \in \mathbb{N}$.*

Definition 4.2. *For two σ-algebras $\mathcal{G}, \mathcal{H}$, we write $\mathcal{G} \overset{\mathsf{P}}{=} \mathcal{H}$ if for every set $A \in \mathcal{G}$ there is a set $B \in \mathcal{H}$ such that $\mathsf{P}(A \triangle B) = 0$, and for every set $A \in \mathcal{H}$ there is a set $B \in \mathcal{G}$ such that $\mathsf{P}(A \triangle B) = 0$.*

Theorem 4.4. *(1) Let P be a probability measure on a Borel space $(\Omega, \mathcal{F})$ and let $\mathcal{G} \subset \mathcal{F}$ be a σ-algebra on Ω. Then there is a set $\Omega' \in \mathcal{G}$ satisfying $\mathsf{P}(\Omega') = 1$ and a probability kernel $\mathsf{P}_{\mathcal{G}}(\cdot, \cdot)$ from*

$(\Omega', \mathcal{G})$ to $(\Omega, \mathcal{F})$ with the following property: if $X \in L^1(\Omega, \mathcal{F}, \mathsf{P})$, then

$$\mathsf{E}(X|\mathcal{G})(\omega) = \int_\Omega X(\sigma)\mathsf{P}_\mathcal{G}(\omega, d\sigma), \quad \omega \in \Omega'. \tag{4.4}$$

(2) Suppose additionally $\mathcal{G}$ is countably generated. Then the set Ω' in the first part of this theorem can be chosen so that for all $\omega \in \Omega'$, $\mathsf{P}_\mathcal{G}(\omega, [\omega]_\mathcal{G}) = 1$, where

$$[\omega]_\mathcal{G} = \bigcap_{A:\omega\in A\in\mathcal{G}} A$$

is the atom of $\mathcal{G}$ containing ω. Moreover, if $\omega_1, \omega_2 \in \Omega'$ and $[\omega_1]_\mathcal{G} = [\omega_2]_\mathcal{G}$, then $\mathsf{P}_\mathcal{G}(\omega_1, \cdot) = \mathsf{P}_\mathcal{G}(\omega_2, \cdot)$.

(3) If $\mathcal{H} \subset \mathcal{F}$ is another σ-algebra such that $\mathcal{G} \overset{\mathsf{P}}{=} \mathcal{H}$, then $\mathsf{P}_\mathcal{G}(\omega, \cdot) = \mathsf{P}_\mathcal{H}(\omega, \cdot)$ for almost all ω.

REMARK 4.1. The kernel P is called *regular conditional probability* with respect to $\mathcal{G}$, see [**Shi96**, Section II.7]

PROOF: Using the definition of Borel spaces, we can assume that Ω is a compact segment of $\mathbb{R}$ and $\mathcal{F}$ is a restriction of Borel σ-algebra onto Ω. Under this assumption we can find a dense countable set in $C(\Omega)$. The vector space R generated by this set over $\mathbb{Q}$ is also a countable set. For every $Y \in R$, we can find a version $G[Y]$ of $\mathsf{E}(Y|\mathcal{G})$. Let us now consider the following countable set of conditions:

(1) $G[pY + qZ](\omega) = pG[Y](\omega) + qG[Z](\omega), \quad p, q \in \mathbb{Q}$, $Y, Z \in R$.
(2) $\min Y \le G[Y](\omega) \le \max Y, \quad Y \in R$.

Each of these conditions is violated on an exceptional set of zero measure belonging to $\mathcal{G}$. Since there are countably many of these conditions, there is a set $A \in \mathcal{G}$ such that $\mathsf{P}(A) = 1$ and all the conditions above are satisfied for all $\omega \in A$. In particular, for each $\omega \in A$, the functional $Y \mapsto G[Y](\omega)$ is a linear functional on R over $\mathbb{Q}$, with norm bounded by 1. So, this functional can be extended by continuity to a continuous linear functional on $C(\Omega)$ over $\mathbb{R}$, in a unique way. Such functionals can be identified with measures by Riesz's theorem (reference????). Thus for each $\omega \in A$, we obtain a measure $\mathsf{P}_\mathcal{G}(\omega, \cdot)$ on $(\Omega, \mathcal{F}, \mathsf{P})$. This measure satisfies

$$\int_\Omega Y(\sigma)\mathsf{P}_\mathcal{G}(\omega, d\sigma) = G[Y](\omega), \quad \omega \in A,\ Y \in C(\Omega). \tag{4.5}$$

In particular the left-hand side of this identity is a $\mathcal{G}$-measurable function for every $Y \in C(\Omega)$.

To see that for every $B \in \mathcal{F}$, $\mathsf{P}_\mathcal{G}(\cdot, B)$ is a $\mathcal{G}$-measurable function it is sufficient to represent indicators $\mathbf{1}_B$ as pointwise limits of continuous uniformly bounded functions Y_n, because then by dominated convergence,

$$\mathsf{P}_\mathcal{G}(\omega, B) = \lim_{n\to\infty} \int_\Omega Y_n(\sigma)\mathsf{P}_\mathcal{G}(\omega, d\sigma),$$

and a limit of $\mathcal{G}$-measurable functions is a $\mathcal{G}$-measurable function.

Let $\mathcal{A}$ be the collection of sets such that their indicators can be represented as pointwise limits of continuous uniformly bounded functions. Then $\mathcal{A}$ is a set algebra and a monotone class (see [**Shi96**, Section II.2] for these notions). Therefore it is a σ-algebra. Since $\mathcal{A}$ contains open sets, it coincides with Borel σ-algebra. Therefore, our claim that for every $B \in \mathcal{F}$, $\mathsf{P}_{\mathcal{G}}(\cdot, B)$ is a $\mathcal{G}$-measurable function hods true.

PROBLEM 4.1. Let $X \in L^1(\Omega, \mathcal{F}, \mathsf{P})$. Prove that there is a sequence $X_n \in C(\Omega), n \in \mathbb{N}$ such that $X \overset{\text{a.s.}}{=} \sum_{n=1}^{\infty} X_n$ and $\sum_{n=1}^{\infty} \|X_n\|_{L^1(\Omega,\mathcal{F},\mathsf{P})} < \infty$.

Using this problem, we can write

$$\mathsf{E}(X|\mathcal{G})(\omega) \overset{\text{a.s.}}{=} \sum_{n=1}^{\infty} \mathsf{E}(X_n|\mathcal{G})(\omega) \overset{\text{a.s.}}{=} \sum_{n=1}^{\infty} \int_{\Omega} X_n(\sigma) \mathsf{P}_{\mathcal{G}}(\omega, d\sigma)$$
$$\overset{\text{a.s.}}{=} \int_{\Omega} \sum_{n=1}^{\infty} X_n(\sigma) \mathsf{P}_{\mathcal{G}}(\omega, d\sigma) \overset{\text{a.s.}}{=} \int X(\sigma) \mathsf{P}_{\mathcal{G}}(\omega, d\sigma).$$

The first identity holds since conditional expectation is a continuous linear operator in L^1. The second identity follows from (4.5) and continuity of X_n. The last identity follows from the definition of X_n, so it remains to explain the change of order of integration and summation in the third identity. This will be justified if we can prove that $\int_{\Omega} \sum_{n=1}^{\infty} |X_n(\sigma)| \mathsf{P}_{\mathcal{G}}(\omega, d\sigma)$ is a.s.-finite. This will follow from

$$\mathsf{E} \int_{\Omega} \sum_{n=1}^{\infty} |X_n(\sigma)| \mathsf{P}_{\mathcal{G}}(\omega, d\sigma) < \infty. \tag{4.6}$$

Identity

$$\mathsf{E} \int_{\Omega} Y(\sigma) \mathsf{P}_{\mathcal{G}}(\omega, d\sigma) = \mathsf{E}Y$$

holds for all $Y \in R$ by definition and thus for all $Y \in C(\Omega)$, so we use it along with the monotone convergence theorem to see that the left-hand side of (4.6) equals

$$\lim_{m\to\infty} \mathsf{E} \int_{\Omega} \sum_{n=1}^{m} |X_n(\sigma)| P(\omega, d\sigma) \le \lim_{m\to\infty} \sum_{n=1}^{m} \mathsf{E}|X_n(\sigma)| \le \sum_{n=1}^{\infty} \|X_n\|_{L^1} < \infty,$$

which completes the proof of part 1 with $\Omega' = A$. We will need to adjust Ω' in the proof of part 2 that follows. First,

$$\mathsf{P}(E_i|\mathcal{G})(\omega) \overset{\text{a.s.}}{=} \mathbf{1}_{E_i}(\omega) \overset{\text{a.s.}}{=} \mathsf{P}_{\mathcal{G}}(\omega, E_i), \quad i \in \mathbb{N},$$

where $\mathsf{P}(E_i|\mathcal{G})(\omega) \overset{\text{a.s.}}{=} \mathbf{1}_{E_i}(\omega)$ follows by definition of conditional expectation, and $\mathsf{P}(E_i|\mathcal{G})(\omega) \overset{\text{a.s.}}{=} \mathsf{P}_{\mathcal{G}}(\omega, E_i)$ is a specific case of (4.4).

Let $N \supset A^c$ be the union of all exceptional sets in this identity for $i \in \mathbb{N}$. Then for all $i \in \mathbb{N}$,

$$\mathsf{P}_{\mathcal{G}}(\omega, E_i) = \begin{cases} 1, & \omega \in E_i \cap N^c, \\ 0, & \omega \in E_i^c \cap N^c, \\ ?, & \omega \in N. \end{cases} \tag{4.7}$$

Under the conditions of part 2

$$[\omega]_{\mathcal{G}} = \bigcap_{i:\omega \in E_i} E_i \cap \bigcap_{i:\omega \notin E_i} E_i^c, \quad \omega \in \Omega. \tag{4.8}$$

Problem 4.2. Prove representation (4.8). Deduce that $[\omega]_{\mathcal{G}}$ belongs to $\mathcal{G}$ and, in fact, can be defined as the smallest element of $\mathcal{G}$ containing ω).

Identities (4.7) and (4.8) now imply $\mathsf{P}_{\mathcal{G}}(\omega, [\omega]_{\mathcal{G}}) = 1$, $\omega \in N^c$.

Since for any $B \in \mathcal{F}$ the map $\omega \mapsto \mathsf{P}_{\mathcal{G}}(\omega, B)$ is $\mathcal{G}$-measurable, $\mathsf{P}_{\mathcal{G}}(\omega_1, B) = \mathsf{P}_{\mathcal{G}}(\omega_2, B)$ for any ω_1, ω_2 satisfying $[\omega_1]_{\mathcal{G}} = [\omega_2]_{\mathcal{G}}$. So part 2 holds with $\Omega' = N^c$.

To prove part 3, we introduce $\mathcal{A} = \sigma(\mathcal{G}, \mathcal{H})$. Then we take a countable dense set $\{X_n\}_{n\in\mathbb{N}}$ in $C(\Omega)$ and notice that for each $n \in \mathbb{N}$, $\mathsf{E}(X_n|\mathcal{G})$ and $\mathsf{E}(X_n|\mathcal{H})$ are versions of $\mathsf{E}(X_n|\mathcal{A})$, so they coincide almost surely. Therefore, there is a set Ω' of full measure such that

$$\int \mathsf{P}_{\mathcal{G}}(\omega, d\sigma) X_n(\sigma) = \int \mathsf{P}_{\mathcal{H}}(\omega, d\sigma) X_n(\sigma), \quad n \in \mathbb{N},\ \omega \in \Omega'.$$

Since measures on Ω are uniquely defined by the integrals of X_n with respect to them, we conclude that $\mathsf{P}_{\mathcal{G}}(\omega, \cdot) = \mathsf{P}_{\mathcal{H}}(\omega, \cdot)$ on Ω'. □

Theorem 4.5. *If P is a probability measure on a Borel space $(\Omega, \mathcal{F})$ and $\mathcal{G} \subset \mathcal{F}$ is a σ-algebra, then there is a countably generated σ-algebra $\mathcal{H}$ such that $\mathcal{G} \overset{\mathsf{P}}{=} \mathcal{H}$.*

Proof: Using the Borel isomorphism we can assume that Ω is a compact interval on $\mathbb{R}$. Then $L^1(\Omega, \mathcal{F}, \mathsf{P})$ is a separable space and thus so is its subset $\{\mathbf{1}_A | A \in \mathcal{G}\}$. Therefore, there is a family $(A_n)_{n\in\mathbb{N}}$ of sets in $\mathcal{G}$ such that all sets from $\mathcal{G}$ are approximated by sets A_n with arbitrary precision. Establishing relation

$$\sigma(A_n, n \in \mathbb{N}) \overset{\mathsf{P}}{=} \mathcal{G} \tag{4.9}$$

will complete the proof. □

Problem 4.3. Prove relation (4.9).

Theorem 4.6. *Let P be a probability measure on a Borel space $(\Omega, \mathcal{F})$, invariant under a measurable transformation θ. Let $\mathsf{P}_{\mathcal{I}}(\cdot, \cdot)$ be a regular conditional probability with respect to $\mathcal{I}$. Then for P-almost every ω, the*

measure $\mathsf{P}_{\mathcal{I}}(\omega,\cdot)$ *is* θ*-invariant and ergodic, and* P *is a mixture or convex combination of those ergodic measures:*

$$\mathsf{P}=\int_\Omega \mathsf{P}(d\omega)\mathsf{P}_{\mathcal{I}}(\omega,\cdot),$$

i.e.,

$$\mathsf{P}(A)=\int_\Omega \mathsf{P}(d\omega)\mathsf{P}_{\mathcal{I}}(\omega,A),\quad A\in\mathcal{F}, \tag{4.10}$$

and, more generally,

$$\mathsf{E}X=\int_\Omega \mathsf{P}(d\omega)\int_\Omega \mathsf{P}_{\mathcal{I}}(\omega,d\sigma)X(\sigma),\quad X\in L^1(\Omega,\mathcal{F},\mathsf{P}). \tag{4.11}$$

Proof: Checking (4.10) or (4.11) is straightforward: according to Theorem 4.4,

$$\mathsf{E}X=\mathsf{E}\mathsf{E}(X|\mathcal{I})=\mathsf{E}\int_\Omega X(\sigma)\mathsf{P}_{\mathcal{I}}(\omega,d\sigma),\quad X\in L^1(\Omega,\mathcal{F},\mathsf{P}). \tag{4.12}$$

From now on we use the Borel isomorphism to assume that $(\Omega,\mathcal{F})$ is a compact segment on $\mathbb{R}$ with Borel σ-algebra. We choose a dense set $D=\{X_n\}_{n\in\mathbb{N}}$ in $C(\Omega)$.

Let us prove invariance of $P_{\mathcal{I}}(\omega,\cdot)$. By Birkhoff's ergodic theorem, there is a set $\Omega_1\in\mathcal{F}$ such that $\mathsf{P}(\Omega_1)=1$, $\Omega_1\subset\Omega'$ (where Ω' is introduced in Theorem 4.6), and for all $n\in\mathbb{N}$ and $\omega\in\Omega_1$,

$$\int_\Omega \mathsf{P}_{\mathcal{I}}(\omega,d\sigma)X_n(\sigma)=\mathsf{E}(X_n|\mathcal{I})(\omega)=\lim_{N\to\infty}\frac{1}{N}\sum_{k=0}^{N-1}X_n(\theta^k\omega)$$

and

$$\int_\Omega \mathsf{P}_{\mathcal{I}}(\omega,d\sigma)X_n(\theta\sigma)=\mathsf{E}(X_n\circ\theta|\mathcal{I})(\omega)=\lim_{N\to\infty}\frac{1}{N}\sum_{k=0}^{N-1}X_n(\theta^{k+1}\omega).$$

Since X_n is bounded, the right-hand sides of the last two identities coincide for $\omega\in\Omega_1$. Since every $X\in C(\Omega)$ can be uniformly approximated by functions from D, we obtain that

$$\int_\Omega \mathsf{P}_{\mathcal{I}}(\omega,d\sigma)X(\sigma)=\int_\Omega \mathsf{P}_{\mathcal{I}}(\omega,d\sigma)X(\theta\sigma) \tag{4.13}$$

holds for $X\in C(\Omega)$ and $\omega\in\Omega_1$. For every set $B\in\mathcal{F}$ there is a sequence of uniformly bounded continuous functions convergent pointwise to $\mathbf{1}_B$. Bounded convergence implies that (4.13) holds for $X=\mathbf{1}_B$ which proves that θ preserves $\mathsf{P}_{\mathcal{I}}(\omega,\cdot)$ for $\omega\in\Omega_1$.

To prove the ergodicity part of the theorem, we will need the following statements:

LEMMA 4.1. *Let $(\Omega, \mathcal{F}, \mathsf{P}, \theta)$ be a metric dynamical system, where $(\Omega, \mathcal{F})$ is a compact subset of $\mathbb{R}$ with Borel σ-algebra on it. Then ergodicity is equivalent to the following statement: there is a countable dense set $\{X_n\}_{n\in\mathbb{N}}$ in $C(\Omega)$ such that*

$$\frac{1}{N}\sum_{i=0}^{N-1} X_n(\theta^i\omega) \xrightarrow{\text{a.s.}} \int_\Omega X_n(\omega)\mathsf{P}(d\omega).$$

PROBLEM 4.4. Prove Lemma 4.1.

LEMMA 4.2. *Let P be a probability measure on a Borel space $(\Omega, \mathcal{F})$. Let $\mathcal{H} \subset \mathcal{F}$ be a σ-algebra on Ω and let $\mathsf{P}_{\mathcal{H}}(\cdot,\cdot)$ be the regular conditional probability with respect to $\mathcal{H}$ constructed in Theorem 4.4. Suppose $B \in \mathcal{F}$ satisfies $\mathsf{P}(B) = 1$. Then, for P-almost all $\omega \in \Omega$, $\mathsf{P}_{\mathcal{H}}(\omega, B) = 1$.*

PROBLEM 4.5. Prove Lemma 4.2

We need to prove ergodicity of almost all measures $\mathsf{P}_{\mathcal{I}}(\omega,\cdot)$. It is sufficient to prove ergodicity of almost all measures $\mathsf{P}_{\mathcal{H}}(\omega,\cdot)$, where $\mathcal{H}$ is a countably generated σ-algebra such that $\mathcal{I} \overset{\mathsf{P}}{=} \mathcal{H}$ (the existence of such $\mathcal{H}$ is guaranteed by Theorem 4.5). The advantage of dealing with $\mathcal{H}$ is that part 2 of Theorem 4.4 applied to $\mathcal{H} = \mathcal{G}$ holds on a set Ω' of full measure.

Let us construct a set of full measure P such that for ω from that set, $(\Omega, \mathcal{F}, \mathsf{P}_{\mathcal{H}}(\omega,\cdot), \theta)$ satisfies the ergodicity criterion provided by Lemma 4.1.

Using Birkhoff's ergodic theorem we can find $\Omega_2 \subset \Omega_1$ such that $\mathsf{P}(\Omega_2) = 1$ and for $\omega \in \Omega_2$ and any $n \in \mathbb{N}$

$$\frac{1}{N}\sum_{i=0}^{N-1} X_n(\theta^i\omega) \to \int_\Omega \mathsf{P}_{\mathcal{H}}(\omega, d\sigma)X_n(\sigma), \quad N \to \infty.$$

This, along with Lemma 4.2, implies that there is a set $\Omega_3 \subset \Omega_2$ such that $\mathsf{P}(\Omega_3) = 1$ and for all $\omega \in \Omega_3$, $P_{\mathcal{H}}(\omega, B) = 1$, where

$$B = \left\{\zeta : \frac{1}{N}\sum_{i=0}^{N-1} X_n(\theta^i\zeta) \to \int_\Omega \mathsf{P}_{\mathcal{H}}(\zeta, d\sigma)X_n(\sigma), \quad N \to \infty, \quad n \in \mathbb{N}\right\}.$$

Applying Lemma 4.2 to Ω_3, we obtain that there is $\Omega_4 \subset \Omega_3$ with $\mathsf{P}_{\mathcal{H}}(\omega, \Omega_3) = 1$ for all $\omega \in \Omega_4$.

Since $\Omega_3 \subset \Omega'$, we have $\mathsf{P}_{\mathcal{H}}(\omega, [\omega]_{\mathcal{H}}) = 1$ for $\omega \in \Omega_3$ (see part 2 of Theorem 4.6). So, $\mathsf{P}_{\mathcal{H}}(\omega, [\omega]_{\mathcal{H}} \cap B \cap \Omega_3) = 1$ for $\omega \in \Omega_4$. On the other hand, since $\omega \mapsto \int \mathsf{P}_{\mathcal{H}}(\omega, d\sigma)X_n(\sigma)$ is $\mathcal{H}$-measurable, we have

$$\int_\Omega \mathsf{P}_{\mathcal{H}}(\zeta, d\sigma)X_n(\sigma) = \int_\Omega \mathsf{P}_{\mathcal{H}}(\omega, d\sigma)X_n(\sigma)$$

if $\omega, \zeta \in \Omega'$ and $\zeta \in [\omega]_{\mathcal{H}}$.

So, if $\omega \in \Omega_4$, then for every $\zeta \in [\omega]_{\mathcal{H}} \cap B \cap \Omega_3$ (which is a $\mathsf{P}_{\mathcal{H}}(\omega, \cdot)$-full measure set), we have

$$\frac{1}{N} \sum_{i=0}^{N-1} X_n(\theta^i \zeta) \to \int_\Omega \mathsf{P}_{\mathcal{H}}(\omega, d\sigma) X_n(\sigma), \quad N \to \infty.$$

We see that the criterion of Lemma 4.1 is satisfied, so measures $\mathsf{P}_{\mathcal{H}}(\omega, \cdot)$ are ergodic for $\omega \in \Omega_4$. □

3. Absolutely continuous invariant measures

Suppose now that $\Omega = (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and θ is a differentiable transformation. In some cases it is reasonable to expect that there is an absolutely continuous invariant measure P for θ. Absolute continuity of P with respect to Lebesgue measure means that there is an $L^1(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \mathrm{Leb})$ function ρ called *density* such that for any set $B \in \mathcal{B}(\mathbb{R}^d)$,

$$\mathsf{P}(B) = \int_B f(\omega) d\omega.$$

Densities are not uniquely defined since they can be modified on a zero measure set.

In what follows, for a differentiable function $f : \mathbb{R}^m \to \mathbb{R}^n$ we denote by $Df(x)$ the Jacobi matrix of f at point x: $(\partial_i f^j(x))_{\substack{i=1,\dots,m \\ j=1,\dots,n}}$.

THEOREM 4.7. *Let $\mathsf{P}(dx) = \rho(x)dx$ be a measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Suppose $\theta : \mathbb{R}^d \to \mathbb{R}^d$ is differentiable and nondegenerate for almost all x (the latter means $\mathrm{Leb}\{x : \det D\theta(x) = 0\} = 0$). Then $\mathsf{P}\theta^{-1}$ is also absolutely continuous with respect to Lebesgue measure and the so called transfer operator or Perron–Frobenius operator $\mathcal{L}$ applied to ρ:*

$$\mathcal{L}\rho(x) = \sum_{y \in \theta^{-1}x} \frac{\rho(y)}{|\det D\theta(y)|}, \quad x \in \mathbb{R}^d, \tag{4.14}$$

gives a density of $\mathsf{P}\theta^{-1}$. Here we adopt the usual convention that summation over an empty set is zero.

Since densities that differ only on a set of zero measure define the same measure, it makes sense to consider their equivalence classes and look at the transfer operator as a transformation in L^1 understood as a space of equivalence classes.

In particular, we have the following statement.

THEOREM 4.8. *Under the conditions of Theorem 4.7, an absolutely continuous measure P is invariant under transformation θ if and only if identity*

$$\mathcal{L}\rho(x) = \rho(x) \tag{4.15}$$

holds for Leb*-almost all $x \in \mathbb{R}^d$. In other words, ρ is a fixed point of the transfer operator $\mathcal{L}$ or, equivalently, its eigenfunction with eigenvalue* 1.

So, the problem of finding invariant densities for smooth maps reduces to solving equation (4.15). Although this equation looks quite innocent, in many concrete cases establishing existence of solutions of (4.15), finding these solutions and exploring their properties poses difficult problems.

Sometimes one can derive existence using conjugation to well-studied dynamical systems. For example, the following is a slightly weakened statement of Denjoy's theorem (1932):

THEOREM 4.9. *Let $\theta : \mathbf{S}^1 \to \mathbf{S}^1$ be an orientation preserving C^2-diffeomorphism with no periodic points. Then there is $\alpha \in \mathbb{R} \setminus \mathbb{Q}$ such that θ is continuously conjugated to θ_α defined by*

$$\theta_\alpha(\omega) = \omega + \alpha.$$

Continuous conjugation means that there is a homeomorphism $\phi : \mathbf{S}^1 \to \mathbf{S}^1$ such that

$$\theta = \phi^{-1} \circ \theta_\alpha \circ \phi.$$

Let us make an additional assumption that the conjugation map ϕ is, in fact, a diffeomorphism. Then, since the Lebesgue measure is a unique invariant probability measure for θ_α, we can construct a unique invariant probability measure for θ as the pushforward of Lebesgue measure under ϕ. This measure is absolutely continuous due to the diffeomorphism property of θ.

There is a series of papers where various sufficient conditions for ϕ to satisfy various smoothness conditions are established. It turns out that there is a set A of full Lebesgue measure such that for all $\alpha \in A$ and every orientation preserving diffeomorphism continuously conjugated to θ_α, the conjugation is, in fact, a diffeomorphism. The set A can be described in number-theoretic terms, namely, via the rate of approximation of α by rational numbers, see, e.g. [**KT09**]

Let us look at one concrete example known as the Gauss map. Let $d = 1$ and $\theta x = \{1/x\}$ for $x \in \mathbb{R}^1$. Let us check that the function

$$\rho(x) = \begin{cases} \frac{1}{\ln 2(1+x)}, & x \in [0,1), \\ 0, & \text{otherwise} \end{cases}$$

is an invariant probability density. Ignoring the normalizing constant $1/\ln 2$, noticing that $\theta^{-1}x = \{1/(x+m) : m \in \mathbb{N}\}$, and computing $|\theta'(1/(x+m))| = (x+m)^2$, we see that we need to check

$$\rho(x) = \sum_{m \in \mathbb{N}} \frac{\rho\left(\frac{1}{x+m}\right)}{(x+m)^2}, \quad x \in [0,1).$$

The right-hand side equals

$$\sum_{m\in\mathbb{N}} \frac{1}{1+\frac{1}{x+m}} \frac{1}{(x+m)^2} = \sum_{m\in\mathbb{N}} \frac{1}{(x+m+1)(x+m)}$$
$$= \sum_{m\in\mathbb{N}} \left(\frac{1}{x+m} - \frac{1}{x+m+1} \right) = \frac{1}{1+x} = \rho(x).$$

PROBLEM 4.6. Suppose an open set $U \subset \mathbb{R}^d$, a point $x_0 \in U$, and a map $\theta : \mathbb{R}^d \to \mathbb{R}^d$ satisfy $\lim_{n\to\infty} \theta^n x = x_0$ for all $x \in U$. Then every absolutely continuous invarant measure μ must satisfy $\mu(U) = 0$.

PROBLEM 4.7. Suppose a differentiable map $\theta : \mathbb{R}^d \to \mathbb{R}^d$ satisfies $|D\theta(x)| < 1$ for all x in a bounded open forward invariant set U. Then every absolutely continuous invarant measure μ must satisfy $\mu(U) = 0$.

One can interpret these results as nonexistence of absolutely continuous invariant measures due to "deregularizing" properties of the transfer operator $\mathcal{L}$. Namely, due to contraction in the phase space, under the iterations of $\mathcal{L}$ the densities become large and tends to abnormally concentrate. In fact, in situations where $\mathcal{L}$ is sufficiently regularizing or smoothening, one can prove existence.

THEOREM 4.10 (A. Rényi, 1953). *Suppose $(\Omega, \mathcal{F}) = ([0,1), \mathcal{B})$. Let $m \in \mathbb{N}$, $m > 1$ and $f : [0,1] \to [0,m]$ satisfies $f(0) = 0$, $f(1) = m$ and $f'(x) > 1$ for all $x \in [0,1]$. Then the transformation θ defined by $\theta x = \{f(x)\}$ has an absolutely continuous invariant measure. If $f \in C^r$ for some $r \geq 2$, then the invariant density has a C^{r-1} version.*

THEOREM 4.11 (A. Lasota and J. A. Yorke, 1973). *Let $\theta : [0,1] \to [0,1]$ be a piecewise C^2 function such that $\inf |\theta'| > 1$. Then there is a θ-invariant measure absolutely continuous with respect to Lebesgue measure and such that its density has a version with bounded variation.*

IDEA OF THE PROOF: It turns out that the following "Lasota–Yorke" inequality holds. There are numbers $N \in \mathbb{N}$, $\alpha > 0$ and $\beta \in (0,1)$ such that for every f with bounded variation $V(f)$,

$$V(\mathcal{L}^N f) \leq \alpha \|f\|_{L^1} + \beta V(f).$$

So, taking any $f \geq 0$ and denoting $f_k = \mathcal{L}^k f$, we obtain

$$V(f_{Nk}) \leq \alpha \|f_{N(k-1)}\|_{L^1} + \beta V(f_{N(k-1)}) \leq \alpha \|f\|_{L^1} + \beta V(f_{N(k-1)}),$$

where we used the fact that if $f \geq 0$ and $f \in L^1$, then $\mathcal{L}f \geq 0$ and $\|\mathcal{L}f\|_{L^1} = \|f\|_{L^1}$. Iterating this inequality, we obtain

$$V(f_{Nk}) \leq \alpha(1 + \beta + \beta^2 + \ldots + \beta^{k-1}) \|f\|_{L^1} + \beta^k V(f).$$

Therefore,

$$\limsup_{k\to\infty} V(f_{Nk}) \leq \alpha(1-\beta)^{-1} \|f\|_{L^1},$$

This along with $\|f_k\|_{L^1} \leq \|f\|_{L^1}$ guarantees that $C = \{f_{Nk}\}_{k\in\mathbb{N}}$ is precompact in L^1 (see [**DS88**, Theorem IV.8.20]). Therefore, $\{f_k\}_{k\in\mathbb{N}}$ is also precompact. Mazur's theorem (see [**DS88**, Theorem V.2.6]) says that in Banach spaces precompactness of a set is equivalent to precompactness of its convex hull, so the set $\{g_n\}_{n\in\mathbb{N}}$, where

$$g_n = \frac{1}{n}\sum_{k=0}^{n-1} \mathcal{L}^k f, \quad n \in \mathbb{N}$$

is also relatively compact. So let us choose a convergent subsequence $g_{n_k} \xrightarrow{L^1} g$ and notice that

$$\mathcal{L}g_{n_k} - g_{n_k} = \frac{1}{n}(\mathcal{L}^{n_k} f - f) \xrightarrow{L^1} 0, \quad k \to \infty.$$

Since $\mathcal{L}$ is a continuous operator in L^1, we conclude that $\mathcal{L}g = g$. □

REMARK 4.2. For many interesting systems, in the absence of absolutely continuous invariant distributions one can still define various relevant physical invariant measures. Examples of such distributions are measures of maximal entropy, Sinai–Ruelle–Bowen (SRB) measures. We refer to [**Wal82**, Chapters 8–9] and [**Bal00**, Chapter 4]. This remark may need extending.

CHAPTER 5

Markov Processes

The goal of this chapter is to develop general ergodic theory of stochastic processes with instantaneous loss of memory.

1. Basic notions

Here we begin studying situations where the state of the system at time $n+1$ is not uniquely determined by the state at time n. We will mostly be interested in processes with instantaneous loss of memory. They are usually called Markov processes and are defined by the following property: given the history of the process X up to time n, the state at time $n+1$ is random and has conditional distribution that depends only on the value X_n.

DEFINITION 5.1. Let $(\mathbb{X}, \mathcal{X})$ be a measurable space. A function $P : \mathbb{X} \times \mathcal{X} \to \mathbb{R}$ is called a *transition probabiliy*, or *transition kernel*, or *probability kernel* if for each $x \in \mathbb{X}$, $P(x, \cdot)$ is a probability measure, and for each $B \in \mathcal{X}$, $P(\cdot, B)$ is an $\mathcal{X}$-measurable function.

Throughout this section we assume that $(\mathbb{X}, \mathcal{X})$ is a Borel space. We also fix a transition kernel $P(\cdot, \cdot)$ on $(\mathbb{X}, \mathcal{X})$.

For any measure ρ on $(\mathbb{X}, \mathcal{X})$ (that will serve as the initial distribution) and time $n \in \mathbb{Z}$, we are going to define a measure on path starting at n, i.e., on the space $(\mathbb{X}^{\{n,n+1,\dots\}}, \mathcal{X}^{\{n,n+1,\dots\}})$.

DEFINITION 5.2. *Let $(\mathbb{X}, \mathcal{X})$ be a Borel space. Let ρ be a probability measure on a $(\mathbb{X}, \mathcal{X})$ and let $P(\cdot, \cdot)$ be a transition kernel on $(\mathbb{X}, \mathcal{X})$. An $(\mathbb{X}, \mathcal{X})$-valued process X defined on a probability space $(\Omega, \mathcal{F}, \mathsf{P})$ with time indexed by $\{n, n+1, n+2, \dots\}$ for some $n \in \mathbb{Z}$ is called a homogeneous Markov process with initial distribution ρ and one-step transition probability $P(\cdot, \cdot)$ if for any $k \geq 0$ and any sets $A_0, A_1, \dots, A_k \in \mathcal{X}$,*

$$\begin{aligned} &\mathsf{P}\{X_n \in A_0, \dots, X_{n+k} \in A_k\} \\ &= \int_{A_0} \rho(dx_0) \int_{A_1} P(x_0, dx_1) \dots \int_{A_{k-2}} P(x_{k-2}, dx_{k-1}) \int_{A_{k-1}} P(x_{k-1}, A_k). \end{aligned} \tag{5.1}$$

The existence and uniqueness of a measure $\mathsf{P} = \mathsf{P}_\rho^{n,\infty}$ on the canonical space $(\mathbb{X}^{\{n,n+1,\dots\}}, \mathcal{X}^{\{n,n+1,\dots\}})$ with finite-dimensional distributions described by (5.1) follows from the Kolmogorov–Daniell consistency theorem.

PROBLEM 5.1. Show that formula (5.1) defines a consistent family of finite-dimensional distributions.

PROBLEM 5.2. Show that Definition 5.2 is equivalent to the following: for any $m \in \mathbb{N}$ and any bounded and $(\mathcal{X}^n, \mathcal{B}(\mathbb{R}))$-measurable function $f : \mathbb{X}^{m+1} \to \mathbb{R}$,

$$\mathsf{E} f(X_n, \dots, X_{n+m}) = \int_{\mathbb{X}} \rho(dx_0) \int_{\mathbb{X}} P(x_0, dx_1) \dots \int_{\mathbb{X}} P(x_{m-1}, dx_m) f(x_0, \dots, x_m). \tag{5.2}$$

It is convenient to work on the canonical probability space. For any $n \in \mathbb{Z}$, the canonical random variables X_m on $(\mathbb{X}^{\{n,n+1,\dots\}}, \mathcal{X}^{\{n,n+1,\dots\}})$ are defined by $X_m(x_n, x_{n+1}, \dots,) = x_m$, $m \geq n$.

In what follows we most often work with the case where $n = 0$. If $n = 0$, we will often write P_ρ for $\mathsf{P}_\rho^{n,\infty}$. Also, if $\rho = \delta_y$ for some $y \in \mathbb{X}$, we write P_y for P_{δ_y}.

For any numbers $m_1, m_2 \in \mathbb{Z}$ such that $m_1 \leq m_2$ we denote

$$\mathcal{X}^{m_1 m_2} = \sigma(X_{m_1}, \dots, X_{m_2}).$$

Let us also denote $\mathcal{F}_m = \mathcal{X}^{0m}$ and $\mathcal{X}^{m\infty} = \sigma(X_m, X_{m+1}, \dots)$ for any $m \geq 0$.

It is clear from the definition (5.1), that ρ is the distribution of X_0 under P_ρ, i.e., $\mathsf{P}_\rho\{X_0 \in A\} = \rho(A)$ for every $A \in \mathcal{X}$. Thus, the measure ρ can be interpreted as the starting distribution of the Markov process. Let us convince ourselves that the function $P(\cdot, \cdot)$ can be interpreted as a conditional transition probability.

THEOREM 5.1 (Markov property). *For any $m \in \mathbb{Z}_+$ and $A \in \mathcal{X}$.*

$$\mathsf{P}_\rho(X_{m+1} \in A | \mathcal{F}_m) \overset{\text{a.s.}}{=} \mathsf{P}_\rho(X_{m+1} \in A | X_m) \overset{\text{a.s.}}{=} P(X_m, A).$$

PROOF: It is sufficient to prove that for any m and any measurable bounded function $f : \mathbb{X}^{m+1} \to \mathbb{R}$,

$$\mathsf{E} P(X_m, A) f(X_0, \dots, X_m) = \mathsf{E} \mathbf{1}_{X_{m+1} \in A} f(X_0, \dots, X_m).$$

To prove this formula, we can directly compute both sides of this identity using (5.2). □

THEOREM 5.2. *For any measure ρ, any $m, k \geq 0$ and any $A_1, \dots, A_k \in \mathcal{X}$,*

$$\begin{aligned} \mathsf{P}_\rho(X_{m+1} \in A_1, \dots, X_{m+k} \in A_k | \mathcal{F}_m) &\overset{\text{a.s.}}{=} \mathsf{P}_\rho(X_{m+1} \in A_1, \dots, X_{m+k} \in A_k | X_m) \\ &\overset{\text{a.s.}}{=} \int_{A_1} P(X_m, dx_1) \int_{A_2} P(x_1, dx_2) \dots \int_{A_{k-2}} P(x_{k-2}, dx_{k-1}) \int_{A_{k-1}} P(x_{k-1}, A_k) \\ &\overset{\text{a.s.}}{=} \mathsf{P}_{X_m}^{m\infty}\{X_{m+1} \in A_1, \dots, X_{m+k} \in A_k\}. \end{aligned}$$

PROOF: To be inserted later □

THEOREM 5.3. *For any measure ρ, any $m, k \geq 0$, any $A_0, A_1, \ldots, A_k \in \mathcal{X}$,*

$$\begin{aligned}\mathsf{P}_\rho(X_m \in A_0, \ldots, X_{m+k} \in A_k | \mathcal{F}_m) &\overset{\text{a.s.}}{=} \mathsf{P}_\rho(X_m \in A_0, \ldots, X_{m+k} \in A_k | X_m) \\ &\overset{\text{a.s.}}{=} \mathbf{1}_{X_m \in A_0} \int_{A_1} P(X_m, dx_1) \int_{A_2} P(x_1, dx_2) \ldots \int_{A_{k-2}} P(x_{k-2}, dx_{k-1}) \int_{A_{k-1}} P(x_{k-1}, A_k) \\ &\overset{\text{a.s.}}{=} \mathsf{P}^{m\infty}_{X_m}(X_m \in A_0, \ldots, X_{m+k} \in A_k).\end{aligned}$$

PROOF: This follows directly from the previous theorem and the fact that $\mathbf{1}_{X_m \in A_0}$ is $\sigma(X_m)$-measurable □

THEOREM 5.4. *For any measure ρ, any $m, k \geq 0$, and any bounded measurable $f : \mathcal{X}^{k+1} \to \mathbb{R}$,*

$$\begin{aligned}\mathsf{E}_{\mathsf{P}_\rho}(f(X_m, \ldots, X_{m+k}) | \mathcal{F}_m) &\overset{\text{a.s.}}{=} \mathsf{E}_{\mathsf{P}_\rho}(f(X_m, \ldots, X_{m+k} | X_m)) \\ &\overset{\text{a.s.}}{=} \int_{\mathbb{X}} P(X_m, dx_1) \int_{\mathbb{X}} P(x_1, dx_2) \ldots \int_{\mathbb{X}} P(x_{k-2}, dx_{k-1}) \int_{\mathbb{X}} P(x_{k-1}, dx_k) f(X_m, x_1, \ldots, x_k). \\ &\overset{\text{a.s.}}{=} \mathsf{E}_{\mathsf{P}^{m\infty}_{X_m}} f(X_m, x_1, \ldots, x_k).\end{aligned}$$

PROOF: To be inserted later □

In the following theorem we use the standard shift operator θ on $(\mathbb{X}^{\mathbb{Z}_+}, \mathcal{X}^{\mathbb{Z}_+})$.

THEOREM 5.5. *Let $m \geq 0$, and let $A \in \mathcal{X}^{\mathbb{Z}_+}$. Then*

$$\mathsf{P}_\rho(\theta^{-m} A | \mathcal{F}_m) \overset{\text{a.s.}}{=} \mathsf{P}_\rho(\theta^{-m} A | X_m) \overset{\text{a.s.}}{=} \mathsf{P}_{X_m}(A).$$

PROOF: If $A \in \mathcal{X}^{\mathbb{Z}_+}$, then

$$\theta^{-m} A = \{(x_0, x_1, \ldots) : (x_m, x_{m+1}, \ldots) \in A\} \in \mathcal{X}^{m\infty}.$$

The theorem follows from approximating $\theta^{-m} A$ by cylinder sets based on coordinates starting with the m-th one, and using Theorem 5.4. □

One can easily priove an extension of this theorem where indicators are replaced by arbitrary bounded functions:

THEOREM 5.6. *Let $m \geq 0$, and let $H : \mathcal{X}^{\mathbb{Z}_+} \to \mathbb{R}$ be a bounded random variable. Then*

$$\mathsf{E}_{\mathsf{P}_\rho}(H \circ \theta^m | \mathcal{F}_m) \overset{\text{a.s.}}{=} \mathsf{E}_{\mathsf{P}_\rho}(H \circ \theta^m | X_m) \overset{\text{a.s.}}{=} \mathsf{E}_{\mathsf{P}_{X_m}} H.$$

For example, for every set $B \in \mathcal{X}^{\mathbb{Z}_+}$,

$$\mathsf{P}_\rho(B) = \int_{\mathbb{X}} \rho(dx_0) \mathsf{P}_{x_0}(B). \tag{5.3}$$

This identity follows from

$$\mathsf{P}_\rho(B) = \mathsf{E}_{\mathsf{P}_\rho} \mathbf{1}_B = \mathsf{E}_{\mathsf{P}_\rho} \mathsf{E}_{\mathsf{P}_\rho}(\mathbf{1}_B | \mathcal{F}_0) = \mathsf{E}_{\mathsf{P}_\rho} \mathsf{P}_{X_0}(B) = \int_{\mathbb{X}} \rho(dx_0) \mathsf{P}_{x_0}(B).$$

We already know that the distribution of X_0 under P_ρ is fiven by ρ. Using (5.1), we can also compute the distribution of X_1,

$$\mathsf{P}_\rho\{X_1 \in A\} = \int_{\mathbb{X}} \rho(dx_0)P(x, A).$$

It is convenient to introduce additional notation: for a mesaure ρ on $(\mathbb{X}, \mathcal{X})$ and a kernel on $(\mathbb{X}, \mathcal{X})$, we define the measure ρP by

$$\rho P(A) = \int_{\mathbb{X}} \rho(dx_0)P(x, A). \tag{5.4}$$

Using this notation we can say that the distribution of X_1 is given by ρP.

Also, for two kernels $P(\cdot, \cdot)$ and $Q(\cdot, \cdot)$, we can define a new kernel $PQ(\cdot, \cdot)$ by

$$PQ(x_0, A) = \int_{\mathbb{X}} P(x_0, dx_1)Q(x_1, A), \quad x \in \mathbb{X}, A \in \mathcal{X},$$

we also can inductively introduce $P^0(x_0, \cdot) = \delta_{x_0}$ and

$$P^{n+1} = P^n P = PP^n, \quad n \in \mathbb{N}.$$

With this notation at hand, we can also write

$$\mathsf{P}_\rho\{X_n \in A\} = \rho P^n(A), \quad A \in \mathcal{X}, n \in \mathbb{N}.$$

In the simplest case where $\mathbb{X} = \{1, \dots, N\}$ for some $N \in \mathbb{N}$, the transition kernel can be identified with transition probability matrix $P_{ij} = P(i, \{j\})$, $i, j = 1, \dots, N$. Then the formulas above can be interpreted via matrix products.

2. Stationary Markov Processes and Invariant Distributions

Suppose now we would like to study statistical properties of the Markov process $(X_m)_{m\geq 0}$. We already know that if a process (X_m) is stationary, then one can view it as a metric dynamical system $(\mathbb{X}^{\mathbb{Z}_+}, \mathcal{X}^{\mathbb{Z}_+}, \mathsf{P}, \theta)$ and apply the ergodic theorem to compute limits of averages like

$$\frac{1}{n}\sum_{k=0}^{n-1} f(X_k).$$

So the first question one could ask is, given the transition probability $P(\cdot, \cdot)$, what are the initial distributions ρ such that P_ρ defines a stationary process?

THEOREM 5.7. *P_ρ defines a stationary process iff*

$$\rho P = \rho, \tag{5.5}$$

PROOF: If P_ρ defines a stationary process, then the distributions of X_0 and X_1 must coincide, so (5.5) holds.

Now suppose that (5.5) holds. Then for all m, we have $\rho P^m = \rho$, i.e., the distribution of X_m is also ρ.

We need to check that for all $m \in \mathbb{Z}_+$, all $k \in \mathbb{N}$,

$$\mathsf{P}_\rho\{X_m \in A_0, \dots, X_{m+k} \in A_k\} = \mathsf{P}_\rho\{X_0 \in A_0, \dots, X_k \in A_k\}. \tag{5.6}$$

We can write

$$\begin{aligned}
&\mathsf{P}_\rho\{X_m \in A_0, \dots, X_{m+k} \in A_k\} = \\
=&\mathsf{E}_{\mathsf{P}_\rho}\mathsf{P}_\rho(X_m \in A_0, \dots, X_{m+k} \in A_k | X_m) \\
=&\mathsf{E}_{\mathsf{P}_\rho}\mathbf{1}_{X_m \in A_0}\mathsf{P}_\rho(X_{m+1} \in A_1, \dots, X_{m+k} \in A_k | X_m) \\
=&\mathsf{E}_{\mathsf{P}_\rho}\mathbf{1}_{X_m \in A_0}\int_{A_1} P(X_m, dx_1)\int_{A_2} P(x_1, dx_2)\dots\int_{A_{k-2}} P(x_{k-2}, dx_{k-1})\int_{A_{k-1}} P(x_{k-1}, A_k).
\end{aligned}$$

Denoting for $y \in \mathbb{X}$

$$f(y) = \mathbf{1}_{x \in A_0}\int_{A_1} P(y, dx_1)\int_{A_2} P(x_1, dx_2)\dots\int_{A_{k-2}} P(x_{k-2}, dx_{k-1})\int_{A_{k-1}} P(x_{k-1}, A_k),$$

we can write

$$\mathsf{P}_\rho\{X_m \in A_0, \dots, X_{m+k} \in A_k\} = \mathsf{E}_{\mathsf{P}_\rho} f(X_m) = \int_{\mathbb{X}} \rho(dy) f(y),$$

since the distribution of X_m is ρ. This expression on the right-hand side does not depend on m, so the proof of (5.6) is complete. □

DEFINITION 5.3. *Any distribution ρ that satisfies* (5.5) *is called P-invariant.*

If $\mathbb{X}$ is finite, say $\mathbb{X} = \{1, \dots, N\}$ for some $N \in \mathbb{N}$, then any measure ρ and any transition kernel P are uniquely defined by their values on single-point sets. Denoting $\rho_i = \rho\{i\}$ and $P_{ij} = P(i, \{j\})$ for all $i, j \in \mathbb{X}$, we obtain that ρ is P-invariant iff

$$\sum_{i \in \mathbb{X}} \rho_i P_{ij} = \rho_j. \tag{5.7}$$

In other words, the identity $\rho P = \rho$ can be understood in the linear algebra sense. Here ρ is a row vector and P is a square matrix.

For example, let $\mathbb{X} = \{1, 2, 3\}$ with $\mathcal{X} = 2^{\mathbb{X}}$, and $P(x, A) = |A \setminus \{x\}|/2$ for any $A \subset \mathbb{X}$ which means that, given that at any time step the system changes its current state to a new one chosen uniformly from the remaining two states. Denoting

$$P_{ij} = P(i, \{j\}) = \begin{cases} \frac{1}{2}, & j \neq i, \\ 0, & j = i, \end{cases}$$

we see that the system (5.7) has a line of solutions $\rho_1 = \rho_2 = \rho_3$. Since we are interested in probability measures, we have to set $\rho_1 = \rho_2 = \rho_3 = 1/3$, and so this transition kernel has a unique invariant distribution.

Another example. Let us consider the following probability kernel on the real line:

$$P(x,A)=\frac{1}{2}\delta_{\frac{x}{2}}(A)+\frac{1}{2}\delta_{\frac{1+x}{2}}(A),\quad x\in\mathbb{R}.$$

In terms of Markov processes, this means that between times n and $n+1$ the process jumps from X_n to either $X_n/2$ or $(1+X_n)/2$, and these values are each chosen with probability $1/2$.

Let us check that the uniform distribution on $[0,1]$ is P-invariant. Let us take any Borel $A\subset[0,1]$ and compute

$$\begin{aligned}\int_{[0,1]}P(x,A)dx&=\frac{1}{2}\int_{[0,1]}\mathbf{1}_A(x/2)dx+\frac{1}{2}\int_{[0,1]}\mathbf{1}_A((1+x)/2)dx\\&=\frac{1}{2}\operatorname{Leb}\left\{x\in[0,1]:\frac{x}{2}\in A\right\}+\frac{1}{2}\operatorname{Leb}\left\{x\in[0,1]:\frac{1+x}{2}\in A\right\}\\&=\frac{1}{2}\operatorname{Leb}((2A)\cap[0,1])+\frac{1}{2}\operatorname{Leb}((2A-1)\cap[0,1])\\&=\operatorname{Leb}(A\cap[0,1/2])+\operatorname{Leb}(A\cap[1/2,1])=\operatorname{Leb}(A).\end{aligned}$$

In fact, there are no other invariant distributions, but checking this is not so easy as in the first example.

Problem 5.3. Use binary decompositions of numbers in $[0,1]$ to give another proof of invariance of the uniform distribution in this example.

Problem 5.4. Find an invariant distribution for the following probability kernel on the real line:

$$P(x,A)=\frac{1}{2}\delta_{\frac{x}{3}}(A)+\frac{1}{2}\delta_{\frac{2+x}{3}}(A),\quad x\in\mathbb{R}.$$

Let us consider another example on the real line. Let us fix two parameters $a\in(0,1)$ and $\sigma^2>0$ and suppose that $P(x,\cdot)$ is a Gaussian measure with mean ax and variance σ^2. Let us find a number $r^2>0$ such that a centered Gaussian distribution with variance r^2 is invariant. We need to make sure that for any Borel $A\subset\mathbb{R}$,

$$\int_A\frac{dx}{\sqrt{2\pi r}}e^{-\frac{x^2}{2r}}=\int_{\mathbb{R}}\frac{dx}{\sqrt{2\pi r}}e^{-\frac{x^2}{2r}}\int_A\frac{dy}{\sqrt{2\pi}\sigma}e^{-\frac{(y-ax)^2}{2\sigma^2}}.$$

After a change of variables $ax=z$, the right-hand side becomes

$$\int_{\mathbb{R}}\frac{dx}{\sqrt{2\pi}ra}e^{-\frac{z^2}{2r^2a^2}}\int_A\frac{dy}{\sqrt{2\pi}\sigma}e^{-\frac{(y-z)^2}{2\sigma^2}}=\int_A p(y)dy,$$

where p is the convolution of densities of two centered Gaussian distributions with variances r^2a^2 and σ^2, i.e., p is a centered Gaussian density with variance $r^2a^2+\sigma^2$. So we will have an invariant Gaussian distribution if and only if $r^2=r^2a^2+\sigma^2$, i.e., $r^2=\sigma^2/(1-a^2)$. We will prove later that there are no other invariant distributions for this Markov semigroup.

Now we address the structure of the set of invariant measures.

We already know that deterministic transformations can fail to have any invariant distributions at all. The same applies to Markov transition kernels. In fact, any deterministic transformation ϕ on a space $(\mathbb{X}, \mathcal{X})$ naturally generates a Markov transition kernel given by $P(x, \cdot) = \delta_{\phi(x)}$, i.e., for any measure ρ on $(\mathbb{X}, \mathcal{X})$, $\rho P = \rho\phi^{-1}$. So basic examples of deterministic transformations without invariant distributions naturally generate examples of Markov kernels without invariant distributions.

If the set of invariant distributions for P is nonempty, then it has a structure similar to that of the set of invariant measures for deterministic transformations. To make more concrete statements we need to introduce several new notions.

DEFINITION 5.4. *Let $P(\cdot, \cdot)$ be a probability kernel on $(\mathbb{X}, \mathcal{X})$ and let ρ be a P-invariant probability measure on $(\mathbb{X}, \mathcal{X})$. A set A is called (P, ρ)-invariant if*

$$\rho\{x \in A : P(x, A^c) > 0\} = 0.$$

The set of all (ρ, P)-invariant sets is denoted by $\mathcal{I}(\rho, P)$.

LEMMA 5.1. *Let $P(\cdot, \cdot)$ be a probability kernel on $(\mathbb{X}, \mathcal{X})$ and let ρ be a P-invariant probability measure on $(\mathbb{X}, \mathcal{X})$. The set $\mathcal{I}(\rho, P)$ is a σ-algebra.*

PROOF: It is obvious that $\Omega \in \mathcal{I}(\rho, P)$ since $P(x, \Omega) = 1$ for all x.

If $A_1, A_2, \ldots \in \mathcal{I}(\rho, P)$, then for $A = \bigcup_{i \in \mathbb{N}} A_i$, we have

$$\begin{aligned}\rho\{x \in A : P(x, A^c) > 0\} &\leq \sum_i \rho\{x \in A_i : P(x, A^c) > 0\} \\ &\leq \sum_i \rho\{x \in A_i : P(x, A_i^c) > 0\} = 0.\end{aligned}$$

Suppose now $A \in \mathcal{I}(\rho, P)$. Let us prove that $A^c \in \mathcal{I}(\rho, P)$. We have

$$\rho(A) = \int_{\mathbb{X}} \rho(dx) P(x, A) = \int_A \rho(dx) P(x, A) + \int_{A^c} \rho(dx) P(x, A).$$

The first term on the right equals $\rho(A)$ because for ρ-almost all $x \in A$, $P(x, A) = 1$. Therefore, the second term is zero, which implies that $P(x, A) = 0$ for ρ-almost all $x \in A^c$. □

DEFINITION 5.5. *Let $P(\cdot, \cdot)$ be a probability kernel on $(\mathbb{X}, \mathcal{X})$ and let ρ be a P-invariant probability measure on $(\mathbb{X}, \mathcal{X})$. We say that the pair (ρ, P) is ergodic if for every (ρ, P)-invariant set A, $\rho(A) \in \{0, 1\}$.*

Ergodicity means that one cannot decompose the system into two systems that can be studied independently. We will also often say that ρ is P-ergodic if (ρ, P) is an ergodic pair.

THEOREM 5.8. *Suppose ρ is an invariant measure for a Markov kernel $P(\cdot, \cdot)$ on $(\mathbb{X}, \mathcal{X})$. A set $B \in \mathcal{X}^{\mathbb{Z}_+}$ belongs to the σ-algebra $\mathcal{I}^*(\mathbb{X}^{\mathbb{Z}_+}, \mathcal{X}^{\mathbb{Z}_+}, \mathsf{P}_\rho, \theta)$*

iff there is $A \in \mathcal{I}(\rho, P)$ *such that*

$$B \overset{\mathsf{P}_\rho}{=} \{X_0 \in A\} = A \times \mathbb{X} \times \mathbb{X} \times \ldots = A \times \mathbb{X}^{\mathbb{N}}. \tag{5.8}$$

If $A \in \mathcal{I}(\rho, P)$*, then condition* (5.8) *is equivalent to*

$$B \overset{\mathsf{P}_\rho}{=} \{X_0 \in A, X_1 \in A, \ldots\} = A \times A \times A \times \ldots = A^{\mathbb{Z}_+}. \tag{5.9}$$

Also $\mathsf{P}_\rho(B) = \rho(A)$.

Proof: First, let us prove the equivalence of conditions (5.8) and (5.9).

For all $n \in \mathbb{Z}_+$, we introduce $B_n = A^{\{0,1,\ldots,n\}} \times \mathbb{X}^{\{n+1,n+2,\ldots\}}$. Then

$$\mathsf{P}_\rho(B_n) = \int_A \rho(dx_0) \int_A P(x_0, dx_1) \ldots \int_A P(x_{k-1}, A) = \rho(A),$$

due to the invariance of A. Since $B_{n+1} \subset B_n$ for all n, and $\bigcap_{n\in\mathbb{Z}_+} B_n = A^{\mathbb{Z}_+}$, we obtain $\mathsf{P}_\rho(A^{\mathbb{Z}_+}) = \rho(A)$. Since $B_0 = A \times \mathbb{X}^{\mathbb{N}}$ and $\mathsf{P}_\rho(B_0) = \rho(A)$, we have $A^{\mathbb{Z}_+} \overset{\mathsf{P}_\rho}{=} A \times \mathbb{X}^{\mathbb{N}}$.

To see that $A \in \mathcal{I}(\rho, \mathsf{P})$ implies $A^{\mathbb{Z}^+} \in \mathcal{I}^*(\mathbb{X}^{\mathbb{Z}_+}, \mathcal{X}^{\mathbb{Z}_+}, \mathsf{P}_\rho, \theta)$, we can either use the forward invariance of A under θ, or write

$$\theta^{-1} A^{\mathbb{Z}_+} = \mathbb{X} \times A^{\mathbb{N}} = A^{\mathbb{Z}_+} \cup (A^c \times A^{\mathbb{N}})$$

and notice that $\mathsf{P}_\rho(A^c \times A^{\mathbb{N}}) = 0$.

Now let us assume that $B = \theta^{-1} B$. We have

$$\mathsf{P}_\rho(B|\mathcal{F}_n) \xrightarrow{\text{a.s.}} \mathsf{P}_\rho(B|\mathcal{X}^{\mathbb{Z}_+}) \overset{\text{a.s.}}{=} \mathbf{1}_B,$$

since the sequence $\mathsf{P}_\rho(B|\mathcal{F}_n)$, $n \in \mathbb{Z}_+$, forms a bounded martingale with respect to the filtration $(\mathcal{F}_n)_{n\in\mathbb{Z}_+}$, $\cup_n \mathcal{F}_n = \mathcal{X}^{\mathbb{Z}_+}$, and Doob's convergence theorem applies.

On the other hand, the invariance of B and the Markov property (more precisely, Theorem 5.5) imply that for all $n \in \mathbb{Z}_+$:

$$\mathsf{P}_\rho(B|\mathcal{F}_n) \overset{\text{a.s.}}{=} \mathsf{P}_\rho(\theta^{-n} B|\mathcal{F}_n) \overset{\text{a.s.}}{=} \mathsf{P}_\rho(\theta^{-n} B|X_n) \overset{\text{a.s.}}{=} \mathsf{P}_{X_n}(B) = \phi(X_n) = f(\theta^n x). \tag{5.10}$$

for some measurable functions $\phi : \mathbb{X} \to [0,1]$ and $f : \mathbb{X}^{\mathbb{Z}_+} \to [0,1]$. These functions do not denpend on n.

We claim that

$$f(x) \overset{\text{a.s.}}{=} \mathbf{1}_B(x). \tag{5.11}$$

This will follow from

$$f(x)\mathbf{1}_B(x) \overset{\text{a.s.}}{=} \mathbf{1}_B(x), \tag{5.12}$$

$$f(x)\mathbf{1}_{B^c}(x) \overset{\text{a.s.}}{=} 0. \tag{5.13}$$

To prove these identities, we will combine (5.10) with the fact that if for uniformly bounded random variables $(\xi_n)_{n\geq 0}$ and η, $\xi_n \xrightarrow{\text{a.s.}} \eta$, then $\xi_n \mathbf{1}_C \xrightarrow{L^1} \eta \mathbf{1}_C$ for any event C. Specifically, we can write

$$f(\theta^n x)\mathbf{1}_B(x) \xrightarrow{L^1} \mathbf{1}_B(x) \cdot \mathbf{1}_B(x) = \mathbf{1}_B(x), \tag{5.14}$$

and

$$f(\theta^n x)\mathbf{1}_{B^c}(x) \xrightarrow{L^1} \mathbf{1}_B(x)\cdot\mathbf{1}_B^c(x) = 0. \tag{5.15}$$

Using the invariance of B, we can rewrite (5.14) and (5.15) as

$$f(\theta^n x)\mathbf{1}_B(\theta x) \xrightarrow{L^1} \mathbf{1}_B(x), \tag{5.16}$$

and

$$f(\theta^n x)\mathbf{1}_{B^c}(\theta x) \xrightarrow{L^1} 0, \tag{5.17}$$

Since

$$\mathsf{E}_{\mathsf{P}_\rho} f(\theta^n x)\mathbf{1}_B(\theta x) = \mathsf{E}_{\mathsf{P}_\rho} f(x)\mathbf{1}_B(x),$$

the convergence (5.16) implies $\mathsf{E}_{\mathsf{P}_\rho} f(x)\mathbf{1}_B(x) = \mathsf{E}_{\mathsf{P}_\rho}\mathbf{1}_B(x)$, and (5.12) follows. Since

$$\mathsf{E}_{\mathsf{P}_\rho} f(\theta^n x)\mathbf{1}_{B^c}(\theta x) = \mathsf{E}_{\mathsf{P}_\rho} f(x)\mathbf{1}_{B^c}(x),$$

the convergence (5.17) implies $\mathsf{E}_{\mathsf{P}_\rho} f(x)\mathbf{1}_B(x) = 0$, and (5.13) follows. The proof of (5.11) is completed, and (5.10) is established.

We can now use (5.10) to write $\mathbf{1}_B \overset{\text{a.s.}}{=} \phi(X_n)$ for all $n \in \mathbb{Z}_+$. We see that ϕ takes values 0 and 1 with probability 1. Therefore, $\phi(X_n) = \mathbf{1}_{X_n\in A}$ for a set $A \in \mathcal{X}$. So, $B \overset{\mathsf{P}_\rho}{=} \{X_n \in A\}$ for all $n \in \mathbb{Z}_+$ and, moreover, $B \overset{\mathsf{P}_\rho}{=} \bigcap\{X_n \in A\} = A^{\mathbb{Z}_+}$.

Let us prove that A is (ρ, P)-invariant. We have

$$\int_A \rho(dx_0)P(x_0, A) = \mathsf{P}_\rho\{X_0 \in A, X_1 \in A\} = \mathsf{P}_\rho(B) = \mathsf{P}\{X_0 \in A\} = \rho(A).$$

Therefore, $P(x_0, A) = 1$ for ρ-a.e. $x_0 \in A$. □

Definition 5.6. *Let us denote by $\mathcal{I}_0$ the σ-subalgebra of $\mathcal{X}^{\mathbb{Z}_+}$ generated by sets $\{X_0 \in A\}$, $A \in \mathcal{I}(\rho, P)$.*

Remark 5.1. Theorem 5.8 may be interpreted as

$$\mathcal{I}(\mathbb{X}^{\mathbb{Z}_+}, \mathcal{X}^{\mathbb{Z}_+}, \mathsf{P}_\rho, \theta) \overset{\mathsf{P}_\rho}{=} \mathcal{I}^*(\mathbb{X}^{\mathbb{Z}_+}, \mathcal{X}^{\mathbb{Z}_+}, \mathsf{P}_\rho, \theta) \overset{\mathsf{P}_\rho}{=} \mathcal{I}_0. \tag{5.18}$$

Lemma 5.2. *Let $g : \mathbb{X} \to \mathbb{R}$ be $\mathcal{X}$-measurable and bounded. For P_ρ-almost every $x \in \mathbb{X}$,*

$$\int_{\mathbb{X}^{\mathbb{Z}_+}} \mathsf{P}_{\rho,\mathcal{I}_0}(x, dy)g(y_0) = \int_{\mathbb{X}} \rho_{\mathcal{I}}(x_0, dy_0)g(y_0), \tag{5.19}$$

where $\rho_{\mathcal{I}}(\cdot,\cdot) = \rho_{\mathcal{I}(\rho,P)}$ is a regular conditional probability with respect to $\mathcal{I}(\rho, P)$.

Proof: Since both sides of (5.19) are $\mathcal{I}_0$-measurable, it suffices to check that for every $B \in \mathcal{I}_0$,

$$\int_B \mathsf{P}_\rho(dx) \int_{\mathbb{X}^{\mathbb{Z}_+}} \mathsf{P}_{\rho,\mathcal{I}_0}(x, dy)g(y_0) = \int_B \mathsf{P}_\rho(dx) \int_{\mathbb{X}} \rho_{\mathcal{I}}(x_0, dy_0)g(y_0). \tag{5.20}$$

Let us denote the left-hand side and right-hand side of (5.20) by L and R, respectively. Let $A \in \mathcal{I}(\rho, P)$ be the projection of B onto the zeroth coordinate, i.e., let A be taken from the representation (5.8) of B.

By the definitions of conditional expectation, regular conditional probability, and of P_ρ, we have

$$L = \mathsf{E}_{\mathsf{P}_\rho} \mathbf{1}_B(X) g(X_0) = \mathsf{E}_{\mathsf{P}_\rho} \mathbf{1}_A(X_0) g(X_0) = \int_A \rho(dx_0) g(x_0).$$

Since the integrand on the right-hand side of (5.20) depends only on x_0,

$$R = \int_A \rho(dx_0) \int_{\mathbb{X}} \rho_{\mathcal{I}}(x_0, dy_0) g(y_0) = \int_A \rho(dx_0) g(x_0),$$

where we used the definitions of conditional expectation and regular conditional probability. So $L = R$, and the proof is complete. □

THEOREM 5.9. *Let ρ be an invariant distribution for a Markov kernel P on a space $(\mathbb{X}, \mathcal{X})$. The pair (ρ, P) is ergodic iff $(\mathbb{X}^{\mathbb{Z}_+}, \mathcal{X}^{\mathbb{Z}_+}, \mathsf{P}_\rho, \theta)$ is ergodic.*

PROOF: This is a direct corollary of Theorem 5.8. □

THEOREM 5.10. *If (ρ_1, P) and (ρ_2, P) are ergodic and $\rho_1 \neq \rho_2$, then ρ_1 and ρ_2 are mutually singular.*

LEMMA 5.3. *Two finite measures ρ_1 and ρ_2 on $(\mathbb{X}, \mathcal{X})$ are not mutually singular if there are measures μ, ν_1, ν_2 such that $\mu(\mathbb{X}) > 0$ and*

$$\rho_1 = \mu + \nu_1, \tag{5.21}$$

$$\rho_2 = \mu + \nu_2. \tag{5.22}$$

PROOF: Suppose decompositions (5.21)–(5.22) hold true. Let us assume that ρ_1 and ρ_2 are mutually singular, i.e., that there is a set A such that $\rho_1(A) = 0$ and $\rho_2(A^c) = \rho_2(0)$. From (5.21), $\mu(A) = 0$. From (5.22), $\mu(A^c) = 0$. Therefore, $\mu(\mathbb{X}) > 0$ cannot hold.

Let now have two measures ρ_1 and ρ_2 that are not mutually sungular. Let $\gamma = \rho_1 + \rho_2$ be a new measure. Then both ρ_1 and ρ_2 are absolutely continuous with respect to γ. Let p_1 and p_2 be densities (Radon–Nikodym derivarives) of ρ_1 and ρ_2 with respect to γ. Let us define $q(x) = p_1(x) \wedge p_2(x)$, $x \in \mathbb{X}$. Let us define $r_1(x) = p_1(x) - q(x)$ and $r_2(x) = p_2(x) - q(x)$. Let μ, ν_1, ν_2 be measures absolutely continuous with respect to γ with densities, respectively, q, r_1, r_2. Then identities (5.21)–(5.22) hold true.

If $\mu(\mathbb{X}) = 0$, then $q(x) = 0$ γ-a.s., and we get $\rho_1(dx) = r_1(x)\gamma(dx)$ and $\rho_2(dx) = r_2(x)\gamma(dx)$. Therefore $\rho_1(A_1) = \rho_1(\mathbb{X})$ and $\rho_2(A_2) = \rho_2(\mathbb{X})$, where $A_1 = \{x : r_1(x) > 0\}$ and $A_2 = \{x : r_2(x) > 0\}$ are two disjoint sets. □

PROOF OF THEOREM 5.10: Suppose that ρ_1 and ρ_2 are not mutually singular. Let μ, ν_1, ν_2 be defined according to Lemma 5.3. Then

$$\mathsf{P}_{\rho_1} = \mathsf{P}_\mu + \mathsf{P}_{\nu_1}, \tag{5.23}$$

$$\mathsf{P}_{\rho_2} = \mathsf{P}_\mu + \mathsf{P}_{\nu_2}. \tag{5.24}$$

Due to Lemma 5.9, measures P_{ρ_1} and P_{ρ_2} are ergodic and hence mutually singular. However, this contradicts Lemma 5.3 and decompositions (5.23)–(5.24). □

THEOREM 5.11. *Let ρ be a probability measure on a Borel space $(\mathbb{X}, \mathcal{X})$, invariant under $P(\cdot,\cdot)$. Let $\rho_{\mathcal{I}}(\cdot,\cdot)$ be a regular conditional probability with respect to $\mathcal{I}(\rho, P)$. Then for ρ-almost every x_0, the measure $\rho_{\mathcal{I}}(x_0,\cdot)$ is P-invariant, forms an ergodic pair with P, and ρ is a mixture or convex combination of those ergodic measures:*

$$\rho = \int_{\mathbb{X}} \rho(dx_0)\rho_{\mathcal{I}}(x_0,\cdot),$$

i.e.,

$$\rho(A) = \int_{\mathbb{X}} \rho(dx_0)\rho_{\mathcal{I}}(x_0, A), \quad A \in \mathcal{X}, \tag{5.25}$$

and, more generally,

$$\int_{\mathbb{X}} \rho(dx_0) f(x_0) = \int_{\mathbb{X}} \rho(dx_0) \int_{\Omega} \rho_{\mathcal{I}}(x_0, dy_0) f(y_0), \quad f \in L^1(\mathbb{X}, \mathcal{X}, \rho). \tag{5.26}$$

PROOF: We start with the ergodic decomposition for the measure P_ρ described in Theorem 4.6. We know that for P_ρ-almost every $x \in \mathbb{X}_+^{\mathbb{Z}}$, the measure $\mathsf{P}_{\rho,\mathcal{I}} = \mathsf{P}_{\rho,\mathcal{I}(\mathbb{X}^{\mathbb{Z}_+},\mathcal{X}^{\mathbb{Z}_+},\mathsf{P}_\rho,\theta)}(x,\cdot)$ is θ-invariant and ergodic. Due to (5.18) and part 3 of Theorem 4.4, we also have that for P_ρ-almost every $x \in \mathbb{X}_+^{\mathbb{Z}}$,

$$\mathsf{P}_{\rho,\mathcal{I}(\mathbb{X}^{\mathbb{Z}_+},\mathcal{X}^{\mathbb{Z}_+},\mathsf{P}_\rho,\theta)}(x,\cdot) = \mathsf{P}_{\rho,\mathcal{I}_0}(x,\cdot).$$

Let us prove that for almost every x, $\mathsf{P}_{\rho,\mathcal{I}_0}(x,\cdot)$ is a Markov measure on $(\mathbb{X}^{\mathbb{Z}_+}, \mathcal{X}^{\mathbb{Z}^+})$. We can assume that $(\mathbb{X}, \mathcal{X})$ is the unit segment with Borel σ-algebra. Let us take a countable set D dense in $C[0,1]$. The values of

$$\mathsf{E}_{\mathsf{P}}[f_0(X_0)f_1(X_1)\dots f_n(X_n)], \quad n \in \mathbb{Z}_+, \quad f_0, f_1, \dots, f_n \in D,$$

uniquely define the measure P on $(\mathbb{X}^{\mathbb{Z}_+}, \mathcal{X}^{\mathbb{Z}^+})$. So let us compute

$$\begin{aligned}\mathsf{E}_{\mathsf{P}_{\rho,\mathcal{I}_0}(x,\cdot)}[f_0(X_0)f_1(X_1)\dots f_n(X_n)] &= \mathsf{E}_{\mathsf{P}_\rho}[f_0(X_0)f_1(X_1)\dots f_n(X_n)|\mathcal{I}_0](x) \\ &= \mathsf{E}_{\mathsf{P}_\rho}[\mathsf{E}_{\mathsf{P}_\rho}[f_0(X_0)f_1(X_1)\dots f_n(X_n)|\mathcal{F}_0]|\mathcal{I}_0](x).\end{aligned}$$

We have

$$\begin{aligned}&\mathsf{E}_{\mathsf{P}_\rho}[f_0(X_0)f_1(X_1)\dots f_n(X_n)|\mathcal{F}_0]\\ =&f_0(X_0)\int_{\mathbb{X}}P(X_0,dy_1)f_1(y_1)\int_{\mathbb{X}}P(y_1,dy_2)f_2(y_2)\dots\int_{\mathbb{X}}P(y_{n-1},dy_n)f_n(y_n)\\ =&g(X_0)=G(X)\end{aligned}$$

for some functions $g:\mathbb{X}\to\mathbb{R}$, $G:\mathbb{X}^{\mathbb{Z}_+}\to\mathbb{R}$. Now

$$\begin{aligned}\mathsf{E}[g(X_0)|\mathcal{I}_0](x)&=\int_{\mathbb{X}^{\mathbb{Z}_+}}\mathsf{P}_{\mathcal{I}_0}(x,dy)G(y)\\ &=\int_{\mathbb{X}^{\mathbb{Z}_+}}\mathsf{P}_{\mathcal{I}_0}(x,dy)g(y_0)=\int_{\mathbb{X}}\rho_{\mathcal{I}}(x_0,dy_0)g(y_0),\end{aligned}$$

where in the last identity we used Lemma 5.2.

We thus obtain that for almost all $x\in\mathbb{X}^{\mathbb{Z}_+}$,

$$\begin{aligned}&\mathsf{E}_{\mathsf{P}_{\rho,\mathcal{I}_0(x,\cdot)}}[f_0(X_0)f_1(X_1)\dots f_n(X_n)]\\ &=\int_{\mathbb{X}}\rho_{\mathcal{I}}(x_0,y_0)f_0(y_0)\int_{\mathbb{X}}P(y_0,dy_1)f_1(y_1)\dots\int_{\mathbb{X}}P(y_{n-1},dy_n)f_n(y_n).\end{aligned}$$

Now, the union of all exceptional sets in this identity for all choices of n and functions $f_i\in D$, $i=0,\dots,n$, still has probability zero, and we obtain that there is a set $\bar\Omega\subset\mathbb{X}^{\mathbb{Z}_+}$ such that $\mathsf{P}_\rho(\bar\Omega)=1$ and for every $x\in\bar\Omega$, $\mathsf{P}_{\rho,\mathcal{I}}(x_0,\cdot)$ defines a Markov process with transition kernel $P(\cdot,\cdot)$ and initial distribution $\rho_{\mathcal{I}}(x_0,\cdot)$. The set $\bar\Omega$ can be also chosen so that all these measures are invariant and ergodic with respect to θ. Therefore, their marginal distributions $\rho_{\mathcal{I}}(x_0,\cdot)$ are P-invariant and form ergodic pairs with P.

Identities (5.25) and (5.26) follow directly from the definitions of conditional expectation and regular conditional probability. □

Corollary 5.1. *If there is a P-invariant measure for a kernel P, there is a P-ergodic measure.*

If there are two distinct P-invariant measures, then there are two distinct P-ergodic measures, i.e., to prove uniqueness of a P-invariant distribution, it suffices to show uniqueness of P-ergodic distribution.

3. Filtrations. Stopping times. Strong Markov property

(I already have used some martingale techniques, so some material has to be reordered.)

Suppose we have a probability space $(\Omega,\mathcal{F},\mathsf{P})$.

Definition 5.7. A family $(\mathcal{F}_n)_{n\in\mathbb{Z}_+}$ of σ-algebras is called a *filtration* if $\mathcal{F}_n\subset\mathcal{F}_{n+1}$ for every $n\in\mathbb{Z}_+$.

The standard interpretation of filtrations is that for every $n \in \mathbb{Z}_+$, $\mathcal{F}_n$ is the collection of events observed on the time interval $\{0, 1, \dots, n\}$. One can fully decide whether an event $A \in \mathcal{F}_n$ is true by observing all available information up to time n.

DEFINITION 5.8. Let $(X_n)_{n\geq 0}$ be a stochastic process. Then the filtration $(\mathcal{F}_n)_{n\geq 0}$ defined by $\mathcal{F}_n = \sigma(X_0, \dots, X_n)$ is the *natural filtration* of $(X_n)_{n\geq 0}$ or *filtration generated by* $(X_n)_{n\geq 0}$.

In this case $\mathcal{F}_n$ cosnsists of all sets that can be described in terms of random variables $X_0, \dots, X_n$, $n \geq 0$. One can decide whether an event $A \in \mathcal{F}_n$ is true by observing the trajectory of the process X up to time n.

DEFINITION 5.9. Let $(X_n)_{n\geq 0}$ be a stochastic process. We say that $(X_n)_{n\geq 0}$ is *adapted* to a filtration $(\mathcal{F}_n)_{n\geq 0}$ if for every n, the random variable X_n is measurable with respect to $\mathcal{F}_n$.

In particular, any process is adapted to its own natural filtration.

DEFINITION 5.10. A random variable $\tau : \Omega \to \mathbb{Z}_+ \cup \{+\infty\}$ is called a *stopping time* with respect to $(\mathcal{F}_n)_{n\in\mathbb{Z}_+}$ if $\{\tau \leq n\} \in \mathcal{F}_n$ for every $n \geq 0$.

PROBLEM 5.5. Check that in the definition above one can replace $\{\tau \leq n\} \in \mathcal{F}_n$ by $\{\tau = n\} \in \mathcal{F}_n$ for every $n \geq 0$.

The following is one of the most useful examples of hitting times. Let $(X_n)_{n\geq 0}$ be an $(\mathbb{X}, \mathcal{X})$-valued stochastic process adapted to a filtration $(\mathcal{F}_n)_{n\geq 0}$. This means that for every n, the random variable X_n is measurable with respect to $\mathcal{F}_n$. Let A be any set in $\mathcal{X}$, and let

$$\tau_A = \inf\{n \in \mathbb{N} : \ X_n \in A\} \in \mathbb{Z}_+ \cup \{+\infty\}. \tag{5.27}$$

Then τ_A is a stopping time since

$$\{\tau_A = n\} = \bigcap_{k=1}^{n-1} \{X_k \notin A\} \cap \{X_n \in A\}.$$

DEFINITION 5.11. The σ-algebra associated to a filtration $(\mathcal{F}_n)_{n\geq 0}$ and a stopping time τ with respect to $(\mathcal{F}_n)_{n\geq 0}$ is defined by $\mathcal{F}_\tau = \{A \in \mathcal{F} : A \cap \{\tau = n\} \in \mathcal{F}_n\}$.

PROBLEM 5.6. Prove that $\mathcal{F}_\tau$ is a σ-algebra. Give an example of a filtration and a random variable τ (that is not a stopping time) such that $\mathcal{F}_\tau$ defined above is not a σ-algebra

The σ-algebra $\mathcal{F}_\tau$ is interpreted as the σ-algebra of events observed on a random time interval $\{0, 1, \dots, \tau\}$. In other words, every event from $\mathcal{F}_\tau$ can be described in terms of the information available up to time τ. If $(\mathcal{F}_n)_{n\geq 0}$ is the natural filtration of $(X_n)_{n\geq 0}$, then events from $\mathcal{F}_\tau$ can be described in terms of the realization of X up to time τ, i.e., on can decide whether an event $A \in \mathcal{F}_\tau$ is true or not based on the trajectory $X_0, X_1, \dots, X_\tau$ (of random length).

PROBLEM 5.7. Let $(\mathcal{F}_n)_{n\geq 0}$ be a filtration. Let $(X_n)_{n\geq 0}$ be adapted to $(\mathcal{F}_n)_{n\geq 0}$. Let τ be a finite stopping time with respect to $(\mathcal{F}_n)_{n\geq 0}$. Prove that τ and X_τ are $\mathcal{F}_\tau$ measurable.

The standard Markov property means that the future depends on the past only through the present for any given time n, see, e.g., Theorems 5.5 and 5.4.

A nontrivial strengthening of the Markov property is obtained if one requires the same with respect to a random time τ. Let us recall that we work with Markov processes on the canonical space $(\mathbb{X}^{\mathbb{Z}_+}, \mathcal{X}^{\mathbb{Z}_+})$ eqauipped with the standard shift operator θ, and its natural filtration $(\mathcal{F}_n)$ that we will also call the canonical filtration. For all initial distributions ρ and all Markov kernels $P(\cdot,\cdot)$ on $(\mathbb{X},\mathcal{X})$, P_ρ is the corresponding measure on the canonical path space $(\mathbb{X}^{\mathbb{Z}_+}, \mathcal{X}^{\mathbb{Z}_+})$

THEOREM 5.12. *Let τ be a stopping time with respect to $(\mathcal{F}_n)_{n\geq 0}$, and let $H : \mathcal{X}^{\mathbb{Z}_+} \to \mathbb{R}$ be a bounded random variable. For every distribution ρ on $(\mathbb{X},\mathcal{X})$, on $\{\tau < \infty\}$ we have*

$$\mathsf{E}_{\mathsf{P}_\rho}(H\circ\theta^\tau|\mathcal{F}_\tau)\overset{\text{a.s.}}{=}\mathsf{E}_{\mathsf{P}_\rho}(H\circ\theta^\tau|X_\tau)\overset{\text{a.s.}}{=}\mathsf{E}_{\mathsf{P}_{X_\tau}}H.$$

DEFINITION 5.12. *The claim of Theorem 5.12 is called the strong Markov property.*

PROOF OF THEOREM 5.12: We need to check that for every $B \in \mathcal{F}_\tau$,

$$\mathsf{E}_{\mathsf{P}_\rho}\left[H\circ\theta^\tau\cdot\mathbf{1}_{B\cap\{\tau<\infty\}}\right]=\mathsf{E}_{\mathsf{P}_\rho}\left[\mathsf{E}_{\mathsf{P}_{X_\tau}}H\cdot\mathbf{1}_{B\cap\{\tau<\infty\}}\right].$$

It is sufficient to check

$$\mathsf{E}_{\mathsf{P}_\rho}\left[H\circ\theta^\tau\cdot\mathbf{1}_{B\cap\{\tau=n\}}\right]=\mathsf{E}_{\mathsf{P}_\rho}\left[\mathsf{E}_{\mathsf{P}_{X_\tau}}H\cdot\mathbf{1}_{B\cap\{\tau=n\}}\right],\quad n\geq 0,$$

or, equivalently,

$$\mathsf{E}_{\mathsf{P}_\rho}\left[H\circ\theta^n\cdot\mathbf{1}_{B\cap\{\tau=n\}}\right]=\mathsf{E}_{\mathsf{P}_\rho}\left[\mathsf{E}_{\mathsf{P}_{X_n}}H\cdot\mathbf{1}_{B\cap\{\tau=n\}}\right],\quad n\geq 0.$$

Since $B\cap\{\tau=n\}\in\mathcal{F}_n$, the last identity is exactly the content of Theorem 5.6. □

CHAPTER 6

Markov Processes on Finite State Spaces

This is not quite proofread. Proceed with caution.

1. Approach based on the abstract theory

In this section we apply the powerful results we have obtained to a relatively simple case where $\mathbb{X}$ is finite.

We fix $N \in \mathbb{N}$ and assume that $\mathbb{X} = \{1, \dots, N\}$, and $\mathcal{X} = 2^{\mathbb{X}}$. Every probability measure ρ on $(\mathbb{X}, \mathcal{X})$ in this section will be identified with a probability vector $(\rho_i)_{i \in \mathbb{X}}$ with $\rho_i = \rho\{i\}$. Every Markov kernel P on $(\mathbb{X}, \mathcal{X})$ will be identified with a transition matrix $(P_{ij})_{i,j=1}^N$ defined by $P_{ij} = P(i, \{j\})$. Let us study the set of all invariant distributions ρ with respect to P.

Of course, the problem is equivalent to finding all solutions of

$$\rho P = \rho \tag{6.1}$$

satisfying $\rho \in \Delta_N$, where

$$\Delta_N = \left\{ p \in \mathbb{R}^N : \ p_1 + \ldots + p_n = 1, \text{ and } p_i \geq 0 \text{ for all } i \in \{1, \dots, N\} \right\}$$

The analysis can be performed with the help of the Perron–Frobenius theorem, but let us use the theory of ergodic decomposition instead.

The simplex Δ_N is compact and convex. The vector subspace defined by (6.1) is also convex. Therefore, the intersection is also compact and convex (we will shortly see that it is non-empty) and can be seen as the convex hull of its extreme points. We know from the abstract ergodic decomposition that these extreme points are ergodic distributions. So let us establish several useful facts.

We say that a set $A \subset \mathbb{X}$ is absorbing with repect to P if $P(i, A) = 1$ for *all* $i \in A$.

LEMMA 6.1. *If $A \subset \mathbb{X}$ is absorbing with respect to P, then there is a P-invariant distribution ρ satisfying $\rho(A) = 1$.*

PROOF: Let us use the Krylov–Bogolyubov approach. Let us take any initial state $i \in A$ and consider the initial distribution δ_i concentrated at i. Consider a sequence of measures (or probability vectors)

$$\rho^n = \frac{1}{n} \sum_{k=0}^{n-1} \delta_i P^k, \quad n \in \mathbb{N}.$$

Since $\rho^n \in \Delta_N$, there is an increasing sequence (n_k) and a vector $\rho \in \Delta_N$ such that $\rho^{n_k} \to \rho$. Now ρ satisfies (6.1) since

$$\rho^{n_k} P - \rho^{n_k} = \frac{1}{n_k}(\delta_i P^{n_k} - \delta_i) \to 0, \quad k \to \infty.$$

It is also clear that due to the absorbing property of A, $\rho^n(A) = 1$ and therefore $\rho(A) = 1$. □

LEMMA 6.2. *For every $i \in \mathbb{X}$, there is at most one ergodic measure ρ such that $\rho_i > 0$.*

PROOF: Had there been two such ergodic measures, they would not be mutually singular. □

For $i, j \in \mathbb{X}$, we write $i \to j$ if there is $n \in \mathbb{N}$ and $i_1, \dots, i_n$ such that $P_{ii_1} P_{i_1 i_2} \dots P_{i_{n-1} i_n} P_{i_n j} > 0$. If $i \to j$ and $j \to i$ we write $i \leftrightarrow j$. This is an equivalence relation and we call any equivalence class a class of communicating states.

LEMMA 6.3. *If ρ is invariant, $\rho_i > 0$, $i \to j$, then $\rho_j > 0$.*

PROOF: Let $P_{ii_1} P_{i_1 i_2} \dots P_{i_{n-1} i_n} P_{i_n j} > 0$. Then

$$\rho_j = \mathsf{P}_\rho\{X_{n+1} = j\} \geq \rho_i P_{ii_1} P_{i_1 i_2} \dots P_{i_{n-1} i_n} P_{i_n j} > 0.$$

□

LEMMA 6.4. *For every class A of communicating states there is at most one ergodic measure ρ such that $\rho(A) > 0$.*

PROOF: The previous lemma implies that every invariant measure positive on A assigns positive weights to all elements of i. The desired statement follows now from Lemma 6.2. □

LEMMA 6.5. *If a set A is an absorbing classs of communicating states, then there is exactly one ergodic measure ρ satisfying $\rho(A) > 0$. This measure ρ satisfies $\rho(A) = 1$.*

PROOF: Existence follows from Lemma 6.1. Uniqueness follows from Lemma 6.4. Ergodicity, the absorbing property of A, and $\rho(A) > 0$ imply $\rho(A) = 1$. □

LEMMA 6.6. *If ρ is P-invariant, then the set $A = \{i \in \mathbb{X} : \rho_i > 0\}$ is an absorbing class. If in addition ρ is ergodic, then A is a class of communicating states.*

PROOF: If A is not absorbing, there are $i \in A$ and $j \in A^c$ such that $P_{ij} > 0$. Then $0 = \rho_j \geq \rho_i P_{ij} > 0$, a contradiction.

Suppose there are states $i, j \in A$ such that $i \not\to j$. The set $B = \{k \in \mathbb{X} : i \to k\}$ is an absorbing set not containing j. We have $\rho(B) \geq \rho_i > 0$ and $\rho(B) \leq 1 - \rho_j < 1$. This contradicts the ergodicity of ρ. □

LEMMA 6.7. *Let $i \in \mathbb{X}$ and $\mathsf{P}_i(B) > 0$, where*

$$B = \{X_n \neq i \text{ for all } n \in \mathbb{N}\}.$$

Then $\rho_i = 0$ for every invariant measure ρ.

PROOF: Due to the ergodic decomposition, it is sufficient to prove that for every ergodic ρ, $\rho_i = 0$. Suppose $\rho_i > 0$. Due to the ergodic theorem, $\mathsf{P}_\rho(A) = 1$, where

$$A = \left\{ \frac{1}{n} \sum_{k=0}^{n-1} \mathbf{1}_{\{X_k = i\}} \to \rho_i \right\}.$$

Since

$$1 = \mathsf{P}_\rho(A) = \int_{\mathbb{X}} \rho(dj) \mathsf{P}_j(A) = \sum_{j \in \mathbb{X}} \rho_j \mathsf{P}_j(A),$$

we conclude that $\mathsf{P}_i(A) = 1$. Clearly, $A \cap B = \emptyset$, so $\mathsf{P}_i(A) = 1$ and $\mathsf{P}_i(B) > 0$ contradict each other. □

Summarizing the results above we obtain the following theorem.

THEOREM 6.1. *Let $i \in \mathbb{X}$. If there is j such that $i \to j$ and $j \not\to i$, then there is no invariant distribution ρ satisfying $\rho_i > 0$. If there is no such state j, then there is a unique ergodic distribution ρ satisfying $\rho_i > 0$. The set $A = \{j \in \mathbb{X} : \ \rho_j > 0\}$ coincides with $B = \{j \in \mathbb{X} : i \to j\}$.*

PROOF: The first part follows from Lemma 6.7. To prove the second part, we note that the set B is an absorbing class of communicating states. Lemma 6.5 implies that there is a unique ergodic measure ρ supported by B. Lemma 6.3 implies $\rho_j > 0$ for every $j \in B$. □

If all states form one class, the kernel P is called irreducible or ergodic. Then there is exactly one invariant distribution. This situation is often called *unique ergodicity*.

Often, to analyze the invariant distribution ρ there is no better way the to solve equation (6.1). However, there are useful representations for the invariant distributions.

One such representation is the following, taken from [**FW12**]. We think of $\mathbb{X}$ as the complete directed graph where each edge (ij) is assigned a weight P_{ij}. To each state i we associate a collection G_i of directed subgraphs g with the following properties: there is no arrow coming out of i, for each $j \neq i$, there is exactly one arrow coming out of j, and there are no cycles. For every collection g of arrows we define

$$\pi(g) = \prod_{(jk) \in g} P_{jk}.$$

For every $i \in \mathbb{X}$, we define $Q_i = \sum_{g \in G_i} \pi(g)$. We also set

$$q_i = \frac{Q_i}{\sum_{j \in \mathbb{X}} Q_j}. \tag{6.2}$$

THEOREM 6.2. *Suppose q is defined in* (6.2). *Then $qP = q$.*

PROOF: It is sufficient to check

$$\sum_{i\in\mathbb{X}} Q_i P_{ij} = Q_j,$$

or, moving $Q_j P_{jj}$ to the right-hand side,

$$\sum_{i\neq j} Q_i P_{ij} = Q_j(1 - P_{jj}) = \sum_{i\neq j} Q_j P_{ji}.$$

It remains to check that both sides are equal to $\sum_{g\in H} \pi(g)$, where H consists of all graphs with the following properties: (i) for every vertex, there is exactly one outgoing arrow, (ii) there is exactly one cycle, (iii) this cycle contains j. □

Often, the following property is useful. We say that the transition kernel P is reversible with respect to a measure ρ if the following condition of detailed balance holds:

$$\rho_i P_{ij} = \rho_j P_{ji}, \quad i, j \in \mathbb{X}. \tag{6.3}$$

This condition automatically implies invariance of ρ since

$$\sum_i \rho_i P_{ij} = \sum_i \rho_j P_{ji} = \rho_j \sum_i P_{ji} = \rho_j, \quad j \in \mathbb{X}.$$

Reversibility, in fact, means that the time-reversed process has the same distribution. To compute the transition probability of the time-reversed process, we write

$$\begin{aligned}
\mathsf{P}_\rho(X_n = i | X_{n+1} = j) &= \frac{\mathsf{P}_\rho\{X_n = i, X_{n+1} = j\}}{\mathsf{P}_\rho\{X_{n+1} = j\}} \\
&= \frac{\mathsf{P}_\rho\{X_n = i\}\mathsf{P}_\rho\{X_{n+1} = j | X_n = i\}}{\mathsf{P}_\rho\{X_{n+1} = j\}} \\
&= \frac{\rho_i P_{ij}}{\rho_j} = \frac{\rho_j P_{ji}}{\rho_j} = P_{ji}.
\end{aligned}$$

One example of a time-reversible Markov chain is the random walk on an undirected graph. Here $P(i, \cdot)$ is the uniform distribution on the neighbors of i. It is easy to see that if one defines $\rho_i = \deg(i)/Z$ where $\deg(i)$ denotes the number of neighbors of i in the graph, and Z is a normalization constant, then the detailed balance holds. In particular, ergodic theorem implies that the average time the random walk spends at a vertex is proportional to the degree of the vertex and does not depend on any other geometric features of the graph.

To check if the detailed balance holds one can start with fixing ρ_i for some i and then sequentially assign weights ρ_j to other vertices using (6.3) hoping that the assigned values will be self-consiistent. This is essentially the idea behind the following Kolmogorov's reversibility criterion:

THEOREM 6.3. *Suppose (P_{ij}) is an irreducible kernel. Then it is reversible with respect to some measure ρ iff for every n and every sequence of vertices $i_1, i_2, \dots, i_n$,*

$$P_{i_1 i_2} P_{i_2 i_3} \dots P_{i_{n-1} i_n} P_{i_n i_1} = P_{i_1 i_n} P_{i_n i_{n-1}} \dots P_{i_2 i_1} \tag{6.4}$$

PROOF: To derive (6.4) from reversibility, it is sufficient to write (6.3) for i_1, i_2, for i_2, i_3, etc, take the product of these identities and cancel $\prod_k \rho_{i_k}$ on both sides.

If we assume that (6.4) holds true for all sequences of states, then we can take an arbitrary $i_0 \in \mathbb{X}$ and for every i define

$$\rho_i = \frac{1}{Z} \frac{P_{i_0 i_1} P_{i_1 i_2} \dots P_{i_{n-1} i_n} P_{i_n i}}{P_{i i_n} P_{i_n i_{n-1}} \dots P_{i_2 i_1} P_{i_1 i_0}},$$

where the sequence $(i_0, i_1, \dots, i_n)$ is chosen so that $P_{i i_n} P_{i_n i_{n-1}} \dots P_{i_2 i_1} P_{i_1 i_0} > 0$, and the normalization constant Z independent of i is to be chosen later.

Condition (6.4) implies that this definition does not depend on the choice of a specific sequence $(i_0, i_1, \dots, i_n)$: If $(i_0, i'_1, \dots, i'_n)$ is another such sequence, we obtain

$$\frac{P_{i_0 i_1} P_{i_1 i_2} \dots P_{i_{n-1} i_n} P_{i_n i}}{P_{i i_n} P_{i_n i_{n-1}} \dots P_{i_2 i_1} P_{i_1 i_0}} = \frac{P_{i_0 i'_1} P_{i'_1 i'_2} \dots P_{i'_{n-1} i'_n} P_{i'_n i}}{P_{i i'_n} P_{i'_n i'_{n-1}} \dots P_{i'_2 i'_1} P_{i'_1 i'_0}}.$$

Also, it is clear from (6.4) that existence of a cycle realizing $i_0 \to i \to i_0$ implies that $\rho_i > 0$ for all i.

Similarly, we see that if $P_{ij} > 0$, then $P_{ji} > 0$, so to check (6.3), we write

$$\rho_i P_{ij} = \frac{1}{Z} \frac{P_{i_0 i_1} P_{i_1 i_2} \dots P_{i_{n-1} i_n} P_{i_n i} \cdot P_{ij}}{P_{i i_n} P_{i_n i_{n-1}} \dots P_{i_2 i_1} P_{i_1 i_0}} \cdot \frac{P_{ji}}{P_{ji}} = \frac{1}{Z} \rho_j P_{ji}.$$

□

In applications, one can use the idea of reversibility to construct Markov chains with a given invariant distribution ρ.

Often in systems with statistical mechanics flavor, there is no easy access to the values ρ_i themselves, but the ratios ρ_i/ρ_j are easy to compute. This happens, for example, for Boltzmann–Gibbs distributions where

$$\rho_i = \frac{e^{-\beta H(i)}}{Z}, \quad i \in \mathbb{X}.$$

Here $H : \mathbb{X} \to \mathbb{R}$ is the energy function and $\beta \geq 0$ is the inverse temperature. Such distributions for various choices of H are ubiquitous in statistical mechanics. Let us fix some H. Some characteristic features of such a family of distributions can be seen for limiting cases $\beta \to 0$ and $\beta \to \infty$. When $\beta = 0$ and the temperature is infinite, then all states are equally probable, and this can be interpreted as a complete disorder. On the other hand, if β is large (the temperature is small), the states with minimal energy ("ground" states) are much more probable than other states, and as $\beta \to \infty$, the system "freezes" to the ground states.

When $\mathbb{X}$ is a large set, computing the values of Z and ρ_i is not easy. However, computing $\rho_i/\rho_j = e^{\beta(H(j)-H(i))}$ is immediate. Let us use this to construct a ρ-reversible Markov chain known as Metropolis–Hastings algorithm. Let us assume for simplicity that $\mathbb{X}$ is endowed with a graph structure and each state is connected to exactly d other states. We denote by E the set of edges in this graph. Then let us choose

$$P_{ij} = \begin{cases} 0, & \{i,j\} \notin E, \\ \frac{1}{d}(1 \wedge \frac{\rho_j}{\rho_i}), & \{i,j\} \in E, \\ 1 - \sum_{j:\{i,j\}\in E} \frac{1}{d}(1 \wedge \frac{\rho_j}{\rho_i}), & j = i. \end{cases}$$

In words, if the system is at state i, it chooses j among the neighboring d states uniformly at random (with probability $1/d$) and then, if $\rho_j \geq \rho_i$, jumps to j. If $\rho_j < \rho_i$, then the jump to j gets approved only with probability ρ_j/ρ_i. In the case where it does not get approved, the system stays in the state i.

To check the detailed balance it is sufficient to note that

$$\frac{P_{ij}}{P_{ji}} = \frac{1 \wedge \frac{\rho_j}{\rho_i}}{1 \wedge \frac{\rho_i}{\rho_j}} = \frac{\rho_j}{\rho_i}, \quad \{i,j\} \in E,$$

irrespective of whether $\rho_j \geq \rho_i$ or vice versa.

2. Perron–Frobenius Theorem

One can also treat ergodic results on Markov kernels on finite state spaces as consequences of the classical Perron–Frobenius theorem to stochastic matrices (i.e., matrices with nonnegative values with sum of elements in each row equal to 1). For completeness, we give several proofs of this theorem or related statements.

There are several statements that pass under this title. Here we follow the exposition in [**KH95**].

THEOREM 6.4. *Let $d \in \mathbb{N}$ and suppose A is a $d \times d$ matrix with nonnegative entries such that for some $n \in \mathbb{N}$ all entries of A^n are positive. Then there is an eigenvector x_0 with all positive components. If x is an eigenvector with nonnegative components, then $x = cx_0$ for some $c > 0$. The eigenvalue λ associated with x_0 is simple and all other eigenvalues are less than λ in absolute value.*

PROOF: For $x \in \mathbb{R}^d$ we will write $|x|_1 = \sum_{i=1}^d |x_i|$. Let us introduce $\mathcal{C} = \{x \in \mathbb{R}^N : x_i \geq 0, i = 1, \dots, d\}$ and $\Delta = \{x \in \mathcal{C} : \sum_{i=1}^d x_i = 1\}$. The simplex Δ is the convex hull of its extreme points $e_1 = (1, 0, 0, \dots,), e_2 = (0, 1, 0, \dots,), \dots, e_d = (0, 0, \dots, 0, 1)$.

For every $x \in \Delta$ we introduce $Tx = Ax/|Ax|_1 \in \Delta$.

Let us study some properties of $\Delta_0 = \bigcap_{k=1}^\infty T^k\Delta$. Images of convex closed sets under T are convex closed sets, so all sets $T^k\Delta$, $k \in \mathbb{N}$ are convex

and closed, and so is their intersection Δ_0. Since $T^{k+1}\Delta \subset T^k\Delta$ for all $k \in \mathbb{N}$, we obtain $\Delta_0 \neq \emptyset$.

The conditions of the theorem imply that

$$T^m\Delta \subset \operatorname{Int}\Delta, \quad m \geq n, \tag{6.5}$$

where $\operatorname{Int}\Delta$ denotes the relative interior of Δ. Therefore, $\Delta_0 \subset \operatorname{Int}\Delta$.

Let us show that Δ_0 has at most d extreme points. Namely, let us choose an increasing sequence $(n_k)_{k\in\mathbb{N}}$ such that $\lim_{k\to\infty} T^{n_k}e_i = x_i$ for some $x_i \in \Delta_0$, $i = 1, \dots, d$, and prove that all $\operatorname{ext}\Delta_0 \subset \{x_1, \dots, x_d\}$.

If $x \in \Delta_0$ then $x \in T^{n_k}\Delta$ for all $k \in \mathbb{N}$ and therefore x is a convex combination of extreme points of $T^{n_k}\Delta$. All these are images of extreme points of Δ under T^{n_k}, so we obtain a representation $x = \sum_{i=1}^d \alpha_i^{(k)} T^{n_k}e_i$ for some $\alpha_i^{(k)} \geq 0$ satisfying $\sum_{i=1}^d \alpha_i^{(k)} = 1$, $k \in \mathbb{N}$. Choosing an increasing sequence $(k(m))_{m\in\mathbb{N}}$ such that $\lim_{m\to\infty} \alpha_i^{(n_{k(m)})} = \alpha_i$ for some α_i, we obtain $x = \sum_{i=1}^d \alpha_i x_i$, so if x is an extreme point it has to coincide with one of x_i, $i = 1, \dots, d$.

Since $\Delta_0 = T\Delta_0$ and $\operatorname{ext}\Delta_0$ is finite, we obtain $\operatorname{ext}\Delta_0 = T\operatorname{ext}\Delta_0$, in other words T acts a permutation on $\operatorname{ext}\Delta_0$. Therefore there is a number $m \in \mathbb{N}$ such that $T^m x = x$ for every $x \in \operatorname{ext}\Delta_0$. So, every $x \in \operatorname{ext}\Delta_0$ is an eigenvector of A^m with a positive eigenvalue.

Let us prove that, in fact, there cannot be two distinct eigenvectors with those properties. There are two cases to exclude: (i) there are $x, y \in \Delta_0$ and $\nu > 0$ such that $x \neq y$ $A^m x = \nu x$ and $A^m y = \nu y$; (ii) there are $x, y \in \Delta_0$ and $\nu > \mu > 0$ such that and $A^m x = \nu x$ and $A^m y = \mu y$.

In case (i), there is a number α such that $z = x + \alpha(y - x) \in \partial\Delta$. Therefore $A^{mn}z = \nu^n z \in \partial\mathcal{C}$, so $T^{mn}z \in \partial\Delta$ which contradicts $T^{mn}\Delta \subset \operatorname{Int}\Delta$.

In case (ii), we can use $y \in \operatorname{Int}\Delta$ to choose $\varepsilon > 0$ small enough to ensure that $z = y - \varepsilon x \in \operatorname{Int}\mathcal{C}$. Then $A^{km}z = \mu^k y - \varepsilon\nu^k x$, so for sufficiently large k, $A^{km}z \notin \mathcal{C}$ which is a contradiction.

We conclude that Δ_0 contains a unique point x_0. Recalling that Δ_0 is T-invariant, we obtain that $Tx_0 = x_0$, i.e., $Ax_0 = \lambda x_0$ for some $\lambda > 0$.

Let us show that if $Ax = \mu x$ for any $x \in \mathcal{C}$ that is not a multiple of x_0 and satisfies $x \neq 0$, then $|\mu| < \lambda$.

If $\mu = \pm\lambda$, then we can find $\alpha \in \mathbb{R}$ such that $z = x_0 + \alpha x \in \partial\mathcal{C} \setminus \{0\}$. Then $A^{2k}z = \lambda^{2k}z$ for all $k \in \mathbb{N}$ which contradicts (6.5).

If $\mu \in \mathbb{R}$ and $|\mu| > \lambda$, then we have already proved that $x \notin \mathcal{C}$. Also we can find $\varepsilon > 0$ small enough to guarantee that $z = x_0 + \varepsilon x \in \operatorname{Int}\mathcal{C}$. Then for large values of m, the direction of $A^{2m+1}z$ will be close to the direction of $\pm x$, so we will have $A^{2m+1}z \notin \mathcal{C}$ which contradicts A-invariance of $\mathcal{C}$.

In general, if $\mu = \rho e^{i2\pi\phi}$ for some $\rho > 0$ and $\phi \in \mathbb{R}$, then there is a plane L such that the action of A on L is multiplication by ρ and rotation by angle ϕ.

If $\phi = k/l$ for some $k \in \mathbb{Z}$ and $l \in \mathbb{N}$, then A^l acts on L as multiplication by ρ^l, and we can apply the above analysis to A^l and conclude that $\rho < \mu$.

If ϕ is irrational, then for any $\delta > 0$, the direction of $A^k x$ will get δ-close to that of x for an infinitely many values of k. Let us denote the set of those values of k by K. So we choose $x \in L \setminus \mathcal{C}$ and ε small enough to ensure $z = x_0 + \varepsilon x \in \operatorname{Int} \mathcal{C}$. If $\rho > \lambda$, then for sufficiently large $k \in K$, the direction of $A^k z$ will be close to that of x, and so $A^k z \notin \mathcal{C}$ for those values of k, a contradiction. If $\rho = \lambda$, then we choose $y \in L$ so that $z = \frac{x_0 + \alpha y}{|x_0 + \alpha y|_1} \in \partial\Delta$. Since ϕ is irrational,

$$T^k z = \frac{x_0 + e^{i2\pi\phi}\alpha y}{|x_0 + e^{i2\pi\phi}\alpha y|_1}$$

will get arbitrarily close to $\partial\Delta$ for infinitely many values of k which contradicts (6.5).

To finish the proof that λ is simple, we must prove that there are no nontrivial Jordan blocks for matrix A. Suppose there is such a block. Then there is a vector x_1 such that $Ax_1 = \lambda x_1 + x_0$. Let $y = \lambda x_1 + x_0$. Then $Ay = \lambda(y + x_0)$. Now we can choose ε such that $z = x_0 - \varepsilon y \in \operatorname{Int} \mathcal{C}$, and compute by induction

$$A^n z = \lambda^n((1 - n\varepsilon)x_0 - \varepsilon y), \quad n \geq 0.$$

We see that for sufficiently large n, $A^n z \notin \mathcal{C}$ which is a contradiction. □

3. Hilbert projective metric approach.

Let us give another proof of the Perron–Frobenius theorem based on contraction in the so called Hilbert projective metric. We follow exposition in [**Bal00**] that, in turn, follows [**Fur60**]

For two points $x, y \in \mathbb{R}^d$ we will write $x \leq y$ and $y \geq x$ if $y - x \in \mathcal{C}$. We write $x \sim y$ if $x = ry$ for some $r > 0$. Let us denote $\mathcal{C}^* = \mathcal{C} \setminus \{0\}$ and introduce

$$\alpha(x, y) = \sup\{r \in \mathbb{R}_+ : rx \leq y\}, \quad x, y \in \mathcal{C}^*.$$

We note that $\alpha(x, y) \in [0, \infty)$ for all $x, y \in \mathcal{C}$ and $\alpha(x, x) = 1$ for all $x \in \mathcal{C}$. Also,

$$\alpha(x, z) \geq \alpha(x, y)\alpha(y, z), \quad x, y, z \in \mathcal{C}^*, \tag{6.6}$$

since $z \geq \alpha(y, z)y$ and $y \geq \alpha(x, y)x$. Next we define $\Gamma(x, y) = \alpha(x, y)\alpha(y, x)$. Inequality (6.6) implies

$$\Gamma(x, z) \geq \Gamma(x, y)\Gamma(y, z), \quad x, y, z \in \mathcal{C}^*. \tag{6.7}$$

Applying this to $z = x$, and noticing that $\Gamma(x, x) = 1$, we obtain

$$0 \leq \Gamma(x, y) \leq 1, \quad x, y \in \mathcal{C}^*.$$

Moreover, $\Gamma(x, y) = 1$ iff $x \sim y$. One part of this statement is obvious since for every $r > 0$, $\alpha(x, rx) = r$. To see the converse implication, we notice that $\alpha(x, y)\alpha(y, x) = 1$ implies that there is $r > 0$ such that $x \geq ry$ and

$y \geq r^{-1}x$. The latter is equivalent to $ry \geq x$, so these two inequalities can hold together only if $x = ry$.

We can now define the Hilbert projective metric by

$$\Theta(x, y) = -\ln \Gamma(x, y) = -\ln \alpha(x, y) - \ln \alpha(y, x) \in [0, \infty].$$

Notice that $\Theta(x, y)$ is infinite if $x, y \in \partial\mathcal{C}$ and $x \not\sim y$. The following triangle inequality follows from (6.7):

$$\Theta(x, z) \leq \Theta(x, y) + \Theta(y, z), \quad x, y, z \in \mathcal{C}^*.$$

From the properties of Γ, we derive that $\Theta(x, y) = \Theta(x, ry)$ for any $x, y \in \mathcal{C}^*$ and $r > 0$, and $\Theta(x, y) = 0$ iff $x \sim y$, so Θ is a pseudo-metric on $\mathcal{C}^*$ and a metric on $\Delta = \mathcal{C}^*/\sim$.

Other ways to represent Θ are

$$\Theta(x, y) = -\ln\left(\min_i \frac{x_i}{y_i} \cdot \min_j \frac{y_j}{x_j}\right) = \ln\left(\max_i \frac{y_i}{x_i} \cdot \max_j \frac{x_j}{y_j}\right) = \ln \max_{i,j} \frac{x_j y_i}{y_j x_i}.$$

Let us prove that the map $T : \Delta \to \Delta$ is nonexpanding. First, we notice that $x - ry \in \mathcal{C}$ implies $A(x - ry) = Ax - rAy \in \mathcal{C}$. Therefore $\alpha(Ax, Ay) \geq \alpha(x, y)$, so $\Theta(Ax, Ay) \leq \Theta(x, y)$.

Assuming now that all the entries of A are positive, let us prove that the contraction coefficient is actually strictly less than 1. First we recall that in this case $T\Delta$ is a compact subset of $\operatorname{Int}\Delta$. Therefore, expressions $\ln(x_j y_i/(y_j x_i))$ are uniformly bounded with respect to $x, y \in \Delta$, and

$$D = \operatorname{diam}(T\Delta) = \sup\{\Theta(x, y) : x, y \in \Delta\} < \infty,$$

or, equivalently

$$\delta = \inf\{\Gamma(x, y) : x, y \in \mathcal{C}^*\} > 0.$$

These numbers are related by $\delta = e^{-D}$. The following theorem establishes the strict contraction property of θ.

THEOREM 6.5. *If A has all positive entries, then for all $x, y \in \Delta$,*

$$\Theta(Tx, Ty) \leq \frac{1 - \sqrt{\delta}}{1 + \sqrt{\delta}} \Theta(x, y). \tag{6.8}$$

PROOF: Let us take two distinct points $x, y \in \Delta$. We can assume that $\Theta(x, y) < \infty$, i.e., both numbers $\alpha_1 = \alpha(x, y)$ and $\alpha_2 = \alpha(y, x)$ are nonzero. Then $y - \alpha_1 x \in \mathcal{C}^*$ and $x/\alpha_2 - y \in \mathcal{C}^*$. Applying the definition of δ to those vectors, we obtain that there are two numbers $\lambda_1, \lambda_2 \geq 0$ such that $\lambda_1\lambda_2 \geq \delta$ and

$$A(x/\alpha_2 - y) \geq \lambda_1 A(y - \alpha_1 x),$$
$$A(y - \alpha_1 x) \geq \lambda_2 A(x/\alpha_2 - y),$$

i.e.,

$$Ax \geq \frac{1 + \lambda_1}{\frac{1}{\alpha_2} + \lambda_1\alpha_1} Ay, \quad Ay \geq \frac{\alpha_1 + \frac{\lambda_2}{\alpha_2}}{1 + \lambda_2} Ax.$$

So,

$$\Theta(Tx,Ty) = \Theta(Ax,Ay) \le -\ln \frac{(1+\lambda_1)(\alpha_1+\frac{\lambda_2}{\alpha_2})}{(\frac{1}{\alpha_2}+\lambda_1\alpha_1)(1+\lambda_2)}$$

$$= -\left(\ln \frac{\frac{1}{\alpha_1\alpha_2}+\frac{1}{\lambda_2}}{\frac{1}{\alpha_1\alpha_2}+\lambda_1} - \ln\frac{1+\frac{1}{\lambda_2}}{1+\lambda_1}\right)$$

$$= \int_0^{1/(\alpha_1\alpha_2)} \frac{\frac{1}{\lambda_2}-\lambda_1}{(t+\frac{1}{\lambda_2})(t+\lambda_1)}dt = \int_0^{\ln(1/(\alpha_1\alpha_2))} \frac{(\frac{1}{\lambda_2}-\lambda_1)e^s}{(e^s+\frac{1}{\lambda_2})(e^s+\lambda_1)}ds$$

$$= \left(\frac{1}{\lambda_1\lambda_2}-1\right)\int_0^{\ln(1/(\alpha_1\alpha_2))} \frac{\lambda_1 e^s}{(e^s+\frac{1}{\lambda_2})(e^s+\lambda_1)}ds$$

Elementary analysis shows that the maximum of the expression $\lambda_1 t/((t+\frac{1}{\lambda_2})(t+\lambda_1))$ over $t \ge 0$ is attained at $t = \sqrt{\lambda_1/\lambda_2}$ and equals $(1+1/\sqrt{\lambda_1\lambda_2})^{-2}$. Therefore,

$$\Theta(Tx,Ty) \le \ln \frac{1}{\alpha_1\alpha_2}\frac{\frac{1}{\lambda_1\lambda_2}-1}{(\frac{1}{\sqrt{\lambda_1\lambda_2}}+1)^2} = \frac{\frac{1}{\sqrt{\lambda_1\lambda_2}}-1}{\frac{1}{\sqrt{\lambda_1\lambda_2}}+1}\Theta(x,y) = \frac{1-\sqrt{\lambda_1\lambda_2}}{1+\sqrt{\lambda_1\lambda_2}}\Theta(x,y).$$

The function $t \mapsto (1-t)/(1+t)$ decays for $t \ge 0$, so the last expression is bounded by the right-hand side of (6.8). □

This strict contraction property implies existence and uniqueness of a fixed point of T on Δ, along with exponential convergence of iterations of $T^n\Delta$ to that fixed point. If the requirement of positivity of all entries is satisfied not for A but for A^m for some $m \in \mathbb{N}$, then the above reasoning still can be applied to A^m, since A itself is non-expanding.

CHAPTER 7

Countable state space Markov chains

1. Strong Markov property with respect to state hitting times

A very useful way to look at Markov chains is to study return times to a state. In this section we assume that $\mathbb{X} = \{1, \dots, N\}$ or $\mathbb{X} = \mathbb{N}$ (in the latter case we set $N = \infty$).

Let us recall the definition (5.27). For an arbitrary $i \in \mathbb{X}$, we define $\tau_i = \tau_{\{i\}}$ and note that $\mathsf{P}_{X_{\tau_i}} = \mathsf{P}_i$ on $\{\tau_i < \infty\}$.

THEOREM 7.1. *Let $\mathsf{P}_\rho\{\tau_i < \infty\} = 1$. Then for any bounded random variable $H : \mathbb{X}^{\mathbb{Z}_+} \to \mathbb{R}$, and every set $B \in \mathcal{F}_{\tau_i}$, i.e.,*

$$\mathsf{E}_{\mathsf{P}_\rho}[H \circ \theta^{\tau_i} \cdot \mathbf{1}_B] = \mathsf{E}_{\mathsf{P}_\rho}[H \circ \theta^{\tau_i}]\mathsf{P}_\rho(B). \tag{7.1}$$

PROOF: For brevity we denote $\tau = \tau_i$ The strong Markov property implies that

$$\begin{aligned}
\mathsf{E}_{\mathsf{P}_\rho}[H \circ \theta^\tau \cdot \mathbf{1}_B] &= \mathsf{E}_{\mathsf{P}_\rho}[\mathsf{E}_{\mathsf{P}_\rho}[H \circ \theta^\tau \cdot \mathbf{1}_B | \mathcal{F}_\tau]] \\
&= \mathsf{E}_{\mathsf{P}_\rho}[\mathbf{1}_B \mathsf{E}_{\mathsf{P}_\rho}[H \circ \theta^\tau | \mathcal{F}_\tau]] \\
&= \mathsf{E}_{\mathsf{P}_\rho}[\mathbf{1}_B \mathsf{E}_{\mathsf{P}_{X_\tau}} H] \\
&= \mathsf{E}_{\mathsf{P}_\rho}[\mathbf{1}_B \mathsf{E}_{\mathsf{P}_i} H] \\
&= \mathsf{P}_\rho(B) \mathsf{E}_{\mathsf{P}_i} H
\end{aligned}$$

This identity with $B = \mathbb{X}^{\mathbb{Z}_+}$ gives

$$\mathsf{E}_{\mathsf{P}_\rho}[H \circ \theta^\tau] = \mathsf{E}_{\mathsf{P}_i} H.$$

The last two identities together imply(7.1). □

Theorem 7.1 means that the future and the past of a Markov process with respect to a visit to a state i are independent of each other. Iterating this argument, we obtain that the realization of the Markov process can be split in *excursions*, each excursion being a path between two consecutive visits to i.

Of course, this is loosely formulated. I hope to add more to this later.

2. Invariant measures and classification of states

Some notions and results for countable state Markov chains can be adapted from the finite state space situation. In particular, Lemmas 6.2–6.7 still hold true for countable $\mathbb{X}$. However, Theorem 6.1 does not hold true for countable state space. The main difference is the following. The only

situation where a state $i \in \mathbb{X}$ cannot belong to the support of any invariant distribution on a finite $\mathbb{X}$ is described by the conditions of Lemma 6.7. This statement fails to be true for a general Markov chain on a countable $\mathbb{X}$.

The central topic in the study of Markov processes on a countable state space is *recurrence*. Let us introduce the corresponding notions.

We recall that $\tau_i = \min\{k \in \mathbb{N} : X_k = i\}$ for $i \in \mathbb{X}$.

DEFINITION 7.1. A state $i \in \mathbb{X}$ is *transient* if $\mathsf{P}_i\{\tau_i < \infty\} < 1$.

DEFINITION 7.2. A state $i \in \mathbb{X}$ is *recurrent* if $\mathsf{P}_i\{\tau_i < \infty\} = 1$.

DEFINITION 7.3. A recurrent state $i \in \mathbb{X}$ is *positive recurrent* if $\mathsf{E}_i\tau_i < \infty$.

DEFINITION 7.4. A recurrent state $i \in \mathbb{X}$ is null-recurrent if $\mathsf{E}_i\tau_i < \infty$.

THEOREM 7.2. *In every communication class, all states are of the same type.*

PROOF: To be inserted □

LEMMA 7.1. *If $i \in \mathbb{X}$ is transient, then for every ergodic invariant distribution ρ, $\rho_i = 0$.*

PROOF: This lemma is simply a version of Lemma 6.7. □

THEOREM 7.3. *If $h \in \mathbb{X}$ is positive recurrent, then there is a unique invariant distribution ρ such that $\rho_h > 0$. This invariant measure ρ is ergodic.*

PROOF: Let $\pi_i, i \in \mathbb{X}$ be the average time spent in i by the process started at h before coming back to h:

$$\pi_i = \mathsf{E}_h \sum_{k=0}^{\infty} \mathbf{1}_{\{X_k=i, k<\tau_h\}} = \sum_{k=0}^{\infty} \mathsf{P}_h\{X_k = i, k < \tau_h\} \tag{7.2}$$

Note that according to this definition, $\pi_h = 1$. Note also that

$$\sum_{i\in\mathbb{X}} \pi_i = \mathsf{E}_h \sum_{i\in\mathbb{X}} \sum_{k=0}^{\infty} \mathbf{1}_{\{X_k=i, k<\tau_h\}} = \mathsf{E} \sum_{k=0}^{\infty} \mathbf{1}_{\{k<\tau_h\}} = \mathsf{E}_h\tau_h < \infty,$$

so the measure π is finite, and its normalized version $\rho_i = \pi_i/\mathsf{E}_h\tau_h$ is a probability measure. Let us check that $\pi P = \pi$. We can write

$$\sum_{j\in\mathbb{X}} \pi_j P_{ji} = \sum_{j\in\mathbb{X}} \sum_{k=0}^{\infty} \mathsf{P}_h\{X_k = j, k < \tau_h\} P_{ji}. \tag{7.3}$$

If $i \neq h$, then the right-hand side equals

$$\sum_{k=0}^{\infty} \mathsf{P}_h\{X_{k+1} = i, k+1 < \tau_h\} = \pi_i.$$

If $i = h$, the the right-hand side of (7.3) equals

$$\sum_{k=0}^{\infty} \mathsf{P}_h\{X_{k+1} = h, k+1 = \tau_h\} = \mathsf{P}\{\tau_h < \infty\}.$$

The invariance of ρ follows, and so does ergodicity and uniqueness of the invariant measure. which completes the proof. □

COROLLARY 7.1. *If $h \in \mathbb{X}$ is positive recurrent and ρ is the ergodic invariant distribution satisfying $\rho_i > 0$, then $\rho_h = 1/\mathsf{E}_i\tau_i$, or, equivalently*

$$\mathsf{E}_h\tau_h = 1/\rho_h. \tag{7.4}$$

Let us prove the converse statement.

THEOREM 7.4. *Suppose an ergodic invariant distribution ρ satisfies $\rho_h > 0$ for some $h \in \mathbb{X}$. Then (7.4) holds.*

PROOF: Let us denote $\tau_h^0 = \inf\{k \in \mathbb{N} : X_k = h\}$ and then recursively,

$$\tau_h^{n+1} = \inf\{k \in \mathbb{N} : X_{\tau_n + k} = h\}, \quad n \in \mathbb{N}.$$

In other words, τ_h^n is the length of the n-th excursion between two cinsecutive visits to h.

We also set $S_0 = 0$ and

$$S_n = \tau_h^1 + \ldots + \tau_h^n, \quad n \in \mathbb{N}.$$

Due to the strong Markov property, random variables $(\tau_h^n)_{n\geq 2}$ form an i.i.d. sequence. Therefore, the strong law of large numbers applies, and we obtain

$$\frac{S_n}{n} \xrightarrow{\text{a.s.}} \mathsf{E}_h\tau_h \in (0, \infty].$$

The number of excursions or visits to h up to time $m \in \mathbb{N}$ is given by

$$N(m) = \max\{n \geq 0 : \ S_n \leq m\}, \quad m \geq 0.$$

We have

$$S_{N(m)} \leq m < S_{N(m)+1}, \quad m \geq 0.$$

Therefore,

$$\frac{S_{N(m)}}{N(m)} \leq \frac{m}{N(m)} < \frac{S_{N(m)+1}}{N(m)}, \quad m \geq 0.$$

By the strong law of large numbers, both l.h.s. and r.h.s. converge to $\mathsf{E}_h\tau_h$ a.s., and so does

$$\frac{m}{N(m)} \xrightarrow{\text{a.s.}} \mathsf{E}_h\tau_h.$$

On the other hand, ergodicity of ρ impies the ergodicity of P_ρ, and the ergodic theorem implies

$$\frac{N(m)}{m} = \frac{1}{m}\sum_{k=0}^{m-1} \mathbf{1}_{\{X_k = h\}} \xrightarrow{\text{a.s.}} \mathsf{E}_\rho \mathbf{1}_{\{X_0 = h\}} = \mathsf{P}_\rho\{X_0 = h\} = \rho_h.$$

Comparing two last displays, we obtain the statement of the theorem. □

COROLLARY 7.2. *Suppose $h \in \mathbb{X}$ is null-recurrent. Then every invariant distribution ρ satisfies $\rho_h = 0$.*

REMARK 7.1. In fact, Theorem 7.4 and its corollary justify the choice of terminology. In the long run, the process spends a positive fraction of time in any positive recurrent state, and it spends asymptotically zero fraction of time in a null recurrent state.

2.1. Lyapunov–Foster criterion for positive recurrence.

THEOREM 7.5. *Let $P(\cdot,\cdot)$ be a transition kernel on a countable space $\mathbb{X}$ such that all states in $\mathbb{X}$ are communicating. Let $Lf = Pf - f$ for any function f such that Pf is well-defined. The process is positive recurrent iff there is a function $V : \mathbb{X} \to \mathbb{R}_+$ such that $A = \{x \in \mathbb{X} : \ LV(x) > -1\}$ is a finite set.*

The function V is called a *Lyapunov function*. In the theory of differential equations, a Lyapunov function is a function that decreases along the trajectories of the system. In the Markov setting, requiring deterministic decrease is often unrealistic. However, the Lyapunov function introduced in the conditions of Theorem 7.5 decreases on average outside of A:

$$PV(x) \le V(x) - 1, \quad x \in A^c. \tag{7.5}$$

PROOF OF THEOREM 7.5: For $x \notin A$ we can write

$$V(x) \ge 1 + PV(x) \ge 1 + \int_{\mathbb{X}} P(x, dx_1)V(x_1) \ge 1 + \int_{A^c} P(x, dx_1)V(x_1).$$

Applying the same estimate to $V(x_1)$ we obtain iteratively

$$\begin{aligned}
V(x) \ge& 1 + \int_{A^c} P(x, dx_1) \left(1 + \int_{A^c} P(x_1, dx_2)V(x_2)\right) \\
\ge& 1 + \int_{A^c} P(x, dx_1) + \int_{A^c}\int_{A^c} P(x, dx_1)P(x_1, dx_2)V(x_2) \\
\ge& \dots \\
\ge& 1 + \int_{A^c} P(x, dx_1) + \int_{A^c}\int_{A^c} P(x, dx_1)P(x_1, dx_2) \\
&+ \dots + \int_{A^c} \dots \int_{A^c} P(x, dx_1)P(x_1, dx_2) \dots P(x_{n-1}, dx_n) \\
&+ \int_{A^c} \dots \int_{A^c} P(x, dx_1)P(x_1, dx_2) \dots P(x_n, dx_{n+1})V(x_{n+1}).
\end{aligned}$$

Omitting the last nonnegative term, we obtain

$$V(x) \ge 1 + \mathsf{P}_x\{\tau_A > 1\} + \mathsf{P}_x\{\tau_A > 2\} + \dots + \mathsf{P}_x\{\tau_A > n\}, \quad n \in \mathbb{N}.$$

Therefore,

$$V(x) \geq \sum_{n=0}^{\infty} \mathsf{P}_x\{\tau_A > n\} = \sum_{n=0}^{\infty} \sum_{k=n+1}^{\infty} \mathsf{P}_x\{\tau_A = k\} = \sum_{k=1}^{\infty} \sum_{n=0}^{k-1} \mathsf{P}_x\{\tau_A = k\}$$
$$= \sum_{k=1}^{\infty} k \mathsf{P}_x\{\tau_A = k\} = \mathsf{E}_x \tau_A.$$

In particular, $\mathsf{P}_x\{\tau_A = \infty\} = 0$, and we also obtain $A \neq \emptyset$.

So we have proved that for all $x \in A^c$,

$$\mathsf{E}_x \tau_A < \infty \tag{7.6}$$

In fact, it immediately follows that (7.6) holds for all $x \in \mathbb{X}$. Let us now prove that there are positive recurrenct states in A. We will do so by constructing a P-invariant finite measure π on $\mathbb{X}$ such that $\pi(A) > 0$.

Let us consider a transition matrix Q_{ij} on A defined by

$$Q_{ij} = \mathsf{P}_i\{X_{\tau_A} = j\}, \quad i, j, \in A.$$

This transition matrix defines a Markov kernel on a finite set A. By Lemma 6.1, there is a Q-invariant probability ν. For all $i \in \mathbb{X}$, we define π_i to be the average time spent at i by the process X during an excursion between two consecutive visits to A, initiated according to the initial distribution ν.

$$\pi_i = \mathsf{E}_\nu \sum_{k=0}^{\infty} \mathbf{1}_{\{X_k = i, k < \tau\}} = \sum_{h \in A} \nu_h \sum_{k=0}^{\infty} \mathsf{P}_h\{X_k = i, k < \tau\}.$$

We claim that this is an invariant measure for P. For $j \notin A$,

$$(\pi P)_j = \sum_{i \in \mathbb{X}} \sum_{h \in A} \nu_h \sum_{k=0}^{\infty} \mathsf{P}_h\{X_k = i, k < \tau\} P_{ij}$$
$$= \sum_{h \in A} \nu_h \sum_{k=0}^{\infty} \mathsf{P}_h\{X_{k+1} = j, k+1 < \tau\} = \pi_j.$$

For $j \in A$,

$$(\pi P)_j = \sum_{i \in \mathbb{X}} \sum_{h \in A} \nu_h \sum_{k=0}^{\infty} \mathsf{P}_h\{X_k = i, k < \tau_A\} P_{ij}$$
$$= \sum_{h \in A} \nu_h \sum_{k=0}^{\infty} \mathsf{P}_h\{X_{k+1} = j, \tau_A = k+1\}$$
$$= \sum_{h \in A} \nu_h \mathsf{P}_h\{X_{\tau_A} = j\} = \sum_{h \in A} \nu_h Q_{hj} = \nu_j.$$

This finishes the proof of the claim since $\nu_j = \pi_j$ for $j \in A$.

Since $\pi(A) > 0$ and all states communicate, $\pi_i > 0$ for all i. Therefore, all states are positive recurrent by Theorem 7.4.

In the opposite direction, if the process is recurrent, then we can take an arbitrary non-empty finite set $A \subset \mathbb{X}$ and define $V(x) = \mathsf{E}_x \tau_A^0$ for any state $x \in \mathbb{X}$, where

$$\tau_A^0 = \inf\{n \in \mathbb{Z}_+ : X_n \in A\} = \begin{cases} 0, & X_0 \in A, \\ \tau_A, & X_0 \notin A. \end{cases}$$

Then for any $x \notin A$

$$V(x) = \mathsf{E}_x \tau_A^0 = \int_{\mathbb{X}} P(x, dy)(1 + \mathsf{E}_y \tau_A^0) = 1 + \int_{\mathbb{X}} P(x, dy) \mathsf{E}_y \tau_A^0$$
$$= 1 + \int_{\mathbb{X}} P(x, dy) V(y) = 1 + PV(x),$$

so $LV(x) = -1$. □

Let us consider an example. Suppose $\mathbb{X} = \mathbb{Z}_+$ and

$$P_{ij} = \begin{cases} p, & j = i - 1,\ i \neq 0 \\ 1 - p, & j = i + 1,\ i \neq 0, \\ 1, & i = 0,\ j = 1, \\ 0, & \text{otherwise}, \end{cases}$$

for some $p \in (0, 1/2)$. Let $v(i) = i$, $i \in \mathbb{Z}_+$. Then for $i \neq 0$,

$$Lv(i) = p(i - 1) + (1 - p)(i + 1) - i = 1 - 2p.$$

Therefore, taking $V(i) = i/(1 - 2p)$, we obtain a Lyapunov function satisfying the conditions of Theorem 7.5.

PROBLEM 7.1. Prove that the process is not positive recurrent if $p \geq 1/2$.

CHAPTER 8

More general state space. Minorization conditions.

In this section we consider an arbitrary state space $(\mathbb{X}, \mathcal{X})$ without making any assumptions on the cardinality of $\mathbb{X}$.

1. The Doeblin condition

We say that a transition probability $P(\cdot, \cdot)$ on $(\mathbb{X}, \mathcal{X})$ satisfies the Doeblin condition if there is a probability measure ν and a number $p \in (0, 1)$ on $(\mathbb{X}, \mathcal{X})$ such that $P(x, A) \geq p\nu(A)$ for all $x \in \mathbb{X}$ and all $A \in \mathcal{X}$.

The measure ν is called a *minorizing* measure for kernel $P(\cdot, \cdot)$. Under the Doeblin condition, we can introduce a kernel $Q(x, \cdot) = \nu(\cdot)$ that is identically equal to ν and define

$$P_\nu(x, \cdot) = P(x, \cdot) - \nu(\cdot) = P(x, \cdot) - pQ(x, \cdot).$$

Then P_ν is a subprobability kernel.

DEFINITION 8.1. A function $R : \mathbb{X} \times \mathcal{X} \to [0, 1]$ is called a *subprobability kernel* or *sub-Markov kernel* if i.e., (i)for every $x \in \mathbb{X}$, $R(x, \cdot)$ is a measure on $(\mathbb{X}, \mathcal{X})$ satisfying $R(x, \mathbb{X}) \leq 1$ and (ii) for every $A \in \mathbb{X}$, $R(\cdot, A)$ is $\mathcal{X}$-measurable.

THEOREM 8.1. *If Doeblin's condition holds true, then there is an invariant measure.*

PROOF: We use Q and P_ν introduced above to define

$$\rho = p \sum_{n=0}^{\infty} \nu P_\nu^n = p \sum_{n=0}^{\infty} \nu (P - pQ)^n. \tag{8.1}$$

This series converges and defines a probability measure since

$$\rho(\mathbb{X}) = p \sum_{n=0}^{\infty} (1 - p)^n = 1.$$

Let us prove that ρ is invariant. We represent $P = pQ + (P - pQ)$, note that $\mu Q = \nu$ for any measure μ on $(\mathbb{X}, \mathcal{X})$, and write

$$\rho P = p\rho Q + \rho(P - pQ) = p\nu + p \sum_{n=0}^{\infty} \nu (P - pQ)^{n+1}$$

$$= p\nu + p \sum_{n=1}^{\infty} \nu (P - pQ)^n = p \sum_{n=0}^{\infty} \nu (P - pQ)^n = \rho.$$

□

Let us try to understand the "resolvent" formula (8.1). Since

$$P(x, dy) = p\nu(dy) + (1-p)R(x, dy),$$

where $R(x, dy) = (P(x, dy) - \nu(dy))/(1-p)$ is a transition probability, we can rewrite (8.1) as

$$\rho = \sum_{n=0}^{\infty} pQ((1-p)R)^n. \tag{8.2}$$

This formula suggests that it is useful to represent the Markov evolution in the following way: one considers an i.i.d. family of Bernoulli variables (κ_i) taking values 1 with probability p and 0 with probability $1-p$. To obtain X_n given X_{n-1} we apply the following procedure: Conditioned on $\kappa_n = 1$, the distribution of X_n is ν, Conditioned on $\kappa_n = 0$, the distribution of X_n is $R(X_{n-1}, \cdot)$.

So let us now find the distribution at time 0. First, let us condition on the event

$$B_n = \{\kappa_{-n} = 1, \kappa_{-n+1} = 0, \dots, \kappa_0 = 0\}.$$

Conditioned on this event, $X_{-n} \sim \nu$, $X_{-n+1} \sim \nu R$,..., $X_0 \sim \nu R^n$. Probability of B_n equals $p(1-p)^n$. Using the complete probability formula gives (8.1).

Also, in analogy with countable state space situation, one can interpret ρ as the average occupation time during one excursion if by excursion we mean the time interval between two times when $\kappa_i = 1$.

The idea is that if

$$\tau_\kappa = \min\{n \in \mathbb{N} : \kappa_n = 0\},$$

then the distribution of X_{τ_κ} is given by ν. In a sense, the process gets started anew with ν being the initial distribution.

(This needs a more detailed explanation)

THEOREM 8.2. *If Doeblin's condition is satisfied, then there is at most one P-invariant distribution ρ.*

PROOF: Due to the ergodic decomposition, we only need to prove that there is at most one ergodic distribution under P. We also know that every two distinct ergodic distributions must be mutually singular. However, for every invariant distribution ρ,

$$\rho(A) = \int_{\mathbb{X}} \rho(dx)P(x, A) \geq \int_{\mathbb{X}} \rho(dx)p\nu(A) \geq p\nu(A),$$

an we obtain a contradiction with Lemma 5.3. □

The Doeblin condition is rather restrictive. For example, if $P(x, \cdot)$ is given by $\mathcal{N}(x/2, 1)$, one cannot find a minorizing measure serving all x simultaneously. However, this condition becomes useful if one considers embedded Markov processes. In analogy with the trick we used when proving

Theorem 7.5, we can try to execute the following program: (i) find a subset A of a Markov process is positive recurrent, (ii) prove the Doeblin condition for the Markov process observed only at times it visits A, (iii)construct an invariant distribution for this "embedded" Markov process, and (iv)use this measure to construct an invariant ditribution for the original process.

How can one ensure that a set A is positive recurrent?

THEOREM 8.3. *Let $P(\cdot,\cdot)$ be a transition kernel on a space $(\mathbb{X},\mathcal{X})$. Suppose there are a fuction $V:\mathbb{X}\to\mathbb{R}_+$ and a set $A\subset\mathcal{X}$ such that for every $x\in A$, $LV(x)\le -1$. Then $\mathsf{E}_x\tau_A\le V(x)<\infty$ for all $x\in A^c$.*

The proof of this theorem literally repeats the first part of the proof of Theorem 7.5.

The set A can often be chosen to be much smaller than $\mathbb{X}$.

Let us consider an ARMA(1) example where $P(x,\cdot)$ is given by $\mathcal{N}(ax,1)$ with $a\in(0,1)$. Let us take $V(x)=x^2$ and compute PV. Let $\xi_x\sim\mathcal{N}(ax,1)$. Then

$$PV(x)=\mathsf{E}\xi_x^2=a^2x^2+1=a^2V(x)+1,$$

One can easily check that $PV(x)-V(x)\le -1$ if $x>\sqrt{2/(1-a^2)}$, so we can conclude that the set $A=\{x\in\mathbb{R}:|x|\le\sqrt{2/(1-a^2)}\}$ is positive recurrent.

Let us introduce the following (subprobability) kernel of the embedded Markov process (observed only when in A):

$$Q(x,B)=\mathsf{P}_x\{\tau_A<\infty,\ X_{\tau_A}\in B\},\quad x\in A,\ B\in\mathbb{X}|_A.$$

This is a probability kernel if $\mathsf{P}_x\{\tau_A<\infty\}=1$ for all $x\in A$. This kernel often satisfies the Doeblin conditions. For example, in our ARMA example,

$$Q(x,B)\ge\mathsf{P}_x\{X_1\in B\}\ge c|B|,\quad x\in A,\ B\in\mathbb{X}|_A$$

where

$$c=\min_{|x|,|y|\in A}\frac{1}{\sqrt{2\pi}}e^{-(x-y)^2/2}>0.$$

Let us now understand how to construct invariant distributions using embedded Markov processes.

THEOREM 8.4. *Let $P(\cdot,\cdot)$ be a probability kernel on $(\mathbb{X},\mathcal{X})$. Let $A\in\mathcal{X}$ and a measure ν on $(A,\mathcal{X}|_A)$ satisfy*

1. *$\mathsf{E}_\nu\tau_A<\infty$.*
2. *ν is Q-invariant, where Q is a probability kernel on $(A,\mathcal{X}|_A)$ defined by*

$$Q(x,B)=\mathsf{P}_x\{\tau_A<\infty,\ X_{\tau_A}\in B\}.$$

Then the measure π defined by

$$\pi(B)=\mathsf{E}_\nu\sum_{k=0}^{\infty}\mathbf{1}_{\{X_k\in B,\ k<\tau_A\}}=\int_{\mathbb{X}}\nu(dx)\sum_{k=0}^{\infty}\mathsf{P}_x\{X_k\in B,\ k<\tau_A\}\quad B\in\mathcal{X},$$

is finite and P-invariant.

PROOF: Finiteness of π follows from (1). To prove P-invariance of π, it is sufficient to check $\pi P(B) = \pi(B)$ for measurable $B \subset A$ and $B \subset A^c$. For any $B \in \mathcal{X}$,

$$\pi P(B) = \int_{\mathbb{X}} \nu(dx) \sum_{k=0}^{\infty} \mathsf{P}_x\{X_k \in dy,\ k < \tau_A\} P(y, B).$$

If $B \subset A$, then we can continue this as:

$$\begin{aligned} \pi P(B) &= \int_{\mathbb{X}} \nu(dx) \sum_{k=0}^{\infty} \mathsf{P}_x\{X_{k+1} \in B,\ \tau_A = k+1\} \\ &= \int_{\mathbb{X}} \nu(dx) Q(x, B) = \nu(B) = \pi(B). \end{aligned}$$

If $B \subset A^c$, then

$$\pi P(B) = \int_{\mathbb{X}} \nu(dx) \sum_{k=0}^{\infty} \mathsf{P}_x\{X_{k+1} \in B, k+1 < \tau_A\} = \pi(B),$$

and the proof is completed. □

We see that the main ingredients in the above construction are: recurrence, a minorization condition, and averaging over excursions.

Note that the proof of the Perron–Frobenius theorem via contraction in Hilbert projective metric (see Section 3) exploits a similar minorization idea.

2. A Harris positive recurrence condition

In this section we introduce a version of the Harris condition originated in [**Har56**].

We will say that a transition probability $P(\cdot, \cdot)$ on $(\mathbb{X}, \mathcal{X})$ defines a Harris (positive) recurrent process if there is a set $A \in \mathcal{X}$, a probability measure ν on $(\mathbb{X}, \mathcal{X})$, numbers $p \in (0, 1)$ and $m \in \mathbb{N}$ such that the following conditions hold true:

(A) $\sup_{x \in A} \mathsf{E}_x \tau_A < \infty$
(B) $P^m(x, B) > p\nu(B)$ for all $x \in A$.

THEOREM 8.5. *If $P(\cdot, \cdot)$ satisfies the Harris recurrence condition, then there is a unique invariant distribution ρ satisfying $\rho(A) > 0$.*

PROOF: First we define

$$\tau_A^m = \min\{n > m :\ X_n \in A\}.$$

Condition (A) implies

$$\sup_{x \in A} \mathsf{E}_x \tau_A^m < \infty. \tag{8.3}$$

Also, the method of constructing invariant measures described in the last session and based on τ_A can be implemented using τ_A^m instead.

So we are going to use the Doeblin condition to prove that the Markov kernel $Q(\cdot,\cdot)$ on $(A,\mathcal{X}|_A)$ defined by

$$Q(x,B)=\mathsf{P}_x\{\tau_A^m<\infty,\ X_{\tau_A^m}\in B\}.$$

has an invariant distribution and then apply the approach of Theorem 8.4. First of all, $Q(x,B)$ is indeed a probability kernel due to (8.3).

Let us define

$$\lambda(B)=\int_{\mathbb{X}}\nu(dy)\mathsf{P}_y\{\tau_A<\infty,\ X_{\tau_A}\in B\},\quad B\in\mathcal{X}|_A.$$

Then λ is a probability measure λ on $(A,\mathcal{X}|_A)$ due to (8.3) and Condition (B).

Since

$$\begin{aligned}Q(x,B)&=\int_{\mathbb{X}}P^m(x,dy)\mathsf{P}_y\{\tau_A<\infty,\ X_{\tau_A}\in B\}\\&\geq\int_{\mathbb{X}}p\nu(dy)\mathsf{P}_y\{\tau_A<\infty,\ X_{\tau_A}\in B\}=p\lambda(B).\end{aligned}$$

the kernel Q satisfies the Doeblin condition and there is a (unique) Q-invariant measure μ. Now to prove the existence part, it remains to prove that the following measure π is finite and P-invariant:

$$\pi(B)=\mathsf{E}_\mu\sum_{k=0}^{\infty}\mathbf{1}_{\{X_k\in B,\ k<\tau_A^m\}}=\int_{\mathbb{X}}\mu(dx)\sum_{k=0}^{\infty}\mathsf{P}_x\{X_k\in B,\ k<\tau_A^m\}\quad B\in\mathcal{X},$$

and the proof of this follows the proof of Theorem 8.4, although some changes are needed.

Problem 8.1. Prove finiteness and P-invariance of π.

Let us prove uniqueness. If there are two distinct P-invariant distributions giving positive weight to A, then there are two different ergodic distributions giving positive weight to A. If ρ is P-invariant, then for any $B\in\mathcal{X}$,

$$\rho(B)=\rho P^m(B)\geq\int_A\rho(dx)P(x,A)=\rho(A)p\nu(B),$$

so these measures cannot be mutually singular, a contradiction. □

Definition 8.2. The set A that appears in the Harris condition is often called a *small set*

Definition 8.3. *We say that $A\in\mathcal{X}$ is accessible from $x\in\mathbb{X}$ if $P^m(x,A)>0$ for some m.*

Let us denote all points $x\in\mathbb{X}$ such that A is accessible from x by $\Phi(A)$. If a distribution ρ is P-invariant, then $\rho(\Phi(A))>0$ implies $\rho(A)>0$. So, one can strengthen Theorem 8.5 in the following way:

THEOREM 8.6. *If $P(\cdot,\cdot)$ satisfies the Harris recurrence condition, then there is a unique invariant distribution ρ satisfying $\rho(\Phi(A)) > 0$. If $\Phi(A) = \mathbb{X}$, i.e., A is accessible from all points, unique ergodicity (i.e., uniqueness of an invariant measure) holds.*

We refer to [**MT09**] and [**Num84**] for a much more systematic exposition of the theory. In particular, using the notion of recurrence with respect to a measure one can characterize existence of small sets. See also [**Bax11**] for a brief exposition of the theory.

ANOTHER CONSTRUCTION OF AN INVARIANT MEASURE UNDER THE HARRIS CONDITION: This construction introduces a *split chain* approach due to Nummelin [**Num78**] and Athreya and Ney[**AN78**]. It is used systematically in [**MT09**] and [**Num84**].

Let us assume for simplicity that the Harris condition holds true with $m = 1$.

We represent the distribution of X_n given $X_{n-1} \in A$ as follows: let (κ_n) be an i.i.d. sequence Bernoulli r.v. taking value 1 with probability p and value 0 with probability $1-p$. Conditioned on $\{\kappa_n = 1\}$, $X_n \sim \nu$. Conditioned on $\{\kappa_n = 0\}$, $X_n \sim R(x,\cdot)$, where $R(x,dy) = (P(x,dy) - \nu(dy))/(1-p)$.

The process (X_n, κ_n) is Markov with values in the extended space $\tilde{\mathbb{X}} = \mathbb{X} \times \{0,1\}$. We denote by $\tilde{\mathsf{P}}_\nu$ the probability on the extended probability space $\tilde{\mathbb{X}}^{\mathbb{Z}_+}$ that includes the information about κ's, and by $\tilde{\mathsf{E}}_\nu$ the corresponding expectation.

We create the following sequence of stopping times:

$$\tau_1 = \min\{n \geq 1 : X_{n-1} \in A, \kappa_{n-1} = 1\},$$
$$\tau_k = \min\{n \geq \tau_{k-1} + 1 : X_{n-1} \in A, \kappa_{n-1} = 1\}.$$

Then (X_{τ_k}) forms a Markov process with invariant distribution ν.

Let $\sigma_k = \tau_k - \tau_{k-1}, k \in \mathbb{N}$. The excursion lenghts $(\sigma_k)_{k\in\mathbb{N}}$ form an i.i.d. sequence under $\tilde{\mathsf{P}}_\nu$.

One can use condition (A) to estimate

$$\tilde{\mathsf{E}}\sigma_1 \leq \sum_{m=1}^{\infty} p(1-p)^{r-1} r \mathsf{E}\tau_A < \infty.$$

Denoting for brevity $\sigma = \sigma_1$, we can define the average occupation measure ρ by

$$\rho(B) = \tilde{\mathsf{E}}_\nu \sum_{r=0}^{\infty} \mathbf{1}_{\{X_r \in B, r < \sigma\}} = \sum_{r=0}^{\infty} \tilde{\mathsf{P}}_\nu\{X_r \in B, r < \sigma\}, \quad B \in \mathcal{X},$$

and use the previous display to show that $\rho(\mathbb{X}) < \infty$.

Let us prove that ρ is invariant. Let us first notice that

$$\begin{aligned}\rho P(B) &= \int_{\mathbb{X}} \rho(dx) P(x,B) \\ &= \int_{A^c} \rho(dx) P(x,B) + (1-p)\int_A \rho(dx) R(x,B) + p\int_A \rho(dx)\nu(B) \\ &= \sum_{r=0}^{\infty} \tilde{\mathsf{P}}_\nu\{X_r \in A^c, X_{r+1} \in B, r+1 < \sigma\} \\ &\quad + \sum_{r=0}^{\infty} \tilde{\mathsf{P}}_\nu\{X_r \in A, X_{r+1} \in B, r+1 < \sigma\} \\ &\quad + \sum_{r=0}^{\infty} \tilde{\mathsf{P}}_\nu\{X_r \in A, X_{r+1} \in B, \sigma = r+1\}.\end{aligned}$$

Now we (i) combine the first two sums into one and (ii) use the invariance of ν under the embedded Markov chain to derive

$$\sum_{r=0}^{\infty} \tilde{\mathsf{P}}_\nu\{X_r \in A, X_{r+1} \in B, \sigma = r+1\} = \nu(B).$$

The result is

$$\begin{aligned}\rho P(B) &= \sum_{r=0}^{\infty} \tilde{\mathsf{P}}_\nu\{X_{r+1} \in B, r+1 < \sigma\} + \nu(B) \\ &= \sum_{r=1}^{\infty} \tilde{\mathsf{P}}_\nu\{X_r \in B, r < \sigma\} + \tilde{\mathsf{P}}_\nu\{X_0 \in B, 0 < \sigma\} \\ &= \sum_{r=0}^{\infty} \tilde{\mathsf{P}}_\nu\{X_r \in B, r < \sigma\} = \rho(B),\end{aligned}$$

where we used $\sigma \geq 1$ in the second indentity. □

3. Coupling and convergence in total variation under the Doeblin condition

In this section we introduce a very important tool called coupling. The idea is to benefit from realizing several Markov processes with the same transition mechanism on one probability space. We will give a different proof of uniqueness of

PROOF OF UNIQUENESS OF AN INVARIANT DISTRIBUTION UNDER THE DOEBLIN CONDITION: Let ρ_1 and ρ_2 be two different invariant distributions. Let us prove that they are equal to each other using the coupling method. Let us organize a Markov process on $(\mathbb{X}\times\mathbb{X}, \mathcal{X}\times\mathcal{X})$ with initial distribution $\rho_1 \times \rho_2$ and special transition probabilities $P((x_1,x_2),\cdot)$ that we proceed to describe. So, we are going to describe the distribution of (X^1_{n+1}, X^2_{n+1}) conditioned on $(X^1_n, X^2_n) = (x_1, x_2)$.

If $x_1 \neq x_2$ then we will need a Bernoulli r.v. κ taking value 1 with probability p and value 0 with probability $1-p$. Let ξ, η_1, η_2 be a r.v.'s independent of κ with distributions ν, $R(x_1,\cdot)$ and $R(x_2,\cdot)$, respectively. Here $R(x,\cdot) = (P(x,\cdot) - p\nu(\cdot))/(1-p)$. Now we define $P((x_1,x_2),\cdot)$ to be the joint distribution of $\mathbb{X}$-valued random variables Y_1 and Y_2 defined by:

$$\begin{aligned} Y_1 =& \kappa\xi + (1-\kappa)\eta_1, \\ Y_2 =& \kappa\xi + (1-\kappa)\eta_2. \end{aligned}$$

If $x_1 = x_2$ we let Y be distributed according to $P(x_1,\cdot)$ and set $P((x_1,x_2),\cdot)$ to be the joint distribution of (Y,Y). This distribution is concentrated on the *diagonal* set $\{(y,y) :\ y \in \mathbb{X}\}$.

Notice that in both cases the transition probabilities $P((x_1,x_2),\cdot)$ project onto $P(x_1,\cdot)$ and $P(x_2,\cdot)$. This is obvious for $x_1 = x_2$, whereas if $x_1 \neq x_2$ then for any $A \in \mathcal{X}$,

$$\begin{aligned} P((x_1,x_2), A\times\mathbb{X}) &= \mathsf{P}\{Y_1 \in A; \kappa = 1\} + \mathsf{P}\{Y_1 \in A; \kappa = 0\} \\ &= p\mathsf{P}\{\xi \in A\} + (1-p)\mathsf{P}\{\eta_1 \in A\} \\ &= p\nu(A) + (1-p)(P(x_1,A) - p\nu(A))/(1-p) = P(x_1,A), \end{aligned}$$

and a similar computation shows $\mathsf{P}\{Y_2 \in A\} = P(x_2,A)$.

Therefore, the Markov process (X_n^1, X_n^2) defined by the initial distribution $\rho_1 \times \rho_2$ and transition probabilities $P((x_1,x_2),\cdot)$ satisfies the following property: the distributions of X_1 and X_2 coincide with distributions of Markov process with initial distribution ρ_1 and ρ_2, respectively, and transition probabilities $P(\cdot,\cdot)$. To define the Markov measure $\mathsf{P}_{\rho_1\times\rho_2}$ one can introduce an auxiliary probability space $(\Omega, \mathcal{F}, \mathsf{P})$ that supports a sequence of i.i.d. random variables $(\kappa_n)_{n\in\mathbb{N}}$ distributed as κ.

Moreover, if $X_n^1 = X_n^2$ for some n, then $X_m^1 = X_m^2$ holds for all $m \geq n$. Therefore,

$$\mathsf{P}_{\rho_1\times\rho_2}\{X_n^1 \neq X_2\} \leq \mathsf{P}_{\rho_1\times\rho_2}\{\tau > n\}, \quad n \in \mathbb{N},$$

where $\tau = \min\{k \in \mathbb{N} :\ X_k^1 = X_k^2\}$.

In particular, for any $A \in \mathcal{X}$

$$\begin{aligned} |\rho_1(A) - \rho_2(A)| &= |\mathsf{P}_{\rho_1}\{X_n \in A\} - \mathsf{P}_{\rho_2}\{X_n \in A\}| \\ &= |\mathsf{P}_{\rho_1\times\rho_2}\{X_n^1 \in A\} - \mathsf{P}_{\rho_1\times\rho_2}\{X_n^2 \in A\}| \\ &\leq \mathsf{P}_{\rho_1\times\rho_2}(\{X_n^1 \in A\} \triangle \{X_n^2 \in A\}) \\ &\leq \mathsf{P}_{\rho_1\times\rho_2}\{X_n^1 \neq X_n^2\} \leq \mathsf{P}_{\rho_1\times\rho_2}\{\tau > n\}. \end{aligned}$$

We have

$$\mathsf{P}_{\rho_1\times\rho_2}\{\tau > n\} \leq \mathsf{P}\{\kappa_1 = \ldots = \kappa_n = 0\} \leq (1-p)^n,$$

and, taking $n \to \infty$, we obtain $\rho_1(A) = \rho_2(A)$ which completes the proof.□

The stopping time τ in the above proof is called the coupling time.

The proof can easily be modified to provide the proof of the following strengthening of Theorem 8.2:

THEOREM 8.7. *Let Doeblin's condition be satisfied. If ρ is the invariant distribution (the uniqueness of which is guaranteed by Theorem 8.2), then for any other initial distribution ν,*

$$\|\nu P^n - \rho\|_{TV} \leq (1-p)^n.$$

□

CHAPTER 9

Taking the topology into account

1. Feller property and existence of invariant measures the via Krylov–Bogolyubov approach

Some regularity properties of transition kernels are convenient to formulate in terms of the operator P on bounded functions or the semigroup $(P^n)_{n\in\mathbb{Z}_+}$ generated by P. Let $(\mathbb{X}, d)$ be a metric space with Borel σ-algebra $\mathcal{X}$.

DEFINITION 9.1. *A transition probability $P(\cdot,\cdot)$ on $(\mathbb{X}, d)$ is Feller if for every bounded continuous $f : \mathbb{X} \to \mathbb{R}$, Pf is also continuous.*

Feller property means that $P(x,\cdot)$ is continuous in x in the topology of weak convergence.

Let us now state an analogue of the Krylov–Bogolyubov theorem for Markov processes.

THEOREM 9.1. *Let ρ be a probability measure on $(\mathbb{X}, \mathcal{X})$ such that the sequence $(\mu_n)_{n\in\mathbb{N}}$ defined by*

$$\mu_n = \frac{1}{n}\sum_{j=0}^{n-1} \rho P^j = \rho \frac{1}{n}\sum_{j=0}^{n-1} P^j, \quad n \in \mathbb{N} \tag{9.1}$$

is tight. Then there is a P-invariant measure. Also, condition (9.1) is guaranteed by tightness of the sequence $(\rho P^n)_{n\in\mathbb{N}}$.

PROOF: By Prokhorov's theorem, tightness implies relative compactness, so there is a sequence $n_k \uparrow \infty$ and a probability measure ν such that μ_{n_k} converges to ν weakly.

To prove that $\nu P = \nu$, it is sufficient to prove

$$\nu P f = \nu f \tag{9.2}$$

for all bounded continuous functions f. For such a function f, the Feller property guarantees that Pf is continuous, so due to the weak convergence,

$\mu f = \lim_{k\to\infty} \mu_{n_k} f$ and $\mu P f = \lim_{k\to\infty} \mu_{n_k} P f$. Therefore,

$$\begin{aligned}
|\nu P f - \nu f| &= \left| \lim_{k\to\infty} \frac{1}{n_k} \sum_{j=0}^{n_k-1} \rho P^j P f - \lim_{k\to\infty} \frac{1}{n_k} \sum_{j=0}^{n_k-1} \rho P^j f \right| \\
&= \left| \lim_{k\to\infty} \frac{1}{n_k} \left(\sum_{j=0}^{n_k-1} \rho P^{j+1} f - \sum_{j=0}^{n_k-1} \rho P^j f \right) \right| \\
&\le \lim_{k\to\infty} \left\| \frac{1}{n_k} (P^{n_k} f - f) \right\|_\infty \le \lim_{k\to\infty} \frac{2\|f\|_\infty}{n_k} = 0,
\end{aligned}$$

so (9.2) holds.

To prove the second claim of the theorem, we note that if a compact set K satisfies $\rho P^n(K^c) \le \varepsilon$ for all n, then $\mu_n(K^c) \le \varepsilon$ for all n. □

THEOREM 9.2. *Suppose $(\mathbb{X}, \mathcal{X})$ is a metric space. Let $V : \mathcal{X} \to \mathbb{R}$ be a measurable function such that $U_r = \{x \in \mathcal{X} : V(x) \le r\}$ is a pre-compact set for any $r > 0$. If a transition probability $P(\cdot,\cdot)$ on $(\mathbb{X}, \mathcal{X})$ satisfies*

$$PV(x) \le \alpha V(x) + \beta, \quad x \in \mathbb{X}, \tag{9.3}$$

for some $\alpha \in (0,1)$ and $\beta \ge 0$, then there is an invariant distribution for P.

PROOF: By iterating (9.3), we obtain

$$P^2 V(x) \le \alpha(\alpha V(x) + \beta) + \beta \le \alpha^2 V + \alpha\beta + \beta,$$

$$P^n V(x) \le \alpha^n V(x) + \beta(\alpha^{n-1} + \ldots + \alpha + 1) \le \alpha^n V(x) + \frac{\beta}{1-\alpha}.$$

Therefore, defining for any $x_0 \in \mathbb{X}$, μ_n by (9.1) with $\rho = \delta_{x_0}$. we get

$$\mu_n V = \delta_{x_0} \frac{1}{n} \sum_{j=0}^{n-1} P^j V \le \delta_{x_0} \left(V + \frac{\beta}{1-\alpha} \right) = V(x_0) + \frac{\beta}{1-\alpha}.$$

Due to Markov's inequality

$$\mu_n\{x : V > r\} \le \frac{\mu_n V}{r} \le \frac{1}{r} \left(V(x_0) + \frac{\beta}{1-\alpha} \right).$$

This estimate on the measure μ_n of the complement of pre-compact set U_r does not depend on n, and can be made arbitrarily small by choosing sufficiently large values of r. Therefore, μ_n is a tight family. Theorem 9.1 implies that there is a P-invariant measure. □

Let us consider an ARMA(1) example where $P(x,\cdot)$ is given by $\mathcal{N}(ax, 1)$ with $a \in (0,1)$. Let us take $V(x) = x^2$ and compute PV. Let $\xi_x \sim \mathcal{N}(ax, 1)$. Then

$$PV(x) = \mathsf{E}\xi_x^2 = a^2 x^2 + 1 = a^2 V(x) + 1,$$

so conditions of Theorem 9.2 hold with $\alpha = a^2$ and $\beta = 1$.

2. Applications to SDEs, Stochastic heat equation, stochastic Navier–Stokes equation

Will fill in later.

3. Strong Feller property and uniqueness.

DEFINITION 9.2. The *resolvent kernel* is defined by

$$Q(y, A) = \sum_{n\in\mathbb{N}} 2^{-n} P^n(y, U), \quad y \in \mathbb{X},\ A \in \mathcal{X}.$$

DEFINITION 9.3. The support of a probability measure μ on $(\mathbb{X}, \mathcal{X})$ denoted by $\operatorname{supp}\mu$ consists of all points $x \in \mathbb{X}$ such that for every open set U containing x, $\mu(U) > 0$

DEFINITION 9.4. A point $x \in \mathbb{X}$ is called *P-acccessible* if for all $y \in \mathbb{X}$, $x \in \operatorname{supp} Q(y, \cdot)$. The set of all P-accessible points will be denoted by $\mathrm{Acc}(P)$.

LEMMA 9.1. *For every P-invariant measure μ, $\mathrm{Acc}(P) \subset \mathrm{supp}(\mu)$.*

PROOF: Since $\mu = \mu P$, we compute

$$\mu Q = \sum_{n\in\mathbb{N}} 2^{-n} \mu P^n = \sum_{n\in\mathbb{N}} 2^{-n} \mu = \mu,$$

i.e., μ is also invariant under Q. Therefore, for every $x \in \mathrm{Acc}(P)$ and every open set U containing x, we obtain

$$\mu(U) = \int_{\mathbb{X}} \mu(dy) Q(y, U) > 0.$$

□

DEFINITION 9.5. *The kernel P is called strong Feller if for every bounded measurable function $f : \mathbb{X} \to \mathbb{R}$, Pf is a continuous function.*

An immediate corollary of this definition is that for a strong Feller kernel P and any set $A \in \mathcal{X}$, $P(x, A) = P\mathbf{1}_A(x)$ is a continuous function. While the usual Feller property means that $P(x, \cdot)$ converges to $P(y, \cdot)$ weakly as $x \to y$, the strong Feller property means convergence almost as strong as total variation convergence. Any continuous deterministic map $\phi : \mathbb{X} \to \mathbb{X}$ generates a Markov kernel $P(x, \cdot) = \delta_{\phi(x)}$ that is Feller but not strong Feller. One can also say that strong Feller transition probabilities improve the regularity by smoothening whereas the usual Feller kernels only preserve the regularity.

One obvious consequence of the strong Feller property is that if $P(x, A) > 0$ for some x and A, then $P(y, A) > 0$ for all y that are sufficiently close to x. In particular $P(x, \cdot)$ and $P(y, \cdot)$ cannot be mutually singular if x and y are close enough.

LEMMA 9.2. *Suppose P is strong Feller. Then for distinct P-invariant and ergodic measures μ and ν,* $\operatorname{supp}(\mu) \cap \operatorname{supp}(\nu) = \emptyset$.

PROOF: There is a set $A \in \mathcal{X}$ such that $\mu(A) = 1$ and $\nu(A) = 0$. Let us define $h(x) = P\mathbf{1}_A(x) = P(x, A)$. By the choice of A, we have $h(x) = 1$ for μ-a.e. x, and $h(x) = 0$ for ν-a.e. x. Since h is continuous due to the strong Feller property, we conclude that if $x \in \operatorname{supp}\mu$, then $h(x) = 1$, and if $x \in \operatorname{supp}\nu$, $h(x) = 0$. Therefore, $\operatorname{supp}\mu$ and $\operatorname{supp}\nu$ are disjoint. □

LEMMA 9.3. *Let P be a strong Feller transition kernel such that* $\operatorname{Acc}(P) \neq \emptyset$. *Then there is at most one P-invariant distribution.*

PROOF: Suppose that there are two distinct invariant distributions. Then there are two distinct ergodic distributions μ and ν. Lemma 9.1 impies

$$\operatorname{Acc}(P) \subset \operatorname{supp}\mu \cap \operatorname{supp}\nu = \emptyset,$$

a contradiction. □

Many finite-dimensional Markov process are strong Feller. The notion was introduced by Girsanov in [**Gir60**], and the first theorem on strong Feller property appeared in[**Mol68**]:

THEOREM 9.3. *The Markov semigroup on a compact manifold generated by a second-order uniformly elliptic operator with C^2 drift and diffusion coefficients is strong Feller.*

The study of the strong Feller property is tightly related to the regularity of transition densities. Many results can be found in [**SV06**].

The strong Feller property is tightly related to convergence to the unique invariant measure in total variation.

Let us now consider an example where there is an obvious unique invariant distribution, but the strong Feller property does not hold. Consider the Markov family associated to the following system of stochastic equations:

$$\begin{aligned} dX &= -X\,dt + dW, \\ dY &= -Y\,dt. \end{aligned}$$

The evolution of X and Y is chosen to be disentangled on purpose. We have $Y(t) = e^{-t}Y(0)$. Therefore the time t transition probability $P^t((x_0, y_0), \cdot)$ is concentrated on the line $\{(x, y) \in \mathbb{R}^2 : \ y = y_0 e^{-t}\}$. In particular, if $y_0 \neq y_0'$, then the measures $P^t((x_0, y_0), \cdot)$ and $P^t((x_0, y_0'), \cdot)$ are mutually singular and therefore P cannot be strong Feller.

Bibliography

[AN78] K. B. Athreya and P. Ney. A new approach to the limit theory of recurrent Markov chains. *Trans. Amer. Math. Soc.*, 245:493–501, 1978.

[Arn98] Ludwig Arnold. *Random dynamical systems*. Springer Monographs in Mathematics. Springer-Verlag, Berlin, 1998.

[Bal00] Viviane Baladi. *Positive transfer operators and decay of correlations*, volume 16 of *Advanced Series in Nonlinear Dynamics*. World Scientific Publishing Co. Inc., River Edge, NJ, 2000.

[Bax11] Peter Baxendale. T. E. Harris's contributions to recurrent Markov processes and stochastic flows. *Ann. Probab.*, 39(2):417–428, 2011.

[BC72] R. M. Blumenthal and H. H. Corson. On continuous collections of measures. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability (Univ. California, Berkeley, Calif., 1970/1971), Vol. II: Probability theory*, pages 33–40, Berkeley, Calif., 1972. Univ. California Press.

[Bil68] Patrick Billingsley. *Convergence of probability measures*. John Wiley & Sons Inc., New York, 1968.

[Bir31] George D. Birkhoff. Proof of the ergodic theorem. *Proceedings of the National Academy of Sciences*, 17(12):656–660, 1931.

[Boh09] Piers Bohl. ber ein in der theorie der skularen strungen vorkommendes problem. *J. f. Math.*, 135:189283, 1909.

[CDF97] Hans Crauel, Arnaud Debussche, and Franco Flandoli. Random attractors. *J. Dynam. Differential Equations*, 9(2):307–341, 1997.

[CF94] Hans Crauel and Franco Flandoli. Attractors for random dynamical systems. *Probab. Theory Related Fields*, 100(3):365–393, 1994.

[Chu67] Kai Lai Chung. *Markov chains with stationary transition probabilities*. Second edition. Die Grundlehren der mathematischen Wissenschaften, Band 104. Springer-Verlag New York, Inc., New York, 1967.

[Doe38] W. Doeblin. Expose de la théorie des chaînes simples constantes de Markoff à un nombre fini d'etats. Rev. Math. Union Interbalkan. 2, 77-105 (1938)., 1938.

[DS88] Nelson Dunford and Jacob T. Schwartz. *Linear operators. Part I*. Wiley Classics Library. John Wiley & Sons, Inc., New York, 1988. General theory, With the assistance of William G. Bade and Robert G. Bartle, Reprint of the 1958 original, A Wiley-Interscience Publication.

[EW11] Manfred Einsiedler and Thomas Ward. *Ergodic theory with a view towards number theory*, volume 259 of *Graduate Texts in Mathematics*. Springer-Verlag London Ltd., London, 2011.

[Fur60] Harry Furstenberg. *Stationary processes and prediction theory*. Annals of Mathematics Studies, No. 44. Princeton University Press, Princeton, N.J., 1960.

[FW12] Mark I. Freidlin and Alexander D. Wentzell. *Random perturbations of dynamical systems*, volume 260 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer, Heidelberg, third edition, 2012. Translated from the 1979 Russian original by Joseph Szücs.

[Gir60] I. V. Girsanov. Strong Feller processes. I. General properties. *Teor. Verojatnost. i Primenen.*, 5:7–28, 1960.

[Hal60] Paul R. Halmos. *Lectures on ergodic theory*. Chelsea Publishing Co., New York, 1960.

[Har56] T. E. Harris. The existence of stationary measures for certain Markov processes. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, vol. II*, pages 113–124. University of California Press, Berkeley and Los Angeles, 1956.

[Kam82] Teturo Kamae. A simple proof of the ergodic theorem using nonstandard analysis. *Israel J. Math.*, 42(4):284–290, 1982.

[KH95] Anatole Katok and Boris Hasselblatt. *Introduction to the modern theory of dynamical systems*, volume 54 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, 1995. With a supplementary chapter by Katok and Leonardo Mendoza.

[Kif86] Yuri Kifer. *Ergodic theory of random transformations*, volume 10 of *Progress in Probability and Statistics*. Birkhäuser Boston Inc., Boston, MA, 1986.

[Kin73] J. F. C. Kingman. Subadditive ergodic theory. *Ann. Probability*, 1:883–909, 1973. With discussion by D. L. Burkholder, Daryl Daley, H. Kesten, P. Ney, Frank Spitzer and J. M. Hammersley, and a reply by the author.

[Kol50] A. N. Kolmogorov. *Foundations of the Theory of Probability (Translation of Grundbegriffe der Wahrscheinlichkeitsrechnung, Springer, Berlin, 1933.)*. Chelsea Publishing Company, New York, N. Y., 1950.

[Koo31] B. O. Koopman. Hamiltonian Systems and Transformations in Hilbert Space. *Proceedings of the National Academy of Science*, 17:315–318, May 1931.

[KS07] Leonid B. Koralov and Yakov G. Sinai. *Theory of probability and random processes*. Universitext. Springer, Berlin, second edition, 2007.

[KT09] K. Khanin and A. Teplinsky. Herman's theory revisited. *Invent. Math.*, 178(2):333–344, 2009.

[Kur66] K. Kuratowski. *Topology. Vol. I*. New edition, revised and augmented. Translated from the French by J. Jaworowski. Academic Press, New York, 1966.

[KW82] Yitzhak Katznelson and Benjamin Weiss. A simple proof of some ergodic theorems. *Israel J. Math.*, 42(4):291–296, 1982.

[Lal10] Steven P. Lalley. Kingman's subadditive ergodic theorem. *http://galton.uchicago.edu/~lalley/Courses/Graz/Kingman.pdf*, 2010.

[Lig05] Thomas M. Liggett. *Interacting particle systems*. Classics in Mathematics. Springer-Verlag, Berlin, 2005. Reprint of the 1985 original.

[Mol68] S. A. Molčanov. The strong Feller property of diffusion processes on smooth manifolds. *Teor. Verojatnost. i Primenen.*, 13:493–498, 1968.

[MT09] Sean Meyn and Richard L. Tweedie. *Markov chains and stochastic stability*. Cambridge University Press, Cambridge, second edition, 2009. With a prologue by Peter W. Glynn.

[Num78] E. Nummelin. A splitting technique for Harris recurrent Markov chains. *Z. Wahrsch. Verw. Gebiete*, 43(4):309–318, 1978.

[Num84] Esa Nummelin. *General irreducible Markov chains and nonnegative operators*, volume 83 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 1984.

[Ohn83] Taijiro Ohno. Asymptotic behaviors of dynamical systems with random parameters. *Publ. Res. Inst. Math. Sci.*, 19(1):83–98, 1983.

[Poi90] Henri Poincaré. Sur le problème des trois corps et les équations de la dynamique. *Acta Mathematica*, 13:1–270, 1890.

[Poi99] Henri Poincaré. *Les méthodes nouvelles de la mécanique céleste, Volume 3*. Gauthier-Villars, Paris, 1899.

[Ros71] Murray Rosenblatt. *Markov processes. Structure and asymptotic behavior*. Springer-Verlag, New York-Heidelberg, 1971. Die Grundlehren der mathematischen Wissenschaften, Band 184.

[Sar09] Omri Sarig. *Lecture Notes on Ergodic Theory. http://www.wisdom.weizmann.ac.il/šarigo/506/ErgodicNotes.pdf*, 2009.

[Sch94] B. Schmalfuss. Stochastische Attraktoren des stochastischen Lorenz-Systems. *Z. Angew. Math. Mech.*, 74(6):t627–t628, 1994.

[Shi96] A. N. Shiryaev. *Probability*, volume 95 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, second edition, 1996. Translated from the first (1980) Russian edition by R. P. Boas.

[Sie10] W. Sierpinski. Sur la valeur asymptotique d'une certaine somme. *Bull Intl. Acad. Polonmaise des Sci. et des Lettres (Cracovie) series A*, page 911, 1910.

[Sin76] Ya. G. Sinai. *Introduction to ergodic theory*. Princeton University Press, Princeton, N.J., 1976. Translated by V. Scheffer, Mathematical Notes, 18.

[SV06] Daniel W. Stroock and S. R. Srinivasa Varadhan. *Multidimensional diffusion processes*. Classics in Mathematics. Springer-Verlag, Berlin, 2006. Reprint of the 1997 edition.

[Tao08] Terence Tao. 254a, lecture 8: The mean ergodic theorem. *http://terrytao.wordpress.com/2008/01/30/254a-lecture-8-the-mean-ergodic-theorem/*, 2008.

[vN32] J. von Neumann. Proof of the quasi-ergodic hypothesis. *Proceedings of the National Academy of Sciences*, 18(1):70–82, 1932.

[Wal82] Peter Walters. *An introduction to ergodic theory*, volume 79 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1982.

[Wey10] Hermann Weyl. Über die Gibbs'sche Erscheinung und verwandte Konvergenzphänomene. *Rendiconti del Circolo Matematico di Palermo*, 330:377407, 1910.

[You86] Lai-Sang Young. Stochastic stability of hyperbolic attractors. *Ergodic Theory Dynam. Systems*, 6(2):311–319, 1986.