Lecture Notes on Discrete State Stochastic Processes

Melih İşeri *

April 23, 2025

Abstract

These lecture notes are written for a one-semester course on stochastic processes at the upper-undergraduate level. In terms of the contents covered, it follows Durrett [3] and Cohen [2]. However, the overall structure of the presentation, including the proofs, is typically different. The contents are mostly proven rigorously; where this is not the case, the necessary steps are mentioned.

*Department of Mathematics, University of Michigan, United States, iseri@umich.edu.

Contents

1 Preliminaries

Up until the section 7.1, where we will introduce the notions in the measure theory, we will rely on the construction of probability theory for the countable state spaces. For this purpose, we will recall the core definitions.

Let Ω be a countable set, endowed with a mapping $p : \Omega \to [0,1]$ satisfying $\sum_{\omega\in\Omega} p(\omega) = 1$. Introduce the probability distribution $\mathbb{P}$ from subsets of Ω to $[0,1]$ as

$$\mathbb{P}(E) := \sum_{\omega\in E} p(\omega), \quad E \subset \Omega$$

(i) Conditional probability

$$\mathbb{P}(A|B) := \frac{1}{\mathbb{P}(B)}\mathbb{P}(A \cap B)$$

(ii) Bayes' Rule

$$\mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A)$$

(iii) Law of total probability

$$\mathbb{P}(A) = \sum_i \mathbb{P}(A \cap B_i) = \sum_i \mathbb{P}(A|B_i)\mathbb{P}(B_i)$$

where $\cup_i B_i = \Omega$ and B_i's are disjoint. It is also useful in computations to note that

$$\mathbb{P}(A|C) = \sum_i \mathbb{P}(A \cap B_i|C) = \sum_i \mathbb{P}(A|B_i \cap C)\mathbb{P}(B_i|C)$$

if B_i's are disjoint and $C \subset \cup_i B_i$.

(iv) We say $\{A_i\}_{i=1}^n$ are mutually independent if for any $1 \le j_1 < \cdots < j_k \le n$

$$\mathbb{P}\Big(\bigcap_{\ell=1}^k A_{j_\ell}\Big) = \prod_{\ell=1}^k \mathbb{P}\big(A_{j_\ell}\big)$$

We call a mapping $X : \Omega \to \mathbb{R}$ a random variable (RV). Let $R_X \subset \mathbb{R}$ be the range of X, which is also countable. Notice that any RV induces a distribution on R_X, which we call the law of X as

$$\mathcal{L}_X(E) := \mathbb{P}\big(\{\omega \in \Omega : X(\omega) \in E\}\big) := \mathbb{P}\big(X^{-1}(E)\big), \quad E \subset R_X$$

Expected of X is defined as

$$\mathbb{E}[X] := \sum_{\omega\in\Omega} X(\omega)\mathbb{P}(\omega) = \sum_{x\in R_X} x\mathcal{L}_X(x)$$

Let us go over a simple example of tossing two coins. Let $\Omega = \{HH, TT, HT, TH\}$, with equal probabilities ($\mathbb{P}$ is a uniform distribution). Define $X(HH) = 1$ and 0 otherwise. Then $\mathcal{L}_X(1) = 1/4$ and $\mathcal{L}_X(0) = 3/4$. Expected is

$$\mathbb{E}[X] = 1 \cdot \mathbb{P}(HH) + 0 \cdot \mathbb{P}(TT) + 0 \cdot \mathbb{P}(HT) + 0 \cdot \mathbb{P}(TH) = 1 \cdot \mathcal{L}_X(1) + 0 \cdot \mathcal{L}_X(0)$$

We can also define the expected value given an event, and decompose an expectation as follows,

$$\mathbb{E}[X] = \sum_i \mathbb{E}[X|B_i]\mathbb{P}(B_i) := \sum_i \Big(\sum_{\omega\in\Omega} X(\omega)\mathbb{P}(\omega|B_i) \Big)\mathbb{P}(B_i)$$

where $\cup_i B_i = \Omega$ and B_i's are disjoint.

We call $\{X_n\}_{n\in\mathbb{N}}$ a stochastic process with state space $\mathcal{S}$, if for all $n \in \mathbb{N}$, X_n is a RV with values in $\mathcal{S}$. i.e. $X_n : \Omega \to \mathcal{S}$. We refer to $n \mapsto X_n(\omega)$ as a path of X.

Notice that even if the state space has only two elements, there are 2^n many potential paths of $(X_1, X_2, \cdots, X_n)$. If the common probability space Ω has only few elements, then one cannot have rich dynamics. Typically, we don't model Ω and concentrate on the state space itself.

1.1 Total Variation Distance

Metric Theory is crucial to have in order to discuss local behaviors and understand convergence situations. Defined for any arbitrary set, the space of all probability distributions can be endowed with a metric. There are many choices, and we will present one that fits well for discrete state spaces;

Definition 1.1. Given two probability distributions μ, ν over the state space $\mathcal{S}$, we introduce the total variation distance as

$$d_{\text{TV}}(\mu, \nu) := \sup_{A\subset\mathcal{S}} |\mu(A) - \nu(A)| \tag{1.1}$$

Note that the total variation distance is defined for any measurable space (see Section 7.1). However, it is typically not the most useful one as it does not require a metric on $\mathcal{S}$ and, hence, does not embed it.

Proposition 1.2. *For a discrete state space $\mathcal{S}$, we have the characterization as*

$$d_{\text{TV}}(\mu, \nu) = \frac{1}{2}\sum_{x\in\mathcal{S}} |\mu(x) - \nu(x)|$$

Proof. First of all, since μ, ν are probability distribution, observe that

$$\sum_{x\in\{\mu(x)\geq\nu(x)\}} |\mu(x) - \nu(x)| = \sum_{x\in\{\mu(x)<\nu(x)\}} |\mu(x) - \nu(x)|$$

Then, to determine the set A maximizing (1.1),

$$\begin{aligned}
|\mu(A) - \nu(A)| &= \Big| \sum_{x\in A} \mu(x) - \nu(x) \Big| \\
&= \left| \sum_{x\in A\cap\{\mu(x)\geq\nu(x)\}} \mu(x) - \nu(x) + \sum_{x\in A\cap\{\mu(x)<\nu(x)\}} \mu(x) - \nu(x) \right| \\
&\leq \left(\sum_{x\in A\cap\{\mu(x)\geq\nu(x)\}} |\mu(x) - \nu(x)| \right) \vee \left(\sum_{x\in A\cap\{\mu(x)<\nu(x)\}} |\mu(x) - \nu(x)| \right) \\
&\leq \left(\sum_{x\in\{\mu(x)\geq\nu(x)\}} |\mu(x) - \nu(x)| \right)
\end{aligned}$$

where the last inequality uses the first observation. Now, it is clear that $A = \{\mu(x) \geq \nu(x)\}$ achieves the upper bound. It is straightforward to conclude by these observations. ∎

Proposition 1.3. *The dual representation of the total variation metric is;*

$$d_{\mathrm{TV}}(\mu,\nu) = \frac{1}{2}\sup\Big\{\sum_{x\in\mathcal{S}} f(x)\mu(x) - \sum_{x\in\mathcal{S}} f(x)\nu(x) \;:\; \max_{x\in\mathcal{S}}|f(x)| = 1\Big\}$$

Proof. Exercise. ■

1.2 Coupling

To understand the relations between two distributions on $\mathcal{S}$, one can consider the product space $\mathcal{S}\times\mathcal{S}$ and all distributions defined on it, with marginals corresponding to these two distributions. Any such distribution is called a coupling, which serves as a powerful tool to study relations between distributions.

Definition 1.4. Given two distributions μ,ν on the state space $\mathcal{S}$, we say a distribution δ on $\mathcal{S}\times\mathcal{S}$ is a coupling of μ and ν, if

$$\delta(\cdot\times\mathcal{S}) = \mu(\cdot) \quad \text{and} \quad \delta(\mathcal{S}\times\cdot) = \nu(\cdot)$$

Equivalently, we call random variables X, Y defined on a single probability space and taking values on $\mathcal{S}$ a coupling of μ and ν if

$$\mathcal{L}_X = \mu \;\; (=: X\sim\mu) \quad \text{and} \quad \mathcal{L}_Y = \nu \;\; (=: Y\sim\nu)$$

Moreover, given two random variables X and Y, we say X', Y' is a coupling of X and Y, if X', Y' is a coupling for $\mathcal{L}_X, \mathcal{L}_Y$.

Example 1.5. Let us consider the easiest case where $\mu = \nu$. It is always possible to couple measures independently, that is,

$$\delta(A\times B) := \mu(A)\mu(B)$$

is always a coupling. In this simpler case, we can also set

$$\delta'(A\times B) := \mu(A\cap B)$$

Note that the coupling δ spreads out two distributions across $\mathcal{S}\times\mathcal{S}$, whereas the coupling δ' concentrates all the mass on the diagonal. In the sense that if the first marginal A does not intersect with the second marginal B, then the probability is 0. For a finite $\mathcal{S}$, when representing the probabilities on $\mathcal{S}\times\mathcal{S}$ as a matrix, this implies that all the mass is concentrated on the diagonal.

If we consider a notion of a distance defined by couplings, observe that

$$\sum_{x,y\in\mathcal{S}} |x-y|\delta(x,y) > \sum_{x,y\in\mathcal{S}} |x-y|\delta'(x,y) = 0$$

In fact, the infimum over all couplings of the above quantity turns out to be a quite important notion of distance, connected to the so-called optimal transport, and is widely used in more general settings.

To demonstrate the definition of coupling in terms of random variables, which is equivalent, on the one hand take independent X and Y with distribution μ, and on the other hand note that X, X is a coupling of (μ,μ) in this simple case. Exactly as above,

$$\mathbb{E}[|X-Y|] > \mathbb{E}[|X-X|] = 0$$

Note that we have discussed a notion of a metric in the example where it embeds the metric on the state space $\mathcal{S}$, which was the absolute value. It turns out that we can represent the total variation metric in terms of couplings, which does not rely on any notion of distance over $\mathcal{S}$;

Proposition 1.6.

$$d_{\mathrm{TV}}(\mu,\nu) = \inf\left\{ \mathbb{P}(X \neq Y) : (X,Y) \textit{ is a coupling of } \mu \textit{ and } \nu \right\}$$

Proof. For any coupling (X,Y),

$$\mu(A) - \nu(A) = \mathbb{P}(X \in A) - \mathbb{P}(Y \in A) \leq \mathbb{P}(X \in A, Y \notin A) \leq \mathbb{P}(X \neq Y)$$

and by symmetry $|\mu(A) - \nu(A)| \leq \mathbb{P}(X \neq Y)$. Taking the supremum over A and infimum over couplings yields one inequality.

For the other inequality, we need to construct the optimal coupling. That is, we need to construct a coupling (X,Y) such that $\mathbb{P}(X \neq Y) = d_{\mathrm{TV}}(\mu,\nu)$. Consider an atomless probability space $(\Omega,\mathcal{F},\mathbb{P})$ together with an event $E \subset \Omega$ that has probability $d_{\mathrm{TV}}(\mu,\nu)$.[1] We will now describe how to construct $X, Y : \Omega \to \mathbb{S}$;

On the set E, let X maps to $\{x \in \mathbb{S} : \mu(x) \geq \nu(x)\}$ and Y maps to $\{x \in \mathbb{S} : \mu(x) < \nu(x)\}$. On E^c we will set $X = Y$, and thus by construction $d_{\mathrm{TV}}(\mu,\nu) = \mathbb{P}(X \neq Y)$. Crucial part is to make sure that their marginal distributions are matching μ and ν.

We set

$$\mathbb{P}(X = x|E) = \frac{\mu(x) - \nu(x)}{d_{\mathrm{TV}}(\mu,\nu)} \mathbf{1}_{\{\mu(x) \geq \nu(x)\}} \quad \text{and} \quad \mathbb{P}(Y = x|E) = \frac{\nu(x) - \mu(x)}{d_{\mathrm{TV}}(\mu,\nu)} \mathbf{1}_{\{\mu(x) < \nu(x)\}}$$

whereas on the complement, we set

$$\mathbb{P}(X = Y = x|E^c) = \frac{\mu(x) \wedge \nu(x)}{1 - d_{\mathrm{TV}}(\mu,\nu)}$$

To observe that these are well defined probability distributions, note that **[Exercise]**

$$\sum_{x \in \mathcal{S}} \mu(x) \wedge \nu(x) = 1 - \sum_{\mu(x) \geq \nu(x)} \mu(x) - \nu(x) = 1 - d_{\mathrm{TV}}(\mu,\nu)$$

Now, we are ready to compute their marginal distributions;

$$\mathbb{P}(X = x) = (1 - d_{\mathrm{TV}}(\mu,\nu)) \frac{\mu(x) \wedge \nu(x)}{(1 - d_{\mathrm{TV}}(\mu,\nu))} + d_{\mathrm{TV}}(\mu,\nu) \frac{\mu(x) - \nu(x)}{d_{\mathrm{TV}}(\mu,\nu)} \mathbf{1}_{\{\mu(x) \geq \nu(x)\}} = \mu(x)$$

Similar computation yields $\mathbb{P}(Y = x) = \nu(x)$ and hence we conclude the result. ∎

Let us note that, in the proof of the Convergence Theorem 2.47, we will construct couplings for Markov Chains and use Proposition 1.6.

Example 1.7 (Distance of Poisson Distributions)**.** Let μ_1, μ_2 be Poisson distributions with parameters $\lambda_1 < \lambda_2$. Recall that the sum of independent Poisson random variables is again a Poisson random variable. Therefore, we can couple μ_1, μ_2 by letting $X \sim \mathrm{Poi}(\lambda_1)$ and $Y \sim \mathrm{Poi}(\lambda_2 - \lambda_1)$ be independent, and observing that $(X_1, X_1 + Y_1)$ is a coupling. Thus, by 1.6,

$$d_{\mathrm{TV}}(\mu_1,\mu_2) \leq \mathbb{P}(X \neq (X+Y)) = \mathbb{P}(Y > 0) = 1 - e^{\lambda_2 - \lambda_1} \leq \lambda_2 - \lambda_1$$

[1] One can simply take the probability space as $([0,1], \mathcal{B}([0,1]), \mathbf{m})$, where $\mathbf{m}$ is the Lebesgue (or uniform) distribution.

2 Markov Chains

The fundamental idea is to characterize the evolution of systems with uncertainty. Essentially, our interest is to be able to model changes that are not known by certainty to us. Then, we can integrate those changes to obtain a description for the future of the system. We remark that almost all sufficiently complicated systems have an uncertain future. Thus, we need to model what we do not know, which is exactly the primary role of probability theory.

The fundamental question of interest is, given that the system is in a particular state, can you predict what will be the next state of the system. That is the information we want to integrate to obtain future distributions. The system might be formed by subatomic particles, galaxies, banks, animals, etc. and one can associate as many state variable as needed, such as position, momentum, balance sheet, connections etc. which characterizes the system well enough that allows better predictions of their potential future states.

Markov Chains (MC) models systems where past states are irrelevant but only the current state of the system determines the probabilities for the next state. Please see Wikipedia for a list of examples. To further simplify, we will consider MCs that does not depend on time n either.

Definition 2.1. We say a stochastic process X_n with state space $\mathcal{S}$ is a temporally homogeneous discrete Markov Chain with transition matrix $p : \mathcal{S} \times \mathcal{S} \to [0, 1]$, if for any $n \in \mathbb{N}$ and $x, y, x_0, \cdots, x_{n-1} \in \mathcal{S}$,

$$\mathbb{P}(X_{n+1} = y | X_n = x, X_{n-1} = x_{n-1}, \cdots, X_0 = x_0) = \mathbb{P}(X_{n+1} = y | X_n = x) = p(x, y)$$

whenever conditional probabilities are defined: $\mathbb{P}(X_n = x, X_{n-1} = x_{n-1}, \cdots, X_0 = x_0) > 0$.

Moreover, any transition matrix $p : \mathcal{S} \times \mathcal{S} \to [0, 1]$ satisfying $\sum_{y \in \mathcal{S}} p(x, y) = 1$ is called stochastic matrix, and any stochastic matrix can be associated with a Markov Chain.

In Section 3, we will introduce the concept of decision making and discuss reinforcement learning, which provides many examples and motivation. For now, we will introduce preliminary examples for which related quantities can be computed by hand.

Example 2.2 (Gambler's Ruin)**.** Consider a gambling game. At each round, you win \$1 with probability 0.4 and lose \$1 with probabiliy 0.6. Game ends when you lose all your money, or reach a fixed value N. To model this as a MC, let $\mathcal{S} = \{0, 1, \cdots, N\}$. Transition matrix is given by

$$\begin{aligned} &p(x, x + 1) = 0.4, \quad p(x, x - 1) = 0.6, \quad 0 < x < N \\ &p(0, 0) = 1, \quad p(N, N) = 1 \end{aligned}$$

[Write down the transition matrix for $N = 3$]

Example 2.3 (Ehrenfest Chain)**.** Consider two boxes connected with a very small opening. Each box contains some particles, and there are N particles in total. Let X_n be the number of particles in one of the boxes. At each step, one particle passes through the opening, and the probability depends on the number of particles in the box. The transition matrix is given by

$$p(i, i + 1) = (N - i)/N, \quad p(i, i - 1) = i/N, \quad p(i, j) = 0,$$

for all $i \in \{0, \cdots, N\}$, and $j \in \{0, \cdots, N\} \setminus \{i - 1, i + 1\}$. In the long run, probability of $X_n = i$ is given by $\frac{1}{2^N}\binom{N}{i}$, independent of the initial condition.

Example 2.4 (Inventory Chain)**.** Let X_n denote the stock in the inventory at the end of day n. At the beginning of the day, if the inventory is less or equal to s, we order enough to bring total stocks to S. Let D_n be the demand throughout day n. Then the dynamics of X_n are given by

$$X_{n+1} = \begin{cases} (X_n - D_{n+1})^+ & \text{if } X_n > s \\ (S - D_{n+1})^+ & \text{if } X_n \leq s \end{cases}$$

Suppose now an electronic store sells a video game. Set $S = 5, s = 1$. Let the demand be

$$\mathbb{P}(D_n = k) = \begin{cases} 0.3 & \text{if } k = 0 \\ 0.4 & \text{if } k = 1 \\ 0.2 & \text{if } k = 2 \\ 0.1 & \text{if } k = 3 \end{cases}$$

Then the transition matrix is given by

	0	**1**	**2**	**3**	**4**	**5**
0	0	0	0.1	0.2	0.4	0.3
1	0	0	0.1	0.2	0.4	0.3
2	0.3	0.4	0.3	0	0	0
3	0.1	0.2	0.4	0.3	0	0
4	0	0.1	0.2	0.4	0.3	0
5	0	0	0.1	0.2	0.4	0.3

Question. Suppose each unit solds for \$12 and storage cost is \$2. What is the long-run profit per day of this inventory policy? How do we choose s and S to maximize profit?

Example 2.5 (Two-Stage Markov Chains)**.** We can easily extend a Markov Chain such that X_{n+1} depends on X_n, X_{n-1}. For example, consider a basketball player who makes a shot with following probabilities:

- $1/2$ if he has missed the last two times
- $2/3$ if he has missed one of his last two shots
- $3/4$ if he hit both shots

Notice that the probability of next shot is given by the previous two. Therefore, let the state space be $\mathcal{S} = \{\mathbf{HH}, \mathbf{MM}, \mathbf{HM}, \mathbf{MH}\}$, $\mathbf{H}, \mathbf{M}$ standing for Hit and Miss. The transition matrix is given by

	HH	**HM**	**MH**	**MM**
HH	$3/4$	$1/4$	0	0
HM	0	0	$2/3$	$1/3$
MH	$2/3$	$1/3$	0	0
MM	0	0	$1/2$	$1/2$

Suggested Exercises: (Durret, 3rd ed.) 1.1,1.2,1.3

2.1 Multistep Transition Probabilities

In this section, we aim to show that

$$p^m(x, y) = \mathbb{P}(X_{n+m} = y | X_n = x)$$

That is, the probability of getting from the state x to y in m-steps is given by the m-th power of the transition matrix. Let us first present two exercises;

Exercise. Show that

$$\mathbb{P}(X_{n+m} = y | X_n = x) = \mathbb{P}(X_{n+m} = y | X_n = x, X_{n-1} = x_{n-1}, \cdots, X_0 = x_0)$$

Exercise. Suppose we have classifications of social classes as l, m, u, standing for lower, middle, and upper class. Transition matrix p denotes the probability of social mobility at each generation. For example, $p(m, u)$ denotes the probability that if the parents are in the middle class, children are in the upper class. **(i):** Suppose your parents are in the middle class. What is the probability that you are in the upper class and your children are lower class? **(ii):** What is the probability that your children are lower class, given that your parents are middle class?

Theorem 2.6. *The m-th step transition probability $\mathbb{P}(X_{n+m} = y | X_n = x)$ is the m-th power of the transition matrix p, computed at (x, y).*

Proof. Suppose the claim is true for $m - 1$. Then

$$\begin{aligned}
&\mathbb{P}(X_{n+m} = y | X_n = x) \\
&= \sum_{k \in \mathcal{S}} \mathbb{P}(X_{n+m} = y | X_{n+m-1} = k, X_n = x) \mathbb{P}(X_{n+m-1} = k | X_n = x) \\
&= \sum_{k \in \mathcal{S}} \mathbb{P}(X_{n+m} = y | X_{n+m-1} = k) \mathbb{P}(X_{n+m-1} = k | X_n = x) \\
&= \sum_{k \in \mathcal{S}} p(k, y) p^{m-1}(x, k) = p^m(x, y)
\end{aligned}$$

■

Let us note that in our discrete setting, Chapman-Kolmogorov Equation is the standard matrix multiplication;

$$p^{m+n}(x, y) = \sum_{k \in \mathcal{S}} p^m(x, k) p^n(k, y) \tag{2.1}$$

Example 2.7 (Gambler's Ruin)**.** Recall the example and its transition matrix. Set $N = 4$. By using computers, we can easily compute the 20-th power of the transition matrix, which gives us the probability distributions after 20 step starting at each state:

$$p^{20} = \begin{pmatrix}
 & \mathbf{0} & \mathbf{1} & \mathbf{2} & \mathbf{3} & \mathbf{4} \\
\mathbf{0} & 1.0 & 0 & 0 & 0 & 0 \\
\mathbf{1} & 0.87655 & 0.00032 & 0 & 0.00022 & 0.12291 \\
\mathbf{2} & 0.69186 & 0 & 0.00065 & 0 & 0.30749 \\
\mathbf{3} & 0.41842 & 0.00049 & 0 & 0.00032 & 0.58437 \\
\mathbf{4} & 0 & 0 & 0 & 0 & 1.0
\end{pmatrix}$$

Notice that with very high probability the game ends with either 0 or 4. Later, we will show that the limit p^n as $n \to \infty$ exists.

2.2 Strong Markov Property

We defined a Markov chain as being independent of the past, given X_n for some n. In this section, we will prove that time can, in fact, be randomized. To achieve this, we introduce the important concept of a stopping time;[2]

Definition 2.8. We say $T : \Omega \to \mathbb{N} \cup \{\infty\}$ is a stopping time with respect to X_n, if $\{T = n\}$ is determined by $\{X_0, \cdots, X_n\}$.

Note that the definition depends on the stochastic process X_n, and aims to identify random times that an event occurs. A typical example is "hitting time", where X_n becomes a particular state for the first time. In this case, for example, $\{T = n\}$ corresponds to all paths that the chain hits this particular state at time n. One can imagine how varying paths might change the hitting time of this particular state.

Theorem 2.9 (Strong Markov Property)**.** *Let T be a stopping time with respect to the Markov chain X_n. For any $k \in \mathbb{N}, x \in \mathcal{S}$, given $\{T < \infty, X_T = x\}$, X_{T+k} is independent of $\{X_0, \cdots, X_T\}$. Moreover,*

$$\mathbb{P}(X_{T+k} = y | T < \infty, X_T = x) = p^k(x, y)$$

Proof. For any $n \geq 1$, let $S(n, x)$ be the set of all sequences $(x_0, \cdots, x_{n-1}, x)$ such that if $X_0 = x_0, \cdots, X_{n-1} = x_{n-1}, X_n = x$, then it holds $T = n$. That is,

$$\{T < \infty, X_T = x\} = \cup_{n \geq 1} \{\vec{X}_n \in S(n, x)\}$$

where $\vec{X}_n := (X_0, \cdots, X_n)$. On the set $\{T < \infty, X_T = x\}$,

$$\begin{aligned}
&\mathbb{P}(X_{T+k} = y | T < \infty, X_T = x, \vec{X}_T) \\
&:= \sum_{n \geq 1} \sum_{\vec{x} \in S(n,x)} \mathbf{1}_{\{\vec{X}_n = \vec{x}\}} \mathbb{P}(X_{n+k} = y | T = n, X_n = x, \vec{X}_n = \vec{x}) \\
&= p^k(x, y) \sum_{n \geq 1} \sum_{\vec{x} \in S(n,x)} \mathbf{1}_{\{\vec{X}_n = \vec{x}\}} = p^k(x, y)
\end{aligned}$$

by the Markov property. This clearly implies the result because

$$\begin{aligned}
&\mathbb{P}(X_{T+k} = y | T < \infty, X_T = x) \\
&= \sum_{n \geq 1} \sum_{\vec{x} \in S(n,x)} \mathbb{P}(X_{T+k} = y | T < \infty, X_T = x, \vec{X}_n = \vec{x}) \mathbb{P}(\vec{X}_n = \vec{x} | T < \infty, X_T = x) \\
&= p^k(x, y) \sum_{n \geq 1} \sum_{\vec{x} \in S(n,x)} \mathbb{P}(\vec{X}_n = \vec{x} | T < \infty, X_T = x) = p^k(x, y) \\
&= \mathbb{P}(X_{T+k} = y | T < \infty, X_T = x, \vec{X}_T)
\end{aligned}$$

■

Remark 2.10. For continuous Markov processes, the strong Markov property and the Markov property are not equivalent. A standard example is a process that waits until an exponential clock rings and then moves in one direction at a constant speed.

To see that it is Markov, note that given $\{X_t = 0\}$, distribution of X_{t+s} is the same as the distribution of X_s, due to the memoryless property of the exponential distribution. Given $\{X_t \neq 0\}$, the distribution of X_{t+s} is a Dirac mass at $X_t + s$.

To see that it is not strong Markov, consider the stopping time $\tau = \inf\{t > 0 : X_t \neq 0\}$. Then, given $X_\tau = 0$, $X_{\tau+t}$ has the distribution Dirac mass at t, which is not the same as the distribution of X_t given $X_0 = 0$.

[2] Once we introduced the measure theoretic probability in section 7.1, we will define stopping times rigorously in section 7.6.

2.3 Classification of States

In this section, we will classify states to distinguish whether or not MC visits them indefinitely. Let us introduce the notation $\mathbb{P}_x(A) := \mathbb{P}(A|X_0 = x)$, and let $\mathbb{E}_x$ denote the expectation under the measure $\mathbb{P}_x$. Introduce the return time;

$$T_y := \min\{n \geq 1 : X_n = y\}, \quad \text{and} \quad \rho_{xy} := \mathbb{P}_x(T_y < \infty)$$

T_y is a stopping time because

$$\{T_y = n\} = \{X_0 \neq y, \cdots, X_{n-1} \neq y, X_n = y\}$$

and thus $\{T_y = n\}$ is determined by $\{X_0, \cdots, X_n\}$. Let us also introduce the k-th return time, which is also a stopping time,

$$T_y^1 := T_y, \quad T_y^k := \min\{n > T_y^{k-1} : X_n = y\}, \quad k \geq 2$$

Because of the strong Markov property, it holds that **[Exercise]**

$$\mathbb{P}_x(T_y^k < \infty) = \rho_{xy}\rho_{yy}^{k-1}$$

Exercise. Argue that $K_y := \max\{n \geq 1 : X_n = y\}$ is not a stopping time.

Before we continue to the task of classification of states, let us first prove an estimate involving the return time;

Lemma 2.11. *Suppose* $\mathbb{P}_x(T_y \leq k) \geq a > 0$ *for all* $x \in \mathcal{S}$. *Then*

$$\mathbb{P}_x(T_y > mk) \leq (1-a)^m$$

Proof. Define Markov chains $X_n^{(1)} := X_{k+n}, \cdots, X_n^{(m-1)} := X_{(m-1)k+n}$, and corresponding return times $T_y^{(1)}, \cdots, T_y^{(m-1)}$.

$$\begin{aligned}
&\mathbb{P}_x(T_y > mk) \\
&= \mathbb{P}_x(T_y > k, T_y^{(1)} > k, \cdots, T_y^{(m-1)} > k) \\
&= \mathbb{P}_x(T_y > k, T_y^{(1)} > k, \cdots, T_y^{(m-1)} > k | X_k \neq y, \cdots, X_{mk} \neq y) \\
&\quad \cdot \mathbb{P}_x(X_k \neq y, \cdots, X_{mk} \neq y) \\
&= \mathbb{P}_x(T_y > k | X_k \neq y)\mathbb{P}_x(T_y^{(1)} > k | X_{2k}, \neq y X_k \neq y) \cdots \mathbb{P}_x(T_y^{(m-1)} > k | X_{mk} \neq y, X_{(m-1)k} \neq y) \\
&\quad \cdot \mathbb{P}_x(X_{mk} \neq y | X_{(m-1)k} \neq y) \cdots \mathbb{P}_x(X_{2k} \neq y | X_k \neq y)\mathbb{P}_x(X_k \neq y) \\
&= \mathbb{P}_x(T_y > k)\Big(\sum_{z \neq y} \mathbb{P}_z(T_y > k)\mathbb{P}_x(X_k = z | X_k \neq y)\Big) \cdots \Big(\sum_{z \neq y} \mathbb{P}_z(T_y > k)\mathbb{P}_x(X_k = z | X_k \neq y)\Big) \\
&\leq (1-a)^m
\end{aligned}$$

We have used the independence on the third equality. For example, given what is X_k, the event $\{T_y > k\} = \{X_1 \neq y, \cdots, X_k \neq y\}$ is independent from the event $\{T_y^{(1)} > k\} = \{X_{k+1} \neq y, \cdots, X_{2k} \neq y\}$. ■

The main definitions that we interested are as follows;

Definition 2.12. We say $y \in \mathcal{S}$ is transient if $\rho_{yy} < 1$ and recurrent if $\rho_{yy} = 1$.

Definition 2.13. We say x communicates with y, and denote it by $x \to y$, if $\rho_{xy} > 0$.

Definition 2.14. A set $A \subset \mathcal{S}$ is called closed if $x \in A$ and $y \notin A$, then $p(x, y) = 0$.

Definition 2.15. A set $A \subset \mathcal{S}$ is called irreducible if $x, y \in A$ then $x \to y$.

We are primarily interested in understanding recurrent versus transient states. This will yield the long term domain of the process, in the sense that, $\lim_{n\to\infty} p^n(x, y) = 0$ for any transient state y. Also, note that $\mathbb{P}_y(T_y^k < \infty) = \rho_{yy}^k \to 0$ as $k \to \infty$ for a transient state y. That is, probability of observing a transient state many times is exponentially small.

Example 2.16. Recall Gambler's Ruin example. States 0 and N are recurrent, while all other states are transient. Closed sets are $\{0\}, \{N\}, \{0, N\}, \mathcal{S}$. Closed and irreducible sets are $\{0\}$ and $\{N\}$.

Suggested Exercises: (Durret, 3rd.) Exercices 1.5,1.6,1.7

Lemma 2.17. *If $x \to y$ and $y \to z$, then $x \to z$. That is, $\to$ is a transitive relation.*

Proof. By assumption, there exists $m, n \in \mathbb{N}$ such that $p^m(x, y) > 0$ and $p^n(y, z) > 0$. Then,

$$p^{m+n}(x, z) = \sum_k p^m(x, k) p^n(k, z) \geq p^m(x, y) p^n(y, z) > 0$$

■

Due to the transitivity, it is important to define $R_x := \{y \in \mathcal{S} : x \to y\}$ for $x \in \mathcal{S}$. This is because, any $y \to z$, $z \in R_x$ if $y \in R_x$, and hence forms a closed set. Note that, if a set A is not closed, then there exists $y \in A, z \in A^c$ where $y \to z$. The next theorem, and its corollary are crucial for classifying states.

Theorem 2.18. *If $\rho_{xy} > 0$ and $\rho_{yx} < 1$, then x is transient.*

Proof. Note that $\rho_{xy} > 0$ implies $\mathbb{P}_x(T_y < T_x) > 0$, and $\rho_{yx} < 1$ means $\mathbb{P}_y(T_x < \infty) < 1$.

$$\begin{aligned}
&\mathbb{P}_x(T_x < \infty) \\
&= \mathbb{P}_x(T_x < \infty | T_y < T_x)\mathbb{P}_x(T_y < T_x) + \mathbb{P}_x(T_x < \infty | T_y > T_x)\mathbb{P}_x(T_y > T_x) \\
&= \mathbb{P}_y(T_x < \infty)\mathbb{P}_x(T_y < T_x) + \mathbb{P}_x(T_x < \infty | T_y > T_x)\mathbb{P}_x(T_y > T_x) \\
&\leq \mathbb{P}_y(T_x < \infty)\mathbb{P}_x(T_y < T_x) + (1 - \mathbb{P}_x(T_y < T_x)) < 1
\end{aligned}$$

■

Corollary 2.19. *If x is recurrent and $x \to y$, then $\rho_{yx} = 1$.*

Now, we state the central theorem of this section. The reason is that it implies the closed and irreducible set R_x we introduced has states that are all recurrent, provided it is finite. This will immediately yield the following decomposition theorem.

Theorem 2.20. *If set $A \subset \mathcal{S}$ is finite, closed and irreducible, then all states in A are recurrent.*

Before we move on to the proof of Theorem 2.20, we will first use it to show the decomposition theorem. This theorem tells us that it suffices to study closed and irreducible sets if the Markov chain is finite.

Theorem 2.21 (Decomposition Theorem)**.** *If the state space $\mathcal{S}$ is finite, then it can be written as a disjoint union $\mathcal{S} = T \cup R_1 \cup \cdots \cup R_k$ where T is the set of transient states, and R_i's are closed irreducible sets of recurrent states.*

Proof. Introduce

$$T := \{x \in \mathbb{S} : \exists y \in \mathbb{S} \text{ s.t. } x \to y, y \nrightarrow x\}$$

By Theorem 2.18, all states in T are transient. Pick any $x \in \mathcal{S} \setminus T$. Let $R_1 := \{y \in \mathcal{S} : x \to y\}$. We have already observed that R_1 is closed. Now, we argue that it is irreducible for any $x \notin T$. Take any $y, z \in R_1$. Since $x \notin T$, $y \to x$, and as $x \to z$, Lemma 2.17 concludes the irreducibility. By Theorem 2.20, we conclude that all states in R_1 are recurrent. One can repeat the same argument to form R_i's until all elements of $\mathcal{S}$ are exhausted. ■

Now, we will study the number of visits to a particular state and its connections to the returning time to prove the Theorem 2.20. First, let us recall a well known relation;

Lemma 2.22. *For any random variable X taking values in $\mathbb{N}$,*

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} \mathbb{P}(X \geq k) \tag{2.2}$$

Proof. Note that $X = \sum_{k=1}^{\infty} \mathbf{1}_{\{X \geq k\}}$ and $\mathbb{E}\mathbf{1}_{\{X \geq k\}} = \mathbb{P}(X \geq k)$. We can change the order of $\mathbb{E}$ and $\sum_{k=1}^{\infty}$ due to Fubini's Theorem. [3] ■

Define N_y be the number of visits to y at times $n \geq 1$. That is

$$N_y := \sum_{n=1}^{\infty} \mathbf{1}_{\{X_n = y\}}$$

Note that it is closely connected to returning time;

$$\{N_y \geq k\} = \{T_y^k < \infty\}, \text{ and } \mathbb{E}_x N_y = \sum_{n=1}^{\infty} \mathbb{P}_x(X_n = y) = \sum_{n=1}^{\infty} p^n(x, y) \tag{2.3}$$

and by relying on this, we further have the following lemma;

Lemma 2.23.

$$\mathbb{E}_x N_y = \frac{\rho_{xy}}{(1 - \rho_{yy})}, \qquad \forall x, y \in \mathcal{S} \tag{2.4}$$

Proof.

$$\mathbb{E}_x N_y = \sum_{k=1}^{\infty} \mathbb{P}_x(N_y \geq k) = \sum_{k=1}^{\infty} \mathbb{P}_x(T_y^k < \infty) = \rho_{xy} \sum_{k=1}^{\infty} \rho_{yy}^{k-1} = \rho_{xy}/(1 - \rho_{yy})$$

■

This allows us to have an another characterization of recurrent states, which is a direct corollary of (2.4);

Theorem 2.24. *A state $x \in \mathcal{S}$ is recurrent if and only if $\sum_{n=1}^{\infty} p^n(x, x) = \mathbb{E}_x N_x = \infty$.*

Having this characterization at hand, we can now prove the two main lemmas;

Lemma 2.25. *If x is recurrent and $x \to y$, then y is recurrent.*

[3] Altough Fubini's Theorem is out of the scope of this course, we will interchange infinite summation with expectation, which is always valid if the integrand is positive.

Proof. Corollary (2.19) implies, by assumption, $\rho_{yx} = 1$. Pick m, n such that $p^m(x,y) > 0$ and $p^n(y,x) > 0$. Notice that

$$p^{m+n+k}(y,y) \geq p^n(y,x)p^k(x,x)p^m(x,y)$$

because

$$\{X_{m+n+k} = y | X_0 = y\} \supset \{X_n = x, X_{n+k} = x, X_{m+n+k} = y | X_0 = y\}$$

Summing over k yields

$$\sum_{\ell=1}^{\infty} p^\ell(y,y) \geq \sum_{k=1}^{\infty} p^{m+n+k}(y,y) \geq p^n(y,x)p^m(x,y)\sum_{k=1}^{\infty} p^k(x,x) = \infty$$

and the Theorem 2.24 concludes the result. ■

Lemma 2.26. *In a finite closed set, there has to be one recurrent state.*

Proof. Denote the finite closed set by A. If all the states are transient, then by (2.4), $\mathbb{E}_x N_y < \infty$ for any $x, y \in A$. It follows, since A is finite,

$$\infty > \sum_{y\in A} \mathbb{E}_x N_y = \sum_{y\in A}\sum_{n=1}^{\infty} p^n(x,y) = \sum_{n=1}^{\infty}\sum_{y\in A} p^n(x,y) = \sum_{n=1}^{\infty} 1 = \infty$$

which is a contradiction. ■

Proof. [of Theorem 2.20] The proof is a direct corollary of the last two lemmas. Namely, if the set is finite and closed, there has to be at least one recurrent element. But since the set is irreducible, every element is recurrent. ■

Suggested Exercise: (Durret, 3rd ed.) 1.8

2.4 Stationary Distributions

As the distribution of a Markov chain evolves over time, modeling our knowledge about where the chain will be in the future, information about the initial starting point diminishes. In fact, we will learn that this distribution converges to a stationary point. Thus, in this section, we will explore the meaning of a stationary distribution and argue its well-posedness.

Definition 2.27. Let π be a probability measure on the state space $\mathcal{S}$. We say π is a stationary distribution if

$$\pi(x) = \sum_{y\in\mathcal{S}} \pi(y)p(y,x), \quad \forall x \in \mathcal{S} \tag{2.5}$$

Remark 2.28. (i): Summation in (2.5) represents the evolution of the distribution π under the transition matrix of the Markov Chain. That is why π satisfying (2.5) is called a stationary distribution.

(ii): Note that, $\mathcal{L}_{X_n}$, the law of X_n, is also a probability measure on the state space $\mathcal{S}$. We may refer to $\mathcal{L}_{X_0}$ as a stationary distribution, if $\mathbb{P}$ induces a stationary distribution.

(iii): If the state space is finite as $\mathcal{S} = \{x_1, \ldots, x_k\}$, then we can write (2.5) in the matrix form;

$$\pi p = \pi$$

where π is taken as a row vector. In this case, the law of the MC is taken as $[\mathbb{P}(X_0 = x_1) \cdots \mathbb{P}(X_0 = x_k)]$.

(iv): We can discretize a distribution. Instead of a distribution π evolving, consider N many "particles" with states $\{x_1, \ldots, x_N\}$ such that

$$\frac{1}{N} \sum_{k=1}^{N} \delta_{x_k} \quad \text{is close to} \quad \pi$$

where $\delta_{\cdot}$ is the Dirac delta distribution. Then, one can imagine that these particles are evolving as a MC starting from these states. Each particle will change state, however, their empirical distribution $\frac{1}{N} \sum_{k=1}^{N} \delta_{X_k}$ will remain almost undisturbed.

Example 2.29. Lets explicitly solve a stationary distribution

$$\begin{bmatrix} \pi_1 & \pi_2 \end{bmatrix} \begin{bmatrix} 0.6 & 0.4 \\ 0.2 & 0.8 \end{bmatrix} = \begin{bmatrix} \pi_1 & \pi_2 \end{bmatrix}$$

Noting that $\pi_1 + \pi_2 = 1$ yields $\pi_1 = 1/3, \pi_2 = 2/3$.

Example 2.30 (General two state transition probability)**.**

$$\begin{bmatrix} \pi_1 & \pi_2 \end{bmatrix} \begin{bmatrix} 1-a & a \\ b & 1-b \end{bmatrix} = \begin{bmatrix} \pi_1 & \pi_2 \end{bmatrix}$$

The solution is given by

$$\pi_1 = \frac{b}{a+b}, \quad \pi_2 = \frac{a}{a+b}$$

In general, it is not easy to compute the stationary distribution. We will work out some particular cases, and the first one is when the transition matrix is also a stochastic matrix when time is reversed;

Definition 2.31. A transition matrix p is called doubly stochastic if $\sum_{x \in \mathcal{S}} p(x, y) = 1$.

Theorem 2.32. *If p is a doubly stochastic transition probability for a Markov chain with N states, then the uniform distribution, $\pi(x) \equiv 1/N$, is a stationary distribution.*

Proof. $\pi p = \frac{1}{N} \sum_{x \in \mathcal{S}} p(x, y) = \pi$ ■

Example 2.33 (Symmetric Reflecting Random Walk on the Line)**.** Set $\mathcal{S} = \{0, \cdots, L\}$. Markov chain moves left or right (or stay on the boundary) with equal probability. For example, if $L = 4$ the transition matrix is

	0	**1**	**2**	**3**	**4**
0	0.5	0.5	0	0	0
1	0.5	0	0.5	0	0
2	0	0.5	0	0.5	0
3	0	0	0.5	0	0.5
4	0	0	0	0.5	0.5

and as it is doubly stochastic, the stationary distribution is the uniform distribution.

2.4.1 Existence of stationary distribution

We now argue the existence of a stationary distribution. Here we state the result for finite state space, and later in Section 2.9 we will show how to cover the countable case.

Theorem 2.34 (Existence of stationary distribution)**.** *Assume $\mathcal{S}$ is finite. For any recurrent $x \in \mathcal{S}$, define*

$$\mu_x(y) := \sum_{n \geq 1} \mathbb{P}_x(X_n = y, T_x \geq n), \qquad \forall y \in \mathcal{S}$$

Then $\mu_x(x) = 1$ and

$$\pi_x(y) := \frac{\mu_x(y)}{\mathbb{E}_x[T_x]}$$

is a stationary distribution.

Proof. First of all, since x is recurrent,

$$\mu_x(x) = \sum_{n \geq 1} \mathbb{P}_x(X_n = x, T_x \geq n) = \sum_{n \geq 1} \mathbb{P}_x(T_x = n) = \mathbb{P}_x(T_x < \infty) = 1$$

and

$$\sum_{y \in \mathcal{S}} \mu_x(y) = \sum_{n \geq 1} \sum_{y \in \mathcal{S}} \mathbb{P}_x(X_n = y, T_x \geq n) = \sum_{n \geq 1} \mathbb{P}_x(T_x \geq n) = \mathbb{E}_x[T_x]$$

which explains the normalization constant. Let us start to show the stationary property.

$$\begin{aligned} &\sum_{y \in \mathcal{S}} \mu_x(y) p(y,z) \\ &= \sum_{n \geq 1} \sum_{y \in \mathcal{S}} \mathbb{P}_x(X_n = y, T_x \geq n) p(y,z) \\ &= \sum_{n \geq 1} \sum_{y \in \mathcal{S} \setminus \{x\}} \mathbb{P}_x(X_n = y, T_x \geq n) p(y,z) + \sum_{n \geq 1} \mathbb{P}_x(X_n = x, T_x \geq n) p(x,z) \end{aligned} \tag{2.6}$$

The second term is simply $p(x,z)$ as we have already shown the summation is 1. For the first term of (2.6), note that

$$\mathbb{P}_x(X_n = y, T_x \geq n) p(y,z) = \mathbb{P}_x(X_n = y, X_{n+1} = z, T_x \geq n)$$

because

$$\begin{aligned} &\mathbb{P}_x(X_n = y, T_x \geq n) p(y,z) \\ &= \mathbb{P}_x(X_n = y, T_x \geq n) \mathbb{P}_x(X_{n+1} = z | X_n = y) \\ &= \mathbb{P}_x(T_x \geq n | X_n = y) \mathbb{P}_x(X_{n+1} = z | X_n = y) \mathbb{P}_x(X_n = y) \\ &= \mathbb{P}_x(X_{n+1} = z, T_x \geq n | X_n = y) \mathbb{P}_x(X_n = y) \end{aligned}$$

where the last equality holds because $\{T_x \geq n\}$ and $\{X_{n+1} = z\}$ are independent given $\{X_n = y\}$. This is immediate, but let us formally show from the Markov property for the sake of getting familiarity;

$$\begin{aligned} &\mathbb{P}(X_{n+1}, T_x \geq n | X_n = y) \\ &= \mathbb{P}(X_{n+1} | X_n = y, T_x \geq n) \mathbb{P}(T_x \geq n | X_n = y) \\ &= \mathbb{P}(X_{n+1} | X_n = y, X_{n-1} \neq x, \cdots, X_0 \neq x) \mathbb{P}(T_x \geq n | X_n = y) \\ &= \mathbb{P}(X_{n+1} | X_n = y) \mathbb{P}(T_x \geq n | X_n = y) \end{aligned}$$

Therefore,

$$\sum_{y\in\mathcal{S}\setminus\{x\}} \mathbb{P}_x(X_n = y, T_x \geq n)p(y,z)$$
$$= \sum_{y\in\mathcal{S}\setminus\{x\}} \mathbb{P}_x(X_n = y, X_{n+1} = z, T_x \geq n) = \mathbb{P}_x(X_{n+1} = z, T_x \geq n+1)$$

and using this in (2.6) yields

$$\sum_{y\in\mathcal{S}} \mu_x(y)p(y,z)$$
$$= \sum_{n\geq 1} \mathbb{P}_x(X_{n+1} = z, T_x \geq n+1) + p(x,z)$$
$$= \mu_x(z) - \mathbb{P}_x(X_1 = z, T_x \geq 1) + p(x,z) = \mu_x(z) - \mathbb{P}_x(X_1 = z) + p(x,z) = \mu_x(z)$$

which concludes μ_x satisfies the stationary condition.

Now, we want to normalize μ_x to obtain a distribution. To do so, we will argue that $\mu_x(y) < \infty$ for all $y \in \mathcal{S}$. Then, since $\mathcal{S}$ is finite, we can normalize by $\sum_{y\in\mathcal{S}} \mu_x(y)$;

If $x \to y$, since x is recurrent $y \to x$. Choose m such that $p^m(y,x) > 0$. It follows

$$\mu_x(y)p^m(y,x) \leq \sum_{z\in\mathcal{S}} \mu_x(z)p^m(z,x) = \mu_x(x) = 1$$

and hence $\mu_x(y) < \infty$. This also shows that $\mathbb{E}_x[T_x] < \infty$ for finite Markov chains. ■

Remark 2.35. Note that we only used finiteness to normalize at the end. Existence of a measure (not normalized to have the mass 1) is always guaranteed when there is a recurrent state. To remove finiteness assumption, one needs to assume all states satisfy $\mathbb{E}_x[T_x] < \infty$, where such states are called positive recurrent states.

2.4.2 Uniqueness of stationary distribution

Theorem 2.36 (Uniqueness of stationary distribution)**.** *If $\mathcal{S}$ is finite and Markov chain is irreducible, then there exists a unique stationary distribution given by*

$$\pi(x) = \frac{1}{\mathbb{E}_x[T_x]} > 0, \qquad \forall x \in \mathcal{S} \tag{2.7}$$

Proof. Given the uniqueness, since the chain is irreducible, (2.7) immmediately follows from Theorem 2.34. Moreover, recall that we have shown $\mathbb{E}_x[T_x] < \infty$ in the proof of Theorem 2.34.

Now, to argue the uniqueness, let ν be an another stationary distribution, and fix any recurrent $x \in \mathcal{S}$.

$$\nu(y) = \sum_{z\in\mathcal{S}} \nu(z)p(z,y) = \nu(x)p(x,y) + \sum_{z\in\mathcal{S}\setminus\{x\}} \nu(z)p(z,y)$$

Iterate the second summation again by using the stationary property,

$$\sum_{z\in\mathcal{S}\setminus\{x\}} \nu(z)p(z,y) = \nu(x) \sum_{z\in\mathcal{S}\setminus\{x\}} p(x,z)p(z,y) + \sum_{z'\in\mathcal{S}\setminus\{x\}} \nu(z') \sum_{z\in\mathcal{S}\setminus\{x\}} p(z',z)p(z,y)$$

which then one notices that repeating this pattern for m times yield

$$\nu(y) = \nu(x) \sum_{n=1}^{m} \mathbb{P}_x(T_x \geq n, X_n = y) + \sum_{\tilde{z} \in \mathcal{S} \setminus \{x\}} \nu(\tilde{z}) \mathbb{P}_{\tilde{z}}(T_x \geq n, X_m = y)$$

By irreducability, since x is recurrent and $x \to \tilde{z}$, $\mathbb{P}_{\tilde{z}}(T_x \geq n, X_n = y) \leq \mathbb{P}_{\tilde{z}}(T_x \geq n) \to 0$ for all $\tilde{z}$, by corollary 2.19. Thus, by the Dominated Convergence Theorem, the second term vanishes as $n \to \infty$. Then,

$$\nu(y) = \nu(x) \sum_{m=1}^{\infty} \mathbb{P}_x(T_x \geq m, X_m = y) = \nu(x) \mu_x(y)$$

That is, ν has to be a constant multiple of μ_x. In order to have a probability distribution, $1 = \sum_{y \in \mathcal{S}} \nu(y) = \nu(x) \sum_{y \in \mathcal{S}} \mu_x(y) = \nu(x) \mathbb{E}_x T_x$ and hence concludes the result. ∎

Remark 2.37. **(i):** Note that we only used finiteness to have $\mathbb{E}_x[T_x] < \infty$. The proof actually shows that any stationary measure (without requiring the total mass to be 1) has to be a constant multiple of μ_x.

(ii): Recall the decomposition theorem 2.21. Each disjoint component has its own unique stationary distribution, except the set of transient states.

Suggested Exercises. Durrett, 3rd edition. 1.10, 1.12.

2.5 Detailed Balance Condition

In this section, we will explore the case where the Markov Chain is, in some sense, symmetric in time. We will explain this symmetry in the following subsections. We characterize this case by its stationary distribution, which is said to satisfy the following condition;

Definition 2.38. A distribution π is said to satisfy the detailed balance condition if

$$\pi(x) p(x, y) = \pi(y) p(y, x), \quad \forall x, y \in \mathcal{S} \tag{2.8}$$

This condition is stronger than being stationary, because if π satisfies (2.8), then

$$\sum_{y \in \mathcal{S}} \pi(y) p(y, x) = \sum_{y \in \mathcal{S}} \pi(x) p(x, y) = \pi(x)$$

which is exactly the condition of being a stationary distribution. Altough it is not typically satisfied, in some cases, it could be much simpler to solve (2.8). We will work it out for two cases below.

Example 2.39 (Birth and Death Chains)**.** Set $\mathcal{S} = \{l, l+1, \cdots, m-1, m\}$. We say a Markov chain is a birth and death chain if jumps can occur only to adjacent states. That is, $p(x, y) = 0$ if $|x - y| > 1$. Let

$$p(x, x+1) = u_x, \quad p(x, x-1) = d_x, \quad p(x, x) = 1 - (u_x + d_x)$$

where $d_l = u_m = 0$. Since jumps can occur only to adjacent states, detailed balance condition becomes

$$\pi(x) u_x = \pi(x+1) d_{x+1} \quad \text{or} \quad \frac{\pi(x+1)}{\pi(x)} = \frac{u_x}{d_{x+1}}$$

Therefore,

$$\frac{\pi(x)}{\pi(l)} = \frac{\pi(l+1)}{\pi(l)} \cdots \frac{\pi(x)}{\pi(x-1)} = \frac{u_l}{d_{l+1}} \cdots \frac{u_{x-1}}{d_x}, \quad x > l$$

Considering the normalization condition,

$$1 = \sum_{x=l}^{m} \pi(x) = \pi(l)\Big(1 + \sum_{x=l+1}^{m} \frac{u_l \cdots u_{x-1}}{d_{l+1} \cdots d_x}\Big) =: \pi(l) Z$$

and we conclude

$$\pi(l) = \frac{1}{Z}, \quad \pi(x) = \frac{1}{Z}\left(\frac{u_l \cdots u_{x-1}}{d_{l+1} \cdots d_x}\right)$$

is a stationary distribution satisfying the detailed balance condition. Notice that we implicitly required $d_x > 0$ for all x. Requiring $u_x > 0$ for all x would similarly work.

Example 2.40 (Random Walks on Graphs)**.** Consider a graph with finite vertices V and symmetric adjacency matrix $A : V \times V \to \{0, 1\}$ which encodes which vertices are connected (undirected). Degree of $x \in V$ is defined as the number of connected vertices as

$$d(x) := \sum_{y \in V} A(x, y)$$

A random walk transitioning uniformly over adjacent vertices is characterized by the transition matrix

$$p(x, y) = \frac{A(x, y)}{d(x)}$$

The corresponding detailed balance condition is

$$\pi(x)\frac{A(x, y)}{d(x)} = \pi(y)\frac{A(y, x)}{d(y)}$$

Therefore, since A is symmetric,

$$\pi(x) = \frac{d(x)}{\sum_{y \in V} d(y)}$$

is a stationary distribution satisfying detailed balance condition.

Suggested Exercises. Durrett, 3rd edition. 1.11.

2.5.1 Reversibility

In this section, we will characterize the transition matrix of a Markov Chain reversed in time.

Theorem 2.41. *Consider a Markov chain with transition matrix p and stationary distribution π. Suppose that the initial distribution is equal to π, i.e. $\mathbb{P}(X_0 = x) = \pi(x)$. Fix n and set $\hat{X}_m = X_{n-m}$ for $0 \le m \le n$. Then $\hat{X}$ is a Markov chain with transition matrix*

$$\hat{p}(x, y) = p(y, x)\frac{\pi(y)}{\pi(x)} = \frac{\pi(y)p(y, x)}{\sum_{z \in \mathcal{S}} \pi(z)p(z, x)} \tag{2.9}$$

Proof. Let us start by writing the Markov condition for $\hat{X}$,

$$\begin{aligned}
&\mathbb{P}(\hat{X}_{m+1} = y | \hat{X}_m = x, \hat{X}_{m-1} = x_{m-1}, \cdots, \hat{X}_0 = x_0) \\
&= \mathbb{P}(X_{n-(m+1)} = y | X_{n-m} = x, X_{n-(m-1)} = x_{m-1}, \cdots, X_n = x_0) \\
&= \frac{\mathbb{P}(X_{n-m} = x, X_{n-(m-1)} = x_{m-1}, \cdots, X_n = x_0 | X_{n-(m+1)} = y)}{\mathbb{P}(X_{n-m} = x, X_{n-(m-1)} = x_{m-1}, \cdots, X_n = x_0)}\pi(y)
\end{aligned}$$

The last equality is due to the stationary property. Notice that the numerator is

$$\mathbb{P}(X_{n-m} = x | X_{n-(m+1)} = y)\mathbb{P}(X_{n-(m-1)} = x_{m-1}, \cdots, X_n = x_0 | X_{n-m} = x)$$

and the denominator is

$$\mathbb{P}(X_{n-(m-1)} = x_{m-1}, \cdots, X_n = x_0 | X_{n-m} = x)\mathbb{P}(X_{n-m} = x)$$

and the result follows again by observing the stationary property. ■

Remark 2.42.

(i): (2.9) is called the dual transition probability. Although $\hat{p}$ is always well-defined, it is important to note that we must start the original chain from its well-posed stationary distribution to ensure that the future distribution is independent of the initial condition, and hence the reversed chain $\hat{X}$ is also a Markov Chain.

(ii): If the Markov chain satisfies the detailed balance condition, then $\hat{p}(x, y) = p(x, y)$.

(iii): If the stationary distribution is uniform as in the case of doubly stochastic transition matrices, then $\hat{p}(x, y) = p(y, x)$.

2.5.2 Kolmogorov Cycle Condition

We will only state this characterization of the detailed balance condition. See Durrett [3] for the proof. In words, the detailed balance condition is equivalent to being time reversible on any cycle of states.

Theorem 2.43. *For an irreducible Markov chain with state space S, there exists a stationary distribution satisfying detailed balance condition if and only if, given any cycle $x_0, x_1, \cdots, x_n = x_0$ it holds*

$$\prod_{i=1}^{n} p(x_{i-1}, x_i) = \prod_{i=1}^{n} p(x_i, x_{i-1})$$

2.6 Limit Behavior

In this section, we study the limit behavior of $p^n(x, y)$. For a transient state y, since $\mathbb{E}_x N_y = \sum_{n=1}^{\infty} p^n(x, y) < \infty$ by Theorem 2.24, we know that $p^n(x, y) \to 0$ for any $x \in \mathcal{S}$. Therefore, we are interested in closed and irreducible components containing recurrent states.

Expectation is that the distribution of X_n to converge to the unique stationary distribution as n tends to infinity. However, one needs to take care of periodicity, which is the only possibility that prevents the convergence. For example, consider the transition matrix on $\mathcal{S} = \{0, 1\}$

$$p(0, 1) = 1, \quad p(1, 0) = 1, \quad p(0, 0) = 0, \quad p(1, 1) = 0$$

It is obvious that $\pi(0) = \pi(1) = 1/2$ is the stationary distribution. However, note that $p^n(0, 0) = 0$ if n is odd and $p^n(0, 0) = 1$ if n is even. Therefore, the limit over n simply does not exists.

To resolve this issue, for a recurrent state x, define

$$I_x := \{n \geq 1 : p^n(x, x) > 0\}$$

and set d_x, the period of x, to be the greatest common divisor of I_x. Since $\sum_{n=1}^{\infty} p^n(x, x) = \infty$, I_x is non empty and d_x is well defined. We first show that the periodicity is in fact a property of an irreducible component;

Lemma 2.44. *If $x \to y$ and $y \to x$, then x and y has the same period.*

Proof. Suppose $d_y < d_x$. Choose n, m such that $p^n(x,y) > 0$ and $p^m(y,x) > 0$. Since

$$p^{n+m}(x,x) \geq p^n(x,y)p^m(y,x) > 0$$

$n + m \in I_x$. Consider any $\ell \in I_y$,

$$p^{n+m+\ell}(x,x) \geq p^n(x,y)p^\ell(y,y)p^m(y,x) > 0$$

Hence $n+m+\ell \in I_x$ too. Since d_x divides $n+m$, d_x divides arbitrary $\ell \in I_y$ too. Which means $d_x \leq d_y$ yielding a contradiction. ■

Definition 2.45. We say a recurrent state is aperiodic if $d_x = 1$. We say an irreducible Markov chain is aperiodic, if one hence all states are aperiodic.

We will now work out the key technical lemma to use the aperiodicity in the the convergence analysis;

Lemma 2.46. *If x is aperiodic, then there is $n_0 \in \mathbb{N}$ such that $n \in I_x$ for all $n \geq n_0$.*

Proof. First, suppose there exists $k \in \mathbb{N}$ such that $k, k+1 \in I_x$. Then, because I_x is closed under addition, $2k, 2k+1, 2k+2 \in I_x$. In general, $mk, mk+1, \cdots, mk+m \in I_x$. Thus, choosing m large enough to satisfy $(m+1)k \leq mk + m$ concludes that the rest of the integers are contained in I_x.

Now, we will show that we can always find $k, k+1 \in I_x$. Since x is recurrent, $\sum_{n\geq 1} p^n(x,x) = \infty$ and hence I_x has infinitely many elements. Take any n_0 and $n_0 + k$ in I_x with $k > 1$. We can find $n_1 \in I_x$ such that k does not divide n_1, simply because x is aperiodic. Let (m, r) satisfy

$$n_1 = mk + r \text{ with } 0 < r < k$$

Again since I_x is closed under addition, both

$$(m+1)(n_0+k) = (m+1)n_0 + n_1 + k - r \quad \text{and} \quad (m+1)n_0 + n_1$$

are in I_x. Since the difference is $k - r < k$, we can iterate this algorithm at most k times to obtain two consecutive numbers in I_x. ■

2.6.1 Convergence Theorem

Please review the first section on total variation distance and coupling for the proof of the following theorem.

Theorem 2.47 (Convergence)**.** *Suppose a finite Markov chain is irreducible and aperiodic with the stationary distribution π. Then,*

$$\lim_{n\to\infty} p^n(x,y) = \lim_{n\to\infty} \mathbb{P}_x(X_n = y) = \pi(y), \quad \forall x, y \in \mathcal{S}$$

Proof. The proof constructs a Markov Chain on the product space $\mathcal{S} \times \mathcal{S}$ to couple the distribution $p^n(x,\cdot)$ with π. To do so, define the transition probabilities on the $\mathcal{S} \times \mathcal{S}$ as follows;

$$\bar{p}((x,x'),(y,y')) = \begin{cases} p(x,y)p(x',y') & \text{if } x \neq x' \\ p(x,y)\mathbf{1}_{\{y=y'\}} & \text{otherwise} \end{cases}$$

That is, whenever the components are not equal, each component is independently evolving under the dynamics of the original Markov Chain. However, once they transition to the same state, they evolve together.

Let (X_n, Y_n) be the Markov Chain corresponding to $\bar{p}$, where initial distributions $X_0 = x$ and $Y_0 \sim \pi$. We will observe that X_n, Y_n is a coupling of $p^n(x, y)$ and $\pi(y)$, and we will only show this for a single step;

$$\begin{aligned}
&\mathbb{P}(X_n = y | X_{n-1} = x) \\
&= \sum_{x' \in \mathcal{S}} \mathbb{P}(X_n = x, Y_n \in \mathcal{S} | X_{n-1} = x, Y_{n-1} = x') \mathbb{P}(Y_{n-1} = x' | X_{n-1} = x) \\
&= \sum_{x', y' \in \mathcal{S}} \bar{p}((x, x'), (y, y')) \mathbb{P}(Y_{n-1} = x' | X_{n-1} = x) \\
&= p(x, y) \mathbb{P}(Y_{n-1} = x) + p(x, y) \sum_{x' \neq x, y' \in \mathcal{S}} p(x', y') \mathbb{P}(Y_{n-1} = x' | X_{n-1} = x) = p(x, y)
\end{aligned}$$

and by symmetry, it also holds for Y_n that their marginal transition probabilities are the same as the original MC. This leads, by proposition 1.6,

$$d_{\text{TV}}(p^n(x, \cdot), \pi) \leq \mathbb{P}(X_n \neq Y_n)$$

Idea is to show that X_n, Y_n eventually becomes equal and hence the $\mathbb{P}(X_n \neq Y_n) \to 0$ as $n \to \infty$. To argue this, we will concentrate on the simpler version of the Markov chain where transition probabilities are simply

$$\hat{p}((x, x'), (y, y')) = p(x, y) p(x', y')$$

with corresponding random variables $\hat{X}_n, \hat{Y}_n$ with the same initial conditions as before. Define the stopping time $\hat{T} = \inf\{n \geq 1 : \hat{X}_n = \hat{Y}_n\}$ and note that we constructed the first chain exactly to have

$$\mathbb{P}(X_n \neq Y_n) = \mathbb{P}(\hat{T} > n)$$

Now, we claim $\hat{p}$ is irreducible, hence all states are recurrent by finiteness. That is $\mathbb{P}(\hat{T} < \infty) \geq \mathbb{P}(T_{(x,x)} < \infty) = 1$, and hence $\mathbb{P}(\hat{T} > n) \to 0$.

To show the irreducibility, take any $(x, x'), (y, y')$. Since p is irreducible, there exists n_1, n_2 such that $p^{n_1}(x, y) > 0$ and $p^{n_2}(x', y') > 0$. Moreover, by lemma 2.46, there exists n_3 such that for any $n \geq n_3$, $p^n(y, y) > 0$ and $p^n(y', y') > 0$. Observe that

$$\begin{aligned}
p^{n_1+n_2+n_3}(x, y) &\geq p^{n_1}(x, y) p^{n_2+n_3}(y, y) > 0, \\
p^{n_1+n_2+n_3}(x', y') &\geq p^{n_2}(x', y') p^{n_1+n_3}(y', y') > 0
\end{aligned}$$

Therefore, since $\hat{p}^n((x, x'), (y, y')) = p^n(x, y) p^n(x', y')$ for any n, $(x, x') \to (y, y')$. ■

Remark 2.48.

(i): We know that states of finite irreducible chain has to be recurrent. (See Theorem 2.20) Therefore, we skip to argue that $\hat{\pi}((x, y)) = \pi(x)\pi(y)$ is an stationary distribution for $\hat{p}$. One can easily argue that $\hat{\pi}$ is a stationary distribution for $\hat{p}$. Then, the Theorem 2.60 implies all states are positive recurrent, and in particular recurrent.

To sum up, with finiteness assumption, we actually do not need to assume existence of the stationary distribution. We can drop the finiteness assumption in the convergence theorem, but then existence of stationary distribution is essential.

(ii): By controlling the probability $\mathbb{P}(\hat{T} > n)$, one can argue that the convergence is exponentially fast.

2.7 Exit Distributions & Exit Times

In this section, we will study the probability of hitting a set of states. This exploration is motivated by various applications. For instance, consider a stochastic process that models the value of a portfolio of financial assets; an investor naturally wants to know the probability that this value reaches a target level (indicative of a successful retirement) before it drops to an unsustainable level (leading to bankruptcy). Similar questions arise in physics (probability of reaching a critical energy level), biology (probability of survival of a species), and other fields.

Definition 2.49. For a given $A \subset \mathcal{S}$, we define the exit time as

$$V_A := \inf\{n \geq 0 : X_n \in A\}$$

If the set A is a singleton $\{a\}$, we write $V_a := V_{\{a\}}$.

Theorem 2.50. *Consider a Markov chain with state space $\mathcal{S}$. Let $A, B \subset \mathcal{S}$ such that $C = \mathcal{S} \setminus (A \cup B)$ is finite. If $\mathbb{P}_c(V_A \wedge V_B < \infty) > 0$ for all $c \in C$, then $h(x) := \mathbb{P}_x(V_A < V_B)$ is the unique bounded solution to*

$$h(a) = 1, \ \forall a \in A, \quad h(b) = 0, \ \forall b \in B, \quad \text{and} \quad h(c) = \sum_{y \in \mathcal{S}} p(c, y)h(y), \ \ \forall c \in C \tag{2.10}$$

Proof. Suppose we have a function $h : \mathcal{S} \to [0, 1]$ satisfying (2.10) and let $T = V_A \wedge V_B$.

We first claim that, since C is finite, Lemma 2.11 implies that $\mathbb{P}_c(T < \infty) = 1$ for all $c \in C$. To argue this, note that it suffices to assume the condition of Lemma 2.11 only for $x \in \mathcal{S} \setminus \{y\}$. Moreover, it does not matter that T is the hitting time of a single element, and the proof works exactly the same for hitting time of a set of states. Thus, Lemma 2.11 applies to $T = V_A \wedge V_B$ if we satisfy the assumption. Now, since C is assumed to be finite, by assumption $\min_{c \in C} \mathbb{P}_c(T < \infty) > a > 0$. Moreover, since

$$\lim_{k \to \infty} \min_{c \in C} \mathbb{P}_c(T \leq k) = \min_{c \in C} \mathbb{P}_c(T < \infty) > a$$

we can find a large enough k for which $\mathbb{P}_c(T \leq k) \geq a > 0$ for all $c \in C$. Thus,

$$\mathbb{P}_c(T = \infty) = \lim_{m \to \infty} \mathbb{P}_c(T > mk) \leq \lim_{m \to \infty} (1 - a)^m = 0$$

With this observation coming from the assumption of the theorem, we are now ready to show that $h(x) = \mathbb{P}_x(V_A < V_B)$;

$$\begin{aligned} h(c) &= \sum_{y \in A \cup B} p(c, y)h(y) + \sum_{y \in C} p(c, y)h(y) \\ &= \sum_{y \in A \cup B} p(c, y)h(y) + \sum_{y \in C} \sum_{z \in \mathcal{S}} p(c, y)p(y, z)h(z) \\ &= \mathbb{E}_c h(X_{T \wedge 2})\mathbf{1}_{\{T=1\}} + \mathbb{E}_c h(X_{T \wedge 2})\mathbf{1}_{\{T \geq 2\}} = \mathbb{E}_c h(X_{T \wedge 2}) \end{aligned}$$

By iterating this n times, one can see that $h(c) = \mathbb{E}_c h(X_{T \wedge n})$. Since h is bounded, and $h(X_{T \wedge n}) \to h(X_T) = \mathbf{1}_{\{X_T \in A\}}$ almost surely, by the Dominated Convergence Theorem,

$$h(c) = \mathbb{E}_c h(X_{T \wedge n}) = \lim_{n \to \infty} \mathbb{E}_c h(X_{T \wedge n}) = \mathbb{E}_c h(X_T) = \mathbb{E}_c \mathbf{1}_{\{X_T \in A\}} = \mathbb{P}_c(V_A < V_B)$$

which concludes the result. ∎

Example 2.51 (Gambler's Ruin)**.** Recall that the state space is $\{0, \cdots, N\}$, where $p(x, x+1) = p$ and the Markov chain stops at 0 or N. Let

$$h(x) = \mathbb{P}_x(V_N < V_0)$$

be the probability of winning. To find this explicitly, we will solve (2.10);

$$h(x) = ph(x+1) + (1-p)h(x-1)$$

Rearrange this equation to get,

$$h(x+1) - h(x) = \frac{1-p}{p}(h(x) - h(x-1)) = \cdots = \Big(\frac{1-p}{p}\Big)^x h(1)$$

Now, sum over x to obtain the value of $h(1)$;

$$1 = h(N) - h(0) = h(1) \sum_{x=0}^{N-1} \Big(\frac{1-p}{p}\Big)^x = h(1) \begin{cases} \frac{1-(\frac{1-p}{p})^N}{1-\frac{1-p}{p}} & \text{if } p \neq 1/2 \\ N & \text{if } p = 1/2 \end{cases}$$

Similarly, summing only up to x, and using the value of $h(1)$, we conclude

$$\mathbb{P}_x(V_N < V_0) = \begin{cases} \frac{1-(\frac{1-p}{p})^x}{1-(\frac{1-p}{p})^N} & \text{if } p \neq 1/2 \\ \frac{x}{N} & \text{if } p = 1/2 \end{cases}$$

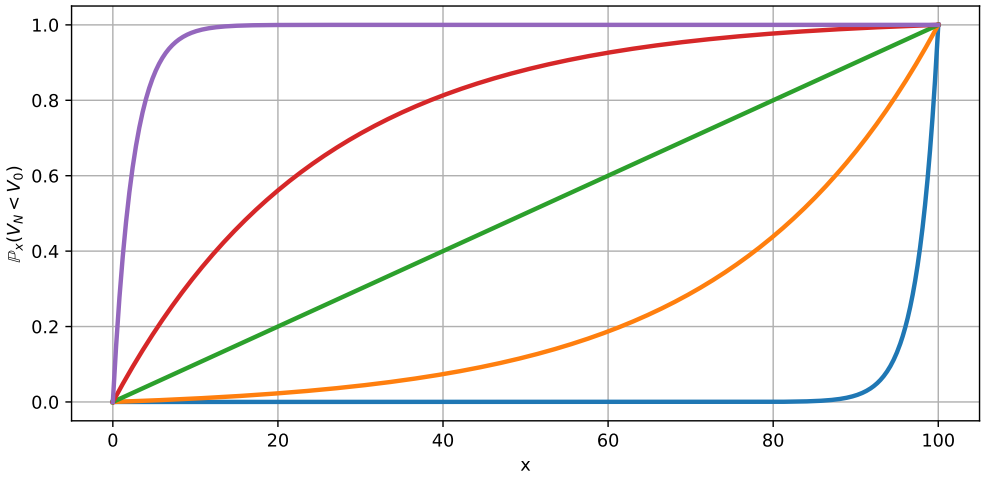

Figure 1: The figure shows the plot of $\mathbb{P}_x(V_N < V_0)$ for $N = 100$ and $p \in \{0.4, 0.49, 0.5, 0.51, 0.6\}$.

Now, it is also quite important to understand the expected time to hit a subset of states. For this purpose, we will now work out the expected of any transition function up to the hitting time;

Theorem 2.52. *Consider a Markov chain with state space $\mathcal{S}$. Let $A \subset \mathcal{S}$ such that $C = \mathcal{S} \setminus A$ is finite, and $f : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ is non-negative. If $\mathbb{P}_c(V_A < \infty) > 0$ for all $c \in C$, then*

$$g(x) := \mathbb{E}_x\Big[\sum_{m=1}^{V_A} f(X_{m-1}, X_m)\Big] = \mathbb{E}_x\Big[\sum_{m=1}^{\infty} f(X_{m-1}, X_m)\mathbf{1}_{\{V_a \geq m\}}\Big]$$

is the unique bounded solution to

$$g(a) = 0, \ \forall a \in A, \quad g(c) = \sum_{y \in \mathcal{S}} p(c,y) f(c,y) + \sum_{y \in \mathcal{S}} p(c,y) g(y), \ \forall c \in C \tag{2.11}$$

In particular, for $f \equiv 1$, *we have* $g(x) = \mathbb{E}_x[V_A]$.

Proof. Since C is finite, Lemma 2.11 similarly implies $\mathbb{P}_c(V_A < \infty) = 1$ for all $c \in C$. Now, suppose a bounded g satisfies (2.11), and then iterate as before to see the pattern;

$$\begin{aligned} g(c) &= \sum_{y \in \mathcal{S}} p(c,y) f(c,y) + \sum_{y \in C} p(c,y) g(y) \\ &= \sum_{y \in \mathcal{S}} p(c,y) f(c,y) + \sum_{y \in C, z \in \mathcal{S}} p(c,y) p(y,z) f(y,z) + \sum_{y \in C, z \in C} p(c,y) p(y,z) g(z) \\ &= \cdots = \mathbb{E}_c\Big[\sum_{m=1}^{n} f(X_{m-1}, X_m) \mathbf{1}_{\{V_A \geq m\}} \Big] + \mathbb{E}_c[g(X_n) \mathbf{1}_{\{V_A > n\}}] \end{aligned}$$

Note that the first equality uses $g(a) = 0$ for all $a \in A$. Since f is non-negative, the Monotone Convergence Theorem yields the convergence of the first term. Since g is bounded, the Dominated Convergence Theorem, together with the fact that $\mathbb{P}_c(V_A < \infty) = 1$ concludes that the second term vanishes in the limit $n \to \infty$. ■

Example 2.53 (Gambler's Ruin)**.** Let us present the expected time of the game, that is when $A = \{0, N\}$,

$$\mathbb{E}_x[V_A] = \begin{cases} x(N-x) & \text{if } p = 1/2 \\ \frac{x}{1-2p} - \frac{N}{1-2p} \frac{1-(\frac{1-p}{p})^x}{1-(\frac{1-p}{p})^N} & \text{if } p \neq 1/2 \end{cases}$$

The case of $p = 1/2$ can be derived from $g(x) = 1 + (1/2)g(x+1) + (1/2)g(x-1)$ by rearranging and telescoping sum. The case of $p \neq 1/2$, although doable, is tedious to derive and we will omit. However, it is straightforward to check the condition (2.11) for the above solution.

Lets look at check some limit behaviors. If the winning probability $p \to 0$, then

$$\mathbb{P}_x(V_N < V_0) = \frac{1 - (\frac{1-p}{p})^x}{1 - (\frac{1-p}{p})^N} \to 0 \quad \text{and} \quad \mathbb{E}_x[V_A] \to x$$

which is expected. If $p > 1/2$ and the target capital $N \to \infty$,

$$\mathbb{P}_x(V_N < V_0) \to 1 - \Big(\frac{1-p}{p}\Big)^x \quad \text{and} \quad \mathbb{E}_x[V_A] \to \infty$$

That is, when winning probability is in favor, reaching infinitely large capital is possible. Not with probability one but strictly positive. Of course, expected time blows up to infinity in this case. On the other hand, if $p < 1/2$, then

$$\mathbb{P}_x(V_N < V_0) \to 0 \quad \text{and} \quad \mathbb{E}_x[V_A] \to \frac{x}{1-2p}$$

meaning that there is no chance to achieve infinitely large target.

As it is somewhat counter-intuitive, let's look at the case $p = 1/2$, $N = 100$ and we start from $x = 99$. Then $\mathbb{P}_{99}(V_{100} < V_0) = 99/100$, which is quite high as expected. However, notice that $\mathbb{E}_{99}[V_{\{0,100\}}] = 99$. That is, just to hit 100 (or sometimes 0) starting from 99, one has to take 99 steps on average.

Suggested Exercises. Durrett, 3rd edition. 1.45, 1.49, 1.57, 1.62, 1.67.

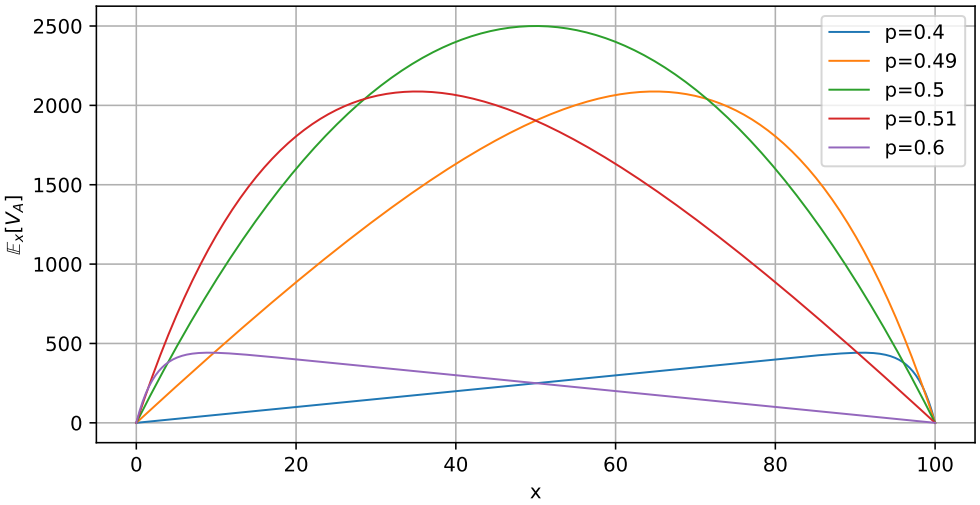

Figure 2: The figure shows the plot of $\mathbb{E}_x V_A$ for $N = 100$ and $p \in \{0.4, 0.49, 0.5, 0.51, 0.6\}$.

2.8 Ergodic Theorem

Theorem 2.54 (Ergodic Theorem)**.** *Suppose the Markov chain is irreducible and has a stationary distribution π. Let $f : \mathcal{S} \to \mathbb{R}$ be integrable with respect to π, that is, $\sum_{x\in\mathcal{S}} |f(x)|\pi(x) < \infty$. Then*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n} f(X_k) = \mathbb{E}_\pi[f] := \sum_{x\in\mathcal{S}} f(x)\pi(x)$$

almost surely under any initial distribution.

Proof. Fix $x \in \mathcal{S}$ and suppose $X_0 = x$ almost surely. Recall the definition of k-th return time T_x^k, and notice that by the strong Markov property,

$$\sum_{m=T_x^k+1}^{T_x^{k+1}} f(X_m)$$

are all independent and identically distributed (i.i.d.). Therefore, we can invoke the Law of Large Numbers, but first, let us compute the expectation

$$\begin{aligned}
&\mathbb{E}_x\Big[\sum_{m=T_x^k+1}^{T_x^{k+1}} f(X_m)\Big] = \mathbb{E}_x\Big[\sum_{m=1}^{T_x} f(X_m)\Big] \\
&= \sum_{m=1}^{\infty} \mathbb{E}_x\big[\mathbf{1}_{\{T_x\geq m\}} f(X_m)\big] = \sum_{m=1}^{\infty} \mathbb{E}_x\Big[\sum_{z\in\mathcal{S}} \mathbf{1}_{\{X_m=z\}}\mathbf{1}_{\{T_x\geq m\}} f(z)\Big] \\
&= \sum_{m=1}^{\infty}\sum_{z\in\mathcal{S}} \mathbb{P}_x(X_m = z, T_x \geq m) f(z) = \sum_{z\in\mathcal{S}} \mu_x(z) f(z)
\end{aligned}$$

where μ_x is defined as in Theorem 2.34. Thus, the Law of Large Numbers implies

$$\frac{1}{n}\sum_{k=0}^{n-1}\Big(\sum_{m=T_x^k+1}^{T_x^{k+1}} f(X_m)\Big) \to \sum_{z\in\mathcal{S}} \mu_x(z) f(z)$$

almost surely.[4] Again by the Law of Large Numbers, since $T_x^n = \sum_{m=1}^n T_x^m - T_x^{m-1}$,

$$\frac{T_x^n}{n} \to \mathbb{E}_x[T_x]$$

almost surely.[5] Therefore,

$$\frac{1}{T_x^n}\sum_{m=1}^{T_x^n} f(X_m) = \frac{n}{T_x^n}\frac{1}{n}\sum_{k=0}^{n-1}\sum_{m=T_x^k+1}^{T_x^{k+1}} f(X_m) \to \frac{1}{\mathbb{E}_x[T_x]}\sum_{z\in\mathcal{S}}\mu_x(z)f(z) = \sum_{z\in\mathcal{S}}\pi(z)f(z)$$

almost surely. This is almost what we want, and completes the core part of the proof. However, we only showed the convergence for a subsequence, not the full sequence.

Suppose $|f| < C$. Then lemma 8.1 implies the result by noting

$$\lim_{n\to\infty}\frac{T_x^{n+1}}{T_x^n} = \lim_{n\to\infty}\frac{T_x^{n+1}}{n+1}\cdot\frac{n+1}{n}\cdot\frac{n}{T_x^n} = 1 \quad \text{almost surely.}$$

Next, truncate f as $f^C(x) := f(x)\mathbf{1}_{\{|f(x)|\le C\}}$. Then, note that $\sum_{x\in\mathcal{S}} f^C(x)\pi(x) \to \sum_{x\in\mathcal{S}} f(x)\pi(x)$ as $C\to\infty$ by the Dominated Convergence Theorem. Hence, it suffices to control

$$\begin{aligned}\left|\frac{1}{n}\sum_{m=0}^n f(X_m) - \frac{1}{n}\sum_{m=0}^n f^C(X_m)\right| &\le \frac{1}{n}\sum_{m=0}^n |f(X_m)|\mathbf{1}_{\{|f(X_m)|>C\}}\\ &\le \frac{1}{n}\sum_{m=0}^{T_x^n} |f(X_m)|\mathbf{1}_{\{|f(X_m)|>C\}}\end{aligned}$$

Multiply the inequality by $\mathbb{E}_x[T_x](n/T_x^n)$ and send $n\to\infty$. By the initial arguments, we know that the limit exists over the sequence of return times, and hence again by Dominated Convergence Theorem, it goes to 0 as $C\to\infty$.

Note that we argue the result for a chain starting from an initial condition x. Suppose $\mathbb{P}$ corresponds to arbitrary initial measure ν,

$$\mathbb{P}\left(\lim_{n\to\infty}\frac{1}{n}\sum_{k=0}^n f(X_k) = \mathbb{E}_\pi f\right) = \sum_{z\in\mathcal{S}}\mathbb{P}_z\left(\lim_{n\to\infty}\frac{1}{n}\sum_{k=0}^n f(X_k) = \mathbb{E}_\pi f\right)\nu(z) = 1$$

■

Theorem 2.55 (Asymptotic Frequency)**.** *Consider an irreducible Markov chain where all the states are recurrent. Define the number of visits to x up to time n,*

$$N_n(x) := \sum_{m=0}^n \mathbf{1}_{\{X_m=x\}}$$

Then,

$$\lim_{n\to\infty}\frac{N_n(x)}{n} = \frac{1}{\mathbb{E}_x[T_x]} \quad \textit{almost surely.}$$

[4] Set $T_x^0 = 0$.

[5] In Theorem 2.60, we will show that since we assumed there exists a stationary distribution, $\mathbb{E}_x[T_x] < \infty$ for all $x\in\mathcal{S}$.

Proof. In the proof of the Ergodic Theorem, we have already observed

$$\frac{T_x^n}{n} \to \mathbb{E}_x[T_x]$$

This holds even if $\mathbb{E}_x[T_x] = \infty$ since $T_x^m - T_x^{m-1}$'s are positive (and $\mathbb{E}[T_x^m - T_x^{m-1}] \in [0, \infty]$ exists). Observe,

$$T_x^k = \min\{n \geq 1 : N_n(x) = k\}$$

hence

$$T_x^{N_n(x)} = \min\{k \geq 1 : N_k(x) = N_n(x)\} \leq n$$

and similarly

$$n < T_x^{N_n(x)+1} = \min\{k \geq 1 : N_k(x) = N_n(x) + 1\}$$

Therefore,

$$\frac{T_x^{N_n(x)}}{N_n(x)} \leq \frac{n}{N_n(x)} \leq \frac{T_x^{N_n(x)+1}}{N_n(x)+1} \frac{N_n(x)+1}{N_n(x)}$$

and sending $n \to \infty$ concludes the result. ∎

Remark 2.56. **(i):** Note that, if we set $f(y) = \mathbf{1}_{\{y=x\}}$ for some $x \in \mathcal{S}$, the Ergodic theorem implies

$$\frac{N_n(x)}{n} \to \pi(x) = \frac{1}{\mathbb{E}_x[T_x]}$$

as expected. However, we do not need the existence of the stationary measure for the asymptotic frequency to hold.

(ii): Taking the expectation and interchanging the limit yields (Bounded Convergence Theorem)

$$\frac{1}{n} \sum_{m=0}^{n} p^m(x, y) \to \frac{1}{\mathbb{E}_y T_y}$$

Therefore, even without aperiodicity, the average of p^m converges. In case π exists, it is the right limit.

Suggested Exercises. Durrett, 3rd edition. 1.21, 1.39.

2.9 Existence and Uniqueness in Countable State Space

In this section, we state the existence and uniqueness theorems without the finiteness assumption. We invite the reader to review their proofs in the finite case, as we have already presented arguments suitable for the countable case.

Let us start with the following proposition first,

Proposition 2.57. *If π is a stationary distribution, then each state x with $\pi(x) > 0$ is recurrent.*

Proof. Suppose $\pi(y) > 0$. By (2.3),

$$\sum_{x \in \mathcal{S}} \pi(x) \mathbb{E}_x N(y) = \sum_{n \geq 1} \sum_{x \in \mathcal{S}} \pi(x) p^n(x, y) = \sum_{n \geq 1} \pi(y) = \infty$$

On the other hand, by (2.4),

$$\infty = \sum_{x \in \mathcal{S}} \pi(x) \mathbb{E}_x N(y) = \sum_{x \in \mathcal{S}} \pi(x) \frac{\rho_{xy}}{(1 - \rho_{yy})} \leq \frac{1}{(1 - \rho_{yy})}$$

hence assuming $\rho_{yy} < 1$ is a contradiction. ∎

Corollary 2.58. *An irreducible Markov chain where all states are transient cannot have a stationary distribution.*

Recall that previously we have used the finiteness assumption to have $\mathbb{E}_x[T_x] < \infty$. In the countable case, we will instead assume this;

Definition 2.59. A recurrent state x is called positive recurrent if $\mathbb{E}_x[T_x] < \infty$, and is called null recurrent if $\mathbb{E}_x[T_x] = \infty$.

Theorem 2.60 (Existence and Uniqueness)**.** *Suppose the Markov chain is irreducible. It has a stationary distribution π if and only if all states are positive recurrent. Moreover, if a stationary distribution exists, then it is unique and*

$$\pi(x) = \frac{1}{\mathbb{E}_x[T_x]}$$

Proof. In fact, we do not need to do any extra work for this proof. Let us show how it follows from previous arguments.

Suppose it has a stationary distribution. Then at least for one state $\pi(x) > 0$, and that x is recurrent. Recall μ_x from the Theorem 2.34, where we argued it is a stationary measure but might not have mass 1. (hence not a distribution in general). Moreover, in the proof of Theorem 2.36 we have shown that any stationary measure has to be a constant multiple of μ_x. (See following remarks of theorems.)

$$\mu_x(y) = c\pi(y) \text{ implies } 1 = \mu_x(x) = c\pi(x)$$

and hence $c = 1/\pi(x) > 0$. Recall that $\mathbb{E}_x[T_x] = \sum_{y\in\mathcal{S}} \mu_x(y)$, and hence

$$\mathbb{E}_x[T_x] = \sum_{y\in\mathcal{S}} \mu_x(y) = \frac{1}{\pi(x)} < \infty$$

That is, x is positive recurrent for any $\pi(x) > 0$. Now, since the chain is irreducible, for any y there exists n such that $p^n(x, y) > 0$. Therefore,

$$\pi(y) = \sum_{z\in\mathcal{S}} \pi(z)p^n(z, y) \geq \pi(x)p^n(x, y) > 0$$

and hence all the states are positive recurrent.

Assuming that all states are positive recurrent, proofs of Theorem 2.34 and 2.36 holds to be true without any modification. Therefore, there exists a stationary distribution π and it is unique. ■

Example 2.61 (Reflecting Random Walk)**.** Let $\mathcal{S} = \mathbb{N}$. Let the transition matrix be

$$p(x, x+1) = p,\ p(x, x-1) = 1-p,\ \forall x \geq 1,\ p(0, 0) = 1-p$$

It is obviously irreducible for $0 < p < 1$. Since $p(0, 0) > 0$, it is aperiodic. By the Lemma 2.44, the whole chain is aperiodic.

Case $p < 1/2$: Let us recall the detailed balance condition for this Markov chain,

$$p\pi(x) = (1-p)\pi(x+1),\ \forall x \geq 0$$

That is,

$$\pi(x) = \left(\frac{p}{1-p}\right)^x \pi(0)$$

Since $p < 1/2$,

$$Z := \sum_{x \geq 0} \left(\frac{p}{1-p}\right)^x = \frac{1-p}{1-2p} < \infty$$

Therefore, $\pi(0) = 1/Z$ and we obtained a stationary distribution. The Convergence Theorem 2.47 implies $\mathbb{P}_x(X_n = y) \to \pi(y)$. (See the following remark for the finiteness assumption.) Moreover, Theorem 2.60 yields,

$$\mathbb{E}_0[T_0] = \frac{1}{\pi(0)} = \frac{1-p}{1-2p}$$

Case $p > 1/2$: In this case, we will argue that 0 is a transient state. Since Markov chain is irreducible, then by the Lemma 2.25, all states are transient. Notice that return time T_0 and exit time V_0 coincide if we do not start from state 0. Therefore,

$$\rho_{x0} = \mathbb{P}_x(T_0 < \infty) = \mathbb{P}_x(V_0 < \infty)$$

We have already computed in the Example 2.51 that

$$\mathbb{P}_x(V_0 < V_N) = 1 - \mathbb{P}_x(V_N < V_0) = \begin{cases} 1 - \frac{1-(\frac{1-p}{p})^x}{1-(\frac{1-p}{p})^N} & \text{if } p \neq 1/2 \\ 1 - \frac{x}{N} & \text{if } p = 1/2 \end{cases}$$

which implies

$$\rho_{x0} = \mathbb{P}_x(V_0 < \infty) = \lim_{N \to \infty} \mathbb{P}_x(V_0 < V_N) = \left(\frac{1-p}{p}\right)^x < 1$$

On the other hand, obviously $\rho_{0x} > 0$. Then the Theorem 2.18 implies 0 is transient. Now, the Corollary 2.58 implies there exists no stationary distribution.

Case $p = 1/2$: By similar arguments as in the previous case,

$$\mathbb{P}_x(V_0 < \infty) = 1, \quad \forall x \in \mathbb{N}$$

Now,

$$\mathbb{P}_0(T_0 < \infty) = (1/2)\mathbb{P}_0(V_0 < \infty | X_1 = 1) + (1/2)\mathbb{P}_0(T_0 < \infty | X_1 = 0) = (1/2) \cdot 1 + (1/2)\mathbb{P}_0(T_0 < \infty)$$

which means $\mathbb{P}_0(T_0 < \infty) = 1$. Since the chain is irreducible, all states are recurrent. However, we will argue that states are in fact null recurrent. To do so,

$$\mathbb{E}_1 V_0 \geq \mathbb{E}_1 V_{\{0,N\}} = 1(N-1) \to \infty$$

hence $\mathbb{E}_1 V_0 = \infty$. Similarly,

$$\mathbb{E}_0[T_0] = (1/2)\mathbb{E}_0[V_0 | X_1 = 1] + (1/2)\mathbb{E}_0[T_0 | X_1 = 0] = \infty$$

Since the chain is irreducible, and 0 is null recurrent, there cannot be a stationary distribution by the Theorem 2.60.

As a side note, symmetric random walk in one and two dimensions are recurrent. However, in three dimension it becomes transient. Roughly speaking $p^n(x,x) \sim 1/n^{d/2}$ where d is the dimension, and hence is not summable for $d = 1, 2$.

Suggested Exercises. Durrett, 3rd edition. 1.70, 1.74, 1.75.

3 Steps to Reinforcement Learning

3.1 Markov Decision Process

In this section, we will learn to formalize decision making where the underlying dynamics are modeled as a Markov Chain. This is the first basic step toward Reinforcement Learning.

It is important to emphasize that every system can be modeled as Markovian. Although it might not always be useful, particularly if the state space is chosen too small, many systems can indeed be well approximated. To motivate this idea, consider a simple example where $X_0, X_1 \in \{0, 1\}$ independently and uniformly, and $X_2 = 1$ if $(X_1 = 1, X_0 = 1)$ and 0 otherwise. If an agent, such as a robot, can only observe X_1, then

$$\mathbb{P}(X_2 = 1 | X_1 = 1) = \frac{1}{2}, \quad \mathbb{P}(X_2 = 0 | X_1 = 0) = 1$$

This is what the agent approximates from the real dynamics. In the agent's experience, the above perfectly defines the dynamics, and the agent can learn to behave optimally. Although the dynamics of X_2 are completely deterministic with sufficient information (i.e., (X_0, X_1)), the agent approximates it by Markovian dynamics. Although this approximation does not guarantee useful optimal solutions, it can be done more or less independent of the system's nature. Additionally, this simple example also motivates that the choice of the state space plays a role in how well we can approximate.

Let us define what a Markov Decision Process (MDP) is, and then we will present various examples.

Definition 3.1. A Markov Decision Process is a 4-tuple $(\mathcal{S}, \mathbb{A}, P, R)$ where $\mathcal{S}$ is the state space of the Markov chain, $\mathbb{A}$ is the set of actions, $P : \mathbb{A} \times \mathcal{S} \times \mathcal{S} \to [0, 1]$ determines the transition probabilities

$$\sum_{y \in \mathcal{S}} P_a(x, y) = 1;$$

and $R : \mathcal{S} \times \mathbb{A} \to \mathbb{R}$ is the immediate reward function.

- **Maneuvering a helicopter** [continuous state & continuous actions]
  – The state space might include parameters of its motor and rotors, wind speed/direction, height, etc.
  – Dynamics are determined by the real world.
  – Actions are determined by the controller of a helicopter.
  – The objective might be to learn to make complicated maneuvers, where the reward could be following given paths and the cost might be a crash.

- **Playing backgammon** [discrete state & discrete actions]
  – The state space could be the positions of checkers.
  – Dynamics are deterministic given the result of the dice, however, they appear random to the player!
  – Actions follow the rules of backgammon.
  – The objective is positive for winning and negative for losing.

- **Other games**
  – One can think of games with discrete action spaces but continuous state spaces (like resources).

- **Controlling robots**
  – Examples include humanoid robots learning to walk, robotic arms making surgical moves, and cars learning to self-drive. There are many different robots/machines that can learn to act optimally.

- **Cooling systems**
  – Google used reinforcement learning to cool down massive data servers. The objective was to minimize energy consumption while maintaining safety protocols, achieving a reduction in consumption by 40%.

- **Traffic control**
  – Traffic optimization can be achieved by controlling traffic lights, which can again be learned through reinforcement learning methods.

- **Finance**
  – One can learn to understand price actions and how to behave optimally, deciding buy/sell actions in an optimal way.

Now, let us continue to detail the MDP. First, since transition probabilities depend on the actions taken, we need to provide a future strategy to be able to define a Markov Chain. To achieve this, let

$$\mathcal{A} := \{\alpha : \mathbb{N} \times \mathcal{S} \to \mathbb{A}\}$$

be the set of so-called Markov policies (or controls). One can consider different choices for the set of admissible controls; for example, it can depend on past states. However, in our simplified setting, it can be shown that the optimal policy takes this form. Given a strategy (or policy, control) $\alpha \in \mathcal{A}$, we define a Markov Chain by

$$\mathbb{P}(X_{n+1} = y | X_n = x) = P_{\alpha(n,x)}(x, y)$$

We denote this chain by X_n^α, representing the future distribution of the state process under the strategy α.

Next, to discuss which strategy is "optimal", we need to define the objective of the agent. To do so, we assign a value (or cost) for each strategy as follows;

$$U(x, \alpha) := \mathbb{E}_x\Big[\sum_{n\geq 0} \lambda^n R(X_n^\alpha, \alpha(n, X_n^\alpha))\Big]$$

where $0 < \lambda < 1$ is the discount factor. Note that if R is bounded, λ essentially sets a horizon for the problem. Indeed, if we do not discount, this expectation would not even depend on the state x.

We are ready to introduce the value function, which plays a crucial role as it satisfies a dynamic backward equation and allows us to construct optimal policies.

$$V(x) = \sup_{\alpha\in\mathcal{A}} U(x, \alpha) \tag{3.1}$$

The equation that the value V satisfies is derived as follows:

$$\begin{aligned}
&\mathbb{E}_x\Big[\sum_{n\geq 0} \lambda^n R(X_n^\alpha, \alpha(n, X_n^\alpha))\Big] \\
&= R(x, \alpha(0, x)) + \mathbb{E}_x\Big[\sum_{n\geq 1} \lambda^n R(X_n^\alpha, \alpha(n, X_n^\alpha))\Big] \\
&= R(x, \alpha(0, x)) + \lambda\mathbb{E}_x\Big[\sum_{n\geq 0} \lambda^n R(X_{n+1}^\alpha, \alpha(n+1, X_{n+1}^\alpha))\Big] \\
&= R(x, \alpha(0, x)) + \lambda \sum_{z\in\mathcal{S}} P_{\alpha(0,x)}(x, z)\mathbb{E}_z\Big[\sum_{n\geq 0} \lambda^n R(X_n^\alpha, \alpha(n+1, X_n^\alpha))\Big]
\end{aligned}$$

Here, X_n^α in the last expectation is using the shifted control, starting from time 1 instead of 0. Considering the supremum over controls α, we obtained the Bellman Equation;

$$V(x) = \sup_{a\in\mathbb{A}} \Big\{R(x, a) + \lambda \sum_{z\in\mathcal{S}} P_a(x, z)V(z)\Big\} \tag{3.2}$$

Theorem 3.2. *Suppose the reward function R is bounded. Then V defined in (3.1) is the unique bounded solution to the Bellman equation (3.2). Moreover, if there exists $I^* : \mathcal{S} \to A$ such that*

$$I^*(x) \in \arg\max_{a\in\mathbb{A}} \left\{ R(x,a) + \lambda \sum_{z\in\mathcal{S}} P_a(x,z)V(z) \right\}$$

Then $\alpha(n,x) := I^*(x)$ is the stationary optimal control.

Proof. We already know that V is a solution to the Bellman equation. Let $B(\mathcal{S};\mathbb{R})$ be the set of bounded functions equiped with the supremum metric. Define $T : B(\mathcal{S};\mathbb{R}) \to B(\mathcal{S};\mathbb{R})$ as

$$T(U)(x) := \sup_{a\in\mathbb{A}} \left\{ R(x,a) + \lambda \sum_{z\in\mathcal{S}} P_a(x,z)U(z) \right\}$$

Note that T is bounded. We will argue that T is a contraction mapping. Then by the Banach Fixed Point Theorem 8.3, V is the unique bounded solution to the Bellman equation. To show the contraction, we first recall a general inequality from Lemma 8.2

$$|\sup_a f(a) - \sup_a g(a)| \le \sup_a |f(a) - g(a)|$$

Then, the contraction property follows as;

$$\begin{aligned}
\sup_{x\in\mathcal{S}} |T(U)(x) - T(\tilde{U})(x)| &\le \sup_{x\in\mathcal{S}} \sup_{a\in\mathbb{A}} \left\{ \lambda \sum_{z\in\mathcal{S}} P_a(x,z)|U(z) - \tilde{U}(z)| \right\} \\
&\le \lambda \Big(\sup_{z\in\mathcal{S}} |U(z) - \tilde{U}(z)| \Big) \sup_{x\in\mathcal{S}} \sup_{a\in\mathbb{A}} \left\{ \sum_{z\in\mathcal{S}} P_a(x,z) \right\} \\
&\le \lambda \sup_{x\in\mathcal{S}} |U(x) - \tilde{U}(x)|
\end{aligned}$$

which concludes T is a contraction mapping.

To show the last claim, referred typically as verification, for $\alpha(n,x) := I^*(x)$ the Bellman equation computed at X_n^α becomes

$$V(X_n^\alpha) = R(X_n^\alpha, I^*(X_n^\alpha)) + \lambda \mathbb{E}_{X_n^\alpha}[V(X_{n+1}^\alpha)] = R(X_n^\alpha, I^*(X_n^\alpha)) + \lambda \mathbb{E}_x[V(X_{n+1}^\alpha)|X_n^\alpha]$$

Multiply this equation by λ^n, sum over n and take the expectation $\mathbb{E}_x$[6];

$$\begin{aligned}
\sum_{n=0}^{\infty} \lambda^n \mathbb{E}_x\big[V(X_n^\alpha)\big] &= \mathbb{E}_x\Big[\sum_{n=0}^{\infty} \lambda^n R(X_n^\alpha, \alpha(n, X_n^\alpha)) \Big] + \mathbb{E}_x\Big[\sum_{n=0}^{\infty} \lambda^{n+1} \mathbb{E}_x[V(X_{n+1}^\alpha)|X_n^\alpha] \Big] \\
&= U(x,\alpha) + \sum_{n=0}^{\infty} \lambda^{n+1} \mathbb{E}_x[V(X_{n+1}^\alpha)]
\end{aligned}$$

by the tower property of the conditional expectation (See Proposition 7.41). That is,

$$V(x) = U(x,\alpha)$$

and α satisfies the supremum in (3.1), concluding that it is an optimal control. ∎

[6] Also, use Dominated Convergence Theorem since V is bounded.

Remark 3.3. (i): Although it is impossible in practice, notice that one can iterate the contraction mapping T, starting from any initial guess for V, and obtain the value function hence the optimal control.

(ii): It is important to note that Bellman equation the (3.2) is not only optimizing the immediate reward, but considers the future potential values. Of course, we cannot just take actions to maximize our immediate reward and it is reflected in the equation.

Example 3.4 (Consumption Problem)**.** Suppose an agent has a random capital X_n, and at each time the agent consumes some amount of it while investing the rest. Suppose the control denotes the amount of consumption. Let $R(x,a)$ be reward function, or utility of agent for having x amount of money and consuming a amount of money. Agent might aim to optimize

$$\mathbb{E}_x\Big[\sum_{n\geq 0}\Big(\frac{1}{1+r}\Big)^n R(X_n^\alpha, \alpha(n, X_n^\alpha))\Big]$$

where r is the interest rate. Notice that, even if the dynamics of Markov chain is independent of the control, Bellman equation would be

$$V(x) = \sup_{a\in\mathbb{A}}\Big\{R(x,a) + \lambda\sum_{z\in\mathcal{S}} p(x-a,z)V(z)\Big\}$$

3.2 Reinforcement Learning

The concept of Markov Decision Process motivates and teaches us how to act optimally given the parameters of the problem. However in real life, we almost never know what the parameters are. We might not know how the process transitions, or what are the rewards. For example, if you are playing chess, what is the reward of transitioning from one state to another? "Values" of pieces might guide you, but of course it is only a primitive guide, mastering chess is way harder than learning the assigned values of pieces. Another example that many of us are familiar is video games with unfamiliar dynamics. As a player, we have to learn how we move.

To attack this problem, Reinforcement Learning (RL) defines a concept of player. We will not get into the technical details, but in words, there are sequences of observations, sequences of potential actions (or strategies) and parameters that player is learning. The choices for what player is learning is endless, it can include transitions and rewards, but beyond that can include estimating other player's behavior, communications, useful embeddings of states, and many more depending on the complexity of the problem. We will consider one of the simplest cases where the player tries to learn a value associated to a pair of state and action (x,a). This is typically referred as Q-learning. Moreover, we assume that there are finitely many states and actions with managable cardinalities. To approximate continuous state or action spaces, one needs to use function approximations like neural networks.

Let us also point out that randomization (or exploration) is the key element of learning. The whole life emerges through molecules exploring infinite amount of structural designs (proteins), presenting the self-replicating long term stable solutions. In our context, since we start by not knowing the Q-function, we shouldn't solely rely on it to construct our strategies and need to introduce ways of promoting exploration.

Now, we discuss a generic structure of an algorithm to learn the Q-function below. Here, we assume that we don't know the transition probabilities P or the reward function R, but observe it as we interact with the environment.[7]

- Initially choose an arbitrary strategy that you will do at any state x. It is better to choose one with random actions instead of deterministic ones, so that you can explore.

[7]Typically, we design the reward function R to lead the player (or agent). However, in the standard Q-learning, this is not known by the agent.

- The aim is to approximate a useful Q. There are two general paths (with many potential variations);
  – Monte Carlo: Generate a lot of paths from your learned model, set Q as the average of rewards you get. (Advantage: versatile, Disadvantage: large variance)
  – TD: Generate only one step from your learned model, set

$$Q(x,a) \leftarrow Q(x,a) + a(R + \lambda Q(z,a') - Q(x,a))$$

  if we started from state x, take the action a, ended up in state z, got a reward R, and planned to use a'. Note that this approach exploits the Markov structure. (Advantage: Low variance, Disadvantage: Requires some structure)

- Given $Q(x,a)$, update your strategy to be more likely as $\arg\max_a Q(x,a)$. But not exactly! Because we don't have the best Q, so what we determine as the current best might be a very bad strategy (which initially is indeed a very bad guess), and if we don't try to explore we might get stuck with it.

- Repeat this procedure as follows. Choose an initial starting position of interest. Generate a lot of paths from your learned model and update Q values of all states from those experiences. Learn a better policy from the new Q. Repeat it again by sampling a lot of paths.

- Notice that we haven't used the functions P or R at all! This is called "model-free" and standard Q-learning refers to this. In a sense, we used samples from the Bellman equation to avoid the use of P and R. Depending on the problem, we can learn these.

  (i): For any state $x \in \mathcal{S}$, $a \in \mathbb{A}$, generate real life experiences (or simulations). That is, generate new states from your model, starting from state x using the action a. By looking at frequency of occurences, we can approximate the $P_a(x,\cdot)$. This is useful even if we know what P_a actually is. If the dynamics are not simple, as we are learning from real experiences, we essentially learn what occurs frequently. If there are vast number of states for which many of them are extremely unlikely to transition, trying to implement the full transition probability P_a might be untrackable.

  (ii): The function $R(x,a)$ might be already known. Even if not, we can use the same logic as in (i) and keep track of expected reward from our simulations. Moreover, we might want to model $R(x,a,z)$ to incorporate which state we are transitioning to.

  Now, since we have access to the functions P and R, we can make updates smoother by using the Bellman equation directly;

$$Q(x,a) \leftarrow \sum_{z \in \mathcal{S}} P_a(x,z)\big[R(x,a,z) + \lambda \max_{\tilde{a}} Q(z,\tilde{a})\big]$$

Let us conclude by pointing out that the design of a player might be remarkably complex, as it can potentially learn and behave as complex as a human. Here, we scratched the surface and introduced a player as just trying to learn a single function Q.

4 Poisson Processes

4.1 Exponential and Poisson Distribution

Let us recall the properties of the Exponential distribution, which has a significant importance as being the unique distribution without a memory. This will play a crucial role to introduce the fundamental jump process, Poisson process, and will allow us to characterize continuous time discrete state Markov processes.

Definition 4.1 (Exponential Distribution)**.** A random variable X is said to have an exponential distribution with parameter λ, denoted as $X \sim \text{Exp}(\lambda)$, if

$$\mathbb{P}(X \leq t) = \left[1 - e^{-\lambda t}\right]\mathbf{1}_{\{t \geq 0\}}$$

The probability density function of X is given by

$$f_X(t) = \partial_t \mathbb{P}(X \leq t) = \lambda e^{-\lambda t}\mathbf{1}_{\{t \geq 0\}}$$

To find the expected value, one can use integration by parts to compute

$$\mathbb{E}[X] = \int_0^\infty t f_X(t) dt = \frac{1}{\lambda}$$

or (2.2) as

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X \geq t) dt = \frac{1}{\lambda}$$

Similarly, one can compute $\mathbb{E}[|X|^2] = 2/\lambda^2$ and hence $\text{Var}(X) = \mathbb{E}[|X|^2] - (\mathbb{E}[X])^2 = 1/\lambda^2$. Let us also note the scaling property: if $X \sim \text{Exp}(\lambda_1)$ and $\lambda_2 > 0$, then

$$\mathbb{P}(X/\lambda_2 \leq t) = \mathbb{P}(X \leq t\lambda_2) = \left[1 - e^{-\lambda_1 \lambda_2 t}\right]\mathbf{1}_{\{t \geq 0\}}$$

that is, $X/\lambda_2 \sim \text{Exp}(\lambda_1 \lambda_2)$. Now, the special property of the exponential distribution is

$$\mathbb{P}(X > t + s | X > t) = \mathbb{P}(X > s) \tag{4.1}$$

That is, it does not matter if an event didn't occured until some time to determine the future probabilities.
Exercise. Exponential distribution is the unique distribution satisfying (4.1).

Let us show that geometric distribution can converge to exponential distribution, which provides a good intuition about memoryless property. Suppose we are flipping a coin with success rate λ/n at each time step $1/n$. Let X_n be the first success time. Then,

$$\mathbb{P}(nX_n = k) = (1 - \lambda/n)^{k-1}(\lambda/n)$$

That is, nX_n is a geometric distribution and we can easily compute the limit as $n \to \infty$;

$$\mathbb{P}(X_n > t) = \mathbb{P}(nX_n > nt) \simeq (1 - \lambda/n)^{nt} \to e^{-\lambda t}$$

One natural event we observe exponential distribution is the particle decay. Given that the radioactive particle has not decayed yet provides no information about when it will decay. There are many other phenomenas that can be modeled by the exponential distributions, such as distance between mutations in DNA, getting a phone call, arrival of costumers etc.

Typically, we use the exponential distribution as a "clock" that rings for events occurring independent of time. It has additional useful properties: the minimum of such clocks, or the first one to ring, is also exponentially distributed, and the identity of the ringing clock is independent of the time at which it rings. Formally, we have the following theorem:

Theorem 4.2. *For $1 \leq i \leq n$, suppose $X_i \sim \text{Exp}(\lambda_i)$ are independent. Let $V := \min_i X_i$ and $I := \text{argmin}_i X_i$. Then*

$$\mathbb{P}(V > t) = e^{-(\lambda_1 + \cdots + \lambda_n)t}, \text{ i.e. } V \sim \text{Exp}(\lambda_1 + \cdots \lambda_n), \text{ and } \mathbb{P}(I = i) = \frac{\lambda_i}{\lambda_1 + \cdots + \lambda_n}$$

Moreover, V and I are independent.

Proof. First, minimum is easy to compute:

$$\mathbb{P}(V > t) = \mathbb{P}(\min_i X_i > t) = \mathbb{P}(X_1 > t, \cdots, X_n > t) = \mathbb{P}(X_1 > t) \cdots \mathbb{P}(X_n > t) = e^{-(\lambda_1 + \cdots + \lambda_n)t}$$

Now, to argue the distribution of the index, we first do it for $n = 2$:

$$\mathbb{P}(I = 1) = \mathbb{P}(X_1 < X_2) = \int_0^\infty f_{X_1}(t)\mathbb{P}(X_2 > t)dt = \int_0^\infty \lambda_1 e^{-\lambda_1 t} e^{-\lambda_2 t} dt = \frac{\lambda_1}{\lambda_1 + \lambda_2}$$

Now, we can generalize by defining $V^{-i} := \min_{j \neq i} X_j$. Note that $V^{-i} \sim \text{Exp}(\sum_{j \neq i} \lambda_j)$ and independent of X_i. Thus,

$$\mathbb{P}(I = i) = \mathbb{P}(X_i < V^{-i}) = \frac{\lambda_i}{\lambda_1 + \cdots + \lambda_n}$$

To show the last independence claim;

$$\begin{aligned}
&\mathbb{P}(V < t, I = i) = \mathbb{P}(X_i < t, X_j > X_i, \forall j \neq i) \\
&= \int_0^t f_{X_i}(s)\mathbb{P}(X_j > s, \forall j \neq i)ds = \int_0^t \lambda_i e^{-(\lambda_1 + \cdots + \lambda_n)s} ds \\
&= \frac{\lambda_i}{\lambda_1 + \cdots + \lambda_i}(1 - e^{(\lambda_1 + \cdots + \lambda_n)t}) = \mathbb{P}(V < t)\mathbb{P}(I = i)
\end{aligned}$$

■

Proposition 4.3. *Let $X_1, X_2, \cdots$ be independent exponential distributions with the same parameter λ. Set $T_n = X_1 + \cdots + X_n$. Then T_n distributed as $\Gamma(n, \lambda)$, that is, the density function is given by*

$$f_{T_n}(t) = \lambda e^{-\lambda t} \frac{(\lambda t)^{n-1}}{(n-1)!} \mathbf{1}_{\{t \geq 0\}}$$

Exercise. Prove it by induction.

Example 4.4. Suppose there are two servers in a store: one provides goods, and the other handles payments. Let's assume their service times are exponentially distributed with rates λ_1 and λ_2.

Now, suppose that when a new customer arrives, there is already one person at the first server. What is the expected time until the new customer leaves the store?

Let C_1^1 and C_2^1 be the times the first customer spends at each server. Similarly, let C_1^2 and C_2^2 be the times for the new customer, given that they are being served. Then, the total time the new customer spends in the system is given by

$$C_1^1 + \max\{C_1^2, C_2^1\} + C_2^2 = C_1^1 + C_1^2 + C_2^1 - \min\{C_1^2, C_2^1\} + C_2^2.$$

Notice that, due to the memoryless property, the distribution of C_1^1 remains unchanged from the perspective of the second customer, despite observing a customer already being served at the first server. The computation of the expected value follows straightforwardly.

Suggested Exercises. Durrett, 3rd edition. 2.1, 2.5, 2.6, 2.45, 2.47, 2.48, 2.49

Let us also recall the definition of the Poisson random variable, which will be connected to the counting of the exponential clocks.

Definition 4.5. We say a random variable X has a Poisson distribution with mean λ, denoted as $X \sim \text{Poi}(\lambda)$, if

$$\mathbb{P}(X = k) = e^{-\lambda}\frac{\lambda^k}{k!},\ k \geq 0$$

Note that if $X \sim \text{Poi}(\lambda)$, then $\mathbb{E}[X] = \text{Var}(X) = \lambda$. Moreover,

Proposition 4.6. *If $X_i \sim \text{Poi}(\lambda_i)$ are independent, then $\sum_{k=1}^n X_i \sim \text{Poi}(\sum_{k=1}^n \lambda_i)$.*

Proof. We will show only for $n = 2$, as the rest is straigtforward by induction.

$$\begin{aligned}\mathbb{P}(X_1 + X_2 = m) &= \sum_{k=0}^{m} \mathbb{P}(X_1 = m-k | X_2 = k)\mathbb{P}(X_2 = k) \\ &= \frac{e^{-(\lambda_1+\lambda_2)}}{m!}\sum_{k=0}^{m}\frac{m!}{k!(m-k)!}(\lambda_1)^{m-k}(\lambda_2)^k = e^{-(\lambda_1+\lambda_2)}\frac{(\lambda_1+\lambda_2)^m}{m!}\end{aligned}$$

■

Proposition 4.7. *Suppose $X_1, X_2, \cdots$ are independent Poisson distributions with parameters $\lambda_1, \lambda_2, \cdots$. Then*

$$\mathbb{P}\Big(X_1 = \ell \Big| \sum_{k=1}^{n} X_k = m\Big) = \binom{m}{\ell}\Big(\frac{\lambda_1}{\sum_{k=1}^n \lambda_k}\Big)^{\ell}\Big(1 - \frac{\lambda_1}{\sum_{k=1}^n \lambda_k}\Big)^{m-\ell}$$

Proof. We will only show the case $n = 2$, as the general case follows by noting that sum of independent Poisson distributions is again Poisson.

$$\begin{aligned}\mathbb{P}(X_1 = \ell | X_1 + X_2 = m) &= \mathbb{P}(X_1 = \ell, X_2 = m-\ell)/\mathbb{P}(X_1 + X_2 = m) \\ &= e^{-\lambda_1}\frac{\lambda_1^\ell}{\ell!}e^{-\lambda_2}\frac{\lambda_2^{m-\ell}}{(m-\ell)!}e^{\lambda_1+\lambda_2}\frac{m!}{(\lambda_1+\lambda_2)^m}\end{aligned}$$

■

4.2 Poisson Process

We are now interested in counting the number of events with exponential distribution, which will be represented by the following definition:

Definition 4.8. Let $X_1, X_2, \cdots$ be independent exponentially distributed random variables with parameter λ. Set $T_n := X_1 + \cdots + X_n$. We call the stochastic process

$$N_t := \sum_{n\geq 1} \mathbf{1}_{\{T_n \leq t\}} = \max\{n \geq 1 : T_n \leq t\}$$

a Poisson process with intensity (rate) $\lambda > 0$.

Intensity λ determines the instantaneous rate for the jump probability in the sense that:

$$h^{-1}\mathbb{P}(N_{t+h} \neq N_t) \to \lambda$$

To see this, observe that

$$\begin{aligned}\mathbb{P}(N_{t+h} \neq N_t) &= 1 - \mathbb{P}(N_{t+h} = N_t) \\ &= 1 - \mathbb{P}(X_{N_t+1} > h + (t - T_{N_t}) | T_{N_t} < t, X_{N_t+1} > (t - T_{N_t})) \\ &= 1 - \mathbb{P}(X_{N_t+1} > h) = 1 - \mathbb{P}(X_1 > h)\end{aligned}$$

due to the memoryless property, and taking the limit is easy.

Theorem 4.9. *If N_t is a Poisson process with intensity λ, then $N_t \sim \mathrm{Poi}(\lambda t)$, $\forall t \geq 0$.*

Proof. Let T_n be the associated clocks, and recall T_n distributed as $\Gamma(n, \lambda)$ by proposition 4.3. Then,

$$\mathbb{P}(N_t = n) = \mathbb{P}(T_n \leq t, T_{n+1} > t) = \int_0^t \mathbb{P}(T_{n+1} > t | T_n = s) f_{T_n}(s) ds$$

$$\int_0^t \mathbb{P}(X_{n+1} > t - s) \lambda e^{-\lambda s} \frac{(\lambda s)^{n-1}}{(n-1)!} ds = \int_0^t e^{-\lambda(t-s)} \lambda e^{-\lambda s} \frac{(\lambda s)^{n-1}}{(n-1)!} ds = e^{-\lambda t} \frac{(\lambda t)^n}{n!}$$

■

Theorem 4.10. *Let N_t be a Poisson process and fix $s \geq 0$. Then, $\tilde{N}_t := N_{t+s} - N_s$ is also a Poisson process with the same intensity, independent of N_r for $0 \leq r \leq s$.*

Proof. Let $X_1, X_2, \ldots$ and $T_n := X_1 + \cdots + X_n$ be associated with the Poisson process N_t as in the definition. Let us observe that, on the event $\{N_s = n\}$,

$$\begin{aligned}N_{t+s} - N_s &= \sum_{m \geq 1} \mathbf{1}_{\{T_m \leq t+s\}} - \sum_{m \geq 1} \mathbf{1}_{\{T_m \leq s\}} = \sum_{m \geq n+1} \mathbf{1}_{\{T_m \leq t+s\}} \\ &= \sum_{m \geq n+1} \mathbf{1}_{\{X_{n+1} - (s - T_n) + X_{n+2} + \cdots + X_m \leq t\}} = \sum_{m \geq 1} \mathbf{1}_{\{\tilde{T}_m \leq t\}}\end{aligned}$$

where we set

$$\tilde{X}_1 = X_{n+1} - (s - T_n), \quad \tilde{X}_k = X_{n+k}, \quad \forall k \geq 2 \quad \text{on } \{N_s = n\}$$

and $\tilde{T}_m := \tilde{X}_1 + \cdots + \tilde{X}_m$. That is,

$$\tilde{N}_t = \sum_{m \geq 1} \mathbf{1}_{\{\tilde{T}_m \leq t\}}$$

It is clear that, for $0 \leq r \leq s$ and $k \geq 2$,

$$\begin{aligned}\mathbb{P}(\tilde{X}_k > t | N_r = m) &= \sum_{n \geq m} \mathbb{P}(X_{n+k} > t | N_s = n, N_r = m) \mathbb{P}(N_s = n | N_r = m) \\ &= \sum_{n \geq m} \mathbb{P}(X_{n+k} > t | N_s = n) \mathbb{P}(N_s = n | N_r = m) = \mathbb{P}(X_1 > t)\end{aligned}$$

since $\{N_r = m\} = \{X_1 + \cdots + X_m < r, X_{m+1} > (r - T_m)\}$, and X_{n+k} is independent of $X_1, \cdots, X_{m+1}$. Thus, $\tilde{X}_k$'s are i.i.d. exponential clock with the same rate and we only need to take care of $\tilde{X}_1$ to conclude that $\tilde{N}_t$ is a Poisson process. Now,

$$\mathbb{P}(\tilde{X}_1 > t | N_s = n) = \mathbb{P}(X_{n+1} > t + (s - T_n) | T_n < s, X_{n+1} > (s - T_n)) = \mathbb{P}(X_{n+1} > t)$$

due to the memoryless property. In particular, $\tilde{X}_1$ is independent of N_s, and similar argument to $k \geq 2$ can be made to notice it is independent of N_r for $0 \leq r \leq s$. ■

Corollary 4.11. *A Poisson process has independent increments. That is,* $N_{t_1} - N_{t_0}, \cdots, N_{t_n} - N_{t_{n-1}}$ *are all independent for any* n *and* $0 \leq t_0 \leq t_1 \leq \cdots t_n < \infty$.

In fact, Theorem 4.9 and Corollary 4.11 characterizes the Poisson process;

Theorem 4.12. *A càdlàg* [8] *stochastic process* N_t *is a Poisson process with intensity* λ *if and only if*

(i) $N_0 = 0$

(ii) $N_t - N_s \sim \text{Poi}(\lambda(t-s))$

(iii) N_t *has independent increments.*

Proof. [Sketch] We have already argued that a Poisson process satisfies (i), (ii), and (iii). Now, let us assume that (i), (ii), and (iii) hold for a càdlàg process N_t. We only need to observe that these conditions determine the finite-dimensional distributions. Since N_t is càdlàg , this suffices to uniquely characterize the law of the process. Now, take any $0 \leq t_1 \leq \cdots \leq t_k < \infty$, together with $n_1 \leq \cdots \leq n_k$. Then

$$\mathbb{P}(N_{t_k} = n_k, \cdots, N_{t_1} = n_1) = \mathbb{P}(N_{t_k} - N_{t_{k-1}} = n_k - n_{k-1}, \cdots, N_{t_1} - N_0 = n_1)$$
$$\mathbb{P}(N_{t_k} - N_{t_{k-1}} = n_k - n_{k-1}) \cdots \mathbb{P}(N_{t_1} - N_0 = n_1)$$

and we are done.

Let us emphasize that we are not constructing $X_1, X_2, \ldots$ to show that N_t is indeed in the form given by the definition, but rather arguing that the law of the process is exactly the same if N_t satisfies (i), (ii), (iii). ■

Suggested Exercises. Durrett, 3rd edition. 2.15, 2.17, 2.31.

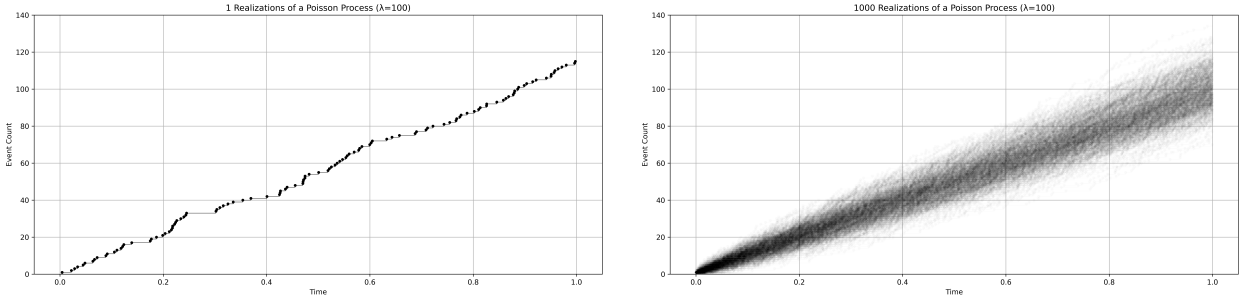

Figure 3: Realizations of Poisson processes with rate 100 in unit time.

We can generalize this definition in a straightforward manner;

Definition 4.13 (Non-homogeneous Poisson Process)**.** A stochastic process N_t is a non-homogeneous Poisson process with rate (or intensity) function $\lambda : \mathbb{R}^+ \to \mathbb{R}^+$, if

(i) $N_0 = 0$

(ii) $N_t - N_s \sim \text{Poi}\big(\int_s^t \lambda(r)dr\big)$

(iii) N_t has independent increments.

[8] càdlàg means paths are right continuous with left limits almost surely.

This definition is quite useful since not all processes maintain the same intensity at all times. One important thing to note is that, if we define the arrival times,

$$\tau_0 := 0, \quad \tau_{n+1} := \inf\{t > 0 : N_{t+\tau_n} \neq N_{\tau_n}\}, \quad \forall n \geq 0 \tag{4.2}$$

then their distributions are neither independent nor identical. They are clearly not identical since the rate changes over time. They are also not independent because, given information about an earlier arrival time, one gains knowledge about which part of the rate function is relevant for the next arrival.

Next, since the Poisson process can be used to count the number of independent events that have occurred, it can also be used to combine i.i.d. random variables;

Definition 4.14 (Compound Poisson Process)**.** A stochastic process S_t is a compound Poisson process if

$$S_t = Y_1 + \cdots + Y_{N_t}$$

where $S_t = 0$ whenever $N_t = 0$, N_t is a Poisson process, and Y_n's are i.i.d. random variables.

It is important to compute the moments of such random variables. To do so, we have the following theorem,

Theorem 4.15. *Let Y_i be i.i.d. random variables, and N be an independent non-negative integer valued random variable. Set $S = Y_1 + \cdots + Y_N$ with $S = 0$ when $N = 0$. Then followings hold,*

(i) If $\mathbb{E}[|Y_1|] < \infty$ and $\mathbb{E}[|N|] < \infty$, then $\mathbb{E}[S] = \mathbb{E}[Y_1]\mathbb{E}[N]$.

(ii) If $\mathbb{E}[|Y_1|^2] < \infty$ and $\mathbb{E}[|N|^2] < \infty$, then $\mathrm{Var}(S) = \mathrm{Var}(Y_1)\mathbb{E}[N] + \mathrm{Var}(N)\mathbb{E}[Y_1]^2$.

(iii) If $N \sim \mathrm{Poi}(\lambda)$, then $\mathrm{Var}(S) = \lambda\mathbb{E}[|Y_1|^2]$.

Proof. (i) is straightforward by noting

$$\mathbb{E}[S] = \sum_{n=0}^{\infty} \mathbb{E}[S|N = n]\mathbb{P}(N = n) = \mathbb{E}[Y_1] \sum_{n=0}^{\infty} n\mathbb{P}(N = n) = \mathbb{E}[Y_1]\mathbb{E}[N]$$

Similarly, we can compute $\mathbb{E}[|S|^2]$,

$$\begin{aligned}
\mathbb{E}[|S|^2] &= \sum_{n=0}^{\infty} \mathbb{E}\big[|S|^2 \mid N = n\big]\mathbb{P}(N = n) \\
&= \sum_{n=0}^{\infty} \sum_{i,j=1}^{n} \mathbb{E}\big[\,\mathrm{Cov}(Y_i, Y_j) + \mathbb{E}[Y_1]^2\big]\mathbb{P}(N = n) \\
&= \sum_{n=0}^{\infty} \Big(n\,\mathrm{Var}(Y_1) + (n\mathbb{E}[Y_1])^2)\Big)\mathbb{P}(N = n) \\
&= \mathrm{Var}(Y_1)\mathbb{E}[N] + \mathbb{E}[Y_1]^2\mathbb{E}[|N|^2]
\end{aligned}$$

and (ii) follows by $\mathrm{Var}(S) = \mathbb{E}[|S|^2] - \mathbb{E}[S]^2$. (iii) follows from (ii) by noting $\mathbb{E}[Y_1] = \lambda = \mathrm{Var}(Y_1)$. ■

Next, we will study the number of arrivals per each event of Y_n's, called thinning.

Theorem 4.16 (Thinning). *Let S_t be a compound Poisson process, where associated i.i.d. random variables Y_n's are taking values in a discrete state space $\mathcal{S}$. Define,*

$$N_t^y = \sum_{n=1}^{N_t} \mathbf{1}_{\{Y_n = y\}}, \quad y \in \mathcal{S}$$

Then $\{N_t^y\}_{y\in\mathcal{S}}$'s are independent Poisson processes with rate $\lambda\mathbb{P}(Y_1 = y)$.

Proof. By definition $N_0^y = 0$. It is also clear that it has independent increments as N_t has it. Now, we can directly compute the distribution of N_t^y for fixed $y \in \mathcal{S}$. Set $p = \mathbb{P}(Y_1 = y)$, $q = 1 - p$.

$$\begin{aligned}
\mathbb{P}(N_t^y = k) &= \sum_{n\geq k} \mathbb{P}(N_t^y = k | N_t = n)\mathbb{P}(N_t = n) \\
&= \sum_{n\geq k} \binom{n}{k} p^k q^{n-k} e^{-\lambda t} \frac{(\lambda t)^n}{n!} \\
&= e^{-\lambda t} \frac{(\lambda tp)^k}{k!} \sum_{n\geq k} \frac{(\lambda tq)^{n-k}}{(n-k)!} = e^{-\lambda t} \frac{(\lambda tp)^k}{k!} e^{\lambda tq} = e^{-\lambda tp} \frac{(\lambda tp)^k}{k!}
\end{aligned}$$

It is interesting that N_t^y's are independent. Computation is straightforward, yet intuition is important. Take any sequence $(k_1, k_2, \cdots)$ where only finitely many of them are non-zero, and let $p_1 = \mathbb{P}(Y_1 = x_1)$, $p_2 = \mathbb{P}(Y_1 = x_2), \cdots$ for $x_i \in \mathcal{S}$.

$$\begin{aligned}
&\mathbb{P}(N_t^{x_1} = k_1, N_t^{x_2} = k_2, \cdots) \\
&= \mathbb{P}(N_t^{x_1} = k_1, N_t^{x_2} = k_2, \cdots | N_t = k_1 + k_2 + \cdots)\mathbb{P}(N_t = k_1 + k_2 + \cdots) \\
&= \frac{(k_1 + k_2 + \cdots)!}{k_1! k_2! \cdots} (p_1^{k_1} p_2^{k_2} \cdots) e^{-\lambda t} \frac{(\lambda t)^{k_1 + k_2 + \cdots}}{(k_1 + k_2 + \cdots)!} \\
&= e^{-\lambda p_1 t} \frac{(\lambda p_1 t)^{k_1}}{k_1!} e^{-\lambda p_2 t} \frac{(\lambda p_2 t)^{k_2}}{k_2!} \cdots = \mathbb{P}(N_t^{x_1} = k_1)\mathbb{P}(N_t^{x_2} = k_2) \cdots
\end{aligned}$$

which shows they are mutually independent, hence also pairwise independent. ■

To further explain the intuition behind independence in thinning, consider flipping a coin each time the Poisson process jumps. Suppose there are two people, H and T, and based on the coin flip, we send a signal to the corresponding one. They cannot communicate with each other. Each of them still sees a Poisson process, but now at half the original rate. Suppose we are told that 100 heads occurred in a unit time. This does not change the distribution of how many tails occurred. Initially, this might seem counterintuitive, one might expect that if 100 heads occurred, the number of tails should also be close to 100. However, consider a case where the original Poisson process has rate 2, meaning we expect 1 event per unit time per person after thinning. Now, if we are given that 100 heads occurred, this is already an extremely unlikely event, but since it has happened, it does not affect the expected number of tails, which remains close to 1. It might be easier to get convinced that

$$\mathbb{P}(N_t = 100 + k, N_t^T = 100) = \mathbb{P}(N_t = 100 + k)\mathbb{P}(N_t^T = 100)$$

which is equivalent in this scenario.

Next, we can also add independent Poisson processes, called superposition.

Theorem 4.17 (Superposition)**.** *Let $N_t^1, \cdots, N_t^n$ be independent Poisson processes with rates $\lambda_1, \cdots, \lambda_n$. Then $N_t^1 + \cdots + N_t^n$ is a Poisson process with rate $\lambda_1 + \cdots + \lambda_n$.*

Proof. (i),(iii) are immediate in the equivalent definition of Poisson process. (ii) follows from the Proposition 4.6. ■

Next, we state a lemma that demonstrates how to superpose Poisson processes and use thinning to recover the distribution of individual Poisson processes. In essence, the lemma only states that one can represent a collection of Poisson processes by constructing a new set of exponential clocks each time one of them rings.

Lemma 4.18. *Consider independent Poisson processes $N_t^1, \cdots, N_t^n$ with rates $\lambda^1, \cdots, \lambda^n$. Set $N_t := N_t^1 + \cdots + N_t^n$. Let $\{\tau_k^1\}_{k=0}^\infty, \cdots, \{\tau_k^n\}_{k=0}^\infty$ be independent exponential random variables with rates $\lambda^1, \cdots, \lambda^n$. and introduce random variables X_k's taking values in $\{1, \cdots, n\}$ as*

$$X_k = \operatorname*{argmin}_{1 \le j \le n} \tau_k^j$$

Then X_k's are independent and

$$\mathbb{P}(X_k = j) = \frac{\lambda^j}{\lambda^1 + \cdots + \lambda^n}$$

Moreover, N_t^j has the same distribution as $\sum_{k=1}^{N_t} \mathbf{1}_{\{X_k = j\}}$.

Proof. It is by assumption that X_k's are independent and the Theorem 4.2 gives the distribution of X_k's. Lastly, by superposition and thinning theorem, $\sum_{k=0}^{N_t} \mathbf{1}_{\{X_k = j\}}$ is a Poisson process with rate

$$(\lambda^1 + \cdots + \lambda^n)\mathbb{P}(X_1 = j) = \lambda^j$$

■

Let us also note that the Proposition 4.7 implies

Corollary 4.19. *Let N_t be a Poisson process. For all $0 \le s \le t$ and $0 \le m \le n$,*

$$\mathbb{P}(N_s = m | N_t = n) = \binom{n}{m} \left(\frac{s}{t}\right)^m \left(1 - \frac{s}{t}\right)^{n-m}$$

That is, on $\{N_t = n\}$, N_s has Binomial distribution with parameters $(n, s/t)$, denoted as $\mathrm{Bin}(n, s/t)$, which is independent of the rate of N_t.

Proof. It suffices to note that

$$\mathbb{P}(N_s = m | N_t = n) = \mathbb{P}(N_s = m | N_s + (N_t - N_s) = n)$$

to invoke the Proposition 4.7. ■

It is a nice fact that, given the number of events that have already occurred, the event times are uniformly distributed. That is,

Theorem 4.20. *For a given Poisson process N_t, let τ_k be arrival times defined as in 4.2. Let $T_k = \tau_1 + \cdots + \tau_k$. Consider independent uniform distributions $U_1, \cdots, U_n$ on $[0, t]$ and their ordered version $U_{(1)}, \cdots, U_{(n)}$. On the set $\{N_t = n\}$, distribution of $T_1, \cdots, T_n$ is equal to $U_{(1)}, \cdots, U_{(n)}$.*

Proof. First, the arrival times τ_k are i.i.d. with an exponential distribution of rate λ, the same as N_t. Although we have not explicitly argued this, it follows directly from the Strong Markov property of the Poisson process.

Now, take any $0 = t_0 < t_1 < \cdots t_n \leq t$. By slightly abusing notation to refer to the density function, we have

$$\begin{aligned}
&\mathbb{P}(T_1 = t_1, \cdots, T_n = t_n | N_t = n) \\
&= \frac{1}{\mathbb{P}(N_t = n)} \mathbb{P}(\tau_1 = t_1 - t_0, \cdots, \tau_n = t_n - t_{n-1}, \tau_{n+1} > t - t_n) \\
&= e^{\lambda t} \frac{n!}{(\lambda t)^n} \big(\lambda e^{-\lambda(t_1 - t_0)}\big) \cdots \big(\lambda e^{-\lambda(t_n - t_{n-1})}\big) e^{-\lambda(t - t_n)} = \frac{n!}{t^n}
\end{aligned}$$

which is independent of $t_1, \cdots, t_n$, hence follows a uniform distribution. Note that the n-dimensional cube has volume t^n and there are $n!$ many potential orderings. Therefore, the volume of the region $\{0 \leq t_1 \leq \cdots \leq t_n \leq t\}$ is $t^n/n!$. ■

Example 4.21. Suppose young and erderly customers arrive to a ticket office, and we can model each of them as independent Poisson processes N_t^y, N_t^e with rates 30 and 20.

Question. Suppose each customer, independently buys 1 ticket with probability $1/2$, or 2 tickets with probability $1/2$. Let Z^k be the number of customers in the first hour that bought $k \in \{1, 2\}$ tickets. What is the joint distribution of (Z^1, Z^2)?

Answer. By the superposition theorem, $N_t = N_t^y + N_t^e$ is a Poisson process with rate 50. Then, by the thinning theorem, we can introduce N_t^1 and N_t^2 where corresponding Y_n's are uniform over $\{1, 2\}$. Then,

$$\mathbb{P}(Z^1 = m, Z^2 = n) = \mathbb{P}(N_1^1 = m, N_1^2 = n) = \mathbb{P}(N_1^1 = m)\mathbb{P}(N_1^2 = n) = e^{-50} \frac{25^m}{m!} \frac{25^n}{n!}$$

Question. What is the probability that the first 3 customers are young?

Answer. We will use the Theorem 4.18. To do so, let $\{\tau_k^y\}_{k \geq 0}$ and $\{\tau_k^e\}_{k \geq 0}$ be independent exponential distributions with parameters 30 and 20. Define the corresponding X_k's taking values in $\{y, e\}$. Let τ_3 be the arrival time of the third customer. We know that

$$\mathbb{P}(N_{\tau_3}^y = 3) = \mathbb{P}\left(\sum_{k=1}^{3} \mathbf{1}_{\{X_k = y\}} = 3\right) = \mathbb{P}(X_1 = y)\mathbb{P}(X_2 = y)\mathbb{P}(X_3 = y) = (30/50)^3$$

Suggested Exercises. Durrett, 3rd edition. 2.35, 2.40, 2.57, 2.61.

5 Renewal Processes

[9] In general, there are many events that occurences are not memoryless. In this section, we will analyse jump processes where arrival times are i.i.d. but not necessarily exponentially distributed as in the Poisson process.

Let τ_i be i.i.d. random variables with the common cumulative distribution F, where $F(0) = 0$. Define $T(n) := T_n := \tau_1 + \cdots + \tau_n$ and

$$N_t := N(t) := \sum_{n\geq 1} \mathbf{1}_{\{T_n \leq t\}} = \max\{n \geq 1 : T_n \leq t\}$$

Example 5.1 (Markov Chains)**.** Let X_n be a Markov chain and set $\tau_n = T_x^n - T_x^{n-1}$. By the strong Markov property, τ_n's are i.i.d. under $\mathbb{P}_x$. Then the corresponding N_t is a renewal process.

Theorem 5.2. *Let $\mu = \mathbb{E}\tau_1$ be the mean interarrival time. If $\mathbb{P}(\tau_1 > 0) > 0$, then*

$$N(t)/t \to 1/\mu \text{ as } t \to \infty \text{ almost surely.}$$

Proof. Note that $\mathbb{P}(\tau_1 > 0) > 0$ implies $\mu > 0$. Otherwise $N(t) = \infty$ for all $t \geq 0$. By Law of Large Numbers, we know that $T_n/t \to \mu$ almost surely. Also, observe that $T(N(t)) \leq t \leq T(N(t)+1)$. Then it is straightforward as

$$\frac{T(N(t))}{N(t)} \leq \frac{t}{N(t)} \leq \frac{T(N(t)+1)}{N(t)+1}\frac{N(t)+1}{N(t)}$$

■

Next, we assign a reward r_i to ith renewal, where all of them are i.i.d. and independent of τ_i. One can further generalize r_i to depend on τ_i. Define the total reward

$$R(t) := \sum_{i=1}^{N(t)} r_i$$

and we have the strong law of large numbers related to reward process,

Theorem 5.3. *Let $\mu = \mathbb{E}\tau_1 > 0$. Then, almost surely it holds $R(t)/t \to \mathbb{E}r_1/\mu$. In words, limit of total reward per time is expected reward divided by expected time.*

Proof.

$$\frac{R(t)}{t} = \frac{N(t)}{t}\frac{1}{N(t)}\sum_{i=1}^{N(t)} r_i \to \frac{\mathbb{E}r_1}{\mu}$$

■

Let us discuss briefly the alternating renewal process. Let $s_1, s_2, \cdots$ be independent with cumulative distribution F and mean μ_F. Similarly, let $u_1, u_2, \cdots$ be independent with cumulative distribution G and mean μ_G. Suppose a system alternates between two states, and times spend in each state is determined by s_i and u_i. For example, we can think about a machine that works until it breaks, and some time passes until it becomes operational again.

[9] Due to time constraints, this section has not yet been taught in class. Thus, it might not be polished.

Theorem 5.4. *For alternating renewal process, limiting fraction of times in states are*

$$\frac{\mu_F}{\mu_F + \mu_G} \quad \textit{and} \quad \frac{\mu_G}{\mu_F + \mu_G}$$

Example 5.5. Suppose you have a machine that works some time with a given distribution (or you estimate with past data). When it brakes down, you call a service and the repair time has another distribution. One can use the result above to determine the long term working time of the machine.

5.1 Queueing systems

In this section, we will briefly discuss queueing theory. Let us show some simple examples of Kendall's notation to explain what is of interest;

- M/G/1: Poisson input, general service time, 1 server
- GI/M/c: General independent interarrival times, exponential service times, c servers.
- GI/G/c: General independent interarrival times, general service times, c servers.

Applications of such systems include telecommunication, traffic engineering, computing, project management, and particularly industrial engineering where it is applied in the design of factories, shops, offices, and hospitals.

We will first start with GI/G/1. Let $\tau_1, \tau_2, \cdots$ be i.i.d. arrival times with mean $1/\lambda$ and $s_1, s_2, \cdots$ be i.i.d. serving times with mean $1/\mu$.

Theorem 5.6. *Assume $\lambda < \mu$ and the queue starts from finite number of customers. Then the queue will eventually empty out almost surely. Furthermore, the limiting fraction of time the server is busy is at most λ/μ.*

Proof. Let $T_n = \tau_1 + \cdots + \tau_n$ be the arrival time of the n-th customer, and $S_n = S_0 + s_1 + \cdots + s_n$ is the service time of $n + k$ customers, where S_0 denoting the total service time of the initial k customers. We will show that strong law of large numbers imply $S_n/T_n \to \lambda/\mu$ almost surely, which shows both claims. First, with probability one, we can find n large enough such that $S_n < T_n$. That is, serving the first $n + k$ customer takes less time then the arrival of n-th customer. Second, ignoring times between arrivals, total time that the queue is busy is at most S_n at time T_n. Now, the claim follows as before,

$$\frac{S_n}{T_n} = \frac{n}{T_n}\Big(\frac{S_0}{n} + \frac{S_n - S_0}{n}\Big) \to \frac{\lambda}{\mu}$$

■

Let X_t be the number of customers at time t in the server, and define the long term average as

$$L = \lim_{t\to\infty} \frac{1}{t}\int_0^t X_t \quad \text{(average number of customers)}, \quad L_q \ \text{(average queue length)}$$

Next, define the long term average of the service times

$$W = \lim_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n} w_n \quad \text{(average time spend)}, \quad W_q \ \text{(average waiting in queue)}$$

where w_n is the total time spend for the nth customer. Lastly, let

$$\lambda = \lim_{t\to\infty} N_t/t \quad \text{(average arrival rate)}$$

be the long term average of customers arriving per time.

Theorem 5.7 (Little's Law)**.** *$L = \lambda W$, and $L_q = \lambda W_q$*

See *A proof for the Queuing Formula: $L = \lambda W$*, John D.C. Little. (1961)

Example 5.8. We will discuss two examples to present the idea.

For example, if you are planning for a hospital, one might have the data of average arrival of patients (λ) and average time they spent (W). Then, one can estimate the average patient number (L) and have a crude estimate of what is necessary to accomodate them.

Another example might be a factory, where the amount of materials needed for production is exactly known and fixed per day (λ), and working capacity of the factory is known (L). Then one can estimate the so called flow time W, which tells how long it takes for the factory to process single item from start to end.

Next, we will consider M/G/1 queue, where costumers arrive as Poisson process with rate λ. Define the probability that k customers arrive during one service time as

$$a_k = \int_0^\infty e^{-\lambda t} \frac{(\lambda t)^k}{k!} dG(t)$$

Then, the average customer arrival is

$$\sum_{k \geq 0} k a_k = \int_0^\infty \lambda t \, dG(t) = \lambda / \mu$$

Let ζ_i be i.i.d. RVs with $\mathbb{P}(\zeta_i = k) = a_k$. Then, we can create a Markov chain as

$$X_{n+1} = (X_n + \zeta_n - 1)^+$$

Note that X_n represents our queue, because Poisson process has independent increments even under random times. Namely, given how many customers arrived during the first service time does not give any information about how many customers will arrive in the second service time and has the same distribution.

Theorem 5.9.

- *If $\lambda < \mu$, then X_n is positive recurrent with $\mathbb{E}_0 T_0 = \mu/(\mu - \lambda)$.*
- *If $\lambda = \mu$, then X_n is null recurrent.*
- *If $\lambda > \mu$, then X_n is transient.*

Proof. Let us use the Little's Law to derive the following relation. Since there exists only one server, queue length is $(X_t - 1)^+$, and hence

$$\sum_{k=0}^{n} (X_k - 1)^+ = \sum_{k=0}^{n} \mathbf{1}_{\{X_k \neq 0\}} (X_k - 1) = \sum_{k=0}^{n} (X_k - 1) dt + \sum_{k=0}^{n} \mathbf{1}_{\{X_k = 0\}}$$

dividing by t and taking the limit yields, by asymptotic frequency,

$$L_q = L - 1 + \frac{1}{\mathbb{E}_0 T_0}$$

On the other hand, we also have that $W - 1/\mu = W_q$. Therefore, by Little's Law

$$L - 1 + 1/\mathbb{E}_0 T_0 = \lambda W - \lambda/\mu \implies 1/\mathbb{E}_0 T_0 = 1 - \lambda/\mu$$

This concludes the first two result right away. However, let us show the null recurrence once more by different arguments. If $\mu = \nu$, then $\mathbb{E}\zeta_1 - 1 = 0$. So if $k > 1$, then $\mathbb{E}_k X_1 = k$. Recall $V := V_{0,N}$ and invoke optional stopping theorem;

$$k = \mathbb{E}_k X_V \geq N\mathbb{P}_k(V_N < V_0) \implies 1 - k/N \leq \mathbb{P}_k(V_0 < V_N)$$

Since

$$\mathbb{P}_k(T_0 < \infty) \geq \mathbb{P}_k(V_0 < V_N) \geq 1 - k/N$$

and sending $N \to \infty$ concludes $\mathbb{P}_k(T_0 < \infty) = 1$ which is sufficient to conclude that the chain is recurrent. Note that, we are using the fact that the chain is irreducible, hence working on the state 0 suffices to characterize recurrence behaviour.

Finally, to argue that the chain is transient,

$$\frac{1}{V_0}\big(X_{V_0} - X_0\big) = -\frac{k}{V_0} = \frac{1}{V_0}\sum_{k=0}^{V_0}\zeta_k - 1$$

holds whenever $V_0 < \infty$. Then, assuming

$$\mathbb{P}_k(V_0 < \infty) = 1$$

yields, by taking the limit as $k \to \infty$ and using the law of large numbers a contradiction. Hence $\mathbb{P}_k(V_0 < \infty) < 1$. ■ We will not give the proof of the following result.

Theorem 5.10 (Pollaczek-Khintchine Formula)**.** *For M/G/1 queue, long term average waiting time in queue is*

$$W_q = \frac{\lambda\mathbb{E}(s_i^2/2)}{1 - \lambda/\mu}$$

6 Continuous Time Markov Chains

In this section, we will extend our index set for time from being discrete as $\mathbb{N}$ to continuous $\mathbb{R}^+$.

Definition 6.1. We say a stochastic process X_t over a countable state space $\mathcal{S}$ is a temporally homogeneous continuous time Markov chain (CTMC) if

$$\mathbb{P}(X_{t+s} = y | X_s = x, X_{s_n} = x_n, \cdots, X_{s_0} = x_0) = \mathbb{P}(X_t = y | X_0 = x)$$

for any $x, y, x_0, \cdots, x_n \in \mathcal{S}$, $0 \le s_0 < \cdots < s_n < s$, and $t \ge 0$. Moreover,

$$p_t(x, y) := \mathbb{P}(X_t = y | X_0 = x) := \mathbb{P}_x(X_t = y)$$

is called the corresponding transition probability of CTMC, and

$$q(x, y) := \lim_{t \to 0} \frac{p_t(x, y) - p_0(x, y)}{t} = \lim_{t \to 0} \frac{p_t(x, y) - \mathbf{1}_{\{x=y\}}}{t}$$

is called the jump rate, if the limit exists for all $x, y \in \mathcal{S}$.

We are going to observe that the jump rate determines the CTMC under regularity assumptions. Thus, the following proposition provides the first representation of all regular CTMCs.

Proposition 6.2. *Let N_t be a Poisson process with intensity λ, and let Y_n be an independent discrete Markov chain on $\mathcal{S}$ with transition probability p_Y. Then $X_t := Y_{N_t}$ is a continuous time Markov chain with the transition probability*

$$p_t(x, y) = \sum_{m=0}^{\infty} \mathbb{P}_x(Y_m = y)\mathbb{P}(N_t = m) = e^{-\lambda t} \sum_{m=0}^{\infty} p_Y^m(x, y) \frac{(\lambda t)^m}{m!}$$

with the convention $p_Y^0(x, y) = \mathbf{1}_{\{x=y\}}$. Moreover, jump rate is given by

$$q(x, y) = -\lambda \mathbf{1}_{\{x=y\}} + \lambda p_Y(x, y)$$

and hence the following relations holds,

$$\sum_{y \in \mathcal{S}} q(x, y) = 0, \quad \lambda = \frac{\sum_{y \neq x} q(x, y)}{1 - p_Y(x, x)}, \quad p_Y(x, y) = \mathbf{1}_{\{x=y\}} + \frac{q(x, y)}{\lambda}$$

Lastly, given $X_0 = x$, first time that X_t leaves x is distributed exponentially with parameter $\lambda(1 - p_Y(x, x))$.

Proof. Take any $x, y, y_0, \cdots, y_n \in \mathcal{S}$ and $t \ge 0$, $0 \le s_0 < \cdots < s_n < s$. For notational simplicity, let us abuse notations slightly and write $\vec{X} = \vec{y}$ in the place of $X_{s_n} = y_n, \cdots, X_{s_0} = y_0$;

$$\begin{aligned}
&\mathbb{P}(X_{t+s} = y | X_s = x, \vec{X} = \vec{y}) \\
&= \sum_{m \ge k} \mathbb{P}(X_{t+s} = y | N_{t+s} = m, N_s = k, X_s = x, \vec{X} = \vec{y}) \mathbb{P}(N_{t+s} = m, N_s = k | X_s = x, \vec{X} = \vec{y}) \\
&= \sum_{m \ge k} \mathbb{P}(Y_m = y | N_{t+s} = m, N_s = k, Y_k = x, \vec{X} = \vec{y}) \mathbb{P}(N_{t+s} = m, N_s = k | X_s = x, \vec{X} = \vec{y}) \\
&= \sum_{m \ge k} \mathbb{P}(Y_{m-k} = y | Y_0 = x) \mathbb{P}(N_{t+s} - N_s = m - k | X_s = x, \vec{X} = \vec{y}) \mathbb{P}(N_s = k | X_s = x, \vec{X} = \vec{y}) \\
&= \sum_{\ell} \mathbb{P}(Y_\ell = y | Y_0 = x) \mathbb{P}(N_{t+s} - N_s = \ell) \sum_{k} \mathbb{P}(N_s = k | X_s = x, \vec{X} = \vec{y})
\end{aligned}$$

where the last sum is 1 and the result follows. Let us note that the first equality is the decomposition onto the values of the Poisson process; the third equality uses the Markov property of the discrete chain Y_n; and the last equality uses the independent increments of the Poisson process, together with the fact

$$\sum_{m\geq 0}\sum_{0\leq k\leq m}\phi(m-k)\varphi(k)=\sum_{k\geq 0}\sum_{m\geq k}\phi(m-k)\varphi(k)=\sum_{k\geq 0}\sum_{\ell\geq 0}\phi(\ell)\varphi(k)$$

Next, we can directly compute $q(x,y)$,

$$q(x,y)=\lim_{h\to 0}\frac{e^{-\lambda h}}{h}\Big(\mathbf{1}_{\{x=y\}}+p_Y(x,y)(\lambda h)+o(h)\Big)-\frac{\mathbf{1}_{\{x=y\}}}{h}$$

and following relations are easy to deduce. Lastly,

$$\begin{aligned}\mathbb{P}_x(X_s=x,\ \forall s\in[0,t])&=\sum_{n\geq 0}\mathbb{P}_x(Y_1=x,\cdots Y_n=x|N_t=n)\mathbb{P}_x(N_t=n)\\&=e^{-\lambda t}\sum_{n\geq 0}\frac{(p_Y(x,x)\lambda t)^n}{n!}=e^{-\lambda(1-p_Y(x,x))t}\end{aligned}$$

which gives the probability the Markov chain leaves x after time t, and notice that the distribution matches with the exponential distribution. ■

Assumption 6.3. *(i): Markov chain is almost surely right continuous. That is,*

$$\mathbb{P}\Big(\lim_{h\downarrow 0}X_{t+h}=X_t\Big)=1$$

(ii-a): Markov chain admits jump rate $q(x,y)$ where $-q(x,x)<\infty$ and $\sum_{y\in\mathcal{S}}q(x,y)=0$.
(ii-b): Markov chain admits jump rate where $-\inf_{x\in\mathcal{S}}q(x,x)<\infty$ and $\sum_{y\in\mathcal{S}}q(x,y)=0$.

Remark 6.4. The assumption $\sum_{y\in\mathcal{S}}q(x,y)=0$ is not arbitrary. One can show that it implies $p_t(x,y)$ is continuously differentiable. Conversely,

$$\sum_{y\in\mathcal{S}}q(x,y)=\lim_{t\to 0}\frac{1}{t}\sum_{y\in\mathcal{S}}p_t(x,y)-\mathbf{1}_{\{x=y\}}=0$$

so if $p_t(x,y)$ is continuously differentiable, one can properly interchange the limit and recover the condition $\sum_{y\in\mathcal{S}}q(x,y)=0$.

Theorem 6.5. *Suppose a continuous time Markov chain satisfies 6.3 (i), and jumps finitely many on any interval almost surely. Then it satisfies 6.3 (ii-a).*

We will not prove this theorem. Without the right continuity assumption, there are processes that makes infinitely many jumps in any finite interval. (due to Blackwell)

Theorem 6.6. *Suppose Markov chain satisfies the assumption 6.3. Let*

$$\lambda_x:=-q(x,x)<\infty$$

Then,

(i) First time Markov chain leaves the state x has exponential distribution with rate λ_x.

(ii) If $\lambda_x = 0$, then Markov chain never leaves x.

(iii) If $\lambda_x > 0$, then Markov chain jumps to the state $y \neq x$ with probability $q(x,y)/\lambda_x$.

Proof. [Sketch of proof.] (i): Let $\tau_x := \inf\{t > 0 : X_t \neq x\}$ be the first time X_t leaves x. Fix $t > 0$ and take a partition of time as $t_k^n = (k/n)t$ for some n.

$$\begin{aligned}
\mathbb{P}(\tau_x > t) &= \mathbb{P}(X_s = x, \forall s \in [0,t]) \\
&= \lim_{n\to\infty} \mathbb{P}(X_s = x, \forall s \in [0,t] | X_{t_n^n} = x, \cdots, X_{t_0^n} = x) \mathbb{P}(X_{t_n^n} = x, \cdots, X_{t_0^n} = x) \\
&= \lim_{n\to\infty} \mathbb{P}(X_{t_n^n} = x, \cdots, X_{t_0^n} = x) \\
&= \lim_{n\to\infty} \mathbb{P}(X_{t_n^n} = x | X_{t_{n-1}^n} = x) \cdots \mathbb{P}(X_{t_1^n} = x | X_{t_0^n} = x) \\
&= \lim_{n\to\infty} \Big(p_{(t/n)}(x,x) \Big)^n \\
&= \lim_{n\to\infty} \left[1 + \frac{p_{(t/n)}(x,x) - 1}{(t/n)} \frac{t}{n} \right]^n \\
&= \lim_{n\to\infty} \left[1 + q(x,x) \frac{t}{n} \right]^n = e^{-\lambda_x t}
\end{aligned}$$

Note that, if the first limit in the second line is not going to 1, that would imply we can find a convergent subsequence on a set of positive measure where X is not càdlàg .

This computation also shows (ii), as if $\lambda_x = 0$, $\mathbb{P}(\tau_x > t) = 1$ for all $t > 0$. Finally, to argue (ii), let f_{τ_x} be the density of τ_x and then,

$$\begin{aligned}
\mathbb{P}_x(X_{\tau_x} = y) &= \int_0^\infty \mathbb{P}(X_s = y | \tau_x = s) f_{\tau_x}(s) ds \\
&= \int_0^\infty \mathbb{P}(X_s = y | X_s \neq x, X_r = x, \forall 0 \leq r < s) f_{\tau_x}(s) ds \\
&= \int_0^\infty \lim_{\varepsilon\to 0} \mathbb{P}(X_s = y | X_s \neq x, X_r = x, \forall 0 \leq r \leq s - \varepsilon) f_{\tau_x}(s) ds \\
&= \int_0^\infty \lim_{\varepsilon\to 0} \mathbb{P}(X_s = y | X_s \neq x, X_{s-\varepsilon} = x) f_{\tau_x}(s) ds \\
&= \int_0^\infty \lim_{\varepsilon\to 0} \frac{p_\varepsilon(x,y)}{1 - p_\varepsilon(x,x)} f_{\tau_x}(s) ds \\
&= \int_0^\infty \lim_{\varepsilon\to 0} \frac{q(x,y)\varepsilon + o(\varepsilon)}{-q(x,x)\varepsilon + o(\varepsilon)} f_{\tau_x}(s) ds = \frac{q(x,y)}{\lambda_x} \int_0^\infty f_{\tau_x}(s) ds
\end{aligned}$$

■

Proposition 6.7. *Suppose jump rate $q : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ is given, where $\sum_{y\neq x} q(x,y) < \infty$ for all $x \in \mathcal{S}$. Then there exists a continuous time Markov chain with this jump rate.*

Proof. [Sketch] Define

$$\lambda_x := \sum_{y\neq x} q(x,y)$$

Take a discrete Markov chain Z_n with transition matrix $p_Z(x,y) := (q(x,y)/\lambda_x)\mathbf{1}_{\{x\neq y\}}$. Let $\tau_0, \tau_1, \cdots$ be independent random variables with distribution Exp(1). Next, let $t_{n+1} := \tau_n/(\lambda_{Z_n})$ for $n \geq 0$, and set

$T_n = t_1 + \cdots + t_n$. We claim that the process

$$X_t := Z_{N_t}, \quad \text{where } N_t := \sum_{n\geq 1} \mathbf{1}_{\{T_n \leq t\}}$$

is the corresponding Markov chain. Given any $\{X_{s_0} = x_0, \cdots X_{s_n} = x_n\}$ (denoted as $\vec{X} = \vec{x}$), $0 \leq s_0 < \cdots < s_n < s$ and $x_0, \cdots, x_n \in \mathcal{S}$,

$$\begin{aligned}
&\mathbb{P}(X_{t+s} = y | X_s = x, \vec{X} = \vec{x}) \\
&= \mathbb{P}(Z_{N_{t+s}} = y | Z_{N_s} = x, \vec{X} = \vec{x}) = \begin{cases} \mathbf{1}_{\{x=y\}} & \text{if} \quad N_{t+s} = N_s \\ \mathbb{P}(Z_{N_{t+s}} = y | Z_{N_s} = x) & \text{if} \quad N_{t+s} > N_s \end{cases} \\
&= \mathbb{P}(X_{t+s} = y | X_s = x) \\
&= \sum_{k\geq 0} \mathbb{P}(Z_{N_t+k} = y | Z_{N_s} = x) \mathbb{P}(N_{t+s} - N_s = k | Z_{N_s} = x) \\
&= \sum_{k\geq 0} \mathbb{P}(Z_k = y | Z_0 = x) \mathbb{P}(T_k \leq t < T_{k+1} | Z_0 = x) =: p_t(x, y)
\end{aligned}$$

To take the derivative at 0, we will simply ignore (hence not really rigorous) the probability that two exponential clock rings in a very small time interval. Then for $x \neq y$, only the $k = 1$ term contributes and roughly looks like

$$\partial_t p_t(x, y)|_{t=0} \sim \frac{q(x, y)}{\lambda_x} \partial_t (1 - e^{-\lambda_x t})|_{t=0} = q(x, y)$$

■

Remark 6.8. **(i):** Given a CTMC, Z_n defined in the proof is called the embedded Markov chain.

(ii): Recall Proposition 6.2. Let

$$\lambda := \sup_x \lambda_x \quad \text{and} \quad p_Y(x, y) := \frac{q(x, y)}{\lambda} \quad \text{if } x \neq y, \quad p_Y(x, x) := 1 - \frac{\lambda_x}{\lambda}$$

Then, we already knew that the corresponding process X_t has the given jump rate. We now have two representations of a continuous-time Markov chain: one using the same exponential clocks and allowing jumps to the same state (under Assumption 6.3 (ii-b)), and one using different exponential clocks and disallowing jumps to the same state (under Assumption 6.3 (ii-a)). One representation may be easier to work with depending on the context of the question.

(ii): Note that we can simulate a Markov chain with given jump rates q by following the construction in the proof. Take any initial state $x \in \mathcal{S}$. Then, draw the time of the jump from an exponential distribution with parameter λ_x, and draw the next state according to the jump probabilities given by $q(x, \cdot)/\lambda_x$. Repeat this procedure at the next state.

Example 6.9 (Poisson Process)**.** Poisson process with intensity λ is a CTMC with

$$q(n, n + 1) = \lambda$$

This follows from Proposition 6.2 by setting $\mathcal{S} = \mathbb{N}$, $p_Y(n, n + 1) = 1$. Then $X_t = N_t$ and hence

$$q(n, n) = -\lambda, \quad q(n, n + 1) = \lambda \tag{6.1}$$

Example 6.10 (Branching Process)**.** Each member of the population independently dies with rate μ or gives birth with rate λ. Let X_t denote the number of members of this population. We have that X_t is a Markov chain. If $X_t = n$, then there are n exponential distributions (clocks) with rate μ and n exponential clocks with rate λ. Let τ_b and τ_d be the minimum of n clocks, standing for birth and death. Then by Theorem 4.2, $\tau_b \sim \text{Exp}(n\lambda)$ and $\tau_d \sim \text{Exp}(n\mu)$. By the Theorem 6.6,

$$\lambda_n := q(n, n+1) + q(n, n-1) = n(\lambda + \mu)$$

which comes from $\min(\tau_b, \tau_d)$. Again by Theorem 4.2,

$$\mathbb{P}(\tau_b = \min(\tau_b, \tau_d)) = \frac{\lambda}{\lambda + \mu}$$

and Theorem 6.6 implies

$$q(n, n+1) = \lambda_n \frac{\lambda}{\lambda + \mu} = n\lambda, \text{ and similarly } q(n, n-1) = n\mu$$

The general idea is that if jumps occur only by size 1, then q is determined by the rate of the exponential clock. Let us use this in the next example.

Example 6.11 (Queue)**.** Suppose customers arrive according to Poisson process with rate λ and there are s tellers, independently serving customers with rate μ. Let X_t denote the number of customers in the bank. Then

$$q(n, n+1) = \lambda, \quad q(n, n-1) = \begin{cases} n\mu, & \text{if } 0 \le n \le s \\ s\mu, & \text{if } n \ge s \end{cases}$$

Before moving on, we will state the Strong Markov property. We defer the proper definition of stopping times to later, see the Definition 7.50.

Theorem 6.12 (Strong Markov Property)**.** *Let T be a stopping time with respect to CTMC X_t with transition probabilities p_t. For any $t \ge 0$, given $\{T < \infty, X_T = x\}$, X_{T+t} is independent of $X_{[0,T]}$. Moreover,*

$$\mathbb{P}(X_{T+t} = y | T < \infty, X_T = x) = p_t(x, y)$$

We will not prove this result, but as a general idea, one can first prove it for Poisson process and then lift it to the CTMC.

6.1 Kolmogorov's Equations

We have seen that given transition rates, we can form a Markov chain. In this section, we will determine the transition probability given the transition rates.

Theorem 6.13 (Chapman-Kolmogorov Equation)**.** *Consider a CTMC with countable state space $\mathcal{S}$ and transition probabilities p_t. Then*

$$p_{t+s} = p_t p_s, \quad \textit{that is,} \quad p_{t+s}(x, y) = \sum_{z \in \mathcal{S}} p_t(x, z) p_s(z, y)$$

Proof. By Markov property,

$$p_{t+s}(x, y) = \sum_{z \in \mathcal{S}} \mathbb{P}(X_{t+s} = y | X_s = z) \mathbb{P}(X_s = z | X_0 = x)$$

■

Definition 6.14 (Generator)**.** For a given transition rates, define the generator Q as

$$Q(x,y) := \begin{cases} q(x,y) & \text{if } x \neq y \\ -\lambda_x := -\sum_{y \neq x} q(x,y) & \text{if } x = y \end{cases}$$

Theorem 6.15 (Kolmogorov's Equations)**.** *Transition probabilities p_t of a CTMC satisfies*

$$p_t' = Qp_t \quad \textit{and} \quad p_t' = p_t Q$$

where the former is called Kolmogorov's Backward Equation and the latter is called Kolmogorov's Forward Equation[10]*. In particular, p_t commutes with the generator Q.*

Proof. For the backward equation,

$$\begin{aligned} \frac{1}{h}[p_{t+h}(x,y) - p_t(x,y)] &= \frac{1}{h}\sum_{z\in\mathcal{S}} p_h(x,z)p_t(z,y) - p_t(x,y) \\ &= \frac{1}{h}\sum_{z\neq x} p_h(x,z)p_t(z,y) - p_t(x,y)(1-p_h(x,x)) \\ &= \sum_{z\neq x} \frac{p_h(x,z)}{h} p_t(z,y) - p_t(x,y)\sum_{z\neq x}\frac{p_h(x,z)}{h} \end{aligned}$$

send $h \to 0$, by assumptions that we omit, it follows

$$p_t'(x,y) = \sum_{z\neq x} q(x,z)p_t(z,y) - \lambda_x p_t(x,y) = (Qp_t)(x,y)$$

For the forward equation,

$$\begin{aligned} \frac{1}{h}[p_{t+h}(x,y) - p_t(x,y)] &= \frac{1}{h}\sum_{z\in\mathcal{S}} p_t(x,z)p_h(z,y) - p_t(x,y) \\ &= \frac{1}{h}\sum_{z\neq y} p_t(x,z)p_h(z,y) - p_t(x,y)(1-p_h(y,y)) \\ &= \sum_{z\neq y} p_t(x,z)\frac{p_h(z,y)}{h} - p_t(x,y)\sum_{z\neq y}\frac{p_h(x,z)}{h} \end{aligned}$$

send $h \to 0$, by assumptions that we omit, it follows

$$p_t'(x,y) = \sum_{z\neq y} p_t(x,z)q(z,y) - p_t(x,y)\lambda(y) = (p_t Q)(x,y)$$

How do we interchange the derivative with summation? In fact, the forward equation is easier to handle. Note that

$$p_h(x,y) \leq 1 - p_h(x,x) \leq 1 - e^{-\lambda_x h} \leq \lambda_x h$$

Therefore,

$$\sum_{z\neq y} p_t(x,z)\frac{p_h(z,y)}{h} \leq \sum_{z\neq y} p_t(x,z)\lambda_z \leq \sup_z \lambda_z < \infty$$

[10] Kolmogorov's Forward Equation is also called Fokker-Planck equation.

Backward equation is harder to handle, as it involves the dynamics of $p_t(z, y)$ on the backward variable. (Hence not immediately integrable). One can show that p_t satisfies an integral equation,

$$p_t(x, y) = \mathbf{1}_{\{x=y\}} e^{-\lambda_x t} + \lambda_x \int_0^t e^{-s\lambda_x} (p_Z p_{t-s})(x, y) ds$$

where p_Z is the transition matrix of the embedded MC. This integral equation is equivalent to p_t being continuously differentiable. That suffices for properly changing the operations. ■

As the solution is given by an exponential, let us recall the definition of the exponential for matrices;

Definition 6.16. Let M be any $n \times n$ matrix. Define

$$e^M = \exp(M) := \sum_{k \geq 0} \frac{M^k}{k!} = \lim_{k \to \infty} \left(I + \frac{M}{k} \right)^k$$

where $M^0 = I$ is the identity matrix.

Theorem 6.17. *For a CTMC with the generator Q, transition probabilities are given by*

$$p_t = \exp(tQ)$$

Proof. [Comment] Uniqueness, given that the initial condition $p_0 = I$, is standard in ODE theory. One can differentiate the summation term by term to see that it is a solution, however, one needs a bit more work to rigorously show it. ■

Example 6.18. (Two state Markov chain) Let $\mathcal{S} = \{0, 1\}$ and consider a general generator

$$Q = \begin{pmatrix} -\lambda & \lambda \\ \mu & -\mu \end{pmatrix}$$

Then the Kolmogorov's backward equation is

$$p_t'(0, 0) = -\lambda(p_t(0, 0) - p_t(1, 0)), \text{ and } p_t'(1, 0) = \mu(p_t(0, 0) - p_t(1, 0))$$

Substract these equations to get

$$[p_t(0, 0) - p_t(1, 0)]' = -(\mu + \lambda)(p_t(0, 0) - p_t(1, 0))$$

this has the unique solution

$$p_t(0, 0) - p_t(1, 0) = e^{-(\mu+\lambda)t}$$

Then we can solve

$$p_t'(0, 0) = -\lambda e^{-(\mu+\lambda)t}, \text{ and } p_t'(1, 0) = \mu e^{-(\mu+\lambda)t}$$

which yields, together with $p_0(0, 0) = 1$, $p_0(1, 0) = 0$,

$$p_t(0, 0) = \frac{\mu}{\lambda + \mu} + \frac{\lambda}{\lambda + \mu} e^{-(\mu+\lambda)t}, \text{ and } p_t(1, 0) = \frac{\mu}{\lambda + \mu} - \frac{\mu}{\lambda + \mu} e^{-(\mu+\lambda)t}$$

Lastly, recall that $p_t(0, 1) = 1 - p_t(0, 0)$ and $p_t(1, 1) = 1 - p_t(1, 0)$. Therefore, $p_t(x, \cdot)$ converges exponentially fast to $(\mu, \nu)/(\mu + \nu)$.

6.2 Limiting Behaviour

Definition 6.19. We say a Markov chain with the generator Q is irreducible, if for any states $x, y \in \mathcal{S}$, there exists $x = x_0, x_1, \cdots, x_n = y$ such that $q(x_k, x_{k+1}) > 0$ for all $0 \leq k < n$.

Lemma 6.20. *If CTMC is irreducible then $p_t(x, y) > 0$ for all $x, y \in \mathcal{S}$ and $t > 0$. Moreover, under the assumption 6.3 (i), (ii)-b, the followings are equivalent,*

- *CTMC is irreducible*
- *Embedded MC is irreducible*
- *$p_t(x, y) > 0$ for some $t > 0$ and for all $x, y \in \mathcal{S}$.*

Proof. We will only show the first part of the lemma. Second part should be intutively true, and it is not hard to rigorously argue.

Assume that the Markov chain is irreducible. We know that the leaving a state has exponential distribution (see the Theorem 6.6), hence

$$p_t(x, x) \geq \mathbb{P}(X_s = x,\ \forall s \in [0, t]) = e^{-\lambda_x t} > 0$$

Now, by assumption, let $x = x_0, \cdots, x_n = y$ be a path with $q(x_k, x_{k+1}) > 0$. Because

$$p_t(x, y) \geq p_{nh}(x, y) p_{t-nh}(y, y)$$

it suffices to show that $p_{nh}(x, y) > 0$ for sufficiently small h. This follows by noting that

$$\begin{aligned} p_{nh}(x, y) &\geq p_h(x_0, x_1) \cdots p_h(x_{n-1}, x_n) \\ &\geq (q(x_0, x_1)h + o(h)) \cdots (q(x_{n-1}, x_n)h + o(h)) > 0 \end{aligned}$$

which holds for sufficiently small h, since n is fixed. ∎

Definition 6.21. We say a distribution π on $\mathcal{S}$ is stationary, if

$$\sum_{y \in \mathcal{S}} \pi(y) p_t(y, x) = \pi(x), \quad \forall t > 0$$

π is also called equilibrium distribution.

Theorem 6.22. *Suppose CTMC is irreducible. π is a stationary distribution if and only if*

$$\sum_{y \in \mathcal{S}} \pi(y) Q(y, x) = 0$$

for all $x \in \mathcal{S}$ where Q is the generator of the Markov chain.

Proof. If π is a stationary distribution, then $\pi p_t = \pi$ for all t. By the Kolmogorov's Forward Equation,

$$\begin{aligned} &\sum_y \pi(y) Q(y, x) = \sum_y \sum_z \pi(z) p_t(z, y) Q(y, x) = \sum_z \pi(z) \sum_y p_t(z, y) Q(y, x) \\ &= \sum_z \pi(z) p_t'(z, x) = \Big(\sum_z \pi(z) p_t(z, x) \Big)' = 0 \end{aligned}$$

where we can interchange summation with differentiation because

$$\sum_z \pi(z)|p_t'|(z,x) = \sum_z \pi(z)\Big|\sum_y Q(z,y)p_t(y,x)\Big| \le 2\sum_z \pi(z)\lambda_z < \infty$$

Now, if it holds that $\pi Q = 0$, then by the Kolmogorov's Backward Equation,

$$\Big(\sum_z \pi(z)p_t(z,x)\Big)' = \sum_z \pi(z)p_t'(z,x)$$
$$= \sum_z \pi(z)\sum_y Q(z,y)p_t(y,x) = \sum_y \Big(\sum_z \pi(z)Q(z,y)\Big)p_t(y,x) = 0$$

which means $\pi p_t = \pi p_0 = \pi$. ■

Theorem 6.23. *Suppose CTMC is irreducible. π is a stationary distribution for a CTMC if and only if $\tilde{\pi}$ is a stationary distribution for the embedded chain, where*

$$\pi(x) = \frac{1}{Z}\frac{\tilde{\pi}(x)}{\lambda_x}$$

with Z being the normalization constant.

Proof. We know that embedded MC has transition probailities given by

$$\tilde{p}(x,y) = \frac{q(x,y)}{\lambda_x}\mathbf{1}_{\{y\neq x\}}$$

therefore, if $\tilde{\pi}$ is a stationary distribution for the embedded MC, then

$$Z\pi(x)\lambda_x = \tilde{\pi}(x) = \sum_{y\neq x}\tilde{\pi}(y)\tilde{p}(y,x) = Z\sum_{y\neq x}\pi(y)q(y,x)$$

which yields $\pi Q = 0$. By reversing the exact same argument, $\pi Q = 0$ implies $\tilde{\pi}$ is a stationary distribution. ■

Note that, for the discrete time Markov chain, if a recurrent state is not aperiodic, then $p^{nk}(x,x) = 0$ for some k and for all n, which prevents the convergence. In the continuous case, $p_t(x,x) > 0$ for all x, hence we cannot capture such aperiodicity. In other words, if we define a discrete time Markov chain by using p_t for some fixed t, it will be aperiodic and will converge to the stationary distribution. Let us use this argument to prove the following theorem;

Theorem 6.24. *Consider an irreducible CTMC with transition probability p_t and stationary distribution π. Then*

$$\lim_{t\to\infty} p_t(x,y) = \pi(y) \quad \textit{and} \quad \lim_{t\to\infty}\frac{1}{t}\int_0^t \mathbf{1}_{\{X_s=y\}} = \pi(y)$$

Proof. Consider a discrete time Markov chain corresponding to $p_h(x,y)$. First of all,

$$p_{2h}(x,y) = \mathbb{P}_x(X_{2h}=y) = \sum_z \mathbb{P}_x(X_{2h}=y|X_h=z)\mathbb{P}_x(X_h=z) = p_h^2(x,y)$$

and obviously it holds for n instead of 2.

By assumption, π is a stationary distribution for any $h > 0$ for this discrete Markov chain. Since

$$p_t(x,y) \geq p_{t-nh}(y,y)p_{nh}(x,y) \geq e^{-\lambda_y h} p_h^n(x,y), \quad \text{for } nh \leq t < (n+1)h$$

it follows

$$\underline{\lim}_{t\to\infty} p_t(x,y) \geq e^{-\lambda_y h}\pi(y) \text{ hence } \underline{\lim}_{t\to\infty} p_t(x,y) \geq \pi(y)$$

$$\overline{\lim}_{t\to\infty} p_t(x,y) = 1 - \underline{\lim}_{t\to\infty} \sum_{z\neq y} p_t(x,z) \leq 1 - \sum_{z\neq y} \underline{\lim}_{t\to\infty} p_t(x,z) \leq 1 - \sum_{z\neq y} \pi(z) = \pi(y)$$

where the first inequality follows by the Fatou's lemma.

Now, we can use the Asymptotic frequency for the discrete Markov chain. Define $N_n^h(y)$ corresponding to p_h, and

$$\lim_{t\to\infty} \frac{1}{t}\int_0^t \mathbf{1}_{\{X_s=y\}} ds = \lim_{t\to\infty}\lim_{h\to 0} \frac{1}{nh}\sum_{k=1}^n \mathbf{1}_{\{X_{hk}=y\}} h = \lim_{h\to 0}\lim_{t\to\infty} \frac{N_n^h(y)}{n} = \pi(y)$$

Here we can interchange the limit due to Moore-Osgood theorem, as the convergence of the Riemann sum is uniform. ■

Example 6.25 (Weather in LA)**.** Suppose weather has 3 states, sunny, smoggy, and rainy. The weather changes as follows

- It is sunny for exponential time, with mean 3, and then gets smoggy.
- It is smoggy for exponential time, with mean 4, and then gets rainy.
- It is rainy for exponential time, with mean 1, and then gets sunny.

Now, let us note that we are given the parameters λ at each state. However, there is only one possibility of jump, so effectively transition rate q's are given. Therefore,

$$Q = \begin{pmatrix} -\frac{1}{3} & \frac{1}{3} & 0 \\ 0 & -\frac{1}{4} & \frac{1}{4} \\ 1 & 0 & -1 \end{pmatrix}$$

Let us find the equilibrium distribution by $\pi Q = 0$. That is,

$$-\pi(1)/3 + \pi(3) = 0, \;\; \pi(1)/3 - \pi(2)/4 = 0, \;\; \pi(2)/4 - \pi(3) = 0,$$

Note that adding the first and third equation yields the second. Hence, we need to use that total mass adds up to 1 to solve it. Solving this yields,

$$\pi(1) = 3/8, \;\; \pi(2) = 4/8, \;\; \pi(3) = 1/8$$

Therefore, long term fractions are given by these probabilities.

6.3 Detailed Balance Condition

Definition 6.26. Assume that a distribution π satisfies

$$\pi(x)Q(x,y) = \pi(y)Q(y,x)$$

for all $x, y \in \mathcal{S}$. Then it is said to satisfy the detailed balance condition.

As expected, if π satisfies the detailed balance condition, it is a stationary distribution.

Example 6.27 (Birth and Death chain)**.** Assume the state space $\mathcal{S} = \mathbb{N}$. Let

$$Q(n, n+1) = \lambda_n,\ n < N, \quad Q(n, n-1) = \mu_n,\ 0 < n$$

For example, if each individual gives birth to one child with rate λ and dies with rate μ, then $\lambda_n = n\lambda$ and $\mu_n = n\mu$. Assuming $\lambda_n, \mu_n > 0$ for all n, the chain is irreducible.

Now, let us find the stationary distribution by the detailed balance condition.

$$\pi(n)\mu_n = \pi(n-1)\lambda_{n-1} \implies \pi(n) = \pi(0)\frac{\lambda_{n-1}\cdots\lambda_0}{\mu_n\cdots\mu_1},\ n > 0$$

As the total mass equals to 1, we get

$$\pi(0) = \frac{1}{1 + \sum_{k=1}^{N} \frac{\lambda_{k-1}\cdots\lambda_0}{\mu_k\cdots\mu_1}}$$

which yields the stationary distribution.

Lemma 6.28. *For a Poisson process N_t with rate λ, it holds that*

$$\lim_{t\to\infty} \frac{N_t}{t} = \lambda$$

almost surely.

Proof. Note that, for integer t,

$$\frac{N_t}{t} = \frac{1}{t}\sum_{k=1}^{t} N_k - N_{k-1}$$

and since N_t has independent increments, we conclude by the Strong Law of Large Numbers. ■

Theorem 6.29. *Let N_t be a Poisson process with rate λ and jump times T_k. Consider an irreducible CTMC X_t with the stationary distribution π, which may depend on N_t. For a state $x \in \mathcal{S}$, define*

$$N_t^x := \sum_{k\geq 1} \mathbf{1}_{\{T_k \leq t\}} \mathbf{1}_{\{X_{T_k-} = x\}}$$

which counts the number of arrivals of the Poisson process during which the Markov chain is in state x immediately before arrivals. Then, for any state x such that either X_t leaves x independently of N_t, or X_t leaves x at the next arrival of N_t, we have

$$\lim_{t\to\infty} \frac{N_t^x}{t} = \lambda\pi(x) \quad \textit{and hence} \quad \lim_{t\to\infty} \frac{N_t^x}{N_t} = \pi(x)$$

Proof. For a given state x, introduce stopping times $\tau_k^{\text{in}}, \tau_k^{\text{out}}$ as the k-th jump time to state x and the k-th leaving time from state x of X_t. Let us also define number of jumps to state x as $n_t = \max\{k \geq 1 : \tau_k^{\text{in}} \leq t\}$. Then,

$$N_t^x := \sum_{k\geq 1} \mathbf{1}_{\{T_k \leq t\}} \mathbf{1}_{\left\{X_{T_k^-} = x\right\}} = \sum_{k=1}^{n_t} N_{\tau_k^{\text{out}} \wedge t} - N_{\tau_k^{\text{in}}}$$

where $X_{t^-} := \lim_{s\uparrow t} X_s$. Now, let us start by (i) and decompose N_t^x/t as;

$$N_t^x/t = \left(\frac{1}{n_t}\sum_{k\geq 1}^{n_t} N_{\tau_k^{\text{out}}\wedge t} - N_{\tau_k^{\text{in}}}\right)\left(\frac{n_t}{\int_0^t \mathbf{1}_{\{X_s=x\}} ds}\right)\left(\frac{\int_0^t \mathbf{1}_{\{X_s=x\}} ds}{t}\right)$$

First, $n_t \to \infty$ as $t \to \infty$ almost surely. Then, the first term is an average of i.i.d. RVs except the very last term. This is due to the fact that X_t leaves x independent of N_t. Thus, by the SLLN,

$$\left(\frac{1}{n_t}\sum_{k\geq 1}^{n_t} N_{\tau_k^{\text{out}}\wedge t} - N_{\tau_k^{\text{in}}}\right) \to \mathbb{E}\left[N_{\tau^x}\right] = \mathbb{E}\left[\mathbb{E}[N_{\tau^x}|\tau^x]\right] = \lambda \mathbb{E}[\tau^x]$$

where τ^x is the leaving time of state x for the chain X_t. For the second term, one observes that the denominator is a sum of n_t many exponential variables, excluding the boundary at t. Thus, again by SLLN,

$$\left(\frac{n_t}{\int_0^t \mathbf{1}_{\{X_s=x\}} ds}\right) \to \frac{1}{\mathbb{E}[\tau^x]}$$

The third term converges to $\pi(x)$ by the Theorem 6.24. Thus, combining all yields

$$N_t^x/t \to \lambda \pi(x)$$

Then, the limit of N_t^x/N_t follows by Lemma 6.28.

Lastly, (ii) follows by the same computations in a simpler manner since $N_{\tau_k^{\text{out}}} - N_{\tau_k^{\text{in}}} = 1$, and since X_t leaves x with the next arrival of the Poisson process, $1/\mathbb{E}[\tau^x] = \lambda$. ∎

Example 6.30 (Barbershop)**.** A barber cuts hair with rate 3 per people. That is, each haircut is exponential random variable with rate 3. The shop has 2 chairs where customers can wait, and no customer is willing to wait standing hence they leave. Customers arrive according to the Poisson process, with rate 2 (in units of hours). Let X_t be the total number of customers in the shop, hence $\mathcal{S} = \{0, 1, 2, 3\}$.

From the verbal description, we claim the generator is given by

$$Q = \begin{pmatrix} -2 & 2 & 0 & 0 \\ 3 & -5 & 2 & 0 \\ 0 & 3 & -5 & 2 \\ 0 & 0 & 3 & -3 \end{pmatrix}$$

State 0 and 3 are clearly matching with the description. For 1 and 2, jump occurs according to the minimum of two exponential distributions, and the Theorem 4.2 together with the Theorem 6.6 shows that we match the given description.

To find the stationary distribution, we can check the detailed balance condition,

$$2\pi(0) = 3\pi(1), \quad 2\pi(1) = 3\pi(2), \quad 2\pi(2) = 3\pi(3)$$

together with the total mass condition, we get

$$\pi(0) = \frac{27}{65}, \quad \pi(1) = \frac{18}{65}, \quad \pi(2) = \frac{12}{65}, \quad \pi(3) = \frac{8}{65}$$

Next, let us find the long run fraction of lost customers. We can use the Theorem 6.29. Note that the customer arrivals is a Poisson process, and X_t leaves the state 3 independent of the customer arrivals. Thus, long run fraction of lost customers is $\pi(3)$.

Suggested Exercises. Durrett, 3rd edition. 4.1, 4.3, 4.5, 4.7, 4.8, 4.12

6.4 Exit Distributions & Exit Times

Definition 6.31. Introduce the hitting time of a set $A \subset \mathcal{S}$ as

$$V_A := \inf\{t \geq 0 : X_t \in A\}$$

and set $V_a := V_{\{a\}}$.

Note that if we are interested in the event $\{V_A < V_B\}$, then we do not need to keep track of the jump times, and hence the embedded Markov chain (see Proposition 6.7) has the exact same events. That is, if we define $\hat{V}$ for the embedded discrete time Markov chain, and take the continuous time Markov chain to be defined by this embedded chain (so that the events can be compared directly), then $\{V_A < V_B\} = \{\hat{V}_A < \hat{V}_B\}$.

Theorem 6.32. *Consider a CTMC with state space $\mathcal{S}$. Let $A, B \subset \mathcal{S}$ such that $C = \mathcal{S} \setminus (A \cup B)$ is finite. If $\mathbb{P}_c(V_A \wedge V_B < \infty) > 0$ for all $c \in C$, then $h(x) = \mathbb{P}_x(V_A < V_B)$ is the unique bounded solution to*

$$h(a) = 1, \quad \forall a \in A, \quad h(b) = 0, \quad \forall b \in B, \quad \text{and} \quad \sum_{y \in \mathcal{S}} Q(c, y)h(y) = 0$$

Proof. Recall that for the embedded Markov chain, transition matrix is given by

$$p_Z(x, y) = \frac{q(x, y)}{\lambda_x} \mathbf{1}_{\{x \neq y\}}$$

We only need to observe that

$$\frac{1}{\lambda_c} \sum_{y \in \mathcal{S}} Q(c, y)h(y) = \sum_{y \neq c} p_Z(c, y)h(y) - h(c)$$

That is,

$$\sum_{y \in \mathcal{S}} Q(c, y)h(y) = 0 \iff h(c) = \sum_{y \in \mathcal{S}} p_Z(c, y)h(y)$$

and hence the Theorem 2.50 implies the result. ∎

Example 6.33 (Barbershop)**.** Let us find what is the probability that the barbershop will be full before it gets empty. That is, $A = \{3\}$, $B = \{0\}$. Recall

$$Q = \begin{pmatrix} -2 & 2 & 0 & 0 \\ 3 & -5 & 2 & 0 \\ 0 & 3 & -5 & 2 \\ 0 & 0 & 3 & -3 \end{pmatrix}$$

hence for $c = \{1, 2\}$, we get

$$3h(0) - 5h(1) + 2h(2) = 0, \;\; 3h(1) - 5h(2) + 2h(3) = 0$$

which yields

$$\mathbb{P}_1(V_3 < V_0) = 4/19, \;\; \mathbb{P}_2(V_3 < V_0) = 10/19$$

Next, we will lift the Theorem 2.52 to the continuous case again by relying on the embedded Markov chain.

Theorem 6.34. *Consider a CTMC with state space $\mathcal{S}$ and jump times $\{T_k\}_{k=0}^{\infty}$. Let $A \subset \mathcal{S}$ such that $C = \mathcal{S} \setminus A$ is finite, and $f : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ where*

$$f(x, y) \geq 0, \quad f(x, x) = 0 \quad \forall x, y \in \mathcal{S}$$

If $\mathbb{P}_c(V_A < \infty) > 0$ for all $c \in C$, then

$$g(x) := \mathbb{E}_x\Big[\sum_{k=1}^{\infty} f(X_{T_{k-1}}, X_{T_k})\mathbf{1}_{\{T_k \leq V_A\}}\Big]$$

is the unique bounded solution to

$$g(a) = 0, \;\; \forall a \in A, \quad \textit{and} \quad \sum_{y \in \mathcal{S}} Q(c, y)g(y) + \sum_{y \in \mathcal{S}} Q(c, y)f(c, y) = 0$$

In particular, if

$$f(x, y) = \frac{1}{\lambda_x}\mathbf{1}_{\{x \neq y\}}$$

then $g(x) = \mathbb{E}_x[V_A]$, and the equation becomes

$$g(a) = 0, \;\; \forall a \in A, \quad \textit{and} \quad \sum_{y \in \mathcal{S}} Q(c, y)g(y) + 1 = 0$$

Proof. As before, consider the representation of CTMC through the embedded Markov chain Z_n with transition matrix p_Z and jump times T_k's as defined in Proposition 6.7. Let $\hat{V}_A$ be the corresponding hitting time for the embedded chain. Note that $\{T_k \leq V_A\} = \{k \leq \hat{V}_A\}$ and $X_{T_k} = Z_{N_{T_k}} = Z_k$. Thus,

$$g(x) = \mathbb{E}_x\Big[\sum_{k=1}^{\infty} f(Z_{k-1}, Z_k)\mathbf{1}_{\{k \leq \hat{V}_A\}}\Big]$$

Then, by the Theorem 2.52, g is the unique bounded solution to $g(a) = 0, \;\forall a \in A$, and

$$\begin{aligned} g(c) &= \sum_{y \in \mathcal{S}} p_Z(c, y)f(c, y) + \sum_{y \in \mathcal{S}} p_Z(c, y)g(y) \\ &= \frac{1}{\lambda_c}\sum_{y \neq c} q(c, y)f(c, y) + \frac{1}{\lambda_c}\sum_{y \neq c} q(c, y)g(y) \end{aligned}$$

Multiplying both sides by λ_c, noting that $f(c, c) = 0$, and rearranging concludes the claim.

Now, let us consider the case $f(x, y) = \lambda_x^{-1}\mathbf{1}_{\{x \neq y\}}$, and recall that

$$\lambda_x^{-1} = \mathbb{E}[T_1 - T_0 | Z_0 = x] = \mathbb{E}_x[T_k - T_{k-1} | Z_{k-1} = x]$$

Thus,

$$\begin{aligned} g(x) &= \mathbb{E}_x\Big[\sum_{k=1}^{\infty} \frac{1}{\lambda_{Z_{k-1}}} \mathbf{1}_{\{k \le \hat{V}_A\}}\Big] \\ &= \mathbb{E}_x\Big[\sum_{k=1}^{\infty} \mathbb{E}_x\Big[T_k - T_{k-1} | Z_{k-1}\Big] \mathbf{1}_{\{k \le \hat{V}_A\}}\Big] \\ &= \mathbb{E}_x\Big[\sum_{k=1}^{\infty} \mathbb{E}_x\Big[(T_k - T_{k-1}) \mathbf{1}_{\{k \le \hat{V}_A\}} | Z_{k-1}\Big]\Big] \qquad \Big(\{k \le \hat{V}_A\} \in \sigma(Z_{k-1})\Big) \\ &= \mathbb{E}_x\Big[\sum_{k=1}^{\infty} (T_k - T_{k-1}) \mathbf{1}_{\{T_k \le V_A\}}\Big] = \mathbb{E}_x[V_A] \end{aligned}$$

Lastly, $\sum_{y \in \mathcal{S}} Q(c,y) f(c,y) = \lambda_c^{-1} \sum_{y \neq c} q(c,y) = 1$ and we are done. ■

Example 6.35 (Barbershop)**.** Let us now compute the expected time for barbershop to get full. That is, $f(x,y) = \lambda_x^{-1} \mathbf{1}_{\{x \neq y\}}$ and $A = \{3\}$. For $x = \{0,1,2\}$, we get

$$-2g(0) + 2g(1) + 1 = 0, \;\; 3g(0) - 5g(1) + 2g(2) + 1 = 0, \;\; 3g(1) - 5g(2) + 1 = 0$$

solving these yields

$$g(0) = 33/8, \; g(1) = 29/8, \; g(2) = 19/8$$

Suggested Exercises. Durrett, 3rd edition. 4.15, 4.16, 4.17, 4.18

7 From Measure Theory to Martingales

7.1 A Tour in Measure Theory

In this section, we will introduce the fundamentals of measure theory and provide an exposure to the core theorems.

The main objective of measure theory is to study functions that assign a real value to subsets of a given set. As a core structural requirement, we aim to understand functions that are additive under disjoint subsets.

In geometry, which is vital for intuition, this notion corresponds to length, area, volume etc. Importantly, measure theory allows the study of integration in a robust way, allowing powerful limit theorems to hold. In our context, probability theory is built upon measure theory.

It turns out that not all subsets of a given set can be considered. Vague intuition is that, a set has no constraints, it is just a collection. And without any requirements, one can construct pathological examples in mathematics. Banach and Tarski proved the following interesting result in 1924:

Let U, V be arbitrary bounded open sets in $\mathbb{R}^m$ for $m \geq 3$. There exists $k \in \mathbb{N}$ and subsets $E_1, \cdots, E_k$, $F_1, \cdots, F_k$ of $\mathbb{R}^m$ such that E_i's partition U, F_i's partition V and E_i is congurent (translation + rotation + reflection) to F_i.

Definition 7.1. A collection of subsets $\mathcal{F}$ of a set Ω is called a σ-algebra, if

- If $\{E_k\}_{k=1}^{\infty} \in \mathcal{F}$, then $\cup_k E_k \in \mathcal{F}$.
- If $E \in \mathcal{F}$, then $E^c \in \mathcal{F}$.

For any set Ω, all subsets 2^{Ω} and $\{\emptyset, \Omega\}$ (trivial) are σ-algebras. Furthermore, intersection of σ-algebras is itself a σ-algebra. Therefore, given any collection of subsets $\mathcal{E}$, we define $\sigma(\mathcal{E})$ as the σ-algebra generated by $\mathcal{E}$.

Exercise 7.2. Show that

(i) $\emptyset$ and Ω is contained in any σ-algebra,

(ii) σ-algebra is closed under countable intersections,

(iii) Intersection of arbitrary family of σ-algebras is a σ-algebra.
Relying on this, define the σ-algebra generated by any collection of subsets $\mathcal{E}$, denoted as $\sigma(\mathcal{E})$.

So, which σ-algebra to take? For any topological space, the most important σ-algebra is the one generated by all the open sets. We call it the Borel σ-algebra. Roughly speaking, Borel σ-algebra contains all the countable intersections and unions of open sets. It is informative to keep in mind (i) geometric objects with partitions, (ii) intervals in (real) numbers, (iii) open balls around continuous functions. (i) serves well for an abstract case, and the discrete nature of it helps the intuition greatly. (ii) is crucial as numbers are, however, as in every area of mathematics, concepts becomes unimaginably powerful once applied to functions and (iii) will serve us as a basis in stochastic processes.

Next proposition is the underlying reason why cumulative distribution function characterizes a probability distribution. The proof (omitted) relies on the fact that every open set in $\mathbb{R}$ is a countable union of open intervals.

Proposition 7.3. *The Borel σ-algebra on $\mathbb{R}$ is generated by* intervals*. Any collection of intervals, such as (a, b), $[a, b]$, $(-\infty, a)$, etc. can be used as a generator of the Borel σ-algebra.*

Definition 7.4. A measure on $(\Omega, \mathcal{F})$ is a function $\mu : \mathcal{F} \to [0, \infty]$ such that

- $\mu(\emptyset) = 0$
- $\mu(\cup_{k=1}^{\infty} E_k) = \sum_{k=1}^{\infty} \mu(E_k)$ for any collection of disjoint subsets E_k in $\mathcal{F}$.

We call $(\Omega, \mathcal{F}, \mu)$ a measure space.

Example 7.5. Let $\mathcal{F} = 2^{\Omega}$[11] for any set Ω, and take any $\rho : \Omega \to [0, \infty]$. Then,

$$\mu(E) := \sum_{x \in E} \rho(x) := \sup \left\{ \sum_{x \in F} \rho(x) \ : \ F \subset E, F \text{ finite} \right\}$$

is a measure on $(\Omega, \mathcal{F})$. In general, ρ might be understood as a density. Two special cases are important:

(i) If $\rho(x) = 1$ for all x, it is called *counting measure*.

(ii) If $\rho(x_0) = 1$ for some $x_0 \in \Omega$ and 0 otherwise, it is called *Dirac measure* or *point mass*.

Next theorem clarifies why probability distributions are characterized by their cumulative distributions.[12]

Theorem 7.6. *Let $F : \mathbb{R} \to \mathbb{R}$ be any increasing, right continuous function. Then there exists a unique measure μ_F on $\mathbb{R}$ with $\mu_F((a, b)) = F(b) - F(a)$. If G is another such function, we have $\mu_F = \mu_G$ if and only if $F - G$ is constant.*

Note that we can generate a significant amount of measures on $\mathbb{R}$ by above theorem. Most important example is the so called Lebesgue measure $\mathbf{m}$ on $\mathbb{R}$, which is the measure associated with $F(x) = x$ where $\mathbf{m}((a, b)) = b - a$. Some basic properties are as follows,

Theorem 7.7. *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space.*

(Monotonicity) *If $E \subset F$, then $\mu(E) \leq \mu(F)$.*

(Subadditivity) $\mu(\cup_k E_k) \leq \sum_k \mu(E_k)$.

(Continuity) *If $E_1 \subset E_2 \subset \cdots$, then $\mu(\cup_k E_k) = \lim_k \mu(E_k)$.*
If $E_1 \supset E_2 \supset \cdots$ and $\mu(E_1) < \infty$ then $\mu(\cap_k E_k) = \lim_k \mu(E_k)$.

Proof. (Monotonicity) Since $\mu(F) = \mu(E \cup (F \setminus E)) = \mu(E) + \mu(F \setminus E)$, and $\mu(\cdot) \in [0, \infty]$, we are done.

(Subadditivity) Set $F_1 = E_1$ and $F_k = E_k \setminus (\bigcup_{j=1}^{k-1} E_j)$. Then, F_k's are disjoint and $\cup_{j=1}^{k} F_j = \cup_{j=1}^{k} E_j$ for all k. Then,

$$\mu(\cup_{k=1}^{\infty} E_k) = \mu(\cup_{k=1}^{\infty} F_k) = \sum_{k=1}^{\infty} \mu(F_k) \leq \sum_{k=1}^{\infty} \mu(E_k)$$

(Continuity from below)

$$\mu(\cup_{k=1}^{\infty} E_k) = \lim_{n \to \infty} \sum_{k=1}^{n} \mu(E_k \setminus E_{k-1}) = \lim_{n \to \infty} \mu(E_n)$$

[11] 2^{Ω} denotes all the subsets.

[12] Probability distribution means $\mu(\Omega) = 1$.

(Continuity from above) Reverse the sequence in the sense that $F_k := E_1 \setminus E_k$. Then $F_1 \subset F_2 \subset \cdots$. Also, $\mu(E_1) = \mu(F_k) + \mu(E_k)$ and $\bigcup_{k=1}^{\infty} F_k = E_1 \setminus (\bigcap_{k=1}^{\infty} E_k)$. Then, by continuity from below,

$$\mu(E_1) = \mu(\cap_{k=1}^{\infty} E_k) + \lim_{k\to\infty} \mu(F_k) = \mu(\cap_{k=1}^{\infty} E_k) + \lim_{k\to\infty} (\mu(E_1) - \mu(E_k))$$

and substract $\mu(E_1)$ from both sides to get the result. ■

Exercise 7.8. (i) Find a sequence E_k such that $\mu(E_1) = \infty$ and continuity from above fails.

(ii) Show that, if $\mu_1, \mu_2, \cdots$ are measures on $(\Omega, \mathcal{F})$, and $a_1, a_2, \cdots \in [0, \infty)$, then $\sum_{k=1}^{\infty} a_k \mu_k$ is a measure on $(\Omega, \mathcal{F})$.

(iii) $(\Omega, \mathcal{F}, \mu)$ is a measure space. Show that $\mu(E) + \mu(F) = \mu(E \cup F) + \mu(E \cap F), \forall E, F \in \mathcal{F}$.

(iv) $(\Omega, \mathcal{F}, \mu)$ is a measure space and fix $E \in \mathcal{F}$. Show that $\mu_E(A) := \mu(A \cap E)$ is a measure.

We say $E \in \mathcal{F}$ is a null set if $\mu(E) = 0$. If a statement is true for all $\omega \in \Omega$ excluding a null set, then we say it holds *almost surely*, or *almost everywhere*.

Next, we will discuss measurable functions. First, recall that any function

$$f : \Omega \to \Lambda$$

induces a mapping

$$f^{-1} : 2^{\Lambda} \to 2^{\Omega} \text{ defined as } f^{-1}(E) = \{\omega \in \Omega : f(\omega) \in E\}$$

which preserves unions, intersection and complements. Therefore,
• If $\mathcal{G}$ is a σ-algebra for Λ, then $\{f^{-1}(E) : E \in \mathcal{G}\}$ is a σ-algebra for Ω.

Definition 7.9. Given two measurable spaces $(\Omega, \mathcal{F})$, $(\Lambda, \mathcal{G})$, a function $f : \Omega \to \Lambda$ is called measurable if $f^{-1}(E) \in \mathcal{F}$ for all $E \in \mathcal{G}$.

Proposition 7.10. *If X, Y are topological spaces, any continuous function is measurable when X, Y are equipped with Borel σ-algebras.*

In fact, measurable functions are closely related to continuous functions but we will not explore this. As an informative example, note that the function

$$f(x) = 1 \text{ for all } x \in [0, 1] \setminus \mathbb{Q} \text{ and } 0 \text{ otherwise}$$

is a measurable function. Since $\mathbf{m}(\mathbb{Q}) = 0$, for arbitrary $\varepsilon > 0$, one can find a domain with measure $1 - \varepsilon$ for which f is continuous.

Introduce the indicator function or characteristic function as

$$\mathbf{1}_{\{E\}}(\omega) := \begin{cases} 1 & \text{if } \omega \in E \\ 0 & \text{if } \omega \notin E \end{cases}$$

which is measurable iff E is in the σ-algebra. Then we have the definition of functions that the integration is build on.

Definition 7.11. We say $\phi : \Omega \to \mathbb{R}$ is simple, if ϕ is measurable and the range is a finite subset of $\mathbb{R}$. The *standard representation* of ϕ is

$$\phi = \sum_{k=1}^{n} x_k \mathbf{1}_{\{E_k\}}, \text{ where } E_k = \phi^{-1}(x_k), \text{ range}(\phi) = \{x_1, \cdots, x_n\}$$

We are ready to talk about the integration now. Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. First, we define the integral of a simple function ϕ with the standard representation as

$$\int \phi d\mu := \sum_k x_k \mu(E_k), \quad \text{and} \quad \int_A \phi d\mu := \int \phi \mathbf{1}_{\{A\}} d\mu := \int \phi \mathbf{1}_{\{A\}}, \forall A \in \mathcal{F} \tag{7.1}$$

Define L^+ as the space of all measurable positive functions,

$$L^+ := \left\{ \text{all measurable } f : \Omega \to \mathbb{R}^+ \right\}$$

We now lift the definition of integral to any $f \in L^+$ as

$$\int f d\mu := \int f := \sup \left\{ \int \phi d\mu \ : \ 0 \le \phi \le f, \ \phi \text{ simple} \right\} \tag{7.2}$$

Since all the functions can be decomposed as $f = f^+ - f^-$[13] to negative and positive parts, we can define integrals if $\int |f| d\mu < \infty$, which we denote all such functions as L^1,

$$L^1 := \left\{ \text{all measurable } f : \Omega \to \mathbb{R} \text{ s.t. } \int |f| d\mu < \infty \right\}$$

Exercise 7.12. Show that,

(i) when f is a simple function, (7.2) agrees with (7.1).

(ii) $c \int f = \int cf$, and if $f \le g$ then $\int f \le \int g$.

Let us note down some expected properties of integration, along with a simple but important observation: integration—when considered as a mapping on subsets—defines a measure.

Proposition 7.13. *Let $\phi, \varphi \in L^+$ be simple functions. Then,*

(i) If $c > 0$, then $\int c\phi = c \int \phi$.

(ii) $\int (\phi + \varphi) = \int \phi + \int \varphi$

(iii) If $\phi \le \varphi$, then $\int \phi \le \int \varphi$.

(iv) The map $A \mapsto \int_A \phi d\mu$ is a measure.

Proof. (i) is obvious. (iii): Let $\sum_k x_k \mathbf{1}_{\{E_k\}}$ and $\sum_\ell y_\ell \mathbf{1}_{\{F_\ell\}}$ are the standard representations of ϕ and φ. First, $\phi \le \varphi$ means $x_k \le y_\ell$ whenever $\mu(E_k \cap F_\ell) \neq 0$. Therefore

$$\int \phi = \sum_{k,\ell} x_k \mu(E_k \cap F_\ell) \le \sum_{k,\ell} y_\ell \mu(E_k \cap F_\ell) = \int \varphi$$

[13] $f^+ = \max(0, f)$ and $f^- = \max(0, -f)$. Also, $\int f := \int f^+ - \int f^-$.

(ii): Next, since $\cup_k E_k = \cup_\ell F_k = \Omega$, $E_k = \bigcup_\ell (E_k \cap F_\ell)$ and $F_\ell = \bigcup_k (F_\ell \cap E_k)$. Therefore,

$$\int \phi + \int \varphi = \sum_{k,\ell} (x_k + y_\ell)\mu(E_k \cap F_\ell) = \int \phi + \varphi$$

(iv): Lastly, let $A_m \in \mathcal{F}$ be a collection of disjoint subsets. Then,

$$\int_{\cup_m A_m} \phi d\mu = \int \phi \mathbf{1}_{\{\cup_m A_m\}} = \sum_k x_k \mu(E_k \cap (\cup_m A_m)) = \sum_{k,m} x_k \mu(E_k \cap A_m) = \sum_m \int_{A_m} \phi$$

■

The following examples demonstrate how the theory we are discussing can unify the idea of integration, and the subsequent theorems will show that the generality (or simplicity) of the definitions allows us to obtain powerful (or general) results concerning the relations between integration and convergence.

Example 7.14 (Summation)**.** Let $\Omega = \mathbb{N}$, $\mathcal{F}$ all subsets of $\mathbb{N}$, and $\mu(E) = |E|$. Then

$$\int f d\mu = \sum_{n \geq 0} f(n)$$

Example 7.15 (Lebesgue Integral)**.** Let $\Omega = [a, b]$, $\mathcal{B}$ be the Borel σ-algebra and $\mathbf{m}$ Lebesgue measure. Then $\int f d\mathbf{m} = \int_a^b f(x)dx$ if f has discontinuities only on a set of measure 0.

Recall the function f that is equal to 1 on $[0, 1]$ except $\mathbb{Q}$. We simply identify this function same as identically 1, and integral is well defined to be 1 too. Recall that we typically characterize Rimeann integral on continuous functions, whereas now we have a larger class of functions for which in particular allows us to 'ignore' zero measure events.

Example 7.16 (Probability)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. That is, $\mathbb{P}(\Omega) = 1$. Then,

$$\mathbb{E}X := \frac{1}{\mathbb{P}(\Omega)} \int_\Omega X d\mathbb{P}$$

for any random variable (i.e. measurable function) X.

We now list three basic convergence theorems, which forms the backbone of the theory. These convergence theorems with measurable functions allows one to carry out analysis, whereas working on continuous functions requires verifications case by case. We will omit the proofs for the sake of this course.

Theorem 7.17 (**Monotone Convergence Theorem**)**.** *If $f_k \in L^+$ and $f_k \leq f_{k+1}$ for all $1 \leq k$, then*

$$\lim_{k\to\infty} \int f_k d\mu = \int \lim_{k\to\infty} f_k d\mu$$

Theorem 7.18 (**Fatou's Lemma**)**.** *If $f_k \in L^+$ for all $1 \leq k$,*

$$\int \underline{\lim}\, f_k d\mu \leq \underline{\lim} \int f_k d\mu$$

Theorem 7.19 (**Dominated Convergence Theorem**)**.** *Suppose $f_k \in L^1$ and*

- $\lim_k f_k = f$ *almost everwhere*

- *there exists $g \in L^1$ such that $|f_k| \leq g$ for all k,*

then

$$\lim_k \int f_k d\mu = \int f d\mu$$

Lastly, we will see two more important theorems, simplified considerably.

Theorem 7.20 (**Fubini-Tonelli**). *If $f \in L^+(\Omega \times \Lambda)$ (Tonelli) or $f \in L^1(\Omega \times \Lambda)$ (Fubini), then*

$$\int_{\Omega\times\Lambda} f(x,y) d(\mu \times \nu)(x,y) = \int_\Omega \Big(\int_\Lambda f(x,y) d\nu(y) \Big) d\mu(x) = \int_\Lambda \Big(\int_\Omega f(x,y) d\mu(x) \Big) d\nu(y)$$

where $\mu \times \nu$ is the product measure on $\Omega \times \Lambda$.

Recall Example 7.14, and Fubini-Tonelli allows us to interchange summations.

Theorem 7.21 (**Radon-Nikodym**). *Let μ, ν be σ-finite*[14] *measures on $(\Omega, \mathcal{F})$ where $\nu(E) = 0$ if $\mu(E) = 0$ (denoted as $\nu \ll \mu$). Then there exists a unique (almost everywhere) integrable function $f : \Omega \to \mathbb{R}$ such that*

$$d\nu = f d\mu, \quad \text{that is,} \quad \nu(E) = \int_E f d\mu$$

- Recall that $E \mapsto \int_E f d\mu$ is a measure by Proposition 7.13.

[14] That is, Ω is a countable union of sets with finite measures.

7.2 Basics of Probability Theory

In probability theory, measures (with total mass 1) are typically denoted by $\mathbb{P}$, and the integral is denoted by $\mathbb{E}$ or $\mathbb{E}^{\mathbb{P}}$. Measurable space is typically called the event space, and we typically do not model it except for the sake of introductory examples. Measurable functions $\Omega \to \mathbb{R}$ are called random variables (RVs), denoted as X, Y, Z etc. We always implicitly consider the Borel sigma algebra $\mathcal{B}(\mathbb{R})$ on $\mathbb{R}$. Moreover, we take the change of variable formula as granted:

$$\mathbb{E}[X] = \int_\Omega X d\mathbb{P} = \int_{\mathbb{R}} x d\mu_X(x)$$

where $\mu_X(A) = \mathbb{P}(X \in A)$ is the *law of* X.

Definition 7.22. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and X be a random variable. We denote the sigma algebra generated by X as $\sigma(X) := \{X^{-1}(A) : A \in \mathcal{B}(\mathbb{R})\}$.

Example 7.23. Consider the event space as $\Omega = \{1, 2, 3, 4, 5, 6\}$, $\mathcal{F} = 2^\Omega$ and let $X = 1$ on 6 and 0 otherwise. Then $\sigma(X) = \{\emptyset, \{6\}, \{1, 2, 3, 4, 5\}, \Omega\}$.

As X is measurable, $\sigma(X) \subset \mathcal{F}$ but it might be strict as above. Also, if X is constant, $\sigma(X) = \{\emptyset, \Omega\}$. Roughly speaking, $\sigma(X)$ characterizes how much information X yields. Note that, if X takes finitely many values, then $\sigma(X)$ is generated by finitely many sets.

In this course, we will work with square integrable random variables (Hilbert space),

$$L^2 := \left\{ \text{all RVs } X : \Omega \to \mathbb{R} \text{ s.t. } \mathbb{E}|X|^2 < \infty \right\}$$

For $X, Y \in L^2$, introduce

$$\begin{aligned}
\mathrm{Var}(X) &:= \mathbb{E}\Big[\big(X - \mathbb{E}[X]\big)^2\Big] = \mathbb{E}[|X|^2] - |\mathbb{E}[X]|^2 \\
\mathrm{Cov}(X, Y) &:= \mathbb{E}\Big[\big(X - \mathbb{E}[X]\big)\big(Y - \mathbb{E}[Y]\big)\Big] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \\
\rho(X, Y) &:= \rho_{X,Y} := \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}, \quad \sigma_\cdot := \sqrt{\mathrm{Var}(\cdot)}
\end{aligned}$$

For random vectors $X = (X_1, \cdots, X_d)^\top : \Omega \to \mathbb{R}^d$, we similarly define $\mu_X(A) := \mathbb{P}(X \in A)$ for A in Borel σ-algebra of $\mathbb{R}^d$ and introduce the cumulative distribution function (cdf) as

$$F_X(x) := \mathbb{P}(X_1 \le x_1, \ldots, X_d \le x_d), \quad x \in \mathbb{R}^d$$

We say random variables $X_1, \ldots, X_n$ are independent by the following equivalent definitions

(i)
$$F_{(X_1,\ldots,X_n)}(x_1, \ldots, x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n)$$

(ii) $\mathbb{E}\big[\prod_{i=1}^n g_i(X_i)\big] = \prod_{i=1}^n \mathbb{E}\big[g_i(X_i)\big]$ for any bounded scalar Borel measurable functions $g_1, \ldots, g_n$.

Remark 7.24. Independent RVs $X_1, \ldots, X_n$ induces a product measure on Ω^n. Let $n = 2$ and recall Fubini-Tonelli Theorem 7.20 with $f(x, y) = xy$. We get that

$$\mathbb{E}[XY] = \int_{\mathbb{R}\times\mathbb{R}} xy(d\mu_X \times d\mu_Y) = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} xy d\mu_X \right) d\mu_Y = \int_{\mathbb{R}} x d\mu_X \int_{\mathbb{R}} y d\mu_Y = \mathbb{E}[X]\mathbb{E}[Y]$$

Also, if $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$, or equivalently, $\mathrm{Cov}(X, Y) = 0$, we say X and Y are uncorrelated.

Definition 7.25. Recall the Radon-Nikodym Theorem 7.21. Suppose the law of X satisfies $\mu_X \ll \mathbf{m}$. Then the Radon-Nikodym derivative f_X is called the density (pdf) of X. In particular,

$$\mu_X((a,b)) = \int_{(a,b)} f d\mathbf{m} = \int_a^b f(x)dx$$

Moreover, existence of density is equivalent to F_X being absolutely continuous. In this case, F_X is almost everywhere differentiable and $\partial_x F_X(x) = f(x)$.

Definition 7.26. Suppose $\{X^n\}_{n\geq 1}$ and X are random variables. We say $X^n \to X$

- almost surely if $\mathbb{P}(\lim_{n\to\infty} X^n = X) = 1$,
- in probability, if $\lim_{n\to\infty} \mathbb{P}(|X^n - X| > \varepsilon) = 0$,
- in distribution, if $\lim_{n\to\infty} F_{X^n}(x) = F_X(x)$ for all x where F_X is continuous at x.
- in L^p, if $\lim_{n\to\infty} \mathbb{E}\big[|X^n - X|^p\big] = 0$ for some $p \geq 1$,
- weakly in L^2, if we are considering $X^n, X \in L^2$, and $\lim_{n\to\infty} \mathbb{E}[X^n\eta] = \mathbb{E}[X\eta]$, $\forall \eta \in L^2$.

Note that $\lim_{n\to\infty} \int f d\mu_{X^n} = \int f d\mu_X$ for all bounded, continuous f is equivalent to convergence in distribution.

Remark 7.27. In this remark, we will try to clarify some differences between these convergences. There are a lot of connections to explore, which we are not aiming to do here. Suppose $\Omega = [0,1]$ with Lebesgue measure $\mathbf{m}$, and consider the following examples:
(i) $X^n(\omega) = n\mathbf{1}_{\{[0,1/n]\}}$,
(ii) $X^n(\omega) = \mathbf{1}_{\{[i/2^k,(i+1)/2^k]\}}$ where $n = 2^k + i$ with $0 \leq i < 2^k$,
(iii) $X^{2n}(\omega) = \omega$, $X^{2n-1}(\omega) = 1 - \omega$, and
(iv) X^n is i.i.d. sequence of uniform distributions with mean 0 and variance 1.

• Now, (i) converges to 0 almost surely (a.s.), however, does not converge in L^1. On the other hand, (ii) converges in L^1 whereas does not converge for any $x \in [0,1]$.

• (iii) obviously converges in distribution to the uniform distribution on $[0,1]$. However, it does not converge in probability.

• (iv) converges to 0 weakly in L^2, whereas it trivially converges to uniform measure in distribution. To see the weak convergence in L^2, note that $\{X^n\}$ is an orthonormal sequence. Given any $\eta \in L^2$, we can write

$$\eta = \sum_{n=1}^{\infty} a_n X^n + \eta^\top, \quad \text{where } a_n := \mathbb{E}[\eta X^n]$$

by Bessel's Inequality (1828), $\sum_{n=1}^{\infty} |a_n|^2 \leq \mathbb{E}[|\eta|^2] < \infty$. In particular, $a_n \to 0$ and this is exactly what we need to conclude X^n converges weakly in L^2.

Exercise 7.28. Prove that, if $X^n \to X$ in L^2, then $X^n \to X$ in probability.
[Hint: Chebyshev's inequality.]

We left reader to recall cdf and pdf of some common distributions:
(i) Bernoulli, (ii) Binomial, (iii) Geometric, (iv) Uniform, (v) Exponential

We say X have normal distribution, denoted as $X \sim \mathcal{N}(\mu, \sigma^2)$, if it has the pdf

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Check that $\mathbb{E}[X] = \mu$ and $\mathrm{Var}[X] = \sigma^2$. Moreover, $aX + b \sim \mathcal{N}(\mu + b, a^2\sigma^2)$ and if $Y \sim \mathcal{N}(\nu, \rho^2)$ independent of X, then $X + Y \sim \mathcal{N}(\mu + \nu, \sigma^2 + \rho^2)$. We say Z have standard normal distribution if $Z \sim \mathcal{N}(0, 1)$.

We say $X = (X_1, \ldots, X_n)^\top$ has a multivariate Gaussian distribution if any linear combination of $X_1, \ldots, X_n$ has normal distribution. In particular, if $X_1, \ldots, X_n$ are independent and have normal distribution, then X have multivariate Gaussian distribution. Also, if $X_1, \ldots, X_n$ have Gaussian distribution, then they are independent if and only if they are pairwise uncorrelated. We say that $Z = (Z_1, \ldots, Z_n)$ has a standard Gaussian distribution if $Z_1, \ldots, Z_n$ has independent standard normal distribution. Equivalently, we may define $X = (X_1, \ldots, X_n)^\top$ has multivariate Gaussian distribution if there exists $m \leq n$, a vector $\mu = (\mu_1 \ldots, \mu_n)$ and a $n \times m$ matrix A such that $X = \mu + AZ$ where $Z = (Z_1, \ldots, Z_m)^\top$ has standard Gaussian distribution. We write $X \sim \mathcal{N}(\nu, \Sigma)$ where

$$\Sigma_{i,j} = \mathrm{Cov}(X_i, X_j) = \mathbb{E}\big[(X_i - \mu_i)(X_j - \mu_j)\big] = A_i^\top \mathbb{E}\big[ZZ^T\big]A_j = A_i^\top A_j$$

that is, $\Sigma = A^\top A$. If the covariance matrix Σ is invertible, then the density of $X \sim \mathcal{N}(\nu, \Sigma)$ is given by

$$f_X(x) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\Big(-\tfrac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\Big)$$

The crucial reason why the normal distribution is fundamental is given by the following theorem.

Theorem 7.29 (**Central Limit Theorem**). *Suppose $\{X_n\}_{n\geq 1}$ are i.i.d. with $\mathbb{E}[X_n] = \mu$ and $\mathrm{Var}(X_n) = \sigma^2$. Denote the sample mean $\bar{X}_n := \frac{1}{n}\sum_{i=1}^n X_i$ and $Z_n := \sqrt{n}\frac{\bar{X}_n - \mu}{\sigma}$. Then, Z_n converges to $\mathcal{N}(0,1)$ in distribution.*

Let us also quickly recall the strong Law of Large Numbers (SLLN).

Theorem 7.30 (**Strong Law of Large Numbers**). *Let $X_1, X_2, \cdots$ be pairwise independent identically distributed random variables, where $\mathbb{E}X_1$ exists.* [15]*. Then the sample mean $\bar{X}_n := \frac{1}{n}\sum_{i=1}^n X_i$ converges to $\mathbb{E}X_1$ almost surely.*

[15] That is, $\mathbb{E}X_1^- < \infty$, where $X_1^- = -\min(0, X_1)$

7.3 Conditional Expectation

We have an important result which characterizes the notion of measurability with respect to the σ-algebra of a measurable function:

Theorem 7.31 (**Doob-Dynkin**). *Let X, Y be random variables. Then Y is measurable with respect to $\sigma(X)$ if and only if $Y = h(X)$ for some (Borel) measurable $h : \mathbb{R} \to \mathbb{R}$.*

Definition 7.32 (Conditional Expectation)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and $X \in L^1$. Consider a sub σ-algebra $\mathcal{H} \subset \mathcal{F}$. We call $\mathbb{E}[X|\mathcal{H}] \in L^1$ the conditional expectation of X given $\mathcal{H}$, satisfying

- $\mathbb{E}[X|\mathcal{H}]$ measurable with respect to $\mathcal{H}$, and
- $\int_H \mathbb{E}[X|\mathcal{H}]d\mathbb{P} = \int_H X d\mathbb{P}$ for all $H \in \mathcal{H}$.

Furthermore, if Y is an another random variable, we denote $\mathbb{E}[X|Y] := \mathbb{E}[X|\sigma(Y)]$.

Remark 7.33.

(i) By Doob-Dynkin lemma, $\mathbb{E}[X|Y] = h(Y)$ for some measurable h.

(ii) Equivalent condition to second condition is

$$\mathbb{E}[Y\mathbb{E}[X|\mathcal{H}]] = \mathbb{E}[YX]$$

for all Y measurable with respect to $\mathcal{H}$. If $\mathcal{H} = \sigma(Z)$ for some Z, then we can further write $Y = g(Z)$ and consider all Borel measurable functions g by Doob-Dynkin lemma.

(iii) Conditional expectation exists and unique by the Radon-Nikodym theorem. Namely, $\mu(H) := \int_H X d\mathbb{P}$ is a measure for $(\Omega, \mathcal{H})$, which satisfies $\mu \ll \mathbb{P}_{|\mathcal{H}}$.

(iv) X itself satisfies the second bullet. However, may not be $\mathcal{H}$ measurable.

Example 7.34. If $\mathcal{H} = \{\emptyset, \Omega\}$, then $\mathbb{E}[X|\mathcal{H}] = \mathbb{E}[X]$.

Example 7.35. If X is independent of $\mathcal{H}$, that is,

$$\mathbb{P}(\{X \in B\} \cap H) = \mathbb{P}(X \in B)\mathbb{P}(H) \text{ for all } B \in \mathcal{B}(\mathbb{R}), H \in \mathcal{H}$$

then $\mathbb{E}[X|\mathcal{H}] = \mathbb{E}[X]$. To see this, constanct functions are always measurable. Hence, take $H \in \mathcal{H}$, and then

$$\int_H X d\mathbb{P} = \mathbb{E}[X\mathbf{1}_{\{H\}}] = \mathbb{E}[X]\mathbb{E}[\mathbf{1}_{\{H\}}] = \int_H \mathbb{E}[X|\mathcal{H}]d\mathbb{P}$$

Example 7.36. If X is measurable with respect to $\mathcal{H}$, then $\mathbb{E}[X|\mathcal{H}] = X$.

Example 7.37. Suppose $\Omega_1, \Omega_2, \cdots$ is a partition of Ω, where $\mathbb{P}(\Omega_k) > 0$ for all $1 \leq k$. Let $\mathcal{H} = \sigma(\Omega_1, \Omega_2, \cdots)$. Then we claim

$$\mathbb{E}[X|\mathcal{H}] = \frac{\mathbb{E}[X\mathbf{1}_{\{X \in \Omega_k\}}]}{\mathbb{P}(\Omega_k)} = \frac{1}{\mathbb{P}(\Omega_k)} \int_{\Omega_k} X d\mathbb{P} \quad \text{on } \Omega_k$$

To see this, note that $\mathbb{E}[X|\mathcal{H}]$ is constant on each Ω_k. Therefore, it is measurable with respect to $\mathcal{H}$. Then we need to check the second condition, and it suffices to check for $H = \Omega_k$, which is trivial.

Remark 7.38. As it follows from the example above, if Y is a random variable with discrete values, then

$$\mathbb{E}[X|Y = y] = \frac{\mathbb{E}[X\mathbf{1}_{\{Y=y\}}]}{\mathbb{P}(Y = y)}$$

Example 7.39. Consider two independent fair coin flips X_1, X_2. Then

$$\mathbb{E}[X_1|X_1+X_2] = \begin{cases} \frac{\mathbb{E}[X_1\mathbf{1}_{\{X_1+X_2=0\}}]}{\mathbb{P}(X_1+X_2=0)} & \text{if } X_1+X_2=0 \\ \frac{\mathbb{E}[X_1\mathbf{1}_{\{X_1+X_2=1\}}]}{\mathbb{P}(X_1+X_2=1)} & \text{if } X_1+X_2=1 \\ \frac{\mathbb{E}[X_1\mathbf{1}_{\{X_1+X_2=2\}}]}{\mathbb{P}(X_1+X_2=2)} & \text{if } X_1+X_2=2 \end{cases} = \frac{X_1+X_2}{2}$$

Intuitively, if X_1+X_2 is 0 or 1, we know the value of X_1. If the sum is 1, we have no information about X_1.

Example 7.40. Consider X, Y with joint density function $f(x,y)$. That is,

$$\mathbb{P}((X,Y)\in B) = \int_B f(x,y)dxdy \text{ for } B\in\mathcal{B}(\mathbb{R}^2)$$

If $\mathbb{E}[|g(X)|]<\infty$, then

$$\mathbb{E}[g(X)|Y] = h(Y), \quad \text{where } h(y) = \frac{1}{\int f(x,y)dx}\int g(x)f(x,y)dx$$

To verify this, since h itself is a measurable function, $h(Y)$ is measurable with respect to $\sigma(Y)$. Now, let $A=\{Y\in B\}$ for some $B\in\mathcal{B}(\mathbb{R})$. Then,

$$\int_A h(Y)d\mathbb{P} = \int h(Y)\mathbf{1}_{\{Y\in B\}}d\mathbb{P} = \int_{\mathbb{R}^2} h(y)\mathbf{1}_{\{y\in B\}}f(x,y)dxdy =$$
$$= \int_{\mathbb{R}} h(y)\mathbf{1}_{\{y\in B\}}\Big(\int_{\mathbb{R}} f(x,y)dx\Big)dy = \int_{\mathbb{R}^2}\mathbf{1}_{\{y\in B\}}g(x)f(x,y)dy = \int_A g(X)d\mathbb{P}$$

Proposition 7.41 (Properties of Conditional Expectation)**.**

(Linear) $\mathbb{E}[aX+Y|\mathcal{H}] = a\mathbb{E}[X|\mathcal{H}]+\mathbb{E}[Y|\mathcal{H}]$.

(Monotone) *If $X\le Y$ (almost surely), then* $\mathbb{E}[X|\mathcal{H}]\le\mathbb{E}[Y|\mathcal{H}]$.

(Jensen's inequality) *If ϕ is convex,* $\mathbb{E}|X|, \mathbb{E}[|\phi(X)|]<\infty$, *then* $\phi(\mathbb{E}[X|\mathcal{H}])\le\mathbb{E}[\phi(X)|\mathcal{H}]$.

(Tower property) *If $\mathcal{H}\subset\mathcal{G}$, then* $\mathbb{E}[\mathbb{E}[X|\mathcal{H}]|\mathcal{G}] = \mathbb{E}[\mathbb{E}[X|\mathcal{G}]|\mathcal{H}] = \mathbb{E}[X|\mathcal{H}]$.

- *If $X\in\mathcal{H}$,* $\mathbb{E}|Y|<\infty$, $\mathbb{E}|XY|<\infty$, *then* $\mathbb{E}[XY|\mathcal{H}] = X\mathbb{E}[Y|\mathcal{H}]$.

Remark 7.42. Let's note some particular cases. If $\phi(x)=x^2$, then $(\mathbb{E}[X|\mathcal{H}])^2\le\mathbb{E}[X^2|\mathcal{H}]$. Since by taking $\mathcal{H}=\{\emptyset,\Omega\}$, we also conclude $(\mathbb{E}[X])^2\le\mathbb{E}[X^2]$. By the same choice of $\mathcal{H}$, $\mathbb{E}[\mathbb{E}[X|\mathcal{G}]] = \mathbb{E}[X]$.

Example 7.43 (Random walk)**.** Let ξ_k be i.i.d. random variables with mean μ. Define $Z_n := \sum_{k=1}^n \xi_k$. Then,

$$\mathbb{E}[Z_{n+1}|\xi_1,\dots,\xi_n] = \mathbb{E}[Z_n+\xi_{n+1}|\xi_1,\cdots,\xi_n] = Z_n+\mu$$

7.4 Stochastic Processes

We now introduce the notion of filtrations, to accomodate stochastic processes.

Definition 7.44 (Filtration)**.** Let $\mathbb{I}$ be either $\mathbb{N}$ or $\mathbb{R}^+$. We say $\mathbb{F} = \{\mathcal{F}_n\}_{n\in\mathbb{I}}$ is a filtration if $\mathcal{F}_k \subset \mathcal{F}_n$ whenever $k \leq n$.

Stochastic process is a collection of random variables, indexed by an ordered set $\mathbb{I}$. We will work in continuous time setting with $\mathbb{I} = [0, T]$, and say that stochastic process X is a mapping $[0, T] \times \Omega \to \mathbb{R}$. Instead of viewing a stochastic process as $\{X_t : 0 \leq t \leq T\}$, it is also typical to view it as family of paths $\{X.(\omega), \omega \in \Omega\}$.

Example 7.45. Typically, filtration is generated by a stochastic process. Let X be a stochastic process. Then

$$\mathbb{F}^X := \{\mathcal{F}_t^X\}_{t\in[0,T]}, \quad \mathcal{F}_t^X := \sigma(X_s : s \leq t)$$

is the filtration generated by X.

Definition 7.46 (Adaptedness)**.** We say a stochastic process X is adapted to the filtration $\mathbb{F} = \{\mathcal{F}_t\}_{t\geq 0}$ if X_t is $\mathcal{F}_t$ measurable.

Remark 7.47.

(i) X is always adapted to its own filtration $\mathbb{F}^X$. Recall the random walk $Z_n = \sum_{k=1}^n \xi_k$. Here, Z is adapted to the filtration formed by $\mathcal{F}_n = \sigma(\xi_1, \ldots, \xi_n)$, which is the same filtration as $\mathbb{F}^Z$.

(ii) We are simplifying the discussion here by considering adapted processes. In fact, one needs to consider progressively measurable processes. We call a process X progressively measurable, if restriction of X onto $[0, t]$ is $\mathcal{B}([0, t]) \times \mathcal{F}_t$-measurable for all t. Similarly, we also omit the discussion around what it means for two process to be equal.

Theorem 7.48 (**Kolmogorov's Extension**)**.** *Let $\mu_{t_1,\ldots,t_n}$ be a family of distributions on $\mathbb{R}^n$ satisfying*

$$\mu_{t_1,\ldots,t_n}(A_1 \times \cdots \times A_{i-1} \times \mathbb{R} \times A_{i+1} \times \cdots \times A_n) = \mu_{t_1,\ldots,t_{i-1},t_{i+1},\ldots,t_n}(A_1 \times \cdots \times A_{i-1} \times A_{i+1} \times \cdots \times A_n)$$

for all i and A_i Borel measurable subsets of $\mathbb{R}$. Then, there exists $(\Omega, \mathcal{F}, \mathbb{P})$ and a stochastic process X where joint distribution of $(X_{t_1}, \ldots, X_{t_n})$ is given by $\mu_{t_1,\ldots,t_n}$.

Theorem 7.49 (**Kolmogorov's Continuity**)**.** *Suppose X is a stochastic process where there exists $\alpha, \beta, C > 0$ such that*

$$\mathbb{E}\left[|X_{t,s}|^\alpha\right] \leq C\,|t - s|^{1+\beta}, \quad \forall s, t \in [0, T] \quad \text{where} \quad X_{t,s} := X_s - X_t$$

Then, for any $\gamma \in (0, \beta/\alpha)$, $X.(\omega)$ is γ-Hölder continuous almost surely.[16]

Definition 7.50 (Stopping time)**.** We say $\tau : \Omega \to [0, T]$ is a $\mathbb{F}$-stopping time if $\{\tau \leq t\} \in \mathcal{F}_t$ for all $t \in [0, T]$. Moreover, we introduce the σ-field corresponding to the stopping time τ as

$$\mathcal{F}_\tau := \left\{A \subset \Omega : A \cap \{\tau \leq t\} \in \mathcal{F}_t, \ \forall 0 \leq t \leq T\right\}$$

Intuitively, being measurable with respect to $\mathcal{F}_\tau$ implies that the function is determined by $(\tau, X_{[0,\tau]})$. Let us also note that $\{[0, t], \forall 0 < t < T\}$ creates a basis for the $\mathcal{B}([0, T])$.

Lemma 7.51. *Suppose $A \subset \mathbb{R}^d$ is closed. Then, $\tau = \inf\{t > 0 : X_t \in A\}$ is a $\mathbb{F}^X$ stopping time.*

In case the filtration is generated by a stochastic process X, which is typically the case, τ is a stopping time means we can determine if 'τ ringed before time t' by knowing the path of X from 0 to t (denoted typically as $X_{[0,t]}$).

[16] To be more precise, one needs to say there exists a modification of X that is γ-Hölder continuous almost surely.

7.5 Markov Processes

As we have discussed the measure-theoretic foundations, let us revisit the definition of Markov processes.

Suppose X is $\mathbb{F} = \{\mathcal{F}_t\}_{t\geq 0}$ adapted process. We say X is a Markov process if, for any $0 \leq s < t \leq T$ and bounded Borel measurable φ, it holds

$$\mathbb{E}\big[\varphi(X_t)|\mathcal{F}_s\big] = \mathbb{E}\big[\varphi(X_t)|X_s\big] \text{ a.s.}$$

Roughly, this means $\{X_t : t \leq s\}$ and $\{X_t : t \geq s\}$ are independent given X_s. By Doob-Dynkin's Lemma 7.31, $\mathbb{E}[\varphi(X_t)|X_s] = \psi(X_s)$ for some Borel measurable ψ.

Moreover, we say X is strong Markov process if, for any two stopping time ρ, τ satisfying $\rho \leq \tau$,

$$\mathbb{E}\big[\varphi(X_\tau)|\mathcal{F}_\rho\big] = \mathbb{E}\big[\varphi(X_\tau)|\sigma(\rho, X_\rho)\big] \text{ a.s.}$$

In this case, $\mathbb{E}\big[\varphi(X_\tau)|\sigma(\rho, X_\rho)\big] = \psi(\rho, X_\rho)$.

To see the independence, let $B \in \mathcal{F}$ and $A \in \mathcal{F}_\rho$ and observe that,

$$\begin{aligned}\mathbb{P}\big(\{X_\tau \in B\} \cap A|\sigma(\rho, X_\rho)\big) &:= \mathbb{E}\big[\mathbf{1}_{\{X_\tau\in B\}}\mathbf{1}_{\{A\}}|\sigma(\rho, X_\rho)\big] \\ &= \mathbb{E}\big[\mathbb{E}[\mathbf{1}_{\{X_\tau\in B\}}|\mathcal{F}_\rho]\mathbf{1}_{\{A\}}|\sigma(\rho, X_\rho)\big] \\ \mathbb{P}\big(X_\tau \in B|\sigma(\rho, X_\rho)\big)\mathbb{P}\big(A|\sigma(\rho, X_\rho)\big) &:= \mathbb{E}\big[\mathbf{1}_{\{X_\tau\in B\}}|\sigma(\rho, X_\rho)\big]\mathbb{E}\big[\mathbf{1}_{\{A\}}|\sigma(\rho, X_\rho)\big]\end{aligned}$$

7.6 Martingales

Now, we are ready to give the definition of martingales. These are, in a rough sense, processes that do not drift deterministically.

Definition 7.52 (Martingale)**.** Let $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ be a filtered probability space. We say a stochastic process M_t is a $(\mathbb{F}, \mathbb{P})$-martingale if

- M is adapted to $\mathbb{F}$.
- $\mathbb{E}[|M_t|] < \infty$ for all t.
- $\mathbb{E}[M_t|\mathcal{F}_s] = M_s$ for all $s < t$.

Exercise 7.53. Show that if M_t is a martingale, $\mathbb{E}[M_t] = \mathbb{E}[M_0]$. In the discrete case, we assume $\mathbb{E}[M_{n+1}|\mathcal{F}_n] = M_n$. Show that $\mathbb{E}[M_{n+k}|\mathcal{F}_n] = M_n$ for any $1 \leq k$.

Next, we define the sub and super martingales. Roughly, increasing and decreasing processes, similar to the definition of martingale.

Definition 7.54. We say a stochastic process M_t is a $(\mathbb{F}, \mathbb{P})$-submartingale (supermartingale) if

- M is adapted to $\mathbb{F}$.
- $\mathbb{E}|M_t| < \infty$ for all t.
- $\mathbb{E}[M_t|\mathcal{F}_s] \geq (\leq) M_s$ for all $s < t$.

We can construct martingales from a given process. We will handle some particular cases, and for the sake of examples, we will consider discrete time.

Example 7.55 (Asymmetric simple random walk)**.** Let ξ_i be 1 with probability p and -1 with probability $q = 1 - p$. Then,

$$\mathbb{E}[Z_{n+1}|\mathcal{F}_n^Z] = Z_n + p - q, \text{ where } Z_n := \sum_{k=1}^{n} \xi_k$$

hence if $p - q \geq 0$ then Z_n is a submartingale, and if $p - q \leq 0$ then Z_n is a supermartingale.

Exercise 7.56. Show that $M_n := Z_n + (p - q)n$ is a martingale.

Exercise 7.57 (Random walk)**.** Suppose ξ_k's are i.i.d. with mean 0 and variance σ^2. Then $Z_n = \sum_{k=1}^{n} \xi_k$ and $Z_n^2 - n\sigma^2$ are both martingales.

Example 7.58. $M_n := (q/p)^{Z_n}$ is a martingale, where Z_n as in Example 7.55. It is obviously adapted to the filtration $\mathcal{F}_n^Z$. Next,

$$\mathbb{E}|M_n| \leq (q/p)^n + (q/p)^{-n} < \infty$$

and

$$\mathbb{E}[M_{n+1}|\mathcal{F}_n^Z] = \mathbb{E}[(q/p)^{Z_n}(q/p)^{\xi_{n+1}}|\mathcal{F}_n^Z] = M_n\Big[(q/p)^1 p + (q/p)^{-1} q\Big] = M_n$$

Let us note further properties of martingales.

Theorem 7.59.

- *If M_n is a martingale, and ϕ a convex function where $\mathbb{E}|\phi(M_n)| < \infty$, then $\phi(M_n)$ is a submartingale. In particular, if $\mathbb{E}|M_n|^2 < \infty$, then M_n^2 is a submartingale.*

- *If M_n is a submartingale, and ϕ is non-decreasing convex function where $\mathbb{E}|\phi(M_n)| < \infty$, then $\phi(M_n)$ is a submartingale.*
- *If M_n is a martingale where $\mathbb{E}|M_n|^2 < \infty$, then for any $0 \le \ell \le k \le m \le n$,*
$$\mathbb{E}[(M_n - M_m)M_k] = 0 \text{ and } \mathbb{E}[(M_n - M_m)(M_k - M_\ell)] = 0$$
- *If M_n is a martingale where $\mathbb{E}|M_n|^2 < \infty$, then*
$$M_n^2 - \sum_{k=1}^{n} \mathbb{E}[(M_k - M_{k-1})^2 | \mathcal{F}_{k-1}]$$
is a martingale.

Proof. **Exercise.** ■

Suggested Exercises. Durrett, 3rd edition. 5.1, 5.2.

Next, we will discuss some connections to Markov chains.

Theorem 7.60. *Let X_n be a discrete Markov chain with state space $\mathcal{S}$ and transition matrix p. Assume $f : \mathbb{N} \times \mathcal{S} \to \mathbb{R}$ satisfies,*

- $\sum_{y \in \mathcal{S}} p^n(x, y)|f(n, y)| < \infty$
- $\sum_{y \in \mathcal{S}} p(x, y) f(n + 1, y) = f(n, x)$

holds for all $x \in \mathcal{S}$ and $0 \le n$. Then $M_n := f(n, X_n)$ is a martingale with respect to the filtration of X_n.

Proof. Exercise. ■

Example 7.61 (Branching Process)**.** Let X_n be the size of the population at n-th generation. Suppose each member gives birth to random offsprings. Let Y_ℓ^n be i.i.d. random variables, with mean μ, for all $\ell, n \in \mathbb{N}$. Let $p(k, m) = \mathbb{P}(\sum_{\ell=1}^{k} Y_\ell^n = m)$, which gives the distribution of the size of the population for the next generation.

Note that
$$\mathbb{E}[X_{n+1}|X_n = k] = \sum_{m \in \mathbb{N}} p(k, m)m = \mathbb{E}\left[\sum_{\ell=1}^{k} Y_\ell^n\right] = \mu k$$
That is, $\mathbb{E}[X_{n+1}|X_n] = \mu X_n$. Moreover,
$$\mathbb{E}_x[X_n] = \mathbb{E}_x[\mathbb{E}_x[X_n|X_{n-1}]] = \mu \mathbb{E}_x[X_{n-1}] = \mu^n x$$
Set
$$M_n := f(n, X_n) := \frac{X_n}{\mu^n}$$
Then M_n is a martingale. To see this, first,
$$\begin{aligned}\sum_{m \in \mathbb{N}} p^n(k, m)\frac{m}{\mu^n} &= \sum_{z \in \mathbb{N}} p^{n-1}(k, z) \sum_{m \in \mathbb{N}} p(z, m)\frac{m}{\mu^n} \\ &= \sum_{z \in \mathbb{N}} p^{n-1}(k, z)\frac{z}{\mu^{n-1}} = k < \infty\end{aligned}$$
and
$$\sum_{m \in \mathbb{N}} p(k, m)\frac{m}{\mu^{n+1}} = \frac{k}{\mu^n} = f(n, k)$$
hence we conclude by the theorem that M_n is a martingale.

7.7 Optional Stopping

We now present an important theorem, which allows us to determine if a martingale M_n is still a martingale if we consider M_τ where τ is a stopping time. In gambling, this result will imply that if you are betting on a martingale, you cannot increase your expected return by the option of leaving the game (stopping). An important implication is that if you create a strategy that almost surely wins in a fair game, such as the doubling strategy, then your account must have a chance to blow up. This is because it cannot be uniformly integrable, as defined below. (See the following remark (iii), (iv))

Theorem 7.62 (Optional Stopping)**.** *Let M_n be a (sub,super) martingale and suppose ρ, τ are stopping times with $\rho \leq \tau$. If*

$$\lim_{R\to\infty} \sup_{n\in\mathbb{N}} \mathbb{E}|M_{\tau\wedge n}\mathbf{1}_{\{|M_{\tau\wedge n}|\geq R\}}| = 0 \qquad (\textit{i.e. } M_{\tau\wedge n} \textit{ is uniformly integrable})$$

then

$$\mathbb{E}[M_\tau|\mathcal{F}_\rho](\geq, \leq) = M_\rho$$

Remark 7.63. **(i):** Let us recall the definition of the σ-algebra generated by a stopping time;

$$\mathcal{F}_\rho := \left\{A \in \cup_{t>0}\mathcal{F}_t : A \cap \{\rho \leq t\} \in \mathcal{F}_t,\ \forall t > 0\right\}$$

and remark that the σ-algebra of a stopping time indeed relies on the filtration of the stochastic process it depends on.

(ii): If $M_{\tau\wedge n}$ is uniformly integrable, then $\mathbb{E}[M_\tau] = \mathbb{E}[M_0]$.

(iii): Note that, if τ is almost surely bounded, then $M_{\tau\wedge n}$ is uniformly integrable.

In particular, as $n \wedge \tau$ is bounded for any τ, $M_{n\wedge\tau}$ is a (sub,super) martingale.

(iv): If $|M_{n\wedge\tau}| < C$ almost surely for all $n \in \mathbb{N}$, again the condition of the Theorem 7.62 is satisfied.

(v): If M_n itself is uniformly integrable, then $M_{n\wedge\tau}$ is uniformly integrable for any stopping time τ.

An important result for applications is as follows;

Theorem 7.64. *Suppose M_n is a (sub, super) martingale and satisfies $\mathbb{E}[|M_{n+1} - M_n||\mathcal{F}_n] < C$ almost surely. If τ is a stopping time with $\mathbb{E}\tau < \infty$, then $M_{n\wedge\tau}$ is uniformly integrable.*

Above theorem is a generalization of Wald's identity. To see this, let $X_1, X_2, \cdots$ be i.i.d. integrable random variables and set $S_n = \sum_{k=1}^n X_k$. We have seen that $M_n = S_n - \mu n$ is a martingale, where $\mathbb{E}[|M_{n+1} - M_n||\mathcal{F}_n] = \mathbb{E}[|X_{n+1} - \mu||\mathcal{F}_n] < \mathbb{E}|X_1| + |\mu|$.

Theorem 7.65 (Wald's Identity)**.** *Let τ be a stopping time with $\mathbb{E}[\tau] < \infty$. Then,*

$$\mathbb{E}[S_\tau] = \mu\mathbb{E}[\tau]$$

Let us also present Wald's Second Identity.

Theorem 7.66 (Wald's Second Identity)**.** *Let τ be a stopping time with $\mathbb{E}[\tau] < \infty$. Then,*

$$\mathbb{E}[(S_\tau - \mu\tau)^2] = \sigma^2\mathbb{E}[\tau]$$

Proof. First, we can assume $\mu = 0$ by considering $\tilde{X} := X - \mu$. Then, we observe that

$$S^2_{n\wedge\tau} = S^2_{(n-1)\wedge\tau} + (2X_nS_{n-1} + X_n^2)\mathbf{1}_{\{\tau\geq n\}}$$

Since $\mathbb{E}[X_n] = 0$, X_n is independent of S_{n-1}, both $X_n, S_{n-1} \in L^2$, and $\mathbf{1}_{\{\tau \geq n\}} = 1 - \mathbf{1}_{\{\tau \leq n-1\}} \in \mathcal{F}_{n-1}$, we get

$$\begin{aligned}\mathbb{E}\big[S^2_{n\wedge\tau}\big] &= \mathbb{E}\big[S^2_{(n-1)\wedge\tau}\big] + 2\mathbb{E}\big[\mathbf{1}_{\{\tau\geq n\}} S_{n-1}\mathbb{E}[X_n|\mathcal{F}_{n-1}]\big] + \mathbb{E}\big[\mathbf{1}_{\{\tau\geq n\}}\mathbb{E}[X_n^2|\mathcal{F}_{n-1}]\big] \\ &= \mathbb{E}\big[S^2_{(n-1)\wedge\tau}\big] + \sigma^2\mathbb{P}(\tau \geq n) \\ &= \sigma^2 \sum_{k=1}^{n} \mathbb{P}(\tau \geq k)\end{aligned}$$

Since $S_{n\wedge\tau} - S_{m\wedge\tau} = \sum_{k=m+1}^{n} X_k$, the same computation above shows that

$$\mathbb{E}\big[(S_{n\wedge\tau} - S_{m\wedge\tau})^2\big] = \sigma^2 \sum_{k=m+1}^{n} \mathbb{P}(\tau \geq k)$$

and hence $S_{n\wedge\tau}$ is a Cauchy sequence in L^2. Now, $\tau < \infty$ almost surely, which means $S_{n\wedge\tau} \to S_\tau$ almost surely. This is the same limit in L^2, which concludes the result. ■

Theorem 7.67 (Doob's supermartingale inequality)**.** *Let X_n be a supermartingale, where $X_n \geq 0$ for all n. Then for $\lambda > 0$,*

$$\mathbb{P}(\sup_n X_n > \lambda) \leq \frac{1}{\lambda}\mathbb{E}X_0$$

Proof. Let $\tau = \inf\{k : X_k \geq \lambda\}$. Then $\{\sup_n X_n > \lambda\} = \{\tau < \infty\}$. Now, by the Optional Stopping Theorem applied with $\tau \wedge n$, and as X_n is a supermartingale,

$$\mathbb{E}X_0 \geq \mathbb{E}X_{n\wedge\tau} = \mathbb{E}[X_\tau \mathbf{1}_{\{\tau\leq n\}}] + \mathbb{E}[X_n \mathbf{1}_{\{\tau>n\}}] \geq \mathbb{E}[X_\tau \mathbf{1}_{\{\tau\leq n\}}] \geq \lambda\mathbb{P}(\tau \leq n)$$

where the second inequality is because $X_n \geq 0$. Let $n \to \infty$ and use continuity of measures to conclude the result. ■

Next, we will discuss Doob's decomposition theorem. We say a stochastic process H_n is $\mathbb{F}$-predictible, if H_0 is $\mathcal{F}_0$ and H_n is $\mathcal{F}_{n-1}$ measurable. As the name suggests the intuition, we can predict the H_n before the time n.

Theorem 7.68 (Doob's Decomposition Theorem)**.** *Let X_n be any submartingale. Then there exists unique martingale M_n and a predictible increasing process H_n with $H_0 = 0$ such that*

$$X_n = M_n + H_n$$

Proof. Exercise. ■

Next, we will consider some examples. First, we will consider a trading strategy. To formalize this, we denote S_n as the price of an asset, and H_n denotes the strategy, namely how much asset we hold. We require that H_n is predictible. Since we observe a price at n, we should have known how much we invest at $n-1$. Now, wealth is given by

$$W_n = W_0 + \sum_{k=1}^{n} H_k(S_k - S_{k-1}) \tag{7.3}$$

Here, H_k denotes the amount of asset we have at S_{k-1} and wealth is updated by the price action.

Proposition 7.69. *Suppose S_n is a martingale, and H_n is a predictible bounded process. Then, (7.3) is a martingale.*

Proof. Exercise. ∎

Suppose we have a stopping time τ to determine when to stop trading the asset, which is almost surely bounded. Then W_τ is a martingale, hence expected return is 0.

Example 7.70 (Simple random walk)**.** Let $X_1, X_2, \cdots$ be i.i.d. random variables taking values $1, -1$ with probabilities $1/2$. Set $S_n^x = x + \sum_{k=1}^n X_k$, which is a martingale. Consider $a \le x \le b$, and define a stopping time

$$\tau_{a,b}^x = \inf\{n \ge 0 : S_n^x \le a \text{ or } b \le S_n^x\} \quad \big(= V_a \wedge V_b\big)$$

Note that $|S_{n\wedge\tau}| \le |a| + |b|$, therefore S_τ is a martingale. To avoid small terms, assume b, a, x are all integers. Now,

$$x = \mathbb{E}S_0^x = \mathbb{E}S_\tau^x = a\mathbb{P}(S_\tau^x = a) + b(1 - \mathbb{P}(S_\tau^x = a))$$

That is, we computed

$$\mathbb{P}(V_a < V_b) = \mathbb{P}(S_\tau^x = a) = \frac{b-x}{b-a}$$

Next, let us compute that $\mathbb{E}\tau_{a,b}^x = \mathbb{E}\tau$ for the symmetric case. Note that we cannot use Wald's identity directly, because it will yield $0 = \mathbb{E}\tau \cdot 0$. Therefore, let $M_n = S_n^2 - n$ which is a martingale **(Exercise)**. Note that

$$|M_{n\wedge\tau}| \le a^2 + b^2 + \tau$$

From the study of Markov chains we know $\mathbb{E}\tau < \infty$, and hence by the Dominated Convergence Theorem,

$$\lim_{n\to\infty} \mathbb{E}[M_{n\wedge\tau}] = \mathbb{E}[M_\tau]$$

On the other hand, by the Optional Stopping Theorem, $x^2 = \mathbb{E}[M_0] = \mathbb{E}[M_{n\wedge\tau}]$ and by letting $n \to \infty$;

$$x^2 = \mathbb{E}M_\tau = a^2\frac{x-b}{a-b} + b^2\frac{a-x}{a-b} - \mathbb{E}\tau$$

That is,

$$\mathbb{E}\tau = (x-a)(b-x)$$

Now, let's handle the asymmetric case. Let X_k takes values $1, -1$ with probabilities p, q respectively. We have seen that

$$M_n = (q/p)^{S_n^x}$$

is a martingale. Let $\tau_{a,b}^x = V_a \wedge V_b$ as before. Then,

$$|M_{n\wedge\tau}| \le (q/p)^a + (q/p)^{-a} + (q/p)^b + (q/p)^{-b}$$

which is uniformly bounded. Therefore,

$$(q/p)^x = \mathbb{E}M_0 = \mathbb{E}M_\tau = (q/p)^a\mathbb{P}(V_a < V_b) + (q/p)^b(1 - \mathbb{P}(V_a < V_b))$$

which yields

$$\mathbb{P}(V_a < V_b) = \frac{(q/p)^x - (q/p)^b}{(q/p)^a - (q/p)^b}$$

Note that $S_n - (p-q)n$ is also a martingale. However, it does not help to solve the problem.

Example 7.71 (Doubling Strategy)**.** Suppose $S_n = \sum_{k=1}^{n} X_k$ is a symmetric simple random walk starting from 0. Let $\tau = \inf\{n \geq 0 : X_n = 1\}$, which tells that the gambler will stop when there is a succesful bet. Set $H_n = 2^{n-1}$, meaning that at each step gambler doubles the bet. Then the wealth is given by

$$W_n = \sum_{k=1}^{n} 2^{k-1} X_k \quad \text{and because} \sum_{k=1}^{n} 2^{k-1} = 2^n - 1, \quad W_\tau = 1$$

Note that $\mathbb{E}W_\tau = 1 \neq 0 = \mathbb{E}W_0$, and hence $W_{n\wedge\tau}$ is not uniformly integrable.

Suggested Exercises. Durrett, 3rd edition. 5.6, 5.7, 5.8.

8 Appendix

Lemma 8.1. *Consider a bounded sequence $|a_i| < C$, and suppose there exists a sequence of indices n_k such that*

$$\frac{a_1 + \cdots + a_{n_k}}{n_k} \to a$$

where $\lim_{k\to\infty} n_{k+1}/n_k \to 1$ *and* $\lim_{k\to\infty} n_k = \infty$. *Then*

$$\lim_{n\to\infty} \frac{a_1 + \cdots + a_n}{n} \to a$$

Proof. To show this, for any n, there exists appropriate k such that $n_k \leq n < n_{k+1}$,

$$\frac{a_1 + \cdots + a_{n_k} + C(n_{k+1} - n_k)}{n_k} \geq \frac{a_1 + \cdots + a_n}{n} \geq \frac{a_1 + \cdots + a_{n_{k+1}} - C(n_{k+1} - n_k)}{n_{k+1}}$$

and taking the limit shows the result. ■

Lemma 8.2. *For any function f and g with range $\mathbb{R}$, it holds*

$$\sup_x (f(x) + g(x)) \leq \sup_x f(x) + \sup_x g(x)$$

and hence

$$|\sup_x f(x) - \sup_x g(x)| \leq \sup_x |f(x) - g(x)|$$

Proof. The first result is trivial since obviously

$$f(x) + g(x) \leq \sup_x f(x) + \sup_x g(x)$$

Now, to see the second one, note that

$$\sup_x f(x) = \sup_x (f(x) - g(x) + g(x)) \leq \sup_x (f(x) - g(x)) + \sup_x g(x)$$

then, since $\sup_x (f(x) - g(x)) \leq \sup_x |f(x) - g(x)|$,

$$\sup_x f(x) - \sup_x g(x) \leq \sup_x |f(x) - g(x)|$$

The right hand side is symmetric with respect to f, g, and thus conclude by changing the roles of f and g. ■

Theorem 8.3 (Banach Fixed Point)**.** *Let $(\mathcal{S}, d)$ be a complete metric space and $T : \mathcal{S} \to \mathcal{S}$ be a contraction mapping with constant $0 < \lambda < 1$, i.e.*

$$d(T(x), T(y)) \leq \lambda d(x, y)$$

Then there exists a unique fixed point x^* of T, i.e.

$$T(x^*) = x^*$$

Proof. Choose arbitrary $x_0 \in \mathcal{S}$. Define the sequence $x_n := T(x_{n-1})$ for all $n \geq 1$. Observe that

$$d(x_{n+1}, x_n) = d(T(x_n), T(x_{n-1})) \leq \lambda d(x_n, x_{n-1}) \leq \lambda^n d(x_1, x_0)$$

We now show that x_n is a Cauchy sequence, and then since $\mathcal{S}$ is complete, by definition it converges to some x^*. By triangle inequality,

$$d(x_m, x_n) \leq \sum_{k=n+1}^{m} d(x_k, x_{k-1}) \leq d(x_1, x_0) \sum_{k=n+1}^{m} \lambda^k = d(x_1, x_0)\lambda^n \frac{1-\lambda^{m-n}}{1-\lambda}$$

which tends to 0 as $n \to 0$, hence x_n is a Cauchy sequence. Next, we claim the limit x^* is the fixed point.

$$\begin{aligned} d(T(x^*), x^*) &\leq d(T(x^*), x_n) + d(x_n, x^*) \\ &= d(T(x^*), T(x_{n-1})) + d(x_n, x^*) \\ &\leq \lambda d(x^*, x_{n-1}) + d(x_n, x^*) \to 0 \end{aligned}$$

Lastly, suppose there exists another fixed point x. Then,

$$d(x^*, x) = d(T(x^*), T(x)) \leq \lambda d(x^*, x)$$

which is a contradiction if $d(x^*, x) > 0$. ■

References

[1] Berestycki, N. and Sousi P., *Applied Probability* lecture notes (2017).

[2] Cohen A., *Lecture notes*.

[3] Durrett R., *Essentials of stochastic processes.* 3rd edition, Springer (2018).

[4] Durrett R., *Probability: theory and examples*, 4th edition, Cambridge University Press (2010).

[5] Folland B. G., *Real analysis: modern techniques and their applications*, 2nd edition, Wiley (2007).

[6] Lalley P. S., *Continuous time Markov chains*, lecture notes.

[7] Little D. C. J., *A proof for the queuing formula:* $L = \lambda W$, Operations Research, Vol 9 No 3, 1961.

[8] Little D. C. J. and Graves, C., *Little's Law*, International Series in Operations Research & Management Science, chapter 5, Springer (2008).

[9] Liggett M. T., *Continuous time Markov processes: An introduction*, American Mathematical Society, 113 (2010).

[10] Schinazi B. R., *Classical and spatial stochastic processes*, Springer Science and Business Media (1999).

[11] Strook W. D., *An introduction to Markov processes*, 2nd edition, Springer (2014).

[12] Tijms C. H., *A first course in stochastic models*, John Wiley & Sons (2003).