Ergodic Properties of Markov Processes

July 29, 2018

Martin Hairer

Lecture given at The University of Warwick in Spring 2006

1 Introduction

Markov processes describe the time-evolution of random systems that do not have any memory. Let us demonstrate what we mean by this with the following example.

Consider a switch that has two states: on and off. At the beginning of the experiment, the switch is on. Every minute after that, we throw a dice. If the dice shows 6, we flip the switch, otherwise we leave it as it is. The state of the switch as a function of time is a **Markov process**. This very simple example allows us to explain what we mean by "does not have any memory". It is clear that the state of the switch has some memory in the sense that if the switch is off after 10 minutes, then it is more likely to be also off after 11 minutes, whereas if it was on, it would be more likely to be on. However, if we know the state of the switch at time n, we can predict its evolution (in terms of random variables of course) for all future times, without requiring any knowledge about the state of the switch at times less than n. In other words, **the future of the process depends on the present but is independent of the past**.

The following is an example of a process which is not a Markov process. Consider again a switch that has two states and is on at the beginning of the experiment. We again throw a dice every minute. However, this time we flip the switch only if the dice shows a 6 but didn't show a 6 the previous time.

Let us go back to our first example and write $x_1^{(n)}$ for the probability that the switch is on at time n. Similarly, we write $x_2^{(n)}$ for the probability of the switch being off at time n. One then has the following recursion relation:

$$x_1^{(n+1)} = \frac{5}{6}x_1^{(n)} + \frac{1}{6}x_2^{(n)}\,, \qquad x_2^{(n+1)} = \frac{1}{6}x_1^{(n)} + \frac{5}{6}x_2^{(n)}\,, \tag{1.1}$$

with $x_1^{(0)} = 1$ and $x_2^{(0)} = 0$. The first equality comes from the observation that the switch is on at time $n+1$ if either it was on at time n and we didn't throw a 6 or it was off at time n and we did throw a 6. Equation (1.1) can be written in matrix form as

$$x^{(n+1)} = Tx^{(n)}\,, \quad T = \frac{1}{6}\begin{pmatrix} 5 & 1 \\ 1 & 5 \end{pmatrix}.$$

We note that T has the eigenvalue 1 with eigenvector $(1,1)$ and the eigenvalue $2/3$ with eigenvector $(1,-1)$. Note also that $x_1^{(n)} + x_2^{(n)} = 1$ for all values of n. Therefore we have

$$\lim_{n\to\infty} x_1^{(n)} = \frac{1}{2}\,, \quad \lim_{n\to\infty} x_2^{(n)} = \frac{1}{2}\,.$$

We would of course have reached the same conclusion if we started with our switch being off at time 0.

2 Elements of probability theory

Recall that a probability space $(\Omega, \mathscr{F}, \mathbf{P})$ consists of a set Ω endowed with a σ-algebra $\mathscr{F}$ and a probability measure $\mathbf{P}$. We have

Definition 2.1 A **σ-algebra** $\mathscr{F}$ over a set Ω is a collection of subsets of Ω with the properties that $\emptyset \in \mathscr{F}$, if $A \in \mathscr{F}$ then $A^c \in \mathscr{F}$ and, if $\{A_n\}_{n>0}$ is a countable collection of elements of $\mathscr{F}$, then $\bigcup_{n>0} A_n \in \mathscr{F}$.

Note that if $\mathscr{G}$ is any collection of subsets of a set Ω, then there always exists a smallest σ-algebra containing $\mathscr{G}$. (Show that this is indeed the case.) We denote it by $\sigma\mathscr{G}$ and call it the σ-algebra generated by $\mathscr{G}$.

Definition 2.2 A **probability measure P** on the measurable space $(\Omega, \mathscr{F})$ is a map $\mathbf{P}\colon \mathscr{F} \to [0,1]$ with the properties

- $\mathbf{P}(\emptyset) = 0$ and $\mathbf{P}(\Omega) = 1$.
- If $\{A_n\}_{n>0}$ is a countable collection of elements of $\mathscr{F}$ that are all disjoint, then one has $\mathbf{P}(\bigcup_{n>0} A_n) = \sum_{n>0} \mathbf{P}(A_n)$.

Throughout this course, we will always consider the case of discrete time. We therefore give the following definition of a stochastic process.

Definition 2.3 A **stochastic process** x with state space $\mathcal{X}$ is a collection $\{x_n\}_{n=0}^{\infty}$ of $\mathcal{X}$-valued random variables on some probability space $(\Omega, \mathscr{F}, \mathbf{P})$. Given n, we refer to x_n as the value of the process at time n. We will sometimes consider processes for which time can take negative values, *i.e.* $\{x_n\}_{n\in\mathbf{Z}}$.

Note that we didn't say anything about the state space $\mathcal{X}$. For the moment, all we need is that the notion of $\mathcal{X}$-valued random variable makes sense. For this, we need $\mathcal{X}$ to be a measurable space, so that an $\mathcal{X}$-valued random variable is a measurable map from Ω to $\mathcal{X}$. We will however always assume that $\mathcal{X}$ is a complete separable metric space, so that for example Fubini's theorem holds.

We will impose more structure on $\mathcal{X}$ further on. Typical examples are:

- A finite set, $\mathcal{X} = \{1, \dots, n\}$.
- $\mathcal{X} = \mathbf{R}^n$ or $\mathcal{X} = \mathbf{Z}^n$.
- Some manifold, for example $\mathcal{X} = S^n$, the n-dimensional sphere or $\mathcal{X} = \mathcal{T}$, the torus.
- A Hilbert space $\mathcal{X} = L^2([0,1])$ or $\mathcal{X} = \ell^2$.

We will always denote by $\mathscr{B}(\mathcal{X})$ the Borel σ-algebra on $\mathcal{X}$, *i.e.* $\mathscr{B}(\mathcal{X})$ is the smallest σ-algebra which contains every open set. We will call a function f between two topological spaces **measurable** if $f^{-1}(A)$ is a Borel set for every Borel set A. If $f\colon \Omega \to \mathcal{X}$, we call f a **random variable**, provided that $f^{-1}(A) \in \mathscr{F}$ for every Borel set A. One actually has:

Proposition 2.4 *Let $f\colon \Omega \to \mathcal{X}$ and suppose that $f^{-1}(A) \in \mathscr{F}$ for every open set A. Then $f^{-1}(A) \in \mathscr{F}$ for every Borel set A.*

Proof. Define $\mathscr{G}_0 = \{f^{-1}(A) \,|\, A \text{ open}\}$ and $\mathscr{G} = \{f^{-1}(A) \,|\, A \text{ Borel}\}$. Since $\mathscr{G}$ is a σ-algebra and $\mathscr{G}_0 \subset \mathscr{G}$, one has $\sigma\mathscr{G}_0 \subset \mathscr{G}$.

Define now $\mathscr{F}_0 = \{A \in \mathscr{B}(\mathcal{X}) \,|\, f^{-1}(A) \in \sigma\mathscr{G}_0\}$. It is straightforward to check that $\mathscr{F}_0$ is a σ-algebra and that it contains all open sets. Since $\mathscr{B}(\mathcal{X})$ is the smallest σ-algebra containing all open

sets, this shows that $\mathscr{F}_0 = \mathscr{B}(\mathcal{X})$ and therefore $\sigma\mathscr{G}_0 = \mathscr{G}$. Since on the other hand $\sigma\mathscr{G}_0 \subset \mathscr{F}$, this shows the claim. □

This proposition is useful because of the following corollary:

Corollary 2.5 *Let $\mathcal{X}$ and $\mathcal{Y}$ be two topological spaces and let $f\colon \mathcal{X} \to \mathcal{Y}$ be continuous. Then f is (Borel) measurable.*

Proof. Since f is continuous, $f^{-1}(A)$ is open for every open set A, so that $f^{-1}(A) \in \mathscr{B}(\mathcal{X})$. The claim then follows from Proposition 2.4. □

Exercise 2.6 You have probably seen Lebesgue measurable functions defined through the property that $f^{-1}(A)$ is Lebesgue measurable for every open set A. Show that every Borel measurable function is also Lebesgue measurable but that the converse is not true in the case of functions from $\mathbf{R}$ to $\mathbf{R}$.

Show that if $f : \mathcal{X} \to \mathcal{Y}$ and $g : \mathcal{Y} \to \mathcal{Z}$ are Borel measurable functions, then $g \circ f$ is also Borel measurable. This property is *not* true for Lebesgue measurable functions. Try to find a *continuous* function $f\colon \mathbf{R} \to \mathbf{R}$ and a Lebesgue measurable function g (you can take an indicator function for g) such that $g \circ f$ is not Lebesgue meausrable. **Hint:** It is useful to remember that every measurable set A of positive Lebesgue measure contains a subset $A' \subset A$ which is not Lebesgue measurable. Another useful ingredient for the construction of f is the Cantor function (also called Devil's staircase).

2.1 Conditional expectations and probabilities

Consider the following situation. You and three friends play Bridge. The dealer dealt the cards but you haven't looked at them yet. At this point, assuming that the cards were perfectly shuffled (*i.e.* every configuration has the same probability), the probability that your partner has the ace of spades is equal to $1/4$. Now look at your cards. If you happen to have the ace of spades, the probability that your partner has it obviously drops to 0. If you don't have it, the probability that your partner has it raises to $1/3$. Note that this probability is now a *function* of the values of the cards in your hand. The possible values of this function depend on the nature of the information that becomes available. This is a simple example of a conditional probability.

The mathematical object that represents information is the σ-algebra. In mathematical terms, if a quantity can be evaluated by using only the information contained in a given σ-algebra, then it is *measurable* with respect to that σ-algebra. It is a good exercise to convince yourself that this intuitive notion of measurability does indeed correspond to the formal definition given above.

As an example, consider the trivial σ-algebra given by $\mathscr{T} = \{\emptyset, \Omega\}$. This is the mathematical equivalent to the statement 'we have no information at all'. A function which is measurable with respect to $\mathscr{T}$ is constant, which means that its value at any given point can be computed without requiring any information at all on that point. One should think of the conditional expectation of a random variable as the best guess one can make for its value (on average) given a certain amount of information. If no information at all is given, the best guess would be the expectation of the random variable, which is indeed a constant.

Let us go back to the example of the Bridge players. In this case, a natural choice for Ω is the set of all possible configuration of cards. When you look at your hand, the σ-algebra encoding this extra information is given by the collection of all subsets A with the property that if one particular configuration of cards belongs to A, then all the other configurations that assign to you the same hand also belong to A.

These considerations motivate the following definition for the conditional expectation of a random variable:

Definition 2.7 Let X be a real-valued random variable on some probability space $(\Omega, \mathscr{F}, \mathbf{P})$ such that $\mathbf{E}|X| < \infty$ and let $\mathscr{F}'$ be a sub σ-algebra of $\mathscr{F}$. Then the **conditional expectation** of X with respect to $\mathscr{F}'$ is the $\mathscr{F}'$-measurable random variable X' such that

$$\int_A X(\omega)\,\mathbf{P}(d\omega) = \int_A X'(\omega)\,\mathbf{P}(d\omega)\,, \tag{2.1}$$

for every $A \in \mathscr{F}'$. We denote this by $X' = \mathbf{E}(X \,|\, \mathscr{F}')$.

Example 2.8 If the only information we know is whether a certain event B happened or not, then it should by now be intuitively clear that the conditional expectation of a random variable X with respect to this information is given by

$$X_B = \frac{1}{\mathbf{P}(B)} \int_B X(\omega)\,\mathbf{P}(d\omega)\,,$$

if B happened and by

$$X_{B^c} = \frac{1}{\mathbf{P}(B^c)} \int_{B^c} X(\omega)\,\mathbf{P}(d\omega)\,,$$

if B didn't happen. (Here we used the notation B^c to denote the complement of B.) It is a straightforward exercise that the conditional expectation of X with respect to the σ-algebra $\mathscr{F}_B = \{\emptyset, \Omega, B, B^c\}$ is indeed given by

$$\mathbf{E}(X \,|\, \mathscr{F}_B)(\omega) = \begin{cases} X_B & \text{if } \omega \in B \\ X_{B^c} & \text{otherwise.} \end{cases}$$

It is a good exercise to compute the conditional expectation of a random variable X with respect to the σ-algebra generated by two events B_1 and B_2 (*i.e.* the smallest σ-algebra containing both B_1 and B_2).

Recall the Radon-Nikodym theorem from measure theory:

Theorem 2.9 (Radon-Nikodym) *Let μ and ν be two finite measures on a space $(\Omega, \mathscr{F})$ such that μ is absolutely continuous with respect to ν (i.e. $\nu(A) = 0$ implies $\mu(A) = 0$) and ν is positive. Then, there exists an essentially unique measurable function $D\colon \Omega \to \mathbf{R}$ such that $\mu(A) = \int_A D(\omega)\,\nu(d\omega)$.*

Here we used the expression essentially unique to say that if D_1 and D_2 are two possible choices for the density of μ with respect to ν, then the set $\{\omega \,|\, D_1(\omega) \neq D_2(\omega)\}$, is of ν-measure 0. Using this theorem, we can prove:

Proposition 2.10 *With the notations as above, the conditional expectation $X' = \mathbf{E}(X \,|\, \mathscr{F}')$ exists and is essentially unique.*

Proof. Denote by ν the restriction of $\mathbf{P}$ to $\mathscr{F}'$ and define the measure μ on $(\Omega, \mathscr{F}')$ by $\mu(A) = \int_A X(\omega)\,\mathbf{P}(d\omega)$ for every $A \in \mathscr{F}'$. It is clear that μ is absolutely continuous with respect to ν. Its density with respect to ν given by the Radon-Nikodym theorem is then the required conditional expectation. The uniqueness follows from the uniqueness statement in the Radon-Nikodym theorem. □

Exercise 2.11 Show that if $\mathscr{F}'$ is the trivial σ-algebra, *i.e.* $\mathscr{F}' = \{\emptyset, \Omega\}$, then X' is constant and equal to the expectation of X. **Hint:** remember that X' being $\mathscr{F}'$-measurable means that the preimage under X' of an arbitrary Borel set is in $\mathscr{F}'$.

Using only (2.1), show that $\mathscr{F}' = \mathscr{F}$ implies $X'(\omega) = X(\omega)$ for almost every $\omega \in \Omega$.

Exercise 2.12 Show the following elementary properties of conditional expectations:

- If $\mathscr{F}_1 \subset \mathscr{F}_2$, then one has $\mathbf{E}(\mathbf{E}(X \,|\, \mathscr{F}_2) \,|\, \mathscr{F}_1) = \mathbf{E}(\mathbf{E}(X \,|\, \mathscr{F}_1) \,|\, \mathscr{F}_2) = \mathbf{E}(X \,|\, \mathscr{F}_1 \wedge \mathscr{F}_2)$.
- If Y is $\mathscr{F}_1$-measurable, then $\mathbf{E}(XY \,|\, \mathscr{F}_1) = Y\, \mathbf{E}(X \,|\, \mathscr{F}_1)$.
- If $\mathscr{F}_1 \subset \mathscr{F}_2$, and Y is $\mathscr{F}_2$-measurable then $\mathbf{E}(Y\mathbf{E}(X \,|\, \mathscr{F}_2) \,|\, \mathscr{F}_1) = \mathbf{E}(XY \,|\, \mathscr{F}_1)$.
- Show by a counterexample that $\mathbf{E}(\mathbf{E}(X \,|\, \mathscr{F}_2) \,|\, \mathscr{F}_1) = \mathbf{E}(\mathbf{E}(X \,|\, \mathscr{F}_1) \,|\, \mathscr{F}_2)$ is *not* true in general.

We define similarly the concept of **conditional probability**.

Definition 2.13 Let X be an $\mathcal{X}$-valued random variable and write χ_A for the characteristic function of a measurable set $A \subset \mathcal{X}$. Let $\mathscr{F}'$ and $(\Omega, \mathscr{F}, \mathbf{P})$ be as above. We define

$$\mathbf{P}(X \in A \,|\, \mathscr{F}') = \mathbf{E}(\chi_A \circ X \,|\, \mathscr{F}')\,,$$

and we call this the conditional probability that X is in A knowing $\mathscr{F}'$.

Remark 2.14 It is in general a non-trivial task to show that the conditional probabilities as defined above yield a $\mathscr{F}'$-measurable function from $\mathcal{X}$ into the space of probability measures on $\mathcal{X}$. Can you imagine where the problem lies?

In many situations, we will describe $\mathscr{F}'$ as the σ-algebra generated by an other random variable Y:

Definition 2.15 Let Y be a $\mathcal{Y}$-valued random variable on a probability space $(\Omega, \mathscr{F}, \mathbf{P})$. We denote by $\mathscr{F}_Y \subset \mathscr{F}$ the σ-algebra consisting of all elements of the form $Y^{-1}(A)$ with $A \in \mathscr{B}(\mathcal{Y})$ and we say that $\mathscr{F}_Y$ is the σ-algebra **generated by** Y.

The following lemma gives a rigorous meaning to the notation $\mathbf{P}(X \in A \,|\, Y = y)$:

Lemma 2.16 *Let X be an $\mathbf{R}$-valued random variable and Y be a $\mathcal{Y}$-valued random variable. Then X is $\mathscr{F}_Y$-measurable if and only if there exists a measurable function $f : \mathcal{Y} \to \mathbf{R}$ such that $X = f \circ Y$.*

Proof. It is clear that $X = f \circ Y$ implies the $\mathscr{F}_Y$-measurability of X, so we only prove the converse.

Consider first the case where X takes only a countable number of values (a_n) and write $A_n = X^{-1}(\{a_n\})$. Since X is $\mathscr{F}_Y$-measurable, there exist sets $B_n \in \mathscr{B}(\mathcal{Y})$ such that $Y^{-1}(B_n) = A_n$. Define now the sets $C_n = B_n \backslash \bigcup_{p<n} B_p$. These sets are disjoint and one has again $Y^{-1}(C_n) = A_n$. Setting $f(x) = a_n$ for $x \in C_n$ and $f(x) = 0$ for $x \in \mathcal{Y} \setminus \bigcup_n C_n$, we see that f has the required property.

In the general case, we can approximate X by a $\mathscr{F}_Y$-measurable random variable $X_N = [NX]/N$. This random variable takes only a countable number of values, so, by the previous part, there is f_N such that $X_N = f_N \circ Y$. Define $f(x) = \lim_{n\to\infty} f_n(x)$ whenever that limit exists and 0 otherwise. Then f is the function we are looking for. Since f is a pointwise limit of measurable functions, f is also measurable. □

Note that Lemma 2.16 is still valid if $\mathbf{R}$ is replaced by $\mathbf{R}^n$. This results allows us to use the notations $\mathbf{E}(X \,|\, Y = y)$ for $\mathbf{R}^n$ valued random variables X and $\mathbf{P}(X \in A \,|\, Y = y)$ for arbitrary random variables.

Definition 2.17 Given two σ-algebras $\mathscr{F}_1$ and $\mathscr{F}_2$, we denote by $\mathscr{F}_1 \vee \mathscr{F}_2$ the smallest σ-algebra containing $\mathscr{F}_1$ and $\mathscr{F}_2$. We denote by $\mathscr{F}_1 \wedge \mathscr{F}_2$ the intersection of $\mathscr{F}_1$ and $\mathscr{F}_2$.

Exercise 2.18 Show that $\mathscr{F}_1 \wedge \mathscr{F}_2$ is indeed again a σ-algebra.

Exercise 2.19 Show that $\mathscr{F}_1 \vee \mathscr{F}_2$ can equivalently be characterised by the expressions:

- $\mathscr{F}_1 \vee \mathscr{F}_2 = \sigma\{A \cup B \,|\, A \in \mathscr{F}_1 \text{ and } B \in \mathscr{F}_2\}$,
- $\mathscr{F}_1 \vee \mathscr{F}_2 = \sigma\{A \cap B \,|\, A \in \mathscr{F}_1 \text{ and } B \in \mathscr{F}_2\}$,

where $\sigma\mathscr{G}$ denotes the smallest σ-algebra containing $\mathscr{G}$.

2.2 Markov processes

With the previous notations, given a stochastic process $\{x_n\}_{n\in\mathbf{N}}$, we define for every $m \geq n$ the σ-algebras $\mathscr{F}_n^m = \sigma(x_n) \vee \sigma(x_{n+1}) \vee \ldots \vee \sigma(x_m)$, where $\sigma(Y)$ denotes the σ-algebra generated by a random variable Y. We also use the abbreviation $\mathscr{F}_n = \sigma(x_n) = \mathscr{F}_n^n$.

With this notation, we define a **Markov process** as follows:

Definition 2.20 A process x is Markov if, for every $n > 0$ and every measurable bounded function $f\colon \mathcal{X} \to \mathbf{R}$ one has

$$\mathbf{E}(f(x_n) \,|\, \mathscr{F}_0^{n-1}) = \mathbf{E}(f(x_n) \,|\, \mathscr{F}_{n-1}) \,,$$

almost surely.

Intuitively, this means that knowing the entire history of the process does not contain any more information than knowing its last value.

Exercise 2.21 Show that if a process is Markov then, for every sequence of times $t_1, \ldots, t_k$, one has

$$\mathbf{E}(f(x_{t_k}) \,|\, \mathscr{F}_{t_1} \vee \mathscr{F}_{t_2} \vee \ldots \vee \mathscr{F}_{t_{k-1}}) = \mathbf{E}(f(x_{t_k}) \,|\, \mathscr{F}_{t_{k-1}}) \,,$$

for every measurable bounded function $f\colon \mathcal{X} \to \mathbf{R}$.

Definition 2.20 has the following consequence:

Proposition 2.22 *Let x be a Markov process and let $\ell \leq m \leq n$. Then, for every measurable function f, $\mathbf{E}(f(x_n) \,|\, \mathscr{F}_\ell \vee \mathscr{F}_m) = \mathbf{E}(f(x_n) \,|\, \mathscr{F}_m)$.*

Proof. Fix ℓ and m and let us prove the claim by recursion on n. If $n = m$, the claim is true since both sides of the equality are simply equal to $f(x_n)$. Let us therefore assume that the claim holds for $n = k-1$ with $k > m$. One then has

$$\mathbf{E}(f(x_k) \,|\, \mathscr{F}_\ell \vee \mathscr{F}_m) = \mathbf{E}(\mathbf{E}(f(x_k) \,|\, \mathscr{F}_0^{k-1}) \,|\, \mathscr{F}_\ell \vee \mathscr{F}_m) = \mathbf{E}(\mathbf{E}(f(x_k) \,|\, \mathscr{F}_{k-1}) \,|\, \mathscr{F}_\ell \vee \mathscr{F}_m) \,.$$

Let us now define g by $g(x_{k-1}) = \mathbf{E}(f(x_k) \,|\, \mathscr{F}_{k-1})$ (such a function exists by Lemma 2.16). One can the use our assumption, so that

$$\begin{aligned}\mathbf{E}(f(x_k) \,|\, \mathscr{F}_\ell \vee \mathscr{F}_m) &= \mathbf{E}(g(x_{k-1}) \,|\, \mathscr{F}_\ell \vee \mathscr{F}_m) = \mathbf{E}(g(x_{k-1}) \,|\, \mathscr{F}_m) \\ &= \mathbf{E}(\mathbf{E}(f(x_k) \,|\, \mathscr{F}_{k-1}) \,|\, \mathscr{F}_m) \,.\end{aligned}$$

On the other hand, $\mathbf{E}(f(x_k) \,|\, \mathscr{F}_{k-1}) = \mathbf{E}(f(x_k) \,|\, \mathscr{F}_0^{k-1})$ by the Markov property and $\mathscr{F}_m \subset \mathscr{F}_0^{k-1}$, so that the right-hand side is equal to $\mathbf{E}(f(x_k) \,|\, \mathscr{F}_m)$. This shows that the claim then holds for $n = k$, so that by induction it holds for every $n \geq m$. □

Theorem 2.23 *Given a process $\{x_n\}_{n\in\mathbf{N}}$, three indices $\ell < m < n$, the following properties are equivalent:*

(i) For every measurable function f, $\mathbf{E}(f(x_n) \,|\, \mathscr{F}_\ell \vee \mathscr{F}_m) = \mathbf{E}(f(x_n) \,|\, \mathscr{F}_m)$.

(ii) For every measurable function g, $\mathbf{E}(g(x_\ell) \,|\, \mathscr{F}_m \vee \mathscr{F}_n) = \mathbf{E}(g(x_\ell) \,|\, \mathscr{F}_m)$.

(iii) For every two measurable functions f and g, one has

$$\mathbf{E}(f(x_n)g(x_\ell) \,|\, \mathscr{F}_m) = \mathbf{E}(f(x_n) \,|\, \mathscr{F}_m)\, \mathbf{E}(g(x_\ell) \,|\, \mathscr{F}_m) \;.$$

Proof. By symmetry, it is enough to prove that *(i)* is equivalent to *(iii)*. We start by proving that *(i)* implies *(iii)*. Given some f and g, it follows from Exercise 2.12 that

$$\begin{aligned}\mathbf{E}(f(x_n)g(x_\ell) \,|\, \mathscr{F}_m) &= \mathbf{E}(\mathbf{E}(f(x_n)g(x_\ell) \,|\, \mathscr{F}_m \vee \mathscr{F}_n) \,|\, \mathscr{F}_m) = \mathbf{E}(f(x_n)\mathbf{E}(g(x_\ell) \,|\, \mathscr{F}_m \vee \mathscr{F}_n) \,|\, \mathscr{F}_m) \\ &= \mathbf{E}(f(x_n)\mathbf{E}(g(x_\ell) \,|\, \mathscr{F}_m) \,|\, \mathscr{F}_m) = \mathbf{E}(g(x_\ell) \,|\, \mathscr{F}_m)\, \mathbf{E}(f(x_n) \,|\, \mathscr{F}_m) \;,\end{aligned}$$

and so *(iii)* holds. To show the converse, fix arbitrary functions f, g and h. One then has

$$\begin{aligned}&\mathbf{E}(g(x_\ell)h(x_m)\mathbf{E}(f(x_n) \,|\, \mathscr{F}_\ell \vee \mathscr{F}_m)) = \mathbf{E}(g(x_\ell)h(x_m)f(x_n)) = \mathbf{E}(h(x_m)\mathbf{E}(g(x_\ell)f(x_n) \,|\, \mathscr{F}_m)) \\ &\quad = \mathbf{E}(h(x_m)\mathbf{E}(g(x_\ell) \,|\, \mathscr{F}_m)\mathbf{E}(f(x_n) \,|\, \mathscr{F}_m)) = \mathbf{E}(\mathbf{E}(h(x_m)g(x_\ell)\mathbf{E}(f(x_n) \,|\, \mathscr{F}_m) \,|\, \mathscr{F}_m)) \\ &\quad = \mathbf{E}(h(x_m)g(x_\ell)\mathbf{E}(f(x_n) \,|\, \mathscr{F}_m)) \;.\end{aligned}$$

Since g and h are arbitrary, this shows that one must have $\mathbf{E}(f(x_n \,|\, \mathscr{F}_\ell \vee \mathscr{F}_m) = \mathbf{E}(f(x_n) \,|\, \mathscr{F}_m)$ (almost surely). □

Intuitively, property *(iii)* means that the future of the process is independent of its past, provided that we know the present.

Remark 2.24 It follows from Exercise 2.21 that every Markov properties satisfies the properties of the last theorem. It was however proven in [FWY00] that the converse is not true, *i.e.* there exist processes that satisfy the three (equivalent) properties above but fail to be Markov.

Definition 2.25 A Markov process is **time-homogeneous** if there exists a measurable map P from $\mathcal{X}$ into $\mathcal{P}(\mathcal{X})$, the space of probability measures on $\mathcal{X}$, such that

$$\mathbf{P}(x_n \in A \,|\, x_{n-1} = a) = (P(a))(A) \;,$$

for every $A \in \mathscr{B}(\mathcal{X})$, almost every $a \in \mathcal{X}$, and every $n > 0$. We will from now on use the notation $(P(a))(A) = P(a, A)$ and we call P the **transition probabilities** for x.

Example 2.26 Let $\mathcal{X} = \mathbf{R}$, let $\{\xi_n\}_{n\geq 0}$ be an i.i.d. sequence of Normally distributed random variables, and let $\alpha, \beta \in \mathbf{R}$ be fixed. Then, the process defined by $x_0 = \xi_0$ and $x_{n+1} = \alpha x_n + \beta\xi_{n+1}$ is Markov. Its transition probabilities are given by

$$P(x, dy) = \frac{1}{\sqrt{2\pi}\beta} \exp\Big(\frac{-(y-\alpha x)^2}{2\beta^2}\Big)\, dy \;.$$

Note that if $\alpha^2 + \beta^2 = 1$, the law of x_n is independent of n.

Example 2.27 Let $F\colon \mathcal{X} \to \mathcal{X}$ be an arbitrary measurable map and consider an arbitrary probability measure μ on $\mathcal{X}$. Then, the stochastic process obtained by choosing x_0 randomly in $\mathcal{X}$ with law μ and defining recursively $x_{n+1} = F(x_n)$ is a Markov process. Its transition probabilities are given by $P(x, \cdot) = \delta_{F(x)}$.

We will only consider time-homogeneous Markov processes from now on.

Exercise 2.28 Let ξ_n be a sequence of real-valued i.i.d. random variables and define x_n recursively by $x_0 = 0$, $x_n = \alpha x_{n-1} + \xi_n$. Sow that x defined in this way is a time-homogeneous Markov process and write its transition probabilities in the cases where (1) the ξ_n are Bernoulli random variables (*i.e.* $\xi_n = 0$ with probability $1/2$ and $\xi_n = 1$ otherwise) and (2) the law of ξ_n has a density p with respect to the Lebesgue measure on $\mathbf{R}$.

In the case (1) with $\alpha < 1/2$, what does the law of x_n look like for large values of n?

The following result is fundamental to the description of Markov processes:

Theorem 2.29 *Let x be a time-homogeneous Markov process with transition probabilities P. Then, one has*

$$\mathbf{P}(x_n \in A \,|\, x_0 = a) = P^n(a, A)\,, \tag{2.2}$$

where P^n is defined recursively by

$$P^1 = P\,, \quad P^n(a, A) = \int_{\mathcal{X}} P(x, A)\, P^{n-1}(a, dx)\,. \tag{2.3}$$

Equation (2.3) is called the Chapman-Kolmogorov equation.

Proof. The proof goes by induction. The statement is true for $n = 1$ by Definition 2.25. Assume that it holds for $n = k \geq 1$. We then have

$$\begin{aligned}\mathbf{P}(x_{k+1} \in A \,|\, \mathscr{F}_0) &= \mathbf{E}(\chi_A(x_{k+1}) \,|\, \mathscr{F}_0) = \mathbf{E}(\mathbf{E}(\chi_A(x_{k+1}) \,|\, \mathscr{F}_0 \vee \mathscr{F}_k) \,|\, \mathscr{F}_0) \\ &= \mathbf{E}(\mathbf{E}(\chi_A(x_{k+1}) \,|\, \mathscr{F}_k) \,|\, \mathscr{F}_0) = \mathbf{E}(P(x_k, A) \,|\, \mathscr{F}_0)\,.\end{aligned}$$

Since, by assumption, the law of x_k conditioned on $x_0 = a$ is given by $P^k(a, \cdot\,)$, the claim follows. □

Exercise 2.30 Check that $P^{n+m}(a, A) = \int_{\mathcal{X}} P^n(x, A)\, P^m(a, dx)$ for every $n, m \geq 1$.

Definition 2.31 Given transition probabilities P, we define a **transition operator** T on $\mathcal{P}(\mathcal{X})$ by

$$(T\mu)(A) = \int_{\mathcal{X}} P(x, A)\, \mu(dx)\,. \tag{2.4}$$

Note that T can be extended to the space of all signed measures by linearity.

Exercise 2.32 Check that the operator T^n obtained by replacing P by P^n in (2.4) is equal to the operator obtained by applying T n times, $T^n = T \circ T \circ \ldots \circ T$.

Exercise 2.33 Show that if the state space $\mathcal{X}$ is countable and T is an arbitrary linear operator on the space of finite signed measures which maps probability measures into probability measures, then T is of the form (2.4) for some P.

Exercise 2.34 ($\star$) Show that the conclusions of Exercise 2.33 still hold under the assumptions that $\mathcal{X}$ is a complete separable metric space and T is continuous in the weak topology.

Hint Use the fact that with these assumptions, every probability measure can be approximated in the weak topology by a finite sum of δ-measures (with some weigths).

We similarly define an operator $T_\star$ on the space of bounded measurable functions from $\mathcal{X}$ to $\mathbf{R}$ by

$$(T_\star f)(x) = \mathbf{E}(f(x_1) \,|\, x_0 = x) = \int_{\mathcal{X}} f(y)\, P(x, dy)\ .$$

Note that one always has $T_\star 1 = 1$.

Exercise 2.35 Check that the operators T and $T_\star$ are each other's dual, *i.e.* that

$$\int_{\mathcal{X}} (T_\star f)(x)\, \mu(dx) = \int_{\mathcal{X}} f(x)\, (T\mu)(dx)$$

holds for every probability measure μ and every bounded function f.

Definition 2.36 We say that a homogeneous Markov process with transition operator $T_\star$ is **Feller** if $T_\star f$ is continuous whenever f is continuous and bounded. It is **strong Feller** if $T_\star f$ is continuous whenever f is measurable and bounded.

It is a legitimate question to ask whether any such function P can be used to construct a corresponding time-homogeneous Markov process. This can be answered affirmatively by making use of the following result from probability theory which will not be proven here:

Theorem 2.37 (Kolmogorov's extension theorem) *Let $\{\mu_n\}$ be a sequence of probability measures on $\mathcal{X}^n$ such that*

$$\mu_n(A_1 \times A_2 \times \ldots \times A_n) = \mu_{n+1}(A_1 \times A_2 \times \ldots \times A_n \times \mathcal{X}) \tag{2.5}$$

for every n and every sequence of Borel sets A_i. Then there exists a unique probabiliy measure μ on $\mathcal{X}^\infty$ such that $\mu_n(A) = \mu(A \times \mathcal{X}^\infty)$.

As a corollary, we get the following result:

Proposition 2.38 *Let P be a measurable map from $\mathcal{X}$ to $\mathcal{P}(\mathcal{X})$ and let μ_0 be a probability measure on $\mathcal{X}$. Then, there exists a (unique in law) Markov process x with transition probabilities P such that the law of x_0 is μ_0.*

Proof. Define the sequence of measures μ_n on $\mathcal{X}^n$ by

$$\mu_{n+1}(A_0 \times \ldots \times A_n) = \int_{A_0}\int_{A_1}\int_{A_2}\cdots\int_{A_{n-2}}\int_{A_{n-1}} P(x_{n-1}, A_n)P(x_{n-2}, dx_{n-1})\cdots P(x_1, dx_2)P(x_0, dx_1)\mu(dx_0)\ .$$

It is easy to check that this sequence of measures satisfies (2.5). By Kolmogorov's extension theorem, there thus exists a unique measure μ on $\mathcal{X}^\infty$ such that the restriction of μ to $\mathcal{X}^n$ is given by μ_n. We now choose $\Omega = \mathcal{X}^\infty$ as our probability space equipped with the probability measure $\mathbf{P} = \mu$. We define the process x as the canonical process, *i.e.* $x_n(w_0, w_1, \ldots) = w_n$.

It is straightforward to check that x is a Markov process with the required transition probabilities (and such that the law of x_0 is μ_0). This concludes the 'existence' part. The uniqueness follows from the 'uniqueness' part of Kolmogorov's extension theorem, since one can show by induction that the law of $(x_0, \ldots, x_n)$ must be equal to μ_{n+1} for every n. □

2.3 Stopping times

Definition 2.39 Given a Markov process x, an integer-valued random variable T is called a **stopping time** for x, if the event $\{T = n\}$ is $\mathscr{F}_0^n$-measurable for every $n \geq 0$. (The value $T = \infty$ is usually allowed as well and no condition is imposed on its measurability.)

Exercise 2.40 Show that the above definition is equivalent to the same definition with $\{T = n\}$ replaced by $\{T \leq n\}$.

Given a stopping time T and a Markov process x we introduce the stopped process $x_{n\wedge T}$ by

$$x_{n\wedge T} = \begin{cases} x_n & \text{if } n \leq T, \\ x_T & \text{otherwise.} \end{cases}$$

We denote by $\mathscr{F}_T = \mathscr{F}_T^T$ the σ-algebra generated by x_T and by $\mathscr{F}_m^T$ the σ-algebra generated by the collection $\{x_{n\wedge T}\}_{n\geq m}$.

Exercise 2.41 Show that this notation is consistent with the one introduced in Section 2.2 in the sense that if $T = m$ almost surely for some m, then one has $\mathscr{F}_n^T = \mathscr{F}_n^m$ and $\mathscr{F}_T = \mathscr{F}_m$.

Proposition 2.42 *Let T be a random variable taking a countable number of values t_i and let X and Y be random variables such that there exist countable families of random variables X_i and Y_i such that $X(\omega) = X_i(\omega)$ if $T(\omega) = t_i$ and $Y(\omega) = Y_i(\omega)$ if $T(\omega) = t_i$. Then, one has $\mathbf{E}(X|Y\&T) = \mathbf{E}(X_i|Y_i\&T)$ on the set $T(\omega) = t_i$.*

Proof. Denote $\Omega_i = \{\omega \,|\, T(\omega) = t_i\}$ and introduce the σ-algebras $\mathcal{F}_i = \sigma\{Y^{-1}(A) \cap \Omega_i\}$. Note that since Y and Y_i coincide on Ω_i, one also has $\mathcal{F}_i = \sigma\{Y_i^{-1}(A) \cap \Omega_i\}$.

Using the notation $X|_A$ for the random variable which is equal to X on A and equal to 0 otherwise, one has

$$\mathbf{E}(X|Y\&T)|_{\Omega_i} = \mathbf{E}(X|_{\Omega_i}|\mathcal{F}_i)\ .$$

On the other hand, one has

$$\mathbf{E}(X_i|Y_i\&T)|_{\Omega_i} = \mathbf{E}(X_i|_{\Omega_i}|\mathcal{F}_i) = \mathbf{E}(X|_{\Omega_i}|\mathcal{F}_i)\ ,$$

which completes the proof. □

The interest of the definition of a stopping time is that if T is a stopping time for a time-homogeneous Markov process x, then the process x_{T+n} is again a Markov process with the same transition probabilities. Stopping times can therefore be considered as times where the process x "starts afresh". More precisely, we have the following theorem:

Theorem 2.43 (strong Markov property) *Let x be a time-homogeneous Markov process with transition probabilities P and let T be a stopping time which is almost-surely finite. Then, the process $\tilde{x}_n = x_{T+n}$ is also Markov with transition probabilities P and one has*

$$\mathbf{E}(f(\tilde{x})\,|\,\mathscr{F}_0^T) = \mathbf{E}(f(\tilde{x})\,|\,\mathscr{F}_T)\ , \tag{2.6}$$

for every measurable $f : \mathcal{X}^\infty \to \mathbf{R}$.

Proof. Let us first show that $\tilde{x}$ is Markov. We have indeed

$$\begin{aligned}\mathbf{P}(\tilde{x}_n \in A \,|\, \tilde{x}_0 = a_0, \dots, \tilde{x}_{n-1} = a_{n-1}) \\ &= \sum_{m\geq 0} \mathbf{P}(\tilde{x}_n \in A \,|\, \tilde{x}_0 = a_0, \dots, \tilde{x}_{n-1} = a_{n-1} \,\&\, T = m)\mathbf{P}(T = m) \\ &= \sum_{m\geq 0} \mathbf{P}(\tilde{x}_n \in A \,|\, x_m = a_0, \dots, x_{m+n-1} = a_{n-1} \,\&\, T = m)\mathbf{P}(T = m) \\ &= \sum_{m\geq 0} \mathbf{P}(x_{n+m} \in A \,|\, x_{n+m-1} = a_{n-1})\mathbf{P}(T = m) = \mathbf{P}(a_{n-1}, A)\;.\end{aligned}$$

Here we used Proposition 2.42 to go from the second to the third line and the Markov property of x to obtain the last line.

For every measurable set $A \subset \mathcal{X}$, one has

$$\begin{aligned}\mathbf{P}(x_{T+1} \in A \,|\, x_T = a) &= \sum_{n\geq 0} \mathbf{P}(x_{T+1} \in A \,|\, x_T = a \;\&\; T = n)\, \mathbf{P}(T = n) \\ &= \sum_{n\geq 0} \mathbf{P}(x_{n+1} \in A \,|\, x_n = a \;\&\; T = n)\, \mathbf{P}(T = n)\;.\end{aligned}$$

Since T is a stopping time, the event $\{x_n = a \;\&\; T = n\}$ is in $\mathscr{F}_0^n$. Furthermore, the Markov property for x ensures that the function $\mathbf{P}(x_{n+1} \in A \,|\, \mathscr{F}_0^n)$ only depends on x_n and is equal to $P(x_n, A)$. Therefore, one has

$$\mathbf{P}(x_{T+1} \in A \,|\, x_T = a) = \sum_{n\geq 0} P(a, A)\, \mathbf{P}(T = n) = P(a, A)\;.$$

Since the above reasoning still holds if, instead of fixing x_T one fixes the whole stopped process $x_{n\wedge T}$, one actually has

$$\mathbf{P}(x_{T+1} \in A \,|\, \mathscr{F}_0^T) = P(x_T, A)\;.$$

This shows that (2.6) holds if f depends only on the first coordinate. The whole argument can however be repeated for any expression of the type

$$\mathbf{P}(x_{T+j} \in A_j \;\forall j \geq 0 \,|\, \mathscr{F}_0^T)\;,$$

thus leading to the stated result. □

Exercise 2.44 Prove that if $T = \infty$ is allowed, then the process x_{T+n} conditioned on $\{T < \infty\}$ (it is undefined outside of that set) is again Markov with transition probabilities P.

3 Finite state space

In this section, we assume that the space $\mathcal{X}$ is finite, so we identify it with $\{1, \dots, N\}$ for some $N > 0$. In this case, the space of signed measures is identified in a natural way with $\mathbf{R}^N$ in the following way. Given a measure μ on $\mathcal{X}$, we associate to it the vector $a \in \mathbf{R}^N$ by $a_i = \mu(\{i\})$. Reciprocally, given $a \in \mathbf{R}^N$, we associate to it a measure μ by $\mu(A) = \sum_{i\in A} a_i$. From now on, we will therefore use the terms "vector" and "measure" interchangeably and use the notation $\mu_i = \mu(i) = \mu(\{i\})$.

The set of probability measures on $\mathcal{X}$ is thus identified with the set of vectors in $\mathbf{R}^N$ which have non-negative entries that sum up to 1. In this context, a transition operator is a linear operator from

$\mathbf{R}^N$ to $\mathbf{R}^N$ which preserves probability measures. Such operators are given by $N \times N$ matrices (P_{ij}) with positive entries such that

$$\sum_{i=1}^{N} P_{ij} = 1 \ , \quad \text{for all } j. \tag{3.1}$$

The number P_{ij} should be interpreted as the probability of jumping from state j to state i.

Definition 3.1 We call a matrix P with positive entries which satisfies (3.1) a **stochastic matrix**.

Exercise 3.2 Given a vector $\mu \in \mathbf{C}^N$, we write $|\mu|$ for the vector with entries $|\mu_i|$ and $\sum(\mu)$ for the number $\sum_{i=1}^{N} \mu_i$. Show that if P is a stochastic matrix, then one has $\sum(P\mu) = \sum(\mu)$ and $\sum(|P\mu|) \leq \sum(|\mu|)$.

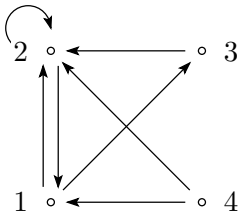


Figure 1: Graph for P.

We can associate to such a matrix P_{ij} an oriented graph, called the **incidence graph** of P by taking $\mathcal{X} = \{1, \ldots, N\}$ as the set of vertices and by saying that there is an oriented edge going from i to j if and only if $P_{ji} \neq 0$. For example, if

$$P = \frac{1}{10} \begin{pmatrix} 0 & 3 & 0 & 2 \\ 5 & 7 & 10 & 8 \\ 5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \tag{3.2}$$

then the associated graph is given by the one in Figure 1. Note that the 4th row of P is zero, which implies that the vertex 4 can not be reached by any walk on the graph that follows the arrows.

3.1 Irreducible matrices

Definition 3.3 We call a transition matrix P **irreducible** if it is possible to go from any point to any point of the associated graph by following the arrows. Otherwise, we call it **reducible**.

At an intuitive level, being irreducible means that every point will be visited by our Markov process. Otherwise, the state space can be split into several sets A_i such a way that if one starts the process in A_i it stays in A_i forever and if one starts it outside of the A_i's it will eventually enter one of them. For example, the matrix given in (3.2) is reducible because it is impossible to reach 4 from any of the other points in the system.

For every state $i = 1, \ldots, N$, we define the set $R(i)$ of **return times** to i by

$$R(i) = \{n > 0 \,|\, (P^n)_{ii} > 0\} \ .$$

In other words, $R(i)$ contains the lengths of all possible paths (on the incidence graph) that connect i to itself. Note that $R(i)$ has the property that if n and m belong to it, then $n + m$ belongs to it as well.

The **period** of the state i is then defined by

$$p(i) = \gcd R(i) \ .$$

We call a stochastic matrix **aperiodic** if $p(i) = 1$ for every i. We call it **periodic** of period p if $p(i) = p > 1$ for every i.

The following result is well-known in number theory:

Proposition 3.4 *There exists $K > 0$ such that $kp(i) \in R(i)$ for every $k \geq K$*

Proof. By dividing everything by $p(i)$, we can assume without loss that $p(i) = 1$. Since $\gcd R(i) = 1$, there exists a finite collection $p_1, \dots, p_n$ in $R(i)$ such that $\gcd\{p_1, \dots, p_n\} = 1$. The Euclidean algorithm implies that there exist integers $a_1, \dots, a_n$ such that $\sum_{i=1}^n a_i p_i = 1$. Set $P = \sum p_i$. Then, for $k = 1, \dots, P$, one has

$$NP + k = \sum_{i=1}^{n} (N + ka_i) p_i \ .$$

This shows that $NP + k \in R(i)$ for every $k \in \{0, \dots, P\}$ and every $N \geq N_0$ with $N_0 = P \max\{|a_1|, \dots, |a_n|\}$. Therefore, the claim holds with $K = N_0 P$. □

As a consequence of this result, we can show

Proposition 3.5 *An irreducible stochastic matrix is either aperiodic or of period p for some p.*

Proof. It suffices to show that $p(i) = p(j)$ for every pair i, j. Since P is irreducible, there exist n and m such that $(P^n)_{ij} > 0$ and $(P^m)_{ji} > 0$. Setting $N = n + m$, this implies that $N \in R(i) \cap R(j)$ and so $p(i)$ divides N and $p(j)$ divides N. Since one has $N + R(i) \subset R(j)$, this implies that $p(j)$ divides $p(i)$. On the other hand, the same is true with i and j exchanged, so that one must have $p(i) = p(j)$. □

Note that a different characterisation of stochastic matrices with period p is the following;

Lemma 3.6 *A stochastic matrix P is periodic with period p if and only if it is possible to write $\{1, \dots, N\}$ as a disjoint union of sets $A_0 \sqcup \dots \sqcup A_{p-1}$ in such a way that if $P_{ji} \neq 0$ for a pair of indices i and j, then $i \in A_n$ and $j \in A_m$ with $m = n + 1$ (mod p).*

Proof. Assume for the moment that P is irreducible and define A_n by

$$A_n = \{j \,|\, \exists\, m = n \ (\text{mod } p) \text{ such that } P^m_{j1} > 0\} \ .$$

The choice of the index 1 is arbitrary, this just determines that $1 \in A_0$. Since Ω is assumed to be irreducible, the union of the A_n is all of $\{1, \dots, N\}$. Furthermore, they are disjoint. Otherwise, one could find j such that $P^n_{j1} > 0$ and $P^m_{j1} > 0$ with $m \neq n$ (mod p). Since P is irreducible, there exists furthermore q such that $P^q_{1j} > 0$, so that $n + q \in R(1)$ and $m + q \in R(1)$. This contradicts the fact that P is periodic of period p. The fact that these sets have the required property is then immediate.

If P is not irreducible, it suffices to note that the definition of periodicity implies that $\{1, \dots, N\}$ can be broken into sets $B_1, \dots, B_k$ with the property that $P_{ij} = 0$ if $i \in B_k$, $j \in B_\ell$ and $k \neq \ell$. The restriction of P to each of these sets is then irreducible, so that the result follows from the irreducible case. □

Exercise 3.7 Let P be irreducible of period p. Show that, for $n \geq 1$, the period q of P^n is given by $q = p/r$, where r is the greatest common divider between p and n. The corresponding partition $\{B_i\}$ of $\{1, \dots, N\}$ is given by $B_i = \bigcup_{n \geq 0} A_{i+nq \,(\text{mod } p)}$, where $\{A_i\}$ is the partition associated to P by Lemma 3.6.

Exercise 3.8 Consider an irreducible stochastic matrix P and an arbitrary partition $\{B_j\}_{j=0}^{q-1}$ of $\{1,\dots,N\}$ such that if $i \in B_n$ and $j \in B_m$ with $m \neq n+1 \pmod q$, then $P_{ji} = 0$. Show that q must be a divider of p and that the partition $\{B_j\}$ is the one associated by Lemma 3.6 to the matrix $P^{p/q}$.

Exercise 3.9 Show that the three following conditions are equivalent:

(a) P is irreducible and aperiodic.

(b) P^n is irreducible for every $n \geq 1$.

(c) There exists $n \geq 1$ such that $(P^n)_{ij} > 0$ for every $i, j = 1, \dots, N$.

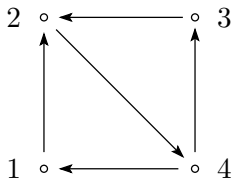


Figure 2: Periodic.

The example given in (3.2) is aperiodic. However the example shown in Figure 2 is periodic with period 3. In this particular case, one can take $A_0 = \{2\}$, $A_1 = \{1,3\}$, and $A_2 = \{4\}$. Note that this choice is unique (up to permutations of course). Note also that even though P is irreducible, P^3 is not. This is a general fact for periodic processes. Stochastic matrices such that the corresponding incidence graph is given by Figure 2 are of the form

$$P = \begin{pmatrix} 0 & 0 & 0 & q \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1-q \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

for some $q \in (0,1)$.

Exercise 3.10 Prove that if there exists j such that $P_{jj} \neq 0$, then the matrix is aperiodic.

Theorem 3.11 (Perron-Frobenius) *If P is irreducible, then there exists exactly one eigenvector π with $P\pi = \pi$. Furthermore, π can be chosen such that all its entries are strictly positive. If P is aperiodic, all other eigenvalues satisfy $|\lambda| < 1$. If P is periodic with period p, there are eigenvalues $\lambda_j = e^{\frac{2i\pi j}{p}}$ and the associated eigenvectors μ_j satisfy*

$$\mu_j(n) = e^{-2i\pi\frac{jk}{p}}\pi(n)\,, \quad \text{if } n \in A_k, \tag{3.3}$$

where π is the (only) eigenvector with eigenvalue 1 and the sets A_k are the ones associated to P by Lemma 3.6.

Proof. Since $\|P\mu\|_1 \leq \|\mu\|_1$ for every vector $\mu \in \mathbf{C}^N$ (see Exercise 3.2), the eigenvalues of P must all satisfy $|\lambda| \leq 1$. Since the vector $\mathbf{1} = \frac{1}{N}(1,1,\dots,1)$ is an eigenvector with eigenvalue 1 for P^T, there exists an eigenvector with eigenvalue 1 for P, let us call it π. Since P is real, we can choose π to be real too. Let us now prove that π can be chosen positive as well.

Define the matrix $T^n = \frac{1}{n}(P + P^2 + \ldots + P^n)$. Clearly T^n is again a stochastic matrix and π is an eigenvector of T^n with eigenvalue 1. Since P is irreducible, there exists n and $\delta > 0$ such that $T^n_{ij} \geq \delta$ for every i and j. Write now π_+ for the positive part of π and π_- for its negative part. We also define $\alpha = \min\{\|\pi_+\|_1, \|\pi_-\|_1\}$. It is clear that one has $T^n\pi_+ \geq \delta\alpha\mathbf{1}$ and $T^n\pi_- \geq \delta\alpha\mathbf{1}$. Therefore,

$$\begin{aligned} \|T^n\pi\|_1 = \|T^n\pi_+ - T^n\pi_-\|_1 &\leq \|T^n\pi_+ - \delta\alpha\mathbf{1}\|_1 + \|T^n\pi_- - \delta\alpha\mathbf{1}\|_1 \\ &\leq \|\pi_+\|_1 + \|\pi_-\|_1 - 2\delta\alpha = \|\pi\|_1 - 2\delta\alpha\,. \end{aligned}$$

Since $T^n\pi = \pi$ and $\delta > 0$, one must have $\alpha = 0$, which implies that π is either entirely positive or entirely negative (in which case $-\pi$ is entirely positive).

From now on, we normalise π in such a way that $\sum(\pi) = \sum(|\pi|) = 1$. All entries of π are strictly positive since $\pi = T^n\pi \geq \delta\mathbf{1}$. The fact that exists only one π (up to multiplication by a scalar) such that $P\pi = \pi$ is now easy. Assume that $P\pi_1 = \pi_1$ and $P\pi_2 = \pi_2$. By the previous argument, we can assume that the entries of the π_i are positive sum to 1. Then the vector $\pi_3 = \pi_1 - \pi_2$ is also an eigenvector with eigenvalue 1 for P. However, since $\sum(\pi_3) = 0$, one must have $\pi_3 = 0$. From now on, we call the unique positive eigenvector with eigenvalue 1 of an irreducible stochastic matrix P the **Perron-Frobenius vector of** P.

It remains to consider eigenvalues with $|\lambda| = 1$ but $\lambda \neq 1$. Denote by ν an eigenvector for the eigenvalue $e^{i\theta}$. We write the components of ν in the form $\nu_i = r_i e^{i\theta_i}$ for $r_i \geq 0$ and we normalise them in such a way that $\sum r_i = 1$. The relation $\sum_{j=1}^N P_{kj}\nu_j = e^{i\theta}\nu_k$ then translates into

$$\sum_{j=1}^N e^{i\theta_j} P_{kj} r_j = e^{i(\theta+\theta_k)} r_k \; . \tag{3.4}$$

Multiplying both sides by $e^{-i(\theta+\theta_k)}$ and summing up yields $\sum_{j,k=1}^N e^{i(\theta_j-\theta_k-\theta)} P_{kj} r_j = 1$. On the other hand, we know that $P_{kj} r_j \geq 0$ and that $\sum_{j,k=1}^N P_{kj} r_j = 1$. This implies that

$$e^{i\theta_k} = e^{i(\theta_j-\theta)} \; , \quad \text{for every } j \text{ and } k \text{ such that } P_{kj} \neq 0. \tag{3.5}$$

Combining this with (3.4) in turn implies that $r = \pi$, the Perron-Frobenius vector. By multiplying ν with a scalar, we can assume that $\theta_1 = 0$. Since P is irreducible, the relation (3.5) then determines every θ_j uniquely provided that we know θ. On the other hand, (3.5) shows that one must have $N\theta = 0 \pmod{2\pi}$ for every $N \in R(i)$ (and for every i). It follows from Proposition 3.4 that $R(i)$ contains every large enough multiple of p, so that one must have $\theta = \frac{2\pi j}{p}$ for some integer j, so that the corresponding eigenvector ν is of the form (3.3). □

Remark 3.12 The Perron-Frobenius vector π has a very specific interpretation. We see that if we construct a Markov process x_n with transition probabilities P and such that the law of x_0 is π, then the law of x_n is π for every $n \geq 0$ as well. For this reason, we will also call it the **invariant measure** of P.

A very important consequence of the Perron-Frobenius theorem is the following

Theorem 3.13 *Let P be irreducible and aperiodic and let π be its Perron-Frobenius vector. Then, for any probability measure $\nu \in \mathbf{R}^N$, one has* $\lim_{n\to\infty} P^n\nu = \pi$.

Proof. Let us denote by $\|\mu\|_1 = \sum_i |\mu_i|$ the L^1-norm of a vector. It follows from Exercise 3.9 that there exist values $n > 0$ and $\delta \in (0,1)$ such that $P^n\eta \geq \delta\|\eta\|_1\mathbf{1}$ for every positive vector η. Write now $(\pi-\nu)_+$ for the positive part of $\pi - \nu$ and similarly for its negative part. Note also that $\|(\pi-\nu)_+\|_1 = \|(\pi-\nu)_-\|_1 = \frac{1}{2}\|\pi-\nu\|_1$. One then has

$$\begin{aligned}
\|P^n\nu - \pi\|_1 &= \|P^n(\pi-\nu)\|_1 = \|P^n(\pi-\nu)_+ - P^n(\pi-\nu)_-\|_1 \\
&\leq \|P^n(\pi-\nu)_+ - \delta\|(\pi-\nu)_+\|_1\mathbf{1}\|_1 + \|P^n(\pi-\nu)_- - \delta\|(\pi-\nu)_-\|_1\mathbf{1}\|_1 \\
&\leq (1-\delta)\|\pi-\nu\|_1 \; .
\end{aligned}$$

Since ν was arbitrary, one gets $\|P^{kn}\nu - \pi\|_1 \leq (1-\delta)^k\|\pi-\nu\|_1$ by iterating this bound. □

Note that Theorem 3.13 also follows immediately from the fact that if P is irreducible and aperiodic, then all eigenvalues of P have modulus strictly smaller than 1, except for the isolated eigenvalue 1 with eigenvector π. The proof given above however has the advantage that it can be generalised in a straightforward way to situations where the state space is not finite.

Exercise 3.14 Show that the conclusion of Theorem 3.13 also hold if one only assumes that $\sum_i \nu_i = 1$.

3.2 The general case

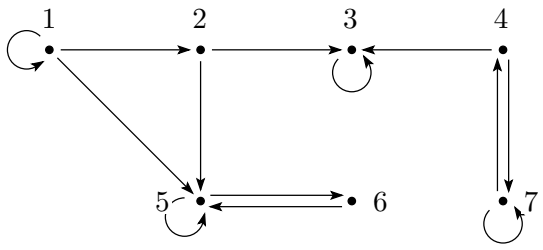
A general stochastic matrix is not irreducible. It can however be broken up into irreducible components in the following way. Fix an arbitrary stochastic matrix P of dimension N and call Γ the associated directed graph. The set $\{1, \dots, N\}$ is then naturally endowed with an equivalence relation by saying that $i \sim j$ if and only if there is a path on Γ going from i to j and back to i (we make it an equivalence relation by writing $i \sim i$ regardless on whether $P_{ii} > 0$ or not). In terms of the matrix, this means that $i \sim j$ if and only if there exist $m, n \geq 0$ such that $(P^m)_{ij} > 0$ and $(P^n)_{ji} > 0$, with the convention that P^0 is the identity matrix.

We denote by $[i]$ the equivalence class of i under this relation and we call it the **communication class** of i. For example, in the case of (3.2), we have $[1] = \{1, 2, 3\}$ and $[4] = \{4\}$. The set of equivalence classes is endowed with a partial order $\leq$ by saying that $[i] \leq [j]$ if and only if there is a path on Γ going from j to i. In the above example, one has $[1] \leq [4]$. Note that this order is not total, so it may happen that one has neither $[i] \leq [j]$ nor $[j] \leq [i]$.

Exercise 3.15 Check that the relation $\leq$ defined above is indeed a partial order.

Definition 3.16 An equivalence class $[i]$ is **minimal** if there is no $[j]$ such that $[j] \leq [i]$ and $[j] \neq [i]$.

Consider a stochastic matrix such that the associated graph is given by



In this case, the communication classes are given by

$$[1] = \{1\}\,, \qquad [2] = \{2\}\,, \qquad [3] = \{3\}\,,$$
$$[4] = \{4, 7\}\,, \quad [5] = \{5, 6\}\,.$$

One furthermore has the relations $[5] < [2] < [1]$, $[3] < [4]$, and $[3] < [2] < [1]$. Note that $[4]$ and $[2]$ for instance are not comparable.

Definition 3.17 A state i such that $[i]$ is minimal is called **recurrent**. All other states are called **transient**.

By construction, we see that every Markov process $\{x_n\}$ with transition probabilities P satisfies $[x_{n+1}] \leq [x_n]$ for every n. It seems therefore reasonable that every Markov process with transition probabilities P eventually ends up in one of the recurrent states. This justifies the terminology "transient" for the other states, since they will only ever be visited a finite number of times. Before we prove this result, we give a definition.

Definition 3.18 An $N \times N$ matrix P with positive entries such that $\sum_i P_{ij} \leq 1$ for all j is a **substochastic** matrix.

Substochastic matrices are typically obtained when we restrict a stochastic matrix to a subset of indices. One has the following:

Lemma 3.19 *Let P be an irreducible substochastic matrix which is not a stochastic matrix. Then, $P^n\mu \to 0$ for every μ and the convergence is exponential. In particular, the eigenvalues of P are all of modulus strictly less than 1 and so $1 - P$ is invertible.*

Proof. It is sufficient to prove the claim for μ positive with norm 1. Define T^n as in the proof of the Perron-Frobenius theorem. Then, since $\|P\mu\|_1 \leq \|\mu\|_1$ for every positive vector μ, one has $\|P^{n+1}\mu\|_1 \leq \|PT^n\mu\|_1$ for every $n > 0$. Choose n such that $T^n\mu \geq \delta\mathbf{1}$ (such an n exists by the irreducibility of P). Since P is not a stochastic matrix, there exists $\alpha > 0$ and an index j such that $\sum_i P_{ij} \leq 1 - \alpha$. Therefore $\|P^{n+1}\mu\|_1 \leq \|PT^n\mu\|_1 \leq (1 - \alpha\delta)\|\mu\|_1$, which concludes the proof. □

This shows that

Theorem 3.20 *Let $\{x_n\}$ be a Markov process with transition probabilities P and let i be a transient state. Then the probability that $x_n \in [i]$ for an infinite number of values n is 0.*

Proof. Recall first the Borel-Cantelli lemma from probability theory:

Lemma 3.21 (Borel-Cantelli) *Let $\{A_n\}_{n\geq 0}$ be a sequence of events in a probability space Ω. If $\sum_n \mathbf{P}(A_n) < \infty$, then the probability that infinitely many of these events happen is 0.*

By the strong Markov property, it is sufficient to prove the theorem for the particular case when $x_0 = j$ for a state $j \in [i]$. We take as A_n the event $\{x_n \in [i]\}$. By the Borel-Cantelli lemma, the claim follows if we can show that

$$\sum_n \mathbf{P}(x_n \in [i]) = \sum_n \sum_{k\in[i]} (P^n)_{kj} < \infty \, .$$

Denote by $\tilde{P}$ the restriction of P to the indices in $[i]$. Then $\tilde{P}$ is an irreducible substochastic matrix and one has $(P^n)_{kj} = (\tilde{P}^n)_{kj}$ for $k, j \in [i]$. The result follows from Lemma 3.19. □

Exercise 3.22 Let P be an arbitrary stochastic matrix. Show that the set of all normalised positive vectors μ such that $P\mu = \mu$ consists of all convex linear combinations of the Perron-Frobenius vectors of the restrictions of P to its recurrent classes.

In order to conclude this subsection, let us give a formula for the probability that, starting from a given transient state, the Markov process will eventually end up in a given recurrence class. In order to somewhat simplify the argument, we assume that the recurrent classes consist of single

points, that the states 1 to R are recurrent, and that the states $R+1$ to $R+T$ are transient (set $N = T + R$). Therefore, the transition matrix P can be written as

$$P = \begin{pmatrix} I & S \\ 0 & Q \end{pmatrix},$$

where I is the identity and Q is some substochastic matrix (so that $(Q - I)$ is invertible).

Define now the matrix A_{ij} with $j \in \{1, \dots, T\}$ and $i \in \{1, \dots, R\}$ as the probability that the process starting at the transient state $R + j$ will eventually end up in the recurrent state i. One has

Proposition 3.23 *The matrix A is given by $A = S(I - Q)^{-1}$.*

Proof. One has

$$\begin{aligned} A_{ij} &= \mathbf{P}(\text{the process reaches } i \text{ eventually} \,|\, x_0 = R + j) \\ &= \sum_{k=1}^{T} Q_{kj}\, \mathbf{P}(\text{the process reaches } i \text{ eventually} \,|\, x_0 = R + k) + S_{ij} \\ &= \sum_{k=1}^{T} A_{ik} Q_{kj} + S_{ij}\,, \end{aligned}$$

where we used the Markov property to go from the first to the second line. In matrix notation, this reads $A = AQ + S$, and therefore $A = S(I - Q)^{-1}$. The invertibility of $(I - Q)$ is an immediate consequence of Lemma 3.19. □

3.3 Return times and the law of large numbers

In this section, we are interested in the following question: given a finite-state Markov process with transition probabilities P starting in a distinguished state i. How long does it take to get back to i? It may be rather surprising that this can easily be computed explicitely:

Theorem 3.24 *Let x be an aperiodic irreducible homogeneous Markov process on a finite state space $\mathcal{X}$ with invariant measure π and satisfying $x_0 = i$ almost surely for some distinguished state i. Let T be the random (stopping) time defined by*

$$T = \min\{n > 0 \text{ such that } x_n = i\}\,.$$

Then, one has $\mathbf{E}T = 1/\pi(i)$.

A closely related result is the Strong Law of Large Numbers for Markov processes. Let us recall the Strong Law of Large Numbers of probability theory:

Theorem 3.25 (Strong Law of Large Numbers) *Let $\{\xi_n\}_{n\geq 1}$ be a sequence of i.i.d. real-valued random variables such that $\mathbf{E}\xi_n = \mu < \infty$ and $\mathbf{E}(\xi - \mu)^2 = \sigma^2 < \infty$. Define $S_N = \frac{1}{N}\sum_{n=1}^{N} \xi_n$. Then, one has $\lim_{N\to\infty} S_N = \mu$ almost surely.*

Its simplest extension to Markov processes states:

Theorem 3.26 *Let x be an aperiodic irreducible homogeneous Markov process on a finite state space $\mathcal{X}$ with invariant measure π and let $f\colon \mathcal{X} \to \mathbf{R}$. Then, one has*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} f(x_n) = \sum_{j\in\mathcal{X}} f(j)\pi(j)\,, \tag{3.6}$$

almost surely.

Before we turn to the proof of these two results, let us make a preliminary calculation.

Lemma 3.27 *Let P be irreducible and let x be a Markov process with transition probabilities P. Fix two states i and j and define the stopping time T_i by*

$$T_i = \inf\{k \geq 0 \,|\, x_k = i\} \; .$$

Then, for every $p \geq 1$, the expectation $\mathbf{E}(T_i^p \,|\, x_0 = j)$ is finite.

Proof. Fix i as in the statement of the Lemma and denote by $\tilde{x}$ the process stopped at T_i. For every $n > 0$, define the quantity

$$Q_n = \sup_{j \neq i} \mathbf{P}(\tilde{x}_n \neq i \,|\, \tilde{x}_0 = j) \; .$$

Note in particular that $\mathbf{P}(T_i > n \,|\, x_0 = j) \leq Q_n$. The irreducibility of P implies that there exist $N > 0$ and $\delta > 0$ such that $Q_N \leq 1 - \delta$.

The strong Markov property yields the following bound:

$$\begin{aligned} Q_{kN} &= \sup_{j \neq i} \sum_{\ell \neq i} \mathbf{P}(\tilde{x}_{kN} \neq i \,|\, \tilde{x}_N = \ell) \mathbf{P}(\tilde{x}_N = \ell \,|\, \tilde{x}_0 = j) \leq \sup_{j \neq i} \sum_{\ell \neq i} Q_{(k-1)N} \mathbf{P}(\tilde{x}_N = \ell \,|\, \tilde{x}_0 = j) \\ &= Q_{(k-1)N} \sup_{j \neq i} \mathbf{P}(\tilde{x}_N \neq i \,|\, \tilde{x}_0 = j) = Q_{(k-1)N} Q_N \leq (1 - \delta) Q_{(k-1)N} \; . \end{aligned}$$

It follows that $Q_{kN} \leq (1 - \delta)^k$, so that there exist constants C and $\gamma > 0$ such that $Q_n \leq C e^{-\gamma n}$ for every $n \geq 0$. Therefore, one has

$$\begin{aligned} \mathbf{E}(T_i^p \,|\, x_0 = j) &= \sum_{n \geq 0} n^p \mathbf{P}(T_i = n \,|\, x_0 = j) \leq \sum_{n \geq 0} n^p \mathbf{P}(T_i > n - 1 \,|\, x_0 = j) \\ &\leq \sum_{n \geq 0} n^p Q_{n-1} \leq C \sum_{n \geq 0} n^p e^{-\gamma n} \; . \end{aligned}$$

This sum always converges, and so the result follows. □

We now give the reasoning that simultaneously proves both results stated at the beginning of this section.

Proof of Theorems 3.24 and 3.26. Let $\chi_i \colon \mathcal{X} \to \mathbf{R}$ be the indicator function for the set $\{i\}$. Since any function on $\mathcal{X}$ can be written as a finite linear combination of such functions, it suffices to consider Theorem 3.26 with $f = \chi_i$, so that the right-hand side is equal to $\pi(i)$.

Note that we have already proven in Theorem 3.13 that $\lim_{n \to \infty} \mathbf{E}\chi_i(x_n) = \pi(i)$ and therefore also that

$$\lim_{N \to \infty} \mathbf{E}\Big(\frac{1}{N} \sum_{n=1}^{N} \chi_i(x_n)\Big) = \pi(i) \; . \tag{3.7}$$

In order to get (3.6) it thus suffices to get rid of the expectation on the left-hand side. Define a sequence of stopping times T_n by $T_0 = -1$ and, recursively,

$$T_{n+1} = \min\{k > T_n \text{ such that } x_k = i\} \; .$$

The strong Markov property implies that (except for $T_1 - T_0$) the sequence of intervals $T_n - T_{n-1}$ consists of i.i.d. random variables with the same law as the time T considered in Theorem 3.24. Since $\mathbf{E}T^2 < \infty$ by Lemma 3.27, It follows from the Law of Large Numbers that

$$\lim_{n \to \infty} \frac{T_n}{n} = \mathbf{E}T \; ,$$

almost surely. Define now $E_n = n\mathbf{E}T$, so that one has $T_n \approx E_n$ for large values of n. Since $T_n \geq n$ by definition, one has $|\frac{E_n}{T_n} - 1| < 1 + \mathbf{E}T$, so that Lebesgue's dominated convergence theorem yields

$$\lim_{n\to\infty} \mathbf{E}\Big|\frac{E_n}{T_n} - 1\Big| = 0 \; . \tag{3.8}$$

Note that the definition of the times T_n yields the relation $\frac{n}{T_n} = \frac{1}{T_n}\sum_{k=0}^{T_n} \chi_i(x_k)$. We can rewrite this as

$$\frac{E_n}{T_n}\,\frac{1}{\mathbf{E}T} = \frac{1}{E_n}\sum_{k=1}^{E_n} \chi_i(x_k) + R_n \; , \tag{3.9}$$

where the rest term R_n satisfies

$$\begin{aligned} |R_n| &= \Big|\frac{1}{T_n}\sum_{k=1}^{T_n} \chi_i(x_k) - \frac{1}{E_n}\sum_{k=1}^{E_n} \chi_i(x_k)\Big| \leq \Big|\frac{1}{T_n}\Big(\sum_{k=1}^{T_n} \chi_i(x_k) - \sum_{k=1}^{E_n} \chi_i(x_k)\Big)\Big| \\ &\quad + \Big|\Big(\frac{1}{T_n} - \frac{1}{E_n}\Big)\sum_{k=1}^{E_n} \chi_i(x_k)\Big| \leq \Big|\frac{E_n - T_n}{T_n}\Big| + \Big|\frac{E_n}{T_n} - 1\Big| = 2\Big|\frac{E_n}{T_n} - 1\Big| \; . \end{aligned} \tag{3.10}$$

Taking expectations on both sides and using (3.7) and (3.8), we see that one has $\mathbf{E}T = 1/\pi(i)$, thus concluding the proof of Theorem 3.24.

On the other hand taking limits on both sides of (3.9) and using the fact that $E_n/T_n \to 1$ almost surely, we see that $\lim_{n\to\infty} \frac{1}{E_n}\sum_{k=1}^{E_n} \chi_i(x_k) = \pi(i)$ almost surely. By the same argument as in (3.10), it follows immeditely that one has $\lim_{N\to\infty} \frac{1}{N}\sum_{k=1}^{N} \chi_i(x_k) = \pi(i)$, thus concluding the proof of Theorem 3.26. □

Exercise 3.28 Show that the assumption that x is aperiodic is not needed in order to prove (3.7). Therefore, Theorems 3.24 and 3.26 hold for general irreducible Markov chains on a finite state space.

3.4 Random walks on finite groups and card shuffling

A very important particular case is that of a random walk on a finite group. Think of card shuffling: there are only a finite number of possible orders for a deck of card, so this is a Markov process on a finite set. However, this set has a natural group structure by identifying a deck of card with an element of the group of permutations and the Markov process respects this group structure in the following sense. The probability of going from e (the identity) to an element g of the permutation group is the same as the probability of going from an arbitrary element h to $g \cdot h$. This motivates the following definition:

Definition 3.29 Consider a group G and a Markov chain with transition matrix P on G. We say that the Markov chain is a **left-invariant random walk** on G if there exists a probability measure $\bar{P}$ on G such that $P_{gh} = \bar{P}(h^{-1}g)$. We call it **right-invariant** if the same statement holds with $P_{gh} = \bar{P}(gh^{-1})$ instead.

It is clear that if the group G happens to be abelian, right-invariant and left-invariant random walks are the same.

Exercise 3.30 Show that if $\{x_n\}$ is a left-invariant random walk, then $\{x_n^{-1}\}$ is a right-invariant random walk and find its transition probabilities.

Because of Exercise 3.30, it suffices to study one of the two types of random walks. Let us choose the left-invariant ones.

Exercise 3.31 Consider a random walk with transition matrix P on a finite group G and define $\Sigma = \{g \in G \,|\, \bar{P}(g) > 0\}$. Show that P is irreducible if and only if Σ generates G.

Exercise 3.32 Show that the normalised counting measure $\pi(g) = 1/|G|$ is an invariant measure for every random walk on G.

The most common example of a random walk on a finite group is card shuffling. Take a deck consisting of n cards. Then, the set of all possible states of the deck can be identified in an obvious way with the symmetric group S_n, *i.e.* the group of all possible permutations of n elements. When identifying a permutation with a bijective map from $\{1, \ldots, n\}$ into itself, the composition law on the group is simply the composition of maps.

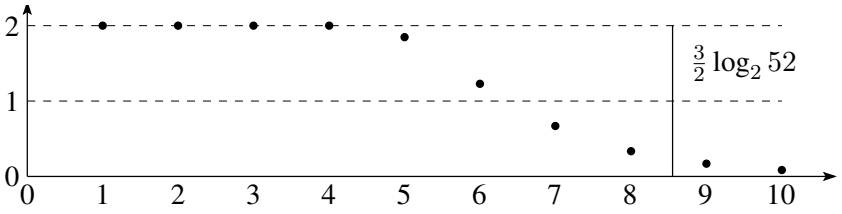
3.5 The Gilbert-Shannon-Reeds shuffling

A quite realistic way of shuffling a deck of n cards is the following. Assign 0 or 1 randomly and independently to each card. Then make a pile with all the cards marked 0 and another one with all the cards marked 1 (without changing the order of the cards within the pile) and put the two piles on top of each other. This is the definition of the inverse of a Gilbert-Shannon-Reeds shuffle. In this section, we will argue why the following result holds:

Theorem 3.33 *It takes about $\frac{3}{2}\log_2 n$ GSR shuffles to mix a deck of n cards.*

The precise formulation of Theorem 3.33 can be found in a 1992 paper by Bayer and Diaconis.

In principle, this approximation holds only for very large values of n. However, if we denote by π the uniform measure, by δ_e the measure concentrated on the identity, and by P the transition matrix associated to the GSR shuffle with 52 cards, one gets the following picture for $\|\pi - P^m\delta_e\|_1$ as a function of m:



Note that $\frac{3}{2}\log_2 52$ is quite a good approximation for the number of shuffles required to mix the deck.

A little thought shows that the inverse of m consecutive GSR shuffles can be constructed as follows. Make space for 2^m piles of cards on the table and place your deck of cards face up. Pick the cards one by one and place each of them face down onto one of the 2^m piles chosen uniformly and independently for each card. Finally, put each of the piles on top of each other starting with the first one. Using this characterisation of the inverse of m consecutive GSR shuffles, we will now give an explicit formula for the probability of m shuffles producing a given permutation σ. In order to state the formula, we introduce the concept of "rising sequences" for a permutation σ.

Definition 3.34 A **rising sequence** for a permutation σ of N elements is a collection of *consecutive* indices $A \subset \{1, \ldots, N\}$ such that σ is increasing on A. A rising sequence is **maximal** if it is not contained in any other rising sequence. The number of rising sequences of a given permutation is denoted by $R(\sigma)$.

Example 3.35 Consider the shuffle that brings an ordered deck of 5 cards in the configuration $(2,4,1,5,3)$. We associate to it the permutation $\sigma(1)=3$, $\sigma(2)=1$, $\sigma(3)=5$, $\sigma(4)=2$, $\sigma(5)=4$. This permutation contains three maximal rising sequences, $\{1\}$, $\{2,3\}$, and $\{4,5\}$, so that $R(\sigma)=3$. Note that even though σ is increasing on $\{2,4,5\}$, this is not a rising sequence because the indices are not consecutive.

Theorem 3.36 *The probability that m GSR shuffles of a deck of n cards produce a given permutation σ is given by*

$$P(\sigma)=\frac{1}{2^{mn}}\binom{2^m+n-R(\sigma)}{n}, \tag{3.11}$$

where we use the convention $\binom{a}{b}=0$ if $a<b$.

Proof. Take the example of $n=5$, $m=2$ and σ as in Example 3.35. In this case, we want to find a sequence of 2 inverse GSR shuffles that map $(2,4,1,5,3)$ into $(1,2,3,4,5)$. An inverse GSR shuffle is characterised in this case by a sequence of numbers $N_i\in\{1,\dots,4\}$ which say in which pile the card i ends up. There are obviously 2^{nm} such inverse shuffles. In order to get a perfectly ordered card deck at the end, one certainly needs that $N_i\le N_j$ if $i\le j$. Furthermore, we need in our example that $N_1\neq N_2$ and that $N_3\neq N_4$. In this particular case, the list of all possible GSR shuffles (written in the format $(N_1N_2N_3N_4N_5)$) that produce the right permutation is thus given by

$$(12233)\quad(12344)\quad(12234)\quad(12244)\quad(13344)\quad(23344)\,.$$

This is consistent with (3.11) which predicts $\binom{4+5-3}{5}=6$.

In the general case, the number of GSR shuffles which yields a given permutation σ is given by the number of increasing functions $N:\{1,\dots,n\}\to\{1,\dots,2^m\}$ that have jumps of size at least 1 at $R(\sigma)-1$ prescribed places. Of course no such function exists if $R(\sigma)>2^m$, which is consistent with the convention taken in (3.11). Subtracting the function that jumps by 1 at these places, this is the same as the number of increasing functions $N:\{1,\dots,n\}\to\{1,\dots,2^m+1-R(\sigma)\}$. If we use the convention $N(0)=1$ and $N(n+1)=2^m+1-R(\sigma)$ and count jumps with multiplicities, such a function has exactly $2^m-R(\sigma)$ jumps. We can therefore represent it by a sequence of n zeroes and $2^m-R(\sigma)$ ones, where having k ones between the ith and the jth zero means that $N(j)-N(i)=k$. The number of such sequences is obviously given by $\binom{2^m+n-R(\sigma)}{n}$ and the result follows since every inverse GSR shuffle is equally likely. □

We can now give the idea of the proof of Theorem 3.33. One has

$$\begin{aligned}\|P^m\delta_e-\pi\|_1&=\frac{1}{n!}\sum_\sigma\Big|1-\frac{n!}{2^{mn}}\binom{2^m+n-R(\sigma)}{n}\Big|=\frac{1}{n!}\sum_\sigma\Big|1-\frac{(2^m+n-R(\sigma))!}{2^{mn}(2^m-R(\sigma))!}\Big|\\&=\frac{1}{n!}\sum_\sigma\Big|1-\prod_{k=1}^n\frac{2^m+k-R(\sigma)}{2^m}\Big|=\frac{1}{n!}\sum_\sigma\Big|1-\prod_{k=1}^n\Big(1+\frac{k-R(\sigma)}{2^m}\Big)\Big|\,.\end{aligned}$$

Since, if m is large, the term $\frac{k-R(\sigma)}{2^m}$ is small, one can arguably use the approximation $\prod_i(1+x_i)\approx 1+\sum_i x_i$, which is valid if the x_i are small. One gets

$$\|P^m\delta_e-\pi\|_1\approx\frac{1}{n!}\sum_\sigma\Big|\sum_{k=1}^n\frac{k-R(\sigma)}{2^m}\Big|\approx\frac{n}{n!}\sum_\sigma\Big|\frac{n/2-R(\sigma)}{2^m}\Big|=\frac{n}{2^m}\mathbf{E}\Big|\frac{n}{2}-R(\sigma)\Big|\,,$$

where the expectation is taken with respect to the uniform measure on the set of all permutations σ.

At this point, it is not obvious how to proceed. It has been proven however that the probability (under the uniform measure) that $R(\sigma) = m$ is exactly given by the probability that the sum of n i.i.d. uniform $[0,1]$-valued random variables is between m and $m+1$. Therefore, the central limit applies and shows that, for large values of n, the expression $\frac{n}{2} - R(\sigma)$ is approximately normal with variance n. This implies that

$$\|P^m \delta_e - \pi\|_1 \approx \frac{n^{3/2}}{2^m} .$$

As a consequence, one needs $m \gg \frac{3}{2}\log_2 n$ to make this distance small, which is exactly the result of Bayer and Diaconis.

4 Invariant measures in the general case

4.1 Reversible and stationary Markov processes

Recall that, given a transition probability P on a space $\mathcal{X}$, we associate to it the operator T acting on finite signed measures on $\mathcal{X}$ by

$$(T\mu)(A) = \int_{\mathcal{X}} P(x, A)\, \mu(dx) .$$

A probability measure π is said to be **invariant** for P if $T\pi = \pi$.

In general, given a transition probability P and a corresponding invariant measure π, one can construct a measure $\mathbf{P}_\pi$ on the space of biinfinite sequences $\mathcal{X}^{\mathbf{Z}}$ in the following way. Given any positive number $n > 0$, we define a measure $\mathbf{P}_\pi^n$ on $\mathcal{X}^{2n+1}$ by

$$\int f(x_{-n}, \dots, x_n)\, \mathbf{P}_\pi^n(dx) = \int_{\mathcal{X}} \int_{\mathcal{X}} \cdots \int_{\mathcal{X}} f(x_{-n}, \dots, x_n)\, P(x_{n-1}, dx_n) \cdots P(x_{-n}, dx_{1-n}) \pi(dx_{-n}) . \tag{4.1}$$

It is an easy, although tedious, exercise to check that the family of measures on $\mathcal{X}^{2n+1}$ defined by (4.1) is consistent, so that it defines a unique measure on $\mathcal{X}^{\mathbf{Z}}$ by Kolmogorov's extension theorem, Theorem 2.37. We define on $\mathcal{X}^{\mathbf{Z}}$ the family $\{\theta_n\}$ of shift maps and the time-reversal map ϱ by

$$(\varrho(x))_k = x_{-k} , \quad (\theta_n(x))_k = x_{k+n} .$$

Note that one has the group property $\theta_k \circ \theta_\ell = \theta_{k+\ell}$, so that the family of maps θ_n induces a natural action of $\mathbf{Z}$ on $\mathcal{X}^{\mathbf{Z}}$. With these two maps at hand, we give the following definitions:

Definition 4.1 A probability measure $\mathbf{P}$ on $\mathcal{X}^{\mathbf{Z}}$ is said to define a **stationary** process if $\theta_n^* \mathbf{P} = \mathbf{P}$ for every $n \in \mathbf{Z}$.

Definition 4.2 A probability measure $\mathbf{P}$ on $\mathcal{X}^{\mathbf{Z}}$ is said to define a **reversible** process if $\varrho^* \mathbf{P} = \mathbf{P}$.

In other words, a stationary process is one where, statistically speaking, every time is equivalent. A reversible process is one which looks the same whether time flows forward or backward. We have the following results:

Lemma 4.3 *The measure $\mathbf{P}_\pi$ defined above defines a stationary Markov process.*

Proof. It is sufficient to check that

$$\int f(x_{-n},\dots,x_{n-1})\,\mathbf{P}_\pi^n(dx) = \int f(x_{1-n},\dots,x_n)\,\mathbf{P}_\pi^n(dx)\;,$$

for every $f\colon \mathcal{X}^{2n} \to \mathbf{R}$. We have

$$\begin{aligned}
&\int f(x_{-n},\dots,x_{n-1})\,\mathbf{P}_\pi^n(dx)\\
&\quad= \int_{\mathcal{X}}\int_{\mathcal{X}}\cdots\int_{\mathcal{X}} f(x_{-n},\dots,x_{n-1})\,P(x_{n-2},dx_{n-1})\cdots P(x_{-n},dx_{1-n})\pi(dx_{-n})\\
&\quad= \int_{\mathcal{X}}\int_{\mathcal{X}}\cdots\int_{\mathcal{X}} f(x_{1-n},\dots,x_n)\,P(x_{n-1},dx_n)\cdots P(x_{1-n},dx_{2-n})\pi(dx_{1-n})\\
&\quad= \int_{\mathcal{X}}\int_{\mathcal{X}}\cdots\int_{\mathcal{X}} f(x_{1-n},\dots,x_n)\,P(x_{n-1},dx_n)\cdots P(x_{1-n},dx_{2-n})P(x_{-n},dx_{1-n})\pi(dx_{-n})\\
&\quad= \int f(x_{1-n},\dots,x_n)\,\mathbf{P}_\pi^n(dx)\;.
\end{aligned}$$

Here we went from the second to the third line by just renaming variables. We went from the third to the fourth line by using the invariance of π, namely that $\int_{\mathcal{X}} P(x,A)\,\pi(dx) = \pi(A)$. □

It turns out that, for Markov processes, there is an easy criteria that allows to check whether a given process is reversible or not. In order to state it, define $\varrho\colon \mathcal{X}^2 \to \mathcal{X}^2$ by $\varrho(x,y) = (y,x)$, and write $P\pi$ for the measure on $\mathcal{X}^2$ given by

$$(P\pi)(A\times B) = \int_A P(x,B)\,\pi(dx)\;. \tag{4.2}$$

With this notation, we have

Theorem 4.4 *Consider a stationary Markov process x with transition probabilities P and invariant measure π. Suppose that there exist transition probabilities Q such that $\varrho^*(P\pi) = Q\pi$. Then, the process $y_n = x_{-n}$ is a stationary Markov process with transition probabilities Q and invariant measure π.*

Proof. Note that the assumption is just another way of saying that

$$\int f(x,y)P(x,dy)\,\pi(dx) = \int f(x,y)Q(y,dx)\,\pi(dy)\;,$$

for every measurable and integrable function $f\colon \mathcal{X}^2 \to \mathbf{R}$. We therefore have

$$\begin{aligned}
&\int f(x_{-n},\dots,x_n)\,\mathbf{P}_\pi^n(dx)\\
&\quad= \int_{\mathcal{X}}\int_{\mathcal{X}}\cdots\int_{\mathcal{X}} f(x_{-n},\dots,x_n)\,P(x_{n-1},dx_n)\cdots P(x_{-n},dx_{1-n})\pi(dx_{-n})\\
&\quad= \int_{\mathcal{X}}\int_{\mathcal{X}}\cdots\int_{\mathcal{X}} f(x_{-n},\dots,x_n)\,P(x_{n-1},dx_n)\cdots P(x_{1-n},dx_{2-n})Q(x_{1-n},dx_{-n})\pi(dx_{1-n})\\
&\quad= \int_{\mathcal{X}}\int_{\mathcal{X}}\cdots\int_{\mathcal{X}} f(x_{-n},\dots,x_n)\,P(x_{n-1},dx_n)\cdots Q(x_{1-n},dx_{-n})P(x_{1-n},dx_{2-n})\pi(dx_{1-n})\\
&\quad= \int_{\mathcal{X}}\int_{\mathcal{X}}\cdots\int_{\mathcal{X}} f(x_{-n},\dots,x_n)\,P(x_{n-1},dx_n)\cdots Q(x_{1-n},dx_{-n})Q(x_{2-n},dx_{1-n})\pi(dx_{2-n})\;.
\end{aligned}$$

Proceeding in the same fashion, we finally arrive at

$$\begin{aligned}\int f(x_{-n},\dots,x_n)\,\mathbf{P}_\pi^n(dx)&\\ &= \int_{\mathcal{X}}\int_{\mathcal{X}}\cdots\int_{\mathcal{X}} f(x_{-n},\dots,x_n)\,Q(x_{1-n},dx_{-n})\cdots Q(x_n,dx_{n-1})\pi(dx_n)\\ &= \int f(x_{-n},\dots,x_n)\,(\varrho^*\mathbf{Q}_\pi^n)(dx)\;,\end{aligned}$$

where we denoted by $\mathbf{Q}_\pi$ the law of the stationary Markov process with transition probabilities Q and invariant measure π. This shows that $\mathbf{P}_\pi = \varrho^*\mathbf{Q}_\pi$ and therefore that $\varrho^*\mathbf{P}_\pi = \mathbf{Q}_\pi$, which is the desired result. □

We get as an immediate corollary:

Corollary 4.5 *The measure $\mathbf{P}_\pi$ defined above defines a reversible Markov process if and only if one has $\varrho^*(P\pi) = P\pi$.*

Proof. It is obvious that the condition is necessary since otherwise the law of (x_0, x_1) would be different from the law of (x_1, x_0) under $\mathbf{P}_\pi$. The sufficiency follows from the above theorem since on can take $Q = P$. □

Note that in the case where $\mathcal{X}$ is countable, the condition (4.2) can be written as

$$P_{ij}\pi_j = P_{ji}\pi_i \tag{4.3}$$

for every pair i, j. Summing over j in (4.3) or choosing $B = \mathcal{X}$ in (4.2), we see that if there exists a probability measure π such that (4.2) holds, then this measure is automatically an invariant measure for P. This allows one to easily 'guess' an invariant measure if one believes that a given process is reversible by using the equality

$$\frac{\pi_i}{\pi_j} = \frac{P_{ij}}{P_{ji}}\;.$$

Closer inspection of this equation allows to formulate the following equivalent characterisation for reversibility:

Lemma 4.6 *An irreducible Markov process on a finite state space with transition probabilities P is reversible with respect to some measure π if and only if one has*

$$P_{i_1i_n}P_{i_ni_{n-1}}\cdots P_{i_3i_2}P_{i_2i_1} = P_{i_ni_1}P_{i_1i_2}\cdots P_{i_{n-2}i_{n-1}}P_{i_{n-1}i_n} \tag{4.4}$$

for every n and every sequence of indices $i_1, \dots, i_n$.

In other words, such a process is reversible if and only if the product of the transition probabilities over any loop in the incidence graph is independent of the direction in which one goes through the loop.

Proof. In order to show that the condition is necessary, let us consider the case $n = 3$. One has

$$P_{i_1i_3}P_{i_3i_2}P_{i_2i_1}\pi_{i_1} = P_{i_1i_3}P_{i_3i_2}P_{i_1i_2}\pi_{i_2} = P_{i_1i_3}P_{i_2i_3}P_{i_1i_2}\pi_{i_3} = P_{i_3i_1}P_{i_2i_3}P_{i_1i_2}\pi_{i_1}\;.$$

Since the process is irreducible, we can divide by π_{i_1} on both sides and get the desired equality. The proof for arbitrary n works in exactly the same way.

Let us now show that the condition is sufficient. Fix one particular point in the state space, say the point 1. Since the process is irreducible, we can find for every index i a path $i_1, \dots, i_n$ in the incidence graph connecting 1 to i (we set $i_1 = 1$ and $i_n = i$). We then define a measure π on the state space by

$$\pi_i = \frac{P_{i_n i_{n-1}}}{P_{i_{n-1} i_n}} \frac{P_{i_{n-1} i_{n-2}}}{P_{i_{n-2} i_{n-1}}} \cdots \frac{P_{i_2 i_1}}{P_{i_1 i_2}} \, .$$

Note that (4.4) ensures that this definition does not depend on the particular path that was chosen. Since our state space is finite, one can then normalise the resulting measure in order to make it a probability measure. Furthermore, one has

$$\frac{P_{ji}\pi_i}{P_{ij}\pi_j} = \frac{P_{ji}}{P_{ij}} \cdot \frac{P_{i_n i_{n-1}}}{P_{i_{n-1} i_n}} \frac{P_{i_{n-1} i_{n-2}}}{P_{i_{n-2} i_{n-1}}} \cdots \frac{P_{i_2 i_1}}{P_{i_1 i_2}} \cdot \frac{P_{j_{n-1} j_n}}{P_{j_n j_{n-1}}} \frac{P_{j_{n-2} j_{n-1}}}{P_{j_{n-1} j_{n-2}}} \cdots \frac{P_{j_1 j_2}}{P_{j_2 j_1}} \, . \tag{4.5}$$

Since we have $i = i_n$, $j = j_n$, and $i_1 = j_1$, the path $i_1, \dots, i_n, j_n, \dots, j_1$ forms a closed loop and the ratio in (4.5) is equal to 1. This shows that the process is indeed reversible with respect to π (and therefore that π is its invariant measure). □

Example 4.7 Let $\alpha \in (0,1)$ and $\beta > 0$ be some fixed constants and let $\{\xi_n\}$ be a sequence of i.i.d. $\mathcal{N}(0,1)$ random variables. Define a Markov process on $\mathbf{R}$ by the recursion relation

$$x_{n+1} = \alpha x_n + \beta \xi_n \, .$$

It is immediate that $\pi = \mathcal{N}(0, \beta^2/(1-\alpha^2))$ is an invariant measure for this process (in fact it is the only one). The measure $P\pi$ is given by

$$\begin{aligned}(P\pi)(dx, dy) &= C \exp\Big(-\frac{(1-\alpha^2)x^2}{2\beta^2} - \frac{(y-\alpha x)^2}{2\beta^2}\Big)\, dx\, dy \\ &= C \exp\Big(-\frac{x^2 + y^2 - 2\alpha x y}{2\beta^2}\Big)\, dx\, dy \, ,\end{aligned}$$

for some constant C. It is clear that this measure is invariant under the transformation $x \leftrightarrow y$, so that this process is reversible with respect to π. This may appear strange at first sight if one bases one's intuition on the behaviour of the determinstic part of the recursion relation $x_{n+1} = \alpha x_n$.

Example 4.8 Let $L > 0$ be fixed and let $\mathcal{X}$ be the interval $[0, L]$ with the idntification $0 \sim L$ (*i.e.* $\mathcal{X}$ is a circle of perimeter L). Let $\{\xi_n\}$ be again a sequence of i.i.d. $\mathcal{N}(0,1)$ random variables and define a Markov process on $\mathcal{X}$ by

$$x_{n+1} = x_n + \xi_n \ (\mathrm{mod}\ L) \, .$$

In this case, an invariant probability measure is given by the multiple of the Lebesgue measure $\pi(dx) = dx/L$, and the transition probabilities are given by

$$P(x, dy) = C \sum_{n \in \mathbf{Z}} \exp\Big(-\frac{(y - x - nL)^2}{2}\Big)\, dy \, .$$

Since this density is symmetric under the exchange of x and y, the process is reversible with respect to the Lebesque measure.

Example 4.9 Let (V, E) be a non-oriented connected graph and let x be a random walk on V defined in the following way. Let us fix a function $p: V \to (0,1)$. If $x_n = v \in V$, then x_{n+1} is equal to v with probability $p(v)$ and to one of the k_v adjacent edges to v with probability $(1 - p(v))/k(v)$. In this case, the measure $\pi(v) = ck(v)/(1 - p(v))$ is invariant and the process is reversible with respect to this measure.

Finally, let us note that if a Markov process with transition probabilities P is reversible with respect to some probability measure π, then the operator $T_\star$ is symmetric when viewed as an operator on $\mathrm{L}^2(\mathcal{X}, \pi)$.

4.2 Existence of invariant measures

In the previous section, we have seen that a Markov process on a finite state space always has (at least) one invariant probability measure π.

In the case of an infinite state space, this is no longer true. Consider for example the simpe random walk on $\mathbf{Z}$. This process is constructed by choosing a sequence $\{\xi_n\}$ of i.i.d. random variables taking the values $\{\pm 1\}$ with equal probabilities. One then writes $x_0 = 0$ and $x_{n+1} = x_n + \xi_n$. A probability measure π on $\mathbf{Z}$ is given by a sequence of positive numbers π_n such that $\sum_{n=-\infty}^{\infty} \pi_n = 1$. The invariance condition for π shows that one should have

$$\pi_n = \frac{\pi_{n+1} + \pi_{n-1}}{2}\,, \tag{4.6}$$

for every $n \in \mathbf{Z}$. A moment of reflection shows that the only positive solution to (4.6) with the convention $\pi_0 = 1$ is given by the constant solution $\pi_n = 1$ for every n (exercise: prove it). Since there are infinitely many values of n, this can not be normalised as to give a probability measure.

Intuitively, this phenomenon can be understood by the fact that the random walk tends to make larger and larger excursions away from the origin. In the following subsection, we make this intuition clear by formulating a condition which guarantees the existence of invariant measures for a Markov process on a general state space.

4.3 Weak convergence and Prokhorov's theorem

Recall first of all the following definition:

Definition 4.10 A metric space $\mathcal{X}$ is called **separable** if it has a countable dense subset.

Example 4.11 Examples of separable spaces are $\mathbf{R}^n$ (take points with rational coordinates) and $\mathrm{L}^p(\mathbf{R}^n)$ for every n and every $p \in [1, \infty)$ (take functions of the form $P(x)e^{-|x|^2}$ where P is a polynomial with rational coefficients).

Remember that a sequence μ_n of probability measures on a topological space $\mathcal{X}$ is said to **converge weakly** to a probability measure μ if

$$\lim_{n\to\infty} \int_{\mathcal{X}} \varphi(x)\, \mu_n(dx) = \int_{\mathcal{X}} \varphi(x)\, \mu(dx)\,, \tag{4.7}$$

for every bounded and continuous function $\varphi: \mathcal{X} \to \mathbf{R}$. Note that the speed of the convergence in (4.7) is allowed to depend on φ.

Example 4.12 If $\{x_n\}$ is a sequence of elements converging to a limit x, then the sequence δ_{x_n} converges weakly to δ_x. In this sense the notion of weak convergence is a natural extension of the notion of convergence on the underlying space $\mathcal{X}$.

The aim of this section is to give a 'compactness' theorem that provides us with a very useful criteria to check whether a given sequence of probability measures has a convergent subsequence. In order to state this criteria, let us first introduce the notion of 'tightness':

Definition 4.13 Let $\mathcal{M} \subset \mathcal{P}(\mathcal{X})$ be an arbitrary subset of the set of probability measures on some topological space $\mathcal{X}$. We say that $\mathcal{M}$ is **tight** if, for every $\varepsilon > 0$ there exists a compact set $K \subset \mathcal{X}$ such that $\mu(K) \geq 1 - \varepsilon$ for every $\mu \in \mathcal{M}$.

To see that this concept is not far-fetched consider the following:

Lemma 4.14 *If $\mathcal{X}$ is a complete separable metric space and $\mathcal{M}$ consists of a single measure μ, then $\mathcal{M}$ is tight.*

Loosely speaking, this lemma says that on every 'reasonable' space $\mathcal{X}$, probability measures concentrate on compact sets.

Proof. Let $\{r_i\}$ be a countable dense subset of $\mathcal{X}$ and denote by $\mathcal{B}(x, r)$ the ball of radius r centred at x. Fix $\varepsilon > 0$ and, for every integer $n > 0$, denote by N_n the smallest integer such that

$$\mu\Big(\bigcup_{k \leq N_n} \mathcal{B}(r_k, 1/n)\Big) \geq 1 - \frac{\varepsilon}{2^n} \, .$$

Note that since $\{r_k\}$ is a dense set, one has $\bigcup_{k>0} \mathcal{B}(r_k, 1/n) = \mathcal{X}$, so that N_n is finite for every n. Define now the set K as

$$K = \bigcap_{n \geq 0} \bigcup_{k \leq N_n} \mathcal{B}(r_k, 1/n) \, .$$

It is clear that $\mu(K) > 1 - \varepsilon$. Furthermore, K is totally bounded, *i.e.* for every $\delta > 0$ it can be covered by a finite number of balls of radius δ (since it can be covered by N_n balls of radius $1/n$). It is a classical result from topology that in complete separable metric spaces, totally bounded sets have compact closure. □

On the other hand, one can show that:

Theorem 4.15 (Prohorov) *Let $\{\mu_n\}$ be a tight sequence of probability measures on a complete separable metric space $\mathcal{X}$. Then, there exists a probability measure μ on $\mathcal{X}$ and a subsequence μ_{n_k} such that $\mu_{n_k} \to \mu$ weakly.*

In order to prove this theorem, we need the following little lemma, which is a special case of Tychonoff's theorem:

Lemma 4.16 *Let $\{x_n\}$ be a sequence of elements in $[0,1]^\infty$. Then, there exists a subsequence n_k and an element $x \in [0,1]^\infty$ such that $\lim_{k\to\infty} x_{n_k}(i) \to x(i)$ for every i.*

Proof. Since $[0,1]$ is compact, there exists a subsequence n_k^1 and a number $x(1) \in [0,1]$ such that $\lim_{k\to\infty} x_{n_k^1}(1) \to x(1)$. Similarly, there exists a subsubsequence n_k^2 of n_k^1 and a number $x(2)$ such that $\lim_{k\to\infty} x_{n_k^2}(2) \to x(2)$. One can iterate this construction to find a family of subsequences n_k^i and numbers $x(i)$ such that

- $x_{n_k^i}$ is a subsequence of $x_{n_k^{i-1}}$ for every i.
- $\lim_{k\to\infty} x_{n_k^i}(i) \to x(i)$ for every i.

It now suffices to define $n_k = n_k^k$. The sequence n_k obviously tends to infinity. Furthermore, for every i, the sequence $\{x_{n_k}(i)\}_{k\geq i}$ is a subsequence of $\{x_{n_k^i}(i)\}_{k\geq 0}$ and therefore converges to the same limit $x(i)$. □

Proof of Prohorov's theorem. We only give a sketch of the proof and only consider the case $\mathcal{X} = \mathbf{R}$. Let r_i be an enumeration of $\mathbf{Q}$ and write F_n for the distribution function of μ_n, *i.e.* $F_n(x) = \mu_n((-\infty, x])$. Note that F_n is automatically right-continuous since $(-\infty, x] = \bigcap_{k>0}(-\infty, x_k]$ for every sequence x_k converging to x from above. (It is not left-continuous in general since if x_k is a sequence converging to x from below, one has $\bigcup_{k>0}(-\infty, x_k] = (-\infty, x)$ which is not the same as $(-\infty, x]$. As a generic counterexample, consider the case $\mu = \delta$ and $x = 0$.) Note that the right-continuity of F_n and the density of the points r_i together imply that one has $F_n(x) = \inf\{F_n(r_i) \,|\, r_i > x\}$ for every x. In other words, the values of F_n at the points r_i are sufficient to determine F_n.

Note furthermore that $F_n(x) \in [0,1]$ for every n and every x since we are considering probability measures, so that we can associate to every function F_n an element $\tilde{F}_n$ in $[0,1]^\infty$ by $\tilde{F}_{n,i} = F_n(r_i)$. Since $[0,1]^\infty$ is compact, there exists a subsequence $\tilde{F}_{n_k}$ and an element $\tilde{F} \in [0,1]^\infty$ such that $\lim_{k\to\infty} \tilde{F}_{n_k,i} = \tilde{F}_i$ for every i. Define a function $F\colon \mathbf{R} \to [0,1]$ by $F(x) = \inf\{\tilde{F}_i \,|\, r_i > x\}$ for every $x \in \mathbf{R}$. Then the function F has the following properties:

1. F is increasing.
2. F is right-continuous.
3. $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$.

The first and second claims follows immediately from the definition of F. Since the sequence of measures $\{\mu_n\}$ is tight by assumption, for every $\varepsilon > 0$ there exists $R > 0$ such that $F_n(R) \geq 1-\varepsilon$ and $F_n(-R) \leq \varepsilon$ for every n. Therefore F satisfies the same equalities so that the third claim follows, so that F is the distribution function of some probability measure μ.

We now show that if F is continuous at some point x, then one actually has $F_{n_k}(x) \to F(x)$. The continuity of F at x implies that, for every $\varepsilon > 0$, we can find rationals r_i and r_j such that $r_i < x < r_j$ and such that $\tilde{F}_i > F(x) - \varepsilon$ and $\tilde{F}_j < F(x) + \varepsilon$. Therefore, there exists N such that $\tilde{F}_{n_k,i} > F(x) - 2\varepsilon$ and $\tilde{F}_{n_k,j} < F(x) + 2\varepsilon$ for every $k \geq N$. In particular, the fact that the functions F_n are increasing implies that $|F_{n_k}(x) - F(x)| \leq 2\varepsilon$ for every $k \geq N$ and so proves the claim.

Denote now by S the set of discontinuities of F. Since F is increasing, S is countable. We just proved that $\mu_{n_k}((a,b]) \to \mu((a,b])$ for every interval $(a,b]$ such that a and b do not belong to S. Fix now an arbitrary continuous function $\varphi\colon \mathbf{R} \to [-1,1]$ and a value $\varepsilon > 0$. We want to show that there exists an N such that $|\int \varphi(x)\mu_{n_k}(dx) - \int \varphi(x)\mu(dx)| < 7\varepsilon$ for every $k \geq N$. Choose R as above and note that the tightness condition implies that

$$\Big|\int \varphi(x)\mu_{n_k}(dx) - \int_{-R}^{R} \varphi(x)\mu_{n_k}(dx)\Big| \leq 2\varepsilon\,, \tag{4.8}$$

for every n. The same bound also holds for the integral against μ. Since φ is uniformly continous on $[-R,R]$, there exists $\delta > 0$ such that $|\varphi(x) - \varphi(y)| \leq \varepsilon$ for every pair $(x,y) \in [-R,R]^2$ such that $|x-y| \leq \delta$. Choose now an arbitrary finite strictly increasing sequence $\{x_m\}_{m=0}^M$ such that $x_0 = -R$, $x_M = R$, $|x_{m+1} - x_m| \leq \delta$ for every m, and $x_m \notin S$ for every m. Define furthermore the function $\tilde{\varphi}$: on $(-R,R]$ by $\tilde{\varphi}(x) = x_m$ whenever $x \in (x_m, x_{m+1}]$. Since $\tilde{\varphi}$ is a finite linear combination of characteristic functions for intervals of the form considered above, there exists N such that $|\int_{-R}^{R} \tilde{\varphi}(x)\mu_{n_k}(dx) - \int_{-R}^{R} \tilde{\varphi}(x)\mu(dx)| < \varepsilon$ for every $k \geq N$. Putting these bounds

together yields

$$\begin{aligned}\Big|\int \varphi(x)\mu_{n_k}(dx) - \int \varphi(x)\mu(dx)\Big| &\le \Big|\int \varphi(x)\mu_{n_k}(dx) - \int_{-R}^{R} \varphi(x)\mu_{n_k}(dx)\Big| \\ &+ \Big|\int \varphi(x)\mu(dx) - \int_{-R}^{R} \varphi(x)\mu(dx)\Big| + \Big|\int_{-R}^{R} \tilde\varphi(x)\mu_{n_k}(dx) - \int_{-R}^{R} \varphi(x)\mu_{n_k}(dx)\Big| \\ &+ \Big|\int_{-R}^{R} \tilde\varphi(x)\mu(dx) - \int_{-R}^{R} \varphi(x)\mu(dx)\Big| + \Big|\int_{-R}^{R} \tilde\varphi(x)\mu_{n_k}(dx) - \int_{-R}^{R} \tilde\varphi(x)\mu(dx)\Big| \\ &\le 2\varepsilon + 2\varepsilon + \varepsilon + \varepsilon + \varepsilon \le 7\varepsilon\;,\end{aligned}$$

for every $k \ge N$, thus concluding the proof. □

This theorem allows us to give a very simple criteria for the existence of an invariant measure for a given Markov process.

Theorem 4.17 (Krylov-Bogolubov) *Let P be a Feller transition probability on a complete separable metric space $\mathcal{X}$. If there exists $x \in \mathcal{X}$ such that the sequence of measures $\{P^n(x,\cdot)\}_{n\ge 0}$ is tight, then there exists an invariant probability measure for P.*

Proof. Fix x as given by the assumptions and consider the sequence Q^n of measures on $\mathcal{X}$ defined by

$$Q^n(A) = \frac{1}{n}\sum_{k=1}^{n} P^k(x,A)\;.$$

It is clear that this sequence is also tight, so it has a subsequence that converges weakly to some probability measure π on $\mathcal{X}$. Note furthermore that one has the equality

$$TQ^n - Q^n = \frac{1}{n}(P^{n+1}(x,\cdot) - P(x,\cdot))\;.$$

Let φ be any continuous function from $\mathcal{X}$ to $\mathbf{R}$ which is bounded by 1 and fix $\varepsilon > 0$. By the definition of weak convergence, there exists a value $n > 1/\varepsilon$ for which $|\int \varphi(x)\,Q^n(dx) - \int \varphi(x)\,\pi(dx)| \le \varepsilon$. Since $T_\star\varphi$ is also continuous by assumption (we assumed that P was Feller), we can choose n sufficiently large so that $|\int T_\star\varphi(x)\,Q^n(dx) - \int T_\star\varphi(x)\,\pi(dx)| \le \varepsilon$ as well. We then have

$$\begin{aligned}\Big|\int \varphi(x)\,(T\pi)(dx) - \int \varphi(x)\,\pi(dx)\Big| &\le \Big|\int \varphi(x)\,(T\pi)(dx) - \int \varphi(x)\,(TQ^n)(dx)\Big| \\ &+ \Big|\int \varphi(x)\,(TQ^n)(dx) - \int \varphi(x)\,Q^n(dx)\Big| + \Big|\int \varphi(x)\,Q^n(dx) - \int \varphi(x)\,\pi(dx)\Big| \\ &\le \Big|\int (T_\star\varphi)(x)\,\pi(dx) - \int (T_\star\varphi)(x)\,Q^n(dx)\Big| \\ &\quad + \frac{1}{n}\Big|\int \varphi(y)\,(P^{n+1})(x,dy) - \int \varphi(y)\,P(x,dy)\Big| + \varepsilon \\ &\le 2\varepsilon + \frac{2}{n} \le 4\varepsilon\;.\end{aligned}$$

Since ε was arbitrary, this means that $|\int \varphi(x)\,(T\pi)(dx) - \int \varphi(x)\,\pi(dx)| = 0$. Since φ was also arbitrary, this in trun implies that $T\pi = \pi$, *i.e.* that π is an invariant measure for our system. □

As an immediate consequence, we have that

Corollary 4.18 *If the space $\mathcal{X}$ is compact, then every Feller semigroup on $\mathcal{X}$ has an invariant probability measure.*

Proof. On a compact space, every family of probability measures is tight. □

Remark 4.19 Note that the completeness of $\mathcal{X}$ is essential in all the previous arguments. Consider for example the Markov process defined on $(0,1)$ by the recursion relation $x_{n+1} = x_n/2$. It obviously doesn't have an invariant measure on the open interval $(0,1)$, even though it defines a perfectly valid Feller semigroup on $(0,1)$ equipped with the topology inherited from $\mathbf{R}$.

One simple way of checking that the tightness condition of the Krylov-Bogolubov theorem holds is to find a so-called Lyapunov function for the system:

Definition 4.20 Let $\mathcal{X}$ be a complete separable metric space and let P be a transition probability on $\mathcal{X}$. A Borel measurable function $V\colon \mathcal{X} \to \mathbf{R}_+ \cup \{\infty\}$ is called a **Lyapunov function** for P if it satisfies the following conditions:

- $V^{-1}(\mathbf{R}_+) \neq \emptyset$, in other words there are some values of x for which $V(x)$ is finite.
- For every $c \in \mathbf{R}_+$, the set $V^{-1}(\{x \leq c\})$ is compact.
- There exists a positive constant $\gamma < 1$ and a constant C such that

$$\int_{\mathcal{X}} V(y)\, P(x, dy) \leq \gamma V(x) + C\,,$$

  for every x such that $V(x) \neq \infty$.

With this definition at hand, it is now easy to prove the following result:

Theorem 4.21 *If a transition probability P is Feller and admits a Lyapunov function, then it also has an invariant probability measure.*

Proof. Let $x \in \mathcal{X}$ be any point such that $V(x) \neq \infty$, and consider the sequence of measures $\{P^n(x,\cdot)\}$. Defining $V_n = \int_{\mathcal{X}} V(y)\, P^n(x, dy)$, we then have the inequalities:

$$V_n \leq \gamma V_{n-1} + C \leq \gamma(\gamma V_{n-2} + C) + C \leq \ldots \leq \gamma^n V(x) + \frac{C}{1-\gamma} \leq V(x) + \frac{C}{1-\gamma}\,. \quad (4.9)$$

Therefore, there exists a constant $\tilde{C}$ such that $V_n \leq \tilde{C}$ for every $n \geq 0$. Let now $\varepsilon > 0$ and denote by K_c the family of compact sets $\{x \,|\, V(x) \leq c\}$. Tchebycheff's inequality shows that $P^n(x, K_c) \geq 1 - \tilde{C}/c$. It thus suffices to choose $K = K_c$ with $c = \varepsilon/\tilde{C}$ to have a compact set such that $P^n(x, K) \geq 1 - \varepsilon$ for every $n \geq 0$. □

It remains to find an effective criteria for the transition probabilities to be Feller. We have the following:

Theorem 4.22 *Let x be a Markov process defined by a recursion relation of the type*

$$x_{n+1} = F(x_n, \xi_n)\,,$$

for $\{\xi_n\}$ a sequence of i.i.d. random variables taking values in a measurable space Ω and $F\colon \mathcal{X} \times \Omega \to \mathcal{X}$. If the function $F(\cdot, \xi)\colon \mathcal{X} \to \mathcal{X}$ is continuous for almost every realisation of ξ, then the corresponding transition semigroup is Feller.

Proof. Denote by $\mathbf{P}$ the law of ξ_n on Ω and by $\varphi\colon \mathcal{X} \to \mathcal{X}$ an arbitrary continuous bounded function. It follows from the definition of the transition semigroup $T_\star$ that

$$(T_\star\varphi)(x) = \int_\Omega \varphi(F(x,\xi))\,\mathbf{P}(d\xi)\;.$$

Let now $\{x_n\}$ be a sequence of elements in $\mathcal{X}$ converging to x. Lebesgue's dominated convergence theorem shows that

$$\begin{aligned}\lim_{n\to\infty}(T_\star\varphi)(x_n) &= \lim_{n\to\infty}\int_\Omega \varphi(F(x_n,\xi))\,\mathbf{P}(d\xi) = \int_\Omega \lim_{n\to\infty}\varphi(F(x_n,\xi))\,\mathbf{P}(d\xi)\\ &= \int_\Omega \varphi(F(x,\xi))\,\mathbf{P}(d\xi) = (T_\star\varphi)(x)\;,\end{aligned}$$

which implies that $T_\star\varphi$ is continuous and therefore that $T_\star$ is Feller. □

Combining all of the above yields:

Corollary 4.23 *Let x be a Markov process defined by a recursion relation of the type*

$$x_{n+1} = F(x_n,\xi_n)\;,$$

for $\{\xi_n\}$ a sequence of i.i.d. random variables taking values in a measurable space Ω and $F\colon \mathcal{X}\times\Omega\to\mathcal{X}$. If there exists a function $V\colon\mathcal{X}\to\mathcal{X}$ with compact level sets and constants $\gamma\in(0,1)$ and $C>0$ such that

$$\int_\Omega V(F(x,\xi))\,\mathbf{P}(d\xi) \le \gamma V(x) + C\;,\quad \forall x\in\mathcal{X}\;,$$

then the process x has at least one invariant probability measure on $\mathcal{X}$.

The proof of the previous theorem suggests that if a Markov process has a Lyapunov function V, then its invariant measures should satisfy the bound $\int V(x)\,\pi(dx) \le C/(1-\gamma)$, where C and γ are the constants appearing in (4.9). This is indeed the case, as shown by the following proposition:

Proposition 4.24 *Let P be a transition probability on $\mathcal{X}$ and let $V\colon\mathcal{X}\to\mathbf{R}_+$ be a measurable function such that there exist constants $\gamma\in(0,1)$ and $C\ge 0$ with*

$$\int_{\mathcal{X}} V(y)\,P(x,dy) \le \gamma V(x) + C\;.$$

Then, every invariant measure π for P satisfies

$$\int_{\mathcal{X}} V(x)\,\pi(dx) \le \frac{C}{1-\gamma}\;.$$

Proof. Let $M\ge 0$ be an arbitrary constant. As a shorthand, we will use the notation $a\wedge b$ to denote the minimum between two numbers a and b. One then has the following chain of inequalities:

$$\begin{aligned}\int_{\mathcal{X}} (V(x)\wedge M)\,\pi(dx) &= \int_{\mathcal{X}} (V(x)\wedge M)\,(T\pi)(dx) = \int_{\mathcal{X}} (T_\star(V\wedge M))(x)\,\pi(dx)\\ &= \int_{\mathcal{X}}\int_{\mathcal{X}} (V(y)\wedge M)\,P(x,dy)\,\pi(dx) \le \int_{\mathcal{X}} ((\gamma V(x)+C)\wedge M)\,\pi(dx)\end{aligned}$$

Iterating this estimate, one finds that

$$\int_{\mathcal{X}} (V(x) \wedge M)\, \pi(dx) \leq \int_{\mathcal{X}} ((\gamma^n V(x) + \tfrac{C}{1-\gamma}) \wedge M)\, \pi(dx)$$

for every $n \geq 0$. Since the function on the right hand side is bounded by M, we can apply the Lebesgue dominated convergence theorem. It yields the bound

$$\int_{\mathcal{X}} (V(x) \wedge M)\, \pi(dx) \leq \frac{C}{1-\gamma}\,,$$

which holds uniformily in M, and the result follows. □

4.4 Uniqueness of the invariant measure due to deterministic contraction

In this section, we give a very simple criteria for the uniqueness of the invariant measure for a given system.

Theorem 4.25 *Consider a Markov process defined by a recursion relation of the type*

$$x_{n+1} = F(x_n, \xi_n)\,, \tag{4.10}$$

for $\{\xi_n\}$ a sequence of i.i.d. random variables taking values in a measurable space Ω and $F\colon \mathcal{X} \times \Omega \to \mathcal{X}$. If there exists a constant $\gamma \in (0,1)$ such that

$$\mathbf{E} d(F(x,\xi), F(y,\xi)) \leq \gamma d(x,y)\,, \tag{4.11}$$

for every pair x, y in $\mathcal{X}$, then the process (4.10) has at most one invariant probability measure.

Proof. Let π_1 and π_2 be any two invariant measures for (4.10) and let x_0 and y_0 be two independent $\mathcal{X}$-valued random variables with respective laws π_1 and π_2. Let $\{\xi_n\}$ be an independent sequence of i.i.d. random variables as in the statement of the theorem and define x_n and y_n recursively via (4.10).

We have the inequality

$$\mathbf{E}(1 \wedge d(x_n, y_n) \,|\, \mathcal{F}_n) = \mathbf{E}(1 \wedge d(F(x_{n-1},\xi), F(y_{n-1},\xi))) \leq 1 \wedge \gamma d(x_{n-1}, y_{n-1})\,.$$

Iterating this bound in the same way as in the proof of Proposition 4.24, we obtain

$$\mathbf{E}(1 \wedge d(x_n, y_n)) \leq \mathbf{E}(1 \wedge \gamma^n d(x_0, y_0))\,. \tag{4.12}$$

Denote now by μ_n the joint law of (x_n, y_n) in $\mathcal{X}^2$ and define the projection maps $G_i\colon \mathcal{X}^2 \to \mathcal{X}$ by $G_1(x,y) = x$ and $G_2(x,y) = y$. Since the measures π_i are invariant, we have $G_i^* \mu_n = \pi_i$ for $i = 1, 2$ and for every $n \geq 0$. In order to show that the sequence μ_n is tight, fix $\varepsilon > 0$. We know from Lemma 4.14 that there exist compact sets K_1 and K_2 in $\mathcal{X}$ such that $\pi_i(K_i) \geq 1 - \varepsilon$ (in other words $\pi_i(\mathcal{X} \setminus K_i) < \varepsilon$). Therefore

$$\mu_n(K_1 \times K_2) = 1 - \mu_n(\mathcal{X}^2 \setminus K_1 \times K_2) \geq 1 - \mu_n(\mathcal{X} \times (\mathcal{X} \setminus K_2)) - \mu_n((\mathcal{X} \setminus K_1) \times \mathcal{X}) \geq 1 - 2\varepsilon\,,$$

so that the sequence μ_n is tight. This implies that there exists a measure μ and a subsequence n_k such that $\mu_{n_k} \to \mu$ weakly. Since $1 \wedge d$ is continuous, one has

$$\int (1 \wedge d(x,y)) \mu(dx, dy) = \lim_{k \to \infty} \int (1 \wedge d(x,y)) \mu_{n_k}(dx, dy) \leq \lim_{k \to \infty} \int (1 \wedge \gamma^{n_k} d(x,y)) \mu_0(dx, dy)\,,$$

where the second inequality is nothing but (4.12). Note now that $1 \wedge d^n$ converges pointwise to 0 and is bounded by 1, so that Lebesgue's dominated convergence theorem yields

$$\int (1 \wedge d(x,y))\mu(dx,dy) = 0 \,,$$

so that $\mu(\Delta) = 1$, where $\Delta = \{(x,x) \,|\, x \in \mathcal{X}\}$ is the 'diagonal' in $\mathcal{X}^2$. Since the G_i are continuous, one has again $G_i^* \mu = \pi_i$, so that

$$\pi_1(A) = \mu(A \times \mathcal{X}) = \mu((A \times \mathcal{X}) \cap \Delta) = \mu(A \times A) = \mu((\mathcal{X} \times A) \cap \Delta) = \pi_2(A) \,,$$

implying $\pi_1 = \pi_2$. Since the π_i were arbitrary invariant measures, this shows that there can be only one of them. □

There are situations (we will see one of them immediately) where (4.11) only holds for x and y in some subset $\mathcal{A}$ of $\mathcal{X}$, but where $\mathcal{A}$ has the property of eventually 'absorbing' every trajectory. This motivates the following discussion.

If there exists a closed set $\mathcal{A} \subset \mathcal{X}$ such that $P(x, \mathcal{A}) = 1$ for every $x \in \mathcal{A}$, then one can restrict the original Markov process to a process on $\mathcal{A}$. In this situation, we say that $\mathcal{A}$ is **invariant** for P. It then suffices to check (4.11) for x and y in $\mathcal{A}$ to conclude that the process has a unique invariant measure in $\mathcal{P}(\mathcal{A})$. In this case, one would like to have a criteria that ensures that every invariant measure for P is in $\mathcal{P}(\mathcal{A})$. Consider the sequence $\mathcal{A}_n$ of sets recursively defined by

$$\mathcal{A}_0 = \mathcal{A} \,, \quad \mathcal{A}_{n+1} = \{x \in \mathcal{X} \,|\, P(x, \mathcal{A}_n) > 0\} \,. \tag{4.13}$$

With these definitions, we have

Proposition 4.26 *Let $\mathcal{A}$ be an invariant set for P and let $\mathcal{A}_n$ be defined as in (4.13). If $\bigcup_{n \geq 0} \mathcal{A}_n = \mathcal{X}$, then every invariant measure π for P is in $\mathcal{P}(\mathcal{A})$.*

Proof. We first show recursively that $P^n(x, \mathcal{A}) > 0$ for every $x \in \mathcal{A}_n$. The statement is true by assumption for $n = 0$. Suppose that it is also true for $n = k-1$ and let x be an arbitrary element in $\mathcal{A}_k$. One then has

$$P^k(x, \mathcal{A}) = \int_{\mathcal{X}} P^{k-1}(y, \mathcal{A})\, P(x, dy) \geq \int_{\mathcal{A}_{k-1}} P^{k-1}(y, \mathcal{A})\, P(x, dy) > 0 \,.$$

The last inequality follows from the fact that the function $y \mapsto P^{k-1}(y, \mathcal{A})$ is strictly positive on $\mathcal{A}_{k-1}$ by construction and $P(x, \mathcal{A}_{k-1}) > 0$ by the definition of $\mathcal{A}_k$.

Suppose now that $\pi(\mathcal{A}) < 1$. Since $\bigcup_{n \geq 0} \mathcal{A}_n = \mathcal{X}$ and obviously $\pi(\mathcal{X}) = 1$, there must exist $n > 0$ such that $\pi(\mathcal{A}_n \setminus \mathcal{A}) > 0$. Since $T^n \pi = \pi$ by the invariance of π, this implies that

$$\pi(\mathcal{A}) = \int_{\mathcal{X}} P^n(x, \mathcal{A})\, \pi(dx) \geq \int_{\mathcal{A}} P^n(x, \mathcal{A})\, \pi(dx) + \int_{\mathcal{A}_n \setminus \mathcal{A}} P^n(x, \mathcal{A})\, \pi(dx) > \pi(\mathcal{A}) \,,$$

where the last inequality follows from the fact that $\pi(\mathcal{A}_n \setminus \mathcal{A}) > 0$ and $P^n(x, \mathcal{A}) > 0$ for every $x \in \mathcal{A}_n$. This is a contradiction, so that one must have $\pi(\mathcal{A}) = 1$. □

Let us conclude this section by a complete treatment of the following example:

Proposition 4.27 *Let x be the Markov process on $\mathbf{R}_+$ such that x_{n+1} is given by the solution at time 1 to the differential equation*

$$\frac{dx}{dt} = \frac{1}{x(t)} - 2 + \xi_n(t)\,, \quad x(0) = x_n\,, \tag{4.14}$$

for a sequence of i.i.d. $\mathcal{C}([0,1],\mathbf{R})$-valued random variables $\{\xi_n\}$ such that $\sup_{t\in[0,1]}|\xi_n(t)| \le 1$ almost surely. Then, this process has a unique invariant measure π. Furthermore, π satisfies $\pi([1/3,1]) = 1$.

Proof. Denote by Φ the solution map to (4.14), so that $x_{n+1} = \Phi(x_n,\xi_n)$. Denote furthermore by Φ_+ the map that solves (4.14) with $\xi_n(t) = 1$ for all t and by Φ_- the map that solves (4.14) with $\xi_n(t) = -1$ for all t. Then, a standard comparison argument shows that $x_{n+1} \in [\Phi_-(x_n), \Phi_+(x_n)]$ almost surely.

Fix $\varepsilon > 0$, and define $\mathcal{A} = [1/3 - \varepsilon, 1 + \varepsilon]$. With this definition, one has $[\Phi_-^{-n}(1/3 - \varepsilon), \Phi_+^{-n}(1+\varepsilon)] \subset \mathcal{A}_n$, where we set $\Phi_-^{-n}(x) = 0$ if x has no preimage under Φ_-^n. Since $\lim_{n\to\infty}\Phi_-^n(x) = 1/3$ and $\lim_{n\to\infty}\Phi_+^n(x) = 1$ for every $x \in \mathbf{R}_+$, it is clear that $\bigcup_{n\ge0}\mathcal{A}_n = \mathbf{R}_+$ so that Proposition 4.26 applies. Since this was true for every $\varepsilon > 0$, one must actually have $\pi([1/3,1]) = 1$.

Denote now by Φ' the derivative of Φ with respect to x. We know from the elementary properties of differential equations that $\Phi'(x_n,\xi_n)$ is the solution at time 1 to the differential equation

$$\frac{dy}{dt} = -\frac{y(t)}{x^2(t)}\,, \quad y(0) = 1\,,$$

where x is the solution to (4.14). This equation can be solved explicitly, so that

$$\Phi'(x_n,\xi_n) = \exp\Big(-\int_0^1 \frac{dt}{x^2(t)}\Big)\,.$$

This shows that the map Φ is continuous in x (actually even differentiable), so that the corresponding transition operator is Feller. Since $[1/3,1]$ is compact, this in turn implies that it has at least one invariant probability measure. Furthermore, one has $|\Phi'(x,\xi)| \le 1/e < 1$ for every $x \le 1$, so that Theorem 4.25 applies. □

4.5 Uniqueness of the invariant measure due to probabilistic effects

In this section, we give another simple criteria for the uniqueness of the invariant measure of a Markov transition operator which is based on completely different mechanisms from the previous section. The result presented in the previous section only used the contractive properties of the map F in order to prove uniqueness. This was very much in the spirit of the Banach fixed point theorem and can be viewed as a purely 'deterministic' effect. The criteria given in this section is much more probabilistic in nature and can be viewed as a strong form of irreducibility.

The criteria in this section will also be based on Banach's fixed point theorem, but this time in the space of probability measures. The 'right' distance between probability measures that makes it work is the **total variation distance** defined in the following way.

Given two positive measures μ and ν on a measurable space Ω, we denote by $\mathcal{D}_\mu$ and $\mathcal{D}_\nu$ their Radon-Nikodym derivatives with respect to the measure $\mu + \nu$. It is easy to check that both μ and ν are absolutely continuous with respect to $\mu+\nu$, so that these derivatives exist. With this notation in mind, we then define

$$\|\mu - \nu\|_{\mathrm{TV}} \equiv \int_\Omega |\mathcal{D}_\mu(w) - \mathcal{D}_\nu(w)|\,(\mu+\nu)(dw)\,. \tag{4.15}$$

Note that this distance does not depend on the choice of reference measure. In other words, if η is an arbitrary positive measure on Ω such that both μ and ν are absolutely continuous with respect to η (with respective derivatives $\tilde{\mathcal{D}}_\mu$ and $\tilde{\mathcal{D}}_\nu$), then one has

$$\|\mu - \nu\|_{\text{TV}} = \int_\Omega |\tilde{\mathcal{D}}_\mu(w) - \tilde{\mathcal{D}}_\nu(w)|\, \eta(dw) \,. \tag{4.16}$$

This follows immediately from the fact that in this case one has $\tilde{\mathcal{D}}_\mu = \mathcal{D}_\mu(\tilde{\mathcal{D}}_\mu + \tilde{\mathcal{D}}_\nu)$ and $\tilde{\mathcal{D}}_\nu = \mathcal{D}_\nu(\tilde{\mathcal{D}}_\mu + \tilde{\mathcal{D}}_\nu)$, and therefore $|\tilde{\mathcal{D}}_\nu - \tilde{\mathcal{D}}_\mu| = |\mathcal{D}_\nu - \mathcal{D}_\mu|(\tilde{\mathcal{D}}_\mu + \tilde{\mathcal{D}}_\nu)$.

Given two positive measures μ and ν, we denote by $\mu \wedge \nu$ the measure obtained by

$$(\mu \wedge \nu)(A) = \int_A \min\{\mathcal{D}_\mu(w), \mathcal{D}_\nu(w)\}\, (\mu + \nu)(dw) \,.$$

Since, for any two positive numbers, one has $|x - y| = x + y - 2\min\{x, y\}$, the definition (4.15) immediately implies that if μ and ν are two probability measures, one has

$$\|\mu - \nu\|_{\text{TV}} = 2 - 2(\mu \wedge \nu)(\Omega) \,. \tag{4.17}$$

Note also that

Lemma 4.28 *The space of probability measures on Ω endowed with the total variation distance $\|\cdot\|_{\text{TV}}$ is complete.*

Proof. Let μ_n be a sequence of probability measures that is Cauchy in the total variation distance and let η be defined by $\eta = \sum_{n>0} 2^{-n}\mu_n$. Then each of the μ_n is absolutely continuous with respect to η. By (4.16), the total variation distance is equal to the L^1 distance between the corresponding Radon-Nikodym derivatives. The result thus follows from the completeness of $\mathrm{L}^1(\Omega, \eta)$. □

We are now in a position to formulate the criteria announced at the beginning of this section.

Theorem 4.29 *Let P be a transition probability on a space $\mathcal{X}$. Assume that there exists $\alpha > 0$ and a probability measure η on $\mathcal{X}$ such that $P(x, \cdot) \geq \alpha\eta$ for every $x \in \mathcal{X}$. Then, P has a unique invariant measure π.*

Proof. Note first that the assumption implies that $T\mu \geq \alpha\eta$ for every probability measure μ on $\mathcal{X}$. We can therefore define probability measures $\bar{T}\mu$ by

$$T\mu = \alpha\eta + (1 - \alpha)\bar{T}\mu \,. \tag{4.18}$$

Let μ and ν now be two arbitrary probability measures on $\mathcal{X}$. Using (4.17), we can define the probability measures $\bar{\mu}$ and $\bar{\nu}$ by

$$\mu = \mu \wedge \nu + \frac{\|\mu - \nu\|_{\text{TV}}}{2}\bar{\mu} \,, \qquad \nu = \mu \wedge \nu + \frac{\|\mu - \nu\|_{\text{TV}}}{2}\bar{\nu} \,.$$

One then has

$$\|T\mu - T\nu\|_{\text{TV}} = \|T\bar{\mu} - T\bar{\nu}\|_{\text{TV}}\frac{\|\mu - \nu\|_{\text{TV}}}{2} \,.$$

It follows from the definition (4.18) that

$$\|T\bar{\mu} - T\bar{\nu}\|_{\text{TV}} = \|\alpha\eta + (1 - \alpha)\bar{T}\bar{\mu} - \alpha\eta - (1 - \alpha)\bar{T}\bar{\nu}\|_{\text{TV}}$$

$$= (1-\alpha)\|\bar{T}\bar{\mu} - \bar{T}\bar{\nu}\| \leq 2(1-\alpha)\;,$$

where we used the fact that the total variation distance between two probability measures can never exceed 2. Combining these bounds yields

$$\|T\mu - T\nu\|_{\text{TV}} \leq (1-\alpha)\|\mu - \nu\|_{\text{TV}}\;,$$

so that T is a contraction. The result now follows from Banach's fixed point theorem. □

5 Structure theorem for invariant measures

In this section, we prove a general structure theorem for Markov processes that gives us a better understanding of what the set of invariant probability measures can look like. Since for any two invariant measures π_1 and π_2 for a given transition operator T, any convex combination of the type $t\pi_1 + (1-t)\pi_2$ with $t \in [0,1]$ is again an invariant measure for T, the set $\mathcal{I}(T)$ of invariant probability measures for T is obviously convex. If T is Feller, then it is a continuous map from $\mathcal{P}(\mathcal{X})$ to $\mathcal{P}(\mathcal{X})$ in the topology of weak convergence. Therefore, if π_n is a sequence of invariant measures converging weakly to a limit π, one has

$$T\pi = T\lim_{n\to\infty}\pi_n = \lim_{n\to\infty} T\pi_n = \lim_{n\to\infty}\pi_n = \pi\;,$$

so that π is again an invariant probability measure for T. This shows that if T is Feller, then the set $\mathcal{I}(T)$ is closed (in the topology of weak convergence).

Remark 5.1 If T is not Feller, it is not true in general that $\mathcal{I}(T)$ is closed. Choose for example an arbitrary measure μ on $\mathbf{R}_+$ and consider the transition probabilities given by

$$P(x,\cdot) = \begin{cases} \delta_x & \text{if } x < 0 \\ \mu & \text{if } x \geq 0. \end{cases}$$

In this case, $\delta_x \in \mathcal{I}(T)$ for every $x < 0$, but $\delta_0 \notin \mathcal{I}(T)$.

Before we get to the "meat" of this section, let us make a short excursion into deterministic ergodic theory.

5.1 Ergodic theory for dynamical systems

Recall that a **dynamical system** consists of a probability space $(\Omega, \mathscr{F}, \mathbf{P})$ and a measurable measure preserving map $\theta\colon \Omega \to \Omega$, *i.e.* a map such that $\mathbf{P}(\theta^{-1}(A)) = \mathbf{P}(A)$ for every $A \in \mathscr{F}$. We will denote as usual by $\mathbf{E}$ expectations with respect to $\mathbf{P}$.

Given such a dynamical system, we define $\mathscr{I} \subset \mathscr{F}$ as the set of subsets such that $\theta^{-1}(A) = A$. It is clear that $\mathscr{I}$ is again a σ-algebra. The perhaps most famous result in the theory of dynamical systems is

Theorem 5.2 (Birkhoff's Ergodic Theorem) *Let $(\Omega, \mathscr{F}, \mathbf{P}, \theta, \mathscr{I})$ be as above and let $f\colon \Omega \to \mathbf{R}$ be such that $\mathbf{E}|f| < \infty$. Then,*

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1} f(\theta^n\omega) = \mathbf{E}(f \,|\, \mathscr{I})$$

almost surely.

remember that a dynamical system is said to be **ergodic** if all sets in $\mathscr{I}$ have either measure 0 or measure 1. Note that this is a property of the map θ as well as of the measure $\mathbf{P}$.

Corollary 5.3 *With the notations of Theorem 5.2, if the dynamical system is ergodic, then*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} f(\theta^n \omega) = \mathbf{E} f$$

almost surely.

Proof of the corollary. By definition, the function $\bar{f} \equiv \mathbf{E}(f \,|\, \mathscr{I})$ is $\mathscr{I}$-measurable. Define the sets $A_+ = \{\omega \in \Omega \,|\, \bar{f}(\omega) > \mathbf{E}\bar{f}\}$, $A_- = \{\omega \in \Omega \,|\, \bar{f}(\omega) < \mathbf{E}\bar{f}\}$, and $A_0 = \{\omega \in \Omega \,|\, \bar{f}(\omega) = \mathbf{E}\bar{f}\}$. All three sets belong to $\mathscr{I}$ and they form a partition of Ω. Therefore, exactly one of them has measure 1 and the other two must have measure 0. If it was A_+, one would have $\mathbf{E}\bar{f} = \int_{A_+} f(\omega)\,\mathbf{P}(d\omega) > \mathbf{E}\bar{f}$, which is a contradiction and similarly for A_-. This implies that $\mathbf{P}(A_0) = 1$, and so $\mathbf{P}(\bar{f} = \mathbf{E}\bar{f}) = 1$. □

Before we trun to the proof of Theorem 5.2, we establish the following important result:

Theorem 5.4 (Maximal Ergodic Theorem) *With the notations of Theorem 5.2, define*

$$S_N(\omega) = \sum_{n=0}^{N-1} f(\theta^n \omega)\,, \quad M_N(\omega) = \max\{S_0(\omega), S_1(\omega), \ldots, S_N(\omega)\}\,,$$

with the convention $S_0 = 0$. Then, $\int_{\{M_N>0\}} f(\omega)\,\mathbf{P}(d\omega) \geq 0$ for every $N \geq 1$.

Proof. For every $N \geq k \geq 0$ and every $\omega \in \Omega$, one has $M_N(\theta\omega) \geq S_k(\theta\omega)$ by definition, and so $f(\omega) + M_N(\theta\omega) \geq f(\omega) + S_k(\theta\omega) = S_{k+1}(\omega)$. Therefore

$$f(\omega) \geq \max\{S_1(\omega), S_2(\omega), \ldots, S_N(\omega)\} - M_N(\theta\omega)\,.$$

Furthermore, $\max\{S_1(\omega), \ldots, S_N(\omega)\} = M_N(\omega)$ on the set $\{M_N > 0\}$, so that

$$\int_{\{M_N>0\}} f(\omega)\,\mathbf{P}(d\omega) \geq \int_{\{M_N>0\}} (M_N(\omega) - M_N(\theta\omega))\,\mathbf{P}(d\omega) \geq \mathbf{E}M_N - \int_{A_N} M_N(\omega)\,\mathbf{P}(d\omega)\,,$$

where $A_N = \{\theta\omega \,|\, M_N(\omega) > 0\}$. The second-to-last inequality follows from the fact that $M_N \geq 0$ and the last inequality follows from the fact that θ is measure-preserving. Since $M_N \geq 0$, $\int_A M_N(\omega)\,\mathbf{P}(d\omega) \leq \mathbf{E}M_N$ for every set A, so that the expression above is greater or equal to 0, which is the required result. □

We can now turn to the

Proof of Birkhoff's Ergodic Theorem. Replacing f by $f - \mathbf{E}(f \,|\, \mathscr{I})$, we can assume without loss of generality that $\mathbf{E}(f \,|\, \mathscr{I}) = 0$. Define $\bar{\eta} = \limsup_{n\to\infty} S_n/n$ and $\underline{\eta} = \liminf_{n\to\infty} S_n/n$. It is sufficient to show that $\bar{\eta} \leq 0$ almost surely, since this implies (by considering $-f$ instead of f) that $\underline{\eta} \geq 0$ and so $\bar{\eta} = \underline{\eta} = 0$.

It is clear that $\bar{\eta}(\theta\omega) = \bar{\eta}(\omega)$ for every ω, so that, for every $\varepsilon > 0$, one has $A^\varepsilon = \{\bar{\eta}(\omega) > \varepsilon\} \in \mathscr{I}$. Define

$$f^\varepsilon(\omega) = (f(\omega) - \varepsilon)\,\chi_{A^\varepsilon}(\omega)\,,$$

and define S_N^ε and M_N^ε accordingly. It follows from Theorem 5.4 that $\int_{\{M_N^\varepsilon > 0\}} f^\varepsilon(\omega)\,\mathbf{P}(d\omega) \geq 0$ for every $N \geq 1$. Note that with these definitions we have that

$$\frac{S_N^\varepsilon(\omega)}{N} = \begin{cases} 0 & \text{if } \bar\eta(\omega) \leq \varepsilon \\ \frac{S_N(\omega)}{N} - \varepsilon & \text{otherwise.} \end{cases} \tag{5.1}$$

The sequence of sets $\{M_N^\varepsilon > 0\}$ increases to the set $B^\varepsilon \equiv \{\sup_N S_N^\varepsilon > 0\} = \{\sup_N \frac{S_N^\varepsilon}{N} > 0\}$. It follows from (5.1) that

$$B^\varepsilon = \{\bar\eta > \varepsilon\} \cap \Big\{\sup_N \frac{S_N}{N} > \varepsilon\Big\} = \{\bar\eta > \varepsilon\} = A^\varepsilon \;.$$

Since $\mathbf{E}|f^\varepsilon| \leq \mathbf{E}|f| + \varepsilon < \infty$, the dominated convergence theorem implies that

$$\lim_{N\to\infty} \int_{\{M_N^\varepsilon > 0\}} f^\varepsilon(\omega)\,\mathbf{P}(d\omega) = \int_{A^\varepsilon} f^\varepsilon(\omega)\,\mathbf{P}(d\omega) \geq 0 \;,$$

and so

$$\begin{aligned} 0 \leq \int_{A^\varepsilon} f^\varepsilon(\omega)\,\mathbf{P}(d\omega) &= \int_{A^\varepsilon} (f(\omega) - \varepsilon)\,\mathbf{P}(d\omega) = \int_{A^\varepsilon} f(\omega)\,\mathbf{P}(d\omega) - \varepsilon \mathbf{P}(A^\varepsilon) \\ &= \int_{A^\varepsilon} \mathbf{E}(f(\omega)\,|\,\mathscr{I})\,\mathbf{P}(d\omega) - \varepsilon \mathbf{P}(A^\varepsilon) = -\varepsilon \mathbf{P}(A^\varepsilon) \;, \end{aligned}$$

where we used the fact that $A^\varepsilon \in \mathscr{I}$ to go from the first to the second line. Therefore, one must have $\mathbf{P}(A^\varepsilon) = 0$ for every $\varepsilon > 0$, which implies that $\bar\eta \leq 0$ almost surely. □

5.2 Structure of the set of invariant measures

Recall the construction from Section 4.1 that associates to every invariant probability measure π of a given transition operator a measure $\mathbf{P}_\pi$ on the space $\mathcal{X}^{\mathbf{Z}}$ of $\mathcal{X}$-valued processes. We furthermore defined the shifts θ_n on $\mathcal{X}^{\mathbf{Z}}$ by

$$(\theta_n x)(m) = x(n+m) \;,$$

and we write $\theta = \theta_1$. By the definition of stationarity, one has:

Lemma 5.5 *The triple $(\mathcal{X}^{\mathbf{Z}}, \theta, \mathbf{P}_\pi)$ defines a continuous dynamical system.*

Proof. It is clear that θ is continuous. It was already checked in Lemma 4.3 that $\mathbf{P}_\pi$ defines a stationary process, *i.e.* that it is invariant under θ. □

In this section, we will often approximate sets belonging to one particular σ-algebra by sets belonging to another σ-algebra. In this context, it is convenient to introduce a notation for the **completion** of a σ-algebra under a given probability measure. Assuming that it is clear from the context what the probability measure $\mathbf{P}$ is, we define the completion $\bar{\mathscr{F}}$ of a σ-algebra $\mathscr{F}$ to be the smallest σ-algebra containing $\mathscr{F}$ with the additional property that if $A \in \bar{\mathscr{F}}$ with $\mathbf{P}(A) = 0$ and $B \subset A$ is any subset of A, then $B \in \bar{\mathscr{F}}$.

Remember also from the theory of dynamical systems that the measure $\mathbf{P}_\pi$ is said to be **ergodic** if every measurable set $A \subset \mathcal{X}^{\mathbf{Z}}$ which is invariant under θ satisfies $\mathbf{P}_\pi(A) \in \{0, 1\}$. As in the previous section, we denote by $\mathscr{I}$ the set of all measurable subsets of $\mathcal{X}^{\mathbf{Z}}$ that are invariant under θ.

Definition 5.6 We say that an invariant measure π of a Markov process with associated transition semigroup T is **ergodic** if the corresponding measure $\mathbf{P}_\pi$ is ergodic for θ.

The main result of this section is the following characterisation of the set of all invariant measure for a given Markov semigroup:

Theorem 5.7 *The set $\mathcal{I}(T)$ of all invariant probability measures for a Markov semigroup T is convex and $\pi \in \mathcal{I}(T)$ is ergodic if and only if it is an extremal of $\mathcal{I}(T)$ (that is it cannot be decomposed as $\pi = t\pi_1 + (1-t)\pi_2$ with $t \in (0,1)$ and $\pi_i \in \mathcal{I}(T)$). Furthermore, any two ergodic invariant probability measures are either identical or mutually singular.*

Before we turn to the proof of Theorem 5.7, we prove the following preliminary lemma,

Lemma 5.8 *Let $\mathbf{P}$ be the law of a stationary Markov process on $\mathcal{X}^{\mathbf{Z}}$. Then, the σ-algebra $\mathscr{I}$ of all subsets invariant under θ is contained (up to sets of $\mathbf{P}$-measure 0) in $\mathscr{F}_0^0$.*

Proof. Consider the collection of events

$$\mathcal{B}_0 = \{A \in \mathcal{B} : \forall \varepsilon > 0 \; \exists N > 0 \,\&\, A_\varepsilon \in \mathscr{F}_{-N}^N \text{ with } \mathbf{P}(A \triangle A_\varepsilon) < \varepsilon\} .$$

We claim that one actually has $\mathcal{B}_0 = \mathscr{B}(\mathcal{X}^{\mathbf{Z}})$. Since $\mathcal{B}_0$ contains all cylindrical sets, it suffices to show that it is a σ-algebra. For this, since $\mathcal{B}_0$ clearly contains $\emptyset$ and $\mathcal{X}^{\mathbf{Z}}$ and is stable under taking complements, it suffices to consider countable unions. For a sequence of events $\{A_j\}_{j\geq 1} \subset \mathcal{B}_0$, we can by assumption find a sequence N_j and events $A_j' \in \mathscr{F}_{-N_j}^{N_j}$ such that $\mathbf{P}(A_j \triangle A_j') \leq \varepsilon 2^{-j}$. Since $\mathbf{P}$ is finite, we can also find J such that, setting $A = \bigcup_{j\geq 1} A_j$, one has $\mathbf{P}(A \triangle \bigcup_{j\leq J} A_j) \leq \varepsilon$. We conclude that $\mathbf{P}(A \triangle \bigcup_{j\leq J} A_j') \leq 2\varepsilon$ so that, since $\bigcup_{j\leq J} A_j' \subset \mathcal{F}_{-N}^N$ for $N = \max\{N_j : j \leq J\}$, the claim follows.

Let now $A \in \mathscr{I}$ and, for every $\varepsilon > 0$, consider $N > 0$ and a set $A_\varepsilon \in \mathscr{F}_{-N}^N$ such that $\mathbf{P}(A \triangle A_\varepsilon) < \varepsilon$, which exists since $A \in \mathcal{B}_0$. By the invariance of A and of $\mathbf{P}$ under shifts, it follows that we also have $\mathbf{P}(A \triangle \theta^{-(k+N)} A_\varepsilon) < \varepsilon$. Since $\theta^{-(k+N)} A_\varepsilon \subset \mathscr{F}_k^\infty$ for every ε, it follows that one has $A \in \bar{\mathscr{F}}_k^\infty$. Since this is true for every k, one actually has $A \in \bar{\mathscr{F}}_\infty^\infty$. The same reasoning but shifting in the other direction shows that one also has $A \in \bar{\mathscr{F}}_{-\infty}^{-\infty}$.

We use from now on the notation $A \sim B$ to signify that A and B differ by a set of $\mathbf{P}$-measure 0. Point *(iii)* of Theorem 2.23 (or rather a slight extension of it) shows that if f and g are two functions that are respectively $\bar{\mathscr{F}}_\infty^\infty$ and $\bar{\mathscr{F}}_{-\infty}^{-\infty}$-measurable, then

$$\mathbf{E}(fg \,|\, \mathscr{F}_0^0) = \mathbf{E}(f \,|\, \mathscr{F}_0^0)\, \mathbf{E}(g \,|\, \mathscr{F}_0^0) .$$

Applying this result with $f = g = \chi_A$, we find that

$$\mathbf{E}(\chi_A^2 \,|\, \mathscr{F}_0^0) = \left(\mathbf{E}(\chi_A \,|\, \mathscr{F}_0^0)\right)^2 .$$

Since on the other hand $\chi_A^2 = \chi_A$ and $\mathbf{E}(\chi_A \,|\, \mathscr{F}_0^0) \in [0,1]$, one has $\mathbf{E}(\chi_A \,|\, \mathscr{F}_0^0) \in \{0,1\}$ almost surely. Let $\hat{A}$ denote the points such that $\mathbf{E}(\chi_A \,|\, \mathscr{F}_0^0) = 1$, so that $\hat{A} \in \mathscr{F}_0^0$ by the definition of conditional expectations. Furthermore, the definition of conditional expectations yields $\mathbf{P}(\hat{A} \cap E) = \mathbf{P}(A \cap E)$ for every set $E \in \mathscr{F}_0^0$ and (using the same reasoning as above for $1 - \chi_A$) $\mathbf{P}(\hat{A}^c \cap E) = \mathbf{P}(A^c \cap E)$ as well. Using this for $E = \hat{A}$ and $E = \hat{A}^c$ respectively shows that $A \sim \hat{A}$, as required. □

Corollary 5.9 *Let again $\mathbf{P}$ be the law of a stationary Markov process. Then, for every set $A \in \mathscr{I}$ there exists a measurable set $\bar{A} \subset \mathcal{X}$ such that $A \sim \bar{A}^{\mathbf{Z}}$.*

Proof. We know by Lemma 5.8 that $A \in \bar{\mathscr{F}}_0^0$, so that the event A is equivalent to an event of the form $\{x_0 \in \bar{A}\}$ for some $\bar{A} \subset \mathcal{X}$. Since $\mathbf{P}$ is stationary and $A \in \mathscr{I}$, the time 0 is not distinguishable from any other time, so that this implies that A is equivalent to the event $\{x_n \in \bar{A}\}$ for every $n \in \mathbf{Z}$. In particular, it is equivalent to the event $\{x_n \in \bar{A} \text{ for every } n\}$. □

Note that this result is crucial in the proof of the structure theorem, since it allows us to relate invariant sets $A \in \mathscr{I}$ to invariant sets $\bar{A} \subset \mathcal{X}$, in the following sense:

Definition 5.10 Let T be a transition operator on a space $\mathcal{X}$ and let π be an invariant probability measure for T. We say that a measurable set $\bar{A} \subset \mathcal{X}$ is π-**invariant** if $P(x, \bar{A}) = 1$ for π-almost every $x \in \bar{A}$.

With this definition, we have

Corollary 5.11 *Let T be a transition operator on a space $\mathcal{X}$ and let π be an invariant probability measure for T. Then π is ergodic if and only if every π-invariant set $\bar{A}$ is of π-measure 0 or 1.*

Proof. It follows immediately from the definition of an invariant set that one has $\pi(\bar{A}) = \mathbf{P}_\pi(\bar{A}^{\mathbf{Z}})$ for every π-invariant set $\bar{A}$.

Now if π is ergodic, then $\mathbf{P}_\pi(\bar{A}^{\mathbf{Z}}) \in \{0, 1\}$ for every set $\bar{A}$, so that in particular $\pi(\bar{A}) \in \{0, 1\}$ for every π-invariant set. If π is not ergodic, then there exists a set $A \in \mathscr{I}$ such that $\mathbf{P}_\pi(A) \notin \{0, 1\}$. By Corollary 5.9, there exists a set $\bar{A} \subset \mathcal{X}$ such that $A \sim \{x_0 \in \bar{A}\} \sim \bar{A}^{\mathbf{Z}}$. The set $\bar{A}$ must be π-invariant, since otherwise the relation $\{x_0 \in \bar{A}\} \sim \bar{A}^{\mathbf{Z}}$ would fail. □

Proof of Theorem 5.7. Assume first that $\pi \in \mathcal{I}(T)$ is not extremal, *i.e.* it is of the form $\pi = t\pi_1 + (1-t)\pi_2$ with $t \in (0, 1)$ and $\pi_i \in \mathcal{I}(T)$. (Note that therefore $\mathbf{P}_\pi = t\mathbf{P}_{\pi_1} + (1-t)\mathbf{P}_{\pi_2}$.) Assume by contradiction that π is ergodic, so that $\mathbf{P}_\pi(A) \in \{0, 1\}$ for every $A \in \mathscr{I}$. If $\mathbf{P}_\pi(A) = 0$, then one must have $\mathbf{P}_{\pi_1}(A) = \mathbf{P}_{\pi_2}(A) = 0$ and smilarly if $\mathbf{P}_\pi(A) = 1$. Therefore, $\mathbf{P}_{\pi_1}$ and $\mathbf{P}_{\pi_2}$ agree on $\mathscr{I}$, so that both $\mathbf{P}_{\pi_1}$ and $\mathbf{P}_{\pi_2}$ are ergodic. Let now $f\colon \mathcal{X}^{\mathbf{Z}} \to \mathbf{R}$ be an arbitrary bounded measurable function and consider the function $f^*\colon \mathcal{X}^{\mathbf{Z}} \to \mathbf{R}$ which is defined by

$$f^*(x) = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} f(\theta^k(x)) \,,$$

on the set E on which this limit exists and by $f^*(x) = 0$ otherwise. Denote by E_i the set of points x such that $f^*(x) = \int f(x)\, \mathbf{P}_{\pi_i}(dx)$. By Corollary 5.3, one has $\mathbf{P}_{\pi_i}(E_i) = 1$, so that in particular $\mathbf{P}_\pi(E_1) = \mathbf{P}_\pi(E_2) = 1$. Since f was arbitrary, one can choose it so that $\int f(x)\, \mathbf{P}_{\pi_1}(dx) \neq \int f(x)\, \mathbf{P}_{\pi_2}(dx)$, which would imply $E_1 \cap E_2 = \emptyset$, thus contradicting the fact that $\mathbf{P}_\pi(E_1) = \mathbf{P}_\pi(E_2) = 1$.

Let now $\pi \in \mathcal{I}(T)$ be an invariant measure that is not ergodic, we want to show that it can be written as $\pi = t\pi_1 + (1-t)\pi_2$ for some $\pi_i \in \mathcal{I}(T)$ and $t \in (0, 1)$. By Corollary 5.11, there exists a set $\bar{A} \subset \mathcal{X}$ such that $\pi(\bar{A}) = t$ and such that $P(x, \bar{A}) = 1$ for π-almost every $x \in \bar{A}$. Furthermore, one has $\pi(\bar{A}^c) = 1 - t$ and the stationarity of π implies that one must have $P(x, \bar{A}^c) = 1$ for π-almost every $x \in \bar{A}^c$. This invariance property immediately implies that the measures π_i defined by

$$\pi_1(B) = \frac{1}{t}\pi(\bar{A} \cap B) \,, \qquad \pi_2(B) = \frac{1}{1-t}\pi(\bar{A}^c \cap B) \,,$$

belong to $\mathcal{I}(T)$ and therefore have the required property.

The last statement follows immediately from Corollary 5.3. Let indeed π_1 and π_2 be two distinct ergodic invariant probability measures. Since they are distinct, there exists a measurable

bounded function $f\colon \mathcal{X} \to \mathbf{R}$ such that $\int f(x)\,\pi_1(dx) \neq \int f(x)\,\pi_2(dx)$. Let us denote by $\{x_n\}$ the Markov process with transition operator T starting at x_0. Then, using the shift map θ in Corollary 5.3, we find that the equality

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} f(x_n) = \int f(x)\,\pi_i(dx)$$

holds almost surely for π_i-almost every initial condition x_0 (which is the same as to say that it holds for $\mathbf{P}_{\pi_i}$-almost every sequence x). Since $\int f(x)\,\pi_1(dx) \neq \int f(x)\,\pi_2(dx)$ by assumption, this implies that π_1 and π_2 are mutually singular. □

This structure theorem allows to draw several important conclusions concerning the set of all invariant probability measures of a given Markov process. For example, we have that

Corollary 5.12 *If a Markov process with transition operator T has a unique invariant measure π, then π is ergodic.*

Proof. In this case $\mathcal{I}(T) = \{\pi\}$, so that π is an extremal of $\mathcal{I}(T)$. □

In a rather analogous way, one has the following extension of Proposition 4.26:

Proposition 5.13 *Let $\mathcal{A}$ be an invariant set for P and let $\mathcal{A}_n$ be defined as in (4.13). If $\bigcup_{n\geq 0} \mathcal{A}_n = \mathcal{X}$ and $\mathcal{A}$ can be written as a disjoint union of closed sets*

$$\mathcal{A} = \bigsqcup_{k=1}^{m} \mathcal{B}_k\,,$$

with the property that every $\mathcal{B}_k$ is invariant for P and the Markov process restricted to $\mathcal{B}_k$ has a unique invariant measure π_k, then the π_k are ergodic and they are the only ergodic invariant measures for that process.

Proof. The ergodicity of the π_k follows from Corollary 5.12. Suppose now that π is an arbitrary invariant measure for the process. It follows from Proposition 4.26 that $\pi(\mathcal{A}) = 1$. Furthermore, it follows as in the proof of the second part of Theorem 5.7 that the restriction of π to $\mathcal{B}_k$ is again an invariant measure for P. Since on the other hand we assumed that the process restricted to $\mathcal{B}_k$ has a unique invariant measure π_k, this shows that $\pi = \sum_k \pi(\mathcal{B}_k)\,\pi_k$. □

Let us finish this course with a final example. Consider a sequence ξ_n of i.i.d. random variables that take the values ± 1 with equal probabilities and fix some small value $\varepsilon > 0$. Define a Markov process $\{x_n\}$ so that, given x_n, x_{n+1} is the solution at time 1 to the differential equation

$$\frac{dx(t)}{dt} = \sin x(t) + \varepsilon\xi_n \sin \tfrac{x(t)}{2}\,, \quad x(0) = x_n\,.$$

It is a good exercise to check the following facts:

- The measures $\delta_{2k\pi}$ with $k \in \mathbf{Z}$ are invariant (and therefore ergodic because they are δ-measures) for this Markov process.
- For ε sufficiently small (how small approximately?), the sets of the form $[(2k+3/4)\pi, (2k+5/4)\pi]$ with $k \in \mathbf{Z}$ are invariant and there exists a unique (and therefore ergodic) invariant measure on each of them.
- The invariant measures that were just considered are the only ergodic invariant measures for this system.

Appendix A Measurable and topological spaces

This section contains some definitions from measure theory that are taken to be granted. They are only included here so that the course is self-contained. It also contains a few notations that are used throughout this course.

Given a set $\mathcal{X}$, denote by $2^{\mathcal{X}}$ the set of all subsets of $\mathcal{X}$.

A **measurable space** $(\mathcal{M}, \mathscr{F})$ consists of a set $\mathcal{M}$ equipped with a σ-algebra $\mathscr{F}$, *i.e.* a subset $\mathscr{F} \subset 2^{\mathcal{M}}$ such that:

- $\emptyset \in \mathscr{F}$ and $\mathcal{M} \in \mathscr{F}$.
- If $A \in \mathscr{F}$, then $A^c \in \mathscr{F}$, where A^c denotes the complement of A.
- If $\{A_0, A_1, \ldots\} \subset \mathscr{F}$, then $\bigcup_{n=0}^{\infty} A_n \in \mathscr{F}$ and $\bigcap_{n=0}^{\infty} A_n \in \mathscr{F}$.

In other words, $\mathscr{F}$ is closed under complementation, countable unions, and countable intersections. Elements of $\mathscr{F}$ are called **measurable sets**. A function between measurable spaces is **measurable** if the preimages of measurable sets are measurable sets.

A **topological space** $(\mathcal{X}, \mathcal{T})$ consists of a set $\mathcal{X}$ equipped with a topology $\mathcal{T}$, *i.e.* a subset $\mathcal{T} \subset 2^{\mathcal{X}}$ such that:

- $\emptyset \in \mathcal{T}$ and $\mathcal{X} \in \mathcal{T}$.
- If $\{A_0, A_1, \ldots, A_N\} \subset \mathcal{T}$, then $\bigcap_{n=0}^{N} A_n \in \mathcal{T}$.
- If $\mathcal{A} \subset \mathcal{T}$, then $\bigcup_{A \in \mathcal{A}} A \in \mathcal{T}$.

In other words, $\mathcal{T}$ is closed under arbitrary unions and finite intersections. Elements of $\mathcal{T}$ are called **open sets**. A function between topological spaces is **continuous** if the preimages of open sets are open sets.

Given a topological space $(\mathcal{X}, \mathcal{T})$, we define $\mathscr{B}(\mathcal{X})$ as the smallest σ-algebra on $\mathcal{X}$ containing $\mathcal{T}$. This particular σ-algebra is called the **Borel σ-algebra** of $\mathcal{X}$. In other words, the Borel σ-algebra is the smallest σ-algebra such that all open sets are measurable. We denote by $\mathcal{B}_b(\mathcal{X})$ the (Banach) space of all Borel-measurable and bounded functions from $\mathcal{X}$ to $\mathbf{R}$ equipped with the norm

$$\|\varphi\|_\infty = \sup_{x \in \mathcal{X}} |\varphi(x)| \; . \tag{A.1}$$

We denote by $\mathcal{C}_b(\mathcal{X})$ the (Banach) space of all continuous and bounded functions from $\mathcal{X}$ to $\mathbf{R}$ equipped with the same norm as in (A.1).

All the measurable spaces we will consider are topological spaces equipped with their Borel σ-algebra.

A.1 Measures

Given a measurable space $(\mathcal{M}, \mathscr{F})$, a **measure** μ on $\mathcal{M}$ is a function from $\mathscr{F}$ to $\mathbf{R}_+$ with the following properties:

- $\mu(\emptyset) = 0$.
- If $\{A_0, A_1, \ldots\} \subset \mathscr{F}$ is a collection of pairwise disjoint sets, then

$$\mu\Big(\bigcup_{n=0}^{\infty} A_n\Big) = \sum_{n=0}^{\infty} \mu(A_n) \; .$$

We will call $\mu(\mathcal{M})$ the **mass** of μ. A **signed measure** μ is a function from $\mathscr{F}$ to $\mathbf{R}$ with the property that there exists two measures μ_+ and μ_- such that $\mu(A) = \mu_+(A) - \mu_-(A)$ for every $A \in \mathscr{F}$.

Given a measure space $(\mathcal{M}, \mathscr{F}, \mu)$, we denote by $\bar{\mathscr{F}}$ the **completion** of $\mathscr{F}$ with respect to μ. The σ-algebra $\bar{\mathscr{F}}$ is defined to be the smallest σ-algebra with the properties that $\mathscr{F} \subset \bar{\mathscr{F}}$ and that if $A \in \bar{\mathscr{F}}$, $\mu(A) = 0$ and $B \subset A$, then $B \in \bar{\mathscr{F}}$. In the particular case where $\mathcal{M} = [0, 1]$, $\mathscr{F}$ consists of the Borel sets, and μ is the Lebesgue measure, $\bar{\mathscr{F}}$ consists precisely of the Lebesgue measurable sets. This is why $\bar{\mathscr{F}}$ is also called the **Lebesgue completion** of $\mathscr{F}$ with respect to μ.

A **probability space** $(\Omega, \mathscr{F}, \mathbf{P})$ consists of a measurable space $(\Omega, \mathscr{F})$ and a **probability measure P** on Ω, *i.e.* a measure on Ω such that $\mathbf{P}(\Omega) = 1$. We denote the set of probability measures on Ω by $\mathcal{P}(\Omega)$.

A.2 Weak, strong, and total variation convergence

Let $\mu_1, \mu_2, \dots$ be a sequence of measures on a topological space $\mathcal{X}$. We say that the sequence converges **weakly** to a limit μ if

$$\lim_{n\to\infty} \int_{\mathcal{X}} f(x)\, \mu_n(dx) = \int_{\mathcal{X}} f(x)\, \mu(dx)\,, \tag{A.2}$$

for every $f \in \mathcal{C}_b(\mathcal{X})$. We say that it converges **strongly** if (A.2) holds for every $f \in \mathcal{B}_b(\mathcal{X})$. We define the **total variation** distance between two measures μ and ν by

$$\|\mu - \nu\|_{\text{TV}} = \frac{1}{2} \sup_{\substack{f\in\mathcal{B}_b(\mathcal{X}) \\ \|f\|_\infty = 1}} \Big| \int_{\mathcal{X}} f(x)\, \mu(dx) - \int_{\mathcal{X}} f(x)\, \nu(dx) \Big|\,, \tag{A.3}$$

Another equivalent definition of the total variation distance is

$$\|\mu - \nu\|_{\text{TV}} = \sup_{A\subset\mathcal{X}} |\mu(A) - \nu(A)|\,,$$

where the supremum runs over all measurable subsets of $\mathcal{X}$. Finally, if we denote by D_μ and D_ν the densities of μ and ν with respect to the measure $\eta = \frac{1}{2}(\mu + \nu)$ (these densities can easily be shown to exist by the Radon-Nikodym theorem), then one has the equality

$$\|\mu - \nu\|_{\text{TV}} = \int_{\mathcal{X}} |D_\mu(x) - D_\nu(x)|\, \eta(dx)\,.$$

We say that a sequence $\{\mu_n\}$ converges in total variation to a limit μ if

$$\lim_{n\to\infty} \|\mu_n - \mu\|_{\text{TV}} = 0\,.$$

Even though it may look at first sight as if convergence in total variation was equivalent to strong convergence, this is not true as can be seen in Example A.5 below.

It is also a fact that under very mild conditions on $\mathcal{X}$ (being a complete separable metric space is more than enough), (A.3) is the same as the seemingly weaker norm,

$$\|\mu - \nu\|_{\text{TV}} = \sup_{\substack{f\in\mathcal{C}_b(\mathcal{X}) \\ \|f\|_\infty = 1}} \Big| \int_{\mathcal{X}} f(x)\, \mu(dx) - \int_{\mathcal{X}} f(x)\, \nu(dx) \Big|\,, \tag{A.4}$$

where the supremum only runs over continuous bounded functions.

A.3 Examples

Example A.1 The interval $[0, 1]$ equipped with its Borel σ-algebra and the Lebesgue measure is a probability space.

Example A.2 The half-line $\mathbf{R}_+$ equipped with the measure

$$\mathbf{P}(A) = \int_A e^{-x}\, dx$$

is a probability space. In such a situation, where the measure has a density with respect to Lebesgue measure, we will also use the short-hand notation $\mathbf{P}(dx) = e^{-x}\, dx$.

Example A.3 Given $a \in \Omega$, the measure δ_a defined by

$$\delta_a(A) = \begin{cases} 1 & \text{if } a \in A, \\ 0 & \text{otherwise.} \end{cases}$$

is a probability measure.

Example A.4 Let $\{a_n\}_{n\geq 0} \subset \mathbf{R}$ be a sequence such that $\lim_{n\to\infty} a_n = a$ exists. Then, the sequence δ_{a_n} converges weakly to δ_a, but does not converge strongly.

Example A.5 Let Ω be the unit interval and define the probability measures

$$\mu_n(dx) = (1 + \sin(2\pi n x))\, dx\ .$$

Then, μ_n converges to the Lebesgue measure weakly and strongly, but not in total variation. (This resut is also called Riemann's lemma and is well-known in Fourier analysis.)

Example A.6 The sequence $\mathcal{N}(1/n, 1)$ of normal measures with mean $1/n$ and variance one converges to $\mathcal{N}(0, 1)$ in total variation (and therefore also weakly and strongly).

References

[Dia88] P. DIACONIS. *Group representations in probability and statistics*. Institute of Mathematical Statistics Lecture Notes—Monograph Series, 11. Institute of Mathematical Statistics, Hayward, CA, 1988.

[DPZ96] G. DA PRATO and J. ZABCZYK. *Ergodicity for Infinite Dimensional Systems*, vol. 229 of *London Mathematical Society Lecture Note Series*. University Press, Cambridge, 1996.

[FWY00] H. FÖLLMER, C.-T. WU, and M. YOR. On weak Brownian motions of arbitrary order. *Ann. Inst. H. Poincaré Probab. Statist.* **36**, no. 4, (2000), 447–487.

[Law95] G. F. LAWLER. *Introduction to stochastic processes*. Chapman & Hall Probability Series. Chapman & Hall, New York, 1995.

[MT94] S. P. MEYN and R. L. TWEEDIE. *Markov Chains and Stochastic Stability*. Springer, New York, 1994.

[Nor98] J. R. NORRIS. *Markov chains*, vol. 2 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998. Reprint of 1997 original.

[Pet89] K. PETERSEN. *Ergodic theory*, vol. 2 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1989. Corrected reprint of the 1983 original.

[RB03] L. REY-BELLET. Ergodic properties of markov processes, 2003. Lecture Notes of the 2003 Grenoble Summer School on Open Quantum Systems.

[Rev84] D. REVUZ. *Markov chains*, vol. 11 of *North-Holland Mathematical Library*. North-Holland Publishing Co., Amsterdam, second ed., 1984.

[SC04] L. SALOFF-COSTE. Random walks on finite groups. In *Probability on discrete structures*, vol. 110 of *Encyclopaedia Math. Sci.*, 263–346. Springer, Berlin, 2004.

[Shi84] A. N. SHIRYAYEV. *Probability*, vol. 95 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1984. Translated from the Russian by R. P. Boas.

[Vil03] C. VILLANI. *Topics in optimal transportation*, vol. 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003.