

Markov Processes

Martin Hairer and Xue-Mei Li

September 12, 2022

©Martin Hairer and Xue-Mei Li (2021). These notes are provided for the personal study of students taking this module.

Contents

1	Introduction	9
1.1	Introduction	9
1.1.1	Evolution of probability distributions	11
1.1.2	Questions	11
1.2	References	12
2	The Markov property	13
2.1	Information and Filtration	14
2.2	The Markov Property	14
2.2.1	Conditional expectation	15
2.2.2	Markov property on discrete state space	16
2.2.3	A random dynamical system	17
2.2.4	Random walk on Z and on Z_+	18
2.2.5	A non-Markov random walk	19
2.2.6	Examples	19
3	Markov Property	21
3.1	The Markov Property	21
3.1.1	Equivalent definitions for the Markov property	23
3.2	Gaussian Measure and Gaussian Processes	26

4	Kolmogorov's Extension Theorem	27
4.1	Pushed forward measures	27
4.2	Finite dimensional distributions	28
4.3	Construction of random variables	29
4.4	Kolmogorov's extension theorem	30
4.5	Canonical stochastic process, canonical probability space	31
4.6	Stationary Processes	31
5	Markov Processes With Transition Probabilities	33
5.1	Transition Probabilities	33
5.1.1	Transition functions and Chapman-Kolmogorov equations	35
5.1.2	Construction of transition function from transition probabilities	35
5.1.3	Markov chain with transition functions/ transition kernels	36
5.1.4	Existence of Markov Chains with given transition probabilities	39
5.2	Examples and Exercises	40
5.3	Transition operators and invariant measures	42
5.3.1	Stationary Markov chain	43
5.4	Example: Markov Chains On Discrete State Spaces	43
5.4.1	Stochastic Matrix	43
5.4.2	N-step Transitions	44
5.4.3	The Markov property is determined by finite dimensional distributions	46
5.4.4	Conditional independence of the future and the past	47
5.4.5	Operator on Measures	48
5.4.6	Example: Two State Markov Chains	48
6	Strong Markov Property	51
6.1	Stopping Times	51
6.2	The Stopped σ -algebra	53

6.3	Strong Markov Property	58
6.3.1	Markov property at finite stopping times	59
6.3.2	Markov property at non-finite stopping times	61
7	Time Homogeneous Markov Chains on Discrete State Spaces	65
7.1	Communication Classes - Lecture 8	66
7.2	Recurrence and Transience	70
7.2.1	Another way for guessing the invariant measures –Lecture 9	71
7.3	Passage times	72
7.4	Recurrence Criterion	77
7.5	Irreducibility and reduced Markov Chains	82
7.6	Construction of invariant measure from recurrent state	83
7.7	The long run probabilities	87
7.7.1	Return Times and Aperiodicity	88
7.7.2	Ergodic Theorem	89
7.7.3	The total variation distance	94
7.7.4	Convergence Theorem in Total Variation	95
7.7.5	Periodic Chains, Cycles	96
7.7.6	Invariant measure for periodic chains	98
7.8	Ergodic Theorem: The law of Large Numbers	99
7.9	Reversible Markov Chains	104
7.9.1	Application : Numerical Simulation	106
7.9.2	Examples	107
7.10	Finite State Markov Chain	108
7.10.1	Characterising aperiodic irreducible chains	108
7.10.2	Perron-Frobenius Theorem	111
7.10.3	The Structure Theorem for Invariant Measures	116
7.10.4	Rate of Convergence	117

7.10.5	The long run probability for reducible chains	120
7.10.6	Examples	121
7.11	A summary	123
8	Invariant measures in the general case	127
8.1	Weak convergence	128
8.2	Feller and Strong Feller Property	130
8.3	Weak convergence and Prokhorov's theorem	131
8.4	Existence of Invariant Measures	132
8.4.1	Lyapunov Function test	135
8.4.2	Application to a random dynamical system	138
8.5	Uniqueness of the invariant measure	139
8.5.1	Properties of couplings	139
8.5.2	Uniqueness due to deterministic contraction	141
8.6	Uniqueness and minorisation	142
8.6.1	Properties of the Total Variation	143
8.6.2	Uniqueness by minorisation	145
8.6.3	Strong Feller	148
8.7	Using P -invariant sets	152
8.7.1	ODEs and Random Dynamical Systems	156
8.7.2	Example	160
9	The Structure Theorem and Ergodic Theorem (Mastery Material)	165
9.1	Ergodic theory for dynamical systems	165
9.2	Dynamical Systems induced by Markov chains	167
9.2.1	Sequence spaces and the shift operator θ	167
9.2.2	Stationary Markov chains as dynamical systems	169
9.2.3	Construction of two sided stationary Markov chains	171

9.2.4	Proof of two sided Markov chains construction	171
9.2.5	Birkhoff's ergodic theorem for Markov Chains	174
9.3	Structure Theorem	176
9.3.1	The statements	177
9.3.2	Proof of the Structure Theorem	177
9.3.3	Proof of Birkhoff's Ergodic Theorem	182
9.3.4	Example	184
10	Appendix	193
10.1	Time reversal on general state space	193
10.1.1	Reversible Process**	194
10.2	Metric and topological spaces: a review	198
10.3	Measures on metric spaces	199
10.3.1	Borel measures and approximations	200
10.3.2	On a compact metric space	201
10.3.3	On a separable metric space	201
10.3.4	On a complete separable metric space	201
10.3.5	Measures on C	202
10.4	The total variation norm	203
10.5	Examples	205
10.6	Proof of Prohorov's theorem	205
10.7	Useful References	207

Chapter 1

Introduction

1.1 Introduction

A stochastic process describes the evolution in time of a stochastic system. For discrete times we tend to use the notation (x_n) where $n = 0, 1, 2, \dots$. The ‘randomness’ comes from the lack of complete information about the system.

We will have:

- an underlying probability space: $(\Omega, \mathcal{F}, \mathbb{P})$,
- a state space \mathcal{X} ,
- the Borel σ -algebra $\mathcal{B}(\mathcal{X})$.

The state space, which we denote by \mathcal{X} , will be assumed to be a separable complete metric space. The Euclidean spaces \mathbf{R}^n are separable complete metric spaces, so can be open sets of \mathbf{R}^n be given a complete metric space which is separable.

A (discrete time) stochastic process (x_n) is a collection of random variable on some probability space, where n is perceived as time. The ‘time’ in a continuous time process (x_t) takes values in an interval. We focus mainly on the case of discrete time processes and therefore give the definition below.

A sample of a process is a function of time (a sequence), by the random nature, we cannot say much about a sample, we can say something about any statistics of its observables, which can be deduced with its probability distributions. To obtain informations on a stochastic process, for example the averages of an observable of the process, e.g $\mathbf{E}[f(x_n)]$, one assumes naturally that the x_n ’s are random variables.

Definition 1.1.1 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 $(\Omega, \mathcal{F}, \mathbb{P})$. Given n , we refer to x_n as the value of the process at time n . We will sometimes consider processes where the time \mathcal{X} -valued random variables can take negative values, *i.e.* $\{x_n\}_{n \in \mathbb{Z}}$.

If the time index I is an interval, then a stochastic process (x_t) where $t \in I$ is again a collection of $\{x_t, t \in I\}$, on which we often make regularity assumptions in t .

Recall that a random variable $X : \Omega \rightarrow \mathcal{X}$ is simply a measurable map, its probability distribution is the pushed forward measure $X_*(\mathbb{P})$:

$$\begin{array}{ccc} X : & \Omega & \longrightarrow \mathcal{X} \\ & \mathbb{P} & \longmapsto X_*\mathbb{P} \end{array}$$

An example of a sequence of random variables 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 Markov process describes the time-evolution of random systems that do not have any memory. 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, given the present, 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:

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

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

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

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

1.1.1 Evolution of probability distributions

The transitions

$$x_0 \rightarrow x_1 \rightarrow x_2 \rightarrow \dots$$

induce a family of probability measures on the state space. There are many interesting questions on Markov processes. We are concerned with the following:

1.1.2 Questions

What is the probability that the Markov chain visit state j at time n given that $x_0 = i$? Does the time n distribution of (x_n) converge to some measure as $n \rightarrow \infty$? In what sense does it converge? What distance does one put on the space of probability measures? At what speed does the convergence happen?

Is there an initial probability distribution μ , such that for each n , x_n is distributed as x_0 ?

Definition 1.1.2 A measure π is called an invariant (probability) measure (invariant distribution) for a Markov chain if $\mathcal{L}(X_n) = \pi$ for any n .

We have discussed whether $\mu_n = (x_n)_* \mathbb{P} \rightarrow \pi$, does the distribution of the stochastic process on $\mathcal{X}^{\mathbb{Z}^+}$ converge to that of the chain with an invariant initial distribution π ? If exists, 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

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

What are the techniques for studying these problems? If we denote by P_μ the probability of the chain with initial μ , we ask $P_\mu \rightarrow P_\pi$?

Ergodic theorems: does it hold?

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

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

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

If x_0, ξ_1, ξ_2 are independent random variables on \mathbf{R} with $x_0 \sim N(0, a)$ and $\xi_i \sim N(0, b)$ for all $i = 1, 2, \dots$. Define for a positive number $M > 0$,

$$M(x_{n+1} - x_n) = -x_n + \xi_{n+1}.$$

It is an easy exercise to find an invariant measure for this Markov chain. See 6.2.5.

Exercise 1.1.1 Find out whether there is an invariant measure for the switch Markov process.

1.2 References

- Markov Chains and Mixing Times, by David A. Levin Yuval Peres Elizabeth L. Wilmer
<https://pages.uoregon.edu/dlevin/MARKOV/markovmixing.pdf>
- Markov Chains, James Norris
- Markov Chains and stochastic stability, Meyn and Tweedie (<http://probability.ca/MT/BOOK.pdf>)
- Markov Processes and Applications Algorithms, Networks, Genome and Finance, E. Pardoux
- Brémaud: Markov chains
- Stroock: An introduction to Markov processes
- Markov Processes, A. Eberle, <https://uni-bonn.sciebo.de/s/kzTUFff5FrWGAay#pdfviewer>

Chapter 2

The Markov property

The Markov property states that given information on its present, any further information on its history does not improve the estimates on the whereabouts of the process at a future time.

A stochastic process describes the evolution of a random system. The ‘randomness’ describes the lack of complete information about the system. We have the set-up:

- an underlying probability space: $(\Omega, \mathcal{F}, \mathbb{P})$, e.g. $([0, 1], \mathcal{B}([0, 1]), dx)$.
- a state space \mathcal{X} , a separable complete metric space.
- the Borel σ -algebra $\mathcal{B}(\mathcal{X})$.

Definition 2.0.1 A stochastic process (x_n) is simply a collections of random variables.

Discrete time stochastic process: the time I set is \mathbf{N} , or \mathbf{N}^0 , or \mathbf{Z} . Continuous time Markov process : $I = \mathbf{R}_+, [0, 1]$.

Given a stochastic process (x_n) , we study the evolution of $\mathcal{L}(x_n)$, among other topics.

Each x_n is a measurable map,

$$\begin{aligned} x_n : \Omega &\longrightarrow \mathcal{X}. \\ \omega &\longrightarrow x_n(\omega) \end{aligned}$$

Its probability distribution, denoted by $\mathcal{L}(x_n)$, is the measure obtained by pushing \mathbb{P} forward by x_n defined below:

$$(x_n)_* \mathbb{P}(A) := \mathbb{P}(\{\omega : x_n(\omega) \in A\}).$$

How are $\{x_0, x_1, x_n\}$ correlated? They are independent if $\mathcal{L}((x_0, \dots, x_n)) = \otimes_{i=1}^n \mathcal{L}(x_i)$. In general we do not expect them independent.

2.1 Information and Filtration

Definition 2.1.1 The information on the stochastic process at time n is the σ -algebra of all possible events at this time:

$$\sigma(X_n) = \sigma(\{x_n^{-1}(A) : A \in \mathcal{B}(\mathcal{X})\}).$$

The set A runs through the Borel subsets of \mathcal{X} and $x_n^{-1}(A) := \{\omega : x_n(\omega) \in A\}$.

If $\mathbb{P}(x_n = \pm 1) = \frac{1}{2}$, the collection of sets

$$\{X_n = 1\}, \quad \{X_n = -1\}, \quad \phi, \quad \Omega$$

contain all the information we can possibly have on the random variable.

Definition 2.1.2 Let $\sigma(x_0, \dots, x_n)$ denote the σ -algebra generated by the random variables inside the bracket, it is generated by sets of the form:

$$\{x_0 \in A_0, \dots, x_n \in A_n\},$$

where $A_i \in \mathcal{B}(\mathcal{X})$. This is the smallest σ -algebra such that each of the random variables are measurable.

The information on the process (x_n) up to time n is the σ algebra generated x_0, \dots, x_n .

Definition 2.1.3 • A family of σ -algebras $\{\mathcal{F}_n\}_{n \geq 0}$, satisfying $\mathcal{F}_n \subset \mathcal{F}_m$ whenever $n < m$ and $\mathcal{F}_n \subset \mathcal{F}$, is a filtration (of σ -algebras).

- A stochastic process x_n is said to be adapted to a filtration \mathcal{F}_n if for every n , x_n is measurable with respect to \mathcal{F}_n .
- The filtration $\mathcal{F}_n^{x_\cdot} := \sigma(x_0, \dots, x_n)$ is the natural filtration for (x_n) .

The natural filtration is the smallest filtration to which the process is adapted.

2.2 The Markov Property

Definition 2.2.1 A stochastic process (x_n) with state space \mathcal{X} is said to have the Markov property (with respect to its natural filtration) if for any Borel measurable set A of \mathcal{X} , any $n \geq 0$,

$$\mathbb{P}(x_{n+1} \in A \mid x_0, \dots, x_n) = \mathbb{P}(x_{n+1} \in A \mid x_n) \quad a.s.$$

The distribution of the random variable x_0 is called the initial distribution. Discrete time Markov processes are also called Markov chains.

Notation.

$$\mathbb{P}(x_{n+1} \in A | x_0, \dots, x_n) := \mathbb{P}(x_{n+1} \in A | \sigma(x_0, \dots, x_n)).$$

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

2.2.1 Conditional expectation

Definition 2.2.2 Let X be a real-valued random variable on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ 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

$$\int_A X(\omega) \mathbb{P}(d\omega) = \int_A X'(\omega) \mathbb{P}(d\omega), \quad (2.1)$$

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

Proposition 2.2.3 *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 \mathbb{P} to \mathcal{F}' and define the measure μ on (Ω, \mathcal{F}') by $\mu(A) = \int_A X(\omega) \mathbb{P}(d\omega)$ 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. \square

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

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

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

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

Exercise 2.2.1 Show that a stochastic process (x_n) with state space \mathcal{X} satisfies the simple Markov property if and only if the following holds for any bounded measurable functions $f : \mathcal{X} \rightarrow \mathbf{R}$ such that

$$\mathbf{E}(f(x_{n+1}) | \sigma(x_0, \dots, x_n)) = \mathbf{E}(f(x_{n+1}) | x_n). \quad (2.2)$$

Example 2.2.1 Let $F' = \{A, A^c, \phi, \Omega\}$. Recall: $\mathbf{E}(X|A) = \frac{\mathbf{E}(X1_A)}{\mathbb{P}(A)}$.

$$\mathbf{E}(X|\mathcal{F}')(\omega) = \begin{cases} \mathbf{E}(X|A), & \text{if } \omega \in A \\ \mathbf{E}(X|A^c), & \text{if } \omega \in A^c. \end{cases}$$

If Y is a random variable on \mathcal{Y} , then $\mathbf{E}(X|Y) = \varphi(Y)$ for some Borel measurable function $\varphi : \mathcal{Y} \rightarrow \mathbf{R}$. We write $\mathbf{E}(X|Y = y)$ for the function $\varphi(y)$.

Example 2.2.2 Suppose that Y takes values in a state space in which case we identify it with \mathbf{N} .

$$\mathbf{E}(X|Y)(\omega) = \sum_{i: \mathbb{P}(\{Y=i\}) > 0} \mathbf{E}(X|\{Y=i\})1_{\{Y=i\}}(\omega).$$

i.e. $\varphi(i) = \mathbf{E}(X|\{Y=i\})$. This justify the notation:

$$\mathbf{E}(X|Y = i) := \mathbf{E}(X|\{Y = i\}).$$

2.2.2 Markov property on discrete state space

If x_n has only a finite number of a countable number of states, then we can define the Markov property using elementary probabilities: $\mathbb{P}(A|B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$ when $\mathbb{P}(B) > 0$.

If \mathcal{X} is a discrete state space, a space with at most a countable number of elements, in which case we may let $\mathcal{X} = \mathbf{N}$. Let (x_n) be a stochastic process on a discrete state space \mathcal{X} . The above discussion allows us to reduce the Markov property

$$\mathbb{P}(x_{n+1} \in A \mid x_0, \dots, x_n) = \mathbb{P}(x_{n+1} \in A \mid x_n) \quad a.s.$$

to a simpler expression. The Markov property is equivalent to the following: for any $n = 1, 2, \dots$ and for any $s_1, \dots, s_{n+1} \in \mathcal{X}$ such that $\mathbb{P}(x_0 = s_0, \dots, x_n = s_n) > 0$, we have

$$\mathbb{P}(x_{n+1} = s_{n+1} \mid x_0 = s_0, \dots, x_n = s_n) = \mathbb{P}(x_{n+1} = s_{n+1} \mid x_n = s_n).$$

We may omit mentioning the condition $\mathbb{P}(x_0 = s_0, \dots, x_n = s_n) > 0$.

Notation:

$$\{x_n = k\} = \{\omega : x_n(\omega) = k\}, \quad \{x_0 = s_0, \dots, x_n = s_n\} = \cap_{i=0}^n \{x_i = s_i\}.$$

The event we are conditioning on is:

$$\{\omega : x_0(\omega) = s_0, \dots, x_n(\omega) = s_n\}$$

If \mathcal{X} is a general complete separable metric space with its Borel σ -algebra, it is conceivable that $\mathbb{P}(x_n = s) = 0$ for any $s \in \mathcal{X}$, and so the elementary probability formulation may fail.

2.2.3 A random dynamical system

We first give a proof for a discrete state space, in this case a more elementary proof is available to us.

Example 2.2.3 Suppose that \mathcal{X} is a discrete space and $(\mathcal{Y}, \mathcal{G})$ a measurable space. Let $F : \mathcal{X} \times \mathcal{Y} \rightarrow \mathcal{X}$ be a measurable map. Suppose that (ξ_n) are independent random variables with state space \mathcal{Y} , and is independent of x_0 , (In other words, x_0, ξ_1, ξ_2, \dots are all independent). Set

$$x_{n+1} = F(x_n, \xi_{n+1}), \quad n \geq 0.$$

Then (x_n) is a Markov chain.

Proof. Let $\mathcal{X} = \mathbf{N}$, $x_n = F(x_{n-1}, \xi_n)$. Then,

$$\begin{aligned} & \mathbb{P}(x_{n+1} = j \mid x_0 = i_0, \dots, x_n = i_n) \\ &= \mathbb{P}(F(x_n, \xi_{n+1}) = j \mid x_0 = i_0, \dots, x_n = i_n) \\ &= \mathbb{P}(F(i_n, \xi_{n+1}) = j \mid x_0 = i_0, \dots, x_n = i_n) \\ &= \mathbb{P}(F(i_n, \xi_{n+1}) = j). \end{aligned}$$

In the final line we use the fact that ξ_{n+1} is independent of $\{x_0, x_1, \dots, x_n\}$, and hence the independence of ξ_{n+1} from x_0, \dots, x_n . Similarly,

$$\begin{aligned} \mathbb{P}(x_{n+1} = j \mid x_n = i_n) &= \mathbb{P}(F(x_n, \xi_{n+1}) = j \mid x_n = i_n) = \mathbb{P}(F(i_n, \xi_{n+1}) = j \mid x_n = i_n) \\ &= \mathbb{P}(F(i_n, \xi_{n+1}) = j). \end{aligned}$$

The Markov property holds. □

One can also compute, opening up the conditioning as below:

$$\begin{aligned} \mathbb{P}(x_{n+1} = i_{n+1} \mid x_0 = i_0, \dots, x_n = i_n) &= \frac{\mathbb{P}(x_{n+1} = i_{n+1}, x_0 = i_0, \dots, x_n = i_n)}{\mathbb{P}(x_0 = i_0, \dots, x_n = i_n)} \\ &= \frac{\mathbb{P}(F(x_n, \xi_{n+1}) = i_{n+1}, x_0 = i_0, \dots, x_n = i_n)}{\mathbb{P}(x_0 = i_0, \dots, x_n = i_n)} \end{aligned}$$

To give a proof for other state spaces, we recall the following

Exercise 2.2.2 Let $X : \Omega \rightarrow \mathcal{X}$ and $Y : \Omega \rightarrow \mathcal{Y}$ be random variables with X measurable with respect to $\mathcal{G} \subset \mathcal{F}$ and Y is independent of \mathcal{G} . If $\varphi : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbf{R}$ is a measurable function such that $\varphi(X, Y)$ is integrable, then

$$\mathbf{E}(\varphi(X, Y) \mid \mathcal{G})(\omega) = \mathbf{E}(\varphi(X(\omega), Y)), \quad a.s.$$

Example 2.2.4 Suppose that $\{\xi_1, \xi_2, \dots\}$ are independent with state space \mathcal{Y} and independent of x_0 on a state space \mathcal{X} . Let $F : \mathcal{X} \times \mathcal{Y} \rightarrow \mathcal{X}$ be a measurable function and define x_n recursively as follows:

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

Then (x_n) is a Markov process.

Proof. Let A be an arbitrary set from $\mathcal{B}(\mathcal{X})$. We only need to show that $\mathbb{P}(x_{n+1} \in A | x_0, \dots, x_n)$, is $\sigma(x_n)$ measurable in which case, since on any $A \in \sigma(x_n) \subset \sigma(x_0, \dots, x_n)$, its average and the average of x_{n+1} are the same, it satisfies the requirement for $\mathbb{P}(x_{n+1} \in A | x_n)$. Then, by the uniqueness of the conditional expectations, it is $\mathbb{P}(x_{n+1} \in A | x_n)$.

For any $A \in \mathcal{B}(\mathcal{X})$, we set $\varphi(x, y) = \mathbf{1}_A(F(x, y))$. Then,

$$\begin{aligned} & \mathbb{P}(x_{n+1} \in A | x_0, \dots, x_n)(\omega) \\ &= \mathbf{E}\left(\mathbf{1}_A(F(x_n, \xi_{n+1})) | x_0, \dots, x_n\right)(\omega) = \mathbf{E}(\mathbf{1}_A(F(x_n(\omega), \xi_{n+1}))). \end{aligned}$$

Set $Y(\omega) = \mathbf{E}(\mathbf{1}_A(F(x_n(\omega), \xi_{n+1})))$. Let μ_{n+1} denote the probability distribution of ξ_{n+1} , set $g(x) = \int_{\mathcal{Y}} \mathbf{1}_A(F(x, y)) \mu(dy)$. Then,

$$Y(\omega) = \int_{\mathcal{Y}} \mathbf{1}_A(F(x_n(\omega), y)) \mu(dy) = g(x_n(\omega)).$$

This concludes that $\mathbf{E}(\mathbf{1}_A(F(x_n(\omega), \xi_{n+1})) | x_0, \dots, x_n)$ is $\sigma(x_n)$ measurable, completing the proof. \square

This proof covers of course the previous example.

2.2.4 Random walk on Z and on Z_+

Let $S_n = \sum_{i=1}^n \xi_i$ where ξ_i are i.i.d.'s with

$$\mathbb{P}(\xi_i = 1) = p, \quad \mathbb{P}(\xi_i = -1) = 1 - p.$$

Then $S_{n+1} = S_n + \xi_{n+1}$. Note that $S_0 = 0$. Then

$$\begin{aligned} & \mathbb{P}(S_{n+1} = k | S_1 = i_1, \dots, S_n = i_n) \\ &= \mathbb{P}(\xi_{n+1} = k - i_n) \\ &= \begin{cases} p, & k = i_n + 1, \\ 1 - p, & k = i_n - 1, \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

A similar computation shows that $\mathbb{P}(S_{n+1} = k | S_n = i_n)$ gives the same value.

The following model is used in queuing theory.

Exercise 2.2.3 Random walk on \mathbf{Z}_+ . Let ξ_n be i.i.d.'s with distribution μ and state space \mathbf{R} . Set

$$x_n = [x_{n-1} + \xi_n]_+.$$

Compute the transition probabilities $\mathbb{P}(x_{n+1} = j | x_n = i)$.

2.2.5 A non-Markov random walk

We have used in an essential way of the independence of ξ_i for proving the Markov property of the random walk. What happens if we remove the independence.

Let us take ξ_1 and ξ_2 be independent with

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

Set for $n > 2$,

$$\xi_{n+1} = \begin{cases} 1, & \text{if } \xi_n = 1, \\ \{\pm 1\} \text{ with probability } \frac{1}{2}, & \text{if } \xi_n \neq 1. \end{cases}$$

This walk $\{S_n\}$ is not a Markov process.

2.2.6 Examples

1. 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 $\mathcal{X} = \{s_1, s_2, \dots, s_{64}\}$, each state is one of the 64 squares.
2. 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 its value before the initial time s .
3. Any deterministic sequence x_n satisfies the simple Markov property. The information on a deterministic random variable is trivial: $\{\phi, \Omega\}$. Note that

$$B := \{x_0 \in A_0, \dots, x_{n-1} \in A_{n-1}\}$$

has probability 1 or 0. When B has non-zero probability, knowing it does not add any information on $x_n \in A$. Take for example, $x_n \equiv 1$. We know its value, there is no need to evaluate its past events. In fact, $\mathbb{P}(x_n \in A | x_0, \dots, x_{n-1}) = \delta_{y_n}(A)$ (I use y_n for the value of x_n to make the distinction.) If the values of the deterministic process are all different, this can be turned into a dynamical system $x_{n+1} = f(x_n)$ where $f(x_n)$ is defined to be x_{n+1} , for all $n \geq 0$, and otherwise a value is arbitrarily assigned.

Consider a sequence with the rule $x_{n+1} = f(x_n + x_{n-1})$. If we fix x_0, x_1 this produces a deterministic sequence, which is a Markov chain as explained earlier. When we vary x_0, x_1 , there is no guarantee for the transition mechanisms being the same. T

In fact there are different notions for a Markov chain. By a Markov chain we often refers to a family of Markov processes with a family of initial conditions x_0 . From any initial condition, its evolution is the same. We often use a subscript to describe the initial condition, e.g. $\mathbb{P}_x(x_n \in A)$ indicates we are discussing the process with initial value x .

4. Let $x_{n+1} = \sin(x_n + x_{n-1})$. We can build a stochastic process, which is manifestly a Markov process, of which x_n is a component. Setting $y_n = (x_{n-1}, x_n)$, then

$$y_{n+1} \equiv (y_{n+1}^{(1)}, y_{n+1}^{(2)}) = (x_n, x_{n+1}) = \left(y_n^{(2)}, \sin(y_n^{(1)} + y_n^{(2)})\right).$$

So (x_n) is realised as a component of a Markov process.

We also present a number of continuous time Markov processes.

1. Brownian motions B_t .
2. solutions of the one dimensional stochastic heat equation with white noise.

$$du_t = \frac{1}{2}\Delta u_t + \dot{\xi}_t.$$

3. Solutions of SDEs:

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

¹This marks the end of Week 1 lectures

Chapter 3

Markov Property

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

3.1 The Markov Property

We will introduce a more general notion of Markov property for which we can just as well assume the process taking value in an index set which is not necessarily discrete set.

Let I be a subset of \mathbf{R} , this is usually an interval in \mathbf{R}_+ , or \mathbf{N}^0 or \mathbf{Z} . A filtration is a family of σ -algebras $(\mathcal{F}_s, s \in I)$ on Ω such that $\mathcal{F}_s \subset \mathcal{F}_t \subset \mathcal{F}$, where $s < t$. If $(x_t, t \in I)$ 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 . 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 is the natural filtration of (x_t) .

Definition 3.1.1 A stochastic process (x_t) is said to have the Markov property (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 $s, t \in I$, $t > s$,

$$\mathbb{P}(x_t \in A \mid \mathcal{F}_s) = \mathbb{P}(x_t \in A \mid x_s) \quad a.s. \quad (3.1)$$

This is the same as the statement that for any $C \in \mathcal{F}_s$,

$$\mathbf{E}(\mathbf{1}_A(x_t)\mathbf{1}_C) = \mathbf{E}(\mathbb{P}(x_t \in A \mid x_s)\mathbf{1}_C). \quad (3.2)$$

Proposition 3.1.2 Suppose that a π -system \mathcal{D} generates $\mathcal{B}(\mathcal{X})$. If for any $s, t \geq 0$, and $A \in \mathcal{D}$, (3.2) holds. Then (3.2) is satisfied by all $A \in \mathcal{B}(\mathcal{X})$. 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 C be fixed. Let

$$\mathcal{A} = \{A \in \mathcal{B}(\mathcal{X}) : A \text{ satisfies (3.2)}\}.$$

The set of A satisfying (3.2) contains \mathcal{D} . It is sufficient to show that \mathcal{A} is a λ -system and hence conclude by the $\pi - \lambda$ theorem that $\mathcal{A} = \mathcal{B}(\mathcal{X})$. (1) \mathcal{X} is in \mathcal{A} . (2) If $A \subset B$, $\mathbf{1}_{B \setminus A} = \mathbf{1}_B - \mathbf{1}_A$. Using linearity on the sides of (3.2) 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, $A \in \mathcal{A}$. \square

Proposition 3.1.3 *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. Let $\mathcal{F}_s^x := \sigma(x_r, 0 \leq r \leq s)$. Since $\sigma(x_s) \subset \mathcal{F}_s^x \subset \mathcal{F}_s$, for any $s < t$, A a measurable set in the state space, the following holds almost surely:

$$\begin{aligned} \mathbb{P}(x_{s+t} \in A \mid \mathcal{F}_s^x) &= \mathbf{E}(\mathbb{P}(x_{s+t} \in A \mid \mathcal{F}_s) \mid \mathcal{F}_s^x) \\ &= \mathbf{E}(\mathbb{P}(x_{s+t} \in A \mid x_s) \mid \mathcal{F}_s^x) \\ &= \mathbb{P}(x_{s+t} \in A \mid x_s), \end{aligned}$$

where we applied the tower property. This completes the proof. \square

Theorem 3.1.4 *If (x_t) is a Markov process, with filtration (\mathcal{F}_t) , then for any bounded Borel measurable $f : \mathcal{X} \rightarrow \mathbf{R}$,*

$$\mathbf{E}(f(x_t) \mid \mathcal{F}_s) = \mathbf{E}(f(x_t) \mid x_s). \quad (3.3)$$

Proof. If $f = \mathbf{1}_A$, where A is a measurable subset of \mathcal{X} , then the required identity is exactly (3.1). Let $f = \sum_{i=1}^n a_i \mathbf{1}_{A_i}$, an arbitrary simple function, then by the linearity of taking conditional expectations, (3.3) holds also. If f is non-negative and bounded, then there exists an increasing sequence of non-negative simple functions f_n with limit f . We apply the conditional monotone convergence theorem,

$$\mathbf{E}(f(x_t) \mid \mathcal{F}_s) = \lim_{n \rightarrow \infty} \mathbf{E}(f_n(x_t) \mid \mathcal{F}_s) = \lim_{n \rightarrow \infty} \mathbf{E}(f_n(x_t) \mid x_s) = \mathbf{E}(f(x_t) \mid x_s).$$

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. \square

Before presenting a remark on the filtration to which the Markov property is tested, we recall the following property of taking conditional expectations.

Exercise 3.1.1 Suppose that X, Y, S are random variables with S independent of X and Y . Suppose that X is integrable, show that

$$\mathbf{E}(X \mid Y, S) = \mathbf{E}(X \mid Y), \quad a.s.$$

Remark 3.1.5 One might ask ‘given a stochastic process, why we bother to introduce another filtration other than its natural filtration?’ One reason is the necessity /convenience to study several stochastic processes on the same filtered probability space, so that \mathcal{F}_t contains information on the related stochastic processes of interest simultaneously. The other reason is that we often assume that \mathcal{F}_t is a σ -algebra with other properties, e.g. it contains all null sets –this is not often satisfied by the natural filtration. For continuous parameter filtrations, we also often assume that \mathcal{F}_t is right continuous in time –which is again not often satisfied by the natural filtration.

We have seen, in Proposition 3.1.3, that 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. By the above exercise we can enrich \mathcal{F}_t with the σ -algebra of an independent random variable without changing the Markov property, while a Markov process with respect to its natural filtration may not have the Markov property with respect to a large σ -algebra.

3.1.1 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.

The role played by the future can be exchanged, see the theorem below. Let $\mathcal{B}_b(\mathcal{X})$ denotes the set of functions $f : \mathcal{X} \rightarrow \mathbf{R}$ that is bounded and Borel measurable.

Theorem 3.1.6 *Given a process $\{x_n\}_{n \in \mathbf{N}}$, three indices $\ell < m < n$, the following properties are equivalent:*

- (i) *For every $f \in \mathcal{B}_b(\mathcal{X})$, $\mathbf{E}(f(x_n) | x_\ell, x_m) = \mathbf{E}(f(x_n) | x_m)$.*
- (ii) *For every $g \in \mathcal{B}_b(\mathcal{X})$, $\mathbf{E}(g(x_\ell) | x_m, x_n) = \mathbf{E}(g(x_\ell) | x_m)$.*
- (iii) *For any $f, g \in \mathcal{B}_b(\mathcal{X})$, 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).

$$\begin{aligned} \mathbf{E}(f(x_n)g(x_\ell) | x_m) &\stackrel{\text{tower}}{=} \mathbf{E}(\mathbf{E}(f(x_n)g(x_\ell) | x_m, x_\ell) | x_m) \\ &\stackrel{\text{taking out known}}{=} \mathbf{E}(g(x_\ell)\mathbf{E}(f(x_n) | x_m, x_\ell) | x_m) \\ &\stackrel{(i)}{=} \mathbf{E}(g(x_\ell)\mathbf{E}(f(x_n) | x_m) | x_m) = \mathbf{E}(g(x_\ell) | x_m) \mathbf{E}(f(x_n) | x_m) , \end{aligned}$$

and so (iii) holds.

To show the converse, we test with functions of the form $g(x_m)h(x_\ell)$ where $g, h \in \mathcal{B}_b$.

$$\begin{aligned} \mathbf{E}(f(x_n)g(x_\ell)h(x_m)) &= \mathbf{E}(h(x_m) \mathbf{E}(f(x_n)g(x_\ell) \mid x_m)) \\ &\stackrel{(iii)}{=} \mathbf{E}(h(x_m) \mathbf{E}(f(x_n) \mid x_m) \mathbf{E}(g(x_\ell) \mid x_m)) \\ &= \mathbf{E}(E(g(x_\ell)h(x_m) \mathbf{E}(f(x_n) \mid x_m) \mid x_m)) \\ &= \mathbf{E}(g(x_\ell)h(x_m) \mathbf{E}(f(x_n) \mid x_m)). \end{aligned}$$

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

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

Remark 3.1.7 Every Markov process satisfies the properties of Theorem 3.1.6. It was however proven in [?] that the converse is not true, *i.e.* there exist processes that satisfy the three (equivalent) properties above but fail to be Markov.

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

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

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:

$$\begin{aligned} \mathbf{E}(f(x_{n+1})h(x_n) \mid \mathcal{G}) &= \mathbf{E}(\overbrace{\mathbf{E}(\mathbf{E}(f(x_{n+1})h(x_n) \mid \mathcal{F}_n) \mid \mathcal{F}_{n-1})}^{\mathcal{F}_n} \mid \mathcal{G}) \\ &= \mathbf{E}(\overbrace{\mathbf{E}(h(x_n)\mathbf{E}(f(x_{n+1}) \mid x_n))}^{\mathcal{F}_n} \mid \mathcal{F}_{n-1}) \mid \mathcal{G}). \end{aligned}$$

We have used the Markov property. Since $Y := \mathbf{E}(f(x_{n+1})h(x_n) \mid \mathcal{F}_n)$ is a function of x_n , we may apply again the Markov property this time conditioning Y on \mathcal{F}_{n-1} to obtain a function of x_{n-1} which is \mathcal{G} -measurable. We may now collapsing the conditional expectation on \mathcal{G} , using the tower property again:

$$\mathbf{E}(f(x_{n+1})h(x_n) \mid \mathcal{G}) = \mathbf{E}(Y \mid x_{n-1}).$$

Finally we collapse the conditioning on \mathcal{F}_{n-1} in Y to conclude. \square

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

Corollary 3.1.9 *Let (x_n) be a Markov process, let $t_1 < t_2 < \dots < t_{m-1} < t_m = n - 1$ where $n > 1$ and $t_i \in \mathbf{N}^0$. Let A be a Borel set, then*

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

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

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

Finally,

Proposition 3.1.10 *A stochastic process $(x_n)_{n=0}^\infty$ is a Markov process with respect to its own filtration with state space \mathcal{X} if and only if one of the following conditions holds.*

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

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

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

$$\mathbf{E}(f(x_n) \mid x_0, x_1, \dots, x_{n-1}) = \mathbf{E}(f(x_n) \mid x_{n-1}). \quad (3.4)$$

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

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

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 , and thus holds on \mathcal{F}_s . See Proposition 3.1.2. The Equivalence of (2) with the Markov property is proved earlier. (iii) obviously implies (i). Using the tower property, the Markov property leads to $\mathbf{E}(\pi_{i=1}^n f_i(x_i)) = \mathbf{E}(\pi_{i=1}^{n-1} f_i(x_i) \mathbf{E}(f_n(x_n) \mid \mathcal{F}_{n-1})) = \mathbf{E}(\pi_{i=1}^{n-1} f_i(x_i) \mathbf{E}(f_n(x_n) \mid x_{n-1}))$, and (iii) follows. \square

3.2 Gaussian Measure and Gaussian Processes

I will take a digression to explain Gaussian measures, this will allow us to work with Exercise 6.2.5.

Definition 3.2.1 A measure μ on \mathbf{R}^n is Gaussian if there exists a non-negative symmetric $n \times n$ matrix K and a vector $m \in \mathbf{R}^n$ such that

$$\int_{\mathbf{R}^n} e^{i\langle \lambda, x \rangle} \mu(dx) = e^{i\langle \lambda, m \rangle - \frac{1}{2} \langle K \lambda, \lambda \rangle}.$$

The Gaussian measure has a density with respect to the Lebesgue measure if and only if K is non-degenerate in which case the density is

$$\frac{1}{\sqrt{(2\pi)^n \det(K)}} e^{-\frac{1}{2} \langle K^{-1}(x-m), x-m \rangle}.$$

The vector m is called its mean, and K is called its covariance operator.

We emphasise that a Gaussian measure is determined by its mean and its covariance operator. A random variable with a Gaussian distribution is called a Gaussian random variable.

Theorem 3.2.2 If X is a Gaussian random variable on \mathbf{R}^d with covariance operator K , and $A : \mathbf{R}^d \rightarrow \mathbf{R}^n$ a linear map, then AX is a Gaussian random variable with covariance AKA^T .

Proof. We only need to identify $\mathbf{E}[e^{i\langle \lambda, AX \rangle}]$ for any $\lambda \in \mathbf{R}^n$:

$$\begin{aligned} \mathbf{E}[e^{i\langle \lambda, AX \rangle}] &= \mathbf{E}[e^{i\langle A^T \lambda, X \rangle}] \\ &= e^{i\langle A^T \lambda, m \rangle - \frac{1}{2} \langle K A^T \lambda, A^T \lambda \rangle} \\ &= e^{i\langle \lambda, Am \rangle - \frac{1}{2} \langle AK A^T \lambda, \lambda \rangle}. \end{aligned}$$

This shows that AX is a Gaussian random variable with mean Am and covariance AKA^T . \square

Example 3.2.1 If (X_1, \dots, X_N) is a Gaussian random variable with each component X_i taking values in \mathbf{R}^d , and $a_i \in \mathbf{R}$, then $\sum_{i=1}^N a_i X_i$ is a Gaussian random variable.

There some examples of a random variable $X = (X_1, X_2)$ such that X_1 and X_2 are Gaussian, but X is not Gaussian.

Exercise 3.2.1 If $\{X_1, \dots, X_N\}$ are independent random variables with each X_i Gaussian on \mathbf{R}^d , and $a_i \in \mathbf{R}$, show that $\sum_{i=1}^N a_i X_i$ is a Gaussian random variable.

Chapter 4

Kolmogorov's Extension Theorem

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. The fundamental about a random variable is its probability distributions.

Definition 4.0.1 If $z : \Omega \rightarrow \mathcal{X}$ 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 $z_*(\mathbb{P})$:

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

This is the probability distribution of z .

4.1 Pushed forward measures

Let (x_n) be a stochastic process with state space \mathcal{X} . Let $\mathcal{X}^{\mathbb{N}_0}$ denote the sequence space:

$$\mathcal{X}^{\mathbb{N}_0} = \prod_{i=0}^{\infty} \mathcal{X} = \{(a_0, a_1, a_2, \dots) : a_i \in \mathcal{X}\}.$$

We may consider (x_n) as a map from Ω to $\mathcal{X}^{\mathbb{N}_0}$:

$$\begin{aligned} \Omega &\rightarrow \mathcal{X}^{\mathbb{N}_0} \\ \omega &\mapsto (x_n(\omega)). \end{aligned}$$

Is this a measurable map? To answer this question we should specify a σ -algebra on the product space.

Given any index set Λ and a collection of sets \mathcal{X}_i , we denote by $\prod_{i \in \Lambda} \mathcal{X}_i$ the product space whose elements are of the form $(a_i)_{i \in \Lambda}$ where $a_i \in \mathcal{X}_i$. We denote by π_m the projection maps:

$$\begin{aligned} \pi_m : \prod_{i=1}^{\infty} \mathcal{X}_i &\rightarrow \mathcal{X}_m \\ (a_i)_{i \in \Lambda} &\mapsto a_m. \end{aligned}$$

Definition 4.1.1 Let Λ be an index set and for each $i \in \Lambda$, $(\mathcal{X}_i, \mathcal{F}_i)$ a measurable space. The product σ -algebra $\otimes_{i \in \Lambda} \mathcal{F}_i$ is the smallest σ -algebra on $\prod_{i=1}^{\infty} \mathcal{X}_i$ such that each π_i is measurable.

In other words,

$$\otimes_{i \in \Lambda} \mathcal{F}_i = \sigma\{\pi_m^{-1}(A_m) : A_m \in \mathcal{F}_m, m \in \Lambda\}.$$

The product σ -algebra is generated by cylindrical sets, those are sets of the form $\{\pi_{n_1} \in A_1, \dots, \pi_{n_m} \in A_m\}$, where $n_1 < n_2 < \dots < n_m$ is a set of times and with $A_i \in \mathcal{B}(\mathcal{X}_i)$. Cylindrical sets are of the form $\prod_{i=1}^{\infty} A_i$ in which only a finite number of A_i 's are not the whole space.

The following can be found in Real Analysis by G. B. Folland, pages 22-23.

Proposition 4.1.2 *If we have a countable product space $\prod_{i=1}^{\infty} \mathcal{X}_i$, each factor with a σ -algebra \mathcal{F}_i . Suppose that $\mathcal{F}_i = \sigma(\mathcal{D}_i)$. Then*

$$\otimes_{i=1}^{\infty} \mathcal{F}_i = \sigma(\prod_{i=1}^{\infty} E_i : E_i \in \mathcal{D}_i).$$

If $\mathcal{X}_1, \dots, \mathcal{X}_n$ are separable metric spaces, then the Borel σ -algebra of the product metric space $\prod_{i=1}^n \mathcal{X}_i$ equals the product σ -algebra:

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

Definition 4.1.3 If (x_n) is a stochastic process on \mathcal{X} ,

$$\begin{aligned} \Omega &\rightarrow \mathcal{X}^{\infty} \\ \omega &\mapsto (x_n(\omega)) \end{aligned}$$

is measurable (assuming the product space is equipped with the product σ -algebra). It induces a probability measure on $(\mathcal{X}^{\infty}, \otimes^{\infty} \mathcal{B}(\mathcal{X}))$, which is the distribution of the process.

4.2 Finite dimensional distributions

Similarly, the first n component of the process (x_1, \dots, x_n) is a measurable map from

$$\Omega \rightarrow \prod_{i=1}^n \mathcal{X}.$$

Their joint probability distribution, μ_n , is a measure on $\otimes_{i=1}^n \mathcal{X}$. The collection of measure $\{\mu_n\}$ are the finite dimensional distributions of the process (x_n) .

A similar definition holds for a continuous time stochastic process:

Definition 4.2.1 If (x_t) is a stochastic process with $t \in I$. Then for any $t_1 < \dots < t_n$, $t_i \in I$ and any n , we define a family of probability measures μ_{t_1, \dots, t_n} on $\mathcal{X}^n = \prod_{i=1}^n \mathcal{X}$ to be the measure pushed forward by $(x_{t_1}, \dots, x_{t_n})$:

$$\mu_{t_1, \dots, t_n}(A_1 \times \dots \times A_n) = \mathbb{P}(x_{t_1} \in A_1, \dots, x_{t_n} \in A_n).$$

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 $\{x_1 = i_1, \dots, x_n = i_n\}$.

4.3 Construction of random variables

Let μ be a probability measure on \mathcal{X} . We set $\Omega = \mathcal{X}$, $\mathcal{F} = \mathcal{B}(\mathcal{X})$ and $\mathbb{P} = \mu$. Then the identity map X , $X(\omega) = \omega$, is a random variable with probability distribution \mathbb{P} on \mathcal{X} :

$$\mathbb{P}(X \in A) = \mathbb{P}(A).$$

In other words, the identity map is the canonical realisation for the probability distribution \mathbb{P} .

Similarly, if μ_n is a probability measure on $(\prod_{i=1}^n \mathcal{X}_i, \mathcal{B}(\prod_{i=1}^n \mathcal{X}_i))$, we take the trio as our underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then (π_1, \dots, π_n) is a random variable with values in $\prod_{i=1}^n \mathcal{X}_i$ with probability distribution μ_n . In deed, for any $A_i \in \mathcal{B}(\mathcal{X}_i)$,

$$\mu_n(\pi_1 \in A_1, \dots, \pi_n \in A_n).$$

Definition 4.3.1 A family of random variables (Y_1, \dots, Y_n) is independent if

$$\mathbb{P}(Y_1 \in A_1, \dots, Y_n \in A_n) = \prod_{i=1}^n \mathbb{P}(Y_i \in A_i),$$

for any measurable sets A_i . In other words the random variables are independent if and only if the pushed forward measures of \mathbb{P} by (Y_1, \dots, Y_n) , is the product of the marginal probability distributions.

Example 4.3.1 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:

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

for any $A_i \in \mathcal{B}(\mathcal{X}_i)$. Any random variable, say (X_1, \dots, X_n) , with $\mu_1 \otimes \dots \otimes \mu_n$ as probability distribution has independent components.

4.4 Kolmogorov's extension theorem

Definition 4.4.1 Let, for $n = 0, 1, 2, \dots$, μ_n be a probability measures on \mathcal{X}^n . They are said to satisfy Kolmogorov's consistency conditions if

$$\mu_{n+1}(A_1 \times A_2 \times \dots \times A_n \times \mathcal{X}) = \mu_n(A_1 \times A_2 \times \dots \times A_n), \quad (4.1)$$

for any $n \geq 0$ and any $A_i \in \mathcal{B}(\mathcal{X})$.

Example 4.4.1 Let (x_n) be a stochastic process on \mathcal{X} and $\{\mu_n\}$ its finite dimensional distributions:

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

Since,

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

$\{\mu_n\}$ satisfies the consistency conditions.

Theorem 4.4.2 (Kolmogorov's extension theorem) *Let, for $n \in \mathbf{N}$, $\{\mu_n\}$ be probability measures on \mathcal{X}^n , satisfying Kolmogorov's consistency conditions. Then there exists a unique probability measure μ on \mathcal{X}^∞ such that*

$$\mu_n(A) = \mu(A \times \mathcal{X}^\infty)$$

for any $n \geq 1$ and for any $A \in \mathcal{B}(\mathcal{X}^n)$.

In other words, if $\text{Proj}_n : \mathcal{X}^\infty \rightarrow \mathcal{X}^n$ is the projection to the first n components, $\mu_n = (\text{Proj}_n)_* \mathbb{P}$.

Applying this to Example 4.4.1, we have the following statement:

Corollary 4.4.3 *The finite dimensional distributions of a stochastic process (x_n) on \mathcal{X} determine uniquely its probability distribution on $\otimes_{i=1}^\infty \mathcal{B}(\mathcal{X})$.*

Remark 4.4.4 This extension theorem is formulated slightly different from the usual one in which you have a complete family $\{\mu_J\}$ where J is any finite ordered sub-index set satisfying consistency conditions. It is clear how to construct μ_J if J is a subindex set of $\{1, \dots, n\}$ so that the consistency conditions hold, the μ_J can be constructed from μ_n by filling any component corresponding to the missing index with the whole set. For example at order 2 we have

$$\mu_{1,2}(A \times B) = \mu_2(A \times B), \quad \mu_{\cdot,2}(A) = \mu(\mathcal{X} \times A), \quad \mu_{1,\cdot}(A) = \mu_1(A) = \mu_2(A \times \mathcal{X}).$$

The dot in the notation indicates the missing indexes, so we have one measure on \mathbf{R}^2 , consistent and with it two measures on \mathbf{R}^1 .

At order 3, we have in addition to the above, also $\mu_3 = \mu_{1,2,3}$,

$$\mu_{1,\cdot,3}(A \times B) = A \times \mathcal{X} \times B, \mu_{\cdot,2,3}(A \times B) = \mu(\mathcal{X} \times A \times B), \mu_{\cdot,\cdot,3}(A) = \mu(\mathcal{X} \times \mathcal{X} \times A).$$

At $n = 4$, one can add further the following: $\mu_{1,2,3,4} = \mu_4$,

$$\mu_{\cdot,2,3,4}, \mu_{1,\cdot,3,4}, \mu_{1,2,\cdot,4}, \mu_{\cdot,\cdot,3,4}, \mu_{\cdot,2,\cdot,4}, \mu_{1,\cdot,\cdot,4}, \mu_{\cdot,\cdot,\cdot,4}.$$

4.5 Canonical stochastic process, canonical probability space

Corollary 4.5.1 *Given any consistent family of probability measure $\{\mu_n\}$, there exists a stochastic process with $\{\mu_n\}$ as its probability distribution.*

Proof. Let $\Omega = \Pi_{i=0}^{\infty} \mathcal{X}$, $\mathcal{F} = \otimes_{i=0}^{\infty} \mathcal{B}(\mathcal{X})$. Let \mathbb{P} denote the probability measure determined by $\{\mu_n\}$ with Kolmogorov's theorem. Then $\{\pi_n\}$ is a stochastic process with μ_n as its finite dimensional distributions. Indeed,

$$\mathbb{P}(z : \pi_0(z) \in A_0, \dots, \pi_n(z) \in A_n) = \mathbb{P}(A_0 \times \dots \times A_n) = \mu_n(A_0 \times \dots \times A_n),$$

as required. □

4.6 Stationary Processes

With the notions introduced earlier, we can understand the concept of stationary processes. We define the shift operators θ_n , where $n \in \mathbf{N}$,

$$\begin{aligned} \theta_n : \mathcal{X}^{\mathbf{N}^0} &\rightarrow \mathcal{X}^{\mathbf{N}^0} \\ (a_0, a_1, \dots, 0) &\mapsto (a_n, a_{n+1}, \dots). \end{aligned}$$

In other words, $\theta_n(a_{\cdot}) = (a_{n+}, \cdot)$.

Definition 4.6.1 A stochastic process (x_n) is said to be a stationary process if for any $n \geq 0$, the probability distributions of the stochastic processes x_{\cdot} and $\theta_n x_{\cdot}$ are the same.

A stochastic process is a stationary process if and only if their finite dimensional distributions are invariant under the shifts. For a Gaussian processes this is equivalent to the statement that $\mathbf{E}x_n$ is a constant of n and the covariances $\text{cov}(x_n, x_{n+m})$ are invariant in n .

Chapter 5

Markov Processes With Transition Probabilities

5.1 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 $\mathbb{P}(x_{n+1} \in A | x_n = x)$, these functions are determined only on a set of full measure with respect to \mathbb{P}_{x_n} , the law of x_n . We now assume the time-homogeneous property: these functions are independent of time. We also assume that we can choose versions of $\mathbb{P}(x_n | x_{n-1} = x)$, 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 5.1.1 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 5.1.2 ** 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

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

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

Example 5.1.1 Let us consider the random dynamical system $x_{n+1} = F(x_n, \xi_{n+1})$, from Example 2.2.4, where ξ_n are independent random variables on \mathcal{Y} with probability distribution μ .

Then,

$$\mathbb{P}(x_{n+1} \in A | x_0, \dots, x_n)(\omega) = \mathbb{P}(F(x_n(\omega), \xi_{n+1}) \in A) = \int_{\mathcal{Y}} \mathbf{1}_A(F(x_n(\omega), y)) \mu(dy).$$

Then $\{P(x, A)\}$, where

$$P(x, A) := \int_{\mathcal{Y}} \mathbf{1}_A(F(x, y)) \mu(dy),$$

are transition probabilities, and

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

Note that the transition mechanism $P(x, A)$ is independent of time n . This is a time homogenous Markov process.

Example 5.1.2 Suppose that $\{P(x, A)\}$ is a family of transition probabilities and x_n is a Markov chain with $\mathbb{P}(x_{n+1} \in A | x_n)(\omega) = P(x_n(\omega), A)$. Then for any $f \in \mathcal{B}_b(\mathcal{X})$,

$$\mathbf{E}(f(x_{n+1}) | x_n) = \int_{\mathcal{X}} f(y) P(x_n, dy). \quad (5.1)$$

Furthermore, for almost surely all ω ,

$$\begin{aligned} \mathbb{P}(x_{n+2} \in A | x_n)(\omega) &= \mathbf{E}(\mathbf{E}(x_{n+2} \in A | x_n, x_{n+1}) | x_n)(\omega) \\ &= \mathbf{E}(P(x_n, A) | x_n)(\omega) = \int_{\mathcal{X}} P(y, A) P(x_n(\omega), dy). \end{aligned}$$

We set

$$P^2(x, A) = \int_{\mathcal{X}} P(y, A) P(x, dy).$$

Then $\mathbb{P}(x_{n+2} \in A | x_n) = P^2(x_n, A)$ and

$$\int_{\mathcal{X}} f(y) P^2(x, dy) = \int_{\mathcal{X}} \int_{\mathcal{X}} f(y) P(z, dy) P(x, dz).$$

$$\begin{aligned} \mathbb{P}(x_{n+3} \in A | x_n)(\omega) &= \mathbf{E}(\mathbf{E}(x_{n+3} \in A | x_n, x_{n+1}, x_{n+2}) | x_n)(\omega) \\ &= \mathbf{E}(P(x_{n+2}, A) | x_n)(\omega) = \int_{\mathcal{X}} P(y, A) P^2(x_n(\omega), dy). \end{aligned}$$

Set

$$P^3(x, y) = \int_{\mathcal{X}} P(y, A) P^2(x, dy).$$

By first conditioning x_n , we expect that

$$P^3(x, y) = \int_{\mathcal{X}} P(y, A) P^2(x, dy) = \int_{\mathcal{X}} P^2(y, A) P(x, dy).$$

We will see later this is the Chapman-Kolmogorov equation.

5.1.1 Transition functions and Chapmann-Kolmogorov equations

Definition 5.1.3 A family of transition probabilities $\{P^n(x, \cdot) : x \in \mathcal{X} \mid n = 0, 1, 2, \dots\}$ is a transition function if

- (1) For any n and each $x \in \mathcal{X}$, $P(x, \cdot)$ is a probability measure on \mathcal{X} ;
- (2) For each $A \in \mathcal{B}(\mathcal{X})$ and n , the function $x \mapsto P^n(x, A)$ is Borel measurable.
- (3) $P^0(x, \cdot) = \delta_x$.
- (4) (Chapmann-Kolmogorov equations): For every $n, m \geq 1$,

$$P^{n+m}(x, A) = \int_{\mathcal{X}} P^n(y, A) P^m(x, dy). \quad (5.2)$$

The probabilities P^n are called n -step transition probabilities.

Remark 5.1.4 The Chapmann-Kolmogorov equation (5.2) holds for every $A \in \mathcal{B}(\mathcal{X})$ is equivalent to the statement that for any $f \in \mathcal{B}_b$,

$$\int_{\mathcal{X}} f(z) P^{n+m}(x, dz) = \int_{\mathcal{X}} \left(\int_{\mathcal{X}} f(z) P^n(y, dz) \right) P^m(x, dy). \quad (5.3)$$

5.1.2 Construction of transition function from transition probabilities

As usual, let $\mathcal{B}_b(\mathcal{X})$ denote the set of bounded Borel measurable real valued functions on \mathcal{X} . We first define P^n , then associate with the movement of the Markov chain.

Definition 5.1.5 Given one step probabilities P , set

1. $P^0(x, \cdot) = \delta_x$,
2. $P^1(x, \cdot) = P(x, \cdot)$,
3. For any $n \geq 1$ and $x \in \mathcal{X}$,

$$P^{n+1}(x, A) = \int_{\mathcal{X}} P(y, A) P^n(x, dy), \quad \forall A \in \mathcal{B}(\mathcal{X}) \quad (5.4)$$

Note that for any $n \geq 1$,

$$\int_{\mathcal{X}} P^n(y, A) P^0(x, dy) = P^n(x, A), \quad \int_{\mathcal{X}} P^0(y, A) P^n(x, dy) = P^n(x, A).$$

Proposition 5.1.6 *Given a family of (one-step) transition probabilities $\{P(x, \cdot)\}$, the family of probability measures $\{P^n(x, \cdot), x \in \mathcal{X}, n = 0, 1, 2, \dots\}$ constructed in Theorem 5.1.5, is a transition function.*

Proof. We only need to show the Chapman-Kolmogorov equations hold. This holds for every $n \geq 1$ and for every $m = 0, 1$. We assume that it holds for all n, m such that $k = n + m$. We show (5.2) holds for $k = n + m + 1$. Let $0 \leq j < n + m$. We first use the definition,

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

In the second step we used that for any $f \in \mathcal{B}_b$,

$$\int_{\mathcal{X}} f(z) P^{n+m}(x, dz) = \int_{\mathcal{X}} \int_{\mathcal{X}} f(z) P^n(y, dz) P^m(x, dy). \quad (5.5)$$

The proof is complete. \square

5.1.3 Markov chain with transition functions/ transition kernels

Definition 5.1.7 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 $n \geq 0$,

$$\mathbb{P}(x_{n+1} \in A | x_n) = P(x_n, A), \quad a.s. \quad (5.6)$$

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

Note. Henceforth, we focus on time-homogeneous Markov chains with transition probabilities, sometimes we drop ‘time-homogeneous’ and/ or ‘with transition probabilities’ for simplicity.

Remark 5.1.8 1. Since (x_n) is a Markov chain, (5.6) is equivalent to

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

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

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

Exercise 5.1.1 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 $\mathbf{E}(Y | \mathcal{F}_1)$.

Remark 5.1.9 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.

Example 5.1.3 Let $\mathcal{X} = \mathbf{N}$. Then the transition probabilities are determined by the numbers:

$$p(i, j) = \mathbb{P}(x_{n+1} = j | x_n = i).$$

They satisfies

$$\sum_{j \in \mathcal{X}} p(i, j) = 1.$$

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 $P_{ij}^m = \mathbb{P}(x_{n+m} = j | x_n = i)$ 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.

———This, together with the example in Section 5.3, marks the end of Week 2 lectures. ———

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

Theorem 5.1.10 *Let (x_n) be a time-homogeneous Markov process with transition probabilities P . Then, one has for every $n, m \geq 0$,*

$$(1) \quad \mathbb{P}(x_{n+m} \in A | x_m) = P^n(x_m, A), \quad (5.7)$$

Note that $\mathbb{P}(x_{n+m} \in A | x_m = x) = P^n(x, A)$ a.s..

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

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

¹William Feller: Non-Markovian processes with the semi-group property. In Ann. Math. Statist. Volum 30, number 4 (1959) pp1252-1253.

<https://projecteuclid.org/download/pdf1/euclid.aoms/1177706110>

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

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

In the last line, we have used induction hypothesis $\mathbf{E}(f(x_{k+m}) | x_m) = \int_{\mathcal{X}} f(z) P^k(x, dz)$ applied to $f = P(\cdot, A)$.

(2)

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

□

Exercise 5.1.2 If (x_n) is a time-homogeneous Markov process with transition probabilities P and initial distribution μ , prove that

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

Proposition 5.1.11 If (x_n) is a Markov process with transition probabilities P , then for any $f_i \in \mathcal{B}_b$,

$$\mathbf{E}(\Pi_{i=0}^n f_i(x_i)) = \overbrace{\int_{\mathcal{X}} \cdots \int_{\mathcal{X}}}^{n+1} \Pi_{i=0}^n f_i(y_i) \Pi_{i=0}^{n-1} P(y_i, dy_{i+1}) \mu(dy_0). \quad (5.8)$$

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

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

The last function involves only $\{x_0, x_1, \dots, x_{n-1}\}$, we may apply the induction hypothesis. The rest follows from induction:

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

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

The last line follows after bring $\Pi_{i=0}^{n-1} f_i(y_i)$ inside the inner most integral. \square

Remark 5.1.12 If (x_n) is a stochastic process such that (5.8) holds for any $n \geq 0$ 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 3.1.10.

Corollary 5.1.13 *If (x_n) is a time homogeneous Markov chain with transition function P , then for any $n \geq 1$ and for any $A_i \in \mathcal{B}(\mathcal{X})$,*

$$\begin{aligned} \mathbb{P}(x_0 \in A_0, x_1 \in A_1, \dots, x_n \in A_n) \\ = \int_{A_0} \int_{A_1} \cdots \int_{A_{n-1}} P(y_{n-1}, A_n) P(y_{n-2}, dy_{n-1}) \cdots P(y_1, dy_2) P(y_0, dy_1) \mu(dy_0). \end{aligned} \quad (5.9)$$

We emphasize that if (x_n) is a stochastic process such that (5.9) holds for any $n \geq 1$ and for any $A_i \in \mathcal{B}(\mathcal{X})$, (5.8) holds and (x_n) is a Markov chain (with transition probability P and initial distribution μ_0).

Corollary 5.1.14 *If x_n is a process with finite dimensional distribution given by*

$$\mu_n(A_1 \times \cdots \times A_n) = \int_{A_0} \int_{A_1} \cdots \int_{A_{n-1}} P(y_{n-1}, A_n) P(y_{n-2}, dy_{n-1}) \cdots P(y_1, dy_2) P(y_0, dy_1) \mu(dy_0),$$

then (x_n) is a time homogeneous Markov process with transition probability P .

Proof. The statement that (5.8) holds for any $f \in \mathcal{B}_b$ is equivalent to (5.9) holds for any $A_i \in \mathcal{B}(\mathcal{X})$. Hence, x_n is a Markov process, and one can check that its transition probabilities is P . \square

5.1.4 Existence of Markov Chains with given transition probabilities

Proposition 5.1.15 *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

$$\mu_n(A_0 \times \cdots \times A_n) =$$

$$\int_{A_0} \int_{A_1} \int_{A_2} \cdots \int_{A_{n-2}} \int_{A_{n-1}} P(y_{n-1}, A_n) P(y_{n-2}, dy_{n-1}) \cdots P(y_1, dy_2) P(y_0, dy_1) \mu(dy_0) .$$

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 \mathbb{P}_μ on \mathcal{X}^∞ such that the restriction of \mathbb{P}_μ 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 \mathbb{P}_μ . Then for (π_n) the canonical process, *i.e.* $\pi_n((w_0, w_1, \dots)) = w_n$,

$$\mathbb{P}_\mu(\pi_0 \in A_0, \dots, \pi_{n+1} \in A_n) = \mathbb{P}_\mu(A_0 \times \cdots \times A_n \times \mathcal{X}^\infty) = \mu_n(A_0 \times \cdots \times A_n).$$

This means that (π_n) has μ_n as its finite dimensional distribution, and by Corollary 5.1.14, it is a Markov process with the required transition probabilities and initial distribution μ_0 . This concludes the 'existence' part. The uniqueness follows from the 'uniqueness' part of Kolmogorov's extension theorem. \square

From the proof, the projection maps $\pi_n : \mathcal{X}^\infty \rightarrow \mathcal{X}$ is a Markov process on $(\mathcal{X}^\infty, \otimes_{i=0}^\infty \mathcal{B}(\mathcal{X}), \mathbb{P}_\mu)$ with state space \mathcal{X} , transition probabilities P and initial distribution μ . Recall that this process is called the canonical process.

It is traditional to denote by \mathbb{P}_x the probability measure on the canonical space \mathcal{X}^∞ 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 \mathbb{P}_μ . Then on the canonical probability space we use \mathbf{E}_x and \mathbf{E}_μ to denote taking expectations w.r.t. \mathbb{P}_x and \mathbb{P}_μ respectively.

Example 5.1.4 Let x_n be a Markov chain with transition probability P , then

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

Remark 5.1.16 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.

5.2 Examples and Exercises

Example 5.2.1 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

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

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

Example 5.2.2 Let $F: \mathcal{X} \rightarrow \mathcal{X}$ 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 5.2.1 Let ξ_n be a sequence of real-valued i.i.d. random variables and define x_n recursively by $x_0 = 0$, $x_n = \alpha x_{n-1} + \xi_n$. Show 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 5.2.3 Let $M_n = \max(x_0, x_1, \dots, x_n)$, 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: $\mathbb{P}(\xi_i = \pm 1) = \frac{1}{2}$. Then (M_n) is not a Markov process.

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 paths σ_1 and 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, concluding that (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 $\mathbb{P}(M_4 = 2 | M_3 = 1) = \frac{1}{3}$. This can be computed using the space of the path of uniform probability as probability space, counting paths or using elementary conditional expectations.

$$\mathbb{P}(M_4 = 2 | (x_0, x_1, x_2, x_3) = \sigma_1) = \frac{\mathbb{P}(M_4 = 2, (x_0, x_1, x_2, x_3) = \sigma_1)}{\mathbb{P}((x_0, x_1, x_2, x_3) = \sigma_1)} = \frac{1}{2}.$$

5.3 Transition operators and invariant measures

Let $(\Omega, \mathcal{F}, \mathbb{P})$ 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 ,

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

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 5.3.1 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

$$(T\mu)(A) = \int_{\mathcal{X}} P(x, A) \mu(dx). \quad (5.10)$$

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

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

Remark: We denote by $Tf = \int f d\mu$, for any measurable function for which the integral exists.

Definition 5.3.2 A measure such that $T\mu = \mu$ is called an invariant measure for P (or for the time homogeneous Markov chain).

The measure assigns every measurable set 0 is the trivial measure. By an invariant measure we mean a non-trivial invariant measure. We are most interested in finite invariant measures, which can then be normalised to a probability measure.

Exercise 5.3.1 Check that the operator T^n obtained by replacing P by P^n in (5.10) is equal to the operator obtained by applying T n times, $T^n = T \circ T \circ \dots \circ T$.

Remark 5.3.3 ** 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 (5.10) 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 $T_\star : \mathcal{B}_b(\mathcal{X}) \rightarrow \mathcal{B}_b(\mathcal{X})$, the space of bounded measurable functions from \mathcal{X} to \mathbf{R} , by

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

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

Exercise 5.3.2 Check that the operators T and T_\star are each other's dual, *i.e.* that

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

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

Exercise 5.3.3 Show that $T_\star 1 = 1$, $T_\star f \geq 0$ if $f \geq 0$.

5.3.1 Stationary Markov chain

Remark 5.3.4 *Invariant probability measures and stationary processes.* Given an invariant measure π , take $x_0 \sim \pi$. Then π and transition probabilities $(P(x, \cdot))$ determine \mathbb{P}_π on \mathcal{X}^∞ , see Proposition 5.1.15. By Markov property, the shifted process $\theta_n x_\cdot$ is a Markov process with transition probabilities $P = (P(x, \cdot))$ and initial distribution $\mathcal{L}(x_n) = \pi$. Hence on \mathcal{X}^∞ , $\mathcal{L}(\theta_n x_\cdot) = \mathbb{P}_\pi$. The process (x_n) is a *stationary process*.

5.4 Example: Markov Chains On Discrete State Spaces

We return to Example 5.1.3 to formulate the discrete state space example in more detail.

5.4.1 Stochastic Matrix

Let $\mathcal{X} = \mathbf{N}$ or have only a finite number of elements $\{1, \dots, N\}$. A probability measure on \mathcal{X} is determined by its value on singleton sets $\{j\}$. Suppose we are given $\nu(i)$ with $\sum_{i \in \mathcal{X}} \nu(i) = 1$

and $\mu(i) \geq 0$, then $\nu(A) = \sum_{i \in A} \mu(i)$ defines a probability measure on \mathcal{X} . Thus we identify a measure on \mathcal{X} with a row vector with entries $\nu(i) \geq 0$ and $\sum_{i \in \mathcal{X}} \nu(i) = 1$.

The transition probabilities in Definition 5.1.1 are determined by $P = (P_{ij})$ where P_{ij} is shorthand for $P(i, \{j\})$. Since $P(i, \cdot)$ is a probability measure,

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

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.

Definition 5.4.1 Suppose for $i, j \in \mathcal{X}$, we are given $P_{ij} \geq 0$ with $\sum_{j \in \mathcal{X}} P_{ij} = 1$. Then $P = (P_{ij})$ is called a stochastic matrix. The sum of each row is 1.

5.4.2 N-step Transitions

If $\mathcal{X} = \{1, \dots, N\}$ then $P = (P_{ij})$ is a $N \times N$ -matrix, otherwise it is a semi-infinite matrix. These matrices determine the n -step transition probabilities defined by iteration by $P_{ij} = P_{ij}^1$, and

$$P^{n+1}(i, \{j\}) = \int_{\mathcal{X}} P(y, \{j\}) P^n(i, dy) = \sum_{k \in \mathcal{X}} P(k, \{j\}) P^n(i, \{k\}) = \sum_{k \in \mathcal{X}} P_{ik}^n P_{kj}, \quad i, j \in \mathcal{X}$$

Let us denote by P^n the matrix with entries $P(i, \{j\})$. Write also $P_{ij}^n = P^n(i, \{j\})$. (The P^n 's are also called n -step transition matrices.)

Exercise 5.4.1 Denote $P_{ij}^n = P^n(i, \{j\})$. Show that

$$P_{ij}^n = \sum_{k_{n-1} \in \mathcal{X}} \cdots \sum_{k_1 \in \mathcal{X}} P_{ik_1} \cdots P_{k_{n-2}k_{n-1}} P_{k_{n-1}j}.$$

This means the n -step transition matrices is in fact $P^n = \overbrace{P \times \cdots \times P}^n$, the matrix multiplication of P by itself n -times. For this reason both $(P^n)_{ij}$ and P_{ij}^n are used for its entry at row i and column j . Observe that P^n is again a stochastic matrix, each row sums to 1.

The following definition is a paraphrase of Definition 5.1.7, with respect to the natural filtration.

Definition 5.4.2 A time homogeneous Markov chain on a countable state space \mathcal{X} with initial distribution μ and transition probabilities (P_{ij}) is a stochastic process such that the following holds:

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

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

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

The following is Corollary 5.1.13 for discrete state space, a separate proof is given here for reader's convenience.

Proposition 5.4.3 *Let (x_n) be a Markov chain with transition probabilities P_{ij} with initial distribution μ . Then, for any state i_j and any $n \geq 0$,*

$$\mathbb{P}(x_0 = i_0, \dots, x_n = i_n, x_{n+1} = i_{n+1}) = \mu(i_0) P_{i_0 i_1} \dots P_{i_{n-1} i_n} P_{i_n i_{n+1}}. \quad (5.11)$$

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

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

Suppose the identity holds on n -times:

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

Then,

$$\begin{aligned} & \mathbb{P}(x_0 = i_0, \dots, x_n = i_n, x_{n+1} = i_{n+1}) \\ &= \mathbb{P}(x_{n+1} = i_{n+1} | x_0 = i_0, \dots, x_n = i_n, x_n = i_n) \mathbb{P}(x_0 = i_0, \dots, x_n = i_n, x_n = i_n) \\ &= P_{i_n, i_{n+1}} \mathbb{P}(x_0 = i_0, \dots, x_n = i_n, x_n = i_n). \end{aligned}$$

In the last line we used the Markov property. The rest follows by the induction hypothesis on $\mathbb{P}(x_0 = i_0, \dots, x_n = i_n, x_n = i_n)$. \square

Corollary 5.4.4

$$\mathbb{P}(x_{n+1} = i_{n+1}, \dots, x_{n+m} = i_{n+m} | x_0 = i_0, \dots, x_n = i_n) = \prod_{k=n}^{n+m-1} P_{i_k i_{k+1}}. \quad (5.12)$$

Exercise 5.4.2 Show that for any $m \geq 0$,

$$\begin{aligned} & \mathbb{P}(x_{n+m} = k_n, x_{n+m-1} = k_{n-1}, \dots, x_{m+1} = k_1 | x_m = i) = P_{i k_1} P_{k_1 k_2} \dots P_{k_{n-1} k_n} \\ & \mathbb{P}(x_{n+m} = k_n, x_{n+m-1} = k_{n-1}, \dots, x_{m+1} = k_1, x_m = i) = P_{i k_1} P_{k_1 k_2} \dots P_{k_{n-1} k_n} \mathbb{P}(x_m = i). \end{aligned} \quad (5.13)$$

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

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

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

Proposition 5.4.5 *For any $n \geq 2, m \geq 0$ and any $i, j \in \mathcal{X}$,*

$$\mathbb{P}(x_{n+m} = j \mid x_m = i) = P_{ij}^n.$$

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

$$\mathbb{P}(x_n = j) = \sum_{k \in \mathcal{X}} \mu(k) P_{kj}^n.$$

Proof. Firstly,

$$\begin{aligned} \mathbb{P}(x_{n+m} = j \mid x_m = i) &= \sum_{k_{m-1} \in \mathcal{X}} \cdots \sum_{k_1 \in \mathcal{X}} \mathbb{P}(x_{n+m} = j, x_{n+m-1} = k_{n-1}, \dots, x_{m+1} = k_1 \mid x_m = i) \\ &= \sum_{k_1 \in \mathcal{X}} \cdots \sum_{k_{n-1} \in \mathcal{X}} P_{ik_1} \cdots P_{k_{n-2}k_{n-1}} P_{k_{n-1}j}. \end{aligned}$$

We have used (5.13) in the last step. Finally,

$$\mathbb{P}(x_n = j) = \sum_{k \in \mathcal{X}} \mathbb{P}(x_n = j, x_0 = k) = \sum_{k \in \mathcal{X}} \mathbb{P}(x_0 = k) \mathbb{P}(x_n = j \mid x_0 = k) = \sum_{k \in \mathcal{X}} \mu(k) P_{kj}^n,$$

completing the proof. \square

Theorem 5.4.6 *(The Chapman-Kolmogorov equation) For any $i, j \in \mathcal{X}$ and $n, m \geq 1$, we have*

$$P_{ij}^{n+m} = \sum_{k \in \mathcal{X}} P_{ik}^n P_{kj}^m.$$

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

$$\begin{aligned} P_{ij}^{n+m} &= \mathbb{P}(x_{n+m} = j \mid x_0 = i) = \sum_{k \in \mathcal{X}} \mathbb{P}(x_{n+m} = j, x_n = k, x_0 = i) / P(x_0 = i) \\ &= \sum_{k \in \mathcal{X}} \mathbb{P}(x_{n+m} = j \mid x_n = k) \frac{\mathbb{P}(x_n = k, x_0 = i)}{\mathbb{P}(x_0 = i)} \\ &= \sum_{k \in \mathcal{X}} P_{ik}^n P_{kj}^m. \end{aligned}$$

We have used Proposition 5.4.5. \square

5.4.3 The Markov property is determined by finite dimensional distributions

The following can be considered to be a converse to Proposition 5.4.3.

Theorem 5.4.7 *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 $i_0, \dots, i_{n+1} \in \mathcal{X}$,*

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

Note this is (5.11). Then (x_n) is a Markov chain with transition probabilities $P = (P_{ij})$ with initial distribution μ .

Proof. Take $n = 1$ in the above, $\mathbb{P}(x_0 = i_0, x_1 \in i_1) = P_{i_0 i_1} \mu(i_0)$, summing up $i_1 \in \mathcal{X}$, we get $\mathbb{P}(x_0 = i_0) = \mu(i_0)$. Also,

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

This means, $P(x_{n+1} = i_{n+1} \mid x_n, \dots, x_0) = P_{x_n, i_{n+1}}$, 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. \square

Remark 5.4.8 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 $n = 1, 2, \dots$.

5.4.4 Conditional independence of the future and the past

Theorem 5.4.9 *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 $x_s = i$, $y_s \equiv x_{s+n}$ is a time homogeneous Markov chain with transition probability $(P_{i,j})$ and initial distribution δ_i ’ means precisely the following:

$$\mathbb{P}(y_0 = i_s, \dots, y_{n+1} = i_{s+n+1} \mid x_s = i) = \delta_{ii_s} P_{i_s i_{s+1}} \dots P_{i_{s+n} i_{s+n+1}} = \prod_{k=s}^{s+n} P_{i_k i_{k+1}} \delta_{ii_s}.$$

The left hand side is $\mathbb{P}(x_s = i_s, x_{s+1} = i_{s+1}, \dots, x_{s+n+1} = i_{s+n+1} \mid x_s = i)$, the identity follows from (5.12). The conditional independent statement means precisely the following: for any $A \in \sigma(x_0, \dots, x_s)$,

$$\mathbb{P}((y_0 = i_s, \dots, y_{n+1} = i_{s+n+1}) \cap A \mid x_s = i) = \mathbb{P}((y_0 = i_s, \dots, y_{n+1} = i_{s+n+1}) \mid x_s = i) \delta_{ii_s} P(A \mid x_s = i). \quad (5.14)$$

It is sufficient to check the above holds for $A = \{x_0 = i_0, \dots, x_s = i_s\}$ (The collection of set of this form is a π -system generating $\sigma(x_0, \dots, x_s)$). The right hand side of (5.14) is

$$\begin{aligned} & \mathbb{P}((y_1 = i_{s+1}, \dots, y_{n+1} = i_{s+n+1}) | x_s = i_s) P(x_0 = i_0, \dots, x_s = i_s | x_s = i_s) \\ &= \prod_{k=s}^{s+n} P_{i_k i_{k+1}} \delta_{i i_s} \delta_{i i_s} P(x_0 = i_0, \dots, x_s = i_s | x_s = i). \end{aligned}$$

By the definition of elementary conditional probability, the left hand side of (5.14) is

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

proving the independence of the future and past given the present. \square

5.4.5 Operator on Measures

Suppose that we are given a stochastic matrix (P_{ij}) on \mathcal{X} . Define:

$$(T^* \nu)(i) = \sum_{k \in \mathcal{X}} \nu(k) P_{ki}.$$

The matrix P acts on the measure ν on the right as matrix multiplication: $T^* \nu = \nu P$.

Since P is a stochastic matrix,

$$\sum_{i \in \mathcal{X}} (T^* \nu)(i) = \sum_{k \in \mathcal{X}} \sum_{i \in \mathcal{X}} P_{ki} \nu(k) = \sum_{k \in \mathcal{X}} \nu(k).$$

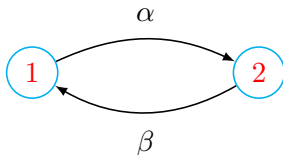
So the total mass of the new measure $T\nu$ is the same as that of ν . To summarise,

Remark 5.4.10 The map $\nu \rightarrow T^* \nu$ is a transformation on probability measures on \mathcal{X} . We sometime use T in place of T^* .

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

5.4.6 Example: Two State Markov Chains

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



Suppose that

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

Then,

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

Let P^0 be the identity matrix, then note $(\nu P^n)^T = (P^T)_n \nu^T$,

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

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

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

Thus,


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

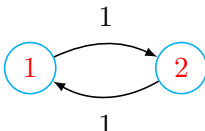
This is β if $\alpha + \beta = 1$. If $\alpha + \beta \neq 1$, iterate this to see,

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

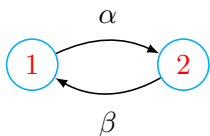
By symmetry,

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

 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 $n \rightarrow \infty$,

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

The rate of convergence is exponential.

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

$$(a, 1 - a)P^n \rightarrow (a, 1 - a) \begin{pmatrix} \frac{\beta}{\alpha + \beta} & \frac{\alpha}{\alpha + \beta} \\ \frac{\beta}{\alpha + \beta} & \frac{\alpha}{\alpha + \beta} \end{pmatrix} = \left(\frac{\beta}{\alpha + \beta}, \frac{\alpha}{\alpha + \beta} \right)$$

Observe that the measure is invariant under the transformation $\nu \mapsto \nu P$.

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^T has eigenvalue 1 and $\lambda = 1 - \alpha - \beta$. Their corresponding eigenvectors are

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

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

$$\left(\frac{\beta}{\alpha + \beta}, \frac{\alpha}{\alpha + \beta} \right),$$

the stationary probability measure! Let

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

Then,

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

Chapter 6

Strong Markov Property

6.1 Stopping Times

As usual, we have a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ 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 $\mathcal{F}_n = \sigma\{x_0, \dots, x_n\}$.

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

Definition 6.1.1 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.)

For a continuous time filtration $(F_t)_{t \geq 0}$, we say that $T : \Omega \rightarrow \mathbf{R}_+ \cup \{\infty\}$ is a stopping time if $\{T \leq t\} \in \mathcal{F}_t$ for every $t \geq 0$.

Exercise 6.1.1 Show that T is an \mathcal{F}_n -stopping time if and only if $\{T \leq n\} \in \mathcal{F}_n$.

Example 6.1.1 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: $\{\tau_A = 0\} = \emptyset \in \mathcal{F}_0$. 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 stopping time.

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

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

concluding that σ_A is a stopping 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 6.1.2 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

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

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 $(x_{T \wedge n}(\omega), n \geq 0)$.

Example 6.1.2 (a) A constant time is a stopping time.

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

Example 6.1.3 The following time is, in general, **not** a stopping time:

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

Proposition 6.1.2 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) $\limsup_{n \rightarrow \infty} T_n$ and $\liminf_{n \rightarrow \infty} T_n$ are stopping times.

Proof. Part (1) follows from the following observations:

$$\{\omega : \max(S, T) \leq n\} = \{S \leq n\} \cap \{T \leq n\} \in \mathcal{F}_n,$$

$$\{\omega : \min(S, T) \leq n\} = \{S \leq n\} \cup \{T \leq n\} \in \mathcal{F}_n.$$

Since

$$\limsup_{n \rightarrow \infty} T_n = \inf_{n \geq 1} \sup_{k \geq n} T_k, \quad \liminf_{n \rightarrow \infty} T_n = \sup_{n \geq 1} \inf_{k \geq n} T_k$$

we only need to prove that if T_n is an increasing sequence, $\sup_n T_n$ is a stopping time; and if S_n is a decreasing sequence of stopping times with limit S , $\inf_n S_n$ is a stopping time. These follows from

$$\{\sup_n T_n \leq n\} = \bigcap_n \{T_n \leq n\} \in \mathcal{F}_n, \quad \{\inf_n S_n \leq n\} = \bigcup_n \{S_n \leq n\} \in \mathcal{F}_n.$$

□

6.2 The Stopped σ -algebra

Let $\mathcal{F}_\infty = \bigvee_{n=0}^\infty \mathcal{F}_n$.

Definition 6.2.1 If T is an \mathcal{F}_n -stopping time, we define the associate σ -algebra to be

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

Note that $A \cap \{T = \infty\} = A \setminus \bigcup_{n=1}^\infty (A \cap \{T = n\})$ where $A \in \mathcal{F}_\infty$ is always in \mathcal{F}_∞ . In particular, $\{T = \infty\} \in \mathcal{F}_\infty$. If $T \equiv n$ is a constant time, then \mathcal{F}_T agrees with \mathcal{F}_n . The continuous time version is:

$$\mathcal{F}_T = \{A \in \bigvee_{t \geq 0} \mathcal{F}_t : A \cap \{T \leq t\} \in \mathcal{F}_t, \forall t \geq 0\}.$$

Lemma 6.2.2 If T is an \mathcal{F}_n -stopping time, T is \mathcal{F}_T measurable.

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

Exercise 6.2.1 Let S, T be $(\mathcal{F}_t, t \geq 0)$ -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) $\mathcal{F}_S \cap \{S \leq T\} \subset \mathcal{F}_{S \wedge T}$.
- (5) $\mathcal{F}_T = \mathcal{F}_t$ on $\{T = t\}$.

Proof. (1) If $A \in \mathcal{F}_S$,

$$A \cap \{T \leq t\} = (A \cap \{S \leq t\}) \cap \{T \leq t\} \in \mathcal{F}_t$$

and hence $A \in \mathcal{F}_T$.

(2) Since $\mathcal{F}_S \subset \mathcal{F}_T$,

$$\{S\mathbf{1}_A + T\mathbf{1}_{A^c} \leq t\} = (\{S \leq t\} \cap A) \cup (\{T \leq t\} \cap A^c) \in \mathcal{F}_T.$$

(3) Let $r, t \in \mathbf{R}$, $\{S \leq r\} \cap \{S \leq t\} = \{S \leq \min(r, t)\} \in \mathcal{F}_t$. Hence $\{S \leq r\} \in \mathcal{F}_r$.

(4) Take $A \in \mathcal{F}_S$ and $t \geq 0$. Then

$$A \cap \{S \leq T\} \cap \{S \wedge T \leq t\} = A \cap \{S \leq t\} \cap \{S \wedge t \leq T \wedge t\} \in \mathcal{F}_t.$$

which follows as $S \wedge t$ and $T \wedge t$ are \mathcal{F}_t -measurable. Hence $A \cap \{S \leq T\} \in \mathcal{F}_{S \wedge T}$.

(5) Let $A \in \mathcal{F}_T$, then $A \cap \{T = t\} \in \mathcal{F}_t$ by the definition. If $A \in \mathcal{F}_t$, $A \cap \{T = t\} \cap \{T \leq s\} \in \mathcal{F}_s$ for any s . Hence $A \cap \{T = t\} \in \mathcal{F}_T$. \square

Below in Lemmas 6.2.3, Lemma 6.2.4, and Proposition 6.2.5, we motivate the definition for \mathcal{F}_T . For this we work with a stopping time taking finite values only.

Lemma 6.2.3 *Let (x_n) be an adapted stochastic process. If $T < \infty$ is a stopping time and (x_n) is adapted, then x_T is \mathcal{F}_T -measurable, and so are $x_{T \wedge m}$ for any $m \in \mathbf{N}$.*

Proof. Let B be a Borel subset of \mathcal{X} . Since $T < \infty$, x_T is well defined. For any $m = 0, 1, 2, \dots$

$$\{x_T \in B\} \cap \{T = m\} = \{x_m \in B\} \cap \{T = m\} \in \mathcal{F}_m,$$

showing that $\{x_T \in B\} \in \mathcal{F}_T$ and x_T is \mathcal{F}_T -measurable. Similarly, for any $n \geq 0$,

$$\{x_{T \wedge m} \in B\} \cap \{T = n\} = \{x_{m \wedge n} \in B\} \cap \{T = n\} \in \mathcal{F}_n,$$

showing that $x_{T \wedge m}$ is \mathcal{F}_T -measurable. \square

Lemma 6.2.4 *Let $\mathcal{F}_n = \vee_{i=0}^n \sigma(x_i)$. If $T < \infty$ is an (\mathcal{F}_n) -stopping time, then for every $k \geq 0$,*

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

Proof. The statement is equivalent to, for any k , $\mathbf{1}_{\{T=k\}}$ is of the form $\varphi_k(\sigma(x_{0 \wedge T}, \dots, x_{T \wedge k}))$, where $\varphi_k \in \mathcal{B}_b(\mathcal{X}^{k+1})$. Firstly, $\{T = 0\} \in \mathcal{F}_0 = \sigma(x_{0 \wedge T})$. We prove this by induction on n assume it holds for $n = k - 1$. Since $\{T = k\} \in \sigma(x_0, x_1, \dots, x_k)$, by the factorisation lemma, we may assume that

$$\mathbf{1}_{\{T=k\}} = \psi(x_0, x_1, \dots, x_k)$$

where $\psi \in \mathcal{B}_b(\mathcal{X}^{k+1})$. Then

$$\mathbf{1}_{\{T=k\}} = \mathbf{1}_{\{T=k\}} \mathbf{1}_{\{T \geq k\}} = \psi_1(x_{0 \wedge T}, \dots, x_{T \wedge k}) \mathbf{1}_{\{T \geq k\}} = \psi(x_{0 \wedge T}, \dots, x_{T \wedge k})(1 - \mathbf{1}_{\{T \leq k-1\}}).$$

By the induction hypothesis, $\mathbf{1}_{\{T \leq k-1\}} = \varphi_{k-1}(x_{0 \wedge T}, \dots, x_{T \wedge (k-1)})$, this completes the proof. \square

Proposition 6.2.5 *Let (x_n) be a stochastic process, $\mathcal{F}_n = \sigma(x_0, \dots, x_n)$. Set*

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

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

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

Proof. By Lemma 6.2.3 $x_{T \wedge m}$ is \mathcal{F}_T -measurable for any $m \in \mathbf{N}$, showing that $\sigma(x^T) \subset \mathcal{F}_T$. For the converse of the theorem, let us take $A \in \mathcal{F}_T$. Then $A \cap \{T = n\} \in \mathcal{F}_n$ which means $\mathbf{1}_{A \cap \{T=n\}} = \Psi(x_0, \dots, x_n)$ for some $\Psi \in \mathcal{B}_b(\mathcal{X}^{n+1})$. Hence

$$\mathbf{1}_{A \cap \{T=n\}} = \Psi(x_0, \dots, x_n) \mathbf{1}_{\{T=n\}} = \Psi(x_{0 \wedge T}, \dots, x_{n \wedge T}) \mathbf{1}_{\{T=n\}},$$

Since $\{T = n\} \in \sigma(x_{T \wedge 0}, x_{T \wedge 1}, \dots, x_{T \wedge n})$ the Lemma 6.2.4, $\mathbf{1}_{A \cap \{T=n\}}$ is measurable w.r.t. $\sigma(x^T)$. We thus conclude that $A \in \bigvee_{n=1}^{\infty} \sigma(x_{n \wedge T})$ and $\mathcal{F}_T \subset \bigvee_{n=1}^{\infty} \sigma(x_{n \wedge T})$. \square

Exercise 6.2.2 Let $x_n = \sum_{i=1}^n \xi_i$ be as in Example 6.1.2. Let $x_0 = 0$, T the first time $x_n = 1$. Is the event $\{x_n = 2\}$ in \mathcal{F}_T ?

This marks the end of lecture 6 content.

Problem Class 1

Exercise 6.2.3 (*Ex. 2.2.2 Notes / Ex. 9 PS1*)

Let $X : \Omega \rightarrow \mathcal{X}$ and $Y : \Omega \rightarrow \mathcal{Y}$ be random variables with X measurable with respect to $\mathcal{G} \subset \mathcal{F}$ and Y is independent of \mathcal{G} . If $\varphi : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbf{R}$ is measurable and bounded, show that

$$\mathbf{E}(\varphi(X, Y) | \mathcal{G})(\omega) = \mathbf{E}(\varphi(X(\omega), Y)), \quad a.s. \quad (6.1)$$

Solution from Piazza Forum. Fix $A \in \mathcal{G}$, define $\psi : \mathcal{X} \times \mathbf{R} \times \mathcal{Y} \rightarrow \mathbf{R}$ by $\psi(x, z, y) = \varphi(x, y)z$. With $Z = \mathbf{1}_A$, we are interested in $\psi(X, Z, Y) = \varphi(X, Y)Z = \varphi(X, Y)\mathbf{1}_A$. Let $\nu = \mathcal{L}((X, Z))$ and $\mu = \mathcal{L}(Y)$, then

$$\begin{aligned} \mathbf{E}[\psi(X, Z, Y)] &= \int_{\mathcal{X} \times \mathbf{R}} \int_{\mathcal{Y}} \psi(x, z, y) \mu(dy) \nu(dx, dz) = \mathbf{E} \left[\int_{\mathcal{Y}} \psi(X, \mathbf{1}_A, y) \mu(dy) \right] \\ &= \mathbf{E} \left[\int_{\mathcal{Y}} \varphi(X, y) \mu(dy) \mathbf{1}_A \right]. \end{aligned}$$

This implies (6.1), since

$$\mathbf{E}[\varphi(X, Y)\mathbf{1}_A] = \mathbf{E}[\psi(X, Z, Y)] = \mathbf{E} \left[\int_{\mathcal{Y}} \varphi(X, y) \mu(dy) \mathbf{1}_A \right].$$

Standard Solution. It is sufficient to show (6.1) for $\varphi(x, y)$ of the form $\varphi(x, y) = f(x)g(y)$ with $f \in \mathcal{B}_b(\mathcal{X})$ and $g \in \mathcal{B}_b(\mathcal{Y})$. For $A \in \mathcal{G}$,

$$\begin{aligned} \mathbf{E}[\varphi(X, Y)\mathbf{1}_A] &= \mathbf{E}[f(X)g(Y)\mathbf{1}_A] = \mathbf{E}[\mathbf{E}[f(X)g(Y)\mathbf{1}_A | \mathcal{G}]] = \mathbf{E}[f(X)\mathbf{E}[g(Y) | \mathcal{G}]\mathbf{1}_A] \\ &= \mathbf{E}[f(X)\mathbf{E}[g(Y)]\mathbf{1}_A], \end{aligned} \quad (6.2)$$

which proves (6.1) for $\varphi(x, y) = f(x)g(y)$. Then by suitable approximation (e.g. via simple functions), (6.1) holds for general φ . \square

Monotone Class Solution.¹ For any $B_1 \in \mathcal{B}(\mathcal{X})$ and $B_2 \in \mathcal{B}(\mathcal{Y})$, the derivation (6.2) (with $f = \mathbf{1}_{B_1}$ and $g = \mathbf{1}_{B_2}$) holds for the function $\varphi(x, y) = \mathbf{1}_{B_1 \times B_2}(x, y) = \mathbf{1}_{B_1}(x)\mathbf{1}_{B_2}(y)$. Then, if we consider

$$\begin{aligned}\mathcal{H} &= \{\varphi \in \mathcal{B}_b(\mathcal{X} \times \mathcal{Y}) : \text{equality (6.1) holds}\}, \\ \mathcal{C} &= \{B_1 \times B_2 : B_1 \in \mathcal{B}(\mathcal{X}), B_2 \in \mathcal{B}(\mathcal{Y})\},\end{aligned}$$

we can see that \mathcal{C} is π -system and \mathcal{H} is a vector space. By the observation above and monotone convergence of expectation, we have

- The constant function $\varphi(x, y) = 1$ belongs to \mathcal{H} ,
- If $B_1 \times B_2 \in \mathcal{C}$, then $\mathbf{1}_{B_1 \times B_2} \in \mathcal{H}$,
- If $f_n \in \mathcal{H}$ is an increasing sequence with $f_n \geq 0$ and $f_n \nearrow f$ with f bounded, then $f \in \mathcal{H}$.

So that by Monotone Class theorem (lemma below), we infer that any bounded $\sigma(\mathcal{C})$ -measurable function φ belongs to \mathcal{H} , concluding the proof.²

Lemma 6.2.6 (Monotone Class theorem) *Suppose that \mathcal{C} is a π -system containing Ω , and \mathcal{H} a class of functions with the property:*

1. $1 \in \mathcal{H}$,
2. If $A \in \mathcal{C}$ then $\mathbf{1}_A \in \mathcal{H}$,
3. \mathcal{H} is a vector space,
4. If $f_n \in \mathcal{H}$ is a non-negative increasing sequence of functions with limit f bounded (resp. f is finite), then $f \in \mathcal{H}$.

We can conclude that every bounded (resp. finite) $\sigma(\mathcal{C})$ -measurable function is in \mathcal{H} .

Exercise 6.2.4 (*Markov Chain practice example*)

Consider the time homogeneous Markov chain on $\mathcal{X} = \{1, 2, 3\}$, with transition probabilities P and initial distribution $x_0 \sim \nu = (\frac{1}{2}, \frac{1}{2}, 0)$. Recall that P is uniquely determined by $P_{ij} = P(i, \{j\}) = \mathbb{P}(x_1 = j | x_0 = i)$. We have

$$P = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & 1 & 0 \end{pmatrix} \tag{6.3}$$

¹Here we provide an alternative proof, using as main tool the powerful Monotone Class theorem, stated below. This proof is longer/more detailed than the standard one you would be expected to provide.

²Note that \mathcal{C} generates the Borel sigma algebra on $\mathcal{X} \times \mathcal{Y}$.

1. Compute the following probabilities:

$$\begin{aligned} \mathbb{P}(x_0 = 1, x_1 = 2, x_2 = 2), \quad \mathbb{P}(x_1 = 2, x_2 = 2 | x_0 = 1), \\ \mathbb{P}(x_0 = 1, x_2 = 2), \quad \mathbb{P}(x_3 = 1 | x_0 = 1), \quad \mathbb{P}(x_0 = 1, x_1 = 2, x_3 = 2). \end{aligned}$$

2. Starting from $x_0 \sim \nu$, compute $\mathbb{P}(x_1 = 1)$. Then compute the law of x_1 , i.e. $\mathcal{L}(x_1) = \mu_1 = (\mu_1(1), \mu_1(2), \mu_1(3))$.

3. Compute the invariant distribution π (which satisfies $\pi P = \pi$).

Exercise 6.2.5 (*Ex 1.1.1 Notes*)

Let x_0, ξ_1, ξ_2 be independent random variables on \mathbf{R} . Let $x_0 \sim N(0, a^2)$, $\xi_i \sim N(0, b^2)$ for all $i = 1, 2, \dots$. Define for a positive number $M > 0$,

$$M(x_{n+1} - x_n) = -bx_n + \xi_{n+1}.$$

Find an invariant measure for this Markov chain.

Remark Note that $(x_{n+1} - x_n)$ can be recognised as the discrete derivative $\frac{d}{dt}(x.)$ and then (x_n) as the discrete version of the Ornstein-Uhlenbeck process.

Solution. Rearranging, we have

$$x_{n+1} = \left(1 - \frac{b}{M}\right)x_n + \frac{1}{M}\xi_{n+1} =: cx_n + \eta_{n+1}, \quad (6.4)$$

so that $x_1 = cx_0 + \eta_{n+1}$ is Gaussian if x_0 is Gaussian. Hence, iteratively we have x_n Gaussian if x_0 Gaussian. So, our guess for invariant probability measure needs to be a mean zero Gaussian. If $x_0 \sim N(0, a^2)$, by (6.4) we have

$$\mathcal{L}(x_1) = N(0, c^2a^2 + b^2/M^2),$$

so that

$$\mathcal{L}(x_1) = \mathcal{L}(x_0) \iff a^2 = c^2a^2 + \frac{b^2}{M^2}.$$

Thus we require

$$a^2 = \frac{b^2}{M^2(1 - c^2)}. \quad (6.5)$$

Given (6.5), we then have $\mathcal{L}(x_0) = \mathcal{L}(x_1) = \mathcal{L}(x_2) = \dots = \mathcal{L}(x_k)$, for all $k \geq 1$ and $\mathcal{L}(x_0) = N(0, a^2)$ is the invariant measure.

The discrete Ornstein-Uhlenbeck process, starting from the invariant probability measure $N(0, a^2)$ computed in Ex. 6.2.5, is a stationary Markov process.

6.3 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 theorems below.

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 $\mathcal{X}^{\mathbf{N}}$.

Note. We emphasize that by a Markov process we mean a time homogeneous Markov process with transition probabilities.

Recall the shift operators θ_t , where $t \in \mathbf{N}$,

$$\begin{aligned} \theta_t : \mathcal{X}^{\mathbf{N}^0} &\rightarrow \mathcal{X}^{\mathbf{N}^0} \\ (a_n, n \geq 0) &\mapsto (a_{t+n}, n \geq 0). \end{aligned}$$

We also define:

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

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} .

Definition 6.3.1 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 measurable function $\Phi : \mathcal{X}^{\mathbf{N}} \rightarrow \mathbf{R}$, the following holds:

$$\mathbf{E}(\Phi(\theta_T x) | \mathcal{F}_T) = \mathbf{E}(\Phi(\theta_T x) | X_T) \quad a.s. \quad (6.6)$$

Remark 6.3.2 Let us consider Borel measurable function on the path space $\mathcal{X}^{\mathbf{N}}$ with the product σ -algebra, which are generated by cylindrical sets of the form $\{\pi_{n_1} \in A_1, \dots, \pi_{n_m} \in A_m\}$ 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 $f : \mathcal{X}^m \rightarrow \mathbf{R}$ Borel measurable,

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

Even simpler, it is sufficient to take $\Phi(\sigma) = \Pi_{i=1}^n f_i(\sigma_{n_i})$ where $f_i \in \mathcal{B}_b(\mathcal{X})$.

Remark 6.3.3 (1) For simplicity let us define $y_n = x_{T+n}$. Set $\mathcal{G}_n = \mathcal{F}_{T+n}$. If (6.6) holds for every finite stopping time T for x_n then for every bounded Borel measurable function

$f : \mathcal{X} \rightarrow \mathbf{R}$, the following holds

$$\begin{aligned} \mathbf{E}(f(y_{n+m})|\mathcal{G}_m) &= \mathbf{E}(f(x_{T+n+m})|\mathcal{F}_{T+m}) \stackrel{(6.6)}{=} \mathbf{E}(f(x_{T+n+m})|x_{T+m}) \\ &= \mathbf{E}(f(y_{n+m})|y_m), \end{aligned}$$

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

$$\mathbb{P}(x_{n+T} \in A | \mathcal{F}_T) = \mathbb{P}^n(x_T, A). \quad (6.7)$$

then for every bounded Borel measurable function f ,

$$\mathbf{E}(f(x_{n+T})|\mathcal{F}_T) = \int f(y) \mathbb{P}^n(x_T, dy). \quad (6.8)$$

Exercise 6.3.1 If $\mathbb{P}(x_{n+T} \in A | \mathcal{F}_T) = \mathbb{P}^n(x_T, A)$ holds for every measurable set A , then

$$\mathbb{P}(x_{n_1+T} \in A_1, \dots, x_{n_m+T} \in A_m | \mathcal{F}_T) = \mathbb{P}(x_{n_1+T} \in A_1, \dots, x_{n_m+T} \in A_m | x_T), \quad (6.9)$$

or equivalently, for bounded measurable functions f_i ,

$$\mathbf{E}(\Pi_{i=1}^m f_i(x_{n_i+T}) | \mathcal{F}_T) = \mathbf{E}(\Pi_{i=1}^m f_i(x_{n_i+T}) | x_T). \quad (6.10)$$

6.3.1 Markov property at finite stopping times

Recall that given any probability distribution μ and any transition probabilities P there exists a unique probability measure \mathbb{P}_μ on \mathcal{X}^∞ which is the probability distribution of a Markov process (we sometimes denote such a process by X^x) with transition probabilities \mathbb{P} and initial distribution \mathbb{P}_μ . If $\mu = \delta_s$, this is denoted by \mathbb{P}_s . If $\Phi : \mathcal{X}^\infty \rightarrow \mathbf{R}$ is bounded and measurable, we write $\mathbf{E}_x[\Phi]$ for the integral of Φ with respect to \mathbb{P}_s ,

$$\mathbf{E}_x[\Phi] = \int_{\mathcal{X}^\infty} \Phi d\mathbb{P}_x.$$

Using the process X^x , this is $\mathbf{E}[\Phi(X^x)]$.

Example 6.3.1 Let (x_n) is the Markov chain with transition probabilities P and initial condition x . If $\Phi(\sigma) = f(\sigma_3, \sigma_7)$ for some $g : \mathcal{X}^2 \rightarrow \mathbf{R}$, then

$$\mathbf{E}_x[\Phi] = \mathbf{E}[f(x_3, x_7)] = \int f(y_1, y_2) P^3(x, dy_1) P^4(y_1, dy_2),$$

If $\Phi = \pi_{i=1}^m f_i \circ \pi_i$, where π are the projections (coordinate mappings) and $f_i \in \mathcal{B}_b(\mathcal{X})$, then

$$\mathbf{E}_x[\Phi] = \int_{\mathcal{X}} \cdots \int_{\mathcal{X}} \Pi_{i=1}^m f_i(y_i) \Pi_{i=1}^m P^{n_i - n_{i-1}}(y_{i-1}, dy_i).$$

The measure $\Pi_{i=1}^m P^{n_i - n_{i-1}}(y_{i-1}, dy_i)$ is the finite dimensional distribution of $(x_{n_1}, \dots, x_{n_m})$.

If y is a random variable, we then denote $\mathbf{E}_y[\Phi]$ for the composition:

$$\mathbf{E}_y[\Phi](\omega) := \mathbf{E}_{y(\omega)}[\Phi].$$

Similarly for $C \in \otimes^\infty \mathcal{B}(\mathcal{X})$, $\mathbb{P}_y(C)(\omega) = \mathbb{P}_{y(\omega)}(C)$.

By a finite stopping time we mean one that is almost surely finite.

Theorem 6.3.4 (Strong Markov property) *Let (x_n) be a time-homogeneous Markov process with transition probabilities P . If T is a finite stopping time, the process $(\theta_T x)_n$ is also a time-homogeneous Markov process with transition probabilities P . Furthermore if $\Phi : \mathcal{X}^\mathbb{N} \rightarrow \mathbf{R}$ is bounded Borel measurable*

$$\mathbf{E}(\Phi(\theta_T x) \mid \mathcal{F}_T) = \mathbf{E}_{x_T}[\Phi] .$$

In particular,

$$\mathbb{P}(x_{n+T} \in A \mid \mathcal{F}_T) = P^n(x_T, A), \text{ a.s.}$$

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

The proof will be given after the lemma.

Lemma 6.3.5 *Let (x_n) be a time-homogeneous Markov process with transition probabilities P and if T is a finite stopping time then*

$$\mathbb{P}(x_{n+T} \in A \mid \mathcal{F}_T) = P^n(x_T, A), \text{ a.s.}$$

for any $n > 0$ and any $A \in \mathcal{B}(\mathcal{X})$.

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} ,

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

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

$$\mathbf{E}(f(x_{n+T}) \mid \mathcal{F}_T) = \int_{\mathcal{X}} f(y) P^n(x_T, dy), \tag{6.11}$$

completing the proof. \square

Proof for Theorem 6.3.4. Let $\mathcal{G}_n = \mathcal{F}_{T+n}$ and $y_n = x_{T+n}$ then y is adapted to \mathcal{G} and $\mathbb{P}(y_{n+m} \in A | \mathcal{G}_n) = P^m(y_n, A)$, so indeed, $y := \theta_T x$ is a time-homogeneous Markov process with transition probabilities P . For the rest of the statement, it is sufficient to take Φ to be the indicator functions of cylindrical sets. Let

$$\Phi(\sigma) = \Pi_{i=1}^m f_i(x_{n_i}),$$

We assume that for any $f_j \in \mathcal{B}(\mathcal{X})$ and any $M \leq k-1$, the required identity

$$\mathbf{E}(\Pi_{i=1}^M f_i(x_{T+n_i}) | \mathcal{F}_T) = \mathbf{E}_{x_T}[\Pi_{i=1}^M f_i] = \int_{\mathcal{X}} \cdots \int_{\mathcal{X}} \Pi_{i=1}^M f_i(y_i) \Pi_{i=1}^M P^{n_i-n_{i-1}}(y_{i-1}, dy_i)$$

holds. We make induct on k , first taking an extra layer of conditional expectation then use (6.11), and then the induction hypothesis:

$$\begin{aligned} & \mathbb{P}(\Pi_{i=1}^k f_i(x_{T+m_i}) | \mathcal{F}_T) \\ &= \mathbf{E}[\Pi_{i=1}^{k-1} f_i(x_{T+m_i}) \mathbf{E}(f_k(x_{T+m_k}) | \mathcal{F}_{T+m_{k-1}}) | \mathcal{F}_T] \\ &= \mathbf{E}\left[\Pi_{i=1}^{k-1} f_i(x_{T+m_i}) \int_{\mathcal{X}} f_k(y_k) P^{m_k-m_{k-1}}(x_{T+m_{k-1}}, dy_k) | \mathcal{F}_T\right] \\ &= \int_{\mathcal{X}} \cdots \int_{\mathcal{X}} \Pi_{i=1}^{k-1} f_i(y_i) \int_{\mathcal{X}} f_k(y_k) P^{m_k-m_{k-1}}(y_{k-1}, dy_k) \Pi_{i=1}^m P^{m_i-m_{i-1}}(y_{i-1}, dy_i) \\ & \quad \int_{\mathcal{X}} \Pi_{i=1}^k f_k(y_k) \Pi_{i=1}^k P^{m_i-m_{i-1}}(y_{i-1}, dy_i). \end{aligned}$$

This complete the proof.

6.3.2 Markov property at non-finite stopping times

Lemma 6.3.6 *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. Restricted to the set $\{T < \infty\}$, x_T is 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 6.3.7 *Let (x_n) be a time-homogeneous Markov process with transition probabilities P and let T be a stopping time. Let $\Phi : \mathcal{X}^{\infty} \rightarrow \mathbf{R}$ be a function. Then on the set $\{T < \infty\}$,*

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

In other words,

$$\mathbf{E}(\Phi(\theta_T x) \mathbf{1}_{\{T < \infty\}} | \mathcal{F}_T) = \mathbf{E}_{x_T}[\Phi] \mathbf{1}_{\{T < \infty\}}.$$

Proof This is left as an exercise. 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 $f : \mathcal{X} \rightarrow \mathbf{R}$ be bounded measurable, and $B \in \mathcal{F}_T$,

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

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

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

We can also interpret this as

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

concluding the proof. \square

As an application to Theorem 6.3.7, we study an example.

Example 6.3.2 Simple Random Walk on \mathbf{Z} . Let ξ be i.i.d. such that $\mathbb{P}(\xi = \pm 1) = 1/2$, and define $S_n = x + \sum_{i=1}^n \xi_i$, letting $x = 0$. Define $T_i = \inf\{n \geq 0, S_n = i\}$ and we use the notation $\mathbb{P}_i(\dots) = \mathbb{P}(\dots | x_0 = i)$. We will show later in Example 7.5.3 that $\mathbb{P}_1(T_1 < \infty) = 1$. Let us use this and the strong Markov property to show that $\mathbb{P}_0(T_1 < \infty) = 1$.

Proof. We give one proof, another proof is given later in Example 7.10.3. Suppose for a contradiction $\mathbb{P}_0(T_1 < \infty) < 1$. Let $B = \{\omega : \exists \text{ path from 1, passing through 0}\}$. If $\omega \in B$, once $S_*(\omega)$ has reached 0, it restarts as a random walk (with $x = 0$) and with positive probability it does not reach 1, since $\mathbb{P}_0(T_1 < \infty) < 1$. Hence we deduce that $\mathbb{P}_1(T_1 < \infty) < 1$ too, contradicting recurrent property. Formally,

$$\begin{aligned} \mathbb{P}_1(T_1 = \infty) &\geq \mathbb{P}_1(T_1 = \infty, T_0 < \infty) = \mathbb{E}_1(\mathbb{P}_1(T_1 = \infty | \mathcal{F}_{T_0}) \mathbf{1}_{T_0 < \infty}) \\ &= \mathbb{E}_1(\mathbb{P}_{S_{T_0}}(T_1 = \infty) \mathbf{1}_{T_0 < \infty}) = \mathbb{E}_1(\mathbb{P}_0(T_1 = \infty) \mathbf{1}_{T_0 < \infty}) \\ &= \mathbb{P}_0(T_1 = \infty) \mathbb{P}_1(T_0 < \infty), \end{aligned}$$

We used the strong Markov property in the second line. Since both $\mathbb{P}_1(T_0 < \infty) > 0$ (path existence) and $\mathbb{P}_0(T_1 = \infty) > 0$ (assumption for contradiction), we infer $\mathbb{P}_1(T_1 = \infty) > 0$, violating the fact that $\mathbb{P}_1(T_1 < \infty) = 1$.

This marks the end of lecture 7.

Chapter 7

Time Homogeneous Markov Chains on Discrete State Spaces

Let us consider time homogenous Markov Chains (THMC) (x_n) on discrete state space \mathcal{X} . If \mathcal{X} has a finite number of elements, we shall consider $\mathcal{X} = \{1, 2, \dots, N\}$, otherwise $\mathcal{X} = \{1, 2, \dots\}$. If the THMC (x_n) has transition probabilities $P = (P_{ij})$ on \mathcal{X} with initial distribution $\nu = \mathcal{L}(x_0)$, then the distribution of x_n is $\nu P^n = \mathcal{L}(x_n)$.

The number P_{ij} should be interpreted as the probability of jumping from state i to state j and $P_{ij}^n = \mathbb{P}(x_n = j | x_0 = i)$. Thus, $\sum_{j \in \mathcal{X}} P_{ij} = 1$ for every i . For a finite state space we have a matrix $P = (P_{ij})$. If \mathcal{X} is a countable space, we have a matrix with infinite entries, and $\sum_{i=1}^{\infty} P_{ij} = 1$.

Definition 7.0.1 We call a matrix P with positive entries which satisfies $\sum_{i=1}^N P_{ij} = 1$ for all j a **stochastic matrix**.

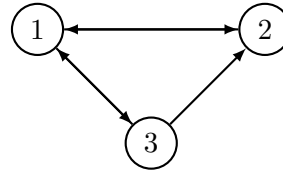
The initial distribution ν is represented by a row vector on $[0, 1]^{\mathcal{X}}$ with i -th entry $\nu(\{i\})$. We use shorthand notation $\nu(i)$ or ν_i for $\nu(\{i\})$. Thus $(\nu P)(i) = \sum_{k \in \mathcal{X}} \nu(k) P_{ki}$. If the size of \mathcal{X} is finite, i.e. $\#(\mathcal{X}) = N$, ν is a row vector, P is an $N \times N$ matrix and νP is a matrix product. 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 $T : \mathcal{P}(\mathcal{X}) \rightarrow \mathcal{P}(\mathcal{X})$ is a linear operator from \mathbf{R}^N to \mathbf{R}^N which preserves probability measures.

7.1 Communication Classes - Lecture 8

We demonstrate this with $\mathcal{X} = \{1, \dots, N\}$. Fixing 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 $P_{ji} \neq 0$. This strategy works well for chains with a manageable number of states.

Example 7.1.1 For example, take P below. If $P_{ij} \neq 0$, we draw an oriented graph and obtain the incidence graph. Following the arrows in the graph, we can reach any vertex from any other.

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

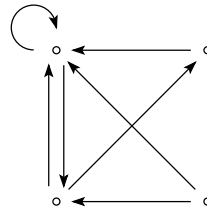


From 2, we can see $P_{21} > 0$ and $P_{13} > 0$, so that

$$P_{23}^2 = \mathbb{P}(x_2 = 3 | x_0 = 2) = P_{21}P_{13} > 0.$$

Example 7.1.2 Another example is

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



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.

Still in this finite state situation, 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. 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. A general stochastic matrix is not irreducible. It can however be broken up into irreducible components in the following way. The set $\{1, \dots, N\}$ 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 definition. For example, the matrix given in (7.1.2) is reducible because it is impossible to reach 4 from any of the other points in the system.

Let us now give the definition that applies to a countable state space., in which case the stochastic matrix has infinite number of rows and columns.

Definition 7.1.1 Let \mathcal{X} be a countable space, we have the following definitions:

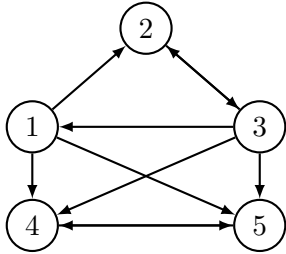
1. We say that j is accessible from i , if $P_{ij}^n > 0$ for some n . This is denoted in symbol by $i \rightarrow j$.
2. Two states i and j are said to communicate with each other, if there exist $m, n \geq 0$ such that $P_{ij}^n > 0$ and $P_{ji}^m > 0$. This is denoted by $i \sim j$.
3. The set of states $[i] = \{j \in \mathcal{X} : j \sim i\}$ is the communication class containing i .
4. A chain is said to be irreducible if there exists only one communication class, otherwise is reducible.

In the case of Example (7.1.2), we have $[1] = \{1, 2, 3\}$ and $[4] = \{4\}$.

Example 7.1.3 Similarly to previous example we can observe from the incidence graph below

that in this case the states $\{1, 2, 3\}$ cannot be reached from 4.

$$P = \frac{1}{5} \begin{pmatrix} 1 & 1 & 0 & 2 & 1 \\ 0 & 0 & 5 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 3 & 2 \\ 0 & 0 & 0 & 3 & 2 \end{pmatrix}$$

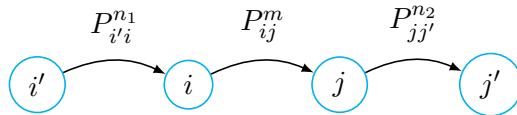


In this example, the chain is reducible as we have two communication classes

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

Exercise 7.1.1 Check \sim defines an equivalence relation.

Lemma 7.1.2 If $i \rightarrow j$, i.e. j is accessible from i , then any $j' \in [j]$ is accessible from any element $i' \in [i]$.



Proof. Fixed $j' \in [j]$ and $i' \in [i]$, we want to show that there exists n with $P_{i'j'}^n > 0$. By assumption, there exist some n_1, n_2, m such that $P_{ij}^m > 0, P_{i'i}^{n_1} > 0$ and $P_{jj'}^{n_2} > 0$.

Then by Chapman Kolmogorov equation,

$$P_{i'j'}^{n_1+m+n_2} = \sum_{k \in \mathcal{X}} \sum_{l \in \mathcal{X}} P_{i'k}^{n_1} P_{kl}^m P_{lj'}^{n_2} \geq P_{i'i}^{n_1} P_{ij}^m P_{jj'}^{n_2} > 0.$$

This implies $i' \rightarrow j'$. □

This means the set of equivalence classes is endowed with a partial order \leq : we say that $[i] \leq [j]$ if we can access an element of $[i]$ from an element of $[j]$. Equivalently, there exist a path from j to i of positive probability.

Remark 7.1.3 For a finite state space, $[i] \leq [j]$ if and only if there is a path on Γ going from j to i . In Example (7.1.2), one has $[1] \leq [4]$. Note that this order is not total, so it may happen that one has neither $[i] \leq [j]$ nor $[j] \leq [i]$. 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 states in the minimal classes (the recurrent states). This justifies the terminology “transient” for the other states, since they will only ever be visited a finite number of times.

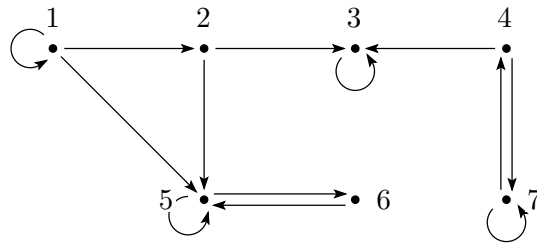
Exercise 7.1.2 Check that the relation \leq defined above is indeed a partial order.

Definition 7.1.4 An equivalence class $[i]$ is said to be minimal if there exists **no** j such that $[j] \leq [i]$ and $[j] \neq [i]$. A minimal class is also said to be closed.

Returning to example 7.1.3, we have $1 \rightarrow 4$ and $[4] \leq [1]$, implying $[4]$ is the minimal class.

Remark 7.1.5 The state \mathcal{X} can be decomposed, it is the disjoint unions of the communication classes. If $[i]$ is closed, there is no path from any $k \in [i]$ to the other communication classes. In other words, for any $k \in [i]$ and $j \notin [i]$, there is no path from k to j .

Example 7.1.4 Consider a stochastic matrix such that the associated graph is given by

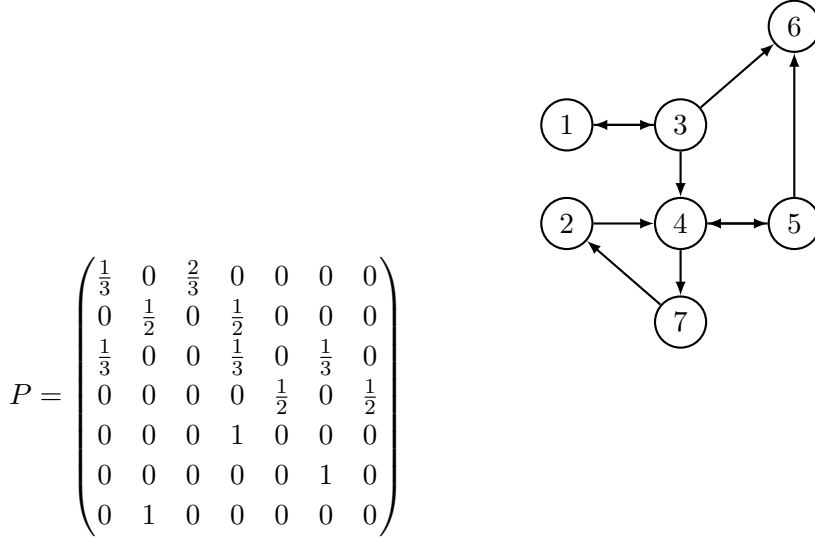


In this case, the communication classes are given by

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

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 7.1.5 Consider the stochastic matrix P with its incidence graph



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.

Definition 7.1.6 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

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

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 7.1.6 Let us return to the two state Markov chain, take $x_0 \sim \mu$, then we use \mathbb{P}_μ to denote the probability concerning the chain with $x_0 \sim \mu$.

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

If $x_0 = \delta_0$,

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

For $n \geq 1$,

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

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

Example 7.1.7 (*A Lazy Walker (Birth and Death Process)*)

Let us consider the Markov Chain on state space $\mathcal{X} = \mathbf{Z}$ with transition P given by

$$P_{ij} = \begin{cases} \frac{1}{2} & \text{if } i = j, \\ \frac{1}{4} & \text{if } j = i - 1, \\ \frac{1}{4} & \text{if } j = i + 1. \end{cases} \Rightarrow P = \frac{1}{4} \begin{pmatrix} \ddots & \ddots & & & \\ & \ddots & 2 & 1 & \\ & & 1 & 2 & 1 \\ & & & 1 & 2 & \ddots \\ & & & & \ddots & \ddots \end{pmatrix}. \quad (7.1)$$

Let $j - i = n$, then, as in the argument of Lemma 7.1.2 we have

$$P_{ij}^n \geq P_{ii+1} \cdots P_{j-1j} \geq \left(\frac{1}{4}\right)^n, \quad P_{ji}^n \geq P_{jj-1} \cdots P_{i+1i} \geq \left(\frac{1}{4}\right)^n.$$

Hence the chain is irreducible.

We can now answer whether there exists an invariant measure for P , which means there exists a solution ν to $\nu P = \nu$. By the note above, $\nu P = \nu$ is equivalent to require for any $j \in \mathcal{X}$ that

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

Since $P_{jj} = \frac{1}{2}$ and $P_{jj+1} = P_{jj-1} = \frac{1}{4}$, all other P_{ij} vanish,

$$\nu(j) = \frac{1}{4}\nu(j-1) + \frac{1}{4}\nu(j+1) + \frac{1}{2}\nu(j) \implies \nu(j) - \nu(j-1) = \nu(j+1) - \nu(j). \quad (7.2)$$

Hence any ν with $\nu(j)$ constant for all j satisfies the relation (7.2). Such a ν is a uniform measure on \mathcal{X} , but not a probability measure.

One may wonder if there exists any other solution to $\nu P = \nu$? Let $\nu(0) = a \geq 0$. Suppose that $b = \nu(j+1) - \nu(j) > 0$. Then $\nu(j)$ will be negative for $j < 0$ sufficiently small. Similarly if $b < 0$, $\nu(j)$ becomes negative for j sufficiently large. So there exists only one solution, up to a multiplicative constant, with $\nu(j) \geq 0$. This is the uniform measure.

Conclusion. For the lazy walk, there exists a measure with $\nu P = \nu$, which is unique up to a multiplicative constant, but no invariant *probability* measure.

7.2 Recurrence and Transience

There are further important properties of THMC communication classes. Let $T_i = \inf\{n \geq 1 : x_n = i\}$.

Definition 7.2.1 A state i is recurrent if $\mathbb{P}(T_i < \infty | x_0 = i) = 1$. Otherwise it is said to be transient.

Notation. For brevity we may use the following notation $\mathbb{P}_i(A) := \mathbb{P}(A|x_0 = i)$, so that $\mathbb{P}_i(T_i < \infty) = \mathbb{P}(T_i < \infty|x_0 = i)$; also we denote $\mathbf{E}_i(Y) := \mathbf{E}[Y|x_0 = i]$ for an integrable random variable Y .

Definition 7.2.2 A Markov chain on \mathcal{X} is recurrent if every state $i \in \mathcal{X}$ is recurrent; it is transient if every state is transient.

We will see later that the existence of a recurrent state implies the existence of an invariant measure. Also, a state i is recurrent if and only if it is visited infinitely often almost surely. And also, being transient/recurrent is a class property, a property of the communication class. At this point we note:

Lemma 7.2.3 Given two states $i, j \in \mathcal{X}$, then $i \rightarrow j$ if and only if $\mathbb{P}_i(T_j < \infty) > 0$. Moreover

$$\mathbb{P}_i(T_j < \infty) \leq \sum_{n=1}^{\infty} P_{ij}^n \quad (7.3)$$

Proof. The (\Rightarrow) direction holds trivially, as $\{T_j < \infty\} = \bigcup_{n=1}^{\infty} \{T_j = n\}$. For the other direction (\Leftarrow) , we can derive (7.3) by

$$\mathbb{P}_i(T_j < \infty) \leq \sum_{n=1}^{\infty} \mathbb{P}_i(T_j = n) \leq \sum_{n=1}^{\infty} \mathbb{P}_i(x_n = j) = \sum_{n=1}^{\infty} P_{ij}^n.$$

Hence $\mathbb{P}_i(T_j < \infty) > 0$ implies that $P_{ij}^n > 0$ for some n . □

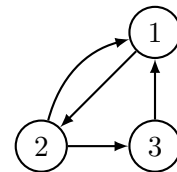
Exercise 7.2.1 Are the states in Example 7.1.1 recurrent?

7.2.1 Another way for guessing the invariant measures –Lecture 9

Whether a state i is recurrent is an important question of its own. It is actually associated with the existence of an invariant probability measure (an equilibrium).

Let us start with an example.

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



Two paths loop back to 1: $\omega_1 : 1 \rightarrow 2 \rightarrow 1$ and $\omega_2 : 1 \rightarrow 2 \rightarrow 3 \rightarrow 1$. Each of these paths occurs with probability $\frac{1}{2}$. The state 2 is visited on both paths, so we give it weight 2. The state 3 is visited on path 2, it is visited on the average $\frac{1}{2}$ times. Let us define $\nu = (1, 1, \frac{1}{2})$. Check this is an invariant measure for P . Do you expect this to hold more generally?

7.3 Passage times

Throughout this section x_n be a time homogeneous Markov chain. For the following lemma, let us define:

$$T_j^0 := 0, \quad T_j^1 = T_j, \quad T_j^{n+1} = \inf\{k > T_j^n : x_k = j\} \quad \text{for } n \geq 1. \quad (7.4)$$

The stopping times T_j^n are also called the passage times to j , the times $T_j^n - T_j^{n-1}$ are the length of the n -th excursion to the state j .

Lemma 7.3.1 *Let $i, j \in \mathcal{X}$. If j is recurrent and $\mathbb{P}_j(T_i < \infty) > 0$, then $\mathbb{P}_i(T_j < \infty) = 1$. If μ is an initial distribution supported on $[j]$, then $\mathbb{P}_\mu(T_j < \infty) = 1$.*

We first give an hand waving argument. If $\mathbb{P}_i(T_j < \infty) < 1$, then exists a set of path A which starts from i never coming back to j . Since $\mathbb{P}_j(T_i < \infty) > 0$, j can access i , there is a path of shortest length m from j to i . We concatenate this path to a path from A that starts from j never visits i , thus obtaining a set of paths from i , never returns to i . This means $\mathbb{P}_i(T_i < \infty) < 1$ contradicting with the assumption.

Proof. It is sufficient to prove this statement for $\mu = \delta_j$. Since $\mathbb{P}_i(T_j < \infty) > 0$, j is accessible from i . Let m be the smallest number such that $P_{ij}^m > 0$, which means that $\mathbb{P}_i(x_1 \neq j, \dots, x_{m-1} \neq j, x_m = j) > 0$. Suppose that $\mathbb{P}_j(T_i = \infty) > 0$, this mean once in j the paths has positive probability not returning to i . Our conclusion follows by the reasoning below:

$$\begin{aligned} \mathbb{P}_i(T_i = \infty) &\geq \mathbb{P}_i(T_i = \infty, T_j = m) = \mathbf{E}[\mathbb{P}(T_i = \infty, T_j = m | \mathcal{F}_{T_j})] \\ &= \mathbf{E}[\mathbb{P}(x_{T_j+\ell} \neq i, \forall \ell \geq 1, T_j = m | \mathcal{F}_{T_j})] = \mathbf{E}[\mathbb{P}(x_{T_j+\ell} \neq i, \forall \ell \geq 1 | \mathcal{F}_{T_j}) \mathbf{1}_{T_j=m}] \\ &= \mathbb{P}_j(T_i = \infty) \mathbb{P}_i(T_j = m) > 0. \end{aligned}$$

This is in contradiction with i recurrent. In the final line we used the Markov property, noting that $\{x_0 = i, T_j = m\} = \{x_1 \neq j, \dots, x_{m-1} \neq j, x_m = j, x_0 = i\}$ and the fact that m is the shortest length of a path from i to j , so a path from i to j in m -steps does not visit i for $n = 1, \dots, m$. \square

An alternative proof with elementary probability is as follows. Let $B = \{x_0 = i, x_k \notin \{i, j\}, k = 1, \dots, m-1, x_m = j\}$. Then m is the shortest length from i to j implies that $\{x_0 = i, T_j = m\} = B$ and

$$\begin{aligned} \mathbb{P}_i(T_i = \infty) &\geq \mathbb{P}_i(T_i = \infty, T_j = m) = \mathbb{P}(x_0 = i, x_k \neq i, \forall k \geq 1, T_j = m) \\ &= \mathbb{P}(\{x_0 = i, T_j = m\} \cap \{x_{m+\ell} \neq i, \forall \ell \geq 1\}) \\ &= \mathbb{P}(\{x_{m+\ell} \neq i, \forall \ell \geq 1\} | x_0 = i, T_j = m) \mathbb{P}_i(T_j = m) \\ &= \mathbb{P}_j(T_i = \infty) \mathbb{P}_i(T_j = m) > 0. \end{aligned}$$

The following lemma on inter-arrival times is primarily interesting for irreducible recurrent chains.

Lemma 7.3.2 *Let x_n be a Markov process starting with initial distribution μ and $\mathbb{P}_\mu(T_j < \infty) = 1$ for a recurrent state j . Then, the intervals $\{T_j^n - T_j^{n-1}\}_{n \geq 1}$ are independent. And for any $k \geq 1$, $m \in \mathcal{X}$,*

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

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

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

By the strong Markov property, for $k \geq 1$,

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

The second line follows from the strong Markov property and that $T^{k+1} - T^k = m$ if and only if $(x_{T^k+1} \neq j, \dots, x_{T^k+m-1} \neq j, x_{T^k+m} = j)$. (From the point of view of the shifted process $\theta_{T^k}x.$, this is the hitting time of j .) Taking expectations we see that

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

To see that $\{T_j^k - T_j^{k-1}\}$ are independent random variables, it is sufficient to observe that $\mathbb{P}(T^{k+1} - T^k = m | \mathcal{F}_{T^k})(\omega)$ is a deterministic event. In fact, for any $A \in \mathcal{F}_{T^k}$, we have

$$\mathbf{E}(\mathbf{1}_A \mathbf{1}_{\{T^{k+1} - T^k = m\}}) = \mathbf{E}(\mathbf{1}_A \mathbb{P}_j(T = m)) = \mathbb{P}(A) \mathbb{P}_j(T = m),$$

where in the first equality we conditioned w.r.t. \mathcal{F}_{T^k} inside expectation. Hence $\{T^{k+1} - T^k = m\}$ is independent of \mathcal{F}_{T^k} and $T_j \in \mathcal{F}_{T^k}$ for any $j = k - 1$. We conclude that $\{T^0, T^2 - T^1, T^3 - T^2, \dots\}$ are independent random variables. \square

Remark 7.3.3 For those with curious mind, let us check the case where the THMC does not reach j with probability one from its initial state.

1. Then by the strong Markov property, Theorem 6.3.7, we see that

$$\mathbb{P}(T_j^{k+1} - T_j^k = m, T_j^k < \infty | \mathcal{F}_{T_j^k}) = \mathbb{P}_j(T_j = m) \mathbf{1}_{T_j^k < \infty},$$

taking expectation to see that $\mathbb{P}(T_j^{k+1} - T_j^k = m, T_j^k < \infty) = \mathbb{P}_j(T_j = m)\mathbb{P}_i(T_j^k < \infty)$, so For any $m \in \mathcal{X}$, and any any initial distribution,

$$\mathbb{P}(T_j^{k+1} - T_j^k = m \mid T_j^k < \infty) = \mathbb{P}_j(T_j = m).$$

In the above statement, we do not specify an initial distribution.

2. Do we maintain the statement that the passage times are independent if j is not recurrent? For any statement of the kind, we must assign a value to $T_j^{k+1} - T_j^k$ when $T_j^k = \infty$. Let us define a family of identically distributed independent random variables $\xi_l : \mathcal{X} \rightarrow \mathbf{N} \cup \{+\infty\}$ with $\mathbb{P}(\xi_l = m) = \mathbb{P}_j(T_j = m)$. Set $\eta_1 = T_j$, and for $k \geq 1$,

$$\eta_{k+1} = \begin{cases} T_j^{k+1} - T_j^k, & \text{if } T_j^k < \infty, \\ \xi_{k+1}, & \text{otherwise.} \end{cases}$$

Claim: $\{\eta_k\}$ are independent random variables. Indeed, let $\mathcal{G}_k = \mathcal{F}_{T_j^k} \vee \sigma\{\xi_2, \dots, \xi_k\}$. Let $k \geq 2$,

$$\begin{aligned} \mathbb{P}(\eta_k = m | \mathcal{G}_k) &= \mathbb{P}(T_j^{k+1} - T_j^k = m, T_j^k < \infty | \mathcal{G}_k) + \mathbb{P}(\xi_{k+1} = m, T_j^k = \infty | \mathcal{G}_k) \\ &= \mathbf{1}_{T_j^k < \infty} \mathbb{P}_j(T_j = m) + \mathbb{P}(\xi_{k+1} = m | \mathcal{G}_k) \mathbf{1}_{T_j^k = \infty} \\ &= \mathbf{1}_{T_j^k < \infty} \mathbb{P}_j(T_j = m) + \mathbb{P}_j(T_j = m) \mathbf{1}_{T_j^k = \infty} = \mathbb{P}_j(T_j = m). \end{aligned}$$

We used that η_{k+1} is independent of \mathcal{G}_k . Since $\mathbb{P}(\eta_k = m | \mathcal{G}_k)$ is a constant, we conclude that η_k is independent of \mathcal{G}_k .

Note that $T_j^n = \sum_{k=1}^n (T_j^k - T_j^{k-1})$ and $\{T_j^n < \infty\} = \cap_{k=0}^n \{T_j^k - T_j^{k-1} < \infty\}$. This motivates the following useful lemma.

Lemma 7.3.4 *For any two states i, j , any natural number $k \geq 1$,*

$$\mathbb{P}_i(T_j^{k+1} < \infty) = \mathbb{P}_i(T_j < \infty) \cdot \mathbb{P}_j(T_j^k < \infty). \quad (7.5)$$

In particular, for any $j \in \mathcal{X}$, and $k = 2, \dots$,

$$\mathbb{P}_j(T_j^k < \infty) = (\mathbb{P}_j(T_j < \infty))^k. \quad (7.6)$$

Proof. Let $\Phi : \mathcal{X}^\infty \rightarrow \mathbf{R}$ be the function defined below:

$$\Phi((a_n)) = \begin{cases} 1, & \text{if } a_{n_1} = j, \dots, a_{n_k} = j \text{ for some } 1 \leq n_1 < n_2 < \dots < n_k, \\ 0, & \text{if otherwise.} \end{cases}$$

Thus, $\Phi = \mathbf{1}_A$ where A contains sequences that visits j at least k -times at some finite time $n \geq 1$.

Then, on $\{T_j < \infty\}$, $T_j^{k+1}(\omega) < \infty$ if and only if

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

We apply the strong Markov property (Theorem 6.3.7 for stopping times that are not necessarily finite) to obtain:

$$\begin{aligned} \mathbf{E}(\mathbf{1}_{\{T_j^{k+1} < \infty\}} \mathbf{1}_{\{T_j < \infty\}} | \mathcal{F}_{T_j}) &= \mathbf{E}(\Phi(\theta_{T_j} x.) \mathbf{1}_{\{T_j < \infty\}} | \mathcal{F}_{T_j}) \\ &= \mathbf{1}_{\{T_j < \infty\}} \mathbf{E}_{x_{T_j}}(\Phi(x.)) = \mathbf{1}_{\{T_j < \infty\}} \mathbb{P}_j(T_j^k < \infty). \end{aligned}$$

Since $x_{T_j} = j$, we take the expectation (conditional on $x_0 = j$) on both sides to obtain that:

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

(Owing to the tower property, the left hand side becomes $\mathbf{E}(\mathbf{1}_{\{T_j^{k+1} < \infty\}} \mathbf{1}_{\{T_j < \infty\}} | x_0 = j)$. The second indicator function can be removed as it poses no restriction: $\mathbf{1}_{\{T_j^{k+1} < \infty\}} \mathbf{1}_{\{T_j < \infty\}} = \mathbf{1}_{\{T_j^{k+1} < \infty\}}$.) We use the fact that $x_{T_j} = j$. The expression $\mathbf{E}_j(\Phi(x.))$ is the same $\mathbf{E}_j[\Phi]$ the latter means integration of Φ with respect to the probability measure of the chain with initial distribution \mathbb{P}_{δ_j} the first is integration on the probability space of the random variable $\Phi \circ x$ and $x_0 = j$. One can now take the expectation (conditional in $x_0 = j$) on both sides to obtain that:

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

The right hand side is obvious, since $\mathbb{P}_j(T_j^k < \infty)$ is non-random. On the left hand side, $\mathbf{1}_{\{T_j^{k+1} < \infty\}} \mathbf{1}_{\{T_j < \infty\}} = \mathbf{1}_{\{T_j^{k+1} < \infty\}}$, and

$$\mathbf{E}(\mathbf{E}_i(\mathbf{1}_{\{T_j^{k+1} < \infty\}} | \mathcal{F}_{T_j})) = \mathbf{E}_i(\mathbf{1}_{\{T_j^{k+1} < \infty\}}) = \mathbb{P}_i(T_j^{k+1} < \infty).$$

Take $i = j$, we see that

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

Inducting on k , we see that $\mathbb{P}_j(T_j^{k+1} < \infty) = \mathbb{P}_j(T_j < \infty)^{k+1}$. □

Example 7.3.1 *This example is not given in the lectures.* Let us look at a Markov model. Suppose that customers arrive independently. We represent their arrival as 1 and their departure as 0, and denote it by X_n . Assume X_n are i.i.d. Bernoulli random variable on $\{0, 1\}$ with $\mathbb{P}(X_i = 1) = p$. Then (X_n) is a Markov process with $\mathbb{P}(X_n = 1 | X_{n-1} = j) = p$ for any $j \in \{0, 1\}$. Let T be a stopping time, e.g. the first arrival time of somebody with surname started from A . Then X_{1+T}, X_{2+T}, \dots is also a Markov process with the same transition probability.

Let S_n denote the number of arrivals, then

$$S_n = X_1 + \dots + X_n$$

then S_n is binomial distributed:

$$\mathbb{P}(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

Let T be the first arrival time, i.e. the first time $S_n = 1$. Then T is geometrically distributed:

$$\mathbb{P}(T = k) = \mathbb{P}(X_0 = 0, \dots, X_{k-1} = 0, X_k = 1) = (1-p)^k p.$$

$\mathbf{E}[T] = \frac{1}{p}$. Since S_n is Markov, fixing m , the first time after m a customer arrives is also geometrically distributed. Let S_1 denote the first time there is an arrival from X_{m+1}, X_{m+2}, \dots . Then

$$\mathbb{P}(S_1 - m = k) = \mathbb{P}(T = k).$$

This is a special case of what we have learnt of the independent properties of the inter-arrival times T_i . Then the averaged arrival time for the k -th event is $\mathbf{E}[T_1 + \dots + T_k] = \frac{k}{p}$.

Elaborating further from this we can even consider a queuing system with maximal size of customers $\mathcal{X} = \{0, 1, \dots, N\}$. Let x_n denote the number of customers at time n . Customers arrive independently and with identically distributed Bernoulli distribution and leaves independently with identical Bernoulli distributions, independent of each other. We assume that at any time one customer arrives with rate $p \in (0, 1)$ and one customer leaves at rate $q \in (0, 1)$. A model for this is as follows: for $k \neq 0, N$, we count whether a customer arrived and whether a customer departed:

$$\mathbb{P}(x_{n+1} = j | x_n = k) = \begin{cases} p(1-q), & \text{if } j = k+1 \\ (1-p)(1-q) + pq, & \text{if } j = k \\ (1-p)q, & \text{if } j = k-1 \end{cases}$$

At position 0, which means there is no customer at time n , then

$$\mathbb{P}(x_{n+1} = 1 | x_n = 0) = p, \quad \mathbb{P}(x_{n+1} = 0 | x_n = 0) = 1 - p$$

At position N , it means the capacity is full, no new customer can arrive,

$$\mathbb{P}(x_{n+1} = N-1 | x_n = N) = q, \quad \mathbb{P}(x_{n+1} = N | x_n = N) = 1 - q.$$

Write for simplicity $a = p(1-q)$, $b = (1-p)(1-q) + pq$, and $c = (q(1-p))$. Then, we can write down this graphically as a stochastic matrix:

$$\begin{pmatrix} 1-p & p & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ b & a & c & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & b & a & c & 0 & \dots & 0 & 0 & 0 \\ & & & \dots & & & & & \\ & & & & \dots & & & & \\ & & & & & \dots & & & \\ 0 & 0 & 0 & 0 & 0 & \dots & b & a & c \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 & q & 1-q \end{pmatrix}.$$

=

7.4 Recurrence Criterion

Definition 7.4.1 Let $\eta_j = \sum_{n=1}^{\infty} \mathbf{1}_{\{x_n=j\}}$, this is the occupation time of state j , i.e. the number of times the chain visits j .

Theorem 7.4.2 (Recurrent-criterion) 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. For $\eta_j = \sum_{n=1}^{\infty} \mathbf{1}_{\{x_n=j\}}$, we have $\mathbf{E}_j[\eta_j] = \sum_{n=1}^{\infty} P_{jj}^n$. Then, with Lemma 7.3.4, we obtain

$$\sum_{n=1}^{\infty} P_{jj}^n = \mathbf{E}_j[\eta_j] = \sum_{n=1}^{\infty} \mathbb{P}_j(\eta_j \geq n) = \sum_{n=1}^{\infty} \mathbb{P}_j(T_j^n < \infty) = \sum_{n=1}^{\infty} (\mathbb{P}_j(T_j < \infty))^n.$$

In conclusion, the geometric series is convergent if and only if $\mathbb{P}_j(T_j < \infty) < 1$, i.e. if and only if j is transient. Hence j is transient if and only if $\sum_{n=1}^{\infty} P_{jj}^n < \infty$, while j is recurrent iff $\sum_{n=1}^{\infty} P_{jj}^n = \infty$. \square

Example 7.4.1 (Simple Random Walk on \mathbf{Z})

Let S_n be the simple random walk with the transition probability $\mathbb{P}(x_n = j | x_{n-1} = i) = \frac{1}{2}$ if $j = i + 1$ or $j = i - 1$. Let ξ_i be i.i.d. such that $\mathbb{P}(\xi = \pm 1) = 1/2$, and let i a given state. Let $S_n = i + \sum_{i=1}^n \xi_i$. Then in this case $T_i = \inf\{n \geq 0, S_n = i\}$ and $\mathbb{P}_i(A) = \mathbb{P}(A | x_0 = i)$. To return to i the walk must go up the same number of steps as it goes down. Hence,

$$\sum_{n=1}^{\infty} P_{ii}^n = \sum_{n=1}^{\infty} P_{ii}^{2n} = \sum_{n=1}^{\infty} \binom{2n}{n} \left(\frac{1}{2}\right)^{2n}$$

Owing to the Stirling formula: $\lim_{n \rightarrow \infty} \frac{n!}{(\frac{n}{e})^n \sqrt{2\pi n}} = 1$, we have

$$\sum_{n=1}^{\infty} \mathbb{P}_i(T_i = n) \sim \frac{1}{\sqrt{\pi}} \sum_{n=1}^{\infty} \frac{1}{\sqrt{n}} = \infty,$$

By the recurrence criterion, every state i is recurrent, leading to the conclusion that the chain is recurrent.

Corollary 7.4.3 Suppose $j \in [i]$, then j and i are either both recurrent or transient. So a state being transient or recurrent is a class property.

Proof. Assume i is recurrent, which by Corollary 7.4.2 meant that $\sum_{k=1}^{\infty} P_{ii}^k = \infty$. Since i and j are accessible from each other, we may choose m_1, m_2 so that $P_{ji}^{m_1} > 0$ and $P_{ij}^{m_2} > 0$. Note that

$$\sum_{k=m_1+m_2+1}^{\infty} P_{jj}^k \geq \sum_{n=1}^{\infty} P_{ji}^{m_1} P_{ii}^n P_{ij}^{m_2} = P_{ji}^{m_1} P_{ij}^{m_2} \sum_{n=1}^{\infty} P_{ii}^n = \infty.$$

Using the recurrence criterion again, Corollary 7.4.2, we see that i is recurrent implies that also j is recurrent, and vice versa. \square

Lemma 7.4.4 *Let $k \in \mathcal{X}$.*

Then, either $\sum_{n=1}^{\infty} P_{ij}^n = \infty$ for any $i, j \in [k]$ or $\sum_{n=1}^{\infty} P_{ij}^n < \infty$ for any $i, j \in [k]$. In particular if $[k]$ has a finite number of elements and is a minimal class, then $\sum_{n=1}^{\infty} P_{ij}^n = \infty$ for any $i, j \in \mathcal{X}$ and every element of $[k]$ is recurrent.

Proof. Suppose $\sum_{n=1}^{\infty} P_{ij}^n = \infty$ and $i, j \in [k]$. If $i', j' \in [k]$, then there exist m_1, m_2 with $P_{i'i}^{m_1} > 0$ and $P_{jj'}^{m_2} > 0$. Then

$$\sum_{n=1}^{\infty} P_{i'j'}^n \geq \sum_{n=1}^{\infty} P_{ij}^n P_{i'i}^{m_1} P_{jj'}^{m_2} = \infty.$$

On the other hand, if $|[k]| < \infty$ and $\sum_{n=1}^{\infty} P_{ij}^n < \infty$, then

$$\sum_{n=1}^{\infty} \sum_{j \in [k]} P_{ij}^n = \sum_{n=1}^{\infty} 1 = \infty,$$

leading to a contradiction. □

Theorem 7.4.5 (Dichotomy statement) *The following dichotomy hold*

1. *j is recurrent iff $P_j(x_n = j, \text{ infinitely often}) = 1$.*
2. *j is transient iff $P_j(x_n = j, \text{ infinitely often}) = 0$.*

Proof. This is a matter of expressing $x_n = j$ for an infinitely number of n 's with η_j .

$$\{x_n = j, \text{ i.o.}\} = \{\eta_j = \infty\}.$$

We begin with the increasing sequence of events $\{\eta_j > m\} = \{T_j^{m+1} < \infty\}$,

$$\lim_{m \rightarrow \infty} \mathbb{P}_j(\eta_j \leq m) = 1 - \lim_{m \rightarrow \infty} \mathbb{P}_j(T_j^{m+1} < \infty) = 1 - \lim_{m \rightarrow \infty} \mathbb{P}_j(T_j < \infty)^{m+1}.$$

Hence j is recurrent means precisely $\mathbb{P}_j(\eta_j < \infty) = 0$ in the last step. We have used Lemma 7.3.4. Likewise j is recurrent if and only if $\mathbb{P}_j(\eta_j < \infty) = 1$. □

Lemma 7.4.6 *Suppose \mathcal{X} is finite. There always exist a recurrent state. Moreover, a state is recurrent if and only if it is in a closed/minimal class.*

Proof. By Lemma 7.4.4, elements of a closed communication class from a finite state space are always recurrent. A minimal class always exists. Suppose $[i]$ is not minimal, $\exists j \in [i], k \notin [i]$ such that $P_{jk}^m > 0$ for some m . the path from j to k cannot return to $[i]$ whence $\mathbb{P}_j(T_j < \infty) < 1 - P_{jk} < 1$ concluding that j and any other element of $[j]$ is transient. □

Proposition 7.4.7 *Suppose that i, j are two states such that j is accessible from i , but i not accessible from j . Then i is transient. In particular, if $[i]$ is not closed/minimal, it contains only transient states.*

Proof. Let m be the smallest number such that $P_{ij}^m > 0$. By the Chapman-Kolmogorov equation, there exist paths from i to j of length m . Such a path from i to j (it is of shortest length) does not return to i before time m . Since i is not accessible from j , such a path cannot return to i either after time m . If x is the chain starting from i , $\{x_m = j\} = \{x_m = j, x_{m-1} \notin \{i, j\}, \dots, x_1 \notin \{i, j\}\}$ and

$$\mathbb{P}_i(T_i = \infty) \geq \mathbb{P}_i(x_m = j, x_{m-1} \neq j, \dots, x_1 \neq j) \geq P_{ij}^m > 0,$$

so j is transient. □

Definition 7.4.8 A recurrent state i is positive recurrent, if $\mathbf{E}_i T_i = \mathbf{E}(T_i | x_0 = i) < \infty$.

The standard argument for the existence of a recurrent state in a finite state space (or in a finite minimal class) is where does it go otherwise? The rigorous proof goes like the following. Let $[i]$ be a minimal class. If i is transient, every state in $[i]$ is transient by Corollary 7.4.3, and $\mathbb{P}_i(x_n = i, \text{ finitely often}) = 1$. Then there exists M such that

$$A_M = \{\omega : x_n(\omega) = i, \text{ at most } M \text{ times}\}, \quad \mathbb{P}_i(A_M) > \frac{1}{2}.$$

If $\omega \in A_M$, the chain progresses to the next state j in $[i]$ and never returns. The same argument applies to j for the set of $\omega \in A_M$. Then in finite time, with positive probability the chain never returns to any element of $[i]$, but this contradicts that $[i]$ is closed (minimal).

Example 7.4.2 Let $x_0, \xi_0, \xi_1, \xi_2, \dots$ be independent r.v.'s with ξ_n taking values in $\{1, 2, 3, \dots\}$, and define $x_{n+1} = x_n + \xi_n$ on $\mathcal{X} = \mathbf{Z}$. Then x_n moves to the rights on \mathbf{Z} , it cannot give charge at 0 since it moves away at one step. Similarly it cannot charge any state $i \in \mathbf{Z}$. Note that (x_n) is a transient walk.

Example 7.4.3 (*Simple Random Walk on \mathbf{Z}*) Consider for example the simple 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

$$\pi_n = \frac{\pi_{n+1} + \pi_{n-1}}{2}, \tag{7.7}$$

for every $n \in \mathbf{Z}$. A moment of reflection shows that the only positive solution to (7.7) with the convention $\pi_0 = 1$ is given by the constant solution $\pi_n = 1$ for every n (exercise: prove it). In fact, this is the only solution. 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.

This marks the end of lecture 9 (week 4) .

Lemma 7.4.9 Recall $\eta_j = \sum_{n=1}^{\infty} \mathbf{1}_{x_n=j}$ is the occupation time of the site j . Then,

$$\sum_{n=1}^{\infty} P_{ij}^n = \frac{\mathbb{P}_i(T_j < \infty)}{1 - \mathbb{P}_j(T_j < \infty)}.$$

Proof.

$$\begin{aligned} \sum_{n=1}^{\infty} P_{ij}^n &= \mathbf{E}_i(\eta_j) = \sum_{k=1}^{\infty} \mathbb{P}_i(\eta_j \geq k) = \sum_{k=1}^{\infty} \mathbb{P}_i(T_j^k < \infty) \\ &= \sum_{k=1}^{\infty} \mathbb{P}_i(T_j < \infty) \mathbb{P}_j(T_j^{k-1} < \infty) \\ &= \sum_{k=1}^{\infty} \mathbb{P}_i(T_j < \infty) (\mathbb{P}_j(T_j < \infty))^{k-1} \\ &= \frac{\mathbb{P}_i(T_j < \infty)}{1 - \mathbb{P}_j(T_j < \infty)}. \end{aligned}$$

In line 2 and line 3 we have applied Lemma 7.3.4. If $\mathbb{P}_j(T_j < \infty) = 1$, then $\sum_{n=1}^{\infty} P_{ij}^n = \infty$. \square

Theorem 7.4.10 If a state j is transient, then

$$\sum_{n=1}^{\infty} P_{ij}^n < \infty \quad \text{and} \quad \lim_{n \rightarrow \infty} P_{ij}^n = 0, \quad \forall i \in \mathcal{X}$$

Proof. For any $i \in \mathcal{X}$, by Lemma 7.4.9

$$\sum_{n=1}^{\infty} P_{ij}^n = \frac{\mathbb{P}_i(T_j < \infty)}{1 - \mathbb{P}_j(T_j < \infty)}.$$

Since $\mathbb{P}_j(T_j < \infty) < 1$ because j is transient, then $\sum_{n=1}^{\infty} P_{ij}^n < \infty$. Hence $\lim_{n \rightarrow \infty} P_{ij}^n = 0$. \square

Remark 7.4.11 We can give a more elementary proof the following fact, which might be more illuminating: any invariant probability measure assigns zero probability to a transient state.

Proof. Suppose π is an invariant probability measure. Suppose that i_0 is transient and with $\pi(i_0) > 0$. Owing to $\sum_{k=1}^{\infty} \pi(k) = 1$, there exists N_0 such that

$$\sum_{k=N_0}^{\infty} \pi(k) < \frac{1}{2} \pi(i_0).$$

Since $\lim_{n \rightarrow \infty} P_{ji_0}^n = 0$, $\forall j$ (by Theorem 7.4.10), there exists \tilde{N} such that for $n > \tilde{N}$, $P_{ji_0}^n < \frac{1}{2}\pi(i_0)$ for any $j \leq N_0$. But by invariance $\pi(i_0) = \pi P^n(i_0)$, so if we take $n > \tilde{N}$,

$$\begin{aligned} \pi(i_0) &= \sum_{j=1}^{N_0-1} \pi(j) P_{ji_0}^n + \sum_{j \geq N_0} \pi(j) P_{ji_0}^n \\ &\leq \sum_{j=1}^{N_0-1} \pi(j) + \frac{1}{2}\pi(i_0) \\ &< \pi(i_0). \end{aligned}$$

This is a contradiction. So π is not an invariant probability measure. \square

Note. If $|\mathcal{X}| < \infty$, $\pi(i_0) = \sum_{j=1}^{|\mathcal{X}|} \pi(j) P_{ji_0}^n \leq \max_{j \in \mathcal{X}} P_{ji_0}^n \rightarrow 0$.

Theorem 7.4.12 *If π is an invariant probability measure and if $\pi(j) > 0$ then j is recurrent.*

Proof. We have $\pi = \pi P^n$ for any $n \geq 1$. Hence for any $j \in \mathcal{X}$ we have

$$\pi(j) = \sum_{l \in \mathcal{X}} \pi(l) P_{lj}^n. \quad (7.8)$$

Summing the RHS of (7.8) over n , and using Lemma 7.4.9,

$$\sum_{n=1}^{\infty} \sum_{l \in \mathcal{X}} \pi(l) P_{lj}^n = \sum_{l \in \mathcal{X}} \pi(l) \sum_{n=1}^{\infty} P_{lj}^n = \sum_{l \in \mathcal{X}} \pi(l) \frac{\mathbb{P}_l(T_j < \infty)}{\mathbb{P}_j(T_j = \infty)} = \frac{\mathbb{P}_\pi(T_j < \infty)}{\mathbb{P}_j(T_j = \infty)}.$$

If $\pi(j) > 0$, then

$$\sum_{n=1}^{\infty} \pi(j) = \infty \iff \mathbb{P}_j(T_j = \infty) = 0,$$

hence j is recurrent. \square

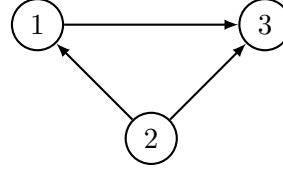
Corollary 7.4.13 *A transient Markov chain has no invariant **probability measure**.*

Example 7.4.4 Suppose $P_{ij} = \gamma_i$ for $j = i+1, \dots, i+n$, and $\sum_{i=1}^n \gamma_i = 1$, over state space $\mathcal{X} = \{1, 2, 3, \dots\}$. Then the associated Markov chain has no invariant probability measure.

Example 7.4.5 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 7.4.6 Let us consider

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



Then states $\{1\}$ and $\{2\}$ are transient states, while state $\{3\}$ is a recurrent state. An invariant measure μ on $\mathcal{X} = \{1, 2, 3\}$ is always finite. We require $\mu(1) = \mu(2) = 0$, and we can take $\mu = (0, 0, 1)$ to have total mass 1. Then μ is an invariant (probability) measure for P .

7.5 Irreducibility and reduced Markov Chains

Let us begin with an example.

Example 7.5.1 (*Restricted Chains*)

Let us consider

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

then $[4] = \{4, 5\}$ is a closed communication class. If the chain starts from $\{4, 5\}$, it stays there. Hence the chain restricts to a chain on $\{4, 5\}$ with

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

Example 7.5.2 Let us consider

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

then the communication classes are $\{1, 2\}$ and $\{3\}$ respectively. We note that $(1/2, 1/2)$ is an invariant measure for $\begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}$. Hence

$$\pi_1 = \left(\frac{1}{2}, \frac{1}{2}, 0\right), \quad \pi_2 = (0, 0, 1),$$

are invariant measures for the Markov chain associated to P . Then, with $a \in [0, 1]$, we have a family $a\pi_1 + (1 - a)\pi_2$ of invariant measures for P .

Definition 7.5.1

- A recurrent state i is positive recurrent, if $\mathbf{E}_i T_i = \mathbf{E}(T_i | x_0 = i) < \infty$.
i.e. The return time to i has first moment.
- A recurrent state with $\mathbf{E}_i T_i = \infty$ is called null-recurrent.

Positive recurrence and null recurrence are class properties.

Example 7.5.3 Let $x_n = x_{n-1} + Y_n$ where Y_i are i.i.d.'s with values in $\{1, -1\}$. Let $p \in (0, 1)$ so $\mathbb{P}(Y = 1) = p$ 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, c.f. Example 7.4.1, not positive recurrent. If $p \neq \frac{1}{2}$, the chain is transient.

To check whether a state i is recurrent, by Corollary 7.4.2 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 so precisely when $p = \frac{1}{2}$ 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 7.6.4 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 7.5.4 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 $d = 1, 2$.

7.6 Construction of invariant measure from recurrent state

Let (x_n) be a time homogeneous Markov chain with transition probabilities P . Let i be a recurrent state. For any $j \in \mathcal{X}$, define

$$\mu(j) = \mathbf{E}_i \left(\sum_{n=0}^{T_i-1} \mathbf{1}_{\{x_n=j\}} \right). \quad (7.9)$$

This is the expected number of visits to j during an excursion from i . Note that $\mu(i) = 1$. Since $\mathbb{P}_i(T_i < \infty) = 1$,

$$\mu(j) = \mathbf{E}_i \left(\sum_{n=0}^{\infty} \mathbf{1}_{\{n < T_i\}} \mathbf{1}_{\{x_n = j\}} \right) = \sum_{n=0}^{\infty} \mathbb{P}_i(x_n = j, T_i > n). \quad (7.10)$$

Theorem 7.6.1 (Existence of invariant measure) *Let i be a recurrent state. Then μ given below, defines an invariant measure.*

$$\mu(j) = \sum_{n=0}^{\infty} \mathbb{P}_i(x_n = j, T_i > n)$$

Proof. We need to show $\mu P = \mu$.

Case $j \neq i$. First we consider $j \neq i$, in this case $\mu(j) = \sum_{n=1}^{\infty} \mathbb{P}_i(x_n = j, T_i > n)$. Then,

$$\begin{aligned} \mu(j) &= \sum_{k \in \mathcal{X}} \sum_{n=1}^{\infty} \mathbb{P}_i(T_i > n, x_n = j, x_{n-1} = k, T_i > n-1) \\ &= \sum_{k \in \mathcal{X}} \sum_{n=1}^{\infty} \mathbb{P}(T_i > n, x_n = j | x_{n-1} = k, T_i > n-1) \mathbb{P}_i(x_{n-1} = k, T_i > n-1) \\ &= \sum_{k \in \mathcal{X}} \sum_{n=1}^{\infty} \mathbb{P}(x_n = j | x_{n-1} = k, T_i > n-1) \mathbb{P}_i(x_{n-1} = k, T_i > n-1) \\ &= \sum_{k \in \mathcal{X}} P_{kj} \mu(k) = (\mu P)(j). \end{aligned}$$

Where in the last line we used that $\{T_i > n-1\} \in \mathcal{F}_{n-1}$ and the definition of $\mu(k)$.

Case $j = i$. It now remains to show that $(\mu P)(i) = \mu(i)$, where

$$(\mu P)(i) = \sum_{k \in \mathcal{X}} \sum_{n=0}^{\infty} \mathbb{P}_i(T_i > n, x_n = k) P_{ki}.$$

On the other hand, we have

$$\begin{aligned} \mathbb{P}_i(T_i = n+1) &= \sum_{k \neq i} \mathbb{P}_i(T_i > n, x_{n+1} = i, x_n = k) \\ &= \sum_{k \neq i} \mathbb{P}_i(x_{n+1} = i | x_n = k, T_i > n) \mathbb{P}_i(T_i > n, x_n = k) \\ &= \sum_{k \neq i} P_{ki} \mathbb{P}_i(T_i > n, x_n = k) = \sum_{k \in \mathcal{X}} \mathbb{P}_i(T_i > n, x_n = k) P_{ki}. \end{aligned}$$

Hence

$$(\mu P)(i) = \sum_{n=0}^{\infty} \mathbb{P}_i(T_i = n+1) = \mathbb{P}_i(T_i < \infty) = 1 = \mu(i).$$

Puttign the two cases together, we showed $\mu P = \mu$. □

Corollary 7.6.2 *The invariant measure constructed in Theorem 7.6.1 has finite mass if the state i is positive recurrent.*

Proof. By (7.10), we note that

$$\sum_{j \in \mathcal{X}} \mu(j) = \sum_{n=0}^{\infty} \mathbb{P}_i(T_i > n) = \mathbf{E}_i[T_i] \quad (7.11)$$

is finite, if i is positive recurrent. \square

Lemma 7.6.3 *Let i be a recurrent state and μ the invariant measure defined at (7.9) Let ν be any other invariant measure, then*

$$\nu(k) \geq \nu(i) \mu_i(k), \quad \forall k \in \mathcal{X}. \quad (7.12)$$

Proof. Note that $\mu(i) = 1$ and equality holds for $k = i$. Let L^n be the last visit to i before n , then we can decompose

$$\Omega = \bigcup_{m=0}^{n-1} \{L^n = m\} \cup A_0, \quad A_0 = \{\text{no visits before } n\}.$$

Then

$$\mathbb{P}_j(x_n = k) \geq \sum_{m=0}^{n-1} \mathbb{P}_j(x_n = k, L^n = m),$$

so that, for $k \neq i$,

$$\begin{aligned} P_{jk}^n &\geq \sum_{m=0}^{n-1} \mathbb{P}_j(x_n = k, x_{n-1} \neq 1, \dots, x_{m+1} \neq i, x_m = i) \\ &= \sum_{m=0}^{n-1} \mathbb{P}(x_n = k, x_{n-1} \neq 1, \dots, x_{m+1} \neq i | x_m = i) P_{ji}^m \\ &= \sum_{m=0}^{n-1} \mathbb{P}_i(x_{n-m} = k, T_i > n - m) P_{ji}^m. \end{aligned}$$

Hence,

$$\begin{aligned} \nu(k) &= (\nu P^n)(k) \geq \sum_{j \in \mathcal{X}} \nu(j) \sum_{m=0}^{n-1} \mathbb{P}_i(x_{n-m} = k, T_i > n - m) P_{ji}^m \\ &= \sum_{m=0}^{n-1} \mathbb{P}_i(x_{n-m} = k, T_i > n - m) \sum_{j \in \mathcal{X}} \nu(j) P_{ji}^m \end{aligned}$$

$$= \sum_{m=0}^{n-1} \mathbb{P}_i(x_{n-m} = k, T_i > n - m) \nu(i), \quad \forall n.$$

Where in the last line we used invariance of ν . Noting

$$\mu(k) = \sum_{l=1}^{\infty} \mathbb{P}_i(x_l = k, T_i > l), \quad \sum_{m=0}^{n-1} \mathbb{P}_i(x_l = k, T_i > l) = \sum_{l=1}^n \mathbb{P}_i(x_{n-m} = k, T_i > n - m),$$

we conclude

$$\nu(k) \geq \mu(k) \nu(i) = \frac{\mu(k)}{\mu(i)} \nu(i)$$

and the proof. \square

Theorem 7.6.4 (Uniqueness of invariant measure) *If the chain is irreducible and recurrent, then the invariant measure is unique up to a multiplication constant.*

Proof. Let ν be any invariant measure and let i be any recurrent state. Let us consider μ defined as in (7.9). Since $\mu(i) = 1$,

$$0 = \nu(i) - \nu(i)\mu(i) = \nu P^n(i) - \nu(i) \mu P^n(i) = \sum_{k \in \mathcal{X}} \overbrace{(\nu(k) - \nu(i)\mu(k))}^{\geq 0} P_{ki}^n.$$

Lemma 7.6.3 implies that

$$(\nu(k) - \nu(i)\mu(k)) P_{ki}^n, \quad \forall n, k.$$

For any k, i , there exists n such that $P_{ki}^n \neq 0$, then we must have

$$\nu(k) = \nu(i)\mu(k), \quad \forall k,$$

concluding the proof. \square

Theorem 7.6.5 (Invariant probability measure and positive recurrent) *Let (x_n) be an irreducible THMC.*

1. *If the THMC has an invariant probability measure π , then $\mathbf{E}_i T_i < \infty$ for all i (i.e. all states are positive recurrent) and*

$$\pi(i) = \frac{1}{\mathbf{E}_i T_i} \quad (> 0).$$

2. *If there exists a positive recurrent state, the THMC has a unique invariant probability measure and every state is positive recurrent.*

Proof. (1) Suppose π is an invariant probability measure, then there exists i with $\pi(i) > 0$, this site i is recurrent by Theorem 7.4.12. By irreducibility, every site is recurrent. For a distinguished i , we define a measure μ as below

$$\mu(j) := \sum_{n=0}^{\infty} \mathbb{P}_i(x_n = j, T_i > n)$$

This is the construction in Theorem 7.6.1, by which we know that μ is an invariant measure. Then summing over j , we have

$$\sum_{j \in \mathcal{X}} \mu(j) = \sum_{j=1}^{\infty} \sum_{n=0}^{\infty} \mathbb{P}(x_n = j, T_i > n) = \sum_{n=0}^{\infty} \mathbb{P}_i(T_i > n) = \mathbf{E}_i T_i.$$

On the other hand, the invariant measure μ is finite by uniqueness (due to Theorem 7.6.4), concluding $\mathbf{E}_i T_i < \infty$. Since $\mu(i) = 1$,

$$\pi(i) = \frac{\mu(i)}{\mathbf{E}_i T_i} = \frac{1}{\mathbf{E}_i T_i}.$$

This procedure can be applied to any $i \in \mathcal{X}$, concluding that every state i is positive recurrent. This procedure can be applied to every state concluding the first part of the theorem.

(2) We assume that i is a positive recurrent state, then μ constructed in Theorem 7.6.1 is a finite measure (see also Corollary 7.6.2). By the previous argument, every state is positive recurrent.

□

This marks the end of lecture 10 (week 5) .

7.7 The long run probabilities

As usual, let x_n be a THMC with transition probabilities, usually denoted by P , on a countable state space \mathcal{X} . On a finite state space, we can compute P^n , in principle. In practice this is horrendous when the size of the state is not so small. We have seen that if $\lim_{n \rightarrow \infty} P_{ij}^n$ exists for any j , then $\nu(j) = \lim_{n \rightarrow \infty} P_{ij}^n$ defines an invariant measure.

Exercise 7.7.1 If $P_{i,j}^n \rightarrow \pi(j)$ as $n \rightarrow \infty$ for every i and j (the rate the Markov chain goes to state j from any other state is $\pi(j)$), show that π is an invariant probability measure.

Does the converse hold? We are now familiar with invariant measures, could we use this to our advantages? The answer is yes, if the chain satisfies a set of suitable conditions.

Firstly, if it has two distinct invariant probability measures, $\lim_{n \rightarrow \infty} P_{ij}^n$ could not agree with both sets of values. We will need to restrict to irreducible Markov chains. For the existence of a probability measure π , positive recurrence is called for.

In Theorem 7.4.10, we showed that if j is a transient state, then $\lim_{n \rightarrow \infty} P_{ij}^n = 0$ for any i . We also showed that, if π is an invariant probability measure, then $\pi(j) = 0$ for a transient state j , and in this case $\lim_{n \rightarrow \infty} P_{ij}^n = \pi(j)$ for any i .

How about the recurrent states?

Example 7.7.1 Let $P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. Since $P^{2n} = Id$ and $P^{2n+1} = P$, the two states are both recurrent. But, $P_{11}^{2n} = 1$, $P_{11}^{2n+1} = 0$, we have an alternating series, $\lim_{n \rightarrow \infty} P_{11}^n$ does not exist.

The chain with transition matrix $P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ of above, is an example of a periodic chain. To obtain a reasonable limit theorem, we exclude also periodic chains.

The main theorem is then Theorem 7.7.6 which states that under the conditions we stated, $P_{ij}^n \rightarrow \pi(j)$ for every j, i . In fact we will show the probability measure μP^n , where μ is the initial distribution, converges to π in total variation.

7.7.1 Return Times and Aperiodicity

For every state i , we define the set $R(i)$ of **return times** to i to be:

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

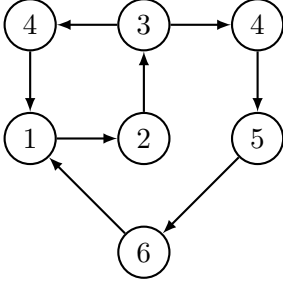
Note that $n \in R(i)$ if and only if there exists a path $i \rightarrow i$ (of positive probability) of length n . If $n, m \in R(i)$, then $n + m \in R(i)$ which follows from the Chapman-Kolmogorov equations: $P_{ii}^{n+m} = \sum_{k \in \mathcal{X}} P_{ik}^n P_{ki}^m \geq P_{ii}^n P_{ii}^m > 0$. If $R(i) \neq \emptyset$, then it is of infinite size, i.e. $|R(i)| = \infty$.

Definition 7.7.1 The **period** of the state i is then defined by

$$d(i) = \begin{cases} \gcd R(i), & \text{if } R(i) \neq \emptyset; \\ +\infty, & \text{if } R(i) = \emptyset. \end{cases}$$

($R(i) = \emptyset$ can only happen if and only if $[i]$ contains a single state from which the chain leaves straightaway and never returns, i.e. $[i] = \{i\}$ and $P_{ii} = 0$.)

Note. The period $d(i)$ may not belong to $R(i)$. It does not even necessarily mean that the chain will necessarily be able to return at time $d(i)$. See example of incidence graph below



We have $R(i) = \{4n, 6m, 4n + 6m, \dots, : n, m \geq 1\}$ and $d(i) = 2$. However the chain does not return at time $2 = d(i)$.

Definition 7.7.2 If $d(i) = 1$, we say that state i is **aperiodic**.

If $d(i) > 1$, we say that state i is **periodic**.

Note. If $P_{ii} > 0$, then $1 \in R(i)$, so that i is aperiodic.

Proposition 7.7.3 If i and j are any two 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 $n + m \in R(i) \cap R(j)$, so that both $d(i)$ and $d(j)$ divide $n + m$, and $P_{ii}^{n+m} \geq P_{ij}^n P_{ji}^m > 0$. If $k \in R(i)$ then $k + n + m \in R(j)$, as

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

Then $d(j)$ divides $n + m$ and $k + n + m$, which implies that $d(j) | k$, for any $k \in R(i)$. Hence

$$d(j) \leq d(i).$$

The same is true with i and j exchanged, i.e. $d(i) \leq d(j)$, so that one must have $d(i) = d(j)$. \square

Definition 7.7.4 A THMC (or stochastic matrix P) is **aperiodic** if $d(i) = 1$ for all $i \in \mathcal{X}$.

A THMC (or stochastic matrix P) is **periodic** of period $d > 1$ if for any $i \in \mathcal{X}$, $d(i) = d$.

As a consequence of Proposition 7.7.3, any two states in the same communication class have the same period and we can conclude the following.

Corollary 7.7.5 An irreducible chain is either periodic or aperiodic.

7.7.2 Ergodic Theorem

Theorem 7.7.6 Assume P be irreducible, aperiodic and positive recurrent. Let π denote its unique invariant probability measure. Then

$$\lim_{n \rightarrow \infty} \sum_{j \in \mathcal{X}} |P_{ij}^n - \pi(j)| = 0, \quad \forall i \in \mathcal{X}.$$

Proof. Let $(x_n, n \geq 0)$ and $(x'_n, n \geq 0)$ be independent time homogeneous Markov processes on $\mathcal{X} = \mathbf{N}$, with transition probabilities P and with initial distribution $x_0 \sim \mu$ and $x'_0 \sim \nu$ respectively.

Claim 1. The stochastic process $z_n := (x_n, x'_n)$ is a THMC on \mathcal{X}^2 with transition probabilities Q and initial distribution $\mu \otimes \nu$, where

$$Q_{(i,i'),(j,j')} = P_{ij}P_{i'j'}, \quad \forall i, i', j, j' \in \mathcal{X}.$$

Let

$$T = \inf_{n \geq 0} \{x_n = x'_n\}$$

be the coalescing time of the stochastic processes x_n and x'_n . This is the first time that z_n reaches the diagonal set $\Delta = \{(i, i) : i \in \mathcal{X}\}$.

Claim 2. We have $\mathbb{P}(T < \infty) = 1$.

Claim 3. $\sum_{i \in \mathcal{X}} |\mathbb{P}(x_n = i) - \mathbb{P}(x'_n = i)| \leq 2\mathbb{P}(T > n)$.

Note that by Claim 2, $\mathbb{P}(T > n) \rightarrow 0$ as $n \rightarrow \infty$. If we take

$$\begin{aligned} x_0 &= i, & \mathbb{P}(x_n = j) &= P_{ij}^n, \\ x'_0 &= \pi, & \mathbb{P}(x'_n = j) &= \pi(j), \end{aligned}$$

we conclude the proof. □

We now go ahead **to prove the claims** employed in Theorem 7.7.6. Claim 3 is proved in Lemma 7.7.9, Claim 1 is Lemma 7.7.10 and Claim 2 is deduced through Lemma 7.7.11-7.7.13 in Lemma 7.7.12.

Definition 7.7.7 By a coupling of two random variables X and Y with state space \mathcal{X} we mean a random variable $Z = (X', Y')$ with state space \mathcal{X}^2 such that

$$\mathcal{L}(X) = \mathcal{L}(X') \quad \text{and} \quad \mathcal{L}(Y) = \mathcal{L}(Y').$$

(So $\mathcal{L}(Z)$ is a coupling of $\mathcal{L}(X)$ and $\mathcal{L}(Y)$)

Of course we can speak of a coupling of two stochastic processes.

Set-up for next lemmas. Let $(x_n, n \geq 0)$ and $(x'_n, n \geq 0)$ be independent time homogeneous Markov processes on $\mathcal{X} = \mathbf{N}$, with transition probabilities P and with respectively initial distributions μ and ν . The process $z_n = (x_n, x'_n)$ is known as “Doeblin coupling”.

Lemma 7.7.8 (Coupling lemma)

Let $z_n = (x_n, x'_n)$ be the Doeblin coupling. Let $T = \inf_{n \geq 0} \{x_n = x'_n\}$ be the coalescing time of

the Markov processes x_n and x'_n . Define

$$y_n = \begin{cases} x_n, & n < T, \\ x'_n, & n \geq T. \end{cases}$$

Then (y_n) is a Markov process with initial distributions $\mu = \mathcal{L}(x_0)$ and transition probabilities P .

Proof. Let $\mathcal{F}_n = \sigma(x_k, k \leq n) \vee \sigma(x'_k, k \leq n)$. Let $f \in \mathcal{B}_b(\mathcal{X})$, then we have

$$\begin{aligned} \mathbf{E}[f(y_{n+1})|\mathcal{F}_n] &= \mathbf{E}[f(y_{n+1})\mathbf{1}_{\{T \leq n\}}|\mathcal{F}_n] + \mathbf{E}[f(y_{n+1})\mathbf{1}_{\{T > n\}}|\mathcal{F}_n] \\ &= \mathbf{1}_{\{T \leq n\}}\mathbf{E}[f(x'_{n+1})|\mathcal{F}_n] + \mathbf{E}[f(x_{n+1})|\mathcal{F}_n]\mathbf{1}_{\{T > n\}} \\ &= \mathbf{1}_{\{T \leq n\}}Pf(x'_n) + \mathbf{1}_{\{T > n\}}Pf(x_n) \\ &= \mathbf{1}_{\{T \leq n\}}Pf(y_n) + \mathbf{1}_{\{T > n\}}Pf(y_n) \\ &= Pf(y_n). \end{aligned}$$

In the second line we used the fact that $T > n$ implies $T \geq n + 1$ (and on $T = n + 1$ $x_{n+1} = x'_{n+1} = y_{n+1}$). We also used a consequence of Exercise 7.7.2 (see below). \square

Exercise 7.7.2 Let $\{\mathcal{G}_n\}$ and $\{\mathcal{G}'_n\}$ be independent σ -algebras. Suppose that if (x_n) is a THMC w.r.t. \mathcal{G}_n , i.e.

$$\mathbf{E}[f(x_{n+1})|\mathcal{G}_n] = Pf(x_n), \quad a.e. \quad \forall f \in \mathcal{B}_b(\mathcal{X}), \forall n \geq 0.$$

Then (x_n) is a THMC w.r.t. $\mathcal{G}_n \vee \mathcal{G}'_n$, i.e.

$$\mathbf{E}[f(x_{n+1})|\mathcal{G}_n \vee \mathcal{G}'_n] = Pf(x_n), \quad a.e.$$

Lemma 7.7.9 (Coupling inequality)

Let $z_n = (x_n, x'_n)$ be the Doeblin coupling. The following inequality (Claim 3) holds

$$\sum_{j \in \mathcal{X}} |\mathbb{P}(x_n = j) - \mathbb{P}(x'_n = j)| \leq 2\mathbb{P}(T > n).$$

Proof. Let $j \in \mathcal{X}$,

$$\begin{aligned} |\mathbb{P}(x_n = j) - \mathbb{P}(x'_n = j)| &= |\mathbb{P}(y_n = j) - \mathbb{P}(x'_n = j)| \quad (\text{by Lemma 7.7.8}) \\ &= |\mathbb{P}(y_n = j) - \mathbb{P}(x'_n = j, n < T) - \mathbb{P}(y_n = j, n \geq T)| \\ &= |\mathbb{P}(y_n = j, n < T) - \mathbb{P}(x'_n = j, n < T)|. \end{aligned}$$

Hence

$$\begin{aligned} \sum_{j \in \mathcal{X}} |\mathbb{P}(x_n = j) - \mathbb{P}(x'_n = j)| &\leq \sum_{j \in \mathcal{X}} \mathbb{P}(y_n = j, n < T) + \sum_{j \in \mathcal{X}} \mathbb{P}(x'_n = j, n < T) \\ &\leq 2\mathbb{P}(T > n). \end{aligned}$$

We proved the required inequality. \square

Lemma 7.7.10 (The Doeblin coupling)

The Doeblin coupling $z_n = (x_n, x'_n)$ is a THMC on \mathcal{X}^2 with transition probabilities Q and initial distribution $\mu \otimes \nu$, where

$$Q_{(i,i'),(j,j')} = P_{ij}P_{i'j'}, \quad \forall i, i', j, j' \in \mathcal{X}.$$

Proof. By the independence of x_0 and x'_0 , $\mathcal{L}(x_0, x'_0) = \mu \otimes \nu$. Similarly, for any $j, j' \in \mathcal{X}$ and n ,

$$\begin{aligned} \mathbb{P}(z_{n+1} = (j, j') | \mathcal{F}_n) &= \mathbb{P}(x_{n+1} = j, x'_{n+1} = j' | \mathcal{F}_n) = \mathbb{P}(x_{n+1} = j | x_n) \cdot \mathbb{P}(x'_{n+1} = j' | x'_n) \\ &= P_{x_n, j} P_{x'_n, j'} = Q_{z_n, (j, j')} \end{aligned}$$

Recall that $P_{x_n, j}$ is $P_{\cdot, j} \equiv \mathbb{P}(\cdot, \{j\})$ composed with x_n , and the others are defined similarly. This proves that $\mathbb{P}(z_{n+1} = (j, j') | \mathcal{F}_n) = \mathbb{P}(z_{n+1} = (j, j') | x_n)$ and that (z_n) is a THMC with transition probability Q . \square

Exercise 7.7.3 Check that

$$\mathbb{P}(x_{n+1} = j, x'_{n+1} = j' | \mathcal{F}_n) = \mathbb{P}(x_{n+1} = j | x_n) \cdot \mathbb{P}(x'_{n+1} = j' | x'_n).$$

Lemma 7.7.11 If P is irreducible, aperiodic, and positive recurrent, then Q is irreducible and positive recurrent.

Proof. Firstly, $\pi \otimes \pi$ is an invariant probability measure for Q . If Q is irreducible, then Q is positive recurrent.

Fact (please verify it):

$$Q_{(i,i'),(j,j')}^n = P_{ij}^n P_{i'j'}^n, \quad \forall n \geq 1.$$

We know there exists n, n' such that $P_{ij}^n > 0$ and $P_{i'j'}^{n'} > 0$. The question is whether we can find a number n such that both simultaneously positive.

Owing to Lemma 7.7.13 (see below), for any $i \in \mathcal{X}$ there exists an N such that $P_{ii}^n > 0$ for any $n > N$. By the irreducibility of P , for any i, j , there exists m with $P_{ij}^m > 0$. Then for any $n > N$,

$$P_{ij}^{n+m} \geq P_{ii}^n P_{ij}^m > 0, \quad \forall n > N.$$

Hence for any i, j , $P_{ij}^n > 0$ for every n sufficiently large. By symmetry, $P_{ji}^n > 0$ for all n sufficiently large.

To summarise, for any two pairs $(i, j), (i', j')$, we can find a common n with

$$P_{ij}^n > 0, P_{i'j'}^n > 0 \Rightarrow Q_{(i,i'),(j,j')}^n > 0,$$

proving that Q is irreducible. \square

Lemma 7.7.12 (Successful coupling)

Let P be irreducible, aperiodic, and positive recurrent. Then,

$$\mathbb{P}(T < \infty) = 1.$$

Proof. Recall coalescing time $T = \inf_{n \geq 0} \{x_n = x'_n\}$. Let

$$T_{(i,i')} = \inf\{n \geq 1 : z_n = (x_n, x'_n) = (i, i')\},$$

then $T \leq T_{(i,i')}$. Since Q is irreducible and recurrent, by Lemma 7.7.11,

$$\mathbb{P}_z(T_{(i,i')} < \infty) = 1, \quad \forall z \in \mathcal{X}^2.$$

Hence

$$\mathbb{P}(T < \infty) \geq \mathbb{P}(T_{(i,i')} < \infty) = 1.$$

This shows that the Doeblin coupling is successful. \square

We finally come back to prove the lemma using a well known result in number theory.

Lemma 7.7.13 Suppose i is aperiodic and recurrent, then $\exists N$ such that $P_{ii}^n > 0$ for every $n > N$.

Proof. Since i is recurrent we have $\sum_{n=1}^{\infty} P_{ii}^n = \infty$ and $R(i) \neq \emptyset$. By the Chinese remainder theorem, see Lemma 7.7.14 below, $\exists N$ such that any $n > N$ belongs to $R(i)$. \square

_____This marks the end of lecture 11 (the end of the week 5 lectures) . _____

A set $S \subset \mathbf{N}$ is said to have the additive property if $n, m \in S$ implies that $n + m \in S$. The following result is well-known in number theory (the Chinese remainder theorem):

Lemma 7.7.14 (A number theory lemma) Let $S \subset \mathbf{N}$ with the additive property, and not empty. Let $d = \gcd(S)$. There exists $K > 0$ such that $kd \in S$ for every $k \geq K$.

Proof. By dividing everything by d , we can assume without loss that $d = 1$. Since $\gcd S = 1$, there exists a finite collection d_1, \dots, d_n in S 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 $k \leq M$, we can choose N_0 such that $N_0 + ka_i \geq 0$ ($N_0 = M \max\{|a_1|, \dots, |a_n|\}$). By additive property of S , this implies that $NM + k \in S$ for every $k \in \{0, \dots, M\}$ and every $N \geq N_0$. Therefore, the claim holds with $K = N_0M$. \square

7.7.3 The total variation distance

Definition 7.7.15 The total variation distance between two probability measures μ and ν (in any measurable space) are

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

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

Remark 7.7.16 This is equivalent to

$$\|\mu - \nu\|_{TV} = \sup_{\substack{f \in \mathcal{B}_b(\mathcal{X}) \\ \|f\|_\infty = 1}} \left| \int_{\mathcal{X}} f(x) \mu(dx) - \int_{\mathcal{X}} f(x) \nu(dx) \right|, \quad (7.13)$$

where the maximum is run over bounded measurable functions.

It is clear that $\|\mu - \nu\|_{TV} = 0$ if and only if $\mu = \nu$. Furthermore, the total variation distance between any two probability measures is smaller or equal to two: $\|\mu - \nu\|_{TV} \leq 2$. If μ and ν are singular, there exists a measurable subset \mathcal{X}_0 such that $\mu(\mathcal{X}_0) = 1$ and $\nu(\mathcal{X}_0) = 0$. Then $\|\mu - \nu\|_{TV} \geq 2\|\mu(\mathcal{X}_0) - \nu(\mathcal{X}_0)\| = 2$ and so $\|\mu - \nu\|_{TV} = 2$. One sees that μ and ν are singular if and only if their total variation distance is the maximum value 2, c.f. Lemma 8.6.5.

Lemma 7.7.17 If μ, ν are probability measures on a discrete space \mathcal{X} , then

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

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

Proof. Let $B = \{i : \mu(i) \geq \nu(i)\}$. Since $\mu(B) - \nu(B) = \nu(B^c) - \mu(B^c)$,

$$\sum_{\{i: \mu(i) \geq \nu(i)\}} (\mu(i) - \nu(i)) = \sum_{\{i: \mu(i) < \nu(i)\}} (\nu(i) - \mu(i)) = \frac{1}{2} \sum_{i \in \mathcal{X}} |\mu(i) - \nu(i)|.$$

Also,

$$\begin{aligned} \sum_{i \in \mathcal{X}} |\mu(i) - \nu(i)| &= \sum_{\{i \in B\}} (\mu(i) - \nu(i)) + \sum_{\{i \in B^c\}} (\nu(i) - \mu(i)) \\ &= \mu(B) - \nu(B) + \nu(B^c) - \mu(B^c) = 2(\mu(B) - \nu(B)) \leq \|\mu - \nu\|_{TV}. \end{aligned}$$

For any $A \subset \mathcal{X}$,

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

Hence, $\|\mu - \nu\|_{TV} \leq \sum_{i \in \mathcal{X}} |\mu(i) - \nu(i)|$. This completes the proof. \square

Exercise 7.7.4 Let $d\mu = f dx$ and $d\nu = g dx$ on \mathbf{R}^d . Show that

$$\|\mu - \nu\|_{TV} = \int_{\mathbf{R}^d} |f(x) - g(x)| dx .$$

Definition 7.7.18 We say that a sequence $\{\mu_n\}$ converges in total variation to a limit μ if

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

Example 7.7.2 Let $\mu_n = \delta_{\frac{1}{n}}$ on \mathbf{R} . Then $\mu_n \rightarrow \delta_0$ 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 $\lim_{n \rightarrow \infty} \mu_n(A) = \mu(A)$ for every measurable set A , this is not true as can be seen in Example 7.7.3 below.

Example 7.7.3 Let Ω be the unit interval and define the probability measures

$$\mu_n(dx) = (1 + \sin(2\pi nx)) dx .$$

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

Example 7.7.4 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).

Example 7.7.5 Let $\mathcal{X} = \{1, 2\}$ and let $P = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}$. 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 $\alpha + \beta = 1$?)

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

Note $P_{11}^n - P_{21}^n = (1 - \alpha - \beta)^n$. So if $\alpha + \beta < 1$,

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

7.7.4 Convergence Theorem in Total Variation

In sight of the definition of the total variation distance between probability measure, we can see

$$\sum_{j \in \mathcal{X}} |P_{ij}^n - \pi(j)| = \|P^n(i, \cdot) - \pi\|_{TV}.$$

If $x_0 \sim \mu$, recall that $\mathbb{P}(x_n = j) = \mu P^n(j) = \sum_{i \in \mathcal{X}} \mu(i) P_{ij}^n$, and we can reformulate Theorem 7.7.6 as follows, proving it for $\mathcal{X} = \mathbf{N}$.

Theorem 7.7.19 *If x_n is an irreducible, aperiodic and positive recurrent THMC with $x_0 \sim \mu$, then*

$$\|\mu P^n - \pi\|_{\text{TV}} \rightarrow 0, \quad \text{as } n \rightarrow \infty. \quad (7.14)$$

Proof. We have shown in Theorem 7.7.6 that

$$\lim_{n \rightarrow \infty} \|P^n(i, \cdot) - \pi\|_{\text{TV}} = 0, \quad \forall i \in \mathcal{X}. \quad (7.15)$$

This is (7.14) in the case $\mu = \delta_i$, the next lemma shows that (7.15) implies (7.14). \square

Lemma 7.7.20 *Let μ be any distribution, then (7.15) implies*

$$\lim_{n \rightarrow \infty} \|\mu P^n - \pi\|_{\text{TV}} = 0. \quad (7.16)$$

Proof. Let us first note that

$$\|\mu P^n - \pi\|_{\text{TV}} = \sum_{j=1}^{\infty} \left| \sum_{i=1}^{\infty} \mu(i) P_{ij}^n - \pi(j) \right| = \sum_{j=1}^{\infty} \left| \sum_{i=1}^{\infty} \mu(i) P_{ij}^n - \sum_{i=1}^{\infty} \mu(i) \pi(j) \right|.$$

Given any $\varepsilon > 0$, we can choose N such that $\sum_{i=N+1}^{\infty} \mu(i) < \frac{\varepsilon}{4}$, so that

$$\sum_{i=N+1}^{\infty} \mu(i) \sum_{j=1}^{\infty} |P_{ij}^n - \pi(j)| < \frac{\varepsilon}{2}. \quad (7.17)$$

Since $\sum_{j=1}^{\infty} |P_{ij}^n - \pi(j)| \leq \sum_{j=1}^{\infty} P_{ij}^n + \sum_{j=1}^{\infty} \pi(j) = 2$. On the other hand, we may choose M such that for all $i \leq N$, $\sum_{j=1}^{\infty} |P_{ij}^n - \pi(j)| \leq \frac{\varepsilon}{4}$ for any $n \geq M$. Hence

$$\sum_{j=1}^{\infty} \sum_{i=1}^N \mu(i) |P_{ij}^n - \pi(j)| < \sum_{i=1}^N \mu(i) \frac{\varepsilon}{4} \leq \frac{\varepsilon}{4}. \quad (7.18)$$

Then (7.17) and (7.18) combined give us (7.16). \square

7.7.5 Periodic Chains, Cycles

An irreducible TH Markov chain decomposes into disjoint union of cycles. A characterisation of periodic chains is the following.

Lemma 7.7.21 *The period of an irreducible stochastic matrix P is the largest value $d \geq 1$ such that it is possible to write \mathcal{X} as a disjoint union of non-empty sets $A_0 \sqcup \dots \sqcup A_{d-1}$ in such a way that if $i \in A_n$ and $P_{ij} > 0$, then $j \in A_{n+1}$. We have identified A_{n+kd} with A_n .*

Proof. 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:

$$A_0 = \{j : P_{1j}^{kd} > 0 \text{ for some } k \in \mathbf{N}\},$$

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

$$A_n = \{j \mid \exists m = 0 \ (P_{1j}^{n+kd} > 0 \text{ for some } k \in \mathbf{N})\}. \quad (7.19)$$

Claim. $\{A_n\}$ are disjoint and $\mathcal{X} = \bigcup_{k=1}^{d-1} A_k$.

Since P is assumed to be irreducible, for any j , there exists n such that $P_{1j}^n > 0$, hence the union of the A_n is all of \mathcal{X} . Furthermore, they are disjoint. Otherwise, one could find j such that it belongs to $A_{n_1} \cap A_{n_2}$. So $P_{1j}^{n_1+k_1d} > 0$ and $P_{1j}^{n_2+k_2d} > 0$ 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 $P_{j1}^q > 0$, so that $n_1 + k_1d + q \in R(1)$ and $n_2 + k_2d + q \in R(1)$. 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.

If such a decomposition exists for $p > 1$, then if $i \in A_n$, a chain starts from i cannot return to A_n in less than p -steps. Since the chain is irreducible, it must have positive probability to return to it, so p is a divisor of d and the largest d with this decomposition is the period of the chain. Note that for this d , P^d must be reducible. \square

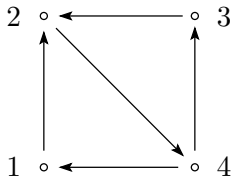


Figure 7.1: Periodic.

The example given in (7.1.2) is aperiodic. However the example shown in Figure 7.1 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 7.1 are of the form

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

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.

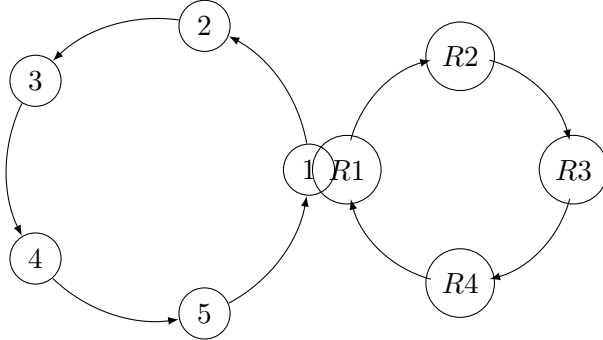
Example 7.7.6 *Simple Random Walk on \mathbf{Z} .* Recall Example 7.4.3. The chain is periodic of period 2, with state space decomposed as

$$\mathbf{Z} = \{2n\} \cup \{2n + 1\}.$$

If we let $y_n = x_{2n}$ (where $x_k = \sum_{i=1}^k \xi_i$, with ξ_i be i.i.d. such that $\mathbb{P}(\xi_i = \pm 1) = 1/2$). Then y_n is a THMC on even integers $\{2n\}$ with

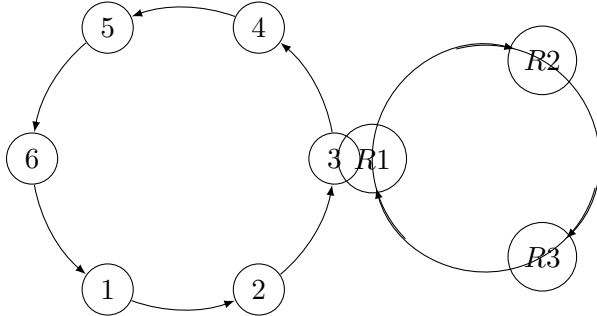
$$\mathbb{P}(y_n = k | \mathcal{F}_{n-1}^y) = \mathbb{P}(x_{2n} = k | x_0, \dots, x_{2n-2}) = \mathbb{P}(x_{2n} = k | x_{2n-2}) = P^2(y_{n-1}, k).$$

Example 7.7.7 Consider a Markov chain on the the states marked below



The chain (with $1 = R1, 6 = R2, 7 = R3, 8 = R4$) is aperiodic with decomposition $\{3n, 5n\}$.

Example 7.7.8 The chain with incidence graph below is periodic with period $d = 3$.



Where $3 = R1, 7 = R2, 8 = R3$. We can decompose state space into

$$A_0 = \{1, 4, R2 = 7\}, \quad A_1 = \{2, 5, R3 = 8\}, \quad A_2 = \{3, 6\}.$$

The process $y_n = x_{3n}$ is a Markov chain on each A_i with restricted t.p. from P^3 .

The THMC on each cycle is irreducible. Suppose it has an invariant measure when restricted to A_0 , could we use it to construct an invariant measure for P ?

7.7.6 Invariant measure for periodic chains

Let us recall the transformation T on measures, which in this setting is given by $T\mu := \mu P$.

Proposition 7.7.22 *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

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

□

Remark 7.7.23 If we have a periodic chain with period d , then $\mathcal{X} = A_0 \cup \dots \cup A_{d-1}$. If there exists an invariant measure μ for the chain on A_0 , then $\hat{\mu} = \frac{1}{d} \sum_{k=1}^d \mu P^k$ is an invariant measure for P .

Exercise 7.7.5 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 .

Define the partition $\{B_i\}$ of $\{1, \dots, N\}$ given by $B_i = \bigcup_{n \geq 0} A_{i+nq \pmod{d}}$, where $\{A_i\}$ is the partition associated to P by Lemma 7.7.21. Then the THMC with t.p. P^n always jumps from B_i to B_{i+1} in one step.

Examples: $\mathcal{X} = \{1, \dots, 6\}$ and $P_{i(i+1)} = 1$. Then the period of P is 6. Note that the chain (y_n) with t.p. P^3 has communication classes $\{1, 4\}$, $\{2, 5\}$, and $\{3, 6\}$. Now $r = 3$, $q = 6/3 = 2$, $B_1 = \{1, 3, 5\}$ and $B_2 = \{2, 4, 6\}$. Note that the chain y_n always jumps from B_1 to B_2 , and also from B_2 to B_1 .

Exercise 7.7.6 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 $P_{ij} = 0$. Show that q must be a divider of d and that the partition $\{B_j\}$ is the one associated by Lemma 7.7.21 to the matrix $P^{d/q}$.

7.8 Ergodic Theorem: The law of Large Numbers

Let us first recall Kolmogorov's strong law of large numbers.

Theorem (Strong Law of Large Numbers). *Let $(\xi_n)_{n \geq 1}$ be a family of real-valued independent and identically distributed random variables. Suppose $\mathbf{E}|\xi_i| < \infty$ and $a = \mathbf{E}\xi_i$, then*

$$\frac{1}{n} \sum_{k=1}^n \xi_k \rightarrow a \quad \text{a.e.}$$

Its simplest extension to Markov processes states:

Theorem 7.8.1 *Let x be an irreducible positive recurrent THMC on $\mathcal{X} = \mathbf{N}$. Let π denote its invariant probability measure. Then for any $f: \mathcal{X} \rightarrow \mathbf{R}$ integrable,*

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n f(x_k) = \sum_{j \in \mathcal{X}} f(j) \pi(j) = \int_{\mathcal{X}} f d\pi, \quad \text{a.e.} \quad (7.20)$$

Remark 7.8.2 If $f = \mathbf{1}_i$ for some state $i \in \mathbf{N}$, and $\mu = \mathcal{L}(x_0)$, then (7.20) becomes

$$\frac{1}{n} \sum_{k=1}^n \mathbf{1}_i(x_k) \xrightarrow{(n \rightarrow \infty)} \pi(i).$$

In the case the chain is aperiodic, we have already proven (Theorem 7.7.19) that $P^n \mu \rightarrow \pi$ for any probability measure μ , in particular $\lim_{n \rightarrow \infty} \mathbf{E}_\mu(\mathbf{1}_i(x_n)) = \lim_{n \rightarrow \infty} \mathbb{P}_\mu(x_n = i) = \pi(i)$ and therefore by the dominated convergence theorem

$$\lim_{n \rightarrow \infty} \mathbf{E}_\mu \left(\frac{1}{n} \sum_{k=1}^n \mathbf{1}_i(x_k) \right) = \lim_{n \rightarrow \infty} \left(\frac{1}{n} \sum_{k=1}^n \mathbf{E}_\mu \mathbf{1}_i(x_k) \right) = \pi(i). \quad (7.21)$$

Remark 7.8.3 It is often asked whether aperiodicity is needed in Theorem 7.8.1 is needed. It is not needed. The proof does not use aperiodicity. In fact, assuming that Theorem 7.8.1 holds for aperiodic irreducible positive recurrent chains, we now deduce it for periodic chains. Indeed, suppose it is periodic of period d , we may take $n = md$ terms where d is the period (if $n = md + k$, $k < d$ one can take an approximation). Then P^d is aperiodic, irreducible and positive recurrent on each cycle. In this $n = md$ case,

$$\frac{1}{n} \sum_{k=1}^n f(x_k) = \frac{1}{d} \sum_{\ell=0}^{d-1} \frac{1}{m} \sum_{k=1}^m f(x_{kd+\ell}).$$

Let $y_k^0 = x_{kd}$. Then

$$\frac{1}{m} \sum_{k=1}^m f(x_{kd}) \rightarrow \int f(x) d\mu_0(x)$$

where $\mu_0(x)$ is the invariant measure on the cycle containing x . Now $y_k^\ell = x_{kd+\ell}$ is the chain with initial condition x_ℓ on the ℓ -th cycle,

$$\frac{1}{m} \sum_{k=1}^{md} f(x_{kd+\ell}) \rightarrow \int f(y) d\mu_\ell(x)$$

where μ_ℓ is the invariant measure of P^d in the ℓ the cycle. In fact

$$(\mu_0 P^\ell) P^d = (\mu_0 P^d) P^\ell = \mu_0 P^\ell,$$

$\mu_\ell = \mu_0 P^\ell$ where $\ell = 0, 1, \dots, d-1$. Since $\mu = \frac{1}{d} \sum_{\ell=0}^{d-1} \mu_0 P^\ell$, .c.f Theorem 7.7.22 this then proves the statement of the law of large numbers.

Proof. Let $i \in \mathcal{X}$ be a distinguished state, let $T := T_i$ and T^k the successive return times to i , as in (7.4) in Section 7.3. Then it can be shown that

$$\left\{ \sum_{l=T^k+1}^{T^{k+1}} f(x_l), \quad k = 1, 2, \dots \right\} \quad \text{are iid's.}$$

Let $f \geq 0$. Note first that

$$\mathbf{E} \left[\sum_{l=T^k+1}^{T^{k+1}} \mathbf{1}_{x_l=j} \right] = \mathbf{E}_i \left[\sum_{l=1}^T \mathbf{1}_{x_l=j} \right] = \mu(j) = \pi(j) \mathbf{E}_i(T), \quad (7.22)$$

where μ is the same as the one defined in (7.9) in Section 7.6. Moreover, we used uniqueness (up to multiplication constant) of Theorem 7.6.4 and the fact that $\sum_{j \in \mathcal{X}} \mu(j) = \mathbf{E}_i(T)$, recalling (7.11). Given (7.22), we then deduce that

$$\begin{aligned} \mathbf{E}_i \left[\sum_{l=T+1}^{T^2} f(x_l) \right] &= \mathbf{E} \left[\sum_{l=T+1}^{T^2} \sum_{j \in \mathcal{X}} f(j) \mathbf{1}_{x_l=j} \right] = \mathbf{E} \left[\sum_{j \in \mathcal{X}} f(j) \sum_{l=T+1}^{T^2} \mathbf{1}_{x_l=j} \right] \\ &= \sum_{j \in \mathcal{X}} f(j) \pi(j) \mathbf{E}_i(T) = \int_{\mathcal{X}} f d\pi \cdot \mathbf{E}_i(T) < \infty. \end{aligned}$$

By Kolmogorov's Strong LLN

$$\frac{1}{n} \sum_{l=0}^{T^n} f(x_l) \rightarrow \int_{\mathcal{X}} f d\pi \cdot \mathbf{E}_i(T) \quad a.s., \quad (7.23)$$

and

$$\lim_{n \rightarrow \infty} \frac{T^n}{n} = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n (T^{k+1} - T^k) = \mathbf{E}_i(T^1 - T^0) = \mathbf{E}_i(T), \quad (7.24)$$

where we used Lemma 7.3.2. Now let us consider $\eta(n) = \eta_i(n) := \sum_{k=1}^n \mathbf{1}_{\{x_k=i\}}$, then

$$T^{\eta(n)} \leq n < T^{\eta(n)+1}.$$

This means that for $f \geq 0$,

$$\frac{1}{\eta(n)} \sum_{l=0}^{T^{\eta(n)}} f(x_l) \leq \frac{1}{\eta(n)} \sum_{l=0}^n f(x_l) \leq \frac{1}{\eta(n)} \sum_{l=0}^{T^{\eta(n)+1}} f(x_l) \quad a.s., \quad (7.25)$$

Since i is recurrent we have $\eta(n) = \sum_{k=1}^n \mathbf{1}_{\{x_k=i\}} \rightarrow \infty$ as $n \rightarrow \infty$ (since state i is visited i.o. almost surely, see Theorem 7.4.5), and by (7.23), both the left and the right hand term converge to the same limit which leads to

$$\lim_{n \rightarrow \infty} \frac{1}{\eta(n)} \sum_{l=0}^n f(x_l) = \int_{\mathcal{X}} f d\pi \cdot \mathbf{E}_i(T).$$

Take $f \equiv 1$, we see

$$\lim_{n \rightarrow \infty} \frac{n}{\eta(n)} = \mathbf{E}_i(T). \quad (7.26)$$

Finally,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{l=0}^n f(x_l) = \lim_{n \rightarrow \infty} \frac{\eta(n)}{n} \frac{1}{\eta(n)} \sum_{l=0}^n f(x_l) = \int_{\mathcal{X}} f d\pi.$$

This holds for any $x_0 = i$ and $f \geq 0$ (The statement holds for each initial point implies it holds for every initial distribution. Take this to the canonical space, since $\mathbb{P}_\mu = \int_{\mathcal{X}} \mathbb{P}_x d\mu$. If the limit holds almost surely under each \mathbb{P}_x , it holds almost surely under \mathbb{P}_μ . For the countable state space, \mathbb{P}_μ is a convex combination of \mathbb{P}_x 's. In other words, $\mathbb{P}_\mu(\lim_{n \rightarrow \infty} \Phi(x) = a) = \int_{\mathcal{X}} \mathbb{P}_x(\lim_{n \rightarrow \infty} \Phi(x) = a) d\mu(x) = 0$. In fact, $\mathbb{P}_\mu = \sum_i \mu(i) \mathbb{P}_i$.) Apply this to $f^+, f^- \geq 0$ for general $f = f^+ - f^-$ to conclude the proof. \square

Remark 7.8.4 By (7.24), $\lim_{n \rightarrow \infty} \frac{T^{\eta(n)}}{\eta(n)} = E_i(T)$, together with (7.26), we see in fact

$$\lim_{n \rightarrow \infty} \frac{T^{\eta(n)}}{n} = 1.$$

Remark 7.8.5 (*Average time spent*) Take $f \equiv \mathbf{1}_{\{j\}}$ in (7.23), we have

$$\frac{1}{n} \sum_{l=1}^{T_i^n} \mathbf{1}_{\{j\}}(x_l) \rightarrow \frac{\pi(j)}{\pi(i)}. \quad (7.27)$$

The ratio $\frac{\pi(j)}{\pi(i)}$ is the average time spent at site j during one excursion.

Alternative Proof for Thm 7.8.1:

Proof. ** Since any function on \mathcal{X} can be written as a finite linear combination of functions $\mathbf{1}_i \stackrel{\text{def}}{=} \mathbf{1}_{\{i\}}$, it suffices to consider Theorem 7.8.1 with $f = \mathbf{1}_i$, so that (7.20) becomes:

$$\lim_{m \rightarrow \infty} \frac{1}{m} \sum_{n=1}^m \mathbf{1}_i(x_n) = \pi(i). \quad (7.28)$$

In order to get (7.20), we should 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 7.3.2) and distributed as T , the first return time to i , we apply to it the law of large numbers,

$$\lim_{n \rightarrow \infty} \frac{T_i^n}{n} = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n (T_i^{k+1} - T_i^k) = \frac{1}{n} (\mathbf{E}_i(T_i^1 - T_i^0)) = \mathbf{E}T, \quad (7.29)$$

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,

$$\lim_{n \rightarrow \infty} \mathbf{E} \left| \frac{n \mathbf{E} T}{T_i^n} - 1 \right| = 0. \quad (7.30)$$

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

$$\frac{n}{T_i^n} = \frac{1}{T_i^n} \sum_{k=0}^{T_i^n} \mathbf{1}_i(x_k). \quad (7.31)$$

We can rewrite this as

$$\frac{n}{T_i^n} = \frac{1}{n \mathbf{E} T} \sum_{k=1}^{n \mathbf{E} T} \mathbf{1}_i(x_k) + R_n, \quad (7.32)$$

where the error term $R_n \rightarrow 0$. Taking expectation of the right hand side of (7.32), taking $n \rightarrow \infty$ and use the LLN in averaged form (7.21), one has

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

and so by (7.29), $\frac{1}{\mathbf{E} T} = \pi(i)$, proving part two of the assertion.

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

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

almost surely and in L_1 .

To show $\frac{1}{m} \sum_{n=1}^m \mathbf{1}_i(x_n)$ converges, we return to (7.32) and take $n \rightarrow \infty$ using the fact that $\frac{n \mathbf{E} T}{T_i^n} \rightarrow 1$ and $R_n \rightarrow 0$ almost surely, with a bit analysis we obtain the required LLN. \square

Example 7.8.1 (Empirical Averages) Let (x_n) be a THMC on $\mathcal{X} = \mathbf{N}_+$ with transition probability P . Define $y_n = (x_n, x_{n+1})$, then y_n is a THMC on \mathcal{X}^2 . Then

$$\mathbb{P}(y_{n+1} = (i, i') | y_0 = (i_0, i'_0), \dots, y_n = (i_n, i'_n), i'_j = i_{j+i}) = P_{ii'} \delta_{i'_n, i}.$$

Let

$$Q_{(i, i'), (j, j')} = \begin{cases} P_{jj'} \delta_{i'j}, & \text{if } P_{ii'} > 0, \\ 0, & \text{otherwise.} \end{cases}$$

We can restrict y_n to the subspace $\mathcal{Y} = \{(i, i'), P_{ii'} > 0\}$.

Exercise: If (x_n) is irreducible and recurrent, so is (y_n) on \mathcal{Y} .

It is useful to observe that

$$\begin{aligned} Q_{(i,i'),(j,j')}^2 &= Q_{(i,i'),(i',j)} Q_{(i',j),(j,j')} = P_{i'j} P_{jj'} \\ Q_{(i,i'),(j,j')}^3 &= \sum_{k=1}^{\infty} Q_{(i,i'),(i',k)} Q_{(i',k),(j,j')}^2 = \sum_{k=1}^{\infty} P_{i'k} P_{kj} P_{jj'} \\ &= P_{i'j}^2 P_{jj'}. \end{aligned}$$

Suppose (x_n) is irreducible with invariant probability measure π . Set $\tilde{\pi}((i, i')) = \pi(i)P_{ii'}$. Then $\tilde{\pi}$ is invariant for (y_n) and (y_n) is positive recurrent. Let $\varphi : \mathcal{X}^2 \rightarrow \mathbf{R}$ be integrable w.r.t. $\tilde{\pi}$, then

$$\frac{1}{m} \sum_{n=0}^m \varphi(x_n, x_{n+1}) \rightarrow \int_{\mathcal{X}^2} \varphi d\tilde{\pi} = \sum_{i,i' \in \mathcal{X}} \varphi(i, i') \pi(i) P_{ii'}.$$

Take e.g. $\varphi(x, y) = \frac{1}{2}$.

Remark 7.8.6 Later when we learnt Birkhoff's ergodic theorem on the path space, we can work with $\frac{1}{m} \sum_{n=0}^m \varphi(x_n, x_{n+1})$ directly.

Application. (*MCMC - Markov Chain Monte Carlo*) Let π a probability measure, we would like to compute

$$\int_{\mathcal{X}} f d\pi.$$

This is approximated by the construction of a THMC (x_n) with invariant measure π . Then

$$\frac{1}{n} \sum_{k=1}^n f(x_k) \approx \int_{\mathcal{X}} f d\pi.$$

The transition probability $P = (P_{ij})$ is to be determined (i.e. devise appropriate algorithm). For example choose P with

$$\pi(i)P_{ij} = \pi(j)P_{ji}.$$

7.9 Reversible Markov Chains

Suppose that 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

$$\mathbb{P}(x_m = i) = \mathbb{P}(\hat{x}_0 = i)$$

$$\mathbb{P}(x_m = i, x_0 = j) = \mathbb{P}(x_0 = i, x_m = j).$$

Summing over j , we have $\mathbb{P}(x_m = i) = \mathbb{P}(x_0 = i)$, and also we see \hat{x}_0 is distributed as π .

We define \hat{P} by

$$\pi(i)P_{ij} = \pi(j)\hat{P}_{ji}.$$

Since $\pi(j) \neq 0$, this is well defined and since $\pi = \pi P$, $\sum_{i \in \mathcal{X}} \hat{P}_{ji} = \sum_{i \in \mathcal{X}} \frac{\pi(i)P_{ij}}{\pi(j)} = 1$. Assume $x_0 \sim \pi$, then

$$\mathbb{P}(x_n = j | x_{n+1} = i) = \frac{\mathbb{P}(x_{n+1} = i | x_n = j) \mathbb{P}(x_n = j)}{\mathbb{P}(x_{n+1} = i)} = \frac{\pi(j)}{\pi(i)} P_{ji} = \hat{P}_{ij}.$$

So (x_n) with time reversed is a THMC (\hat{x}_n) with t.p. \hat{P} .

Theorem 7.9.1 *Suppose that (x_n) is an irreducible time positive recurrent homogeneous Markov chain with stochastic matrix P and with initial distribution the invariant probability measure π . 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*

$$\hat{P}_{ji} = P_{ij} \frac{\pi(i)}{\pi(j)}.$$

Proof. Note that $\pi(i) > 0$ for every i and $\hat{x}_0 = x_M$ is distributed as π . For \hat{x} to be a Markov process with stochastic matrix \hat{P} and initial distribution π it is sufficient to compute its distribution,

$$\begin{aligned} \mathbb{P}(\hat{x}_0 = i_0, \dots, \hat{x}_n = i_n) &= \mathbb{P}(x_M = i_0, \dots, x_{M-n} = i_n) \\ &= \pi(i_n) P_{i_n i_{n-1}} \dots P_{i_1 i_0} \\ &= \frac{\pi(i_n)}{\pi(i_{n-1})} P_{i_n i_{n-1}} \dots \frac{\pi(i_1)}{\pi(i_0)} P_{i_1 i_0} \pi(i_0) \\ &= \hat{P}_{i_{n-1}, i_n} \dots \hat{P}_{i_0, i_1} \pi(i_0). \end{aligned}$$

In view of Theorem 5.4.7, (\hat{x}_n) is a THMC. □

Theorem 7.9.2 *If $\pi(i)P_{ij} = \pi(j)P_{ji}$ 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 7.9.3 • The relation

$$\pi(i)P_{ij} = \pi(j)P_{ji}, \quad \forall i, j \tag{7.33}$$

is called detailed balance.

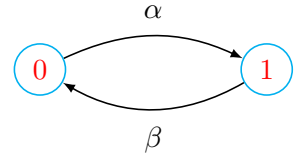
- A Markov chain is said to be reversible if the new Markov chain (x_{m-n}) is again a time homogeneous Markov chain with stochastic matrix P and with initial distribution π .

Summing over j in (7.33), 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

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

Example 7.9.1 Let us consider a Markov chain on two states $\{1, 2\}$ with $P = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}$.



Then $\pi(1)P_{12} = \pi(2)P_{21}$ means $\alpha\pi(1) = \beta\pi(2)$. So π is proportional to (β, α) .

Exercise 7.9.1 Let us define $\langle f, g \rangle_\pi = \int f g d\pi = \sum_i f(i)g(i)\pi(i)$. Then P is reversible w.r.t. π if and only if

$$\langle Pf, g \rangle = \langle f, Pg \rangle.$$

7.9.1 Application : Numerical Simulation

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 combinatorial 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:

$$\sum_{i \in \mathcal{X}} f(i) \pi(i) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n f(x_k).$$

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 Q we may define

$$P_{ij} = Q_{ij} \wedge \frac{\pi(j)}{\pi(i)} Q_{ji}, \quad i \neq j,$$

$$P_{ii} = 1 - \sum_{j \neq i} P_{ij}$$

Then, $\pi(i)P_{ij} = \pi(i)Q_{ij} \wedge \pi(j)Q_{ji} = \pi(j)P_{ji}$, so that P is reversible w.r.t. π . We write

$$P_{ij} = Q_{ij} - \alpha_{ij},$$

where $\alpha_{ij} = \min\left(1, \frac{\pi(j)}{\pi(i)} \frac{Q_{ji}}{Q_{ij}}\right)$ is called the acceptance probability (for accepting the state j proposed by the matrix Q).

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 Q on a non-oriented graph. Then there are two standard choices for Q , they are known respectively as the Metropolis algorithm and the Gibbs sampler.

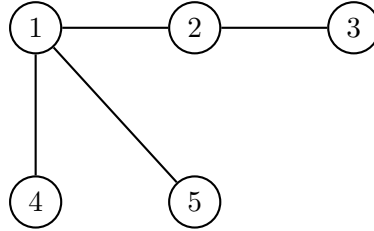
Observe the difference of MCMC versus MC is that we may start from any initial distribution, when time runs its course we will arrive approximately the invariant probability distribution, while Monte Carlo method uses the invariant probability distribution as the initial distribution.

7.9.2 Examples

Example 7.9.2 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:

$$P_{ij} = \frac{w(i, j)}{w(i)}, \quad w(i) = \sum_{j \sim i} w(i, j).$$



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

$$P_{ij} = \begin{cases} \frac{1}{d(i)}, & \text{if } i \text{ and } j \text{ are connected by an edge,} \\ 0, & \text{otherwise.} \end{cases}$$

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.)

7.10 Finite State Markov Chain

Let us now consider finite state space $\mathcal{X} = \{1, 2, \dots, N\}$. In view of previous chapter, \mathcal{X} is the union of disjoint communication classes. The minimal/closed classes consist of recurrent states, the non-minimal ones consist of transient states.

7.10.1 Characterising aperiodic irreducible chains

This follows from Lemma 7.7.14 and similarly the following proposition.

Proposition 7.10.1 *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 $\delta_{n_0} > 0$.
 (Incidentally $\delta_n \geq \delta_{n_0}$ 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) Suppose there is n_0 such that $\min_{i,j=1,\dots,N} (P^{n_0})_{ij} = \delta_{n_0} > 0$. Hence 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 $j, k \leq N$,

$$P_{jk}^{n_0+m} = \sum_{l=1}^N P_{jl}^m P_{lk}^{n_0} \geq \delta_{n_0} \sum_{l=1}^N P_{jl}^m = \delta_{n_0}, \quad \forall m.$$

Similarly we see that $\{\delta_n\}_{n \geq 1}$ is increasing, since for all $j, k \leq N$

$$P_{jk}^{n+1} = \sum_{l=1}^N P_{jl} P_{lk}^n \geq \delta_n \sum_{l=1}^N P_{jl} = \delta_n.$$

Hence $\delta_{n+1} \geq \delta_n$, thus inductively we see that for all $n \geq n_0$, P^n has all entries strictly positive. Therefore $n_0, n_0 + 1, \dots \in R(i)$ and $d(i) = 1$, for any i . Hence P is aperiodic and P^n is irreducible for $n \geq n_0$. We note that for $n < n_0$, $(P^{kn})_{ij} > 0$ for some k sufficiently large, then P^n is irreducible. Hence P^n is irreducible for every $n \geq 0$. Thus (c) implies (b).

(a) \implies (c): Suppose P is irreducible and aperiodic. By the number theory Lemma 7.7.14, 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 $N_0 = \max_{i \in \mathcal{X}} \{k_i\}$. Since P is aperiodic, it implies that $d(i) = 1$ for all i , and therefore

$$P_{ii}^n > 0, \quad \forall i \in \mathcal{X}, \forall n \geq N_0.$$

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

$$P_{ij}^{n+m(i, j)} \geq P_{ii}^n P_{ij}^{m(i, j)} > 0.$$

Let $M = \max_{1 \leq j, i \leq N} m(i, j) < \infty$, and define

$$n_0 = N_0 + M.$$

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

$$P_{ij}^n = \sum_{k=1}^N P_{ik}^{m(i, j)} P_{kj}^{n-m(i, j)} \geq P_{ij}^{m(i, j)} P_{jj}^{n-m(i, j)} > 0.$$

Where for the last inequality we used that $P_{ij}^{m(i,j)} > 0$ and that $n - m(i, j) \geq N_0 \geq k_i$ and thus $P_{jj}^{n-m(i,j)} > 0$. Taking minimum over all $i, j \in \mathcal{X}$ in the above inequality gives us $\delta_n \geq \delta_{n_0} > 0$. \square

Lemma 7.10.2 *Let (x_n) be an aperiodic and irreducible THMC on a finite state space with t.p. P . Then for any two states $i, j \in \mathcal{X}$,*

$$\mathbf{E}_j[(T_i)^\alpha] < \infty \quad \text{for any } \alpha \geq 1.$$

Proof. Let $i, j \in \mathcal{X}$, then $\mathbb{P}_j(T_i < \infty) = 1$.

$$\mathbf{E}_j[(T_i)^\alpha] = \sum_{n \geq 0} n^\alpha \mathbb{P}_j(T_i = n) \leq \sum_{n \geq 0} n^\alpha \mathbb{P}_j(T_i > n - 1). \quad (7.34)$$

Let n_0 be the number such that $\delta_{n_0} = \min_{i,j \in \mathcal{X}} (P_{ij}^{n_0}) > 0$ (see Proposition 7.10.1), then

$$\begin{aligned} \mathbb{P}(x_{n_0(k+1)} \neq i | x_{n_0k} \neq i) &\leq \sum_{l \neq i} \mathbb{P}(x_{n_0(k+1)} \neq i | x_{n_0k} = l) \frac{\mathbb{P}(x_{n_0k} = l)}{\mathbb{P}(x_{n_0k} \neq i)} \\ &\leq \sum_{l \neq i} (1 - \delta_{n_0}) \frac{\mathbb{P}(x_{n_0k} = l)}{\mathbb{P}(x_{n_0k} \neq i)} \leq 1 - \delta_{n_0} \end{aligned}$$

Hence

$$\begin{aligned} \mathbb{P}_j(T_i > n_0(k+1)) &\leq \mathbb{P}_j(x_{n_0(k+1)} \neq i, \dots, x_{2n_0} \neq i, x_{n_0} \neq i) \\ &= \mathbb{P}(x_{n_0(k+1)} \neq i | x_{n_0k} \neq i) \mathbb{P}_j(x_{n_0k} \neq i, \dots, x_{n_0} \neq i) \\ &\leq (1 - \delta_{n_0}) \mathbb{P}_j(T_i > n_0k) \leq \dots \leq (1 - \delta_{n_0})^{k+1}. \end{aligned}$$

For any n , there exists some k such that $n \in [n_0k, n_0(k+1)]$. Then

$$\mathbb{P}_j(T_i > n - 1) \leq \mathbb{P}_j(T_i > n_0k - 1) \leq (1 - \delta_{n_0})^k.$$

Hence, from (7.34), we can estimate

$$\begin{aligned} \mathbf{E}_j[(T_i)^\alpha] &\leq \sum_{n \geq 0} n^\alpha (1 - \delta_{n_0})^k \leq \sum_{k=0}^{\infty} [n_0(k+1)]^\alpha (1 - \delta_{n_0})^k \\ &= n_0^\alpha \sum_{k=1}^{\infty} k^\alpha (1 - \delta_{n_0})^{k-1} < \infty, \end{aligned}$$

hence T_i has moments of all order. \square

By Proposition 7.10.1, we have:

Lemma 7.10.3 *Let $\mathcal{X} = \{1, \dots, N\}$. If P is irreducible and aperiodic, show that there exists $n > 0$ and $\delta > 0$ such that $\eta P^n \geq \delta \|\eta\|_1 \mathbf{1}$ for every vector $\eta \in \mathbf{R}_+^N$ with entries $\eta_i \geq 0$. Here $\mathbf{1}$ is the row vector with every entry being 1 and $\|\eta\|_1 = \sum_i |\eta_i|$.*

Proof. Simply note that $\eta P^n(j) = \sum_{i \in \mathcal{X}} \eta(i) P_{ij}^n \geq \min_{i,j \in \mathcal{X}} P_{ij}^n \sum_{i \in \mathcal{X}} \eta(i)$. \square

Lemma 7.10.4 *Suppose that P is an irreducible stochastic matrix. Let $T^n = \frac{1}{n}(P + P^2 + \dots + P^n)$. Then there exists a number n_0 s.t. T^n has positive entries. There exists $\delta > 0$ such that*

$$\min_{i,j=1,\dots,N} T_{ij}^n \geq \delta.$$

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

Proof. This is Proposition 7.10.1 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 $j \in A_{k+1}$, $P_{i_0 j} > 0$ for some $i_0 \in A_k$. Thus $P_{ij}^{n_0 d+1} \geq P_{i_0 j}^{n_0 d} P_{i_0 j} > 0$. This shows that $P_{ij}^{n_0 d} + P_{ij}^{n_0 d+1} > 0$ for $i, j \in A_k \cup A_{k+1}$. By induction, this proves $\min_{i,j=1,\dots,N} T_{ij}^n \geq \delta$. (The final part follows again from $\eta(i) = \sum_{j=1}^N T_{ij}^n \eta(j) \geq \delta \sum_{j=1}^N \eta(j) = \delta \|\eta\|_1$.) \square

7.10.2 Perron-Frobenius Theorem

In this section we will focus on the Perron-Frobenius theorem. First, let us recall some notions on square matrices of finite size.

Spectrum of a stochastic matrix. The spectrum of a matrix is the set of its eigenvalues (and information on their eigenvectors if anything can be deduced). In particular, for a given stochastic matrix $P = (P_{ij})_{1 \leq i,j \leq N}$, we are interested in finding (non-zero) row vector $\pi \in \mathbf{R}^N$ that satisfies equation

$$\pi P = \pi. \tag{7.35}$$

Then π is a left eigenvector of P with eigenvalue 1. Equivalently $P^T \pi^T = \pi^T$, and π^T is an eigenvector of P^T with corresponding eigenvalue 1. Let us note that under transposition the determinant is preserved, i.e.

$$|P^T - \lambda I| = |P - \lambda I|.$$

Then P^T and P have the same eigenvalues. Let $\mathbf{1} = (1, \dots, 1)$, then $(P\mathbf{1}^T)_k = \sum_{j=1}^N P_{kj} = 1$, which implies that $P\mathbf{1} = \mathbf{1}$. Hence 1 is an eigenvalue also for P^T , thus (7.35) has a non trivial solution.

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

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

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

$$\mu_- = \left((-\mu(1)) \vee 0, \dots, (-\mu(N)) \vee 0 \right).$$

By $\mu \geq 0$, we mean $\mu_- = 0$. Then we have the following

Lemma 7.10.5 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: $\|\mu_1 + \mu_2\|_1 = \|\mu_1\|_1 + \|\mu_2\|_1$.

Lemma 7.10.6 Let P be a stochastic matrix, then

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

(2) $\|\mu P\|_1 \leq \|\mu\|_1$. 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 $\mu \in \mathbf{C}^N$.

The first statement is obvious. For (2) just observe that,

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

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

Theorem 7.10.7 (Perron-Frobenius) Let P be an $N \times N$ irreducible stochastic matrix on finite state space \mathcal{X} . Then the following hold.

(A) The real number 1 is a (left) eigenvalue for P , and there exists exactly one (left) eigenvector π (up to multiplication by a constant) with $\pi P = \pi$.

Furthermore, π can be chosen such that all its entries are strictly positive and with $\sum_{i=1}^N \pi(i) = 1$.

This unique eigenvector is called the **Perron-Frobenius vector of P** .

(B) Every eigenvalue of P must satisfy $|\lambda| \leq 1$. 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

$$\mu_j(n) = \lambda_j^{-k} \pi(n), \quad \text{if } n \in A_k, \quad (7.36)$$

where π is the Perron-Frobenius vector of P and the sets A_k are the ones associated to P by Lemma 7.7.21.

Since $\|\mu P\|_1 \leq \|\mu\|_1$ for every vector $\mu \in \mathbf{C}^N$ (see Exercise 7.10.6), the eigenvalues of P must all satisfy $|\lambda| \leq 1$.

Proof. Step 1. Since $\|\mu P\|_1 \leq \|\mu\|_1$ for every vector $\mu \in \mathbf{C}^N$ (see Exercise 7.10.6), the eigenvalues of P must all satisfy $|\lambda| \leq 1$. This can be seen by the fact that $\mu P = \lambda\mu$, implies $\|\mu P\|_1 = |\lambda|\|\mu\|_1$. Then $|\lambda| \leq 1$ since

$$\|\mu P\|_1 = \sum_{i \in \mathcal{X}} \sum_{j \in \mathcal{X}} |\mu(i)P^{ij}| \leq \sum_{i \in \mathcal{X}} |\mu(i)| = \|\mu\|_1.$$

– Proof of (A).

Step 2. Since the vector $\mathbf{1} = \frac{1}{N}(1, 1, \dots, 1)$ is an eigenvector with eigenvalue 1 for P^T , there exists an eigenvector with eigenvalue 1 for P , let us call it π . Since P is real, we can choose π to be real too. Let us now prove that π can be chosen positive as well.

Step 3. Define the matrix $T^n = \frac{1}{n}(P + P^2 + \dots + P^n)$. Clearly T^n is again a stochastic matrix and π is an eigenvector of T^n with eigenvalue 1, i.e. $\pi T^n = \pi$. If either $\|\pi_+\|_1 = 0$ or $\|\pi_-\|_1 = 0$, the proof is complete.

Otherwise define $\alpha = \min\{\|\pi_+\|_1, \|\pi_-\|_1\}$. Since P is irreducible, by Lemma 7.10.4 there exists n such that T^n has strictly positive entries. This implies $\exists \delta > 0$ such that $\pi_+ T^n \geq \delta \alpha \mathbf{1}$ and $\pi_- T^n \geq \delta \alpha \mathbf{1}$. Therefore,

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

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

Step 4. From now on, we normalise π in such a way that $\pi \geq 0$ and it has mass 1: $\sum_{i=1}^N \pi(i) = 1$. All entries of π are strictly positive since $\pi = \pi T^n$ and hence (again by Lemma 7.10.4)

$$\pi(i) = \pi T^n(i) = \sum_{j \in \mathcal{X}} \pi(j) T_{ji}^n \geq \delta \sum_{j \in \mathcal{X}} \pi(j) > \delta > 0.$$

Step 5. The fact that exists only one π (up to multiplication by a scalar) such that $\pi P = \pi$ is now easy. Assume that $\pi, \tilde{\pi}$ satisfy $\tilde{\pi} P = \tilde{\pi}$ and $\pi P = \pi$ with nonnegative entries and mass 1. Then the vector $\nu = \pi - \tilde{\pi}$ is also an eigenvector with eigenvalue 1 for P , i.e. $\nu P = \nu$. By the previous argument, we can assume that $\nu = \pi - \tilde{\pi} \geq 0$. But

$$0 = \sum_{i \in \mathcal{X}} \pi(i) - \sum_{i \in \mathcal{X}} \tilde{\pi}(i) = \sum_{i \in \mathcal{X}} \underbrace{(\pi(i) - \tilde{\pi}(i))}_{\geq 0}.$$

Hence $\pi(i) = \tilde{\pi}(i)$ for all i , thus uniqueness holds. This completes the proof of (A).

(The rest of the proof is not given in class and not examinable)

To part (B) and (C), we show that $\lambda_j = e^{2\pi \frac{j}{p}}$, $p = 0, \dots, p-1$, 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 $\nu = (r_1 e^{i\theta_1}, \dots, r_N e^{i\theta_N})$ one of its 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 $\mu P = \lambda \mu$, $\sum_{j=1}^N \nu_j P_{jk} = e^{i\theta} \nu_k$, then translates into

$$\sum_{j=1}^N e^{i\theta_j} r_j P_{jk} = e^{i(\theta+\theta_k)} r_k. \quad (7.37)$$

Multiplying both sides by $e^{-i(\theta+\theta_k)}$ and summing up yields $\sum_{j,k=1}^N e^{i(\theta_j-\theta_k-\theta)} r_j P_{jk} = 1$. On the other hand, we know that $r_j P_{jk} \geq 0$ and that $\sum_{j,k=1}^N r_j P_{jk} = 1$. (Indeed, let $a_{j,k}$ be the real part of $e^{i(\theta_j-\theta_k-\theta)}$, we must have $\sum_{j,k} a_{j,k} r_j P_{jk} = 1$. Since $a_{j,k} \leq 1$, $\sum_{j,k=1}^N r_j P_{jk}$ can only be one if the multiple $a_{j,k}$ before a non-zero $r_j P_{jk}$ must be 1.) This implies that

$$e^{i(\theta_j-\theta_k-\theta)} = 1, \quad \text{for every } j \text{ and } k \text{ such that } r_j P_{jk} \neq 0. \quad (7.38)$$

Combining this with (7.37) in turn implies that $r = \pi$. (then $r_j \neq 0$ for any j) Indeed for every k ,

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

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

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

$$e^{i\theta_j} = e^{i\theta n + i\theta_k}, \quad \text{for every } j \text{ and } k \text{ such that } P_{jk}^n \neq 0. \quad (7.39)$$

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 $\theta Kp = 0 \pmod{2\pi}$, $\theta(K+1)p = 0 \pmod{2\pi}$ which implies that $\theta p = 0 \pmod{2\pi}$. Thus all possible eigenvalues with $|\lambda| = 1$ are of the form $\lambda_j = e^{2\pi \frac{j}{p}}$.

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

-We find μ which satisfies $\mu P = \lambda \mu$. By multiplying μ with a scalar, we can assume that $\theta_1 = 0$. The relation (7.38) 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, $P_{1k} \neq 0$, 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 $\theta_k = \lambda^{-n \pmod{2\pi}}$ 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 (7.37) can be written as

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

Fix k , on its left hand side, the only non-zero term P_{jk} 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

$$\sum_{j \in A_{p-1}} \pi(j) P_{j1} = \pi(1),$$

this is the identity for Perron-Frobenius vector, observing that $\sum_{j \in A_{p-1}} \lambda^1 \pi(j) P_{j1} = \sum_{j=1}^N \lambda^1 \pi(j) P_{j1}$. This is true for all $k \in A_0$. The rest of the relations can be verified similarly. \square

We emphasize that an irreducible stochastic matrix P always has left eigenvalue 1, whose eigenspace is one dimensional .

Since every irreducible Markov chain on a finite state space has an invariant probability measure, (since $\pi(i) > 0$ for all i) we see that

Corollary 7.10.8 *An irreducible Markov chain on a finite state space is positive recurrent.*

This marks the end of lecture 14 (week 6) .

We remark that an application of the Perron-Frobenius theorem is to give a direct proof for the ergodic theorem (Theorem 7.7.6) for a finite state Markov chain. 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 . We give another proof below, which explores the fact that $P^n \eta \geq \delta \|\eta\|_1 \mathbf{1}$.

Exercise 7.10.1 Let P be irreducible and aperiodic and let π be its Perron-Frobenius vector. Prove that, without refereeing to the general theorem for MC on a countable state space, for any probability measure $\nu \in \mathbf{R}^N$, one has $\lim_{n \rightarrow \infty} \nu P^n = \pi$.

Proof. It follows from Lemma 7.10.3 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

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

$$\begin{aligned}
&\leq \|(\pi - \nu)_+\|_1 + \|(\pi - \nu)_-\|_1 - \delta N \|\pi - \nu\|_1 \\
&\leq (1 - \delta N) \|\pi - \nu\|_1 .
\end{aligned}$$

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 $k \rightarrow \infty$ to conclude (observe that $\|P^m(\nu - \pi)\|_1$ decreases with m .) \square

Under the conditions of the theorem, each row of P^n converges to π (just take ν to be the j th basis vector). We see that $\lim_{n \rightarrow \infty} P_{ji}^n = \pi(i)$ for every i .

Exercise 7.10.2 Show that the conclusion of Exercise 7.10.1 also hold if one only assumes that $\sum_i \nu_i = 1$.

7.10.3 The Structure Theorem for Invariant Measures

If π_i , $i = 1, \dots, k$ are invariant probability measures of P and if a_i are positive numbers with $\sum_{i=1}^k a_i = 1$, then $\sum_{i=1}^k a_i \pi_i$ is an invariant probability measure of P . The measure $\sum_{i=1}^k a_i \pi_i$ is called a convex combination of $\{\mu_1, \dots, \mu_k\}$.

Theorem 7.10.9 *Let P be an arbitrary stochastic matrix. The set of all invariant probability measures of P consists of all convex linear combinations of the Perron-Frobenius vectors of the restrictions of P to its recurrent communication classes.*

Proof. Any convex linear combinations of the Perron-Frobenius vectors of the restrictions of P to its recurrent communication classes is an invariant probability measure for P .

For the other way around, let A_0 be the collections of sites not in one of the minimal classes, and A_1, \dots, A_k denote the minimal classes. The matrix P can be written as

$$P = \begin{pmatrix} T & S_1 & S_2 & \dots & S_k \\ 0 & P_1 & 0 & \dots & 0 \\ 0 & 0 & P_2 & \dots & 0 \\ & & \dots & & \\ 0 & 0 & 0 & \dots & P_k \end{pmatrix} . \tag{7.40}$$

Let $\mu = (v_0, v_1, \dots, v_k)$ be an invariant probability measure, where v_0 is a vector corresponds to the transient states and, for $i \in \{1, \dots, k\}$, each v_i is a vector corresponding to states in the

closed communications class A_i . Then,

$$\begin{aligned}\mu P &= \begin{pmatrix} v_0 & v_1 & v_2 & \dots & v_k \end{pmatrix} \begin{pmatrix} T & S_1 & S_2 & \dots & S_k \\ 0 & P_1 & 0 & \dots & 0 \\ 0 & 0 & P_2 & \dots & 0 \\ & & \dots & & \\ 0 & 0 & 0 & \dots & P_k \end{pmatrix} \\ &= \begin{pmatrix} v_0 T, v_0 S_1 + v_1 P_1, \dots, v_0 S_k + v_k P_k \end{pmatrix}\end{aligned}$$

Since an invariant probability measure does not charge transient states, $v_0 = 0$. Consequently,

$$\begin{pmatrix} 0, v_1 P_1, \dots, v_k P_k \end{pmatrix} = (0, v_1, \dots, v_k).$$

Since $v_i P_i = v_i$, for $i = 1, \dots, k$ and each P_i is irreducible, by Perron-Frobenius theorem, v_i is a multiple of the Perron-Frobenius vectors π_i of P_i . Then $\mu = (0, \alpha_1 \pi_1, \dots, \alpha_k \pi_k) = \sum_{i=1}^k \alpha_i v_i$. Since μ is a probability measure, with $\sum \alpha_i = 1$ and $v_i > 0$, concluding the proof. \square

7.10.4 Rate of Convergence

Theorem 7.10.10 (Minorisation) *Suppose that there exist j_0 and a number $\delta > 0$ s.t. $P_{ij_0} \geq \delta$ for all i . If μ is any probability vector in \mathcal{X} , then μP^n is a Cauchy sequence. Denote by π its limit, then π is invariant. Furthermore, $\pi(j_0) \geq \delta$ and $\|\mu P^n - \pi\|_1 \leq 2(1 - \delta)^n$.*

Proof. Let η be the row column with a single non-zero entry $\eta_{j_0} = 1$. If μ is a probability measure on \mathcal{X} , $\mu P \geq \delta \eta$ following from the assumption that $P_{ij_0} \geq \delta$ for all i . Let ν be a vector with $\sum \nu(i) = 0$ then $\nu A = 0$. Since $2\|\nu\|_1 = \|\nu^+\|_1 = \|\nu^-\|_1$, then $\tilde{\nu}^+ = \frac{\nu^+}{\frac{1}{2}\|\nu\|_1}$ and $\tilde{\nu}^- = \frac{\nu^-}{\frac{1}{2}\|\nu\|_1}$ are probability measures.

$$\begin{aligned}\|\nu P\|_1 &= \frac{1}{2}\|\nu\|_1 \left\| \frac{\nu}{\frac{1}{2}\|\nu\|_1} \cdot P \right\|_1 = \frac{1}{2}\|\nu\|_1 (1 - \delta) \left\| \frac{\tilde{\nu}^+ P - \delta \eta}{1 - \delta} - \frac{\tilde{\nu}^- P - \delta \eta}{1 - \delta} \right\|_1 \\ &\leq (1 - \delta)\|\nu\|_1.\end{aligned}$$

In the last step we use triangle inequality for the L_1 norm, and that the L^1 norm for a probability measures is 1.

Since both μP^m and μ are probability vectors,

$$\sum_{i \in \mathcal{X}} (\mu P^m(i) - \mu(i)) = 0.$$

Letting $\nu = \mu P^m - \mu$ in the previous estimate,

$$\|\mu P^{n+m} - \mu P^n\| \leq \|(\mu P^m - \mu) P^n\| \leq (1 - \delta)^n \|\mu P^m - \mu\| \rightarrow 0.$$

Thus $\{\mu P^n\}$ is indeed a Cauchy sequence and has limit π which is indeed a probability vector (use positive preserving of the limit, and that the sum of entries of μP^n is always 1). If π is the limit,

$$\pi P = \lim_{n \rightarrow \infty} (\mu P^n) P = \lim_{n \rightarrow \infty} \mu P^{n+1} = \pi.$$

Similarly

$$\|\mu P^n - \pi\|_1 = \|\mu P^n - \pi P^n\|_1 \leq (1 - \delta)^n \|\mu - \pi\|_1 \leq 2(1 - \delta)^n.$$

Take $\mu = e_i$, for any j , $P_{ij}^n = e_i P^n(j) \rightarrow \pi(j)$. Since $P_{ij_0} \geq \delta$ for all i , $P_{ij_0}^2 = \sum_k P_{ik} P_{kj_0} \geq \delta \sum_{k \in \mathcal{X}} P_{ik} = \delta$. By induction $P_{ij_0}^n \geq \delta$.

$$\pi(j_0) = \sum_k \pi(k) P_{kj_0}^n \geq \delta \sum_k \pi(k) = \delta.$$

This completes the proof. \square

Sub-stochastic matrices

Definition 7.10.11 An $N \times N$ matrix P with positive entries such that $\sum_i P_{ji} \leq 1$ for all j is a **substochastic** matrix.

Substochastic matrices are typically obtained when we restrict a stochastic matrix to a subset of indices.

Example 7.10.1 Consider the stochastic matrix from Example 7.1.5, and its restrictions to the communication classes $\{2, 4, 5, 7\}$, $\{1, 3\}$ and $\{6\}$.

$$P = \begin{pmatrix} \frac{1}{3} & 0 & \frac{2}{3} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{3} & 0 & 0 & \frac{1}{3} & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad P_1 = \begin{pmatrix} \frac{1}{3} & \frac{2}{3} \\ \frac{1}{3} & 0 \end{pmatrix}, \quad P_2 = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}, \quad P_6 = (1).$$

One has the following:

Lemma 7.10.12 Let P be an irreducible substochastic matrix which is not a stochastic matrix. Then, $\mu P^n \rightarrow 0$ for every probability measure¹ μ and the convergence is exponential. More specifically there exists $\lambda > 0$ such that

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

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

¹For any positive measure this is $\|\mu\|_1 e^{-\lambda n}$, an exponential rate $e^{-\lambda' n}$ (some λ') for large n .

Proof. It is sufficient to prove the claim for μ positive with norm 1 (unless $\mu = 0$). Define $T^n = \frac{1}{n}(P + \dots + P^n)$ as in the proof of the Perron-Frobenius theorem. For a positive vector μ , one has $\|\mu P\|_1 \leq \|\mu\|_1$ and one has also

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

for every $n > 0$. Choose n_0 such that T^{n_0} has positive entries (such an n exists by the irreducibility of P). Let $\delta = \min_{ij} T_{ij}^{n_0} > 0$. Then, $\mu T^{n_0} \geq \delta \|\mu\|_1 \mathbf{1}$. Since P is not a stochastic matrix, there exists $\alpha > 0$ and an index i_0 such that $\sum_{j \in \mathcal{X}} P_{i_0 j} = 1 - \alpha$. Let $e_{i_0} = (0, \dots, 1, \dots, 0)$ denote the unit vector with entries 1 at i_0 th entry and with 0 at other entries. Therefore

$$\begin{aligned} \|\mu P^{n_0+1}\|_1 &\leq \|\mu T^{n_0} P\|_1 = \|\mu T^{n_0} (P - P\delta e_{i_0}) + \mu P\delta e_{i_0}\|_1 = \|\mu P T^{n_0} - \delta \mu P e_{i_0}\|_1 + \delta \|P e_{i_0}\|_1 \\ &\leq \|\mu T^{n_0}\|_1 - \delta \|e_{i_0}\|_1 + \delta(1 - \alpha) \\ &= \|\mu T^{n_0}\|_1 - \delta \cdot \|e_{i_0}\|_1 + \delta(1 - \alpha) \leq (1 - \alpha\delta) = (1 - \alpha\delta)\|\mu\|_1. \end{aligned}$$

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

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

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

Recall first the Borel-Cantelli lemma from probability theory:

Lemma 7.10.13 (Borel-Cantelli) *Let $\{A_n\}_{n \geq 0}$ be a sequence of events in a probability space Ω . If $\sum_n \mathbb{P}(A_n) < \infty$, then the probability that infinitely many of these events happen is 0. Equivalently this implies that $\mathbb{P}(\cap_{n=1}^\infty \cup_{k=n}^\infty A_n) = 0$.*

Exercise 7.10.3 Let $\{x_n\}$ be a Markov process with transition probabilities P and let i be from a non-minimal class. Show that without referring to the general theorem on countable state space, that the probability that $x_n \in [i]$ for an infinite number of values n is 0.

Proof. 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

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

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 7.10.12. \square

7.10.5 The long run probability for reducible chains

If the chain is reducible and aperiodic 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. We know $\lim_{n \rightarrow \infty} \mathbb{P}_\mu(x_n = i) = \pi(i)$, in particular, $\lim_{n \rightarrow \infty} P_{ji}^n = \mathbb{P}_j(x_n = i) = \pi(i)$ for every state j . Such limit can also be computed when the chain is reducible.

Let 0 stand for a sink (with $[0]$ containing only one single element 0 which is a recurrent state) and let

$$B = \{ \text{the chain eventually ends at site } 0 \}.$$

Since 0 is a sink,

$$B = \{ \omega : \text{there exists } n_0 \text{ s.t. if } n \geq n_0 \text{ } x_n(\omega) = i \}.$$

Set $f(j) = \mathbb{P}_j(B)$. This is the probability that the chain starts from j ends at 0 eventually. Then

$$\begin{aligned} f(j) &= \mathbf{E}_j(\mathbf{E}(\text{the chain eventually ends at site } 0 \mid x_1)) \\ &= \mathbf{E}_j(f(x_1)) = \sum_{k \in \mathcal{X}} f(k) \mathbb{P}(x_1 = k \mid x_0 = j) \\ &= \sum_{k \in \mathcal{X}} P_{jk} f(k) = (Pf)(j). \end{aligned}$$

Remark 7.10.14 We note that if 0 is a sink, $f(j) = \lim_{n \rightarrow \infty} P_{j0}^n$. Indeed, since the chain stays at 0 when arrived at 0, $B_n = \{ \omega : x_n(\omega) = 0 \}$ is increasing set with limit $B = \cup_n B_n$. Hence $\mathbb{P}_{j0}^n = \mathbb{P}_j(B_n) \rightarrow \mathbb{P}_j(B) = f(j)$.

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 whereabouts is not observable and of no relevance. The probability we calculated is then the probability it enters the minimal classes. The probability $\lim_{n \rightarrow \infty} P_{ji}^n$, where i is in the minimal class, is then obtained according to the unique invariant probability distribution of the reduced chain.

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. Suppose that all recurrent communication classes consist of singletons (called sinks).

In order to somewhat simplify the argument, we assume that the recurrent classes consist of *single points*, that the states $1, \dots, q$ are transient, $q+1, \dots, q+k$ are recurrent. Therefore, the

transition matrix P can be written as

$$P = \begin{pmatrix} T & S \\ 0 & \tilde{P} \end{pmatrix}, \quad (7.41)$$

where T is some sub-stochastic matrix and \tilde{P} is a stochastic matrix.

Define now the matrix A_{ij} the probability that the process starting at the transient state i will eventually end up in the recurrent state j .

Proposition 7.10.15 *The matrix A is given by $A = (I - T)^{-1}S$.*

Proof. The invertibility of $(I - T)$ is an immediate consequence of Lemma 7.10.12. One has

$$\begin{aligned} A_{ij} &= \sum_{k \in \mathcal{X}} \mathbb{P}(\text{the process reaches } j \text{ eventually}, x_1 = k \mid x_0 = i) \\ &= \sum_{k=1}^q \mathbb{P}(\text{the process reaches } j \text{ eventually} \mid x_1 = k, x_0 = i) \mathbb{P}_i(x_1 = k) + P_i(x_1 = j) \\ &= \sum_{k=1}^T A_{kj} T_{ik} + S_{ij}, \end{aligned}$$

where we used the Markov property to go from the first to the second line. In matrix notation, this reads $A = TA + S$, and therefore $A = (I - T)^{-1}S$. \square

7.10.6 Examples

Example 7.10.2 Let $\mathcal{X} = \{0, 1, 2, 3, 4, 5\}$ and let (x_n) be the time homogeneous Markov chain with stochastic matrix

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

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 $f = (f(0), f(1), \dots, f(5))^T$. We solve

$$f = Pf,$$

with the boundary conditions $f(0) = 1$, $f(5) = 0$ (starting from 5, never ends at 0). Then $f(1) = \frac{1}{3}f(0) + \frac{1}{3}f(1) + \frac{1}{3}f(2)$, i.e. $2f(1) = 1 + f(2)$. Similarly, we work out the equations for $f(2), f(3), f(4)$.

$$2f(1) = 1 + f(2)$$

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

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

We give another example whose proof using this conditioning recursive method (the statement itself has a number of (quick) proofs).

Example 7.10.3 Simple Random Walk on \mathbf{Z} . Let ξ be i.i.d. such that $\mathbb{P}(\xi = \pm 1) = 1/2$, and define $S_n = x + \sum_{i=1}^n \xi_i$, letting $x = 0$. Define $T_i = \inf\{n \geq 0, S_n = i\}$ and we use the notation $\mathbb{P}_i(\dots) = \mathbb{P}(\dots | x_0 = i)$. Show that $\mathbb{P}_0(T_1 < \infty) = 1$.

Proof. For $k \in \mathbf{Z}$, let

$$f_k := \mathbb{P}(T_1 < \infty | x_0 = k), \quad f_k \in [0, 1].$$

If $x_0 = 1$, $T_1 = 0$, so $f_1 = \mathbb{P}(T_1 < \infty | x_0 = 1) = 1$. Note that unless $x_0 = 1$:

$$T_1 = n \implies T_1 \circ \theta_1 = n - 1.$$

Firstly, we have

$$f_k = \mathbb{P}(T_1 \circ \theta_1 < \infty, x_1 = k + 1 | x_0 = k) + \mathbb{P}(T_1 \circ \theta_1 < \infty, x_1 = k - 1 | x_0 = k). \quad (7.42)$$

Then for $k \neq 1$, we have

$$\begin{aligned}
f_k &= \mathbb{P}(T_1 \circ \theta_1 < \infty | x_0 = k, x_1 = k + 1) \mathbb{P}(x_1 = k + 1 | x_0 = k) \\
&\quad + \mathbb{P}(T_1 \circ \theta_1 < \infty | x_0 = k, x_1 = k - 1) \mathbb{P}(x_1 = k - 1 | x_0 = k) \\
&= \mathbb{P}_{k+1}(T_1 < \infty) \frac{1}{2} + \mathbb{P}_{k-1}(T_1 < \infty) \frac{1}{2}.
\end{aligned}$$

This means that

$$f_k = \frac{1}{2} f_{k+1} + \frac{1}{2} f_{k-1}, \quad \text{for } k \neq 1,$$

implying that

$$f_{k+2} - f_{k+1} = f_{k+1} - f_k = \dots = f_1 - f_0 = 1 - f_0.$$

If $f_0 \neq 1$, $1 - f_1 > 0$. Then f_k eventually becomes greater than 1 for k large enough, in contradiction with its definition (7.42). Hence we have $f_1 = f_0 = 1$. (f_k eventually becomes less than 0 for $-k$ sufficiently large and we conclude $f_k = 1$ for all k).

—End of Lecture 15—

Example: Random walks on finite groups

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 7.10.16 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 7.10.4 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

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

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

Exercise 7.10.4 The Markov chain is left invariant if and only if there exists a probability measure \hat{P} on G such that $\mathbb{P}(x_{n+1} = g | x_n = h) = \hat{P}(h^{-1}g)$. The Markov chain is right invariant if and only if there exists a probability measure \hat{P} on G such that $\mathbb{P}(x_{n+1} = g | x_n = h) = \hat{P}(gh^{-1})$.

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.

7.11 A summary

By a Markov chain or a Markov process we mean a time homogeneous Markov chain with a transition matrix (THMC) unless otherwise stated.

A state i is recurrent if $P_i(T_i < \infty) = 1$, it is positive recurrent if $\mathbf{E}_i T_i < \infty$. The chain is recurrent or positive recurrent if every state has the property. The following dichotomy hold, Theorem 7.4.5,

1. j is recurrent iff $P_j(x_n = j, \text{ infinitely often }) = 1$.
2. j is transient iff $P_j(x_n = j, \text{ infinitely often }) = 0$.

If the chain is irreducible, then either $\sum_{n=1}^{\infty} P_{ij}^n = \infty$ for any i, j and every state is recurrent or $\sum_{n=1}^{\infty} P_{ij}^n < \infty$ for any i, j and every state is transient (see Lemma 7.4.4).

If the time homogeneous Markov chain on a countable state space is irreducible then all states are simultaneously recurrent and transient (Corollary 7.4.3), also all states are simultaneously positive recurrent or not (Theorems 7.6.5 and 7.6.4).

On a finite state space, the family of transition probabilities $P(x, \cdot)$ is determined by a stochastic matrix. Every time homogeneous Markov chain on a finite state space has an invariant probability measure (Theorem 7.10.7) and for an irreducible chain there exists precisely one invariant probability measure. Every recurrent state is positive recurrent (Corollary 7.10.8). See also Lemma 7.10.2 for the existence of the p th moments of the return times for aperiodic irreducible THMC's. The set of all invariant probability measures of a stochastic matrix P consists of all convex linear combinations of the Perron-Frobenius vectors of the restrictions of P to its recurrent communication classes, see Theorem 7.10.9).

On a countable state space, an invariant measure may not have finite mass. If the time homogeneous Markov chain has a recurrent state i , we can construct an invariant measure (Theorem 7.6.1). This measure has finite mass if and only if the site i is positive recurrent. If the chain is irreducible and recurrent, there is at most one invariant measure (up to a constant multiple, Theorem 7.6.4). If the chain is irreducible and positive recurrent for one site, then $\mathbf{E}_i T_i < \infty$ for all sites (and there exists a unique invariant probability measure). For an irreducible chain, the existence of an invariant probability measure is in fact equivalent to that all states are positively recurrent (Theorem 7.6.5).

Theorem 7.11.1 1. If P has a recurrent state, it has an invariant measure.

2. If π is an invariant probability measure and if $\pi(j) > 0$ then j is recurrent (Theorem 7.4.12).
3. If the chain is irreducible and recurrent, then there exists precisely one invariant measure (unique up to a multiplication constant). The invariant measure is finite if and only if the chain is positive recurrent.
4. If P is irreducible with stationary probability measure π , then $\mathbf{E}_i T_i < \infty$ for all i and

$$\pi(i) = \frac{1}{\mathbf{E}_i T_i}.$$

5. *If the chain is irreducible and has a positive recurrent state, there exists an invariant probability measure by π . Also,*

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

Chapter 8

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

Example 8.0.1 • 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.

- \mathbf{R}^n with its usual metric (The set of points with rational coordinates is a dense subset)
- \mathcal{X} a smooth complete finite dimensional Riemannian manifold M (Hausdorff and second countable).
- The Wiener space $C([0, 1]; \mathbf{R}^n)$ with its uniform distance. Also $\mathbf{C}([0, 1]; M)$ where M is as above.
- $\mathcal{L}^p(\mathbf{R}^n)$ 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

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

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

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

The subscript \star will be from time to time omitted. Note $\int f d(T\mu) = \int T f d\mu$ for any $f \in \mathcal{B}_b$ and T determines also the transition probabilities $\{P(x, \cdot)\}$.

A probability measure π is said to be **invariant** for P if $T\pi = \pi$. How do we go about finding the invariant probability measures?

Typically $P(x, \cdot)$ assigns null measure to singleton sets, the communication classes, very effective for discrete state space, is not suitable as it is. The same story with the concept of irreducibility based on communication classes. There are other coccepts of irreducibility, eg assume that $\mathbb{P}_x(T_A) > 0$ where T_A si the first return time to a set A which is not too small. The size of A can be fro example measured by an auxiliary measure. We cannot cover this material in these lectures. See the book of Myan and Tweedie.

8.1 Weak convergence

One useful construction for invariant measure is by averaging: 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. What notions of convergence should one use? For finite state space, a probability measure is identified with a vector in \mathbf{R}^n , so all notions of convergence is equivalent.

Definition 8.1.1 A sequence μ_n of probability measures on a topological space \mathcal{X} is said to **converge weakly** to a probability measure μ if

$$\lim_{n \rightarrow \infty} \int_{\mathcal{X}} \varphi(x) \mu_n(dx) = \int_{\mathcal{X}} \varphi(x) \mu(dx) , \quad (8.1)$$

for every bounded and continuous function $\varphi : \mathcal{X} \rightarrow \mathbf{R}$.

Note that the speed of the convergence in (8.1) is allowed to depend on φ .

This following lemma is behind the notion of ‘weak convergence’.

Lemma 8.1.2 [Paratharathy, page 39] Let μ, ν be measures on a metric space \mathcal{X} . If for all bounded real valued uniformly continuous function $f : \mathcal{X} \rightarrow \mathbf{R}$,

$$\int f d\mu = \int f d\nu$$

then $\mu = \nu$.

In fact, the space of probability measures $P(\mathcal{X})$ can be given a topology, called the weak topology. Recall topology defines the concept of continuity of functions and convergence.

Remark 8.1.3 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.

If d is the metric that metrizes $\mathbb{P}(\mathcal{X})$, then $\mu_n \rightarrow \mu$ weakly if and only if $d(\mu_n, \mu) \rightarrow 0$. Also, weak convergence describes the weak topology. so $\mu_n \rightarrow \mu$ means $\mu_n \rightarrow \mu$ in the weak topology.

Example 8.1.1 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 a measure μ , it must be a Dirac measure δ_x and $x_n \rightarrow x$. 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 $\delta_{x_n} \rightarrow \mu$ 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 $\varphi_{\epsilon_n}(x)$ that equals 1 on $B_{\epsilon_n/2}(x)$ and are supported in $B_{\epsilon_n}(x)$. Similarly we can take ψ_{ϵ_n} which equals 1 on $B_{\epsilon_n/2}(y)$ and supported on $B_{\epsilon_n}(y)$. 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 8.1.2 If \mathcal{X} is a discrete state space, any function $f : \mathcal{X} \rightarrow \mathbf{R}$ is continuous. Hence $\mu_n \rightarrow \mu$ weakly if and only if $\mu_n(A) \rightarrow \mu(A)$ for an subset A of \mathcal{X} .

Proposition 8.1.4 Let $\mathcal{X} = \mathbf{R}$. Let $F(x) = \mu((-\infty, x])$ and $F_n(x) = \mu_n((-\infty, x])$. Then $\mu_n \rightarrow \mu$ weakly if and only if $F_n(x) \rightarrow F(x)$ for all x such that F is continuous at x .

This follows from Portmanteau Theorem.

Example 8.1.3 If Y_n are random variables distributed as μ_n and Y is distributed as μ and $Y_n \rightarrow Y$ in probability then $\mu_n \rightarrow \mu$ weakly (i.e. Y_n converges to Y in distribution). The converse does not hold, take for example $Y = c$ a deterministic function and $\mathbb{P}(Y_n = \pm \frac{1}{2}) = \frac{1}{2}$.

8.2 Feller and Strong Feller Property

One distinct feature for state space \mathcal{X} that is not countable is that not every function $f : \mathcal{X} \rightarrow \mathbf{R}$ is continuous (or measurable).

Definition 8.2.1 We say that a homogeneous Markov process with transition operator T is **Feller** if Tf is continuous whenever f is continuous and bounded. It is **strong Feller** if Tf is continuous whenever f is measurable and bounded.

We also extend these terminology to its transition probabilities P and to the transition operators T .

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.

The statement that T 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,

$$\lim_{n \rightarrow \infty} \int f(y)P(x_n, dy) = \int f(y)P(x, dy)$$

whenever $x_n \rightarrow x$.

Example 8.2.1 Let $x_0 \in \mathcal{X}$, set $P(x, dy) = \delta_{x-x_0}$. Then $Tf(x) = \int_{\mathcal{X}} f(y)P(x, dy) = f(x - x_0)$ is Feller.

Example 8.2.2 (Not Feller) Let $P(x, A)$ be a family of transition probabilities on \mathbf{R} given below

$$P(x, \cdot) = \begin{cases} \delta_1, & \text{if } x > 0 \\ \delta_0, & \text{if } x \leq 0. \end{cases}$$

Then

$$Tf(x) = \int_{\mathbf{R}} f(y)P(x, dy) = \begin{cases} f(1) & \text{if } x > 0 \\ f(0), & \text{if } x \leq 0, \end{cases}$$

and Tf fails to be continuous at 0 for continuous functions f with $f(1) \neq f(0)$.

Example 8.2.3 Let x_n be a random walk on \mathbf{R} with $x_n = x_{n-1} + Y_n$ 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 $\mathbb{P}(Y = 1) = \frac{1}{2}$ and $\mathbb{P}(Y = -1) = \frac{1}{2}$, 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 $T_*f(x) = \frac{1}{\sqrt{2\pi}} \int_{\mathbf{R}} f(y) e^{-\frac{|y-x|^2}{2}} dy$ has the strong Feller property. Indeed this follows from properties of Gaussian densities (parabolic PDE theory) or by properties of convolutions.

8.3 Weak convergence and Prokhorov's theorem

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 mean the total mass can be almost packed into a compact set.

Lemma 8.3.1 *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 $\bigcup_{k \geq 0} \mathcal{B}(r_k, 1/n) = \mathcal{X}$ for every n . Fix $\varepsilon > 0$ and, for every integer $n > 0$, denote by N_n the smallest integer such that

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

Since $\bigcup_{k \geq 0} \mathcal{B}(r_k, 1/n) = \mathcal{X}$, the number N_n is finite for every n . Define now the set K as

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

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. \square

Definition 8.3.2 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 $\mu(K) \geq 1 - \varepsilon$ for every $\mu \in \mathcal{M}$.

By Lemma 8.3.1, 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 $\mu_{n_k} \rightarrow \mu$ weakly.

Theorem 8.3.3 (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 8.3.1 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.

Example 8.3.1 Let M be a subset of $\mathcal{P}(\mathbf{R})$. Suppose that there exists a non-decreasing function $\varphi : [0, \infty) \rightarrow [0, \infty)$ such that $\lim_{x \rightarrow \infty} \varphi(x) = \infty$ and $C = \sup_{\mu \in M} \int_{\mathcal{X}} \varphi(|x|) d\mu < \infty$, then M is tight.

Proof. Observe that

$$\begin{aligned} \mu(|x| \geq n) &= \int_{|x| \geq n} d\mu = \int_{|x| \geq n} \frac{\varphi(|x|)}{\varphi(n)} d\mu \leq \frac{1}{\varphi(n)} \int_{|x| \geq n} \varphi(|x|) d\mu \\ &\leq \frac{C}{\varphi(n)}. \end{aligned}$$

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. \square

Remark 8.3.4 (not examinable) A topological space \mathcal{X} is said to be a Polish space, if there exists a complete metric on \mathcal{X} whose metric topology agrees with the topology of \mathcal{X} . Since both ‘relative compact’ and ‘compact sets’ and therefore ‘tightness’ of a family of measures are topological concepts, Prohorov’s theorem holds if \mathcal{X} is a Polish space.

Note that $(0, 1)$ is a Polish space although it is not a complete metric space with respect to the inherited metric from \mathbf{R} . The open sets and therefore the topology of $(0, 1)$ is understood to be that induced by \mathbf{R} . Let $\varphi(x) = \tan(\pi x - \frac{\pi}{2})$, then $\varrho(x, y) = |\varphi(x) - \varphi(y)|$ defines a complete metric on $(0, 1)$. Its collection of open balls can be characterised with that by the distance function $d(x, y) = |x - y|$.

Let $\mu_n = \delta_{\frac{x}{2^n}}$ where x is a given point $(0, 1)$. Then $\{\mu_n\}$ is not tight. However μ_n , considered as measures on \mathbf{R} , converges weakly to δ_0 . It is not relatively compact as measures on $(0, 1)$: it does not have a subsequence with limit a measure on $(0, 1)$.

8.4 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 8.4.1 (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

$$\mu_N(A) = \frac{1}{N} \sum_{n=1}^N P^n(x_0, A). \quad (8.2)$$

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 $n_k \rightarrow \infty$ such that $\mu_{n_k} \rightarrow \pi$ 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 $T\pi = \pi$, by Lemma 8.1.2, we only need to 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:

$$\begin{aligned} \int_{\mathcal{X}} \varphi d(T\pi) &= \int T\varphi d\pi = \lim_{k \rightarrow \infty} \int T\varphi d\mu_{n_k} \\ &= \lim_{k \rightarrow \infty} \frac{1}{n_k} \sum_{n=1}^{n_k} \int T\varphi P^n(x_0, dy) = \lim_{k \rightarrow \infty} \frac{1}{n_k} \sum_{n=1}^{n_k} \int \varphi P^{n+1}(x_0, dy) \\ &= \lim_{k \rightarrow \infty} \int_{\mathcal{X}} \varphi \left(d\mu_{n_k} + \frac{1}{n_k} P^{n_k+1}(x_0, dy) - \frac{1}{n_k} P(x_0, dy) \right) \\ &= \int \varphi d\pi + \lim_{k \rightarrow \infty} \frac{1}{n_k} \int_{\mathcal{X}} \varphi P^{n_k+1}(x_0, dy) - \lim_{k \rightarrow \infty} \frac{1}{n_k} \int_{\mathcal{X}} \varphi P(x_0, dy) = \int \varphi d\pi. \end{aligned}$$

Since φ was also arbitrary, this in turn implies that $T\pi = \pi$, *i.e.* that π is an invariant measure for our system. \square

This marks the end of lecture 16 (week 7) content.

Krylov-Bogoliubov Theorem holds on Polish spaces.

Example 8.4.1 Consider the Markov process defined on $(0, 1)$ by the recursion relation $x_{n+1} = x_n/2$. Note also $T\varphi(x) = \mathbf{E}(\varphi(x_1)|x_0 = x) = \varphi(\frac{x}{2})$, $P(x, dy) = \delta_{\frac{x}{2}}$. Note that the law of $\{x_n\}$ is not tight. Since x_{n+1} will eventually does not charge any Borel subset of $(0, 1)$, the Markov chain does not 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} .

As an immediate consequence of Theorem 8.4.1, we have that

Corollary 8.4.2 *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. \square

Example 8.4.2 Let $\mathcal{X} = \{1, \dots, N\}$, then (by Corollary 8.4.2) any Markov chain on \mathcal{X} has an invariant probability measure

Example 8.4.3 Let $\Phi : \mathcal{X} \times \mathcal{Y} \rightarrow \mathcal{X}$ be continuous and bounded. Define the Markov chain by $x_{n+1} = \Phi(x_n, \xi_{n+1})$, with $\mu \sim \xi_k$ iid random variables and $\{x_0, \xi_k, k \geq 1\}$ independent. Then if $f \in C_b(\mathcal{X})$,

$$Tf(x) = \mathbf{E}[f(\Phi(x, \xi_{n+1}))] = \int f \circ \Phi(x, y) \mu(dy),$$

then Tf is continuous. Hence (x_n) is Feller and has an invariant probability measure. An example is $x_{n+1} = \sin(x_n + \xi_{n+1})$.

Example 8.4.4 Consider (x_n) a Markov chain on \mathbf{R}^n with initial position x_0 . Assume P (equivalently T) is Feller, then there exists an invariant probability measure if any of the following holds:

- 1) $\sup_{n \geq 0} \mathbf{E}[|x_n|^p] < \infty$ for some $p > 0$.
- 2) $\sup_{n \geq 0} \mathbf{E}[\log(|x_n| + 1)] < \infty$.

Proof. In these settings we have $P^n(x_0, \cdot) = \mathcal{L}(x_n)$, and tightness for 2) follows from below¹

$$P^n(x_0, (B_M)^c) = \mathbb{P}(|x_n| > M) \leq \sup_{n \geq 0} \frac{\mathbf{E} \log(|x_n| + 1)}{\log(M + 1)} \rightarrow 0, \quad \text{as } M \rightarrow \infty,$$

where B_M is the closed ball of radius M centred at 0. The proof for 1) is similar. \square

Example 8.4.5 (*Tightness*) Suppose $\{\xi_n\}$ are iid and independent of x_0 , with $\mathbf{E}|x_0| < \infty$ and Markov chain $x_{n+1} = \frac{1}{2}x_n + \xi_{n+1}$. Assume also $\mathbf{E}|\xi_k| = a < 1$. The chain is Feller (check as in Example 8.4.4). The following arguments shows that the probability distribution of $\{x_n\}$ is tight. For all $n \geq 1$

$$\begin{aligned} \mathbf{E}|x_{n+1}| &\leq \frac{1}{2}\mathbf{E}|x_n| + \mathbf{E}|\xi_{n+1}| \leq \frac{1}{2}\left(\frac{1}{2}\mathbf{E}|x_{n-1}| + a\right) + a \\ &= a + \frac{1}{2}a + \frac{1}{4}(|x_{n-2}| + a) \\ &\leq a + \frac{1}{2}a + \frac{1}{4}a + \dots + \frac{1}{2^{n+1}}a + \mathbf{E}|x_0| \end{aligned}$$

¹Using Markov's inequality with non-negative monotone function $u \mapsto \log(u + 1)$.

$$\leq 2a + \mathbf{E}|x_0|.$$

Hence $\sup_{n \geq 0} \mathbf{E}|x_n| < \infty$ and the system has an invariant probability measure. We remark since we only need to show $\{P^n(x_0, \cdot)\}$ is tight for some x_0 , we can simply start the chain from a fixed point.

The Lyapunov test function method allows us to use this reasoning for more general systems.

8.4.1 Lyapunov Function test

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. A Lyapunov function is allowed to take the value $+\infty$. 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 $TV(x) \leq \gamma V(x) + C$ implies that $P(x, \mathcal{X}_0) = 1$ for every x with $V(x) < \infty$.

Lemma 8.4.3 *Let P be a transition function on \mathcal{X} and let $V : \mathcal{X} \rightarrow \mathbf{R}_+ \cup \{\infty\}$ be a Borel measurable function. Suppose there exist a positive constant $\gamma \in (0, 1)$ and a constant $C > 0$ such that*

$$TV(x) \leq \gamma V(x) + C,$$

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

$$T^n V(x) \leq \gamma^n V(x) + \frac{C}{1 - \gamma}. \quad (8.3)$$

Proof. This is a simple consequence of the Chapman-Kolmogorov equations:

$$\begin{aligned} T^n V(x) &= \int_{\mathcal{X}} V(y) P^n(x, dy) = \int_{\mathcal{X}} TV(y) P^{n-1}(x, dy) = \int_{\mathcal{X}} \int_{\mathcal{X}} V(y) P(z, dy) P^{n-1}(x, dz) \\ &\leq C + \gamma \int_{\mathcal{X}} V(z) P^{n-1}(x, dz) \leq \dots \\ &\leq C + C\gamma + \dots + C\gamma^n + \gamma^n V(x) \leq \gamma^n V(x) + \frac{C}{1 - \gamma}, \end{aligned}$$

completing the proof. □

Typically, $V(x) = |x|^p$ or $V(x) = \log |x|$, etc... These allow us to control $\mathbf{E}|x_n|^p$ etc. Note the following:

- If V is bounded $\mathbf{E}V(x_n) < \infty$ provides no information on tightness of the law of $\{x_n\}$. To avoid this assume $V^{-1}([0, a]) := \{y : V(y) \leq a\}$ is compact.

- We can allow $V = +\infty$ where x_n does not visit. But V should not be $+\infty$ everywhere, i.e. $V^{-1}(\mathbf{R}_+) \neq \emptyset$.

Definition 8.4.4 Let \mathcal{X} be a complete separable metric space and let P be a transition probability on \mathcal{X} . A Borel measurable function $V: \mathcal{X} \rightarrow \mathbf{R}_+ \cup \{\infty\}$ is called a **Lyapunov function** for P if it satisfies the following conditions:

- $V^{-1}(\mathbf{R}_+) \neq \emptyset$.
- For every $a \in \mathbf{R}_+$, the set $\{y : V(y) \leq a\}$ is compact.
- There exist a positive constant $\gamma < 1$ and a constant C such that

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

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

With this definition at hand, it is now easy to prove the following results.

Theorem 8.4.5 (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,

$$T^n V(x_0) = \int_{\mathcal{X}} V P^n(x_0, dy) \leq \gamma^n V(x_0) + \frac{C}{1 - \gamma}.$$

Tchebycheff's inequality shows that

$$\begin{aligned} P^n(x_0, (K_a)^c) &= \int_{\{V(y) > a\}} P^n(x_0, dy) \leq \int_{\{V(y) > a\}} \frac{V(y)}{a} P^n(x_0, dy) \leq \frac{1}{a} T^n V(x_0) \\ &\leq \frac{1}{a} \left(\gamma^n V(x_0) + \frac{C}{1 - \gamma} \right). \end{aligned}$$

We have used Lemma 8.4.3 and the fact that $\gamma < 1$. The results follows from convergence of the right hand side, as $a \rightarrow \infty$, with rate uniform in n . (More precisely, for every $\varepsilon > 0$ we can now choose $a \geq \frac{1}{\varepsilon} \left(V(x_0) + \frac{C}{1 - \gamma} \right)$, then $P^n(x, K_a) \geq 1 - \varepsilon$ for every $n \geq 0$.) We can now use Krylov-Bogoliubov theorem to conclude. \square

The proof the previous theorem suggests that a Lyapunov function V for T allows us to deduce information on its invariant measures. E.g. if $V(x) = |x|^2$ we expect to deduce that π has second moment and the second moment bound $C/(1 - \gamma)$, where C and γ are the constants appearing in (8.3). This is indeed the case, as shown by the following proposition:

Proposition 8.4.6 *Let P be a transition probability on \mathcal{X} and let $V: \mathcal{X} \rightarrow \mathbf{R}_+$ 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:

$$\begin{aligned} \int_{\mathcal{X}} V_M(x) \pi(dx) &= \int_{\mathcal{X}} V_M(x) (T^n \pi)(dx) = \int_{\mathcal{X}} T^n V_M(x) \pi(dx) \\ &\leq \int_{\mathcal{X}} (\gamma^n V_M(x) + \frac{C}{1 - \gamma}) \pi(dx) \end{aligned}$$

We have used Jensen's inequality. Since the function on the right hand side is bounded by M , we can apply the Lebesgue dominated convergence theorem. It yields the bound

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

which holds uniformly in M , and the result follows. \square

We complete this section with a couple of inequalities which can be handy for applying Lyapunov function methods.

Lemma 8.4.7 *For any $p \geq 1$ and any $\delta > 0$ there exists a constant $K > 1$ such that*

$$|1 + x|^p \leq K|x|^p + 1 + \delta .$$

Note that if $x \leq 0$, $|x + 1|^p \leq 1 + |x|^p$.

Proof. This is clear if $x < 0$. For p an integer, this can also be obtained by apply Young's inequality to terms $|x|^{p'} |y|^{p-p'}$ in the expansion of $|x + y|^p$.

Now we assume $x \geq 0$. Let $f(x) = |1 + x|^p$. Let $g(x) = K|x|^p + 1 + \delta$. Note that $g(0) > f(0)$. If $f(x) \geq g(x)$ for some x , then by the intermediate value theorem there exists a point where they have equal value. Let x_0 be the first point they are equal. Then $x_0 > 0$. Choose $K = \left(\left| \frac{1}{x_0} \right|^p + 1 \right)^p$. Then $f(x) = |x|^p \left(\left| \frac{1}{x} \right|^p + 1 \right)^p \leq K|x|^p$ for any $x \geq x_0$. \square

Young's inequality: for any $\alpha, \beta > 0$ with $\frac{1}{\alpha} + \frac{1}{\beta} = 1$,

$$ab \leq \frac{(\epsilon a)^\alpha}{\alpha} + \frac{b^\beta}{\beta \epsilon^\beta} .$$

8.4.2 Application to a random dynamical system

In this section, let (x_n) be a Markov process defined by a recursion relation of the type

$$x_{n+1} = F(x_n, \xi_n) , \quad (8.4)$$

for $\{\xi_n\}$ a sequence of independent and identically distributed random variables taking values in a measurable space \mathcal{Y} , and all independent of x_0 , and $F: \mathcal{X} \times \mathcal{Y} \rightarrow \mathcal{X}$ a Borel measurable function. Then for any $V \in \mathcal{B}_b(X)$,

$$TV(x) = \mathbf{E}[V(F(x, \xi_n))].$$

An effective criteria for the transition probabilities to be Feller is as follows:

Theorem 8.4.8 *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 $F: \mathcal{X} \times \mathcal{Y} \rightarrow \mathcal{X}$. If the function $F(\cdot, \xi_n): \mathcal{X} \rightarrow \mathcal{X}$ 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 $\mathbb{P}(\xi_n \in A) = 1$ does not depend on n .), then the corresponding transition semigroup is Feller.

Proof. Denote by $\hat{\mathbb{P}}$ the law of ξ_n on \mathcal{Y} and by $\varphi: \mathcal{X} \rightarrow \mathcal{X}$ an arbitrary continuous bounded function. It follows from the definition of the transition semigroup T that

$$(T\varphi)(x) = \mathbf{E}(\varphi(x_{n+1}) | x_n = x) = \mathbf{E}\varphi(F(x, \xi_n)) = \int_{\mathcal{Y}} \varphi(F(x, y)) \hat{\mathbb{P}}(dy) .$$

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

$$\begin{aligned} \lim_{n \rightarrow \infty} (T\varphi)(z_n) &= \lim_{n \rightarrow \infty} \int_{\mathcal{Y}} \varphi(F(z_n, y)) \hat{\mathbb{P}}(dy) = \int_{\mathcal{Y}} \lim_{n \rightarrow \infty} \varphi(F(z_n, y)) \hat{\mathbb{P}}(dy) \\ &= \int_{\mathcal{Y}} \varphi(F(z, y)) \hat{\mathbb{P}}(dy) = (T\varphi)(z) , \end{aligned}$$

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

If F is continuous in the first variable for each y , then the Markov process is Feller.

Theorem 8.4.9 *Suppose that the function $F(\cdot, \xi_n): \mathcal{X} \rightarrow \mathcal{X}$ is continuous for almost every realisation of ξ_n . If, furthermore, there exists a Borel measurable function $V: \mathcal{X} \rightarrow \mathbf{R}$ with compact sub-level sets and constants $\gamma \in (0, 1)$ and $C \geq 0$ such that*

$$\int_{\mathcal{Y}} V(F(x, y)) \hat{\mathbb{P}}(dy) \leq \gamma V(x) + C , \quad \forall x \in \mathcal{X} ,$$

where $\hat{\mathbb{P}}$ 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 8.4.8. 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. \square

This marks the end of lecture 17.

8.5 Uniqueness of the invariant measure

In this section, we give a very simple criteria for the uniqueness of the invariant measure, due to deterministic contraction. We first review some general results regarding coupling of probability measures and criterion to establish uniqueness

8.5.1 Properties of couplings

Let π_1 and π_2 be two probability measures on metric space (\mathcal{X}, d) . Let μ a coupling of π_1 and π_2 , i.e. $\mu \in \mathcal{P}(\mathcal{X}^2)$ with $(proj_1)_*\mu = \pi_1$ and $(proj_2)_*\mu = \pi_2$. Where $(proj_i) : \mathcal{X} \times \mathcal{X} \rightarrow \mathcal{X}_0$ are the projections to first and second component respectively ($i = 1, 2$).

Example 8.5.1 Let $\pi_1 = N(0, 1)$ and $\pi_2 = N(0, 1)$. Let X, Y be independent $N(0, 1)$ random variables. Then the joint probability distribution of $(aX + bY, cX + dY)$ where $a^2 + b^2 = 1$, $c^2 + d^2 = 1$ is a coupling of π_1 and π_2 . This is a two dimensional Gaussian distribution with covariance matrix $\begin{pmatrix} 1 & ac + bd \\ ac + bd & 1 \end{pmatrix}$.

Lemma 8.5.1 Define $\Delta \subset \mathcal{X} \times \mathcal{X}$ to be the diagonal $\Delta = \{(x, x) : x \in \mathcal{X}\}$. If there exists a coupling μ of π_1 and π_2 such that $\mu(\Delta) = 1$, then $\pi_1 = \pi_2$. In particular $\pi_1 = \pi_2$ if

$$\int_{\mathcal{X} \times \mathcal{X}} 1 \wedge d(x, y) \mu(dx, dy) = 0 \tag{8.5}$$

Proof. Let $A \subset \mathcal{X}$ be a Borel measurable set. Then using assumptions

$$\begin{aligned} \pi_1(A) &= \mu(A \times \mathcal{X}) = \mu((A \times \mathcal{X}) \cap \Delta) \\ &= \mu((\mathcal{X} \times A) \cap \Delta) = \mu(\mathcal{X} \times A) = \pi_2(A). \end{aligned}$$

Hence $\pi_1 = \pi_2$.

Also note that $\{(x, y) : 1 \wedge d(x, y) = 0\} = \Delta$. Then

$$\int_{\mathcal{X} \times \mathcal{X}} 1 \wedge d(x, y) \mu(dx, dy) = 0 \implies \mu(\Delta) = \mu(\{(x, y) : 1 \wedge d(x, y) = 0\}) = 1.$$

Hence (8.5) implies $\mu(\Delta) = 1$ and then $\pi_1 = \pi_2$ by above. \square

Lemma 8.5.2 *Let $\{\mu_n\}$ be a family of couplings of $\pi_1, \pi_2 \in \mathcal{P}(\mathcal{X})$, then $\{\mu_n\}$ is tight.*

Proof. Since π_1, π_2 are probability measures, for any $\varepsilon > 0$ there exists compact sets $K_1, K_2 \subset \mathcal{X}$ such that

$$\pi_1(K_1) > 1 - \frac{\varepsilon}{2}, \quad \pi_2(K_2) > 1 - \frac{\varepsilon}{2}.$$

Then we infer tightness of $\{\mu_n\}$ since for any n :

$$\begin{aligned} \mu_n(\mathcal{X}^2 \setminus K_1 \times K_2) &\leq \mu_n((\mathcal{X} \setminus K_1) \times \mathcal{X}) + \mu_n(\mathcal{X} \times (\mathcal{X} \setminus K_2)) \\ &= \pi_1(K_1^C) + \pi_2(K_2^C) < \varepsilon. \end{aligned}$$

\square

Lemma 8.5.3 *If μ_n is a sequence of measures on \mathcal{X} converging weakly to a measure μ , then for any continuous map $\varphi : \mathcal{X} \rightarrow \mathcal{Y}$, $\varphi_*\mu_n$ converges to $\varphi_*\mu$.*

Proof. Let $f : \mathcal{Y} \rightarrow \mathbf{R}$ be bounded continuous, since $f \circ \varphi$ is bounded continuous, then

$$\int_{\mathcal{Y}} f d\varphi_*\mu_n = \int_{\mathcal{X}} f\varphi d\mu_n \rightarrow \int_{\mathcal{X}} f\varphi d\mu = \int_{\mathcal{Y}} f d\varphi_*\mu,$$

so $\varphi_*\mu_n \rightarrow \varphi_*\mu$. \square

Lemma 8.5.4 *If $\{\mu_n\}$ is a family of couplings of $\pi_1, \pi_2 \in \mathcal{P}(\mathcal{X})$, then so is any of its accumulation points.*

Proof. This due to the fact that $(proj_i)$ is continuous and $(Proj_i)_*\mu = \pi_i$. Suppose $\lim_{n \rightarrow \infty} \mu_n = \mu$ (weakly), then by Lemma 8.5.3, for $i = 1, 2$ and any $f : \mathcal{X} \rightarrow \mathbf{R}$ bounded continuous,

$$\int_{\mathcal{X}} f d((proj_i)_*\mu) = \int_{\mathcal{X}} f d\pi_i.$$

Since measures are determined by bounded continuous functions, we deduce

$$(proj_i)_*\mu = \pi_i,$$

concluding that μ is a coupling of π_1, π_2 . \square

8.5.2 Uniqueness due to deterministic contraction

Consider now the random dynamical system of previous section. Let $x_{n+1} = F(x_n, \xi_{n+1})$ be a Markov chain, where $\{\xi_i\}_{i \geq 1}$ are i.i.d random variables. Then we have the following uniqueness criterion.

In this section, we give a very simple criteria for the uniqueness of the invariant measure, due to deterministic contraction.

Theorem 8.5.5 *If there exists a constant $\gamma \in (0, 1)$ such that*

$$\mathbf{E}d(F(x, \xi_1), F(y, \xi_1)) \leq \gamma d(x, y) \quad \forall x, y \in \mathcal{X}, \quad (8.6)$$

then the process (8.4) has at most one invariant probability measure.

Proof. Let π_1 and π_2 be any two invariant probability measures for (8.4). Let x_0 and y_0 be two independent \mathcal{X} -valued random variables, independent of $\{\xi_i\}$ and with laws $\mathcal{L}(x_0) = \pi_1$ and $\mathcal{L}(y_0) = \pi_2$. Then define x_n and y_n as follows:

$$\begin{aligned} x_1 &= F(x_0, \xi_1), & x_{n+1} &= F(x_n, \xi_{n+1}), & n &\geq 1; \\ y_1 &= F(y_0, \xi_1), & y_{n+1} &= F(y_n, \xi_{n+1}), & n &\geq 1. \end{aligned}$$

Then $\mathcal{L}(x_n) = \pi_1$, $\mathcal{L}(y_n) = \pi_2$ for all $n \geq 0$, and $\mu_n := \mathcal{L}((x_n, y_n))$ is a coupling of π_1 and π_2 . By Lemma 8.5.2 and 8.5.4, and Prohorov's Theorem 8.3.3, there exists a weakly convergent subsequence μ_{n_k} , whose limit μ is a coupling of π_1 and π_2 . If

$$\lim_{k \rightarrow \infty} \int_{\mathcal{X}} 1 \wedge d(x, y) d\mu_{n_k} = 0, \quad (8.7)$$

so that, by the boundedness and the continuity of $1 \wedge d(\cdot, \cdot)$,

$$\int_{\mathcal{X}} 1 \wedge d(x, y) d\mu = 0.$$

Then we deduce $\pi_1 = \pi_2$ by Lemma 8.5.1.

Therefore it remains to show (8.7) to conclude the proof, this is proved in the next Lemma. \square

Definition 8.5.6 We say that (x_n, y_n) is a synchronized coupling if x_0, y_0 are independent, and $x_{n+1} = F(x_n, \xi_{n+1})$ and $y_{n+1} = F(y_n, \xi_{n+1})$ are defined by iteration (with the same noise ξ_n).

The following lemma shows that the synchronised coupling (x_n, y_n) having its mass concentrate more and more on the diagonal.

Lemma 8.5.7 *Let $\mu_n := \mathcal{L}((x_n, y_n))$ be a synchronized coupling. Assume the conditions of Theorem 8.5.5, then,*

$$\lim_{n \rightarrow \infty} \mathbf{E}(1 \wedge d(x_n, y_n)) = \lim_{n \rightarrow \infty} \int_{\mathcal{X}} 1 \wedge d(x, y) d\mu_n = 0.$$

Proof. Note that contraction assumption is about fixed starting point. Here we do not impose integrability of $d(x_0, y_0)$, nor on $d(x_n, y_n)$. This means we have to tread carefully, and hence the introduction of $1 \wedge d$. Let $c > 0$. Since $\varphi(t) = 1 \wedge ct$ is concave we can use conditional Jensen's inequality and the Markov property to derive the following

$$\mathbf{E}(1 \wedge c d(x_n, y_n)) \leq \mathbf{E} \mathbf{E}(\varphi \circ d(x_n, y_n) \mid x_{n-1}, y_{n-1}) \leq \mathbf{E} \varphi(\mathbf{E}(d(x_n, y_n) \mid x_{n-1}, y_{n-1})).$$

Using $x_n = F(x_{n-1}, \xi_n)$, $y_n = F(y_{n-1}, \xi_n)$, by independence of ξ_n and (x_{n-1}, y_{n-1}) ,

$$\begin{aligned} \mathbf{E}(d(x_n, y_n) \mid x_{n-1}, y_{n-1}) &= \mathbf{E}(d(F(x, \xi_n), F(y, \xi_n))_{x=x_{n-1}, y=y_{n-1}}) \\ &\leq \mathbf{E}(\gamma d(x_{n-1}, y_{n-1})). \end{aligned}$$

where the inequality is nothing but (8.6) We have the inequality for any $a > 0$, and any n ,

$$\mathbf{E}(1 \wedge c d(x_n, y_n)) \leq \mathbf{E} \varphi(\mathbf{E}(d(x_{n-1}, y_{n-1}))) = \mathbf{E}(1 \wedge c \gamma d(x_{n-1}, y_{n-1})).$$

By iteration,

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

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

$$\mathbf{E}(1 \wedge d(x_n, y_n)) \longrightarrow 0,$$

proving the Lemma. □

This marks the end of first half lecture 18.

8.6 Uniqueness and minorisation

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** as introduced at Definition 7.7.15 in Section 7.7.3.

8.6.1 Properties of the Total Variation

In this subsection we review some definitions and useful properties in sight of the main Theorem 8.6.10. Recall the following

1. The total variation distance between two probability measures μ, ν on Ω is

$$\|\mu - \nu\|_{\text{TV}} = 2 \sup_A |\mu(A) - \nu(A)|.$$

where the supremum runs over all measurable subsets .

2. The duality formulation (recall Remark 7.7.16) equivalently gives us

$$\|\mu - \nu\|_{\text{TV}} = \sup_{\substack{f \in \mathcal{B}_b(\mathcal{X}) \\ \|f\|_\infty = 1}} \left| \int_{\mathcal{X}} f(x) \mu(dx) - \int_{\mathcal{X}} f(x) \nu(dx) \right|,$$

where the maximum is run over bounded measurable functions.

Remark 8.6.1 (1) It is also a fact that under very mild conditions on \mathcal{X} (being a complete separable metric space is more than enough), (10.6) is the same as the seemingly weaker norm,

$$\|\mu - \nu\|_{\text{TV}} = \sup_{\substack{f \in \mathcal{C}_b(\mathcal{X}) \\ \|f\|_\infty = 1}} \left| \int_{\mathcal{X}} f(x) \mu(dx) - \int_{\mathcal{X}} f(x) \nu(dx) \right|, \quad (8.8)$$

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

$$\sup_{\substack{f \in \mathcal{C}_b(\mathcal{X}) \\ \|f\|_\infty = 1}} \left| \int_{\mathcal{X}} f(x) \mu(dx) - \int_{\mathcal{X}} f(x) \nu(dx) \right| = \frac{1}{2} \sup_{\substack{f \in \mathcal{C}_b(\mathcal{X}) \\ \text{Oso}(f) = 1}} \left| \int_{\mathcal{X}} f(x) \mu(dx) - \int_{\mathcal{X}} f(x) \nu(dx) \right|,$$

where $\text{Oso}(f) = \sup(f) - \inf(f)$.

Another equivalent definition for the total variation is as follows, this will be the one we use in the formulation. If μ and ν are two positive measures, we use $\mu \ll \eta$ to indicate that μ is absolutely continuous with respect to ν .

Definition 8.6.2 Given two (positive) measures μ and ν on a measurable space Ω . Let η be a (positive) measure such that $\mu \ll \eta$ and $\nu \ll \eta$, define

$$\|\mu - \nu\|_{\text{TV}} = \int_{\Omega} \left| \frac{d\mu}{d\eta} - \frac{d\nu}{d\eta} \right| d\eta. \quad (8.9)$$

(Where we denote by $\frac{d\mu}{d\eta}$ and $\frac{d\nu}{d\eta}$ their Radon-Nikodym derivatives with respect to η .)

Remark 8.6.3

1. Given two positive measures μ and ν on a measurable space Ω , there always exists η such that $\mu \ll \eta$ and $\nu \ll \eta$. For example, it is easy to check that both μ and ν are absolutely continuous with respect to $\mu + \nu$.

2. The definition (8.9) is independent of the choice of η . In fact, since

$$\frac{d\mu}{d\eta} = \frac{d\mu}{d(\mu + \nu)} \frac{d(\mu + \nu)}{d\eta}, \quad \frac{d\nu}{d\eta} = \frac{d\nu}{d(\mu + \nu)} \frac{d(\mu + \nu)}{d\eta},$$

then

$$\begin{aligned} \int \left| \frac{d\nu}{d\eta} - \frac{d\mu}{d\eta} \right| d\eta &= \int \frac{d(\mu + \nu)}{d\eta} \left| \frac{d\nu}{d(\mu + \nu)} - \frac{d\mu}{d(\mu + \nu)} \right| d\eta \\ &= \int \left| \frac{d\nu}{d(\mu + \nu)} - \frac{d\mu}{d(\mu + \nu)} \right| d(\mu + \nu). \end{aligned}$$

Hence we can simply define:

$$\|\mu - \nu\|_{\text{TV}} = \int_{\Omega} \left| \frac{d\nu}{d(\mu + \nu)} - \frac{d\mu}{d(\mu + \nu)} \right| d(\mu + \nu). \quad (8.10)$$

Remark 8.6.4 Definition (8.9) means that the total variation distance between μ and ν is given by the $L_1(\eta)$ norm of the corresponding Radon-Nikodym derivatives.

Example 8.6.1 Consider μ, ν measures on \mathbf{R} that have densities w.r.t. the Lebesgue measure, i.e. $d\mu = f dx$ and $d\nu = g dx$. Then

$$\|\mu - \nu\|_{\text{TV}} = \int_{\mathbf{R}} |f - g| dx = \|f - g\|_{L^1(\mathbf{R})}.$$

Since, for any two positive numbers, one has $|x - y| = x + y - 2 \min\{x, y\}$, then we have

$$\left| \frac{d\mu}{d\eta} - \frac{d\nu}{d\eta} \right| = \frac{d\mu}{d\eta} + \frac{d\nu}{d\eta} - 2 \left(\frac{d\mu}{d\eta} \wedge \frac{d\nu}{d\eta} \right). \quad (8.11)$$

Notation. If μ and ν are two positive measures, we denote by $\mu \wedge \nu$ the measure obtained by

$$(\mu \wedge \nu)(A) = \int_A \min \left\{ \frac{d\mu}{d(\mu + \nu)}, \frac{d\nu}{d(\mu + \nu)} \right\} d(\mu + \nu).$$

Lemma 8.6.5 Assuming $\mu, \nu \in \mathcal{P}(\Omega)$, we have a useful identity:

$$\|\mu - \nu\|_{\text{TV}} = 2(1 - \mu \wedge \nu(\Omega)). \quad (8.12)$$

Proof. By plugging (8.11) into the definition (8.10), we immediately see that

$$\|\mu - \nu\|_{\text{TV}} = \mu(\Omega) + \nu(\Omega) - 2\mu \wedge \nu(\Omega) = 2(1 - \mu \wedge \nu(\Omega)),$$

as required. \square

Exercise 8.6.1 Show that if η is an positive measure on Ω such that $\mu \ll \eta$ and $\nu \ll \eta$, then

$$(\mu \wedge \nu)(A) = \int_A \min \left\{ \frac{d\mu}{d\eta}, \frac{d\nu}{d\eta} \right\} d\eta.$$

Lemma 8.6.6 *The space of probability measures $\mathcal{P}(\Omega)$ endowed with the total variation distance $\|\cdot\|_{\text{TV}}$ is complete.*

Proof. Let $\mu_n \in \mathcal{P}(\mathcal{X})$ be a sequence of probability measures that is Cauchy in the total variation distance. Let η be defined by $\eta = \sum_{n=1}^{\infty} \frac{1}{2^n} \mu_n$. Then $\mu_n \ll \eta$ for each n . By (8.9), the total variation distance is equal to the \mathcal{L}^1 distance between the corresponding Radon-Nikodym derivatives:

$$\|\mu_n - \mu_m\| = \int_{\mathcal{X}} \left| \frac{d\mu_n}{d\eta} - \frac{d\mu_m}{d\eta} \right| d\eta = \left\| \frac{d\mu_n}{d\eta} - \frac{d\mu_m}{d\eta} \right\|_{L^1(\eta)}.$$

Then $\{\mu_n\}$ is Cauchy in TV if and only if $\{\frac{d\mu_n}{d\eta}\}$ is a Cauchy sequence in $L^1(\eta)$. But $\mathcal{L}^1(\Omega, \eta)$ is complete and so $\frac{d\mu_n}{d\eta} \xrightarrow{L^1} \bar{f} : \Omega \rightarrow \mathbf{R}_+$ (with $\bar{f} \geq 0$ a function with unit L^1 -norm). Then $\mu_n \rightarrow \bar{\mu} \in \mathcal{P}(\Omega)$ in total variation, where $\frac{d\bar{\mu}}{d\eta} = \bar{f}$ and the probability measure $\bar{\mu}$ is given by $\bar{\mu}(A) = \int_A \bar{f} d\eta$. \square

8.6.2 Uniqueness by minorisation

Lemma 8.6.7 *Let μ, ν be two probability measures on \mathcal{X} . Let $\bar{\mu}$ and $\bar{\nu}$ be defined as follows*

$$\bar{\mu} = \frac{\mu - \mu \wedge \nu}{\frac{1}{2}\|\mu - \nu\|_{\text{TV}}}, \quad \bar{\nu} = \frac{\nu - \mu \wedge \nu}{\frac{1}{2}\|\mu - \nu\|_{\text{TV}}}. \quad (8.13)$$

Then $\bar{\mu}$ and $\bar{\nu}$ are probability measures and the following equality holds

$$\mu - \nu = \frac{1}{2}\|\mu - \nu\|_{\text{TV}}(\bar{\mu} - \bar{\nu}). \quad (8.14)$$

Proof. By Lemma 8.6.5 (with $\Omega = \mathcal{X}$) we have

$$\frac{1}{2}\|\mu - \nu\|_{\text{TV}} = 1 - (\mu \wedge \nu)(\mathcal{X}) = (\mu - \mu \wedge \nu)(\mathcal{X}).$$

Hence we see that

$$\begin{aligned} \mu &= \frac{1}{2}\|\mu - \nu\|_{\text{TV}} \bar{\mu} + \mu \wedge \nu \\ \mu &= \frac{1}{2}\|\mu - \nu\|_{\text{TV}} \bar{\nu} + \mu \wedge \nu \end{aligned}$$

Subtracting the two, we obtain $\mu - \nu = \frac{1}{2}\|\mu - \nu\|_{\text{TV}}(\bar{\mu} - \bar{\nu})$. \square

Lemma 8.6.8 *Let μ, ν be two probability measures on \mathcal{X} and T a transition operator. Then*

$$\begin{aligned} \|T\mu - T\nu\|_{\text{TV}} &= \frac{1}{2} \|\mu - \nu\|_{\text{TV}} \cdot \|T\bar{\mu} - T\bar{\nu}\|_{\text{TV}} \\ &\leq \|\mu - \nu\|_{\text{TV}}. \end{aligned} \quad (8.15)$$

Proof. Applying the operator T to (8.14) we have

$$T\mu - T\nu = \frac{1}{2} \|\mu - \nu\|_{\text{TV}} (T\bar{\mu} - T\bar{\nu}).$$

Then (8.15) follows, noting that $\|T\bar{\mu} - T\bar{\nu}\|_{\text{TV}} \leq 2$. \square

Definition 8.6.9 Let $\eta \in \mathcal{P}(\mathcal{X})$. We say transition probability $P = (P(x, \cdot))$ is minorized by η if there exists $a > 0$ such that

$$P(x, \cdot) \geq a\eta, \quad \forall x \in \mathcal{X}.$$

A family of transition probabilities for which the above condition holds is also said to satisfy the Doeblin's Condition.)

Note. For a finite state chain this means

$$P(i, j) \geq a\eta(j), \quad \forall i, j \in \mathcal{X}. \quad (8.16)$$

For example, if we take $\eta = (0, \dots, 1, \dots, 0) = \delta_{j_0}$, where every entry but the j_0 th vanishes, then (8.16) $\Leftrightarrow P(i, j_0) \geq a$ for all i . This is equivalent to having the j_0 -th column of P bounded below by $a(1, 1, \dots, 1)^T$. We will show a convergence theorem in the total variation distance (cf. Proposition 7.10.1, Theorem 7.10.10, Theorem 7.7.19.).

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

Theorem 8.6.10 *Let P be a transition probability on a space \mathcal{X} . Assume that P is minorized by a probability measure η on \mathcal{X} , so there exists $\alpha > 0$ such that $P(x, \cdot) \geq \alpha\eta$ for every $x \in \mathcal{X}$. Then*

(1) *P has a unique invariant probability measure π .*

(2) *Furthermore for any $\mu, \nu \in P(\mathcal{X})$,*

$$\|T^{n+1}\mu - T^{n+1}\nu\|_{\text{TV}} \leq (1 - \alpha)^n \|\mu - \nu\|_{\text{TV}}.$$

Proof. For any measure $m \in P(\mathcal{X})$,

$$Tm = \int_{\mathcal{X}} P(x, \cdot) dm \geq \alpha\eta,$$

it follows that $Tm - \alpha\eta$ is a positive measure with total mass $(Tm - \alpha\eta)(\mathcal{X}) = 1 - \alpha$. Thus,

$$\|Tm - T\bar{m}\|_{\text{TV}} \leq (1 - \alpha) \left\| \frac{Tm - \alpha\eta}{1 - \alpha} - \frac{T\bar{m} - \alpha\eta}{1 - \alpha} \right\|_{\text{TV}} \leq 2(1 - \alpha).$$

In sight of the equality in (8.15), this implies (using $m = \bar{\mu}$, $\bar{m} = \bar{\nu}$)

$$\|T\mu - T\nu\|_{\text{TV}} = \frac{1}{2} \|\mu - \nu\|_{\text{TV}} \|T\bar{\mu} - T\bar{\nu}\|_{\text{TV}} \leq (1 - \alpha) \|\mu - \nu\|_{\text{TV}}.$$

Hence T is a strict contraction, By Banach's fixed point theorem on $(\mathcal{P}(\mathcal{X}), \|\cdot\|_{\text{TV}})$ we have that $\mu \mapsto T\mu$ has a unique fixed point. Finally,

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

by iteration. □

Taking ν to be an invariant measure, so $T^n\pi = \pi$, the following lemma follows immediately:

Corollary 8.6.11 *Assume the conditions of Theorem 8.6.10. Let μ be any probability measure, and π the unique invariant probability measure. Then*

$$\|T^n\mu - \pi\|_{\text{TV}} \leq (1 - \alpha)^n \|\mu - \pi\|_{\text{TV}}.$$

Compare the results with Theorem 7.10.10 in Section 7.10.4. The minorisation assumption can be weakened to hold for some integer n_0 .

Exercise 8.6.2 Assume there exists $n_0, \alpha \in (0, 1)$ and $\eta \in \mathcal{P}(\mathcal{X})$ such that

$$P^{n_0}(x, \cdot) \geq \alpha\eta, \quad \forall x \in \mathcal{X}_0.$$

Show there exists a unique invariant probability measure π and there exists $a \in (0, 1)$ such that

$$\|T^n\mu - T^n\nu\|_{\text{TV}} \leq a^n \|\mu - \nu\|_{\text{TV}}.$$

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 8.6.8, and use the notation there, we obtain

$$\|T^{n_0}\mu - T^{n_0}\nu\|_{\text{TV}} = \frac{1}{2} \|\mu - \nu\|_{\text{TV}} \cdot \|T^{n_0}\bar{\mu} - T^{n_0}\bar{\nu}\|_{\text{TV}}.$$

$$= (1 - \alpha) \|\mu - \nu\|_{\text{TV}} \cdot \frac{1}{2} \left\| \frac{T^{n_0} \bar{\mu} - \alpha \eta}{1 - \alpha} - \frac{T^{n_0} \bar{\nu} - \alpha \eta}{1 - \alpha} \right\|_{\text{TV}} \leq (1 - \alpha) \|\mu - \nu\|_{\text{TV}}.$$

Thus T^{n_0} is a contraction. If $n, m \geq n_0 k$ where $k \in \mathbf{N}$, we use the property that T does not increase the total variation norm (Lemma 8.6.8),

$$\begin{aligned} \|T^n \mu - T^m \nu\|_{\text{TV}} &\leq \|T^{n_0 k} T^{n-n_0 k} \mu - T^{n_0 k} T^{m-n_0 k} \nu\|_{\text{TV}} \\ &\leq (1 - \alpha)^k \|T^{n-n_0 k} \mu - T^{m-n_0 k} \nu\|_{\text{TV}} \leq 2(1 - \alpha)^k, \end{aligned}$$

So $T^n \mu$ is a Cauchy sequence and converges to a probability measure π , by the completeness of $\mathbb{P}(\mathcal{X})$. Since T is continuous, we see that π is an invariant measure. The uniqueness follows from the contraction.

Example 8.6.2 Let $\{\xi_n\}$ be independent identically distributed real valued random variables with transition probabilities probability distribution

$$\Gamma(A) = \frac{1}{\sqrt{2\pi}} \int_A e^{-\frac{(y-g(x))^2}{2}} dy, \quad \forall A \in \mathcal{B}(\mathbf{R}).$$

where

$$g(x) = 4 \cos(x),$$

so

$$Tf(x) = \frac{1}{\sqrt{2\pi}} \int_{\mathbf{R}} f(y) e^{-\frac{(y-g(x))^2}{2}} dy.$$

Then the minorisation condition is satisfied. Take $e^{-\frac{(y-g(x))^2}{2}} \geq p(y) := e^{-\frac{|y+4|^2}{2}}$ for $y > 4$, this is $e^{-\frac{|y+4|^2}{2}}$ and for $y < -4$ it bounded from below by $e^{-\frac{|y-4|^2}{2}}$, both are integrable in y . Let us define $p(y) = \inf_x e^{-\frac{(y-g(x))^2}{2}}$ for $y \in [-4, 4]$, as above for $|y| > 4$. Let η be the probability measure with density cp for a normalising constant c . Then the minorisation condition is satisfied.

Exercise 8.6.3 Is the family of probability measures $c \sin(x+y)^2 dy$ on $[0, 1]$ where c is a normalising constant on $[0, 1]$ minorised by the δ measure at $\frac{1}{2}$? Here x runs through $[0, 1]$.

[This marks the end of lecture 18.](#)

8.6.3 Strong Feller

Definition 8.6.12 Let \mathcal{X} be a separable metric space. Then for any probability measure $\mu \in \mathcal{P}(\mathcal{X})$, there exists a closed set A such that A is the smallest closed subset of full measure. This is the support of μ .

Lemma 8.6.13 The support of a measure on \mathcal{X} is the set of points with the property that any open set containing it has positive measure, i.e. $\text{supp}(\mu) = \{x \in \mathcal{X} : \mu(B(x, \varepsilon)) > 0, \forall \varepsilon > 0\}$.

Example 8.6.3 (*Measures, support and mutual singularity*)

1. Consider $\mu = \sum_{i=1}^n \delta_{x_i}$, then $\text{supp}(\mu) = \{x_1, \dots, x_n\}$.
2. If $\mu = N(0, 1)$ is a standard normal random variable on \mathbf{R} , then $\text{supp}(\mu) = \mathbf{R}$.
3. If $\mu = \mathbf{1}_{[0,1]}(x)dx + \mathbf{1}_{(2,3]}(x)dx$, then $\text{supp}(\mu) = [0, 1] \cup [2, 3]$.
4. Now consider $\mu = \delta_0$ and $\nu = \mathbf{1}_{(0,1]}dx$. Then μ and ν are mutually singular, since $\nu((0, 1]) = 1$ and $\delta_0((0, 1]) = 0$. On the other hand

$$\text{supp}(\mu) \cap \text{supp}(\nu) = \{0\} \cap [0, 1] = \{0\} \neq \emptyset.$$

Even if μ and ν are mutually singular, their supports are not disjoint.

Theorem 8.6.14 *Let $\mu, \nu \in \mathcal{P}(\mathcal{X})$ be mutually singular invariant probability measures for a transition operator T . Suppose T has the Strong Feller property. Then*

$$\text{supp}(\mu) \cap \text{supp}(\nu) = \emptyset.$$

Proof. By assumption ($\mu \perp \nu$) there exists $F \subset \mathcal{X}$ Borel measurable such that $\mu(F) = 1$ and $\nu(F) = 0$. Let $\varphi := \mathbf{1}_F$. Then $T\varphi$ is continuous by the Strong Feller property, and $T\varphi \in [0, 1]$. Since $\nu(F) = 0$, $\varphi = \mathbf{1}_F$ and invariance of ν ,

$$0 = \int_{\mathcal{X}} \varphi(y) \nu(dy) = \int_{\mathcal{X}} \varphi(y) T\nu(dy) = \int_{\mathcal{X}} T\varphi(y) \nu(dy).$$

Since $T\varphi \geq 0$, then $\nu(\{y : T\varphi(y) = 0\}) = 1$ (i.e. $T\varphi = 0$ ν -a.e., but we do not know necessarily which elements are in the exceptional sets.). With the assumption $T\varphi$ is continuous, coming from the Strong property, we may assert that $\{y : T\varphi(y) = 0\}$ is closed and therefore $\text{supp}(\nu) \subset \{y : T\varphi(y) = 0\}$. If $x \in \text{supp}(\nu)$, then $T\varphi(x) = 0$ necessarily. Also

$$1 = \int_{\mathcal{X}} \varphi(y) \mu(dy) = \int_{\mathcal{X}} T\varphi(y) \mu(dy).$$

Which implies $T\varphi(y) = 1$ for μ -a.e. y . Then for $z \in \text{supp}(\mu)$, $T\varphi(z) = 1$. Hence $\text{supp}(\mu) \cap \text{supp}(\nu) = \emptyset$. \square

Remark 8.6.15

1. Suppose T has the strong Feller property, and we can identify a common point in the support of every invariant probability measure, then by Part (b) of Theorem 9.3.3 there exists only one invariant probability measure.
2. We will see (cf. Theorem 9.3.3) that all ergodic invariant probability measure for a transition probability are either identical or mutually singular.

Definition 8.6.16 If $x \in \mathcal{X}$ is a point such that for every neighbourhood A of x and for every $y \in \mathcal{X}$, $P(y, A) > 0$, we say that x is accessible.

Exercise 8.6.4 Let P be a time homogeneous transition probability on a complete separable metric space \mathcal{X} with an accessible point x . Show that if P is strong, then P can have at most one invariant probability measure.

Strong Feller property for transition probabilities with kernels

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 $T_*f(x) = \int f(y)p(y-x) = f * p$. First, let us review below Lemma 8.6.17 regarding convolution of functions.

Lemma 8.6.17 *Let $f : \mathbf{R}^n \rightarrow \mathbf{R}$ be bounded Borel measurable, and $g : \mathbf{R}^n \rightarrow \mathbf{R}$ be in $L^1(dx)$. Then the convolution*

$$f * g = \int_{-\infty}^{\infty} f(y)g(x-y)dy$$

is a bounded continuous function.

Proof. (a) First suppose that $g \in C_K^\infty$ is smooth with compact support. For any $x \in \mathbf{R}^n$,

$$f * g(x) - f * g(x') = \int f(y)(g(x-y) - g(x'-y))dy \rightarrow 0, \quad \text{as } x' \rightarrow x. \quad (8.17)$$

Since $g(x-y) - g(x'-y) \rightarrow 0$ when $x' \rightarrow x$, and $|f(y)(g(x-y) - g(x'-y))| \leq 2|f|_\infty|g|_\infty$, so that the continuity follows from the dominated convergence theorem.

(b) If $g \in L^1(\mathbf{R})$, it can be approximated by functions $g_k \in C_K^\infty$ with compact support, i.e. $g_k \xrightarrow{L^1} g$. Pick x and let $x' \rightarrow x$, we need to show $f * g(x') \rightarrow f * g(x)$.

The difference $f * g(x) - f * g(x')$ can be split up into 3 terms, such that

$$|f * g(x) - f * g(x')| \leq |f * g(x) - f * g_k(x)| + |f * g_k(x) - f * g_k(x')| + |f * g_k(x') - f * g(x')| \quad (8.18)$$

For the first term (recall $f * h = h * f$ by the translation invariance of the Lebesgue measure)

$$|f * g(x) - f * g_k(x)| \leq \int |f(y-x)| |g_k(y) - g(y)| dy \leq |f|_\infty |g_k - g|_{L^1}.$$

The upper bound does not depend on the points x , and the convergence is uniform in x , and therefore the third term has the bound also: $|f * g_k(x') - f * g(x')| \leq |f|_\infty |g_k - g|_{L^1}$. Hence, given $\varepsilon > 0$, there exists an N such that for $k \geq N$,

$$|f * g(x) - f * g_k(x)| + |f * g_k(x') - f * g(x')| < \frac{2}{3}\varepsilon.$$

For the second term of inequality (8.18), we can apply (8.17) to g_N such that for some $\delta = \delta(\varepsilon) > 0$

$$|f * g_N(x) - f * g_N(x')| < \frac{\varepsilon}{3}, \quad |x - x'| < \delta.$$

Hence using $k = N$ in (8.18), we deduce continuity since for $|x - x'| < \delta$ we have

$$|f * g(x) - f * g(x')| < \varepsilon.$$

Hence $f * g$ is continuous. □

From the Lemma, when $Tf = (f * g)(x)$ is given by convolution, the Strong Feller Property follows immediately:

Proposition 8.6.18 *Let $\mathcal{X} = \mathbf{R}^n$, $g : \mathbf{R}^n \rightarrow \mathbf{R}_+$ be Borel measurable and in $L^1(dx)$ such that $\int_{\mathbf{R}^n} g(x) dx = 1$, with $dx = \text{Lebesgue measure}$ (i.e. $g(x)dx$ is a probability measure). Suppose*

$$Tf(x) = \int f(y)g(x-y) dy.$$

Then T has the strong Feller property.

Proof. We can normalise g to have mass 1 and apply the Lemma. □

If T is defined by a density which is not homogeneous, strong Feller property can still be proved, but there is not such a beautiful statement. Below we explore some situations. Let $B_a(x)$ stand for the open ball centred at x with radius a .

Example 8.6.4 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 more generally (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 .

Proof. Let $f : \mathcal{X} \rightarrow \mathbf{R}$ be bounded measurable function, and let $x_n \rightarrow x$.

$$|T_*f(x_n) - T_*f(x)| \leq \left| \int f(y)(p(x_n, y) - p(x, y)) \mu(dy) \right|.$$

Since $p(x_n, y) \rightarrow p(x, y)$ and for x_n near x , $|p(x_n, y) - p(x, y)| \leq \sup_{x \in B_a(x)} p(x, y)$ and the latter in L^1 , by the dominated convergence theorem, we may take the limit $n \rightarrow \infty$ inside the integration sign. Concerning the alternative assumption (2b), uniform integrability will allow us to take the limit inside the integral. □

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 convergence

$$\lim_{x \rightarrow x_0} \sup_A |T_t \mathbf{1}_A(x) - T_t \mathbf{1}_A(x_0)| = \lim_{x \rightarrow x_0} \sup_A |P(x, A) - P(x_0, A)| = 0,$$

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 notes by Martin Hairer and notes by Jan Sedler.

This marks the end of lecture 19 - Week 8.

8.7 Using P -invariant sets

We have seen the uniqueness of invariant probability measure due to the deterministic contraction (8.6). There are situations (we will see one of them immediately) where (8.6), i. e. $\mathbf{E}d(F(x, \xi_1), F(y, \xi_1)) \leq \gamma d(x, y)$, 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 8.7.1 Let $P = (P(x, \cdot), x \in \mathcal{X})$ 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 $x \in A$.

Example 8.7.1 If \mathcal{X} is finite, a closed communication class is a P -invariant set.

Remark 8.7.2 If A is a P -invariant set and $x_0 \sim \pi$, then

$$\mathbb{P}(x_0 \in A, x_1 \in A, \dots, x_n \in A) = \pi(A).$$

This is a consequence of the Chapman-Kolmogorov equations:

$$\begin{aligned} & \mathbb{P}(x_0 \in A_0, x_1 \in A_1, \dots, x_n \in A_n) \\ &= \int_{A_0} \int_{A_1} \cdots \int_{A_{n-1}} \int_{A_n} P(x_{n-1}, dx_n) P(x_{n-2}, dx_{n-1}) \cdots P(x_1, dx_2) P(x_0, dx_1) \pi(dx_0) \\ &= \int_{A_0} \int_{A_1} \cdots \int_{A_{n-1}} P(x_{n-1}, A_n) P(x_{n-2}, dx_{n-1}) \cdots P(x_1, dx_2) P(x_0, dx_1) \pi(dx_0) \\ &= \mathbb{P}(x_0 \in A, x_1 \in A, \dots, x_{n-1} \in A) = \cdots = \mathbb{P}(x_0 \in A) \\ &= \pi(A). \end{aligned}$$

Where $A_i = A$ to keep track of integrals.

Remark 8.7.3 Let P_π be the stationary measure on $\mathcal{X}^{\mathbf{N}}$ and A is a P -invariant set. Let $A^{\mathbf{N}} = A \times A \times \cdots \times A \in \mathcal{B}(\mathcal{X}^{\mathbf{N}})$. Then $P_\pi(A^{\mathbf{N}}) = \pi(A)$.

Example 8.7.2 Let $\mathcal{X} = \mathbf{R}$, and define the transition probabilities as follows

$$P(x, A) = \begin{cases} \int_0^1 \mathbf{1}_A(x) dx & x \geq 0, \\ \int_{-1}^0 \mathbf{1}_A(x) dx & x < 0. \end{cases} \quad \forall A \in \mathcal{B}(\mathbf{R}).$$

Then both $[0, 1]$ and $[-1, 0)$ are P -invariant sets. So are $(0, 1)$ and $(-1, 0)$ (absolute continuity of Lebesgue measure). But $[-1, 0]$ is not a P -invariant set.

More generally $[0, 1 + a^2]$, $[-1 - a^2, 0)$ are p -invariant sets for all $a > 0$.

Note. We can also take $[0, 1] \cap \mathbf{Q}^C$, however restrictions to an open and closed set fits in with our set up easily.

Restrictions of Markov Chain. Given a P -invariant set $A \subset \mathcal{X}$, so that $\forall x \in A, P(x, A) = 1$ then $\{P(x, \cdot), x \in A\}$ are transition probabilities on A . Also, for any Borel set B ,

$$P(x, B \cap A) = P(x, B), \quad \text{when } x \in A.$$

If there exists a closed P -invariant set $A \subset \mathcal{X}$, one can restrict P to A to obtain a Markov process on the complete separable metric space A , and Krylov-Bugoliubov criterion (Theorem 8.4.1) can be applied. (Moreover one may be able to check (8.6) for x and y in A to establish uniqueness of invariant measure π).

In example 8.7.2 above, we can restrict the chain to $[0, 1]$, a compact metric space (so that Krylov-Bugoliubov criterion can be verified).

Lemma 8.7.4 Let A be P -invariant, where P is a t.p. on \mathcal{X} . Let $\pi^0 \in \mathcal{P}(A)$ be a probability measure on A . Define a probability measure π on \mathcal{X} by

$$\pi(B) := \pi^0(B \cap A), \quad \forall B \in \mathcal{B}(\mathcal{X}).$$

1. If π^0 is invariant for P restricted to A , then π is invariant for P on \mathcal{X} .
2. Conversely, if π is invariant for P , its restriction π^0 on A is invariant for P on A .

Proof. 1. Since π^0 is invariant for P restricted to A , we have

$$\pi^0(C) = (T\pi^0)(C) = \int_A P(x, C) \pi^0(dx), \quad \text{for any } C \in A \cap \mathcal{B}(\mathcal{X}) = \{B : B \subset A\}. \quad (8.19)$$

Then, given any Borel set $B \subset \mathcal{X}$, since $\text{supp}(\pi) \subset A$ we have

$$\begin{aligned} (T\pi)(B) &= \int_{\mathcal{X}} P(x, B) \pi(dx) = \int_A P(x, B) \pi(dx) = \int_A P(x, B \cap A) \pi(dx) \\ &= \pi^0(B \cap A) = \pi(B). \end{aligned}$$

Where we used $B \cap A \subset A$ and (8.19) to pass to the last line. This proves one direction. In the other direction, suppose that $\pi(C) = \int_{\mathcal{X}} p(x, C) \pi(dx)$ for any $C \in \mathcal{B}(\mathcal{X})$. Let π^0 denote the restriction of π on A . Then for any Borel set $B \subset A$,

$$\pi(B) = \int_{\mathcal{X}} p(x, B) \pi(dx) = \int_A p(x, B) \pi(dx) + \int_{A^c} p(x, B) \pi(dx).$$

Also, by the invariance of A ,

$$\pi(A) = \int_A p(x, A)\pi(dx) + \int_{A^c} p(x, A)\pi(dx) = \pi(A) + \int_{A^c} p(x, A)\pi(dx),$$

concluding $\int_{A^c} p(x, A)\pi(dx) = 0$ and $\int_{A^c} p(x, B)\pi(dx) = 0$ for $B \subset A$. Inserting this into the previous line, we conclude

$$\pi(B) = \int_A p(x, B)\pi(dx) + \int_{A^c} p(x, B)\pi(dx) = \int_A p(x, B)\pi(dx),$$

completing the proof. \square

Theorem 8.7.5 *Let P (with associated operator T) be a Feller transition probabilities on \mathcal{X} . Suppose that A is a compact P -invariant set, then the restriction of P to A is also Feller. Consequently, there exists an invariant probability measure for P (on \mathcal{X}).*

Proof. Let P^0 be the restriction of P to A with corresponding transition operator T^0 . Since A is compact we only need to show T^0 is Feller to conclude the existence of an invariant probability measure for P^0 , using Theorem 8.4.2. Then, by Lemma 8.7.4 there exists an invariant $\pi \in \mathcal{P}(\mathcal{X})$ for P , with $\pi|_A = \pi^0$.

Let $f : A \rightarrow \mathbf{R}$ be any bounded continuous function, Tietze's theorem's allow an extension of f to a bounded continuous function $\bar{f} : \mathcal{X} \rightarrow \mathbf{R}$. Then, if T^0 is the transition operator on A associated to P^0 , given $x \in A$,

$$T^0 f(x) = \int_A f(y) P^0(x, dy) = \int_A \bar{f}(y) P(x, dy) = \int_{\mathcal{X}} \bar{f}(y) P(x, dy) = T\bar{f}(x).$$

Where in the second-last equality we used $P(x, A) = 1$ for $x \in A$. This concludes that T^0 is Feller. \square

Motivation. Could we use a P -invariant set for uniqueness? 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 $\mathcal{P}(\mathcal{A})$ (i.e. has support in \mathcal{A}). Such criteria may be satisfied if invariant set $\mathcal{A} \subset \mathcal{X}$ is sufficiently absorbing, as formalised in Proposition 8.7.7.

Given a P -invariant set \mathcal{A} , consider the sequence \mathcal{A}_n of sets recursively defined by

$$\begin{aligned} \mathcal{A}_0 &= \mathcal{A}, \\ \mathcal{A}_1 &= \{x \in \mathcal{X} : P(x, \mathcal{A}) > 0\}, & (\text{can enter } \mathcal{A} = \mathcal{A}_0) \\ \mathcal{A}_2 &= \{x \in \mathcal{X} : P(x, \mathcal{A}_1) > 0\}, & (\text{can enter } \mathcal{A}_1) \\ \mathcal{A}_{n+1} &= \{x \in \mathcal{X} : P(x, \mathcal{A}_n) > 0\}, & n \geq 0. \end{aligned} \tag{8.20}$$

Observe that $A_0 \subset \mathcal{A}_1$ since \mathcal{A}_0 is invariant. In fact, by induction

$$\mathcal{A}_0 \subset \mathcal{A}_1 \subset \mathcal{A}_2 \subset \dots$$

Since if we assume that $\mathcal{A}_{n-1} \subset \mathcal{A}_n$, and by the definition we have $P(x, \mathcal{A}_{n-1}) > 0$, then

$$P(x, \mathcal{A}_n) \geq P(x, \mathcal{A}_{n-1}) > 0, \quad \forall x \in \mathcal{A}_n. \quad (8.21)$$

So that for all $x \in \mathcal{A}_n$, we must have $x \in \mathcal{A}_{n+1}$ by definition.

With these definitions, we have

Lemma 8.7.6 *Let $A \subset \mathcal{X}$ be P -invariant. For every $n \geq 1$, for any $x \in \mathcal{A}_n$, $P^n(x, \mathcal{A}) > 0$.*

Proof. The statement is true by definition 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

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

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 . \square

Proposition 8.7.7 *Let \mathcal{A} be an invariant set for P and let \mathcal{A}_n be defined as in (8.20). Suppose that $\bigcup_{n \geq 0} \mathcal{A}_n = \mathcal{X}$. Then every invariant probability measure π for P is in $\mathcal{P}(\mathcal{A})$, i.e. is an invariant probability measure for P on \mathcal{A} (i.e. $\pi(\mathcal{A}) = 1$).*

Proof. Assume we have $\pi \in \mathcal{P}(\mathcal{X})$ with $T\pi = \pi$. Suppose (for a contradiction) that $\pi(\mathcal{A}) < 1$. Since $\pi(\bigcup_{n \geq 0} \mathcal{A}_n) = \pi(\mathcal{X}) = 1$, by the assumption, $\lim_{n \rightarrow \infty} \pi(\mathcal{A}_n) = 1$. There must exist $n > 0$ such that $\pi(\mathcal{A}_n \setminus \mathcal{A}) > 0$. Since $T^{n_0}\pi = \pi$ by the invariance of π , this implies that

$$\begin{aligned} \pi(\mathcal{A}) &= T^{n_0}\pi(\mathcal{A}) = \int_{\mathcal{X}} P^{n_0}(x, \mathcal{A}) \pi(dx) \geq \int_{\mathcal{A}_{n_0}} P^{n_0}(x, \mathcal{A}) \pi(dx) \\ &\geq \int_{\mathcal{A}} P^{n_0}(x, \mathcal{A}) \pi(dx) + \int_{\mathcal{A}_{n_0} \setminus \mathcal{A}} P^{n_0}(x, \mathcal{A}) \pi(dx) > \pi(\mathcal{A}), \end{aligned}$$

where the last inequality follows from the fact that

$$\int_{\mathcal{A}} P^{n_0}(x, \mathcal{A}) \pi(dx) = \int_{\mathcal{A}} \pi(dx) = \pi(\mathcal{A}) \quad \text{and} \quad \int_{\mathcal{A}_{n_0} \setminus \mathcal{A}} P^{n_0}(x, \mathcal{A}) \pi(dx) > 0$$

(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 $P^{n_0}(x, \mathcal{A}) > 0$ for every $x \in \mathcal{A}_{n_0}$, c.f. (8.21). This is a contradiction, so that one must have $\pi(\mathcal{A}) = 1$. \square

Combining Theorem 8.7.5, with deterministic contraction Theorem 8.5.5 (Lec 18), we obtain:

Corollary 8.7.8 *Suppose that \mathcal{A} is a P -invariant set, where P is a transition probability on \mathcal{X} with Feller property. If \mathcal{A} is compact, $\bigcup_{n \geq 0} \mathcal{A}_n = \mathcal{X}$ and $\exists \gamma < 1$ such that*

$$\mathbf{E}d(F(x, \xi_1), F(y, \xi_1)) \leq \gamma d(x, y) \quad \forall x, y \in \mathcal{A}. \quad (8.22)$$

Then there exists a unique invariant probability measure $\pi \in \mathcal{P}(\mathcal{X})$ for P .

Proof. The restriction P^0 of P to \mathcal{A} is also Feller (by Theorem 8.7.5). Since \mathcal{A} is compact, Theorem 8.4.2 gives the existence of an invariant probability measure for P^0 , then the deterministic contraction condition (8.22) leads to the uniqueness for P^0 (Theorem 8.5.5). The existence and uniqueness of invariant probability measure π for P on \mathcal{X} follows from Lemma 8.7.4 and Proposition 8.7.7 respectively. \square

Remark 8.7.9 To determine existence and uniqueness on \mathcal{A} one may also use any other available criterions (e.g. Lyapunov test function method, if \mathcal{A} not compact, and minorisation in place of (8.22)).

This marks the end of lecture 20 - Week 9.

8.7.1 ODEs and Random Dynamical Systems

In the next section we construct and consider an example of random dynamical system, for which we will determine existence and uniqueness of an invariant probability measure, and the P -invariant set on which is supported (see Proposition 8.7.16). Here we start by reviewing some ODE settings/results and how one may then construct a Markov Chain.

A Review of ODEs

Let $g : \mathbf{R}^n \rightarrow \mathbf{R}^n$ and $f : \mathbf{R} \rightarrow \mathbf{R}^n$ be measurable functions, later we will consider system of this form:

$$\begin{cases} \dot{x}(t) = g(x(t)) + f(t) \\ x(t_0) = x_0 \end{cases} \quad (8.23)$$

Definition 8.7.10 Given the settings above, by a solution to (8.23) we mean a continuous function $(x(t), t \in (a, b))$ such that

$$x(t) = x_0 + \int_{t_0}^t g(x(s)) ds + \int_{t_0}^t f(s) ds. \quad (8.24)$$

If g is locally Lipschitz continuous and f is continuous, then $\frac{dx}{dt}$ exists and $x(t)$ is continuously differentiable.

Proposition 8.7.11 *Consider the system (8.23), the following hold:*

1. Maximal solution. *If we assume that g is locally Lipschitz continuous and f is locally bounded, then for every initial point x , there exists a unique maximal (local) solution.*
2. Growth condition. *If furthermore $\exists C$ such that*

$$\langle x, g(x) \rangle \leq C(1 + |x|^2), \quad \forall x \in \mathbf{R}^n. \quad (8.25)$$

Then there is no explosion / equation is complete (i.e. has global + unique solution). The condition (8.25) is called one sided linear growth condition.

3. Flow property. *Suppose for every $x_0 = x \in \mathbf{R}^n$, $t_0 \in \mathbf{R}$, there exists a unique global solution to integral equation (8.24). Denote by $\varphi_{t_0, t}(x)$ the solution, then the following semigroup/flow property holds*

$$\varphi_{u, t}(x) = \varphi_{s, t}(\varphi_{u, s}(x)), \quad \text{for } u < s < t. \quad (8.26)$$

Notation: *we will use $\varphi_t(x) := \varphi_{0, t}(x)$, and also denote $x_t = \varphi_t(x_0)$.*

Proof. Part (1) is covered by Piccard's theorem (local version).

Part (2). We want to show that $|x(t)|^2$ is finite for any t . By the chain rule

$$\frac{d}{dt} |\varphi_t(x_0)|^2 = 2 \langle \varphi_t(x_0), \frac{d}{dt} \varphi_t(x_0) \rangle_{\mathbf{R}^n} = 2 \langle \varphi_t(x_0), g(\varphi_t(x_0)) + f(t) \rangle_{\mathbf{R}^n}.$$

We integrate both sides of the identity from 0 to t and also denote $x_t = \varphi_t(x_0)$ for simplicity, we obtain

$$\begin{aligned} |x_t|^2 &= |x_0|^2 + 2 \int_0^t \langle x_s, g(x_s) \rangle ds + 2 \int_0^t \langle f(s), x_s \rangle ds \\ &\leq |x_0|^2 + 2 \int_0^t c(1 + |x_s|^2) ds + \sup_{s \leq t} |f(s)|^2 + \int_0^t |x(s)|^2 ds. \end{aligned}$$

Re-arrange,

$$|x_t|^2 \leq |x_0|^2 + 2ct + (2c + 1) \int_0^t |x_s|^2 ds + \sup_{s \leq t} |f(s)|^2.$$

Hence by Gronwall's inequality,

$$|x_t|^2 \leq (|x_0|^2 + 2ct + \sup_{s \leq t} |f(s)|^2) e^{2ct+t}.$$

Part (3). The proof for the flow property is standard. □

Exercise 8.7.1 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 denote df the Jacobian of f , v_t solves

$$dv_t = df_{x_t}(v_t).$$

One may consider more general ODE systems, we state the following which may be used for system (8.29) in Theorem 8.7.13 below.

Proposition 8.7.12 *Let $G : \mathbf{R}_+ \times \mathbf{R}^d \rightarrow \mathbf{R}^d$ be continuous and global Lipschitz in space, i.e.*

$$|G(t, x) - G(t, y)| \leq K|x - y|, \quad \forall x, y \in \mathbf{R}^d, \quad \forall t.$$

Consider the system

$$\begin{cases} \dot{x}(t) = G(t, x(t)) \\ x(t_0) = x_0 \end{cases} \quad (8.27)$$

Then for any initial data $x_0 \in \mathbf{R}^d$, there exists a unique global solution which is differentiable in time and satisfies (8.27). In particular $x(t) = x_0 + \int_{t_0}^t G(s, x(s)) ds$.

*Also, if $\varphi_{t_0, t}(x)$ denotes the solution starting from x (i.e. $\varphi_{t_0, t}(x) = x + \int_{t_0}^t G(s, x(s)) ds$), then for any $t > 0$ the function $x \mapsto \varphi_{t_0, t}(x)$ is **differentiable**.*

Indeed, let $x_0 \in \mathbf{R}^n$ and U an open set containing x_0 and suppose that for $t \leq \delta$, the solutions $\varphi(t, x)$ are defined for every $x \in U$. Then from

$$\varphi(t, x) = x + \int_0^t f(\varphi(s, x)) ds + f(t), \quad \varphi(t, y) = y + \int_0^t f(\varphi(s, y)) ds + f(t),$$

we see that

$$\begin{aligned} |\varphi(t, x) - \varphi(t, y)| &\leq |x - y| + \int_0^t |f(\varphi(s, x)) - f(\varphi(s, y))| ds \\ &\leq |x - y| + K \int_0^t |\varphi(s, x) - \varphi(s, y)| ds. \end{aligned}$$

Thus, $|\varphi(t, x) - \varphi(t, y)| \leq K|x - y|e^{Kt}$.

We know go back to our system (8.23) of interest for later example.

Notation. If $g : \mathbf{R}^n \rightarrow \mathbf{R}^n$ is a differential function, we denote by $(Dg)(x)(v)$ or $(Dg)_x(v)$ its derivative at x in the direction of v . In components, if $g = (g_1, \dots, g_n)$ where $g_i : \mathbf{R}^n \rightarrow \mathbf{R}$, then for any $x, v \in \mathbf{R}^n$,

$$Dg_i(x)(v) = \sum_{k=1}^n \frac{\partial g_i}{\partial x_k}(x) v_k.$$

Then

$$Dg(x)(v) = ((Dg_1)(x)(v), \dots, (Dg_n)(x)(v)) = J(x)v,$$

where $J(x)$ is the Jacobian of g at x :

$$J(x) = \begin{pmatrix} \frac{\partial g_1}{\partial x_1}(x) & \dots & \frac{\partial g_1}{\partial x_n}(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial g_n}{\partial x_1}(x) & \dots & \frac{\partial g_n}{\partial x_n}(x) \end{pmatrix}.$$

We denote: $|Dg|_x = \sup_{|v|=1, v \in \mathbf{R}^n} |(Dg)_x(v)|$. Let $|Dg|_\infty = \sup_x |Dg|_x$.

Theorem 8.7.13 *Suppose the functions $g : \mathbf{R}^n \rightarrow \mathbf{R}^n$ and $f : \mathbf{R} \rightarrow \mathbf{R}^n$ satisfy the following:*

$$f \text{ bounded measurable,} \quad g \in C^1 \quad \text{and} \quad |g(x) - g(y)| \leq K|x - y|, \quad \forall x, y \in \mathbf{R}^n.$$

Then the solution (8.24) is unique and global. Furthermore, we have $(t, x) \mapsto \varphi_t(x)$ continuous and, for any fixed t , $x \mapsto \varphi_t(x)$ differentiable.

Let

$$v(t) := (D\varphi_t)_{x_0}(v_0), \quad \text{for } x_0, v_0 \in \mathbf{R}^n. \quad (8.28)$$

Then $v(t)$ is the solution to the time dependent linear equation

$$\begin{cases} \dot{v}(t) = (Dg)_{\varphi_t(x_0)}(v(t)), \\ v(0) = v_0. \end{cases} \quad (8.29)$$

Note, the function $(x, v) \mapsto (Dg)_x(v)$ is continuous (linear in v) and Lipschitz (by Proposition 8.7.12 applied to (8.29)).

Corollary 8.7.14 *Assume the conditions of Theorem 8.7.13. If for all $v \in \mathbf{R}^n$,*

$$\langle (Dg)_x(v), v \rangle \leq -c(x)|v|^2. \quad (8.30)$$

Then, letting $v_t = (D\varphi_t)_{\varphi_t(x)}(v)$ (i.e. $v(t)$ in (8.28)), for any starting $x, v \in \mathbf{R}^n$ we have

$$|v_t| \leq e^{-\int_0^t c(\varphi_s(x)) ds}. \quad (8.31)$$

Proof. In view of the equation (8.29) satisfied by v_t and assumption (8.30), we can estimate

$$\frac{d}{dt}|v_t|^2 = 2\langle v_t, \frac{d}{dt}v_t \rangle = 2\langle v_t, (Dg)_{\varphi_t(x)}(v_t) \rangle \leq -2c(\varphi_t(x))|v_t|^2.$$

This implies that $|v_t|^2 \leq e^{-2\int_0^t c(\varphi_s(x)) ds}$, hence (8.31) follows. \square

Remark 8.7.15 If we also have that for any $x \in \mathbf{R}^n$, $c(x) \geq c > 0$ with positive constant c . Then we may use the following to derive contraction of the system

$$|\varphi_t(x) - \varphi_t(y)| \leq |D\varphi_t|_\infty |x - y| \leq e^{-ct} |x - y|. \quad (8.32)$$

Example 8.7.3 Consider $\dot{x}(t) = -x(t) + f(t)$ (a special case of system above). Then, by looking at equation (8.28) in this case, we have

$$\dot{v}(t) = -v(t), \quad v(t) = v(0)e^{-t}.$$

Hence we have the following contraction $|\varphi_t(x) - \varphi_t(y)| \leq e^{-t} |x - y|$.

Construction of Markov Chains from Random Differential Equations

Let us conclude this section by a complete treatment of an example of random dynamical system.

Settings and notation. We are going to vary f (in (8.23)) along the dynamics, we denote by $\varphi_t(x_0, f)$ the solution to

$$\begin{cases} \dot{x}(t) = g(x(t)) + f(t), \\ x(0) = x. \end{cases}$$

We assume that for any initial data x_0 , there exists a unique global solution (in the sense of Definition 8.7.10). Let us consider the solution at time $t = 1$ and denote it by Φ :

$$\Phi(x, f) = \varphi_1(x, f). \quad (8.33)$$

Then we shall consider $\{\xi_n\}$ continuous iid stochastic processes, which will be the f contribution in the ODE at each step, i.e.

$$\begin{cases} \dot{x}(t) = g(x(t)) + \xi_n(t, \omega), \\ x(0) = x. \end{cases}$$

We extract from it a discrete time dynamics as follows:

$$x_0 := x, \quad x_1 = \Phi(x, \xi_1), \quad \dots \quad x_n = \Phi(x_{n-1}, \xi_n). \quad (8.34)$$

The process (x_n) is a Markov chain.

8.7.2 Example

We will focus on the following example.

Proposition 8.7.16 *Let $\{\xi_n\}$ be a sequence of i.i.d. $\mathcal{C}([0, 1], \mathbf{R})$ -valued random variables such that $\sup_{t \in [0, 1]} |\xi_n(t)| \leq 1$ almost surely. Let $\varphi_t(x, f)$ be the solution of the differential equation on $(0, \infty)$ of*

$$\begin{cases} \frac{d}{dt}x(t) = \frac{1}{x(t)} - 2 + f(t), \\ x(0) = x. \end{cases} \quad (8.35)$$

Let $\Phi(x, f) = \varphi_1(x, f)$ and then define (x_n) by setting $x_0 = x$, $x_1 = \Phi(x, \xi_1)$ and recursively:

$$x_{n+1} = \Phi(x_n, \xi_{n+1}), \quad n \geq 0.$$

Then the Markov chain (x_n) has a unique invariant probability measure π on $(0, \infty)$. Furthermore, π satisfies $\pi([\frac{1}{3}, 1]) = 1$.

Preliminaries. We have the existence and uniqueness of a solution to

$$\dot{x}(t) = g(x(t)) + f(t),$$

for g Lipschitz continuous (see 1. in Proposition 8.7.11).

In our case $g(x) = \frac{1}{x}$ is not locally Lipschitz on \mathbf{R} , but it is on $\mathcal{X} = (0, \infty)$. We need to check that the evolution is well-defined within \mathcal{X} , i.e. we need to show solution $\varphi_t(x, f)$ of $\dot{x}(t) = \frac{1}{x(t)} - 2 + f(t)$ starting with $x(0) = x > 0$ remains positive.

Since $f \geq -1$, by the comparison theorem for ODEs, it is sufficient to consider

$$\dot{y}(t) = \frac{1}{y(t)} - 3, \quad y(0) = x > 0. \quad (8.36)$$

Then $x(t) \geq y(t)$ for all t . But the velocity $\frac{1}{y(t)} - 3 > 0$ on $(0, \frac{1}{3})$, hence $y(t) > 0$ for all time.

Remark 8.7.17 Consider the simpler equation

$$\dot{z}(t) = \frac{1}{z(t)} - 1, \quad z(0) = x > 0, \quad (8.37)$$

which is the equation (8.35) for $x(t)$ with $f \equiv 1$. Note that $z(t) \equiv 1$ is a solution (in fact a stable fixed point). Then one may check that the solution satisfies

$$z(t) + \log |z(t) - 1| = c - t, \quad (8.38)$$

where $c = z(0) + \log |1 - z(0)|$, for $z(0) \neq 1$. This does not seem to help with understanding of the solution and our Markov chain asymptotics.²

Proof. In our settings $f \in C([0, 1]; \mathbf{R})$, recall that we denote by $\varphi_t(x, f)$ the solution to

$$\frac{dx}{dt} = \frac{1}{x(t)} - 2 + f(t), \quad x(0) = x.$$

So that $\varphi_0(x, f) = x$ and solves the differential equation. Let $\Phi(x, f) := \varphi_1(x, f)$, then our Markov chain is defined by $x_{n+1} = \Phi(x_n, \xi_{n+1})$, where $\xi_i \in [-1, 1]$, *a.s.* (we omit the dependence on ω) and we define the following extremal (deterministic) maps

$$\Phi_+(x) := \Phi(x, 1), \quad \Phi_-(x) := \Phi(x, -1).$$

By comparison we have

$$x_{n+1} \in [\Phi_-(x_n), \Phi_+(x_n)], \quad a.s. \quad (8.39)$$

By analysing the equations (8.36)-(8.37), for initial data $x > 0$ we have the following:

²Note however that taking exponential of (8.38), if $z(0) \neq 1$, we obtain the relation $|z(t) - 1|e^{z(t)} = e^{c-t} \rightarrow 0$ as $t \rightarrow \infty$. This then suggests that we necessarily have $|z(t) - 1| \xrightarrow{t \rightarrow \infty} 0$, hence $z^* = 1$ is an attracting fixed point (unique limit point).

- $\dot{z}(t) = \frac{1}{z(t)} - 1$, then $z(t) \equiv 1$ is an equilibrium solution, and

$$\lim_{t \rightarrow \infty} \varphi_t(x, 1) = \lim_{t \rightarrow \infty} z(t) = 1.$$

- $\dot{y}(t) = \frac{1}{y(t)} - 3$, then $y(t) \equiv \frac{1}{3}$ is an equilibrium solution, and

$$\lim_{t \rightarrow \infty} \varphi_t(x, -1) = \lim_{t \rightarrow \infty} y(t) = \frac{1}{3}.$$

P-invariance and absorbing property. Fix $\epsilon > 0$ small, let $A = [\frac{1}{3} - \epsilon, 1 + \epsilon]$, we will show that this is P -invariant. Note the following

$$(\Phi_-)^{-1}(\frac{1}{3} - \epsilon) < \frac{1}{3} - \epsilon, \quad (\Phi_+)^{-1}(1 + \epsilon) > 1 + \epsilon.$$

Then we have $A \subset [(\Phi_-)^{-1}(\frac{1}{3} - \epsilon), (\Phi_+)^{-1}(1 + \epsilon)]$, which implies that (in view of (8.39)):

$$x_0 \in [(\Phi_-)^{-1}(\frac{1}{3} - \epsilon), (\Phi_+)^{-1}(1 + \epsilon)] \implies x_1 \in [\frac{1}{3} - \epsilon, 1 + \epsilon] = A.$$

Hence A is P -invariant. Now define $a_+^n = (\Phi_+)^{-n}(1 + \epsilon)$ and $a_-^n = (\Phi_-)^{-n}(\frac{1}{3} - \epsilon)$. Recalling Definition 8.20, we have $A_{n+1} := \{x \in \mathcal{X} : P(x, A_n) > 0\}$, then iterating the above argument we have

$$[a_-^n, a_+^n] \subset A_n \quad \text{and} \quad \bigcup_n A_n = \mathcal{X} = (0, \infty). \quad (8.40)$$

This holds for all ϵ sufficiently small. To check this one uses $\lim_{n \rightarrow \infty} \varphi_n(x, -1) = \frac{1}{3}$ and $\lim_{n \rightarrow \infty} \varphi_n(x, 1) = 1$ and the flow property e.g. $\varphi_2(x, 1) = \Phi_+ \circ \Phi_+(x)$. Then, by Proposition 8.7.7, any invariant probability measure π has $\pi([\frac{1}{3} - \epsilon, 1 + \epsilon]) = 1$, for any $\epsilon > 0$, hence $\pi([\frac{1}{3}, 1]) = 1$.

Feller and Contraction. Through examining the associated derivative flow, we will establish deterministic contraction. Let $v(t)$ denote the derivative of $\varphi_t(x, f)$ w.r.t. x . Then this solves

$$\frac{d}{dt}v(t) = -\frac{1}{x^2(t)}v(t), \quad v(0) = 1.$$

The solution is given by

$$v(t) = \exp\left(-\int_0^t \frac{ds}{x^2(s)}\right).$$

Since $\Phi(x, \xi_i)$ is differentiable in x , hence continuous in x , the process (x_n) is Feller by Theorem 8.4.8. Since $A = [\frac{1}{3}, 1]$ is compact, there exists an invariant probability measure on A by Corollary 8.4.2 (Krylov-Bogoliubov criterion). Also

$$\Phi'(x_n, \xi_{n+1}) = e^{-\int_0^1 \frac{ds}{x^2(s)}} \leq e^{\frac{1}{1+\epsilon}} < 1.$$

This implies that the map is a contraction

$$\mathbf{E}|\Phi(x, \xi_n) - \Phi(y, \xi_n)| \leq e^{\frac{1}{1+\epsilon}}|x - y|.$$

Then uniqueness holds on A by Theorem 8.5.5. Given P -invariance of A and $\bigcup_n A_n = \mathcal{X}$, we see that the Markov chain (x_n) on $(0, \infty)$ has a unique invariant probability measure. \square

This marks the end of lecture 21 - Week 9.

Chapter 9

The Structure Theorem and Ergodic Theorem (Mastery Material)

A stationary time homogenous Markov chain induces a dynamical system. With Birkhoff's ergodic theorem we can state and prove a structure theorem of invariant probability measures.

9.1 Ergodic theory for dynamical systems

In this small section we introduce/recall some core notions of dynamical systems, these will connect to stationary Markov chains viewed on the canonical path space $\mathcal{X}^{\mathbb{N}}$ or two-sided path space $\mathcal{X}^{\mathbb{Z}}$.

Definition 9.1.1 A **dynamical system** consists of a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a measure preserving measurable map $\theta: \Omega \rightarrow \Omega$, *i.e.* a map such that $\mathbb{P}(\theta^{-1}(A)) = \mathbb{P}(A)$ for every $A \in \mathcal{F}$ (*i.e.* $\theta_*\mathbb{P} = \mathbb{P}$).

We will denote by \mathbf{E} expectations with respect to \mathbb{P} as usual. In the following parts we will be interested in the sets invariant under the transformation θ on Ω .

Definition 9.1.2 Given a measurable transformation θ on $(\Omega, \mathcal{F}, \mathbb{P})$, a set with $\theta^{-1}(A) = A$ is called an invariant set for θ (or θ -invariant). Then the invariant σ -algebra $\mathcal{I} \subset \mathcal{F}$ is defined as

$$\mathcal{I} = \{A \in \mathcal{F} : \theta^{-1}(A) = A\}.$$

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 9.1.3 A measurable function $f : \Omega \rightarrow \mathbf{R}$ is said to be θ -invariant (or simply invariant) if $f \circ \theta = f$.

Note. Let $f = \mathbf{1}_A$, then f is invariant iff A is invariant (w.r.t. θ), i.e.

$$f \circ \theta = f \iff \mathbf{1}_A = \mathbf{1}_{\{\omega : \theta\omega \in A\}} = \mathbf{1}_{\theta^{-1}A}.$$

Exercise 9.1.1 Let $f : \Omega \rightarrow \mathbf{R}$ 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} .

Definition 9.1.4 Given a dynamical system $(\Omega, \mathcal{F}, \mathbb{P})$ and θ . We say θ is ergodic if any θ -invariant set has either measure 0 or measure 1. Note that this is a property of the map θ as well as of the measure \mathbb{P} . We also say \mathbb{P} is ergodic (w.r.t. θ).

Proposition 9.1.5 *The following statements are equivalent.*

1. \mathbb{P} is ergodic (θ is ergodic);
2. Every invariant integrable function f is almost surely a constant.
3. Every invariant bounded function is almost surely a constant.

Proof. From (2) to (3) is trivial. It remains to show (3) \Rightarrow (1), and (1) \Rightarrow (2).

(3) \Rightarrow (1). Assume that (3) holds. Let $f = \mathbf{1}_A$ where A is an invariant set. Then $\mathbf{1}_A$ is invariant and $\mathbf{1}_A = 1$ or 0 a.e., hence $\mathbf{1}_A = \mathbb{P}(A) \in \{0, 1\}$ and \mathbb{P} is ergodic.

(1) \Rightarrow (2). Suppose that \mathbb{P} is ergodic, i.e. $\mathbb{P}(A) = 1$ or 0 for any $A \in \mathcal{I}$. Let function f be integrable and invariant, then f is measurable with respect to \mathcal{I} .¹ We prove that $f = \mathbf{E}f$ a.e. . Note that the following sets

$$A_+ = \{\omega \in \Omega \mid f(\omega) > \mathbf{E}f\}, \quad A_- = \{\omega \in \Omega \mid f(\omega) < \mathbf{E}f\}, \quad A_0 = \{\omega \in \Omega \mid f(\omega) = \mathbf{E}f\},$$

are invariant sets and form a partition of Ω . Therefore, by ergodicity, exactly one of them has measure 1 and the other two must have measure 0. Suppose $\mathbb{P}(A_+) = 1$, then

$$0 = \int_{\Omega} (f - \mathbf{E}f) d\mathbb{P} = \int_{A_+} (f - \mathbf{E}f) d\mathbb{P}.$$

Then $f - \mathbf{E}f = 0$ a.s. on A_+ , which is a contradiction. Similarly if $\mathbb{P}(A_-) = 1$, we also have $f = \mathbf{E}f$ a.e., hence we must have $\mathbb{P}(A_0) = 1$. \square

¹See Exercise 3 of Problem Sheet 8.

Theorem 9.1.6 (Birkhoff's Ergodic Theorem) *Let $(\Omega, \mathcal{F}, \mathbb{P}, \theta, \mathcal{I})$ be as above and let $f: \Omega \rightarrow \mathbf{R}$ be such that $\mathbf{E}|f| < \infty$. Then,*

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

almost surely.

Note that if f is invariant, both sides are equal to $f(\omega)$.

The limit function $\mathbf{E}(f | \mathcal{I})$ in Birkhoff's ergodic theorem is \mathcal{I} -measurable. Hence ergodicity of the dynamical system implies that $\mathbf{E}(f | \mathcal{I})$ is a.e. a constant. This leads to the following corollary.

Corollary 9.1.7 *If the dynamical system in Theorem 9.1.6 is ergodic, then*

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=0}^{N-1} f(\theta^n \omega) = \mathbf{E}f, \quad a.s.$$

9.2 Dynamical Systems induced by Markov chains

9.2.1 Sequence spaces and the shift operator θ

If we have a semi-infinite sequence (a_0, a_1, a_2, \dots) with $a_i \in \mathcal{X}$, we define the shift operator

$$\theta(a_0, a_1, a_2, \dots) = (a_1, a_2, a_3, \dots). \quad (9.1)$$

Similarly θ can be defined on $\mathcal{X}^{\mathbf{Z}}$ in the same way. Note that we may use the inverse θ^{-1} instead of θ .

Examples of invariant sets. Recall Definition 9.1.2 of an invariant set, the following are examples of invariant sets (w.r.t. θ or θ^{-1} defined above) on sequence spaces $\mathcal{X}^{\mathbf{Z}}$ or $\mathcal{X}^{\mathbf{N}}$.

1. Any constant sequence (a, a, a, \dots) with $a \in \mathcal{X}$, is an invariant set.
2. Similarly, for $A \in \mathcal{B}(\mathcal{X})$, the following are invariant sets in $\mathcal{X}^{\mathbf{N}}$ and $\mathcal{X}^{\mathbf{Z}}$ respectively,

$$A \times A \times \dots \subset \mathcal{X}^{\mathbf{N}}, \quad \dots \times A \times A \times A \times \dots \subset \mathcal{X}^{\mathbf{Z}}.$$

3. The set $\{\underline{a}, \underline{b}, \underline{c}\} \in \mathcal{B}(\mathcal{X}^{\mathbf{Z}})$ is an invariant set, composed of the following sequences:

$$\begin{aligned} \underline{a} &= (\dots, a, b, c, a, b, c, \dots), \\ \underline{b} &= (\dots, b, c, a, b, c, a, \dots), \\ \underline{c} &= (\dots, c, a, b, c, a, b, \dots), \end{aligned} \quad a, b, c \in \mathcal{X}. \quad (9.2)$$

4. Similarly, we can have set-valued sequences, then the following is an invariant set on $\mathcal{X}^{\mathbf{N}}$:

$$\{A \times B \times C \times A \times B \times C \times \dots\} \cup \{B \times C \times A \times B \times C \times A \times \dots\} \cup \{C \times A \times B \times C \times A \times B \times \dots\}.$$

5. The set $\{(a_n) : \exists N \text{ s.t. } a_n \in B, \forall n \geq N\}$ where $B \in \mathcal{B}(\mathcal{X})$, is a θ -invariant sub set of $\mathcal{X}^{\mathbf{Z}}$.

Example 9.2.1 Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let (x_n) be a Markov chain with state space \mathcal{X} countable and transition matrix $P = (P_{ij})$.

a. Let $\varphi : \mathcal{X}^{\mathbf{N}} \rightarrow \mathbf{R}$ be a bounded invariant function (i.e. $\varphi \circ \theta = \varphi$). Define

$$Y = \varphi(x_0, x_1, \dots).$$

By the invariance, for every $n \geq 1$,

$$Y = \varphi(x_0, x_1, \dots) = \varphi \circ \theta^n(x_0, x_1, \dots) = \varphi(x_n, x_{n+1}, \dots), \quad \forall n.$$

Let us define $f(j) = \mathbf{E}(Y | x_0 = j)$. Then

$$\begin{aligned} f(j) &= \mathbf{E}(\mathbf{E}(Y | x_0, x_1) | x_0 = j) = \mathbf{E}(\mathbf{E}(\varphi(x_1, x_2, \dots) | x_0, x_1) | x_0 = j) \\ &= \mathbf{E}(\mathbf{E}(\varphi(x_1, x_2, \dots) | x_1) | x_0 = j) \\ &= \mathbf{E}(f(x_1) | x_0 = j) = Pf(j). \end{aligned}$$

We have used consecutively the tower property, the invariant property of φ , and the Markov property of (x_n) . This means $f = Pf$, i.e.

$$f(j) = \sum_k P_{jk} f(k).$$

** In fact, f is a ‘harmonic function’ and $f(x_n)$ is a ‘martingale’.

b. Given a measurable set $B \subset \mathcal{X}$, we can also take $\varphi = \mathbf{1}_{\hat{B}}$, where

$$\hat{B} = \{x. : (x_k)_{k \geq n} \text{ eventually will be in a set } B\}.$$

Example 9.2.2 Now we suppose that $\mathcal{X} = C_0 \cup \bigcup_{k=1}^M C_k$, 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 $f(j) = \mathbb{P}(B | x_0 = j)$, this is the probability that x_n from j eventually lands in C_1 . We may then solve the equations

$$f(j) = \sum_k P_{jk} f(k)$$

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

$$g(j) = \mathbb{P}(x_n \in C_0 \text{ for all } n | x_0 = j),$$

are not all zero. We seek for the minimal solution.

9.2.2 Stationary Markov chains as dynamical systems

To obtain a dynamical system on the sequence space (either $\mathcal{X}^{\mathbf{N}}$ or $\mathcal{X}^{\mathbf{Z}}$), we need a measure \mathbb{P} which is θ -invariant (i.e. $\theta^*\mathbb{P} = \mathbb{P}$).

Example 9.2.3 (*Stationary measures on sequence spaces*)

1. The stationary measure, \mathbb{P}_π , induced on $\mathcal{X}^{\mathbf{N}}$ by a Markov chain with transition probabilities P and with initial distribution an invariant probability measure π , is θ -invariant.
2. Similarly, the stationary measure on \mathbb{P}_π on $\mathcal{X}^{\mathbf{Z}}$ induced by a family of one step transition probabilities P and an initial probability distribution π invariant for P , is θ -invariant. The measure \mathbb{P}_π is the probability distribution of a two sided stationary Markov process, which we introduce later in section 9.2.3.

From now on let θ be the shift operator on $\mathcal{X}^{\mathbf{Z}}$, i.e. $\theta(x)(n) = x(n+1)$, so that

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

and we write $\theta = \theta_1$ and $\theta^{-1} = \theta_{-1}$. As in previous section, we denote by \mathcal{I} the set of all measurable subsets of $\mathcal{X}^{\mathbf{Z}}$ that are invariant under θ ,

$$\mathcal{I} = \{C \in \mathcal{B}(\mathcal{X}^{\mathbf{Z}}) : \theta^{-1}C = C\}.$$

Also let $P = (P(x, \cdot), x \in \mathcal{X})$ be a family of transition probabilities and a probability measure $\pi \in \mathbb{P}(\mathcal{X})$ satisfying $\pi = \int_{\mathcal{X}} P(x, \cdot) \pi(dx)$.

By the definition of stationarity, one has:

Lemma 9.2.1 *The triple $(\mathcal{X}^{\mathbf{Z}}, \mathcal{B}(\mathcal{X}^{\mathbf{Z}}), \mathbb{P}_\pi, \theta)$ defines a dynamical system, and θ is continuous (This is called a continuous dynamical system).*

Proof. It is clear that θ is continuous (with respect to the product topology). The product topology is the coarsest topology such that each projection map $\pi_i : \Pi\mathcal{X} \rightarrow \mathcal{X}$ is continuous. We only need to test with open sets of the form $\pi_i^{-1}(U)$. It is clear that $\theta^{-1}(\pi_i^{-1}(U))$ is an open set. We have already seen that θ is \mathbb{P}_π -invariant (see Lemma 9.2.7 for details). \square

Let d denote the metric on \mathcal{X} then $\varrho((a_n), (b_n)) = \sum_{n=1}^{\infty} \frac{1}{2^n} \frac{d(a_n, b_n)}{1+d(a_n, b_n)}$ is a metric inducing the product topology.

Remark 9.2.2 Given a family of one step transition probabilities $(P(x, \cdot), x \in \mathcal{X})$ with an invariant $\pi \in \mathcal{P}(\mathcal{X})$, working on $\mathcal{X}^{\mathbf{Z}}$ leads to stronger results than on $\mathcal{X}^{\mathbf{N}}$. Think along the lines of Birkhoff's ergodic Theorem 9.1.6, the collection of functions $\{f : \mathcal{X}^{\mathbf{N}} \rightarrow \mathbf{R}\}$ contains less information than the collections $\{f : \mathcal{X}^{\mathbf{Z}} \rightarrow \mathbf{R}\}$.

Remember that the measure \mathbb{P}_π is **ergodic** if every $A \in \mathcal{I}$ has $\mathbb{P}_\pi(A) \in \{0, 1\}$.

Definition 9.2.3 We say that an invariant measure π of a Markov process with associated transition semigroup T is **ergodic** if the corresponding measure \mathbb{P}_π is ergodic for θ .

Recall that a measurable subset \bar{A} of \mathcal{X} is said to be P -invariant if $P(x, \bar{A}) = 1$ for all $x \in \bar{A}$. Then the θ -invariant set $\Pi_{i=0}^\infty \bar{A}$ has measure

$$\mathbb{P}_\pi(\Pi_{i=0}^\infty \bar{A}) = \pi(\bar{A}). \quad (9.3)$$

This is the content of Remark 8.7.2. Examining the proof for Remark 8.7.3, we see the statement (9.3) holds if the invariance is relaxed to hold for almost surely starting point x from \bar{A} , the almost sure property is with respect to an invariant probability measure π . This prompts the following definition.

Definition 9.2.4 A measurable subset \bar{A} of \mathcal{X} is said to be π -invariant if $P(x, \bar{A}) = 1$ for π -a.s. every $x \in \bar{A}$.

Question. If \mathbb{P}_π is ergodic, then $\mathbb{P}_\pi(\Pi_{i=0}^\infty \bar{A}) \in \{0, 1\}$. If furthermore \bar{A} is π -invariant, $\pi(\bar{A}) \in \{0, 1\}$. How about the other way around? To understand the structure of invariant probability measures, we are going to relate shift-invariant subsets of $\mathcal{X}^{\mathbb{Z}}$ with π -invariant subsets of \mathcal{X} .

Before closing this section, we state Birkhoff's ergodic Theorem 9.1.6 for the dynamical system $(\mathcal{X}^{\mathbb{Z}}, \mathcal{B}(\mathcal{X}^{\mathbb{Z}}), \mathbb{P}_\pi, \theta)$ (analogous theorem holds also for the chain on $\mathcal{X}^{\mathbb{N}}$) applied to functions of one time:

Corollary 9.2.5 Let $f : \mathcal{X} \rightarrow \mathbf{R}$ be integrable and define $\tilde{f} : \mathcal{X}^{\mathbb{Z}} \rightarrow \mathbf{R}$ by $\tilde{f}((a_0, a_1, \dots)) := f(a_0)$. Then $\tilde{f}(\theta^n a) = f(a_n)$, so that we have

$$\frac{1}{n} \sum_{k=1}^n f(a_k) \xrightarrow{n \rightarrow \infty} \mathbf{E}_{\mathbb{P}_\pi}(\tilde{f} | \mathcal{I}) \quad \mathbb{P}_\pi - a.s. \quad (9.4)$$

If π is ergodic (as in Definition 9.2.3), then

$$\frac{1}{n} \sum_{k=1}^n f(a_k) \xrightarrow{n \rightarrow \infty} \int_{\mathcal{X}} f \, d\pi \quad \mathbb{P}_\pi - a.s. \quad (9.5)$$

Hence time average (LHS) is approximately equal to spatial average (RHS).

Just observe that $\mathbf{E}_{\mathbb{P}_\pi} \tilde{f} = \int_{\mathcal{X}} f \, d\pi$.

9.2.3 Construction of two sided stationary 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 $(x_n, n \in \mathbf{Z})$, which defines a probability measure \mathbb{P}_π on the space $\mathcal{X}^{\mathbf{Z}}$ of \mathcal{X} -valued sequences. Let $P = (P(x, \cdot), x \in \mathcal{X})$ denote the transition probabilities. Let $\pi \in \mathbb{P}(\mathcal{X})$ with $\pi = \int_{\mathcal{X}} P(x, \cdot) \pi(dx)$.

The finite dimensional distribution approach. We construct a probability measure \mathbb{P}_π on $\mathcal{X}^{\mathbf{Z}}$ by specifying its finite dimensional distributions and Kolmogorov's Theorem 4.4.2. The process $(x_n, n \in \mathbf{Z})$ with \mathbb{P}_π as its probability distribution is a stationary Markov process with t.p. P . Let $\mu_{n,m}$ denote the distribution of $(x_{-n}, \dots, x_{-1}, x_0, x_1, \dots, x_m)$ given by

$$P(z_{m-1}, dz_m) \cdots P(z_0, dz_1) P(z_{-1}, dz_0) \cdots P(z_{-n}, dz_{-n+1}) \pi(dz_{-n}) = \Pi_{k=-n}^{m-1} P(z_k, dz_{k+1}) \pi(dz_{-n}).$$

Then $\{\mu_{n,m}\}$ is a consistent family of probability measures. Therefore (through Theorem 4.4.2) defines \mathbb{P}_π on $\mathcal{X}^{\mathbf{Z}}$, and a (stationary) Markov chain $(x_n, n \in \mathbf{Z})$ with transition probabilities P and $\mathcal{L}(x_n) = \pi$ for any $n \in \mathbf{Z}$. See section 9.2.4 for details.

The time shift approach.

An alternative strategy is to start with a Markov process $(x_n, n \geq 0) \in \mathcal{X}^{\mathbf{N}}$, with invariant probability measure π as initial distribution, and push it back. Let $(y_n^{(m)}, n \geq -m)$, such that $(y_{-1}^{(1)}, y_0^{(1)}, y_1^{(1)}, \dots) = (x_0, x_1, x_2, \dots)$, and $y^{(n+1)}$ is obtained from $y^{(n)}$ in a similar manner. Then $y^{(m)}$ has a limit, this limit is the required two sided stationary process. Just need to check the finite dimensional distributions for these processes are eventually the same.

This marks the end of lecture 22 - Week 10.

9.2.4 Proof of two sided Markov chains construction

This is not given in the lectures. We begin by defining a probability measure on \mathcal{X}^k as follows. Given any positive number $n, m > 0$, we define a measure $\mathbb{P}_\pi^{n,m}$ (earlier denoted for brevity $\mu_{n,m}$) on \mathcal{X}^{n+m+1} in the following way. For $x = (x_{-n}, \dots, x_m)$,

$$\begin{aligned} \int_{\mathcal{X}^{n+m+1}} f(x_{-n}, \dots, x_m) \mathbb{P}_\pi^{n,m}(dx) \\ = \int_{\mathcal{X}} \cdots \int_{\mathcal{X}}^{\overbrace{n+m+1}} f(x_{m-1}, \dots, x_m) P(x_{m-1}, dx_m) \cdots P(x_{-n}, dx_{-n+1}) \pi(dx_{-n}). \end{aligned} \tag{9.6}$$

In addition, we define $\int_{\mathcal{X}} f(x_0) \mathbb{P}_\pi^{0,0}(dx) = \int_{\mathcal{X}} f(x_0) \pi(dx)$, 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) .

Note that

$$\mathbb{P}(x_n \in A) = \int_{\mathcal{X}} \cdots \int_{\mathcal{X}}^{\overbrace{\quad}^{n+m+1}} \mathbf{1}_A(x_{-n}) P(x_{m-1}, dx_m) \cdots P(x_{-n}, dx_{-n+1}) \pi(dx_{-n}) = \int_{\mathcal{X}} \mathbf{1}_A(x_{-n}) \pi(dx_{-n}),$$

so x_n is distributed as π for all n .

It's worth to have in mind that the canonical process on the measurable space $\mathcal{X}^{\mathbb{Z}}$ 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 $\mathbb{P}_{\pi}^{n,m}(dz)$ as the finite dimensional probability distribution of the two sided Markov chain, to be constructed.

Theorem 9.2.6 *Let P be transition probabilities with invariant π . Then the measures $\mathbb{P}^{n,m}$ defined by (9.6) are consistent and extends by Kolmogorov's theorem to a measure \mathbb{P}_{π} on $\mathcal{X}^{\mathbb{Z}}$. The corresponding Markov chain is called the two sided Markov chain associated with P and π .*

Proof. I do not go over this proof in class. It is an easy, although tedious, exercise to check that the family of measures on \mathcal{X}^{2n+1} defined by (9.6) is consistent, so that it defines a unique measure on $\mathcal{X}^{\mathbb{Z}}$ by Kolmogorov's extension theorem, Theorem 4.4.2. 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 $n \in \mathbb{Z}$,

$$\begin{aligned} \int_{\mathcal{X}^2} f(x_{n+1}) \mathbb{P}_{\pi}^{n,n+1}(dx) &\stackrel{\text{def}}{=} \int_{\mathcal{X}} \int_{\mathcal{X}} f(x_{n+1}) P(x_n, dx_{n+1}) \pi(dx_n) \stackrel{\text{invariance}}{=} \int_{\mathcal{X}} f(x_{n+1}) \pi(dx_{n+1}) \\ \int_{\mathcal{X}^2} f(x_n) \mathbb{P}_{\pi}^{n,n+1}(dx) &= \int_{\mathcal{X}} \left(f(x_n) \int_{\mathcal{X}} P(x_n, dx_{n+1}) \right) \pi(dx_n) = \int_{\mathcal{X}} f(x_n) \pi(dx_n). \end{aligned}$$

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 then move to multiple times:

$$\begin{aligned} \int_{\mathcal{X}^{n+m+2}} f(x_{-n}, \dots, x_m) \mathbb{P}_{\pi}^{n,m+1}(dx) \\ &= \int_{\mathcal{X}} \cdots \int_{\mathcal{X}}^{\overbrace{\quad}^{n+m+2}} f(x_{-n}, \dots, x_m) P(x_m, dx_{m+1}) P(x_{m-1}, dx_m) \cdots P(x_{-n+1}, dx_{-n}) \pi(dx_{-n}) \\ &= \int_{\mathcal{X}} \cdots \int_{\mathcal{X}}^{\overbrace{\quad}^{n+m+1}} \left(\int_{\mathcal{X}} P(x_m, dx_{m+1}) \right) f(x_{m-1}, \dots, x_m) P(x_{m-1}, dx_m) \cdots P(x_{-n+1}, dx_{-n}) \pi(dx_{-n}) \end{aligned}$$

$$= \int_{\mathcal{X}} \cdots \int_{\mathcal{X}}^{\overbrace{\quad}^{n+m+1}} f(x_{-n}, \dots, x_m) P(x_{m-1}, dx_m) \cdots P(x_{-n+1}, dx_{-n}) \pi(dx_{-n}).$$

Also,

$$\begin{aligned} & \int_{\mathcal{X}^{n+m+2}} f(x_{-n}, \dots, x_m) \mathbb{P}_{\pi}^{n+1, m}(dx) \\ &= \int_{\mathcal{X}} \cdots \int_{\mathcal{X}}^{\overbrace{\quad}^{n+m+2}} f(x_{-n}, \dots, x_m) P(x_{m-1}, dx_m) \cdots P(x_{-n+1}, dx_{-n}) P(x_{-n-1}, dx_{-n}) \pi(dx_{-n-1}) \\ &= \int_{\mathcal{X}} \int_{\mathcal{X}} \left(\int_{\mathcal{X}} \cdots \int_{\mathcal{X}}^{\overbrace{\quad}^{n+m}} f(x_{-n}, \dots, x_m) P(x_{m-1}, dx_m) \cdots P(x_{-n}, dx_{-n+1}) \right) P(x_{-n-1}, dx_{-n}) \pi(dx_{-n-1}) \\ &= \int_{\mathcal{X}} \left(\int_{\mathcal{X}} \cdots \int_{\mathcal{X}}^{\overbrace{\quad}^{n+m}} f(x_{-n}, \dots, x_m) P(x_{m-1}, dx_m) \cdots P(x_{-n}, dx_{-n+1}) \right) \pi(dx_{-n}), \end{aligned}$$

the consistency for this case then follows from Fubini's theorem. Further consistency relations will follow by inductions. \square

We have the following results:

Lemma 9.2.7 *The measure \mathbb{P}_{π} defined in Theorem 9.2.6 defines a stationary Markov process.*

Proof. Let (x_n) be the Markov process with probability distribution \mathbb{P}_{π} . By the construction, the finite dimensional projections of \mathbb{P}_{π} , to the $(-n, \dots, m)$ coordinates are

$$\mathbb{P}(x_n \in A_{-n}, \dots, x_m \in A_m) = \int_{A_{-n}} \cdots \int_{A_m} P(x_{m-1}, dx_m) \cdots P(x_{-n}, dx_{-n+1}) \pi(dx_{-n}).$$

This relation on the right hand side is the same for the coordinate maps $(-n+1, \dots, m+1)$. So $(x_n, n \in \mathbf{Z})$ is stationary. \square

In fact \mathbb{P}_{π} is invariant under both θ_1 and its inverse θ_{-1} .

The defining equation (9.6) is in principle the same as the following

$$\int f(x_{-n}, \dots, x_m) \mathbb{P}_{\pi}^{n, m}(dx) = \int_{\mathcal{X}} \cdots \int_{\mathcal{X}}^{\overbrace{\quad}^{n+m+1}} f(x_0, \dots, x_{2n}) P(x_{2n-1}, dx_{2n}) \cdots P(x_0, dx_1) \pi(dx_0),$$

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 trivial to see that $\mathbb{P}_{\pi}^{(n, m)} = \mathbb{P}_{\pi}^{(n+1, m+1)} = \mathbb{P}_{\pi}^{(n-1, m-1)}$: they are defined by the same relations.

9.2.5 Birkhoff's ergodic theorem for Markov Chains

Throughout this section $P = (P(x, \cdot), x \in \mathcal{X})$ is a family of transition probabilities with transition operator $T\mu(\cdot) = \int_{\mathcal{X}} P(x, \cdot)\mu(dx)$. Let

$$I_P = \{\pi \in \mathcal{P}(\mathcal{X}) : T\pi = \pi\}.$$

Let (x_n) denote a stationary THMC on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with t.p. P and initial distribution $\pi \in I_P$.

Let us expand Birkhoff's ergodic theorem a bit more, which hold for the dynamical system $(\mathcal{X}^{\mathbb{Z}}, \mathcal{B}(\mathcal{X}^{\mathbb{Z}}), \mathbb{P}_\pi, \theta)$. In particular we restate Birkhoff's theorem so the statement will be in terms of rather than \mathbb{P}_π a.e. sequences. Let $f : \mathcal{X}^{\mathbb{Z}} \rightarrow \mathbf{R}$ is $L^1(\mathbb{P}_\pi)$ and define

$$E := \left\{ a. \in \mathcal{X}^{\mathbb{Z}} : \frac{1}{n} \sum_{k=1}^n f(\theta^k a.) \xrightarrow{n \rightarrow \infty} \mathbf{E}_{\mathbb{P}_\pi}(f|\mathcal{I}) \right\}.$$

Then $\mathbb{P}_\pi(E) = 1$ by Birkoff's ergodic theorem 9.1.6. By definition

$$\mathbb{P}(\{\omega : x.(\omega) \in E\}) = \mathbb{P}_\pi(E) = 1.$$

Therefore, for any $\omega \in \Omega$ such that $x.(\omega) \in E$, the average $\frac{1}{n} \sum_{k=1}^n f(\theta^k x.(\omega))$ converges (\mathbb{P} -a.e. ω). This leads to the following equivalent statement:

Theorem 9.2.8 *Let $(x_n)_{n \in \mathbb{Z}}$ be a stationary Markov process with $x_0 \sim \pi$, where π is an invariant probability measure. Then the following hold:*

1. *For any integrable function $f : \mathcal{X}^{\mathbb{Z}} \rightarrow \mathbf{R}$, setting $\bar{f} = \mathbf{E}_{\mathbb{P}_\pi}(f|\mathcal{I})$, then*

$$\frac{1}{n} \sum_{k=1}^n f(\theta^k x.(\omega)) \xrightarrow{n \rightarrow \infty} \bar{f}(x.(\omega)), \quad \mathbb{P}\text{-a.e. } \omega.$$

2. *If furthermore π is ergodic, then*

$$\frac{1}{n} \sum_{k=1}^n f(\theta^k x.(\omega)) \xrightarrow{n \rightarrow \infty} \int_{\mathcal{X}^{\mathbb{Z}}} f d\mathbb{P}_\pi \quad \mathbb{P}\text{-a.e. } \omega.$$

We can also establish the result with a fixed starting point (depending on the support of π).

Theorem 9.2.9 *Let $P = P(x, \cdot)$ be a transition probability with an invariant probability measure π . Let $(x_n)_{n \in \mathbb{Z}}$ be a time homogeneous Markov process with t.p. P and initial position $x_0 = x$. Then for π -almost every $x \in \mathcal{X}$, the following statements hold:*

1. *For any integrable function $f : \mathcal{X}^{\mathbb{Z}} \rightarrow \mathbf{R}$,*

$$\frac{1}{n} \sum_{k=1}^n f(\theta^k x.(\omega)) \quad \text{converges for } \mathbb{P}\text{-a.e. } \omega.$$

2. If furthermore π is ergodic,

$$\frac{1}{n} \sum_{k=1}^n f(\theta^k x_0(\omega)) \xrightarrow{n \rightarrow \infty} \int_{\mathcal{X}} f d\mathbb{P}_\pi \quad \mathbb{P}\text{-a.e. } \omega.$$

Proof. There are many proofs for this, here we illustrate the use of stopping times. First let $x_0 \sim \pi$ (then the Markov chain with initial condition x_0 is stationary). By Theorem 9.2.8, we have

$$\frac{1}{n} \sum_{k=1}^n f(\theta^k x_0) \longrightarrow \bar{f}(x_0), \quad \mathbb{P}\text{-a.e. } \omega.$$

Then by the dominated convergence theorem

$$\mathbf{E} \left[\frac{1}{n} \sum_{k=1}^n f(\theta^k x_0) \mid \sigma(x_0) \right] \xrightarrow{n \rightarrow \infty} \mathbf{E}[\bar{f}(x_0) \mid \sigma(x_0)], \quad \mathbb{P}\text{-a.e. } \omega.$$

From this we deduce that for π -almost every x ,

$$\mathbf{E} \left[\frac{1}{n} \sum_{k=1}^n f(\theta^k x_0) \mid x_0 = x \right] \xrightarrow{n \rightarrow \infty} \mathbf{E}[\bar{f}(x) \mid x_0 = x], \quad \mathbb{P}\text{-a.e. } \omega.$$

This can be seen by testing the conditional expectation in the previous line with functions of the form $\varphi(x_0)$ and turn the expectation into integration with respect to x_0 . \square

Example 9.2.4 Let P be a transition probability with an ergodic invariant probability measure π . Let $g : \mathcal{X} \rightarrow \mathbf{R}$ in $L^1(\pi)$.

$$\frac{1}{n} \sum_{k=1}^n g(x_k) \xrightarrow{n \rightarrow \infty} \int_{\mathcal{X}} g d\pi \quad \mathbb{P}\text{-a.e. } \omega.$$

Proof. Define $\tilde{g} : \mathcal{X}^{\mathbf{Z}} \rightarrow \mathbf{R}$ by setting $\tilde{g}((y_i)) := g(y_0)$. Then we apply Theorem 9.2.9 with function \tilde{g} . The result follows, by noting $\int_{\mathcal{X}^{\mathbf{Z}}} \tilde{g} d\mathbb{P}_\pi = \int_{\mathcal{X}} g d\pi$. \square

Example 9.2.5 Suppose that a TMMC (x_n) starts with π , an invariant probability distribution, and with transition probabilities $P(x, dy)$. Let \mathbb{P}_π denote the invariant distribution on $\mathcal{X}^{\mathbf{N}}$. Then, for a bounded measurable function $g : \mathcal{X} \times \mathcal{X} \rightarrow \mathbf{R}$ $\int_{\mathcal{X}^{\mathbf{N}}} g(y_1, y_2) d\mathbb{P}_\pi = \int_{\mathcal{X}} \int_{\mathcal{X}} g(y_1, y_2) P(y_1, dy_2) \pi(dy_1)$ (Note $P_\pi(y_0 \in A_0, y_1 \in A_1) = \int_{\mathcal{X}} P(x, dy) \pi(dx)$.) Then one can work out a law of large numbers for sums of the form

$$\frac{1}{n} \sum_{k=1}^n g(x_k, x_{k+1}).$$

Proposition 9.2.10 Two ergodic invariant probability measures for a THMC are equal or mutually singular.

Proof. Let π_1 and π_2 be two distinct ergodic invariant probability measures. Let $f : \mathcal{X} \rightarrow \mathbf{R}$ be bounded measurable such that $\int_{\mathcal{X}} f d\pi_1 \neq \int_{\mathcal{X}} f d\pi_2$ (which exists).

Let (x_n) be a Markov chain with initial condition x . For $i = 1, 2$, let

$$E_i = \{x : x_0 = x, \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n f(x_k) = \int_{\mathcal{X}} f d\pi_i \quad \mathbb{P} - a.e.\}.$$

By Corollary 9.2.4, the limit exists and equals $\int_{\mathcal{X}} f d\pi_i$ for π_i -a.e. $x \in \mathcal{X}$. Then $\pi_1(E_1) = 1$ and $\pi_2(E_2) = 1$. But $E_1 \cap E_2 = \emptyset$, hence $\pi_1(E_2) = 0$ and π_1, π_2 are mutually singular. \square

9.3 Structure Theorem

In this section, we introduce a general structure theorem (Theorem 9.3.3) for Markov processes that gives us an overview of the set of invariant probability measures. Throughout the section P denotes a fixed t.p. with associated transition operator T and $I_P = \{\pi \in \mathcal{P}(\mathcal{X}) : T\pi = \pi\}$. If π_1 and π_2 are in I_P , then any of their convex combination is in I_P also (these are measures of the form $t\pi_1 + (1-t)\pi_2$ with $t \in [0, 1]$), i.e. I_P is 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 probability measures converging weakly to a limit π , one has

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

so that π is again an invariant probability measure for P . This shows that if T is Feller, then the set I_P is closed (in the topology of weak convergence).

Remark 9.3.1 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 $\delta_0 \notin \mathcal{I}(T)$.

Previously we have defined an invariant set \mathcal{A} by the property that $\mathbb{P}(x, \mathcal{A}) = 1$ for all x , this extends to the π -invariance for $\pi \in I_P(\mathcal{X})$: We say that a measurable set $\mathcal{A} \in \mathcal{B}(\mathcal{X})$ is **π -invariant** if $P(x, \mathcal{A}) = 1$ for π -almost every $x \in \mathcal{A}$. We will show in Corollary 9.3.12 that π is ergodic if and only if any π -invariant set has π measure 0 or 1. This will help us to conclude the proof of Theorem 9.3.3, for which we will also need the following definition:

Definition 9.3.2 A probability measure $\pi \in I_P$ is an extremal, of I_P , if π cannot be decomposed as $\pi = t\pi_1 + (1-t)\pi_2$ with $t \in (0, 1)$ and $\pi_i \in I_P$ are distinct.

9.3.1 The statements

The importance of invariant measures can be seen in the following structural theorem, which is a consequence of Birkhoff's ergodic theorem:

Theorem 9.3.3 *Given a time homogeneous transition probability P , with corresponding transition operator T . With I_P denoting the set of probability measures invariant w.r.t. P , set*

$$\mathcal{E} = \{ \pi \in \mathcal{P}(\mathcal{X}) : T\pi = \pi, \pi \text{ is ergodic} \} \subset I_P.$$

Then the following statements hold.

- (a) *The set I_P is convex and \mathcal{E} is precisely the set of its extremal points.*
- (b) *Any two ergodic invariant probability measures are either identical or mutually singular.*
- (c) *Furthermore, every invariant probability measure $\pi \in I_P$ is a convex combination of ergodic invariant probability measures, i.e. for every invariant measure $\mu \in \mathcal{I}$, there exists a probability measure \mathcal{Q}_μ on \mathcal{E} such that*

$$\mu(A) = \int_{\mathcal{E}} \nu(A) \mathcal{Q}_\mu(d\nu).$$

Remark 9.3.4 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 9.3.5 *If a time homogeneous Markov process has a unique invariant measure π , then π is ergodic.*

Proof. In this case $I_P = \{\pi\}$, so that π is an extremal of I_P . □

9.3.2 Proof of the Structure Theorem

Let us start by reviewing the symmetric difference of two sets and some basic properties.

Definition 9.3.6 Given two measurable sets A and B , we use the notation $A \sim B$ to signify that A and B differ by a set of \mathbb{P} -measure 0, i.e. $\mathbb{P}(A \triangle B) = 0$. Where

$$A \triangle B = (A \setminus B) \cup (B \setminus A) = (A \cap B^C) \cup (B \cap A^C).$$

Remark 9.3.7 Let us recall properties of the symmetric differences $A \triangle B$ of two sets. Firstly,

$$A \triangle B = A \cup B \setminus (A \cap B).$$

Thus $A^c \triangle B^c = A \triangle B$. Also, for any collection of sets $\{A_\alpha, B_\alpha\}$,

$$\left(\bigcup_{\alpha} A_{\alpha}\right) \triangle \left(\bigcup_{\alpha} B_{\alpha}\right) \subset \bigcup_{\alpha} (A_{\alpha} \triangle B_{\alpha}).$$

Also if $f : \Omega \rightarrow \Omega$ is any measurable function, then

$$f^{-1}(A \triangle B) = f^{-1}(A) \triangle f^{-1}(B).$$

Furthermore

$$(A \triangle B) \triangle (B \triangle C) = A \triangle C.$$

Finally, we have

$$\mathbb{P}(A \triangle B) = 0 \quad \implies \quad \mathbb{P}(A) = \mathbb{P}(B),$$

since we can decompose $A = (A \setminus B) \cup (B \setminus (B \setminus A)) = (A \setminus B) \cup (A \cap B)$.

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 \mathbb{P} is, we define the completion of a σ -algebra as follows.

Definition 9.3.8

1. A σ -algebra \mathcal{F} is complete with respect to probability measure μ if whenever $B \in \mathcal{F}$ and $\mu(B) = 0$, then any subset $A \subset B$ belongs to \mathcal{F} .
2. The completion $\bar{\mathcal{F}}$ of a σ -algebra \mathcal{F} is the smallest σ -algebra containing \mathcal{F} with the additional property that if $A \in \bar{\mathcal{F}}$ with $\mathbb{P}(A) = 0$ and $B \subset A$ is any subset of A , then $B \in \bar{\mathcal{F}}$.

Note. Suppose $\mathcal{G} \subset \mathcal{F}$. If $D \in \mathcal{G}$, $E \in \mathcal{F}$ with $\mu(E) = 0$, then $D \setminus E \in \bar{\mathcal{G}}$.²

Notation. We consider $(\mathcal{X}^{\mathbb{Z}}, \mathcal{B}(\mathcal{X}^{\mathbb{Z}}), \mathbb{P}_{\pi})$, both θ and θ^{-1} are measure preserving transformations on $\mathcal{X}^{\mathbb{Z}}$. We write $\mathbb{P} = \mathbb{P}_{\pi}$ when there is no confusion and we use the following σ -algebras of finite number of projections

$$\mathcal{F}_n^m := \vee_{k=-n}^m \sigma(x_k) \subset \mathcal{B}(\mathcal{X}^{\mathbb{Z}}).$$

Also our invariant sets of reference are $\mathcal{I} = \{A \in \mathcal{B}(\mathcal{X}^{\mathbb{Z}}) : \theta^{-1}A = A\}$. We start with proving that sets in $\mathcal{B}(\mathcal{X}^{\mathbb{Z}})$ can be approximated by cylindrical sets.

²This is because $D \setminus E = D \setminus (D \cap E)$, and by the definition of completion $\bar{\mathcal{G}}$.

Lemma 9.3.9 *Let $A \in \mathcal{B}(\mathcal{X}^{\mathbf{Z}})$, then for any $\epsilon > 0$, there exists $N > 0$ and $A_\epsilon \in \mathcal{F}_{-N}^N$ such that*

$$\mathbb{P}(A \triangle A_\epsilon) < \epsilon.$$

Proof. We want to show that

$$\mathcal{B}(\mathcal{X}^{\mathbf{Z}}) = \{A \in \mathcal{B}(\mathcal{X}^{\mathbf{Z}}) : \forall \epsilon > 0, \exists N > 0 \text{ \& } A_\epsilon \in \mathcal{F}_{-N}^N \text{ with } \mathbb{P}(A \triangle A_\epsilon) < \epsilon\}.$$

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 $\mathbb{P}(A_j \triangle A'_j) \leq \epsilon 2^{-j}$. Since \mathbb{P} is finite, we can also find J such that, setting $A = \bigcup_{j \geq 1} A_j$, one has $\mathbb{P}(A \triangle \bigcup_{j \leq J} A_j) \leq \epsilon$. We conclude that

$$\begin{aligned} \mathbb{P}(A \triangle \bigcup_{j \leq J} A'_j) &= \mathbb{P}\left(\left(A \triangle \bigcup_{j \leq J} A_j\right) \triangle \left(\bigcup_{j \leq J} A_j \triangle \bigcup_{j \leq J} A'_j\right)\right) \\ &\leq \mathbb{P}(A \triangle \bigcup_{j \leq J} A_j) + \mathbb{P}\left(\bigcup_{j \leq J} A_j \triangle \bigcup_{j \leq J} A'_j\right) \\ &\leq \epsilon + \mathbb{P}\left(\bigcup_{j \leq J} (A_j \triangle A'_j)\right) \leq \sum_{j \leq J} 2\epsilon \end{aligned}$$

Since $\bigcup_{j \leq J} A'_j \in \mathcal{F}_{-N}^N$ for $N = \max\{N_j : j \leq J\}$, the claim follows. \square

Lemma 9.3.10 *For any $A \in \mathcal{I}$, for any $l \in \mathbf{Z}$, there exists $\hat{A}_l \in \sigma(x_l)$ such that $A \sim \hat{A}_l$.*

Proof. Let $A \in \mathcal{I}$. By Lemma 9.3.9, given any $\epsilon > 0$ there exists $N = N(\epsilon) > 0$ and $A_\epsilon \in \mathcal{F}_{-N}^N$ such that $\mathbb{P}(A \triangle A_\epsilon) < \epsilon$. Since

$$\theta^{-1}(A \triangle A_\epsilon) = \theta^{-1}(A) \triangle \theta^{-1}(A_\epsilon) = A \triangle \theta^{-1}(A_\epsilon)$$

and \mathbb{P} is θ -invariant, then

$$\mathbb{P}(A \triangle \theta^{-k} A_\epsilon) < \epsilon, \quad \forall k \geq 0. \quad (9.7)$$

For this N and for any fixed k , $\theta^{-(N+k)} A_\epsilon \in \mathcal{F}_k^{2N+k} \subset \mathcal{F}_k^\infty$ holds for any ϵ .

Fix k and set $\epsilon_m = \frac{1}{m}$, then define

$$D_n^\epsilon = \theta^{-(N+k)} A_{\frac{\epsilon}{2^n}} \in \mathcal{F}_k^\infty, \quad D = \bigcap_{m \geq 1} \bigcup_{n=1}^{\infty} D_n^{\epsilon_m} \in \mathcal{F}_k^\infty. \quad (9.8)$$

Note that by (9.7) we have $\mathbb{P}(A \triangle D_n^\epsilon) < \frac{\epsilon}{2^n}$, then

$$\mathbb{P}\left(\bigcap_{n=1}^{\infty} (A \setminus D_n^{\epsilon_m})\right) = \lim_{m \rightarrow \infty} \mathbb{P}(A \setminus D_n^{\epsilon_m}) \leq \lim_{m \rightarrow \infty} \frac{\epsilon_m}{2^n} = 0. \quad (9.9)$$

On the other hand for any m ,

$$\mathbb{P}(D \setminus A) \leq \mathbb{P}\left(\bigcup_{n=1}^{\infty} D_n^{\varepsilon_m} \setminus A\right) = \mathbb{P}\left(\bigcap_{n=1}^{\infty} (A \setminus D_n^{\varepsilon_m})\right) \leq \frac{\varepsilon_m}{2^n} \leq \frac{1}{m}, \implies \mathbb{P}(D \setminus A) = 0. \quad (9.10)$$

It remains to show $\mathbb{P}(A \setminus D) = 0$ to have $A \sim D$, using (9.9)

$$\mathbb{P}(A \setminus D) = P\left(A \setminus \bigcap_{m \geq 1} \bigcup_{n=1}^{\infty} D_n^{\varepsilon_m}\right) = \mathbb{P}\left(\bigcup_{m \geq 1} \bigcap_{n=1}^{\infty} (A \setminus D_n^{\varepsilon_m})\right) = 0. \quad (9.11)$$

Hence for any k we found $D^{(k)} := D$ such that

$$\mathbb{P}(A \triangle D^{(k)}) = 0 \quad \text{and} \quad D^{(k)} \in \mathcal{F}_k^{\infty}. \quad (9.12)$$

Similarly, using θ^{-1} in place of θ we found $D^{(-k)} \in \mathcal{F}_{-\infty}^{-k}$, with $\mathbb{P}(A \triangle D^{(-k)}) = 0$.

Then for any l such that $-k < l < k$, by independence of past/future given present (an extension of Theorem 3.1.6), we obtain the following

$$\begin{aligned} \mathbf{E}[\mathbf{1}_A \mid \sigma(x_l)] &= \mathbf{E}[\mathbf{1}_A^2 \mid \sigma(x_l)] = \mathbf{E}[\mathbf{1}_{D^{(-k)}} \mathbf{1}_{D^{(k)}} \mid \sigma(x_l)] = \mathbf{E}[\mathbf{1}_{D^{(-k)}} \mid \sigma(x_l)] \mathbf{E}[\mathbf{1}_{D^{(k)}} \mid \sigma(x_l)] \\ &= (\mathbf{E}[\mathbf{1}_A \mid \sigma(x_l)])^2. \end{aligned}$$

Hence $\mathbf{E}[\mathbf{1}_A \mid \sigma(x_l)](\omega) = 1$ or 0 almost surely. Let

$$\hat{A} := \{\omega \in \mathcal{X}^{\mathbb{Z}} : \mathbf{E}[\mathbf{1}_A \mid \sigma(x_l)](\omega) = 1\} \in \sigma(x_l). \quad (9.13)$$

Then $\mathbf{E}[\mathbf{1}_A \mid \sigma(x_l)] = \mathbf{1}_{\hat{A}}$, and also $\mathbf{E}[\mathbf{1}_{A^C} \mid \sigma(x_l)] = \mathbf{1}_{\hat{A}^C}$. Then for any $E \in \sigma(x_l) \subset \mathcal{X}^{\mathbb{Z}}$, we have

$$\mathbb{P}(A \cap E) = \mathbf{E}(\mathbf{E}[\mathbf{1}_A \mid \sigma(x_l)] \mathbf{1}_E) = \mathbb{P}(\hat{A} \cap E).$$

In particular $\mathbb{P}(A \cap \hat{A}^C) = \mathbb{P}(\phi) = 0$ and similarly $\mathbb{P}(\hat{A} \cap A^C) = 0$. Hence

$$A \triangle \hat{A} = (A \cap \hat{A}^C) \cup (\hat{A} \cap A^C) \quad \text{has measure zero.}$$

Then $\hat{A}_\ell := \hat{A}$ is the $\sigma(x_\ell)$ -measurable function which differs from A by a measure zero set. \square

This marks end of lecture 23 - Week 10.

Proposition 9.3.11 *For any $A \in \mathcal{I}$, there exists $\bar{A} \in \mathcal{B}(\mathcal{X})$ such that $A \sim \Pi_{i \in \mathbb{Z}} \bar{A}$*

Proof. By Lemma 9.3.10, there exists $\hat{A} \in \sigma(x_0)$ with $A \sim \hat{A}$. Let $\bar{A} \in \mathcal{B}(\mathcal{X})$ such that $\hat{A} = \{\omega \in \mathcal{X}^{\mathbb{Z}} : x_0(\omega) \in \bar{A}\}$. Then, by invariance of A and \mathbb{P} (w.r.t. θ) $\mathbb{P}(A \triangle \theta^{-n} \hat{A}) = \mathbb{P}(\theta^{-n}(A \triangle \hat{A})) = 0$, so that

$$\mathbb{P}\left(\bigcup_{n=-\infty}^{\infty} A \triangle \theta^{-n} \hat{A}\right) = 0.$$

Note that $\theta^{-n}\hat{A} = \{\omega : x_n(\omega) \in \bar{A}\}$, then for any n we can check

$$\bigcap_{k=0}^n \theta^{-k}\hat{A} = \{\omega : x_0 \in \bar{A}, x_1 \in \bar{A}, \dots, x_n \in \bar{A}\} \sim A.$$

Hence conclude $\{x_i \in \bar{A}, i \in \mathbf{Z}\} = \Pi_{i \in \mathbf{Z}} \bar{A} \sim A$. \square

Given $\pi \in \mathcal{P}(\mathcal{X})$, recall that $\bar{A} \subset \mathcal{X}$ is π -invariant if $P(x, A) = 1$ for π -almost every $x \in \bar{A}$.

Corollary 9.3.12 *Let π be an invariant probability measure for P . Then π is ergodic if and only if every π -invariant set \bar{A} is of π -measure 0 or 1.*

Proof. (\Rightarrow). If \bar{A} is π -invariant (cf. (9.3), Remark 8.7.2, and Remark 8.7.3) then

$$\mathbb{P}_\pi(\Pi_{i \in \mathbf{Z}} \bar{A}) = \pi(\bar{A}). \quad (9.14)$$

Assume \mathbb{P}_π is ergodic so any θ -invariant set has measure 0 or 1. Since $\dots \times \bar{A} \times \bar{A} \times \bar{A} \times \dots \subset \mathcal{X}^{\mathbf{Z}}$ is θ -invariant then $\mathbb{P}_\pi(\Pi_{i \in \mathbf{Z}} \bar{A}) \in \{0, 1\}$ and we conclude that $\pi(\bar{A}) = 0$ or 1.

(\Leftarrow). For any $A \in \mathcal{I}$, by Proposition (9.3.11) there exists $\bar{A} \in \mathcal{B}(\mathcal{X})$ such that $A \sim \Pi_{i \in \mathbf{Z}} \bar{A}$. If $\mathbb{P}_\pi(A) = 0$ then as a projection of A , $\pi(\bar{A}) = 0$. Otherwise $\mathbb{P}_\pi(A) = 1$ and \bar{A} must be π -invariant. Indeed, then $\mathbb{P}(x_1 \in \bar{A}) = 1 = \int_{\mathcal{X}} P(z, \bar{A})\pi(dz) = 1$ which implies that $P(z, A) = 1$ a.e. z . Hence, by assumption $\pi(\bar{A}) \in \{0, 1\}$. Then, applying (9.14), we get $\mathbb{P}_\pi(A) = 0$ or 1. \square

Proposition 9.3.13 *Let π be an invariant probability measure. Then π is ergodic if and only if π is an extremal of $I_P = \{\nu \in \mathcal{P}(\mathcal{X}) : T\nu = \nu\}$.*

Proof. (\Rightarrow). Suppose π is not an extremal, i.e. $\pi = t\pi_1 + (1-t)\pi_2$, where $t \in (0, 1)$ and $\pi_1 \neq \pi_2 \in \mathcal{P}(\mathcal{X})$. Then $\mathbb{P}_\pi = t\mathbb{P}_{\pi_1} + (1-t)\mathbb{P}_{\pi_2}$. We show by contradiction π is not ergodic. Suppose π is ergodic, then for any θ -invariant set A ,

$$t\mathbb{P}_{\pi_1}(A) + (1-t)\mathbb{P}_{\pi_2}(A) \in \{0, 1\}.$$

This implies either $\mathbb{P}_{\pi_1}(A) = \mathbb{P}_{\pi_2}(A) = 0$ or $\mathbb{P}_{\pi_1}(A) = \mathbb{P}_{\pi_2}(A) = 1$. Therefore π_1 and π_2 are ergodic. By Proposition 9.2.10, π_1, π_2 are mutually singular. Then there exists a measurable set E such that $\pi_1(E) = 1, \pi_2(E) = 0$ which means $\mathbb{P}_{\pi_1}(\Pi_{i \in \mathbf{Z}} E) = 1$ and $\mathbb{P}_{\pi_2}(\Pi_{i \in \mathbf{Z}} E) = 0$. In particular, $P_\pi(\Pi_{i \in \mathbf{Z}} E) = t\mathbb{P}_{\pi_1}(\Pi_{i \in \mathbf{Z}} E) + (1-t)\mathbb{P}_{\pi_2}(\Pi_{i \in \mathbf{Z}} E) = t < 1$. This is in contradiction with π ergodic.

(\Leftarrow). Suppose π is not ergodic. There exists a π -invariant set F with $0 < \pi(F) = t < 1$ (c.f. Corollary 9.3.12). Let $\pi_1, \pi_2 \in \mathcal{P}(\mathcal{X})$ be defined as

$$\pi_1(B) = \frac{1}{t}\pi(B \cap F), \quad \pi_2(B) = \frac{1}{1-t}\pi(B \cap F^C).$$

Then we can write $\pi = t\pi_1 + (1-t)\pi_2$ where π_1 and π_2 are probability measures. We will show that both π_1, π_2 are invariant measures.

By π -invariance, $P(x, F) = 1$ for π -a.e. $x \in F$. By Lemma 8.7.4, the restriction of an invariant measure π to a π -invariant invariant set is invariant, the statement in that Lemma is for restrictions to P -invariant sets (the proof there shows the statement holds for π -invariant set). On the other hand

$$\pi(F^C) = \int_F P(x, F^C) d\pi + \int_{F^C} P(x, F^C) d\pi = \int_{F^C} P(x, F^C) d\pi.$$

This implies $P(x, F^C) = 1$, for π -a.e. $x \in F^C$. Hence π_2 , as the restriction of π on invariant set F^C is itself invariant. Therefore both $\pi_1, \pi_2 \in I_P$. \square

The proof of the Structure Theorem is concluded, now we apply Corollary 9.3.5 to obtain:

Proposition 9.3.14 *Let $A \subset \mathcal{X}$ be a P -invariant set. Let $A_0 = A$ and defined recursively $A_n = \{x \in \mathcal{X}, P(x, A_{n-1}) > 0\}$. Suppose*

$$\mathcal{X} = \bigcup_{n=1}^{\infty} A_n \quad \text{and} \quad A = \bigcup_{k=1}^m B_k,$$

where $\{B_k\}$ are disjoint closed sets with each B_k P -invariant. If the THMC restricts to B_k has unique invariant measure π_k , then π_k are ergodic and they are the only ergodic invariant probability measures.

Proof. Restricting to B_k , the ergodicity follows from Corollary 9.3.5. If $\pi \in I_P$, since $\mathcal{X} = \bigcup_{n=1}^{\infty} A_n$, then $\pi(A) = 1$ (cf. Proposition 8.7.7). The restriction of π on the π -invariant sets B_k is an invariant probability measure for P (see second half of last proof, Proposition 9.3.13). Since on B_k there exists a unique invariant measure, then we can uniquely decompose

$$\pi = \sum_{k=1}^m \pi(B_k) \pi_k$$

concluding the proof. \square

This marks end of lecture 24 - Week 10.

9.3.3 Proof of Birkhoff's Ergodic Theorem

This is not covered in the lectures. Before we turn to the proof of Theorem 9.1.6, we establish the following important result:

Theorem 9.3.15 (Maximal Ergodic Theorem) *Let $(\Omega, \mathcal{F}, \mathbb{P}, \theta)$ be a dynamical system with invariant σ -algebra and let $f: \Omega \rightarrow \mathbf{R}$ be such that $\mathbf{E}|f| < \infty$. Define*

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

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

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

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

Furthermore, $M_N(\omega) = 0 \vee \max\{S_1(\omega), \dots, S_N(\omega)\} = \max\{S_1(\omega), \dots, S_N(\omega)\}$ on the set $\{M_N > 0\}$ and on this set $f(\omega) \geq M_N(\omega) - M_N(\theta\omega)$ so that

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

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

We can now turn to the Proof of Birkhoff's Ergodic Theorem.

Proof. Replacing f by $f - \mathbf{E}(f \mid \mathcal{I})$, we assume without loss of generality that $\mathbf{E}(f \mid \mathcal{I}) = 0$. Let

$$\bar{\eta} = \limsup_{n \rightarrow \infty} \frac{S_n}{n}, \quad \underline{\eta} = \liminf_{n \rightarrow \infty} \frac{S_n}{n}.$$

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

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

and define S_N^ε and M_N^ε accordingly. It follows from Theorem 9.3.15 that

$$\int_{\{M_N^\varepsilon > 0\}} f^\varepsilon(\omega) \mathbb{P}(d\omega) \geq 0$$

for every $N \geq 1$. The sequence of sets $\{M_N^\varepsilon > 0\} = \{\max\{S_0^\varepsilon(\omega), S_1^\varepsilon(\omega), \dots, S_N^\varepsilon(\omega)\} > 0\}$ increases to the set

$$B^\varepsilon \equiv \{\sup_N S_N^\varepsilon > 0\} = \{\sup_N \frac{S_N^\varepsilon}{N} > 0\}.$$

Note that with these definitions we have that

$$\frac{S_N^\varepsilon(\omega)}{N} = \begin{cases} 0 & \text{if } \bar{\eta}(\omega) \leq \varepsilon \\ \frac{S_N(\omega)}{N} - \varepsilon & \text{if } \bar{\eta}(\omega) > \varepsilon \end{cases} \quad (9.15)$$

It follows from (9.15) that, and $\bar{\eta} = \limsup_{n \rightarrow \infty} \frac{S_n}{n}$,

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

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

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

and so

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

where we used the fact that $A^\varepsilon \in \mathcal{I}$ to go from the first to the second line, and the assumption $\mathbf{E}(f | \mathcal{I})(\omega) = 0$ in the last line. Therefore, one must have $\mathbb{P}(A^\varepsilon) = 0$ for every $\varepsilon > 0$, which implies that $\bar{\eta} \leq 0$ almost surely. \square

9.3.4 Example

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

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

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.

The key is to observe that the points $(2k+1)\pi$ are stable stationary solutions for the ODE $\frac{dx(t)}{dt} = \sin x(t)$.

This marks the end of Week 10 lectures.

Revision Class

We studied so far Markov chains $(x_n, n \geq 0)$, and (Time homogeneous) Markov chains with a family of transition probabilities $P = (P(x, \cdot), x \in \mathcal{X})$. The state space \mathcal{X} is a complete separable metric space. **Examples:** $\mathcal{X} = \mathbf{R}^n$, $\mathcal{X} = \mathbf{R}$, \mathcal{X} countable or finite.

While revising it is a good idea to check what a specific statement means for discrete state spaces (i.e. at most countable \mathcal{X} space with discrete topology). Also important to have clear the concepts of *finite dimensional distributions*, *stopping times*, *strong Markov property*.

Countable state space \mathcal{X} . For $i_j \in \mathcal{X}$ (with \mathcal{X} finite or countable) and $x_0 \sim \mu = \mathcal{L}(x_0)$, then the THMC (x_n) satisfies

$$\mathbb{P}(x_0 = i_0, x_1 = i_1, \dots, x_n = i_n) = \mu(i_0)P_{i_0 i_1} \dots P_{i_{n-1} i_n}.$$

For general state space one has (5.9) instead.

Example If (x_n) is a THMC with t.p. P on \mathbf{R} and $x_0 \sim \mu$, then

$$\mathbb{P}(x_0 \leq a_0, x_1 \leq a_1, \dots, x_n \leq a_n) = \int_{-\infty}^{a_0} \int_{-\infty}^{a_1} \dots \int_{-\infty}^{a_n} P(y_{n-1}, dy_n) \dots P(y_0, dy_1) \mu(dy_0).$$

Example If the t.p. on \mathbf{R} is given by $P(x, dy) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(y-x)^2}{2}} dy$, $x_0 \sim \mu$, then

$$\mathbb{P}(x_0 \leq a_0, x_1 \leq a_1, \dots, x_n \leq a_n) = \int_{-\infty}^{a_0} \int_{-\infty}^{a_1} \dots \int_{-\infty}^{a_n} \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\sum_{i=1}^n \frac{(y_{i+1}-y_i)^2}{2}} dy_n \dots dy_1 \mu(dy_0).$$

Chapman-Kolmogorov equations. Given one-step t.p. P , we define recursively

$$P^{n+1}(x, A) := \int_{\mathcal{X}} P(y, A) P^n(x, dy).$$

We showed (Proposition 5.1.6) that then we obtain transition function $\{P^n(x, \cdot) : x \in \mathcal{X}, n \geq 0\}$ satisfying Chapman-Kolmogorov equations, i.e. for every $n, m \geq 1$,

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

In the case of countable state space we can restrict to examine t.p. between single states $P_{ij} = P(i, \{j\})$ and $P_{ij}^n = P^n(i, \{j\})$ and exploit matrix formulation for CK equations:

$$P_{ij}^2 = \sum_{k \in \mathcal{X}} P_{ik} P_{kj}, \quad P_{ij}^{n+m} = \sum_{k \in \mathcal{X}} P_{ik}^n P_{kj}^m = (P^n P^m)_{ij}.$$

Transition operators. We have two transition operators $T^* : \mathcal{P}(\mathcal{X}) \rightarrow \mathcal{P}(\mathcal{X})$ and $T_* : \mathcal{B}_b(\mathcal{X}) \rightarrow \mathcal{B}_b(\mathcal{X})$, both denoted by T when there is no confusion and defined by

$$T\mu(A) = \int_{\mathcal{X}} P(x, A) \mu(dx), \quad Tf(x) = \mathbf{E}[f(x_{n+1}) | x_n = x] = \int_{\mathcal{X}} f(y) P(x, dy).$$

If \mathcal{X} is countable, using $\mu = (\mu(1), \mu(2), \dots) \in \mathcal{P}(\mathcal{X})$ vector characterisation

$$T\mu(i) = \sum_{k \in \mathcal{X}} \mu(k)P_{ki} = \mu P(i), \quad Tf(i) = \sum_{k \in \mathcal{X}} f(k)P_{ik} = Pf(i).$$

Feller property. The THMC (with associated transition operator T) is Feller if $x \mapsto Tf(x)$ is continuous for any $f \in C_b(\mathcal{X})$. Strong Feller if it holds for $f \in \mathcal{B}_b(\mathcal{X})$.

Note: On a discrete state space \mathcal{X} , the discrete metric is

$$d(x, y) = \begin{cases} 1, & x \neq y, \\ 0, & x = y. \end{cases}$$

So every $f : \mathcal{X} \rightarrow \mathbf{R}$ is continuous (w.r.t. this metric). In particular Feller/Strong Feller is automatic for discrete space with discrete topology.

Note: If $|\mathcal{X}| = \infty$, \mathcal{X} is not compact with respect to this metric.

Tightness. A family of probability measures, say $\{\mu_n\}$, is tight if for any $\varepsilon > 0$ there exists a compact set $K = K(\varepsilon)$ such that $\mu_n(K) \geq 1 - \varepsilon$.

Example: consider THMC (x_n) with t.p. on \mathbf{R}^n . Let $x_0 \in \mathbf{R}^n$, then $P^n(x_0, \cdot) = \mathcal{L}(x_n)$. Then $\{P^n(x_0, \cdot)\}$ is tight if for every ε , $\exists M$ such that

$$\mathbb{P}(|x_n| > M) < \varepsilon, \quad \forall n.$$

This is because $B_M = \{x \in \mathbf{R}^n : |x| \leq M\}$ is a compact set and

$$P^n(x_0, \mathbf{R}^n \setminus B_M) = \mathbb{P}(|x_n| > M).$$

By Markov's inequality

$$\mathbb{P}(|x_n| > M) \leq \frac{\mathbf{E}|x_n|}{M}.$$

Then, if $\sup_{n \geq 0} \mathbf{E}|x_n| < \infty$, $\{P^n(x_0, \cdot)\}$ is tight.

Compact sets with $\mathcal{X} = \mathbf{R}^n$. A subset of \mathbf{R}^n is compact if and only if it is closed and bounded. The following are some examples:

$$K_1 = \{x : |x| \leq M\}, \quad K_2 = \{(x_1, \dots, x_n) : |x_1| \leq a_1, \dots, |x_n| \leq a_n\}.$$

A finite union of compact sets is compact.

Invariant Measure. Given t.p. P with transition operator T , $\mu \in \mathcal{P}(\mathcal{X})$ is an invariant probability measure ($\mu \in I_P$) if

$$T\mu = \mu.$$

If $x_0 \sim \mu$, then $T\mu = \mathcal{L}(x_1)$ and $T^n\mu = \mathcal{L}(x_n)$ for all $n \geq 1$. Then, if μ is invariant, $\mathcal{L}(x_n) = T^n\mu = \mu$ for all n . In the case of \mathcal{X} discrete, using vector characterisation $\mu = (\mu(i))_{i \in \mathcal{X}} \in \mathcal{P}(\mathcal{X})$, invariance is equivalent to

$$T\mu(i) = \sum_{k \in \mathcal{X}} \mu(k)P_{ki} = \mu P(i) = \mu.$$

Important Questions

After some time, does $\mathcal{L}(x_n)$ reach an equilibrium? More in details:

1. If $\mu_0 = \mathcal{L}(x_0)$, does $\mu_n := T^n \mu$ converge as $n \rightarrow \infty$?
2. How does μ_n converge? (Weakly, in Total Variation distance)
3. How fast does it converge?
4. How about observables, numerics, statistics?

For example, considering empirical time averages, for *a.e.* $\omega \in \Omega$ do we have the following:

$$\begin{aligned} f : \mathcal{X} \rightarrow \mathbf{R}, \quad & \frac{1}{n} \sum_{k=1}^n f(x_k(\omega)) \approx \int_{\mathcal{X}} f \, d\pi ? \\ \varphi : \mathcal{X}^{\mathbf{Z}} \rightarrow \mathbf{R}, \quad & \frac{1}{n} \sum_{k=1}^n \varphi(\theta^k x) \approx \int_{\mathcal{X}} \varphi \, dP_{\pi} ? \end{aligned}$$

Ergodicity and Structure Theorem. The convergence questions are closely associated with ergodicity of invariant measures, i.e. the ergodicity of associated \mathbb{P}_{π} on canonical path space with shift map $\theta : \mathcal{X}^{\mathbf{N}} \rightarrow \mathcal{X}^{\mathbf{N}}$ or $\theta : \mathcal{X}^{\mathbf{Z}} \rightarrow \mathcal{X}^{\mathbf{Z}}$ (to form a dynamical system). Recall

$$(\theta a.)(n) = a(n+1).$$

For example, with $\mathcal{X} = \mathbf{R}$, then

$$\theta : (0, 1, 2, 3, 4, \dots) \mapsto (1, 2, 3, 4, \dots).$$

Also recall that a measure \mathbb{P} is θ -invariant (equiv. map θ is \mathbb{P} invariant / measure preserving) if

$$\mathbb{P}(\theta^{-1}(A)) = \mathbb{P}(A), \quad \forall A \in \mathcal{B}(\Omega).$$

If π is invariant for P , i.e. $\pi \in I_P$, then \mathbb{P}_{π} is θ -invariant. Recall the notation

$$I_P = \{\pi \in \mathcal{P}(\mathcal{X}) : \pi = T\pi\}, \quad \mathcal{E} = \mathcal{E}_P = \{\pi \in I_P : \pi \text{ is ergodic}\}.$$

We say P_{π} is ergodic if $\mathbb{P}_{\pi}(A) \in \{0, 1\}$ whenever $\theta_{-1}(A) = A$, and $\pi \in I_P$ is ergodic if P_{π} is ergodic.

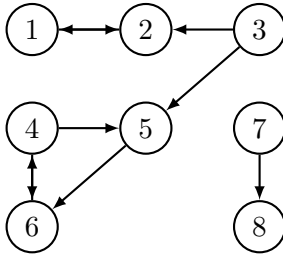
We have two notions for invariant subsets $E \in \mathcal{B}(\mathcal{X})$. The set E is said P -invariant if $P(x, E) = 1$ for every $x \in E$. Whereas, it is said π invariant if $P(x, E) = 1$ for π -a.e. $x \in E$. If

$\pi \in I_P$ and $\text{supp } \pi \supset E$, then E being P -invariant implies π -invariant. We have established the following characterisation (Corollary 9.3.12) of ergodic invariant probability measures:

Corollary Let $\pi \in I_P$. Then $\pi \in \mathcal{E}$ if and only if $\pi(E) \in \{0, 1\}$ for every π -invariant set E .

Note: If $|\mathcal{E}| = 1$, then $\pi \in \mathcal{I}_P$ is ergodic.

Example 1.



The communication classes are $A_1 = \{1, 2\}$, $A_2 = \{3\}$, $A_3 = \{4, 5, 6\}$, $A_4 = \{7\}$ and $A_5 = \{8\}$. These are related by

$$[1] \leq [3], \quad [4] \leq [3], \quad [8] \leq [7].$$

The collection of P -invariant sets

$$\tilde{I} = \{A_1, A_3, A_5, A_1 \cup A_2 \cup A_3, A_4 \cup A_5, \text{unions of those}\}.$$

Let $\pi_1, \pi_2, \pi_3 \in I_P$ be supported on A_1, A_3 and A_5 . Transient states have measure 0 (where was this proved?), then $\forall \pi \in I_P$ has $\pi(A_2 \cup A_4) = 0$. The set of invariant probability measures is

$$I_P = \{t_1\pi_1 + t_2\pi_2 + t_3\pi_3, t_1 + t_2 + t_3 = 1, t_i \geq 0\}.$$

So for any subset $T \subset \{\text{transient states}\}$ and $A \in \tilde{I}$, the set $A \cup T$ is also π -invariant. One can check $\pi_i(A \cup T) \in \{0, 1\}$, then verify π_i are ergodic and show $\mathcal{E} = \{\pi_1, \pi_2, \pi_3\}$.

Remark. (*Long Run Probabilities*) Let $\mathcal{X} = \{1, \dots, N\}$, we are interested in long time behaviour of $P^n(i, \cdot) = T^n \delta_i$. Recall Theorem 7.7.6, if P be irreducible and aperiodic (positive recurrence follows from finite state space) then

$$P_{ij}^n \xrightarrow{(n \rightarrow \infty)} \pi(j), \quad \forall i,$$

where π is the unique invariant probability measure. **Note:** If P is reducible or periodic, the conclusion above may not hold. In the example above (P reducible) we have for all n ,

$$P_{88}^n = 1, \quad P_{18}^n = 0.$$

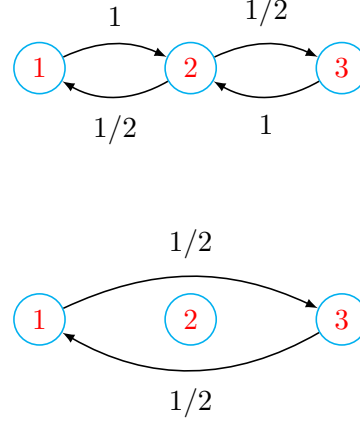
Example 2. (*3 States Markov Chain*)

Consider

$$P = \begin{pmatrix} 0 & 1 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 1 & 0 \end{pmatrix}, \quad P^2 = \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 1 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix}$$

Then the chain is 2-periodic and we have the graph representation on the right. In particular $P_{22}^{2n} = 1$ for all n , $\pi = (\frac{1}{4}, \frac{1}{2}, \frac{1}{2})$ is the invariant probability measure, but

$$\lim_{n \rightarrow \infty} P_{22}^n \neq \pi(2).$$



Birkhoff's Theorem. We already established (through Strong Law of Large numbers) Theorem 7.8.1 relating time averages and spatial averages. This is related to the application of Birkhoff's ergodic theorem for Markov chains (section 9.2.5). In particular (by Theorem 9.2.9), for initial point $x_0 = x \in \mathcal{X}$ and for any ergodic $\pi \in I_P$ we have

$$\frac{1}{n} \sum_{k=1}^n f(x_k) \longrightarrow \int_{\mathcal{X}} f d\pi, \quad \pi\text{-a.e. } x.$$

Applying this to Example 1, we have that π_1 is ergodic and $\pi_1(\{1, 2\}) = 1$. Then, $\pi_1(1) > 0$ and $\pi_1(2) > 0$, so that

$$\frac{1}{n} \sum_{k=1}^n f(x_k) \longrightarrow \int_{\mathcal{X}} f d\pi_1, \quad \text{if } x_0 = 1 \text{ or } 2.$$

Also $\pi_3 = \delta_8$ is ergodic and invariant, with $\pi_3(\{8\}) = 1$. Hence

$$\frac{1}{n} \sum_{k=1}^n f(x_k) \longrightarrow \int_{\mathcal{X}} f d\pi_3, \quad \text{if } x_0 = 8.$$

However, if $x_0 = 3$ (transient state), $\pi(3) = 0$ for any invariant (then also ergodic) probability measure $\pi \in I_P$. Therefore no conclusion can be drawn.

Existence and Uniqueness of invariant probability measures

In general state space, the two main criterion to establish existence are Krylov-Bogoliubov and Minorisation (which also gives uniqueness).

Krylov-Bogoliubov If for some $x_0 \in \mathcal{X}$,

$$\{P^n(x_0, \cdot), n \geq 1\} \text{ is tight} \quad + \quad \text{MC Feller} \quad \implies \quad I_P \neq \emptyset,$$

i.e. there exists an invariant probability measure. **Note:** both conditions hold automatically if \mathcal{X} is finite (c.f. Perron-Frobenius theorem) or if the chain is +ve recurrent + irreducible with \mathcal{X} countable.

Minorisation Suppose there exists $\eta \in \mathcal{P}(\mathcal{X})$ such that t.p. P is minorized, $\exists c > 0$ such that

$$\begin{aligned} P(x, \cdot) \geq c\eta, \forall x &\implies \|T^n\mu - T^n\nu\|_{\text{TV}} \xrightarrow{n \rightarrow \infty} 0, \forall \mu, \nu \in \mathcal{P}(\mathcal{X}) \\ &\implies |I_P| = 1 \quad \text{and} \quad \|T^n\mu - \pi\|_{\text{TV}} \xrightarrow{n \rightarrow \infty} 0, \forall \mu \in \mathcal{P}(\mathcal{X}), \pi \in I_P. \end{aligned}$$

Hence minorisation implies there exists unique $\pi \in I_P$.

Example In the case of countable state space, by Theorem 7.10.10, it suffices to establish with respect to a single state j_0 :

$$\exists j_0 \text{ s.t. } P_{ij_0} \geq \delta, \forall i \in \mathcal{X} \implies \exists! \pi \in I_P \quad \text{and} \quad \|\mu P^n - \pi\|_1 \rightarrow 0, \forall \mu \in \mathcal{P}(\mathcal{X}).$$

Note that minorization $P(i, \cdot) \geq c\eta > 0$ by η clearly implies $P_{ij_0} \geq \delta > 0$ for some $j_0 \in \mathcal{X}$.

There are other criteria/techniques to establish existence and uniqueness in general state space, but also to examine long run probabilities. Notable ones are the Lyapunov function test, the use of P -invariant sets (both for existence and uniqueness), couplings and deterministic contraction.

Techniques

Lyapunov Function test The test consists in finding a function $V : \mathcal{X} \rightarrow \mathbf{R}_+ \cup \{\infty\}$ such that $\{y : V(y) \leq a\}$ is compact for all a and $V^{-1}(\mathbf{R}_+) \neq \emptyset$, which satisfies for some $\gamma \in (0, 1)$

$$\begin{aligned} TV(x) &= \int V(y) P(x, dy) = \mathbf{E}[V(x_{n+1}) | x_n = x] \\ &\leq \gamma V(x) + C, \end{aligned}$$

for x such that $V(x) \neq \infty$. This allows to establish tightness to apply Krylov-Bogoliubov (cf. Thm 8.4.5), there are plenty of examples where this can be useful.

Coupling The method of Doeblin coupling was used in the case of countable state space to establish long run probabilities Ergodic theorem 7.7.6. Couplings was also used to establish Deterministic contraction criterion (Thm 8.5.5) for uniqueness of invariant probability measures:

Deterministic Contraction If $x_{n+1} = F(x_n, \xi_{n+1})$ and $\exists \gamma \in (0, 1)$ such that

$$\mathbf{E}d(F(x, \xi_1), F(y, \xi_1)) \leq \gamma d(x, y) \quad \forall x, y \in \mathcal{X}.$$

Then the process x_n has at most one invariant probability measure.

Invariant sets method This was treated in Section 8.7. It allows to establish existence of invariant probability measures by restricting to P -invariant sets, and also uniqueness if the set is sufficiently “absorbing”. This method was used in the random ODE example (cf. Proposition 8.7.16) with $x_n = \Phi(x_{n-1}, \xi_n)$, where $\Phi(x, f)$ is the (unique) solution at time $t = 1$ of

$$\dot{x}(t) = g(x(t)) + f(t), \quad x(0) = x.$$

Where ξ_n are iid $C([0, 1], \mathbf{R})$ -valued random variables.

Countable state space. Recall conclusions from Theorem 7.11.1 summarising properties of countable state space THMC.

1. Irreducible + recurrent $\implies \exists$ invariant (possibly non-finite mass) measure (unique up to multiplication constants).
2. Irreducible + recurrent case: This invariant measure has finite mass \iff the chain is +ve recurrent ($\iff \mathbf{E}_i T_i < \infty$).
3. In that case, we have the relation $\pi(i) = \frac{1}{\mathbf{E}_i T_i}$.

In the case of X finite, then recurrence implies +ve recurrence. (c.f. also Lemma 7.10.2)

Good luck!

Chapter 10

Appendix

10.1 Time reversal on 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 $\omega \rightarrow (x_n(\omega))$ 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 10.1.1 We define on $\mathcal{X}^{\mathbf{Z}}$ the family $\{\theta_n\}$ of shift maps and the time-reversal map ϱ by

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

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

Definition 10.1.2 A probability measure \mathbb{P} on $\mathcal{X}^{\mathbf{Z}}$ is said to define a **stationary** process if $\theta_n^* \mathbb{P} = \mathbb{P}$ for every $n \in \mathbf{Z}$; it is said to define a **reversible** process if $\varrho^* \mathbb{P} = \mathbb{P}$.

In other words, a stationary process is one where, statistically speaking, every time is equivalent.

10.1.1 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 :

$$(P^{(2)}\pi)(A \times B) = \int_A P(x, B) \pi(dx) = \mathbb{P}(x_0 \in A, x_1 \in B). \quad (10.1)$$

Observe that $(P^{(2)}\pi)(A \times B)$ is the two time probability distribution of the chain. Let us define $\varrho^{(2)}: \mathcal{X}^2 \rightarrow \mathcal{X}^2$ by $\varrho^{(2)}(x, y) = (y, x)$.

With this notation, we have

Theorem 10.1.3 *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 \mathbb{P}_π defined in Theorem 9.2.6 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 10.1.4 Observe that $(\varrho^{(2)})_*(P^{(2)}\pi) = P^{(2)}\pi$ is equivalent to: for every measurable and integrable function $f: \mathcal{X}^2 \rightarrow \mathbf{R}$,

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

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 \mathbb{P}_π . 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 $f: \mathcal{X}^2 \rightarrow \mathbf{R}$. We apply this to a function on \mathcal{X}^{n+m+1} and flip two consecutive coordinates successively:

$$f(x_{-n}, x_1, \dots, x_{m-1}, x_m) \rightarrow f(x_{-n}, \dots, x_m, x_{m-1}) \rightarrow \dots \rightarrow f(x_m, x_{m-1}, \dots, x_{-n+1}, x_{-n}).$$

It is then evident that the flipping of 2-coordinates is sufficient to determine the time reversal on $\mathcal{X}^{\mathbb{Z}}$. More precisely we have,

$$\begin{aligned}
& \int f(x_{-n}, \dots, x_m) \mathbb{P}_\pi(dx) \\
&= \int_{\mathcal{X}} \cdots \int_{\mathcal{X}}^{\overbrace{\quad}^{n+m+1}} f(x_{-n}, \dots, x_m) P(x_{m-1}, dx_m) \cdots P(x_{-n}, dx_{1-n}) \pi(dx_{-n}) \\
&= \int_{\mathcal{X}} \int_{\mathcal{X}} \left(\int_{\mathcal{X}} \cdots \int_{\mathcal{X}}^{\overbrace{\quad}^{n+m-1}} f(x_{-n}, \dots, x_m) P(x_{m-1}, dx_m) \cdots P(x_{-n+1}, dx_{-n+2}) \right) P(x_{-n}, dx_{1-n}) \pi(dx_{-n}) \\
&= \int_{\mathcal{X}} \int_{\mathcal{X}} \left(\int_{\mathcal{X}} \cdots \int_{\mathcal{X}}^{\overbrace{\quad}^{n+m-1}} f(x_{-n}, \dots, x_m) P(x_{m-1}, dx_m) \cdots P(x_{-n+1}, dx_{-n+2}) \right) Q(x_{1-n}, dx_{-n}) \pi(dx_{1-n}) \\
&= \int_{\mathcal{X}} \cdots \int_{\mathcal{X}}^{\overbrace{\quad}^{n+m+1}} f(x_{-n}, \dots, x_m) P(x_{m-1}, dx_m) \cdots Q(x_{1-n}, dx_{-n}) P(x_{1-n}, dx_{2-n}) \pi(dx_{1-n}) \\
&= \int_{\mathcal{X}} \cdots \int_{\mathcal{X}}^{\overbrace{\quad}^{n+m+1}} f(x_{-n}, \dots, x_m) P(x_{m-1}, dx_m) \cdots Q(x_{1-n}, dx_{-n}) Q(x_{2-n}, dx_{1-n}) \pi(dx_{2-n}) .
\end{aligned}$$

Proceeding in the same fashion, we finally arrive at

$$\begin{aligned}
& \int f(x_{-n}, \dots, x_m) \mathbb{P}_\pi(dx) \\
&= \int_{\mathcal{X}} \cdots \int_{\mathcal{X}}^{\overbrace{\quad}^{n+m+1}} f(x_{-n}, \dots, x_m) Q(x_{1-n}, dx_{-n}) \cdots Q(x_m, dx_{m-1}) \pi(dx_m) \\
&= \int f(x_{-n}, \dots, x_m) (\varrho^* \mathbf{Q}_\pi)(dx) ,
\end{aligned}$$

where we denoted by \mathbf{Q}_π 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 $\mathbb{P}_\pi = \varrho^* \mathbf{Q}_\pi$ and therefore that $\varrho^* \mathbb{P}_\pi = \mathbf{Q}_\pi$, which is the desired result. \square

Note that in the case where \mathcal{X} is countable, the condition (10.1) can be written as the detailed balance relation

$$P_{ij}\pi_j = P_{ji}\pi_i \quad (10.2)$$

for every pair i, j . Summing over j in (10.2) or choosing $B = \mathcal{X}$ in (10.1), we see that if there exists a probability measure π such that (10.1) 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 10.1.5 *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*

$$P_{i_1 i_n} P_{i_n i_{n-1}} \cdots P_{i_3 i_2} P_{i_2 i_1} = P_{i_n i_1} P_{i_1 i_2} \cdots P_{i_{n-2} i_{n-1}} P_{i_{n-1} i_n} \quad (10.3)$$

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

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

Proof. In order to show that the condition is necessary, let us consider the case $n = 3$. If the process is reversible, by the detailed balance relation 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, \dots, i_n in the incidence graph connecting 1 to i (we set $i_1 = 1$ and $i_n = i$). We then define a measure π on the state space by

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

Note that (10.3) 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

$$\frac{P_{ji} \pi_i}{P_{ij} \pi_j} = \frac{P_{ji}}{P_{ij}} \cdot \frac{P_{i_n i_{n-1}}}{P_{i_{n-1} i_n}} \frac{P_{i_{n-1} i_{n-2}}}{P_{i_{n-2} i_{n-1}}} \cdots \frac{P_{i_2 i_1}}{P_{i_1 i_2}} \cdot \frac{P_{j_{n-1} j_n}}{P_{j_n j_{n-1}}} \frac{P_{j_{n-2} j_{n-1}}}{P_{j_{n-1} j_{n-2}}} \cdots \frac{P_{j_1 j_2}}{P_{j_2 j_1}}. \quad (10.4)$$

Since we have $i = i_n$, $j = j_n$, and $i_1 = j_1$, the path $i_1, \dots, i_n, j_n, \dots, j_1$ forms a closed loop and the ratio in (10.4) 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). \square

Example 10.1.1 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 $\alpha x + \beta \xi_1 \sim N(\alpha x, \beta^2)$,

$$P(x, A) = \mathbb{P}(\alpha x + \beta \xi_1 \in A) = \int_A \frac{1}{\sqrt{2\pi\beta^2}} e^{-\frac{(y-\alpha x)^2}{2\beta^2}} dy.$$

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 $\pi = \mathcal{N}(0, \frac{\beta^2}{1-\alpha^2})$ 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

$$\mathbb{P}(x_0 \in A, x_1 \in B) = \int_A \int_B P(x, dy) \pi(dx) = \int_A \int_B \frac{1}{\sqrt{2\pi\beta^2}} e^{-\frac{(y-\alpha x)^2}{2\beta^2}} \frac{\sqrt{1-\alpha^2}}{\sqrt{2\pi\beta^2}} e^{-\frac{x^2(1-\alpha^2)}{2\beta^2}} dx dy.$$

Then $\mathbb{P}(x_0 \in \mathbf{R}, x_1 \in B) = \pi$, verifying that π is an invariant measure. To summarise,

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

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 10.1.2 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

$$x_{n+1} = x_n + \xi_n \pmod{L} .$$

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

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

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

Example 10.1.3 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 $p: V \rightarrow (0, 1)$. 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 $\mathcal{L}^2(\mathcal{X}, \pi)$.

10.2 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 U 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 subsets of \mathcal{X} containing A , it is the complement of the union of all open subsets of \mathcal{X} disjoint from A . 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 $d(x_n, x) \rightarrow 0$.

Definition 10.2.1 (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{E} is compact).

A metric space is discrete if for every point $x \in \mathcal{X}$ there exists a ball $B(x, r)$ containing 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 10.2.2 (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 10.2.3 *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 10.2.4 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 $\{A_0, A_1, \dots, A_N\} \subset \mathcal{T}$, 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 $f : \mathcal{X} \rightarrow \mathbf{R}$. On a metric space this concept of continuity agree with the usual continuity: For any $\epsilon > 0$ there exists $\delta > 0$ such that if $d(y, x) < \delta$, $|f(y) - f(x)| < \epsilon$.

10.3 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. \mathbf{Z} and \mathbf{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 converge. Suppose the limit is \bar{x} . If $x_n \not\rightarrow \bar{x}$, 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.

10.3.1 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 10.3.1 *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 10.3.2 *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 10.3.3 (Lusin's Theorem) *Let μ be a probability measure on a metric space \mathcal{X} and $f : \mathcal{X} \rightarrow \mathbf{R}$ 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 10.3.4 *If f is lower semi-continuous and non-negative, e.g. the indicator function of an open set, and μ a probability measure, then*

$$\int f d\mu = \sup \left\{ \int \varphi d\mu : 0 \leq \varphi \leq f, \varphi \in C_c(\mathcal{X}) \right\}.$$

10.3.2 On a compact metric space

A linear functional on $C(\mathcal{X})$ is a linear map $L : \mathbf{C}(X) \rightarrow \mathbf{R}$, it is said to be positive if $L(f) \geq 0$ whenever $f \geq 0$.

Theorem 10.3.5 *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 $f \in C(\mathcal{X})$.*

10.3.3 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 10.3.6 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 $\mathbb{P}(\mathcal{X})$ has the following neighbourhood basis. For any finite set of continuous functions $\{\varphi_i, i = 1, \dots, n\}$, any $n \in \mathbf{N}$ and $\varphi_i \in C_b(\mathcal{X})$, and $\mu_0 \in \mathbb{P}(\mathcal{X})$,

$$\left\{ \mu \in \mathbb{P}(\mathcal{X}) : \left| \int \varphi_i d\mu - \int \varphi_i d\mu_0 \right| \leq \epsilon, \forall \varphi_i \right\}.$$

Proposition 10.3.7 *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 $\int \varphi_k d\mu_n \rightarrow \int \varphi_k d\mu$ for every k .*

10.3.4 On a complete separable metric space

If $\mathcal{X}_1, \mathcal{X}_2$ are complete separable metric spaces and $\varphi : \mathcal{X}_1 \rightarrow \mathcal{X}_2$ 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 10.3.8 *Every probability measure on a complete separable metric space is tight.*

Proposition 10.3.9 *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 $P(\mathcal{X})$ 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.

10.3.5 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 is 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 10.3.10 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 10.3.11 Let \mathcal{X} be a separable metric space and μ_n be any sequence of probability measures on \mathcal{X} . Then $\mu_n \rightarrow \mu$ weakly if and only if

$$\lim_{n \rightarrow \infty} \sup_{f \in A} \left| \int f d\mu_n - \int f d\mu \right| = 0$$

for every family $A \subset C(\mathcal{X})$ which is equi-continuous at all points of \mathcal{X} and uniformly bounded.

Proposition 10.3.12 A subset A of $C([0, 1]; \mathbf{R})$ is relatively compact if it is sufficient and necessary that the following two conditions are satisfied:

$$1. \sup_{x \in A} |x(0)| < \infty$$

2.

$$\lim_{\delta \rightarrow 0} \sup_{x \in A} \omega_x(\delta) = 0,$$

$$\text{where } \omega_x(\delta) = \sup_{|s-t| < \delta} |x(s) - x(t)|.$$

Theorem 10.3.13 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 $\lambda : \mathbf{R}_+ \rightarrow \mathbf{R}_+$ which decreases to zero (M and λ may depend on ϵ),

(1) such that

$$\mu(x \in C : |x(0)| \leq M) \geq 1 - \epsilon, \quad \forall \mu \in \mathcal{M};$$

(2)

$$\mu(\{x : \omega_x(\delta) \leq \lambda(\delta) \text{ for all } \delta\}) > 1 - \epsilon; \quad \forall \mu \in \mathcal{M}.$$

Theorem 10.3.14 *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

$$\mu(x \in C : |x(0)| \leq M) \geq 1 - \epsilon, \quad \forall \mu \in \mathcal{M}$$

(2') For any $\epsilon > 0$ and $\delta > 0$ there exist a number $\eta > 0$ (which may depend on ϵ and δ) such that

$$\mu(\{x : \omega_x(\eta) \leq \delta\}) > 1 - \epsilon, \quad \forall \mu \in \mathcal{M}.$$

Let μ_1, μ_2, \dots be a sequence of measures on a topological space \mathcal{X} . We say that the sequence converges **weakly** to a limit μ if

$$\lim_{n \rightarrow \infty} \int_{\mathcal{X}} f(x) \mu_n(dx) = \int_{\mathcal{X}} f(x) \mu(dx), \quad (10.5)$$

for every $f \in \mathcal{C}_b(\mathcal{X})$. We say that it converges **strongly** if (10.5) holds for every $f \in \mathcal{B}_b(\mathcal{X})$.

10.4 The total variation norm

We define the total variation norm on the set of finite signed measures to be

$$\|\mu\|_{TV} = 2 \sup_{A \in \mathcal{B}(\mathcal{X})} \|\mu(A)\|.$$

Let μ^+ and μ^- are the measures in its Jordan-Hahn decomposition: $\mu = \mu^+ - \mu^-$. Define the measure $|\mu|(A) = \mu^+(A) + \mu^-(A)$. We also define the norm:

$$\|\mu\| = \mu^+(\mathcal{X}) + \mu^-(\mathcal{X}) = |\mu|(\mathcal{X}).$$

We are primarily interested in the difference between two probability measures which is a signed finite measure with $\mu(\mathcal{X}) = 0$. For such measures the two norms are the same.

Proposition 10.4.1 *If μ is a finite signed measure with $\mu(\mathcal{X}) = 0$, then*

$$\|\mu\|_{TV} = |\mu|(\mathcal{X}).$$

Proof. To see this, first note that if $\mu(\mathcal{X}) = 0$ then $\mu^+(\mathcal{X}) = \mu^-(\mathcal{X})$. Then letting $\mathcal{X} = \mathcal{X}^+ + \mathcal{X}^-$ be the Hahn decomposition of \mathcal{X} , we have

$$\|\mu\| := \mu^+(\mathcal{X}^+) + \mu^-(\mathcal{X}^-) = 2\mu^+(\mathcal{X}^+),$$

and

$$\|\mu\|_{TV} \geq 2\mu^+(\mathcal{X}^+) = \|\mu\|.$$

On the other hand one has

$$\mu(A) = \mu(A \cap \mathcal{X}^+) + \mu(A \cap \mathcal{X}^-) = \mu^+(A \cap \mathcal{X}^+) - \mu^-(A \cap \mathcal{X}^-) \leq \mu^+(\mathcal{X}^+) = \frac{1}{2}\|\mu\|,$$

$$\mu(A) \geq -\mu^-(A \cap \mathcal{X}^-) \geq -\mu^-(\mathcal{X}^-) = -\frac{1}{2}\|\mu\|,$$

and so $2|\mu(A)| \leq \|\mu\|$ for every measurable set A and $\|\mu\|_{TV} = 2\sup_A |\mu(A)| \leq \|\mu\|$ concluding $\|\mu\|_{TV} = \|\mu\|$. \square

Proposition 10.4.2 *If μ is a finite signed measure, then*

$$\|\mu\|_{TV} = \sup_{\substack{f \in \mathcal{B}_b(\mathcal{X}) \\ \|f\|_\infty = 1}} \left| \int_{\mathcal{X}} f(x) \mu(dx) \right|, \quad (10.6)$$

where the maximum is run over bounded measurable functions.

Proof. Firstly take the decompositions $\mu = \mu^+ - \mu^-$ and $\mathcal{X} = \mathcal{X}^+ \cup \mathcal{X}^-$ with $\mu^-(\mathcal{X}^+) = 0$ and $\mu^+(\mathcal{X}^-) = 0$. Then,

$$\sup_{\substack{f \in \mathcal{B}_b(\mathcal{X}) \\ \|f\|_\infty = 1}} \left| \int_{\mathcal{X}} f(x) \mu(dx) \right| \geq \int_{\mathcal{X}} (\mathbf{1}_{\mathcal{X}^+} - \mathbf{1}_{\mathcal{X}^-}) d\mu = \mu(\mathcal{X}^+) + \mu^-(\mathcal{X}^-) = \|\mu\|_{TV}.$$

Also for any measurable f with $\|f\|_\infty \leq 1$,

$$\int_{\mathcal{X}} f(x) \mu(dx) = \int_{\mathcal{X}^+} f(x) \mu^+(dx) - \int_{\mathcal{X}^-} f(x) \mu^-(dx) \leq \int_{\mathcal{X}^+} 1 \mu^+(dx) - \int_{\mathcal{X}^-} (-1) \mu^-(dx) = \|\mu\|_{TV}.$$

Similarly,

$$\int_{\mathcal{X}} f(x) \mu(dx) \geq -\|\mu\|_{TV}.$$

And $\left| \int_{\mathcal{X}} f(x) \mu(dx) \right| \leq \|\mu\|_{TV}$ for all such f , concluding the proof. \square

This also agrees with Rudin's definition:

Proposition 10.4.3 *If μ is a finite signed measure, then*

$$\|\mu\|_{TV} = \sup_{\pi} \sum_{A \in \pi} |\mu(A)|$$

where π is a partition of the σ -algebra.

Proof. Firstly for each π ,

$$\sum_{A \in \pi} |\mu(A)| = \sum_{A \in \pi} |\mu(A \cap \mathcal{X}^+) - \mu(A \cap \mathcal{X}^-)| \leq \mu(\mathcal{X}^+) + \mu(\mathcal{X}^-).$$

Then we see the partition $\{\mathcal{X}^+, \mathcal{X}^-\}$ maximise the quantity:

$$\sum_{A \in \{\mathcal{X}^+, \mathcal{X}^-\}} |\mu(A)| \geq \mu(\mathcal{X}^+) + \mu(\mathcal{X}^-) = \|\mu\|_{TV}.$$

□

10.5 Examples

Example 10.5.1 The interval $[0, 1]$ equipped with its Borel σ -algebra and the Lebesgue measure is a probability space.

Example 10.5.2 The half-line \mathbf{R}_+ equipped with the measure

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

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 $\mathbb{P}(dx) = e^{-x} dx$.

Example 10.5.3 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.

10.6 Proof of Prohorov's theorem

Theorem 10.6.1 (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 10.6.2 *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 $\lim_{k \rightarrow \infty} x_{n_k}(i) \rightarrow x(i)$ 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 $\lim_{k \rightarrow \infty} x_{n_k^1}(1) \rightarrow x(1)$. Similarly, there exists a subsequence n_k^2 of n_k^1 and a number $x(2)$ such that $\lim_{k \rightarrow \infty} x_{n_k^2}(2) \rightarrow x(2)$. 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 .
- $\lim_{k \rightarrow \infty} x_{n_k^i}(i) \rightarrow x(i)$ 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 $x(i)$. \square

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 $F_n(x) = \inf\{F_n(r_i) \mid r_i > x\}$ 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 $\lim_{k \rightarrow \infty} \tilde{F}_{n_k,i} = \tilde{F}_i$ for every i . Define a function $F: \mathbf{R} \rightarrow [0, 1]$ by $F(x) = \inf\{\tilde{F}_i \mid r_i > x\}$ for every $x \in \mathbf{R}$. Then the function F has the following properties:

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

The first and second claims follow 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 $F_{n_k}(x) \rightarrow F(x)$. 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 $\mu_{n_k}((a, b]) \rightarrow \mu((a, b])$ for every interval $(a, b]$ such that a and b do not belong to S . Fix now an arbitrary continuous function $\varphi: \mathbf{R} \rightarrow [-1, 1]$ and a value $\varepsilon > 0$. We want to show that there exists an N such that $|\int \varphi(x) \mu_{n_k}(dx) - \int \varphi(x) \mu(dx)| < 7\varepsilon$ for every $k \geq N$. Choose R as above and note that the tightness condition implies that

$$\left| \int \varphi(x) \mu_{n_k}(dx) - \int_{-R}^R \varphi(x) \mu_{n_k}(dx) \right| \leq 2\varepsilon, \quad (10.7)$$

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

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

for every $k \geq N$, thus concluding the proof. \square

10.7 Useful References

- 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
- Foundations of modern probability, O. Kallenberg.