Markov Processes

Martin Hairer and Xue-Mei Li

April 29, 2020

Contents

Chapter 1

Introduction

Module: Markov Processes 2019-2020 (MATH96062/MATH97216/MATH97220)
Lecturer: Xue-Mei Li
Office hour: Tuesdays 11:30, office 6M51
CW1 (5% each): submit to office 11/Feb
CW2 (5% each): submit to office 10/March

Lectures

Thursday 12:00-2:00pm, (HXLY 130 first week, then HXLY 140)

Fridays 1:00pm-2:00pm (HXLY 144 first week, then HXLY 342).

References:

- Markov Processes, A. Eberle, on line and book form,

  https://wt.iam.uni-bonn.de/fileadmin/WT/Inhalt/people/Andreas_Eberle/MarkovProcesses1920 /MarkovProcesses1920.pdf

- Markov Chains, James Norris

- Markov Chains and stochastic stability, Meyn and Tweedie (http://probability.ca/MT/BOOK.pdf)

- Markov Chains and Mixing Times, by David A. Levin Yuval Peres Elizabeth L. Wilmer https://pages.uoregon.edu/dlevin/MARKOV/markovmixing.pdf

- Markov Processes and Applications Algorithms, Networks, Genome and Finance, E. Pardoux

  https://doc.lagout.org/science/0_Computer

1.1 Prologue

A stochastic process is an evolution in time of a stochastic system. The state of the process at time t is denoted by x_t. So we can have a stochastic process (x_t), $t \in [0, \infty)$ for example. For discrete times we tend to use the notation (x_n) where $n = 0, 1, 2, \ldots$. The state space, which we denote by $\mathcal{X}$, will be assumed to be a separable complete metric space. A sample of a process is a function of time (a sequence, in case of discrete time). To obtain informations on a stochastic process, for example the averages or the averages of a function of the process, e.g $\mathbf{E}f(x_n)$, one assumes naturally that the x_n's are random variables (i.e. for each n, $x_n : \Omega \to \mathcal{X}$ is measurable).

Markov processes describe the time-evolution of random systems that do not have any memory. A Markov process moves by a rule that may depend on its current position, but never depend on its previous positions.

Consider a deterministic sequence:

$$(x_1, x_2, x_3, \ldots,).$$

The rule $x_n = x_n^2$ determines a Markov chain (x_n); the rule $x_{n+1} = \frac{1}{2}(x_n + x_{n-1})$ implies that (x_n) is not a Markov chain. (However setting $y_n = (x_{n-1}, x_n)$, then (y_n) is a Markov chain.)

$$y_{n+1} = (x_n, x_{n+1}) = (x_n, \frac{1}{2}x_n) + (0, x_{n-1})$$

is determined by y_n, so y_n is a Markov chain.)

Similarly the solution of the ODE $\dot{x}_t = f(x_t)$ is a deterministic Markov process: given the initial point at the initial time we know its future value $x_t = x + \int_s^t f(x_r)dr$, we do not need to know tits before the initial time s. A Markov chain moves to the next step according to the probability distribution determined by its current position. For example let us move a chess piece on an empty chessboard in the following manner: it moves to one of its nearest neighbours in equal probability. This is a Markov chain with state space $\mathcal{X} = \{s_1, s_2, \ldots s_{64}\}$, each state is one of the 64 squares.

Let us demonstrate what we mean by memoryless 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.

1.2 Introduction

A stochastic process (x_n) is a collection of random variables on some probability space, where $n \in \mathbf{N}\cup\{0\}$ is perceived as time. Denote the probability space by $(\Omega, \mathcal{F}, \mathbf{P})$. The 'time' in a continuous time process (x_t) takes values in an interval. We focus mainly on the case of discrete time and therefore give the following definition of a stochastic process.

Definition 1.2.1 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 $(\Omega, \mathcal{F}, \mathbf{P})$. Given n, we refer to x_n as the value of the process at time n. We will sometimes consider processes where the time can take negative values, *i.e.* $\{x_n\}_{n\in\mathbf{Z}}$.

If the time index I is an interval, then a stochastic process (x_t) where $t \in I$ is again a collection a collection $\{x_t, t \in I\}$ of $\mathcal{X}$-valued random variables.

An example of a sequence are independent random variables. In general the random variables are correlated, how are the random variables correlated? More importantly how does one deduce information on a future time x_t from its past up to time s where $s < t$?

A stochastic process is said to have the 'Markov Property' , if its future values depend on its history only through its present values (not on the full past path). Discrete time Markov processes are known as Markov chains. Just imagine we are given a set of rules that depend only on the value of the current values, not any earlier historical values. For example, a sequence of independent random variables (Y_n) is a Markov process. If x_n evolves with a set of rules, given by recursive formula based on information on x_{n-1}, and at each step an independent noise, then (x_n) is also a Markov process.

The transitions

$$x_0 \to x_1 \to x_2 \to \dots$$

induce a family of probability measures on the state space. We are concerned with the following questions: what is the probability that the Markov chain visit state j at time n given that $x_0 = i$? Is there an initial probability distribution μ, such that for each n, x_n is distributed as x_0? Such a probability distribution is called an invariant distribution (or an invariant probability measure). Is such an invariant distribution unique? Starting from different initial distributions, do the Markov chain look alike after some time? If P_x^n and P_y^n denote the probability distributions of the chain at time n with initial points x, y, does the two probability measures get close? In other words

$$|P_x^n - P_y^n| \to 0?$$

What are the techniques for studying these problems?

1.2.1 Definition of Markov chains

If x_n has only a finite number of a countable number of states, then $\sum_{j \in \mathcal{X}} \mathbf{P}(x_n = j) = 1$, we can define the Markov property using elementary probabilities: If $P(B) \neq 0$, then one defines $\mathbf{P}(A|B) := \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)}$.

Definition 1.2.2 Suppose that (x_n) takes values in a finite or a countable set $\mathcal{X}$. It is a Markov chain if for any $n = 1, 2, \dots$ and for any $s_1, \dots s_{n+1} \in \mathcal{X}$ such that $\mathbf{P}(x_0 = s_0, \dots, x_n = s_n) > 0$, we have

$$\mathbf{P}\left(x_{n+1} = s_{n+1} | x_0 = s_0, \dots, x_n = s_n\right) = \mathbf{P}\left(x_{n+1} = s_{n+1} | x_n = s_n\right).$$

Notation: By $\{x_n = k\}$ we mean $\{\omega : x_n(\omega) = k\}$, and $\{x_0 = s_0, \ldots, x_n = s_n\}$ is another expression for $\cap_{i=0}^{n}\{x_i = s_i\}$.

The event we are conditioning is:

$$\{x_0 = s_0, \ldots, x_n = s_n\} = \cap_{j=0}^{n}\{\omega : x_j(\omega) = s_j\}.$$

If $\mathcal{X}$ is a general complete separable metric space with its Borel σ-algebra, it is conceivable that each $\mathbf{P}(x_n = s\} = 0$ for any state s, we use the concept of conditional expectation of an intergable random variables, as below.

1.2.2 Conditional expectation

Definition 1.2.3 Let X be a real-valued random variable on some probability space $(\Omega, \mathcal{F}, \mathbf{P})$ such that $\mathbf{E}|X| < \infty$ and let $\mathcal{F}'$ be a sub σ-algebra of $\mathcal{F}$. Then the **conditional expectation** of X with respect to $\mathcal{F}'$ is a $\mathcal{F}'$-measurable random variable X' such that

$$\int_A X(\omega)\,\mathbf{P}(d\omega) = \int_A X'(\omega)\,\mathbf{P}(d\omega)\;, \tag{1.2}$$

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

Proposition 1.2.4 *With the notations as above, the conditional expectation $X' = \mathbf{E}(X \,|\, \mathcal{F}')$ exists and is essentially unique (in the sense any two such variables are equal almost surely).*

Proof. Denote by ν the restriction of $\mathbf{P}$ to $\mathcal{F}'$ and define the measure μ on $(\Omega, \mathcal{F}')$ by $\mu(A) = \int_A X(\omega)\,\mathbf{P}(d\omega)$ for every $A \in \mathcal{F}'$. 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. □

Notation. If $A \in \mathcal{F}$ we dfine:

$$\mathbf{P}(A \,|\mathcal{F}') := \mathbf{E}\left(\mathbf{1}_A \,|\mathcal{F}'\right).$$

Also if $\mathcal{F}' = \sigma(Y)$, be that generated by a random variable. Then we use the notation:

$$\mathbf{E}(X|Y) := \mathbf{E}\left(X \,|\sigma(Y)\right).$$

1.2.3 Markov chain in general state space

Definition 1.2.5 Let $\sigma(x_0, \ldots, x_n)$ denote the σ-algebra generated by the random variables inside the bracket, this is the smallest σ-algebra such that each of the random variables are measurable.

Definition 1.2.6 A stochastic process (x_n) with state space $\mathcal{X}$ is a Markov chain if for any Borel measurable set A of $\mathcal{X}$, any $n \geq 0$,

$$\mathbf{P}\left(x_{n+1} \in A \mid \sigma(x_0, \ldots x_n)\right) = \mathbf{P}\left(x_{n+1} \in A \mid \sigma(x_n)\right) \quad a.s.$$

The distribution of the random variable x_0 is called the initial distribution.

Notation.

$$\mathbf{P}\left(x_{n+1} \in A \mid x_n\right) := \mathbf{P}\left(x_{n+1} \in A \mid \sigma(x_n)\right).$$

Intuitively, the best estimates based on information obtained from $x_0, x_1, \ldots, x_n$, is the same as the best estimates based on information obtained from x_n alone.

Proposition 1.2.7 *A stochastic process (x_n) with state space $\mathcal{X}$ is a Markov chain if and only if the following holds for any bounded measurable functions $f : \mathcal{X} \to \mathbf{R}$ such that*

$$\mathbf{E}(f(x_{n+1})|\sigma(x_0, \ldots, x_n)) = \mathbf{E}(f(x_{n+1})|x_n). \tag{1.3}$$

Exercise: prove it. By the proposition, (1.3) can be used to give an equivalent definition for the Markov process.

Remark 1.2.8 By the Markovian property, $\mathbf{E}(f(x_{n+1})|\sigma(x_0, \ldots, x_n))$ is $\sigma(x_n)$ measurable then there exists a Borel measurable function G such that $\mathbf{E}(f(x_{n+1})|\sigma(x_0, \ldots, x_n)) = G(x_n)$ a.s.

Let us further explore the meaning of the Markovian property. If $A \in \mathcal{B}(\mathcal{X})$, $B = \{x_n \in A\}$ is a set in $\sigma(x_0, \ldots x_n)$. By the definition of conditional expectations,

$$\int_B f(x_{n+1}) d\mathbf{P} = \int_B \mathbf{E}\left(f(x_{n+1}) \mid \sigma(x_0, \ldots x_n)\right) d\mathbf{P},$$

Since the above holds for every set $B \in \sigma(x_n)$, if furthermore $\mathbf{E}\left(f(x_{n+1})| \, \sigma(x_0, \ldots x_n)\right)$ is $\sigma(x_n)$-measurable, then $\mathbf{E}(f(x_{n+1})|\sigma(x_0, \ldots, x_n)) = \mathbf{E}(f(x_{n+1})|x_n)$ a.s.

1.3 Markov chains on discrete state spaces

Let us assume that $\{Y_n\}$ are independent random variables with state space $\mathcal{X}_i$, by this we mean if $A_i \in \sigma(Y_i)$,

$$\mathbf{P}(A_1 \cap \cdots \cap A_n) = \Pi_{i=1}^n \mathbf{P}(A_i).$$

In other words, for any $f_i : \mathcal{X}_i \ to \mathbf{R}$ bounded measurable, then

$$\mathbf{E}\Pi_i f(Y_i) = \Pi_i \mathbf{E} f(Y_i).$$

Example 1.3.1 (A Random Dynamical System) Suppose that $\mathcal{X}$ is a discrete space and $(\mathcal{Y}, \mathcal{G})$ a measurable space. Let $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{X}$ be a bounded measurable map. Suppose that (Y_n) are independent random variables with state space $\mathcal{Y}$, and is independent of x_0, (In other words, $x_0, Y_1, Y_2, \dots$ are all independent). Set

$$x_{n+1} = f(x_n, Y_{n+1}), \quad n \ge 0.$$

Then (x_n) is a Markov chain.

Proof. Since f is bounded it is integrable. We compute this,

$$\begin{aligned}
\mathbf{P}\left(x_{n+1} = s_{n+1} | x_0 = s_0, \dots, x_n = s_n\right) &= \frac{\mathbf{P}\left(x_{n+1} = s_{n+1}, x_0 = s_0, \dots, x_n = s_n\right)}{\mathbf{P}(x_0 = s_0, \dots, x_n = s_n)} \\
&= \frac{\mathbf{P}\left(f(x_n, Y_{n+1}) = s_{n+1}, x_0 = s_0, \dots, x_n = s_n\right)}{\mathbf{P}(x_0 = s_0, \dots, x_n = s_n)} \\
&= \frac{\mathbf{P}\left(f(s_n, Y_{n+1}) = s_{n+1}, x_0 = s_0, \dots, x_n = s_n\right)}{\mathbf{P}(x_0 = s_0, \dots, x_n = s_n)} \\
&= \mathbf{P}\left(f(s_n, Y_{n+1}) = s_{n+1}\right).
\end{aligned}$$

In the final line we use the fact that Y_{n+1} is independent of $\{x_0, x_1, \dots, x_n\}$. The same arguments shows that

$$\mathbf{P}\left(x_{n+1} = s_{n+1} | x_n = s_n\right) = \frac{\mathbf{P}\left(f(x_n, Y_{n+1}) = s_{n+1}, x_n = s_n\right)}{\mathbf{P}(x_n = s_n)} = \mathbf{P}\left(f(s_n, Y_{n+1}) = s_{n+1}\right).$$

□

Example 1.3.2 Random walks: let ξ_i be independent identically distributed random variable. Define $S_0 = 0$, $S_n = \sum_{i=1}^n \xi_i$ for $n = 1, 2, \dots$. Then (S_n) is a Markov chain.

1.3.1 Transition Probabilities for time homogeneous Markov chains

By the initial distribution of the stochastic process $(x_n)_{n=0}^\infty$, we mean the distribution of x_0.

Definition 1.3.1 A time homogeneous Markov chain on a countable state space $\mathcal{X}$ with initial distribution μ and transition probabilities P_{ji} is a stochastic process such that the following holds:

(1) $\mathbf{P}(x_0 = i) = \mu(i)$,

(2) for any $i_j \in \mathcal{X}$ and $n = 1, 2, \dots$,

$$\mathbf{P}\left(x_{n+1} = i_{n+1} | x_0 = i_0, \dots, x_n = i_n\right) = \mathbf{P}\left(x_{n+1} = i_{n+1} | x_n = i_n\right) = P_{i_{n+1} i_n}.$$

Note that

$$\sum_{j\in\mathcal{X}} P_{ji} = 1.$$

It turns out these set of numbers P_{ij} will determine the probability that a Markov chain takes a particular path, which we describe below.

The following elementary fact will be used in the discussion later. If $\{C_i\}_{i=1}^{\infty}$ is a partition of Ω, then

$$\mathbf{P}(A|B) = \sum_{i=1}^{\infty} \mathbf{P}(A\cap C_i|A).$$

If x_i is a random variable then $\{x_i = k\}$ where $k\in\mathcal{X}$ is a partition of Ω.

Proposition 1.3.2 *Let (x_n) be a Markov chain with transition probabilities P_{ji} with initial distribution μ. Then, for any state i_j and any $n\ge 0$,*

$$\mathbf{P}\left(x_0 = i_0, \dots, x_n = i_n, x_{n+1} = i_{n+1}\right) = P_{i_{n+1},i_n} P_{i_n,i_{n-1}} \dots P_{i_1 i_0}\, \mu(i_0). \tag{1.4}$$

Proof. We prove it by induction on the time n for which the identity holds.

$$\mathbf{P}\left(x_0 = i_0, x_1 = i_1\right) = \mathbf{P}(x_1 = i_1|x_0 = i_0)P(x_0 = i_0) = P_{i_1,i_0}\mu(i_0).$$

Suppose the identity holds on n-times:

$$\mathbf{P}\left(x_0 = i_0, \dots, x_n = i_n, x_n = i_n\right) = P_{i_n,i_{n-1}} \dots P_{i_1 i_0}\mu(i_0).$$

Then,

$$\begin{aligned}&\mathbf{P}\left(x_0 = i_0, \dots, x_n = i_n, x_{n+1} = i_{n+1}\right)\\ &= \mathbf{P}\left(x_{n+1} = i_{n+1}|x_0 = i_0, \dots, x_n = i_n, x_n = i_n\right)\mathbf{P}(x_0 = i_0, \dots, x_n = i_n, x_n = i_n)\\ &\overset{\text{Markov Property}}{=} \quad P_{i_{n+1},i_n}\,\mathbf{P}(x_0 = i_0, \dots, x_n = i_n, x_n = i_n).\end{aligned}$$

The rest follows by the induction hypothesis on $\mathbf{P}(x_0 = i_0, \dots, x_n = i_n, x_n = i_n)$. □

Corollary 1.3.3

$$\mathbf{P}\left(x_{n+1} = i_{n+1}, \dots x_{n+m} = i_{n+m}\,|\, x_0 = i_0, \dots, x_n = i_n\right) = \Pi_{k=n}^{n+m-1} P_{i_{k+1}i_k}. \tag{1.5}$$

Exercise 1.3.1 Show that for any $m\ge 0$,

$$\begin{aligned}&\mathbf{P}(x_{n+m} = k_n, x_{n+m-1} = k_{n-1}, \dots, x_{m+1} = k_1\,|\, x_m = i) = \mathbf{P}_{k_n k_{n-1}} P_{k_{n-1}k_{n-2}} \dots P_{k_1 i}\\ &\mathbf{P}(x_{n+m} = k_n, x_{n+m-1} = k_{n-1}, \dots, x_{m+1} = k_1, x_m = i) = \mathbf{P}_{k_n k_{n-1}} P_{k_{n-1}k_{n-2}} \dots P_{k_1 i}\,\mathbf{P}(x_m = i).\end{aligned} \tag{1.6}$$

Proposition 1.3.4 *For any $n \geq 2, m \geq 0$ and any $i, j \in \mathcal{X}$,*

$$\mathbf{P}(x_{n+m} = j \mid x_m = i) = \sum_{k_{n-1\in\mathcal{X}}} \cdots \sum_{k_1\in\mathcal{X}} P_{jk_{n-1}} P_{k_{n-1}k_{n-2}} \dots P_{k_1 i}.$$

Proof. Firstly,

$$\begin{aligned}
\mathbf{P}(x_{n+m} = j \mid x_m = i) &= \frac{\mathbf{P}(x_{n+m} = j, x_m = i)}{P(x_m = i)} \\
&= \sum_{k_{m-1\in\mathcal{X}}} \cdots \sum_{k_1\in\mathcal{X}} \mathbf{P}(x_{n+m} = j, x_{n+m-1} = k_{n-1}, \dots, x_{m+1} = k_1, x_m = i) \frac{1}{P(x_m = i)} \\
&= \sum_{k_{n-1\in\mathcal{X}}} \cdots \sum_{k_1\in\mathcal{X}} P_{jk_{n-1}} P_{k_{n-1}k_{n-2}} \dots P_{k_1 i}
\end{aligned}$$

We have used (1.6) in the last step. Finally,

$$\mathbf{P}(x_n = j) = \sum_{k\in\mathcal{X}} \mathbf{P}(x_n = j, x_0 = k) = \sum_{k\in\mathcal{X}} \mathbf{P}(x_n = j | x_0 = k)\mathbf{P}(x_0 = k) = \sum_{k\in\mathcal{X}} P^n_{jk}\mu(k).$$

□

Consequently, we may denote the right hand side by P^n_{ji}:

$$P^n_{ji} = \sum_{k_{n-1\in\mathcal{X}}} \cdots \sum_{k_1\in\mathcal{X}} P_{jk_{n-1}} P_{k_{n-1}k_{n-2}} \dots P_{k_1 i}.$$

We emphasize the definition below:

Definition 1.3.5 If (x_n) is time homogeneous Markov chain on $\mathcal{X} = \{1, 2, \dots, \}$. We define for $n \geq 1$, $P^n_{ji} = \mathbf{P}(x_{n+m} = j \mid x_m = i)$.

Note: $P_{ij} = P^1_{ij}$.

Corollary 1.3.6 *For any $n \geq 1$, $k \in \mathcal{X}$,*

$$\mathbf{P}(x_n = j) = \sum_{k\in\mathcal{X}} P^n_{jk}\mu(k).$$

Theorem 1.3.7 *(The Champman-Kolmogorov equation) For any $i, j \in \mathcal{X}$ and $n, m \geq 1$, we have*

$$P^{n+m}_{ji} = \sum_{k\in\mathcal{X}} P^m_{jk} P^n_{ki}.$$

Proof. Since $\cup_{k\in\mathcal{X}}\{x_n = k\} = \Omega$, we have

$$\begin{aligned}
P^{n+m}_{ji} &= \mathbf{P}(x_{n+m} = j \mid x_0 = i) = \sum_{i\in\mathcal{X}} \frac{\mathbf{P}(x_{n+m} = j, x_n = k, x_0 = i)}{P(x_0 = i)} \\
&= \sum_{k\in\mathcal{X}} \frac{\mathbf{P}(x_{n+m} = j|x_n = k, x_0 = i)\mathbf{P}(x_n = k, x_0 = i)}{\mathbf{P}(x_0 = i)} \\
&\overset{\text{Markov Property+tower property}}{=} \sum_{k\in\mathcal{X}} \mathbf{P}(x_{n+m} = j|x_n = k)\frac{\mathbf{P}(x_n = k, x_0 = i)}{\mathbf{P}(x_0 = i)} \\
&= \sum_{k\in\mathcal{X}} P^m_{jk}P^n_{ki}.
\end{aligned}$$

□

We have used tower property to invoke the whole history, please complete it (for more detail see Lemma 2.1.7).

If $\mathcal{X} = \{1, \dots, N\}$ then $P = (P_{ij})$ is a $n \times n$-matrix.

Proposition 1.3.8 *If (x_n) is a time-homogeneous Markov chain on $\{1, \dots, N\}$ with transition probabilities P_{ji}, show that*

$$\mathbf{P}(x_{n+m} = j \,|\, x_m = i) = (P^n)_{ji}.$$

Where $P^n = \overbrace{P \times \cdots \times P}^{n}$ denotes matrix multiplication.

Definition 1.3.9 Suppose for $i, j \in \mathcal{X}$, we are given $P_{ij} \geq 0$ with $\sum_{i\in\mathcal{X}} P_{ij} = 1$. Then $P = (P_{ij})$ is called a stochastic matrix.

The sum of each column is 1. Observe that P^n is again a stochastic matrix, whose columns sum to 1: it's the sum of probabilities to reach one of the states, in n-steps.

Every measure on the discrete space $\mathcal{X}$ is determined by its value on the individual point $\mathcal{X}$. Suppose we are given $\nu(i)$, then $\nu(A) = \sum_{i\in A} \mu(i)$ defines a measure. Thus we identify a measure on $\mathcal{X}$ with a column vector with entries $\nu(i)$. Evidently $\nu(i) \geq 0$. Also ν is a probability measure if and only if $\sum_{i\in\mathcal{X}} \nu(i) = 1$.

Suppose that we are given are given (P_{ij}) where $i, j \in \mathcal{X}$. The matrix P acts on the measure ν as matrix multiplication:

$$(P\nu)(i) = \sum_{k\in\mathcal{X}} P_{ik}\nu(k).$$

Thus P transforms a measure ν to a measure $P\nu$. Then, if P is a stochastic matrix,

$$\sum_{i\in\mathcal{X}} (P\nu)(i) = \sum_{k\in\mathcal{X}} \sum_{i\in\mathcal{X}} P_{ik}\nu(k) = \sum_{k\in\mathcal{X}} \nu(k).$$

So the total mass of the new measure $P\nu$ is the same as that of ν.

If (x_n) is a Markov chain with transition matrix P and initial distribution ν, $P\nu$ is the distribution of x_1, …, $P^n\nu$ is the distribution of x_n.

The following can be considered to be a converse to Proposition 1.3.2.

Theorem 1.3.10 *Suppose we are given a stochastic matrix P, a probability measure μ and a stochastic process (x_n). Suppose that the following relation holds for any $n \geq 1$, and for any $i_0, \dots, i_{n+1} \in \mathcal{X}$,*

$$\mathbf{P}\left(x_0 = i_0, \dots, x_n = i_n, x_{n+1} = i_{n+1}\right) = P_{i_{n+1}, i_n} P_{i_n, i_{n-1}} \dots P_{i_1 i_0}\ \mu(i_0).$$

Note this is (1.4). Then (x_n) is a Markov chain with transition probabilities P_{ji} with initial distribution μ.

Proof. Take $n = 1$ in the above, $\mathbf{P}\left(x_0 = i_0, x_i \in i_1\right) = P_{i_1 i_0}\ \mu(i_0)$, summing up $i_1 \in \mathcal{X}$, we get $\mathbf{P}(x_0 = i_0) = \mu(i_0)$. Also,

$$P(x_{n+1} = i_{n+1} \mid x_n = i_n, \dots, x_0 = i_0) = \frac{P(x_{n+1} = i_{n+1}, x_n = i_n, \dots, x_0 = i_0)}{P(x_n = i_n \dots, x_0 = i_0)} = P_{i_{n+1}, i_n}.$$

This means, $P(x_{n+1} = i_{n+1} \mid x_n, \dots, x_0) = P_{i_{n+1}, x_n}$, the right hand side is measurable w.r.t. $\sigma(x_i)$, which means

$$P(x_{n+1} = i_{n+1} \mid x_n, \dots, x_0) = P(x_{n+1} = i_{n+1} \mid x_n),$$

proving the Markov property and that P is its (time-independent) transition probabilities. □

Remark 1.3.11 The Markov property of a stochastic processes is entirely determined by the probability distributions of the family of random variables $(x_0, x_1, \dots, x_n)$ where $n = 1, 2, \dots$.

1.3.2 Pushed forward measures

If $z : \Omega \to \mathcal{X}$ is a measurable function, we may push forward the measure on Ω to a measure on $\mathcal{X}$ as follows. The measure is denoted by $z_*(\mathbf{P})$. For any measurable sets on $\mathcal{X}$,

$$z_*(\mathbf{P})(A) = \mathbf{P}(\{\omega : z(\omega) \in A\}),$$

This is called the probability distribution of z.

Let $\mathcal{X}$ be a metric space, which we assume to be separable, and $\mathcal{B}(\mathcal{X})$ the Borel σ-algebra, the smallest σ-algebra generated by the collection of open subsets.

Definition 1.3.12 Let $\hat{\mathcal{X}} = \pi_{i\in I}\mathcal{X}_i$, each $\mathcal{X}_i$ has a σ-algebra $\mathcal{F}_i$. We define the product σ-algebra to be the smallest σ-algebra on $\hat{\mathcal{X}}$ such that each projection $\pi_i : \hat{\mathcal{X}} \to \mathcal{X}_i$ is measurable. We denote this product σ-algebra by $\otimes_{i=1}^n \mathcal{F}_i$.

The following can be found in Real analysis by Folland, pages 22-23.

Proposition 1.3.13 *If we have a countable product space, each with a σ-algebra $\mathcal{F}_\alpha$. This product σ-algebra is generated by the collection of all rectangles: $\Pi_{i=1}^\infty E_i$ where $E_i \in \mathcal{F}_i$.*

Proposition 1.3.14 *Let $\mathcal{X}_i$ be separable metric spaces. Then,*

$$\mathcal{B}(\pi_{i=1}^n \mathcal{X}_i) = \otimes_{i=1}^n \mathcal{B}(\mathcal{X}_i).$$

The notation $\mathcal{B}(\pi_{i=1}^n \mathcal{X}_i)$ denotes the Borel σ-algebra of the product space endowed with the product metric.

Definition 1.3.15 If (x_t) is a stochastic process with $t \in I$. Then for any $t_1 < \ldots, < t_n$, $t_i \in I$ and any n, we define a family of probability measures $\mu_{t_1,\ldots,t_n}$ on $\mathcal{X}^n = \Pi_{i=1}^n \mathcal{X}$ to be the measure pushed forward by $(x_{t_1}, \ldots, x_{t_n})$. In fact,

$$\mu_{t_1,\ldots,t_n}(A_1 \times \cdots \times A_n) = \mathbf{P}(x_{t_1} \in A_1, \ldots, x_{t_n} \in A_n\}).$$

These are called finite dimensional distributions of the stochastic processes.

For discrete state space and discrete time stochastic processes, it is sufficient to work with $\{x_1 = i_1, \ldots, x_n = i_n)$.

Definition 1.3.16 A family of random variables $(y_1, \ldots, y_n)$ is independent if

$$\mathbf{P}(y_1 \in A_1, \ldots, y_n \in A_n) = \Pi_{i=1}^n \mathbf{P}(y_i \in A_i),$$

for any measurable sets A_i. This is equivalent to say:

$$(y_1, \ldots, y_n)_*(\mathbf{P}) = \otimes_{i=1}^n((y_i)_*(\mathbf{P}).$$

In other words the random variables are independent if and only if their pushed forward measures are products of the probability distributions.

1.4 Questions

Does the distribution of (x_n) converge to some measure? In what sense does it converge? What distance does one put on the space of probability measures? At what speed does the convergence happen?

Definition 1.4.1 A measure π is called an invariant measure if $P\pi = \pi$.

We may expect that $\mu_n = (x_n)_* \mathbf{P} \to \pi$. Does π exist?

Does the distribution of the process on $\mathcal{X}^{Z_+}$ converge to that of the chain with initial distribution π? If we denote by P_μ the probability of the chain with initial μ, we ask $P_\mu \to P_\pi$? Erogodic theorems:

$$\frac{1}{n}\sum_{k=0}^{n-1} f(x_k) \to \int_{\mathcal{X}} f d\pi?$$

Let $\theta(\omega)_k = \omega_{k+1}$, this induces a shit on the sequences,

$$\frac{1}{n}\sum_{k=0}^{n-1} F(x_k, x_{k+1}, \dots,) \to \int_{\mathcal{X}^{Z_+}} F\, dP_\pi?$$

Exercises

Exercise 1.4.1 Let $(\mathcal{X}, \mathcal{F})$ be a measurable space. Let $(x_0, x_1, \dots, x_n)$ be a family of random variables with probability distribution Q on $\pi_{i=0}^n \mathcal{X}$. Let $\hat{\Omega} = \pi_{i=0}^n \mathcal{X}$, with the product σ-algebra, and probability measure $\hat{\Omega}$. Let π_i denotes the projections from Ω to $\mathcal{X}$. Show that $(\pi_1, \dots, \pi_n)$ and $(x_0, x_1, \dots, x_n)$ have the same probability distribution.

Exercise 1.4.2 Let Y_n be i.i.d.'s with $P(Y_n = 1) = \frac{1}{2}$ and $P(Y_n = -1) = -\frac{1}{2}$. Set $x_{n+1} = x_n + Y_n$. (This is the symmetric random walk.) Then x_n takes values in $\mathbf{Z}$. Write down the transition probabilities $\mathbf{P}\left(x_{n+1} = s_{n+1} | x_n = s_n\right)$.

The following model is used in queuing theory.

Exercise 1.4.3 Random walk on $\mathbf{Z}_+$. Set

$$x_n = [x_{n-1} + Y_n]_+$$

where Y_n are i.i.d.'s with distribution $\hat{P}$. Write down the transition probabilities for x_n.

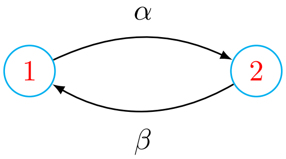
Exercise 1.4.4 If $\{Y_n\}_{n=0}^\infty$ are independent random variables with values in $\{0, 1, 2, \dots, 19\}$. Define

$$x_0 = Y_0, \qquad x_1 = Y_1, \qquad x_{n+1} = x_n + x_{n-1} + y_n, \quad n > 1.$$

- Is (x_n) necessarily a Markov chain? Prove it or give a counter example.
- Let $z_n = (x_n, x_{n-1})$. Is z_n a Markov chain? Justify your answers.

1.5 Two state Markov chains

Let us consider a time-homogeneous Markov chain with two state $\mathcal{X} = \{1, 2\}$ and let $P = \begin{pmatrix} 1-\alpha & \beta \\ \alpha & 1-\beta \end{pmatrix}$. What's the probability that the chain starts from 1 returns to 1 in n-steps? i.e. what is the approximate value of $P_{11}^n = \mathbf{P}(x_n = 1 \,|\, x_0 = 1)$?



Suppose that

$$\mathbf{P}(x_0 = 1) = \nu(1), \quad \mathbf{P}(x_0 = 2) = \nu(2).$$

Then,

$$\mathbf{P}(x_n = 1) = P_{11}^n \nu(1) + P_{12}^n \nu(2), \qquad \mathbf{P}(x_n = 2) \ = P_{21}^n \nu(1) + P_{22}^n \nu(2).$$

Let P^0 be the identity matrix, then

$$\begin{pmatrix} \mathbf{P}(x_n = 1) \\ \mathbf{P}(x_n = 2) \end{pmatrix} = \begin{pmatrix} 1-\alpha & \beta \\ \alpha & 1-\beta \end{pmatrix}^n \begin{pmatrix} \nu(1) \\ \nu(2) \end{pmatrix} = \begin{pmatrix} 1-\alpha & \beta \\ \alpha & 1-\beta \end{pmatrix} \begin{pmatrix} \mathbf{P}(x_{n-1} = 1) \\ \mathbf{P}(x_{n-1} = 2) \end{pmatrix}.$$

Set the initial measure to be $(1, 0)^T$. Then $\mathbf{P}(x_n = 1) = P_{11}^n$, $\mathbf{P}(x_n = 2) = 1 - P_{11}^n$, and

$$\begin{pmatrix} P_{11}^n \\ 1 - P_{11}^n \end{pmatrix} = \begin{pmatrix} 1-\alpha & \beta \\ \alpha & 1-\beta \end{pmatrix} \begin{pmatrix} P_{11}^{n-1} \\ 1 - P_{11}^{n-1} \end{pmatrix} \quad .$$

Thus,

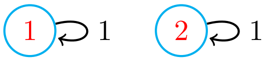
$$P_{11}^n = (1-\alpha-\beta) P_{11}^{n-1} + \beta = (1-\alpha-\beta)((1-\alpha-\beta)P_{11}^{n-2} + \beta) + \beta.$$

This is β if $\alpha + \beta = 1$. If $\alpha + \beta \neq 1$, iterate this to see,

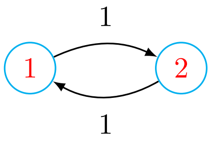
$$\begin{aligned} P_{11}^n &= (1-\alpha-\beta)^n + (1-\alpha-\beta)^{n-1}\beta + \cdots + (1-\alpha-\beta)^\beta + \beta \\ &= \frac{\beta}{\alpha+\beta} + \frac{\alpha}{\alpha+\beta}(1-\alpha-\beta)^n \,. \end{aligned} \quad .$$

By symmetry,

$$P^n = \frac{1}{\alpha+\beta} \begin{pmatrix} \beta + \alpha(1-\alpha-\beta)^n & \beta - \beta(1-\alpha-\beta)^n \\ \alpha - \alpha(1-\alpha-\beta)^n & \alpha + \beta(1-\alpha-\beta)^n \end{pmatrix}.$$

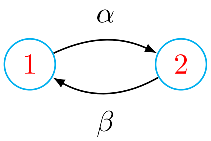


Case 1. $\alpha = \beta = 0$. Then $P = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ is the identity matrix and the chain reduced to two single state Markov chains.



Case 2. $1 = \alpha = \beta$, then $P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. The chain hops from one state to another. It returns to its original state in two steps. This is a 2-periodic Markov chain.

If we define $y_n = x_{2n}$. Then (y_n) is a Markov chain with stochastic matrix given by $P^2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, which can be reduced to two separate Markov chains on $\{1\}$ and $\{2\}$ respectively.



Case 3. $\alpha = 0$, $\beta \neq 0$, then eventually the chain arrives at 1. Similarly if $\beta = 0, \alpha \neq 0$ case.

Case 3. Aperiodic and irreducible. We have $|1 - \alpha - \beta| < 1$.

1. Then as $n \to \infty$,

$$P^n \to \frac{1}{\alpha + \beta} \begin{pmatrix} \beta & \beta \\ \alpha & \alpha \end{pmatrix}.$$

The rate of convergence is exponential.

2. For any initial distribution $(a, 1 - a)$,

$$P^n \begin{pmatrix} a \\ 1 - a \end{pmatrix} \to \begin{pmatrix} \frac{\beta}{\alpha+\beta} & \frac{\beta}{\alpha+\beta} \\ \frac{\alpha}{\alpha+\beta} & \frac{\alpha}{\alpha+\beta} \end{pmatrix} \begin{pmatrix} a \\ 1 - a \end{pmatrix} = \begin{pmatrix} \frac{\beta}{\alpha+\beta} \\ \frac{\alpha}{\alpha+\beta} \end{pmatrix}$$

Observe that

$$P \begin{pmatrix} \frac{\beta}{\alpha+\beta} \\ \frac{\alpha}{\alpha+\beta} \end{pmatrix} = \begin{pmatrix} \frac{\beta}{\alpha+\beta} \\ \frac{\alpha}{\alpha+\beta} \end{pmatrix}.$$

The above convergence indicates that from any initial distribution, the distribution of the chain at time n convergence to the invariant probability distribution (there exists only one such measure), this is ergodicity.

We now repeat this by working out the eigenvalues. It is easy to work out that P has eigenvalue 1 and $\lambda = 1 - \alpha - \beta$. Their corresponding eigenvectors are

$$v_1 = \begin{pmatrix} \beta \\ \alpha \end{pmatrix}, \qquad v_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

Let

$$R = \begin{pmatrix} \beta & 1 \\ \alpha & -1 \end{pmatrix}, \qquad R^{-1} = -\frac{1}{\alpha+\beta} \begin{pmatrix} -1 & -1 \\ -\alpha & \beta \end{pmatrix}.$$

Then,

$$P^n = R \begin{pmatrix} 1 & 0 \\ 0 & \lambda^n \end{pmatrix} R^{-1} \frac{1}{\alpha+\beta} \begin{pmatrix} \beta+\alpha\lambda^n & \beta-\beta\lambda^n \\ \alpha-\alpha\lambda^n & \alpha+\beta\lambda^n \end{pmatrix}.$$

Let us normalise the eigenvector corresponding to the eigenvalue 1 so that the entries sum to 1, then we have

$$\begin{pmatrix} \frac{\beta}{\alpha+\beta} \\ \frac{\alpha}{\alpha+\beta} \end{pmatrix}.$$

This is the stationary probability measure!

1.5.1 Exploring the Markov property

Theorem 1.5.1 *Let (x_n) be a time homogeneous Markov chain with transition probability (P_{ij}) and initial distribution ν, and let s be a given time. Then conditioning on $x_s = i$, (x_{s+n}) is a time homogeneous Markov chain with transition probability (P_{ij}) and initial distribution δ_i, and is independent of $\{x_0, x_1, \dots, x_s\}$*

Proof. The statement 'conditioning on $x_s = i$, (x_{s+n}) is a time homogeneous Markov chain with transition probability $(P_{i,j})$ and initial distribution δ_i' means precisely the following: Let $y_n = x_{s+n}$, then

$$\mathbf{P}(y_0 = i_s, \dots, y_{n+1} = i_{s+n+1} | x_s = i) = P_{i_{s+n+1} i_{s+n}} \dots P_{i_{s+1} i_s} \delta_{i i_s} = \Pi_{k=s}^{s+n} P_{i_{k+1} i_k} \delta_{i i_s}.$$

This follows from (1.5) and

$$\mathbf{P}\left(x_{n+1} = i_{n+1}, \dots x_{n+m} = i_{n+m} \,|\, x_0 = i_0, \dots, x_n = i_n\right) = \mathbf{P}\left(x_{n+1} = i_{n+1}, \dots x_{n+m} = i_{n+m} \,|\, x_n = i_n\right). \tag{1.7}$$

(Check the latter!) The conditional independent parts means for any $A \in \sigma(x_0, \dots, x_s)$,

$$\mathbf{P}((y_0 = i_s, \dots, y_{n+1} = i_{s+n+1}) \cap A | x_s = i) = \mathbf{P}((y_0 = i_s, \dots, y_{n+1} = i_{s+n+1}) | x_s = i) \delta_{i i_s} P(A | x_s = i).$$

It is sufficient to check the above holds for A of the form $A = \{x_0 = i_0, \dots, x_s = i_s\}$ (this is a π-system and generates $\sigma(x_0, \dots, x_s)$. Then the right hand side is $\Pi_{k=s}^{s+n} P_{i_{k+1} i_k} \delta_{i i_s} \delta_{i i_s} P(A | x_s = i)$. The left hand side is

$$\mathbf{P}(y_0 = i_s, \dots, y_{n+1} = i_{s+n+1}, x_0 = i_0, \dots, x_s = i_s | x_s = i) = \frac{\Pi_{k=0}^{n+s} P_{i_{k+1}, i_k} \delta_{i, i_s} \nu(i_0)}{\mathbf{P}(x_s = i)},$$

this equals to the left hand side. □

Chapter 2

General Markov chains

Throughout the chapter the state space for the Markov chain is a separable metric space.

2.1 Basics

Let I be a subset of $\mathbf{R}$, an interval in $\mathbf{R}_+$ or $\mathbf{N} \cup \{0\}$ or Z.

Definition 2.1.1 Let $(\mathcal{F}_s, s \in I)$ be a family of σ-algebra on Ω such that $\mathcal{F}_s \subset \mathcal{F}_t \subset \mathcal{F}$, where $s < t$. This is called a filtration. If (x_t) is a process such that, for each t, x_t is measurable with respect to $\mathcal{F}_t$, we say x_t is adapted to $\mathcal{F}_t$.

If we have a stochastic process indexed by I, for $t \in I$, we denote by $\mathcal{F}_t^0 = \sigma\{x_s : 0 \leq s \leq t, s \in I\}$ the smallest σ-algebra with respect to which each x_s, with $s \leq t$ and $s \in I$, is measurable. This called the natural filtration of the stochastic process. Any stochastic process is adapted to its natural filtration. If x_t is adapted to $\mathcal{F}_t$ then $\mathcal{F}_t$ contains $\mathcal{F}_t^0$.

If we have a discrete time process (x_n), the natural filtration is $\mathcal{F}_n^0 = \sigma\{x_0, \dots, x_n\}$. We would be working with the natural filtration, unless otherwise stated.

Definition 2.1.2 A stochastic process (x_t) is said to be a Markov process (with respect to a filtration $\mathcal{F}_t$) if x_t is adapted to $\mathcal{F}_t$ and if for any measurable subset A of $\mathcal{X}$, any $s, t \in I$, $t > 0$,

$$\mathbf{P}\left(x_{t+s} \in A \mid \mathcal{F}_s\right) = \mathbf{P}\left(x_{t+s} \in A \mid \sigma(x_s)\right) \quad a.s. \tag{2.1}$$

Note that (2.1) means exactly the following: for any $C \in \mathcal{F}_s$,

$$\mathbf{E}\left(\mathbf{1}_A(x_{t+s})\mathbf{1}_C\right) = \mathbf{E}\left(\mathbf{E}(\mathbf{1}_A(x_{t+s}) \mid \sigma(x_s))\mathbf{1}_C\right). \tag{2.2}$$

Theorem 2.1.3 *To show (x_t) satisfies (2.2) for all $A \in \mathcal{B}(\mathbf{R})$ it is sufficient to show it holds for a collection $\mathbf{C}$ of Borel subsets of $\mathbf{R}$ generating $\mathcal{B}(\mathbf{R})$. Similarly one can test with C from a sub-collection of sets of $\mathcal{F}_s$, which is a π-system generating $\mathcal{F}_s$.*

Proof. We only prove the first statement. Let $\mathcal{A} = \{A \in \mathcal{B}(\mathbf{R}) : (2.2)\}$ holds, it contains $\mathbf{C}$. We show that $\mathcal{A}$ is a λ-system. (1) $\mathbf{R}$ is in $\mathbf{R}$ as both sides are $\mathbf{P}(C)$. (2) If $A \subset B$, $\mathbf{1}_{B\setminus A} = \mathbf{1}_B - \mathbf{1}_A$. Using linearity on the sides of (2.2) we see that $B\setminus A \in \mathcal{B}(\mathbf{R})$. (3) If $\mathcal{A}_n$ is an increasing sequence of sets in $\mathcal{A}$ increases to A, then by the monotone convergence theorem, $A \in \mathcal{A}$. By the $\pi - \lambda$ theorem $\mathcal{A} \supset \sigma(\mathbf{C}) = \mathcal{B}(\mathbf{R})$. □

Example 2.1.1 If (x_t) takes values in a discrete space, it sufficient to test with C of the form $\{x_0 = i, \dots, x_n = i\}$.

Proposition 2.1.4 *If x_t is a Markov process with respect to any filtration $\mathcal{F}_t$, it is a Markov process w.r.t. its natural filtration.*

Proof. Since $\sigma(x_s) \subset \sigma(x_r, 0 \le r \le ts) \subset \mathcal{F}_s$, for any $s < t$, A a measurable set in the state space, we apply the tower property, the following holds almost surely:

$$\begin{aligned}\mathbf{P}\left(x_{s+t} \in A \mid \sigma(x_r, 0 \le r \le s)\right) &= \mathbf{E}\left(\mathbf{P}\left(x_{s+t} \in A \mid \ \mathcal{F}_s\right) \mid \sigma(x_r, 0 \le r \le s)\right)\\ &= \mathbf{E}\left(\mathbf{P}\left(x_{s+t} \in A \mid x_s\right) \mid \sigma(x_r, 0 \le r \le s)\right)\\ &= \mathbf{P}\left(x_{s+t} \in A \mid x_s\right),\end{aligned}$$

completting the proof. □

Theorem 2.1.5 *If (x_t) is a Markov process, with filtration $(\mathcal{F}_t)$, then for any bounded Borel measurable $f : \mathcal{X} \to \mathbf{R}$,*

$$\mathbf{E}\big(f(x_t)\,|\mathcal{F}_s\big) = \mathbf{E}\Big(f(x_t) \mid x_s\Big)\,. \tag{2.3}$$

Proof. Let A be a measurable subset of $\mathcal{X}$. Let $f = \mathbf{1}_A$, (2.4) is exactly (2.1). Let $f = \sum_{i=1}^n a_i \mathbf{1}_{A_i}$, an arbitrary simple function, then by the linearity of taking conditional expectations, (2.4) holds also. If f is furthermore non-negative and bounded, then there exists an increasing sequence of non-negative simple functions f_n with limit f, now we apply the conditional monotone convergence theorem (c.f. Prop. 8.2.3 of Measure and Integration notes),

$$\mathbf{E}\big(f(x_t)\,|\mathcal{F}_s\big) = \lim_{n\to\infty} \mathbf{E}\big(f_n(x_t)\,|\mathcal{F}_s\big) = \lim_{n\to\infty} \mathbf{E}\Big(f_n(x_t) \mid x_s\Big) = \mathbf{E}\Big(f(x_t) \mid x_s\Big)\,.$$

Finally let $f = f^+ - f^-$, then $|f| = f^+ - f^-$, and so both f^+ and f^- are non-negative bounded functions, we now apply linearity to conclude. □

2.1.1 Equivalent Definitions for the Markov property

Let us now return to discrete time and $\mathcal{X}$ being a separable metric space. There are a number of other equivalent conditions for the Markov property. We list the more frequently used ones here. We typically set $\mathcal{F}_n = \sigma(x_0, x_1, \dots, x_n) = \vee_{i=0}^n \sigma(x_i)$.

Proposition 2.1.6 *A stochastic process $(x_n)_{n=0}^\infty$ is a Markov process with state space $\mathcal{X}$, if and only if one of the following conditions holds.*

1. *For any $A_i \in \mathcal{B}(\mathcal{X})$, $i = 0, \dots, n$,*

$$\mathbf{P}\Big(x_0 \in A_0, \dots, x_n \in A_n\Big) = \int_{\cap_{i=0}^{n-1}\{x_i \in A_i\}} \mathbf{P}\big(x_n \in A_n \mid x_{n-1}\big)\, d\mathbf{P}.$$

2. *For every $n \in \mathbf{N}$ and for every bounded measurable function $f \colon \mathcal{X} \to \mathbf{R}$ one has*

$$\mathbf{E}\big(f(x_n) \,|\, x_0, x_1, \dots, x_{n-1}\big) = \mathbf{E}\Big(f(x_n) \mid x_{n-1}\Big) \,. \tag{2.4}$$

3. *For any $f_i : \mathcal{X} \to \mathbf{R}$ bounded Borel measurable and for any $n \in \mathbf{N}$,*

$$\mathbf{E}\left(\Pi_{i=1}^n f_i(x_i)\right) = \mathbf{E}\left(\Pi_{i=1}^{n-1} f_i(x_i)\, \mathbf{E}(f_n(x_n)|x_{n-1})\right).$$

Proof. (1) Let $C = \{x_0 \in A_0, \dots, x_{n-1} \in A_{n-1}\}$, then the LHS is $\mathbf{E}(\mathbf{1}_C \mathbf{1}_{A_n}(x_n))$, and so the Markov property holds for these sets, which is a π-system generating $\mathcal{F}_s$.

(2) proved earlier.

(3) Apply tower property and Markov property: $\mathbf{E}(\pi_{i=1}^n f_i(x_i)) = \mathbf{E}(\pi_{i=1}^{n-1} f_i(x_i)\mathbf{E}(f_n(x_n|\mathcal{F}_{n-1})) = \mathbf{E}(\pi_{i=1}^{n-1} f_i(x_i)\mathbf{E}(f_n(x_n|\xi_{n-1}))$. □

The Markov property states conditioning on the whole history up to present time $n-1$ is equivalent to conditioning on x_{n-1}. Below we show that we may replace the whole history by part of the history.

Lemma 2.1.7 *Let x_n be a Markov process. Let $0 \le t_1 < t_2 < \cdots < t_{m-1} < t_m = n-1$ where $n > 1$ and $t_i \in \mathbf{N} \cup \{0\}$. Let $f, h : \mathcal{X} \to \mathbf{R}$ be bounded Borel measurable functions. Then*

$$\mathbf{E}\left(f(x_{n+1})h(x_n) \,|\, x_{t_1}, \dots, x_{t_{m-1}}, x_{n-1}\right) = \mathbf{E}\Big(f(x_{n+1})h(x_n)|x_{n-1}\Big).$$

Proof. Let $\mathcal{G} = \sigma(x_{t_1}, \dots, x_{t_{m-1}}, x_{n-1})$, Since $\mathcal{G} \subset \mathcal{F}_{n-1} \subset \mathcal{F}_n$, we use the tower property to insert a couple of extra conditional expectations:

$$\mathbf{E}\Big(f(x_{n+1})h(x_n) \,|\, \mathcal{G}\Big) = \mathbf{E}\left(\mathbf{E}\left(\overbrace{\mathbf{E}\Big(f(x_{n+1})h(x_n) \,|\, \mathcal{F}_n\Big)} \,|\mathcal{F}_{n-1}\right) |\, \mathcal{G}\right).$$

Since

$$Y := \mathbf{E}\Big(f(x_{n+1})h(x_n)\,|\,\mathcal{F}_n\Big) = \mathbf{E}\Big(f(x_{n+1})\,|\,x_n\Big)h(x_n),$$

by the Markov property and is therefore a function of x_n, we may apply again the Markov property this time conditioning Y on $\mathcal{F}_{n-1}$ and obtain

$$\mathbf{E}\left(\mathbf{E}\left(\mathbf{E}\Big(f(x_{n+1})h(x_n)|\,\mathcal{F}_n\Big)\,|\,x_{n-1}\right)\,|\mathcal{G}\right) = \mathbf{E}\left(\mathbf{E}\Big(f(x_{n+1})h(x_n)|\,\mathcal{F}_n\Big)\,|\,x_{n-1}\right).$$

Finally we use again the tower property to collapse the three conditional expectations to the smallest σ algebra, this completes the proof. □

Take $h \equiv 1$ and f an indicator function we obtain immediately the following:

Corollary 2.1.8 *Let (x_n) be a Markov process, let $t_1 < t_2 < \cdots < t_{m-1} < t_m = n-1$ where $n > 1$ and $t_i \in \mathbf{N} \cup \{0\}$. Let A be a Borel set, then*

$$\mathbf{P}\Big(x_{n+1} \in A\,|\,x_{t_1}, \dots, x_{t_{m-1}}, x_{n-1}\Big) = \mathbf{P}\Big(x_{n+1} \in A\,|\,x_{n-1}\Big).$$

This means that the gap between the future variable and the past need not be 1. Also the same method allow us to work with multi-time points in the future and multiple time points in the past, none needs to consists of consecutive numbers. So by induction we should see a more general statement (see the exercise below).

Exercise 2.1.1 Let x_n be a Markov process. Let $s_1 < s_2 < \cdots < s_m < t_1 < \cdots < t_n$. Let $f_i : \mathcal{X} \to \mathbf{R}$ be bounded Borel measurable, then

$$\mathbf{E}\big(\Pi_{i=1}^n f_i(x_{t_i})\,|\,x_{s_1}, \dots, x_{s_m}\Big) = \mathbf{E}\big(\Pi_{i=1}^n f_i(x_{t_i})\,|\,x_{s_m}\Big).$$

The role played by the future can be exchanged, see the theorem below.

Theorem 2.1.9 *Given a process $\{x_n\}_{n\in\mathbf{N}}$, three indices $\ell < m < n$, the following properties are equivalent:*

(i) For every bounded measurable function f, $\mathbf{E}(f(x_n)\,|\,\sigma(x_\ell) \vee \sigma(x_m)) = \mathbf{E}(f(x_n)\,|\,x_m)$.

(ii) For every bounded measurable function g, $\mathbf{E}(g(x_\ell)\,|\sigma(x_m) \vee \sigma(x_n)) = \mathbf{E}(g(x_\ell)\,|\,x_m)$.

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

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

Proof. By symmetry, it is enough to prove that *(i)* is equivalent to *(iii)*. We start by proving that *(i)* implies *(iii)*.

$$\mathbf{E}(f(x_n)\,g(x_\ell)\,|\,x_m) \overset{tower}{=} \mathbf{E}\left(\mathbf{E}\Big(f(x_n)g(x_\ell)\,|\,\sigma(x_m) \vee \sigma(x_l)\Big)\,|\,x_m\right)$$

$$\overset{\text{taking out known}}{=} \mathbf{E}\left(g(x_\ell)\mathbf{E}\Big(f(x_n)\,|\,\sigma(x_m)\vee\sigma(x_\ell)\Big)\,|\,x_m\right)$$
$$\overset{(i)}{=} \mathbf{E}(g(x_\ell)\mathbf{E}(f(x_n)\,|\,x_m)\,|\,x_m) = \mathbf{E}(g(x_\ell)\,|\,x_m)\,\mathbf{E}(f(x_n)\,|\,x_m)\ ,$$

and so *(iii)* holds. To show the converse, we see both sides of (iii) are $\sigma(x_m)$ measurable and we only need to test them against functions of the form $h(x_m)$.

Given (iii) we show that $\mathbf{E}(f(x_n)\,|\,x_m)$ is the conditional expectation of $f(x_n)$ w.r.t. $\sigma(x_m)\vee\sigma(x_\ell)$. Since it is already measurable w.r.t. $\sigma(x_\ell)\vee\sigma(x_m)$ it is sufficient to test it and $f(x_n)$ w.r.t. $g(x_m)h(x_\ell)$ where $g, h \in \mathcal{B}_b$, showing that

$$\mathbf{E}\left(f(x_n)g(x_\ell)h(x_m)\right) = \mathbf{E}\Big(\mathbf{E}(f(x_n)|x_m)\ g(x_\ell)h(x_m)\Big).$$

(Since last identity implies that $\int_A f(x_n)d\mathbf{P} = \int_A \mathbf{E}(f(x_n)\,|\,x_m)d\mathbf{P}$ for $A = A_1 \cap A_2$ where $A_1 \in \sigma(x_\ell)$ and $A_2 \in \sigma(x_m)$, this proves (i).)

$$\begin{aligned}
&\mathbf{E}\left(f(x_n)g(x_\ell)h(x_m)\right) = \mathbf{E}\Big(\mathbf{E}(f(x_n)g(x_\ell)\ |\,x_m)\,h(x_m)\Big)\\
&\overset{(iii)}{=} \mathbf{E}\left(\mathbf{E}(f(x_n)|x_m)\mathbf{E}\Big(g(x_\ell)\,|\,x_m\Big)\,h(x_m)\right)\\
&=\mathbf{E}\left(\mathbf{E}\Big(g(x_\ell)\mathbf{E}(f(x_n)|x_m)\,h(x_m)\,|\,x_m\Big)\right)\\
&=\mathbf{E}\Big(g(x_\ell)\mathbf{E}(f(x_n)|x_m)\,h(x_m)\Big).
\end{aligned}$$

completing the proof of (iii) to (i). □

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

Remark 2.1.10 Every Markov process satisfies the properties of Theorem 2.1.9. It was however proven in [1] that the converse is not true, *i.e.* there exist processes that satisfy the three (equivalent) properties above but fail to be Markov.

2.2 Markov processes with transition probabilities

Suppose that (x_n) is a Markov process, for each Borel set A and for each n we obtain a family of functions $\mathbf{P}(x_{n+1} \in A|x_n = x)$, these functions are determined only on a set of full measure with respect to $\mathbf{P}_{x_n}$. We now assume the time-homogeneous property: these functions are independent of time. We also assume that we can choose versions of $\mathbf{P}(x_n|x_{n-1} = x)$, denote it by $P(x, A)$ which is independent of n by time homogeneity, in a nice way (the meaning of the nicety is explained below) and denote this function by $P(x, A)$. They indicate the probability to move from x to A in one step.

Definition 2.2.1 We say that $P \equiv \{P(x,A) : x \in \mathcal{X}, A \in \mathcal{B}(\mathcal{X})\}$ is a family of transition probabilities if

- for each $x \in \mathcal{X}$, $P(x,\cdot)$ is a probability measure on $\mathcal{X}$;
- For each $A \in \mathcal{B}(\mathcal{X})$, the function $x \mapsto P(x,A)$ is Borel measurable.

Remark 2.2.2 ** This is equivalent to the statement that there exists a measurable map P from $\mathcal{X}$ into $\mathcal{P}(\mathcal{X})$, the space of probability measures on $\mathcal{X}$, such that

$$\big(P(x)\big)(A) = P(x,A)$$

for every $A \in \mathcal{B}(\mathcal{X})$ and $x \in \mathcal{X}$.

2.2.1 The Chapman-Kolmogorov equation

Let us denote by $\mathcal{B}_b$ bounded Borel measurable real valued functions on $\mathcal{X}$. We first define P^n, then associate with the movement of the Markov chain.

Definition 2.2.3 Given one step probabilities P, denote $P^0(x,\cdot) = \delta_x$, $P^1 = P$, we define recursively the n-step transition probabilities P^n as below. For every $x \in \mathcal{X}$, any $n \geq 1$,

$$P^{n+1}(x,A) = \int_{\mathcal{X}} P(y,A)\, P^n(x,dy) \ , \forall A \in \mathcal{B}(\mathcal{X}) \tag{2.5}$$

Note that $\int_{\mathcal{X}} P(y,A)\, P^0(x,dy) = P(x,A)$ so the above holds for $n = 0$ also.

Equation (2.5) holds for every $A \in \mathcal{B}(\mathcal{X})$ is equivalent to the statement that for any $f \in \mathcal{B}_b$,

$$\int_{\mathcal{X}} f(z) P^{n+1}(x,dz) = \int_{\mathcal{X}} \left(\int_{\mathcal{X}} f(z) P(y,dz) \right) P^n(x,dy). \tag{2.6}$$

Proposition 2.2.4 *For every $n, m \geq 1$,*

$$P^{n+m}(x,A) = \int_{\mathcal{X}} P^n(y,A)\, P^m(x,dy) \tag{2.7}$$

for every $n, m \geq 1$. These are also called the Chapman-Kolmogorov equations.

Observe that for any $f \in \mathcal{B}_b$,

$$\int_{\mathcal{X}} f(z) P^{n+m}(x,dz) = \int_{\mathcal{X}} \int_{\mathcal{X}} f(z) P^n(y,dz)\, P^m(x,dy). \tag{2.8}$$

Proof. This holds for every $n \ge 1$ and for every $m = 0, 1$. We assume that it holds for all n, m such that $k = n + m$. We show (2.7) holds for $k = n + m + 1$. Let $0 \le j < n + m$. We first use the definition and then (2.8),

$$\begin{aligned} P^{n+m+1}(x, A) &= \int_{\mathcal{X}} P(y, A)\, P^{n+m}(x, dy) \\ &= \int_{\mathcal{X}} \left(\int_{\mathcal{X}} P(z, A)\, P^j(y, dz)\right) P^{n+m-j}(x, dy) \\ &= \int_{\mathcal{X}} P^{1+j}(y, dz) P^{n+m-j}(x, dy). \end{aligned}$$

□

2.2.2 Markov chain with transition kernels

Definition 2.2.5 The transition probabilities $P = \{P(x, A) : x \in \mathcal{X}, A \in \mathcal{B}(\mathcal{X})\}$ are the transition probabilities for a *Markov chain* (x_n) if for each $A \in \mathcal{B}(\mathcal{X})$, and each $n \ge 0$,

$$\mathbf{P}(x_{n+1} \in A | x_n) = P(x_n, A), \qquad a.s. \tag{2.9}$$

The Markov chain is then said to be a **time-homogeneous** Markov chain. These transition probabilities are also called the one step probabilities.

Remark 2.2.6 1. Since (x_n) is a Markov chain, (2.9) is equivalent to

$$\mathbf{P}(x_{n+1} \in A | \mathcal{F}_n) = P(x_n, A), \qquad \forall A \in \mathcal{B}(\mathcal{X}).$$

2. This is also equivalent to the statement that for every $f : \mathcal{X} \to \mathbf{R}$ bounded Borel measurable,

$$\mathbf{E}\left(f(x_{n+1}) | \mathcal{F}_n\right) = \int_{\mathcal{X}} f(y)\ P(x_n, dy), \quad a.s.$$

In which case,

$$\mathbf{E}\left(f(x_{n+1}) | x_n\right) = \int_{\mathcal{X}} f(y)\ P(x_n, dy), \quad a.s..$$

Exercise 2.2.1 If Y is an integrable random variable, $\mathcal{F}_1 \subset \mathcal{F}_2$ are sub-algebras, show that if $\mathbf{E}(Y|\mathcal{F}_2)$ is $\mathcal{F}_1$-measurable, then it is $E(Y|\mathcal{F}_1)$.

Remark 2.2.7 There exists a stochastic process (x_n) and transition probabilities P with the relation

$$P(x_{n+1} \in A | x_n) = P(x_n, A),$$

and (x_n) is not a Markov process. This is why we insist on our process is a Markov chain with transition probabilities.

Let (x_n) be a stochastic process satisfying the following: for every triplet of natural numbers i, j, m, there exist numbers p_{ij}^m such that $p_{ij}^m = \mathbf{P}(x_{n+m} = j | x_n = i)$ and for all states i, j and all natural numbers n, m, the Chapman-Kolmogorov relation

$$p_{ij}^{n+m} = \sum_{k=1}^N p_{ik}^n p_{kj}^m,$$

holds. Then (x_n) is not necessarily a Markov process, for an example we refer to a paper by William Feller [1].

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

Theorem 2.2.8 *Let (x_n) be a time-homogeneous Markov process with transition probabilities P. Then, one has for every $n, m \geq 1$,*

(1)

$$\mathbf{P}(x_{n+m} \in A \,|\, x_m) = P^n(x_m, A) \;, \tag{2.10}$$

Also,

$$\mathbf{E}\Big(f(x_{n+m}) | x_m\Big) = \int_{\mathcal{X}} f(z) P^n(x_m, dz).$$

(2) If $x_0 \sim \mu$,

$$\mathbf{P}(x_n \in A) = \int_{\mathcal{X}} P^n(x, A) \, \mu(dx).$$

Proof. (1) The required identity holds for any m and $n = 1$. By induction, we assume one holds for all m and all $n \leq k$. Let $n = k+1$, we begin with inserting conditioning on $\mathcal{F}_m$ and use the Markov property and the induction hypothesis,

$$\begin{aligned}\mathbf{P}(x_{k+m+1} \in A | x_m) &= \mathbf{E}\left(\mathbf{E}\big(\mathbf{1}_{x_{k+m+1} \in A} \,|\, \mathcal{F}_{m+k}\big) | x_m\right) \\ &= \mathbf{E}\left(\mathbf{E}\big(\mathbf{1}_{x_{k+m+1} \in A} \,|\, x_{m+k}\big) | x_m\right) = \mathbf{E}\left(P(x_{m+k}, A) \,| x_m\right) \\ &= \int_{\mathcal{X}} P(z, A) \, P^k(x_m, dz) = P^{k+1}(x_m, A), \quad \forall A \in \mathcal{B}(\mathcal{X}).\end{aligned}$$

We have used induction hypothesis applied first to $f = P(\cdot, A)$.

(2)

$$P(x_n \in A) = \mathbf{E}\left(\mathbf{E}\Big(1_{x_n \in A} | x_0\Big)\right) = \mathbf{E}(P^n(x_0, A)) = \int_{\mathcal{X}} P^n(z, A) \mu(dz).$$

□

[1] William Feller: Non-Markovian processes with the semi-group property. In Ann. Math. Statust. Volum 30, number 4 (1959) pp1252-1253.
https://projecteuclid.org/download/pdf$_1$/euclid.aoms/1177706110

Exercise 2.2.2 If (x_n) is a Markov process with transition probabilities P and initial distribution μ, prove that

$$\mathbf{P}(x_{n+1} \in A, x_n \in B) = \int_{\mathcal{X}} \int_B P(y, A) P^n(z, dy) \mu(dz).$$

Proposition 2.2.9 *If (x_n) is a Markov process with transition probabilities P, then for any $f_i \in \mathcal{B}_b$,*

$$\mathbf{E}\left(\Pi_{i=0}^n f_i(x_i)\right) = \overbrace{\int_{\mathcal{X}} \cdots \int_{\mathcal{X}}}^{n+1} \Pi_{i=0}^n f_i(y_i) \Pi_{i=0}^{n-1} P(y_i, dy_{i+1}) \mu(dy_0). \tag{2.11}$$

Proof. Let us assume that this holds for $k \le n-1$. Then

$$\begin{aligned}\mathbf{E}\left(\Pi_{i=0}^n f_i(x_i)\right) &\overset{tower}{=} \mathbf{E}\left(\mathbf{E}\left(\Pi_{i=0}^n f_i(x_i) | \mathcal{F}_{n-1}\right)\right)\\ &= \mathbf{E}\left(\Pi_{i=0}^{n-1} f_i(x_i) | \mathbf{E}\left(f_n(x_n) \mathcal{F}_{n-1}\right)\right)\\ &\overset{Markov}{=} \mathbf{E}\left(\Pi_{i=0}^{n-1} f_i(x_i) \mathbf{E}\left(f_n(x_n) | x_{n-1}\right)\right)\\ &= \mathbf{E}\left(\Pi_{i=0}^{n-1} f_i(x_i) \int_{\mathcal{X}} f_n(y_n) P(x_{n-1}, dy_n)\right)\end{aligned}$$

The last function involves only $\{x_0, x_1, \dots, x_n\}$ and we can apply the induction hypothesis The rest follows from induction:

$$\begin{aligned}RHS &= \overbrace{\int_{\mathcal{X}} \cdots \int_{\mathcal{X}}}^{n} \left(\Pi_{i=0}^{n-1} f_i(y_i) \int_{\mathcal{X}} f_n(y_n) P(y_{n-1}, dy_n)\right) \Pi_{i=0}^{n-2} P(y_i, dy_{i+1}) \mu(dy_0)\\ &= \overbrace{\int_{\mathcal{X}} \cdots \int_{\mathcal{X}}}^{n+1} \Pi_{i=0}^n f_i(y_i)\, \Pi_{i=0}^{n-1} P(y_i, dy_{i+1}) \mu(dy_0).\end{aligned}$$

We have used Fubini theorem to exchange the order of integrations. □

Remark 2.2.10 If (x_n) is a stochastic process such that (2.11) holds for any $n \ge 0$ and any $f_i \in \mathcal{B}_b$ then (x_n) is a Markov process. Indeed tracing back the steps in the proof, we see $\mathbf{E}(\Pi_{i=0}^n f_i(x_i)) = \mathbf{E}\pi_{i=1}^{n-1} f_i(x_i) \mathbf{E}(f_n(x_n)|x_{n-1})$, then the Markovian property follows from part (iii) of Proposition 2.1.6.

Corollary 2.2.11 *If (x_n) is a Markov chain with transition function P, then for any $n \ge 1$ and for any $A_i \in \mathcal{B}(\mathcal{X})$,*

$$\begin{aligned}&\mathbf{P}(x_0 \in A_0, x_1 \in A_1, \dots, x_n \in A_n)\\ &= \textstyle\int_{A_0} \int_{A_1} \cdots \int_{A_{n-1}} P(x_{n-1}, A_n) P(x_{n-2}, dx_{n-1}) \cdots P(x_1, dx_2) P(x_0, dx_1) \mu(dx_0).\end{aligned} \tag{2.12}$$

We emphasize that if (x_n) is a stochastic process such that (2.12) holds for any $n \ge 1$ and for any $A_i \in \mathcal{B}(\mathcal{X})$, (2.11) holds and (x_n) is a Markov chain (with transition probability P and initial distribution μ_0).

2.2.3 Kolmogorov's extension theorem

We are concerned with the probability distributions of x_n and the joint probability distributions of the first n variables $(x_1, \dots, x_n)$ which takes its value in the product space $\mathcal{X}^n$. The latter are *finite dimensional distributions.*

Set $\mathcal{X} = \Pi_{i=1}^\infty \mathcal{X}_i$, it is the space of sequences $(a_1, a_2, \dots,)$. Denote by $\pi_j : \Pi_{i=1}^\infty \mathcal{X}_i \to X_j$ the projection to the jth component, so $\pi_j(a_1, a_2, \dots) = a_j$. Recall the product σ-algebra is the σ-algebra generated by the pre-images $\pi_j^{-1}(A_j)$, $A_j \in \mathcal{B}_j$. This is the smallest σ-algebra such that each projection π_j is measurable. Moreover,

$$\otimes_{i=1}^n \mathcal{B}(X_i) = \sigma(\{A_1 \times \cdots \times A_n, A_i \in \mathcal{B}_i, i = 1, \dots, n\}).$$

Sets of the form $\Pi_{i=1}^\infty A_i$, with $A_i \in \mathcal{B}_i$ and with only a finite number of A_i's not $\mathcal{X}_i$, are called *cylindrical sets.* The tensor σ-algebra $\otimes_{i=1}^\infty \mathcal{B}_i$ is generated by cylindrical sets.

If μ_i are finite measures on $\mathcal{B}(\mathcal{X}_i)$, we may define a product measure on the tensor Borel σ-algebra as follows:

$$\mu_1 \otimes \cdots \otimes \mu_n(A_1 \times \cdots \times A_n) = \Pi_{i=1}^n \mu_i(A_i)$$

for any $A_i \in \mathcal{B}(\mathcal{X}_i)$.

Example 2.2.1 Suppose that for $i = 0, 1, \dots, n$, $x_i : \Omega \to \mathcal{X}_i$ are random variables. Then they push the measure $\mathbf{P}$ forward to give a measure on $\otimes_{i=0}^\infty \mathcal{B}_i$, any $n \in \mathbf{N}$. Indeed, for any $A_i \in \mathcal{B}(\mathcal{X}_i)$,

$$\mu_n(A_0 \times A_1 \times \cdots \times A_n) = \mathbf{P}(x_0 \in A_0, x_1 \in A_1, \dots, x_n \in A_n).$$

Note that

$$\mathbf{P}(x_0 \in A_0, x_1 \in A_1, \dots, x_n \in A_n, x_{n+1} \in \mathcal{X}) = \mathbf{P}(x_0 \in A_0, x_1 \in A_1, \dots, x_n \in A_n).$$

So $\mu_{n+1}(A_0 \times A_1 \times \cdots \times A_n \times \mathcal{X}) = \mu_n(A_0 \times A_1 \times \cdots \times A_n)$, μ_n are consistent with each other in this sense.

Definition 2.2.12 If $(x_n)_{n=0}^\infty$ is a stochastic process on $\mathcal{X}$, then $\{\mu_n\}_{n=0}^\infty$ defined above are called the finite dimensional distributions.

We can consider $(\Pi_{i=0}^n \mathcal{X}_i, \mathcal{B}(\mathcal{X}_i), \mu_n)$ as a probability space. Let $\pi_i : \Pi_{i=0}^n \mathcal{X}_i \to \mathcal{X}_i$ denote the projections, measurable maps, then for any $A_i \in \mathcal{B}(\mathcal{X}_i)$, $(\pi_0 \in A_0, \dots, \pi_{n+1} \in A_{n+1}) = \Pi_{i=0}^n A_i$. Thus,

$$\mu_n \left(\pi_0 \in A_0, \pi_1 \in A_1, \dots, \pi_n \in A_n\right) = \mu_n(\Pi_{i=0}^n A_i) = \mathbf{P}(x_0 \in A_0, x_1 \in A_1, \dots, x_n \in A_n).$$

The same story holds on the infinite produce spaces, which is the natural space for a stochastic process.

Theorem 2.2.13 (Kolmogorov's extension theorem) *Let $\{\mu_n\}$ be a sequence of probability measures on $\mathcal{X}^n$ such that*

$$\mu_n(A_1 \times A_2 \times \ldots \times A_n) = \mu_{n+1}(A_1 \times A_2 \times \ldots \times A_n \times \mathcal{X}) \tag{2.13}$$

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

A family of measures $\{\mu_n\}$ satisfying (2.13) is called a consistent family of probability measures. As a consequence, for every such consistent family of probability measure, there exists a stochastic process (x_n) having them as its finite dimensional distributions.

Corollary 2.2.14 *Given a stochastic process (x_n) on $\mathcal{X}$, their finite dimensional distributions determine a unique probability measure on $\otimes_{i=1}^\infty \mathcal{B}(\mathcal{X})$.*

2.2.4 The canonical picture and the existence of Markov Chains

Proposition 2.2.15 *Given a family of transition probabilities P on $\mathcal{X}$ and a probability measure μ_0 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

$$\begin{aligned}&\mu_n(A_0 \times \ldots \times A_n) = \\ &\quad \int_{A_0}\int_{A_1}\int_{A_2}\cdots\int_{A_{n-2}}\int_{A_{n-1}} P(y_{n-1}, A_n)P(y_{n-2}, dy_{n-1})\cdots P(y_1, dy_2)P(y_0, dy_1)\mu(dy_0)\ .\end{aligned}$$

With equation (2.12) it is easy to check that this sequence of measures satisfies the consistence condition in Kolmogorov's extension theorem, by this theorem we conclude that there exists a unique measure $\mathbf{P}_\mu$ on $\mathcal{X}^\infty$ such that the restriction of $\mathbf{P}_\mu$ to $\mathcal{X}^n$ is given by μ_n. (The subscript μ indicates the initial distribution).

We now choose $\Omega = \mathcal{X}^\infty$ as our probability space equipped with the probability measure $\mathbf{P}_\mu$. Then for (π_n) the canonical process, *i.e.* $\pi_n((w_0, w_1, \ldots)) = w_n$,

$$\begin{aligned}\mathbf{P}_\mu(\pi_0 \in A_0, \ldots, \pi_{n+1} \in A_{n+1}) &= \mathbf{P}_\mu(A_0 \times \cdots \times A_{n+1}) \\ &= \textstyle\int_{A_0}\int_{A_1}\cdots\int_{A_{n-1}} P(y_{n-1}, A_n)\Pi_{i=0}^{n-1} P(y_i, dy_{i+1})\, \mu(dy_0).\end{aligned}$$

This means that (π_n) is a Markov process with the required transition probabilities and initial distribution μ_0 (see the remark after Corollary 2.2.11). 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 $(\pi_0, \ldots, \pi_n)$ must be equal to μ_{n+1} for every n. □

Remember that $\mathbf{P}_\mu$ restricts to $\mathcal{X}^n$ is the finite dimensional distribution of (π_n),

$$\mathbf{P}_\mu(\pi_0 \in A_0, \dots, \pi_{n+1} \in A_{n+1}) = \int_{\cap_{i=0}^n A_i} P(\pi_n, A_{n+1}) d\mathbf{P}_\mu.$$

It is traditional to denote by $\mathbf{P}_x$ the probability measure on the canonical space $\mathcal{X}^\infty$ induced by the Markov process with transition probabilities P and initial distribution δ_x. That induced by the Markov process with initial distribution μ is denoted by $\mathbf{P}_\mu$. Then on the canonical probability space we use $\mathbf{E}_x$ and $\mathbf{E}_\mu$ to denote taking expectations w.r.t. $\mathbf{P}_x$ and $\mathbf{P}_\mu$ respectively.

Example 2.2.2 Let x_n be a Markov chain with transition probability P, then

$$\begin{aligned}\mathbf{P}_\mu(x_1 \in B) =& \mathbf{E}(P(x_1 \in B|x_0)) = \mathbf{E}P(x_0, B) = \int_{\mathcal{X}} P(y, B)\mu(dy)\\ \mathbf{P}_a(x_1 \in B) =& \mathbf{E}(P(x_1 \in B|x_0)) = \mathbf{E}[P(x_0, B)] = \int P(y, B)\delta_a(dy) = P(a, B).\end{aligned}$$

Definition 2.2.16 From the proof, the projection maps $\pi_n : \mathcal{X}^\infty \to \mathcal{X}$ is a Markov process on $(\mathcal{X}^\infty, \otimes\mathcal{B}(CX)^\infty, \mathbf{P}_\mu)$ with state space $\mathcal{X}$, transition probabilities P and initial distribution μ. This process is called the canonical process.

Remark 2.2.17 We fix the transition probabilities P, and then for every $x \in \mathcal{X}$ we have a Markov process with the initial distribution x. We emphasise that we have a family of Markov process and we can start from everywhere a Markov process with the given transition probability.

2.3 Examples and exercises

Example 2.3.1 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 2.3.2 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 2.3.1 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?

Example 2.3.3 Let $M_n = \max(x_0, x_1, \ldots, x_n)$, where (x_n) is a simple random walk starting from 0, $x_n = \sum_{i=0}^n \xi_i$ where $\xi_0 = 0$ and $\xi_i, i \geq 1$ are iid Bernoulli random variables: $\mathbf{P}(\xi_i = \pm 1) = \frac{1}{2}$. There are three paths with $M_3 = 1$, which has probability $\frac{1}{8}$ being taken:

$$\sigma_1 : x_0 = 0, x_1 = 1, x_2 = 0, x_3 = 1,$$

$$\sigma_2 : x_0 = 0, x_1 = 1, x_2 = 0, x_3 = -1,$$

$$\sigma_3 : x_0 = 0, x_1 = -1, x_2 = 0, x_3 = 1.$$

For the path σ_1, σ_3, $M_4 = 2$ with probability $\frac{1}{2}$. For the path σ_2, $M_4 = 2$ with probability 0. The probability of $M_4 = 2$ depends not just on M_3 it depends on the actual path. (M_n) is not a Markov process. (This can be computed also with elementary probability, since $M_4 = 2$ whether the walk goes up and comes down. Also $\mathbf{P}(M_4 = 2 | M_3 = 1) = \frac{1}{3}$. This can be computed using the space of the path of uniform probability as probability space, counting paths or using elementary conditional expectations.

2.4 Stopping times

As usual, we have a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ and $\mathcal{X}$ a complete separable metric space. When we discuss a stochastic process (x_n) (with state space $\mathcal{X}$), we usually take the filtration to be $\mathcal{F}_n = \sigma\{x_0, \ldots, x_n\}$.

Notation. $a \wedge b = \min(a, b)$, $a \vee b = \max\{a, b\}$.

Definition 2.4.1 An integer-valued random variable T is called a $\mathcal{F}_n$**-stopping time**, if the event $\{T = n\}$ is $\mathcal{F}_n$-measurable for every $n \geq 0$. (The value $T = \infty$ is usually allowed as well and no condition is imposed on its measurability.)

Exercise 2.4.1 Show that T is an $\mathcal{F}_n$-stopping time if and only if $\{T \leq n\} \in \mathcal{F}_n$.

Example 2.4.1 Let $A \in \mathcal{B}(\mathcal{X})$, $\tau_A = \inf\{n \geq 1 : x_n \in A\}$, and $\sigma_A = \inf\{n \geq 0 : x_n \in A\}$. Then both are stopping times. Proof: $\{\tau_A = 0\} = \phi \in \mathcal{F}_0$. For $n \geq 1$.

$$\{\tau_A = n\} = \cap_{i=1}^{n-1}\{x_i \in A^c\} \cap \{x_n \in A\} \in \mathcal{F}_n,$$

This concludes that τ_A is a stoping time.

$\{\sigma_A = 0\} = \{x_0 \in A\} \in \mathcal{F}_0$. For $n \geq 1$.

$$\{\tau_A = n\} = \cap_{i=0}^{n-1}\{x_i \in A^c\} \cap \{x_n \in A\} \in \mathcal{F}_n,$$

concluding that σ_A is a stoping time.

Given a stopping time T and a Markov process x we introduce the stopped process, which is denoted by (x_n^T) or by $(x_{T\wedge n})$:

$$x_n^T(\omega) \equiv x_{n\wedge T}(\omega) = \begin{cases} x_n(\omega) & \text{if } n \leq T(\omega), \\ x_{T(\omega)}(\omega) & \text{otherwise.} \end{cases}$$

Exercise 2.4.2 Let us consider the simple random walk $x_n = \sum_{i=1}^n \xi_i$ on $\mathbf{Z}$, where $\{\xi_i\}$ are i.i.d's such that

$$\mathbf{P}(\xi_i = 1) = \frac{1}{2}, \quad \mathbf{P}(\xi_i = -1) = \frac{1}{2}.$$

Let τ be the first time after $n = 1$ that $x_n = 2$. If ω is a sample such that $(x_0(\omega) = 0, x_1(\omega) = 1, x_2(\omega) = 2, x_3(\omega) = 1, x_4(\omega) = 0, \dots$. Write out the entire sequence $(x_{T\wedge n}(\omega), n \geq 0)$.

Set

$$\mathcal{F}_\infty = \sigma(x_\cdot) = \vee_{n=0}^\infty \sigma(x_n).$$

Definition 2.4.2 If T is a finite $\mathcal{F}_n$-stopping time, we define the associate σ-algebra to be

$$\mathcal{F}_T = \{A \in \mathcal{F}_\infty : A \cap \{T = n\} \in \mathcal{F}_n, \forall n \in \mathbf{N}\}.$$

If $T \equiv n$ is a constant time, then $\mathcal{F}_T$ agrees with $\mathcal{F}_n$.

Example 2.4.2 T is $\mathcal{F}_T$ measurable.

For any integers $m, n \geq 0$, $\{T = m\} \cap \{T = n\}$ is either an empty set in case $m \neq n$ or is $\{T = n\} \in \mathcal{F}_n$ incase $m = n$. Also, This shows for any m, the pre-image $\{T = m\}$ is in $\mathcal{F}_T$.

Exercise 2.4.3 If $T < \infty$, show the following random variables are $\mathcal{F}_T$-measurable: x_T and $x_{T\wedge m}$ for any $m \in \mathbf{N}$.

Lemma 2.4.3 *If T is a finite $\mathcal{F}_n = \vee_{i=0}^n \sigma(x_i)$-stopping time, then for every $n \geq 0$,*

$$\{T = n\} \in \sigma(x_{T\wedge 0}, \dots, x_{T\wedge n}).$$

Proof. This is left as exercise. □

Proposition 2.4.4 *Let*

$$\sigma(x_\cdot^T) = \sigma(x_{T\wedge n}, n \geq 0) = \vee_{n=0}^\infty \sigma(x_{n\wedge T}).$$

If T is a finite $\mathcal{F}_n$-stopping time, then $\mathcal{F}_T$ is the σ-algebra generated by the collection $\{x_{n\wedge T}\}_{n\geq 0}$:

$$\mathcal{F}_T = \sigma(x_\cdot^T).$$

Proof. exercise □

2.4.1 Examples and exercises

Example 2.4.3 The following time is, in general, **not** a stopping time:

$$T = \inf\{n \geq 0 : n \text{ is the last time that } x_n = 1\}.$$

Example 2.4.4 (a) A constant time is a stopping time.

(b) $T(\omega) \equiv \infty$ is also a stopping time.

Exercise 2.4.4 Let S, T, T_n be stopping times.

(1) Then $S \vee T = \max(S,T)$, $S \wedge T = \min(S,T)$ are stopping times.

(2) $\limsup_{n\to\infty} T_n$ and $\liminf_{n\to\infty} T_n$ are stopping times.

Exercise 2.4.5 Let $x_n = \sum_{i=1}^n \xi_i$ be as in Example 2.4.2. Let $x_0 = 0$, T the first time $x_n = 1$. Is the event $\{x_n = 2\}$ in $\mathcal{F}_T$?

Exercise 2.4.6 Let S, T be stopping times.

(1) If $S \leq T$ then $\mathcal{F}_S \subset \mathcal{F}_T$.

(2) Let $S \leq T$ and $A \in \mathcal{F}_S$. Then $S\mathbf{1}_A + T\mathbf{1}_{A^c}$ is a stopping time.

(3) S is $\mathcal{F}_S$ measurable.

(4) $\mathcal{F}_S \cap \{S \leq T\} \subset \mathcal{F}_{S\wedge T}$.

2.5 Strong Markov Property

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". This is stated more precisely in the theorem below. Recall the shift operator

$$(\theta_T x.)_n = x_{T+n}$$

This means for $\omega \in \Omega$ and $n \geq 0$, $(\theta_T x.)_n$ is a random variable given by $(\theta_T x.)_n(\omega) = x_{T(\omega)+n}(\omega)$. The shift Markov process starts from x_T. Observe that x_{T+n} is measurable with respect to $\mathcal{F}_{T+n}$.

Remark 2.5.1 An element of $\mathcal{X}^{\mathbf{N}}$ is a sequence $(\sigma_0, \sigma_1, \dots)$ with $\sigma_i \in \mathcal{X}$. Let us consider Borel measurable function on the path space $\mathcal{X}^{\mathbf{N}}$ with the product σ-algebra. What is a measurable set in this σ-algebra? They are generated by cylindrical sets of the form $\{\sigma_{n_1} \in A_1, \dots, \sigma_{n_m} \in A_m\}$ where $n_1 < n_2 < \dots < n_m$ is a set of times, and A_i are measurable sets from $\mathcal{X}$. The collections of such cylindrical sets is a π-system. A property on measurable functions on $\mathcal{X}$ are typically determined by that of the functions of the following form (called cylindrical functions): for $n_1 < n_2 < \dots < n_m$ and $f : \mathcal{X}^m \to \mathbf{R}$ Borel measurable,

$$\Phi(\sigma) = f(\sigma_{n_1}, \dots, \sigma_{n_m}).$$

Notation. If (x_n) is a stochastic process, for each ω, we have a sequence $(x_0(\omega), x_1(\omega), \dots)$ in $\mathcal{X}^{\mathbf{N}}$. This is also denoted by $x.(\omega)$. The process is a map from Ω to $\mathcal{X}$, thus can be denoted as $x.$, where the script dot means we think $x.(\omega)$ as an element of $\mathcal{X}^{\mathbf{N}}$.

Definition 2.5.2 A time-homogeneous Markov process (x_n) with transition probabilities P is said to have the strong Markov property if for every finite stopping time T and for every bounded Borel measurable function $\Phi : \mathcal{X}^{\mathbf{N}} \to \mathbf{R}$, the following holds:

$$\mathbf{E}\big(\Phi(\theta_T x.) \,|\, \mathcal{F}_T\big) = \mathbf{E}\big(\Phi(\theta_T x.) \,|\, \sigma(X_T)\big) \,, \tag{2.14}$$

Remark 2.5.3 (1) For simplicity let us define $y_n = x_{T+n}$. Set $\mathcal{G}_n = \mathcal{F}_{T+n}$. If (2.14) holds for every finite stopping time T for x_n then for every bounded Borel measurable function $f : \mathcal{X} \to \mathbf{R}$, the following holds

$$\begin{aligned}\mathbf{E}(f(y_{n+m})|\mathcal{G}_m) &= \mathbf{E}(f(x_{T+n+m})|\mathcal{F}_{T+m}) \overset{(2.14)}{=} \mathbf{E}(f(x_{T+n+m})|x_{T+m}) \\ &= \mathbf{E}(f(y_{n+m})|y_m),\end{aligned}.$$

(2) Note that if the following holds for any measurable subset A of $\mathcal{X}$ and any time n:

$$\mathbf{P}(x_{n+T} \in A|\mathcal{F}_T) = \mathbf{P}^n(x_T, A). \tag{2.15}$$

then for every bounded Borel measurable function f,

$$\mathbf{E}(f(x_{n+T})|\mathcal{F}_T) = \int f(y)\mathbf{P}^n(x_T, dy). \tag{2.16}$$

Exercise 2.5.1 If $\mathbf{P}(x_{n+T} \in A|\mathcal{F}_T) = \mathbf{P}^n(x_T, A)$ holds for every measurable set A, then

$$\mathbf{P}(x_{n_1+T} \in A_1, \dots x_{n_m+T} \in A_m|\mathcal{F}_T) = \mathbf{P}(x_{n_1+T} \in A_1, \dots x_{n_m+T} \in A_m|x_T), \tag{2.17}$$

or equivalently, for bounded measurable functions f_i,

$$\mathbf{E}(\Pi_{i=1}^m f_j(x_{n_j+T})|\mathcal{F}_T) = \mathbf{E}(\Pi_{i=1}^m f_j(x_{n_j+T})|x_T). \tag{2.18}$$

2.5.1 Strong Markov property at a finite stopping time

We first consider a stopping that that is almost surely finite.

Theorem 2.5.4 (Strong Markov property) *Let (x_n) be a time-homogeneous Markov process with transition probabilities P and if T is a stopping time which is almost-surely finite then the process $(\theta_T x)_n$ is also Markov with transition probabilities P which means*

$$\mathbf{P}(x_{n+T} \in A|\mathcal{F}_T) = P^n(x_T, A)$$

for any $n > 0$ and any $A \in \mathcal{B}(\mathcal{X})$. It follows that x has the strong Markov property.

Proof. Since $T < \infty$ a.s., $\Omega = \cup_{n=0}^{\infty}\{T = n\} \cup C$ where $C = \{T = \infty\}$ has measure zero. For any f bounded measurable from $\mathcal{X}$ to $\mathbf{R}$,

$$\begin{aligned}
\int_B f(x_{n+T})\,d\mathbf{P} &= \sum_{m=0}^{\infty} \int_{B\cap\{T=m\}} f(x_{n+m})\,d\mathbf{P} \\
&= \sum_{m=0}^{\infty} \int_{B\cap\{T=m\}} \mathbf{E}\left(f(x_{n+m})\,|\,\mathcal{F}_m\right)\,d\mathbf{P} \\
&= \sum_{m=0}^{\infty} \int_{B\cap\{T=m\}} \int f(y)P^n(x_m, dy)\,d\mathbf{P} = \sum_{m=0}^{\infty} \int_{B\cap\{T=m\}} \int_{\mathcal{X}} f(y)P^n(x_T, dy)\,d\mathbf{P} \\
&= \int_B \int_{\mathcal{X}} f(y)P^n(x_T, dy)\,d\mathbf{P}.
\end{aligned}$$

In the second line we have used the fact that $B \cap \{T = m\} \in \mathcal{F}_m$. This shows that

$$\mathbf{E}\left(f(x_{n+T})\,|\mathcal{F}_T\right) = \int_{\mathcal{X}} f(y)P^n(x_T, dy).$$

So indeed, P is the transition probability for x_{n+T}.

Also, if $\Phi : \mathcal{X}^{\mathbf{N}} \to \mathbf{R}$ is bounded Borel measurable, then

$$\mathbf{E}\big(\Phi(\theta_T x.) \,|\, \mathcal{F}_T\big) = \mathbf{E}\big(\Phi(\theta_T x.) \,|\, \sigma(X_T)\big) \ .$$

To show this it is sufficient to prove it for 'cylindrical functions', i.e. functions of the form $\Phi((a_1, a_2, \dots, a_n, \cdot)) = \Pi_{i=1}^k f_i(a_i)$. For this we replace $f(x_{n+T})$ in the above by $\Pi_{i=1}^k f_i(x_{i+T})$ and the conclusion follows in the same way. Thus $x.$ has the strong Markov property. □

2.5.2 Markov property at non-finite stopping times *

Lemma 2.5.5 *If A is any subset of Ω and $\mathcal{F}$ a σ-algebra then $\mathcal{F} \cap A = \{B \cap A : B \in \mathcal{F}\}$ is a σ-algebra on A. This is called the trace σ-algebra.*

Going over the proof for the strong Markov property, we observe that we used the assumption that T is finite in two ways: (1) $\cup_{n=0}^\infty \{T = n\} = \Omega$, (ii) x_T can be defined. This proof can be modified to yield a corresponding result for stopping times that is not necessarily finite. In this case $\cup_{n=0}^\infty \{T = n\} = \{T < \infty\}$ and so we have to limit ourselves on this set. If we do restrict to the set $\{T < \infty\}$, then x_T is again defined.

Let T be a stopping time. Then $\{T < \infty\}$ is a subset of $\mathcal{F}_T$, we may condition on $\mathcal{F}_T \cap \{T < \infty\}$. Now we can state the modified theorem:

Theorem 2.5.6 *Let (x_n) be a time-homogeneous Markov process with transition probabilities P and let T be a stopping time. Let $\Phi : \mathcal{X}^\infty \to \mathbf{R}$ be a function. Then on the set $\{T < \infty\}$,*

$$\mathbf{E}\left(\Phi(\theta_T x.) \,|\, \mathcal{F}_T\right) = \mathbf{E}_{x_T} \Phi(\theta_T x.).$$

Proof. ** We demonstrate the proof for Φ depending only one coordinate, the proof for Φ depending on a finite number of coordinates is the same. Let $f : \mathcal{X} \to \mathbf{R}$ be bounded measurable, and $B \in \mathcal{F}_T$,

$$\begin{aligned}
\int_{B\cap\{T<\infty\}} f(x_{n+T}) d\mathbf{P} &= \sum_{m=0}^\infty \int_{B\cap\{T=m\}} f(x_{n+T})\, d\mathbf{P} = \sum_{m=0}^\infty \int_{B\cap\{T=m\}} f(x_{n+m})\, d\mathbf{P} \\
&= \sum_{m=0}^\infty \int_{B\cap\{T=m\}} \mathbf{E}\left(f(x_{n+m}) \,|\, \mathcal{F}_m\right)\, d\mathbf{P} = \sum_{m=0}^\infty \int_{B\cap\{T=m\}} \mathbf{E}\left(f(x_{n+m}) \,|\, x_m\right)\, d\mathbf{P} \\
&= \sum_{m=0}^\infty \int_{B\cap\{T=m\}} \mathbf{E}\left(f(x_{n+T}) \,|\, x_T\right)\, d\mathbf{P} = \int_{B\cap\{T<\infty\}} \mathbf{E}\left(f(x_{n+T}) \mathbf{1}_{\{T<\infty\}} \,|\, x_T\right)\, d\mathbf{P}.
\end{aligned}$$

We may conclude that on the set $\mathbf{1}_{\{T<\infty\}}$

$$\mathbf{E}(f(x_{n+T})\mathbf{1}_{\{T<\infty\}} | \mathcal{F}_T) = \mathbf{E}\left(f(x_{n+T})\mathbf{1}_{\{T<\infty\}} \,|\, x_T\right).$$

We can also interpret this as

$$\mathbf{1}_{\{T<\infty\}}\mathbf{E}(f(x_{n+T})|\mathcal{F}_T \cap \{T<\infty\}) = \mathbf{1}_{\{T<\infty\}}\mathbf{E}\left(f(x_{n+T})\,|\,x_T\right),$$

concluding the proof. □

2.5.3 Transition operators and invariant measures

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, (x_n) a time-homogenous Markov process with transition probabilities P. Then $P(x_n \in A|x_0 = x) = P^n(x, A)$. If the chain has initial distribution μ, the distribution of x_n is denoted by $P_\mu(x_n \in A)$. Then

$$P_\mu(x_n \in A) = \int_{\mathcal{X}} P^n(x, A)\mu(dx).$$

If x_0 is distributed as δ_a, we denote the probability x_n in A by $P_x(x_n \in A)$. Then

$$P_x(x_n \in A) = \int_{\mathcal{X}} P^n(x, A)\delta_a(dx) = P^n(a, A) = P(x_n \in A|x_0 = x).$$

The subscript plays the role of noting the initial distribution and agrees with our notation for the canonical sequence space picture.

Given a transition probability P we transfer a measure μ to the probability distribution of x_1,

$$\mu \to \int_{\mathcal{X}} P(x, \cdot)\mu(dx).$$

the same mechanics then send this to the distribution of x_2.

Let $\mathcal{P}(\mathcal{X})$ denote the space of probability measures on $\mathcal{X}$.

Definition 2.5.7 Given transition probabilities P, we define a **transition operator** T^* on $\mathcal{P}(\mathcal{X})$, which will be denoted by T if there is no risk of confusion, by

$$(T\mu)(A) = \int_{\mathcal{X}} P(x, A)\,\mu(dx)\,. \tag{2.19}$$

Note that T can be extended to the space of all finite signed measures by linearity. Note if $f : \mathcal{X} \to \mathbf{R}$ is bounded measurable,

$$\int_{\mathcal{X}} d(T\mu) = \int_{\mathcal{X}}\int_{\mathcal{X}} f(y)P(x, dy)dx.$$

Definition 2.5.8 A measure such that $T\mu = \mu$ is called an invariant measure for P (or for the time homogeneous Markov chain).

We are most interested in finite invariant measures, which can then be normalised to a probability measure.

Exercise 2.5.2 Check that the operator T^n obtained by replacing P by P^n in (2.19) is equal to the operator obtained by applying T n times, $T^n = T \circ T \circ \ldots \circ T$.

Remark 2.5.9 ** 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 (2.19) for some P. This conclusion holds under the assumptions that $\mathcal{X}$ is a complete separable metric space and T is continuous in the weak topology. This can be proved using 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 weights).

We similarly define an operator $T_\star : \mathcal{B}_b(\mathcal{X}) \to \mathcal{B}_b(\mathcal{X})$, the space of bounded measurable functions from $\mathcal{X}$ to $\mathbf{R}$, by

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

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

Exercise 2.5.3 Check that the operators T and $T_\star$ are each other's dual, *i.e.* that

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

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

Exercise 2.5.4 Show that $T_* 1 = 1$, $T_* f \geq 0$ if $f \geq 0$.

2.5.4 A remark on continuous time Markov processes

A Brownian motion is a strong Markov process (with continuous trajectories) whose probability distribution is determined by the heat equation:

$$\frac{\partial u_t}{\partial t} = \frac{1}{2}\Delta u_t.$$

If $P_t(x,\cdot)$ are its transition probabilities then $u_t(x) = \int_{\mathcal{X}} u_0(y) P_t(x, dy)$.

Solutions of SDEs of the form

$$dx_t = \sigma(x_t) dB_t + \sigma_0(x_t) dt$$

are Markov processes. So are solutions of many SPDE's. Markov processes can be used to model problems in engineering, biology, medicine, stochastic filtering, etc.

Given a time continuous Markov process, we can extract a Markov chain by evaluations at a discrete set of times t_i: $\{x_{t_1}, x_{t_2}, \dots, \}$. It also happens a Markov chain can sometimes be extracted from a stochastic processes that itself is not a Markov process.

Chapter 3

Finite State and countable state Markov Chains

In this chapter, we assume that the space $\mathcal{X}$ is finite, in which case we identify it with $\{1, \ldots, N\}$ for some $N > 0$, or countable. In the finite 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 $T : \mathcal{P}(\mathcal{X}) \to \mathcal{P}(\mathcal{X})$ 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 $P = (P_{ij})$. If μ is the initial distribution for the chain, the probability distribution of (x_n) is given by $P^n \mu$, where P^n denotes matrix multiplication. For example,

$$P_\mu(x_n = i) = \sum_{j=1}^{n} P^n_{ij} \mathbf{P}(x_0 = j) = \sum_{j=1}^{n} P^n_{ij} \mu(j).$$

Taking $n = 1$, $\mu = \delta_j$, then we see

$$P_{ij} = \mathbf{P}(x_1 = i \,|\, x_0 = j). \tag{3.1}$$

The number P_{ij} should be interpreted as the probability of jumping from state j to state i. If x_1 is a probability measure, then

$$\sum_{i=1}^{N} P_{ij} = 1 \;, \quad \text{for all } j. \tag{3.2}$$

If $\mathcal{X}$ is a countable space $\mathcal{X} = \{1, 2, \ldots, \}$, we have a matrix with infinite entries, and $\sum_{i=1}^{\infty} P_{ij} = 1$.

Definition 3.0.1 We call a matrix P with positive entries which satisfies (3.2) a **stochastic matrix**.

Remark 3.0.2 On the other hand, a time homogeneous Markov chain is given by transition probabilities, which in our case can be represented by a matrix, the transition matrix $p = (p(i,j))$, this is traditionally written as: $p(i,j) = \mathbf{P}(x_1 = j \,|\, x_0 = i)$. A transition matrix p has non-negative entries, with each row sum up to 1: $\sum_{j=1}^n p(i,j) = 1$. It is clear that the stochastic matrix we defined earlier and the transition matrix here are transpose of each other. $P = p^T$. Thus $P^n = (p^n)^T$. Also if π is an invariant measure if $P\pi = \pi$, or equivalently if $\pi^T p = \pi^T$. We can therefore use P or p depending on which one is more convenient to use.

Convention. To conform with standard notations in linear algebra and the notation for transition operators, we use the notation P, and refer to it a stochastic matrix.

We recall the definition of a time homogeneous Markov chain on a countable state space by elementary conditional probabilities.

Exercise 3.0.1 Let P be a stochastic matrix. Define $a_n^j := \min_{1\le i\le N}\{P_{ji}^n\}$ and $b_n^j := \max_{1\le i\le N}\{P_{ji}^n\}$. Show that for every j, $\{a_n^j\}$ increases with n, and $\{b_n^j\}$ decreases with n.

Definition 3.0.3 If P is a stochastic matrix, we call an eigenvector corresponding to the eigenvalue 1, normalised so the entries sum up to one, the Perron-Frobenius vector for P.

We will see later a stochastic matrix P always has an eigenvalue 1, whose eigen-space is one dimensional and we can always select an eigenvector with positive entries, and normalised to sum to one.

3.1 Classification of Chains

In this section we consider $\mathcal{X} = \{1, \dots, N\}$, unless otherwise stated. Fix an arbitrary stochastic matrix P of dimension N we can associate to such a matrix P_{ij} an oriented graph, called the **incidence graph** of P by taking $\mathcal{X} = \{1, \dots, N\}$ 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

$$P = \frac{1}{10}\begin{pmatrix} 0 & 3 & 0 & 2 \\ 5 & 7 & 10 & 8 \\ 5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \tag{3.3}$$

then the associated graph is given by the one in Figure 3.1. 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.

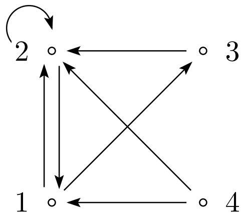


Figure 3.1: Graph for P.

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**. This is to make sure that the chain is really one single chain. This definition applies to chains on a countable state space., in which case the stochastic matrix has infinite number of rows and columns.

Definition 3.1.1 We say that j is accessible from i, if $P_{ji}^n > 0$ for some n. A stochastic matrix is irreducible if for any i, j there exists $n, m \geq 0$ such that $P_{ij}^n > 0$ and $P_{ji}^m > 0$. (These applies also to countable state space)

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 in such a way that if one starts the process in some minimal sets 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 (3.3) is reducible because it is impossible to reach 4 from any of the other points in the system.

A general stochastic matrix is not irreducible. It can however be broken up into irreducible components in the following way. The set $\{1, \ldots, N\}$ is 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, with the convention that P^0 is the identity matrix, we make the following defintion.

Definition 3.1.2 Let $\mathcal{X}$ be a countable space. We say two states are said to communicate with each other, if there exist $m, n \geq 0$ such that $(P^m)_{ij} > 0$ and $(P^n)_{ji} > 0$. This is denoted by $i \sim j$. We denote by $[i]$ the equivalence class of i under this relation and we call it the **communication class** of i. A chain is said to be irreducible if there exists only one communication class.

For example, in the case of (3.3), we have $[1] = \{1, 2, 3\}$ and $[4] = \{4\}$.

Lemma 3.1.3 *If j is accessible from i, then any $j' \in [j]$ is accessible from any $i' \in [i]$.*

Proof. Suppose that $P_{ji}^m > 0$. We want to show for $j' \in [j]$ and $i' \in [i]$, there exists n with $P_{j'i'}^n > 0$. Let $n_1, n_2 \geq 0$ such that $P_{j'j}^{n_1} > 0$ and such that $P_{ii'}^{n_2} > 0$. Then by Chapman

Kolmogorov equation,

$$P^{m+n_1+n_2}_{j'i'} \geq P^{n_1}_{j'j} P^m_{ji} P^{n_2}_{ii'} > 0.$$

Indeed,

$$P^{m+n_2}_{ji'} = \sum_{k\in\mathcal{X}} P^{n_2}_{jk} P^m_{ki'} \geq P^m_{ji} P^{n_2}_{ii'} > 0.$$

Similarly, $P^{m+n_1+n_2}_{j'i'} \geq P^{n_1}_{j'j} \left(P^m_{ji} P^{n_2}_{ii'} \right) > 0.$ □

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 $[j] \leq [i]$.

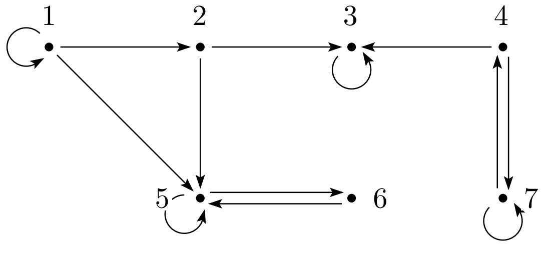
Exercise 3.1.1 Check that the relation $\leq$ defined above is indeed a partial order.

Definition 3.1.4 An equivalence class $[i]$ is **minimal** if there is no $[j]$ such that $[j] \leq [i]$ and $[j] \neq [i]$.

Definition 3.1.5 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.

Example 3.1.1 Consider a stochastic matrix such that the associated graph is given by

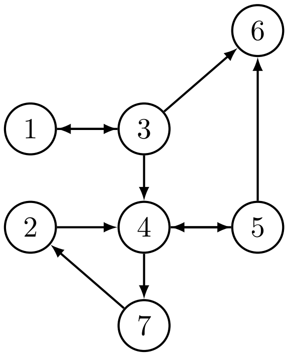


In this case, the communication classes are given by

$$\begin{aligned} &[1] = \{1\}\ , \quad &&[2] = \{2\}\ , \quad &&[3] = \{3\}\ , \\ &[4] = \{4,7\}\ , \quad &&[5] = \{5,6\}\ . \end{aligned}$$

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

Example 3.1.2 Consider the stochastic matrix P with its incidence graph



$$P = \begin{pmatrix} \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 1 \\ \frac{2}{3} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{3} & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{3} & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 \end{pmatrix}$$

The communication classes are: $[1] = \{1,3\}, [6] = \{6\}$, $[2] = \{2,4,5,7\}$. The partial orders are: $[6] \leq [1], [2] \leq [1]$. Thus, $[6]$ and $[2]$ are minimal classes.

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.

Definition 3.1.6 Let $T_i = \inf\{n \geq 1 : x_n = i\}$. If $x_0 = i$, T_i is the first return time to site i.

Note that

$$\{T_i \leq n\} = \cup_{k=1}^{k} \{x_k = n\}.$$

We are interested in the question with what probability a chain starting from i returns to i, or whether a chain from j can reach i with positive probability.

Example 3.1.3 Let us return to the two state Markov chain, take $x_0 \sim \mu$, then we use $\mathbf{P}_\mu$ to denote the probability concerning the chain with $x_0 \sim \mu$.

$$\mathbf{P}_\mu(T_0 = 1) = \mathbf{P}_\mu(x_1 = 0) = \mu(0)(1-\alpha) + \mu(1)\beta.$$

If $x_0 = \delta_0$,

$$\mathbf{P}_0(T_0 = 1) = 1 - \alpha,$$

For $n \geq 1$,

$$\mathbf{P}_0(T_0 = n) = \mathbf{P}(x_1 = 1, \dots, x_{n-1} = 1, x_n = 0 | x_0 = 0) = \alpha(1-\beta)^{n-2}\beta.$$

$$\mathbf{P}_0(T_0 < \infty) = \sum_{n=1}^{\infty} \mathbf{P}_0(T_0 = n) = (1-\alpha) + \sum_{n=2}^{\infty} \alpha(1-\beta)^{n-2}\beta = 1.$$

We will see that many of the conclusions in this simple example holds for more complicated chains.

Recall the notation $\mathbf{P}_i(A) = \mathbf{P}(A|x_0 = i)$ where A is an event. The following holds for a Markov chain in a countable state space.

Proposition 3.1.7 *Let $j \neq i$. Show that j is accessible from i if and only if $\mathbf{P}_i(T_j < \infty) > 0$. Moreover*

$$\mathbf{P}_i(T_j < \infty) \leq \sum_{k=1}^{\infty} P_{ji}^k.$$

Proof. Recall that j is accessible from i, if $P_{ji}^n > 0$ for some n. Since $\{T < \infty\} = \cup_{n=1}^{\infty}\{x_n = j\}$, for any n, $\mathbf{P}(T_j < \infty \,|\, x_0 = i) \geq P_{ji}^n$ the only if part follows. Also

$$\mathbf{P}_i(T_j < \infty) = \mathbf{P}_i(\cup_{k=1}^{\infty}\{X_k = j\}) \leq \sum_{k=1}^{\infty} \mathbf{P}_i\left(X_k = j\right) = \sum_{k=1}^{\infty} P_{ji}^k.$$

So $\mathbf{P}(T_j < \infty \,|\, x_0 = i) = 0$ implies that $P_{ji}^n = 0$ for all n. □

It follows that $i \sim j$ if and only if

$$\mathbf{P}_i(T_j < \infty) > 0, \quad \mathbf{P}_j(T_i < \infty) > 0.$$

Definition 3.1.8 Let $\mathcal{X}$ be countable. We say that a state is recurrent if $\mathbf{P}_i(T_i < \infty) = 1$. Otherwise we say i is transient.

A transient state may have the property that for an infinite number of n, $P_{ii}^n > 0$, but these probabilities are not big (in fact $\sum_{n=1}^{\infty} P_{ii}^n < \infty$ for transient states, see Corollary 3.3.12). Observe that $\sum_{n=1}^{\infty} P_{ii}^n = 0$ if and only if $[i]$ contains no other state, and $P_{ii} = 0$. It is of course a transient state.

Let $\eta_j = \sum_{n=1}^{\infty} \mathbf{1}_{\{x_n = j\}}$ be the number of times the chain visits j (also called the occupation time of $\{j\}$). Recall we denote by $\mathbf{E}_i(\eta_j)$ the expectation $\mathbf{E}(\eta_j \,|\, x_0 = i)$ with initial state i.

Definition 3.1.9 Let $\mathcal{X} = \mathbf{N}$ be a discrete state space. Then the average number of visits to state j, started from i, is

$$U(i,j) := \mathbf{E}(\eta_j \,|\, x_0 = i) = \sum_{n=1}^{\infty} P_{ji}^n.$$

Exercise 3.1.2 Let $\mathcal{X}$ be a countable state space. If the chain is irreducible show that either $\sum_{n=1}^{\infty} P_{ji}^n = \infty$ for all $(i,j) \in \mathcal{X}$, or it is finite for every $i,j \in \mathcal{X}$. Conclude that if $\mathcal{X}$ is finite, then $\sum_{n=1}^{\infty} P_{ji}^n = \infty$ for every (i,j) and that every state is recurrent.

Hint. If $U(i,j) = \infty$, prove that $U(i',j') = \infty$ by obtaining paths from i' to j' from those paths from i to j, by gluing a path from i' to i of positive probability and another path from j to j' of positive probability.

3.1.1 Periodic and aperiodic chains

Let $\mathcal{X}$ be a countable state space. For every state i, we define the set $R(i)$ of **return times** to i by

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

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. If the set $R(i)$ is not empty, it contains infinite numbers.

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

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

If $R(i) = \phi$, we define $d(i) = +\infty$. (This can only happen if $[i]$ contains a single state from which the chain leaves straightaway and never returns.) The period of i is not necessarily the time a chain from i returns to i, it does not even necessarily mean that the chain will necessarily be able to return at time $d(i)$. See Example 3.1.4

Proposition 3.1.10 *Let $\mathcal{X}$ be a countable state space, Suppose that i and j are distinct states with $i \sim j$, then $d(i) = d(j) < \infty$.*

Proof. Since i and j communicate with each other, there exist n and m such that $P_{ij}^n > 0$ and $P_{ji}^m > 0$. This implies that $n+m \in R(i) \cap R(j)$, both $d(i)$ and $d(j)$ divide $n+m$. If $k \in R(i)$ then

$$P_{jj}^{n+m+k} \geq P_{ji}^m P_{ii}^k P_{ij}^n > 0.$$

So one has $n+m+R(i) \subset R(j)$, this implies that $d(j)$ divides every element of $R(i)$. Consequently, $d(j) \leq d(i)$. On the other hand, the same is true with i and j exchanged, so that one must have $d(i) = d(j)$. □

Definition 3.1.11 We call a stochastic matrix **aperiodic** if $d(i) = 1$ for every i. We call it **periodic** of period d if $d(i) = d > 1$ for every i.

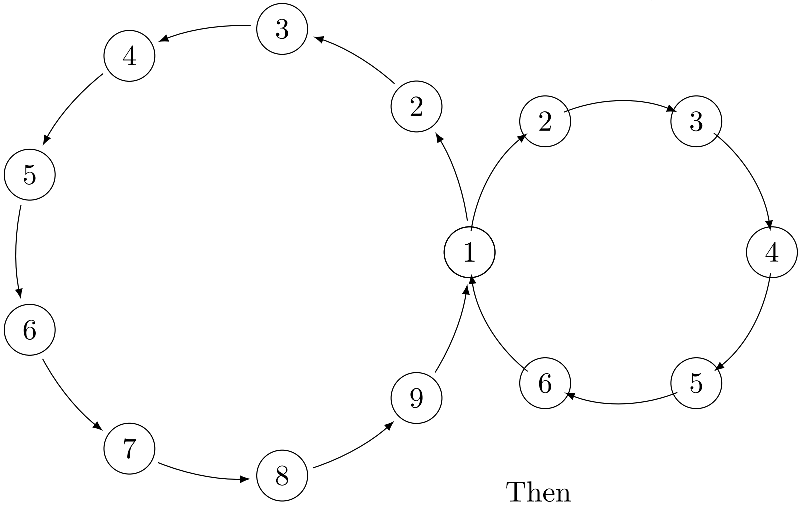
As a consequence of Proposition 3.1.10, any two states in the same communication class have the same period and we can conclude the following.

Corollary 3.1.12 *Let $\mathcal{X}$ be a countable state space. If P is irreducible then there exists d such that $d(i) = d$ for all states i. Thus, P is either aperiodic or of period d for some d.*

Exercise 3.1.3 Prove that if P is irreducible and if there exists j such that $P_{jj} \neq 0$, then the matrix is aperiodic.

If an irreducible stochastic matrix is periodic, it is possible to break it into disjoint sets, say $\{A_i\}$, with an order, so that all transitions from one set go into the next set.

Example 3.1.4 Consider a Markov chain on the the states marked below, the chain moves follow the arrow deterministically, except for at the common node 1 of the left and right circles: at one it picks up either a left 2 or a right 2. Then we have the state space $\mathcal{X} = \{1, 2, 3, 4, 5, 6, 7, 8, 9, R2, R3, R4, R5, R6\}$. (Sorry for the complication, I have not mastered how to label the numbers on the circles at will.) This is a periodic chain with period 3.



Then

$$\{1, 2, 3, 4, 5, 6, 7, 8, 9, R2, R3, R4, R5, R6\} = A_0 \sqcup A_1 \sqcup A_2$$

in which the cycles A_i are given as below:

$$A_0 = \{1, R4, 4, 7\},$$

$$A_1 = \{R2, R5, 2, 5, 8\},$$

$$A_2 = \{R3, R6, 3, 6, 9\}.$$

The chain moves periodically from sets to sets: $A_0 -> A_1 -> A_2 -> A_0 -> \dots$. Observe that it takes exactly 6 steps for the chain starting from i to return to state 1. *The period* 3 *is the time the chain starting from* 1 *returns to the cycle,* A_0*, that contains* 1.

Exercise 3.1.4 Let $y_n = x_{3n}$. Show that y_n is a time homogeneous Markov chain with stochastic matrix P^3.

Observe that

$$P(y_1 = R4|y_0 = 1) = \mathbf{P}(x_3 = R4|x_0 = 1) = P_{R2,1}P_{R3,R2}P_{R4,R3} = \frac{1}{2} \cdot 1 \cdot 1 = 1.$$

Similarly we can work out other transition probabilities and obtain the stochastic matrix for (y_n) restricted to one of the sets A_k. On A_0 it is: $\begin{pmatrix} 0 & 1 & 0 & 1 \\ \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 1 & 0, \end{pmatrix}$ where the first column refers to outlet from state 1, the second column refers to state $R4$, and column 3 refers to state 4, and the last columns refers to state 7. By working out the two loops from 1 to 1, we see the average time for y_n returns to 1 starting from 1 is $\frac{5}{2}$. In general from the probability measure $(\frac{2}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5})$ we can work out the average return time to a state i, which is in inverse proportion with the equilibrium probability of the site.

Exercise 3.1.5 Let us consider a rectangular cube with 6 faces and 8 vertices. For convenience we label them by $\mathcal{X} := \{1, 2, 3, 4, 5, 6, 7, 8\}$. There is a small robot on the cube, jumping from vertices to vertices in the following way: it jumps to its three nearest neighbours with equal probability. Write down the stochastic matrix P for the robot whose position at n is denoted by (x_n).

Show that P is periodic of period 2. Find two disjoint sets A_0, A_1 so that $\mathcal{X} = A_0 \sqcup A_1$ and in one step the chain always jumps from one set to the other. Let $y_n = x_{2n}$. What is the average time for the chain (y_n) (and for the chain x_n respectively) setting off from 1 and return to it?

Finite State Markov Chains

For the rest of the section we assume that $\mathcal{X} = \{1, \dots, N\}$.

In Example 3.1.4, $\mathcal{X}$ decomposes into disjoint subsets of cycles, this is s typical feature of an irreducible periodic chain. Assume that P has period d. We begin with the element 1 (the choice of the index 1 is arbitrary), the cycle contains 1 is:

$$A_0 = \{j : P_{j1}^{kd} > 0 \text{ for some } k \in \mathbf{N}\},$$

Similarly for $n = 1, \dots, d-1$, we define A_n by

$$A_n = \{j \,|\, \exists\ m = 0\ (P^{kd+n}_{j1} > 0 \text{ for some } k \in \mathbf{N}\}\ . \tag{3.4}$$

A characterisation of stochastic matrices with period d is the following.

Lemma 3.1.13 *Let $\mathcal{X} = \{1, \dots, N\}$. Let $d > 1$ be a natural number. An irreducible stochastic matrix P is periodic with period d if and only if it is possible to write $\mathcal{X} = \{1, \dots, N\}$ as a disjoint union of sets $A_0 \sqcup \ldots \sqcup A_{d-1}$ in such a way that if $i \in A_n$ for some n and if j is such that $P_{ji} \neq 0$ then $j \in A_{n+1+kd}$ for some $k \in \mathbf{N}$. We have identified A_{n+kd} with A_n.*

Proof. If such a decomposition exists for $d > 1$, it is clear that a chain starts from 1 cannot return to it in less than d-steps or in k mode d steps for $d \neq 0$. Since the chain is irreducible, it must has positive probability to return to it in d steps. So the chain has period d.

For the converse we define A_n by (3.4). Since P is assumed to be irreducible, the union of the A_n is all of $\{1, \dots, N\}$ (for any j, there exists n such that $P^n_{j1} > 0$.) Furthermore, they are disjoint. Otherwise, one could find j such that it belongs to $A_{n_1} \cap A_{n_2}$. So $P^{n_1+k_1d}_{j1} > 0$ and $P^{n_2+k_2d}_{j1} > 0$ with $k_1, k_2 \in \mathbf{N}$, $n_1, n_2 \in \{0, 1, \dots, d-1\}$. Since P is irreducible, there exists furthermore q such that $P^q_{1j} > 0$, so that $n_1 + k_1 d + q \in R(1)$ and $n_2 + k_2 d + q \in R(1)$. Thus d can divide $n_1 - n_2$ which is only possible when $n_1 = n_2$. The fact that these sets have the required property is then immediate. □

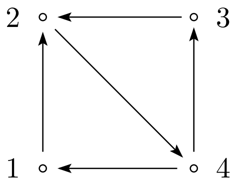


Figure 3.2: Periodic.

The example given in (3.3) is aperiodic. However the example shown in Figure 3.2 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 3.2 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 $q \in (0, 1)$.

The period does not refer to the minimal time for the chain to return to a particular state, it is the time for it to return to its own cycle.

The following result is well-known in number theory:

Lemma 3.1.14 (A number theory lemma) *Let $\mathcal{X} = \{1, \dots, N\}$. Suppose that $R(i) \neq \phi$. There exists $K > 0$ such that $kd(i) \in R(i)$ for every $k \geq K$.*

Proof. By dividing everything by $d(i)$, we can assume without loss that $d(i) = 1$. Since $\gcd R(i) = 1$, there exists a finite collection $d_1, \dots, d_n$ in $R(i)$ such that $\gcd\{d_1, \dots, d_n\} = 1$. The Euclidean algorithm implies that there exist integers $a_1, \dots, a_n$ such that $\sum_{i=1}^n a_i d_i = 1$. Set $M = \sum_{i=1}^n d_i$. Then, for $k = 1, \dots, M$, one has

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

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

Exercise 3.1.6 Let $\mathcal{X} = \{1, \dots, N\}$. Let P be irreducible of period d. Show that, for $n \geq 1$, the period q of P^n is given by $q = d/r$, where r is the greatest common divider between d and n. The corresponding partition $\{B_i\}$ of $\{1, \dots, N\}$ is given by $B_i = \bigcup_{n \geq 0} A_{i+nq \,(\mathrm{mod}\, d)}$, where $\{A_i\}$ is the partition associated to P by Lemma 3.1.13.

Exercise 3.1.7 Let $\mathcal{X} = \{1, \dots, N\}$. 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 d and that the partition $\{B_j\}$ is the one associated by Lemma 3.1.13 to the matrix $P^{d/q}$.

This follows from Lemma 3.1.14 and similarly the following proposition.

Proposition 3.1.15 *Let $\mathcal{X} = \{1, \dots, N\}$. The three following conditions are equivalent:*

(a) P is irreducible and aperiodic.

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

(c) Let $\delta_n = \min_{i,j=1,\dots N} (P^n)_{ij}$. Then there exists $n_0 \geq 1$ such that $\delta_{n_0} > 0$. (Inceidentally $\delta_n \geq \delta_{n_0}$ for $n \geq n_0$.)

Proof. If P is periodic of period d and irreducible, then P^d is reducible and (b) trivially implies (a).

$(c) \implies (a)$ Since there is n_0 such that P^{n_0} has strictly positive entries this clearly implies that P is irreducible since there is always a path of length n_0 between any two vertices. Now from the Chapman-Kolmogorov equation, we get that for all $j, k \leq N$,

$$P_{jk}^{n+1} = \sum_{i=1}^N P_{ji}^n P_{ik} \geq \delta_n \sum_{i=1}^N P_{ik} = \delta_n,$$

since $\sum_{i=1}^{N} P_{ik} = 1$ for all k and $P_{ji}^{n_0} \geq \delta_n$ for all i, j. This implies that

$$\delta_{n+1} \geq \delta_n .$$

Inductively we see that for all $n \geq n_0$, P^n has all entries positive. Therefore for all $n \geq b_0$,

$$n \in R(1) = \{k | \, P_{11}^k > 0\}.$$

We must have period of vertex 1 is 1. Since we showed that P is irreducible this implies that it is aperiodic.

In fact for any $n \leq n_0$ there exists k such that P^{nk} with strictly positive entries, which means P^n is irreducible for every $n \geq 0$. Thus (c) implies (b).

$(a) \implies (c)$: By the number theorem lemma, Lemma 3.1.14, for all $1 \leq i \leq N$ there exists k_i such that $kd(i) \in R(i)$ for all $k \geq k_i$. Let $K = \max_{i \in \mathcal{X}}\{k_i\}$. Since matrix is aperiodic it implies that $d(i) = 1$ for all i and therefore

$$P_{ii}^k > 0, \quad \forall i \in \mathcal{X}, k \geq K.$$

Because P is irreducible then for all $1 \leq j, i \leq N$ there exists $k(j, i) \in \mathbf{N}$ such that $P_{ji}^{k(j,i)} > 0$. Let

$$n_0 = K + \max_{1 \leq j,i \leq N} k(j, i) \ .$$

Since all the entries of P are non-negative, for $n \geq n_0$,

$$P_{ji}^n = \sum_{k=1}^{N} P_{jk}^{k(j,i)} P_{ki}^{n-k(j,i)} \geq P_{ji}^{k(j,i)} P_{ji}^{n-k(j,i)} > 0 \ ,$$

where for the last inequality we used that $P_{ji}^{k(j,i)} > 0$ and that $n - k(j, i) \geq K \geq K_i$ and thus $P_{ji}^{n-k(j,i)} > 0$. Taking minimum over all $1 \leq j, i \leq N$ in the above inequality gives us $\delta_n > 0$. □

Exercise 3.1.8 Consider the simple symmetric random walk on Z, it is irreducible *Why?*. Does the conclusion of part (c) hold? What is its period?)(It moves either to its left neighbour or to its right neighbour, the walk returns to the same site only in even number of steps, so it has period 2.)

By Proposition 3.1.15, we have:

Exercise 3.1.9 Let $\mathcal{X} = \{1, \dots, N\}$. If P is irreducible and aperiodic, show that there exists $n > 0$ and $\delta > 0$ such that $P^n\eta \geq \delta\|\eta\|_1 \mathbf{1}$ for every vector η with entries $\eta_i \geq 0$. Here $\mathbf{1}$ is the vector with every entry being 1 and $\|\eta\|_1 = \sum_i |\eta_i|$.

Lemma 3.1.16 *Suppose that P is an irreducible stochastic matrix. Let $T^n = \frac{1}{n}(P + P^2 + \ldots + P^n)$. Then there exists a number n_0 s.t. T^n has positive entries. There exists $\delta > 0$ such that*

$$\min_{i,j=1,\ldots,N} T^n_{ij} \geq \delta.$$

Thus if η is a vector with non-negative entries, $T^n\eta \geq \delta\mathbf{1}\|\eta\|_1$.

Proof. This is Proposition 3.1.15 if P is aperiodic. If P has period $d > 1$, then P^d is aperiodic and $\mathcal{X}$ decomposes into the union of disjoint blocks A_i. On A_k, P^d is irreducible and so $P^{n_0 d} > 0$ for some n_0. Also for $j \in A_{k+1}$, $P_{ij} > 0$ for some $i_0 \in A_k$. Thus $P^{n_0d+1}_{ij} \geq P^{n_0d}_{ii_0} P_{i_0 j} > 0$. This shows that $P^{n_0d}_{ij} + P^{n_0d+1}_{ij} > 0$ for $i, j \in A_k \cup A_{k+1}$. By induction, this proves $\min_{i,j=1,\ldots,N} T^n_{ij} \geq \delta.$. (The final part follows again from $\eta(i) = \sum_{j=1}^N T^n_{ij}\eta(j) \geq \delta \sum_{j=1}^N \eta(j) = \delta\|\eta\|_1$.) □

3.1.2 Invariant measure for periodic chains

The following proposition holds for discrete time Markov chain on any state space.

Proposition 3.1.17 *Suppose that $T^n\mu = \mu$ for some fixed n. Let $\hat{\mu} = \frac{1}{n}\sum_{k=1}^n T^k\mu$. Then $\hat{\mu}$ is an invariant measure for T.*

Proof. Let A be a Borel measurable set. Then

$$T\hat{\mu}(A) = \frac{1}{n}\sum_{k=1}^{n} T^{k+1}\mu(A) = \frac{1}{n}\sum_{k=1}^{n-1} T^{k+1}\mu(A) + \frac{1}{n}T^{n+1}\mu(A) = \frac{1}{n}\sum_{k=2}^{n} T^{k}\mu(A) + \frac{1}{n}T\mu(A) = \hat{\mu}.$$

□

Remark 3.1.18 If we have a periodic chain with period d, then $\mathcal{X} = A_0 \cup \cdots \cup A_{d-1}$. we may construct an invariant measure for the irreducible P^d on A_k. (by the Perron-Frobenius theorem). Then $\hat{\mu} = \frac{1}{d}\sum_{k=1}^d P^k\mu$ is an invariant measure for P.

3.2 Perron-Frobenius theorem

Let $\mathcal{X} = \{1, \ldots, N\}$ in this section. The L_1-norm on $\mathbf{R}^N$ is defined by $\|\nu\|_1 = \sum_{i=1}^N |\mu(i)|$. Write

$$\mu_+ = (\mu(1) \vee 0, \ldots, \mu(N) \vee 0)$$

for the positive part of μ and similarly μ_- for its negative part

$$\mu_- = \Big((-\mu(1)) \vee 0, \ldots, (-\mu(N)) \vee 0)\Big).$$

Then

Lemma 3.2.1 *1.* $\|\mu\|_1 = \|\mu_+\|_1 + \|\mu_-\|_1$.

2. If $\sum_{i=1}^N \mu(i) = 0$, *then* $\|\mu_+\|_1 = \|\mu_-\|_1 = \frac{1}{2}\|\mu\|_1$.

3. And, if μ_1 *and* μ_2 *are positive vectors (i.e. all entries are non-negative), then the triangle inequality becomes equality:* $\|\mu_1 + \mu_2\|_1 = \|\mu_1\|_1 + \|\mu_2\|_1$.

Exercise 3.2.1 Let P be a stochastic matrix, then

(1) P preserves the mass of a positive measure: $\sum_{i=1}^N (P\mu)(i) = \sum_{i=1}^N \mu(i)$.

(2) $\|P\mu\|_1 \le \|\mu\|_1$. If $\mu \in \mathbf{R}^N$ is a positive vector, the equality holds.

We can let P act on $\mathbf{C}^N$, the above inequality holds for $\mu \in \mathbf{C}^N$.

(1) is obvious. For (2) just observe that,

$$\|P\mu\|_1 = \sum_{i=1}^N \Big| \sum_{j=1}^N P_{i,j}\mu(j)\Big| \le \sum_{j=1}^N \sum_{i=1}^N P_{i,j}\Big|\mu(j)\Big| = \sum_{j=1}^N \Big|\mu(j)\Big| = \|\mu\|_1.$$

We write $|\mu|$ for the vector with entries $|\mu_i|$ and $\sum(\mu)$ for the number $\sum_{i=1}^N \mu_i$.

By Proposition 3.1.15, we have:

Exercise 3.2.2 If P is irreducible and aperiodic, show that there exists $n > 0$ and $\delta > 0$ such that $P^n\eta \ge \delta\mathbf{1}\|\eta\|_1$ for every vector η with entries $\eta_i \ge 0$. Here $\mathbf{1}$ is the vector with every entry being 1.

Theorem 3.2.2 (Perron-Frobenius) *Let* P *be an irreducible stochastic matrix on a finite state space.*

(A) Then 1 *is an eigenvalue for* P, *and there exists exactly one eigenvector* π *(up to multiplication by a constant) with* $P\pi = \pi$. *Furthermore,* π *can be chosen such that all its entries are strictly positive and with* $\sum_{i=1}^N \pi(i) = 1$. *(The unique positive eigenvector with eigenvalue* 1 *of an irreducible stochastic matrix* P *the* ***Perron-Frobenius vector of*** P.*)*

(B) Every eigenvalue of P *must satisfy* $|\lambda| \le 1$. *If* P *is furthermore aperiodic, all other eigenvalues satisfy* $|\lambda| < 1$.

*(C) ** If* P *is periodic with period* p, *there are eigenvalues* $\lambda_j = e^{\frac{2i\pi j}{p}}$ *with associated eigenvector*

$$\mu_j(n) = \lambda_j^{-k}\pi(n) \ , \quad \text{if } n \in A_k, \tag{3.5}$$

where π *is the Perron-Frobenius vector of* P *and the sets* A_k *are the ones associated to* P *by Lemma 3.1.13.*

Proof. Since $\|P\mu\|_1 \leq \|\mu\|_1$ for every vector $\mu \in \mathbf{C}^N$ (see Exercise 3.2.1), the eigenvalues of P must all satisfy $|\lambda| \leq 1$.

– (A). Since the vector $\mathbf{1} = \frac{1}{N}(1, 1, \ldots, 1)$ 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. We define $\alpha = \min\{\|\pi_+\|_1, \|\pi_-\|_1\}$. Since P is irreducible, by Lemma 3.1.16 there exists n and $\delta > 0$ such that $T^n\pi_+ \geq \delta\alpha\mathbf{1}$ and $T^n\pi_- \geq \delta\alpha\mathbf{1}$. Therefore,

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

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 it has mass 1: $\sum_{i=1}^N \pi(i) = 1$. All entries of π are strictly positive since $\pi = T^n\pi \geq \delta\mathbf{1}$, again by Lemma 3.1.16.

–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 $P\pi_2 = \pi_2$. Then the vector $\pi_3 = \pi_1 - \pi_2$ is also an eigenvector with eigenvalue 1 for P. By the previous argument, we can assume that the entries of the π_i are positive summing to 1. However, since $\sum_{i=1}^N \pi_3(i) = 0$, one must have $\pi_3 = 0$.

**(The rest of the proof is no given in class) To part (B) and (C), we show that $\lambda_j = e^{2\pi\frac{j}{p}}, p = 0, \ldots, p-1$, where $d = d(i)$ for some state i and therefore for all, are the only eigenvalues on the unit circle of $\mathbf{C}^N$, centred at 0.

-Consider an eigenvalue with $|\lambda| = 1$ but $\lambda \neq 1$. Let $\lambda = e^{i\theta}$ and $\nu = \Big(r_1 e^{i\theta_1}, \ldots, r_N e^{i\theta_N},\Big)$ one of its an eigenvectors. We can choose the phases in such a way that $r_i \geq 0$, and we normalise them in such a way that $\sum_{i=1}^N r_i = 1$. The relation $P\mu = \lambda\mu$, $\sum_{j=1}^N P_{kj}\nu_j = e^{i\theta}\nu_k$, then translates into

$$\sum_{j=1}^N e^{i\theta_j} P_{kj} r_j = e^{i(\theta+\theta_k)} r_k \ . \tag{3.6}$$

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

$$e^{i(\theta_j-\theta_k-\theta)} = 1 \ , \quad \text{for every } j \text{ and } k \text{ such that } P_{kj} \neq 0. \tag{3.7}$$

Combining this with (3.6) in turn implies that $r = \pi$. Indeed for every k,

$$\sum_{j=1}^N e^{i(\theta+\theta_k)} P_{kj} r_j = e^{i(\theta+\theta_k)} r_k, \quad i.e. \sum_{j=1}^N P_{kj} r_j = r_k,$$

so $r = (1, \ldots, r_N)$ is the Perron-Frobenius vector π.

–Since P^n is a stochastic matrix with eigenvalue $\lambda^n = e^{i\theta}$, repeat the previous arguments shows that

$$e^{i\theta_j} = e^{i\theta n + i\theta_k} \ , \quad \text{for every } j \text{ and } k \text{ such that } P^n_{kj} \neq 0. \tag{3.8}$$

Since P is irreducible, $R(i)$ contains every integer of the form kp, where $k \geq K$ for some K and so we can take $k = j$ above, and $n = Kp$ and $n = (K+1)p$. Then $\theta Kp = 0(\text{mod} 2\pi)$, $\theta(K+1)p = 0(\text{mod}2\pi)$ which implies that $\theta p = 0 \pmod{2\pi}$. This all possible eigenvalues with $|\lambda = 1|$ are of the form $\lambda_j = e^{i2\pi \frac{j}{p}}$.

– In particular if P is aperiodic, 1 is then the only eigenvalue with modulus 1.

–We find μ which satisfies $P\mu = \lambda\mu$. By multiplying u with a scalar, we can assume that $\theta_1 = 0$. The relation (3.7) allow us to assign θ_j for $\lambda = e^{i\theta}$ in the following way. If $k \in A_0$, A_0 being the cycle containing 1, then we set $\theta_k = \theta_1 = 0$. For $k \in A_1$, the next cycle, $P_{k1} \neq 0$, we set $e^{i\theta_k} = e^{-i\theta} = \lambda^{-1}$. Iterating this procedure and moving to the next cycle A_n we may define $\theta_k = \lambda^{-n \,(\text{mod}\, 2\pi)}$ for every $k \in A_n$, thus defining every θ_k. We can verify as follows that this is an eigenvector associated to λ. Equation (3.6) can be written as

$$\sum_{j=1}^{N} e^{i(\theta_j - \theta - \theta_k)} P_{kj} \pi(j) = \pi(k).$$

Fix k, on its left hand side, the only non-zero term P_{kj} are those j in cycle class flowing into that of k. For example for $k = 1$, $\theta_1 = 0$, $\theta_j = \lambda^{-(p-1)}$ for $j \in A_{p-1}$, this is

$$\sum_{j \in A_{p-1}} P_{1j} \pi(j) = \pi(1),$$

this is the identity for Perron-Frobenius vector, observing that $\sum_{j \in A_{p-1}} \lambda^1 P_{1j}\pi(j) = \sum_{j=1}^N \lambda^1 P_{1j}\pi(j))$. This is true for all $k \in A_0$. The rest of the relations can be verified similarly. □

Remark 3.2.3 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.

3.3 The structure of invariant measures and ergodic theorems

Definition 3.3.1 The total variation distance between two probability measures μ and ν (n any measurable space) are

$$\|\mu - \nu\|_{\text{TV}} = 2 \sup_{A \subset \mathcal{X}} |\mu(A) - \nu(A)| \ ,$$

where the supremum runs over all measurable subsets of $\mathcal{X}$.

Definition 3.3.2 We say that a sequence $\{\mu_n\}$ converges in total variation to a limit μ if

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

Example 3.3.1 Let $\mathcal{X} = \{1, 2\}$ and let $P = \begin{pmatrix} 1-\alpha & \beta \\ \alpha & 1-\beta \end{pmatrix}$. Then $\pi = \begin{pmatrix} \frac{\beta}{\alpha+\beta} \\ \frac{\alpha}{\alpha+\beta} \end{pmatrix}$. Let $\mu_0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$. Then $\mu_0 - \pi = \frac{\alpha}{\alpha+\beta} \begin{pmatrix} 1 \\ -1 \end{pmatrix}$ and $\|\mu_0 - \pi\|_{TV} = \frac{2\alpha}{\alpha+\beta}$. Now for $\alpha + \beta \neq 1$, (what happens if $\alpha + \beta = 1$?)

$$\begin{aligned} P^n\mu_0 - \pi = P^n(\mu_0 - \pi) &= \frac{\alpha}{\alpha+\beta} \begin{pmatrix} P^n_{11} & P^n_{12} \\ P^n_{21} & P^n_{22} \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} \\ &= \frac{\alpha}{\alpha+\beta} (P^n_{11} - P^n_{12}) \begin{pmatrix} 1 \\ -1 \end{pmatrix} \end{aligned}$$

Note $P^n_{11} - P^n_{12} = (1 - \alpha - \beta)^n$. So if $\alpha + \beta < 1$,

$$\|P^n\mu_0 - \pi\| = (1 - \alpha - \beta)^n \|\mu_0 - \pi\|_{TV} \to 0.$$

Lemma 3.3.3 *If μ, ν are probability measures on a discrete space $\mathcal{X}$, then*

$$\|\mu - \nu\|_{\mathrm{TV}} = \sum_{i\in\mathcal{X}} |\mu(i) - \nu(i)| = \|\mu - \nu\|_1.$$

Also, $\|\mu - \nu\|_{\mathrm{TV}} = 2\sum_{\{i:\mu(i)\geq\nu(i)\}}(\mu(i) - \nu(i))$.

Proof. (This is left as exercise) Let $B = \{i : \mu(i) \geq \nu(i)\}$. For any $A \subset \mathcal{X}$,

$$\begin{aligned} |\mu(A) - \nu(A)| &= |\mu(A\cap B) - \nu(A\cap B) - (\nu(A\cap B^c) - \mu(A\cap B^c))\,| \\ &\leq \max\left(|\mu(A\cap B) - \nu(A\cap B)|, |\mu(A\cap B^c) - \nu(A\cap B^c)|\right) \\ &\leq \max\left(|\mu(B) - \nu(B)|, |\mu(B^c) - \nu(B^c)|\right) \\ &= |\mu(B) - \nu(B)| = \sum_{\{i:\mu(i)\geq\nu(i)\}} (\mu(i) - \nu(i)), \end{aligned}$$

This last line follows from $\mu(B) - \nu(B) = \nu(B^c) - \mu(B^c)$ which follows from $\mu(\Omega) - \nu(\Omega) = 0$ and

$$2|\mu(B) - \nu(B)| = 2\sum_{\{i:\mu(i)\geq\nu(i)\}} (\mu(i) - \nu(i)).$$

This completes together. □

For the rest of the section we let $\mathcal{X} = \{1, \ldots, N\}$. A very important consequence of the Perron-Frobenius theorem is the following.

Theorem 3.3.4 *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. ** (This proof uses techniques we used already, so not given in class) It follows from Exercise 3.2.2 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 $a = \|(\pi-\nu)_+\|_1 = \|(\pi-\nu)_-\|_1 = \frac{1}{2}\|\pi-\nu\|_1$. One then has

$$\begin{aligned}\|P^n\nu - \pi\|_1 &= \|P^n(\pi-\nu)\|_1 = \|P^n(\pi-\nu)_+ - P^n(\pi-\nu)_-\|_1\\ &\leq \left\|P^n(\pi-\nu)_+ - \delta a\cdot\mathbf{1}\right\|_1 + \left\|P^n(\pi-\nu)_- - \delta\, a\cdot\mathbf{1}\right\|_1\\ &= \|P^n(\pi-\nu)_+\|_1 - \delta\,N\,a + \|P^n(\pi-\nu)_-\|_1 - \delta N\ a\\ &\leq \|(\pi-\nu)_+\|_1 + \|(\pi-\nu)_-\|_1 - \delta N\|\pi-\nu\|_1\\ &\leq (1-\delta N)\|\pi-\nu\|_1\ .\end{aligned}$$

Since ν was arbitrary, one gets $\|P^{kn}\nu - \pi\|_1 \leq (1-\delta)^k\|\pi-\nu\|_1$ by iterating this bound, we then take $k\to\infty$ to conclude (observe that $\|P^m(\nu-\pi)\|_1$ decreases with m.) □

Note that Theorem 3.3.4 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.

Note under the conditions of the theorem, each column of P^n converges to π. Just take ν to be the jth basis vector, we see that $\lim_{n\to\infty} P^n_{ji} = \pi(j)$ for every i.

Exercise 3.3.1 Show that the conclusion of Theorem 3.3.4 also hold if one only assumes that $\sum_i \nu_i = 1$.

One has the following:

Lemma 3.3.5 *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. More specifically there exists $\lambda > 0$ such that*

$$\|P^n\mu\|_1 \leq e^{-\lambda t}.$$

In particular, the eigenvalues of P are all of modulus strictly less than 1 and so $1-P$ is invertible.

Proof. ** (not given in class) It is sufficient to prove the claim for μ positive with norm 1 (unless $\mu = 0$). Define $T^n = \frac{1}{n}(P + \cdots + P^n)$ as in the proof of the Perron-Frobenius theorem. For a positive vector μ, one has $\|P\mu\|_1 \leq \|\mu\|_1$ and one has also

$$\|P^{n+1}\mu\|_1 = \frac{1}{n}(n\|P^{n+1}\mu\|_1) \leq \frac{1}{n}(\|P^2\mu\|_1 + \cdots + \|P^{n+1}\mu\|_1) = \|PT^n\mu\|_1$$

for every $n > 0$. Choose n such that $T^n\mu \geq \delta\mathbf{1}\|\mu\|_1$ (such an n exists by the irreducibility of P). Since P is not a stochastic matrix, there exists $\alpha > 0$ and an index j_0 such that $\sum_i P_{ij} = 1 - \alpha$. Let $e_{j_0} = (0, \ldots, 1, \ldots, 0)$ denote the unit vector with entries 1 at j_0th entry and with 0 at other entries. Therefore

$$\begin{aligned}\|P^{n+1}\mu\|_1 \leq \|PT^n\mu\|_1 &= \|P(T^n\mu - \delta e_{j_0}) + \delta e_{j_0}\|_1 = \|P(T^n\mu - \delta e_{j_0})\|_1 + \delta\|Pe_{j_0}\|_1 \\ &\leq \|T^n\mu - \delta e_{j_0}\|_1 + \delta(1-\alpha) \\ &= \|T^n\mu\|_1 - \delta \cdot \|e_{j_0}\|_1 + \delta(1-\alpha) \leq (1-\alpha\delta) = (1-\alpha\delta)\|\mu\|_1.\end{aligned}$$

Choose and fix a natural number n_0 such that the above inequality holds, then

$$\|P^{(n_0+1)k}\mu\|_1 \leq (1-\delta\alpha)^k\|\mu\|_1 \overset{(k\to\infty)}{\to} 0,$$

which concludes the convergence. The rate of P^n convergence is at least λ^n where $\lambda = (1 - \delta\alpha)^{\frac{1}{n_0+1}}$, thus concluding the proof. □

Recall first the Borel-Cantelli lemma from probability theory:

Lemma 3.3.6 (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. Equivalently this implies that $\mathbf{P}\left(\cap_{n=1}^{\infty} \cup_{k=n}^{\infty} A_n\right) = 0$.*

If $[i]$ is in a recurrent state, by which we mean the minimal class, then it has to visit one of the states in $[i]$ infinitely often, call this state k, and therefor returns to itself with positive probability, following from that $P_{ik}^n > 0$ for some n.

Theorem 3.3.7 *Let $\{x_n\}$ be a Markov process with transition probabilities P and let i be from a non-minimal class. Then the probability that $x_n \in [i]$ for an infinite number of values n is 0.*

Proof. **(not given in class) 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

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

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 3.3.5. □

Theorem 3.3.8 *Let P be an arbitrary stochastic matrix. 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.*

Proof. **(not given in class) Let $A_i, i = 1, \dots, k$ denote the minimal classes, and A_{k+1} the collections of sites not in one of the minimal classes. Let $\mu = (a_1, \dots, a_k, a_{k+1})$ be an invariant probability measure, where a_{k+1} is a vector corresponds to the transient states, the rest of each of the a_i's is vector corresponding to k minimal (=closed) communications classes. The matrix P can be written as

$$P = \begin{pmatrix} P_1 & 0 & \dots & 0 & S_1 \\ 0 & P_2 & \dots & 0 & S_2 \\ 0 & 0 & \dots & P_k & S_k \\ 0 & 0 & \dots & 0 & Q \end{pmatrix}, \tag{3.9}$$

Then,

$$P = \begin{pmatrix} P_1 & 0 & \dots & 0 & S_1 \\ 0 & P_2 & \dots & 0 & S_2 \\ 0 & 0 & \dots & P_k & S_k \\ 0 & 0 & \dots & 0 & Q \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \dots \\ a_k \\ a_{k+1} \end{pmatrix} = \begin{pmatrix} P_1 a_1 + S_1 a_{k+1} \\ P_2 a_2 + S_2 a_{k+1} \\ \dots \\ P_k a_k + S_k a_{k+1} \\ Q a_{k+1} \end{pmatrix} \tag{3.10}$$

Since $P^n \mu = \mu$, $Q^n a_{k+1} = a_{k+1}$. But $Q^n \to 0$, this is only possible if $a_{k+1} = 0$ and $P_i a_i = a_i$, so a_i are multiples of the Perron-Frobenius vectors π_i of P_i. Then $\mu = (\alpha_1 \pi_1, \dots, \alpha_k \pi_k, 0)$, with $\sum \alpha_i = 1$ and $a_i > 0$, concluding the proof. □

3.3.1 Exercises

Exercise 3.3.2 If $P^n_{i,j} \to \pi(i)$ for every i (the rate the Markov chain goes to state i from any other state is $\pi(i)$), show that π is an invariant probability measure.

If a time homogeneous Markov chain is aperiodic and irreducible with finite state space then we know

$$\lim_{n \to \infty} \mathbf{P}_\mu(x_n = i) = \pi(i).$$

In particular, $\lim_{n \to \infty} P^n_{ij} = \mathbf{P}_j(x_n = i) = \pi(i)$ for every state j. If the chain is reducible we can also work out the probability that the chain eventually ends in a particular state. For example if i is a transient state, this is 0. Suppose that all recurrent communication classes consist of singletons. Let be i be a recurrent state (with $[i]$ containing only one single element i) and let $B_0 = \lim_{n \to \infty} x_n = 0$. Set $f(j) = \mathbf{P}_j(B_0) = \mathbf{P}_0($ the chain eventually ends at site 0). This is the probability that the chain starts from j ends at 0 eventually. Then

$$\begin{aligned} f(j) &= \mathbf{E}_j(\mathbf{E}(\text{ the chain eventually ends at site } 0 \,|\, x_1)) = \mathbf{E}_j(f(x_1)) \\ &= \sum_{i=0}^{5} f(i) \mathbf{P}(x_1 = i | x_0 = j) = \sum_{i=0}^{5} f(i) P_{ij} = (fP)(j). \end{aligned}$$

If the minimal classes are not singletons, we may amalgamate elements of each minimal class together and treat such classes as singletons, work out the ratio of the probability flowing into each of the minimal classes, and then redistribute this probability among their elements according to the ratio of their Perron-Frobenius vectors. This amalgamating method can be done because once the chain enters it, it never returns. To the rest of the chains, its exact whereabout is not observable and of no relevance. The probability we calculated is then the probability it enters the minimal classes. The probability it eventually enters a state in the minimal class is then distributed according to the unique invariant probability distribution of the reduced chain.

Exercise 3.3.3 Let $\mathcal{X} = \{0, 1, 2, 3, 4, 5\}$ and let (x_n) be the time homogeneous Markov chain with stochastic matrix

$$P = \begin{pmatrix} 1 & \frac{1}{3} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 \\ 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 \\ 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \\ 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{3} & 1 \end{pmatrix}$$

What is the probability that the chain starting from 3 ends at 0? Here the minimal classes are $\{0\}$ and $\{5\}$. Let $f(j)$ be the probability that the chain starting from j ends at 0. Let $f = (f(0), f(1), \ldots, f(5))$. We solve the left eigenvalue problem

$$f = fP,$$

with the boundary conditions $f(0) = 1$, $f(5) = 0$ (starting from 5, never ends at 0). Then

$$\begin{aligned} 2f(1) &= 1 + f(2) \\ 2f(2) &= f(1) + f(3) \\ 2f(3) &= f(2) + f(4) \\ 2f(4) &= f(3) + 0. \end{aligned}$$

Solving this we obtain: $f = (1, \frac{4}{5}, \frac{3}{5}, \frac{2}{5}, \frac{1}{5}, 0)$. The answer is $f(3) = \frac{2}{5}$.

3.3.2 Where do transient states end up ?**

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

$$P = \begin{pmatrix} I & S \\ 0 & Q \end{pmatrix} , \tag{3.11}$$

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

Proof. One has

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

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 3.3.5. □

3.3.3 Return times and alternative criterion for recurrent/transient

Definition 3.3.10 Let $\mathcal{X}$ be countable. We say that a state is recurrent if $\mathbf{P}_i(T_i < \infty) = 1$. Otherwise we say i is transient.

Let i be a state, let

$$T_j^0 = 0, \qquad T_j^1 \equiv T_j = \inf\{k \ge 1 \,|\, x_k = j\},$$

and recursively we define

$$T_j^{n+1} = \inf\{k > T_j^n \text{ such that } x_k = j\}.$$

Recall that j is recurrent if $\mathbf{P}_j(T_j < \infty) = 1$.

Lemma 3.3.11 *1. If j is recurrent, the sequence of intervals $\{T_j^n - T_j^{n-1}\}_{n\ge1}$ are independent. And for any $k \ge 1$, $m \in \mathcal{X}$,*

$$\mathbf{P}\Big(T_j^{k+1} - T_j^k = m\Big) = \mathbf{P}_j(T_j = m).$$

2. For any two state i, j, any natural number $k \ge 1$,

$$\mathbf{P}_i(T_j^{k+1} < \infty) = \mathbf{P}_i(T_j < \infty) \cdot \mathbf{P}_j(T_j^k < \infty).$$

Proof. (1) Suppose that a state j is recurrent, i.e. $P_j(T_j < \infty) = 1$. We fix this j and set $T = T_j$ and for $T^k = T_j^k$ for $k \geq 2$ for simplicity. Then

$$\theta_{T^k}(x.) = \Big(x_{T^k}, x_{T^k+1}, \dots, x_{T^k+2}, \dots, \Big).$$

By the strong Markov property,

$$\begin{aligned}\mathbf{P}\Big(T^{k+1} - T^k = m | \mathcal{F}_{T^k}\Big)(\omega) &= \mathbf{P}_{x_{T^k}(\omega)}\Big(T^{k+1} - T^k = m\Big)\\ &= \mathbf{P}_{x_{T^k}(\omega)}\Big(x_1 \neq j, \dots, x_{m-1} \neq j, x_m = j\Big)\\ &= \mathbf{P}_j\Big(x_1 \neq j, \dots, x_{m-1} \neq j, x_m = j\Big).\end{aligned}$$

(The second line follows from : $T^{k+1} - T^k = m$ if and only if $(x_{T^k+1} \neq j, \dots, x_{T^k+m-1} \neq j, x_{T^k+m} = j)$.) Taking expectations we see that

$$\mathbf{P}\Big(T^{k+1} - T^k = m\Big) = \mathbf{P}_j\Big(x_1 \neq j, \dots, x_{m-1} \neq j, x_m = j\Big) = \mathbf{P}_j(T = m).$$

To see that $\{T_j^k - T_j^{k-1}\}$ are independent random variables, we observe that $\mathbf{P}(T^{k+1} - T^k = m | \mathcal{F}_{T^k})(\omega)$ is a deterministic event, so $\{T^{k+1} - T^k = m\}$ is independent of $\mathcal{F}_{T^k}$ and $T_j \in \mathcal{F}_{T^k}$ for any $j = k - 1$.

**For (2) let $\Phi : \mathcal{X}^\infty \to \mathbf{R}$ be the function that $\Phi((a_n)) = 1$ if $a_n = j$ for some $n \geq 1$ and $\Phi((a_n)) = 0$ otherwise. (so $\Phi = \mathbf{1}_A$ where A contains sequences that visits j at some finite time $n \geq 1$). Then on $\{T_j < \infty\}$, $T^{k+1}(\omega) < \infty$ if and only if

$$\Phi\Big(\theta_{T_k} x.(\omega)\Big) = 1.$$

On the set $\{T_j < \infty\}$, we apply the strong Markov property:

$$\mathbf{E}\Big(\Phi(\theta_{T_k} x.) \mathbf{1}_{\{T_j < \infty\}} | \mathcal{F}_{T_j}\Big) = \mathbf{1}_{\{T_j < \infty\}} \mathbf{E}_{x_T}\Big(\Phi(\theta_{T_k} x.)\Big).$$

Since $x_T = j$, we take expectation to obtain that:

$$\mathbf{P}_i(T^{k+1} < \infty) = \mathbf{E}(\Phi(\theta_{T_k} x.)) = \mathbf{P}_j(T_1 < \infty) \mathbf{P}_i(T^k < \infty).$$

The rest of the proof follows from induction. □

Theorem 3.3.12 (Recurrent-criterion)

1. *If j is a recurrent state, it is visited an infinite number of times with probability one.*

2. *A state j is transient if and only if $\sum_{n=1}^\infty P_{jj}^n < \infty$. Equivalently, a state j is recurrent if and only if $\sum_{n=1}^\infty P_{jj}^n = \infty$.*

Proof. Part (1) follows from that the inter-arrival times to j are i.i.d.'s, $\mathbf{P}_j(T_{n+1} - T_n < \infty) = 1$. We introduce η_j as the number of visits to site j (the occupation time of the site j): $\eta_j = \sum_{n=1}^\infty \mathbf{1}_{\{x_n=j\}}$. Then,

$$\sum_{n=1} P_{jj}^n = \mathbf{E}_j(\sum_{n=1}^\infty \mathbf{1}_{\{x_n=j\}}) = \sum_{n=1}^\infty \mathbf{P}_j(\eta_j \ge n) = \sum_{n=1}^\infty \mathbf{P}_j(T_j^n < \infty) = \sum_{n=1}^\infty (\mathbf{P}_j(T_j < \infty))^n .$$

This geometric series is convergent if and only if $\mathbf{P}_j(T_j < \infty) < 1$, i.e. if and only if j is transient. □

Corollary 3.3.13 *Suppose that $i \sim j$, then i is recurrent implies j is recurrent. So a state being transient or recurrent is a class property.*

Proof. Assume now i is recurrent, so $\sum_{k=1}^\infty P_{ii}^k = \infty$ by Corollary 3.3.12. Since i and j are accessible from each other, we may choose m_0, n_0 so that $P_{ij}^{n_0} P_{ji}^{m_0} > 0$, also

$$\sum_{k=1}^\infty P_{jj}^{m+n+k} \ge \sum_{k=1}^\infty P_{ji}^{m_0} P_{ij}^k P_{ij}^{n_0} = P_{ji}^{m_0} P_{ij}^{n_0} \sum_{k=1}^\infty P_{ii}^k = \infty,$$

The assertion that j is recurrent follows from Corollary 3.3.12. □

We can now classify the states of a finite state Markov chain as either recurrent or transient.

Corollary 3.3.14 *Let $\mathcal{X}$ be a finite state space. Every state in the minimal communication class is visited infinitely often (and so recurrent) and every state in the other communication classes is visited only finitely often (and is transient)*

Proof. The statement about transient states is the content of Theorem 3.3.7.

If a stochastic matrix is irreducible then every state is visited infinitely often almost surely. Otherwise for at least for one i, and therefore for all i, $\sum_{n=1}^\infty P_{ii}^n < \infty$. Then almost every path visits every state only finitely often, which is impossible for the infinite path. If i is in a closed state, when the chain starts from i, it does not see any other states rather than states in its own communication class and so visits each infinite often. □

Remark 3.3.15 ** The following is the occupation time of j for the chain starting from i:

$$U(i,j) = \mathbf{E}_i(\eta_j) = \sum_{n=1}^\infty \mathbf{P}_i(x_n = j) = \sum_{n=1}^\infty P_{ji}^n. \tag{3.12}$$

Exercise 3.3.4 If i is recurrent and $\mathbf{P}_i(T_j < \infty) > 0$ then $\mathbf{P}_j(T_i < \infty) = 1$.

Proof. Suppose that $\mathbf{P}(T_i < \infty | x_0 = j) < 1$, then there is a set of paths of positive probability starting from j never visits i. Since $\mathbf{P}(T_j < \infty | x_0 = i) > 0$, there is a set of paths of positive probability leading from i arriving at j at some finite time, we concatenate this path to the previously mentioned set of path starting from j never visits i, thus obtaining a set of paths from i, never returns to i. Thus $\mathbf{P}_i(T_i = \infty) \geq \delta > 0$ and i is not recurrent, contradicting with the assumption. □

3.4 Expected return times and the law of large numbers

In this section, $\mathcal{X} = \{1, \dots, N\}$, we are interested in the following question: given a finite-state Markov process (x_n) with transition probabilities P starting in a distinguished state i. Let T be the random (stopping) time defined by

$$T_i = \inf\bigl\{n \geq 1 \text{ such that } x_n = i\bigr\}.$$

Thus is called the (first) return time to i. What is the average time it take to get back to i? It may be rather surprising that this can easily be computed explicitly: $\mathbf{E}T_i = \frac{1}{\pi(i)}$.

Let us make a preliminary calculation. If (x_n) is a Markov chain with $x_0 = i$ almost surely, we denote $\mathbf{E}_j(T_i^p)$ the expectation of the T_i^p.

Lemma 3.4.1 *Let (x_n) be an aperiodic and irreducible Markov process on a finite state space with transition probabilities P. Then for any two states i and j and for any $\alpha \geq 1$, the expectation $\mathbf{E}_j\bigl(T_i^\alpha\bigr)$ is finite.*

Proof. ** Since $\mathbf{P}_j(T_i < \infty) = 1$,

$$\mathbf{E}_j(T_i^\alpha) = \sum_{n\geq 0} n^\alpha \mathbf{P}_j\bigl(T_i = n\bigr) \leq \sum_{n \geq 0} n^\alpha \, \mathbf{P}_j\bigl(T_i > n-1\bigr).$$

By Proposition 3.1.15 there exist $n_0 > 0$ s.t. $\delta = \inf_{i,j\in\mathcal{X}} P^{n_0}_{i,j} > 0$.

$$\begin{aligned}
&\mathbf{P}_j\bigl(T_i > n_0(k+1)\bigr) \leq \mathbf{P}_j\bigl(x_{n_0} \neq i, x_{2n_0} \neq i, \dots x_{n_0(k+1)} \neq i\bigr)\\
&= \mathbf{P}\Bigl(x_{n_0(k+1)} \neq i \,|\, x_{n_0} \neq i, x_{2n_0} \neq i, \dots, x_{n_0k} \neq i\Bigr)\mathbf{P}_j\bigl(x_{n_0} \neq i, x_{2n_0} \neq i, \dots x_{n_0k} \neq i\bigr)\\
&= \mathbf{P}\Bigl(x_{n_0(k+1)} \neq i \Bigm| x_{n_0k} \neq i\Bigr)\mathbf{P}_j\bigl(x_{n_0} \neq i, x_{2n_0} \neq i, \dots x_{n_0k} \neq i\bigr)\\
&\leq (1-\delta)\mathbf{P}_j\bigl(x_{n_0} \neq i, x_{2n_0} \neq i, \dots x_{n_0k} \neq i\bigr)\\
&\leq \dots \leq (1-\delta)^{k+2}.
\end{aligned}$$

We now put the natural numbers in blocks of n_0 integers. On each block, $n \in [n_0k, n_0(k+1)]$, we observe that $\mathbf{P}_j(T_i > n-1) \leq \mathbf{P}_j(T_i > n_0k - 1) \leq (1-\delta)^k$. We then use the bound

$n^\alpha \le (n_0(k+1))^\alpha$. Picking up the earlier estimates, this yields the following:

$$\mathbf{E}_j(T_i^\alpha) \le \sum_{n\ge 0} n^\alpha\, \mathbf{P}_j\bigl(T_i > n-1\bigr) \le n_0 \sum_{k=0}^\infty ((k+1)n_0)^\alpha (1-\delta)^k \le n_0^{\alpha+1} \sum_{k=1}^\infty k^\alpha (1-\delta)^{k-1} < \infty,$$

completing the proof. □

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 3.4.2 (Strong Law of Large Numbers) *Let $\{\xi_n\}_{n\ge1}$ be a sequence of i.i.d. real-valued random variables such that $\mathbf{E}\xi_n = \mu < \infty$ and $\mathbf{E}(\xi-\mu)^2 = \sigma^2 < \infty$. Then, one has*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^n \xi_k = \mu, \qquad\qquad \textit{almost surely.}$$

Its simplest extension to Markov processes states:

Theorem 3.4.3 *Let x be an aperiodic irreducible homogeneous Markov process on a finite state space $\mathcal{X}$ with invariant probability measure π.*

1. *Let $f\colon \mathcal{X} \to \mathbf{R}$. Then, one has the law of large numbers (LLN):*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^n f(x_k) = \sum_{j\in\mathcal{X}} f(j)\pi(j) = \int_{\mathcal{X}} f d\pi \text{ ,a.s..} \tag{3.13}$$

2. *Suppose that $x_0 = i$ almost surely for some distinguished state i, then one has $\mathbf{E}T_i = \frac{1}{\pi(i)}$.*

Proof. ** Since any function on $\mathcal{X}$ can be written as a finite linear combination of functions $\mathbf{1}_i \overset{\text{def}}{=} \mathbf{1}_{\{i\}}$, it suffices to consider Theorem 3.4.3 with $f = \mathbf{1}_i$, so that (3.13) becomes:

$$\lim_{m\to\infty} \frac{1}{m}\sum_{n=1}^m \mathbf{1}_i(x_n) = \pi(i). \tag{3.14}$$

We have already proven in Theorem 3.3.4 that $P^n\mu \to \pi$ for any probability measure μ, in particular $\lim_{n\to\infty} \mathbf{E}_\mu(\mathbf{1}_i(x_n)) = \lim_{n\to\infty} \mathbf{P}_\mu(x_n = i) = \pi(i)$ and therefore by the dominated convergence theorem

$$\lim_{m\to\infty} \mathbf{E}_\mu\Bigl(\frac{1}{m}\sum_{n=1}^m \mathbf{1}_i(x_n)\Bigr) = \lim_{m\to\infty}\Bigl(\frac{1}{m}\sum_{n=1}^m \mathbf{E}_\mu \mathbf{1}_i(x_n)\Bigr) = \pi(i)\ . \tag{3.15}$$

In order to get (3.13) it thus suffices to get rid of the expectation on the left-hand side.

We take $x_0 = i$. Let $T_0^i = 0$. Since $\{T_i^{k+1} - T_i^k, k = 0, 1, 2, \dots\}$ are independent i.i.d.'s with second moments (Lemma 3.4.1) and distributed as T, the first return time to i, we apply to it the law of large numbers,

$$\lim_{n\to\infty} \frac{T_i^n}{n} = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} (T_i^{k+1} - T_i^k) = \frac{1}{n}(\mathbf{E}_i(T_i^1 - T_i^0)) = \mathbf{E}T \ , \tag{3.16}$$

almost surely (so far, we have three notations : $T = T_i = T_i^1$ and $\mathbf{E}T = \mathbf{E}_i T$.)

Since $T_i^n \geq n$ by definition, $|\frac{n}{T_i^n}| \leq 1$. The above converges holds also in L_1 by the Lebesgue's dominated convergence theorem,

$$\lim_{n\to\infty} \mathbf{E}\Big|\frac{n\mathbf{E}T}{T_i^n} - 1\Big| = 0 \ . \tag{3.17}$$

Since $x_{T_i^n} = i$, the definition of the times T_i^n yields the relation

$$\frac{n}{T_i^n} = \frac{1}{T_i^n} \sum_{k=0}^{T_i^n} \mathbf{1}_i(x_k). \tag{3.18}$$

We can rewrite this as

$$\frac{n}{T_i^n} = \frac{1}{n\mathbf{E}T} \sum_{k=1}^{n\mathbf{E}T} \mathbf{1}_i(x_k) + R_n \ , \tag{3.19}$$

where the error term $R_n \to$. Taking expectation of the right hand side of (3.19), taking $n \to \infty$ and use the LLN in averaged form (3.15), one has

$$\lim_{n\to\infty} \mathbf{E}\frac{n}{T_i^n} = \lim_{n\to\infty} \frac{1}{n\mathbf{E}T} \sum_{k=1}^{n\mathbf{E}T} \mathbf{P}(x_k = i) + \lim_{n\to\infty} \mathbf{E}R_n = \pi(i) + \lim_{n\to\infty} \mathbf{E}R_n = \pi(i) \ ,$$

and so by (3.16), $\frac{1}{\mathbf{E}T} = \pi(i)$, proving part two of the assertion.

To show $\mathbf{R}_n \to 0$ a.s. and in L_1, we estimate:

$$\begin{aligned} |R_n| &= \Big| \left(\frac{1}{T_i^n} \sum_{k=n\mathbf{E}T}^{T_i^n} + \frac{1}{T_i^n} \sum_{k=1}^{n\mathbf{E}T} - \frac{1}{n\mathbf{E}T} \sum_{k=1}^{n\mathbf{E}T} \right) \mathbf{1}_i(x_k) \Big| \\ &\leq \Big| \frac{T_i^n - n \cdot \mathbf{E}T}{T_i^n} \Big| + \Big| \frac{1}{T_i^n} - \frac{1}{n\mathbf{E}T} \Big| n\mathbf{E}T = 2\Big|1 - \frac{n \cdot \mathbf{E}T}{T_i^n}\Big| \to 0, \end{aligned}$$

almost surely and in L_1.

To show $\frac{1}{m}\sum_{n=1}^{m} \mathbf{1}_i(x_n)$ converges, we return to (3.19) and take $n \to \infty$ using the fact that $\frac{n\mathbf{E}T}{T_i^n} \to 1$ and $R_n \to 0$ almost surely, with a bit analysis we obtain the required LLN. □

Remark 3.4.4 Show that the assumption that x is aperiodic is not needed in order to prove (3.15). Therefore, Theorems 3.4.3 hold for general irreducible Markov chains on a finite state space.

3.5 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:

The left translations on G are the maps: $h \in G \mapsto gh \in G$ where $g \in G$.

Definition 3.5.1 Consider a group G and a time homogeneous Markov chain with transition matrix P on G. We say that the Markov chain is a **left-invariant random walk** on G if and only if the left translations preserve the matrix P: i.e. $P_{gh_1,gh_2} = P_{h_1,h_2}$ for any $g, h_1, h_2 \in G$. (It is right invariant random walk if $P_{h_1g,h_2g} = P_{h_1,h_2}$ for any $g, h_1, h_2 \in G$.)

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

Example 3.5.1 Random walk on Z. Let $x_n = x_{n-1} + Y_n$ where Y_i are i.i.d.'s with values in Z. Let $\hat{P}$ be the probability distribution of Y_i. Then

$$\mathbf{P}(x_n = j|x_{n-1} = i) = \mathbf{P}(Y_n = j - i) = \hat{P}(j - i).$$

Then $\mathbf{P}(x_n = j|x_{n-1} = i) = \mathbf{P}(x_n = j + k|x_{n-1} = i + k)$. This is an invariant random walk.

Exercise 3.5.1 The Markov chain is left invariant if and only if there exists a probability measure $\hat{P}$ on G such that $\mathbf{P}(x_{n+1} = g\,|\,x_n = h) = \hat{P}(h^{-1}g)$. The Markov chain is right invariant if and only if there exists a probability measure $\hat{P}$ on G such that $\mathbf{P}(x_{n+1} = g\,|\,x_n = h) = \hat{P}(gh^{-1})$.

Proof. We give the proof for the left-invariant case. Suppose that the stochastic matrix P satisfies $P_{gh_1,gh_2} = P_{h_1,h_2}$. We define a probability measure $\hat{P}(g)$ on the group by $\hat{P}(\{g\}) = P_{g,e}$ where e is the identity in G. Then,

$$P_{h_1,h_2} = P_{h_2^{-1}h_1,e} = \hat{P}(h_2^{-1}h_1).$$

If, on the other hand, $P_{g,h} = \hat{P}(gh^{-1})$ fro some probability measure $\hat{P}$ on G, then

$$P_{g'g,g'h} = \hat{P}(\{h^{-1}(g')^{-1}g'g\}) = \hat{P}(\{h^{-1}g\} = P_{g,h},$$

completing the proof. □

Exercise 3.5.2 Show that if $\{x_n\}$ is a left-invariant random walk on G with transition probability P, then $\{x_n^{-1}\}$ is a right-invariant random walk and find its transition probabilities.

Because of Exercise 3.5.2, it suffices to study one of the two types of random walks. Let us choose the left-invariant ones and use 'random walk' to indicate a left invariant random walk.

Exercise 3.5.3 Consider a random walk with transition matrix P on a finite group G and define $\Sigma = \{g \in G \,|\, \hat{P}(g) > 0\}$. Show that P is irreducible if and only if Σ generates G.

Exercise 3.5.4 Show that the normalised counting measure $\pi(g) = 1/|G|$ is an invariant measure for every left invariant random walk on G.

Hint: $\sum_{h\in G} \hat{P}(h^{-1}g) = 1$, so we have a doubly stochastic matrix: the rows sum up to one as well.

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 $\{1, \dots, n\}$ into itself, the composition law on the group is simply the composition of maps.

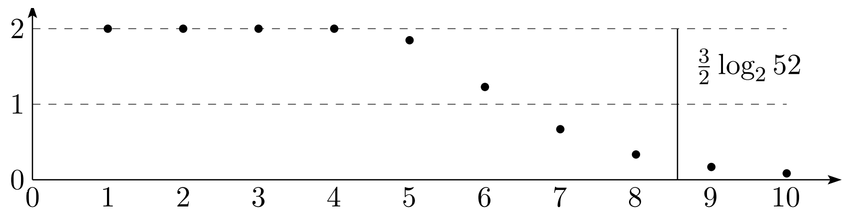
3.5.1 The Gilbert-Shannon-Reeds shuffling**

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

The precise formulation of Theorem 3.5.2 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 3.5.3 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 3.5.2 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 3.5.4 *The probability that m GSR shuffles of a deck of n cards produce a given permutation σ is given by*

$$P(\sigma) = \frac{1}{2^{mn}} \binom{2^m + n - R(\sigma)}{n} , \tag{3.20}$$

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 3.5.2. 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 $N_i \in \{1, \ldots, 4\}$ 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 $N_1 \ne N_2$ and that $N_3 \ne N_4$. In this particular case, the list of all possible GSR shuffles (written in the format $(N_1 N_2 N_3 N_4 N_5)$) that produce the right permutation is thus given by

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

This is consistent with (3.20) 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 (3.20). 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 3.5.2. One has

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

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

3.6 Countable state spaces

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.

Example 3.6.1 We can make even simpler examples. Let $x_0 = 0$, $x_{n+1} = x_n + \xi_{n+1}$, where ξ_n are uniform distributed random variables with values in $\{1, 2, 3, 4, \dots\}$. Then x_n moves to the right on the integer lattice, and cannot have any invariant measure (It cannot give charge at 0, for it moves away in one step. Similarly it cannot charge any state.

Example 3.6.2 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

$$\pi_n = \frac{\pi_{n+1} + \pi_{n-1}}{2} , \tag{3.21}$$

for every $n \in \mathbf{Z}$. A moment of reflection shows that the only positive solution to (3.21) 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.

3.6.1 A summary

Recall that a state i is recurrent if $\mathbf{P}_i(T_i < \infty) = 1$. It is otherwise transient.

Definition 3.6.1 A recurrent state i is positive recurrent, if $\mathbf{E}_i T_i := \mathbf{E}(T_i | x_0 = i) < \infty$. A recurrent state is null recurrent if $\mathbf{E}_i T_i = \infty$.

Being recurent, transient, or positive recurrent is a class property.

Definition 3.6.2 A chain is recurrent/transient if all states are recurrent/transient. A chain is positive recurrent if all states are positive recurrent.

On a finite state space we have seen that every state for an irreducible Markov chain is positively recurrent, c.f. Lemma 3.4.1.

Theorem 3.6.3 *1. If P has a recurrent state, it has an invariant measure.*

2. If the chain is irreducible and recurrent, then there exists an invariant measure, unique up to a constant. The invariant measure is finite if and only if the chain is positive recurrent (i.e. $\mathbf{E}_k T_k < \infty$ for every k).

3. If the chain is irreducible and positive recurrent, we denote the invariant probability measure by π. Then the mean return time to k is:

$$\mathbf{E}_k T_k = \frac{1}{\pi_k}.$$

Also,

$$\mathbf{E}_i(\textit{number of visits to } j \textit{ before returning to } i) = \frac{\pi(j)}{\pi(i)}.$$

4. If π is an invariant probability measure and if $\pi(j) > 0$ then j is recurrent.

5. If P is irreducible with stationary probability measure π then $\mathbf{E}_i T_i < \infty$ for all i (i.e. all states are positive recurrent) and $\pi(i) = \frac{1}{\mathbf{E}_i T_i}$.

Summary. If the time homogeneous Markov chain is irreducible then all states are simultaneously recurrent and transient (Corollary 3.3.13), also all states are simultaneously simultaneously positive recurrent or not (Theorem 3.6.12). On a countable state space, an invariant measures may not have finite mass. If the time homogeneous Markov chain has a recurrent state, we can construct an invariant measure which has the property that its value at a site is the frequency it returns to the site and only the states in the same communication class have positive mass. This measure has finite mass if and only if the site is positive recurrent. If the chain is irreducible and recurrent, there is at most one invariant measure (up to a multiple of constants). If the chain is irreducible and positive recurrent for one site, then $\mathbf{E}_i T_i < \infty$ for all sites and there exists an invariant probability measure (up to a multiple of constants). For an irreducible chain, the existence of an invariant probability measure is in fact equivalent to that all states are positively recurrent.

Example 3.6.3 Let x_n be a simple random walk on Z, where $x_n = x_{n-1} + Y_n$ where Y_i are i.i.d.'s with values in $\{1, -1\}$. Let $p \in (0,1)$ so $\mathbf{P}(Y = 1) = p$ and

$$P(x_1 = i+1 | x_0 = i) = p, \quad P(x_1 = i-1 | x_0 = i) = 1-p.$$

If $p = \frac{1}{2}$, the chain is recurrent, not positive recurrent. If $p \neq \frac{1}{2}$, the chain is transient.

To check whether a state i is recurrent, by Corollary 3.3.12 we only need to verify that $\sum_{n=1}^{\infty} P_{i,i}^n = \infty$. But

$$\sum_{n=1}^{\infty} P_{ii}^n = \sum_{k=1}^{\infty} P_{ii}^{2k} = \sum_{k=1}^{\infty} \binom{2k}{k} p^k (1-p)^k.$$

If $4p(1-p) < 1$ we can apply ratio test to see this is convergent. If $4p(1-p) = 1$, this is precisely the case $p = \frac{1}{2}$ the is infinite , this can be proved with the help of sterling's formula: $k! \sim \sqrt{2\pi k}(k/e)^k$, then $\binom{2k}{k} p^k (1-p)^k \sim \frac{1}{\sqrt{k}}(4pq)^k = \frac{1}{\sqrt{k}}$, thus every state is recurrent for $p = \frac{1}{2}$.

Since it is doubly stochastic, $\mu(i) = 1$ defines an invariant measure. The uniform measure on Z is an invariant measure, not finite. Since recurrent irreducible chain has at most one invariant measure, Theorem 3.6.9 below, it does not have an invariant probability measure.

If $p \neq \frac{1}{2}$, there exists another invariant measure: $\nu(i) = (\frac{p}{1-p})^i$. One can verify that it satisfies the equation: $\sum_j P_{ij}\mu(j) = \mu(i)$, which means $\mu(i-1)p + \mu(i+1)(1-p) = \mu(i)$.

Example 3.6.4 The nearest neighbour random walk on Z^d, which has probability $\frac{1}{2d}$ to jump to one of its $2d$ nearest neighbour, is transient for every $d \neq 1, 2$. It is null recurrent for $d = 1, 2$.

3.6.2 Proof**

We briefly outline some of the main theorems, the construction theorem is of special interest.

Theorem 3.6.4 *If i is a recurrent state then the following defines an invariant measure:*

$$\mu(j) = \mathbf{E}_i \left(\sum_{n=0}^{T_i - 1} \mathbf{1}_{\{x_n = j\}} \right).$$

Remark 3.6.5 It is clear that $\mu(i) = 1$ and $\mu(j)$ is the number of visits to j before the chain starting from i returns to i. Since $\mathbf{P}_i(T_i < \infty) = 1$, we see

$$\mu(j) = \sum_{n=0}^{\infty} \mathbf{P}_i(x_n = j, T_i > n).$$

It is also clear that $\mu(j) = 0$ if $\mathbf{P}(T_j < \infty | x_0 = i) = 0$, so the chain only charges those states in the communication class $[i]$. (c.f. Proposition 3.3.4). Also

$$\sum_{j \in \mathcal{X}} \mu(j) = \mathbf{P}_i \left(\sum_{n=0}^{\infty} T_i > n \right) = \mathbf{E}_i T_i,$$

so it is a probability measure if and only if $\mathbf{E}_i T_i < \infty$.

This is the average number of returns to site before the chain starting from i returns to i. It records the relative frequency for the chain visiting j.

Proof. We show that $P\mu = \mu$. For every k we compute,

$$\begin{aligned}\sum_{j=1}^{\infty} P_{kj}\mu(j) &= \sum_{j=1}^{\infty} P_{kj} \sum_{n=0}^{\infty} \mathbf{P}_i(x_n = j, T_i > n) = \sum_{j=1}^{\infty} P_{kj} \sum_{n=1}^{\infty} \mathbf{P}_i(x_n = j, T_i > n) + \sum_{j=1}^{\infty} P_{kj}\mathbf{P}(x_0 = i, T_i > 0)\\
&= \sum_{n=1}^{\infty}\sum_{j\neq i} P_{kj}\mathbf{P}_i(x_n = j, T_i > n) + P_{ki} = \sum_{n=1}^{\infty}\sum_{j\neq i} P_{kj}\mathbf{P}_i(x_n = j, T_i > n-1) + P_{ki}\\
&= \sum_{n=1}^{\infty}\sum_{j\neq i} \mathbf{P}(x_{n+1} = k \,|\, x_n = j)\mathbf{P}_i(T_i > n-1 | x_n = j)\mathbf{P}(x_n = j) + P_{ki}\\
&= \sum_{n=1}^{\infty}\sum_{j\neq i} \mathbf{P}(x_{n+1} = k, x_n = j, T_i > n-1) + P_{ki}\\
&= \sum_{n=1}^{\infty}\sum_{j\neq i} \mathbf{P}(x_{n+1} = k, x_n = j, T_i > n) + P_{ki}\end{aligned}$$

On line 4 we have used conditional independence of the future and the past on $\{x_n = j\}$. For $k \neq i$, $\mathbf{P}_i(x_0 = k, T_i > 0) = 0$, we obtain that

$$\begin{aligned}P\mu(k) = \sum_{j=1}^{\infty} P_{kj}\mu(j) &= \sum_{n=1}^{\infty}\sum_{j\neq i} \mathbf{P}(x_{n+1} = k, x_n = j, T_i > n+1) + P_{ki}\\
&= \sum_{m=2}^{\infty} \mathbf{P}_i(x_m = k, T_i > m) + P_{ki} = \sum_{m=0}^{\infty} \mathbf{P}_i(x_m = k, T_i > m) = \mu(k).\end{aligned}$$

If $k = i$,

$$P\mu(i) = \sum_{n=1}^{\infty}\sum_{j\neq i} \mathbf{P}_i(x_{n+1} = i, x_n = j, T_i > n) + P_{ii} = \sum_{n=1}^{\infty} \mathbf{P}_i(T_i = n+1) + P_{ii} = 1 = \mu(i).$$

Finally we show that $\mu(k) < \infty$ ffor every k. Firstly we observe that $\sum_{k=1}^{\infty} P_{i,k}^n \mu(k) = 1$ and $P_{i,k}^n > 0$ is equivalent to $\mathbf{P}(T_i < \infty | x_0 = k) > 0$. Consequently, $\mathbf{P}(T_i < \infty | x_0 = k) > 0$ implies that $\mu(k) < \infty$. If $\mathbf{P}(T_i < \infty | x_0 = k) = 0$, then $P(T_k < \infty | x_0 = i) = 0$, the latter implies that $\mu(k) = 0$ from the definition. (The former assertion follows from Proposition 3.3.4: i is recurrent and $\mathbf{P}(T_i < \infty | x_0 = k) > 0$ implies that $\mathbf{P}(T_k < \infty | x_0 = i) > 0$.) □

Remark 3.6.6 Without knowing that the invariant measure is a probability measure, we cannot start the chain from ν. We start it from any j, then waiting until it hits i before a specific time n(call A_0 the event it does not hit i). Let L be the last time it visits i before n. Then $l = n - m$ are disjoint events and

$$\Omega = A_0 \sqcup \sum_{m=1}^{n} \{L = n - m\}$$

$$= A_0 \sqcup \{x_{n-1} = i\} \sqcup \{x_{n-1} \neq i, x_{n-2} = i\} \sqcup \cdots \sqcup \{x_{n-1} \neq i, x_{n-2} \neq i, x_1 \neq i, x_0 = i\}.$$

Also, we could set $\mathbf{P}_\nu(x_n = k) = \sum_{j \in \mathcal{X}} \mathbf{P}_j(x_n = k)\nu(j)$ which equals $\nu(k)$ by the invariance. In other words, we begin with ν push it forward by the chain, the evolution $P^n \nu = \nu$. If ν is a probability measure, then this is the same as the statement that the probability distribution of x_n with initial ν is ν.

Lemma 3.6.7 *Let ν be an invariant measure. Let i be a recurrent state and μ the invariant measure defined in Theorem 3.6.4, where $\mu_i(j) = \sum_{n=0}^\infty \mathbf{P}(x_n = j, T_i > n | x_0 = i)$. Then*

$$\nu(k) \geq \nu(i)\, \mu_i(k), \forall k. \tag{3.22}$$

In fact, $\nu(k) \geq \sum_{m=1}^n \mathbf{P}_i(x_m = k, T_i > m) P_{i,j}^{n-m}$.

Proof. ** We first decompose the space according to the value of the last time the chain visits i before time n, then use the Markov property

$$\begin{aligned}
\mathbf{P}_j(x_n = k) &\geq \sum_{m=1}^n \mathbf{P}_j(x_n = k, x_{n-1} \neq i, \ldots, x_{n-m+1} \neq i, x_{n-m} = i) \\
&= \sum_{m=1}^n \mathbf{P}(x_m = k, x_{m-1} \neq i, \ldots, x_1 \neq i | x_{n-m} = i) \mathbf{P}_j(x_{n-m} = i) \\
&= \sum_{m=1}^n \mathbf{P}_i(x_m = k, T_i > m) \mathbf{P}_j(x_{n-m} = i).
\end{aligned}$$

Multiply the identity by $\nu(k)$ and summing over all j's and use $P^{n-m}\nu = \nu$, we obtain:

$$\begin{aligned}
\sum_{j \in \mathcal{X}} \mathbf{P}_j(x_n = k)\nu(j) &\geq \sum_{m=1}^n \mathbf{P}_i(x_m = k, T_i > m) \sum_{j \in \mathcal{X}} \mathbf{P}_j(x_{n-m} = i)\nu(j) \\
&= \sum_{m=1}^n \mathbf{P}_i(x_m = k, T_i > m)\nu(i).
\end{aligned}$$

Take $n \to \infty$, we have

$$\nu(k) \geq \lim_{n \to \infty} \sum_{m=1}^n \mathbf{P}_i(x_m = k, T_i > m-1)\, \nu(i) = \nu(i)\, \mu_i(k).$$

□

Remark 3.6.8 The above lemma can also be obtained, recursively, using

$$\nu(k) = \sum_{j \in \mathcal{X}} P_{kj}\nu(j) = P_{ki}\nu(i) + \sum_{j \neq i} P_{kj}\nu(j)$$

So on the second iteration, $\mu(k) = P_{ki}\nu(i) + \sum_{j \neq i} P_{kj}P_{ji}\nu(i) + \sum_{j_1 \neq i}\sum_{j_2 \neq i} P_{kj_1}P_{j_1 i}\nu(i)$, i.e.

$$\mu(k) = \mathbf{P}(x_1 = k | x_0 = i)\nu(i) + \mathbf{P}(x_2 = k, x_1 \neq j, x_0 = i)\nu(i) + \sum_{j_1 \neq i}\sum_{j_2 \neq i} P_{kj_1}P_{j_1 i}\nu(i).$$

Recall that for an irreducible chain, all states are simultaneously recurrent or simultaneously transient. If all states are recurrent we say the chain is recurrent.

Theorem 3.6.9 *If the chain is irreducible and recurrent, then the invariant measure is unique up to a constant.*

Proof. Let ν be an invariant measure and let i be a distinguished point, Lemma 3.6.7 shows that for any $k \in \mathcal{X}$: $\nu(k) \geq \nu(i)\,\mu_i(k)$. Observe that $\nu - \mu_i$ is an invariant measure, so the following holds by the invariance:

$$0 = \nu(i) - \nu(i)\mu_i(i) = \sum_{k\in\mathcal{X}} P_{ik}^n \nu(k) - \nu(i) \sum_{k\in\mathcal{X}} P_{ik}^n \mu_i(k) = \sum_{k\in\mathcal{X}} P_{ik}^n(\nu(k) - \nu(i)\mu_i(k)) \geq 0.$$

Since every term in the summation is non-negative, the equality can only holds if

$$P_{ik}^n(\nu(k) - \nu(i)\mu_i(k)) = 0$$

for every k. For every k there exists n with $P_{ik}^n > 0$ (irreducible), it follows that $\nu(k) = \mu_i(k)$ for all k. □

If the chain is irreducible and transient, we may no longer have uniqueness.

We know that $\sum_{n=1}^\infty P_{ji}^n$ are simultaneously finite or infinite for all i, j.

Lemma 3.6.10 *Recall* $\eta_j = \sum_{n=1}^\infty \mathbf{1}_{x_n=j}$ *is the occupation time of the site* j*. Then,*

$$\sum_{n=1}^\infty P_{ji}^n = \frac{\mathbf{P}_i(T_j < \infty)}{1 - \mathbf{P}_i(T_j < \infty)}.$$

Proof.

$$\begin{aligned}\sum_{n=1}^\infty P_{ji}^n = \mathbf{E}_i(\eta_j) = \sum_{k=1}^\infty \mathbf{P}_i(\eta_j \geq k) &= \sum_{k=1}^\infty \mathbf{P}_i(T_j^k < \infty) \\ &= \sum_{k=1}^\infty \mathbf{P}_i(T_j < \infty)\mathbf{P}_j(T_j^{k-1} < \infty) \\ &= \sum_{k=1}^\infty \mathbf{P}_i(T_j < \infty)(\mathbf{P}_j(T_j < \infty))^{k-1} \\ &= \frac{\mathbf{P}(T_j < \infty | x_0 = i)}{1 - \mathbf{P}(T_j < \infty | x_0 = j)}.\end{aligned}$$

In line 2 and line 3 we have applied Lemma 3.3.11. □

Theorem 3.6.11 *If* π *is an invariant probability measure and if* $\pi(j) > 0$ *then* j *is recurrent.*

Proof. Since $P^n\pi = \pi$, then $\sum_{i=1}^{\infty} P_{ji}^n \pi(i) = \pi(j)$ for every n. Summing over n:

$$\sum_{i=1}^{\infty}\sum_{n=1}^{\infty} P_{ji}^n \pi(i) = \sum_{n=1}^{\infty} \pi(j).$$

If $\pi(j) \neq 0$, then LHS is also infinite, but the left hand side is:

$$\sum_{i=1}^{\infty} \pi(i) \frac{\mathbf{P}(T_j < \infty | x_0 = i)}{1 - \mathbf{P}(T_j < \infty | x_0 = j)} = \frac{\mathbf{P}_\pi(T_j < \infty)}{1 - \mathbf{P}(T_j < \infty | x_0 = j)},$$

which is finite unless $\mathbf{P}(T_j < \infty | x_0 = j) = 1$ (i.e. j is recurrent). We have applied Lemma 3.6.10. □

Theorem 3.6.12 *If P is irreducible with stationary probability measure π then $\mathbf{E}_i T_i < \infty$ for all i (i.e. all states are positive recurrent) and $\pi(i) = \frac{1}{\mathbf{E}_i T_i}$.*

Proof. There always exists i with $\pi(i) > 0$, this site i is recurrent by Theorem 3.6.11. By irreducibility, every site is recurrent. For a distinguished i, define

$$\mu(j) = \sum_{n=0}^{\infty} \mathbf{P}_i(x_n = j, T_i > n).$$

Then summing over j, we have

$$\sum_{j=1}^{\infty}\sum_{n=0}^{\infty} \mathbf{P}(x_n = j, T_i > n) = \sum_{n=0}^{\infty} \mathbf{P}_i(T_i > n) = \mathbf{E}_i T_i < \infty.$$

Thus i is positive recurrent. Hence the unique invariant measure is given by $\pi(j) = \frac{\mu(j)}{\mathbf{E}_i T_i}$ and $\pi(i) = \frac{1}{\mathbf{E}_i T_i}$. This procedure can be applied to every state and the conclusion follows. □

Chapter 4

Time Reversal

4.1 Discrete state space

Suppose that $\mathcal{X}$ is countable and π is a probability measure and (x_n) a time homogeneous Markov chain. Fix a time $m > 0$, set $\hat{x}_n = x_{m-n}$. Then $(\hat{x}_n)$ is a Markov chain, this follows from the equivalence of the Markov property and the independence of its future and past when conditioned on the present. However $\hat{x}_n$ may not be a time-homogeneous Markov chain unless its initial distribution is an invariant distribution. To have $\hat{x}$ to be a copy of x, x_0 should start from an invariant probability measure π, for

$$\mathbf{P}(x_m = i) = \mathbf{P}(\hat{x}_0 = i)$$

$$\mathbf{P}((x_m = i, x_0 = j) = \mathbf{P}(x_0 = i, x_m = j).$$

Summing over j, we have $\mathbf{P}(x_m = i) = \mathbf{P}(x_0 = i)$, and also we see $\hat{x}_0$ is distributed as π.

Theorem 4.1.1 *Suppose that (x_n) is an irreducible time homogeneous Markov chain with stochastic matrix P and with initial distribution the invariant probability measure π (which we assume to exist). Then $(\hat{x}_n)$ is again a time homogeneous Markov chain with with initial distribution the invariant probability measure π and with the stochastic matrix $\hat{P}$ given by*

$$\hat{P}_{ji} = P_{ij}\frac{\pi(j)}{\pi(i)}.$$

Proof. By Theorem 3.6.11 , the chain is positive recurrent and $\pi(i) > 0$ for every i. Observe also that $\hat{x}_0 = x_M$ is distributed as π. It is easy to guess the stochastic matrix for $\hat{M}$ as following:

$$\hat{P}_{ji} = \mathbf{P}(\hat{x}_{n+1} = j \,|\, \hat{x}_n = i) = \frac{\mathbf{P}(\hat{x}_{n+1} = j,\, \hat{x}_n = i)}{\mathbf{P}(\hat{x}_n = i)}$$

$$= \frac{\mathbf{P}(x_{M-n} = i,\, x_{M-n-1} = j)}{P(x_{M-n} = i)} = \frac{\mathbf{P}(\, x_{M-n-1} = j | x_{M-n} = i)\mathbf{P}(x_{M-n} = i)}{\pi(i)} = P_{ij}\frac{\pi(j)}{\pi(i)}.$$

Observe $\sum_{j\in\mathcal{X}} \hat{P}_{ji} = 1$, so $\hat{P}$ is a stochastic matrix. For $\hat{x}$ to be a Markov process with stochastic matrix $\hat{P}$ and initial distribution π it is sufficient to show that

$$\mathbf{P}(\hat{x}_0 = i_0, \dots, \hat{x}_n = i_n) = \hat{P}_{i_n, i_{n-1}} \dots \hat{P}_{i_1, i_0}\pi(i_0).$$

This can be verified by turning the LHS to $\mathbf{P}(x_0 = i_0, \dots, x_n = x_0)$ and use the corresponding expression for (x_n). □

Theorem 4.1.2 *If $P_{ji}\pi(i) = P_{ij}\pi(j)$ for all i, j, then $\hat{x}_n$ is also a time homogeneous Markov chain with stochastic matrix P and with initial distribution π.*

Definition 4.1.3 The relation

$$P_{ji}\pi(i) = P_{ij}\pi(j), \quad \forall i, j \tag{4.1}$$

is called detailed balance. A Markov chain is said to be reversible if the new Markov chain (x_{m-n}) is again a time homogeneous Markov chain with stochastic matrix P and with initial distribution π.

Summing over j in (4.4), we see that π is automatically an invariant measure for P. If this holds we say the Markov process (x_n) is reversible (with respect to π.) For a reversible chain, $P_{ij} \neq 0$ implies $P_{ji} \neq 0$, so the arrows between any two sites in the incidence graph must be in both directions. For irreducible chains, we can always rotate the sites, so that the stochastic matrix has the property: its lower diagonal has non-zero entry everywhere. Then we may want to multiply the rows by a number so that the the upper triangle equals the lower triangle and check the resulting matrix is symmetric.

The detailed balance relation allows one to easily 'guess' an invariant measure if one believes that a given process is reversible by using the equality

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

Exercise 4.1.1 Let us define $\langle f, g\rangle_\pi = \int fg d\pi = \sum_i f(i)g(i)\pi(i)$. Then P is reversible w.r.t. π if and only if

$$\langle Pf, g\rangle = \langle f, Pg\rangle.$$

4.1.1 A remark on application in numerical computations**

Suppose that we want to estimate the average of a function f with respect to a probability measure π, which is $\sum_{i\in\mathcal{X}} f(i)\pi(i)$. We may choose i.i.d. random variable with common probability distribution π. However in many situations, such as in statistical physics, $\mathcal{X}$ is very large and the π is only known up to a multiple, e.g. $\sum_{i\in\mathcal{X}}\pi(i)$ is very large and often involving combinatory factors which are difficult to add up and so it is often impossible to compute $\sum_{i\in\mathcal{X}}\pi(i)$ precisely. Then we might use the Mont Carlo Markov chain method (MCMC). This started with Metropolis (1953).

The Mont Carlo Markov chain method (MCMC) for computing the average of a function f with respect to a probability measure π is to construct a finite state irreducible Markov chain with invariant measure π, then use the law of large numbers for the estimation:

$$\sum_{i\in\mathcal{X}} f(i)\pi(i) = \lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^n f(x_k).$$

The convergence rate is quite good. If we can construct a Markov chain which is time reversible then it is sufficient to know π up to a constant. For such processes $P_{ij}\pi_j = P_{ji}\pi_i$, and so the total mass of the finite invariant measure disappears in the ratio.

This relation is not sufficient to construct the stochastic matrix. However if we start any irreducible Markov chain R we may define

$$\begin{aligned} P_{ji} =& R_{ij} \vee \frac{\pi(j)}{\pi(i)} R_{ij}, \quad i \neq g, \\ P_{ii} =& 1 - \sum_{j\neq i} P_{ji} \end{aligned}$$

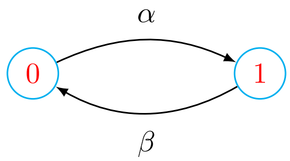
Exercise 4.1.2 Show that P is a time reversible chain w.r.t. π.

This construction does not necessarily produce an irreducible chain (and so in particular there might be other invariant measures, to which the chain may converge to when a wrong initial date is used.) To produce non-irreducible chain, we start with R on a non-oriented graph. Then there are two standard choices for R, they are known respectively as the Metropolis algorithm and the Gibbs sampler.

Observe the difference of MCMC versus MC is that we may start from any initial distribution, when time runs its course we will arrive approximately the invariant probability distribution, while Monte Carlo method uses the invariant probability distribution as the initial distribution.

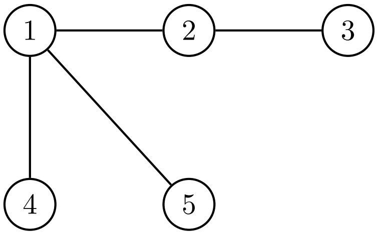
4.1.2 Examples

Example 4.1.1 Let us consider a Markov chain on two states $\{0,1\}$ with $P = \begin{pmatrix} 1-\alpha & \beta \\ \alpha & 1-\beta. \end{pmatrix}$.



Then $P_{10}\pi(0) = P_{01}\pi(1)$ means $\alpha\pi(0) = \beta\pi(1)$. So π is proportional to (β, α).

Example 4.1.2 Let us consider a graph (V, E) with V the set of vertices and E the set of edges. We will assume that the graph is undirected (non-oriented) and connected.



If i, j are connected by an edge, we write $i \sim j$ and say they are adjacent vertices. We assume that there is a weight function w on E, $0 < w(i,j) = w(j,i) < \infty$ if (i,j) is an edge. Let V be the state space of a Markov chain with transition mechanism given by:

$$P_{ji} = \frac{w(i,j)}{w(i)}, \qquad w(i) = \sum_{j \sim i} w(i,j).$$

Let $w = \sum_i w(i)$. Then

$$\pi(i) = \frac{w(i)}{w}$$

defines a probability measure and the chain is reversible with respect π.

It is clear that the chain is irreducible if and only if the graph is connected.

We may also assign a degree to a vertex: $d(i)$ is the number of edges from i, and define

$$P_{ij} = \begin{cases} \frac{1}{d(j)}, & \text{if } i \text{ and } j \text{ are connected by an edge,} \\ 0, & \text{otherwise .} \end{cases}$$

Then $\pi(j) = \frac{d(j)}{2|E|}$ where $|E|$ denotes the number of edges.

Consider a chessboard with only one pieces. Let this piece moves on the otherwise empty chessboard by at every timestep choosing with equal probability the eligible moves. Then it is simple to compute the average time it returns to its initial position i: it is $\frac{2|E|}{d(i)}$. Then it is a matter to count the eligible moves. A standard chessboard has 64 squares. A king piece can move to any one of the square adjacent to it, the graph is connected. A knight's eligible moves are: two steps horizontally and one step vertically. Then umber of edges for the knight move to be 168. (The pawn's graph is not undirected, the bishop's graph is not connected.)

4.2 General state space

Given a time homogeneous Markov chain $(x_n, n \geq 0)$ on a general state space $\mathcal{X}$ with transition probability P and a corresponding invariant measure π, we may start the chain from the initial distribution π, then x_n is distributed as π for every $n \geq 0$. One can say more: the random function $\omega \to (x_\cdot(\omega))$ with state space the sequence space $\mathcal{X}^{\{0\}\cup\mathbf{N}}$ is stationary. Since the multi-time marginals determine a probability measure on $\mathcal{X}^{\{0\}\cup\mathbf{N}}$, one can say this probability measure is stationary. It is convenient to extend this to construct a 2-sided stochastic process $(x_n, n \in \mathbf{Z})$ and so it determines a probability measure on the space of bi-infinite sequences $\mathcal{X}^{\mathbf{Z}}$ (this is the canonical space for two sided Markov chains) with the property that is invariant under shifting θ_n. We also like it to be also invariant under time reversal, this is not always possible, when it does we say the chain is time reversible.

Definition 4.2.1 We define on $\mathcal{X}^{\mathbf{Z}}$ the family $\{\theta_n\}$ of shift maps and the time-reversal map ϱ by

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

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 4.2.2 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 $n \in \mathbf{Z}$; it 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.

4.3 Construction of two sided stationary Markov chains**

We begin be defining a probability measures on $\mathcal{X}^k$ as follows. Given any positive number $n, m > 0$, we define a measure $\mathbf{P}_\pi^{n,m}$ on $\mathcal{X}^{n+m+1}$ in the following way. For $x = (x_{-n}, \dots, x_m)$,

$$\int_{\mathcal{X}^{n+m+1}} f(x_{-n}, \dots, x_m)\, \mathbf{P}_\pi^{n,m}(dx) \tag{4.2}$$

$$=\overbrace{\int_{\mathcal{X}}\cdots\int_{\mathcal{X}}}^{n+m+1} f(x_{m-1},\dots,x_m)\,P(x_{m-1},dx_m)\cdots P(x_{-n},dx_{-n+1})\pi(dx_{-n}).$$

In addition there are $\int_{\mathcal{X}} f(x_0)\mathbf{P}_\pi^{0,0}(dx)=\int_{\mathcal{X}} f(x_0)\pi(dx)$, and similarly $P^{n,0}$ denotes the integration w.r.t. to the coordinates $(x_{-n},\dots,x_0)$ and $P^{0,m}$ denotes integration with respect to the coordinates $(x_0,\dots,x_m)$.

It's worth to have in mind that the canonical process on the measurable space $\mathcal{X}^{\mathbf{Z}}$ with its product σ-algebras is the evaluation of an bi-infinite sequence at a specific time n:

$$(\dots,x_{-2},x_{-1},x_0,x_1,x_2,\dots,)\mapsto x_n.$$

We view $\mathbf{P}_\pi^{n,m}(dz)$ as the finite dimensional probability distribution of the two sided Markov chain, to be constructed.

Theorem 4.3.1 *Let P be transition probabilities with invariant π. Then the measures $\mathbf{P}^{n,m}$ defined by (4.2) are consistent and extends by Kolmogorov's theorem to a measure $\mathbf{P}_\pi$ on $\mathcal{X}^z$. The corresponding Markov chain is called the two sided Markov chain associated with P and π.*

Proof. ** It is an easy, although tedious, exercise to check that the family of measures on $\mathcal{X}^{2n+1}$ defined by (4.2) is consistent, so that it defines a unique measure on $\mathcal{X}^{\mathbf{Z}}$ by Kolmogorov's extension theorem, Theorem 2.2.13. We first recall that π is an invariant measure means if $T\pi=\pi$, i.e. $\int_{\mathcal{X}} P(x,A)\pi(dy)=\pi(A)$ which by the duality relation means that

$$\int_{\mathcal{X}}\int_{\mathcal{X}} f(y)P(z,dy)\pi(dz)=\int_{\mathcal{X}} f(y)\pi(dy).$$

To see the consistency relation more clearly, let us spell out the first cases, the rest can be proved by induction. For any $n\in Z$,

$$\int_{\mathcal{X}^2} f(x_{n+1})\mathbf{P}_\pi^{n,n+1}(dx)\overset{def}{=}\int_{\mathcal{X}}\int_{\mathcal{X}} f(x_{n+1})P(x_n,dx_{n+1})\pi(dx_n)\overset{invariance}{=}\int_{\mathcal{X}} f(x_{n+1})\pi(dx_{n+1})$$
$$\int_{\mathcal{X}^2} f(x_n)\mathbf{P}_\pi^{n,n+1}(dx)=\int_{\mathcal{X}}\left(f(x_n)\int_{\mathcal{X}} P(x_n,dx_{n+1})\right)\pi(dx_n)=\int_{\mathcal{X}} f(x_n)\pi(dx_n).$$

The first equation follows from the invariance of π, the second uses the identity $\int_{\mathcal{X}} P(x_n,dx_{n+1})=1$. We can them move to multiple times:

$$\int_{\mathcal{X}^{n+m+2}} f(x_{-n},\dots,x_m)\,\mathbf{P}_\pi^{n,m+1}(dx)$$
$$=\overbrace{\int_{\mathcal{X}}\cdots\int_{\mathcal{X}}}^{n+m+2} f(x_{-n},\dots,x_m)\,P(x_m,dx_{m+1})P(x_{m-1},dx_m)\dots P(x_{-n+1},dx_{-n})\pi(dx_{-n})$$

$$=\overbrace{\int_{\mathcal{X}}\cdots\int_{\mathcal{X}}}^{n+m+1}\left(\int_{\mathcal{X}}P(x_m,dx_{m+1})\right)f(x_{m-1},\dots,x_m)\,P(x_{m-1},dx_m)\cdots P(x_{-n+1},dx_{-n})\pi(dx_{-n})$$

$$=\overbrace{\int_{\mathcal{X}}\cdots\int_{\mathcal{X}}}^{n+m+1}f(x_{-n},\dots,x_m)\,P(x_{m-1},dx_m)\cdots P(x_{-n+1},dx_{-n})\pi(dx_{-n}).$$

Also,

$$\int_{\mathcal{X}^{n+m+2}}f(x_{-n},\dots,x_m)\,\mathbf{P}_\pi^{n+1,m}(dx)$$

$$=\overbrace{\int_{\mathcal{X}}\cdots\int_{\mathcal{X}}}^{n+m+2}f(x_{-n},\dots,x_m)\,P(x_{m-1},dx_m)\dots P(x_{-n+1},dx_{-n})P(x_{-n-1},dx_{-n})\pi(dx_{-n-1})$$

$$=\int_{\mathcal{X}}\int_{\mathcal{X}}\left(\overbrace{\int_{\mathcal{X}}\cdots\int_{\mathcal{X}}}^{n+m}f(x_{-n},\dots,x_m)\,P(x_{m-1},dx_m)\dots P(x_{-n},dx_{-n+1})\right)P(x_{-n-1},dx_{-n})\pi(dx_{-n-1})$$

$$=\int_{\mathcal{X}}\left(\overbrace{\int_{\mathcal{X}}\cdots\int_{\mathcal{X}}}^{n+m}f(x_{-n},\dots,x_m)\,P(x_{m-1},dx_m)\dots P(x_{-n},dx_{-n+1})\right)\pi(dx_{-n}),$$

the consistency for this case then follows from Fubini's theorem. Further consistency relations will follow by inductions. □

We have the following results:

Lemma 4.3.2 *The measure $\mathbf{P}_\pi$ defined in Theorem 4.3.1 defines a stationary Markov process.*

Proof. It is sufficient to check that it is invariant under θ_1 or θ_{-1}. The defining equation (4.2) is in principle the same as the following

$$\int f(x_{-n},\dots,x_m)\,\mathbf{P}_\pi^{n,m}(dx)=\overbrace{\int_{\mathcal{X}}\cdots\int_{\mathcal{X}}}^{n+m+1}f(x_0,\dots,x_{2n})P(x_{2n-1},dx_{2n})\dots P(x_0,dx_1)\pi(dx_0),$$

(The reason we did not use this as the definition is that we will have to relabel these coordinates for every pair of (n,m) to have them embedded in the bi-infinite sequential space.) It is then trivial to see that $\mathbf{P}_\pi^{(n,m)}=\mathbf{P}_\pi^{(n+1,m+1)}=\mathbf{P}_\pi^{(n-1,m-1)}$: they are defined by the same relations. □

4.4 Reversible Process**

A reversible process is one which looks the same whether time flows forward or backward. It turns out that, for Markov processes, there is an easy criteria that allows to check whether a given process is reversible or not: it is sufficient to work flip two adjacent coordinates and work with the two times marginal $P^{(2)}\pi$ on $\mathcal{X}^2$:

$$\big(P^{(2)}\pi\big)(A \times B) = \int_A P(x, B)\,\pi(dx) = \mathbf{P}(x_0 \in A, x_1 \in B)\;. \tag{4.3}$$

Observe that $\big(P^{(2)}\pi\big)(A \times B)$ is the two time probability distribution of the chain. Let us define $\varrho^{(2)} \colon \mathcal{X}^2 \to \mathcal{X}^2$ by $\varrho^{(2)}(x, y) = (y, x)$.

With this notation, we have

Theorem 4.4.1 *Consider a stationary Markov process (x_n) with transition probabilities P and invariant measure π.*

(1) Suppose that there exist transition probabilities Q such that $(\varrho^{(2)})_*(P^{(2)}\pi) = Q^{(2)}\pi$. Then the process $y_n = x_{-n}$ is also a stationary Markov process, with transition probabilities Q and invariant measure π.

(2) The measure $\mathbf{P}_\pi$ defined in Theorem 4.3.1 defines a reversible Markov process if and only if one has $(\varrho^{(2)})_*(P^{(2)}\pi) = P^{(2)}\pi$, i.e. the two time marginals are invariant under the flipping map.

Remark 4.4.2 Observe that $(\varrho^{(2)})_*(P^{(2)}\pi) = P^{(2)}\pi$ is equivalent to: for every measurable and integrable function $f \colon \mathcal{X}^2 \to \mathbf{R}$,

$$\int_{\mathcal{X}} \int_{\mathcal{X}} f(x, y) P(x, dy)\,\pi(dx) = \int_{\mathcal{X}} \int_{\mathcal{X}} f(x, y) P(y, dx)\,\pi(dy).$$

Similarly, $(\varrho^{(2)})^*(P^{(2)}\pi) = Q^{(2)}\pi$ implies that $P^{(2)}\pi(A \times B) = Q^{(2)}\pi(B \times A)$ and also π is an invariant probability measure for Q.

Proof. For part (2), 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 part (1). since on can take $Q = P$. For part (1), note that the assumption $(\varrho^{(2)})^*(P^{(2)}\pi) = Q^{(2)}\pi$ is just another way of saying that

$$\int_{\mathcal{X}} \int_{\mathcal{X}} f(x, y) P(x, dy)\,\pi(dx) = \int_{\mathcal{X}} \int_{\mathcal{X}} f(x, y) Q(y, dx)\,\pi(dy)\;,$$

for every measurable and integrable function $f\colon \mathcal{X}^2 \to \mathbf{R}$. We apply this to a function on $\mathcal{X}^{n+m+1}$ and flip two consecutive coordinates successively:

$$f(x_{-n}, x_1, \ldots, x_{m-1}, x_m) \to f(x_{-n}, \ldots, x_m, x_{m-1}) \to \cdots \to f(x_m, x_{m-1}, \ldots, x_{-n+1}, x_{-n}).$$

It is then evident that the flipping of 2-coordinates is sufficient to determine the time reversal on $\mathcal{X}^Z$. More precisely we have,

$$\begin{aligned}
&\int f(x_{-n}, \ldots, x_m)\, \mathbf{P}_\pi(dx)\\
&= \overbrace{\int_{\mathcal{X}} \cdots \int_{\mathcal{X}}}^{n+m+1} f(x_{-n}, \ldots, x_m)\, P(x_{m-1}, dx_m) \cdots P(x_{-n}, dx_{1-n}) \pi(dx_{-n})\\
&= \int_{\mathcal{X}} \int_{\mathcal{X}} \left(\overbrace{\int_{\mathcal{X}} \cdots \int_{\mathcal{X}}}^{n+m-1} f(x_{-n}, \ldots, x_m)\, P(x_{m-1}, dx_m) \cdots P(x_{-n+1}, dx_{-n+2}) \right) P(x_{-n}, dx_{1-n}) \pi(dx_{-n})\\
&= \int_{\mathcal{X}} \int_{\mathcal{X}} \left(\overbrace{\int_{\mathcal{X}} \cdots \int_{\mathcal{X}}}^{n+m-1} f(x_{-n}, \ldots, x_m)\, P(x_{m-1}, dx_m) \cdots P(x_{-n+1}, dx_{-n+2}) \right) Q(x_{1-n}, dx_{-n}) \pi(dx_{1-n})\\
&= \overbrace{\int_{\mathcal{X}} \cdots \int_{\mathcal{X}}}^{n+m+1} f(x_{-n}, \ldots, x_m)\, P(x_{m-1}, dx_m) \cdots Q(x_{1-n}, dx_{-n}) P(x_{1-n}, dx_{2-n}) \pi(dx_{1-n})\\
&= \overbrace{\int_{\mathcal{X}} \cdots \int_{\mathcal{X}}}^{n+m+1} f(x_{-n}, \ldots, x_m)\, P(x_{m-1}, dx_m) \cdots Q(x_{1-n}, dx_{-n}) Q(x_{2-n}, dx_{1-n}) \pi(dx_{2-n})\ .
\end{aligned}$$

Proceeding in the same fashion, we finally arrive at

$$\begin{aligned}
&\int f(x_{-n}, \ldots, x_m)\, \mathbf{P}_\pi(dx)\\
&= \overbrace{\int_{\mathcal{X}} \cdots \int_{\mathcal{X}}}^{n+m+1} f(x_{-n}, \ldots, x_m)\, Q(x_{1-n}, dx_{-n}) \cdots Q(x_m, dx_{m-1}) \pi(dx_m)\\
&= \int f(x_{-n}, \ldots, x_m) \left(\varrho^* \mathbf{Q}_\pi\right)(dx)\ ,
\end{aligned}$$

where we denoted by $\mathbf{Q}_\pi$ the law of the stationary Markov process with transition probabilities Q and invariant measure π. Since this holds for every pairs of (n, m) for which the finite dimensional distributions are defined, 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. □

Note that in the case where $\mathcal{X}$ is countable, the condition (4.3) can be written as the detailed balance relation

$$P_{ij}\pi_j = P_{ji}\pi_i \tag{4.4}$$

for every pair i, j. Summing over j in (4.4) or choosing $B = \mathcal{X}$ in (4.3), we see that if there exists a probability measure π such that (4.3) 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 4.4.3 *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*

$$P_{i_1 i_n} P_{i_n i_{n-1}} \cdots P_{i_3 i_2} P_{i_2 i_1} = P_{i_n i_1} P_{i_1 i_2} \cdots P_{i_{n-2} i_{n-1}} P_{i_{n-1} i_n} \tag{4.5}$$

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

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 $n = 3$. If the process is reversible, by the detailed balance rerlation one has

$$P_{i_1 i_3} P_{i_3 i_2} P_{i_2 i_1} \pi_{i_1} = P_{i_1 i_3} P_{i_3 i_2} P_{i_1 i_2} \pi_{i_2} = P_{i_1 i_3} P_{i_2 i_3} P_{i_1 i_2} \pi_{i_3} = P_{i_3 i_1} P_{i_2 i_3} P_{i_1 i_2} \pi_{i_1} \;.$$

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 $i_1, \ldots, i_n$ 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 (4.5) 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

$$\frac{P_{ji}\pi_i}{P_{ij}\pi_j} = \frac{P_{ji}}{P_{ij}} \cdot \frac{P_{i_n i_{n-1}}}{P_{i_{n-1} i_n}} \frac{P_{i_{n-1} i_{n-2}}}{P_{i_{n-2} i_{n-1}}} \cdots \frac{P_{i_2 i_1}}{P_{i_1 i_2}} \cdot \frac{P_{j_{n-1} j_n}}{P_{j_n j_{n-1}}} \frac{P_{j_{n-2} j_{n-1}}}{P_{j_{n-1} j_{n-2}}} \cdots \frac{P_{j_1 j_2}}{P_{j_2 j_1}} \;. \tag{4.6}$$

Since we have $i = i_n$, $j = j_n$, and $i_1 = j_1$, the path $i_1, \ldots, i_n, j_n, \ldots, j_1$ forms a closed loop and the ratio in (4.6) is equal to 1. This shows that the detailed balance relation holds and the process is indeed reversible with respect to π (and therefore that π is its invariant measure). □

Example 4.4.1 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 (with values in $\mathbf{R}$). Define a Markov process on $\mathbf{R}$ by the recursion relation,

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

Since $\alpha x + \beta \xi_1 \sim N(\alpha x, \beta^2)$,

$$P(x,A) = \mathbf{P}(\alpha x + \beta\xi_1 \in A) = \int_A \frac{1}{\sqrt{2\pi\beta^2}} e^{-\frac{(y-\alpha x)^2}{2\beta^2}} dy.$$

Since $x_{n+1} = \alpha x_n + \beta \xi_n$ is distributed as a Gaussian random variable with expectation 0 (if x_n is Gaussian with mean zero) and variance $\alpha \mathbf{E} x_n^2 + \beta^2$. To determine a steady state measure we set $\sigma^2 = \mathbf{E} x_n^2 + \beta^2$, then $\sigma^2 = \frac{\beta^2}{1-\alpha^2}$. It is immediate that $\pi = \mathcal{N}\big(0, \frac{\beta^2}{1-\alpha^2}\big)$ is an invariant measure for this process (in fact it is the only one). Let $x_0 \sim \pi$. The measure $P^{(2)}\pi$ is given by

$$\mathbf{P}(x_0 \in A, x_1 \in B) = \int_A \int_B P(x,dy)\pi(dx) = \int_A \int_B \frac{1}{\sqrt{2\pi\beta^2}} e^{-\frac{(y-\alpha x)^2}{2\beta^2}} \frac{\sqrt{1-\alpha^2}}{\sqrt{2\pi\beta^2}} e^{-\frac{x^2(1-\alpha^2)}{2\beta^2}} dxdy.$$

Then $\mathbf{P}(x_0 \in \mathbf{R}, x_1 \in B) = \pi$, verifying that π is an invariant measure. To summarise,

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

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 deterministic part of the recursion relation $x_{n+1} = \alpha x_n$.

Example 4.4.2 Let $L > 0$ be fixed and let $\mathcal{X}$ be the interval $[0, L]$ with the identification $0 \sim L$ (*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 4.4.3 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 $p\colon V \to (0,1)$. 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 $\mathcal{L}^2(\mathcal{X}, \pi)$.

Chapter 5

Invariant measures in the general case

In this chapter we are concerned with time-homogeneous Markov processes (i.e. Markov chains) on a complete separable metric space $\mathcal{X}$. Recall first of all the following definition:

Definition 5.0.1 A metric space $\mathcal{X}$ is called **separable** if it has a countable dense subset.

Example 5.0.1 Examples of separable spaces are $\mathbf{R}^n$ (The set of points with rational coordinates is a dense subset) and $\mathcal{L}^p(\mathbf{R}^n)$ 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).

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 $T\pi = \pi$. We also defined an operator $T_\star : \mathcal{B}_b(\mathcal{X}) \to \mathcal{B}_b(\mathcal{X})$, the space of bounded measurable functions from $\mathcal{X}$ to $\mathbf{R}$, by

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

The subscript $\star$ will be from time to time omitted.

5.1 Feller and Strong Feller Property

One distinct feature for state space $\mathcal{X}$ that is not countable is that not every function $f : \mathcal{X} \to \mathbf{R}$ is continuous (or measurable).

Definition 5.1.1 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.

If $\mathcal{X}$ is a discrete space, we may use the following distance function

$$d(x,y)=\begin{cases} 1, & \text{if } x\neq y\\ 0, & \text{if } x=y. \end{cases}$$

to describe its power set as the set of all open sets. Indeed any subset of $\mathcal{X}$ is open, close, and Borel measurable. Any functions on discrete space is therefore continuous, and any time homogeneous Markov process is a good process: both Feller property Feller and strong Feller property hold.

Example 5.1.1 (Not Feller) Let $P(x,A)$ be a family of transition probabilities on $\mathbf{R}$ given below

$$P(x,\cdot)=\begin{cases} \delta_1, & \text{if } x>0\\ \delta_0, & \text{if } x\le 0. \end{cases}$$

Then

$$T_\star f(x)=\int_{\mathbf{R}} f(y)P(x,dy)=\begin{cases} f(1) & \text{if } x>0\\ f(0), & \text{if } x\le 0, \end{cases}$$

and $T_\star f$ fails to be continuous at 0 for continuous functions f with $f(1)\neq f(0)$.

Example 5.1.2 Let x_n be a random walk on $\mathbf{R}$ with probability distribution $x_n=x_{n-1}+Y_n$ and Y_n are i.i.d. random variables with probability distribution Γ. Then

$$T_* f(x)=\mathbf{E}f(x+Y_1)=\int_{\mathbf{R}} f(x+y)\Gamma(dy).$$

If $\mathbf{P}(Y=1)=\frac{1}{2}$ and $\mathbf{P}(Y=-1)=\frac{1}{2}$, then $T_*f(x)=\mathbf{E}f(x+Y_1)=\frac{1}{2}f(x+1)+\frac{1}{2}f(x-1)$. Then T_* has Feller property, not strong Feller property.

If Y is standard Gaussian distributed, then $T_*f(x)=\frac{1}{2\pi}\int_{\mathbf{R}} f(y)e^{-\frac{|y-x|^2}{2}}dy$ has the strong Feller property. Indeed this follows from properties of Gaussian densities (parabolic PDE theory) or directly.

The strong Feller property holds more generally when the probability distribution of Y has a density $p(x)$ with respect to the Lebesgue measure, since then $T_\star f(x)=\int f(y)p(y-x)=f*p$. See Lemma 5.1.2 below.

Lemma 5.1.2 *Suppose that $f,g:\mathbf{R}\to\mathbf{R}$ are Borel measurable functions, such that f is bounded and g is Lebesgue integrable, then the convolution $f*g$ is a bounded continuous function. (Recall that $f*g(x)=g*f(x)=\int_{-\infty}^{\infty} f(y)g(x-y)dy$.)*

Proof. (a) First suppose that g is smooth with compact support. Since

$$f * g(x) - f * g(x') = \int f(y)(g(y-x) - g(y-x'))dy,$$

$g(y-x) - g(y-x') \to 0$ when $x \to x'$, and since also $|f(y)g(y-x) - g(y-x'))| \le 2|f|_\infty |g|_\infty$, then the continuity follows from the dominated convergence theorem.

(b) If $g \in L_1(\mathbf{R})$, it can be approximated by continuous functions g_k with compact support. Then $g * f(x_n) - g * f(x_0)$ can be split up into 3 terms,

$$(g * f(x_n) - g_k * f(x_n)) + (g_k * f(x_n) - g_k * f(x)) + (g_k * f(x) - g * f(x)).$$

We use the translation invariance of the Lebesgue measure:

$$\begin{aligned} |g * f(x_n) - g_k * f(x_n)| &= \left| \int_{-\infty}^{\infty} f(y) g(x_n - y) dy - \int_{-\infty}^{\infty} f(y) g_k(x_n - y) dy \right| \\ &\le \int_{\mathbf{R}} |f(y)||g_k(x_n - y) - g(x_n - y)| dy = |f|_\infty \int_{\mathbf{R}} |g_k(y) - g(y)| dy. \end{aligned}$$

Since $g_k \to g$ in L_1, the left hand side converges to zero as $k \to \infty$ uniformly in n. This applies with x_n replaced by x, so for any $\epsilon > 0$, there exists K such that for any $k > K$, and for any n, $|g * f(x_n) - g_k * f(x_n)| + |g * f(x) - g_k * f(x)| < \epsilon/2$. This allows to conclude the first and the last term converges. We then fix a $k > K$, g_k is continuous and has compact support, for n sufficiently large, $|g_k * f(x_n) - g_k * f(x)| < \epsilon/2$ by part (a) of the proof. □

From the Lemma, we see:

Proposition 5.1.3 *Let $T_\star f(x) = \int f(y)p(x-y)dy$ where dy is Lebesgue measure and $p(y)dy$ is probability measure. Then $T_\star$ is strong Feller.*

Let $B_a(x)$ stands for the open ball centred at x with radius a.

Example 5.1.3 Suppose that the transition probabilities have densities with respect to a common measure μ, $P(x, dy) = p(x, y)\mu(dy)$. We suppose also the following conditions:

(1) For every y, $x \mapsto p(x, y)$ is continuous

(2a) For every x there exists $a > 0$ such that $\sup_{x \in B_a(x)} p(x, y)$ is integrable w.r.t. μ.

(Or (2b): for every x, there exists $a > 0$ such that $\{p(z, y), z \in B_a(x)\}$ is uniformly integrable w.r.t. μ).

Then the strong Feller property holds for $T_\star$.

Proof. Let $f : \mathcal{X} \to \mathbf{R}$ be bounded continuous function, and let $x_n \to x$.

$$|T_\star f(x_n) - T_\star f(x)| \le \left| \int f(y)(p(x_n, y) - p(x, y)) \, \mu(dy) \right| .$$

Since $p(x_n, y) \to p(x, y)$ and for x_n near x, $|p(x_n, y) - p(x, y)| \le \sup_{x \in B_a(x)} p(x, y)$ and the latter in L^1, by the dominated convergence theorem, we may take the limit $n \to \infty$ inside the integration sign. Concerning the alternative assumption (2b), uniformly integrability will allow us to take the limit inside the integral. □

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

Theorem 5.1.4 *Let (x_n) 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 $\mathcal{Y}$ and $F \colon \mathcal{X} \times \mathcal{Y} \to \mathcal{X}$. If the function $F(\cdot, \xi_n) \colon \mathcal{X} \to \mathcal{X}$ is continuous for almost every realisation of ξ (If A is the set of y such that $x \mapsto F(x, y)$ is continuous, then the property that $\mathbf{P}(\xi_n \in A) = 1$ does not depend on n.), then the corresponding transition semigroup is Feller.

Proof. Denote by $\hat{\mathbf{P}}$ the law of ξ_n on $\mathcal{Y}$ 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

$$\big(T_\star \varphi\big)(x) = \mathbf{E}\big(\varphi(x_1) \,|\, x_0 = x\big) = \mathbf{E}\varphi(F(x, Y_1)) = \int_\Omega \varphi\big(F(x, y)\big) \, \hat{\mathbf{P}}(dy) \;.$$

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

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

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

If F is continuous in the first variable for each y, then the Markov process is Feller.

5.1.1 Remark and Examples**

This is a remark on Markov process with continuous time parameter. A time homogeneous Markov process $(x_t, t \ge 0)$ on a general state space may not have the strong Markov property.

See Example 5.1.4. On the positive direction, any Feller process has a version with strong Markov property. A Feller process is a Markov process such that its transition operators (which forms a semi-group of linear operators acting on bounded measurable functions) satisfies the following properties: $T_t(C_0) \subset C_0$ and $\lim_{t\to 0} T_t f(x) = f(x)$. Here C_0 denotes the space of continuous functions vanishing at ∞.

A discrete time parameter Markov processes has automatically the strong Markov property, c.f. Theorem 2.5.4. Strong Markov property requires that we can start the process from any point and so in particular we have a family of Markov processes with x_0 distributed as δ_x where $x \in \mathcal{X}$, this essentially demanded the employ of transition probabilities as 'regularized' conditional expectations.

Let us give an example of a Markov process that does not have strong Markov property, the counter example of course has to be continuous time Markov process.

Example 5.1.4 Let us consider a family of transition probability measures on $\mathbf{R}$. For $x \neq 0$, let us take $P_t(x, A)$ to be the transition probability of a Brownian motion on $\mathbf{R}$, so in particular $\lim_{t\downarrow 0} P_t(x, A) = \delta_x(A)$. If $x = 0$, we set $P_t(0, \cdot) = \delta_0(A)$ for all $t \geq 0$. One can show Chapman-Kolmogorov equation holds:

$$\int_{\mathbf{R}} P_s(y, A) P_t(x, dy) = P_{t+s}(x, A), \quad t, s \geq 0, A \in \mathcal{B}(\mathbf{R}), x \in \mathbf{R}.$$

This hods obviously if $x \neq 0$: the value of the integral does not change if we modify the value of the integrand at one point. For $x = 0$, $\int_{\mathbf{R}} P_s(y, A) P_t(0, dy) = P_s(0, A) = \delta_0(A)$, $P_{t+s}(0, A) = \delta_0(A)$. So, if $x \neq 0$, we have a Brownian motion. But when it hits zero (it does in finite time), it gets stuck at 0: from this stopping time, this is no longer a Brownian motion. More precisely, if $\tau = \inf_{t>0}\{x_t = 0\}$, then $x_{\tau+t} = 0$ for all t. The Markov property would require that $x_{t+\tau}$ to behave as a Brownian motion starting from 0.

This Markov process is not Feller!! Let f be a continuous and bounded function, then

$$P_t f(0) = f(0), \qquad P_t f(x) = \int_{\mathbf{R}} f(y) p_t(x, y)\, dy,$$

where p_t is the Gaussian kernel for $N(x, t)$. For $t > 0$, $\lim_{x\to 0} P_t f(x) \neq f(0)$ in general. Take for example $f(y) = y^2$.

There is one more subtlety, suppose we consider any one of the Markov processes (x_t) starting from a point, which denoted by a. Then,

$$\mathbf{P}(x_{n+T} \in A \,|\, \mathcal{F}_T) = \mathbf{P}(x_{n+T} \in A | x_T)$$

obviously holds.

This is a weaker notion, then the strong Markov property meaning the stopped processes is again Markov processes with the same transition probabilities.

Example 5.1.5 Let $\mu_0 = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$, corresponding to coin tossing: let $x_0 = 1$ in case of head; $x_0 = 0$ in case of tail. We consider again a Markov chain with the transition probability in the previous example. Let $\tau = \inf_{t \ge 0}\{x_t = 0\}$. Observe that $\tau = 0$ if $x_0 = 0$. Also $\sigma(x_\tau) = \{\phi, \Omega\}$ since x_τ is a constant. Thus,

$$\mathbf{P}(x_{\tau+n} \in A | \sigma(x_\tau)) = \mathbf{P}(x_{\tau+n} \in A).$$

At time $\tau = 0$, $x_\tau = 0$ we do not know what does the process do without further information. If we look into the history, we immediately know whether $x_0 = 0$ or $x_0 = 1$. In the first case, it will remains zero for all time, in the second case, $x_0(\omega) = 1$, it will continue to behave like a Brownian motion. So history provided more information then the random variable x_τ. This means that Strong Markov property in the 'narrow sense', for a single Markov process with initial distribution fixed, does not hold at τ.

5.2 Metric and topological spaces: a review **

To ease into the next section, we briefly review some of the useful facts concerning metric spaces. This is for self-study only. Let $\mathcal{X}$ be a metric space with distance d. A subset U is open if every point of E is contained in an open ball $B(x, r)$ and $B(x, r) \subset U$. A closed set is the complement of an open set. The closure of a subset A is the intersection of all closed subset of A, it is the complement of the union of all open subsets of A^c. In other words it is the smallest closed set containing A, and is denoted by $\bar{A}$. A sequence x_n is said to converge to x if $d(x_n, x) \to 0$.

Definition 5.2.1 (1) A metric space $\mathcal{X}$ is said to be compact if any cover of it by open sets has a finite sub-covering (The Heine-Borel property).

(b) A subset of a metric space is compact if it is compact as a metric space with the inherited metric.

(c) It is separable if it has a dense countable subset A (dense means $\bar{A} = \mathcal{X}$).

(d) A subset E of $\mathcal{X}$ is relatively compact if its closure $\bar{A}$ is compact).

A metric space is discrete if for every point $x \in \mathcal{X}$ there exists a ball $B(x, r)$ containg no other point (thus every singleton set $\{x\}$ is an open, so is any subset of $\mathcal{X}$). If a metric space is discrete, then the discrete distance function (i.e. the distance between any two distinct points to be 1) defines a metric which is equivalent to the original one. A discrete space is compact if and only if it is finite. A separable discrete space has no more than a countable number of points.

Definition 5.2.2 (a) A metric space E (or its subset) is complete if every Cauchy sequence from it converges to a point in the set. A complete subset of $\mathcal{X}$ is closed.

(b) A metric space is totally bounded if for any $\epsilon > 0$, $\mathcal{X}$ has a finite covering of open balls (or closed balls) of radius ϵ.

A closed subset of a complete metric space is complete, a complete subset of a metric space is closed.

Proposition 5.2.3 *Let K be a subset of a metric space $(\mathcal{X}, d)$. The following are equivalent.*

- *It is compact.*
- *(Bolzano-Weierstrass property) Every sequence from it has a convergent subsequence, the limit is necessarily in K.*
- *It is complete and totally bounded.*

The second property is also called 'sequential compactness'. A subset E of $\mathcal{X}$ is relatively compact if its closure is a compact set. It is equivalent to the property that every sequence from it has a convergent subsequence (the limit does not necessarily belong to E).

Definition 5.2.4 A topological space is a set $\mathcal{X}$ with a collection of subsets, called a topology. Every set from the topology is called an open set. The topology must contain $\mathcal{X}$ and the empty set, and closed under arbitrary unions and finite intersections.

- $\phi \in \mathcal{T}$ 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 $\bigcup_{A \in \mathcal{A}} A \in \mathcal{T}$.

A metric space and its open sets defines a topological space. A topological space $\mathcal{X}$ is metrisable if there exists a metric on $\mathcal{X}$ such that its open sets agree with the topology on $\mathcal{X}$. We can detect the topology by the convergence of sequences. Are there distinct topologies on a space $\mathcal{X}$ such that any sequence converging in one topology also converge in the other? In general yes. However, if a space is metrisable, the topology is determined by convergences of sequences (see Kelley: General Topology), which explains we sometimes only define the concept of convergence, without explicitly mention the topology. The notion of weak convergence of probability measures on a complete separable metric space will be directly linked to the 'weak topology'.

A function between topological spaces is **continuous** if the pre-images of open sets are open sets. We would be interested in the continuity of a real valued function $f : \mathcal{X} \to \mathbf{R}$. On a metric space this concept of continuity agree with the usual continuity: For any $\epsilon > 0$ there exists $\delta > 0$ such that if $d(y, x) < \delta$, $|f(y) - f(x)| < \epsilon$.

5.3 Weak convergence and Prokhorov's theorem

Lemma 5.3.1 (Paratharathy, page 39) *Let μ, ν be measures on a metric space $\mathcal{X}$. If for all bounded real valued uniformly continuous function $f : \mathcal{X} \to \mathbf{R}$,*

$$\int f d\mu = \int f d\nu$$

then $\mu = \nu$.

This lemma is behind the notion of 'weak convergence'.

In fact, the space of probability measures $P(\mathcal{X})$ can be given a topology, called the weak topology. We would not need details of the weak topology, this will be the one we use unless otherwise stated. Recall topology defines the concept of continuity of functions and convergence, this convergence will be described further below.

Definition 5.3.2 A sequence μ_n of probability measures on a topological space $\mathcal{X}$ is said to **converge weakly** to a probability measure μ if

$$\lim_{n\to\infty} \int_{\mathcal{X}} \varphi(x)\, \mu_n(dx) = \int_{\mathcal{X}} \varphi(x)\, \mu(dx)\ , \tag{5.1}$$

for every bounded and continuous function $\varphi\colon \mathcal{X} \to \mathbf{R}$.

Note that the speed of the convergence in (5.1) is allowed to depend on φ. If d is the metric that metrizes $\mathbf{P}(\mathcal{X})$, then μ_n to μ weakly if and only if $d(\mu_n, \mu) \to 0$. Also, weak convergence describes the weak topology. so $\mu_n \to \mu$ means $\mu_n \to \mu$ in the weak topology.

Remark 5.3.3 Suppose that $\mathcal{X}$ is a separable complete metric space. Then the topological space $P(\mathcal{X})$ is metrisable as a separable metric space (e.g. with the Prohorov metric). One can choose this metric such that $P(\mathcal{X})$ is a *separable complete metric space.* Also $P(\mathcal{X})$ is compact if and only if $\mathcal{X}$ is.

The statement that $T_\star$ has the Feller property (or equivalently it preserves the space of bounded continuous functions) holds is equivalent to the statement that $P(x, dy)$ is continuous in the weak topology, which precisely means for any f bounded and continuous,

$$\lim_{n\to\infty} \int f(y) P(x_n, dy) = \int f(y) P(x, dy)$$

whenever $x_n \to x$.

Let $x_0 \in \mathcal{X}$, set $P(x, dy) = \delta_{x-x_0}$. Then $T_\star f(x) = \int_{\mathcal{X}} f(y) P(x, dy) = f(x - x_0)$ is Feller.

Example 5.3.1 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 $\mathcal{X}$.

If $\mathcal{X}$ is a separable complete metric space, a sequence of Dirac measures, δ_{x_n}, converges weakly to an measure μ, it must be a Dirac measure δ_x and $x_n \to x$. This follows from the fact that a probability measure on a separable complete metric space is not a Dirac measure must have at least two points in its support. We use the definition that x is in the support of a measure then any of its neighbourhood (open set containing x) must have positive measure. It is a theorem that the support of a probability measure on a separable complete metric space has full measure. Then μ is a Dirac measure if and only if there is one point in the support of the measure. If $\delta_{x_n} \to \mu$ and μ has two distinct points x, y in its support, we can choose ϵ_n small so that $B_{\epsilon_n}(x)$ and $B_{\epsilon_n}(y)$ are disjoint. We take $\varphi_\epsilon(x)$ that equals 1 on $B_{\epsilon/2}(x)$ and are supported in $B_\epsilon(x)$. Similarly we can take ψ_ϵ which equals 1 on $B_{\epsilon/2}(y)$ and supported on $B_\epsilon(y)$. Then $\int \varphi d\mu \neq 0$ for any functions φ_{ϵ_n} and ψ_{ϵ_n}. Then there are two subsequences of x_n, one of which converges to x, the other to y. This means δ_{x_n} does not converges, giving a contradiction.

Example 5.3.2 Let $\mathcal{X} = \mathbf{R}$. Let $F(x) = \mu((-\infty, x])$ and $F_n(x) = \mu_n((-\infty, x])$. Then $\mu_n \to \mu$ weakly if and only if $F_n(x) \to F(x)$ for all x such that F is continuous at x. If Y_n are random variables distributed as μ_n and Y is distributed as μ and $Y_n \to Y$ in probability then $\mu_n \to \mu$ weakly (i.e. Y_n converges to Y in distribution). The converse does not hold, take for example $Y = c$ a deterministic function and $\mathbf{P}(Y_n = \pm\frac{1}{2}) = \frac{1}{2}$.

For x fixed define:

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

If μ has a limit point then this is potentially an invariant measure.

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'.

By tightness we mean that the measure is tightly packed into a small space, by 'small' we we mean the total mass can be almost packed into a compact set.

Lemma 5.3.4 *If $\mathcal{X}$ is a complete separable metric space, and μ a probability measure. Then for every $\varepsilon > 0$ there exists a compact set $K \subset \mathcal{X}$ such that $\mu(K) \geq 1 - \varepsilon$.*

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. Note that since $\{r_k\}$ is a dense set, one has $\bigcup_{k>0} \mathcal{B}(r_k, 1/n) = \mathcal{X}$ for every n. Fix

$\varepsilon > 0$ and, for every integer $n > 0$, denote by N_n the smallest integer such that

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

Since $\bigcup_{k>0} \mathcal{B}(r_k, 1/n) = \mathcal{X}$, the number N_n is finite for every n. Define now the set K as

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

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. □

Definition 5.3.5 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 (uniformly) **tight** if, for every $\varepsilon > 0$ there exists a compact set $K \subset \mathcal{X}$ such that $\mu(K) \ge 1 - \varepsilon$ for every $\mu \in \mathcal{M}$.

By Lemma 5.3.4, every finite family of probability measures on a complete separable metric space is tight. One can show that: if $\{\mu_n\}$ is a tight sequence of probability measures on a complete separable metric space, then there exists a probability measure μ on $\mathcal{X}$ and a subsequence μ_{n_k} such that $\mu_{n_k} \to \mu$ weakly.

Theorem 5.3.6 (Prohorov) *Let $\mathcal{X}$ be a complete separable metric space. Then a family of probability measures on $\mathcal{X}$ is relatively compact if and only if it is tight.*

Exercise 5.3.1 If $\{\mu_n\} \subset P(\mathcal{X})$ is tight and such that every convergence sub-sequence converges to the same limit, then the sequence converges.

Exercise 5.3.2 Let M be a subset of $\mathcal{P}(\mathbf{R})$. Suppose that there exists a non-decreasing function $\varphi : [0,\infty) \to [0,\infty)$ such that $\lim_{x\to\infty} \varphi(x) = \infty$ and $C = \sup_{\mu\in M} \int_{\mathcal{X}} |\varphi(|x|)\mu(dx) < \infty$, then M is tight.

Proof. Observe that

$$\begin{aligned}\mu(|x| \ge n) &= \int_{|x|\ge n} d\mu \le \int_{\mathcal{X}} \frac{\varphi(x)}{\varphi(n)}\, d\mu = \frac{1}{\varphi(n)} \int_{\mathcal{X}} \varphi(|x|) d\mu \\ &\le \frac{C}{\varphi(n)}.\end{aligned}$$

The quantity on the right hand side is the same for all $\mu \in M$, it converges to 0 uniform in $\mu \in M$, and tightness follows. □

5.4 Existence of Invariant Measures

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

Theorem 5.4.1 (Krylov-Bogoliubov) *Let P be a Feller transition probability on a complete separable metric space $\mathcal{X}$. If there exists $x_0 \in \mathcal{X}$ such that the sequence of measures $\{P^n(x_0, \cdot)\}_{n\geq 0}$ is tight, then there exists an invariant probability measure for P.*

Proof. Fix x_0 as given by the assumptions and let μ_N be the sequence of probability measures defined by

$$\mu_N(A) = \frac{1}{N}\sum_{n=1}^{N} P^n(x_0, A) \; . \tag{5.2}$$

Since our assumption immediately implies that $\{\mu_N\}_{N\geq 1}$ is tight, there exists at least one accumulation point π and a sequence n_k with $n_k \to \infty$ such that $\mu_{n_k} \to \pi$ weakly. Furthermore from

$$TP^n(x_0, \cdot) = \int_{\mathcal{X}} P(y, \cdot)P^n(x_0, dy) = P^{n+1}(x_0, \cdot).$$

To check $T\pi = \pi$, we show that $\int \varphi d(T\pi) = \int \varphi d\pi$ for any $\varphi \in \mathcal{C}_b(\mathcal{X})$. Since T is Feller, $T\varphi$ is a continuous function, since it is also bounded, the dominated convergence theorem can be used:

$$\begin{aligned}\int_{\mathcal{X}} \varphi d(T\pi) &= \int T\varphi d\pi = \lim_{k\to\infty} \int T\varphi d\mu_{n_k} \\ &= \lim_{k\to\infty} \frac{1}{n_k}\sum_{n=1}^{n_k} \int T\varphi \, P^n(x_0, dy) = \lim_{k\to\infty} \frac{1}{n_k}\sum_{n=1}^{n_k} \int \varphi \, P^{n+1}(x_0, dy) \\ &= \lim_{k\to\infty} \int_{\mathcal{X}} \varphi \left(d\mu_{n_k} + \frac{1}{n_k}P^{n_k+1}(x_0, dy) - \frac{1}{n_k}P(x_0, dy)\right) \\ &= \int \varphi d\mu + \lim_{k\to\infty} \frac{1}{n_k}\int_{\mathcal{X}} \varphi \, P^{n_k+1}(x_0, dy) - \lim_{k\to\infty} \frac{1}{n_k}\int_{\mathcal{X}} \varphi \, P(x_0, dy) = \int \varphi d\mu.\end{aligned}$$

Since φ was also arbitrary, this in turn implies that $T\pi = \pi$, *i.e.* that π is an invariant measure for our system. □

As an immediate consequence, we have that

Corollary 5.4.2 *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 5.4.3 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 $x_{n+1} = x_n/2$. Note that $(0,1)$ with the inherited metric is not a complete metric space. Note also $T\varphi(x) = \mathbf{E}(\varphi(x_1)|x_0 = x) = \varphi(\frac{x}{2})$, $P(x,dy) = \delta_{\frac{x}{2}}$. Since x_{n+1} will eventually does not charge any Borel subset of $(0,1)$, the Markov chain 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-Bogoliubov theorem holds is to find a so-called Lyapunov function for the system:

Definition 5.4.4 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:

- There is some value of x for which $V(x)$ is finite.
- For every $a \in \mathbf{R}_+$, the set $V^{-1}\big(\{x \le a\}\big) = \{y : V(y) \le a\}$ is compact.
- There exist a positive constant $\gamma < 1$ and a constant C such that
$$T_\star V(x) = \int_{\mathcal{X}} V(y)\, P(x,dy) \le \gamma V(x) + C\;,$$
for every x such that $V(x) \neq \infty$.

We clarify what does it mean to integrate a function that might take the value $+\infty$. Let $\mathcal{X}_0 = \{x : V(x) < \infty\}$. If μ is a measure on $\mathcal{X}$ with $\mu(\mathcal{X}_0) = 1$, we define $\int_{\mathcal{X}} V d\mu = \int_{\mathcal{X}_0} V d\mu$, otherwise we set $\int_{\mathcal{X}} V d\mu = \infty$. In particular the assumption that $T_\star V(x) \le \gamma V(x) + C$ implies that $P(x,\mathcal{X}_0) = 1$ for every x with $V(x) < \infty$.

With this definition at hand, it is now easy to prove the following results.

Lemma 5.4.5 *Let $V\colon \mathcal{X} \to \mathbf{R}_+ \cup \{\infty\}$ be a Borel measurable function such that there exist a positive constant $\gamma < 1$ and a constant C such that*
$$T_\star V(x) \le \gamma V(x) + C\;,$$
for every x such that $V(x) \neq \infty$. Then
$$T_\star^n V(x) \le \gamma^n V(x) + \frac{C}{1-\gamma}. \tag{5.3}$$

Proof. This is a simple consequence of the Chapman-Kolmogorov equations:
$$T_\star^n V(x) = \int_{\mathcal{X}} V(y)\, P^n(x,dy) = \int_{\mathcal{X}} TV(y) P^{n-1}(x,dy) = \int_{\mathcal{X}}\int_{\mathcal{X}} V(y)\, P(z,dy)\, P^{n-1}(x,dz)$$

$$
\begin{aligned}
&\leq C + \gamma \int_{\mathcal{X}} V(z)\, P^{n-1}(x, dz) \leq \ldots \\
&\leq C + C\gamma + \ldots + C\gamma^n + \gamma^n V(x) \leq \gamma^n V(x) + \frac{C}{1-\gamma}\;.
\end{aligned}
$$

□

Theorem 5.4.6 (Lyapunov function test) *If a transition probability P is Feller and admits a Lyapunov function, then it has an invariant probability measure.*

Proof. Let $x_0 \in \mathcal{X}$ be any point such that $V(x_0) \neq \infty$, we show that the sequence of measures $\{P^n(x_0, \cdot)\}$ is tight. For every $a > 0$, let $K_a = \{y \,|\, V(y) \leq a\}$, a compact set. By the lemma above,

$$T^n V(x_0) = \int_{\mathcal{X}} V P^n(x, dy) \leq \gamma^n V(x) + \frac{C}{1-\gamma}.$$

Tchebycheff's inequality shows that

$$
\begin{aligned}
P^n(x_0, (K_a)^c) = \int_{\{V(y)>a\}} P^n(x_0, dy) &\leq \int_{\{V(y)>a\}} \frac{V(y)}{a} P^n(x_0, dy) \leq \frac{1}{a} T^n V(x_0) \\
&\leq \frac{1}{a}(V(x_0) + \frac{C}{1-\gamma}).
\end{aligned}
$$

We have used Lemma 5.4.5 and the fact that $\gamma < 1$. The results follows from convergence of the right hand side, as $a \to \infty$, with rate uniform in n. (More precisely, for every $\varepsilon > 0$ we can now choose $a \geq \frac{1}{\epsilon}\left(V(x_0) + \frac{C}{1-\gamma}\right)$, then $P^n(x, K_a) \geq 1 - \varepsilon$ for every $n \geq 0$.) We can now use Krylov-Bogoliubov theorem to conclude. □

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 $\int V(x)\, \pi(dx) \leq C/(1-\gamma)$, where C and γ are the constants appearing in (5.3). This is indeed the case, as shown by the following proposition:

Proposition 5.4.7 *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 $C \geq 0$ with*

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

Then, every invariant measure π for P satisfies

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

Proof. Let $M \geq 0$ 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. Let $V_M = V \wedge M$. For every $n \geq 0$, one then has the following chain of inequalities:

$$\begin{aligned}\int_{\mathcal{X}} V_M(x)\,\pi(dx) &= \int_{\mathcal{X}} V_M(x)\,\big(T^n\pi\big)(dx) = \int_{\mathcal{X}} T_\star^n V_M(x)\,\pi(dx) \\ &\leq \int_{\mathcal{X}} (\gamma^n V_M(x) + \frac{C}{1-\gamma})\,\pi(dx)\end{aligned}$$

Since the function on the right hand side is bounded by M, we can apply the Lebesgue dominated convergence theorem. It yields the bound

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

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

5.5 A random dynamical system

In this section let x be a Markov process defined by a recursion relation of the type

$$x_{n+1} = F(x_n, \xi_n)\;, \tag{5.4}$$

for $\{\xi_n\}$ a sequence of i.i.d. random variables taking values in a measurable space $\mathcal{Y}$, and all independent of x_0, and $F\colon \mathcal{X} \times \mathcal{Y} \to \mathcal{X}$ a measurable function.

Theorem 5.5.1 *Suppose that the function $F(\cdot, \xi_n)\colon \mathcal{X} \to \mathcal{X}$ is continuous for almost every realisation of ξ_n. If, furthermore, there exists a Borel measurable function $V\colon \mathcal{X} \to \mathcal{X}$ with compact level sets and constants $\gamma \in (0,1)$ and $C \geq 0$ such that*

$$\int_\Omega V\big(F(x,y)\big)\,\hat{\mathbf{P}}(dy) \leq \gamma V(x) + C\;, \quad \forall x \in \mathcal{X}\;,$$

where $\hat{\mathbf{P}}$ is the distribution of ξ_n, then the process x has at least one invariant probability measure.

Proof. Indeed,

$$P(x,A) = \mathbf{E}(x_1 \in A | x_0 = x) = \mathbf{E}(F(x_0,\xi_0) \in A | x_0 = x) = \int \mathbf{1}_A(F(x,y))\hat{P}(dy).$$

Then P is Feller follows from Theorem 5.1.4. Then the left hand side of the given inequality is TV and V is a Lyapunov function. The existence of an invariant probability measure now follows from the Lyapunov function test. □

5.5.1 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, due to deterministic contraction.

Theorem 5.5.2 *If there exists a constant $\gamma \in (0,1)$ such that*

$$\mathbf{E}d\big(F(x,\xi_1),F(y,\xi_1)\big) \le \gamma d(x,y)\;, \tag{5.5}$$

for every pair x, y in $\mathcal{X}$, then the process (5.4) has at most one invariant probability measure.

Proof. Let π_1 and π_2 be any two invariant measures for (5.4) 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 $x_{n+1} = F(x_n,\xi_n)$ and $y_{n+1} = F(y_n,\xi_n)$ recursively. Since the measures π_i are invariant,, then x_n has distribution π_1, and y_n has distribution π_2 for every n.

We seek a probability measure μ on $\mathcal{X}^2$ with marginals π_i and has mass concentrated on the diagonals, from which to conclude $\pi_1 = \pi_2$. Suppose that such a measure μ exists then

$$\begin{aligned}\pi_1(A) &= \mu(A\times\mathcal{X}) = \mu\big((A\times\mathcal{X})\cap\Delta\big)\\ &= \mu\big((A\times A)\cap\Delta\big) = \mu\big((\mathcal{X}\times A)\cap\Delta\big)\\ &= \pi_2(A)\;,\end{aligned}$$

implying $\pi_1 = \pi_2$. Since the π_1, π_2 were arbitrary invariant probability measures, this shows that there can be only one of them.

This measure μ can be built from a limit point of the joint probability distributions μ_n of (x_n,y_n) on $\mathcal{X}^2$. Define the projection maps $\mathrm{Proj}_i\colon \mathcal{X}^2 \to \mathcal{X}$ by

$$\mathrm{Proj}_1(x,y) = x, \qquad \mathrm{Proj}_2(x,y) = y.$$

We have $\mathrm{Proj}_i^*\mu_n = \pi_i$ for $i = 1,2$ and for every $n \ge 0$, so μ_n is a coupling of π_1 and π_2 for each n. In order to show that the sequence μ_n is tight, fix $\varepsilon > 0$. We know from Lemma 5.3.4 that there exist compact sets K_1 and K_2 in $\mathcal{X}$ such that $\pi_i(K_i) \ge 1-\varepsilon$. Therefore

$$\begin{aligned}\mu_n(\mathcal{X}^2\setminus K_1\times K_2) &\le \mu_n(\mathcal{X}\times(\mathcal{X}\setminus K_2)) + \mu_n((\mathcal{X}\setminus K_1)\times\mathcal{X})\\ &= \pi_2(\mathcal{X}\setminus K_2) + \pi_1(\mathcal{X}\setminus K_1) \le 2\varepsilon\;,\end{aligned}$$

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

$$\int\big(1\wedge d(x,y)\big)\mu(dx,dy) = \lim_{k\to\infty}\int\big(1\wedge d(x,y)\big)\mu_{n_k}(dx,dy) = \lim_{k\to\infty}\mathbf{E}\left(1\wedge d(x_{n_k},y_{n_k})\right)$$

$$
\begin{aligned}
&= \mathbf{E}\left(\mathbf{E}\big(1\wedge\, d(x_{n_k}, y_{n_k})\,|\, x_{n_k-1}, y_{n_k-1}\big)\right)\\
&= \mathbf{E}\left(\mathbf{E}\big(1\wedge\, d(F(x_{n_k-1},\xi), F(y_{n_k-1},\xi)\,|\, x_{n_k-1}, y_{n_k-1}\big)\right)\\
&\le \lim_{k\to\infty} \mathbf{E}\big(1\wedge \gamma d(x_{n_k-1}, y_{n_k-1})\big)\\
&\le \lim_{k\to\infty} \int \big(1\wedge \gamma^{n_k} d(x,y)\big)\mu_0(dx,dy)\ ,
\end{aligned}
$$

where the penultimate inequality is nothing but (5.6) and the last line follows from induction, Detail is as below. We have the inequality for any $a>0$, and any n,

$$
\begin{aligned}
\mathbf{E}\mathbf{E}\big(1\wedge a\, d(x_n,y_n)\,|\,x_{n-1},y_{n-1}\big) &= \mathbf{E}\big(1\wedge a\, d\big(F(x_{n-1},\xi), F(y_{n-1},\xi)\big)\big)\\
&\le 1\wedge a\mathbf{E}\big(d\big(F(x_{n-1},\xi), F(y_{n-1},\xi)\big)\big)\\
&\le 1\wedge a\gamma\, \mathbf{E} d(x_{n-1},y_{n-1})\ .
\end{aligned}
$$

We have applied Jensen inequality to the function $\varphi(x) = 1\wedge ax$. Iterating this bound in the same way as in the proof of Proposition 5.4.7, we obtain

$$
\mathbf{E}\big(1\wedge d(x_n,y_n)\big) \le \mathbf{E}\big(1\wedge \gamma^n d(x_0,y_0)\big)\ . \tag{5.6}
$$

Note now that $1\wedge\gamma^{n_k} d$ converges pointwise to 0 and is bounded by 1, so that Lebesgue's dominated convergence theorem yields

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

so that $\mu(\{(x,y)\,:\,1\wedge d(x,y)\neq 0\}) = 0$. In particular $\mu\{(x,y)\,:\,d(x,y)>1\} = 0$ and $\mu(\{(x,y)\,:\,d(x,y)\neq 0, d(x,y)\le 1\}) = 0$, and so $\mu(\Delta)=1$, where $\Delta = \{(x,x)\,|\,x\in\mathcal{X}\}$ is the 'diagonal' in $\mathcal{X}^2$. Since the Proj_i are continuous, one has again $\mathrm{Proj}_i^*\mu = \pi_i$, so that μ is a coupling of π_1 and π_2. □

5.6 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. The main result in this section is Theorem 5.6.6.

5.6.1 Total Variation distance

We define the **total variation** distance between two measures μ and ν by

$$\|\mu - \nu\|_{\mathrm{TV}} = 2 \sup_{A \in \mathcal{B}(\mathcal{X})} |\mu(A) - \nu(A)| \;,$$

where the supremum runs over all measurable subsets of $\mathcal{X}$. This is equivalent to

$$\|\mu - \nu\|_{\mathrm{TV}} = \sup_{\substack{f \in \mathcal{B}_b(\mathcal{X}) \\ \|f\|_\infty = 1}} \left| \int_{\mathcal{X}} f(x)\,\mu(dx) - \int_{\mathcal{X}} f(x)\,\nu(dx) \right| \;, \tag{5.7}$$

where the maximum is run over bounded measurable functions.

Definition 5.6.1 We say that a sequence $\{\mu_n\}$ converges in total variation to a limit μ if

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

Note that the total variation distance between any two probability measures is smaller or equal to one. If μ and ν are singular, there exists a measurable subset $\mathcal{X}_0$ such that $\mu(\mathcal{X}_0) = 1$ and $\nu(\mathcal{X}_0) = 1$. Then $\|\mu - \nu\|_{TV} \geq 2\|\mu(\mathcal{X}_0) - \nu(\mathcal{X}_0)\| = 2$. It is then easy to see that μ and ν are singular if and only if their total variation distance is the maximum value 2.

Example 5.6.1 Let $\mu_n = \delta_{\frac{1}{n}}$ on $\mathbf{R}$. Then $\mu_n \to \delta_0$ weakly, but not in the total variation norm. In fact the distance $\|\mu_n - \delta_0\|_{TV} = 2$.

Even though it may look at first sight as if convergence in total variation was equivalent to strong convergence, by strong convergence we mean $\lim_{n \to \infty} \mu_n(A) = \mu(A)$ for every measurable set A, this is not true as can be seen in Example 7.2.5 below.

Remark 5.6.2 (1) It is also a fact that under very mild conditions on $\mathcal{X}$ (being a complete separable metric space is more than enough), (5.7) is the same as the seemingly weaker norm,

$$\|\mu - \nu\|_{\mathrm{TV}} = \sup_{\substack{f \in \mathcal{C}_b(\mathcal{X}) \\ \|f\|_\infty = 1}} \left| \int_{\mathcal{X}} f(x)\,\mu(dx) - \int_{\mathcal{X}} f(x)\,\nu(dx) \right| \;, \tag{5.8}$$

where the supremum only runs over continuous bounded functions.

(2) It is also standard to define the total variation distance to be $\frac{1}{2}$ of our total variation distance, i.e. $\|\mu - \nu\|_{\mathrm{TV}} = \sup_{A \subset \mathcal{X}} |\mu(A) - \nu(A)|$. Note also,

$$\sup_{\substack{f \in \mathcal{C}_b(\mathcal{X}) \\ \|f\|_\infty = 1}} \left| \int_{\mathcal{X}} f(x)\,\mu(dx) - \int_{\mathcal{X}} f(x)\,\nu(dx) \right| = \frac{1}{2} \sup_{\substack{f \in \mathcal{C}_b(\mathcal{X}) \\ \mathrm{Oso}(f) = 1}} \left| \int_{\mathcal{X}} f(x)\,\mu(dx) - \int_{\mathcal{X}} f(x)\,\nu(dx) \right|,$$

where $\mathrm{Oso}(f) = \sup(f) - \inf(f)$.

5.6.2 Uniqueness by minorization

Another equivalent definition for the total variation is as follows, this will be the one we use in the formulation. Given two positive measures μ and ν on a measurable space Ω, let $\eta = \mu + \nu$. Then it is easy to check that both μ and ν are absolutely continuous with respect to $\mu + \nu$(we denote this by $\mu << \eta$ and $\nu << \eta$). We denote by $\frac{d\mu}{d\eta}$ and $\frac{d\nu}{d\eta}$ their Radon-Nikodym derivatives with respect to the measure η. Then,

$$\|\mu - \nu\|_{\mathrm{TV}} \equiv \int_\Omega \left|\frac{d\mu}{d\eta} - \frac{d\nu}{d\eta}\right| \, d\eta \; . \tag{5.9}$$

If η is any arbitrary positive measure on Ω such that both $\mu \ll \eta$ and $\nu \ll \eta$, then one has also

$$\|\mu - \nu\|_{\mathrm{TV}} = \int_\Omega \left|\frac{d\mu}{d\eta} - \frac{d\nu}{d\eta}\right| \, d\eta \; . \tag{5.10}$$

In other words this formulation does not depend on the choice of reference measure. This follows immediately from the fact that in this case one has

$$\frac{d\mu}{d\eta} = \frac{d\mu}{d(\mu+\nu)}(\frac{d\mu}{d\eta} + \frac{d\nu}{d\eta}), \qquad \frac{d\nu}{d\eta} = \frac{d\nu}{d(\mu+\nu)}(\frac{d\mu}{d\eta} + \frac{d\nu}{d\eta})$$

and therefore

$$|\frac{d\nu}{d\eta} - \frac{d\mu}{d\eta}| = \left|\frac{d\nu}{d(\mu+\nu)} - \frac{d\mu}{d(\mu+\nu)}\right|(\frac{d\mu}{d\eta} + \frac{d\nu}{d\eta}).$$

Remark 5.6.3 This means that the total variation distance is given by the L_1 norm of the Radon-Nikodym derivatives.

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

$$\big(\mu \wedge \nu\big)(A) = \int_A \min\left\{\frac{d\mu}{d\eta}, \frac{d\nu}{d\eta}\right\} \, d\eta, \qquad \eta = \mu + \nu$$

Since, for any two positive numbers, one has $|x - y| = x + y - 2\min\{x, y\}$, we have

$$\left|\frac{d\mu}{d\eta} - \frac{d\nu}{d\eta}\right| = \frac{d\mu}{d\eta} + \frac{d\mu}{d\eta} - 2\frac{d\mu}{d\eta} \wedge \frac{d\nu}{d\eta}.$$

The definition (5.9) immediately implies that if μ and ν are two probability measures, one has

$$\|\mu - \nu\|_{\mathrm{TV}} = 2 - 2\big(\mu \wedge \nu\big)(\Omega) \; . \tag{5.11}$$

Exercise 5.6.1 Show that if η is an positive measure on Ω such that $\mu \ll \eta$ and $\nu \ll \eta$,

$$\big(\mu \wedge \nu\big)(A) = \int_A \min\left\{\frac{d\mu}{d\eta}, \frac{d\nu}{d\eta}\right\} \, d\eta \; .$$

Note also that

Lemma 5.6.4 *The space of probability measures on Ω endowed with the total variation distance $\|\cdot\|_{\mathrm{TV}}$ 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 $\eta=\sum_{n=1}^{\infty}\frac{1}{2^n}\mu_n$. Then $\mu_n << \eta$ for each n. By (5.10), the total variation distance is equal to the $\mathcal{L}^1$ distance between the corresponding Radon-Nikodym derivatives. Thus $\{\frac{d\mu_n}{d\eta}\}$ is a Cauchy sequence in $L^1(\eta)$: indeed,

$$\Big|\frac{d\mu_m}{d\eta}-\frac{d\mu_n}{d\eta}\Big|_{L^1(\eta)}=\int_{\mathcal{X}}\Big|\frac{d\mu_m}{d\eta}-\frac{d\mu_n}{d\eta}\Big|\,d\eta=\Big\|\mu^m-\mu^n\Big\|_{TV}.$$

The result thus follows from the completeness of $\mathcal{L}^1(\Omega,\eta)$ and so $\frac{d\mu_n}{d\eta}$ converges to a density function $\bar f:\mathcal{X}\to\mathbf{R}_+$ (i.e. an L^1 function), and μ_n converges to the measure $\bar\mu$ given by $\bar\mu(A)=\int_A \bar f d\eta$. □

Lemma 5.6.5 *Let μ,ν be two probability measures on $\mathcal{X}$. Let P be a transition probability.*

(1) The following inequality always holds:

$$\|T\mu-T\nu\|_{\mathrm{TV}}\le\|\mu-\nu\|_{\mathrm{TV}}. \tag{5.12}$$

(2) For any $n\in\mathbf{N}$,

$$\|T^n\mu-T^n\nu\|_{\mathrm{TV}}=\frac{1}{2}\|\mu-\nu\|_{\mathrm{TV}}\cdot\|T^n\bar\mu-T^n\bar\nu\|_{\mathrm{TV}} \tag{5.13}$$

where

$$\bar\mu=\frac{\mu-\mu\wedge\nu}{\frac{1}{2}\|\mu-\nu\|_{\mathrm{TV}}},\qquad \bar\nu=\frac{\nu-\mu\wedge\nu}{\frac{1}{2}\|\mu-\nu\|_{\mathrm{TV}}}.$$

Proof. For consistency, we use the formulation for the total variation norm defined by (5.10). Since $\mu\wedge\nu\le\mu$, we may define the 'compensator' of μ with respect to $\mu\wedge\nu$ to be $\mu-\mu\wedge\nu$. This is a positive measure. Since for $\mu\ne\nu$,

$$1-\big(\mu\wedge\nu\big)(\mathcal{X})=\frac{1}{2}\|\mu-\nu\|_{\mathrm{TV}}\ne 0,$$

we may normalise it to be a probability measure and obtain

$$\bar\mu(A)=\frac{\mu(A)-\mu\wedge\nu(A)}{1-\mu\wedge\nu(\mathcal{X})}.$$

Similarly, $\bar{\nu}(A) = \frac{\nu(A)-\mu\wedge\nu(A)}{1-\mu\wedge\nu(\mathcal{X})}$ defines a probability measure. We have

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

One then has

$$\|T^n\mu - T^n\nu\|_{\mathrm{TV}} = \frac{1}{2}\|\mu-\nu\|_{\mathrm{TV}} \cdot \|T^n\bar{\mu} - T^n\bar{\nu}\|_{\mathrm{TV}}\;.$$

Since the total variation distance between two probability measures can never exceed 2, in particular $\|T^n\bar{\mu} - T^n\bar{\nu}\|_{\mathrm{TV}} \le 2$, then (2) follows by taking $n = 1$. □

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

Theorem 5.6.6 *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 $P(x,\cdot\,) \ge \alpha\eta$ for every $x \in \mathcal{X}$.*

(1) Then, P has a unique invariant probability measure π.

(2) Furthermore for any $\mu,\nu \in P(\mathcal{X})$, $\|T^{n+1}\mu - T^{n+1}\nu\|_{\mathrm{TV}} \le (1-\alpha)^n\|\mu-\nu\|_{\mathrm{TV}}$.

Proof. For any measure $m \in P(\mathcal{X})$,

$$Tm = \int_{\mathcal{X}} P(x,\cdot)dm \ge \alpha\eta.$$

Thus,

$$Tm = \alpha\eta + (1-\alpha)\,\frac{Tm-\alpha\eta}{1-\alpha}$$

and $\frac{Tm-\alpha\eta}{1-\alpha}$ is a probability measure. If μ,ν are two probability measures,

$$\|T\mu - T\nu\|_{TV} \le 1-\alpha)\left\|\frac{T\mu-\alpha\eta}{1-\alpha} - \frac{T\nu-\alpha\eta}{1-\alpha}\right\|_{TV} < 2(1-\alpha), \tag{5.14}$$

We have used again the fact that the total variation distance between two probability measures can never exceed 2. We then use the identity (5.13) and apply (5.14) to the two probability measures μ,ν,

$$\begin{aligned}\|T\mu - T\nu\|_{\mathrm{TV}} &= \frac{1}{2}\|\mu-\nu\|_{\mathrm{TV}} \cdot \|T\bar{\mu} - T\bar{\nu}\|_{\mathrm{TV}} \\ &\le (1-\alpha)\|\mu-\nu\|_{\mathrm{TV}},\end{aligned}$$

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

Corollary 5.6.7 *Assume the conditions of Theorem 5.6.6. Let μ be any probability measure, and π the invariant probability measure. Then*

$$\|T^n\mu - \pi\|_{\mathrm{TV}} = \|T^n\mu - T^n\pi\|_{\mathrm{TV}} \le 2(1-\alpha)^{n-1}\;,$$

The conclusions of the theorem holds if we weaken the condition $P(x,\cdot\,) \geq \alpha\eta$ is weakened to $P^{n_0}(x,\cdot\,) \geq \alpha\eta$ for some n_0. There is also exponential convergence to the equilibrium (with a different exponential rate). More precisely,

Remark 5.6.8 Let P be a transition probability on a space $\mathcal{X}$. Assume that there exists $\alpha > 0$, $n_0 \in \mathbf{N}$, and a probability measure η on $\mathcal{X}$ such that

$$P^{n_0}(x,\cdot\,) \geq \alpha\eta$$

for every $x \in \mathcal{X}$. Then, P has a unique invariant probability measure π.

The proof is almost identical to the earlier theorem. For any $m \in P(\mathcal{X})$,

$$T^{n_0}m = \int_{\mathcal{X}} P^{n_0}(x,\cdot)d\mu \geq \alpha\eta.$$

Write

$$T^{n_0}m = \alpha\eta + (1-\alpha)\,\frac{T^{n_0}m - \alpha\eta}{1-\alpha}.$$

Observe that $\bar{m} := \frac{T^{n_0}m-\alpha\eta}{1-\alpha}$ is a probability measure. (Thus any two probability measures, becomes non-singular after an evolution of time n_0.) Apply Lemma 5.6.4, and use the notation there, we obtain

$$\begin{aligned}\|T^{n_0}\mu - T^{n_0}\nu\|_{\mathrm{TV}} &= \frac{1}{2}\|\mu-\nu\|_{\mathrm{TV}} \cdot \|T^{n_0}\bar{\mu} - T^{n_0}\bar{\nu}\|_{\mathrm{TV}} \ .\\ &= (1-\alpha)\|\mu-\nu\|_{\mathrm{TV}} \cdot \frac{1}{2}\left\|\frac{T^{n_0}\bar{\mu}-\alpha\eta}{1-\alpha} - \frac{T^{n_0}\bar{\nu}-\alpha\eta}{1-\alpha}\right\|_{\mathrm{TV}} \leq (1-\alpha)\|\mu-\nu\|_{\mathrm{TV}}.\end{aligned}$$

Thus T^{n_0} is a contraction. If $n, m \geq n_0k$ where $k \in \mathbf{N}$, we use the property that T does not increase the total variation norm (Lemma 5.6.4),

$$\begin{aligned}\|T^n\mu - T^m\nu|_{TV} &\leq \|T^{n_0k}T^{n-n_0k}\mu - T^{n_0k}T^{m-n_0k}\nu\|_{\mathrm{TV}}\\ &\leq (1-\alpha)^k\|T^{n-n_0k}\mu - T^{n-n_0k}\nu\|_{\mathrm{TV}} \leq 2(1-\alpha)^k\ ,\end{aligned}$$

So $T^n\mu$ is a Cauchy sequence and converges to a probability measure π, by the completeness of $\mathbf{P}(\mathcal{X})$. Since T is continuous, we see that π is an invariant measure. The uniqueness follows from the contraction.

5.7 Invariant sets

We have seen the uniqueness of invariant probability measure due to the deterministic contraction (5.5). There are situations (we will see one of them immediately) where (5.5),i. e. $\mathbf{E}d\big(F(x,\xi_1), F(y,\xi_1)\big) \leq \gamma d(x,y)$, 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.

Definition 5.7.1 Let P be a family of transition probabilities on $\mathcal{X}$. A Borel set A is said to be P-**invariant** if $P(x, A) = 1$ for every $x \in A$.

Remark 5.7.2 If A is a P-invariant set and π an invariant probability measure, and $x_0 \sim \pi$, then

$$\mathbf{P}(x_0 \in A, x_1 \in A, \dots, x_n \in A) = \pi(A).$$

This is a consequence of the Chapman-Kolmogorov equations:

$$\begin{aligned}&\mathbf{P}(x_0 \in A_0, x_1 \in A_1, \dots, x_n \in A_n)\\ &= \textstyle\int_{A_0} \int_{A_1} \cdots \int_{A_{n-1}} P(x_{n-1}, A_n) P(x_{n-2}, dx_{n-1}) \cdots P(x_1, dx_2) P(x_0, dx_1) \mu(dx_0).\end{aligned}$$

If P_π is the stationary measure defining the two-sided stationary process, c.f. Theorem 4.3.1, and A^Z the infinite product of A by itself, then $P_\pi(A^Z) = \pi(A)$. This is related to Corollary 5.9.11.

If there exists a closed P-invariant set $\mathcal{A} \subset \mathcal{X}$, then one can restrict the original Markov process to a process on $\mathcal{A}$. It then suffices to check (5.5) for x and y in $\mathcal{A}$ to conclude that the process has a unique invariant measure π in $\mathcal{P}(\mathcal{A})$.

Remark 5.7.3 Suppose that $\mathcal{A}$ is an invariant set for a time homogeneous transition probability P, Since $\mathcal{A}$ is invariant, we have a Markov chain restricted to $\mathcal{A}$. If π is an invariant measure for the restricted chain,

$$\int_{\mathcal{A}} P(x, A) \pi(dx) = \pi(A).$$

Since π is supported in $\mathcal{A}$, this is the same as

$$\int_{\mathcal{X}} P(x, A) \pi(dx) = \pi(A),$$

Then $\tilde{\pi}(B) = \pi(B \cap \mathcal{A})$ defines an invariant probability measure for the original P on $\mathcal{X}$. And vice versa, if $\tilde{\pi}$ is an invariant probability measure supported on $\mathcal{A}$, it is an invariant probability measure for the restricted Markov chain.

To show the Markov chain on $\mathcal{X}$ has a unique invariant probability measure, 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

$$\mathcal{A}_0 = \mathcal{A}\;, \quad \mathcal{A}_{n+1} = \big\{x \in \mathcal{X} \,|\, P(x, \mathcal{A}_n) > 0\big\}\;. \tag{5.15}$$

Observe that $A_0 \subset \mathcal{A}_1$ since $\mathcal{A}_0$ is invariant. In fact

$$\mathcal{A}_0 \subset \mathcal{A}_1 \subset \mathcal{A}_2 \subset \dots.$$

By induction assume that $\mathcal{A}_{n-1} \subset \mathcal{A}_n$, and by the definition we have $\mathbf{P}(x, \mathcal{A}_{n-1}) > 0$, consequently

$$P(x, \mathcal{A}_n) \geq \mathbf{P}(x, \mathcal{A}_{n-1}) > 0, \qquad \forall x \in \mathcal{A}_n. \tag{5.16}$$

With these definitions, we have

Lemma 5.7.4 *For every $n \geq 1$, $P^n(x, \mathcal{A}) > 0$ for every $x \in \mathcal{A}_n$.*

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

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

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 $\mathcal{A}_k$. □

Proposition 5.7.5 *Let $\mathcal{A}$ be an invariant set for P and let $\mathcal{A}_n$ be defined as in (5.15). Suppose that $\bigcup_{n\geq 0} \mathcal{A}_n = \mathcal{X}$, then every invariant probability measure π for P is in $\mathcal{P}(\mathcal{A})$.*

Proof. Suppose now that $\pi(\mathcal{A}) < 1$. Since $\pi(\bigcup_{n\geq 0} \mathcal{A}_n) = \pi(\mathcal{X}) = 1$, by the assumption, $\lim_{n\to\infty} \pi(\mathcal{A}_n) = 1$. 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

$$\int_{\mathcal{A}} P^n(x, \mathcal{A})\, \pi(dx) = \int_{\mathcal{A}} \pi(dx) = \pi(\mathcal{A}),$$

(since $\mathcal{A}$ is an invariant set and so $P^n(x, \mathcal{A}) = 1$) and we used $\pi(\mathcal{A}_n \setminus \mathcal{A}) > 0$ and $P^n(x, \mathcal{A}) > 0$ for every $x \in \mathcal{A}_n$, c.f. (5.16). This is a contradiction, so that one must have $\pi(\mathcal{A}) = 1$. □

Corollary 5.7.6 *Suppose that $\mathcal{A}$ is an invariant set for a time homogeneous transition probability P. If $\mathcal{A}$ is compact and P is Feller, there exists an invariant probability measure for P.*

Proof. Since $\mathcal{A}$ is invariant, we have a Markov chain restricted to $\mathcal{A}$. If π is an invariant measure for the restricted chain, it extends to $\tilde{\pi}$, an invariant probability measure for the original Markov chain. If $\mathcal{A}$ is furthermore compact, then every continuous function on $\mathcal{A}$ extends to a bounded continuous function on $\mathcal{X}$ (Tietze's theorem) and the restricted chain is Feller. Krylov-Bogoliubov theorem applies to give an invariant measure π for the transition probability restricted to $\mathcal{A}$, and therefore for P itself. □

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

Proposition 5.7.7 *Let x be the Markov process on $(0,\infty)$ such that x_{n+1} is given by the solution at time 1 to the differential equation*

$$\frac{dx}{dt} = \frac{1}{x(t)} - 2 + \xi_n(t)\;, \quad x(0) = x_n\;, \tag{5.17}$$

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. Let $f \in C([0,1];\mathbf{R})$ and denote by $\varphi(t,x,f)$ the solution map to

$$\frac{dx}{dt} = \frac{1}{x(t)} - 2 + f(t)\;, \quad x(0) = x,$$

so $\varphi(0,x,f) = x$, and $\varphi(t,x,f)$ solves the equations. Let $\Phi(x,f) = \varphi(1,x,f)$, then for x_0 a random variable independent of $\{\xi_i\}$, x_n are defined recursively by the formula:

$$x_{n+1}(\omega) = \Phi(x_n(\omega), \xi_n(\omega)).$$

Denote furthermore by $\Phi_+ = \Phi(\cdot,1)$ the map that solves the equation with $f(t) = 1$ for all t and by $\Phi_- = \Phi(\cdot,-1)$ the map that solves the equation with $f(t) = -1$ for all t. Then, a standard comparison argument shows that $x_{n+1} \in [\Phi_-(x_n), \Phi_+(x_n)]$ almost surely.

Observe that $\dot x = \frac{1}{x} - 1$ has only one equilibrium point $x = 1$ as $t \to \infty$, and its solution with any initial value converges to 1. Similarly the solutions to $\dot x = \frac{1}{x} - 3$ converges to $\frac{1}{3}$.

Fix $\varepsilon > 0$, and define $\mathcal{A} = \mathcal{A}_0 = [1/3 - \varepsilon, 1 + \varepsilon]$. Set $\mathcal{A}_{n+1} = \bigl\{x \in \mathcal{X} \,|\, P(x, \mathcal{A}_n) > 0\bigr\}$. 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. (Observe that $\Phi_-^{-1}(1/3 - \varepsilon) < \frac{1}{3} - \varepsilon$, and $\Phi_+^{-n}(1 + \varepsilon > 1 + \epsilon$, in particular $\mathcal{A}$ is an invariant set.)

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 5.7.5 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 (5.17). 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 (Theorem 5.1.4). Since $[1/3, 1]$ is compact, this in turn implies that it has at least one invariant probability measure (c.f. Corollary 5.7.6). Furthermore, one has $|\Phi'(x, \xi)| \le 1/e < 1$ for every $x \le 1$, so that Theorem 5.5.2 applies, we conclude that there exists exactly one invariant probability measure for the Markov chain. □

5.7.1 A useful theorem for ODE**

Let $g : \mathbf{R}^n \to \mathbf{R}^n$ and $f : \mathbf{R}_+ \to \mathbf{R}$ be measurable functions. Suppose that g is locally Lipschitz continuous. Consider $\dot{x}(t) = g(x(t)) + f(t)$.

(1) Suppose that f is locally bounded. Then for every initial point x_0, there exists a unique local solution $\varphi(t, x_0)$, by which we mean it solves the equation on a time interval $[0, \delta]$.

(2) *Non-explosion/ completeness.* Suppose that f is bounded and g satisfies the one sides linear growth condition. The latter means that there exists c such that

$$\langle x, g(x)\rangle \le C(1 + |x|^2).$$

Then every solution is global and $|\varphi(t, x)|^2 \le c|x|^2 e^{ct}$ for some constant c.

(3) *Existence of a global solution flow.* Suppose that there exists a number K such that for every pair of $x, y \in \mathbf{R}^n$

$$|g(x) - g(y)| \le K|x - y|.$$

Then for every f, for which there exists a global solution from every starting point, $x \mapsto \varphi(t, x)$ is continuous for every time $t \ge 0$. Furthermore,

$$|\varphi(t, x) - \varphi(t, y)| \le K|x - y|e^{Kt}.$$

Proof. Part (1) is covered by Piccard's theorem (local version).

Part (2). By the chain rule $\frac{d}{dt}|\varphi(t, x_0)|^2 = 2\langle\varphi(t, x_0), \frac{d}{dt}\varphi(t, x_0)\rangle$ and so,

$$\begin{aligned} |\varphi(t, x_0)|^2 =& |x_0|^2 + \int_0^t \langle x_s, g(x(s))\rangle ds + \langle f(t), x(t)\rangle \\ \le& |x_0|^2 + \int_0^t c(1 + |x_s|^2) ds + \frac{1}{2}|f(t)|^2 + \frac{1}{2}|x(t)|^2. \end{aligned}$$

Re-arrange,

$$|\varphi(t,x_0)|^2 \le 2|x_0|^2 + 2ct + 2c\int_0^t |x_s|^2 ds + |f(t)|^2.$$

Hence by Gronwall's inequality,

$$|\varphi(t,x_0)|^2 \le (2|x_0|^2 + 2ct + |f(t)|^2)e^{2ct}.$$

(3) Let $x_0 \in \mathbf{R}^n$ and U an open set containing x_0 and suppose that for $t \le \delta$, the solutions $\varphi(t,x)$ are defined for every $x \in U$. Then from

$$\varphi(t,x) = x + \int_0^t f(\varphi(s,x))ds + f(t), \qquad \varphi(t,y) = y + \int_0^t f(\varphi(s,y))ds + f(t),$$

we see that

$$\begin{aligned}|\varphi(t,x) - \varphi(t,y)| \le& |x-y| + \int_0^t |f(\varphi(s,x)) - f(\varphi(s,y))|\, ds\\ \le& |x-y| + K\int_0^t |\varphi(s,x) - \varphi(s,y)|ds.\end{aligned}$$

Thus, $|\varphi(t,x) - \varphi(t,y)| \le K|x-y|e^{Kt}$. □

Exercise 5.7.1 Check that, if we replace f by a random variable ξ with values in $C(\mathbf{R}_+, \mathbf{R})$, and $\mathbf{E}|\xi|^2 < \infty$ then for almost surely every ξ there exists a global solution.

If $\varphi(t,x)$ is a global smooth flow for the ODE, let $v_t = d\varphi(t,x_0)(v_0)$ denotes its derivative in the direction v_0 at x_0. Then v_t solves

$$dv_t = \langle \text{ grad} f(x_t), v_t\rangle.$$

5.8 Structure theorem for invariant measures

In this section, we introduce 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 5.8.1 If T is not Feller, it is not true in general that $\mathcal{I}(T)$ is closed. Choose for example an arbitrary measure $\mu \neq \delta_0$ 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 $\delta_0 \notin \mathcal{I}(T)$.

5.8.1 Ergodic theory for dynamical systems

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

Recall that a **dynamical system** consists of a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ 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 $A \in \mathcal{F}$. We will denote as usual by $\mathbf{E}$ expectations with respect to $\mathbf{P}$.

Definition 5.8.2 Given such a dynamical system, a set with $\theta^{-1}(A) = A$ is called an invariant set We define $\mathcal{I} \subset \mathcal{F}$ as the set of subsets such that $\theta^{-1}(A) = A$. It is called the Invariant σ-algebra.

It is clear that $\mathcal{I}$ is again a σ-algebra. In order to emphasise the invariance with respect to θ, we may refer an invariant set as a θ-invariant set.

Definition 5.8.3 A measurable function $f : \Omega \to \mathbf{R}$ is said to be θ-invariant (or simply invariant) if $f \circ \theta = f$.

Exercise 5.8.1 Let $f : \Omega \to \mathbf{R}$ be an $\mathcal{F}$- measurable function. Then f is invariant if and only if f is measurable with respect to the invariant σ-algebra $\mathcal{I}$.

Hint. $\mathbf{1}_A(\theta\omega) = \mathbf{1}_{\theta^{-1}}(\omega)$.

The perhaps most famous result in the theory of dynamical systems is

Theorem 5.8.4 (Birkhoff's Ergodic Theorem) *Let* $(\Omega, \mathcal{F}, \mathbf{P}, \theta, \mathcal{I})$ *be as above and let* $f \colon \Omega \to \mathbf{R}$ *be such that* $\mathbf{E}|f| < \infty$. *Then,*

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

almost surely.

Definition 5.8.5 A dynamical system is said to be **ergodic** if all sets in $\mathcal{I}$ have either measure 0 or measure 1. Note that this is a property of the map θ as well as of the measure $\mathbf{P}$.

Proposition 5.8.6 *The following statements are equivalent.*

1. *θ is ergodic;*

2. *every invariant integrable function f is almost surely a constant.*

3. *every invariant bounded function is almost surely a constant.*

Furthermore, the constant is $\mathbf{E}f$.

Proof. Firstly we show that (1) implies (2). Suppose that θ is ergodic. Let f be an integrable invariant function, we prove that $f = \mathbf{E}f$. Define the sets $A_+ = \{\omega \in \Omega \,|\, f(\omega) > \mathbf{E}f\}$, $A_- = \{\omega \in \Omega \,|\, f(\omega) < \mathbf{E}f\}$, and $A_0 = \{\omega \in \Omega \,|\, f(\omega) = \mathbf{E}f\}$. All three sets belong to $\mathcal{I}$ 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

$$\mathbf{E}f = \int_{\mathcal{X}} f(\omega)\mathbf{P}(d\omega) = \imath = \int_{A_+} f d\mathbf{P} = \int_{A_+} (f - \mathbf{E}f) d\mathbf{P} + \int_{\mathcal{X}} \mathbf{E}f d\mathbf{P} = \int_{A_+} (f - \mathbf{E}f) d\mathbf{P} + \mathbf{E}f,$$

But $\int_{A_+} (f - \mathbf{E}f) d\mathbf{P} > 0$ which leads to a contradiction. This proves that $\mathbf{P}(A_+) = 0$ and and similarly for A_-. It remains the only possibility that $\mathbf{P}(A_0) = 1$, and so $\mathbf{P}(f = \mathbf{E}f) = 1$.

From (2) to (3) is trivial. Assume that (3) holds. Let $f = \mathbf{1}_A$ where A is an invariant set. Then $\mathbf{1}_A = \mathbf{P}(A)$ almost surely. Hence $\mathbf{P}(A) = 1$ or 0, and thus θ is ergodic. □

The limit function $\bar{f} \equiv \mathbf{E}\big(f \,|\, \mathcal{I}\big)$ in Birkhoff's ergodic theorem is $\mathcal{I}$-measurable. This leads to the following corollary.

Corollary 5.8.7 *With the notations of Theorem 5.8.4, 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.

5.9 Dynamical system induced by Markov chains

Given a family of transition probability measures and an invariant probability measure π for it, we can construct a two sided Markov chain $(x_n, n \in Z)$, which defines a measure $\mathbf{P}_\pi$ on the space $\mathcal{X}^{\mathbf{Z}}$ of $\mathcal{X}$-valued sequences. See §4.3. We furthermore defined the shifts θ_n on $\mathcal{X}^{\mathbf{Z}}$ by

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

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

Lemma 5.9.1 *The triple* $(\mathcal{X}^{\mathbf{Z}}, \mathcal{B}(\mathcal{X}^Z), \mathbf{P}_\pi, \theta)$ *defines a continuous dynamical system.*

Proof. It is clear that θ is continuous. It was already checked in Lemma 4.3.2 that $\mathbf{P}_\pi$ defines a stationary process, *i.e.* that it is invariant under θ. □

Remember also 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 $\mathcal{I}$ the set of all measurable subsets of $\mathcal{X}^{\mathbf{Z}}$ that are invariant under θ.

Definition 5.9.2 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 simplest examples of invariant sets are: (1) sets contains constant sequences. (2) A consists of limit cycles, e.g. $A = \{\underline{a}, \underline{b}, \underline{c}\}$, where for $a, b, c \in \mathcal{X}$, $\underline{a} = (\dots, a, b, c, a, b, c, \dots,)$ $\underline{b} = (\dots, b, c, a, b, c, a, \dots)$ and $\underline{c} = (\dots, c, a, b, c, a, b, \dots)$. Also let x_n be random variables on a state space $\mathcal{X}$, B a Borel measurable set, then $A = \{\omega : x_n(\omega) \text{ eventually in } B\}$ is an invariant set. If x_n is a Markov chain on $\{1, \dots, N\}$, the event that the chain starts from state 2 eventually ends at state 1 is an invariant set.

Example 5.9.1 Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space. Let (x_n) be a Markov chain with stationary distribution and with state space $\mathcal{X}$ countable. Let $\varphi : \mathcal{X}^{\mathbf{N}_+} \to \mathbf{R}$ be a bounded invariant function. Define

$$Y = \varphi(x_0, x_1, \dots).$$

By the invariance, for every $n \geq 1$,

$$Y = \varphi(x_n, x_{n+1}, \dots).$$

Let us define $f(j) = \mathbf{E}(Y | x_0 = j)$. Then

$$\begin{aligned} f(x_0) = \mathbf{E}(\mathbf{E}(Y \,|\, x_0, x_1) \,|\, x_0)) &= \mathbf{E}(\mathbf{E}(\varphi(x_1, x_2, \dots) \,|\, x_0, x_1) \,|\, x_0) \\ &= \mathbf{E}(\mathbf{E}(\varphi(x_1, x_2, \dots) \,|\, x_1) \,|\, x_0) \\ &= \mathbf{E}(f(x_1) \,|\, x_0). \end{aligned}$$

We have used consecutively the tower property, the invariant property of φ, and the Markov property of (x_n). This means $f = fP$, i.e.

$$f(j) = \sum_k f(k) P_{kj}.$$

** In fact, f is a 'harmonic function' and $f(x_n)$ is a 'martingale'.

Now we suppose that $\mathcal{X} = C_0 \cup_{k=1}^M C_k$, the sets C_k are disjoint, C_0 is the set of transient states, and C_k, for each $k \neq 0$, is a minimal communication class. Let B denote the subset of $\mathcal{X}^{\mathbf{N}_+}$ whose elements (a_n) has the property that a_n eventually belongs to C_1. Then B is an invariant set, and $\mathbf{1}_B$ an invariant function. Let $Y = \mathbf{1}_B(x_0, x_1, \dots)$. Then as before, we set $f(j) = \mathbf{P}(B|x_0 = j)$, this is the probability that x_n from j eventually lands in C_1. We may then solve the equations

$$f(j) = \sum_k f(k) P_{kj}$$

subject to the following boundary conditions: $f(j) = 1$ if $j \in C_1$ and $f(j) = 0$ if $j \in C_0, C_2, \dots, C_M$. This system of equations may have more than one solution, if the probabilities that x_n stays all the time in the transient states

$$g(j) = \mathbf{P}(x_n \in C_0 \text{ for all n}|x_0 = j),$$

are not all zero. We seek for the minimal solution.

5.9.1 The structure theorem

Previously we have defined an invariant set A by the property that $\mathbf{P}(x, A) = 1$ for all x. Now with an invariant probability measure at hand we would weaken this concept to one that is more useful for the current discussion.

Definition 5.9.3 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 $x \in \bar{A}$.

The importance of invariant measures can be seen in the following structural theorem, which is a consequence of Birkhoff's ergodic theorem:

Theorem 5.9.4 *Given a time homogeneous transition probability P, denote by $\mathcal{I}$ the set of all invariant probability measures for P and by $\mathcal{E} \subset \mathcal{I}$ the set of all those that are ergodic. Then the following statements hold.*

(a) The set $\mathcal{I}$ is convex and $\mathcal{E}$ is precisely the set of its extremal points. (that is it cannot be decomposed as $\pi = t\pi_1 + (1-t)\pi_2$ with $t \in (0,1)$ and $\pi_i \in \mathcal{I}(T)$).

(b) Any two ergodic invariant probability measures are either identical or mutually singular.

(c) Furthermore, every invariant measure is a convex combination of ergodic invariant measures.

Remark 5.9.5 As a consequence, if a Markov process admits more than one invariant measure, it does admit at least two ergodic (and therefore mutually singular) ones. This leads to the intuition that, in order to guarantee the uniqueness of its invariant measure, it suffices to show that a Markov process explores its state space 'sufficiently thoroughly'.

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 5.9.6 *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 $\mathcal{I}(T)$. □

5.9.2 Proofs**

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

Theorem 5.9.7 *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.*

For the proof, we will 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 $\bar{\mathcal{F}}$ of a σ-algebra $\mathcal{F}$ to be the smallest σ-algebra containing $\mathcal{F}$ with the additional property that if $A \in \bar{\mathcal{F}}$ with $\mathbf{P}(A) = 0$ and $B \subset A$ is any subset of A, then $B \in \bar{\mathcal{F}}$.

Definition 5.9.8 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, in other words $\mathbf{P}(A \bigtriangleup B) = 0$.

Let us recall properties of the symmetric differences $A \bigtriangleup B = A \setminus \cup B \setminus A$ of two sets. Firstly,

$$A \bigtriangleup B = A \cup B \setminus (A \cap B).$$

Thus $A^c \bigtriangleup B^c = A \bigtriangleup B$. Also, for any collection of sets $\{A_\alpha, B\alpha\}$,

$$\Big(\bigcup_\alpha A_\alpha\Big) \bigtriangleup \Big(\bigcup_\alpha B_\alpha\Big) \subset \bigcup_\alpha (A_\alpha \bigtriangleup B_\alpha)\,.$$

Also if $f : \Omega \to \Omega$ is any measurable function, then

$$f^{-1}(A \triangle B) = f^{-1}(A \triangle B).$$

Furthermore

$$(A \triangle B) \triangle (B \triangle C) = A \triangle C.$$

Before we turn to the proof of Theorem 5.9.7, we prove the following preliminary lemma, Let $\mathcal{F}_n^m = \vee_{k=n}^m \sigma(x_k)$.

Lemma 5.9.9 *Let $\mathbf{P}$ be the law of a stationary Markov process on $\mathcal{X}^{\mathbf{Z}}$. Then, the σ-algebra $\mathcal{I}$ of all subsets invariant under θ is contained (up to sets of $\mathbf{P}$-measure 0) in $\mathcal{F}_0^0$.*

Proof. We first prove that sets in $\mathcal{B}(\mathcal{X}^{\mathbf{Z}})$ can be approximated by cylindrical sets, we claim that

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

Denote the collections of sets on the right hand side by $\mathcal{B}_0$, which contains all cylindrical sets. It suffices to show that $\mathcal{B}_0$ is a σ-algebra. For this, since $\mathcal{B}_0$ clearly contains ϕ 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 $A_j' \in \mathcal{F}_{-N_j}^{N_j}$ 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

$$\begin{aligned}
\mathbf{P}\left(A \triangle \bigcup_{j\leq J} A_j'\right) &= \mathbf{P}\left(\left(A \triangle \bigcup_{j\leq J} A_j\right) \triangle \left(\bigcup_{j\leq J} A_j \triangle \bigcup_{j\leq J} A_j'\right)\right) \\
&\leq \mathbf{P}\left(A \triangle \bigcup_{j\leq J} A_j\right) + \mathbf{P}\left(\bigcup_{j\leq J} A_j \triangle \bigcup_{j\leq J} A_j'\right) \\
&\leq \varepsilon + \mathbf{P}\left(\bigcup_{j\leq J} (A_j \triangle A_j')\right) \leq \sum_{j\leq J} 2\varepsilon
\end{aligned}$$

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 $A \in \mathcal{I}$. For every $\varepsilon > 0$, consider $N > 0$ and a set $A_\varepsilon \in \mathcal{F}_{-N}^N$ such that $\mathbf{P}(A \triangle A_\varepsilon) < \varepsilon$. Then by the invariance of A and of $\mathbf{P}$ under shifts, for any n,

$$\mathbf{P}(A \triangle A_\varepsilon) = \int_\Omega \mathbf{1}_{A \triangle A_\varepsilon} \circ \theta^n d\mathbf{P} = \int_\Omega \mathbf{1}_{\theta^{-n}(A) \triangle \theta^{-n}(A_\varepsilon)} d\mathbf{P}.$$

By the invariance of A, it follows that we also have for any k, $\mathbf{P}(A \triangle \theta^{-(k+N)} A_\varepsilon) < \varepsilon$. Since $\theta^{-(k+N)} A_\varepsilon \subset \mathcal{F}_k^\infty$ for every ε, it follows that one has $A \in \bar{\mathcal{F}}_k^\infty$. Since this is true for every k, one actually has $A \in \bar{\mathcal{F}}_{\infty}^{\infty}$. The same reasoning but shifting in the other direction shows that one also has $A \in \bar{\mathcal{F}}_{-\infty}^{-\infty}$.

Point *(iii)* of Theorem 2.1.9 (or rather a slight extension of it) shows that if f and g are two functions that are respectively $\bar{\mathcal{F}}_{\infty}^{\infty}$ and $\bar{\mathcal{F}}_{-\infty}^{-\infty}$-measurable, then

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

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

$$\mathbf{E}\big(\chi_A^2 \,|\, \mathcal{F}_0^0\big) = \big(\mathbf{E}(\chi_A \,|\, \mathcal{F}_0^0)\big)^2\;.$$

Since on the other hand $\chi_A^2 = \chi_A$ and $\mathbf{E}(\chi_A \,|\, \mathcal{F}_0^0) \in [0,1]$. Let $\hat{A}$ denotes those points this equals one:

$$\hat{A} = \{\mathbf{E}(\chi_A \,|\, \mathcal{F}_0^0) = 1\}.$$

Then $\hat{A} \in \mathcal{F}_0^0$, and by the definition of conditional expectations, for any $E \in \mathcal{F}_0^0$, $\mathbf{P}(\hat{A} \cap E) = \mathbf{P}(A \cap E)$ 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 and $\hat{A}$ differ by a set of $\mathbf{P}$-measure 0, as required. □

The time 0 is not distinguished and can be replaced by any other time.

Corollary 5.9.10 *Let again* $\mathbf{P}$ *be the law of a stationary Markov process. Then, for every set* $A \in \mathcal{I}$ *there exists a measurable set* $\bar{A} \subset \mathcal{X}$ *such that* $A \sim \bar{A}^{\mathbf{Z}}$.

Proof. We know by Lemma 5.9.9 that $A \in \bar{\mathcal{F}}_0^0$, 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 $A \in \mathcal{I}$, 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 $\{x_n \in \bar{A}$ for every $n\}$. □

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

Let T be a transition operator on a space $\mathcal{X}$ and let π be an invariant probability measure for T. Recall 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 5.9.11 *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 $A \in \mathcal{I}$ such that $\mathbf{P}_\pi(A) \notin \{0,1\}$. By Corollary 5.9.10, 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 5.9.7. 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 $A \in \mathcal{I}$. 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 $\mathcal{I}$, 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 5.8.7, 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 $E_1 \cap E_2 = \phi$, 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 5.9.11, 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 5.8.7. 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 5.8.7, 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. □

In a rather analogous way, one has the following extension of Proposition 5.7.5:

Proposition 5.9.12 *Let $\mathcal{A}$ be an invariant set for P and let $\mathcal{A}_n$ be defined as in (5.15). 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 5.9.6. Suppose now that π is an arbitrary invariant measure for the process. It follows from Proposition 5.7.5 that $\pi(\mathcal{A}) = 1$. Furthermore, it follows as in the proof of the second part of Theorem 5.9.7 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 $\pi = \sum_k \pi(\mathcal{B}_k)\,\pi_k$. □

5.9.3 Proof of Birkhoff's Ergodic Theorem**

Before we turn to the proof of Theorem 5.8.4, we establish the following important result:

Theorem 5.9.13 (Maximal Ergodic Theorem) *With the notations of Theorem 5.8.4, 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

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

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 $f - \mathbf{E}\big(f \,|\, \mathcal{I}\big)$, we can assume without loss of generality that $\mathbf{E}\big(f \,|\, \mathcal{I}\big) = 0$. 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 $A^\varepsilon = \{\bar\eta(\omega) > \varepsilon\} \in \mathcal{I}$. Define

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

and define S_N^ε and M_N^ε accordingly. It follows from Theorem 5.9.13 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

$$\frac{S_N^\varepsilon(\omega)}{N} = \begin{cases} 0 & \text{if } \bar\eta(\omega) \leq \varepsilon \\ \frac{S_N(\omega)}{N} - \varepsilon & \text{otherwise.} \end{cases} \tag{5.18}$$

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 (5.18) 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

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

where we used the fact that $A^\varepsilon \in \mathcal{I}$ 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. □

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.

5.10 An overview

Chapter 6

Appendix: Continuous time Markov processes

This is the mastery material. Some specific references are given at the end of the sections, otherwise use the general references at the end of the notes. The state space is always assumed to be a separable complete metric space.

Guide: Pay attention to the Markov property, martingales, invariant measures, generators, and Markov semi-groups. Familiarise yourself with the standard examples in the notes. We are mainly concerned with processes whose sample paths have the continuity property. Poisson processes are given at the end of the section as an important classes of pure jump processes.

6.1 Introduction

A collection of random variables, $(x_t, t \in I)$, indexed by a parameter t in some interval I is a continuous time stochastic process. The collection of random variables is no longer countable, we will often add further restrictions to eliminate technical problems such as unions of sets of measure zeros. By the sample paths of a stochastic process we mean the functions of time: $t \mapsto x_t(\omega)$ for ω fixed. Suitable regularity assumptions on the sample paths allow us to conclude properties of the stochastic process with its restriction to a countable number of times. For the same (countability) reason instead of using the filtration generated by the processes we often use a modified version that satisfies additional assumptions, and $x.$ is sometimes assumed to be measurable as a function on the product space $I \times \Omega$.

Let $\mathcal{X}$ be a complete separable metric space as before. Let $\mathcal{D}(\mathbf{R}_+, \mathcal{X})$ denote the set of functions $f : \mathbf{R}_+ \to \mathcal{X}$ that is continuous from the right and with limits on the left at every point $t \in \mathbf{R}_+$. This is called the Skorohod space and functions there are said to be càdlàg (right-continuous with left limits). By a càdlàg process we mean that almost surely all its sample paths are càdlàg.

A continuous time Markov process is no longer described by a single Markov transition kernel P, but by a family of transition kernels P_t satisfying the semigroup property $P_{s+t} = P_s P_t$ and such that P_0 is the identity: $P_0(x, \cdot) = \delta_x$. Without further restrictions, a continuous-time Markov process could have very pathological properties. We will therefore always assume that $t \mapsto P_t(x, A)$ is measurable for every pair of x and measurable set A and that, for every initial condition $x \in \mathcal{X}$, the process admits a version that is càdlàg as a function of time. In other words, we will assume that for every $x \in \mathcal{X}$, there exists a probability measure $\mathbf{P}_x$ on $\mathcal{D}(\mathbf{R}_+, \mathcal{X})$ such that its marginals on $\mathcal{X}^n$ at any finite number of times $t_1 < \ldots < t_n$ are given by the probability measure

$$P_{t_1}(x, dx_1) P_{t_2 - t_1}(x_1, dx_2) \cdots P_{t_n - t_n - 1}(x_{n-1}, dx_n) \; .$$

6.2 Markov processes, Transition probabilities

6.2.1 Standard terminologies

For every ω, the function $t \mapsto x_t(\omega)$ is called a sample path of the stochastic process. A function on $\mathbf{R}$ said to be cádlág if it is right continuous and has left limit at every point.

Definition 6.2.1 We say a stochastic process is (sample) continuous (resp. right continuous, cádlág, etc.) if almost surely all of its samples paths are continuous (resp. right continuous, cádlág, etc.)

Unless specifically stated otherwise, all stochastic processes in this chapter are cádlág.

Let $\theta_s : \mathcal{X}^{\mathbf{R}_+} \to \mathcal{X}^{\mathbf{R}_+}$ be the shift map: $\theta_s a_\cdot(t) = a_{s+t}$.

Definition 6.2.2 A stochastic process (x_t) is said to be stationary if for every $s > 0$, the stochastic process $\theta_s x_\cdot$ has the same distribution. This is the same as the statements that for any $s > 0$, $n \geq 1$, $t_1 < t_2 < \cdots < t_n$, and $A_i \in \mathcal{B}(\mathcal{X})$,

$$\mathbf{P}\left(x_{t_1} \in A_1, \ldots, x_{t_n} \in A_n\right) = \mathbf{P}\left(x_{s+t_1} \in A_1, \ldots, x_{s+t_n} \in A_n\right).$$

As for the discrete time case, by a filtration we mean a collection of σ-algebras $(\mathcal{F}_t, t \in I)$ with the property that $\mathcal{F}_s \subset \mathcal{F}_t$ whenever $s \leq t$ and $s, t \in I$. We use $\vee_{t \geq 0} \mathcal{F}_t$ for the smallest σ-algebra containing the union $\cup_{t \geq 0} \mathcal{F}_t$. Let $\mathcal{F}_\infty = \vee_{t \geq 0} \mathcal{F}_t$. If $\mathcal{F}_t$ is a filtration of σ-algebras, we define $F_t^+ := \cap_{h > 0} \mathcal{F}_{t+h}$. Then $(\mathcal{F}_t^+,)$ is also a filtration.

The natural filtration of a stochastic process (x_t) is the smallest σ-algebra such that x_s is measurable for every $0 \le s \le t$. It is often convenient to use a larger filtration, which is convenient for study several stochastic processes simultaneously and sometimes we wish to include subsets of measure zero sets in the σ-algebras. For this reason we introduce the following concepts:

Definition 6.2.3 A stochastic process $(x_t, t \in I)$ is said to be **adapted to a filtration** $\mathcal{F}_t$ (also called $\mathcal{F}_t$-adapted) if x_t is $\mathcal{F}_t$-measurable for every $t \in I$.

Definition 6.2.4 A filtration $(\mathcal{F}_t, t \ge 0)$ is right continuous if $\mathcal{F}_{t+} = \mathcal{F}_t$.

Exercise 6.2.1 If $\mathcal{F}_t$ is right continuous, then T is a stopping time if and only if $\{T \le t\} \in \mathcal{F}_t$.

Given any filtration $\mathcal{F}_t$, the filtration $\{\mathcal{G}_t : \mathcal{G}_t = \mathcal{F}_{t+}\}$ is always right continuous. The natural filtration of a continuous stochastic process is not necessarily right continuous. If (X_t) is an right continuous $(\mathcal{F}_t)$-adapted stochastic process, the hitting time of an open set is an $\mathcal{F}_{t+}$-stopping time. See the book of Revuz-Yor for a proof.

Standard assumption. Unless otherwise stated we assume that $\mathcal{F}_t = \mathcal{F}_t^+$.

Definition 6.2.5 A random time $T : \Omega \to \mathbf{R} \cup \{\infty\}$ is said to be an $\mathcal{F}_t$- stopping time, if $\{T \le t\} \in \mathcal{F}_t$ for every $t \in I$.

Often we drop the reference to the filtration and simply refer T as a stopping time. The first hitting time of closed set by a continuous $(\mathcal{F}_t)$-adapted stochastic process is an $\mathcal{F}_t$- stopping time. If (X_t) is a right continuous $(\mathcal{F}_t)$-adapted stochastic process, the hitting time of an open set is an $\mathcal{F}_t^+$-stopping time.

Theorem 6.2.6 *For a right continuous filtration $(\mathcal{F}_t, t \ge 0)$, the following statements on $T : \Omega \to \mathbf{R}$ are equivalent: (1) $\{T < t\} \in \mathcal{F}_t$ for every $t \ge 0$; (2) $\{T \le t\} \in \mathcal{F}_t$ for every $t \ge 0$.*

6.2.2 General Markov Processes

We may now define a general Markov process with respect to a filtration $\mathcal{F}_t$. This filtration needs not be the natural filtration of the Markov process.

Definition 6.2.7 A family of $\mathcal{F}_t$ adapted stochastic process X_t with values in $\mathcal{X}$ is an $\mathcal{F}_t$-Markov process (or a Markov process with respect to $\mathcal{F}_t$) if for all bounded Borel measurable function $f : \mathcal{X} \to \mathbf{R}$ and for all $0 \le s \le t$,

$$\mathbf{E}(f(X_t)|\mathcal{F}_s) = \mathbf{E}(f(X_t)|X_s).$$

It is strong Markov if for all finite stopping time τ and $t \ge 0$, $\mathbf{E}(f(X_{\tau+t})|\mathcal{F}_\tau) = \mathbf{E}(f(X_{\tau+t})|X_\tau)$.

Let x_t be a Markov process on $\mathcal{X}$ with respect to its natural filtration. Suppose that $\Phi : \mathbf{R} \times \mathcal{X} \to \mathcal{Y}$ is a measurable map where $\mathcal{Y}$ is another separable metric space. Suppose for each t, $\Phi(t, \cdot)$ has a measurable inverse, then $y_t = \Phi(t, x_t)$ and x_t have the same filtration and y_t is also a Markov process.

6.2.3 Time homogeneous Markov processes with transition probabilities

To eliminate unnecessary technical problems, we will however restrict ourselves to the class of Markov processes for which regular conditions expectations exist. More precisely we assume that $\mathbf{E}(f(X_t)|\mathcal{F}_s)$ are given by transition probabilities.

Definition 6.2.8 A family or probability measures $P = \{P_s(x, \cdot), s \in \mathbf{R}_+, x \in \mathcal{X}\}$ on $\mathcal{X}$ is said to be a time homogeneous Markov transition function (or time homogeneous transition probabilities), if the following conditions hold:

1. $P_0(x, \cdot) = \delta_x$, for every $x \in \mathcal{X}$,
2. $(s, x) \mapsto P_s(x, A)$ is measurable, for every $A \in \mathcal{B}(\mathcal{X})$, .
3. Chapman-Kolmogorov equations:
$$P_{s+t}(x, A) = \int_{\mathcal{X}} P_t(y, A)\ P_s(x, dy), \qquad \forall s, t > 0, \forall A \in \mathcal{B}(\mathcal{X}), \forall x \in \mathcal{X}.$$
4. For every $x \in \mathcal{X}$, $\lim_{t\to 0} P_t(x, \cdot) = \delta_x$ (weakly).

Assumption 4, which is on the continuity of P_t at 0, is sometimes not included in the definition. It is nevertheless regularly assumed as an additional assumption.

The Chapman-Kolmogorov equations is equivalent to the following: for any bounded measurable functions $f : \mathcal{X} \to \mathbf{R}$,

$$\int_{\mathcal{X}} f(z) P_{t+s}(x, dz) = \int_{\mathcal{X}} \int_{\mathcal{X}} f(z) P_t(y, dz)\ P_s(x, dy). \tag{6.1}$$

As with the discrete time case, a Markov processes with these transition probabilities can be constructed using Kolmogorov's extension theorem.

Definition 6.2.9 An $\mathcal{F}_t$ adapted stochastic process (X_t) is said to be a time homogeneous Markov process with Markov transition probabilities P, if for all $0 \le s, t$ and for all bounded measurable function $f : \mathcal{X} \to \mathbf{R}$

$$\mathbf{E}(f(X_{t+s})|\mathcal{F}_s) = \int_{\mathcal{X}} f(y) P_t(X_s, dy). \tag{6.2}$$

It is strong Markov if for any finite stopping time τ,

$$\mathbf{E}\big(f(X_{t+\tau})|\mathcal{F}_\tau\big) = \int_{\mathcal{X}} f(y) P_t(X_\tau, dy).$$

We observe that the following identity holds automatically for the time homogeneous Markov process with Markov transition probabilities P:

$$\mathbf{E}(f(X_{t+s})|X_s) = \int_{\mathcal{X}} f(y) P_t(X_s, dy).$$

In other notation, for any $s, t \ge 0$,

$$\mathbf{E}(f(X_{t+s})|X_s = x) = \int_{\mathcal{X}} f(y) P_t(x, dy).$$

As we explained before, the finite dimensional distribution of a time homogenous cádlág Markov process is determined by its initial distribution and the transition probabilities. In other words the time homogeneous Markov process induces a probability measure $\mathbf{P}_\nu$ on the path space $\mathcal{D}(\mathbf{R}_+, \mathcal{X})$ such that its marginals on $\mathcal{X}^n$ at any finite number of times $0 < t_1 < \ldots < t_n$ are given by the probability measure

$$\nu(dx) P_{t_1}(x, dx_1) P_{t_2 - t_1}(x_1, dx_2) \cdots P_{t_n - t_{n-1}}(x_{n-1}, dx_n)\ .$$

If the initial distribution is δ_x we write $\mathbf{P}_x$ for $\mathbf{P}_{\delta_x}$. We summarise this below.

Proposition 6.2.10 *Let (X_t) be a Markov process with probability transition probabilities $\mathbf{P}$ and initial distribution ν. Prove that for $A_i \in \mathcal{B}(\mathcal{X})$ and for $0 \le t_0 < t_1 < \cdots < t_k$,*

$$\begin{aligned}&\mathbf{P}\left(X_0 \in A_0, X_{t_1} \in A_1, \ldots, X_{t_k} \in A_k\right)\\ &\int_{A_0} \cdots \int_{A_k} P_{t_k - t_{k-1}}(x_{k-1}, dx_k) \ldots P_{t_2 - t_1}(x_1, dx_2) P_{t_1}(x_0, dx_1) \nu(dx_0).\end{aligned} \tag{6.3}$$

If f a bounded measurable function, we define for $x \in \mathcal{X}$,

$$P_t f(x) = \int f(y) P_t(x, dy). \tag{6.4}$$

Equivalently, $P_t f(x) = \mathbf{E}\left(f(X_t)|X_0 = x\right)$. If x is the non-random initial value,

$$P_t f(x) = \mathbf{E}[f(x_t)]. \tag{6.5}$$

Definition 6.2.11 Let $\mathcal{B}_b$ be the space of bounded measurable functions from $\mathcal{X} \to \mathbf{R}$. A family of bounded linear operators $P_t : \mathcal{B}_b \to \mathcal{B}_b$ is called a Markov semigroup if

1. P_0 is the identity map,
2. Semi-group property: $P_{s+t} = P_s P_t$ for any $0 \le s, t$,
3. Positivity preserving $P_t f \ge 0$ whenever $f \ge 0$,
4. Conservative: $P_t 1 = 1$, where 1 denotes the function taking the constant value 1.

Exercise 6.2.2 Show that P_t defined by (6.4) is a linear operator on $\mathcal{B}_b$ and a Markov semi-group of bounded linear operators, on the space of bounded measurable functions.

Definition 6.2.12 We say P_t, defined by (6.4) is the Markov semigroup associated with the Markov process. If $t = 1$, this is denoted by P, which are also called transition operators.

We may also define Feller and strong Feller properties.

Definition 6.2.13 We say that a homogeneous Markov process (or its Markov semigroup P_t) is **Feller** if $P_t f$ is continuous whenever f is continuous and bounded for every $t > 0$. It is **strong Feller** if $P_t f$ is continuous whenever f is measurable and bounded for every $t > 0$.

6.3 Invariant Measure

As with the discrete time case, we define the evolution of measures as follows:

$$P_t \mu(A) = \int_{\mathcal{X}} P_t(x, A) \mu(dx). \tag{6.6}$$

We have used the same notation for both the linear operators on functions and for the linear operators on measures. We again have the duality:

$$\int_{\mathcal{X}} P_t f(y) \mu(dy) = \int_{\mathcal{X}} f(y)(P_t \mu)(dy).$$

Definition 6.3.1 A measure μ is invariant for the transition probabilities P_t, if

$$P_t \mu = \mu, \quad \forall t \ge 0.$$

Observe that $P_t \pi = \pi$ is equivalent to $\int_{\mathcal{X}} P_t f(y) \pi(dy) = \int_{\mathcal{X}} f(y) \pi(dy)$ for every bounded measurable function f.

Exercise 6.3.1 If π is an invariant probability measure for P_t then for any $p \ge 1$,

$$\int |P_t f|^p d\pi \le \int |f|^p d\pi.$$

This property is called L_p contraction property.

Note that in theory, it is always possible to restrict oneself to the case of discrete time in the study of the existence and uniqueness of an invariant measure:

Proposition 6.3.2 *Let P_t be a Markov semigroup over $\mathcal{X}$. If $P_T\mu = \mu$ for some fixed $T > 0$, then the measure $\hat{\mu}$ defined by*

$$\hat{\mu}(A) = \frac{1}{T}\int_0^T P_s\mu(A)\,ds$$

is invariant for the semigroup P_t. □

Proof. Suppose that $P_T\mu = \mu$. Let $t \in [0,T]$. Then,

$$\begin{aligned}(P_t\hat{\mu}) &= \frac{1}{T}\int_0^T P_{t+s}\mu\,ds = \frac{1}{T}\int_t^{t+T} P_s\mu\,ds = \frac{1}{T}\left(\int_t^T P_s\mu\,ds + \int_T^{t+T} P_s\mu\,ds\right)\\ &= \frac{1}{T}\left(\int_t^T P_s\mu\,ds + \int_0^t P_{s+T}\mu\,ds\right).\end{aligned}$$

Since $P_{s+T}\mu = P_sP_T\mu = P_s\mu$, the right hand side is $\hat{\mu}$. Similarly, if $t > T$,

$$\begin{aligned}(P_t\hat{\mu}) &= \frac{1}{T}\int_t^{t+T} P_s\mu\,ds = \frac{1}{T}\left(\int_T^{t+T} P_s\mu\,ds - \int_T^t P_s\mu\,ds\right)\\ &= \frac{1}{T}\int_0^t P_s\mu\,ds - \frac{1}{T}\int_T^t P_s\mu\,ds = \hat{\mu}.\end{aligned}$$

□

Remark 6.3.3 The converse is not true at this level of generality. This can be seen for example by taking $P_t(x,\cdot) = \delta_{x+t}$ with $\mathcal{X} = S^1$.

6.4 Martingale

Here I is an interval or $\{0,1,2,\dots\}$ or more generally,

Definition 6.4.1 Let $\mathcal{F}_t$ be a filtration on $(\Omega, \mathcal{F}, P)$. An adapted stochastic process $(X_t, t \in I)$ is a -martingale if $\mathbf{E}|X_t| < \infty$ and

$$\mathbf{E}(X_t|\mathcal{F}_s) = X_s, \qquad \forall s \le t.$$

Note that if M_t is a martingale and $\mathbf{E}M_0 = 0$ then $\mathbf{E}M_t = 0$.

Example 6.4.1 Let $f \in L^1$ and $f_t = \mathbf{E}\{f|\mathcal{F}_t\}$ then f_t is a martingale.

Example 6.4.2 If ξ_i are i.i.d. with mean zero and $x_n = \sum_{k=1}^n \xi_i$ is the random walk. Then $\mathbf{E}(x_n|\mathcal{F}_m) = x_m$.

Definition 6.4.2 Let (X_t) be an $\mathcal{F}_t$-adapted sample continuous stochastic process with $X_0 = 0$. If there exists a sequence of stopping times $\{T_n\}$ with $T_n \le T_m$ for $n \le m$ and $\lim_{n\to\infty} T_n = \infty$ a.s. and the property that for each n, $(X_t^{T_n}\}, t \ge 0)$ is a bounded martingale, we say that (X_t) is a local martingale and that T_n reduces X.

Note that we took $X_0 = 0$ to make the definition simple. A martingale is a local martingale. Conversely to conclude from the local martingale property, which implies

$$\mathbf{E}(X_t^{T_n} | \mathcal{F}_s) = X_s^{T_n},$$

that (X_t) is a martingale we need to assume that $X_t^{T_n}$ is uniformly integrable, sufficient if there exists a number K such that $|X_s| \le K$ for any $\in [0, t]$

6.5 Markov generators

Let f be a bounded measurable function and x_t a Markov process. We define

$$\mathcal{L}f(x) = \lim_{t\to 0} \frac{P_t f(x) - f(x)}{t}$$

if the limit exists for every $x \in \mathcal{X}$, and in which case we say that f is in the domain of the generator $\mathcal{L}$. The generator $\mathcal{L}$, also known as the Markov generator or the infinitesimal generator, is a linear map from a function space on which it is defined. This space is its domain. Determining the domain of the generator is in general tedious, however it is often sufficient to determine a 'large' subset of the domain. For the standard Brownian motion, the domain of the generator contains domain contains smooth functions with compact supports, and on such functions $\mathcal{L}f(x) = \frac{1}{2} f''(x)$. If the Markov process has continuous sample paths (called diffusions), its generator $\mathcal{L}$ is a second order differential operator. This leads to the following definition.

Exercise 6.5.1 Let C be a real number and let $x_t = x + Ct$ denote translation at a speed C. What is the generator of x_t?

6.5.1 Kolmogorov's backward and forward equations

The condition on P_t near zero, given by its generator, describes also the Markov process for all time. Let x_t be a Markov process and f a bounded measurable function, then

$$P_{t+s} f(x) = \mathbf{E}\left(f(x_{t+s} \,|\, x_0 = x\right) = \mathbf{E}\left(\mathbf{E}\left(f(x_{t+s} \,|\, x_s\right) \,|\, x_0 = x\right) = P_s\, P_t f(x).$$

We may then subtract $P_t f$ from both sides, divides the result by s, finally take $s \to 0$. Assume $P_t f$ is in the domain of $\mathcal{L}$, then

$$\frac{\partial}{\partial t} P_t f(x) = \lim_{s\to 0} \frac{P_{t+s} f(x) - P_t f(x)}{s} = \mathcal{L}(P_t f)(x).$$

Definition 6.5.1 The Kolmogorov's backward equation for P_t is:

$$\frac{\partial}{\partial t} = \mathcal{L}. \tag{6.7}$$

The differentiation is with respect to x, which is the initial condition of the Markov process ($\mathbf{E}(f(X_t)|X_0 = x)$) and considered as the backward variable.

Definition 6.5.2 The adjoint $\mathcal{L}^*$ is determined by the following relation

$$\int_{\mathbf{R}} \mathcal{L}f(x)g(x)dx = \int_{\mathbf{R}} f(y)\mathcal{L}^*g(x)dx,$$

where f, g are smooth functions with compact supports.

Consider again

$$\lim_{s\to 0}\frac{1}{s}\left(\int_{\mathcal{X}} f(y)P_{t+s}(x,dy) - \int_{\mathcal{X}} f(y)P_t(x,dy)\right) = \int_{\mathcal{X}} \mathcal{L}f(y)P_t(x,dy) = \int_{\mathcal{X}} f(y)\ \mathcal{L}^*P_t(x,dy).$$

Suppose $\mathcal{X} = \mathbf{R}$ and $P_t(x,dy) = p_t(x,y)dy$ for some measurable function $p_t(x,y)$. The above identity becomes:

$$\int_{\mathbf{R}} f(y)\frac{\partial}{\partial t}p_t(x,y)dy = \int_{\mathbf{R}} \mathcal{L}f(y)p_t(x,y)dy = \int f(y)\mathcal{L}^*p_t(x,y)dy.$$

If this holds for a sufficiently large class of functions f, $\frac{\partial}{\partial t}p_t(x,y) = \mathcal{L}^*p_t(x,y)$.

Definition 6.5.3 The Fokker-Plank equation (forward Kolmogorov equation) is:

$$\frac{\partial}{\partial t}p_t(x,y) = \mathcal{L}^*p_t(x,y).$$

Observe that $\mathcal{L}^*$ is applied to the y variable, the forward variable (WE think of $p_t(x,y)$ to be the probability density of finding a diffusion particle at x moved to y after an overlap time t.)

6.6 Examples

In this section we study some examples.

Definition 6.6.1 We say that a stochastic process $\tilde{x}_t$ is a version of x_t, if $\mathbf{P}(x_t = x'_t) = 1$ for every t.

6.6.1 Brownian motion

It is always possible to choose a version of B_t with continuous sample paths, thus we always assume that Brownian motions are continuous. There are multiple equivalent definitions for Brownian motions (BM). For example, the definition which does not use advanced concept, contains a lot of information and is also easy to use, is the following:

Definition 6.6.2 A stochastic process $(B_t : t \geq 0)$ on $\mathbf{R}^1$ is the standard Brownian motion if the following holds:

(1) [Independent increments property] For each n and times $0 = t_0 \leq t_1 < \cdots \leq t_n$, $\{B_{t_i} - B_{t_{i-1}}, i = 1, \dots, n\}$ are independent random variables,

(2) For any $0 \leq s < t$, the probability distribution of $B_t - B_s$ is the Gaussian measure $N(0, t-s)$,

(3) (B_t) is sample continuous.

Recall that a general Gaussian measure on $\mathbf{R}$ is either a Dirac measure at a point a or a probability measure with a density of the following form with respect to the Lebesgue measure:

$$p_t(a,x) \overset{\text{def}}{=} \frac{1}{\sqrt{2\pi t}} \exp\left(-\frac{(x-a)^2}{2t}\right).$$

The parameters a and t have the following interpretation:

$$a = \int_{-\infty}^{\infty} x\, p_t(a,x)\, dx, \qquad t = \int_{-\infty}^{\infty} (x-a)^2\, p_t(a,x)\, dx$$

A Gaussian random variable is a random variable with a Gaussian distribution. Sums of independent Gaussian random variables are Gaussian random variables.

Brownian motion is a Gaussian process, i.e. for any $n \in \mathbf{N}$ and any numbers $0 \leq t_1 < \cdots < t_n$, the distribution of the random variable $(B_{t_1}, \dots B_{t_n})$, with values in $\mathbf{R}^n$, is Gaussian. In fact their joint probability distribution is:

$$p_{t_1}(0, y_1) \cdot p_{t_2 - t_1}(y_1, y_2) \dots p_{t_n - t_{n-1}}(y_{n-1}, y_n)\, dy_1\, dy_2 \dots dy_n.$$

The measure of the Brownian motion on the path space is called the Wiener measure (its existence can be proved by Kolmogorov's theorems.)

It is an important factor that for almost surely all ω, $t \mapsto B_t(\omega)$ is Hölder continuous of any order up $\alpha < \frac{1}{2}$. With probability one Brownian paths are not Hölder continuous of order $\alpha > \frac{1}{2}$ at any time and so has no derivatives in time.

Definition 6.6.3 An n-dimensional Brownian motion is defined to be $(B_t^1, \dots, B_t^n)$ in which the coordinate processes $\{B_\cdot^i, i = 1, \dots, n\}$ are independent Brownian motions. It is called the Brownian motion started at x if $B_0 = x$, x is a point of $\mathbf{R}^n$. It is called the standard Brownian motion if $x_0 = 0$.

Unless otherwise stated by a Brownian motion we mean a one dimensional Brownian motion. Condition (2) indicates the increments of Brownian motions are stationary: the probability distribution of the increment $x_t - x_s$ depends only on the difference $t - s$. It is a good exercise to compute the covariance $\mathbf{E}(B_t B_s) = s \wedge t \overset{\text{def}}{=} \min(s,t)$.

Theorem 6.6.4 (Time inversion) *Suppose that (B_t) is a Brownian motion with $B_0 = 0$. Then the process $(x_t : t \ge 0)$ defined by $x_0 = 0$ and $x_t = tB_{\frac{1}{t}}$ for $t \ne 0$ is also a Brownian motion. This is called the inversion.*

From this one sees that $\lim_{t\to\infty} \frac{1}{t} B_t = 0$ a.s.

Exercise 6.6.1 Let $x_t = e^t B_{e^{-2t}}$ (this is a Ornstein-Uhlenbeck process). Show that x_t is a Markov process and compute the probability distribution of x_t. (Hint: Consider the natural filtration of x_t)

Exercise 6.6.2 (scale invariant) For any $\lambda > 0$, the process $\frac{1}{\lambda} B_{\lambda^2 t}$ is also a standard Brownian motion.

In this section let $\mathcal{F}_s := \sigma\{B_r : 0 \le r \le s\}$, the natural filtration of (B_t).

Theorem 6.6.5 (Increment is independent of the past) *For every $t \ge 0$ and $s \ge 0$, the stochastic process $\{B_{t+s} - B_s, t \ge 0\}$ is independent of $\mathcal{F}_s^+$.*

Proof. For any s, the processes $\{B_{t+s} - B_s, t \ge 0\}$ is independent of $\sigma\{B_u, 0 \le u \le s\}$. Take a family of s_n decreasing to s. Then $X_t^n = \{B_{t+s_n} - B_{s_n}, t \ge 0\}$ is independent of $\sigma\{B_u, 0 \le u \le s_n\}$, the latter contains $\mathcal{F}_s^+$. letting $n \to \infty$, the limiting process $\{B_{t+s} - B_s, t \ge 0\}$ is also independent of $\mathcal{F}_s^+$. □

Proposition 6.6.6 (Martingale property) *A Brownian motion is an $\mathcal{F}_t^+$-martingale.*

Proof. For any $s < t$,

$$\mathbf{E}(B_t \,|\, \mathcal{F}_s^+) = \mathbf{E}(B_t - B_s + B_s \,|\, \mathcal{F}_s^+) = \mathbf{E}(B_t - B_s \,|\, \mathcal{F}_s^+) + B_s.$$

By Theorem 6.6.5, $B_t - B_s$ is independent of $\mathcal{F}_s^+$, $\mathbf{E}(B_t - B_s \,|\, F_s^+) = \mathbf{E}(B_t - B_s) = 0$. □

Exercise 6.6.3 It is easy to see that B_t is an $\mathcal{F}_t^+$ Markov process.

Since $\mathcal{F}_t \subset \mathcal{F}_t^+$, x_t is $\mathcal{F}_t^+$ measurable. Since the increment increment $B_t - B_s$ is independent of the past $\mathcal{F}_s^+$,

$$\mathbf{E}(f(B_t)|\mathcal{F}_s^+) = \mathbf{E}(f(B_t - B_s + B_s)|\mathcal{F}_s^+) = \int_{\mathbf{R}} f(y + B_s) \frac{1}{\sqrt{2\pi t}} e^{-\frac{y^2}{2(t-s)}}\, dy.$$

Since the right hand side is a function of B_s, $\mathbf{E}(f(B_t)|\mathcal{F}_s^+) = \mathbf{E}(f(B_t)|B_s)$, and (B_t) is a Markov process.

Theorem 6.6.7 (Markov Property) *Let B_t be a Brownian motion (started at x). Then for every $s > 0$, $(B_{t+s} - B_s, t \geq 0)$ is a Brownian motion and is independent of the process $\sigma\{B_u, 0 \leq u \leq s\}$.*

Verify the theorem!

The following theorem states that the germ σ-algebra$\mathcal{F}_0^+$ is trivial.

Theorem 6.6.8 (Blumenthal's 0-1 law) *Let $A \in \mathcal{F}_0^+$. Then $\mathbf{P}(A) \in \{0, 1\}$.*

Proof. By Theorem 6.6.5, $\sigma\{B_t : t \geq 0\}$ is independent of $\mathcal{F}_0^+$. Since $\sigma\{B_t : t \geq 0\}$ contains $\mathcal{F}_0^+$, a set A from $\mathcal{F}_0^+$ is independent of itself, so $\mathbf{P}(A) = \mathbf{P}(A \cap A) = \mathbf{P}(A)\mathbf{P}(A)$, and hence has probability zero or probability 1. □

Let us consider the tail- σ algebra: $\bigcap_{t \geq 0} \sigma\{B_s : s \geq t\}$.

Theorem 6.6.9 *Kolmogorov's 0-1 law. The tail σ-algebra $\bigcap_{t \geq 0} \sigma\{B_s : s \geq t\}$ is trivial.*

Proof. The time inversion $x_t = tB_{\frac{1}{t}}$ is a Brownian motion. By the time reversal, the tail σ-algebra of B_t is in $\mathcal{F}_0^+$, the latter is trivial by Blumenthal's 0-1 law. □

Theorem 6.6.10 (Strong Markov property) *For every almost surely finite stopping time T, the process $(B_{t+T} - B_T : t \geq 0)$ is a Brownian motion, independent of $\mathcal{F}_T^+$.*

Proof. We have proved the strong Markov property for discrete time Markov processes. The same proof shows that a continuous time Markov process has the strong Markov property at any stopping times that takes only a countable numbers of values. We can then approximate T by stopping times taking values on the dyadic subsets of the real numbers: $\frac{k}{2^n}$. Set

$$T_n(\omega) = \frac{k+1}{2^n}, \qquad \text{if } T(\omega) \in \left[\frac{k}{2^n}, \frac{k+1}{2^n}\right).$$

Then T_n decreases to T. Since T_n takes at most a countable number of values, $B_{t+T_n} - B_{T_n}$ is independent of $\mathcal{F}_{T_n}^+$. Consequently, $B_{t+T_n} - B_{T_n}$ is independent of $\mathcal{F}_T^+$, as the latter is a subset of $\mathcal{F}_{T_n}^+$. By the sample continuity, $B_{t+T}-B_T$ is independent of $\mathcal{F}_T^+$ and also the above procedure shows that the increments $B_{t+T}-B_T$ are normal distributed, the independent increments and sample continuity property of $B_{t+T}-B_T$ can also be verified by the same limiting procedure. □

By now we should be able to appreciate section 5.1.1 on the the examples of Markov processes without the Feller property, please go back to read it!

Remark 6.6.11 An equivalent definition for a Brownian motion is: A Brownian motion $(B_t, t \geq 0)$, with values in $\mathbf{R}$ is a sample continuous strong Markov process with transition probabilities:

$$P_t(x,dy) = \frac{1}{\sqrt{2\pi t}} e^{-\frac{(x-y)^2}{2t}} dy, \quad t > 0$$

where dy is the Lebesgue measure on $\mathbf{R}$.

Remark 6.6.12 As with the discrete time case, it is possible to construct a two-sided Brownian motion $(B_t, t \in \mathbf{R})$ with the property that it has continuous sample paths, independent increments, and that the increments $B_t - B_s$ are distributed as $N(0, t-s)$.

Remark 6.6.13 For the Brownian motion $\mathcal{L}^* f = \mathcal{L} f = \frac{1}{2} f''$. This can be proved using Itô's formula. We illustrates as follows. Let $f : \mathbf{R} \to \mathbf{R}$ be C^2, Taylor expansion gives:

$$f(y) = f(x) + (y-x)f'(x) + \frac{1}{2}(y-x)^2 f''(x) + R(x,y)(y-x)^2$$

where $R(x,y)$ is a function that vanishes at the diagonal $x=y$ and continuous when $y \to x$. Then

$$f(x+B_t) = f(x) + B_t f'(x) + \frac{1}{2}(B_t)^2 f''(x) + R(x, x+B_t)(B_t)^2.$$

Taking expectations. If all terms converge, (which can be easily seen to hold for functions with compact supports, the remainder term R is then given by the evaluation of f''' at a point between x and $x+B_t$), then

$$\mathcal{L} f = \lim_{t \to 0} \frac{P_t f(x) - f(x)}{t} = \frac{1}{2} f''(x) = \lim_{t \to 0} \frac{\mathbf{E}(f(x+B_t)) - f(x)}{t} = \frac{1}{2} f''(x).$$

References. The book 'Brownian motion' by Möeters and Peres (esp. Chapter 2), is a good reference for Brownian motions.

6.6.2 Itô Processes and Itô formula

Given an adapted real valued stochastic process H_s with $\mathbf{E}(\int_0^t (H_s)^2 ds) < \infty$ (called an L_2 process), then an Itô integral $I_t := \int_0^t H_s dB_s$ can be defined and (I_t) is again an L_2 process and a martingale. (This integral can also be defined for any sample continuous stochastic process H_s, then the integral is always a local martingale) .

It is known the following holds:

1. Linearity: $\int_0^t (af + bg) dB_s = a \int_0^t f_s dB_s + b \int_0^t g_s dB_s$;

2. For any $r < t$,
$$\int_r^t dB_s = B_t - B_r.$$

3. Product formula:
$$(\int_0^t H_s dB_s)(\int_0^t K_s dB_s) = \int_0^t \left(\int_0^r H_s dB_s \right) K_r dB_r + \int_0^t \left(\int_0^r H_s dB_s \right) K_r dB_r + \int_0^t H_s K_s ds.$$
(Product formula is a form of 'integration by parts'.)

4. Itô Isometry: let H_s, K_s be L_2 processes. Then,
$$\mathbf{E} \left(\int_0^t H_s dB_s \int_0^t K_s \, dB_s \right) = \int_0^t \mathbf{E}(H_s K_s) ds.$$
$$\mathbf{E} \left(\int_0^t H_s dB_s \int_0^t K_s \, ds \right) = 0.$$

5. Let $M_t := \int_0^t H_s dB_s$, then M_t is a local martingale. If H_s is an L_2 process or a bounded process, then it is a martingale.

Definition 6.6.14 An Itô process is of the form
$$\int_0^t H_s dB_s + \int_0^t b_s ds,$$
where H_s is an L_2 process and b_s is an adapted stochastic process such that $\int_0^t b_s(\omega) ds < \infty$ for almost surely every ω.

Definition 6.6.15 **Itô's formula** Let $f : \mathbf{R} \to \mathbf{R}$ be C^2, and $I_t = \int_0^t H_s dB_s + \int_0^t b_s ds$ as defined earlier. Then
$$f(I_t) = f(I_0) + \int_0^t f'(I_s) H_s dB_s + \int_0^t f'(I_s) b_s ds + \frac{1}{2} \int_0^t f''(x_s) ds.$$

6.6.3 Stochastic differential equation

Let $\sigma : \mathbf{R} \to \mathbf{R}$ and $b : \mathbf{R} \to \mathbf{R}$ be C^1. By a solution to the stochastic differential equation:

$$dx_t = \sigma(x_t)dB_t + b(x_t)dt,$$

with initial value x we mean an adapted stochastic process satisfying the identity:

$$x_t = x + \int_0^t b(x_s)ds + \int_0^t \sigma(x_s)dB_s.$$

Solutions to the SDE are time-homogeneous Markov processes. To work out their Markov generators, we apply Itô's formula to x_t.

$$f(x_t) = f(x) + \int_0^t f'(x_s)\sigma(x_s)dB_s + \int_0^t f'(x_s)b(x_s)ds + \frac{1}{2}\int_0^t f''(x_s)(\sigma(x_s)^2 ds.$$

Since $\int_0^t f'(x_s)\sigma(x_s)dB_s$ is a local martingale, it is a martingale if it is uniformly bounded in t for t on finite time intervals $t \in [0, T]$. Then for any $f \in C^2$ bounded with bounded derivatives, we see

$$\mathcal{L}f(x) = \lim_{t\to 0} \frac{P_t f(x) - f(x)}{t} = f'(x)b(x) + \frac{1}{2}(\sigma(x))^2 f''(x).$$

We record this below as this is important :

Proposition 6.6.16 *The Markov generator for the solution is:*

$$\mathcal{L} = \frac{1}{2}(\sigma)^2 \frac{d}{dx} + b\frac{d}{dx}.$$

The simplest of those equations are those with σ a constant. Then by a solution we mean a stochastic process x_t satisfying

$$x_t(\omega) = x_0 + \int_0^t b(x_s(\omega))ds + \sigma B_t(\omega).$$

This means, fixing ω, the identity holds (for almost all ω).

6.6.4 The Ornstein-Uhlenbeck process

The Ornstein Uhlenbeck equation on $\mathbf{R}$ is the solution of the equation

$$dx_t = -\beta x_t\, dt + \sigma\, dB_t$$

where β and σ are two real numbers and B_t is a Brownian motion. (It is a Markov process by standard theory on stochastic differential equations.) This stochastic differential equation with additive noise can be interpreted in the following way without using Itô integration. By the solution we mean that for each ω, $x_t(\omega)$ solves the integral equation:

$$x_t(\omega) = x_0 - \beta \int_0^t x_s(\omega) ds + \sigma B_t(\omega), \tag{6.8}$$

where $x_0 \in \mathbf{R}$ is the initial condition, which we may assume to be a constant. The solution is a Markov process, called the Ornstein-Uhlenbeck process.

Proposition 6.6.17 *The Ornstein-Uhlenbeck process given by*

$$x_t = e^{-\beta t} x_0 + \sigma B_t - \sigma\beta \int_0^t B_s e^{-\beta(t-s)} ds \tag{6.9}$$

solves the Ornstein-Uhlenbeck equation.

Proof. We compute $\int_0^t x_r dr$ for x_t given in (6.9) Then

$$\int_0^t x_r \, dr = \int_0^t e^{-\beta r} x_0 \, dr + \sigma \int_0^t B_r dr - \int_0^t \sigma\beta \int_0^r e^{-\beta(r-s)} B_s \, ds dr.$$

We then apply integration by part to the last term,

$$\begin{aligned} \int_0^t \sigma\beta \int_0^r e^{-\beta(r-s)} B_s \, ds dr &= \sigma \int_0^t \left(\int_s^t \beta e^{-\beta(r-s)} dr \right) B_s \, ds \\ &= -\sigma \int_0^t \left(e^{-\beta(t-s)} - 1 \right) B_s \, ds. \end{aligned}$$

Put them together we have:

$$\int_0^t x_r \, dr = \int_0^t e^{-\beta r} x_0 \, dr + \sigma \int_0^t e^{-\beta(t-s)} \, ds = \frac{1}{\beta}(x_0 - e^{-\beta t} x_0) + \sigma \int_0^t e^{-\beta(t-s)} \, ds.$$

Thus,

$$\beta \int_0^t x_r \, dr = x_0 - e^{-\beta t} x_0 + \sigma\beta \int_0^t e^{-\beta(t-s)} \, ds = x_0 - x_t + \sigma B_t$$

In the last line we use again (6.9). This means that x_t solves (6.8). □

Remark 6.6.18 By the product formula

$$\sigma B_t - \sigma\beta \int_0^t B_s e^{-\beta(t-s)} ds = \sigma \int_0^t e^{-\beta(t-s)} dB_s,$$

the latter integral is Itô integral. By the Itô isometry,

$$\mathbf{E} \left(\sigma \int_0^t e^{-\beta(t-s)} dB_s \right)^2 = \sigma^2 \int_0^t e^{-2\beta(t-s)} ds = \frac{\sigma^2}{2\beta}(1 - e^{-2\beta t}).$$

The probability distribution of x_t is Gaussian (Because it is obtained from Gaussian random variable by a linear operation). Recall that Gaussian distributions on $\mathbf{R}$ are determined by its mean and variance. Observe that $\mathbf{E}x_t = e^{-\beta t}x_0$. By the remark, its variance $\tilde{\sigma}(t)^2$ is

$$\tilde{\sigma}(t)^2 = \frac{\sigma^2}{2\beta}(1 - e^{-2\beta t}).$$

Then x_t is distribution as $N\left(e^{-\beta t}x_0, \tilde{\sigma}(t)^2\right)$, and for $f : \mathbf{R} \to \mathbf{R}$ bounded measurable,

$$P_t f(x_0) = \mathbf{E}f(x_t) = \int_{\mathbf{R}} f(y) \frac{1}{\sqrt{2\pi}\tilde{\sigma}(t)} e^{-\frac{(y-e^{-\beta t}x_0)^2}{2\tilde{\sigma}(t)^2}} dy.$$

Invariant measure for the Ornstein-Uhlenbeck process

Very often we would like to find a candidate or a formula for invariant measures. One way to find an ansatz is to use the generator $\mathcal{L}$. Suppose that its invariant measure is absolutely continuous w.r.t the Lebesgue measure, say $\pi = g(x)dx$ with $g(x)$ the density (i.e. the Radon-Nikodym derivative). Then

$$\mathcal{L}^* g = 0$$

where $\mathcal{L}^*$ is the L^2 adjoint of $\mathcal{L}$ w.r.t. dx.

Exercise 6.6.4 Let $\mathcal{L}f(x) = -\beta x \frac{d}{dx} + \frac{1}{2}\sigma^2 \frac{d^2}{dx^2}$. Show that

$$\mathcal{L}^* g(x) = -\frac{d(\beta x g(x))}{dx} + \frac{1}{2}\sigma^2 \frac{d^2 g(x)}{dx^2}.$$

Problem. Verify that $N(0, \frac{\sigma^2}{2\beta})$ is the invariant probability measure for the Ornstein-Uhlenbeck process.

6.7 Ergodic Theorem

In the case of continuous-time Markov processes, it is however often convenient to formulate Lyapunov-Foster type conditions in terms of the generator $\mathcal{L}$ of the process. Formally, one has $\mathcal{L} = \partial_t P_t|_{t=0}$, but it turns out that the natural domain of the generator with this definition may be too restrictive for our usage. We therefore take a rather pragmatic view of the definition of the generator $\mathcal{L}$ of a Markov process, in the sense that writing

$$\mathcal{L}F = G\ ,$$

is considered to be merely a shorthand notation for the statement that
the process $F(x_t, t) - \int_0^t G(x_s, s)\, ds$ is a martingale for every initial condition x_0. Similarly,

$$\mathcal{L}F \leq G\ ,$$

is a shorthand notation for the statement that $F(x_t, t) - \int_0^t G(x_s, s)\, ds$ is a supermartingale for every x_0.

Definition 6.7.1 An adapted stochastic process $(X_t, t \in I)$ is an (integrable) super-martingale if $\mathbf{E}|X_t| < \infty$ and $\mathbf{E}(X_t|\mathcal{F}_s) \leq X_s$ for all $s \leq t$.

Remark 6.7.2 It is possible to have $\mathcal{L}F \leq G$ even in situations where there does not exist any function H such that $\mathcal{L}F = H$. Think of the case $F(x) = -|x|$ when the process x_t is a Brownian motion. There, one has $\mathcal{L}F \leq 0$, but F does not belong to the domain of the generator, even in the weakened sense described above.

It is often sufficient to study P_t at time $t = 1$.

Notation. We will write $P = P_1$.

Suppose that P satisfies the following geometric drift condition:

Assumption 6.7.1 *There exists a function $V : \mathcal{X} \to [0, \infty)$ and constants $K \geq 0$ and $\gamma \in (0, 1)$ such that*

$$(PV)(x) \leq \gamma V(x) + K \;, \tag{6.10}$$

for all $x \in \mathcal{X}$.

Remark 6.7.3 One could allow V to also take the value $+\infty$. However, since we do not assume any particular structure on $\mathcal{X}$, this case can immediately be reduced to the present case by simply replacing $\mathcal{X}$ by $\{x : V(x) < \infty\}$.

A sufficient condition for Assumption 6.7.1 to hold is that there exists a measurable function $V : \mathcal{X} \to [0, \infty)$ and positive constants c, K such that

$$\mathcal{L}V \leq K - cV.$$

One might ask which of the two conditions, $\mathcal{L}V \leq K - cV$ and $(PV)(x) \leq \gamma V(x) + K$, are easier to verify? The answer to this really depends on the problem at hand.

Exercise 6.7.1 Find a Lyapunov function for the Ornstein-Uhlenbeck process.

Hint Try x^2.

Assumption 6.7.1 ensures that the dynamic enters the “centre” of the state space regularly with tight control on the length of the excursions from the centre. We now assume that a sufficiently large level set of V is sufficiently “nice” in the sense that we have a uniform “minorisation” condition reminiscent of Doeblin’s condition, but localised to the sublevel sets of V.

Döeblin's condition is: There exists a constant $\alpha > 0$ such that $\|P(x,\cdot) - P(y,\cdot)\|_{\mathrm{TV}} \le 2(1-\alpha)$ for every x and y.

Assumption 6.7.2 *[local Doeblin's condition] For every $R > 0$, there exists a constant $\alpha > 0$ so that*

$$\|P(x,\cdot) - P(y,\cdot)\|_{\mathrm{TV}} \le 2(1-\alpha)\;, \tag{6.11}$$

for all x, y such that $V(x) + V(y) \le R$.

An alternative way of formulating (6.11) is to say that the bound

$$|P\varphi(x) - P\varphi(y)| \le 2(1-\alpha)\;,$$

holds uniformly over all functions φ with absolute value bounded by 1.

Exercise 6.7.2 Show that for the Brownian motion, Assumption 6.7.2 holds if V has bounded sub-level sets.

We note that if μ_i have densities ϱ_i with respect to Lebesgue measure such that $\varrho_i \ge c > 0$ on $[0,1]$ (the choice of the interval is not important), then $\|\mu_1 - \mu_2\|_{\mathrm{TV}} \le 2(1-c)$ by (5.10). The BM has densities given by $P(x,dz) = C\exp(-\frac{|z-x|^2}{2})$, which is continuous and strictly positive. It therefore achieves a strictly positive minimum over any bounded set, whence the claim follows.

Remark 6.7.4 Observe that if the transition probabilities has a density w.r.t. the same reference measure, the total variation norm $\|P(x,\cdot) - P(y,\cdot)\|_{\mathrm{TV}}$ can be computed by the L_1 norm of the difference of their densities, for this please refer to the alternative definition (5.10) for the total variation norm.

$$\|\mu - \nu\|_{\mathrm{TV}} = \int_\Omega \left|\frac{d\mu}{d\eta}(w) - \frac{d\nu}{d\eta}(w)\right| \eta(dw)\;.$$

This leads to the convenient formula given in (5.11):

$$\|\mu - \nu\|_{\mathrm{TV}} = 2 - 2\big(\mu \wedge \nu\big)(\Omega)\;.$$

where $\mu \wedge \nu$ can be defined via any positive measure η on $\mathcal{X}$ s.t. $\mu \ll \eta$ and $\nu \ll \eta$ as follows:

$$\big(\mu \wedge \nu\big)(A) = \int_A \min\left\{\frac{d\mu}{d\eta}(w), \frac{d\nu}{d\eta}(w)\right\} \eta(dw)\;.$$

This is convenient for measures on $\mathbf{R}^n$ with densities with respect to the Lebesgue measure.

In order to state the version of Harris' theorem under consideration, we introduce the following weighted supremum norm:

$$\|\varphi\| = \sup_x \frac{|\varphi(x)|}{1 + V(x)}\;. \tag{6.12}$$

With this notation at hand, one has:

Theorem 6.7.5 *If Assumption 6.7.1 (Lyapunov function condition) and Assumption 6.7.2 (local Doeblin condition) hold, then P admits a unique invariant measure π. For every x, $P_t(x,\cdot)$ converges to π in the total variation norm.*

Furthermore, there exist constants $C>0$ and $\varrho \in (0,1)$ such that the bound

$$\left\|P^n\varphi(\cdot) - \int_{\mathcal{X}} \varphi\, d\pi\right\| \le C\varrho^n \left\|\varphi - \int_{\mathcal{X}} \varphi\, d\pi\right\| \tag{6.13}$$

holds for every measurable function $\varphi\colon \mathcal{X} \to \mathbf{R}$ such that $\|\varphi\| < \infty$.

For a proof see for example Section 15 in Meyn-Tweedie's book on Markov chain's. A clean proof is given in the notes on 'Convergence of Markov Processes' by Martin Hairer (online).

Remark 6.7.6 In the duality definition for the total variation norm, the supremum is taken over bounded measurable functions with bound 1. Since $\|\varphi\| \le |\varphi|_\infty$, (6.13) implies that the transition probabilities converge in the total variation norm:

$$\|P_n(x,\cdot) - \pi\|_{TV} \le C\varrho^n\,(1+V(x)) \sup_{|\varphi|_\infty \le 1} \left\|\varphi - \int_{\mathcal{X}} \varphi\, d\pi\right\|,$$

for every $x \in \mathcal{X}$. Thus for every x, the convergence in the total variation norm is exponentially fast.

Remark 6.7.7 The inequality (6.13) can be passed to P_t, where $t \in \mathbf{R}_+$ as below.

$$\left\|P_t\varphi(\cdot) - \int_{\mathcal{X}} \varphi\, d\pi\right\| \le C\varrho^{[t]} \left\|\varphi - \int_{\mathcal{X}} \varphi\, d\pi\right\| \tag{6.14}$$

Indeed, to pass from the convergence of P_n to P_t, we use the properties of the semi-group. Firstly, $P_t = P_{t-[t]}P_{[t]}$, where $[t]$ is the integer part of t, and then

$$\left|P_t\varphi(x) - \int_{\mathcal{X}} \varphi\, d\pi\right| = \left|P_{t-[t]}\left(P_{[t]}\varphi(x) - \int_{\mathcal{X}} \varphi\, d\pi\right)\right| \le \left|P_{[t]}\varphi(x) - \int_{\mathcal{X}} \varphi\, d\pi\right| \to 0.$$

Remark 6.7.8 If local Döeblin's condition is strengthened to Doeblin's condition, then the conclusion becomes:

$$\sup_{x\in\mathcal{X}} \left|P^n\varphi(x) - \int_{\mathcal{X}} \varphi\, d\pi\right| \le C\varrho^n \cdot \left|\varphi - \int_{\mathcal{X}} \varphi\, d\pi\right|_\infty .$$

In particular, $\|P_n(x,\cdot) - \varphi\|_{TV} \le C\varrho^n \sup_{|\varphi|_\infty \le 1} |\varphi - \int_{\mathcal{X}} \varphi\, d\pi|_\infty$.

We close this section by remarking that there is also a structure theorem and Birkhoff's ergodic theorem for continuous time Markov processes. A rich family of valuable examples of sample continuous Markov processes on continuous state space come from stochastic differential equations.

Reference. ' Convergence of Markov Processes' by Martin Hairer (online).

Chapter 7

Appendix

7.1 Measures on metric spaces

A metric space is compact if any covering of it by open sets has a sub-covering of finite open sets. A discrete metric space (whose subsets are all open sets) is compact if and only if it is finite. (e.g. **Z** and **N** with the usual distance $d(x, y) = |x - y|$ is not compact.) A subset of a metric space is compact if it is compact as a metric space with the induced metric. It is relatively compact if its closure is compact. A metric space is sequentially compact if every sequence of its elements has a convergent sub-sequence (with limit in the metric space of course). It is totally bounded if for any $\epsilon > 0$ it has a finite covering by open balls of side ϵ. A metric space is complete if every Cauchy sequence converges.

It is a theorem that a metric space is compact if and only if it is complete and totally bounded. A metric space is compact if and only if it is sequentially compact.

A subset of a metric space is relatively compact if it is sequentially compact (the limit may not need to belong to the subset).

If $\{x_n\}$ is sequentially compact with common limit, then it must converges. Suppose the limit is $\bar{x}$. If $x_n \not\to \bar{x}$, then there exists $\epsilon > 0$ such that for any k, there exists $n_k > k$, with $d(x_{n_k}, \bar{x}) >\geq \epsilon$. No subsequence of $\{x_{n_k}\}$ would converge to $\bar{x}$! Hence the contradiction.

7.1.1 Borel measures and approximations

One nice property of the metric space is the fact that any Borel probability measure μ on it is regular: if A is a Borel set then

$$\mu(A) = \sup\{\mu(F) : F \subset A \text{ and } F \text{ is closed}\} \ = \inf\{\mu(U) : A \subset U \text{ and } U \text{ is open}\}.$$

Theorem 7.1.1 *Let μ and ν be two probability measures on a metric space such that*

$$\int f d\mu = \int f d\nu$$

for every bounded uniformly continuous function f on $\mathcal{X}$, then $\mu = \nu$.

Theorem 7.1.2 *Let $1 \leq p < \infty$ and μ a probability measure on a metric space. The set, $C_c(\mathcal{X})$, of continuous functions with compact support, is dense in $L_p(\mathcal{X})$.*

Theorem 7.1.3 (Lusin's Theorem) *Let μ be a probability measure on a metric space $\mathcal{X}$ and $f : \mathcal{X} \to \mathbf{R}$ is a measurable function that vanishes outside of a set of full measure. Then for any $\epsilon > 0$, there exists a continuous function φ_ϵ with compact support such that φ_ϵ agree with f on a set of measure $1 - \epsilon$. If f is bounded we can choose φ_ϵ with $|\varphi_\epsilon|_\infty \leq |f|_\infty$.*

Theorem 7.1.4 *If f is lower semi-continuous and non-negative, e.g. the indicator function of an open set, and μ a probability measure, then*

$$\int f d\mu = \sup \left\{ \int \varphi d\mu : 0 \leq \varphi \leq f, \varphi \in C_c(\mathcal{X}) \right\}.$$

7.1.2 On a compact metric space

A linear functional on $C(\mathcal{X})$ is a linear map $L : \mathbf{C}(X) \to \mathbf{R}$, it is said to be positive if $L(f) \geq 0$ whenever $f \geq 0$.

Theorem 7.1.5 *Let $\mathcal{X}$ be a compact metric space and L a positive linear functional on $\mathcal{X}$ with the property that $L(1) = 1$. Then there exists a unique Borel probability measure μ on $\mathcal{X}$ such that $L(f) = \int f d\mu$ for all $f \in C(\mathcal{X})$.*

7.1.3 On a separable metric space

If $\mathcal{X}$ is a separable metric space, there exists a countable family of open sets $\mathcal{C}$ such that every open set is the union of sets from $\mathcal{C}$, in particular $\mathcal{B}(\mathcal{X}) = \sigma(\mathcal{C})$.

Definition 7.1.6 If $\mathcal{X}$ is a separable metric space, then for any probability measure on X there exists a closed set A such that A is the smallest closed set of full measure. Furthermore A is the set of points with the property that any open set containing it has positive measure. This set is called the support of μ.

The topology of weak convergence on $\mathbf{P}(\mathcal{X})$ has the following neighbourhood basis. For any finite set of continuous functions $\{\varphi_i, i = 1, \ldots, n\}$, any $n \in \mathbf{N}$ and $\varphi_i \in C_b(\mathcal{X})$, and $\mu_0 \in \mathbf{P}(\mathcal{X})$,

$$\left\{ \mu \in \mathbf{P}(\mathcal{X}) : \left| \int \varphi_i d\mu - \int \varphi_i d\mu_0 \right| \leq \epsilon, \forall \varphi_i \right\}.$$

Proposition 7.1.7 *Let $\mathcal{X}$ be a complete separable metric space. Then we can construct an equivalent metric on $\mathcal{X}$ such that there exists a sequence of bounded uniformly continuous functions $\{\varphi_k\}$ with the following property: for any sequence of probability measures μ_n, μ_n converges to μ weakly if and only if $\int \varphi_k d\mu_n \to \int \varphi_k d\mu$ for every k.*

7.1.4 On a complete separable metric space

If $\mathcal{X}_1$, $\mathcal{X}_2$ are complete separable metric spaces and $\varphi : \mathcal{X}_1 \to \mathcal{X}_2$ a one to one measurable map then the image of a Borel subset of $\mathcal{X}_1$ by φ is a Borel subset of $\mathcal{X}_2$.

Theorem 7.1.8 *Every probability measure on a complete separable metric space is tight.*

Proposition 7.1.9 *Let $P(\mathcal{X})$ denotes the space of probability measures on a metric space $\mathcal{X}$.*

1. *The space $P(\mathcal{X})$ with the weak topology, is metrizable as a separable metric space if and only if $\mathcal{X}$ is a separable metric space.*

2. *If $\mathcal{X}$ is a separable metric space, then $P(X)$ is complete as a topological space if and only if $\mathcal{X}$ is complete.*

3. *Also, $P(\mathcal{X})$ is compact if and only if $\mathcal{X}$ is.*

For further reading we refer to the brilliant book by K. R. Parthasarathy.

7.1.5 Measures on C

A special interesting space for those working with stochastic processes with continuous time and with sample continuous paths is the space of continuous functions with the uniform norm over a set $\mathcal{X}$ is an infinite dimensional space. The unit ball in a metric space is compact if and only if the space if finite dimensional. A subset of $C(\mathcal{X})$ is compact if and only if it is totally bounded and equi-continuous. This follows from the Arzela-Ascoli theorem that states: if a family of continuous functions are bounded and equi-continuous, then it has a uniformly convergent sub-sequence.

Definition 7.1.10 A subset A of $C(\mathcal{X})$ is said to be equicontinuous at a point x if for any $\epsilon > 0$ there exists $a > 0$ such that

$$|f(y) - f(x)| \leq \epsilon$$

for every $f \in A$ and for every $y \in B_a(x)$.

Proposition 7.1.11 *Let $\mathcal{X}$ be a separable metric space and μ_n be any sequence of probability measures on $\mathcal{X}$. Then $\mu_n \to \mu$ weakly if and only if*

$$\lim_{n\to\infty} \sup_{f\in A} \left| \int f d\mu_n - \int f d\mu \right| = 0$$

for every family $A \subset \mathbf{C}(\mathcal{X})$ which is equi-continuous at all points of $\mathcal{X}$ and uniformly bounded.

Proposition 7.1.12 *A subset A of $C([0,1];\mathbf{R})$ is relatively compact it is sufficient and necessary that the following two conditions are satisfied:*

1. $\sup_{x\in A} |x(0)| < \infty$
2. $$\lim_{\delta\to 0} \sup_{x\in A} \omega_x(\delta) = 0,$$
   where $\omega_x(\delta) = \sup_{|s-t|<\delta} |x(s) - x(t)|$.

Theorem 7.1.13 *Let $\mathcal{M}$ be a family of probability measures on $C([0,1];\mathbf{R})$. Then $\mathcal{M}$ is compact if and only if the following conditions are satisfied. For any $\epsilon > 0$ there exists a number M and a function $\lambda : \mathbf{R}_+ \to \mathbf{R}_+$ which decreases to zero (M and λ may depend on ϵ),*

(1) such that

$$\mu(x \in C : |x(0)| \le M) \ge 1 - \epsilon, \quad \forall \mu \in \mathcal{M};$$

(2)

$$\mu\left(\{x : \omega_x(\delta) \le \lambda(\delta) \textit{ for all } \delta\}\right) > 1 - \epsilon; \quad \forall \mu \in \mathcal{M}.$$

Theorem 7.1.14 *Let A be a family of probability measures on $C([0,1];\mathbf{R})$. Then A is compact if and only if the following conditions are satisfied.*

(1) For any $\epsilon > 0$ there exists a number M such that

$$\mu(x \in C : |x(0)| \le M) \ge 1 - \epsilon, \quad \forall \mu \in \mathcal{M}$$

(2') For any $\epsilon > 0$ and $\delta > 0$ there exist a number $\eta > 0$ (which may depend on ϵ and δ) such that

$$\mu\Big(\{x : \omega_x(\eta) \le \delta\} \Big) > 1 - \epsilon, \quad \forall \mu \in \mathcal{M}.$$

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

$$\lim_{n\to\infty} \int_{\mathcal{X}} f(x)\, \mu_n(dx) = \int_{\mathcal{X}} f(x)\, \mu(dx) \;, \tag{7.1}$$

for every $f \in \mathcal{C}_b(\mathcal{X})$. We say that it converges **strongly** if (7.1) holds for every $f \in \mathcal{B}_b(\mathcal{X})$.

7.2 Examples

Example 7.2.1 The interval $[0,1]$ equipped with its Borel σ-algebra and the Lebesgue measure is a probability space.

Example 7.2.2 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 7.2.3 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 7.2.4 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 7.2.5 Let Ω be the unit interval and define the probability measures

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

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 7.2.6 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).

7.3 Proof of Prohorov's theorem

Theorem 7.3.1 (Prohorov) *A sequence of probability measures on a complete separable metric space $\mathcal{X}$ is relatively compact if and only if it is tight.*

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

Lemma 7.3.2 *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 subsequence 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 $\bigl|\int \varphi(x)\mu_{n_k}(dx) - \int \varphi(x)\mu(dx)\bigr| < 7\varepsilon$ for every $k \geq N$. Choose R as above and note that the tightness condition implies that

$$\Bigl|\int \varphi(x)\mu_{n_k}(dx) - \int_{-R}^{R} \varphi(x)\mu_{n_k}(dx)\Bigr| \leq 2\varepsilon\;, \tag{7.2}$$

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 $\tilde\varphi\colon$ 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 $\bigl|\int_{-R}^{R} \tilde\varphi(x)\mu_{n_k}(dx) - \int_{-R}^{R} \tilde\varphi(x)\mu(dx)\bigr| < \varepsilon$ for every $k \geq N$. Putting these bounds together yields

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

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

7.4 Strong Feller

If a Markov transition function $P(x,dy)$ is continuous in the total variation norm, then the transition semigroup is strong Feller. The former is stronger, because the convergenece

$$\lim_{x\to x_0} \sup_A |T_t \mathbf{1}_A(x) -_t \mathbf{1}_A(x_0))| = \lim_{x\to x_0} \sup_A |P(x,A) - P(x_0,A)| = 0,$$

is uniform in the set A.

There is a theorem which states that the composition of two strong Feller Markov kernels is continuous in the total variation norm. By the Chapman-Kolmogorov equations, a continuous time strong Feller Markov semigroup is continuous in total variation norm as soon as the time is positive. There are counter example of strong Feller Markov processes not continuous in the total variation norm. See noted by Martin Hairer and notes by Jan Sedler.

Bibliography

[1] Hans Föllmer, Ching-Tang Wu, and Marc Yor. On weak Brownian motions of arbitrary order. *Ann. Inst. H. Poincaré Probab. Statist.*, 36(4):447–487, 2000.

- Real Analysis by Royden, Chapter 11 (especially section 3) in the third edition for integration.
- Probability by Leo Breiman, Conditional Expectation is in Chapter 4.
- Probability measures on metric spaces, K. R. Parthasarathy.
- Real Analysis, G. B. Folland
- Measures, Integrals and Martingales, R. Schilling

- Markov Chains and Stochastic Stability by Meyn and Tweedie
- Markov Chain by J. Norris
- Markov Chains and Mixing Times, second edition by Levin Peres (available on line)
- Markov processes and applications by E. Pardoux
- Markov processes : Characterization and convergence by Ethier and Kurtz (Very advanced)