DISCRETE EVENT STOCHASTIC PROCESSES
Lecture Notes for an Engineering Curriculum

Anurag Kumar
Department of Electrical Communication Engineering
Indian Institute of Science

Contents

Preface

Over a period of 15 years, I taught a course titled *Stochastic Processes and Queueing Theory* to classes mainly comprising communication engineers, and a few computer scientists. The course (popularly called "SPQT" by the students), was aimed primarily at providing material that prepares students for graduate thesis work in communication networking, an area that draws heavily from the tools of stochastic modeling, optimisation, and control. These notes are essentially a transcription of a part of the material I delivered during my lectures. I have dropped "Queueing Theory" from the title, since I have included here only the material on discrete event stochastic processes, with queues being given as important and useful examples.

The emphasis of the course derives mainly from the textbook by Wolff [17]. It is from this source that the course derives its essentially *renewal theoretic* emphasis, which distinguishes it from most traditional courses in random processes and queueing theory taught in electrical sciences curricula. The latter typically comprise discrete and continuous time Markov chains (using a primarily matrix algebraic treatment), followed by the analysis of standard queueing models, such as M/M/1, M/G/1, etc. We have found the renewal theoretic approach very appropriate for two reasons:

1. The generality of the approach permits the student to understand and conceive of stochastic models more general than those provided by standard text-book queueing theory. Standard queueing models have been found to be of limited utility in applications such as communication networking; see, for example, Kumar et al. [12]. On the other hand semi-Markov models, regenerative models, and Markov regenerative models have been found to be very useful, and are essentially developed out of renewal theory.

2. A renewal theory training provides the student with technical ability that allows him/her to be comfortable with the stochastic analysis that accompanies more advanced techniques, such as sequential statistical analysis, and semi-Markov decision theory.

In these notes, several technical details and proofs have been taken from Wolff [17] and from the texts by Çinlar [5], and Karlin and Taylor [10]. Chapter 5 also depends heavily on Kulkarni [11].

Students who attended this course were also taking, or had taken, a first course on probability, random variables, and random processes, from a book such as the classic by Papoulis [15]. With this background, the material presented in these notes can be easily covered in about 28 lectures, each of 1.5 hours duration. After a review of probability theory in Chapter 1, Chapter 2 treats the topic of discrete time Markov chains (DTMCs) in a mainly traditional manner, though some proofs are deferred to results in the following chapter on renewal theory. The reason for taking this approach is that I have found that engineering students develop a physical feel for DTMCs far more easily than for renewal theory. Chapter 3, on renewal theory, always took a substantial amount of time, as it presents somewhat abstract material that students take time to absorb. This is followed by Chapter 4 on continuous time Markov chains, and Chapter 5 on Markov renewal processes. Each chapter is accompanied by several class-tested problems. A solution manual is also available.

Readers looking for other compact treatments of topics covered in this book might wish to consider the books by Ross [16] or Gallager [8].

While this course takes the training of a student interested in stochastic modeling beyond that of a first course in probability, random variables, and random processes, there are several important advanced topics that are needed to round off the stochastic systems training of an engineering researcher. These are martingale theory, weak convergence, and diffusion processes, topics that could reasonably comprise the syllabus of a third course.

Acknowledgements: I lectured from these notes over one and a half decades before deciding to typeset them as a book. Over this period, hundreds of Masters and PhD students in IISc served as sounding boards, testing my presentation with their many questions. Our lab secretary, Chandrika Sridhar, typed up a "raw" version from my handwritten notes. After polishing up this raw version, I handed these notes out to students, many of whom helped as proof-readers, marking up errors and typos on copies of the printed notes. In the recent years, I began to use parts of these notes in our first random processes course, for which I had several teaching assistants, namely, P.T. Akhil, M. Ashok Kumar, K.P. Naveen, K. Premkumar, Chandramani K. Singh, and Vineeth B.S. Chandramani K. Singh was particularly helpful in going through all the problem sets, and checking the correctness of the problems and the solutions. When it was nearing completion, the manuscript was read in its entirety by Arpan Chattopadhyay, who spotted several errors. I am grateful to all these individuals who have helped bring these notes to a wider audience. Finally, of course, the responsibility for their correctness rests with me, and I welcome readers to report any errors to me by email.

Anurag Kumar
IISc, Bangalore

Chapter 1

A Review of Some Basics

1.1 Axioms of Probability

Probability provides a mathematical framework for reasoning under uncertainty. Although there are several such frameworks, we will be exclusively concerned with the one motivated by the relative frequency interpretation, and codified in Kolmogorov's axioms of probability.

In any situation of probabilistic inference we limit our domain of reasoning to a set of possible basic/elementary outcomes, which we denote by Ω, and usually call the *sample space*.

For example if a coin is tossed once then $\Omega = \{Heads, Tails\}$. Abbreviating Heads to H and Tails to T, if a coin is tossed twice then $\Omega = \{HH, HT, TH, TT\}$. Elements of the set Ω are denoted by ω, and the empty set is customarily denoted by $\emptyset$.

It should be noted that, in general, Ω may remain abstract, i.e., we do not necessarily have to pin it down in every problem (for example, the sample space of a complex queueing network may be very hard to describe, and we may proceed to analyse this queueing network without stopping to identify the exact structure of Ω).

Probability is assigned to events or subsets of Ω. P is a real-valued set function on Ω, defined on a collection of subsets $\mathcal{F}$ of Ω.

Definition 1.1. *The set function $P : \mathcal{F} \to \mathbb{R}$, where $\mathcal{F}$ is a collection of subsets of Ω, is said to be a probability measure on $(\Omega, \mathcal{F})$ if the following hold.*

I. $\mathcal{F}$ is a $\sigma-$algebra, i.e., $\mathcal{F}$ is closed under complements, finite unions and countable unions.

II. (a) $P(\Omega) = 1$ (i.e., P is normalised),

(b) for every $A \in \mathcal{F}$, $P(A) \geq 0$ (i.e., P is nonnegative), and

(c) for every countable collection of sets $\{A_i \in \mathcal{F},\ i \geq 1\}$*, such that, for all* $i \neq j,\ A_i \cap A_j = \emptyset$*, we have* $P\left(\cup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$ *(i.e.,* P *is countably additive).*

■

We say that the 3-tuple $(\Omega, \mathcal{F}, P)$ is a probability space. In all probability models there is an underlying probability space. In carrying out calculations, it is not always necessary to explicitly describe the underlying probability space, but knowing how to do so often proves useful.

The following theorem is easily proved by using the additivity and the nonnegativity axioms.

Theorem 1.1. *P is a monotone set function on $\mathcal{F}$, i.e., if $A, B \in \mathcal{F}$, with $A \subset B$, then* $P(A) \leq P(B)$. ■

The second part of the following theorem is called the *union bound*.

Theorem 1.2. *(i) For every* $A, B \in \mathcal{F},\ P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

(ii) If $A_i \in \mathcal{F},\ i \geq 1$*, then*

$$P\left(\cup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P\left(A_i \cap \left(\cup_{j=1}^{i-1} A_j\right)^c\right) \ \leq \ \sum_{i=1}^{\infty} P(A_i)$$

■

Exercise 1.1.

Prove Theorems 1.1, and 1.2. Hint: Theorem 1.1 follows from the additivity and the nonnegativity axioms, and the first part of Theorem 1.2 follows from finite additivity. The second part of Theorem 1.2 follows from the countable additivity axiom after defining a sequence of sets $E_n = A_n \cap (\cup_{k=1}^{n-1} A_k)^c$, for $n \geq 1$, noticing that the sequence of sets $\{E_n, n \geq 1\}$ partitions $\cup_{k=1}^{\infty} A_k$ into disjoint sets, and then applying the monotonicity property. Note that for $n = 1$, $\cup_{k=1}^{n-1} A_k$ is the union of an empty collection of sets, which is the empty set. ■

Example 1.1. ***Bernoulli trials***

Consider an infinite sequence of coin tosses in each of which the probability of heads (and tails) is 0.5. Let us denote the outcome of heads by 1 and that of tails by 0. The sample space then becomes $\Omega = \{0, 1\}^{\infty}$, i.e., all countably infinite binary sequences. Note that such a sample space will arise in any repeated experiment where we are interested only in two outcomes. For example, die tossing where we only observe whether the toss is even

or odd, or packet transmission in a communication network where we are only interested in whether the packet is received correctly or is not. We often call such an experiment Bernoulli trials.

Observe that the probability of any sequence of outcomes, ω, is $\lim_{n\to\infty}\left(\frac{1}{2}\right)^n = 0$. By placing a decimal point to the left any $\omega \in \Omega$ we can think of each ω as the binary representation of a number in $[0,1]$. In this viewpoint each rational number in $[0,1]$ will correspond to two $\omega \in \Omega$, the recurring and the nonrecurring representation. For example, 0.5 corresponds to $100000\cdots$, and also to $011111\cdots$. Let us take the recurring representation in each such case. By doing this we can associate each number in $[0,1]$ with all but countably many elements of Ω. The set we have left out will have probability 0 (why?). Thus Bernoulli trials can be viewed as yielding an outcome in $[0,1]$. Now consider the interval $[0, 0.5]$; this corresponds to all the outcomes in which the first trial yields tails. Hence the probability of this interval is 0.5. Consider the interval $(0.5, 0.75]$; this corresponds to all the outcomes with heads in the first trial and tails in the second trial, and hence has probability 0.25. We see that the probability measure we obtain is the uniform distribution on $[0,1]$. It turns out that the appropriate σ-algebra of events is the smallest σ-algebra containing all the intervals in $[0,1]$. This is a complicated collection of events to describe and we will not pursue this point further in this book. An important point that this example illustrates is that while each elementary outcome has probability 0, sets of outcomes can have positive probability. Obviously such positive probability sets must have an uncountable infinity of elements (why?). ■

1.1.1 Continuity of Probability

Consider $A_i \in \mathcal{F},\ i \geq 1$, such that $A_1 \subset A_2 \subset A_3 \cdots$ (or, respectively, $A_1 \supset A_2 \supset A_3 \cdots$) then $\lim A_i := \cup_{i=1}^{\infty} A_i$ (respectively, $\cap_{i=1}^{\infty} A_i$) and we say that $A_i \uparrow \cup_{i=1}^{\infty} A_i$ (respectively, $A_i \downarrow \cap_{i=1}^{\infty} A_i$). Consider the sequence of numbers $\{P(A_i), i \geq 1\}$. By Theorem 1.1 this is a nondecreasing sequence, and also the sequence is bounded above by 1. Hence $\lim_{i\to\infty} P(A_i)$ exists. Similarly, if $A_i \downarrow \cap_{i=1}^{\infty} A_i$ then $\lim_{i\to\infty} P(A_i)$ exists. The following theorem asserts that in either case the limit of the probabilities of A_i is the probability of the limit set; i.e., probability is a continuous set function.

Definition 1.2. *A monotone set function Φ is said to be continuous from above if for every sequence $\{A_i, i \geq 1\}$, such that $A_i \downarrow A$,*

$$\lim_{n\to\infty} \Phi(A_n) \ \downarrow \ \phi(A)$$

*and similarly for continuity from below. $\Phi(\cdot)$ is **continuous** if it is continuous from above and from below.* ■

Theorem 1.3. *A probability measure P is a continuous set function.*

Proof: Continuity from below: Given that $\{A_i, i \geq 1\}$ is such that $A_i \uparrow A$, define $E_1 = A_1,\ E_i = A_i - A_{i-1},\ i \geq 2$ (where $A_i - A_{i-1}$ means $A_i \cap A_{i-1}^c$). Notice that $P(E_i) = P(A_i) - P(A_{i-1})$. It is easily seen that $E_i \cap E_j = \emptyset$, and that $\cup_{i=1}^{\infty} A_i = \cup_{i=1}^{\infty} E_i$. The following sequence of equalities can then be written down

$$\begin{aligned} P\left(\cup_{i=1}^{\infty} A_i\right) &= P\left(\cup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i) = \lim_{n\to\infty} \sum_{i=1}^{n} P(E_i) \\ &= \lim_{n\to\infty} \left(P(A_1) + P(A_2) - P(A_1) + \cdots + P(A_n) - P(A_{n-1})\right) \\ &= \lim_{n\to\infty} P(A_n)\ . \end{aligned}$$

In the above calculation, the first equality follows from set equality, the second from countable additivity, and the third is just the meaning of $\sum_{i=1}^{\infty} \cdot$. For continuity from above, consider $A_i \downarrow A$. Define $B_i = \Omega - A_i$. Then $B_i \uparrow B$ and we can use continuity from below to reach the desired conclusion. ■

The following example is a simple illustration of some of the above concepts of limits of sets, continuity of $P(\cdot)$, and countable additivity.

Example 1.2.

A close study of this example will help the reader to understand the basic concepts behind countable additivity and continuity. Consider the sample space generated by an infinite sequence of coin tosses, where we denote heads by 1 and tails by 0; i.e., $\Omega = \{0,1\}^{\infty}$. Denote by $\boldsymbol{\omega}$ an element of Ω, with ω_k denoting the kth element of $\boldsymbol{\omega}$, i.e., the kth outcome in the sequence of outcomes denoted by $\boldsymbol{\omega}$.

Let, for $n \geq 1$,

$$\begin{aligned} A_n &:= \{\boldsymbol{\omega} : \omega_1 = 0, \omega_2 = 0, \cdots, \omega_{n-1} = 0, \omega_n = 1\} \\ &= \{\boldsymbol{\omega} : \omega_k = 0 \text{ for all } k < n,\ \omega_n = 1\} \end{aligned}$$

Thus, for each $n \geq 1$, A_n is the set of outcomes in which the first 1 occurs at the nth trial. Define

$$B := \{\boldsymbol{\omega} : \text{there exists } n \text{ s.t. } \omega_n = 1\}$$

It is evident that the A_i are disjoint. Also, $B = \cup_{n\geq 1} A_n$, since for each $\omega \in B$ there is some n such that the first 1 in ω occurs at that n. Hence $P(B) = \sum_{n\geq 1} P(A_n)$ by $\sigma-$additivity. Let us obtain this conclusion in an alternate way. To this end, define

$$B_n = \{\boldsymbol{\omega} :\ \text{there exists } i \leq n \ \text{ s.t. } \omega_i = 1\}$$

Now, clearly $B_1 \subset B_2 \subset B_3 \cdots$, $B_n = \cup_{i=1}^{n} A_i$, and also $B = \cup_{i=1}^{\infty} B_i$. Using the continuity of $P(\cdot)$ from below, we obtain

$$
\begin{aligned}
P(B) &= \lim_{n\to\infty} P(B_n) && (1.1)\\
&= \lim_{n\to\infty} \sum_{i=1}^{n} P(A_i) && (1.2)\\
&= \sum_{i=1}^{\infty} P(A_i) && (1.3)
\end{aligned}
$$

where the second equality is obtained by finite additivity of $P(\cdot)$.

■

1.2 Random Variables

Definition 1.3. *A (real valued) random variable X on $(\Omega, \mathcal{F})$ is a function $X : \Omega \to \mathbb{R}$ such that, for every $x \in \mathbb{R}$, $\{\omega : X(\omega) \leq x\} \in \mathcal{F}$.* ■

Remark: $\{\omega : X(\omega) \leq x\}$ can be read as "the set of all $\omega \in \Omega$ that are mapped by X into the semi-infinite interval $(-\infty, x]$". This subset of Ω can also be written as the inverse image of X, or $X^{-1}((-\infty, x])$. The requirement that, for every $x \in \mathbb{R}$, $X^{-1}((-\infty, x]) \in \mathcal{F}$ is also called the *measurability* condition, and a random variable is said to be a *measurable function* from the sample space to $\mathbb{R}$. The following example illustrates why this condition is imposed.

Example 1.3.

Consider a single toss of a die. Then $\Omega = \{1, 2, 3, 4, 5, 6\}$. When forming the probability space we are only told whether the die fell even or odd. Hence we are only able to assign probabilities to events in

$$\mathcal{F} = \{\emptyset, \{2, 4, 6\},\ \{1, 3, 5\},\ \Omega\}$$

Now consider the following mapping from Ω to $\mathbb{R}$.

$$X(\omega) = \begin{cases} 1 & \text{if } \omega \text{ is divisible by 3} \\ 0 & \text{otherwise} \end{cases}$$

X is not a random variable, as defined above, since $\{\omega : X(\omega) \in (-\infty, 0]\} = \{1, 2, 4, 5\} \notin \mathcal{F}$. ■

Remark: Thus a random variable is basically a *question* that we ask about the experiment (e.g., X asks: "Was the die toss outcome divisible by 3?") and should be answerable based on the available information. In the above example the available information permitted us to only assign probabilities to the events in $\mathcal{F}$. Hence X is not a valid question.

Definition 1.4. *If X is a random variable on $(\Omega, \mathcal{F})$ then the function $F : \mathbb{R} \to [0,1]$ defined by, for $x \in \mathbb{R}$, $F(x) = P(X \leq x)$, i.e., $F(x) = P\{\omega : X(\omega) \leq x\}$, is called the* distribution function *of X or the* cumulative distribution function (c.d.f.) *of X.* ■

Theorem 1.4. *For a random variable X, its c.d.f. $F(\cdot)$ satisfies the following:*

1. *For all $x \in \mathbb{R}$, $F(x) \geq 0$, i.e., $F(\cdot)$ is a nonnegative function.*
2. *For all $x \in \mathbb{R}$, $\lim_{\epsilon \downarrow 0} F(x+\epsilon) = F(x)$, i.e., $F(\cdot)$ is continuous from the right.*
3. *If $x_1, x_2 \in \mathbb{R}$, $x_1 \leq x_2$, then $F(x_1) \leq F(x_2)$, i.e., $F(\cdot)$ is a monotone increasing function.*
4. *If $P(-\infty < X < +\infty) = 1$, then $\lim_{x \to -\infty} F(x) = 0$, and $\lim_{x \to +\infty} F(x) = 1$.*
5. *Let $D := \{x \in \mathbb{R} : x$ is a point of discontinuity of $F(\cdot)\}$, then D is countable, i.e., $D = \{x_1, x_2, x_3, \cdots, \}$ under some indexing of the points of discontinuity ("jump" points) of $F(\cdot)$.*

■

Remarks 1.1.

1. The proofs of the first four parts of Theorem 1.4 are easy exercises, the second part following from the continuity from above of $P(\cdot)$. If $P(X = -\infty) > 0$ (respectively, $P(X = +\infty) > 0$), then the limits in the fourth part will be > 0 (respectively, < 1). If $P(-\infty < X < +\infty) = 1$ then we say that X is a *proper* random variable, otherwise we say that X is *defective*. Correspondingly, we also say that the c.d.f. $F(\cdot)$ of the random variable X is proper or defective.

2. For any $x \in \mathbb{R}$, let us define $F(x+) = \lim_{\epsilon \downarrow 0} F(x+\epsilon)$ and $F(x-) = \lim_{\epsilon \downarrow 0} F(x-\epsilon)$. By the second part of Theorem 1.4, we see that $F(x) = F(x+)$. By the monotone increasing property, it is clear that, $F(x-) \leq F(x)$. At a discontinuity x_i of $F(\cdot)$, define $p_i := F(x_i) - F(x_i-) > 0$. We call p_i the *point mass* at x_i. If $F(\cdot)$ is defective then there would be point masses at $+\infty$, or at $-\infty$, or both. It can shown that the set of discontinuities of a c.d.f. $F(\cdot)$ is countable; hence, the point masses can be indexed $p_i, i = 1, 2, \cdots$. In general, $\sum_{i=1}^{\infty} p_i \leq 1$. If $\sum_{i=1}^{\infty} p_i = 1$ then X is called a *discrete* random variable.

3. Given a c.d.f. $F(\cdot)$, it can essentially (in the sense of computation of probabilities, expectations, etc.) be decomposed as follows. There is a function $f : \mathbb{R} \to [0, \infty)$ such that, for all $x \in \mathbb{R}$, $F(x) = \sum_{i: x_i \leq x} p_i + \int_{-\infty}^{x} f(u)du$. Clearly, $\int_{-\infty}^{x} f(u)du \leq 1$. This function $f(\cdot)$ is called the *density* of the continuous part of the distribution. In general $f(\cdot)$ need not be less than 1; as an example, consider the uniform distribution over $[0, 0.5]$. Also, observe that $f(u)du$ can be interpreted as $P(X \in (u, u+du))$.

■

Remark: At this point, the reader should review the concept of independence of events and of random variables from a textbook on basic probability.

1.2.1 Expectation

A random variable X such that $X(\omega) \geq 0$ for all $\omega \in \Omega$, is called a *nonnegative* random variable.

Definition 1.5. *For a nonnegative random variable X, with distribution $F(\cdot)$, with point mass p_i at x_i, $i \geq 1$, and density $f(\cdot)$, we define the expectation $\mathsf{E}(X)$ by*

$$\mathsf{E}(X) = \sum_{i=1}^{\infty} x_i p_i + \int_0^{\infty} x\, f(x) dx$$

■

The above general expression for $\mathsf{E}(X)$ can be written compactly in terms of the Riemann-Stieltjes integral,[1] as follows

$$\mathsf{E}(X) = \int_0^{\infty} x\, dF(x) \tag{1.4}$$

Expectations of continuous functions of random variables can be treated in an analogous manner.[2] Thus, the Laplace-Stieltjes Transform (LST) of the c.d.f. $F(\cdot)$, for all $s \in \mathbb{C}$, is given by

$$\tilde{F}(s) := \mathsf{E}\big(e^{-sX}\big) = \int_0^{\infty} e^{-sx}\, dF(x).$$

Exercise 1.2.

As an exercise in using Equation (1.4), establish the following useful expressions for the expectation of a nonnegative random variable:

(i) $\mathsf{E}(X) = \int_0^{+\infty} (1 - F(x)) dx$

(ii) If $f(x) = 0, x \in \mathbb{R}$, then $\mathsf{E}(X) = \sum_{k=1}^{\infty} P(X \geq k)$

[1] Readers unfamiliar with the Riemann-Stieltjes integral, but familiar with the Dirac-δ function, can think formally in terms of writing $dF(x) = (\sum_{i=1}^{\infty} p_i \delta(x - x_i) + f(x))dx$. Further, as usual, $\int_0^{\infty} x\, dF(x) = \lim_{a \to \infty} \int_0^a x\, dF(x)$, and this limit can be ∞.

[2] Technical difficulties that arise in more general cases are handled by the rigorous definition of expectation via the Lebesgue-Stieltjes integral.

(Hint: Write $\int_0^\infty x\ dF(x) = \int_0^\infty \int_0^x dy\ dF(x)$, and then interchange the order of integration.) ■

For a real valued random variable X, we denote $X^+ = XI_{\{X\geq 0\}}$, and $X^- = -XI_{\{X\leq 0\}}$; X^+ and X^- are both nonnegative random variables, and are called the *positive part* and the *negative part* of X. Clearly, we can write $X = X^+ - X^-$, and $|X| = X^+ + X^-$

Definition 1.6. *For a random variable X, $\mathsf{E}(X) = \mathsf{E}(X^+) - \mathsf{E}(X^-)$, provided at least one of $\mathsf{E}(X^+)$ or $\mathsf{E}(X^-)$ is finite.* ■

Example 1.4.

Consider the discrete random variable $X \in \{\cdots, -3, -2, -1, +1, +2, +3, \cdots\}$, which takes the values $+k$ and $-k$, each with probability $\frac{3}{\pi^2 k^2}$. Also, consider the random variable $Y \in \{+1, +2, +3, \cdots\}$, which takes the value $+k$ with probability $\frac{6}{\pi^2 k^2}$. Since $\sum_{k=1}^\infty \frac{1}{k^2} = \frac{\pi^2}{6}$, we see that X and Y are proper random variables. Since $\sum_{k=1}^\infty (k \cdot \frac{1}{k^2}) = +\infty$, by Definition 1.6, we conclude that the expectation of X is not defined, whereas $\mathsf{E}(Y) = +\infty$. ■

The following inequality provides a bound on the tail of the distribution of $|X|$ based on the kth absolute moment of X, i.e., $\mathsf{E}\big(|X|^k\big)$.

Theorem 1.5 (Markov Inequality)**.** *For every $k \geq 0,\ \epsilon > 0$*

$$P\{|X| \geq \epsilon\} \leq \frac{\mathsf{E}\big(|\ X\ |^k\big)}{\epsilon^k}$$

Proof: Let $A = \{\omega : |\ X(\omega)\ | \geq \epsilon\}$, and write $|X|^k = |X|^k I_A + |X|^k I_{A^c}$.

$$\begin{aligned}
\mathsf{E}\Big(|X|^k\Big) &= \mathsf{E}\Big(|X|^k\, I_A\Big) + \mathsf{E}\Big(|X|^k\, I_{A^c}\Big) \\
&\geq \mathsf{E}\Big(|X|^k I_A\Big) \text{ (because each term is nonnegative)} \\
&\geq \epsilon^k \mathsf{E}(I_A) \qquad \text{(because } |X|^k I_A \geq \epsilon^k I_A\text{)} \\
&= \epsilon^k P(A)
\end{aligned}$$

Therefore

$$P(A) \leq \frac{\mathsf{E}\big(|X|^k\big)}{\epsilon^k}$$

■

Corollary 1.1 (Chebychev Inequality)**.** *If X is a random variable with finite variance, then*

$$P\{\omega :\ |X - EX| \geq \epsilon\} \leq \frac{\mathsf{Var}(X)}{\epsilon^2}$$

Proof: Let $Y = X - EX$ and apply the Markov Inequality (Theorem 1.5) with $k = 2$. ■
Remark: Note that if a random variable has finite variance then it has finite mean. In general, if $\mathsf{E}\big(|X|^k\big) < \infty$ for some $k \geq 1$, then $\mathsf{E}\big(|X|^j\big) < \infty$ for all $j \leq k$. Further, $\mathsf{E}(|X|) < \infty$ if and only if $\mathsf{E}(X) < \infty$.

Corollary 1.2. *If $X \geq 0$ with probability 1, then $P(X \geq \epsilon) \leq \frac{\mathsf{E}(X)}{\epsilon}$.* ■

Example 1.5.

The above corollary implies that for a nonnegative random variable X, $P(X \geq 10EX) \leq 0.1$, i.e., the probability that a nonnegative random variable is more than 10 times its average value is less than 0.1. Thus expectation itself can be used to obtain a simple bound (usually very weak) on the tail probability of X. ■

1.3 Stochastic Processes

Definition 1.7. *A collection of random variables $\{X_t, t \in T\}$ defined on the same probability space $(\Omega, \mathcal{F})$ is called a random process.* ■

T is called the parameter set or index set. When T is a finite set, then we just have a random vector. In general, T is infinite.

- If T is a countable set then $\{X_t\}$ is said to be a *discrete* parameter process.
- If T is uncountable then $\{X_t\}$ is said to be a *continuous* parameter process. In this case we may also write the parameter as an argument rather than as a subscript, i.e., $X(t)$.

We note that T may not only have the interpretation of time, discrete or continuous. The parameter set could represent, for example, individuals in a population, or points in one dimensional or two dimensional space.
Remark: It is important to note that, given a stochastic process $\{X_t, t \in T\}$, by definition, for each $t \in T$, X_t is a random variable, but for each given $\omega \in \Omega$, $X_t(\omega), t \in T$, is a real valued function over the parameter set T, and is called the *sample path* of the process corresponding to the sample point ω.

Example 1.6.

The following are some examples of stochastic processes (see Figure 1.1).

1. $\{X_i, i \geq 1\}$ is a sequence of independent and identically distributed (i.i.d.) random variables, with common distribution function $F(\cdot)$. By "independent" we mean that the random variables in the collection are mutually independent, i.e., any finite subset of this collection is a set of independent random variables. Thus for any $m \in$

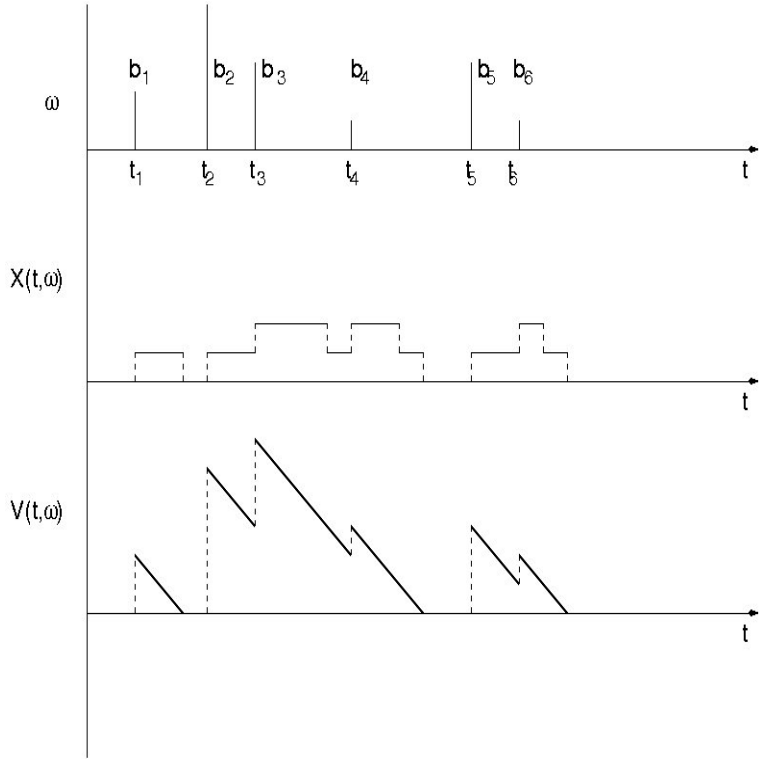


Figure 1.1: A sample path of the queue length process $X(t)$ and the work in system process $V(t)$ for the sample point ω shown at the top. In the depiction of ω, the service requirement of customer k is shown as a vertical line with height b_k.

$\{1, 2, 3, \cdots\}$, and $k_1, k_2, \cdots, k_m \in \{1, 2, 3, \cdots\}$, $F_{k_1,k_2,\cdots,k_m}(x_1, x_2, \cdots, x_m) = \Pi_{j=1}^{m} F(x_j)$. It easily follows that these finite dimensional distributions are consistent. $\{X_i, i \geq 1\}$ is a discrete parameter stochastic process, that can take continuous values.

2. *Processes in a Single Station Queue:* Customers arrive to a waiting room at random times, and bring random amounts of work (expressed in units of time). The customers wait in line in first-come-first-served (FCFS) order. A server works at the rate of 1 second per second on the head-of-the-line (HOL) customer. The server is *nonidling*, which is to mean that it does not idle when there is work to be done. It is easy to see that the evolution of such a queueing system is completely specified when we specify the arrival instants of the customers (say, $(t_1, t_2, t_3, \cdots)$), and the amounts of time required to serve each customer (say, $(b_1, b_2, b_3, \cdots)$). Thus we can take a realisation or outcome of the experiment to

be $\omega = ((t_1, b_1), (t_2, b_2), (t_3, b_3), \cdots)$, and the sample space Ω to be the collection of all such elementary outcomes (see Figure 1.1). The event space $\mathcal{F}$ can also be correspondingly defined, but we will not attempt to do so. Now several useful stochastic processes can be define on this set up.

(a) *The queue length process ($\{X(t),\ t \geq 0\}$):* For each ω, $X(t, \omega)$ is the number of customers in the system at time t. This is a continuous parameter (time) process that takes non-negative integer values (i.e., values in $\mathbb{Z}^+$) (see Figure 1.1).

(b) *The arrival process ($\{A(t), t \geq 0\}$):* For each ω, $A(t, \omega)$ is the number of customers that arrive to the system in the interval $[0, t]$. $A(t)$ is a continuous time discrete values process, taking values in $\mathbb{Z}^+$ (see Figure 1.1).

(c) *The work in the system ($\{V(t), t \geq 0\}$):* For each ω, $V(t, \omega)$ is the total number of seconds of work remaining to be done on the customers in the queue at time t. This is the sum of the residual work to be done on the customer in service, and the total service requirements of the customers that have not yet begun service. $V(t)$ is a continuous time and continuous values process taking values in $\mathbb{R}^+$.

(d) *The sojourn times of the customers ($W_k, k \in \{1, 2, 3, \cdots\}$):* For each ω, $W_k(\omega)$ is the amount of time that the kth customer spends in the system from the instant it arrives until the instant it leaves. W_k is a discrete parameter process (indexed by the customers) that takes continuous values (in $\mathbb{R}^+$). In the example of Figure 1.1 $W_1 = b_1$ and, since $b_1 < t_2 - t_1$, $W_3 = (b_2 - (t_3 - t_2)) + b_3$.

(e) *The number of customers found by each arrival ($X_k^{(a)}, k \in \{1, 2, 3, \cdots\}$)* : For each ω, $X_k^{(a)}(\omega)$ is the number of customers found in the system by the kth arrival. This is a discrete parameter (the indices of the arriving customers) and discrete valued process. In the example of Figure 1.1 $X_1^{(a)} = 0$ and $X_3^{(a)} = 1$.

■

1.3.1 Finite Dimensional Distributions

Definition 1.8. *For any $m \in \{1, 2, 3, \cdots\}$ and a finite set $\{t_1, t_2, \cdots, t_m\} \subset T$*

$$F_{t_1, t_2, \cdots, t_m}(x_1, x_2, \cdots, x_m) = P(X_{t_1} \leq x_1, X_{t_2} \leq x_2, \cdots, X_{t_m} \leq x_m)$$

is called a finite dimensional distribution of the process $X_t, t \in T$. The collection of all such finite dimensional distributions (i.e., for all $m \in \{1, 2, 3, \cdots\}$ and all finite sets $\{t_1, t_2, \cdots, t_m\} \subset T$) is called the set of finite dimensional distributions of the process $\{X_t, t \in T\}$. ■

Given a real valued random variable X on $(\Omega, \mathcal{F})$, its distribution $F(\cdot)$ can be used to obtain an appropriate probability measure on $(\Omega, \mathcal{F})$. In a similar way, a stochastic process can be characterised by the collection of its finite dimensional distributions. But clearly an arbitrary collection of such distributions will not do. To see this, note that if (X, Y) is a random vector then the joint distribution $F_{(X,Y)}(x, y)$ and the individual distributions $F_X(x)$ and $F_Y(y)$ cannot be an arbitrary collection of distributions, but need to be *consistent*; i.e., it must hold that $F_Y(y) = F_{(X,Y)}(\infty, y)$, and $F_X(x) = F_{(X,Y)}(x, \infty)$. In the same way, the finite dimensional distributions of a stochastic process must satisfy the following condition.

Definition 1.9. *A collection of finite dimensional distributions on T is said to be consistent if for all finite sets $T_m \subset T_n \subset T$,*

$$F_{T_m}(x_1, x_2, \cdots, x_m) = F_{T_n}(\infty, \cdots, \infty, x_1, \infty, \cdots, \infty, x_2, \infty, \cdots, \infty, x_m, \infty, \cdots, \infty),$$

where the dimensions that are not in T_m are shown as ∞. ■

In this course we will often define real valued random processes, $X_t, t \in T$, via structural properties. For example, one such structural property, which we will encounter in Chapter 2, is the Markov property. Such properties can be used to prove that the process has consistent finite dimensional distributions. The question then arises: "Is there a probability space $(\Omega, \mathcal{F}, P)$, and a random process, $X_t, t \in T$, whose finite dimensional distributions are the ones derived from the structural properties?" Remarkably, the Kolmogorov Extension Theorem states that if the finite dimensional distributions that are derived are consistent, then, with $\Omega = \mathbb{R}^T$ and an appropriately defined σ-algebra and probability measure, there exists a random process with the same finite dimensional distributions.

We will be concerned only with the situation in which two processes that have the same finite dimensional distributions are essentially the same for all practical purposes. Hence when we define a new kind of stochastic process, one of the first programs that we will have is to determine its finite dimensional distributions. If we are able to do this, it will mean that we have a well defined process. Also, carrying out this program will demonstrate how a definition of a process can end up in completely specifying its finite dimensional distributions, and hence its essential properties.

1.4 Convergence of Random Sequences

Definition 1.10. *A sequence of random variables $\{X_n, n \geq 1\}$ is said to converge* in probability *to a random variable X if, for all $\epsilon > 0$, $\lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0$, and we denote this by $X_n \xrightarrow{p} X$.* ■

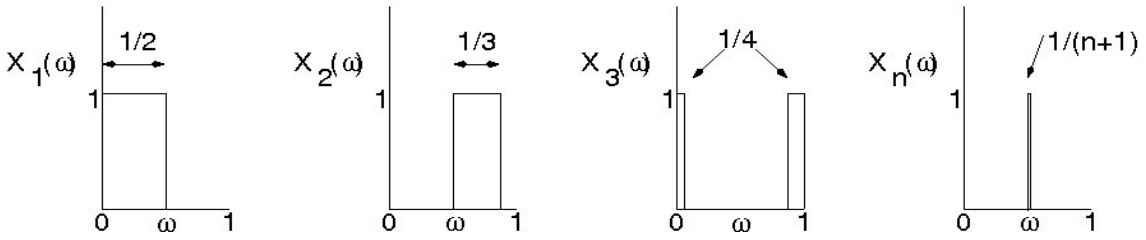


Figure 1.2: The random variables $X_1, X_2, \cdots$, in Example 1.7.

Definition 1.11. *A sequence of random variables* $\{X_n, n \geq 1\}$ *is said to converge* with probability one *or* almost surely *to a random variable* X *if* $P\{\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\} = 1$*, and we denote this by* $X_n \overset{w.p.\ 1}{\longrightarrow} X$*, or by* $X_n \overset{a.s.}{\longrightarrow} X$.

■

We will see that $X_n \overset{\text{w.p. } 1}{\longrightarrow} X$ implies that $X_n \overset{p}{\to} X$. However, the following example shows that the reverse is not true.

Example 1.7.

Consider the sample space $\Omega = [0, 1]$, with the uniform probability over it, i.e., the probability of an interval in $[0, 1]$ is the length of the interval. Define random variables $X_1, X_2, \cdots$ as depicted in Figure 1.2. Notice that the random variable X_n is nonzero only over an interval of width $\frac{1}{n+1}$. Thus for every $\epsilon, 0 < \epsilon < 1$, $P(|X_n - 0| > \epsilon) = \frac{1}{n+1} \to 0$, as $n \to \infty$; i.e., $X_n \overset{p}{\to} 0$. On the other hand, recalling that $\sum_{i=1}^{\infty} \frac{1}{n} = \infty$, we can see that for every ω, for every $m \geq 1$, there is an $n \geq m$, such that $X_n(\omega) = 1$. It follows that $\{\omega : X_n(\omega) \to 0\}$ is empty, and hence X_n does not converge almost surely. ■

1.4.1 Convergence of Expectation

If $X_n, n \geq 1$, is a sequence of random variables converging almost surely to the random variable X, we often need to examine whether $\lim_{n\to\infty} \mathsf{E}(X_n) = \mathsf{E}(X)$. Note that this can be viewed as the validity of the following exchange of limit and expectation:

$$\lim_{n\to\infty} \mathsf{E}(X_n) = \mathsf{E}\Big(\lim_{n\to\infty} X_n\Big)$$

In general, such an exchange is not valid, as is illustrated by the following important example.

Example 1.8.

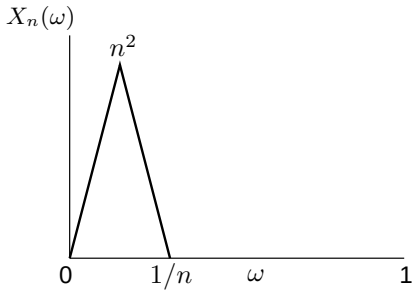


Figure 1.3: The random variable X_n, defined in Example 1.8, shown as a function from $\Omega = [0, 1]$ to $\mathbb{R}$.

Consider the sample space $\Omega = [0, 1]$, with the uniform probability over it. Define random variables $X_1, X_2, \cdots$ as depicted in Figure 1.3. The random variable X_n is zero over the interval $[\frac{1}{n}, 1]$. Evidently, for each $\omega, 0 < \omega \leq 1$, we have $\lim_{n\to\infty} X_n(\omega) = 0$. Defining X to be the random variable that takes value 0 over $[0, 1]$, it follows that

$$X_n \overset{\text{w.p. } 1}{\longrightarrow} X$$

We observe, however, that $\mathsf{E}(X_n) = \frac{1}{2} \times \frac{1}{n} \times n^2 = \frac{n}{2}$. Thus $\lim_{n\to\infty} \mathsf{E}(X_n) = \infty$, whereas $\mathsf{E}(\lim_{n\to\infty} X_n) = 0$. ■

The following two results provide conditions that ensure the validity of the exchange of expectation and almost sure limit of random variables.

Theorem 1.6 (Monotone Convergence)**.** *If $X_n, n \geq 1$, is a sequence of random variables that converges almost surely to the random variable X, i.e., $X_n \overset{\text{w.p. } 1}{\longrightarrow} X$, and the following two conditions hold:*

(i) X_n is non-negative for every $n \geq 1$, and

(ii) $X_1 \leq X_2 \leq \cdots \leq X_n \cdots$ (i.e., the sequence of random variables is monotonically non-decreasing),

then

$$\lim_{n\to\infty} \mathsf{E}(X_n) = \mathsf{E}(X)$$

■

We note that in the situation of Theorem 1.6, it is possible that $\mathsf{E}(X) = \infty$.

Theorem 1.7 (Dominated Convergence)**.** *If $X_n, n \geq 1$, is a sequence of random variables that converges almost surely to the random variable X, i.e., $X_n \overset{\text{w.p. } 1}{\longrightarrow} X$, and the following two conditions hold:*

(i) there is a random variable Y such that $|X_n| \leq Y$ (i.e., the sequence $X_n, n \geq 1$, is dominated by the random variable Y), and

(ii) $\mathsf{E}(Y) < \infty$,

then $\mathsf{E}(X) < \infty$, *and*

$$\lim_{n\to\infty} \mathsf{E}(X_n) = \mathsf{E}(X)$$

■

In the dominated convergence theorem, when Y is a finite constant then the result is called the *bounded convergence theorem.*

1.5 Laws of Large Numbers

We now turn to the simplest stochastic process discussed in Section 1.3 and develop fundamental theorems that are called laws of large numbers. From a foundational point of view these results have the following importance. One of the interpretations of probability is that the probability of an event is the fraction of trials in which the event occurs in independently repeated trials of the experiment, as the number of trials becomes large. This interpretation easily motivates the axioms of probability. For example, for disjoint events A and B, if in n trials n_A is the number of times that event A occurs and n_B is the number of times that event B occurs, then the fraction of time that either A or B occurs is $\frac{n_A}{n} + \frac{n_B}{n}$; this is just the additivity axiom of probability. It is natural to expect, that, when we have set up the mathematical structure of probability, this structure is self consistent in the sense that there is probability close to one that the relative frequency of occurrence of an event in repeated trials converges to the probability of the event.

It is easy to see that such a convergence should take place for random variables as well. Consider repeated independent tosses of a die with the outcomes in $\Omega = \{1, 2, 3, 4, 5, 6\}^\infty$; we write a generic outcome as $\boldsymbol{\omega} = (\omega_1, \omega_2, \omega_3, \cdots)$. Then consider the i.i.d. sequence of random variables defined by $X_i(\boldsymbol{\omega}) = 1$ if ω_i is even and $X_i(\boldsymbol{\omega}) = 2$ if ω_i is odd. Clearly $\mathsf{E}(X_i) = 1.5$. We observe that $\frac{1}{n}\sum_{i=1}^n X_i(\boldsymbol{\omega}) = \frac{1}{n}\sum_{i=1}^n 1 \cdot I_{\{\omega_i \text{ even}\}} + \frac{1}{n}\sum_{i=1}^n 2 \cdot I_{\{\omega_i \text{ odd}\}} \to_{n\to\infty} 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{2} = \mathsf{E}(X_1)$, where the limit is obtained since the fraction of repetitions in which the outcome is odd converges to 0.5 as does the fraction of repetitions in which the outcome is even. Thus the "time" averages of the sequence of i.i.d. random variables converges to the expectation; this is just a generalisation of the convergence of the relative frequency of an event to its probability. We will discuss the laws of large numbers in this generality.

Let us begin by considering a probability space $(\Omega, \mathcal{F}, P)$ and an i.i.d. sequence of random variables $\{X_i,\ i \geq 1\}$ with finite mean $\mathsf{E}(X_1)$. Define for every $n \geq 1$ and $\epsilon > 0$ the events

$$A_n(\epsilon) = \left\{\omega : \left|\frac{1}{n}\sum_{i=1}^n X_i - \mathsf{E}(X_1)\right| \geq \epsilon\right\}$$

Note that, by virtue of the $X_i, i \geq 1$, being random variables on $(\Omega, \mathcal{F})$ the events $A_n(\epsilon)$ are in $\mathcal{F}$. The event $A_n(\epsilon)$ contains all the ωs such that the average value of the stochastic process up to the nth step exceeds the expectation $\mathsf{E}(X_1)$ by at least ϵ. We would expect that the probability of this event is small for large n. This is the content of the following law of large numbers.

Theorem 1.8 (Weak Law of Large Numbers (WLLN))**.** *$X_i, i \geq 1$, is a sequence of i.i.d. random variables such that $\mathsf{Var}(X_i) = \sigma^2 < \infty$, then, for every $\epsilon > 0$, $\lim_{n\to\infty} P(A_n(\epsilon)) = 0$, in other words $\frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{p} \mathsf{E}(X_1)$.*

Remark: This finite variance version is not the most general form of the result, but it is very easy to prove. Note that the finite variance assumption ensures that the expectation exists.

Proof:

$$\begin{aligned} P(A_n(\epsilon)) &\leq \frac{\mathsf{Var}\left(\frac{1}{n}\sum_{i=1}^n (X_i - \mathsf{E}(X_1))\right)}{\epsilon^2} \text{ (by the Chebychev Inequality, Corollary 1.1)} \\ &= \frac{\sigma^2}{n\epsilon^2} \longrightarrow 0 \text{ as } n \to \infty \text{ for each } \epsilon > 0 \end{aligned}$$

where the last step uses the fact that the variance of the sum of independent random variables is the sum of their variances. ■

We will see why this law of large numbers is called "weak" in the course of studying the strong law of large numbers (SLLN), which we now develop. Given $\{X_i, i \geq 1\}$, i.i.d. random variables, consider the event

$$\{\omega : \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n X_i(\omega) = \mathsf{E}(X_1)\}$$

i.e., an ω is in this set if the sample path average of the random process converges to $\mathsf{E}(X_1)$. The strong law of large numbers asserts that, under certain conditions, this event has probability 1; in other words $\frac{1}{n}\sum_{i=1}^n X_i \overset{\text{w.p. } 1}{\longrightarrow} \mathsf{E}(X_1)$. Using the definition of the limit of a sequence of real numbers, we rewrite this event as

$$\left\{\omega : \text{ for every } k \geq 1, \text{ there exists } m \geq 1, \text{ for every } n \geq m, \ \left|\frac{1}{n}\sum_{i=1}^n X_i(\omega) - \mathsf{E}(X_1)\right| < \frac{1}{k}\right\}$$

Taking the complement of this event, we obtain

$$\left\{\omega : \text{there exists } k \geq 1, \text{ for every } m \geq 1, \text{there exists } n \geq m, \ \left|\frac{1}{n}\sum_{i=1}^n X_i(\omega) - \mathsf{E}(X_1)\right| \geq \frac{1}{k}\right\}$$

Using set notation this event can then be written as follows

$$\underbrace{\underbrace{\cup_{k\geq 1}\ \cap_{m\geq 1}\ \cup_{n\geq m}\ \underbrace{\left\{\omega : \left|\frac{1}{n}\sum_{i=1}^{n} X_i(\omega) - \mathsf{E}(X_1)\right| \geq \frac{1}{k}\right\}}_{A_n\left(\frac{1}{k}\right)}}_{A_n\left(\frac{1}{k}\right) \text{ infinitely often}}}_{\text{there exists } k\geq 1 \text{ such that } A_n\left(\frac{1}{k}\right) \text{ infinitely often}}$$

where we have also shown how to read the expression in term of our earlier defined events $A_n(\epsilon)$. Thus the strong law of large numbers would require that this event has probability 0. We note that the phrase "infinitely often" is usually abbreviated to "i.o."

Suppose we can show that, for every $\epsilon > 0$, $A_n(\epsilon)$ i.o. with probability 0 (i.e., $P(\cap_{m\geq 1}\ \cup_{n\geq m} A_n(\epsilon)) = 0$ for every $\epsilon > 0$). Then, using the union bound (Theorem 1.2), we obtain

$$\begin{aligned} P\left(\cup_{k\geq 1}\ A_n\left(\frac{1}{k}\right) \text{ i.o.}\right) &\leq \sum_{k\geq 1} P\left(A_n\left(\frac{1}{k}\right) \text{ i.o.}\right) \\ &= 0 \end{aligned}$$

which will prove SLLN. It can also be seen that this will also imply WLLN, thus showing that WLLN is indeed weaker; observe the following. We have, for every k,

$$P\left(A_n\left(\frac{1}{k}\right) \text{ i.o.}\right) = 0$$

This implies that

$$\begin{aligned} 0 &= P\left(\cap_{m\geq 1}\ \cup_{n\geq m}\ A_n\left(\frac{1}{k}\right)\right) \\ &= \lim_{m\to\infty} P\left(\cup_{n\geq m}\ A_n\left(\frac{1}{k}\right)\right) \\ &\geq \lim_{m\to\infty} P\left(A_m\left(\frac{1}{k}\right)\right) \\ &\geq 0 \end{aligned}$$

where the second equality follows from the continuity of probability (notice that $\cup_{n\geq m} A_n\left(\frac{1}{k}\right) \downarrow \cap_{m\geq 1}\ \cup_{n\geq m}\ A_n\left(\frac{1}{k}\right)$), the first inequality follows since $A_m\left(\frac{1}{k}\right) \subset \cup_{n\geq m}\ A_n\left(\frac{1}{k}\right)$, and the last limit establishes WLLN.

Lemma 1.1 (Borel-Cantelli)**.** *(i) For a sequence of events $A_n, n \geq 1$, if $\sum_{n=1}^{\infty} P(A_n) < \infty$ then $P(A_n$ i.o.$) = 0$.*

(ii) For a sequence of mutually independent events $A_n, n \geq 1$, if $\sum_{n=1}^{\infty} P(A_n) = \infty$ then $P(A_n \text{ i.o.}) = 1$.

Remark: For the case in which the events $A_n, n \geq 1$, are mutually independent, the two parts of this lemma, taken together, are called the *Borel Zero-One Law.* The term Borel-Cantelli Lemma is often used to refer to just Part (i) of the result, and we will adopt this terminology in our discussions.

Proof:

(i) Define the random variable N as follows

$$N(\omega) = \sum_{n=1}^{\infty} I_{A_n}(\omega)$$

i.e., $N(\omega)$ is the number of events in the sequence $A_n, n \geq 1$, to which ω belongs. We would like to show that $\sum_{n=1}^{\infty} P(A_n) < \infty$ implies that $\mathsf{E}(N) < \infty$, which in turn will imply that $P(N = \infty) = 0$, which implies that $\underbrace{P\left(\omega : \sum_{n=1}^{\infty} I_{A_n}(\omega) = \infty\right)}_{P(A_n \text{ i.o.}) = 0} = 0$, thus completing the proof. Thus we need the following argument to hold

$$\begin{aligned}
\mathsf{E}(N) &= \mathsf{E}\left(\sum_{n=1}^{\infty} I_{A_n}\right) \\
&= \mathsf{E}\left(\lim_{m\to\infty} \sum_{n=1}^{m} I_{A_n}\right) \\
&= \lim_{m\to\infty} \mathsf{E}\left(\sum_{n=1}^{m} I_{A_n}\right) \\
&= \lim_{m\to\infty} \sum_{n=1}^{m} P(A_n) \\
&= \sum_{n=1}^{\infty} P(A_n) \\
&< \infty
\end{aligned}$$

In this argument, the only step that requires justification is the third equality, which states that the interchange $\mathsf{E}(\lim_{m\to\infty} \cdot) = \lim_{m\to\infty} \mathsf{E}(\cdot)$ holds true. Here the exchange is permitted by Theorem 1.6 (monotone convergence theorem) when applied to the sequence of random variables $\sum_{n=1}^{m} I_{A_n}$ which are nonnegative and $\sum_{n=1}^{m} I_{A_n} \uparrow N$ with probability 1.

(ii)

$$\begin{aligned}
P(A_n \text{ i.o.}) &= P(\cap_{m=1}^{\infty} \cup_{n=m}^{\infty} A_n) \\
&= \lim_{m\to\infty} P(\cup_{n=m}^{\infty} A_n) \\
&= \lim_{m\to\infty} \lim_{k\to\infty} P(\cup_{n=m}^{k} A_n) \\
&= \lim_{m\to\infty} \lim_{k\to\infty} (1 - P(\cap_{n=m}^{k} A_n^c)) \\
&= \lim_{m\to\infty} \lim_{k\to\infty} \left(1 - \prod_{n=m}^{k} P(A_n^c)\right) \\
&= \lim_{m\to\infty} \lim_{k\to\infty} \left(1 - \prod_{n=m}^{k} (1 - P(A_n))\right) \\
&\geq \lim_{m\to\infty} \lim_{k\to\infty} \left(1 - \prod_{n=m}^{k} e^{-P(A_n)}\right) \\
&= \lim_{m\to\infty} \lim_{k\to\infty} \left(1 - e^{-\sum_{n=m}^{k} P(A_n)}\right) \\
&= 1
\end{aligned}$$

where we have used continuity of probability in the second and third equalities, independence of the events in the fifth equality, the fact that $(1-x) \leq e^{-x}$ for all real x, in establishing the inequality, and the hypothesis that $\sum_{n=1}^{\infty} P(A_n) = \infty$ in the last step.

■

Theorem 1.9 (Strong Law of Large Numbers (SLLN))**.** *For a sequence $X_1, X_2, \ldots$ of i.i.d. random variables with finite variance σ^2 (and, hence, finite mean $\mathsf{E}(X_1)$)*

$$P\left\{\omega : \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_n(\omega) = \mathsf{E}(X_1)\right\} = 1,$$

i.e., $\frac{1}{n}\sum_{i=1}^{n} X_i \overset{w.p.\,1}{\longrightarrow} \mathsf{E}(X_1)$.

Remark: The following proof is provided for completeness; its understanding is not essential for reading the later parts of these notes. The student should understand the role played by the Borel-Cantelli Lemma, Lemma 1.1.

Proof: As before let $A_n\left(\frac{1}{k}\right) = \left\{\omega : \left|\frac{1}{n}\sum_{i=1}^{n} X_i(\omega) - \mathsf{E}(X_1)\right| \geq \frac{1}{k}\right\}$. We wish to show that for every k, $P\left(A_n\left(\frac{1}{k}\right) \text{ i.o.}\right) = 0$.

We will use the Borel-Cantelli Lemma to establish this result. If we were to use the Chebychev inequality, we would obtain

$$\sum_{n=1}^{\infty} P\left(A_n\left(\frac{1}{k}\right)\right) \leq \sum_{n=1}^{\infty} \frac{\sigma^2 k^2}{n}$$

but the sum in the right hand side diverges; hence this is not a useful approach. Instead, let us look at the event sequence only at the indices $m^2, m \geq 1$. This yields

$$\sum_{m=1}^{\infty} P\left(A_{m^2}\left(\frac{1}{k}\right)\right) \leq \sum_{m=1}^{\infty} \frac{\sigma^2 k^2}{m^2} < \infty$$

It then follows from Borel-Cantelli Lemma that, for every $k \geq 1$,

$$P\left(A_{m^2}\left(\frac{1}{k}\right) \text{ i.o. }\right) = 0$$

This implies that

$$\frac{1}{m^2} \sum_{i=1}^{m^2} X_i \stackrel{\text{w.p. } 1}{\longrightarrow} \mathsf{E}(X_1)$$

We have thus proved the result for a subsequence of the desired sequence $\frac{1}{n} \sum_{i=1}^{n} X_i$. The following argument shows that this is sufficient.
Note that the desired result can also be written as

$$\frac{1}{n} \sum_{i=1}^{n} (X_i - \mathsf{E}(X_1)) \stackrel{\text{w.p. } 1}{\longrightarrow} 0,$$

and the intermediate result we have can be written as

$$\frac{1}{m^2} \sum_{i=1}^{m^2} (X_i - \mathsf{E}(X_1)) \stackrel{\text{w.p. } 1}{\longrightarrow} 0 \tag{1.5}$$

Now observe that for any $n \geq 1$, there exists an $m \geq 1$ such that

$$m^2 \leq n < (m+1)^2 \text{ , i.e., } 0 \leq n - m^2 \leq 2m.$$

We need to show that for n such that $m^2 \leq n < (m+1)^2$, $\sum_{i=1}^{n} (X_i - EX)$ does not differ much from $\sum_{i=1}^{m^2} (X_i - EX)$. Let, for every $m \geq 1$,

$$\begin{aligned} M_m &= \max_{\{n: m^2 \leq n < (m+1)^2\}} \left| \sum_{i=1}^{n} (X_i - \mathsf{E}(X_1)) - \sum_{i=1}^{m^2} (X_i - \mathsf{E}(X_1)) \right| \\ &= \max_{\{n: m^2 \leq n < (m+1)^2\}} \left| \sum_{i=m^2+1}^{n} (X_i - \mathsf{E}(X_1)) \right| \end{aligned}$$

Therefore

$$\begin{aligned}\mathsf{E}\,(M_m)^2 &= \mathsf{E}\left(\left(\max_{\{n:m^2\le n<(m+1)^2\}}\left|\sum_{i=m^2+1}^{n}(X_i-\mathsf{E}(X_1))\right|\right)^2\right)\\ &= \mathsf{E}\left(\max_{\{n:m^2\le n<(m+1)^2\}}\left|\sum_{i=m^2+1}^{n}(X_i-\mathsf{E}(X_1))\right|^2\right)\end{aligned}$$

where the second equality follows because square is monotone increasing over nonnegative terms and all the terms in the $\max$ are nonnegative. For n such that $m^2 \le n < (m+1)^2$, define

$$Y_n = \left|\sum_{i=m^2+1}^{n}(X_i-\mathsf{E}(X_1))\right|^2$$

then

$$\mathsf{E}(Y_n) = (n-m^2)\sigma^2$$

since the random variables X_i are i.i.d.; in fact, notice that *we need the random variables only to be uncorrelated.* It follows that

$$\begin{aligned}\mathsf{E}\big(Y_{(m+1)^2-1}\big) &= 2m\sigma^2\\ &\ge \mathsf{E}(Y_n)\end{aligned}$$

where the last inequality holds for n such that $m^2 \le n < (m+1)^2$.

Now fix m, and for each ω, let $\hat{n}(\omega)$ achieve the maximum in $\max_{\{n:m^2\le n<(m+1)^2\}} Y_n(\omega)$; if there are several indices in $\{n : m^2 \le n < (m+1)^2\}$ that achieve the maximum then we break ties by, say, taking the largest such index. For $n \in \{n : m^2 \le n < (m+1)^2\}$, let us define the events

$$B_n = \{\omega : \hat{n}(\omega) = n\}$$

Hence

$$\begin{aligned}\mathsf{E}\,(M_m)^2 &= \mathsf{E}\left(\left(\sum_{n=m^2}^{(m+1)^2-1} Y_n I_{B_n}\right)\right)\\ &\le \sum_{n=m^2}^{(m+1)^2-1} EY_n\\ &\le (2m+1)2m\sigma^2\\ &\le 6m^2\sigma^2\end{aligned}$$

where we have used the fact that $m \geq 1$. Then applying Markov Inequality (Theorem 1.5), we get that, for all $\epsilon > 0$,

$$\begin{aligned} P(M_m > m^2\epsilon) &\leq \frac{6m^2\sigma^2}{(m^2\epsilon)^2} \\ &= \frac{6\sigma^2}{m^2\epsilon^2} \end{aligned}$$

Hence we see that

$$\sum_{m=1}^{\infty} P\left(\frac{M_m}{m^2} > \epsilon\right) < \infty$$

from which it follows, as before, that

$$\frac{M_m}{m^2} \overset{\text{w.p. } 1}{\longrightarrow} 0 \tag{1.6}$$

For notational ease let us now write $S_n = \sum_{i=1}^{n} (X_i - \mathsf{E}(X_1))$. In terms of this notation, our aim is to establish that $\frac{S_n}{n} \overset{\text{w.p. } 1}{\longrightarrow} 0$. Now, for each n, define $m(n)$ such that $m(n)^2 \leq n < (m(n)+1)^2$. By the definition of M_m, we then have

$$M_{m(n)} \geq \left|S_n - S_{m(n)^2}\right| \geq |S_n| - \left|S_{m(n)^2}\right|$$

Therefore

$$|S_n| \leq \left|S_{m(n)^2}\right| + M_{m(n)}$$

and, because $m(n)^2 \leq n$,

$$\frac{|S_n|}{n} \leq \frac{\left|S_{(m(n))^2}\right| + M_{m(n)}}{(m(n))^2} \to 0$$

with probability one, where in the last step we have combined Eqn. (1.5) and Eqn. (1.6). ■

The following SLLN requires only that the i.i.d. sequence have a finite mean, and is a powerful result that is used very frequently.

Theorem 1.10 (Kolmogorov SLLN (KSLLN))**.** *If $\{X_i, i \geq 1\}$ is a sequence of i.i.d. random variables with $\mathsf{E}(X_1) < \infty$, then $\frac{1}{n}\sum_{i=1}^{n} X_i \overset{a.s.}{\to} \mathsf{E}(X_1)$.* ■

1.6 Notation

This chapter has provided a quick overview of some basic concepts in probability theory and random processes, and cannot replace a pre-requisite course in probability theory. This chapter also served to introduce some of the notation that we will use in the remainder of this book. Notation related to concepts not covered in this chapter are provided in the following list.

$\amalg$ independent; For A and B independent events, or independent random variables, we write $A \amalg B$

1.7 Notes on the Bibliography

A first course on probability from a book such as the classic by Papoulis [15] would be sufficient prerequisite for reading these notes. A more mathematical but very accessible treatment of probability is provided by Bremaud [3]. Sophisticated topics such as the Borel-Cantelli Lemma, the Kolmogorov Consistency Theorem, and a proof of the Kolmogorov Strong Law of Large Numbers, along with a wealth of material on advanced probability theory are available in the classic two volumes by Loeve [13, 14], and in the more recent text by Athreya and Lahiri [2].

1.8 Problems

1.1. Given a sequence of sets $\{A_1, A_2, A_3, \cdots\} \subset \Omega$, denote

$$\begin{aligned} \underline{A} &= \cup_{m\geq 1} \cap_{n\geq m} A_n, \\ \overline{A} &= \cap_{m\geq 1} \cup_{n\geq m} A_n. \end{aligned}$$

a. Show that $\omega \in \underline{A}$ if and only if (iff), for some $m(\omega)$, $\omega \in A_n$ for all $n \geq m(\omega)$ (i.e., for each ω, the m will depend on ω).

b. Show that $\underline{A} \subset \overline{A}$.

1.2. $((X_1, Y_1),\ (X_2, Y_2),\ \cdots)$ is a sequence of random vectors such that $P(X_k \geq Y_k) = \alpha^k$ where $0 < \alpha < 1$. Show that $P(\{X_k \geq Y_k\} \text{ i.o.}) = 0$.

1.3. $(X_n,\ n \geq 1)$ is a sequence of random variables taking values in $\{0, 1\}$. Let $p_n = P(X_n = 1)$. Show that $\sum_{n=1}^{\infty}\ p_n < \infty$ implies that $P(\lim_{n\to\infty}\ X_n = 0) = 1$.

1.4. *Borel Strong Law of Large Numbers:* $\{X_n,\ n \geq 1\}$ is a sequence of i.i.d. random variables with $X_n \in \{0, 1\}$ such that $P(X_n = 1) = p$ for all $n \geq 1$. Define $S_n = \sum_{i=1}^{n}\ X_i$. Use the Markov inequality and the Borel-Cantelli Lemma to prove that $P(\lim_{n\to\infty}\ \frac{S_n}{n} = p) = 1$. Do not just apply the strong laws of large numbers stated in the chapter. (Hint: Show that, $\forall \epsilon > 0$, $\sum_{n=1}^{\infty}\ P\left(\left|\frac{S_n}{n} - p\right| \geq \epsilon\right) < \infty$, using Markov inequality with $k = 4$. Then use the Borel-Cantelli Lemma.)

1.5. $\{X_i, i = 1, \cdots, n\}$ are n identically distributed nonnegative random variables with a c.d.f. $F_X(\cdot)$, and are not necessarily independent. Let

$$Z = \max_{1\leq i\leq n} X_i$$

a. Show that $1 - F_Z(z) \leq \min\{1, n(1 - F_X(z))\}$. (Hint: use the "union bound".)

b. Hence obtain an upper bound for $\mathsf{E}(Z)$.

1.6. X is an integer valued random variable i.e., $X \in \{\cdots, -3, -2, -1, 0, 1, 2, 3\cdots\}$. Prove using the Markov inequality that $\mathsf{E}\left(X^2\right) < \infty\ \Rightarrow\ \mathsf{E}(|X|) < \infty$.

Chapter 2

Discrete Time Markov Chains

In Chapter 1 we studied i.i.d. random processes. In this chapter we will study processes that have the Markov property, the simplest dependence structure, and the one that is the most important in applications.

2.1 Conditional Independence

Given a probability space $(\Omega, \mathcal{F}, P)$ we need the notion of conditional independence between two events A and B, given the event C.

Definition 2.1. *The events A and B are defined to be conditionally independent, given the event C (denoted as $(A \amalg B)|C$), if $P(A \cap B|C) = P(A|C)\ P(B|C)$, or, equivalently, if $P(A|BC) = P(A|C)$.* ■

The equivalence of the two definitions can be seen from the following calculation: $P(A|BC) = \frac{P(AB|C)}{P(B|C)} = \frac{P(A|C)\cdot P(B|C)}{P(B|C)} = P(A|C)$. One way to view the above defined concept is to define a probability P_C as follows. For an event $A \in \mathcal{F}$, $P_C = P(A|C)$, where $P(\cdot)$ is the original probability.

Exercise 2.1. *Show that $P_C(\cdot)$ as defined is a probability measure on $(\Omega, \mathcal{F})$.*

■

It is then easy to see that conditional independence with respect to the event C is just ordinary independence in the new probability measure P_C. The statements in Definition 2.1 then become $(A \amalg B)|C$ if $P_C(A \cap B) = P_C(A)P_C(B)$, or $P_C(A|B) = P_C(A)$.

Exercise 2.2.

The random variable X_0 takes either of the values 0.1 or 0.9 with equal probabilities. When X_0 takes the value $p \in \{0.1, 0.9\}$ then X_1 and X_2 are i.i.d. 0-1 random variables that take the value 1 with probability p. Define the events $A = \{X_1 = 1\}$ and $B = \{X_2 = 1\}$, and $C = \{X_0 = 0.1\}$. Show that $(A \amalg B)|C$, but A is not independent of B, *unconditionally*. ■

We conclude from this exercise that conditional independence does not imply unconditional independence. Thinking in terms of coin tossing, X_0 is like choosing one of two coins, and X_1 and X_2 are like the outcomes of two tosses of the chosen coin, with a 1 outcome being viewed as the coin falling heads. Then the intuition is that, if we do not know which coin is being tossed, and the first toss falls heads, then we obtain some information about which coin was tossed, which changes the probabilities we assign to the next toss.

2.2 The Markov Property

Definition 2.2. *A stochastic process $\{X_n, n \geq 0\}$, taking values in the countable set $\mathcal{S}$, is called a **discrete time Markov chain (DTMC)** if it has the **Markov property**, i.e., if for all $i_0, i_1, \cdots, i_{n-1}, i_n = i, j \in S$ and $n \geq 0$*

$$P\{X_{n+1} = j | X_0 = i_0, X_1 = i_1, \ldots, X_n = i\} = P\{X_{n+1} = j | X_n = i\}$$

■

At time n, X_n is the "present", $X_0, X_1, \cdots, X_{n-1}$ form the "past", and $X_{n+1} = j$ is a question about the "future." The Markov property states that *given* the present, the future and the past are independent.

Remark: The term "chain" comes from the fact that $\{X_n, n \geq 0\}$ takes values in a *denumerable* set $\mathcal{S}$. The values taken by the process are also called *states* of the process; thus, the set $\mathcal{S}$ is also called the *state space*.

Example 2.1.

Consider repeated tosses of a coin with *given* probability $p, 0 < p < 1$, of the coin falling heads. Let, for $n \geq 0$, $X_n = 0$ if the outcome of the nth toss is tails, and $X_n = 1$ if the outcome is heads. We know that the sequence of random variables X_n are i.i.d., hence, the Markov property holds trivially. Now considering the same coin tossing experiment, define $Y_0 = 0$, and, for $n \geq 1$, $Y_n = \sum_{i=0}^{n-1} X_i$, i.e., Y_n is the number of heads until (not including) the nth toss. Clearly, $Y_n \in \mathbb{Z}^+$. Now observe that

$$\begin{aligned} P(Y_{n+1} = j | Y_0 = i_0, Y_1 = i_1, \cdots, Y_n = i) &= \begin{cases} 0 & \text{if } j \notin \{i, i+1\} \\ p & \text{if } j = i+1 \\ 1-p & \text{if } j = i \end{cases} \\ &= P(Y_{n+1} = j | Y_n = i) \end{aligned}$$

Thus $Y_n, n \geq 0$, is a DTMC on $\mathbb{Z}^+$. ■

Remark: It is important to note that if $X_n, n \geq 0$, is a DTMC then the Markov property states that $(X_{n+1} \amalg (X_0, \cdots, X_{n-1}))|X_n$, but this does not mean that $(X_{n+1} \amalg (X_0, \cdots, X_{n-1}))$, i.e., the future and the past are independent conditional on the present, but, in general, the future and the past need not be independent *unconditionally*. This can be illustrated from Example 2.1. The sequence $X_n, n \geq 0$, is i.i.d., and any two disjoint subsets of the random sequence are independent. However, notice that Y_{n+1} and Y_{n-1} are independent only if Y_n is given, but are *dependent* unconditionally. To see this just observe that if Y_{n-1} is known to be, say, 10 (10 heads have appeared before the $(n-1)$th toss), then Y_{n+1} is constrained to lie in the set $\{10, 11, 12\}$.

Exercise 2.3.

If $X_n, n \geq 0$, is a DTMC on $\mathcal{S}$, then, for any $\mathcal{A}_0, \mathcal{A}_1, \cdots, \mathcal{A}_{n-1} \subset \mathcal{S}$, and $i, j \in \mathcal{S}$, show that

$$P(X_{n+1} = j | X_0 \in \mathcal{A}_0, X_1 \in \mathcal{A}_1, \cdots, X_{n-1} \in \mathcal{A}_{n-1}, X_n = i) = P(X_{n+1} = j | X_n = i)$$

Hint: Observe that

$$\begin{aligned} &P(X_0 \in \mathcal{A}_0, X_1 \in \mathcal{A}_1, \cdots, X_{n-1} \in \mathcal{A}_{n-1}, X_n = i, X_{n+1} = j) \\ &= \sum_{i_0 \in \mathcal{A}_0, i_1 \in \mathcal{A}_1, \cdots, i_{n-1} \in \mathcal{A}_{n-1}} P(X_0 = i_0, X_1 = i_1, \cdots, X_{n-1} = i_{n-1}, X_n = i, X_{n+1} = j) \end{aligned}$$

and then use the Markov property, defined earlier. ■

Proposition 2.1. *If $X_n, n \geq 0$, is a DTMC on $\mathcal{S}$, then for $n < n_1 < n_2 < \ldots < n_m$, and $i_0, i_1, \cdots, i_{n-1}, i, j_1, \ldots j_m \in \mathcal{S}$*

$$\begin{aligned} &P\left(X_{n_k = j_k},\ 1 \leq k \leq m | X_0 = i_0, X_1 = i_1 \cdots, X_n = i\right) \\ &\qquad = \quad P\left(X_{n_k = j_k},\ 1 \leq k \leq m | X_n = i\right) \end{aligned}$$

Remark: The Markov property in Definition 2.2 appears to be limited to the immediate next step after n. This result says that any finite subset of random variables in the future is jointly independent of the past given the present.

Proof: The following calculation illustrates the idea of the proof. The detailed proof is

left as an exercise.

$$\begin{aligned} P\left(X_{n+2}=k|X_0,\cdots,X_n=i\right) &= \sum_{j\in\mathcal{S}} P\left(X_{n+1}=j,\ X_{n+2}=k|X_0,\cdots,X_n=i\right) \\ &= \sum_{j\in\mathcal{S}} P\left(X_{n+1}=j|X_0,\cdots,X_n=i\right)\cdot \\ &\quad P\left(X_{n+2}=k|X_0,\cdots,X_n=i,X_{n+1}=j\right) \\ &= \sum_{j\in\mathcal{S}} P\left(X_{n+1}=j|X_n=i\right)\cdot P\left(X_{n+2}=k|X_{n+1}=j\right) \\ &= \sum_{j\in\mathcal{S}} P\left(X_{n+1}=j,X_{n+2}=k|X_n=i\right) \\ &= P\left(X_{n+2}=k|X_n=i\right) \end{aligned}$$

where the first equality follows by summing over all possible states at step $n+1$, the second equality follows from the chain rule of probability, the third equality uses the result in Exercise 2.3, and the fourth equality is just the additivity property of probability. ■

We shall only consider the case of time homogeneous Markov chains, which is the most common case in applications. The following definition states what this case means.

Definition 2.3. *(i) $p_{ij}(n) := P\{X_{n+1} = j|X_n = i\}$ is called the **transition probability** of the Markov Chain, at step n.*

*(ii) If $p_{ij}(n) = p_{ij}$, for every $n \geq 0$, then the Markov Chain is said to be **time homogeneous**. For a time homogeneous Markov chain, we denote the **n-step transition probabilities** by $p_{ij}^{(n)} = P\{X_n = j|X_0 = i\}$.*

*(iii) The $|\mathcal{S}| \times |\mathcal{S}|$ matrix $\mathbf{P}$ whose (i,j)th element is p_{ij} is called the **transition probability matrix (t.p.m.)** of the Markov Chain.*

■

Definition 2.4. *A square matrix $\mathbf{P}$ with its rows and columns indexed by elements of the set $\mathcal{S}$, and with entries p_{ij} is a **stochastic matrix** if*

(i) for all i, j $p_{ij} \geq 0$, and

(ii) for all $i, \sum_{j\in\mathcal{S}} p_{ij} = 1$

■

The matrix $\mathbf{P}^{(n)}$ with elements $p_{ij}^{(n)}$ will be called the n step transition probability matrix. Since at each step the Markov chain takes values in $\mathcal{S}$, it is evident that for each, $n \geq 0$, $\mathbf{P}^{(n)}$ is a stochastic matrix. In particular, we write $\mathbf{P} = \mathbf{P}^{(1)}$. Further, we define $p_{jj}^{(0)} = 1$, and $p_{ij}^{(0)} = 0$, for $i \neq j$; or, equivalently, $\mathbf{P}^{(0)} = \mathbf{I}$ where $\mathbf{I}$ denotes the $|\mathcal{S}| \times |\mathcal{S}|$ identity matrix.

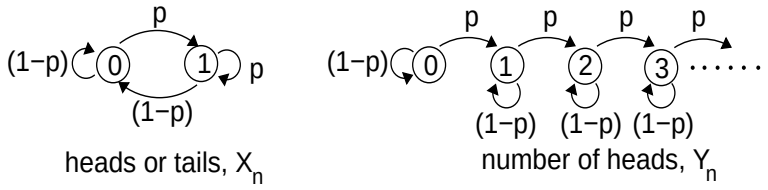


Figure 2.1: Transition probability diagrams of the two DTMCs in Example 2.1.

Example 2.2.

The following are the transition probability matrices of the two DTMCs in Example 2.1. The t.p.m. of $X_n \in \{0, 1\}$ is

$$\begin{bmatrix} 1-p & p \\ 1-p & p \end{bmatrix}$$

and the t.p.m. of $Y_n \in \mathbb{Z}^+$ is

$$\begin{bmatrix} 1-p & p & 0 & 0 & \cdots & \cdots \\ 0 & 1-p & p & 0 & \cdots & \cdots \\ 0 & 0 & \ddots & \ddots & 0 & \cdots \\ \vdots & \vdots & 0 & \ddots & \ddots & \ddots \end{bmatrix}$$

■

It is often convenient to represent the transition probability matrix of a DTMC by means of the **transition probability diagram**. For the two DTMCs in Example 2.1, the transition probability diagrams are depicted in Figure 2.1; their structure is self-evident.

Proposition 2.2. *If $X_n, n \geq 0$, is a DTMC on $\mathcal{S}$, then for $n < n_1 < n_2 < \ldots < n_m$, and $i_0, i_1, \cdots, i_{n-1}, i_n, j_1, \ldots j_m \in \mathcal{S}$*

$$P(X_{n_1} = j_1, \cdots, X_{n_m} = j_m | X_0 = i_0, \cdots, X_n = i_n) = p_{i_n j_1}^{(n_1-n)} p_{j_1 j_2}^{(n_2-n_1)} \cdots p_{j_{m-1} j_m}^{(n_m-n_{m-1})} \tag{2.1}$$

Proof: By Proposition 2.1, the left hand side of Equation 2.1 can be written (using the chain rule for probability) as

$$\begin{aligned} P(X_{n_1} = j_1, \cdots, X_{n_m} = j_m | X_n = i_n) &= P(X_{n_1} = j_1 | X_n = i_n) \cdot \\ &\quad P(X_{n_2} = j_2 | X_n = i_n,\ X_{n_1} = j_1) \cdot \\ &\quad \cdots P(X_{n_m} = j_m | X_n = i_n, \cdots, X_{n_{m-1}} = j_{m-1}) \\ &= P(X_{n_1} = j_1 | X_n = i_n) \cdot \\ &\quad P(X_{n_2} = j_2 | X_{n_1} = j_1) \cdot \\ &\quad \cdots P(X_{n_m} = j_m | X_{n_{m-1}} = j_{m-1}) \end{aligned}$$

where we have used Exercise 2.3 in the second equality. The last expression, using the notation for multistep transition probabilities, yields the right hand side of Equation 2.1. ■

The next question we ask relates to whether it is necessary to specify the n-step transition probability matrices, for each n, or whether it suffices to specify just the one step transition probability matrix.

Theorem 2.1 (Chapman Kolmogorov)**.** *For every n,*

$$\mathbf{P}^{(n)} = \mathbf{P}^n$$

or $p_{ij}^{(n)} = (\mathbf{P}^n)_{ij}$, i.e., the n step transition probability from i to j, is the $(ij)^{th}$ element of the nth power of $\mathbf{P}$.

Proof: For $i, j \in \mathcal{S}$, and $n, l \geq 0$, we can write

$$\begin{aligned}
p_{ij}^{(n+l)} &= P(X_{n+l} = j | X_0 = i) \\
&= \sum_{k \in \mathcal{S}} P(X_n = k, X_{n+l} = j | X_0 = i) \\
&= \sum_{k \in \mathcal{S}} P(X_n = k | X_0 = i) \cdot P(X_{n+l} = j |\ X_0 = i, X_n = k) \\
&= \sum_{k \in \mathcal{S}} P(X_n = k | X_0 = i) \cdot P(X_{n+l} = j | X_n = k) \\
&= \sum_{k \in \mathcal{S}} p_{ik}^{(n)} p_{kj}^{(l)} \\
&= (\mathbf{P}^{(n)} \mathbf{P}^{(l)})_{ij}
\end{aligned}$$

where the second equality is obtained by summing over all possible states at the nth step, and the result in Exercise 2.3 is used in the fourth equality. We have thus shown that

$$\mathbf{P}^{(n+l)} = \mathbf{P}^{(n)} \times \mathbf{P}^{(l)}$$

A simple inductive argument now yields the desired result as follows

$$\begin{aligned}
\mathbf{P}^{(n)} &= \mathbf{P} \times \mathbf{P}^{(n-1)} \\
&= \mathbf{P} \times \left(\mathbf{P} \times \mathbf{P}^{(n-2)} \right) \\
&= \mathbf{P}^2 \times \mathbf{P}^{(n-2)} \\
&= \cdots \\
&= \mathbf{P}^n
\end{aligned}$$

■

2.2.1 Finite Dimensional Distributions

Can we now write down the (unconditional) finite dimensional distributions of a DTMC $X_n, n \geq 0$, on $\mathcal{S}$? To this end, consider, for any $0 \leq n_1, n_2, \cdots, n_m$, and $i_1, i_2, \cdots, i_m \in \mathcal{S}$,

$$\begin{aligned} & P(X_{n_1} = i_1, X_{n_2} = i_2, \cdots, X_{n_m} = i_m) \\ &= \sum_{i_0 \in \mathcal{S}} P(X_0 = i_0, X_{n_1} = i_1, X_{n_2} = i_2, \cdots, X_{n_m} = i_m) \\ &= \sum_{i_0 \in \mathcal{S}} P(X_0 = i_0) P(X_{n_1} = i_1, X_{n_2} = i_2, \cdots, X_{n_m} = i_m | X_0 = i_0) \\ &= \sum_{i_0 \in \mathcal{S}} P(X_0 = i_0) p_{i_0 i_1}^{(n_1)} p_{i_1 i_2}^{(n_2 - n_1)} \cdots p_{i_{(m-1)} i_m}^{(n_m - n_{(m-1)})} \end{aligned}$$

Therefore, it follows (using Theorem 2.1) that the finite dimensional distributions of a DTMC are completely specified in terms of the t.p.m., $\mathbf{P}$, and the initial probability distribution $P(X_0 = i), i \in \mathcal{S}$. *For a given t.p.m., $\mathbf{P}$, we obtain different stochastic processes for different initial probability distributions.*

2.3 The Strong Markov Property

Given a DTMC $X_n, n \in \{0, 1, 2, \cdots\}$, we have so far asserted only that given X_n at a fixed time n the future of the process is independent of the past. Now suppose T is a random time, i.e., T itself is a random variable taking values in the index set of the DTMC, i.e., $T : \Omega \to \{0, 1, 2, \cdots\}$.

Example 2.3.

For $j \in \mathcal{S}$, define T_j to be the random time when the process first visits state j, i.e., $T_j(\omega) = k$ if $X_0(\omega) \neq j, X_1(\omega) \neq j, \cdots, X_{k-1}(\omega) \neq j, X_k(\omega) = j$, and if for ω there is no k such that $X_k(\omega) = j$ then $T_j(\omega) = \infty$. It can be checked that T_j satisfies the second requirement of a random variable (i.e., being a measurable function), since $\{\omega : T_j(\omega) = k\} = \{\omega : X_0(\omega) \neq j, X_1(\omega) \neq j, \cdots, X_{k-1}(\omega) \neq j, X_k(\omega) = j\}$ and $(X_1, X_2, \cdots, X_k)$ is a random vector. ■

Suppose we are given X_T. Note here that by X_T we mean that $X_T(\omega) = X_{T(\omega)}(\omega)$, i.e., the state of the random process at the random time T. Can we say that for $k \geq 1$, $P\left(X_{T+k} = j | X_0 = i_0, \cdots, X_T = i\right) = p_{ij}^{(k)}$; i.e., does the Markov property hold w.r.t. the random time T? The following example shows that this may not be true for any arbitrary random time.

Example 2.4.

Define $U^{(j)}$ to be the time of the second visit to state j, and let $T = U^{(j)} - 1$, i.e., $T(\geq 0)$ is one step before the second visit to state j. Now, for some $i, k \in \mathcal{S}$,

$$P\left(X_{T+1} = k | X_0 = i_0, \cdots, X_T = i\right) \quad = \quad \begin{cases} 1 & \text{if } k = j \\ 0 & \text{if } k \neq j \end{cases}$$

and the right hand side is in general not equal to p_{ik}. We can see that the random time T is *anticipatory* since if we know that we are at the time T we have some information about the future, i.e., that at the next step the process must visit j. ■

Why might we be interested in random times, and the existence of the Markov property with respect to random times? Suppose that the DTMC X_n is the model for a queue. One of the measures we would be interested in is the distribution of the sojourn time of a random arrival. The arrival instant is a random time T, and in order to study the sojourn time of the arriving customer we will need to study the future of the queueing process after the random time. Thus, in analysing discrete event stochastic processes we will often need to condition on random times, and then study the evolution of the process after such times.

The following specialisation of random times eliminates the difficulty illustrated in Example 2.4.

Definition 2.5. *A random time T is said to be a **stopping time** for the random process $\{X_n \in \mathcal{S}, n \geq 0\}$ if for all $n \geq 0$ there exists a function $f_n : \mathcal{S}^{(n+1)} \to \{0, 1\}$ such that, for all n, $I_{\{T \leq n\}}(\omega) = f_n(X_0(\omega), \cdots, X_n(\omega))$, i.e., to answer the question that $T(\omega) \leq n$ we need only look at $X_0(\omega), \cdots, X_n(\omega)$.* ■

Example 2.5.

(i) $T^{(j)}$= the time of the first visit to j. This is a stopping time since we have $f_n(i_0, i_1, \cdots, i_n) = 1$ if $j \in \{i_0, i_1, \cdots, i_n\}$, and 0 otherwise.

(ii) $Z^{(j)}$= the time of the last visit to j. This is not a stopping time since we cannot determine whether $Z^{(j)}(\omega) \leq n$ by looking only at $X_0(\omega), \cdots, X_n(\omega)$.

(iii) The random time T defined in Example 2.4 is not a stopping time, since $\{\omega : T(\omega) \leq n\} = \{\omega : U^{(j)}(\omega) \leq n + 1\}$, hence $I_{\{\omega: T(\omega) \leq n\}}$ cannot be determined by $X_0, X_1, \cdots, X_n$ alone.

(iv) $T = m$ is a stopping time since we can take $f_n(i_0, \cdots, i_n) = \begin{cases} 1 & \text{if } m \leq n \\ 0 & \text{otherwise} \end{cases}$. ■

Theorem 2.2. *$\{X_n, n \geq 0\}$ is a DTMC and T is a stopping time such that $P(T < \infty) = 1$. Then the DTMC has the **strong Markov property** with respect to T, i.e., $P\left(X_{T+s} = j | X_0 = i_0, \cdots, X_T = i\right) = p_{ij}^{(s)}$.*

Proof: Using the hypothesis $P(T < \infty) = 1$ in the first equality below, we can write

$$\begin{aligned}
&P\left(X_0 = i_0, \cdots, X_T = i, X_{T+s} = j\right) \\
&= \sum_{t=0}^{\infty} P\left(X_0 = i_0, \cdots, X_t = i, X_{t+s} = j,\ T = t\right) \\
&= \sum_{t=0}^{\infty} P\left(X_0 = i_0, \cdots, X_t = i\right)\ P\left(X_{t+s} = j | X_0 = i_0 \cdots X_t = i\right) \\
&\qquad P\left(T = t | X_0 = i_0, \cdots, X_t = i, X_{t+s} = j\right) \\
&= \sum_{t=0}^{\infty} P\left(X_0 = i_0, \cdots, X_t = i\right)\ p_{ij}^{(s)}\ \underbrace{P(T = t | X_0 = i_0, \cdots, X_t = i)}_{\text{since } T \text{ is a stopping time}} \\
&= p_{ij}^{(s)} \sum_{t=0}^{\infty} P\left(X_0 = i_0, \cdots, X_t = i, T = t\right) \\
&= p_{ij}^{(s)}\ P\left(X_0 = i_0, \cdots, X_T = i\right)
\end{aligned}$$

and the result follows, by dividing on both sides by the term $P\left(X_0 = i_0, \cdots, X_T = i\right)$ (provided, of course, that this term is positive). Notice that in the third equality we have used the Markov property (with respect to the fixed time t). In this equality we also used the fact that T is a stopping time, and hence, given $X_0 = i_0, \cdots, X_t = i$, the event $T = t$ is independent of X_{t+s}. ■

The following statement is also evident from the above derivation. If T is a proper stopping time, then, for $s_1 < s_2 < \cdots < s_n$, and $i_0, \cdots, i, j_1, j_2, \cdots, j_n \in \mathcal{S}$,

$$\begin{aligned}
P(X_{T+s_1} = j_1, X_{T+s_2} = j_2, \cdots, X_{T+s_n} = j_n | X_0 = i_0, \cdots X_T = i) = \\
P(X_{s_1} = j_1, \cdots, X_{s_n} = j_n | X_0 = i)
\end{aligned}$$

As an example, consider the stopping time $T^{(j)}$, the time of the first visit to j. Then the strong Markov property implies that when the Markov chain hits the state j it statistically restarts, with initial state j, and its evolution thereafter is independent of its past before T_j.

2.4 Hitting Times and Recurrence

In applications, we are usually concerned with questions about the steady state or long run behaviour of the DTMC. We will see that the answers to these questions are intimately related to the recurrence properties of the states of the DTMC.

2.4.1 First Passage Time Distribution

Given that $X_n, n \geq 0$, is a DTMC on $\mathcal{S}$, for $i, j \in \mathcal{S}$, and $n \geq 1$, define $f_{ij}^{(n)} = P\left(X_1 \neq j, X_2 \neq j, \cdots, X_{n-1} \neq j, X_n = j | X_0 = i\right)$, i.e., $f_{ij}^{(n)}$ is the probability that the DTMC *hits* j for the first time at the nth step, given that it starts off in state i. If $j = i$ then $f_{jj}^{(n)}$ is the probability of first *return* to state j at the nth step. Further, define

$$\begin{aligned} f_{ij} &= P(\text{there exists } n \geq 1, \text{such that } X_n = j | X_0 = i) \\ &= P(\cup_{n=1}^{\infty}(X_k \neq j, 1 \leq k < n, X_n = j) | X_0 = i) \\ &= \sum_{n=1}^{\infty} f_{ij}^{(n)} \end{aligned}$$

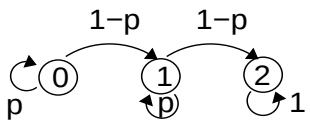
where the last step follows by countable additivity of probability. In other words, f_{ij} is the probability of *ever* hitting state j if the process starts off in state i. In general, $f_{ij} \leq 1$. For $i, j \in \mathcal{S}$, $f_{ij}^{(n)}$ is called the *first passage time distribution* for hitting j starting from state i; i.e., $f_{ij}^{(n)}$ is a probability mass function on the positive integers. This distribution, in general, can be *defective* (i.e., the probability mass function can sum to less than 1). If $f_{ij} < 1$ then $1 - f_{ij} > 0$ is the probability that the DTMC never visits j starting from state i.

When $f_{jj} = 1$, the mean time to return, or the *mean recurrence time* is defined to be

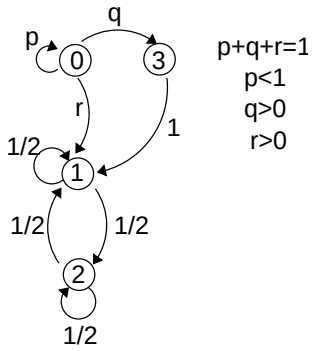
$$\nu_{jj} = \sum_{n=1}^{\infty} n f_{jj}^{(n)}$$

It should be clear that the first passage time distribution depends only on the transition probability matrix of the DTMC, and is basically a property of this stochastic matrix. For completeness, we define, $f_{jj}^{(0)} = 1$, and, for $i \neq j$, $f_{ij}^{(0)} = 0$.

Example 2.6.



a. For the transition structure shown in this diagram, observe that $f_{00}^{(1)} = p$, and, for $n \geq 2$, $f_{00}^{(n)} = 0$; this is because, starting from state 0, the DTMC can return to state 0 in 1 step, or not at all. Thus, $f_{00} = p$, and if $p < 1$, then there is a positive probability of never returning to 0. Also, observe that, for $n \geq 1$, $f_{01}^{(n)} = p^{n-1}(1-p)$, so that $f_{01} = 1$. Similarly, we can see that $f_{02} = 1$, and $f_{22} = 1$.



b. In this example we see that $f_{01}^{(1)} = r$, and, for $k \geq 2$, $f_{01}^{(k)} = p^{k-1}r + p^{k-2}q$. It follows that $f_{01} = r + \sum_{k=2}^{\infty}(p^{k-1}r + p^{k-2}q) = r + p\frac{r}{1-p} + \frac{q}{1-p} = 1$, which is evident from the diagram, since, starting from state 0, eventually the DTMC must visit state 1. ■

Definition 2.6. *For a DTMC $X_n, n \geq 0$, on $\mathcal{S}$, a state $j \in \mathcal{S}$ is called **transient**, if $f_{jj} < 1$, **recurrent**, if $f_{jj} = 1$, **positive recurrent**, if $f_{jj} = 1$ and the mean recurrence time is finite, i.e., $\nu_{jj} < \infty$, and **null recurrent**, if $f_{jj} = 1$ and the mean recurrence time is infinite, i.e., $\nu_{jj} = \infty$.* ■

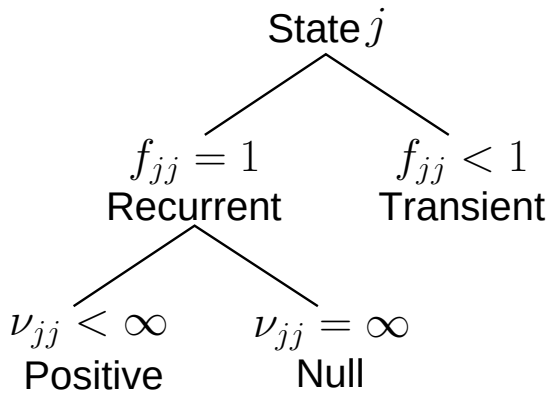


Figure 2.2: A depiction of Definition 2.6.

2.4.2 Number of Returns to a State

Let $M_j = \sum_{n=1}^{\infty} I_{\{X_n=j\}}$, i.e., M_j is the number of visits to state j for $n \geq 1$. Then

$$\begin{aligned} \mathsf{E}(M_j|X_0 = j) &= \mathsf{E}\left(\left(\sum_{n=1}^{\infty} I_{\{X_n=j\}}\right)|X_0 = j\right) \\ &= \sum_{n=1}^{\infty} p_{jj}^{(n)} \end{aligned} \tag{2.2}$$

Notice that in the second equality we have used the monotone convergence theorem. Thus if $\sum_{n=1}^{\infty} p_{jj}^{(n)} < \infty$ then the mean number of returns to j is finite. The following exercise is then a simple observation.

Exercise 2.4.

Show using the Borel Cantelli Lemma (Lemma 1.1) that if $\sum_{n=1}^{\infty} p_{jj}^{(n)} < \infty$ then j occurs infinitely often with probability 0. ■

We can also obtain the distribution of M_j conditioned on $X_0 = j$, as follows. Let $T_k, k \geq 1$, denote the instants of successive visits to the state j. Then we can write

$$\begin{aligned} &P(M_j = m|X_0 = j) \\ &= P(T_1 < \infty,\ T_2 - T_1 < \infty, \cdots, T_m - T_{m-1} < \infty, T_{m+1} - T_m = \infty|X_0 = j) \\ &= P(T_1 < \infty|X_0 = j) \cdot \\ &\quad P(T_2 - T_1 < \infty, \cdots, T_m - T_{m-1} < \infty,\ T_{m+1} - T_m = \infty|X_0 = j,\ T_1 < \infty) \\ &= f_{jj} \cdot \\ &\quad P(T_2 - T_1 < \infty, \cdots, T_m - T_{m-1} < \infty,\ T_{m+1} - T_m = \infty|X_0 = j,\ T_1 < \infty) \end{aligned}$$

We see that T_1 is a stopping time and would like to claim that

$$\begin{aligned} P(T_2 - T_1 < \infty, \cdots, T_m - T_{m-1} < \infty,\ T_{m+1} - T_m = \infty|X_0 = j,\ T_1 < \infty) = \\ P(M_j = m - 1|X_0 = j) \end{aligned}$$

i.e., conditioning on $(X_0 = j,\ T_1 < \infty)$ and asking a question about the future after T_1 is equivalent to starting at time 0 in the state j, and asking the same question about the future. We can see that this is just the strong Markov property; explicitly, it can be seen to follow from the result in the following exercise.

Exercise 2.5.

Show that

$$P(X_{T_1+s} = k|X_0 = j,\ T_1 < \infty) = p_{jk}^{(s)}$$

Hint: The argument is essentially identical to the proof of Theorem 2.2, after making the following observation

$$\begin{aligned} & P\left(X_{T_1+s}=k,\ X_0=j,\ T_1<\infty\right) \\ &= \sum_{t=1}^{\infty} P\left(X_{t+s}=k,\ X_0=j,\ X_1\neq j,\cdots,X_{t-1}\neq j,\ X_t=j,\ T_1=t\right) \end{aligned}$$

■

Continuing the argument recursively we see that

$$\begin{aligned} P\left(M_j=m|X_0=j\right) &= f_{jj}^m P(T_1=\infty|X_0=j) \\ &= f_{jj}^m\left(1-f_{jj}\right) \end{aligned}$$

Thus we have the following result

Theorem 2.3. *For a DTMC X_n on $\mathcal{S}$, if the state $j\in\mathcal{S}$ is transient ($f_{jj}<1$) then, for $m\in\{0,1,2,\cdots\}$,*

$$P\left(M_j=m|X_0=j\right)=f_{jj}^m\left(1-f_{jj}\right)$$

and if j is recurrent ($f_{jj}=1$) then, for $m\in\{0,1,2,\cdots\}$,

$$P\left(M_j=m|X_0=j\right)=0$$

■

Using this result we can conclude that if j is transient then

$$\begin{aligned} \mathsf{E}(M_j|X_0=j) &= \sum_{m=1}^{\infty} m f_{jj}^m\left(1-f_{jj}\right) \\ &= \frac{f_{jj}}{(1-f_{jj})} \\ &< \infty \end{aligned}$$

and, if j is recurrent then, $\mathsf{E}(M_j|X_0=j)=\infty$.

Combining this observation with Equation 2.2, we can further conclude that $\sum_{n=1}^{\infty} p_{jj}^{(n)}<\infty$ for j transient, and hence that, for j transient, $\lim_{n\to\infty} p_{jj}^{(n)}=0$. This says that eventually there is no probability on a transient state j. We have established the following result

Theorem 2.4. *(i) A state j is transient if and only if the expected number of returns is finite. Thus, for a recurrent state j, the expected number of returns to j is infinite, i.e., $\sum_{n=1}^{\infty} p_{jj}^{(n)}=\infty$.*

(ii) For a transient state j, $\lim_{n\to\infty} p_{jj}^{(n)}=0$.

■

2.5 Communicating Classes and Class Properties

Definition 2.7. *Given a DTMC $X_n, n \geq 0$, on the countable state space $\mathcal{S}$, we say that*

(i) j *is* reachable *from* i *if there exists* $n \geq 0$ *such that* $p_{ij}^{(n)} > 0$ *and we denote this by* $i \to j$*, and that*

(ii) i *and* j communicate *if* $i \to j$ *and* $j \to i$*, and denote this by* $i \leftrightarrow j$

■

Proposition 2.3. *The relation $\leftrightarrow$ on $\mathcal{S}$ is an equivalence*

Proof: The following conditions for $\leftrightarrow$ to be an equivalence can easily be checked.

Reflexive: $i \leftrightarrow i$, since $p_{ii}^{(0)} = 1$ (by definition).
Symmetric: $i \leftrightarrow j \Rightarrow j \leftrightarrow i$.
Transitive: $i \leftrightarrow j$ and $j \leftrightarrow k \Rightarrow i \leftrightarrow k$.

■

It follows that $\leftrightarrow$ partitions $\mathcal{S}$ into equivalence classes.

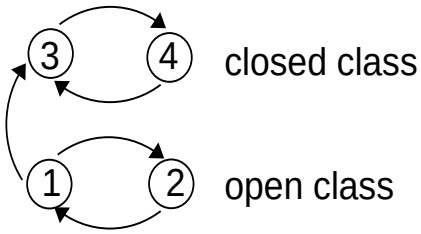
Definition 2.8. *Given a DTMC $X_n, n \geq 0$, on $\mathcal{S}$, with t.p.m. $\mathbf{P}$, we define the following.*

- *The classes into which $\leftrightarrow$ partitions $\mathcal{S}$ are called* communicating classes*.*
- *A communicating class $\mathcal{C}$ is said to be* closed *if, for all $i \in \mathcal{C}$ and $j \notin \mathcal{C}$, $p_{ij} = 0$.*
- $\mathbf{P}$ *is said to be* irreducible *if the entire $\mathcal{S}$ is one communicating class. We also say that the Markov chain is irreducible.*

■

Evidently, an irreducible Markov chain is closed. It should also be clear that the communicating classes (and whether each is closed or not) depend only on the positions of the positive elements of the t.p.m. $\mathbf{P}$, not on their precise values. The transition probability diagram without the transition probability values can be called the *transition graph*.

Example 2.7.



In this transition graph all the transitions shown have positive transition probabilities. Clearly, there are two communicating classes, $\{1, 2\}$, and $\{3, 4\}$. The class $\{1, 2\}$ is open, whereas the class $\{3, 4\}$ is closed. ■

We will use the following theorem repeatedly in the remainder of this chapter, but it is a consequence of the Elementary Renewal Theorem which will be proved in the next chapter (see Section 3.2.1).

Theorem 2.5. *If state j is recurrent (i.e., $f_{jj} = 1$) and $f_{ij} = 1$ then*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} p_{ij}^{(k)} = \gamma_j \begin{cases} = 0 & \text{if } j \text{ is null recurrent} \\ > 0 & \text{if } j \text{ is positive recurrent} \end{cases}$$

■

Remark: We have $\mathsf{E}\big(\sum_{k=1}^{n} I_{\{X_k=j\}} | X_0 = j\big) = \sum_{k=1}^{n} p_{jj}^{(k)}$. When j is transient $\sum_{k=1}^{n} p_{jj}^{(k)} \to_{n\to\infty} \mathsf{E}(M_j) < \infty$. On the other hand when j is recurrent $\sum_{k=1}^{n} p_{jj}^{(k)} \to_{n\to\infty} \infty$. This theorem provides the more refined result that, if j is null recurrent then $\sum_{k=1}^{n} p_{jj}^{(k)}$ is $o(n)$ (i.e., $\sum_{k=1}^{n} p_{jj}^{(k)}$ grows to ∞ slower than n), and when j is positive recurrent then $\sum_{k=1}^{n} p_{jj}^{(k)}$ is $\Theta(n)$ (i.e., $\sum_{k=1}^{n} p_{jj}^{(k)}$ grows to ∞ proportional to n).

■

Theorem 2.6 (Class Property Theorem)**.** *The states in a communicating class are either*

(i) all transient,

(ii) all null recurrent, or

(iii) all positive recurrent.

Proof: Suppose j and k are two states in a communicating class, i.e., $j \to k$ and $k \to j$. Hence there exist r and s such that $p_{jk}^{(r)} > 0$ and $p_{kj}^{(s)} > 0$. Now let us observe that, for all $n \geq 0$,

$$p_{jj}^{(r+n+s)} \geq p_{jk}^{(r)} p_{kk}^{(n)} p_{kj}^{(s)}$$

This can be seen as follows:

$$\begin{aligned} p_{jj}^{(r+n+s)} &= P\{X_{r+n+s} = j | X_0 = j\} \\ &\geq P\{X_r = k,\ X_{r+n} = k,\ X_{r+n+s} = j | X_0 = j\} \\ &= p_{jk}^{(r)}\ p_{kk}^{(n)}\ p_{kj}^{(s)} \end{aligned}$$

Now suppose j is transient. Then it follows from Theorem 2.4 that $\sum_{n=1}^{\infty}\ p_{jj}^{(n)} < \infty$. Now we observe that

$$\begin{aligned} \sum_{n=1}^{\infty} p_{jj}^{(n)} &\geq \sum_{n=1}^{\infty} p_{jj}^{(r+n+s)} \\ &\geq \sum_{n=1}^{\infty} p_{jk}^{(r)}\ p_{kk}^{(n)}\ p_{kj}^{(s)} \end{aligned}$$

Since $p_{jk}^{(r)} > 0$ and $p_{kj}^{(s)} > 0$, it follows that $\sum_{n=1}^{\infty}\ p_{kk}^{(n)} < \infty$, and, hence, by Theorem 2.4, k is transient. Thus if any state in a class is transient, all states in that class are transient.

Suppose j is recurrent null then, by Theorem 2.4, $\sum_{m=1}^{\infty}\ p_{jj}^{(m)} = \infty$. Hence k cannot be transient, else we will have a contradiction by the previous part of the proof. Also by Theorem 2.5

$$\lim_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} p_{jj}^{(m)} = 0$$

Now, for large enough n,

$$\begin{aligned} \frac{1}{n} \sum_{m=0}^{n} p_{jj}^{(m)} \geq \frac{1}{n} \sum_{m=r+s}^{n} p_{jj}^{(m)} &= \frac{1}{n} \sum_{\ell=0}^{n-(r+s)} p_{jj}^{(r+\ell+s)} \\ &\geq p_{jk}^{(r)} \left(\frac{1}{n} \sum_{\ell=0}^{n-(r+s)} p_{kk}^{(\ell)} \right) p_{kj}^{(s)} \\ &= p_{jk}^{(r)} \left(\frac{n-(r+s)+1}{n} \right) \left(\frac{1}{n-(r+s)+1} \sum_{\ell=0}^{n-(r+s)} p_{kk}^{(\ell)} \right) p_{kj}^{(s)} \\ &\geq 0 \end{aligned}$$

Now since the leftmost expression in the above series of inequalities goes to 0, and since $p_{jk}^{(r)} > 0$ and $p_{kj}^{(s)} > 0$ it follows that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} p_{kk}^{(m)} = 0$$

Hence, by Theorem 2.5, k is not positive. It was already asserted above that k cannot be transient (since that would contradict the recurrence of j). Hence k is also recurrent null.

Finally, it is clear from the foregoing that if j is positive recurrent then k cannot be transient or null recurrent, and hence must be positive recurrent. ■

Lemma 2.1. *If j is recurrent and $j \to i$, then $i \to j$ and $f_{ij} = 1$.*

Proof: We observe that, if $j \to i$ then there exists $n \geq 1$ such that $a_{ji}^{(n)} := P(X_1 \neq j, X_2 \neq j, \cdots, X_{n-1} \neq j, X_n = i | X_0 = j)$ is positive ; $a_{ji}^{(n)}$ is the probability of visiting i at the nth step, starting in state j, and without hitting j in between. Suppose $f_{ij} < 1$, then

$$1 - f_{jj} \geq a_{ji}^{(n)} \ (1 - f_{ij}) > 0$$

hence j is not recurrent, a contradiction. Hence $f_{ij} = 1$ and obviously $i \to j$. ■

Corollary 2.1. *If i, j belong to a recurrent class then $f_{ij} = 1$ and $f_{ji} = 1$.* ■

Theorem 2.7. *Open communicating classes are transient.*

Proof: If $\mathcal{C}$ is an open communicating class, there exists $j \notin \mathcal{C}$ and $i \in \mathcal{C}$ such that $p_{ij} > 0$. Since $\mathcal{C}$ is a communicating class, $j \notin \mathcal{C}$, and $i \in \mathcal{C}$ it follows that $f_{ji} = 0$. Now notice that (using the Markov Property in the third equality)

$$\begin{aligned}
f_{ii} &= P(\text{there exists } k \geq 1 \text{ such that } X_k = i | X_0 = i) \\
&= p_{ii} + \sum_{\ell \neq i} P(X_1 = \ell, \text{there exists } k \geq 2 \text{ such that } X_k = i | X_0 = i) \\
&= p_{ii} + \underbrace{p_{ij} \ f_{ji}}_{=0} + \sum_{\ell : \ell \neq j, \ell \neq i} p_{i\ell} \ f_{\ell i} \\
&\leq \sum_{\ell \neq j} p_{i\ell} < 1
\end{aligned}$$

where in the last step we have used the fact that $f_{\ell i} \leq 1$ for all $\ell \in \mathcal{S}$, and that $p_{ij} > 0$. Hence i is transient, and all states in $\mathcal{C}$ are transient by Theorem 2.6. ■

Theorem 2.8. *Finite closed communicating classes are positive recurrent.*

Proof: Since $\mathcal{C}$ is closed, for every $i \in \mathcal{C}$ and for all $n \geq 0$, $P(X_n \in \mathcal{C} | X_0 = i) = 1$, i.e., $\sum_{j \in \mathcal{C}} p_{ij}^{(n)} = 1$. If $\mathcal{C}$ were transient or null then, by Theorem 2.5, for all $j \in \mathcal{C}$,

$$\lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} p_{ij}^{(k)} = 0$$

Therefore (crucially using the fact that $\mathcal{C}$ is a finite set in the second equality of the following equalities; see the Remark following this proof) we see that, for $i \in \mathcal{C}$, if $\mathcal{C}$ is transient or null

$$\begin{aligned} 0 = \sum_{j\in\mathcal{C}} \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} p_{ij}^{(k)} &= \lim_{n\to\infty} \frac{1}{n} \sum_{j\in\mathcal{C}} \sum_{k=1}^{n} p_{ij}^{(k)} \\ &= \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} 1 = 1, \end{aligned}$$

which is absurd. Hence $\mathcal{C}$ must be a positive recurrent class. ■

Remark: The importance of taking care when exchanging infinite sums and limits, or exchanging limits, is underlined by the proof of the above theorem. The exchange of the infinite sum and the limit in $\sum_{j\in\mathcal{C}} \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} p_{ij}^{(k)} = \lim_{n\to\infty} \frac{1}{n} \sum_{j\in\mathcal{C}} \sum_{k=1}^{n} p_{ij}^{(k)}$ is justified when $\mathcal{C}$ is a finite set. If $\mathcal{C}$ is not finite and we still exchange the sum and the limit (assuming this exchange to be valid, in general) we would have "proved" a result that is wrong. In general, infinite closed communicating classes can be transient, or null recurrent or positive recurrent; see Section 2.8 for an example.

2.6 Positive Recurrence and the Invariant Probability Vector

In this section we develop an important condition for the positive recurrence of a communicating class. By Theorem 2.7 we know that we need to look only at closed communicating classes, in which case we can just think of a Markov chain restricted to each of its closed communicating classes. In other words it is sufficient to study irreducible DTMCs.

Theorem 2.9. *An irreducible DTMC is positive recurrent iff there exists a probability mass function (or measure) $\boldsymbol{\pi}$ on $\mathcal{S}$ such that $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}$, with $\pi_i > 0$ for all $i \in \mathcal{S}$. Such a $\boldsymbol{\pi}$ is unique.*

Remarks 2.1.

Before proving the theorem, we will discuss the system of equations $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}$ and some concepts related to its solutions.

a. The matrix expression $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}$ yields one equation for each state in $\mathcal{S}$. The equation for state j is

$$\pi_j = \sum_{i\in\mathcal{S}} \pi_i p_{ij}$$

with the right hand expression being the vector dot product of $\boldsymbol{\pi}$ and the jth column of $\mathbf{P}$. The theorem asks for a positive solution of these equations. Since $\mathbf{P}$ is a stochastic matrix, one of its eigenvalues is 1. It follows that the theorem seeks a left eigenvector of $\mathbf{P}$ for the eigenvalue 1.

b. If $\boldsymbol{\pi}$ is a probability measure on $\mathcal{S}$ such that $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}$. Then by recursing on this equation we see that, for all $n \geq 1$, $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}^n$.

c. If $\boldsymbol{\pi}$ is a probability measure on $\mathcal{S}$ such that $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}$, and we take $P(X_0 = i) = \pi_i$, we see (from $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}^n$) that, for all n, $P(X_n = i) = \pi_i$. For this reason a probability vector that solves $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}$ is also called an *invariant probability vector*. The reader is encouraged to verify, as an exercise, that with this initial distribution the process $X_n, n \geq 1$, is a (strictly) stationary process (just check that the finite dimensional distributions are shift invariant).

d. Suppose that $\mathbf{P}$ is irreducible and $\boldsymbol{\pi}$ is such that $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}$, with $\pi_i > 0$ for *some* $i \in \mathcal{S}$. Now consider any $j \in \mathcal{S}$. Since $\mathbf{P}$ is irreducible, there exists $m \geq 1$, such that $p_{ij}^{(m)} > 0$. It then follows, from $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}^m$, that $\pi_j = \sum_{l \in \mathcal{S}} \pi_l p_{lj}^{(m)}$, which can be seen to be positive. Hence if $\mathbf{P}$ is irreducible and $\boldsymbol{\pi}$ is such that $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}$, and if any element of $\boldsymbol{\pi}$ is positive then the entire vector is positive.

e. Evidently, any positive multiple of a solution of $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}$ is also a solution. This theorem implies that, for an irreducible positive recurrent DTMC, these solutions are *summable* (i.e., $\sum_{i \in \mathcal{S}} \pi_i < \infty$) and hence can be normalised to yield a (unique) probability mass function on $\mathcal{S}$. On the other hand if the DTMC is irreducible recurrent but not positive, then positive solutions of $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}$ exist but will not be summable; see Theorem 3.19.

Proof: *Only if part:* Assume that the Markov chain is positive recurrent. Letting, for $s \geq 0$, $a_i^{(s)} = P\left(X_s = i\right)$, define

$$\mathbf{b}^{(n)} = \frac{1}{n} \sum_{s=1}^{n} \mathbf{a}^{(s)} \text{ and } \mathbf{c}^{(n)} = \frac{1}{n} \sum_{s=1}^{n} \mathbf{a}^{(s-1)}$$

Being time averages of probability mass functions on $\mathcal{S}$, for each $n \geq 1$, $\mathbf{b}^{(n)}$ and $\mathbf{c}^{(n)}$ are also probability mass functions on $\mathcal{S}$. Now $\mathbf{a}^{(s)} = \mathbf{a}\mathbf{P}^s$, where $a_i = P(X_0 = i)$. Clearly, $\mathbf{b}^{(n)} = \mathbf{a}\frac{1}{n} \sum_{s=1}^{n} \mathbf{P}^s$, or

$$b_j^{(n)} = \sum_{i \in \mathcal{S}} a_i \frac{1}{n} \sum_{s=1}^{n} p_{ij}^{(s)}$$

Using Theorem 2.5, and the assumed positive recurrence, we know that $\frac{1}{n} \sum_{s=1}^{n} p_{ij}^{(s)} \to \gamma_j > 0$. We now wish to take $\lim_{n \to \infty}$ in the previous equation, i.e.,

$$\lim_{n \to \infty} b_j^{(n)} = \lim_{n \to \infty} \sum_{i \in \mathcal{S}} a_i \frac{1}{n} \sum_{s=1}^{n} p_{ij}^{(s)}$$

We can exchange the $\lim_{n\to\infty}$ and $\sum_{i\in\mathcal{S}}$ in the right hand side by using the bounded convergence theorem; see Theorem 1.7 (we use the fact that $\frac{1}{n}\sum_{s=1}^{n} p_{ij}^{(s)} \leq 1$). This yields

$$\begin{aligned}\lim_{n\to\infty} b_j^{(n)} &= \sum_{i\in\mathcal{S}} a_i \lim_{n\to\infty} \frac{1}{n}\sum_{s=1}^{n} p_{ij}^{(s)} \\ &= \sum_{i\in\mathcal{S}} a_i \gamma_j \\ &= \gamma_j\end{aligned}$$

In other words,

$$\lim_{n\to\infty} \mathbf{b}^{(n)} = \boldsymbol{\gamma}$$

where $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \cdots,)$. Now let $\boldsymbol{\pi} = \boldsymbol{\gamma}$, and we also observe that $\mathbf{c}^{(n)} = \mathbf{a}\frac{1}{n}\sum_{s=1}^{n} \mathbf{P}^{s-1}$, and hence that

$$\lim_{n\to\infty} \mathbf{c}^{(n)} = \boldsymbol{\pi}$$

It can also be seen that, for every n, $\mathbf{b}^{(n)} = \mathbf{c}^{(n)}\mathbf{P}$. Taking limits on both sides (this does require some care, but we will skip the details) we finally obtain that

$$\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}$$

We thus have a solution $\boldsymbol{\pi} > 0$, of the system of equations $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}$. The remaining questions, in this part of the proof, are (i) whether $\sum_{i\in\mathcal{S}} \pi_i = 1$, and (ii) whether the solution obtained is unique.

We have seen that $\lim_{n\to\infty} \mathbf{b}^{(n)} = \boldsymbol{\pi}$. We had observed that, since $\mathbf{b}^{(n)}$ is the average of probability measures, it is also a probability measure. Hence, for every n, $\sum_{i\in\mathcal{S}} b_i^{(n)} = 1$. This, however, does not imply that $\sum_{i\in\mathcal{S}} \pi_i = 1$, since asserting this right away would require the following exchange of limits to be valid: $\sum_{i\in\mathcal{S}} \lim_{n\to\infty} b_i^{(n)} = \lim_{n\to\infty} \sum_{i\in\mathcal{S}} b_i^{(n)}$. Such an exchange is not always valid; in fact, in general, the limit of a sequence of measures could be defective. We adopt the following argument to establish our desired result. Since $\pi_i > 0$, for all $i \in S$, and, for all $k \geq 1$, $\sum_{i=1}^{k} b_i^{(n)} \leq 1$, it follows that, for every k, $0 < \sum_{i=1}^{k} \pi_i \leq 1$. Hence $0 < \sum_{i=1}^{\infty} \pi_i \leq 1$. Let $\alpha = \sum_{i=1}^{\infty} \pi_i > 0$, and consider the probability vector $\frac{\boldsymbol{\pi}}{\alpha}$. Then, clearly, $\frac{\boldsymbol{\pi}}{\alpha} = \frac{\boldsymbol{\pi}}{\alpha}\mathbf{P}$; hence we have the desired probability vector. With a slight abuse of notation, let us call this positive probability vector also $\boldsymbol{\pi}$.

Let us now prove uniqueness of such a solution. Let $\boldsymbol{\pi}$ be any solution of $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}$, $\sum_{i\in\mathcal{S}} \pi_i = 1$. Then, for every k,

$$\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}^k$$

It then follows that

$$\begin{aligned}\frac{1}{n}\sum_{k=1}^{n}\boldsymbol{\pi} &= \boldsymbol{\pi}\frac{1}{n}\sum_{k=1}^{n}\mathbf{P}^k \\ \boldsymbol{\pi} &= \boldsymbol{\pi}\frac{1}{n}\sum_{k=1}^{n}\mathbf{P}^k\end{aligned}$$

Taking the limit as $n \to \infty$, in an identical way in which the limit of $\mathbf{b}^{(n)}$ was taken above, we obtain

$$\boldsymbol{\pi} = \boldsymbol{\gamma}$$

since $\boldsymbol{\pi}$ is a probability vector. Thus the proposed solution $\boldsymbol{\pi}$ is unique and, in fact, this also shows that if $\gamma_i > 0$ then $\sum_{j\in\mathcal{S}}\ \gamma_j = 1$.
If part: Suppose the states were transient or null then, again by Theorem 2.5, it would be true that, for every i and j,

$$\frac{1}{n}\sum_{k=1}^{n} p_{ij}^{(k)} \to 0$$

Now we are given that there exists $\boldsymbol{\pi} > 0$, such that $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}$. Then, as in the earlier part of the proof we can write, for every n,

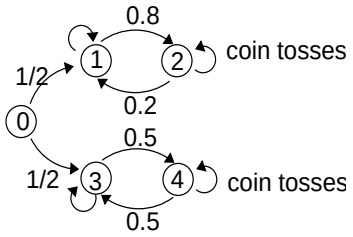
$$\boldsymbol{\pi} = \boldsymbol{\pi}\frac{1}{n}\sum_{k=1}^{n}\mathbf{P}^k$$

Taking the limit as $n \to \infty$, the right hand expression converges to 0, whereas $\boldsymbol{\pi} > 0$, yielding a contradiction. ■

Remarks 2.2.

a. Consider a DTMC with multiple positive recurrent communicating classes, $\mathcal{C}_1, \mathcal{C}_2, \cdots, \mathcal{C}_m$. If we consider the restriction of the DTMC to each of these classes, we will obtain positive probability vectors $\mathbf{a}_k, 1 \le k \le m$. By adding zeros where necessary, expand each such into a probablity vector over the entire state space, and call these probability vectors $\boldsymbol{\pi}_k, 1 \le k \le m$. It is then easily seen that $\boldsymbol{\pi}_k = \boldsymbol{\pi}_k\mathbf{P}$, for every k, and any $\boldsymbol{\pi}$ that is a convex combination of these $\boldsymbol{\pi}_k, 1 \le k \le m$, is a probability vector that also solves $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}$. Thus a DTMC with multiple positive recurrent classes will have a convex set of invariant probability vectors, whose extreme points are the vectors $\boldsymbol{\pi}_k, 1 \le k \le m$.

b. **Example**



This transition probability diagram corresponds to the following transition matrix

$$\mathbf{P} = \begin{bmatrix} 0 & 0.5 & 0 & 0.5 & 0 \\ 0 & 0.2 & 0.8 & 0 & 0 \\ 0 & 0.2 & 0.8 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0 & 0.5 & 0.5 \end{bmatrix}$$

Consider the probability vector

$$\boldsymbol{\pi} = \left(0, \frac{1}{10}, \frac{4}{10}, \frac{1}{4}, \frac{1}{4}\right)$$

It can easily be checked that $\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}$. With the initial probability distribution equal to $\boldsymbol{\pi}$, the DTMC is stationary and the ensemble average will be $\boldsymbol{\pi}$ (show this). But clearly the DTMC is not ergodic. There are two closed communicating classes $\{1, 2\}$ and $\{3, 4\}$. Depending on which of these classes the DTMC gets trapped in, along sample paths we will get averages $\left(0, \frac{1}{5}, \frac{4}{5}, 0, 0\right)$ or $\left(0, 0, 0, \frac{1}{2}, \frac{1}{2}\right)$. Note that that if the initial probability distribution is $\boldsymbol{\pi}$ then the

$$\begin{aligned} \mathsf{E}(X_k) &= \left(0 \cdot 0 + \frac{1}{10} \cdot 1 + \frac{4}{10} \cdot 2 + \frac{1}{4} \cdot 3 + \frac{1}{4} \cdot 4\right) \\ &= 0.1 + 0.8 + 0.75 + 1 = 2.65 \end{aligned}$$

whereas $P\left\{\omega : \frac{1}{n}\sum_{i=0}^{n-1} X_i(\omega) \to 2.65\right\} = 0$. Thus, for the process with initial probability $\boldsymbol{\pi} = \left(0, \frac{1}{10}, \frac{4}{10}, \frac{1}{4}, \frac{1}{4}\right)$, we have a process for which time averages along the sample paths do not permit us to estimate the expected value of the process state (i.e., we have a nonergodic process).

c. Notice that we can view the sequence of vectors $\mathbf{b}(n)$ defined above as follows

$$b_i^{(n)} = E\frac{1}{n}\sum_{s=1}^{n} I_{\{X_s=i\}},$$

i.e., $b_i^{(n)}$ is the expected frequency of visits to i. Thus we see that, for a positive class, the expected frequency of visits to a state i is $\pi_i > 0$.

■

2.7 Transience: A Criterion

Consider a DTMC $X_n, n \geq 0$, on $\mathcal{S}$, with transition probability matrix $\mathbf{P}$. Let $\mathcal{T}$ be a strict subset (possibly infinite) of $\mathcal{S}$. Define, for $n \geq 1$, and $i \in \mathcal{T}$

$$y_i^{(n)} = P(X_1 \in \mathcal{T}, X_2 \in \mathcal{T}, \cdots, X_{n-1} \in \mathcal{T}, X_n \in \mathcal{T} | X_0 = i)$$

Thus $y_i^{(n)}$ is the probability that the DTMC stays in the set of states $\mathcal{T}$ for n steps, given that the initial state is $i \in \mathcal{T}$. Equivalently, we can write,

$$y_i^{(n)} = P(\cap_{k=1}^{n} \{X_k \in \mathcal{T}\} | X_0 = i)$$

It is clear then, that $y_i^{(n)}$ is nonincreasing as n increases, and since $y_i^{(n)} \geq 0$, we conclude that the sequence $y_i^{(n)}, n \geq 1$, converges. Let us denote the limit by y_i. This same argument holds for each $i \in \mathcal{T}$, yielding $y_i, i \in \mathcal{T}$. Consider an ordering of the elements of $\mathcal{S}$ so that all the elements of $\mathcal{T}$ are contiguously indexed from 1 to $|\mathcal{T}|$. Then denote the vector $(y_1, \cdots, y_{|\mathcal{T}|})$ by $\mathbf{y}$. Also denote the restriction of the t.p.m. $\mathbf{P}$ to $\mathcal{T}$ by $\mathbf{Q}$.

Theorem 2.10. $\mathbf{y}$ *is the maximal solution of*

$$\mathbf{y} = \mathbf{Q}\mathbf{y} \quad \mathbf{0} \leq \mathbf{y} \leq \mathbf{1}$$

In addition, either $\mathbf{y} = 0$ *or* $\sup_{i \in \mathcal{T}} \ y_i = 1$.

Remark: The result makes two assertions. First, that the vector of probabilities $\mathbf{y}$ is a solution to the system of linear equations $\mathbf{y} = \mathbf{Q}\mathbf{y}$. This says that $y_i = \sum_{j \in \mathcal{T}} \ q_{ij} \ y_j$, which is intuitive since, starting in state i, the DTMC never leaves $\mathcal{T}$ if and only if, its first transition keeps it in $\mathcal{T}$ and then subsequently the DTMC never leaves $\mathcal{T}$. Second, that if there is any other vector $\mathbf{x}$ such that $\mathbf{x} = \mathbf{Q}\mathbf{x}$, with $\mathbf{0} \leq \mathbf{x} \leq \mathbf{1}$, then $\mathbf{x} \leq \mathbf{y}$. Further the theorem asserts that, either the DTMC leaves the set of states $\mathcal{T}$ with probability 1, or there are states in $\mathcal{T}$ such that starting in them there is a probability arbitrarily close to 1 of never leaving $\mathcal{T}$.

Proof: We first observe that

$$\begin{aligned} y_i^{(n)} &= P(X_1 \in \mathcal{T}, X_2 \in \mathcal{T}, \cdots, X_n \in \mathcal{T} | X_0 = i) \\ &= \sum_{j \in \mathcal{T}} P(X_1 = j, \ X_2 \in \mathcal{T}, \cdots, X_n \in \mathcal{T} | X_0 = i) \\ &= \sum_{j \in \mathcal{T}} q_{ij} \ y_j^{(n-1)} \end{aligned}$$

where we have used the Markov property. Hence, denoting $(y_1^{(n)}, y_2^{(n)}, \cdots, y_{|\mathcal{T}|}^{(n)})$ as $\mathbf{y}^{(n)}$, we can write compactly

$$\mathbf{y}^{(n)} = \mathbf{Q}\mathbf{y}^{(n-1)}$$

Defining $\mathbf{y}^{(0)} = \mathbf{1}$, and recursing this matrix equation, we obtain

$$\mathbf{y}^{(n)} = \mathbf{Q}^n \mathbf{1}$$

a result we need later in the proof.
We proceed by taking limits on both sides in $\mathbf{y}^{(n)} = \mathbf{Q}\mathbf{y}^{(n-1)}$, i.e.,

$$\lim_{n\to\infty} y_i^{(n)} = \lim_{n\to\infty} \sum_{j\in\mathcal{T}} q_{ij}\, y_j^{(n-1)}$$

The order of $\lim_{n\to\infty} \sum_{j\in\mathcal{T}}$ can be exchanged by the bounded convergence theorem; see Theorem 1.7 (the $y_j^{(n-1)}$ are bounded between 0 and 1). This yields

$$y_i = \sum_{j\in\mathcal{T}} q_{ij}\, y_j$$

Expressing the result in vector form we can write

$$\mathbf{y} = \mathbf{Q}\mathbf{y}$$

as desired.

Let $\mathbf{x}$ be any other solution with $\mathbf{0} \le \mathbf{x} \le \mathbf{1}$. We must show that $\mathbf{x} \le \mathbf{y}$. Now $\mathbf{x} \le \mathbf{1} \Rightarrow \mathbf{x} = \mathbf{Q}^n\mathbf{x} \le \mathbf{Q}^n\mathbf{1} = \mathbf{y}^{(n)}$. But $\mathbf{y}^{(n)} \downarrow \mathbf{y}$. Hence $\mathbf{x} \le \mathbf{y}$.

Turning now to the second assertion in the theorem, suppose $\mathbf{y} \neq 0$. Then let $\sup_{i\in\mathcal{T}}\ y_i = c > 0$. Clearly $0 < c \le 1$ since $\mathbf{0} \le \mathbf{y} \le \mathbf{1}$ and $\mathbf{y} \neq \mathbf{0}$. Now, since $\mathbf{y} \le c \cdot \mathbf{1}$, we have $\mathbf{y} = \mathbf{Q}^n\mathbf{y} \le \mathbf{Q}^n(c \cdot \mathbf{1}) = c\mathbf{y}^{(n)}$. Taking the limit, we see that $\mathbf{y} \le c\mathbf{y}$. Now since there exists i such that $y_i > 0$, we cannot have $c < 1$, as then we will have the absurd result that $y_i < y_i$. Hence it must be the case that $c = 1$. ■

The above result yields the following criterion for the transience of an irreducible DTMC (or of a closed communicating class of a DTMC).

Theorem 2.11. *Consider a state of an irreducible DTMC. Call it* 0 *and let* $\mathcal{T} = \mathcal{S} - \{0\}$. *Let* $\mathbf{Q}$ *denote the restriction of the transition probability matrix of the DTMC to* $\mathcal{T}$. *The DTMC is recurrent iff the only solution to* $\mathbf{y} = \mathbf{Q}\mathbf{y},\ 0 \le \mathbf{y} \le 1$, *is* $\mathbf{y} = 0$.

Proof: *if part:* The intuition is that starting in the state 0 if the DTMC leaves 0 then it must enter $\mathcal{T}$ from where it will return with probability 1. Formally, we have

$$f_{00} = p_{00} + \sum_{j\in\mathcal{T}} p_{0j}(1 - y_j) = 1$$

since each $y_j = 0$. Hence 0 is recurrent, and the DTMC is recurrent since it is irreducible. *only if part*: Given that the irreducible DTMC is recurrent, consider a state j such that $y_j > 0$; i.e., $f_{j0} < 1$. Now starting in j the chain will never leave $\mathcal{T}$ with positive probability. Hence 0 cannot be recurrent, yielding a contradiction. Note that we have used Corollary 2.1. ■

2.8 An Example: The Discrete Time M/M/1 Queue

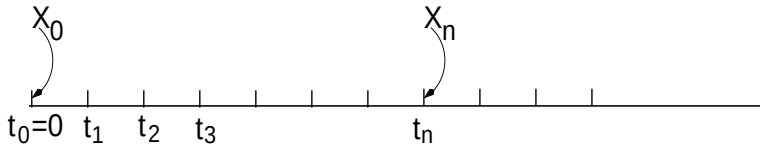


Figure 2.3: The discrete time-line over which the discrete time M/M/1 queue evolves.

In this section we consider a simple single server queue that evolves in discrete time, i.e., arrivals and departures occur in this queue only at the periodic discrete time instants $t_0, t_1, \cdots$, shown in Figure 2.3. The number of customers in the queue at time t_n+ (i.e., just after the time instant t_n) is denoted by X_n, with $X_0 = 0$. The successive interarrival times constitute an i.i.d. sequence of positive integer valued random variables $A_1, A_2, \cdots$, with common distribution

$$P(A_i = k) = (1-\lambda)^{k-1}\lambda \quad k \geq 1$$

where $0 < \lambda < 1$. Notice that since the interarrival time is at least 1, at each instant there can be at most 1 arrival. Also, we see that the interarrival time distribution is geometric.

Customers are indexed by their order of arrival, and wait in the service buffer in a first-come-first-served (FCFS) order. When the ith customer enters service, it take B_i time steps (or "slots") to complete its service, after which the customer departs. The sequence of random variables $B_i, i \geq 1$, is i.i.d. with common distribution

$$P(B_i = k) = (1-\mu)^{k-1}\mu \;\; k \geq 1$$

where $0 < \mu < 1$. Thus it takes at least one time interval to serve a customer. Notice, again, that the service time distribution is geometric.

We are interested in analysing the queue length process X_n. It is clear from the above description that the evolution of X_n is fixed once we are given a sample of the interarrival times and also of the service times of all the customers. We now provide an alternate characterisation of the evolution of the queue length process. Define two independent Bernoulli sequences, $\alpha_n, n \geq 1$, and $\delta_n, n \geq 1$, as follows

$$\begin{aligned} \alpha_n &= \begin{cases} 1 & \text{w.p. } \lambda \\ 0 & \text{w.p. } (1-\lambda) \end{cases} \\ \delta_n &= \begin{cases} 1 & \text{w.p. } \mu \\ 0 & \text{w.p. } (1-\mu) \end{cases} \end{aligned}$$

Let us take an arrival to occur at t_n if $\alpha_n = 1$. Notice that the interarrival times are i.i.d. and with the correct geometric distribution. When the queue is nonempty, let us take the

customer at the head-of-the-line (HOL) position to depart at the next discrete time instant t_m at which $\delta_m = 1$. If at some t_k the queue is empty and $\delta_k = 1$ then this is ignored. Notice that this correctly models the service time sequence. When customers depart in succession (in a busy period) this is clear. When a customer arrives to an empty queue, owing to the memoryless property of the geometric distribution the number of steps until the next instant t_m at which $\delta_m = 1$ is geomerically distributed independent of anything else. Thus we have an alernate discription of the same queueing system in terms of a Bernoulli arrival process, and a Bernoulli *virtual service* process. The service process is called virtual since each "success" in the Bernoulli process is a service only if the queue is nonempty. By an analogy with a continous time counterpart (involving Poisson processes) to be studied later, we can call this the discrete time M/M/1 queue. The first M refers to the arrival process, the second M to the service process and the 1 refers to a single server.

The above construction in terms of the Bernoulli arrival process and the Bernoulli virtual service process immediately permits us to interpret the parameters λ and μ as the *arrival rate* and the *service rate*, respectively. The fact that the arrival rate is λ follows from the strong law of large numbers applied to the Bernoulli arrival process, and similarly for the virtual service process.

With the queue length process embedded at $t_n+, n \geq 0$, the arrival process embedded at $t_n, n \geq 1$, and the virtual service process embedded at t_n- (just before t_n), we can write the following recursion

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

It is then evident that

$$\begin{aligned} & P(X_{n+1} = j | X_0 = i_0, X_1 = i_1, \cdots, X_n = i) \\ &= P((X_n - \delta_{n+1})^+ + \alpha_{n+1} = j | X_0 = i_0, X_1 = i_1, \cdots, X_n = i) \\ &= P((X_n - \delta_{n+1})^+ + \alpha_{n+1} = j | X_n = i) \\ &=: p_{ij} \end{aligned}$$

since the sequences α_n and δ_n are i.i.d. and do not depend on any of the past evolution of the queue length process. This establishes that the process X_n is a DTMC on $(0, 1, 2, \cdots,)$, with transition probabilities

$$p_{ij} = \begin{cases} 0 & \text{unless } j \in \{(i-1)^+, i+1\} \\ (1-\lambda) & \text{for } i = 0, j = 0 \\ \lambda & \text{for } i = 0, j = 1 \\ \mu(1-\lambda) & \text{for } i \geq 1, j = i-1 \\ \lambda(1-\mu) & \text{for } i \geq 1, j = 1+1 \\ 1 - (\mu(1-\lambda) + \lambda(1-\mu)) & \text{for } i \geq 1, j = i \end{cases}$$

These transition probabilities are depicted in the transition probability diagram shown in Figure 2.4. Let $\mathbf{P}$ denote the transition probability matrix.

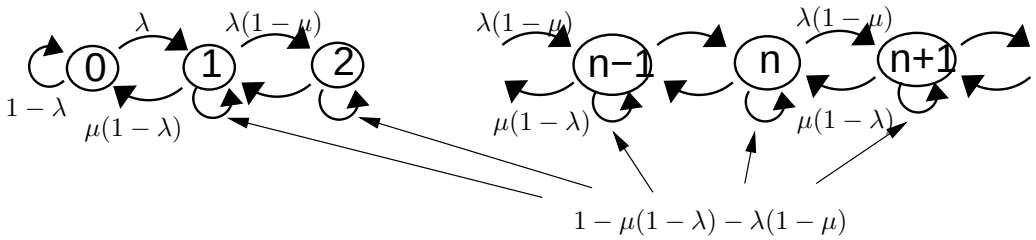


Figure 2.4: Transition probability diagram of the queue length process of the discrete time M/M/1 queue.

We now analyse this DTMC. It is evident that there is a single closed communicating class for $\lambda > 0,\ \mu > 0$, i.e., the DTMC is irreducible for $\lambda > 0,\ \mu > 0$. Next we would be interested in whether the queue is "stable" in some sense under the load offered to it, the load being parameterised by λ and μ. We recall from Theorem 2.4 (and Theorem 2.6) that if an irreducible DTMC is transient then eventually there is no positive probability on any finite state. Hence if the queue process is transient, eventually the queue "blows up" to ∞. Additionally, from Theorem 2.5 we infer that even if an irreducible DTMC is recurrent, unless it is positive recurrent, the mean frequency of visiting finite states is 0; we will elaborate on this in Section 3.10. Thus at the very least it will be important to understand the conditions under which the queue length DTMC is positive recurrent.

We start by applying Theorem 2.9. Hence we seek a probability vector $\boldsymbol{\pi}$ such that $\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}$. There is one equation for each state as follows

$$\begin{aligned} \pi_0 &= \pi_1(1-\lambda)\mu + \pi_0(1-\lambda) \\ \text{i.e., } \pi_0\lambda &= \pi_1(1-\lambda)\mu \Rightarrow \pi_1 = \frac{\pi_0}{1-\mu} \cdot \frac{\lambda/(1-\lambda)}{\mu/(1-\mu)} \\ \pi_1 &= \pi_0\lambda + \pi_1(1-(1-\lambda)\mu - (1-\mu)\lambda) + \pi_2\,\mu(1-\lambda) \end{aligned}$$

which, on using $\pi_1 = \frac{\pi_0}{1-\mu} \cdot \frac{\lambda/(1-\lambda)}{\mu/(1-\mu)}$, yields

$$\pi_2 = \pi_1 \frac{\lambda/(1-\lambda)}{\mu/(1-\mu)}$$

Similarly, for $n \geq 1$

$$\pi_n = \frac{\lambda/(1-\lambda)}{\mu/(1-\mu)}\,\pi_{n-1}$$

From these derivations we finally conclude that every solution of $\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}$ satisfies, for $n \geq 1$,

$$\pi_n = \left(\frac{\lambda/(1-\lambda)}{\mu/(1-\mu)}\right)^n \frac{\pi_0}{1-\mu}$$

However, we seek a *probability* vector that satisfies $\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}$; i.e., we want to make $\sum_{i\geq 0}\ \pi_i = 1$

$$\pi_0 \left[1 + \frac{1}{1-\mu}\ \sum_{i=1}^{\infty} \left(\frac{\lambda/(1-\lambda)}{\mu/(1-\mu)}\right)^i\right] = 1$$

Now, for $\lambda < \mu$, we can sum the infinite series, yielding,

$$\pi_0 \left[1 + \frac{1}{(1-\mu)}\ \left(\frac{1}{1-\frac{\lambda}{\mu}\frac{1-\mu}{1-\lambda}} - 1\right)\right] \ = \ 1$$

which simplifies to $\pi_0 \left[1 + \frac{\lambda}{\mu-\lambda}\right] = 1$ or $\pi_0 = 1-\rho$, where we define $\rho := \frac{\lambda}{\mu}$. Hence $\pi_0 > 0$ for $\lambda < \mu$ ($\rho < 1$) and so is $\pi_n > 0,\ n \geq 1$. Summarising the solution

$$\begin{aligned} \pi_0 &= 1-\rho \\ \pi_n &= \rho^n(1-\rho)\left(\frac{1-\mu}{1-\lambda}\right)^n\ \frac{1}{1-\mu} \end{aligned}$$

We conclude from Theorem 2.9 that for $\lambda < \mu$ (i.e., arrival rate less than service rate) the queue length DTMC is positive recurrent.

For $\frac{\lambda}{\mu} \geq 1$, consider any solution of $\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}$. As shown above, it must be of the form $\pi_n = \frac{\left(\frac{\lambda}{1-\lambda}\right)^n}{\left(\frac{\mu}{1-\mu}\right)^n} \cdot \frac{\pi_0}{1-\mu}$, where now $\frac{\left(\frac{\lambda}{1-\lambda}\right)}{\left(\frac{\mu}{1-\mu}\right)} \geq 1$. Thus, if $\pi_0 = 0$ then $\pi_i = 0$ for all i, and if $\pi_0 > 0$ then $\sum_i \pi_i = \infty$. Hence, with $\frac{\lambda}{\mu} \geq 1$, there is no solution of $\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}$ with $\pi > 0,\ \sum_i \pi_i = 1$. Hence, by the "only if" part of Theorem 2.9, the queue length DTMC is not positive for $\frac{\lambda}{\mu} \geq 1$.

To further analyse the case $\frac{\lambda}{\mu} \geq 1$, let us consider the t.p.m. $\mathbf{P}$ restricted to $(1, 2, 3, \cdots,)$, and denote this restriction by the matrix $\mathbf{Q}$. Motivated by Theorem 2.11, let us consider solutions of the system of equations $\mathbf{y} = \mathbf{Q}\mathbf{y}$, with $\mathbf{0} \leq \mathbf{y} \leq \mathbf{1}$. For notational convenience, define $p = \lambda(1-\mu)$ and $q = \mu(1-\lambda)$. Since we are considering the case $\frac{\lambda}{\mu} \geq 1$, we have $p \geq q$. Then writing out the equations in $\mathbf{y} = \mathbf{Q}\mathbf{y}$, we obtain

$$y_1 = (1-(p+q))y_1 + py_2$$

which, on rearrangement, yields

$$p(y_2 - y_1) = qy_1 \ \ \Rightarrow \ \ y_2 - y_1 = \frac{q}{p} \cdot y_1$$

Further, for $j \geq 2$, we obtain

$$y_j = y_{j-1}q + (1-(p+q))y_j + py_{j+1}$$

i.e., $q(y_j - y_{j-1}) = p(y_{j+1} - y_j)$. Defining $z_j = y_{j+1} - y_j$ for $j \geq 1$, we obtain,

$$\begin{aligned} z_1 &= \frac{q}{p}\, y_1 \\ z_j &= \frac{q}{p}\, z_{j-1} \text{ for } j \geq 2 \end{aligned}$$

Observe that we can write

$$y_j = y_j - y_{j-1} + y_{j-1} - y_{j-2} + \cdots + y_1$$

Then, using the definition of $z_j, j \geq 1$, and using the above derived expressions for z_j, we conclude that every solution of $\mathbf{y} = \mathbf{Q}\mathbf{y}$ satisfies

$$\begin{aligned} y_j &= z_{j-1} + \cdots + z_1 + y_1 \\ &= \left(\left(\frac{q}{p} \right)^{j-1} + \cdots + \frac{q}{p} + 1 \right) y_1 \end{aligned} \tag{2.3}$$

Let us now consider $\lambda = \mu$, which implies that $p = q$. Then, using Equation 2.3, we find that $y_j = j y_1$ for all $j \geq 2$. Now if $y_1 > 0$ then there exists k such that, for $j > k$, $y_j > 1$. Hence, we cannot have a solution $0 \leq y_j \leq 1$, for all j, if $y_1 > 0$. It follows that, for $\lambda = \mu$ the only solution of $\mathbf{y} = \mathbf{Q}\mathbf{y}$, with $\mathbf{0} \leq \mathbf{y} \leq \mathbf{1}$, is $\mathbf{y} = \mathbf{0}$. We conclude from Theorem 2.11 that the queue length DTMC is recurrent for $\lambda = \mu$; since we know that it cannot be positive recurrent, therefore it must be null recurrent.

Now let us consider the case $\lambda > \mu$, i.e., $q < p$. Now we have, from Equation 2.3, for $j \geq 1$,

$$y_j = \left(\frac{1 - (q/p)^j}{1 - q/p} \right) y_1$$

Choosing $y_1 = 1 - q/p$, we have a solution with $y_j \leq 1$ for all $j \geq 1$ (and notice that $\sup_{i \geq 1} y_j = 1$; see Theorem 2.10). Hence the queue length Markov chain is transient if $\lambda > \mu$.

Summarising our conclusions about the DTMC X_n, with $\rho = \frac{\lambda}{\mu}$,

$$\begin{aligned} \rho < 1 \quad & X_n \text{ is positive recurrent} \\ \rho = 1 \quad & X_n \text{ is null recurrent} \\ \rho > 1 \quad & X_n \text{ is transient} \end{aligned}$$

Recall that $\rho < 1$ is the case in which the arrival rate is *strictly* less than the service rate.

2.9 Mean Drift Criteria

In Section 2.8 we provided a detailed analysis of the discrete time M/M/1 queue. We sought to obtain conditions on the arrival rate and service rate so that the queue length

Markov chain was positive, null or transient. Our approach involved the detailed solution of certain linear equations. Such an approach can become quite cumbersome in more complicated problems. Fortunately, there are available theorems that permit the investigation of the recurrence properties of a Markov chain via a mean drift analysis. In applying these theorems the analyst needs to "guess" a certain test function (also called a stochastic Lyapunov function). A study of the mean drift of the value of this function from each state of the Markov chain then allows one to make conclusions about the recurrence properties of the process.

As we present the various results of this type, we will apply them to the analysis of the following example, which, as one can easily see, is closely related to the discrete time M/M/1 queue.

Example 2.8 (Random Walk)**.**

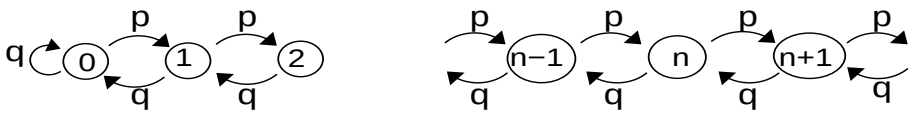


Figure 2.5: Transition probability diagram of the random walk, with $p + q = 1$.

A DTMC $X_n, n \geq 0$, on $\{0, 1, 2, \cdots\}$, with the transition probabilites shown in Figure 2.5 is called a (1-dimensional) random walk. Such a process can be obtained in the following way. Let $Z_k, k \geq 1$, be an i.i.d. sequence of random variables taking values in $\{-1, +1\}$, with $P(Z_k = 1) = p = 1 - P(Z_k = -1)$. Define $X_0 = 0$, and, for $n \geq 1$,

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

It can easily be seen that the process $X_n, n \geq 0$, is a DTMC with the t.p.d. shown in Figure 2.5.

We see that for $0 < p < 1$, the DTMC is irreducible. We wish to obtain conditions on p for which the DTMC is positive, null and recurrent. ■

Theorem 2.12. *An irreducible DTMC X_n, on $\mathcal{S} = \{0, 1, 2, \cdots\}$, is* recurrent *if and only if there exists a* positive *function y on $\mathcal{S}$, such that $y(j) \to \infty$ as $j \to \infty$, and a* finite *set $\mathcal{A} \subset \mathcal{S}$ such that*

$$\mathsf{E}(y(X_{m+1}) - y(X_m)|X_m = i) \leq 0$$

for all $i \notin \mathcal{A}$. ■

Remarks 2.3.

a. Note that $\mathsf{E}(y(X_{m+1}) - y(X_m)|X_m = i) \leq 0$ can be interpreted as the mean *drift* in the value of the function $y(\cdot)$ in one transition of the DTMC starting in the state i. This condition then asserts that the mean drift of the value of the function $y(\cdot)$ is not positive.

b. The expression $\mathsf{E}(y(X_{m+1}) - y(X_m)|X_m = i) \leq 0$ can also be seen to be written as

$$\sum_{j=0}^{\infty} p_{ij} y(j) \leq y(i)$$

where p_{ij} are the transition probabilities of X_n; i.e., the requirement is that the average value of the function $y(\cdot)$ in one transition reduces, starting in state i.

■

Example 2.9 (Random Walk (continued))**.**

Applying Theorem 2.12 to the random walk, take $y(j) = j, j \in \{0, 1, 2, \cdots\}$, and $\mathcal{A} = \{0\}$. Of course, $y(j) \to \infty$, as $j \to \infty$. Now, for $i \notin \mathcal{A}$ (i.e., $i \geq 1$), we have

$$\begin{aligned} \mathsf{E}(y(X_{m+1}) - y(X_m)|X_m = i) &= p - q \\ &\leq 0 \end{aligned}$$

for $p \leq q$. Hence the hypotheses of Theorem 2.12 are met and we conclude that X_n is recurrent for $p \leq q$, i.e., $p \leq \frac{1}{2}$. ■

Theorem 2.13. *An irreducible DTMC X_n, on $\mathcal{S} = \{0, 1, 2, \cdots\}$, is* transient *if and only if there exists a* positive *function y on $\mathcal{S}$, and a set $\mathcal{A} \subset \mathcal{S}$ such that*

(i) there exists a $j \notin \mathcal{A}$ such that $y(j) < \inf_{i \in \mathcal{A}} y(i)$, and

(ii) for all $i \notin \mathcal{A}$

$$\mathsf{E}(y(X_{m+1}) - y(X_m)|X_m = i) \leq 0$$

■

Remark: We can view the hypotheses as follows. There is some j outside $\mathcal{A}$ such that the value of $y(j)$ is less than the smallest value on $\mathcal{A}$, and the mean drift outside $\mathcal{A}$ is negative. Hence the process will tend to drift away from the set $\mathcal{A}$.

Example 2.10 (Random Walk (continued))**.**

Let $\mathcal{A} = \{0\}$ and $y(j) = \alpha^j$ for some $\alpha, 0 \leq \alpha \leq 1$. For $\alpha < 1$ and $j \geq 1$, $y(j) = \alpha^j < \alpha^0 = y(0) = \inf_{i \in \mathcal{A}} y(i)$. Now consider, for $j \geq 1$,

$$\begin{aligned} \mathsf{E}(y(X_{m+1})|X_m = j) - y(j) &= p\alpha^{j+1} + q\alpha^{j-1} - \alpha^j \\ &= \alpha^{j-1}\left(q + \alpha^2 p - \alpha\right) \\ &= \alpha^{j-1}(1 - \alpha)(q - p\alpha) \end{aligned}$$

The term $(1 - \alpha)((1 - p) - p\alpha)$ is shown in Figure 2.6 as a function of α. We see that for $q < p$ there exists α, $\frac{q}{p} < \alpha < 1$, such that $\alpha^{j-1}(1 - \alpha)(q - p\alpha) < 0$ for all $j \geq 1$.Hence, applying Theorem 2.13, $\{X_n\}$ is transient for $q < p$, i.e., $p > \frac{1}{2}$. ■

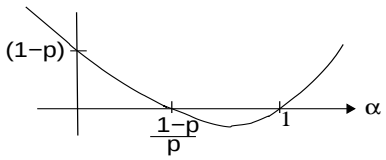


Figure 2.6: Determining the range of α in the transience analysis of the random walk.

Theorem 2.14. *An irreducible DTMC X_n, on $\mathcal{S} = \{0, 1, 2, \cdots\}$, is* positive recurrent *if and only if there exists a* positive *function y on $\mathcal{S}$, a number $\epsilon > 0$, and a* finite *set $\mathcal{A} \subset \mathcal{S}$ such that*

(i) For all $i \notin \mathcal{A}$

$$\mathsf{E}(y(X_{m+1}) - y(X_m)|X_m = i) \leq -\epsilon$$

(ii) For all $i \in \mathcal{A}$

$$\mathsf{E}(y(X_{m+1})|X_m = i) < \infty$$

■

Example 2.11 (Random Walk (continued))**.**

Let $\mathcal{A} = \{0\}$ and $y(j) = j$. Now, for $j \geq 1$,

$$\begin{aligned} \mathsf{E}(y(X_{m+1}) - y(X_m)|X_m = j) &= p - q \\ &\leq -\epsilon \end{aligned}$$

for an appropriately chosen $\epsilon > 0$, provided $p < q$. For example, take $\epsilon = \frac{q-p}{2}$. Further, $\mathsf{E}(y(X_{m+1})|X_m = 0) = p < \infty$. Hence, by Theorem 2.14, X_n is positive recurrent for $p < (1 - p)$, i.e., $p < \frac{1}{2}$. ■

Remark: It might appear from Theorems 2.12 and 2.14 that positive mean drift with an increasing test function would imply transience or lack of positive recurrence. This intuition is, in general, not correct as the following exercise shows. In addition, the importance of hypothesis (ii) in Theorem 2.14 is underlined by Problem 2.18.

Exercise 2.6.

X_n is a Markov chain on $(0, 1, 2, \cdots)$ with the transition probabilities shown in Figure 2.7. We have $p_{0i} = 2^{-i}, i \geq 1$. Show that this DTMC satisfies conditions (i) of Theorem 2.15, that it does not satisfy condition (ii), and that it is positive recurrent. (Hint: Write $\nu_{i0} = \sum_{n=1}^{\infty} n f_{i0}^{(n)}$, i.e., the mean time to reach state 0, starting from state i. Then observe that we can write the mean time to return to state 0 as $\nu_{00} = 1 + \sum_{i=1}^{\infty} p_{0i}\nu_{i0}$. Hence show that the mean time to return to state 0 is finite). ■

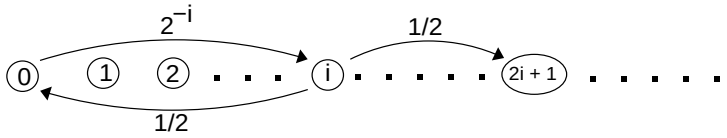


Figure 2.7: A DTMC that has positive mean drift, but is positive recurrent.

The following result provides a sufficient condition that ensures that a DTMC is *not* positive recurrent when it has positive mean drift.

Theorem 2.15. *An irreducible DTMC X_n, on $\mathcal{S} = \{0, 1, 2, \cdots\}$, is* not positive recurrent *if there exists an integer $N \geq 0$ and a real number $B \geq 0$, such that*

(i) For all $i \geq N$,

$$\mathsf{E}(X_{k+1} - X_k | X_k = i) > 0$$

(ii) For all $i \geq N$,

$$\mathsf{E}\big((X_k - X_{k+1})^+ | X_k = i\big) \leq B$$

■

Remark: The expectations in the two conditions in this theorem can be seen to be mean drifts for the function $y(j) = j$. The first condition in Theorem 2.15 states that the mean drift is strictly positive for all $i \geq N$, whereas the second condition states that the mean drift *downwards* is bounded for all $i \geq N$. Thus a positive mean drift implies that the DTMC is *not* positive recurrent if in addition the mean drift downwards is bounded.

Corollary 2.2. *An irreducible DTMC X_n on $\{0, 1, 2, \cdots\}$ is not positive recurrent if it satisfies conditions (i) of Theorem 2.15, and for which, for some $m > 0$, $p_{ij} = 0$ for $j \leq i - m$ for all $i \geq N$.* ■

Remark: The last phrase in the statement of this corollary states that for all states $i \geq N$ the state cannot reduce by more than m in one step. Clearly, this corollary is a simple consequence of Theorem 2.15, since condition (ii) also holds with $B = m$. This result can be very useful in the following way. Suppose X_n has the following evolution

$$X_{n+1} = (X_n - m)^+ + A_{n+1}$$

For example, X_n could be the number in a discrete time queueing system in which in the nth slot at most m services can take place and A_n is an i.i.d. sequence of arrivals. Then X_n is a DTMC. Theorem 2.14 applies, with $y(j) = j$ and $\mathcal{A} = \{0, 1, \cdots, m\}$, and asserts that if $\mathsf{E}(A_1) < m$ then the DTMC is positive recurrent. On the other hand Corollary 2.2 applies when $\mathsf{E}(A_1) > m$ to assert that then the DTMC is not positive recurrent.

2.10 Notes on the Bibliography

The material up to Section 2.7 has been developed from Wolff [17] and Çinlar [5]. In Section 2.9 the three main results, Theorems 2.12, 2.13, and 2.14, have been taken from Fayolle et al. [7]. A compact and elegant treatment of this topic is also provided by Asmussen [1, Chapter 1]. Theorem 2.15, a converse to Theorem 2.14, is a corollary of the main theorem in Kaplan [9].

2.11 Problems

2.1. $\{X_n, n \geq 0\}$ is a stochastic process taking values in the discrete state space S.

a. Show that if $\{X_n, n \geq 0\}$ is stationary then for all $n, m \geq 0$, and for all $i \in S$,

$$P(X_n = i) = P(X_m = i)$$

i.e., the distribution of X_n is invariant with n.

b. If $\{X_n, n \geq 0\}$ is a time homogeneous Markov chain then show that invariance of the distibution of X_n with n is sufficient for $\{X_n, n \geq 0\}$ to be stationary. Is this true for a general stochastic process $\{X_n, n \geq 0\}$?

2.2. Let $p_{ij},\ i, j \in S$, denote the transition probabilities of a Markov chain. If $X_0 = i$, let $R_i = k$ if $X_1 = i, X_2 = i, \cdots, X_{k-1} = i, X_k \neq i$, i.e., R_i is the exit time from (or residence time in) state i. Observe that $R_i \geq 1$.

a. Find $P(R_i = n | X_0 = i)$.

b. Find $\mathsf{E}(R_i | X_0 = i)$.

c. If the Markov chain has only two states (say, 0 and 1), and $X_0 = 0$, find the mean time until the Markov chain first exits 1 to enter 0. (assume $p_{01} > 0, p_{10} > 0$)

2.3. $\{X_n,\ n \geq 0\}$ is a DTMC on a state space S with transition probability matrix $\mathbf{P}$. $X_0 = i$, and for $j \neq i$, define $T_j = n\ (n \geq 1)$ if $X_0 \neq j,\ X_1 \neq j, \cdots, X_{n-1} \neq j,\ X_n = j$. Show that

$$P(X_{T_j+m} = k | X_0 = i,\ T_j < \infty) = (\mathbf{P}^m)_{jk}$$

2.4. $\{X_n\}$ is a DTMC with transition probability matrix $\mathbf{P}$. Either prove that $p_{ij}^{(k)} \geq f_{ij}^{(k)}$ or provide a counter-example.

2.5. $\{X_n, n \geq 0\}$ is a random process taking values in $\mathcal{S}$, a discrete set. Show that $\{X_n\}$ is a time homogeneous DTMC if and only if there exists a function $f(i, j)$ such that

$$P(X_{n+1} = j \mid X_0 = i_0,\ X_1 = i_1, \cdots, X_n = i) = f(i, j)$$

(Hint: for the "if" part, given the hypothesis, try to obtain an expression for $P(X_{n+1} = j \mid X_n = i)$)

2.6. Consider the Markov chain $\{X_n,\ n \geq 0\}$ on the state space $\{0, 1, 2\}$ with transition probabilities $p_{00} = p\ (0 < p < 1),\ p_{01} = 1 - p$, and $p_{12} = 1 = p_{20}$.

a. Is this Markov chain positive recurrent? Explain.

b. Find $f_{ii}^{(k)},\ i \in \{0,1,2\}$, for all $k \geq 1$.

c. Find the invariant probability distribution $\boldsymbol{\pi}$ of $\{X_n\}$.

Define $T_1 = \min\{j \geq 1 : X_j \in \{1,2\}\}$, and for all $k \geq 2, T_k = T_{k-1} + \min\{j \geq 1 : X_{T_{k-1}+j} \in \{1,2\}\}$, i.e., $\{T_k, k \geq 1\}$ are the hitting times of the set of states $\{1,2\}$. Let $Y_k = X_{T_k}, k \geq 1$.

d. Let $X_0 = 0$. Find $P(Y_k = i),\ i \in \{1,2\}$, for all $k \geq 1$.

e. Show that $\{Y_k, k \geq 1\}$ is a Markov chain, and display its transition diagram.

f. Find the invariant measure $\boldsymbol{\xi}$ for $\{Y_k\}$.

g. Suppose $P(X_0 = i) = \pi_i$; observe that $P(X_k = i) = \pi_i$, for all $k \geq 1$. Find $P(Y_k = i),\ i \in \{1,2\}$.

2.7. $\{X_n, n \geq 0\}$ is an irreducible Markov chain on a finite state space $\{0,1,2,\cdots,M\}$. For $X_0 = i \neq 0$, let $T_i = \min\{n \geq 1 :\ X_n = 0\}$. Let $\mathbf{Q}$ be the transition probability matrix restricted to $\{1,2,\cdots,M\}$.

a. Find $\mathsf{E}(T_i)$ in terms of $\mathbf{Q}$.

b. Hence show that $\mathbf{I} - \mathbf{Q}$ is nonsingular.

2.8. $\{X_n, n \geq 0\}$ is an irreducible, positive recurrent DTMC on $S = \{0,1,2,3,\cdots\}$. Let $X_0 = j$. Fix an $i \in S,\ i \neq j$, and for $k \geq 1$, define the random variable $V_k^{(i)}$ as the number of visits to i between the $(k-1)^{th}$ and k^{th} return to j. Show that $\{V_k^{(i)},\ k \geq 1\}$ is an i.i.d. sequence.

2.9. $\{X_n,\ n \geq 0\}$ is a DTMC on $\{0,1,2\cdots\}$ with $p_{0i} = \left(\frac{1}{2}\right)^i$ for $i \in \{1,2,\cdots\}$, and for $i \geq 1,\ p_{i0} = \frac{1}{2},\ p_{i,i+1} = \frac{1}{2}$.

a. Show the state transition diagram for this DTMC.

b. Is this DTMC irreducible?

c. Find $f_{00}^{(n)}$ for $n \geq 1$.

d. Hence conclude that this DTMC is positive recurrent.

2.10. $\{X_k\}$ is a time homogeneous DTMC on $S = \{1,2,\ldots,M\}$ with transition probability matrix $\mathbf{P}$, which is irreducible. For fixed N, $Y_0 = X_0,\ Y_1 = X_N,\ Y_2 = X_{2N},\ \ldots, Y_k = X_{kN}$.

a. Show that $\{Y_k\}$ is a DTMC.

b. What is the transition probability matrix $\mathbf{Q}$ of $\{Y_k\}$ in terms of $\mathbf{P}$?

c. Is $\{Y_k\}$ irreducible? Is $\{Y_k\}$ positive ? If $\{Y_k\}$ is irreducible positive what is the invariant probability vector ?

2.11. Consider a discrete time system, evolving at the times $t_0 = 0,\ t_1 = 1,\ t_2 = 2, \ldots,$ in which parts of types a and b arrive at an assembly station that produces a product ab. The arrival processes are independent and Bernoulli with rates λ_a and λ_b. The assembly time can be assumed to be 0 (zero). Let A_k and B_k denote the number of parts of types a and b at time $t_k, k \geq 0$, with $A_0 = B_0 = 0$. The arrivals occur at times $t_1-, t_2-, t_3-, \ldots$. Observe that both A_k and B_k cannot be nonzero.

a. Define $X_k = A_k - B_k$. Show that X_k is a time homogeneous DTMC, and display its transition probabilities.

b. Write down the conditions on λ_a and λ_b under which $\{X_k\}$ is irreducible.

c. When $\{X_k\}$ is irreducible, obtain conditions on λ_a and λ_b for the DTMC to be positive, null or transient. Long calculations are not necessary, but argue precisely.

2.12. Consider the DTMC $\{X_n\}$ on $\mathcal{S} = \{0, 1\}$, with $p_{01} = \alpha$ and $p_{10} = \beta$, with $0 < \alpha < 1$ and $0 < \beta < 1$. Obtain the following limits

$$\lim_{n\to\infty}\ \frac{1}{n}\sum_{k=1}^{n} p_{01}^{(k)}$$

$$\lim_{n\to\infty}\ \frac{1}{n}\sum_{k=1}^{n} p_{11}^{(k)}$$

and

$$\lim_{n\to\infty}\ \frac{1}{n}\sum_{k=1}^{n} p_{10}^{(k)}$$

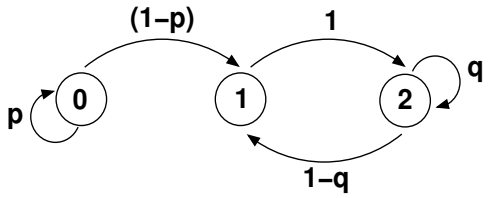
2.13. $\{X_n, n \geq 0\}$ is an irreducible DTMC on $\mathcal{S} = \{0, 1, 2, \cdots\}$. For $j \in \mathcal{S}$, let M_j denote the number of visits to state j for $n \geq 1$.

a. Write down an expression for $P(M_j = k \mid X_0 = i)$ in terms of f_{ij} and f_{jj} (no need to derive the strong Markov property).

b. Obtain $\mathsf{E}(M_j \mid X_0 = i)$ in terms of f_{ij} and f_{jj}.

c. Hence show that if $\{X_n\}$ is transient, then for all $i, j, \in \mathcal{S}$, $\lim_{k\to\infty}\ p_{ij}^{(k)} = 0$.

2.14.



In the displayed transition probability diagram, $0 < p < 1$ and $0 < q < 1$.

a. Obtain f_{00}, and hence obtain $\lim_{k\to\infty} p_{00}^{(k)}$.

b. Obtain $\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} p_{11}^{(k)}$ and $\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} p_{22}^{(k)}$.

c. What is the interpretation of the results in (a) and (b) ?

2.15. $\{X_i, i \geq 1\}$ is a sequence of i.i.d. random variables with $P(X_i = 1) = 1/2$, and $P(X_i = -1) = 1/2$. $\{S_n, n \geq 0\}$ is a process defined as follows. For some given integer k,

$$S_0 = k$$

and for $n \geq 1$,

$$S_n = S_0 + \sum_{i=1}^{n} X_i$$

a. Show that $\{S_n, n \geq 1\}$ is a Markov chain. Display its transition probability diagram.

b. Classify the Markov chain $\{S_n\}$ as follows

   i. reducible or irreducible,

   ii. transient or null or positive.

   Explain your arguments clearly.

2.16. There are N empty boxes and an infinite collection of balls. At each step, a box is chosen at random and a ball is placed in it. Let X_n be the number of empty boxes after the n^{th} ball has been placed.

a. Show that $\{X_n, n \geq 0\}$ is a Markov chain.

b. Display its state space and transition probabilities.

c. Classify the states as transient, null or positive.

2.17. $\{Y_n,\ n \geq 1\}$ is a sequence of i.i.d. random variables with $P(Y_n = 2) = p = 1 - P(Y_n = -1)$. $\{X_n,\ n \geq 0\}$ is defined as follows

$$X_0 = 0$$
$$X_{n+1} = (X_n + Y_{n+1})^+$$

a. Show that $\{X_n\}$ is a DTMC.

b. Display its transition probability diagram.

c. Obtain a sufficient condition for the positive recurrence of $\{X_n\}$.

d. Let Z_k = no. of times $\{X_n\}$ visits 0 over the interval $[0,\ k]$. How will you obtain

$$\lim_{n\to\infty} \frac{1}{n} \mathsf{E}(Z_n)$$

for the parameter values obtained in part (c)? (No need to derive the actual expressions.)

2.18. $\{X_n\}$ is a DTMC on $\{0, 1, 2, \cdots\}$ such that $p_{0i} = \frac{\alpha}{i^2}$ for all $i \geq 1$, with α such that $\sum_{i=1}^{\infty} \frac{\alpha}{i^2} = 1$, and for $i \geq 1$, $p_{i,i+1} = p = 1 - p_{i,i-1}$ with $p < 1/2$.

a. Display the transition probability diagram.

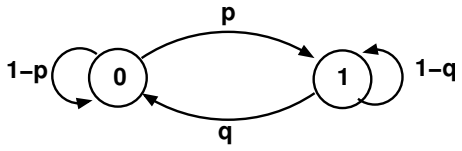
b. Show that $f_{00} = 1$. What does this imply?

c. Show that there is an $\epsilon > 0$ such that $\mathsf{E}(X_{n+1} - X_n | X_n = i) < -\epsilon$ for all $i \geq 1$, and yet the DTMC is null recurrent.

d. Does (c) violate Theorem 2.14?

(Hint: Notice that on $s \in \{1, 2, \cdots\}$ we have a simple random walk. Further, observe that $\nu_{00} = \sum_{k=1}^{\infty} \frac{\alpha}{k^2} \left(1 + \sum_{j=k}^{\infty} f_{k0}^{(j)} \cdot j\right)$, and try lower bounding this.)

2.19. Consider a DTMC $\{X_n,\ n \geq 0\}$ with the transition diagram shown in the figure, with $0 < p < 1$ and $0 < q < 1$. Define $Y = \min\{k \geq 1 : X_k \neq X_0\}$, i.e., the time spent in the initial state.



Obtain, for $i \in \{0, 1\}$,

a. $P(Y = k \mid X_0 = i)$, and name the resulting distribution,

b. f_{ii},

c. $\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} p_{ii}^{(k)}$, and explain the meaning of this limit.

2.20. $\{X_n\}$ is a random walk on $\{0, 1, 2, \ldots\}$ with $p_{00} = 1 - p,\ p_{01} = p$ and, for $i \geq 1$, $p_{i(i-1)} = 1 - p - a$, $p_{ii} = a > 0$, and $p_{i(i+1)} = p$, with $0 < p < 1$ and $a + p < 1$.

a. For $j \geq 1$, obtain $s_j^{(k)} = P\left(X_1 = j,\ X_2 = j, \cdots, X_{k-1} = j,\ X_k \neq j \mid X_0 = j\right)$.

b. Show that the DTMC is positive recurrent for $p < \frac{1-a}{2}$. (Hint: Use Theorem 2.14)

c. For the condition in (b) define, for $j \geq 1$,

$$f_{jj}^{(k)}(j-1) = P(k \text{ returns to } j \text{ without hitting } j-1 \mid X_0 = j)$$

Obtain $\sum_{k=0}^{\infty} k f_{jj}^{(k)}(j-1)$.

Chapter 3

Renewal Theory

Consider an irreducible recurrent DTMC $Y_n, n \geq 0$, with $Y_0 = i$, and consider visits to the state j. Let X_1 denote the time until the first visit, and let $X_k, k \geq 2$, denote the subsequent intervisit times. We recall from Chapter 2 that $P(X_1 = m|Y_0 = i) = f_{ij}^{(m)}$, and, for $k \geq 2$, $P(X_k = m) = f_{jj}^{(m)}$. Also, let $Z_k, k \geq 1$, denote the time of the kth visit to state j. Now for any $k \geq 1$, Z_k is a stopping time for the process Y_n. It then easily follows that (see Exercise 2.5)

$$P(X_{k_1} = m_1, X_{k_2} = m_2, \cdots, X_{k_n} = m_n|Y_0 = i) \quad = \quad \begin{cases} f_{ij}^{(m_1)} f_{jj}^{(m_2)} \cdots f_{jj}^{(m_n)} & \text{for } k_1 = 1 \\ f_{jj}^{(m_1)} f_{jj}^{(m_2)} \cdots f_{jj}^{(m_n)} & \text{for } k_1 > 1 \end{cases}$$

Thus, we see that the sequence of intervisit times, $X_k, k \geq 1$, to the state j in an irreducible recurrent DTMC is a sequence of mutually independent random variables, with the $X_k, k \geq 2$, being a sequence of i.i.d. random variables. Such a sequence of times often arises in discrete event processes, and we call them *renewal life-times*. The associated instants Z_k are called *renewal instants.* This terminology is motivated by the analogy of a component in a system being repeatedly replaced after it gets worn out (for example, the batteries in a portable electronic device).

3.1 Definition and Some Related Processes

In general, a renewal process can be defined as follows.

Definition 3.1. *Given a sequence of mutually independent nonnegative real valued random variables, $X_k, k \geq 1$, with the random variables $X_k, k \geq 2$, also being identically distributed,*

a. this sequence of random variables is called the sequence of life-times *of the renewal process,*

b. for $k \geq 1$, $Z_k = \sum_{i=1}^{k} X_i$ *is the kth* renewal instant*; define* $Z_0 = 0$.

c. For $t \geq 0$, $M(t) = \sup\{k \geq 0 : Z_k \leq t\}$, *the number of renewals in* $[0, t]$, *is called the* renewal process*, and,*

d. for $t \geq 0$, $m(t) = \mathsf{E}(M(t))$ *is called the* renewal function.

■

Remarks 3.1.

a. In general, Z_k is a discrete parameter random process that takes nonnegative real values. Also, $M(t)$ is a continuous time random process that takes values in $\{0, 1, 2, 3, \cdots\}$. On the other hand $m(t)$ is just a nonnegative, real valued, nondecreasing, (deterministic) function of time.

b. Clearly, the processes Z_k and $M(t)$ are determined when a sample path of the life-times $X_k, k \geq 1$, is given. Thus we can think of each such sample path of the life-time process as an elementary outcome ω.

c. Note, from this definition, that $M(t)$ stays flat between renewal instants and increases in jumps at renewal instants. A renewal at time t is included in the process $M(t)$. Thus the sample paths of $M(t)$ are nondecreasing step functions that are right continuous at the jumps. It is possible to have multiple jumps at an instant; for example, it is possible that $Z_k = Z_{k+1}$ in which case $M(Z_{k+1}) = M(Z_k)$, and $M(Z_{k+1}) - M(Z_k-) = 2$.

■

Several interesting related processes can be defined, and some relationships between the processes can be observed.

a. $Z_{M(t)}$ is the instant of the last renewal in the interval $[0, t]$. Notice $M(t)$ is a random time (or random index) for the process Z_k, and $Z_{M(t)}$ is the value of the process Z_k at the random time $M(t)$. Hence, for a sample path ω, $Z_{M(t)}$ means $Z_{M(t,\omega)}(\omega)$. By definition, $Z_{M(t)} \leq t$.

b. $M(t)+1$ is the index of the first renewal after t. Hence, $Z_{M(t)+1}$ is the first renewal instant (strictly) after t, i.e., the first renewal instant in (t, ∞). We have $Z_{M(t)+1} > t$.

c. For $t \geq 0$, $Y(t) = Z_{M(t)+1} - t$ is the *residual life-time* at time t. Thus $Y(0) = X_1$, $Y(t)$ decreases at 1 unit per unit time until Z_1, then $Y(t)$ jumps up by X_2, and so on. The sample paths of $Y(t)$ are right continuous at the jump instants $Z_k, k \geq 1$, and $Y(Z_k-) = 0, k \geq 1$.

d. For $t \geq 0$, define the *age* process $U(t)$ by $U(t) = t - Z_{M(t)}$. If $M(t) = 0$, since, by definition, $Z_0 = 0$, we have $U(t) = t$.

e. Notice that, for all $n \geq 1$, and $t \geq 0$, $(Z_n \leq t) = (M(t) \geq n)$, i.e., the event that the nth renewal occurs at or before t is the same as the event that there are at least n renewals in $[0, t]$. This is easily seen by checking that each ω that is in the event on the left hand side is in the event on the right hand side, and vice versa.

We now see how the renewal function, $m(t)$, can be expressed in terms of the life-time distributions. Let $A(\cdot)$ be the distribution of X_1 and $F(\cdot)$ that of $\{X_2, X_3, \ldots, \}$. It then follows that

$$\begin{aligned} P(Z_1 \leq t) &= A(t) \\ P(Z_2 \leq t) &= \int_0^t F(t-u) dA(u) \\ &= (A \star F)(t) \end{aligned}$$

where, as usual, $\star$ denotes the convolution of the distributions A and F[1]. Continuing, we have

$$\begin{aligned} P(Z_3 \leq t) &= \int_0^t F(t-u)\, d(A \star F)(u) \\ &= (A \star F \star F)(t) = (A \star F^{(2)})(t) \end{aligned}$$

and, for $n \geq 1$,

$$P(Z_n \leq t) = \left(A \star F^{(n-1)}\right)(t)$$

Thus we conclude that

$$\begin{aligned} P(M(t) \geq n) &= P(Z_n \leq t) \\ &= \left(A \star F^{(n-1)}\right)(t) \end{aligned}$$

Hence we can write the renewal function as follows

$$\begin{aligned} m(t) &= EM(t) \\ &= \sum_{n=1}^{\infty} P(M(t) \geq n) \end{aligned}$$

[1] If X and Y are independent random variables, with c.d.f.s $F(x)$ and $G(y)$, then the c.d.f. of $Z := X + Y$ is obtained as

$$P(Z \leq z) = P(X + Y \leq z) = \int_{u=0}^{\infty} F(z-u)\, dG(u) =: (F \star G)(z),$$

where $\star$ denotes convolution.

Thus we obtain the following expression for $m(t)$ in terms of the life-time distributions.

$$m(t) = \sum_{n=1}^{\infty} \left(A \star F^{(n-1)} \right)(t) \tag{3.1}$$

Based on this observation we can establish the following lemma that will be used in several results to follow.

Lemma 3.1. *If* $\mathsf{E}(X_j) > 0, j \geq 2$, *then* $m(t) < \infty$ *for all* t.

Remark: Note that for the nonnegative random variables X_j, $\mathsf{E}(X_j) > 0$ is equivalent to the statement $P(X_j > 0) > 0$, or $F_j(0^+) < 1$. Observe that this result implies that, under the hypotheses, $M(t) < \infty$ with probability 1.

Proof: We first observe that, for any $n \geq 1$, and $0 \leq m \leq n$,

$$\begin{aligned} F^{(n)}(t) &= \int_0^t F^{(n-m)}(t-u)\, dF^{(m)}(u) \\ &\leq F^{(n-m)}(t) \int_0^t dF^m(u) \\ &\leq F^{(n-m)}(t) F^{(m)}(t) \end{aligned} \tag{3.2}$$

where in the two inequalities we have used the fact that a distribution function is nondecreasing. Hence, applying the previous inequality recursively, for any $n \geq 1, r \geq 1, k \geq 0$,

$$\begin{aligned} F^{(nr+k)}(t) &\leq F^{((n-1)r+k)}(t) \cdot F^{(r)}(t) \\ &\leq \left(F^{(r)}(t) \right)^n F^{(k)}(t) \end{aligned} \tag{3.3}$$

Now, from Equation 3.1, we have, for any $r \geq 1$,

$$\begin{aligned} m(t) &= \sum_{n=1}^{\infty} \left(A \star F^{(n-1)} \right)(t) \\ &\leq \sum_{n=0}^{\infty} F^{(n)}(t) \\ &= \sum_{m=0}^{\infty} \underbrace{\sum_{k=0}^{r-1} F^{(mr+k)}(t)}_{\leq r F^{(mr)}(t)} \\ &\leq r \sum_{m=0}^{\infty} \left(F^{(r)}(t) \right)^m \end{aligned}$$

where the first inequality uses the same calculation that let to (3.2) and the second inequality uses (3.3). Now, for each t choose r such that $F^{(r)}(t) < 1$; then we see, from the last expression, that $m(t) < \infty$. Such a choice of t is possible by virtue of the hypothesis that $\mathsf{E}(X_2) > 0$. For then there exists $\epsilon > 0$ such that $F(\epsilon) < 1$. Now, for every n, $(1 - F^{(n)}(n\epsilon)) \geq (1 - F(\epsilon))^n > 0$; i.e., $F^{(n)}(n\epsilon) < 1$. It follows that $F^{\lceil \frac{t}{\epsilon} \rceil}(\epsilon \lceil \frac{t}{\epsilon} \rceil) < 1$. Hence, $F^{\lceil \frac{t}{\epsilon} \rceil}(t) < 1$. ■

3.2 The Elementary Renewal Theorem (ERT)

Theorem 3.1. *Given a sequence of mutually independent nonnegative random variables (life-times) $X_k, k \geq 1$, with $X_k, k \geq 2$, being identically distributed, such that,*

(i) for $k \geq 1$, $P(X_k < \infty) = 1$ (all the life-time random variables are proper),

(ii) $0 \leq \mathsf{E}(X_1) \leq \infty$ (we allow the mean of the first life-time to be 0 and also ∞), and,

(iii) for $k \geq 2$, $0 < \mathsf{E}(X_k) \leq \infty$ (the mean life-times after the first are positive, and possibly infinite, and, of course, identical to $\mathsf{E}(X_2)$). Defining $\mathsf{E}(X_2) = \frac{1}{\mu}$, this hypothesis is equivalent to $0 \leq \mu < \infty$.

Then the following conclusions hold

(a) $\lim_{t\to\infty} \frac{M(t)}{t} = \mu$ almost surely

(b) $\lim_{t\to\infty} \frac{m(t)}{t} = \mu$

■

Remark:

a. Conclusion (a) of the theorem states that the rate of renewals converges almost surely to μ. This is intuitive since, after the first renewal, which occurs in finite time with probability one, the subsequent interrenewal times have mean $\frac{1}{\mu}$. Since $t \to \infty$ the effect of the first life-time eventually vanishes.

b. In the second conclusion we simply have a limit of numbers (unlike in the first, where there is a sequence of random variables). Note that we cannot say that Conclusion (a) implies Conclusion (b) since this would involve the interchange of expectation and limit which is not always legal; see Section 1.4.1

Proof: *Part (a)*

We first observe that, since, for all $j \geq 1$, $P(X_j < \infty) = 1$, hence, for all $k \geq 1$, $P(\cap_{j=1}^{k} \{X_j < \infty\}) = 1$. Further, we observe that

$$\{Z_k < \infty\} = \cap_{j=1}^{k} \{X_j < \infty\}$$

which can be verified by checking that a sample point ω is in the event on the left if and only if it is in the event on the right. Hence it follows that, for all $k \geq 1$,

$$P(Z_k < \infty) = 1$$

i.e., every renewal occurs in finite time with probability one. From this observation it follows that

$$P(\lim_{t\to\infty} M(t) = \infty) = 1$$

for if this were not the case we would have positive probability that a renewal occurs at infinity. Let us now consider the case $\mu > 0$, i.e., $\mathsf{E}(X_2) < \infty$. Observe that

$$\begin{aligned} \frac{Z_n}{n} &= \frac{1}{n}\sum_{i=1}^{n} X_i \\ &= \frac{X_1}{n} + \frac{1}{n}\sum_{i=2}^{n} X_i \end{aligned}$$

Now since $P(X_1 < \infty) = 1$, it follows that $P(\frac{X_1}{n} \to 0) = 1$. Also, since $X_j, j \geq 2$, are i.i.d. with finite mean, by Theorem 1.10, $\frac{1}{n}\sum_{i=2}^{n} X_i \to \frac{1}{\mu}$ with probability 1. Since the intersection of two sets with probability 1 also has probability 1, it follows that, with probability 1,

$$\frac{Z_n}{n} \to \frac{1}{\mu}$$

Also, we saw that, $P(M(t) \to \infty) = 1$. Hence it further follows that

$$P\left(\frac{Z_{M(t)}}{M(t)} \to \frac{1}{\mu}\right) = 1$$

We have seen earlier that

$$Z_{M(t)} \leq t < Z_{M(t)+1}$$

Dividing across by $M(t)$ (for t large enough so that $M(t) > 0$), we have

$$\frac{Z_{M(t)}}{M(t)} \leq \frac{t}{M(t)} < \frac{Z_{M(t)+1}}{M(t)}$$

But $\frac{Z_{M(t)}}{M(t)} \to \frac{1}{\mu}$ with probability 1. Hence $\frac{t}{M(t)} \to \frac{1}{\mu}$ almost surely, from which the desired conclusion follows. We have proved the first conclusion of the theorem for the case $\mathsf{E}(X_2) < \infty$.

For the case $\mathsf{E}(X_2) = \infty$, i.e., $\mu = 0$, Theorem 1.10 cannot be directly used, and so we need to proceed as in Exercise 3.1.

To prove the second part of the theorem we need Wald's Lemma (Lemma 3.2). ■

Exercise 3.1.

Complete the proof of the first part of Theorem 3.1 for the case in which $\mathsf{E}(X_2) = \infty$. (Hint: use the truncated lifetime sequence $X_k^{(c)}, k \geq 1$, as defined in the proof of Part (b) of Theorem 3.1, provided later.) ■

Lemma 3.2 (Wald's Lemma)**.** *Let N be a stopping time for an infinite sequence of mutually independent random variables $X_i, i \geq 1$. If*

(i) $\mathsf{E}(N) < \infty$,

(ii) $\mathsf{E}(|X_n|) < B$, *a constant, for all* $n \geq 1$, *and*

(iii) $\mathsf{E}(X_n) = \mathsf{E}(X_1)$ *for all* $n \geq 1$,

then

$$\mathsf{E}\left(\sum_{n=1}^{N} X_n\right) = \mathsf{E}(X_1) \cdot \mathsf{E}(N)$$

Remarks 3.2.

a. We emphasise that the sequence of random variables, $X_i, i \geq 1$, in the statement of the lemma are not necessarily nonnegative. Further, the random variables need not have the same distribution. They, however, have the same mean. Note that a sequence of i.i.d. nonnegative random variables with finite common expectation (as would arise in a renewal theory application) satisfies conditions (ii) and (iii) of the lemma.

b. Let us also observe that a seemingly obvious calculation does not always work. Let $p_k = P(N = k), k \geq 1$, and define $S_N := \sum_{n=1}^{N} X_n$. Now

$$\begin{aligned} \mathsf{E}(S_N) &= \mathsf{E}\left(\mathsf{E}\left(\sum_{n=1}^{N} X_n | N\right)\right) \\ &= \sum_{k=1}^{\infty} p_k \, \mathsf{E}\left(\sum_{n=1}^{N} X_n | N = k\right) \end{aligned}$$

Now if we could write $\mathsf{E}\left(\sum_{n=1}^{N} X_n | N = k\right) = k\mathsf{E}(X_1)$ then it would immediately follow that $\mathsf{E}(S_N) = \mathsf{E}(X_1)\,\mathsf{E}(N)$. However, in general, this step is incorrect, because conditioning on $N = k$ may change the joint distribution of $X_i, 1 \leq i \leq k$. In fact, given that $N = k$ the random variables $X_i, 1 \leq i \leq k$, may also have become dependent.

c. As an illustration of the previous remark, consider a renewal process with i.i.d. lifetimes $X_i, i \geq 1$, and, for given $t > 0$, define the random time $N = M(t)$. Now look at $\sum_{n=1}^{M(t)} X_n$, and observe that, given $M(t) = k$, it must be that $\sum_{n=1}^{k} X_n \leq t$, so that the random variables $X_1, X_2, \cdots, X_k$ are conditionally dependent and also are conditionally bounded between 0 and t.

Proof: We can write

$$\begin{aligned} S_N &= \sum_{n=1}^{N} X_n \\ &= \sum_{n=1}^{\infty} X_n \, I_{\{n \leq N\}} \end{aligned}$$

Hence

$$\mathsf{E}(S_N) = \mathsf{E}\left(\sum_{n=1}^{\infty} X_n \, I_{\{n \leq N\}} \right) \tag{3.4}$$

Suppose we could exchange $\mathsf{E}(\cdot)$ and $\sum_{n=1}^{\infty}$ in the right hand side (we will justify this before ending the proof). This will yield

$$\mathsf{E}(S_N) = \sum_{n=1}^{\infty} \mathsf{E}\big(X_n I_{\{n \leq N\}}\big)$$

Now observe that

$$\begin{aligned} I_{\{n \leq N\}} &= 1 - I_{\{N \leq (n-1)\}} \\ &= f(X_1, X_2, \cdots, X_{n-1}) \end{aligned}$$

for some function $f(\cdots)$, since N is a stopping time for the sequence $X_i, i \geq 1$. But the $X_i, i \geq 1$, are mutually independent. It therefore follows that X_n is independent of $I_{\{n \leq N\}}$, and we obtain

$$\mathsf{E}(S_N) = \sum_{n=1}^{\infty} \mathsf{E}(X_n) \cdot P(N \geq n)$$

Using the fact that $\mathsf{E}(X_n) = \mathsf{E}(X_1)$, we obtain

$$\mathsf{E}(S_N) = \mathsf{E}(X_1)\, \mathsf{E}(N)$$

and the result is proved.

We finally turn to the justification of the exchange of $\mathsf{E}(\cdot)$ and $\sum_{n=1}^{\infty}$ in Equation 3.4. Define $Y = \sum_{n=1}^{\infty} |X_n| \, I_{\{n \leq N\}}$. It can be seen that, for all $m \geq 1$,

$|\sum_{n=1}^{m} X_n\, I_{\{n\le N\}}| \le Y$. Then using the facts that $\mathsf{E}(|X_n|) < B$ and $\mathsf{E}(N) < \infty$, and that N is a stopping time, it follows that $\mathsf{E}(Y) < \infty$. Dominated convergence theorem then applies and it follows that we can interchange $\mathsf{E}(\cdot)$ and $\sum_{n=1}^{\infty}$ in the right hand side of Equation 3.4.

■

Example 3.1 (The Poisson Process)**.**

Later in this chapter we will study an important renewal process called the Poisson process. This is a renewal process with i.i.d. life-times that are exponentially distributed with mean $\frac{1}{\lambda}$. Since $\sum_{i=1}^{M(t)} X_i \le t$ and there is a positive probability that $\sum_{i=1}^{M(t)} X_i \le t_1 < t$ (e.g., take $t_1 = \frac{t}{2}$), we can see that $\mathsf{E}\Big(\sum_{i=1}^{M(t)} X_i\Big) < t$. On the other hand, we will see that $M(t)$ is a Poisson distributed random variable with mean λt. Hence $\mathsf{E}(M(t)) \cdot \mathsf{E}(X_1) = \lambda t\, \frac{1}{\lambda} = t$, and the conclusion of Wald's Lemma does not hold. The reason is that $M(t)$ is not a stopping time for the renewal process (while all the other hypotheses of the lemma hold). To see this note that, for any $n \ge 1$,

$$\begin{aligned} I_{\{M(t)\le n\}} &= 1 - I_{\{M(t)\ge n+1\}} \\ &= 1 - I_{\{Z_{n+1}\le t\}} \\ &= 1 - I_{\{\sum_{i=1}^{n+1} X_i \le t\}} \\ &= I_{\{\sum_{i=1}^{n+1} X_i > t\}} \end{aligned}$$

Hence to determine if $M(t) \le n$ we need to look at X_{n+1} as well. Hence $M(t)$ is not a stopping time. This becomes evident when we observe that $M(t)$ is the index of the *last* complete life-time before t. ■

Corollary 3.1. *Let $M(t)$ be a renewal process with i.i.d. life-times $X_i, i \ge 1$, and with $0 < \mathsf{E}(X_1) < \infty$. Then $\mathsf{E}\big(Z_{M(t)+1}\big) = \mathsf{E}(X_1)\,(m(t)+1)$.*

Proof: Define the random time $N = M(t) + 1$. Now observe that

$$\begin{aligned} I_{\{N\le n\}} &= 1 - I_{\{N\ge n+1\}} \\ &= 1 - I_{\{M(t)\ge n\}} \\ &= 1 - I_{\{Z_n\le t\}} \\ &= I_{\{Z_n > t\}} \\ &= I_{\{\sum_{i=1}^{n} X_i > t\}} \end{aligned}$$

Hence N is a stopping time for the life-times $X_i, i \ge 1$. Since $0 < \mathsf{E}(X_1)$, from Lemma 3.1 it follows that $m(t) < \infty$ for every t. Applying Wald's Lemma 3.2 we obtain

the desired result, as follows

$$\mathsf{E}\big(Z_{M(t)+1}\big) = \mathsf{E}\left(\sum_{i=1}^{M(t)+1} X_i\right) = \mathsf{E}(X_1)\,(m(t)+1)$$

■

Example 3.2 (The Poisson Process (continued))**.**

Let us again consider the Poisson process. We will see later in this chapter that in a Poisson process at any time t the remaining time until the next renewal is exponentially distributed with mean $\frac{1}{\lambda}$. This is simply a consequence of the memoryless property of the exponential distribution. It follows that

$$\mathsf{E}\left(\sum_{i=1}^{M(t)+1} X_i\right) = \mathsf{E}(t + Y(t)) = t + \frac{1}{\lambda}$$

Further, since $M(t)$ is Poisson distributed with mean λt, we get

$$\mathsf{E}(M(t)+1)\,\mathsf{E}(X_1) = (\lambda t + 1)\frac{1}{\lambda} = t + \frac{1}{\lambda}$$

thus verifying that Wald's Lemma holds in this case. ■

We now continue the proof of Theorem 3.1.

Proof: Theorem 3.1, *Part (b)*

We first take the case: $\mathsf{E}(X_1) < \infty$ and $\mathsf{E}(X_2) < \infty$ (i.e., $\mu > 0$), and consider

$$Z_{M(t)+1} = \sum_{j=1}^{M(t)+1} X_j$$

Following arguments identical to the ones in the proof of Wald's Lemma, we can write (see the proof of Lemma 3.2)

$$\begin{aligned}
\mathsf{E}\big(Z_{M(t)+1}\big) &= \sum_{j=1}^{\infty} \mathsf{E}(X_j) \cdot P((M(t)+1) \geq j) \\
&= \mathsf{E}(X_1) \cdot P(M(t) \geq 0) + \sum_{j=2}^{\infty} \mathsf{E}(X_j) \cdot P(M(t)+1 \geq j) \\
&= \mathsf{E}(X_1) + \mathsf{E}(X_2)\,\mathsf{E}(M(t)) \\
&= \mathsf{E}(X_1) + m(t)\mathsf{E}(X_2) \\
&= \mathsf{E}(X_1) + \frac{m(t)}{\mu}
\end{aligned}$$

where in the first equality we have already used the fact that $M(t)+1$ is a stopping time for the i.i.d. sequence $X_j, j \geq 2$. We have also used the fact that $\mathsf{E}(X_2) < \infty$, and that $\mathsf{E}(X_2) > 0$, which in turn implies (using Lemma 3.1) that $\mathsf{E}(M(t)) < \infty$.

Now we observe that $Z_{M(t)+1} > t$, by definition. Hence

$$\frac{m(t)}{\mu} > t - \mathsf{E}(X_1)$$

i.e.,

$$\frac{m(t)}{t} > \mu - \mu\frac{\mathsf{E}(X_1)}{t}$$

We conclude that

$$\liminf_{t\to\infty} \frac{m(t)}{t} \geq \mu$$

We will be done if we show that $\limsup_{t\to\infty} \frac{m(t)}{t} \leq \mu$. For this purpose, for each $c > 0$, define $X_j^{(c)}$ as

$$X_j^{(c)} = \begin{cases} X_j & \text{if} \quad X_j \leq c \\ c & \text{if} \quad X_j > c \end{cases}$$

Further, define $\mu^{(c)}$ by $\mathsf{E}\left(X_2^{(c)}\right) = \frac{1}{\mu^{(c)}}$. Now consider the renewal process generated by $\{X_j^{(c)},\ j \geq 1\}$; i.e., for any realisation of the life-times $X_j, j \geq 1$, this new renewal process has life-times obtained by truncating the life-times $X_j, j \geq 1$, as shown above. Use the superscript $^{(c)}$ to denote any process associated with this new renewal process. Then, clearly, $Z^{(c)}_{M^{(c)}(t)+1} \leq t + c$ since $X^{(c)}_{M^{(c)}(t)+1} \leq c$. Proceeding as above, we now obtain

$$\begin{aligned} \mathsf{E}\left(X_1^{(c)}\right) + \frac{m^{(c)}(t)}{\mu^{(c)}} &\leq t + c \\ \frac{m^{(c)}(t)}{t} &\leq \mu^{(c)} + \underbrace{\left(\frac{c - \mathsf{E}\left(X_1^{(c)}\right)}{t}\right)}_{\geq 0} \mu^{(c)} \end{aligned}$$

hence $\limsup_{t\to\infty} \frac{m^{(c)}(t)}{t} \leq \mu^{(c)}$.

But, for all $j \geq 1$, $X_j^{(c)} \leq X_j$, hence $Z_n^{(c)} \leq Z_n$, which implies that $M^{(c)}(t) \geq M(t)$ and hence that $m^{(c)}(t) \geq m(t)$. Thus we have

$$\limsup_{t\to\infty} \frac{m(t)}{t} \leq \limsup_{t\to\infty} \frac{m^{(c)}(t)}{t} \leq \mu^{(c)}$$

But $\mu^{(c)} \to \mu$ as $c \to \infty$, hence

$$\limsup_{t\to\infty} \frac{m(t)}{t} \leq \mu$$

now

$$\mu \leq \liminf_{t\to\infty} \frac{m(t)}{t} \leq \limsup_{t\to\infty} \frac{m(t)}{t} \leq \mu$$

hence

$$\lim_{t\to\infty} \frac{m(t)}{t} = \mu$$

If $\mathsf{E}(X_1) < \infty$ but $\mathsf{E}(X_2) = \infty$ then $\mu = 0$ and the last part of the proof holds, i.e., we can still show $0 \leq \limsup_{t\to\infty} \frac{m(t)}{t} \leq 0$ hence

$$0 \leq \liminf_{t\to\infty} \frac{m(t)}{t} \leq \limsup_{t\to\infty} \frac{m(t)}{t} \leq 0$$

We skip the proof for $\mathsf{E}(X_1) = \infty$. ■

3.2.1 Application to DTMCs

We now turn to the proof of Theorem 2.5. Consider a DTMC $Y_n, n \geq 0$, on $\mathcal{S} = 0, 1, 2, \cdots$, with $Y_0 = i$. In the theorem statement, state j is given to be recurrent. As discussed in the beginning of this chapter, the visits to the state j will define a renewal process, with life-times $X_j, j \geq 1$. Note that here all the lifetimes are integer valued random variables. Since f_{ij} is given to be 1, and j is recurrent, we have $P(X_j < \infty) = 1$, for all $j \geq 1$. Further, $\mathsf{E}(X_2) = \nu_j$, the mean return time to the state j, with $\nu_j < \infty$ if j is positive recurrent, and $\nu_j = \infty$ if j is null recurrent.

Proof: Theorem 2.5

We apply the expectation version of ERT, Theorem 3.1, to the renewal process of visits to state j. For this renewal process, we see that, for $n \geq 1$,

$$\begin{aligned}
\frac{m(n)}{n} &= \mathsf{E}\left(\left(\frac{M(n)}{n}\right) \mid Y_0 = i\right) \\
&= \mathsf{E}\left(\left(\frac{1}{n}\sum_{k=1}^{n} I_{\{Y_k=j\}}\right) \mid Y_0 = i\right) \\
&= \frac{1}{n}\sum_{k=1}^{n} p_{ij}^{(k)}
\end{aligned}$$

Now, $P(X_j < \infty) = 1$, $0 \leq \mathsf{E}(X_1) \leq \infty$, and $1 \leq \mathsf{E}(X_2) = \nu_j$ (it takes at least one transition to return to j). Applying Theorem 3.1, we obtain

$$\frac{1}{n}\sum_{k=1}^{n} p_{ij}^{(k)} = \frac{m(n)}{n} \to_{t\to\infty} \frac{1}{\nu_j} \begin{cases} = 0 & \text{if } \nu_j = \infty \text{ (i.e., } j \text{ is null recurrent)} \\ > 0 & \text{if } \nu_j < \infty \text{ (i.e., } j \text{ is positive recurrent)} \end{cases}$$

■

Remark: With the above notation in mind, recall from Theorem 2.9 that when a DTMC is positive recurrent then the invariant measure $\pi_j, j \in \mathcal{S}$, will have the values $\pi_j = \frac{1}{\nu_j}$. Thus, for a positive recurrent class, the mean time to return to a state is the reciprocal of the invariant probability of that state. Also, we see that π_j has the interpretation of the mean rate of visiting the state j.

3.3 Renewal Reward Processes

Consider a renewal process with i.i.d. life times $X_j, j \geq 1$. Associated with each X_j is a "reward" R_j, such that $R_j, j \geq 1$, is also an i.i.d. sequence, however, R_j may depend on X_j (and, in general, *will* be dependent). We say R_j is the reward during cycle j. Note that we can think of the life-times or cycle times, together with the rewards, as the i.i.d. sequence of random vectors $(X_j, R_j), j \geq 1$.

Example 3.3.

Consider a DTMC $Y_k, k \geq 0$, taking values in $\{0, 1, 2, \cdots\}$. Let $Y_0 = j$ and consider returns to the state j. This defines a renewal process with life-times (or cycle times) $X_k, k \geq 1$, the times between the successive returns to j. Now let $i \neq j$ be another state, and, for $k \geq 1$, let R_k be the number of visits to state i during the duration between the $k-1$th visit to j and the kth visit to j, i.e., in the kth cycle time. It is then easily seen that $R_k, k \geq 1$, are i.i.d. Of course, R_j and X_j are dependent random variables, as one can expect that the longer is the time between two visits to j, the more often would the DTMC visit i between these two visits. ■

With the reward R_j being "earned" in cycle j, let, for $t \geq 0$, $C(t)$ be the net reward until time t (including any reward at t; i.e., like our other processes, $C(t)$ is also right continuous). Now several cases can arise, depending on how the rewards accrue.

- R_j may be obtained at the end of the cycle j. Then

$$C(t) = \sum_{j=1}^{M(t)} R_j$$

  i.e., the total reward until time t is the sum of the rewards in cycles *completed* until time t.

- R_j may be obtained at the beginning of cycle j. Then

$$C(t) = \sum_{j=1}^{M(t)+1} R_j$$

i.e., the total reward until time t is the sum of the rewards in cycles *begun* until time t.

- R_j may be earned over the cycle j (continuously or in discrete parts). Suppose at time instant t, $R(t)$ is defined as the partial reward earned until t in the current cycle. Then, clearly,

$$C(t) = \sum_{j=1}^{M(t)} R_j + R(t)$$

Theorem 3.2 (Renewal-Reward Theorem (RRT)). *If $(X_j, R_j), j \geq 1$, constitute an i.i.d. renewal reward process, with $\mathsf{E}(R_1) < \infty$, and $0 < \mathsf{E}(X_1) < \infty$, and $C(t)$ is the total reward accumulated during $[0, t]$, then*

(a) $\lim_{t\to\infty} \frac{C(t)}{t} = \frac{\mathsf{E}(R_1)}{\mathsf{E}(X_1)}$ *w.p.1*

(b) $\lim_{t\to\infty} \frac{\mathsf{E}(C(t))}{t} = \frac{\mathsf{E}(R_1)}{\mathsf{E}(X_1)}$

Remark: Note that if at the end of each cycle we obtain the reward $R_j = 1$ then the conclusions of this theorem are the same as those of Theorem 3.1, i.e., ERT.

Proof: *of Part (a).*

Consider the case

$$C(t) = \sum_{j=1}^{M(t)} R_j$$

Then we can write

$$\frac{C(t)}{t} = \frac{\sum_{j=1}^{M(t)} R_j}{M(t)} \cdot \frac{M(t)}{t}$$

Now, $\mathsf{E}(X_1) < \infty$ implies that $P(X_1 < \infty) = 1$. Hence, as we saw in the proof of Theorem 3.1, $M(t) \to \infty$, w.p. 1. Then, using the fact that $\mathsf{E}(R_1) < \infty$, Theorem 1.10 implies that $\frac{\sum_{j=1}^{M(t)} R_j}{M(t)} \to \mathsf{E}(R_1)$, w.p. 1. Further, by Theorem 3.1, $\frac{M(t)}{t} \to \frac{1}{\mathsf{E}(X_1)}$, w.p. 1. It follows that, w.p. 1,

$$\lim_{t\to\infty} \frac{C(t)}{t} = \frac{\mathsf{E}(R_1)}{\mathsf{E}(X_1)}$$

An identical argument works for

$$C(t) = \sum_{j=1}^{M(t)+1} R_j$$

after writing this as

$$\frac{C(t)}{t} = \frac{\sum_{j=1}^{M(t)+1} R_j}{M(t)+1} \, \frac{M(t)+1}{t}$$

Turning to the case where the rewards in a cycle accrue gradually over the cycle, write the reward $R_j = R_j^+ - R_j^-$, where $x^+ = \max\{x, 0\}$, and $x^- = \max\{-x, 0\}$, i.e., we split the reward as the net "gain" minus the net "loss." Also write

$$C(t) = C^+(t) - C^-(t)$$

i.e., even the net reward until t is split as the net gain minus the net loss. Then, we can write

$$\sum_{j=1}^{M(t)} R_j^+ \leq C^+(t) \leq \sum_{j=1}^{M(t)+1} R_j^+$$

Note that this inequalty does not hold if we replace $C^+(t)$ with $C(t)$ and R_j^+ with R_j, since, in general, rewards need not be positive. It follows then, from the cases already proved, that, w.p. 1,

$$\lim_{t\to\infty} \frac{C^+(t)}{t} = \frac{\mathsf{E}\big(R_1^+\big)}{\mathsf{E}(X_1)}$$

In an identical fashion, we also obtain

$$\lim_{t\to\infty} \frac{C^-(t)}{t} = \frac{\mathsf{E}\big(R_1^-\big)}{\mathsf{E}(X_1)}$$

It follows that, w.p. 1,

$$\lim_{t\to\infty} \frac{C(t)}{t} = \frac{\mathsf{E}\big(R_1^+\big) - \mathsf{E}\big(R_1^-\big)}{\mathsf{E}(X_1)} = \frac{\mathsf{E}(R_1)}{\mathsf{E}(X_1)}$$

Proof of Part (b).
Consider the case

$$C(t) = \sum_{j=1}^{M(t)+1} R_j$$

We know that $M(t) + 1$ is a stopping time for $X_j,\ j \geq 1$, i.e., $I_{\{M(t)+1\geq n\}}$, or, equivalently, $I_{\{M(t)+1\leq(n-1)\}}$, is determined by $(X_1, X_2, \cdots, X_{n-1})$. Since $(R_j,\ j \geq n)$ is independent of $(X_1, X_2, \cdots, X_{n-1})$, it follows that $I_{\{M(t)+1\geq n\}}$ is independent of R_n. Then, following an argument exactly as in the proof of Lemma 3.2 (Wald's Lemma), we obtain

$$\mathsf{E}(C(t)) = \mathsf{E}(R_1)\,\mathsf{E}(M(t) + 1)$$

i.e.,

$$\mathsf{E}(C(t)) = \mathsf{E}(R_1)\,(m(t) + 1)$$

Now, dividing by t, and using the expectation part of ERT (Theorem 3.1), it follows that

$$\lim_{t\to\infty} \frac{\mathsf{E}(C(t))}{t} = \frac{\mathsf{E}(R_1)}{\mathsf{E}(X_1)}$$

We omit the proofs of the remaining cases. ■

3.3.1 Application to Time Averages

Let $Y(t), t \geq 0$, be the residual life process of a renewal process with i.i.d. life-times. The common distribution of the life-times is denoted by $F(\cdot)$. Then consider, for a fixed $y \geq 0$,

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t P(Y(u) \leq y)\, du$$

Remark: This expression can be interpreted in two ways. Consider a random observer "arriving" uniformly over the interval $[0, t]$. Then the expression inside the limit, i.e,

$$\int_0^t P(Y(u) \leq y)\, \frac{1}{t}\, du$$

can be interpreted as the distribution of the residual life seen by the random observer; since $\frac{1}{t}du$ is the probability that the observer arrives in the infinitesimal interval $(u, u+du)$, and, conditional on this happening, $P(Y(u) \leq y)$ is the probability of the observer seeing a residual life-time $\leq y$. On the other hand we can write the expression (inside the $\lim_{t\to\infty}$) as

$$\mathsf{E}\left(\frac{1}{t}\int_0^t I_{\{Y(u)\leq y\}} du\right)$$

i.e., the expected fraction of time over $[0, t]$ during which the residual life is $\leq y$. In either case we are asking for the limit of the expression as $t \to \infty$.

Theorem 3.3. *For a renewal process with i.i.d. life-times, such that* $0 < \mathsf{E}\left(X_1^2\right) < \infty$*, the following hold:*

(i) For fixed $y \geq 0$, $\lim_{t\to\infty} \frac{1}{t}\int_0^t P(Y(u) \leq y)du = \frac{1}{\mathsf{E}(X_1)}\int_0^y (1 - F(x))dx$

(ii) With probability 1, $\lim_{t\to\infty} \frac{1}{t}\int_0^t Y(u)du = \frac{\mathsf{E}\left(X_1^2\right)}{2\mathsf{E}(X_1)}$.

Remarks 3.3.

Before we prove Theorem 3.3, we make some observations about its conclusions. Note that, since X_1 is a non-negative random variable, $\mathsf{E}\left(X_1^2\right) > 0$ implies that $\mathsf{E}(X_1) > 0$.

a. We first observe that the right hand side of Conclusion (i) in this theorem (i.e., $\frac{1}{\mathsf{E}(X_1)}\int_0^y (1 - F(x))dx$), as a function of y, is a distribution. To see this, note that

this expression is nonnegative, nondecreasing with y, and also recall that $\int_0^\infty (1 - F(x))dx = \mathsf{E}(X_1)$. Given a distribution $F(\cdot)$ of a nonnegative random variable, with finite mean, define, for all y,

$$F_e(y) = \frac{1}{\mathsf{E}(X_1)} \int_0^y (1 - F(x))dx$$

$F_e(\cdot)$ is called the *excess* distribution corresponding to the (life-time) distribution $F(\cdot)$.

b. Notice that, for each t, the expression $\frac{1}{t}\int_0^t P(Y(u) \leq y)du$, as a function of y, is also a distribution. Thus, the first conclusion of the theorem states that the time average distribution of the residual life process converges to the excess distribution of the life-time distribution $F(\cdot)$.

c. The second conclusion of the theorem states that the time-average of the residual life process $Y(t)$ converges almost surely to the number $\frac{\mathsf{E}(X_1^2)}{2\mathsf{E}(X_1)}$. It can easily be verified that

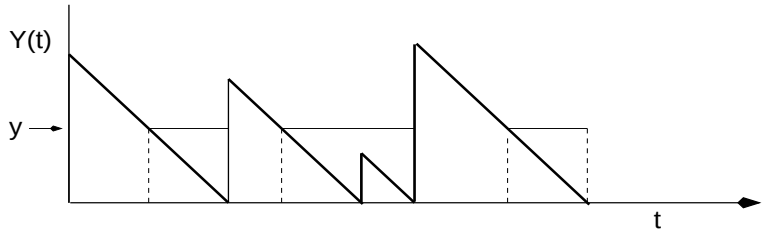
$$\int_0^\infty (1 - F_e(y))dy = \frac{\mathsf{E}(X_1^2)}{2\mathsf{E}(X_1)}$$

To show this, substitute the expression for $F_e(y)$, interchange the order of integration, and use the fact that $\int_0^\infty 2x(1 - F(x))dx = \mathsf{E}(X_1^2)$). Thus, we see that the limit of the time average of the residual life process is the expectation of its time average limiting distribution. A little later in this chapter, we will see that this kind of a result holds more generally. An application of such a result can be that if we are able to obtain the limiting distribution, and we know that such a result holds, then taking the expectation of the limit distribution provides us with the limit of the time average of the process.

Proof: Theorem 3.3, *Part (i).*
Define, for fixed $y \geq 0$, the reward process

$$C(t) = \int_0^t I_{\{Y(u) \leq y\}}\, du$$

Thus $C(t)$ is the total time in $[0, t]$ that the proces $Y(t)$ is below y (the integral is the area under the thin horizontal lines in the figure). We see that we have a renewal reward process



with the reward R_j in the jth cycle being given by

$$R_j = \int_{Z_{j-1}}^{Z_j} I_{\{Y(u)\leq y\}} du$$

where, as before, Z_j is the jth renewal instant. Since the life-times are i.i.d. it follows that $\{R_j,\ j \geq 1\}$ are also i.i.d. Further, observe that

$$R_j = \left\{ \begin{array}{lll} X_j & \text{if} & X_j < y \\ y & \text{if} & X_j \geq y \end{array} \right\} = \min(X_j, y)$$

Thus $0 \leq R_j \leq X_j$. Further, $0 < \mathsf{E}\big(X_1^2\big) < \infty$ implies that $0 < \mathsf{E}(X_1) < \infty$, and, hence, that $\mathsf{E}(R_1) < \infty$. Thus the conditions of RRT (Theorem 3.2) are met. Hence

$$\lim_{t\to\infty} \mathsf{E}\left(\frac{C(t)}{t}\right) \quad = \quad \frac{\mathsf{E}(R_1)}{\mathsf{E}(X_1)}$$

In order to obtain $\mathsf{E}(R_1)$ let us obtain the complementary c.d.f. of R_1. For $r \geq 0$,

$$\begin{array}{rcl} P(R_1 > r) & = & P(\min\{X_1, y\} > r) \\ & = & P(X_1 > r, y > r) \\ & = & \left\{ \begin{array}{ll} P(X_1 > r)(= 1 - F(r)) & \text{for} \quad r < y \\ 0 & \text{for} \quad r \geq y \end{array} \right. \end{array}$$

It follows that

$$\mathsf{E}(R_1) = \int_0^y (1 - F(r))dr$$

We conclude that

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t P(Y(u) \leq y)du = \frac{1}{\mathsf{E}(X_1)} \int_0^y (1 - F(u))du$$

Proof of Part (ii). Let us take the cumulative reward over $[0, t]$ to be

$$C(t) = \int_0^t Y(u)du$$

i.e., cumulative "area" under $Y(t)$ over $[0, t]$. Then the reward in the jth cycle becomes

$$R_j = \int_{Z_{j-1}}^{Z_j} Y(u)du$$

Therefore,

$$R_j = \frac{1}{2}X_j^2$$

It is evident that $R_j, j \geq 1$, is a sequence of i.i.d. random variables. Further, since we are given that $\mathsf{E}\big(X_1^2\big) < \infty$, it follows that $\mathsf{E}(R_1) < \infty$. Thus the conditions of RRT (Theorem 3.2) are met, and we conclude that, with probability 1,

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t Y(u)du = \frac{\mathsf{E}\big(X_1^2\big)}{2\mathsf{E}(X_1)}$$

■

3.3.2 Length and Batch Biasing

We can interpret the second part of Theorem 3.3 as asserting that in a renewal process with i.i.d. life-times the mean residual time seen by a random observer is given by $\frac{\mathsf{E}\big(X_1^2\big)}{2\mathsf{E}(X_1)}$. The following (fallacious) argument gives a different (and wrong) answer. "If a randomly arriving observer arrives in a life-time of duration x, then (since he arrives uniformly over this interval) he sees a mean residual time of $\frac{x}{2}$. The distribution of life-times is $F(\cdot)$. Hence, unconditioning on the length of the life-time into which the observer arrives, the mean residual life seen by a random observer should be given by $\int_0^\infty \frac{x}{2}\, dF(x) = \frac{1}{2}\mathsf{E}(X_1)$." In fact, we notice that, since $\mathsf{E}\big(X_1^2\big) \geq (\mathsf{E}(X_1))^2$ (with equality only if the random variable X_1 is constant with probability 1),

$$\frac{\mathsf{E}\big(X_1^2\big)}{2\mathsf{E}(X_1)} \geq \frac{1}{2}\mathsf{E}(X_1)$$

with strict inequality if the variance of X_1 is positive. Thus, the correct mean residual life will typically be larger than half the mean life-time. What is the flaw in the argument (in quotes) above? The error is in the use of the distribution $F(\cdot)$ when unconditioning on the distribution of the life-time into which the random observer arrives. In fact, we should use a *length biased* distribution as we shall see next.

Theorem 3.4. *If the life-time distribution has finite mean* $\mathsf{E}(X_1)$*, then, for given* $x \geq 0$*,*

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t P(X(u) \leq x) du$$
$$= \frac{1}{EX_1} \left(xF(x) - \int_0^x F(u) du \right) := F_s(x)$$

Proof: Consider, for given $x \geq 0$, the cumulative reward process

$$C(t) = \int_0^t I_{\{X(u)\leq x\}} du$$

and the reward sequence R_j defined by

$$R_j = \begin{cases} 0 & \text{for} \quad X_j > x \\ X_j & \text{for} \quad X_j \leq x \end{cases}$$

Since $R_j \leq X_j$, $\mathsf{E}(R_1) \leq \mathsf{E}(X_1) < \infty$. Hence, applying RRT (Theorem 3.2) we conclude that

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t P(X(u) \leq x) du = \frac{\mathsf{E}(R_1)}{\mathsf{E}(X_1)}$$

In order to evaluate $\mathsf{E}(R_1)$ we first obtain the complementary c.d.f. of R_j. Clearly, $P(R_j > x) = 0$. Further, for $0 \leq u \leq x$,

$$\begin{aligned} P(R_j > u) &= P(X_j \leq x, X_j > u) \\ &= P(X_j \leq x) - P(X_j \leq u) \\ &= F(x) - F(u) \end{aligned}$$

Hence

$$\begin{aligned} \mathsf{E}(R_1) &= \int_0^\infty P(R_1 > u)\, du \\ &= \int_0^x (F(x) - F(u))\, du \\ &= xF(x) - \int_0^x F(u)\, du \end{aligned}$$

from which the result follows. ■

Remarks 3.4.

a. Given a life-time distribution $F(\cdot)$, the distribution $F_s(\cdot)$ obtained in Theorem 3.4 is called the *spread distribution* of $F(\cdot)$. Compare this with the earlier definition of the excess distribution $F_e(\cdot)$.

b. As in the case of $F_e(\cdot)$, the distribution $F_s(\cdot)$ can be viewed as the distribution of the *total* life-time seen by a random observer. When $F(\cdot)$ has a density, then it is easily seen, by differentiating the distribution $F_s(\cdot)$, that the density of the spread distribution is given by

$$f_s(x) = \frac{1}{\mathsf{E}(X_1)}(xf(x))$$

Thus we see that the density is different from $f(\cdot)$. Now following the discussion at the beginning of this section, if we obtain the mean residual life by taking $f_s(\cdot)$ to be the distribution of life-time seen by a random observer, then we will obtain

$$\int_0^\infty \frac{x}{2}\frac{1}{\mathsf{E}(X_1)}(xf(x))dx = \frac{\mathsf{E}(X_1^2)}{2\mathsf{E}(X_1)}$$

the desired answer. What is the intuition behind $f_s(\cdot)$ being the correct density to use, and not $f(\cdot)$? The following rough argument will make this clear. Consider the time interval $[0, T]$, where T is suitably large. Over this time there are roughly $\frac{T}{\mathsf{E}(X_1)}$ renewal cycles; this follows from ERT. Out of these cycles the number that have durations in the interval $(x, x + dx)$ is approximately $\frac{T}{\mathsf{E}(X_1)}f(x)dx$. The amount of time in the interval $[0, T]$ that is covered by such intervals is $x\frac{T}{\mathsf{E}(X_1)}f(x)dx$, where we have ignored terms of order smaller than dx. Hence a random arrival over $[0, T]$ arrives in an interval of length $(x, x + dx)$ with probability $\frac{x\frac{T}{\mathsf{E}(X_1)}f(x)dx}{T} = f_s(x)dx$. Hence $f_s(\cdot)$ is the density of the life-time seen by a random observer. It should now be clear why $f_s(\cdot)$ is called the *length biased* density.

c. The above discussion illustrates the phenomenon of *length biased sampling*. Let $X_k, k \geq 1$, be i.i.d. life-time samples from the c.d.f. $F(\cdot)$ (with nonnegative support). Suppose we take a large number n of samples and place them in a bin and then draw a sample, the distribution of the sample will be $F(\cdot)$. On the other hand if we place the life-time samples "side-by-side" (thinking of them as line segments) on the positive real line (starting at the origin), and then "draw" a sample by randomly picking a point of the positive real line, yielding the sample in which the point falls, then evidently our sampling will be biased towards the larger values, as these cover more of the line.

■

The same phenomenon obviously occurs when the random variables $X_k, k \geq 1$, take only nonnegative integer values, but in this context it is called *batch size biasing*. Suppose families either have 1 child or 2 children, with probability 0.5 for each case. If mothers are asked how many children they have then roughly half will reply "one", and other half will reply "two," yielding an average number of children equal to 1.5. On the other hand if children are asked how many siblings they are, roughly $\frac{1}{3}$rd will answer "one", and the rest will answer "two," yielding the batch biased average: $\frac{5}{3}$. The biasing occurs since in the population of children more children come from the larger families and hence we more often get the larger answer. In some queueing systems, customers arrive in batches, and if a customer is picked at random and asked the size of its batch a batch biased answer will result[2]. We can again study this using the renewal reward theorem.

Let the batch sizes be $\{X_j,\ j \geq 1\}$; these are i.i.d. nonnegative integer random variables representing the number of customers in the batches indexed by $j \geq 1$. Let $P(X_1 = k) = p_k, k \geq 1$. Let us index the customers by $n \in \{1, 2, 3, \cdots\}$, so that customers 1 to X_1 are in the first batch, from $X_1 + 1$ to $X_1 + X_2$ are in the second batch, and so on. Note that we can view this as a renewal process in discrete "time," with the batches corresponding to "life-times." Let $X(n)$ denote the batch of the nth customer; this notation is similar to the notation $X(t)$ in continuous time renewal theory introduced earlier. Now, for fixed $k \geq 1$, consider

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} I_{\{X(i)=k\}}$$

i.e., each customer is asked if its batch size is k, and we seek the fraction of customers who respond with a "yes." View the cumulative "reward" until the nth customer as $\sum_{i=1}^{n} I_{\{X(i)=k\}}$, and for $j \geq 1$, define

$$\begin{aligned} R_j &= 0 \text{ if } X_j \neq k \\ &= k \text{ if } X_j = k \end{aligned}$$

Evidently $(X_j, R_j), j \geq 1$, are i.i.d. and we have a renewal reward process. We have

$$\mathsf{E}(R_j) = kp_k < \infty$$

Hence, using Theorem 3.2, w.p. 1

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} I_{\{X(j)=k\}} = \frac{kp_k}{\mathsf{E}(X_1)}$$

[2]One consequence of this is the following. Consider a single server queue with batch arrivals. If the batches of customers are served in first-come-first-served order, then the delay of a typical customer is the sum of the delay of the batch in which it arrives, and the total service time of the customers in its own batch that are served before it. This latter number will need to be obtained by using the batch biased distribution.

which is the biased distribution of the batch size. It can similarly be seen that

$$\lim_{t\to\infty} \frac{1}{n} \sum_{j=1}^{n} P(X(j) = k) = \frac{kp_k}{\mathsf{E}(X_1)}$$

and, w.p. 1,

$$\lim_{t\to\infty} \frac{1}{n} \sum_{j=1}^{n} X(j) = \frac{\sum_{k=1}^{\infty} k^2 p_k}{\mathsf{E}(X_1)} = \frac{\mathsf{E}(X_1^2)}{\mathsf{E}(X_1)}$$

which is the average of replies received if each customer is asked "What is your batch size?"

Remarks 3.5.

Finally, in the context of the renewal reward theorem, it is important to observe that the form of the result obtained, i.e., that the limit of the average reward rate is $\frac{\mathsf{E}(R_1)}{\mathsf{E}(X_1)}$, incorporates the effect of length biasing. In particular, one could ask why the limit was not $\mathsf{E}\left(\frac{R_1}{X_1}\right)$ (in fact, this expression has sometimes been erroneously used). For example, suppose $(X_k, R_k), k \geq 1$, are i.i.d.. Suppose that $(X_k, R_k) = (1, 10)$ with probability 0.5, and $(X_k, R_k) = (10, 1000)$ with probability 0.5. So, on the average, in half the intervals the reward rate is 10, and in half the intervals the reward rate is 100. One might want to say that the average reward rate is $0.5 \times 10 + 0.5 \times 100 = 55$; this would be the answer one would get if the formula $\mathsf{E}\left(\frac{R_1}{X_1}\right)$ is used. Yet, the theorem declares that the answer is $\frac{\mathsf{E}(R_1)}{\mathsf{E}(X_1)} = \frac{0.5\times10+0.5\times1000}{0.5\times1+0.5\times10} = 91.82$. It is easily checked that this is the answer that we will get if we use the length biased distribution for the cycle time.

■

3.4 The Poisson Process

A point process $N(t), t \geq 0$, is a random process taking values in $0, 1, 2, \cdots$, such that $N(0) = 0$, and, for each ω, $N(t, \omega)$ is a nondecreasing, right continuous step function. A point process is essentially characterised by a random distribution of points on $\mathbb{R}^+$ (i.e., the jump instants), and a sequence of integer random variables corresponding to the jump at each point. If we think of a batch of arrivals at each jump, then $N(t)$ is the number of arrivals in the interval $[0, t]$.

Definition 3.2 (Poisson Process)**.** *A point process $N(t), t \geq 0$, is called a Poisson process if*

(i) *all jumps of $N(t)$ are of unit size, with probability 1,*

(ii) *for all $t, s \geq 0$, $(N(t + s) - N(t)) \amalg \{N(u) : u \leq t\}$, and*

(iii) *for all* $t, s \geq 0$*, distribution of* $N(t+s) - N(t)$ *does not depend on* t*.*

■

Remarks 3.6.

a. For a point process $N(t)$, given $t \geq 0$ and $\tau \geq 0$, $N(t+\tau) - N(t)$ is called the *increment* of $N(t)$ over the interval $(t, t+\tau]$, i.e., $N(t+\tau) - N(t)$ is the cumulative value of the jumps in the interval $(t, t+\tau]$.

b. Thus Definition 3.2 (ii) asserts that any increment of a Poisson process is independent of the past of the process. We can conclude more. Consider $t_1 < t_2 < t_3 \cdots < t_n$ then, by property (ii) in the definition,

$$N(t_1) \amalg (N(t_2) - N(t_1))$$

Further, again using the same property,

$$N(t_3) - N(t_2) \amalg (N(t_1), N(t_2))$$

or, equivalently,

$$(N(t_3) - N(t_2)) \amalg (N(t_1), N(t_2) - N(t_1))$$

Thus, it follows that,

$$(N(t_3) - N(t_2)), (N(t_2) - N(t_1)), N(t_1) \text{ are mutually independent}$$

Thus we conclude that the increments of a Poisson process (over disjoint intervals) are independent. This is known as the *independent increment* property.

c. Definition 3.2 (iii) states that the distribution of an increment depends only on the width of the interval over which it is taken, not the location of the interval in time. This is called the *stationary increment* property.

d. Thus we can say that a Poisson process is a point process with stationary and independent increments. In addition, at each point of a Poisson process there is a unit jump.

■

Definition 3.2 defines a stochastic process. We have seen that a stochastic process is characterised in terms of its finite dimensional distributions (recall Section 1.3.1). Are the properties specified in the definition of a Poisson process sufficient to specify its finite dimensional distributions? The answer is indeed "Yes" as we now proceed to show. For

$t_1 < t_2 < t_3 < \cdots < t_n$, and $i_1 \leq i_2 \leq i_3 \leq \cdots \leq i_n$, using the properties of the Poisson process, we find that the finite dimensional distribution can be computed as follows

$$\begin{aligned} & P(N(t_1) = i_1, N(t_2) = i_2, \cdots, N(t_n) = i_n) \\ &= P(N(t_1) = i_1, N(t_2) - N(t_1) = i_2 - i_1, \cdots, N(t_n) - N(t_{n-1}) = i_n - i_{n-1}) \\ &= P(N(t_1) = i_1) \cdot P(N(t_2 - t_1) = i_2 - i_1) \cdots P(N(t_n - t_{n-1}) = i_n - i_{n-1}) \end{aligned}$$

where, in writing the second equality, we have used the stationary and independent increment properties. Thus we would have the finite dimensional distributions if we could obtain $P(N(t) = k)$ for each $t \geq 0$ and $k \in \{0, 1, 2, \cdots\}$. This distribution is obtained via the following lemmas.

Lemma 3.3. *There exists a $\lambda, 0 \leq \lambda < \infty$, such that, for all $t \geq 0$, $P\{N(t) = 0\} = e^{-\lambda t}$.*

Remark: This is equivalent to the assertion that the time until the first jump in a Poisson process is exponentially distributed with mean $\frac{1}{\lambda}$.
Proof: Using the stationary and independent increment property, we can write

$$\begin{aligned} P(N(t+s) = 0) &= P(N(t) = 0, N(t+s) - N(t) = 0) \\ &= P(N(t) = 0) P(N(t+s) - N(t) = 0) \end{aligned}$$

Let us write, for $t \geq 0$, $f(t) = P(N(t) = 0)$. Thus, we have established that $f(\cdot)$ satisfies the *functional equation:* for all $s, t \geq 0$,

$$f(t+s) = f(t)\, f(s)$$

Since $N(0) = 0$, we have $f(0) = 1$. Define $T_1 := \inf\{t > 0 : N(t) \neq 0\}$, the first "jump" time of $N(t)$. Clearly, $f(t) = P(T_1 > t)$. Thus, $f(t)$ is the complementary c.d.f. of a nonnegative random variable; hence, $f(t), t \geq 0$, is right continuous and nonincreasing. It then follows that (see Theorem 3.22 in the Appendix of this chapter) the only nonzero solution of the functional equation is

$$f(t) = e^{-\lambda t}$$

for some $\lambda, 0 \leq \lambda < \infty$.

■

Lemma 3.4.

$$\lim_{t \to 0} \frac{1}{t} P(N(t) \geq 2) = 0$$

Remark: This result states that $P(N(t) \geq 2) = o(t)$, i.e., that the probability of there being 2 or more points of the process in an interval of length t decreases to 0 faster than t, as t decreases to 0.
Proof: We skip the proof of this result, and only note that the proof utilises the property (i) in Definition 3.2. ■

Lemma 3.5.

$$\lim_{t\to 0} \frac{1}{t} P(N(t) = 1) = \lambda,$$

where λ is as obtained in Lemma 3.3.

Remark: In other words, $P(N(t) = 1)$ can be approximated as $\lambda t + o(t)$ as $t \to 0$.
Proof:

$$P(N(t) = 1) = 1 - P(N(t) = 0) - P(N(t) \geq 2)$$

Using Lemma 3.3, we can write

$$\lim_{t\to 0} \frac{1}{t} P(N(t) = 1) = \lim_{t\to 0} \left(\frac{1 - e^{-\lambda t}}{t} - \frac{P(N(t) \geq 2)}{t} \right)$$

From which the result is obtained after using Lemma 3.4. ■

Theorem 3.5. *If $N(t), t \geq 0$, is a Poisson process then, for all $t \geq 0$, and $k \in \{0, 1, 2, \cdots\}$,*

$$P(N(t) = k) = \frac{e^{-\lambda t}(\lambda t)^k}{k!},$$

where λ is as obtained in Lemma 3.3.

Remark: Thus an increment of a Poisson process over an interval of length t is Poisson distributed with parameter λt.
Proof: For $0 < \alpha < 1$, define $G(t) = \mathsf{E}\big(\alpha^{N(t)}\big)$, i.e., $G(t)$ is the moment generating function of the random variable $N(t)$. For notational simplicity, we have not retained α as an argument of the moment generating function. Now, using the stationary and independent increment property, we obtain a functional equation for $G(t)$ as follows

$$\begin{aligned} G(t+s) &= \mathsf{E}\Big(\alpha^{N(t+s)}\Big) \\ &= \mathsf{E}\Big(\alpha^{N(t)} \cdot \alpha^{(N(t+s)-N(t))}\Big) \\ &= G(t)\, G(s) \end{aligned}$$

Since $N(0) = 0$, $G(0) = 1$; since $N(t)$ increases with t, and $0 < \alpha < 1$, we also conclude that $G(t)$ is nonincreasing with t. Also, by Lemmas 3.3, 3.5, and 3.4, we can write

$$\begin{aligned} \lim_{t\to 0} G(t) &= \lim_{t\to 0} \mathsf{E}\Big(\alpha^{N(t)}\Big) \\ &= \lim_{t\to 0} \left(1 \cdot e^{-\lambda t} + \alpha \cdot (\lambda t + o(t)) + \sum_{k=2}^{\infty} \alpha^k P(N(t) = k) \right) \\ &= 1 + 0 + \lim_{t\to 0} o(t) \\ &= 1 \end{aligned}$$

establishing that $G(t)$ is continuous from the right at $t = 0$. Now using Theorem 3.22 (in the Appendix of this chapter) we conclude that the unique solution to this functional equation is

$$G(t) = e^{g(\alpha)t}$$

for some constant $g(\alpha)$. To obtain $g(\alpha)$, we observe that

$$\begin{aligned} g(\alpha) &= \lim_{t\to 0} \frac{G(t) - G(0)}{t} \\ &= \lim_{t\to 0} \left(\frac{1}{t} \left[P(N(t) = 0) - 1 \right] + \frac{\alpha \cdot P(N(t) = 1)}{t} \right. \\ &\quad \left. + \frac{1}{t} \sum_{k=2}^{\infty} \alpha^k P(N(t) = k) \right) \end{aligned}$$

which, on using Lemmas 3.4 and 3.5, yields $g(\alpha) = -\lambda + \alpha\lambda$. Thus we find that

$$G(t) = e^{-\lambda t + \lambda t \alpha}$$

i.e.,

$$\begin{aligned} G(t) &= \sum_{k=0}^{\infty} \frac{e^{-\lambda t}(\lambda t)^k}{k!} \alpha^k \\ &= \sum_{k=0}^{\infty} P(N(t) = k) \alpha^k \end{aligned}$$

It follows that, for $k \in \{0, 1, 2, \cdots\}$,

$$P(N(t) = k) = \frac{e^{-\lambda t}(\lambda t)^k}{k!}$$

■

Remarks 3.7.

a. Continuing the argument begun before Lemma 3.3, we see that the finite dimensional distributions of a Poisson process are completely characterised in terms of a single parameter λ. Thus we can now use the term: "a Poisson process with parameter λ". A little later we shall see that λ is the *rate* of the Poisson point process. We can then refer to a Poisson process with rate λ.

b. It can easily be verified that $\mathsf{E}(N(t)) = \lambda t$, and $\mathsf{Var}(N(t)) = \lambda t$. It is useful to remember that the variance to mean ratio of a Poisson process is 1. Thus the Poisson process is often used as a "benchmark" for the variability in arrival processes. An arrival process with a higher variance to mean ratio is said to be *burstier* than Poisson, and an arrival process with variance to mean ratio less than Poisson can be viewed as being smoother than Poisson.

■

Theorem 3.6. *$N(t)$ is a Poisson process with rate λ. Let $0 = t_0 < t_1 < t_2 < \cdots < t_n = t$. Then for all $k, k_1, k_2, \cdots, k_n$ nonnegative integers such that $\sum_{i=1}^{n} k_i = k$ we have*

$$P(N(t_1) = k_1, N(t_2) - N(t_1) = k_2, \cdots, N(t_n) - N(t_{n-1}) = k_n | N(t) = k) = \frac{k!}{\prod_{i=1}^{n} k_i!} \prod_{i=1}^{n} \left(\frac{t_i - t_{i-1}}{t} \right)^{k_i}$$

Remarks 3.8.

a. Note that the time points $t_i, 1 \leq i \leq n-1$, partition the interval $[0, t]$ into consecutive subintervals. The question being asked is that, given that exactly k Poisson points are known to have occurred in the interval $[0, t]$, what is the probability that k_i of them fell in the ith interval, $1 \leq i \leq n$.

b. This result states that conditioned on there being k Poisson arrivals in an interval $[0, t]$, these k points are distributed over the interval as if each one of them was independently and uniformly distributed over the interval. With this explanation, the form of the right hand side of the conclusion in the theorem becomes self evident; it is the multinomial probability distribution with k "trials" and n alternatives in each trial, the ith alternative being that a point falls in the ith subinterval, and the probability of this alternative is $\frac{t_i - t_{i-1}}{t}$.

c. Conversely, it can be seen that the Poisson process is obtained if we uniformly distributed points on $\mathbb{R}^+$ in the following way. Let us uniformly and independently distribute n points on the interval $[0, \frac{n}{\lambda}]$. Consider the interval $[0, t]$. Let n be large enough so that $t < \frac{n}{\lambda}$. Then the number of points that fall in the interval $[0, t]$ is distributed as Binomial$(n, \frac{t}{\frac{n}{\lambda}})$. As $n \to \infty$, it then follows that the distribution of the number of points in $[0, t]$ converges to Poisson(λt). Thus the Poisson process can be viewed as the limit of a uniform distribution of points (the reader should try to provide a complete proof; what remains is to show the independent increment property).

Proof: The proof is a simple matter of applying the conditional probability formula, using the stationary and independent increment property, and the Poisson distribution of each increment. The details are left as an exercise. ■

3.4.1 Stopping Times

Definition 3.3. *A random time T is called a stopping time for a Poisson process $N(t), t \geq 0$, if $\{T \leq u\} \amalg \{N(s) - N(u),\ s \geq u\}$.* ■

Remarks 3.9.

a. Thus a random time T is a stopping time for a Poisson process $N(t)$ if the question "Is $T \leq u$?" can be answered independently of the increments of $N(t)$ after the time u. Unlike the earlier Definition 2.5, here we do not require that the event $\{T \leq u\}$ be determined by $N(t), t \leq u$. Note that this property of independence of the future is all that was required in proving Wald's Lemma (Lemma 3.2).

b. From the point of view of applications, this definition of stopping time is what will be more useful. In queueing applications the Poisson process will typically model an arrival process, and various random times will be determined by the arrival process in conjunction with other random processes, such as the sequence of service requirements of customers. Thus the only reasonable requirement of a random time T would be that $\{T \leq u\}$ is independent of future arrivals.

■

Example 3.4.

Let us consider an M/G/1 queue. This is a queueing system in which customers arrive to an infinite buffer and wait for service by a single server (the trailing "1" in the notation). The customer arrival instants constitute a Poisson process (the "M" in the notation), and the service times of the customers is a sequence of i.i.d. random variables (the "G" in the notation, meaning generally distributed services times). Let $X(t), t \geq 0$, be the number of customers at time t. Let $A(t)$ denote the arrival process, and let $T_k, k \geq 1$, denote the successive arrival instants. Suppose that $X(t) = 0$, and, for a sample point ω, let $Z(\omega) := \inf\{t \geq 0 : t \geq T_1(\omega), X(t, \omega) = 0\}$; i.e., Z is the random time at which the queue becomes empty for the first time after once becoming nonempty. The random interval $[0, Z]$ is called an idle-busy period of the queue. It is easily seen that $\{Z \leq z\}$ is determined by the arrivals in the interval $(0, z]$ and the service times of these arrivals, and hence is independent of future increments of the arrival process, i.e., of $A(z + s) - A(z), s \geq 0$. ■

The following is an important property of the Poisson process in relation to stopping times.

Theorem 3.7. *Let T be a stopping time for the Poisson process $N(t)$, with $P(T < \infty) = 1$, then $N(T+s)-N(T),\ s \geq 0$, is a Poisson process independent of T and of $\{N(t),\ t \leq T\}$.*

Remark: This results asserts that if T is a proper stopping time for the Poisson process $N(t)$ then the increments of $N(t)$ starting from T form a Poisson process, and this process is independent of T as well as of the past of $N(t)$ prior to T. With reference to Example 3.4, we can now assert that if Z_1 is the end of the first idle-busy cycle then the process $X(t + Z_1), t \geq 0$, is again an M/G/1 queue length process that starts with an

empty queue. This is because, by virtue of Z_1 being a stopping time for the arrival process, $A(t+Z_1)-A(Z_1), t \geq 0$, is again a Poisson process, and the successive service times are in any case i.i.d. Thus $X(t+Z_1), t \geq 0$, is a statistical replica of $X(t), t \geq 0$. Moreover, since $A(t+Z_1)-A(Z_1), t \geq 0$, is independent of Z_1 and of $A(t), t \leq Z_1$, we conclude that $X(t+Z_1), t \geq 0$, is independent of Z_1 and of $X(t), t \leq Z_1$.

Proof: This result follows from the strong Markov property of continuous time Markov chains; see Theorem 4.3. ■

Exercise 3.2.

Given a Poisson process $N(t), t \geq 0$, show that the following assertions hold

a. The jump instants $T_k, k \geq 1$, are stopping times.

b. If T_k is a jump instant and $\epsilon > 0$, then $T_k - \epsilon$ is not a stopping time.

c. For each $t \geq 0$, $T = t$ is a stopping time.

■

Corollary 3.2. *Given a Poisson process $N(t)$ with rate λ, and T a stopping time with $P(T < \infty) = 1$,*

$$P(N(T+s) - N(T) = 0 | N(s), s \leq T) = e^{-\lambda s}$$

Proof: It follows from Theorem 3.7 that

$$P(N(T+s) - N(T) = 0 | N(s), s \leq T) = P(N(s) = 0) = e^{-\lambda s}$$

■

Remarks 3.10.

a. We saw in Exercise 3.2 that all jump times of a Poisson process are stopping times. It follows from this corollary that the successive interjump times are i.i.d. exponential with parameter λ. Hence we learn that the Poisson process is a renewal process with i.i.d. exponential life-times. This is, in fact, an alternate characterisation of a Poisson process.

b. We also saw in Exercise 3.2 that any time $t \geq 0$ is a stopping time. It follows that the residual life process $Y(t)$ of this renewal process is exponentially distributed for every $t \geq 0$; i.e., $P(Y(t) > s) = e^{-\lambda s}$ for all $t \geq 0$. Also we note here that for any renewal process the residual life process $\{Y(t),\, t \geq 0\}$ is a continuous time continuous state Markov Process. This along with the just observed fact that for a Poisson process the distribution of $Y(t)$ is invariant with t, shows that, for a Poisson process, $\{Y(t),\, t \geq 0\}$ is a stationary process. Later in this chapter will explore this notion of stationarity for more general renewal processes.

c. It should be intuitively clear that these results are a consequence of the memoryless nature of the exponential distribution; i.e., if X has distribution Exp(λ) then $P(X > x + y \mid X > x) = e^{-\lambda y}$. Hence not only is the residual life-time exponentially distributed, it is also independent of the elapsed life.

■

3.4.2 Other Characterisations

In Definition 3.2 a Poisson process was defined as a point process with stationary and independent increments, and with unit jumps, with probability 1. In applications it is often useful to have alternate equivalent characterisations of the Poisson process. The following are two such characterisations.

Theorem 3.8. *A point process $N(t), t \geq 0$, is a Poisson process if and only if*

a. for all $t_0 < t_1 < t_2 < \ldots < t_n$, the increments $N(t_i) - N(t_{i-1}), 1 \leq i \leq n$, are independent random variables, and

b. there exists $\infty > \lambda \geq 0$ such that $P(N(t+s) - N(t) = k) = \frac{(\lambda s)^k \, e^{-\lambda s}}{k!}$.

Proof: The "only if" assertion follows since it has already been shown that the original definition of the Poisson process (i.e., Definition 3.2) implies these properties (see Theorem 3.5). As for the "if part" of the proof, note that the stationary and independent increment property follows immediately. Also the Poisson distribution of the increments implies that the time between successive jumps is 0 with zero probability, thus completing the proof. ■

Theorem 3.9. *A point process $N(t)$, $t \geq 0$, is a Poisson process with parameter λ if and only if the successive jump times $T_k, k \geq 0$, are renewal instants with i.i.d. exponentially distributed inter-renewal times with mean $\frac{1}{\lambda}$.*

Remark: It follows from ERT (Theorem 3.1) that, with probability 1, $\lim_{t\to\infty} \frac{N(t)}{t} = \lambda$. Hence λ is the *rate* of the Poisson point process.

Proof: The "only if" part has already been established in Remarks 3.10. For the "if part," (i) the almost surely unit jump property follows from the exponentially distributed inter-renewal times, (ii) the stationary increment property will follow when we study stationary renewal processes in Section 3.7, and (iii) the independent increment property follows from the memoryless property of the exponentially distributed inter-renewal times. ■

3.4.3 Splitting and Superposition

Consider a Poisson process $N(t)$ with rate λ, and denote its jump instants by $\{T_1, T_2, \ldots\}$. Consider an independent Bernoulli process $Z_k, k \geq 1$,; i.e., such that Z_k are i.i.d. with

$$Z_k = \begin{cases} 1 & \text{w.p.} \quad p \\ 0 & \text{w.p.} \quad 1-p \end{cases}$$

Now define two new point processes $N^{(1)}(t)$ and $N^{(2)}(t)$ as follows: each point $T_k, 1 \leq k \leq N(t)$, is a point of $N^{(1)}(t)$ if $Z_k = 1$, else T_k is a point of $N^{(2)}(t)$. Thus each point of the Poisson process $N(t)$ is assigned to either $N^{(1)}(t)$ or to $N^{(2)}(t)$, with probability p or $1-p$, respectively, and the assignment is independent from across the points of $N(t)$. If $N(t)$ is an arrival process into a queue, it as if Z_k is used to split the process into two arrival processes.

Theorem 3.10. *$N^{(1)}(t)$ and $N^{(2)}(t)$ are independent Poisson processes with rates $p\lambda$ and $(1-p)\lambda$, respectively.*

Remark: The hypothesis of Bernoulli sampling is crucial. If there is dependence between the selection of successive points of $N(t)$ then the resulting process will not be Poisson. As an elementary counterexample, consider splitting the points of $N(t)$ so that alternate points are assigned to $N^{(1)}(t)$ and $N^{(2)}(t)$. Now we see that the interarrival instants in each of $N^{(1)}(t)$ and $N^{(2)}(t)$ are convolutions of exponentially distributed random variables (i.e., they are Erlang distributed, or gamma distributed), and hence neither $N^{(1)}(t)$ nor $N^{(2)}(t)$ can be Poisson. The Bernoulli nature of the splitting is essential for retaining the memoryless property of the life-times.

Proof: Invoking Theorem 3.8, we need to prove the following three assertions

a. $N^{(1)}(t)$ and $N^{(2)}(t)$ are independent processes.

b. $N^{(1)}(t)$ has independent increments, and so does $N^{(2)}(t)$.

c. $N^{(1)}(t)$ has Poisson distributed increments with parameter $p\lambda$, and $N^{(2)}(t)$ has Poisson distributed increments with parameter $(1-p)\lambda$.

Assertion 1 and 2 are established if we show that for $t_2 > t_1$

$$\begin{aligned} & P\left(N^{(1)}(t_2) - N^{(1)}(t_1) = k_1, N^{(2)}(t_2) - N^{(2)}(t_1) = k_2\right) \\ = \quad & \frac{(\lambda p(t_2 - t_1))^{k_1}\, e^{-\lambda p(t_2 - t_1)}}{k_1!} \cdot \\ & \frac{(\lambda(1-p)(t_2 - t_1))^{k_2}\, e^{-\lambda(1-p)(t_2 - t_1)}}{k_2!} \end{aligned}$$

This is easily seen as follows

$$
\begin{aligned}
& P\left(N^{(1)}(t_2) - N^{(1)}(t_1) = k_1,\ N^{(2)}(t_2) - N^{(2)}(t_1) = k_2\right) \\
&= P(N(t_2) - N(t_1) = k_1 + k_2) \cdot \frac{(k_1+k_2)!}{k_1!k_2!}\, p^{k_1}(1-p)^{k_2} \\
&= \frac{(\lambda(t_2-t_1))^{k_1+k_2}\, e^{-\lambda(t_2-t_1)}}{(k_1+k_2)!}\, \frac{(k_1+k_2)!}{k_1!k_2!}\, p^{k_1}(1-p)^{k_2} \\
&= \frac{(\lambda p(t_2-t_1))^{k_1}\, e^{-\lambda p(t_2-t_1)}}{k_1!} \cdot \\
&\quad \frac{(\lambda(1-p)(t_2-t_1))^{k_2}\, e^{-\lambda(1-p)(t_2-t_1)}}{k_2!}
\end{aligned}
$$

To establish the independent increment property for $N^{(1)}(t)$ (or for $N^{(2)}(t)$) (over disjoint intervals) we observe that such increments of $N(t)$ are independent, and each increment of $N^{(1)}(t)$ is obtained by independently selecting points from the increments of $N(t)$ (i.e., the selection of points from disjoint increments of $N(t)$ are done independently by virtue of the Bernoulli sampling). The elementary details are left as an exercise. ■

Let us now consider the superposition (or the merging) of two Poisson processes. Let $N^{(1)}(t)$ and $N^{(2)}(t)$ be two independent Poisson processes with rates λ_1 and λ_2. Define the point process $N(t)$ by

$$N(t) := N^{(1)}(t) + N^{(2)}(t)$$

i.e., each point of $N^{(1)}(t)$ and of $N^{(2)}(t)$ is assigned to $N(t)$.

Theorem 3.11. *$N(t)$ is a Poisson process with rate $\lambda = \lambda_1 + \lambda_2$.*

Proof: Invoking Theorem 3.8, we need to prove that increments of $N(t)$ over disjoint intervals are independent, and an increment over the interval $(t_1, t_2]$ is Poisson distributed with mean $\lambda(t_2 - t_1)$, where $\lambda = \lambda_1 + \lambda_2$. The independent increment property follows easily from the corresponding property of the Poisson processes $N^{(1)}(t)$ and $N^{(2)}(t)$, and we leave the details as an exercise. Turning to the distribution of the increments, observe

that

$$
\begin{aligned}
&P(N(t_2) - N(t_1) = k) \\
&= \sum_{i=0}^{k} P(N^{(1)}(t_2) - N^{(1)}(t_1) = i) \cdot P(N^{(2)}(t_2) - N^{(2)}(t_1) = k - i) \\
&= \sum_{i=0}^{k} \frac{(\lambda_1(t_2 - t_1))^i \, e^{-\lambda_1(t_2 - t_1)}}{i!} \cdot \frac{(\lambda_2(t_2 - t_1))^{k-i} \, e^{-\lambda_2(t_2 - t_1)}}{(k-i)!} \\
&= \frac{((\lambda_1 + \lambda_2)(t_2 - t_1))^k \, e^{-(\lambda_1 + \lambda_2)(t_2 - t_1)}}{k!} \cdot \sum_{i=0}^{k} \frac{k!}{i!(k-i)!} \left(\frac{\lambda_1}{\lambda_1 + \lambda_2}\right)^i \left(\frac{\lambda_2}{\lambda_1 + \lambda_2}\right)^{k-i} \\
&= \frac{(\lambda(t_2 - t_1))^k \, e^{-\lambda(t_2 - t_1)}}{k!}
\end{aligned}
$$

i.e., the increment of $N(t)$ over the interval $(t_1, t_2]$ is Poisson distributed with mean $\lambda(t_2 - t_1)$. ■

It is opportune to state the following result, even though it depends on the concept of a stationary renewal process which we will introduce in Section 3.7. The "only if" part of this result (which we state without proof) can be viewed as the converse to Theorem 3.11.

Theorem 3.12. *The superposition of two independent stationary renewal processes is renewal iff they are both Poisson.*

Remarks 3.11.

a. For this result to hold, the necessity of the two independent renewal processes being Poisson can be seen intuitively as follows. Consider a point in the superposition. It belongs to one of the independent processes, and the time until the next renewal in this process is independent of the past. Therefore, for the time until the next renewal in the superposition to be independent of the past, it is necessary that the residual life-time of the other renewal process is independent of the past, which requires that the life-times of the component processes are exponential.

b. It is not difficult to construct an example that demonstrates that the superposition of two dependent renewal processes can be renewal. Take an ordinary renewal process with non-exponentially distributed life times, and split it using Bernoulli splitting (as in the beginning of this section). The resulting two point processes are each a renewal process, but are dependent.

■

3.5 Regenerative Processes

Let us recall Example 3.4 and the remark following Theorem 3.7. We had the queue length process, $X(t)$, of an M/G/1 system with $X(0) = 0$, and we saw that there exists a random time Z_1 such that $X(t + Z_1), t \geq 0$, (i.e., the evolution of the queue length process after Z_1) (i) is *statistically* identical to the evolution of the process $X(t), t \geq 0$, and (ii) is independent of Z_1 and the past of the queue length process up to Z_1. We say that the process $X(t)$ regenerates at Z_1, and call such a process a regenerative process.

Definition 3.4. *A process $X(t), t \geq 0$, is called a* regenerative process *if there exists a stopping time T_1 such that*

a. $X(t + T_1),\ t \geq 0$, and $X(t),\ t \geq 0$, have the same probability law (i.e., are statistically identical), and

b. $\{X(t + T_1),\ t \geq 0\}$ II $\{T_1, \text{and } X(u),\ u < T_1\}$.

■

Example 3.5.

a. Consider a DTMC $X_n, n \geq 0$, with $X_0 = j$, then the time of first return to j is a stopping time. By the strong Markov property, we see that X_n is a regenerative process.

b. As we saw in Example 3.4 and the remark following Theorem 3.7, the queue length process $X(t)$ of an M/G/1 queue (that starts off empty) is a regenerative process.

Remarks 3.12.

a. It is important to note that the property that $X(t + T_1), t \geq 0$, is independent of T_1 is a rather special property. In general, the evolution after a stopping time need not be independent of the stopping time. As an exercise, the reader is encouraged to construct an example that illustrates this lack of independence, in general.

b. With reference to Definition 3.4, we say that $X(t)$ *regenerates* at T_1. Since $X(t + T_1), t \geq 0$, is statistically identical to $X(t), t \geq 0$, there must be a stopping time $T_2 \geq T_1$, such that the process again regenerates at T_2. By the fact that the successive regenerations are independent, it follows that the sequence of regeneration points $T_1, T_2, T_3, \cdots$, are the renewal points in a renewal process. We call this the *embedded* renewal process. Also, the successive intervals into which these renewals divide time are called *regeneration cycles*.

c. In general, we have what are called *delayed regenerative processes*. For such a process, there exists a stopping time T_1 such that Property 2 in Definition 3.4 holds but $X(t+T_1), t \geq 0$, has a statistical behaviour different from $X(t), t \geq 0$. However, $X(t+T_1), t \geq 0$, is a regenerative process, as defined in Definition 3.4. Thus it as if the regenerations in the process $X(t)$ have been delayed by the time T_1. We observe now that the sequence of random times $T_1, T_2, \cdots$, are the points of a renewal process whose first life-time has a different distribution from the subsequent life-times. As an example, consider a DTMC $X_k, k \geq 0$, with $X_0 = i$, and consider visits to state $j \neq i$. The visits to j constitute regeneration times, but clearly this is a delayed regenerative process. Thus the process defined in Definition 3.4 can be called an *ordinary regenerative process*.

■

3.5.1 Time Averages of a Regenerative Process

Let us consider a regenerative process $X(t), t \geq 0$. We are interested in evaluating limits of the following type

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t X(u)du$$

i.e., the time average of the process $X(t)$, or, for $b \in \mathbb{R}$,

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t I_{\{X(u)\leq b\}} du$$

i.e., the fraction of time the process value is in the interval $(-\infty, b]$, or

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t P\{X(u) \leq b\} du$$

i.e., the limit of the expectation of the fraction of time the process value is in the interval $(-\infty, b]$. Considering the first limit, we can proceed by thinking of $X(t)$ as a "reward" rate at time t. Then the cumulative reward until t is

$$C(t) = \int_0^t X(u)du$$

and we need, the time average reward rate, i.e.,

$$\lim_{t\to\infty} \frac{C(t)}{t}$$

As before, denote the regeneration times of $X(t)$ by $T_1, T_2, \cdots$, and let $T_0 = 0$. Define the reward in the jth cycle, $j \geq 1$, by

$$R_j = \int_{T_{j-1}}^{T_j} X(u)du$$

By the properties of the regenerative process, the sequence of random variable $R_j, j \geq 1$, are mutually independent, and for an ordinary regenerative process this random sequence is i.i.d. Thus, along with the renewal instants $T_k, k \geq 1$, we are now in the renewal-reward frame work. Let us consider the case of an ordinary regenerative process. If $\mathsf{E}\Big(\int_0^{T_1} |X(u)|du\Big) < \infty$, and $\mathsf{E}(T_1) < \infty$, Theorem 3.2 immediately applies and we conclude that, w.p. 1,

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t X(u)du = \frac{\mathsf{E}\Big(\int_0^{T_1} X(u)du\Big)}{\mathsf{E}(T_1)}$$

and, w.p. 1,

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t I_{\{X(u)\leq b\}}du = \frac{\mathsf{E}\Big(\int_0^{T_1} I_{\{X(u)\leq b\}}du\Big)}{\mathsf{E}(T_1)}$$

and

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t P(X(u) \leq b)du = \frac{\mathsf{E}\Big(\int_0^{T_1} I_{\{X(u)\leq b\}}du\Big)}{\mathsf{E}(T_1)}$$

Note that, in order to make the last two assertions, we only need that $\mathsf{E}(T_1) < \infty$, since

$$\mathsf{E}\left(\int_0^{T_1} I_{\{X(u)\leq b\}}du\right) \leq \mathsf{E}\left(\int_0^{T_1} du\right) = \mathsf{E}(T_1) < \infty$$

Thus for the limiting time average distribution to exist it suffices that the cycle times of the regenerative process have finite mean.

Let us observe that $\frac{\mathsf{E}\Big(\int_0^{T_1} I_{\{X(u)\leq b\}}du\Big)}{\mathsf{E}(T_1)}$ is a cumulative distribution function in the argument b, for it is nonnegative, nondecreasing in b, and the limit as $b \to \infty$ is 1. Note also that in the expression $\mathsf{E}\Big(\int_0^{T_1} I_{\{X(u)\leq b\}}\Big)\, du$ we cannot bring the expectation inside the integral, since the upper limit of integration is a random variable, i.e., T_1.

It is usually easier to obtain such limiting distributions. For example, for a positive recurrent DTMC we know that (recall the proof of Theorem 2.9)

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} p_{ij}^{(k)} \to \pi_j$$

Can we use this result to obtain the limiting time average of the process, i.e., $\lim_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n-1} X_k$? The following result asserts that, under certain conditions, the limiting time average of the process can be obtained by taking the expectation of the limiting time average distribution.

Theorem 3.13. *Consider a regenerative process $X(t)$ with $\mathsf{E}(T_1) < \infty$ and let X_∞ denote a random variable with the limiting time average distribution, i.e., $P(X_\infty \leq b) = \frac{\mathsf{E}\left(\int_0^{T_1} I_{\{X(u)\leq b\}} du\right)}{\mathsf{E}(T_1)}$.*

a. If $\mathsf{E}\left(\int_0^{T_1} |X(u)|du\right) < \infty$ then

$$\lim_{t\to\infty} \frac{1}{t}\int_0^t X(u)du \stackrel{a.s.}{=} \mathsf{E}(X_\infty)$$

b. If $X(u) \geq 0$ then

$$\lim_{t\to\infty} \frac{1}{t}\int_0^t X(u)du \stackrel{a.s.}{=} \mathsf{E}(X_\infty)$$

and this limit could be finite or infinite.

Remark: Thus, continuing the example that we were discussing just before the theorem, we can conclude that for a positive recurrent DTMC $X_k \in \mathcal{S}$ (that takes nonnegative values), w.p. 1,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} X_k = \sum_{j\in\mathcal{S}} j\pi_j$$

where π_j is the invariant distribution of the DTMC.

Proof: We omit the proof of this theorem. ■

3.6 The Renewal Equation

We have so far only examined the limiting time averages of various processes and functions. For example, for a renewal process we obtained the limiting time average of the renewal function $m(t)$, or of the distribution of the residual life process $P(Y(t) > y)$. We now study how to characterise the transient behaviour of these processes and functions, and hence to obtain their limiting behaviour. For example, if we can show that for a renewal process

$$\lim_{t\to\infty} P(Y(t) > y) = G^c(y)$$

where $G(\cdot)$ is a continuous distribution, then we have shown the convergence in distribution of the residual life process $Y(t)$.

Example 3.6.

Consider a renewal process with i.i.d. life-times with c.d.f. $F(\cdot)$. Define, for $y \geq 0$,

$$K(t) = P(Y(t) > y)$$

Now we can break up this event into three disjoint parts depending on whether the first renewal occurs in either of the three intervals $(t+y, \infty)$, $(t, t+y]$, $[0, t]$. This yields

$$\begin{aligned} K(t) &= P(Y(t) > y,\ X_1 > t+y) + P(Y(t) > y,\ t < X_1 \leq t+y) \\ &\quad + P(Y(t) > y,\ X_1 \leq t) \\ &= (1 - F(t+y)) + 0 + \int_0^t K(t-x)dF(x) \end{aligned}$$

where, in the second equality, the first term is $(1-F(t+y))$ because the event $\{X_1 > t+y\}$ implies the event $\{Y(t) > y\}$, and the second term is 0 because $\{t < X_1 \leq t+y\}$ implies that $Y(t) \leq y$. In the third term we integrate over all the possible points in $[0, t]$ at which the first renewal can occur. With the first renewal having occurred at x, the time remaining until t is $t - x$, and hence the desired probability is $K(t-x)$. This argument is called a *renewal argument*. Thus we find that, for fixed $y \geq 0$, the function $K(t)$ satisfies the equation

$$K(t) = F^c(t+y) + \int_0^t K(t-x)dF(x) \tag{3.5}$$

or, equivalently,

$$K(t) = a(t) + (K \star F)(t)$$

where $a(t) = F^c(t+y)$, and, as before, $\star$ denotes the Riemann-Stieltjes convolution.

■

The equation for $K(t)$ obtained in the previous example is of the form

$$H(t) = a(t) + \int_0^t H(t-x)dF(x)$$

or

$$H(t) = a(t) + (H \star F)(t)$$

where $a(\cdot)$ is a given function and $F(\cdot)$ is a cumulative distribution function. Such an equation is called a *renewal equation*[3]. More compactly, taking the time argument as being understood, we also write

$$H = a + H \star F$$

The following result provides a formal solution for the renewal equation, under some fairly general conditions.

[3]In general, an equation in which the unknown function appears inside an integral is called an *integral equation*, and, in particular, the renewal equation is called a *Volterra integral equation of the second kind.*

Theorem 3.14. *If the c.d.f. $F(\cdot)$ has positive mean (i.e., $\int_0^\infty (1-F(x))dx > 0$), and if $a(t)$ is a bounded function, then the unique solution of $H = a + H \star F$ that is bounded on finite intervals is*

$$H(t) = a(t) + \int_0^t a(t-u)dm(u)$$

i.e.,

$$H = a + a \star m$$

where $m(t)$ is the renewal function for an ordinary renewal process with life-time c.d.f. $F(\cdot)$, i.e., $m(t) = \sum_{k=1}^{\infty} F^{(k)}(t)$.

Proof: We need to show that the proposed solution

(i) is indeed a solution,

(ii) is bounded on finite intervals, and

(iii) is unique.

(i) To check that the proposed solution satisfies the renewal equation, we substitute it in the renewal equation to obtain

$$\begin{aligned} H &= a + (a + a \star m) \star F \\ &= a + (a \star F + a \star (m \star F)) \\ &= a + a \star (F + m \star F) \\ &= a + a \star m \end{aligned}$$

where we have used the easy observation that $m = F + m \star F$. Thus the proposed solution satisfies the renewal equation.
(ii) To verify that the proposed solution is bounded on finite intervals consider, for a finite $T > 0$,

$$\begin{aligned} \sup_{0 \le t \le T} |a(t) + (a \star m)(t)| &\le \sup_{0 \le t \le T} |a(t)| + \sup_{0 \le t \le T} \left| \int_0^t a(t-u)dm(u) \right| \\ &\le \sup_{0 \le t \le T} |a(t)| + \int_0^T \sup_{0 \le y \le T} |a(y)| dm(u) \\ &= \left(\sup_{0 \le t \le T} |a(t)| \right) (1 + m(T)) < \infty \end{aligned}$$

where the finiteness follows since $m(t)$ is bounded (by $\int_0^\infty (1 - F(x))dx > 0$ and Lemma 3.1), and $a(t)$ is given to be bounded.

(iii) To establish uniqueness, assume to the contrary that H_1 and H_2 are two solutions bounded on finite intervals and let

$$D = H_1 - H_2$$

Hence D is also bounded on finite intervals, then

$$\begin{aligned} D \star F &= H_1 \star F - H_2 \star F \\ &= (H_1 - a) - (H_2 - a) \\ &= D \end{aligned}$$

where the second equality is obtained since H_1 and H_2 both are solutions to $H = a + H \star F$. It follows, by recursion, that, for all $n \geq 1$,

$$D = D \star F^{(n)}$$

Hence for all $n \geq 1$

$$\begin{aligned} |D(t)| &= \left| \int_0^t D(t-x) dF^{(n)}(x) \right| \\ &\leq \int_0^t |D(t-x)| dF^{(n)}(x) \\ &\leq \left(\sup_{u \in [0,t]} |D(u)| \right) F^{(n)}(t) \end{aligned}$$

Now the first term in the last expression is finite, since D is bounded on finite intervals. Also, for each fixed t, $F^{(n)}(t) \to 0$ as $n \to \infty$ since $\int_0^\infty (1 - F(x)) dx > 0$ (see the proof of Lemma 3.1). It follows that, for all t,

$$|D(t)| = 0$$

We conclude that $H_1 = H_2$, i.e., the proposed solution is unique. ■

Remarks 3.13.

a. Let us now apply Theorem 3.14 to solve the renewal Equation 3.5 for $K(t) = P(Y(t) > y)$. This yields

$$K(t) = (1 - F(t+y)) + \int_0^t (1 - F(t+y-x)) \, dm(x)$$

b. There is an intuitive way to think about the solution of the renewal equation provided we appropriately interpret $dm(x)$. To this end, consider the following formal "calculation."

$$\begin{aligned} P(\text{a renewal occurs in the interval } (x, x+dx)) &= \sum_{k=1}^{\infty} P(Z_k \in (x, x+dx)) \\ &= \sum_{k=1}^{\infty} dF^{(k)}(x) \\ &= d(\sum_{k=1}^{\infty} F^{(k)}(x)) \\ &= dm(x) \end{aligned}$$

Thus we can interpret $dm(x)$ as the probability that there is a renewal in the interval $(x, x+dx)$. For this reason $\frac{dm(x)}{dx}$ is called the *renewal density*. As an example, consider the Poisson process, which we now know is a renewal process with exponentially distributed life-times. We also know that for a Poisson process of rate λ the renewal function $m(t) = \lambda t$. Hence the renewal density is λ. Now Lemma 3.5 confirms our interpretation of $dm(x)$ in the case of the Poisson process. With this interpretation of $dm(x)$ let us now view the second term in the solution of the renewal equation as conditioning on the *last* renewal in $[0, t]$. This will yield

$$\begin{aligned} K(t) &= (1 - F(t+y)) \\ &\quad + \int_0^t P(\text{a renewal occurs in } (x, x+dx), \\ &\qquad \text{the next renewal occurs after } t, \\ &\qquad \text{and the residual life-time at } t \text{ is greater than } y) \\ &= (1 - F(t+y)) + \int_0^t (1 - F((t-x)+y))\, dm(x) \end{aligned}$$

■

Let us now consider a delayed renewal process.

Example 3.7.

Let us redo Example 3.6 for a delayed renewal process with the c.d.f. of X_1 being $A(\cdot)$ and that of $X_j, j \geq 2$, being $F(\cdot)$. Again, for fixed $y \geq 0$, define $K(t) = P(Y(t) > y)$, where $Y(t)$ is the residual life process. In an identical fashion as in Example 3.6, we can write

$$K(t) = (1 - P(X_1 \leq t+y)) + \int_0^t K_o(t-x)\, dA(x)$$

where $K_o(t) = P(Y(t) > y)$ for the ordinary renewal process with life-time distribution $F(\cdot)$. Thus we have

$$K(t) = A^c(t+y) + (K_0 \star A)(t)$$

where, in turn, $K_o(t)$ satisfies

$$K_o(t) = F^c(t+y) + (K_0 \star F)(t)$$

■

Thus, in situations involving a delayed renewal process, we obtain a pair of equations of the form

$$H = a + H_0 \star A$$

in conjunction with

$$H_o = a_o + H_o \star F$$

Of course, we know that the equation for $H_o(\cdot)$ has the solution

$$H_o = a_o + a_o \star m_0$$

where

$$m_o = \sum_{k=1}^{\infty} F^{(k)}$$

The solution of the delayed renewal equation is given in the following result, and can be intuitively appreciated based on the interpretation of $\frac{dm(t)}{dt}$ as the renewal density for the delayed renewal process.

Theorem 3.15. *The solution to the general renewal equation is*

$$H = a + a_0 \star m$$

where

$$m = \sum_{k=0}^{\infty} A \star F^{(k)}$$

Proof: It is easy to check that the proposed solution indeed satisfies the delayed renewal equation. We omit the proof of uniqueness. ■

3.7 Stationary Renewal Process

Obviously, a renewal counting process $M(t)$ cannot be stationary since it is increasing with time. What then is meant by a stationary renewal process? Consider a general renewal process with X_j distributed as $F(\cdot)$ for $j \geq 2$, and X_1 distributed as $F_e(\cdot)$ where $F_e(t) =$

$\frac{1}{\mathsf{E}(X_2)} \int_0^t (1 - F(x))dx$. Recall that $F_e(\cdot)$ is the time average equilibrium distribution for a renewal process with i.i.d. life-times with distribution $F(\cdot)$. Denote the renewal function for this delayed renewal process by $m_e(t)$.

Let us apply the result in Theorem 3.15 to calculate $P(Y(t) > y)$ for this general renewal process (for some fixed $y > 0$). We obtain

$$P(Y(t) > y) = F_e^c(t + y) + \int_0^t F^c(t + y - x)\, dm_e(x)$$

Hence we need to determine $m_e(t)$. This is easily done by writing a renewal equation for $m_e(\cdot)$. We observe that the following holds

$$m_e(t) \quad = \quad \mathsf{E}(M_e(t)) = \int_0^t (1 + m_o(t - x))dA(x)$$

This can be understood as follows. If the first arrival occurs in (t, ∞) then the mean number of arrivals in $[0, t]$ is 0. Hence we condition on the first renewal in $[0, t]$. We count 1 for this renewal, and then from here on we have an ordinary renewal process. If the first renewal occurs at x, the remaining mean number of renewals in $[x, t]$ is $m_o(t - x)$. Thus we can write, compactly,

$$m_e(t) = A(t) + (m_o \star A)(t)$$

Taking Laplace Stieltjes Transforms (LSTs) across this equation (see Section 1.2.1), we obtain,

$$\begin{aligned} \tilde{m}_e(s) &= \tilde{A}(s) + \tilde{m}_o(s) \cdot \tilde{A}(s) \\ &= \tilde{A}(s) + \frac{\tilde{F}(s)}{1 - \tilde{F}(s)} \cdot \tilde{A}(s) \\ &= \frac{\tilde{A}(s)}{1 - \tilde{F}(s)} \end{aligned}$$

where the second term in the second equality is obtained by applying the first equality to $m_o(\cdot)$. Further, it can be seen by taking the LST that

$$\tilde{A}(s) = \frac{1 - \tilde{F}(s)}{s\mathsf{E}(X_2)}$$

Substituting, we obtain

$$\tilde{m}_e(s) = \frac{1}{s\mathsf{E}(X_2)}$$

which, on inversion, yields

$$m_e(t) = \frac{t}{\mathsf{E}(X_2)}$$

Remark: This result should be compared with what we know for the Poisson process of rate λ. The Poisson process is a renewal process with exponentially distributed life-times with c.d.f. $F(x) = 1 - e^{-\lambda x}$. It can be verified that $F_e(x) = 1 - e^{-\lambda x}$. Thus the Poisson process automatically satisfies the assumption in the previous derivation. Hence, applying the conclusion of the previous derivation, the renewal function will be $\frac{t}{\mathsf{E}(X_2)} = \lambda t$, which we already know to be true.

Let us now return to obtaining $P(Y(t) > y)$ for the renewal process that we had constructed above. Substituting $m_e(\cdot)$ into Equation 3.6, we obtain (via self evident manipulations)

$$\begin{aligned}
P(Y(t) > y) &= F_e^c(t+y) + \int_0^t F^c(t+y-x)\frac{dx}{\mathsf{E}(X_2)} \\
&= F_e^c(t+y) + \frac{1}{\mathsf{E}(X_2)} \int_y^{t+y} F^c(u)du \\
&= F_e^c(t+y) + \frac{1}{\mathsf{E}(X_2)} \left(\int_0^{t+y} F^c(u)du - \int_0^y F^c(u)du \right) \\
&= F_e^c(t+y) + F_e(t+y) - F_e(y) \\
&= 1 - F_e(y) = F_e^c(y)
\end{aligned}$$

Thus for the renewal process that we constructed, for all $t \geq 0$,

$$P(Y(t) > y) = F_e^c(y)$$

i.e., the marginal distribution of the residual life process is invariant with time. It is easy to see that $Y(t)$ is a Markov process. It follows that, with the marginal distribution being stationary, the $Y(t)$ process is itself a stationary process. It is for this reason that the process we have constructed is called a *stationary* renewal process.

Now, denoting the counting process for this renewal process by $M_e(\cdot)$, let us consider, for $x > 0$, and $t > 0$, $M_e(x)$ and $M_e(t+x) - M_e(t)$. Considering t as the starting time, since the distribution of $Y(t)$ is the same as that of $Y(0)$, we have another renewal process statistically identical to the one that started at time 0. Hence we conclude that $M_e(x)$ and $M_e(t+x) - M_e(t)$ have the same distributions, and thus $M_e(t)$ has *stationary increments*.
Remark: It is important to note, however, that $M_e(t)$ need not have independent increments. As an example, suppose that $X_j, j \geq 2$, are deterministic random variables taking the value T with probability 1. Then $F_e(t)$ is the uniform distribution over $[0, T]$, and $M_e(t)$ is a stationary increment process. But we notice that in $[0, T]$ there is exactly

one renewal. Hence $M(T/2)$ and $M(T) - M(T/2)$ are not independent, even though they have the same distribution.

Definition 3.5. *A point process with stationary and independent increments is called a* compound *or* batch *Poisson process.*

Remark: Basically a batch Poisson arrival process comprises a Poisson process, at the points of which i.i.d. batches of arrivals occur. Each batch has at least one arrival. If the batch size is exactly one then we are back to a Poisson process. Thus, in summary, a general stationary renewal process has stationary increments. If the property of independent increments is added, we obtain a batch Poisson point process. If, in addition, each batch is exactly of size 1, we have a Poisson point process.

3.8 From Time Averages to Limits

Until this point in our discussions, we have focused on obtaining long run time averages of processes and certain time varying quantities associated with these processes. Results such as the elementary renewal theorem, or the renewal reward theorem were the tools we used in establishing the existence of and the forms of such limits. In this section we turn to the important question of studying the limits of the quantities themselves. The approach is by obtaining a renewal equation for the desired quantity, solving this renewal equation and then taking the limit as $t \to \infty$ in the solution. The Key Renewal Theorem is an important tool in this approach.

The following example shows a typical situation in which the time average may exist, with the limit of the associated quantity failing to exist.

Example 3.8.

Consider an ordinary renewal process with life-times $X_i, i \geq 1$, such that $X_i = T$ for all $i \geq 1$, where $T > 0$ is given. It is elementary to see that, given $y \geq 0$,

$$P(Y(t) \leq y) = \begin{cases} 0 & (Tk \leq t < T(k+1) - y \;\; k \geq 0 \\ 1 & \text{otherwise} \end{cases}$$

Hence $P(Y(t) \leq y)$ is an "oscillatory" function of time and does not converge with t. However, as we already know, for $0 \leq y \leq T$,

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t P(Y(t) \leq y) dy = \frac{y}{T}$$

the uniform distribution over $[0, T]$. ■

The life-time in this example has a property that is captured in the following definition.

Definition 3.6. *A random variable X, such that $P(|X| < \infty) = 1$, is called* lattice *if there exists $d > 0$ such that*

$$\sum_{n=-\infty}^{\infty} P(X = nd) = 1$$

otherwise X is called nonlattice. The largest d with this property is called the span *of X. We also say that the distribution of X is lattice.* ■

Notice that in the example above the renewal life-time was lattice with $P(X_1 = T) = 1$.

We state the following result without proof. This result helps in the understanding of the Key Renewal Theorem which we will state and discuss next.

Theorem 3.16 (Blackwell's Renewal Theorem (BRT))**.** *(i) For a general (or delayed) renewal process with arbitrary A and nonlattice F, and any $h > 0$.*

$$\lim_{t\to\infty} [m(t+h) - m(t)] = h\mu$$

where $\frac{1}{\mu} = \int_0^\infty (1 - F(x))dx$.

(ii) For an ordinary renewal process, i.e., $A = F$, with lattice F with span d, for $h = d, 2d, 3d, \cdots$,

$$\lim_{t\to\infty} [m(t+h) - m(t)] = \mu h$$

Remarks 3.14.

a. We first note that part (ii) of the theorem deals with the case of a lattice F, and in this case the result is restricted to the situation in which $A = F$. This is because without such a restriction on the "initial" life-time, different conclusions can be obtained depending on the choice of A. For example, if $A = F_e$ then (recall from Section 3.7) we have a stationary renewal process and $m(t) = \mu t$ for all t, thus yielding $m(t+h) - m(t) = \mu h$ for every t and $h \geq 0$.

b. We also note that, in the same sense as the convergence of a sequence of numbers is stronger than the convergence of the averages of the sequence, BRT is stronger than the expectation version of ERT. This can be seen by the following simple argument that shows that BRT implies ERT. Consider the nonlattice case of BRT (part (i), above). We observe that, for every n, with $m(0) = 0$,

$$m(n) = \sum_{k=0}^{n-1} (m(k+1) - m(k))$$

Now, by BRT, $m(k+1) - m(k) \to \mu$, as $k \to \infty$. Hence, the average of this sequence also converges, i.e.,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} (m(k+1) - m(k)) = \mu$$

i.e.,

$$\lim_{n\to\infty} \frac{m(n)}{n} = \mu$$

which establishes ERT for the case when we pass to the limit via integer values of time. But we also see that

$$\frac{\lfloor t \rfloor}{t} \frac{m(\lfloor t \rfloor)}{\lfloor t \rfloor} \leq \frac{m(t)}{t} \leq \frac{m(\lfloor t \rfloor + 1)}{\lfloor t \rfloor + 1} \frac{\lfloor t \rfloor + 1}{t}$$

Since the upper and lower bounds both converge to μ, we conclude that $\lim_{t\to\infty} \frac{m(t)}{t} = \mu$, which is the expectation version of ERT.

■

We have seen that the solution of the renewal equation

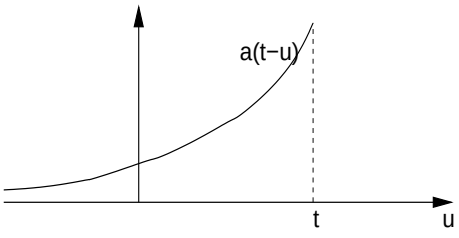
$$H = a + H \star F$$

is

$$H = a + a \star m$$

where $m(t)$ is the renewal function of the ordinary renewal process with life-time distribution F. Our aim next is to understand how to use this solution to obtain $\lim_{t\to\infty} H(t)$. It is usually straightforward to obtain $\lim_{t\to\infty} a(t)$; recall that, in the examples we have seen, $a(t)$ is defined by the tail of a distribution function and hence goes to 0 as $t \to \infty$. The problem remains of obtaining $\lim_{t\to\infty}(a \star m)(t)$. So let us consider

$$(a \star m)(t) = \int_0^t a(t-u)\, dm(u)$$



The diagram on the left shows the way a convolution is performed between a function $a(\cdot)$, which is 0 for negative arguments, and the renewal function $m(t)$. For fixed $t > 0$, the diagram shows $a(t-u)$ as a function of u. Multiplying $a(t-u)$ with $dm(u)$ for $u \geq 0$ and integrating up to t yields the value of the convolution at t.

Now suppose that for large u, $dm(u) \approx \mu du$, as would be suggested by BRT. Suppose, for a moment that $a(\cdot)$ is 0 for large arguments. Then, for large enough t, over the range of integration, the approximation $dm(u) \approx \mu du$ will be better and better as t increases. To see this, look at the picture and note that if $a(\cdot)$ is 0 for large arguments, for large enough t all of the nonzero part of it will be "pushed" into the positive quadrant, and larger values of

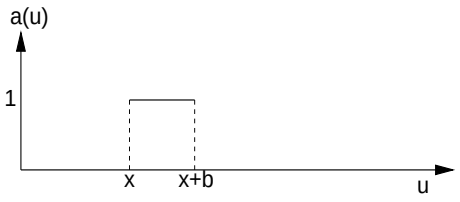
t will push this nonzero part further to the right, thus multiplying it with $dm(u)$ for larger and larger u. Thus, we can write, for large t,

$$\begin{aligned}(a \star m)(t) &\approx \int_0^t a(t-u)\,\mu du \\ &= \mu \int_0^t a(x)dx\end{aligned}$$

This suggests the following result

$$\lim_{t\to\infty} (a \star m)(t) = \mu \int_0^\infty a(x)dx$$

In general, we would like to permit $a(\cdot)$ to be nonzero over all $\mathbb{R}^+$. The Key Renewal Theorem provides an important technical condition that constrains the behaviour of $a(t)$ for large t, and permits the above limiting result to hold. We develop an intuition into this condition before stating the theorem.



Consider the function $a(u)$ shown on the left, i.e.,

$$a(u) = \begin{cases} 1 & \text{if} \quad x \le u \le x+b \\ 0 & \text{otherwise} \end{cases}$$

Consider $\int_0^t a(t-u)dm(u)$. As discussed earlier, this converges to $\mu \int_0^\infty a(u)du = \mu b$. Note that this is also a consequence of BRT (Theorem 3.16).

Next we consider a bounded function, $a(t)$, defined for $t \ge 0$. Define the following function, for all $i \ge 1$,

$$\underline{a}_i(u) = \begin{cases} \inf_{x\in[(i-1)b,ib)} a(x) & \text{for } u \in [(i-1)b, ib) \\ 0 & \text{otherwise} \end{cases}$$

and consider

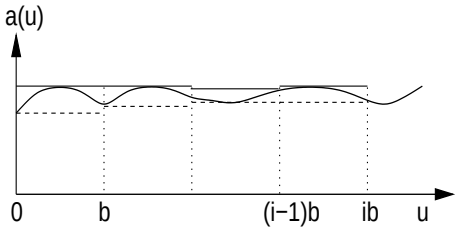
$$\underline{a}(u) = \sum_{i=1}^\infty \underline{a}_i(u)$$

Similarly, define, for all $i \ge 1$,

$$\overline{a}_i(u) = \begin{cases} \sup_{x\in[(i-1)b,ib)} a(x) & \text{for } u \in [(i-1)b, ib) \\ 0 & \text{otherwise} \end{cases}$$

and consider

$$\overline{a}(u) = \sum_{i=1}^{\infty} \overline{a}_i(u)$$



We show on the left a general function $a(t)$. The positive x-axis is partitioned into equal intervals of width b. The piece-wise flat function drawn with thin solid lines is $\overline{a}(\cdot)$ and the piece-wise flat function drawn with thin dashed lines is $\underline{a}(\cdot)$.

By the definitions of $\underline{a}(u)$ and $\overline{a}(u)$, it is clear that, for all t,

$$\int_0^t \underline{a}(t-u)dm(u) \leq \int_0^t a(t-u)dm(u) \leq \int_0^t \overline{a}(t-u)dm(u)$$

Thus, if it can be shown that the upper and lower bounds in this expression converge to the same quantity, as $t \to \infty$, then that will be $\lim_{t\to\infty} \int_0^t a(t-u)dm(u)$.

Now we can expect the following to hold

$$\lim_{t\to\infty} \int_0^t \underline{a}(t-u)dm(u) = \lim_{t\to\infty} \int_0^t \sum_{i=1}^{\infty} \underline{a}_i(t-u)dm(u) = \mu b \sum_{i=1}^{\infty} \inf_{x\in[(i-b),\, ib)} a(x)$$

and, similarly,

$$\lim_{t\to\infty} \int_0^t \overline{a}(t-u)dm(u) = \lim_{t\to\infty} \int_0^t \sum_{i=1}^{\infty} \overline{a}_i(t-u)dm(u) = \mu b \sum_{i=1}^{\infty} \sup_{x\in[(i-1)b,\, ib)} a(x)$$

Notice that the expressions $b \sum_{i=1}^{\infty} \inf_{x\in[(i-b),\, ib)} a(x)$ and $b \sum_{i=1}^{\infty} \sup_{x\in[(i-1)b,\, ib)} a(x)$ are like lower and upper Riemann partial sums, except that they are taken over all $[0, \infty)$. Now letting $b \to 0$, if the upper and lower partial sums over $[0, \infty)$ converge to the same number then that must be equal to $\int_0^\infty a(u)du$ the Riemann integral, and a is said to be Directly Riemann Integrable (DRI). Thus if $a(\cdot)$ is DRI, the above argument suggests that $\lim_{t\to\infty} \int_0^t a(t-u)dm(u) = \mu \int_0^\infty a(u)du$.

Remarks 3.15.

a. Note that in standard Riemann integration we obtain $\int_0^\infty a(u)du = \lim_{t\to\infty} \int_0^t a(u)du$, where $\int_0^t a(u)du$ is the Riemann integral over $[0, t]$, obtained by taking partial sums

over a partition of $[0, t]$ and then letting the partition width (say, b) go to 0. Thus standard Riemann integrability over $[0, \infty)$ requires

$$\lim_{t\to\infty}\lim_{b\to 0} b\sum_{i=1}^{\lceil t/b\rceil}\inf_{x\in[(i-b),\,ib)} a(x) = \lim_{t\to\infty}\lim_{b\to 0} b\sum_{i=1}^{\lceil t/b\rceil}\sup_{x\in[(i-b),\,ib)} a(x)$$

On the other hand in direct Riemann integration over $[0, \infty)$ we set up the partial sums by partitioning all of $[0, \infty)$, and then letting the partition width $b \to 0$. Thus direct Riemann integrability over $[0, \infty)$ requires

$$\lim_{b\to 0}\lim_{t\to\infty} b\sum_{i=1}^{\lceil t/b\rceil}\inf_{x\in[(i-b),\,ib)} a(x) = \lim_{b\to 0}\lim_{t\to\infty} b\sum_{i=1}^{\lceil t/b\rceil}\sup_{x\in[(i-b),\,ib)} a(x)$$

Note that the two requirements differ in the order in which the limits with respect to b and t are taken.

b. It can be shown that if $a(u)$ is DRI over $[0, \infty)$ then it is Riemann integrable but a Riemann integrable function over $[0, \infty)$ need not be DRI. Here is an example of a Riemann integrable function that is not DRI. Define $a(u) = 1$ for $u \in [n - \frac{1}{2n^2}, n + \frac{1}{2n^2}]$, for $n = 1, 2, 3, \cdots$, and $a(u) = 0$ for all other $u \in \mathbb{R}^+$. Thus, the graph of $a(u)$ comprises "pulses" of height 1 and width $\frac{1}{n^2}$ centred at the positive integers. The Riemann integral of $a(u)$ exists since $\lim_{N\to\infty}\sum_{n=1}^{N}\frac{1}{n^2} < \infty$. However, $a(u)$ is not DRI since, for each b, there exists N_b, such that for all $n > N_b$, $\frac{1}{n^2} < b$, and hence the upper Riemann sum over $[0, \infty)$ will have an infinite number of 1s contributed by the pulses centered at $n > N_b$ and is, therefore, ∞ for all b.

■

In general, the DRI property may not be easy to recognise, but the following result is useful in many applications. We state the result without proof.

Lemma 3.6. *If a function $a : \mathbb{R}^+ \to \mathbb{R}$ is (i) nonnegative, (ii) monotone nonincreasing, and (iii) Riemann integrable, then it is directly Riemann integrable.*

■

Finally, we state the following important result without proof.

Theorem 3.17 (Key Renewal Theorem (KRT))**.** *Let A and F be proper distribution functions with $F(0^+) < 1$ and $\int_0^\infty (1 - F(u))du = \frac{1}{\mu}$ with $\mu = 0$ if the integral is infinite. Suppose $a(\cdot)$ is directly Riemann integrable. Then the following hold.*

(i) *If F is non lattice then*

$$\lim_{t\to\infty}\int_0^t a(t-u)dm(u) = \mu\int_0^\infty a(u)du$$

(ii) *If $A = F$, and if F is lattice with span d, then for all $t > 0$*

$$\lim_{n\to\infty} \int_0^{t+nd} a(t+nd-u)dm(u) = \mu d \sum_{n=0}^{\infty} a(t+nd)$$

■

Example 3.9.

Consider a delayed renewal process with initial life-time distribution $A(\cdot)$ and the subsequent life-time distribution being $F(\cdot)$ with mean $\mathsf{E}(X_2)$. We recall that, for given $y \geq 0$, the residual life distribution is given by

$$P(Y(t) > y) = A^c(t+y) + \int_0^t F^c(t+y-x)dm(x)$$

where $m(\cdot)$ is the renewal function. Suppose F is nonlattice and $A(\cdot)$ is a proper distribution. It follows from KRT that

$$\begin{aligned} \lim_{t\to\infty} P(Y(t) > y) &= 0 + \frac{1}{\mathsf{E}(X_2)} \int_0^{\infty} F^c(u+y)du \\ &= \frac{1}{\mathsf{E}(X_2)} \int_y^{\infty} (1-F(x))dx \\ &= F_e^c(y) \end{aligned}$$

which, as would be expected, is the same as the time average result. Notice that this result states that the random process $Y(t)$ converges in distribution to a random variable with distribution F_e. ■

3.9 Limits for Regenerative Processes

What we have learnt above can be applied to obtain a fairly general condition for the convergence in distribution of regenerative processes. Consider a delayed regenerative process $B(t),\ t \geq 0$. We are interested in $\lim_{t\to\infty} P(B(t) \leq b)$. There is a random time X_1 at which the process regenerates. Let the law of the process obtained after the regeneration be denoted by $P_o(\cdot)$, where the subscript "o" relates to the term "ordinary." To appreciate this notation, notice that we can write $A(t) = P(X_1 \leq t)$, and $F(t) = P_o(X_1 \leq t)$; thus for this delayed regenerative process $A(\cdot)$ is the distribution of the first cycle time and $F(\cdot)$ is the distribution of the cycle times after the first regeneration.

It is now easy to see that the following renewal equation is obtained for any $b \in \mathbb{R}$

$$P(B(t) \leq b) \;=\; P(B(t) \leq b,\ X_1 > t) + \int_0^t P_o(B(t-x) \leq b)\ dA(x)$$

where $A(\cdot)$ is the distribution of the time until the first regeneration instant. Let us denote $a(t) = P(B(t) \leq b, X_1 > t)$, and $a_o(t) = P_o(B(t) \leq b, X_1 > t)$. Then it can be seen that the renewal equation has the following solution

$$P(B(t) \leq b) = a(t) + \int_0^t a_o(t-x)dm(x)$$

where $m(t)$ is the renewal function of the delayed renewal process induced by the regenerative process. Now, if $a_o(t)$ is directly Riemann integrable, and $F(\cdot)$ is nonlattice then, by Theorem 3.17,

$$\lim_{t\to\infty} \int_0^t a_o(t-x)dm(x) = \mu \int_0^\infty a_o(u)du$$

where $\frac{1}{\mu} = \int_0^\infty (1 - F(x))dx$. Further, if $A(\cdot)$ is proper,

$$\lim_{t\to\infty} a(t) = 0$$

Therefore, in such situation, we can conclude that

$$\lim_{t\to\infty} P(B(t) \leq b) = \mu \int_0^\infty P_o(B(u) \leq b,\ X_1 > u)du$$

Hence, if $\mu > 0$ we find that the regenerative process converges in distribution to a proper distribution (why is it proper?). It remains to obtain a condition that ensures that $a_o(\cdot)$ is DRI. We need the following definition

Definition 3.7. *a. A real valued function is said to belong to $D[0,\infty)$ if it is right continuous and has left hand limits.*

b. A stochastic process $B(t),\ t \geq 0$, is said to belong to $\mathcal{D}$ (written $B(t) \in \mathcal{D}$) if $P\{w : B(t,w) \in D[0,\infty)\} = 1$

■

The following is the main result, which we state without proof. We state only the nonlattice version.

Theorem 3.18. *For a generalised regenerative process $\{B(t), t \geq 0\}$, with finite mean cycle length $\mathsf{E}(X_i) < \infty,\ i \geq 2$, and with $A(\infty) = 1,\ B(t), t \geq 0$, converges in distribution to*

$$\mu \int_0^\infty P_o(B(t) \leq b,\ X_1 > t)dt$$

if $B(t) \in \mathcal{D}$ and $F(\cdot)$ is nonlattice. ■

Remarks 3.16.

a. The above theorem provides a very simple but powerful tool for proving the stability of a stochastic process that can be shown to be regenerative. Under a fairly general condition, from the point of view of applications, it suffices to show that the mean cycle time is finite.

b. Let us look at the form of the limiting distribution and observe that

$$\int_0^\infty I_{\{B(t,w)\le b,\ X_1(w)>t\}}dt = \int_0^{X_1(w)} I_{\{B(t,w)\le b\}}dt$$

Taking expectation on both sides with respect to the probability law $P_o(\cdot)$, we obtain

$$\int_0^\infty P_o(B(t)\le b,\ X>t)dt = \mathsf{E}_o\left(\int_0^{X_1} I_{\{B(t)\le b\}}dt\right)$$

where $\mathsf{E}_o(\cdot)$ denotes expectation with respect to $P_o(\cdot)$. Then multiplying with μ on both sides we find that the limit provided by Theorem 3.18 is consistent with that obtained in Section 3.5.1.

3.10 Some Topics in Markov Chains

We now turn to applying some of the results in this chapter to DTMCs.

3.10.1 Relative Rate of Visits

Consider a positive recurrent irreducible Markov Chain $X_k, k \ge 0$, taking values in the discrete set $\mathcal{S}$. Suppose $X_0 = i$, and consider any other state j. Consider first the instants of visits to state j. The Markov chain is now a delayed regenerative process with respect to these instants. By the elementary renewal theorem, with probability 1,

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n-1} I_{\{X_k=j\}} = \frac{1}{\nu_j}$$

where $\nu_j = \sum_{n=1}^\infty n f_{jj}^{(n)}$, the mean recurrence time of j. Also, since the DTMC is positive recurrent, denoting by $\pi_j, j\in\mathcal{S}$, its stationary probability vector, we know that $\pi_j = \frac{1}{\nu_j}, j\in\mathcal{S}$. Consider now visits to i. These visits constitute an ordinary renewal process. Let the reward in each renewal interval be the number of visits to j. Then, by the renewal reward theorem, with probability 1,

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n-1} I_{\{X_k=j\}} = \frac{\mathsf{E}(V_{ij})}{\nu_i}$$

where V_{ij} is the random variable of the number of visits to j between visits to i. We have obtained the same limit in two forms, and these must be equal. Thus we obtain

$$\frac{\mathsf{E}(V_{ij})}{\nu_i} = \frac{1}{\nu_j}$$

or,

$$\mathsf{E}(V_{ij}) = \frac{\pi_j}{\pi_i}$$

Thus we conclude that in a positive recurrent DTMC the mean number of visits to state j between consecutive visits to the state i is the ratio of the stationary probabilities of these two states. This result is, however, not limited to positive recurrent DTMCs but holds more generally under the condition of recurrence alone. We state the following result which we shall have occasion to use in the theory of continuous time Markov chains.

Theorem 3.19. *For a recurrent irreducible Markov chain with transition probability matrix* $\mathbf{P}$*, there exist solutions to* $\mathbf{u} = \mathbf{uP}$*,* $\mathbf{u} > 0$*. The vector space of such solutions has dimension 1. Thus, for all* i *and* j*, the ratios* $\frac{u_j}{u_i}$ *are uniquely determined and*

$$\mathsf{E}(V_{ij}) = \frac{u_j}{u_i}$$

■

Remark: Note that if the recurrent DTMC is positive then the solutions of $\mathbf{u} = \mathbf{uP}$, $\mathbf{u} > 0$, will also be *summable* thus yielding a probability measure on $\mathcal{S}$. In general, however, for a recurrent DTMC the solutions of $\mathbf{u} = \mathbf{uP}$, $\mathbf{u} > 0$, will not be summable, but the ratios of their components, and hence the relative frequency of visits to states, are uniquely determined.

3.10.2 Limits of DTMCs

Consider again an irreducible recurrent DTMC $X_n, n \geq 0, X_n \in \mathcal{S}$. Let us view the DTMC as evolving over multiples of a time-step, say 1. In this view point, all times between visits to various states are lattice random variables. For example, letting T_j denote the random variable for the time to return to j, we observe that $P(T_j = n) = f_{jj}^n$, and hence

$$f_{jj} = \sum_{n=1}^{\infty} P(T_j = n) = \sum_{n=1}^{\infty} f_{jj}^{(n)} = 1$$

since we have assumed a recurrent DTMC. Thus T_j is a lattice random variable. Let d_j be the span of T_j. If $d_j > 1$ then we can say that j is periodic, as first returns to j only occur at a number of steps that is a multiple of $d_j > 1$. In general, for $f_{jj} \leq 1$, we have the following definition.

Definition 3.8. *For a state j of a Markov Chain if $d_j := g.c.d.\{n : f_{jj}^{(n)} > 0\} > 1$ then j is called periodic, otherwise j is called aperiodic.* ■

The following lemma is an easy exercise that is left for the reader to prove.

Lemma 3.7.

$$g.c.d.\{n : f_{jj}^{(n)} > 0\} = g.c.d.\{n : p_{jj}^{(n)} > 0\}$$

Proof: Exercise. Hint: note that each element in the set $\{n : p_{jj}^{(n)} > 0\}$ is a sum of one or more elements of the set $\{n : f_{jj}^{(n)} > 0\}$. ■

In the following result we learn that the period of a state is also a class property.

Theorem 3.20. *All states in a communicating class have the same period or are all aperiodic.*

Proof: Consider two states j and k in a communicating class. Then there exist $r > 0$ and $s > 0$ such that $p_{jk}^{(r)} > 0$ and $p_{kj}^{(s)} > 0$. Hence

$$p_{jj}^{(r+s)} > 0 \text{ and } p_{kk}^{(r+s)} > 0$$

Hence, reading $m|n$ as "m divides n," and using Lemma 3.7, we have

$$d_j|(r+s) \text{ and } d_k|(r+s)$$

Further, for all $n \geq 1$,

$$p_{jj}^{(r+n+s)} \geq p_{jk}^{(r)}\, p_{kk}^{(n)}\, p_{kj}^{(s)}$$

Hence, for all $n \geq 1$,

$$p_{kk}^{(n)} > 0 \text{ implies } p_{jj}^{(r+n+s)} > 0$$

Now consider

$$\begin{aligned} N_j &:= \{n :\, p_{jj}^{(n)} > 0\} \\ N_k &:= \{n :\, p_{kk}^{(n)} > 0\} \\ N_k' &:= \{r+n+s :\, n \in N_k\} \end{aligned}$$

It follows that $N_j \supset N_k'$. Now, since $d_j = \text{g.c.d. } N_j$, we conclude that $d_j|N_k'$, where we read $m|A$ as "m divides every element of the set A." But we also know that $d_j|(r+s)$. Then it must be true that $d_j|N_k$. Hence $d_j \leq d_k$. Similarly, $d_k \leq d_j$. Hence $d_j = d_k$. ■

Remark: We infer from this theorem and Lemma 3.7 that if $p_{jj} > 0$ for some j in a communicating class $\mathcal{C}$, then $d_i = 1$ for all $i \in \mathcal{C}$.

Theorem 3.21. *For a recurrent irreducible Markov chain*

a. For all j, $\lim_{n\to\infty} p_{jj}^{(nd)} = \frac{d}{\nu_j}$, where d is the period.

b. If the Markov chain is aperiodic then $\lim_{n\to\infty} p_{ij}^{(n)} = \frac{1}{\nu_j}$

Remark: When the period $d > 1$, then ν_j is a multiple of d. The first part of the result says that starting in j at time 0, if we look at instants that are multiples of d then the probability of finding the process in j at such instants converges to $\frac{d}{\nu_j}$. This makes sense if we think of $\frac{\nu_j}{d}$ as the time to return to j in "units" of the period d. The second part of the result says that, in the aperiodic case, irrespective of the initial state the probability of being in j converges to $\frac{1}{\nu_j}$. Of course, in the null recurrent case, all these limits are 0, as expected.
Proof:

(i) Consider $X_0 = j$ and then visits to j form a renewal process whose life-times have a lattice distribution with span d. Observe that

$$\begin{aligned} p_{jj}^{(nd)} &= \mathsf{E}\big(I_{\{X_{nd}=j\}}|X_0 = j\big) \\ &= \mathsf{E}\left(\left(\sum_{k=1}^{nd} I_{\{X_k=j\}} - \sum_{k=1}^{nd-1} I_{\{X_k=j\}}\right)|X_0 = j\right) \\ &= \mathsf{E}\left(\left(\sum_{k=1}^{nd} I_{\{X_k=j\}} - \sum_{k=1}^{(n-1)d} I_{\{X_k=j\}}\right)|X_0 = j\right) \\ &= m_o(nd) - m_o((n-1)d) \end{aligned}$$

where the second equality is obtained since $I_{\{X_k=j\}} = 0$ for $k = (n-1)d+1, (n-1)d+2, \cdots, (n-1)d+(d-1)$, since j has period d, and $X_0 = j$. Applying the lattice part of Blackwell's renewal theorem (Theorem 3.16, Part (ii)), we obtain

$$\lim_{n\to\infty} p_{jj}^{(nd)} = \frac{d}{\nu_j}$$

(ii) By conditioning on the first visit to j

$$p_{ij}^{(n)} = \sum_{k=1}^{n} f_{ij}^{(k)}\, p_{jj}^{(n-k)}$$

Taking limits on both sides

$$\lim_{n\to\infty} p_{ij}^{(n)} = \lim_{n\to\infty} \sum_{k=1}^{n} f_{ij}^{(k)}\, p_{jj}^{(n-k)}$$

Define

$$Z_n = \begin{cases} p_{jj}^{(n-k)} & \text{w.p.} \quad f_{ij}^{(k)} \text{ if } 1 \le k \le n \\ 0 & \qquad\quad \text{if } k > n \end{cases}$$

Hence, by part (i) of this theorem, $Z_n \to Z \left(= \frac{1}{\nu_j}\right)$ a constant. Further $0 \leq Z_n \leq 1$ for all n. Hence, using the bounded convergence theorem (see Theorem 1.7)

$$\mathsf{E}(Z_n) \left(= \sum_{k=1}^{n} f_{ij}^{(k)} \, p_{jj}^{(n-k)} \right) \to \mathsf{E}(Z) = \frac{1}{\nu_j}$$

■

Remarks 3.17.

a. Consider a DTMC $X_k, k \geq 0$, with $X_k \in \{0, 1\}$, such that $X_0 = 0$, $p_{0,1} = 1$, and $p_{1,0} = 1$. Now this is a positive recurrent DTMC with stationary probability $\pi_0 = 0.5 = \pi_1$, and $\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} p_{01}^{(k)} = 0.5$. On the other hand, we see that $p_{01}^{(k)} = 0$ for k even, and $p_{01}^{(k)} = 1$ for k odd. Hence $p_{01}^{(k)}$ itself does not converge. We see that the period of both the states is 2, and $p_{00}^{(2k)} = 1$ which is consistent with the first part of the above theorem.

b. We have earlier shown (see Section 3.2.1) that for any recurrent discrete time Markov chain

$$\frac{1}{n} \sum_{k=1}^{n} p_{ij}^{(k)} = \frac{1}{\nu_j}$$

which matches the limit in the aperiodic case.

■

3.11 Appendix

Theorem 3.22. *$f(x), x \geq 0$, is right continuous at $x = 0$, and nonincreasing in x, and for each $x \geq 0, y \geq 0$, $f(x + y) = f(x)f(y)$. Then, either $f(x) = 0, x \geq 0$, or there exists $\lambda, 0 \leq \lambda < \infty$, such that $f(x) = e^{-\lambda x}$, for $x \geq 0$.*

Remark: The problem is one of finding the solution to the *functional equation* $f(x+y) = f(x)f(y)$, given some additional restrictions on $f(\cdot)$. The functional equation $g(x + y) = g(x) + g(y)$ is called the *Cauchy* functional equation.
Proof: Setting $x = y = 0$ in the functional equation for $f(\cdot)$, we see that

$$f(0) = (f(0))^2$$

from which it follows that either $f(0) = 0$ or $f(0) = 1$. If $f(0) = 0$, then for every $x > 0$, $f(x) = f(0 + x) = f(0)f(x)$, implying that $f(x) = 0$, for all $x \geq 0$. The alternative

is that $f(0) = 1$. Suppose now that $f(z) = 0$, for some $z > 0$. Then, for all $y \geq 0$, $f(z + y) = f(z)f(y) = 0$. We conclude that $f(x) = 0$ for $x \geq z$. Hence, define

$$b = \sup\{x : f(x) > 0\}$$

Suppose that $0 < b < \infty$. Consider $0 < x < b, 0 < y < b$, with $x + y > b$. Using the functional equation for $f(\cdot)$, and the definition of b, $0 = f(x + y) = f(x)f(y)$, implying that either $f(x) = 0$ or $f(y) = 0$, contradicting the definition of b. Hence, either $b = 0$ or $b = \infty$. If $b = 0$, we have $f(0) = 1$ and $f(x) = 0, x > 0$, contradicting the right continuity of $f(\cdot)$ at $x = 0$. Hence, with $f(0) = 1$, the only possibility consistent with the hypotheses is that $b = \infty$, or $f(x) > 0, x > 0$.

With $f(0) = 1$, and $f(x) > 0$ for $x \geq 0$, we can take the natural logarithm on both sides of our functional equation, yielding

$$\ln f(x + y) = \ln f(x) + \ln f(y)$$

Define, for all $x \geq 0$, $g(x) = -\ln f(x)$. By the hypothesis that $f(\cdot)$ is nonincreasing, $1 \geq f(x) > 0$ for $x \geq 0$, hence we have $0 \leq g(x) < \infty$, for $x \geq 0$. Recalling the Remark above, we see that $g(\cdot)$ satisfies the Cauchy functional equation. Since $f(x)$ is nonincreasing in x, $g(x)$ is nondecreasing in x. Now, for every $x \geq 0$, and m a nonnegative integer, we observe that

$$g(mx) = mg(x) \tag{3.6}$$

This also implies that $g(0) = 0$ (we get a contradiction by taking $g(0) \neq 0$). Now consider $x = \frac{m}{n}z$, where m and n are nonnegative integers, and $x \geq 0, z \geq 0$. It follows, using Equation 3.6, that

$$ng(x) = mg(z)$$

i.e., $g(\frac{m}{n}z) = \frac{m}{n}g(z)$, for all $z \geq 0$. Define $\lambda := g(1)$. Then $\lambda \geq 0$; also, $\lambda < \infty$ since, as observed earlier, $g(x) < \infty$ for $x \geq 0$. We conclude that, for any rational $r \geq 0$,

$$g(r) = \lambda r$$

Now consider any $x > 0$, and determine rational sequences $u_n, v_n, n \geq 1$, such that $u_1 \leq u_2 \leq u_3 \leq \cdots \leq u_n \uparrow x$, and $v_1 \geq v_2 \geq v_3 \geq \cdots \geq v_n \downarrow x$. Using the monotonicity of $g(\cdot)$, we observe that, for every $n \geq 1$,

$$\lambda u_n = g(u_n) \leq g(x) \leq g(v_n) = \lambda v_n$$

Letting $n \to \infty$, we conclude that, for $x \geq 0$, $g(x) = \lambda x$ ($g(0) = 0$ having been observed earlier). Finally, we have, for $x \geq 0$,

$$f(x) = e^{-\lambda x}$$

■

3.12 Notes on the Bibliography

This chapter has been developed mainly from Wolff [17], with Çinlar [5] and Karlin and Taylor [10] having been used as references. The tiny volume by Cox ([6]) remains a classic on the topic of this chapter.

3.13 Problems

3.1. $\{X_i, i \geq 1\}$ is a sequence of i.i.d. random variables with $P(X_i = 1) = 1/2$, and $P(X_i = -1) = 1/2$. $\{S_n, n \geq 0\}$ is a process defined as follows.

$$S_0 = -1$$

and for $n \geq 1$,

$$S_n = S_0 + \sum_{i=1}^{n} X_i$$

Let

$$N = \min\{j \geq 1 : \ S_j = 1\}$$

i.e., N is the hitting time of the state 1.

a. Show that N is a stopping time for the process $\{S_n, n \geq 0\}$.

b. Clearly,

$$S_N = -1 + \sum_{i=1}^{N} X_i$$

Show that an anomaly arises if we "apply" Wald's Lemma to get $\mathsf{E}(S_N)$. Explain the anomaly. (Hint: a hypothesis of Wald's Lemma fails.)

3.2. Consider a discrete time ordinary renewal process with life time distribution $p_k, \ k \geq 0$. Let m_n denote the discrete time renewal function. Show that $m_n = \sum_{k=1}^{n} r_k$ where r_k is the probability that a renewal occurs at k.

3.3. Consider an ordinary renewal process with life time distribution $F(\cdot)$. Let $U(t)$ and $Y(t)$ denote the age and residual life processes. Show that, for given $u \geq 0, \ y \geq 0$,

$$\lim_{t\to\infty} P(U(t) > u, \ Y(t) > y) = \frac{1}{\mathsf{E}(X_1)} \int_{u+y}^{\infty} (1 - F(x))dx$$

3.4. Customers arrive in a Poisson process, $N(t), t \geq 0$, with rate λ to two queues Q1 and Q2. The first customer is assigned to Q1, the next is assigned to Q2, and so on the customers are assigned alternately to the two queues.

a. Show that the arrival process of customers to Q1 and Q2 are renewal processes, and find their life-time distributions.

b. Show that these renewal processes are *not* independent.

3.5. $\{N_1(t)\}$ and $\{N_2(t)\}$ are two independent Poisson processes; $\{N_1(t) + N_2(t)\}$ is their superposition. Consider an observer at $T_n^{(1)}$, the nth epoch of $\{N_1(t)\}$.

a. Obtain the distribution of the time until the next epoch of $\{N_2(t)\}$.

b. Obtain the probability that the next epoch of $\{N_1(t) + N_2(t)\}$ is an epoch in $\{N_1(t)\}$.

c. Obtain the distribution of the time until the next epoch of $\{N_1(t) + N_2(t)\}$.

d. Obtain the mean number of epochs of $\{N_1(t)\}$ that the observer sees before the next epoch of $\{N_2(t)\}$.

3.6. By example show that the superposition of two dependent Poisson processes need not be Poisson.

3.7. Consider a DTMC $\{X_k\}$, $X_k \in \{0, 1, 2, \cdots\}$, with $p_{01} = p_{12} = \frac{1}{2} = p_{00} = p_{10}$, and, for $i \geq 2$, $p_{i(i+1)} = \frac{i-1}{i+1} = 1 - p_{i0}$.

(i) Show that $\{X_n\}$ is positive recurrent.

(ii) Obtain the stationary measure of $\{X_n\}$, and hence show that, with probability 1,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n} X_k = \infty$$

3.8. $\{X_1, X_2, X_3, \cdots\}$ are i.i.d. random variables with $X_i \sim \text{Exp}(\lambda)$, where $\lambda > 0$. $\{Z_1, Z_2, Z_3, \cdots\}$ are i.i.d. integer valued random variables (also independent of $\{X_1, X_2, X_3, \cdots\}$) with $P(Z_i = 1) = p = 1 - P(Z_i = 0)$, where $0 < p < 1$. Define, for $k \geq 1$, $T_k = \sum_{i=1}^{k} X_i$.

a. Define $N(t) = \sup\{k : T_k \leq t\}$. For $0 < t_1 < t_2$, and k_1, k_2 nonnegative integers, write down $P(N(t_1) = k_1, N(t_2) - N(t_1) = k_2)$. No derivation is required, but justify your answer by stating theorem(s).

b. Define $M(t) = \sum_{i=1}^{N(t)} Z_i$.

   i. Derive $P(M(t) = k)$, and provide an interpretation for your answer.

   ii. For $0 < t_1 < t_2$, write down $P(M(t_1) = k_1, M(t_2) - M(t_1) = k_2)$, and justify the answer.

3.9. A person arrives at a pedestrian crossing line on a road. Assume that c seconds are required to cross the road, that the crossing is a zero width line, that the pedestrian can judge the speed of the cars, and that he/she waits until the time to the next vehicle crossing the line is $> c$ seconds.

a. First consider a one way, single lane road. The instants at which vehicles cross the line constitute a Poisson process of rate λ.

   i. Write down an expression for the distribution of the time until the first vehicle crosses the road after the arrival of the pedestrian.

   ii. Obtain the probability that the pedestrian waits for k cars before he crosses the road.

b. Repeat (a) for a two way road with independent Poisson rate λ traffic in each direction, with the pedestrian requiring $2c$ seconds to cross the road (and the road must be crossed without waiting at the lane separating line!).

3.10. Given $\lambda > 0$, a point process on $[0, \infty)$ is constructed as follows: n points are uniformly and independently placed on $\left[0, \frac{n}{\lambda}\right]$, and then n is taken to ∞. Denote the resulting point process by $N(t)$.

a. Derive $P(N(t) = k)$ for $t \geq 0$.

b. Show that, for $t_1 < t_1 + t < t_2 < t_2 + u$, $N(t_1 + t) - N(t_1)$ and $N(t_2 + u) - N(t_2)$ are independent. (Hint: write the joint moment generation function and show it factors, where for a random vector (X_1, X_2) the joint moment generation function is $\mathsf{E}\left(z_1^{X_1} z_2^{X_2}\right)$)

c. What can you conclude about the process $N(t)$?

3.11. Consider a renewal process with i.i.d. life times $\{X_i,\ i \geq 1\}$ with c.d.f. F. Let $\{U(t), t \geq 0\}$ denote the age process, i.e., $U(t) = t - Z_{M(t)}$.

a. Show using the renewal reward theorem that, $\forall u \geq 0$,

$$\lim_{t \to \infty} \frac{1}{t} \int_o^t P(U(\tau) \leq u) d\tau = \frac{1}{\mathsf{E}(X_1)} \int_0^u (1 - F(x)) dx$$

b. Show by formulating a renewal equation for $U(t)$ that

$$\lim_{t \to \infty} P(U(t) > u) = \frac{1}{\mathsf{E}(X_1)} \int_u^\infty (1 - F(x)) dx$$

3.12. $\{X_n, n \geq 0\}$ is a DTMC.

a. Find the distribution of the number of steps for which $\{X_n\}$ stays in a state $j \in S$.

b. Write down a (discrete time) renewal type equation for $K_{jj}(n) = P\left(X_n = j | X_0 = j\right)$.

3.13. $\{V_i,\ i \geq 1\}$ and $\{W_i,\ i \geq 1\}$ are independent sequences of non-negative i.i.d. random variables with distributions $H(\cdot)$ and $G(\cdot)$. Intervals of lengths $\{V_i\}$ and $\{W_i\}$ are placed alternately on the positive real line starting from the origin in the order $(V_1, W_1, V_2, W_2, V_3, W_3, \cdots)$. Let $X_i = V_i + W_i$, and observe that we have an ordinary renewal process embedded at the beginning epochs of the V_i intervals. Define the process $Z(t)$ such that, $Z(t) = 1$ if t is in a V_i interval and 0 otherwise. Obtain expressions for

a. $$\lim_{t\to\infty} \frac{1}{t} \int_0^t I_{\{Z(u)=1\}}\, du$$

b. $$\lim_{t\to\infty} P(Z(t) = 1)$$

Show all your analytical steps, and state your assumptions.

Chapter 4

Continuous Time Markov Chains

In Chapter 2 we studied DTMCs, which are discrete time processes, $X_n, n \geq 0$, that take values in a discrete state space $\mathcal{S}$ and have the Markov property. A DTMC models the evolution of a discrete valued random process in discrete time. In a sense we had a model for the evolution of the *state* of a system, without any notion of the time spent in a state. In this chapter we will study continuous time Markov chains (CTMCs), which are processes that have the Markov property, take values in a discrete state space (hence the term "chains"), but evolve in continuous time. For CTMCs, the time spent in a state will be seen to be exponentially distributed, which is essential for their Markov property.

Definition 4.1. *A process* $\{X(t), t \geq 0\}$, $X(t) \in \mathcal{S}$, *a discrete set of* states*, satisfies the Markov property if, for all* $t \geq 0, s \geq 0$*, for all* $j \in \mathcal{S}$

$$P(X(t+s) = j | X(u), u \leq s) \quad = \quad P(X(t+s) = j | X(s))$$

Such a process is called a continuous time Markov chain. ■

Remark: The idea of the Markov property is exactly the same as in the case of DTMCs: given the past and the present, the future is independent of the past. Unconditionally, however, the future and the past are not independent, in general.

Example 4.1.

Let $N(t), t \geq 0$, be a Poisson (counting) process of rate λ. $N(t)$ takes values in $\{0, 1, 2, \cdots\}$. Further, we have, for $t \geq 0, s \geq 0$, and fixing the value of $N(s) = i$ $(\leq j)$,

$$\begin{aligned} & P(N(t+s) = j | N(u), u \leq s, N(s) = i) \\ &= P(N(t+s) - N(s) = j - i | N(u), u \leq s, N(s) = i) \\ &= P(N(t+s) - N(s) = j - i) \\ &= P(N(t+s) = j | N(s) = i) \end{aligned}$$

where the second equality follows by the independent increment property of the Poisson process. Thus we see that the process $N(t), t \geq 0$, is a CTMC. ■

4.1 Transition Probability Function

Let $X(t), t \geq 0$, be a CTMC. As in the case of DTMCs, we shall assume the property of *time homogeneity*, i.e., $P(X(t+s) = j | X(s) = i)$ does not depend on s, and is denoted by $p_{ij}(t)$. The transition probability matrix over time is denoted by $\mathbf{P}(t)$ and has elements $p_{ij}(t), i, j \in \mathcal{S}$. We can think of this as a family of matrices indexed by $t \geq 0$, or as a matrix valued function with a time argument. As in the case of DTMCs, we observe that

a. For each t, $\mathbf{P}(t)$ is a stochastic matrix. Thus, for all i, j, and $t > 0$, $p_{ij}(t) \geq 0$, and for all $i \in \mathcal{S}$, $\sum_{j \in \mathcal{S}} p_{ij}(t) = 1$. We define $p_{ij}(0) = 0$ if $i \neq j$, and $p_{jj}(0) = 1$, i.e., $\mathbf{P}(0) = \mathbf{I}$.

b. For all $t \geq 0, s \geq 0$, $\mathbf{P}(t+s) = \mathbf{P}(t)\mathbf{P}(s)$. These are called the Chapman-Kolmogorov equations; the derivation is exactly the same as that for DTMCs in Theorem 2.1.

In addition to defining $\mathbf{P}(0) = \mathbf{I}$, let us assume that $\lim_{t \downarrow 0} \mathbf{P}(t) = \mathbf{I}$; i.e., the transition probability function $\mathbf{P}(t), t \geq 0$, is assumed to be right continous at 0. The Chapman-Kolmorov equations can then be used to show that the transition probability function is continuous for all $t > 0$. The right continuity at any $t > 0$ can be seen immediately by writing $\mathbf{P}(t+h) - \mathbf{P}(t) = \mathbf{P}(t)(\mathbf{P}(h) - \mathbf{I})$; the assumed continuity at 0 implies right continuity at any $t > 0$.

Note that, unlike DTMCs, there is no "one step" transition probability matrix that determines the $\mathbf{P}(t)$ for all other t. The following results are obtained exactly as in the case of DTMCs and are presented without proof.

Theorem 4.1. *The following hold for a CTMC $X(t), t \geq 0$.*

(i) For all $0 < t_1 < \cdots < t_n$, and all $i_0, i_1, i_2 \cdots \in \mathcal{S}$.

$$\begin{aligned} & P(X(t_1) = i_1, X(t_2) = i_2, \cdots, X(t_n) = i_n | X(0) = i_0) \\ = \;& p_{i_0 i_1}(t_1) p_{i_1 i_2}(t_2 - t_1) \cdots p_{i_{n-1} i_n}(t_n - t_{n-1}) \end{aligned}$$

(ii) If $P(X(0) = i) = \pi(i)$ then

$$\begin{aligned} & P(X(t_1) = i_1, X(t_2) = i_2, \cdots, X(t_n) = i_n) \\ = \;& \sum_{i_0 \in \mathcal{S}} \pi(i_0)\, p_{i_0 i_1}(t_1) p_{i_1 i_2}(t_2 - t_1) \cdots p_{i_{n-1} i_n}(t_n - t_{n-1}) \end{aligned}$$

■

Remark: Thus, in order to specify the finite dimensional distributions of a CTMC, in general, we need the initial distribution and the transition probability function $\mathbf{P}(t)$, for all $t \geq 0$. Note the contrast with a DTMC, where the stochastic process is completely specified by the initial distribution and the one step transition probability matrix. The characterisation can be simplified for *regular* CTMCs, as we shall see in Section 4.4.

4.2 Sojourn Time in a State

Just as in the case of the Poisson process, the simple independence assumptions (i.e., the Markov property) result in some very specific results. Let us view a CTMC, $X(t), t \geq 0$, as moving randomly from state to state, spending random amounts of time in each state. For all $t \geq 0$, define

$$Y(t) = \inf\{s > 0 : \ X(t+s) \neq X(t)\}$$

i.e., $Y(t)$ is the remaining time that the process spends in the state that it is in at time t.

Theorem 4.2. *For all $i \in \mathcal{S}$ and $t \geq 0, u \geq 0$,*

$$P(Y(t) > u | X(t) = i) = e^{-a_i u}$$

for some real number $a_i \in [0, \infty]$.

Remark: This result states that given that $X(t)$ is in the state i at time t the remaining time in i is exponentially distributed with a parameter that depends only on i. Intuitively, it can be seen that this result is necessary for the Markov property to hold. The remaining time should depend only on the current state, and also should not depend on how much time has already been spent in the state. The latter observation points to the memoryless property of the exponential distribution.

Proof: Define

$$g_i(u) = P(Y(t) > u | X(t) = i)$$

where the lack of dependence on t in the term on the left hand side is because of the property of time homogeneity. We can now write

$$\begin{aligned}
g_i(u+v) &= P(X(s) = i, t \leq s \leq t+u+v | X(t) = i) \\
&= P(X(s) = i, t \leq s \leq t+u; \ X(s) = i, \ t+u < s \leq t+u+v | X(t) = i) \\
&= P(X(s) = i, s \in [t, t+u] \mid X(t) = i) \cdot \\
&\quad P(X(s) = i, s \in (t+u, t+u+v] \mid X(t) = i; X(s) = i, s \in [t, t+u]) \\
&= g_i(u) P(X(s) = i, s \in [t+u, t+u+v] \mid X(t+u) = i) \\
&= g_i(u) g_i(v)
\end{aligned}$$

where we used the Markov property in the fourth equality, and time homogeneity in the fifth equality. By definition, $g_i(\cdot)$ is a complementary c.d.f., hence is right continuous and nonincreasing. It follows, from Theorem 3.22, that $g_i(\cdot)$ must be of the form

$$g_i(u) = e^{-a_i u}$$

for some $0 \leq a_i < \infty$, or $g_i(u) = 0$, for all $u \geq 0$. For compact notation, we denote the latter case by $g_i(u) = e^{-a_i u}$, with $a_i = \infty$. ■

Definition 4.2. *Given a CTMC $X(t)$, a state i is called*

(i) *absorbing if $a_i = 0$,*

(ii) *stable if $0 < a_i < \infty$, and*

(iii) *instantaneous if $a_i = \infty$.*

Remark: We will make the simplifying assumption that $X(t)$ has *no instantaneous states*, i.e., we will focus on *pure jump* CTMCs that evolve by moving randomly from state to state, and, with probability 1, spending positive amounts of time in each state.

4.3 Structure of a Pure Jump CTMC

The content of the following definition is identical to that of Definition 2.5.

Definition 4.3. *Given a CTMC $X(t)$, a random variable $T : \Omega \to [0, \infty]$ is a* stopping time *for $X(t)$, if for every $t \geq 0$, $I_{\{T \leq t\}} = f(X(u), u \leq t)$, where $f(\cdot)$ takes values in $\{0, 1\}$.* ■

Remark: If T_n is the nth jump instant of a CTMC, then T_n is a stopping time. For any given t, $T = t$ is a stopping time.

Definition 4.4. *A Markov Chain is said to be* strong Markov *if, for any stopping time T, for all $i_1, i_2, \cdots, i_n \in \mathcal{S}$, and $0 < t_1 < t_2 < \cdots < t_n$,*

$$P(X(T + t_1) = i_1, \ldots, X(T + t_n) = i_n \mid X(u), u \leq T) = p_{X(T)i_1}(t_1)p_{i_1 i_2}(t_2 - t_1) \cdots p_{i_{n-1} i_n}(t_n - t_{n-1}) \tag{4.1}$$

over $\{T < \infty\}$. ■

Remark: The qualification "over $\{T < \infty\}$" requires explanation. In the left hand side of (4.1), the conditional probability is a random variable. The assertion is that this random variable takes the value on the right hand side in those realisations in which $\{T < \infty\}$; i.e., when the stopping time is finite, the evolution of the CTMC beyond the stopping time is as if it statistically restarts in the state it is in at the stopping time. In particular, we can conclude that

$$P(X(T + t_1) = i_1, \ldots, X(T + t_n) = i_n \mid X(u), u \leq T, X(T) = i_0, T < \infty) = P(X(t_1) = i_1, \ldots, X(t_n) = i_n | X(0) = i_0)$$

We state the following important result without proof.

Theorem 4.3. *A pure-jump CTMC is strong Markov.* ■

We are now ready to develop the jump structure of a pure-jump CTMC. Let $T_0 = 0, T_1, T_2, \cdots$ be the successive jump instants of the CTMC. Let $X_n = X(T_n), n \geq 0$, i.e., X_n is the process *embedded* at the jumps. $X_n, n \geq 0$, is also called the *jump chain*; as might be expected, we will see that it is a DTMC on $\mathcal{S}$.

Theorem 4.4. *For a CTMC $X(t), t \geq 0$, on $\mathcal{S}$, for every n, $i, j \in \mathcal{S}$, and $u \geq 0$, whenever $T_n < \infty$,*

$$P(T_{n+1} - T_n > u, X_{n+1} = j | X_0, \cdots, X_n = i;\ T_0, \cdots, T_n\} = p_{ij}\, e^{-a_i u}$$

where

$$\sum_{j \in \mathcal{S}} p_{ij} = 1,\ p_{ij} \geq 0. \textit{ Further, } a_i > 0 \Rightarrow p_{ii} = 0, \textit{ and } a_i = 0 \Rightarrow p_{ii} = 1$$

Proof: Clearly T_n is a stopping time. Now, for a pure jump process, conditioning on $X_0, \cdots, X_n, T_0, \cdots, T_n$ is the same as conditioning on $X(t), t \leq T_n$. Hence the left hand side in the expression in the theorem is

$$P(X(T_{n+1}) = j, T_{n+1} - T_n > u \mid X(u), u \leq T_n,\ X_{T_n} = i)$$

We can now write (from Theorem 4.3, when $T_n < \infty$)

$$\begin{aligned}
&P(X(T_{n+1}) = j, T_{n+1} - T_n > u \mid X(u), u \leq T_n,\ X_{T_n} = i) \\
&= P(X(T_1) = j,\ T_1 > u \mid X(0) = i) \\
&= P(T_1 > u \mid X(0) = i) \cdot P(X(T_1) = j \mid X_0 = i, T_1 > u) \\
&= e^{-a_i u}\, P(X(u + Y(u)) = j \mid X(s) = i, s \leq u) \\
&= e^{-a_i u}\, P(X(u + Y(u)) = j \mid X(u) = i) \\
&= e^{-a_i u}\, P(X(T_1) = j \mid X_0 = i) \\
&= e^{-a_i u}\, p_{ij}
\end{aligned}$$

where $p_{ij} := P(X(T_1) = j \mid X_0 = i)$. In writing the above sequence of equalities, in the first equality we have used the strong Markov property, in the fourth equality we have used the Markov property, and in the fifth we have used time homogeneity. When $a_i > 0$, by the definition of T_1, $X(T_1) \neq X(0)$, hence $p_{ii} = 0$. Also, if $a_i = 0$ then the state i is never exited; hence, $p_{ij} = 0$ for $j \neq i$, and $p_{ii} = 1$. ■

Remarks 4.1.

a. Thus we see that the state that a pure jump CTMC enters at a jump is independent of how much time was spent in the previous state.

b. Setting $u = 0$ in both sides of the expression in the theorem, we conclude that

$$P(X_{n+1} = j | X_0, \cdots, X_n = i, T_0, \cdots, T_n) = p_{ij}$$

From this it can be easily shown (as an exercise) that $X_n, n \geq 0$, is a DTMC on $\mathcal{S}$, with transition probability matrix $\mathbf{P}$ whose elements are p_{ij}. Thus $X_n, n \geq 0$, is called the *embedded Markov chain* of the DTMC $X(t), t \geq 0$.

Example 4.2 (Poisson Process)**.**

We saw in Example 4.1 that a Poisson process $N(t), t \geq 0$, is a CTMC. It is also a simple observation that for $N(t)$, for all $i \geq 0$, $a_i = \lambda$, and $p_{i(i+1)} = 1$. ■

Example 4.3 (The M/M/1 Queue)**.**

An M/M/1 queue is a single server queue with Poisson arrivals and i.i.d. exponentially distributed service times. Let the arrival rate be λ and the parameter of the exponential service time be μ, i.e., the mean service time is $\frac{1}{\mu}$. Let $X(t)$ be the number of customers in the system at time t; hence $X(t)$ is a jump process that takes values in $\{0, 1, 2, 3, \cdots\}$. Let us first observe that

$$P(X(t+s) = j \mid X(u), u \leq t, X(t) = i) = P(X(s) = j \mid X(0) = i)$$

This is because, given that $X(t) = i$, the future increments of the arrival process are independent of the past, and also the residual service time of the customer in service is independent of the past. Thus we conclude that $X(t), t \geq 0$, is a CTMC. Now given that $X(t) = i > 0$, the remaining sojourn time in this state is exponentially distributed with parameter $\lambda + \mu$; further, the next state is $i + 1$ with probability $\frac{\lambda}{\lambda+\mu}$. If $X(t) = 0$, then the state does not change until an arrival occurs. The time until this arrival is exponentially distributed with parameter λ. In summary, we find that

$$a_i = \begin{cases} \lambda & \text{for } i = 0 \\ \lambda + \mu & \text{for } i > 0 \end{cases}$$

and

$$p_{ij} = \begin{cases} 1 & \text{for } i = 0, j = 1 \\ \frac{\lambda}{\lambda+\mu} & \text{for } i > 0, j = i + 1 \\ \frac{\mu}{\lambda+\mu} & \text{for } i > 0, j = i - 1 \\ 0 & \text{otherwise} \end{cases}$$

■

4.4 Regular CTMC

In Section 4.3 we found that the jump structure of a CTMC is characterised in terms of the jump transition probability matrix $\mathbf{P}$, and the vector of rates of the exponentially distributed sojourn times in each state, denoted by the vector $\mathbf{a}$. Given a transition probability matrix $\mathbf{P}$ and a vector of nonnegative numbers $\mathbf{a}$, we could ask whether these suffice to completely describe a CTMC. In other words, whether a unique transition probability function $\mathbf{P}(\mathbf{t}), t \geq 0$, can be derived. So the question is the following. Given the embedded Markov chain $X_n, n \geq 0$, and the jump instants $T_k, k \geq 1$, let us define a random process $X(t)$ by $X(t) = X_n$ for $t \in [T_n, T_{n+1})$. Does this define $X(t)$ for all $t \geq 0$? The answer to this question is "yes" only if for each t there exists an $n \geq 0$ such that $t \in [T_n, T_{n+1})$. Let us define

$$\xi := \sum_{n=0}^{\infty} (T_{n+1} - T_n)$$

Suppose that for some ω, $\xi(\omega) < \infty$, then for this ω, for $t > \xi(\omega)$, the above construction of $X(t)$ fails. In addition, if $P(\xi < \infty) > 0$, then with positive probability the construction fails. The following is an example of a pair $\mathbf{P}$ and $\mathbf{a}$ for which this happens.

Example 4.4.

Let the state space be $\mathcal{S} = (0, 1, 2, \cdots)$. For $i \geq 0$, let $p_{i(i+1)} = 1$ and $a_i = \frac{1}{\lambda^i}$ where $0 < \lambda < 1$. With $X_0 = 0$, the embedded Markov chain jumps successively from state 0 to 1 to 2 and so on, staying in the successive states for exponentially distributed times with means $\lambda^i, i \geq 0$. We can then see that, with ξ defined earlier,

$$\begin{aligned} \mathsf{E}(\xi) &= \sum_{i=0}^{\infty} \lambda^i \\ &= \frac{1}{1-\lambda} \\ &< \infty \end{aligned}$$

This implies that $P(\xi < \infty) = 1$. Thus, in this case, $X(t)$ as constructed above will not define a process for all t. ■

Remark: With reference to the above example, we note that there are several ways to continue the construction beyond the random time $Z_1 := \xi$. One possibility is to "reset" the process to 0 at ξ. The process will again evolve through the states $0, 1, 2, \cdots$, and there will be a subsequent time Z_2, distributed identically to ξ, at which again the process will need to be reset to 0. Another possibility is to reset the process at Z_1 to some randomly chosen state according to some distribution. We note that each such construction yields a different CTMC, and in fact yields a *different* transition probability function $\mathbf{P}(t)$. Thus, in general, the pair $\mathbf{P}$ and $\mathbf{a}$ do not uniquely specify a CTMC.

We will limit ourselves to the following important restriction that suffices for most practical applications.

Definition 4.5. *A pure jump CTMC $X(t), t \geq 0$, is called* regular *if $\xi = \infty$ with probability 1.* ■

Remark: This definition basically says that a regular Markov process can be constructed, as we did earlier, from its embedded Markov chain and the state sojourn times. Later we will see how to obtain the transition probability function, $\mathbf{P}(t)$, from the parameters $\mathbf{P}$ and $\mathbf{a}$ of the jump structure.

The following results provide simple checks to determine whether a CTMC is regular.

Theorem 4.5. *For a pure jump CTMC, if there exists $\nu > 0$, such that $a_i \leq \nu$, for all $i \in \mathcal{S}$, then the CTMC is regular.*

Proof: Let $\mathbf{P} = [p_{ij}]$ denote the transition probability matrix of the EMC of the CTMC. We use the important idea of *uniformisation*. Consider a Poisson process $N(t)$, of rate ν. We construct the CTMC, $X(t)$, by embedding its jumps at the points of the Poisson process, as follows. Let $Z_0 = 0$, and let $Z_k, k \geq 1$, denote the successive points of the Poisson process. Let $X(0) = i$, and let $X(t) = i$ until Z_1. At Z_1 the CTMC jumps with probability $\frac{a_i}{\nu}$, and continues to stay in i with probability $1 - \frac{a_i}{\nu}$. If the CTMC jumps, the next state is j with probability p_{ij}. Let T_1 denote the instant of the first jump of the CTMC. Note that we can write

$$T_1 = \sum_{k=1}^{M} W_k$$

where M is a random variable, with $P(M = m) = (1 - \frac{a_i}{\nu})^{m-1} \frac{a_i}{\nu}$, for $m \geq 1$, $W_k, k \geq 1$, are i.i.d. exponentially distributed with mean $\frac{1}{\nu}$, and M is independent of the sequence $W_k, k \geq 1$. It can easily be shown that T_1 is exponentially distributed with mean $\frac{1}{a_i}$. It follows that the above construction correctly captures the evolution of the jump structure of the CTMC. Now observe that as long as there are jumps in the Poisson process, $N(t)$, there will continue to be jumps in this construction. But, since $\nu > 0$, with probability 1, $\lim_{k\to\infty} Z_k = \infty$. Hence there is no finite time at which the last jump of $X(t)$ occurs. Hence, $X(t)$ is a regular CTMC. ■

The following corollary is an immediate consequence of the previous result.

Corollary 4.1. *A CTMC with a finite number of states is regular.* ■

In Chapter 2 we have learnt how to recognise that a DTMC is recurrent. These techniques can be used in applying the following result.

Theorem 4.6. *A pure jump CTMC whose EMC is recurrent is regular.*

Proof: Since the EMC is recurrent, there must be a state j with $a_j > 0$. Now j occurs infinitely often, and in each visit the time spent in j is exponentially distributed with mean $\frac{1}{a_j}$. Let us now extract out these times. We have a Poisson process of rate $a_j > 0$. Since the total time covered by the "interarrival" times of this Poisson process is ∞, it follows that $\xi = \infty$ with probability 1, and, hence, that the CTMC is regular.

■

4.5 Communicating Classes

Definition 4.6. *For a CTMC $X(t)$ on $\mathcal{S}$, and $i, j \in \mathcal{S}$, we say that j is* reachable *from i if, for some $t > 0$, $p_{ij}(t) > 0$, and we denote this by $i \to j$. When $i \to j$ and $j \to i$, we say that i and j* communicate *and denote this by $i \leftrightarrow j$.* ■

Remark: As in the case of DTMCs, $\leftrightarrow$ is an equivalence relation on $\mathcal{S}$, and it partitions $\mathcal{S}$ into equivalence classes called *communicating classes*.

We recall that the communication structure of a DTMC is basically a property of the transition probability matrix, and state the following simple correspondence between a regular CTMC and its EMC.

Theorem 4.7. *A regular CTMC and its EMC have the same communicating classes.*

Proof: It suffices to show that $i \to j$ in the CTMC if and only if $i \to j$ in the EMC. Suppose $i \to j$ in the EMC, then there is a path $i, i_1, i_2, \cdots, i_{n-1}, j$ of positive probability along which j can be reached from i in the EMC. Now $a_i > 0, a_{i_1} > 0, \cdots, a_{i_{n-1}} > 0$, since, otherwise, one of the states in the path would be absorbing and j could not be reached from i along this path. It follows that the total time along this path in the CTMC is a proper random variable, being the sum of n independent exponentially distributed random variables with means $\frac{1}{a_i}, \frac{1}{a_{i_1}}, \frac{1}{a_{i_2}}, \cdots, \frac{1}{a_{n-1}}$. Denote the cumulative distribution function of this random variable by $A(\cdot)$. Now, for a regular CTMC, we can observe the following

$$
\begin{aligned}
& P(X(t) = j | X(0) = i) \\
& \quad \geq \quad P(X_1 = i_1, X_2 = i_2, \cdots, X_{n-1} = i_{n-1}, X_n = j, X(t) = j | X(0) = i) \\
& \quad = \quad P(X_1 = i_1, X_2 = i_2, \cdots, X_{n-1} = i_{n-1}, X_n = j | X(0) = i) \\
& \qquad \cdot P(X(t) = j | (X_0 = i, X_1 = i_1, X_2 = i_2, \cdots, X_{n-1} = i_{n-1}, X_n = j)) \\
& \quad \geq \quad P(X_1 = i_1, X_2 = i_2, \cdots, X_{n-1} = i_{n-1}, X_n = j | X(0) = i) \int_0^t e^{-a_j(t-u)} dA(u) \\
& \quad > \quad 0
\end{aligned}
$$

In this computation, the second inequality follows because, given $(X_0 = i, X_1 = i_1, X_2 = i_2, \cdots, X_{n-1} = i_{n-1}, X_n = j)$, the c.d.f. of T_n is $A(\cdot)$, and one way that $X(t) = j$ is

that $T_n < t$, and the sojourn time in j, the state entered at time T_n, exceeds $t - T_n$. We conclude that, in the CTMC, $i \to j$.

Also, clearly, for a regular CTMC, if j is not reachable from i in the EMC then j cannot be reachable from i in the CTMC. ■

4.6 Recurrence and Positivity

Definition 4.7. *For a CTMC $X(t), t \geq 0$, on $\mathcal{S}$, for $j \in \mathcal{S}$, and with $X(0) = j$, define S_{jj} as follows*

$$S_{jj} = \inf\{t \geq 0 : t > Y(0),\ X(t) = j\},$$

i.e., S_{jj} is the time to return to state j after once leaving j (recall that $Y(0)$ is the sojourn time in the state that the process is in at time 0). Then, the state j is said to be recurrent in the CTMC if $P(S_{jj} < \infty) = 1$. An irreducible CTMC is called recurrent if each state $j \in \mathcal{S}$ is recurrent. ■

Remark: We recall, from Section 3.10.2, that there is a notion of periodicity in DTMCs; i.e., an irreducible, periodic DTMC returns to a state only in a number of steps that is a multiple of the period $d > 1$. In a CTMC, however, for every $i \in \mathcal{S}$ and $t > 0$, $p_{ii}(t) > e^{-a_i t} > 0$, hence there is no basic time-step in multiples of which the CTMC is found in the state i, having started in the state i. Hence, there is *no notion of periodicity* for CTMCs. It is easy to see, however, that a CTMC can have a periodic EMC.

Theorem 4.8. *An irreducible regular CTMC is recurrent if and only if its EMC is recurrent.*

Proof: If there is only one state, there is nothing to prove. So let $|\mathcal{S}| \geq 2$. Denote the CMTC by $X(t)$, and its EMC by X_k. Suppose that X_k is recurrent. Then we know (Theorem 2.6) that every state is recurrent in X_k. Consider $j \in \mathcal{S}$, and let T_j denote the recurrence time of j, and we have $P(T_j < \infty) = 1$. Since $X_k, k \geq 0$, is recurrent (and $|\mathcal{S}| \geq 2$) there cannot be any absorbing states. Hence, for all $i \in \mathcal{S}$, $0 < a_i < \infty$. Thus, if $T_j < \infty$, S_{jj} is the sum of a finite number of exponentially distributed state sojourn times, each with finite mean, and hence is finite with probability 1. It follows that S_{jj} is a proper random variable, for each j, and hence that $X(t)$ is recurrent. By the same token, if the EMC is transient, i.e., there is a state j such that $P(T_j = \infty) > 0$, then $P(S_{jj} = \infty) > 0$, and the CTMC is also not recurrent. ■

Remark: Thus we see that the recurrence/transience of an irreducible regular CTMC can be inferred from the same properties of its (irreducible) EMC. However, we will see below, that a CTMC and its EMC do not share the property of positivity or nullity. For example, a CTMC can be positive, while its EMC is null, and vice versa.

Theorem 4.9. *For an irreducible recurrent CTMC, whose state sojourn times have parameters* $\mathbf{a}$*, and whose EMC has transition probability matrix* $\mathbf{P}$*,*

$$\mathsf{E}(S_{jj}) = \frac{1}{u_j} \sum_{i \in \mathcal{S}} \frac{u_i}{a_i}$$

where $\mathbf{u} > 0$ *is a solution of the system of linear equations* $\mathbf{u} = \mathbf{uP}$*.*

Proof: Denote the CTMC by $X(t)$, and let $X(0) = j$. Since the CTMC is recurrent, so is the EMC; hence, by Theorem 3.19, there exists a positive vector $\mathbf{u}$, such that $\mathbf{u} = \mathbf{uP}$. Further, for each $i, j \in \mathcal{S}$, $\frac{u_i}{u_j} = \mathsf{E}(V_{ji})$, where V_{ji} is random number of visits to state i between successive visits to j in the EMC. Let $W_k^{(i)}, k \geq 1$ denote the successive sojourn times in state $i \in \mathcal{S}$; this is a sequence of i.i.d. random variables with an Exponential(a_i) distribution. Now, we observe that,

$$\mathsf{E}(S_{jj}) = \frac{1}{a_j} + \mathsf{E}\left(\sum_{i \in \mathcal{S}, i \neq j} \sum_{k=1}^{V_{ji}} W_k^{(i)} \right) \tag{4.2}$$

where, on the right hand side, the first term is the mean time until the CTMC leaves j, and the second term is the time spent in the other states before return to j. Note that, for each i, V_{ji} is independent of $W_k^{(i)}, k \geq 1$, the sequence of times spent in the state i. Hence

$$\mathsf{E}\left(\sum_{k=1}^{V_{ji}} W_k^{(i)} \right) = \frac{u_i}{u_j} \frac{1}{a_i}$$

which is actually a trivial case of Wald's Lemma. Further, in Equation 4.2, since the terms inside the first sum in the second term on the right hand side are all positive, we can use monotone convergence theorem (Theorem 1.6), to conclude

$$\mathsf{E}(S_{jj}) = \frac{1}{a_j} + \sum_{i \in \mathcal{S}, i \neq j} \frac{u_i}{u_j} \frac{1}{a_i}$$

which is the same as the displayed formula in the theorem statement. ■

Theorem 4.10. *Let* $\{X(t), t \geq 0\}$ *be an irreducible recurrent CTMC.*

(i) For all $i, j \in \mathcal{S}$*,* $\lim_{t \to \infty} p_{ij}(t)$ *exists and is independent of* i*. Denote this limit by* $\pi_j, j \in \mathcal{S}$*.*

(ii) Then $\pi_j > 0$*, for all* $j \in \mathcal{S}$*, if and only if* $\sum_{i \in \mathcal{S}} \frac{u_i}{a_i} < \infty$*, where* $\mathbf{u} > 0$ *is a solution of the system of linear equations* $\mathbf{u} = \mathbf{uP}$*; in this case* $\pi_j = \frac{1/a_j}{\mathsf{E}(S_{jj})}$*, and we say that the CTMC is positive. On the other hand, if* $\sum_{i \in S} \frac{u_i}{a_i} = \infty$ *then* $\pi_j = 0$*, for all* $j \in S$*, and the CTMC is null.*

Remarks 4.2.

a. Note that the result asserts that for an irreducible recurrent CTMC the limit $\lim_{t\to\infty} p_{ij}(t)$ always exists. This is in contrast to the case of DTMCs where the existence of the limit of $p_{ij}^{(k)}$ requires the condition of aperiodicity (see Section 3.10.2).

b. We notice from Theorem 4.9 that the requirement that $\sum_{i\in S} \frac{u_i}{a_i} < \infty$ is equivalent to $\mathsf{E}(S_{jj}) < \infty$ for all $j \in \mathcal{S}$. It is then easily seen, from the renewal reward theorem (Theorem 3.2), that

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t p_{ij}(u)du = \frac{1/a_j}{\mathsf{E}(S_{jj})}$$

which, as expected, is the form of $\pi_j, j \in \mathcal{S}$ displayed in the theorem statement.

c. In the positive case, the expression for $\pi_j, j \in \mathcal{S}$, can also be written as

$$\pi_j = \frac{u_j/a_j}{\sum_{i\in\mathcal{S}} \frac{u_i}{a_i}}$$

showing that π_j is proportional to $\frac{u_j}{a_j}$.

d. Observe that the EMC is positive if and only if $\sum_{i\in S} u_i < \infty$, where $\mathbf{u} > 0$ is a solution of the system of linear equations $\mathbf{u} = \mathbf{uP}$. Thus we see that the condition for the EMC to be positive recurrent and for the CTMC to be positve both involve a vector $\mathbf{u}$, whereas the condition for the CTMC to be positive also involves the state sojourn time parameter vector $\mathbf{a}$. Hence, in general, a recurrent CTMC and its (necessarily, recurrent) EMC do not share the property of positivity or nullity. We will see an example a little later.

Proof: Let $F_{jj}(\cdot)$ denote the c.d.f. of S_{jj}. Now we can write the following renewal equation for $p_{jj}(t)$.

$$p_{jj}(t) = e^{-a_j t} + \int_0^t p_{jj}(t-u)\, dF_{jj}(u)$$

Here the first term corresponds to the case of being in j at time t because of never leaving j until t. The second term corresponds to the case of the CTMC leaving j before t; then, in order to be in j at time t, the process must return to j; the renewal argument is with respect to the time of first return, which occurs in the interval $(u, u+du)$ with probability $dF_{jj}(u)$. Solving this renewal equation we obtain

$$p_{jj}(t) = e^{-a_j t} + \int_0^t e^{-a_j(t-u)}\, dm_{jj}(u)$$

where $m_{jj}(t)$ is the renewal function of the ordinary renewal process comprising visits to the state j with the initial state being j. We know that F_{jj} is nonlattice (being the sum of a random number of exponentially distributed random variables); also since $e^{-a_j t}$ is Riemann integrable and monotone decreasing, it is directly Riemann integrable. Hence, applying KRT (Theorem 3.17), we obtain

$$\lim_{t\to\infty} p_{jj}(t) = \frac{1}{\mathsf{E}(S_{jj})}\frac{1}{a_j}$$

Similarly, we obtain, for $i \neq j$,

$$p_{ij}(t) = \int_0^t e^{-a_j(t-u)}\, dm_{ij}(u)$$

where $m_{ij}(t)$ is the renewal function of the delayed renewal process comprising visits to the state j, with the initial state being i. Again, applying KRT, we obtain

$$\lim_{t\to\infty} p_{ij}(t) = \frac{1}{\mathsf{E}(S_{jj})}\frac{1}{a_j}$$

The result follows after recalling the expression for $\mathsf{E}(S_{jj})$. ■

Definition 4.8. *For a regular CTMC, for all $i, j \in \mathcal{S}, i \neq j$, define $q_{ij} = a_i\, p_{ij}$, and, for each $i \in \mathcal{S}$, define $q_{ii} = -a_i$. The $|\mathcal{S}| \times |\mathcal{S}|$ matrix $\mathbf{Q}$ with elements q_{ij} is called the* transition rate matrix *of the CTMC.* ■

Remarks 4.3.

a. Notice that the row sums of $\mathbf{Q}$ are 0. Also, given $\mathbf{Q}$ we can obtain the transition probability matrix, $\mathbf{P}$, of the EMC, and the rate vector $\mathbf{a}$. The diagonal elements of $\mathbf{Q}$ yield $\mathbf{a}$, and then each off-diagonal term of $\mathbf{Q}$ yields the corresponding element of $\mathbf{P}$.

b. The elements of $\mathbf{Q}$ have an important interpretation. Consider the evolution of the CTMC, $X(t)$, and let us separate out all the times during which the process is in the state i. These will constitute a sequence of exponentially distributed times with mean $\frac{1}{a_i}$. If we "string" all these times together we obtain a Poisson process of rate a_i. This is the *conditional time* during which the process is in the state i. Each point of this Poisson process corresponds to an entry or exit from state i in the original process. Thus $a_i(= -q_{ii})$ can be interpreted as the rate of leaving i conditioned on being in i. Now let us consider those transitions out of i that cause a jump to j. Each transition out of i is independently a jump to j with probability p_{ij} and a jump to some other state (not in $\{i, j\}$) with probability $1 - p_{ij}$. Look at the corresponding points of the conditional time Poisson process. By virtue of the sequence of choices of whether or not to jump to j constituting a Bernoulli process, these points of the Poisson process now constitute a Poisson process of rate $a_i p_{ij} = q_{ij}$. Thus q_{ij} is the conditional rate of leaving i to enter j, where the conditioning is that of being in i.

Theorem 4.11. *An irreducible recurrent CTMC is positive if and only if there exists a positive probability vector π, with elements $\pi_j, j \in \mathcal{S}$, that solves the system of linear equations $\boldsymbol{\pi}\mathbf{Q} = 0$. Such a probability vector is unique.*

Proof: Since the CTMC is irreducible recurrent, by Theorem 3.19, there exists a positive vector $\mathbf{u}$ such that $\mathbf{u} = \mathbf{uP}$, where $\mathbf{P}$ is the transition probability matrix of the EMC. Such a $\mathbf{u}$ is unique up to a multiplicative constant. Also, since the CTMC is irreducible, for all $i \in \mathcal{S}$, $a_i > 0$ and the diagonal elements of $\mathbf{P}$ are 0, i.e., $p_{ii} = 0$. Let us now observe that

$$\begin{aligned}
\boldsymbol{\pi}\mathbf{Q} = 0,\ \boldsymbol{\pi} > 0,\ \boldsymbol{\pi}\cdot\mathbf{1} = 1 \quad \text{iff} \quad & \text{for all } j \in \mathcal{S}, -\pi_j q_{jj} = \sum_{i\neq j} \pi_i q_{ij},\ \boldsymbol{\pi} > 0,\ \boldsymbol{\pi}\cdot\mathbf{1} = 1 \\
\text{iff} \quad & \text{for all } j \in \mathcal{S}, \pi_j a_j = \sum_{i\in\mathcal{S}} \pi_i a_i p_{ij},\ \boldsymbol{\pi} > 0,\ \boldsymbol{\pi}\cdot\mathbf{1} = 1 \\
\text{iff} \quad & \text{for all } j \in \mathcal{S}, \text{defining } u_j := \pi_j a_j, \\
& u_j = \sum_{i\in\mathcal{S}} u_i p_{ij} \text{ (i.e., } \mathbf{u} = \mathbf{uP}), \mathbf{u} > 0,\ \sum_{i\in\mathcal{S}} \frac{u_i}{a_i} = 1 < \infty
\end{aligned}$$

Thus

$$\begin{aligned}
\boldsymbol{\pi}\mathbf{Q} = 0,\ \boldsymbol{\pi} > 0,\ \boldsymbol{\pi}\cdot\mathbf{1} = 1 \quad \Rightarrow \quad & \text{there exists } \mathbf{u} > 0, \mathbf{u} = \mathbf{uP},\ \sum_{i\in\mathcal{S}} \frac{u_i}{a_i} < \infty \\
\Rightarrow \quad & \text{the CTMC is positive by Theorem 4.10}
\end{aligned}$$

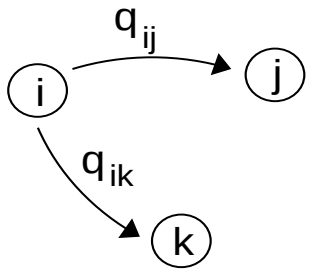
On the other hand, using Theorem 4.10,

$$\begin{aligned}
\text{if the CTMC is positive} \quad \Rightarrow \quad & \text{there exists } \mathbf{u} > 0, \mathbf{u} = \mathbf{uP}, \sum_{i\in\mathcal{S}} \frac{u_i}{a_i} = 1 \\
\Rightarrow \quad & \text{defining, for all } j \in \mathcal{S}, \pi_j := \frac{u_j}{a_j}, \\
& \boldsymbol{\pi} \text{ satisfies } \boldsymbol{\pi}\mathbf{Q} = 0,\ \boldsymbol{\pi} > 0,\ \boldsymbol{\pi}\cdot\mathbf{1} = 1
\end{aligned}$$

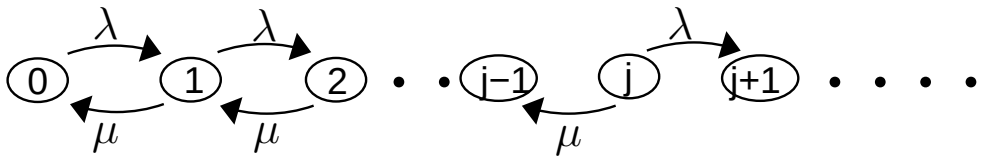
Next we show that when there is a solution to $\boldsymbol{\pi}\mathbf{Q} = 0,\ \boldsymbol{\pi} > 0,\ \boldsymbol{\pi}\cdot\mathbf{1}$, it must be unique. Suppose there are two solutions $\boldsymbol{\pi}^{(1)}$ and $\boldsymbol{\pi}^{(2)}$. Define the vectors $\mathbf{u}^{(1)}$ and $\mathbf{u}^{(2)}$ by $u_i^{(1)} = \pi_i^{(1)} a_i$ and $u_i^{(2)} = \pi_i^{(2)} a_i$. As seen above, $\mathbf{u}^{(1)} = \mathbf{u}^{(1)}\mathbf{P}$ and $\mathbf{u}^{(2)} = \mathbf{u}^{(2)}\mathbf{P}$. Now, by Theorem 3.19, there exists $\alpha > 0$, such that, for every $i \in \mathcal{S}$, $\frac{u_i^{(1)}}{u_i^{(2)}} = \alpha$. Hence $\frac{\pi_i^{(1)} a_i}{\pi_i^{(2)} a_i} = \alpha$, i.e., $\frac{\pi_i^{(1)}}{\pi_i^{(2)}} = \alpha$; since the vectors $\boldsymbol{\pi}^{(1)}$ and $\boldsymbol{\pi}^{(2)}$ are positive probability vectors, this implies that $\alpha = 1$, and hence they are the same. ■

Remark: The equations $\boldsymbol{\pi}\mathbf{Q} = 0,\ \boldsymbol{\pi} > 0,\ \boldsymbol{\pi}\cdot\mathbf{1}$, have an important interpretation. We note that $\pi_j = \frac{1/a_j}{\mathsf{E}(S_{jj})}$ is the long run fraction of time that the CTMC is in state j, and $\pi_j a_j = \frac{1}{\mathsf{E}(S_{jj})}$ is the unconditional rate of entering or leaving j. Now, for each $j \in \mathcal{S}$, we

have $\pi_j a_j = \sum_{i \neq j} \pi_i q_{ij}$. The left hand side of the equation for j is the unconditional rate of leaving j. Similarly, the ith term in the sum on the right hand side is the unconditional rate of leaving i to enter j, so the sum is the rate of entering j. Hence the equations $\boldsymbol{\pi}\mathbf{Q} = 0$ express the equality of rates of leaving j and entering j in equilibrium. For each j the equation $\pi_j a_j = \sum_{i \neq j} \pi_i q_{ij}$ is also called the *global rate balance* equation at j, or simply the *global balance* equation at j. The term "global" is used since the equation expresses the rate balance between *all* other states and j. Together the set of equations $\boldsymbol{\pi}\mathbf{Q} = 0$ are called *global balance equations.*



Transition rate diagram: The transition rate matrix, $\mathbf{Q}$, of a regular CTMC can be depicted pictorially via a transition rate diagram. The figure on the left shows a fragment of such a diagram. Note that we show the off-diagonal terms $q_{ij}, i \neq j$, and we do not show q_{ii} since $q_{ii} = -a_i$ and $a_i = \sum_{j \neq i} q_{ij}$. The fact that the CTMC stays in a state for an exponentially distributed time is implicit. Thus, the following is the transition rate diagram of the queue length process of an M/M/1 queue with arrival rate λ, and service rate μ.



4.7 Birth and Death Processes

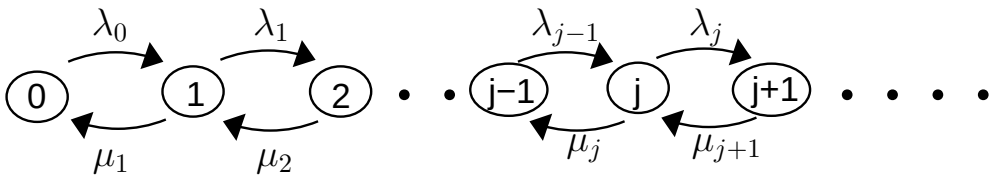
An important example of a pure jump CTMC on $\mathcal{S} = \{0, 1, 2, \ldots, \}$ is provided by the following transition rates

$$\begin{aligned} q_{j(j+1)} &= \lambda_j \text{ for } j \geq 0 \\ q_{j(j-1)} &= \mu_j \text{ for } j \geq 1 \\ q_{jk} &= 0 \text{ otherwise} \end{aligned}$$

Thus, for all $j \geq 1$,

$$q_{jj} = -(\lambda_j + \mu_j)$$

and $q_{00} = -\lambda_0$. These transition rates are depicted in the following transition rate diagram



Exercise 4.1.

Show that the number in system process $X(t)$ of an M/M/c system, where $c\ (>1)$ denotes the number of servers, is a birth and death process, and display the transition rate diagram. ■

This CTMC is irreducible if, for all $j \geq 0$, $0 < \lambda_j < \infty$, and for all $j \geq 1$, $0 < \mu_j < \infty$. Let us denote the transition probabilities of the EMC by

$$\begin{aligned} p_{j(j+1)} &= u_j = \frac{\lambda_j}{\lambda_j + \mu_j} \text{ for } j \geq 1 \\ p_{01} &= u_0 = 1 \\ p_{j(j-1)} &= d_j = \frac{\mu_j}{\lambda_j + \mu_j} \text{ for } j \geq 1 \end{aligned}$$

The other transition probabilities are 0. Denote the transition probability matrix by $\mathbf{P}$. Denote by $\tilde{\mathbf{P}}$ the matrix obtained by deleting the first row and first column of $\mathbf{P}$. By Theorem 2.11 if the only solution of $\mathbf{y} = \tilde{\mathbf{P}}\mathbf{y}$, with $\mathbf{0} \leq \mathbf{y} \leq \mathbf{1}$, is $\mathbf{y} = \mathbf{0}$ then the EMC is recurrent. Hence we now proceed to examine the solutions of $\mathbf{y} = \tilde{\mathbf{P}}\mathbf{y}$. The first equation is

$$y_1 = u_1 y_2$$

or, equivalently, since $d_1 + u_1 = 1$,

$$d_1 y_1 = u_1(y_2 - y_1) \text{ or } y_2 - y_1 = \frac{d_1}{u_1} y_1$$

Continuing, for $j \geq 2$, we obtain

$$y_j = y_{j-1} d_j + y_{j+1} u_j$$

or,

$$u_j y_j + d_j y_j = y_{j-1} d_j + y_{j+1} u_j$$

and, therefore, for $j \geq 2$,

$$(y_{j+1} - y_j) = \frac{d_j}{u_j}(y_j - y_{j-1})$$

Define, for $j \geq 1$, $z_j = y_{j+1} - y_j$. Then the above expressions become

$$\begin{aligned} z_1 &= \frac{d_1}{u_1} y_1 \\ z_j &= \frac{d_j}{u_j} z_{j-1} \text{ for } j \geq 2 \end{aligned}$$

Now, for $j \geq 1$, we can write

$$y_j = z_{j-1} + z_{j-2} + \cdots + z_1 + y_1$$

It follows that any solution to $\mathbf{y} = \tilde{\mathbf{P}}\mathbf{y}$ must satisfy

$$y_j = \left(\frac{d_{j-1} \cdots d_1}{u_{j-1} \cdots u_1} + \frac{d_{j-1} \cdots d_1}{u_{j-1} \cdots u_1} + \cdots + \frac{d_1}{u_1} + 1 \right) y_1$$

Hence there exists no positive bounded solution if and only if

$$1 + \sum_{j=1}^{\infty} \frac{d_1 \cdots d_j}{u_1 \cdots u_j} = \infty$$

or, substituting the expressions for d_j and u_j, we obtain the following necessary and sufficient condition for the recurrence of the EMC and hence of the birth and death process

$$\sum_{j=1}^{\infty} \frac{\mu_1 \cdots \mu_j}{\lambda_1 \cdots \lambda_j} = \infty$$

After verifying that the CTMC is recurrent we can take recourse to Theorem 4.11 to determine if it is also positive. Hence we consider solutions to $\boldsymbol{\pi}\mathbf{Q} = \mathbf{0}$, $\boldsymbol{\pi} > 0$, and $\boldsymbol{\pi}\mathbf{1} = 1$.

Exercise 4.2.

Show that for a birth and death process the global balance equations are *equivalent* to the following equations: for all $j \geq 0$,

$$\pi_j \lambda_j = \pi_{j+1} \mu_{j+1}$$

Hence conclude that any positive solution of $\boldsymbol{\pi}\mathbf{Q} = \mathbf{0}$ is summable if and only if

$$1 + \sum_{j=1}^{\infty} \frac{\lambda_0 \lambda_1 \cdots \lambda_{j-1}}{\mu_1 \mu_2 \cdots \mu_j} < \infty$$

■

In the case in which there is a positive summable solution of $\boldsymbol{\pi}\mathbf{Q} = \mathbf{0}$, the above exercise also yields the form of the solution. We have established the following result

Theorem 4.12. *A birth and death process is recurrent if and only if*

$$\sum_{j=1}^{\infty} \frac{\mu_1 \cdots \mu_j}{\lambda_1 \cdots \lambda_j} = \infty$$

In that case, if $\lambda_j + \mu_j < \infty$ for all j, then the process is also regular. The process is positive if and only if

$$\sum_{j=1}^{\infty} \frac{\lambda_0 \cdots \lambda_{j-1}}{\mu_1 \cdots \mu_j} < \infty$$

and then

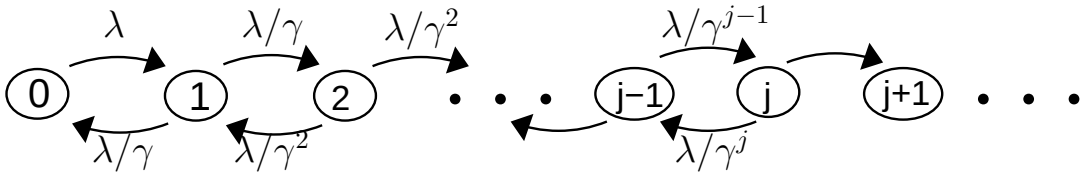
$$\pi_0 = \left(1 + \sum_{j=1}^{\infty} \frac{\lambda_0 \cdots \lambda_{j-1}}{\mu_1 \cdots \mu_j} \right)^{-1}$$

and, for $j \geq 1$,

$$\pi_j = \left(\frac{\lambda_0 \cdots \lambda_{j-1}}{\mu_1 \cdots \mu_j} \right) \cdot \pi_0$$

■

Example 4.5.



Consider a birth and death process with $\lambda_j = \mu_j$, for $j \geq 1$, so that $d_j = u_j = \frac{1}{2}$. Hence the EMC is a recurrent null random walk. Suppose, for $j \geq 0$, $\frac{\lambda_j}{\mu_{j+1}} = \gamma$, with $0 < \gamma < 1$. The transition rate diagram is displayed above. Then it easily follows from Theorem 4.12 that the CTMC is positive. Hence here we have an example of a CTMC that is positive but its EMC is null. ■

4.8 Differential Equations for $\mathbf{P}(t)$

In the case of DTMCs, the k step transition probabilities $\mathbf{P}^{(k)}, k \geq 0$, are completely specified in terms of the one step transition probability matrix $\mathbf{P}$, hence the initial probability distribution $\boldsymbol{\pi}_0$ and $\mathbf{P}$ determine the finite dimensional distributions of a DTMC. We have seen that, in general, for a pure jump CTMC the transition probability matrix of the EMC, $\mathbf{P}$, and the state sojourn parameter vector, $\mathbf{a}$, do not determine the

transition matrix function $\mathbf{P}(t), t \geq 0$. For a regular CTMC, however, the following results show that the transition rate matrix, $\mathbf{Q}$, indeed determines $\mathbf{P}(t), t \geq 0$, and hence, along with an initial probability vector $\boldsymbol{\pi}_0$, the finite dimensional distributions are also determined.

Theorem 4.13. *Let $X(t), t \geq 0$, be a pure jump, regular CTMC with rate matrix $\mathbf{Q}$. For $t \geq 0$,*

$$\begin{aligned} \frac{d}{dt}\mathbf{P}(t) &= \mathbf{Q}\mathbf{P}(t), && \textit{(Backward Equations)} \\ \frac{d}{dt}\mathbf{P}(t) &= \mathbf{P}(t)\mathbf{Q}, && \textit{(Forward Equations)} \end{aligned}$$

Remark: We note that the backward equation is obtained by conditioning on the first jump in $[0, t]$ and the forward equation is obtained by conditioning on the last jump in $[0, t]$. For a regular CTMC since there is a finite number of jumps in any interval, there is a last jump in $[0, t]$, with probability 1.

Proof: *Backward Equations:* Let the time until the first jump be denoted by T_1. Considering two cases: (i) whether the first jump occurs after t, or (ii) the first jump occurs in $[0, t]$, and then, in the latter case, conditioning on the time of this jump and using the strong Markov property, we obtain

$$\begin{aligned} p_{ij}(t) &= P(X(t) = j | X(0) = i) \\ &= P(T_1 > t)\delta_{ij} + \int_0^t \sum_{k \in \mathcal{S}, k \neq i} P(T_1 \in du, X_1 = k, X(t) = j | X(0) = i) \\ &= e^{-a_i t}\delta_{ij} + \int_0^t \left\{ \sum_{k \in \mathcal{S}, k \neq i} p_{ik}\, p_{kj}(t-u) \right\} a_i e^{-a_i u}\, du \end{aligned}$$

where $\delta_{ij} = 1$ if $i = j$, and $\delta_{ij} = 0$ otherwise. Letting $t - u = v$ in the integral, we get

$$p_{ij}(t) = e^{-a_i t}\left[\delta_{ij} + \int_0^t \left\{ \sum_{k \in \mathcal{S}, k \neq i} p_{ik}\, p_{kj}(v) \right\} a_i e^{+a_i v}\, dv \right]$$

Note that when $a_i = 0$, $p_{ij}(t) = \delta_{ij}$, as expected. We now wish to differentiate this expression for $p_{ij}(t)$ with respect to t. This will require the differentiation of the second term which is a definite integral with upper integration limit t. Differentiability of this term requires continuity of its integrand. From the assumptions made in the beginning of Section 4.1, we recall that the transition probability function is continuous for all $t \geq 0$; i.e., the functions $p_{kj}(t)$ are continuous for all $t \geq 0$. Being probabilities, these are bounded by 1. Hence, by the bounded convergence theorem (see Theorem 1.7), we

conclude that the term $\sum_{k\in\mathcal{S},k\neq i}\ p_{ik}\ p_{kj}(v)$ is continuous, and hence the entire integrand is continuous. Differentiating, we obtain

$$\begin{aligned}\frac{d}{dt}\ p_{ij}(t) &= -a_i p_{ij}(t) + e^{-a_i t}\cdot a_i \cdot e^{a_i t}\sum_{k\in\mathcal{S},k\neq i} p_{ik}\ p_{kj}(t)\\ &= \sum_{k\in S} q_{ik}\ p_{kj}(t)\\ &= [\mathbf{QP}(t)]_{ij}\end{aligned}$$

Forward Equations: Letting U denote the random time of the last jump in $[0,t]$, we obtain

$$p_{ij}(t) = e^{-a_i t}\ \delta_{ij} + \int_0^t P(U\in du,\ X(U)=j|X(0)=i) + \alpha_i(t)$$

where the first term is the probability that there is no jump in $[0,t]$, the second term is the probability that the last jump in $[0,t]$ is in $(u,u+du)$, and this jump is into the state j, and $\alpha_i(t)$ is the probability that the number of jumps in $[0,t]$ is infinite. For a regular CTMC, $\alpha_i(t)=0$. Hence we obtain

$$p_{ij}(t) = e^{-a_i t}\ \delta_{ij} + \int_0^t \left(\sum_{k\in\mathcal{S},k\neq j} p_{ik}(u) q_{k,j} du e^{-a_j(t-u)}\right) + 0$$

where the integrand in the second term is understood as follows: $p_{ik}(u)$ is the probability that at time u the process is in the state $k\neq j$, $q_{k,j}du$ is the probability that the process jumps to j in the interval $(u,u+du)$ (see the remarks below), and then $e^{-a_j(t-u)}$ is the probability that the process stays in j at least until t. Differentiating this expression yields the Forward Equations. ■

Remarks 4.4.

a. Letting $t\to 0$ in the backward equation, we observe that $\lim_{t\to 0}\mathbf{P}'(t)=\mathbf{Q}$, since $\mathbf{P}(0)=\mathbf{I}$.

b. We also observe that we can write, as $u\to 0$,

$$p_{ij}(u) = \delta_{ij}(1-a_i u) + a_i u p_{ij} + o(u)$$

where the first term corresponds to no jump in time u, the second term corresponds to one jump, and the remaining terms are $o(u)$. Noting that $a_i p_{ii}=0$, we can write this expression compactly as

$$\mathbf{P}(u) = \mathbf{I} + \mathbf{Q}u + o(u)$$

as $u\to 0$. Since $\mathbf{P}(0)=\mathbf{I}$, we can use this expression to again conclude that $\mathbf{P}'(0)=\mathbf{Q}$.

■

We state the following result without proof.

Theorem 4.14. *For a regular CTMC, the unique solution of the Backward Equations and Forward Equations is*

$$\mathbf{P}(t) = e^{\mathbf{Q}t}$$

where

$$e^{\mathbf{Q}t} = \sum_{k=0}^{\infty} \frac{(\mathbf{Q}t)^k}{k!}$$

Theorem 4.15. $X(t), t \geq 0$, *is an irreducible regular CTMC. There exists a stationary measure* $\boldsymbol{\pi}$ *such that* $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}(t)$ *for all* $t \geq 0$ *iff* $\boldsymbol{\pi}\mathbf{Q} = 0$.

Proof: From Theorem 4.14, $\boldsymbol{\pi}\mathbf{Q} = 0 \Rightarrow \boldsymbol{\pi}\mathbf{P}(t) = \boldsymbol{\pi}$ since

$$\boldsymbol{\pi}\frac{\mathbf{Q}^k t^k}{k!} = 0 \text{ for } k \geq 1$$

Further

$$\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}(t),\ t \geq 0 \Rightarrow \boldsymbol{\pi}\mathbf{P}'(t) = 0$$

and from the Forward Equations, which hold for a regular CTMC,

$$\boldsymbol{\pi}\mathbf{P}(t)\mathbf{Q} = 0 \Rightarrow \boldsymbol{\pi}\mathbf{Q} = 0$$

■

Remarks 4.5.

The following are two consequences of Theorem 4.15.

(i) For an irreducible regular Markov Chain consider the positive vector $\boldsymbol{\pi}$ such that $\boldsymbol{\pi}\mathbf{Q} = 0,\ \boldsymbol{\pi} \cdot \mathbf{1} = 1$. If $P(X(0) = j) = \pi_j$, then, for all $t \geq 0,\ \mathbf{P}(X(t) = j) = \pi_j,\ j \in S$; i.e., the CTMC is stationary.

(ii) The following is a variation of Theorem 4.11. An irreducible *regular* CTMC $X(t), t \geq 0$, is positive if and only if there exists a vector $\boldsymbol{\pi}$ such that $\boldsymbol{\pi}\mathbf{Q} = 0$. The "if part" of this statement is: an irreducible *regular* CTMC $X(t), t \geq 0$, is positive if there exists a vector $\boldsymbol{\pi}$ such that $\boldsymbol{\pi}\mathbf{Q} = 0$. Then, by Theorem 4.15, for all $t \geq 0$, $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}(t)$. Because $\boldsymbol{\pi} \cdot \mathbf{1} = 1$, there exists j with $\pi_j > 0$. Since the CTMC is irreducible, for all k there exists t_k such that $p_{jk}(t_k) > 0$. But then $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}(t_k)$ implies that $\pi_k = \sum_{i \in \mathcal{S}} \pi_i p_{ik}(t_k) > 0$. We conclude that $\boldsymbol{\pi} > 0$. Now fix a $t_1 > 0$, and consider the DTMC $Y_k = X(kt_1), k \in \{0, 1, 2, \cdots\}$. Y_k has the transition probability matrix $\mathbf{P}_1 = \mathbf{P}(t_1)$. For this DTMC, π is a positive probability vector that solves $\pi = \pi\mathbf{P}_1$. Hence, Y_k is a positive DTMC. It follows that $X(t), t \geq 0$,

is recurrent. It then follows from Theorem 4.11 that $X(t)$ is positive. The "only if" part of the statement is: An irreducible *regular* CTMC $X(t), t \geq 0$, is positive only if there exists a vector $\boldsymbol{\pi}$ such that $\boldsymbol{\pi}\mathbf{Q} = 0$. But since positivity implies recurrence this is just a direct consequence of Theorem 4.11.

Theorem 4.11 first requires us to verify that the given CTMC is recurrent. The above variation only asks for regularity. Consider an M/M/1 queue with arrival rate $\lambda, 0 < \lambda < \infty$, and service rate $\mu, 0 < \mu < \infty$. The queue length process is regular. Hence, in order to ascertain positivity of the process, it suffices to look for solutions of $\boldsymbol{\pi}\mathbf{Q} = 0$ where $\mathbf{Q}$ is the transition rate matrix of the M/M/1 queue.

■

4.9 Notes on the Bibliography

This chapter was developed from the material on continuous time Markov chains provided in Çinlar [5] and Wolff [17]. An excellent modern treatment of discrete time and continuous time Markov chains is provided by Bremaud [4].

4.10 Problems

4.1. $\{X(t),\ t \geq 0\}$ and $\{Y(t),\ t \geq 0\}$ are independent Poisson processes of rates λ_1 and λ_2 ($\lambda_1 \neq \lambda_2$). Define $Z(t) = X(t) - Y(t)$.

a. Show that $\{Z(t), t \geq 0\}$ is a CTMC, and display its transition structure.

b. Classify $\{Z(t)\}$ as recurrent, transient, positive or null. Explain.

4.2. Consider a birth-and-death process with $\lambda_{i,i+1} = \lambda$ for $0 \leq i \leq N-1$, $\mu_{j,j-1} = \mu$ for $1 \leq j \leq N$, and $\lambda_{ij} = \mu_{ij} = 0$ otherwise. $(0 < \lambda < \infty,\ 0 < \mu < \infty)$

a. Show that this CTMC is positive recurrent and find the invariant probability measure.

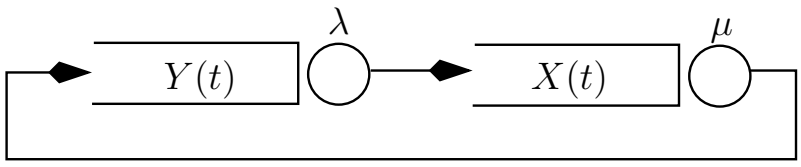
b. Let n be a fixed number such that $0 < n < N$, and let $M(t)$ be the number of $n \to n+1$ transitions in $[0,\ t]$. Obtain an expression for $\lim_{t\to\infty} \frac{1}{t} M(t)$. (Hint: consider visits to state n and use the renewal reward theorem (Theorem 3.2).)

4.3. a. Consider a (λ, μ) birth and death process on $\{0, 1, 2, \cdots, \}$. Assume that it is positive recurrent. For a given $n \geq 1$ find the mean time spent in states $\geq n+1$ between successive visits to n.

b. In (a), if the process is in state n and a down transition occurs, then the state is set back to n with probability p. Show that the resulting process is again a birth-and-death process. What is the condition for its positive recurrence, and what is the associated stationary distribution?

4.4. Construct an example of a birth and death process that is null recurrent while its embedded Markov chain is positive recurrent.

4.5. Two customers are trapped in a system comprising two queues as shown in the following figure. $\{X(t)\}$ and $\{Y(t)\}$ are the queue length processes as shown. The service times in the two queues are exponentially distributed with parameters $\lambda > 0$ and $\mu > 0$ as shown.

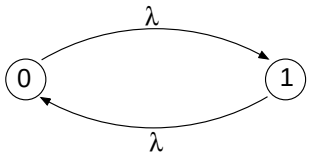


a. Argue that $\{X(t)\}$ is a CTMC; display its transition rate structure. Do the same for $\{Y(t)\}$.

b. For the process $\{X(t)\}$ find $\lim_{t\to\infty}\ p_{01}(t)$; do the same for $\{Y(t)\}$.

c. What is the limiting joint distribution of the process $\{(X(t),\ Y(t))\}$?

4.6. Consider the CTMC with transition rate diagram shown in the figure below.



Define $h(t) = p_{11}(t)$.

a. Write down a renewal equation for $h(t)$.

b. Solve the renewal equation using Laplace transforms. (Hint: $\mathcal{L}^{-1}\left(\frac{\lambda+s}{(\lambda+s)^2-\lambda^2}\right) = e^{-\lambda t}\,\mathcal{L}^{-1}\left(\frac{1}{2}\left(\frac{1}{s+\lambda}+\frac{1}{s-\lambda}\right)\right)$, where $\mathcal{L}^{-1}$ denotes the inverse Laplace transform).

c. Find $\lim_{t\to\infty}\ h(t)$, and compare the result with that obtained from the Key Renewal Theorem (Theorem 3.17).

4.7. For an irreducible CTMC, show that if there exists $a\ >\ 0$, such that for all $i \in S,\ -q_{ii}\ >\ a$, then the CTMC is positive if its EMC is positive. (That the reverse implication does not hold is demonstrated by Example 4.5)

4.8. Consider an M/M/1/1 queue, with arrival rate λ and service rate μ, and let $\{X(t),\ t \geq 0\}$ be the queue length process, with $X(0) = 0$. Note that arrivals that find the queue full are lost.

a. Obtain $\lim_{t\to\infty}\ P(X(t) = 1)$.

b. Obtain the rate of accepted arrivals.

4.9. For the Markov chain $X(t)$ on $S = \{0, 1\}$, with rate matrix elements $q_{01} = \lambda, q_{10} = \mu$, let $\mathbf{P}(t)$ denote the transition probability function. You are given that $0 < \lambda < \infty$ and $0 < \mu < \infty$.

a. Obtain $\lim_{t\to\infty}\ \frac{1}{t}\mathsf{E}\Big(\int_0^t\ I_{\{X(u)=1\}}\ du \mid X(0) = 0\Big)$.

b. Obtain $\lim_{t\to\infty}\ \frac{1}{t}\ \int_0^t\ X(u)du$, with $X(0) = 0$.

c. Obtain a probability measure $\boldsymbol{\pi}$ such that $\boldsymbol{\pi}$=$\boldsymbol{\pi}\mathbf{P}(t)$ for all $t \geq 0$.

d. Show that if $P(X(0) = i) = \pi_i$, then $\{X(t), t \geq 0\}$ is strictly stationary.

4.10. Consider an irreducible regular CTMC on the state space S with rate matrix $\mathbf{Q}$. Show what happens if you take limits as $t \to \infty$ in the forward and backward differential equations.

4.11. Parts A and B arrive to an assembly station in independent Poisson processes of rates λ and μ. The parts are queued in separate buffers. Let $A(t)$ be the number of parts of type A at time t, and similarly define $B(t)$. Assume that assembly time is 0, and hence observe that $A(t)$ and $B(t)$ cannot be both positive. Define $X(t) = A(t) - B(t)$.

a. Show that $X(t)$ is a CTMC and obtain the sojourn rate vector $\mathbf{a}$ and the transition probability matrix of the EMC $\mathbf{P}$.

b. Is $X(t)$ regular? Display the transition rate matrix and the transition rate diagram.

c. Determine λ and μ for which $X(t)$ is
(i) irreducible, (ii) recurrent.

d. The system is modified so that parts B are discarded if there is no part A waiting. Show that $X(t)$ is still a CTMC, and obtain the condition on λ and μ for $X(t)$ to be positive recurrent.

4.12. Calls arrive in a Poisson process of rate $\lambda, 0 < \lambda < \infty$, to a single channel (which can serve only one call at a time). An accepted call holds the channel for an exponentially distributed time with mean $\mu^{-1}, 0 < \mu < \infty$. A call arriving to a busy channel is blocked. Let $X(t) \in \{0, 1\}$ denote the number of calls in the channel.

a. Show that $\{X(t), t \geq 0\}$ is a CTMC.

b. Define S_i to be the time taken to return to state i after once leaving $i \in \{0, 1\}$. Obtain $\mathsf{E}(S_i)$.

c. Classify the CTMC as transient or recurrent, and if recurrent, null or positive.

4.13. Packets arrive into a buffer of size K in a Poisson process $A(t)$ of rate $0 < \lambda < \infty$. There is an independent Poisson process $B(t)$ of rate $0 < \mu < \infty$ at the points of which a packet is removed from the buffer if the buffer is nonempty. Let $X(t)$ denote the number of packets in the buffer at time t.

a. Argue that $X(t)$ is a regular CTMC, and sketch its transition rate matrix.

b. Show that $X(t)$ is a recurrent CTMC.

c. Show that $X(t)$ is positive. (Hint: One does not need to solve $\boldsymbol{\pi Q} = \mathbf{0}$.)

Chapter 5

Markov Renewal Theory

We recall that a pure jump CTMC $X(t), t \geq 0$, on the discrete state space $\mathcal{S}$, has the following jump structure. There is a sequence of random vectors $(X_n, T_n), n \geq 0$, where $T_0(=0) \leq T_1 \leq T_2 \leq \cdots$, is a sequence of random times, and $X_n(= X(T_n)), n \geq 0$, is the embedded DTMC. Further, given $X(u), u \leq T_n$, the sojourn time in the state X_n, i.e., $T_{n+1} - T_n$, is exponentially distributed with a mean that depends only on X_n. A Markov renewal process generalises the above jump structure.

5.1 Markov Renewal Sequences

Let us denote $\tau_n = T_n - T_{n-1}, n \geq 1$.

Definition 5.1. *The sequence $(X_n, T_n), n \geq 0$, is called a Markov renewal sequence if*

$$P(\tau_{n+1} \leq t, X_{n+1} = j | (X_0, T_0), (X_1, T_1), \cdots, (X_n = i, T_n)) = P(\tau_{n+1} \leq t, X_{n+1} = j | X_n = i)$$

for all $n \geq 1,\ i, j \in \mathcal{S},\ t \geq 0$.

■

The generalisation is that, unlike the jump structure of a CTMC, we do not require that the sojourn time in a state and the next state be independent, nor do we require that the sojourn time in a state be exponentially distributed. We say that a Markov renewal sequence is *time homogeneous* if

$$P(\tau_{n+1} \leq t, X_{n+1} = j | X_n = i) = G_{ij}(t)$$

Assuming that $P(\tau_1 < \infty | X_0 = i) = 1$, let

$$p_{ij} = \lim_{t \to \infty} G_{ij}(t)$$

i.e.,

$$p_{ij} = P(X_{n+1} = j | X_n = i)$$

Note that we allow the possibility that $p_{ii} > 0$. It also follows that

$$\sum_{j \in S} p_{ij} = 1$$

Theorem 5.1. *In a Markov renewal sequence $(X_n, T_n), n \geq 0$, the sequence $X_n,\ n \geq 1$, is a Markov Chain on $\mathcal{S}$.*

Proof: In the defining expression (see Definition 5.1), let $t \to \infty$, yielding

$$P(X_{n+1} = j | (X_0, T_0), (X_1, T_1), \cdots, (X_n = i, T_n)) = P(X_{n+1} = j | X_n = i)$$

From this, it can be easily seen that

$$P(X_{n+1} = j | X_0, X_1, \cdots, X_n = i) = P(X_{n+1} = j | X_n = i)$$

■

In the time homogeneous case, $X_n, n \geq 0$, is a time homogeneous DTMC with transition probability matrix $\mathbf{P} = [p_{ij}]$.

Now, using the chain rule for conditional probability, we have

$$P(X_{n+1} = j,\ \tau_{n+1} \leq t | X_n = i) = P(X_{n+1} = j | X_n = i)\, P(\tau_{n+1} \leq t | X_n = i, X_{n+1} = j)$$

Define the second term on the right as $H_{ij}(t)$, i.e.,

$$H_{ij}(t) = \frac{G_{ij}(t)}{p_{ij}}$$

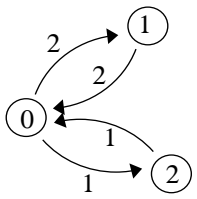
$H_{ij}(t)$ is the conditional distribution of the time between two consecutive states given the two states. We assume that, for every $i, j \in \mathcal{S}$, $\lim_{t \to \infty} H_{ij}(t) = 1$, and that the expectation of this distribution is finite, denoted by η_{ij}.

Theorem 5.2.

$$P(\tau_1 \leq t_1, \ldots, \tau_n \leq t_n | X_0, \cdots X_n) = \prod_{i=1}^{n} H_{X_{i-1}, X_i}(t_i)$$

Proof: Follows easily by writing out $P(X_0 = i_0, X_1 = i_1, \cdots, X_n = i_n, \tau_1 \leq t_1, \cdots, \tau_n \leq t_n)$ using the chain rule, and then applying the properties of a Markov renewal sequence. ■

Remarks 5.1.



a. We conclude that the successive state sojourn times are conditionally independent given the Markov Chain. Thus, we can construct a sample path of the Markov renewal sequence as follows. First obtain a sample sequence of the DTMC $X_n, n \geq 0$. Then use the previous theorem to draw independent samples of $\tau_1, \tau_2, \cdots$, using the distributions $H_{X_0,X_1}(t), H_{X_1,X_2}(t), \cdots$.

b. The above transition diagram depicts a Markov renewal sequence with 3 states[1]. If the transition is from 0 to 1 or 1 to 0, the time taken is 2 units, whereas if the transition is between 0 and 2, then the time taken is 1 unit. The transition probabilities are not shown. Let us take $p_{01} = 0.5 = p_{02}$. Then we see that $P(\tau_{n+1} = 2|X_n = 0) = 0.5$. Further, $P(\tau_{n+1} = 2, X_{n+1} = 1|X_n = 0) = 0.5$. Thus, unlike a CTMC, τ_{n+1} and X_{n+1} are not independent given X_n. It is also evident from this example that, in a Markov renewal sequence, successive sojourn times need *not* be *unconditionally* independent.

■

Further define, for each $i \in \mathcal{S}$

$$\begin{aligned} H_i(t) &= \sum_{j \in \mathcal{S}} G_{ij}(t) \\ &= \sum_{j \in \mathcal{S}} p_{ij} H_{ij}(t) \end{aligned}$$

i.e., $H_i(t)$ is the distribution of the sojourn time spent in the state i (irrespective of the state visited next). Let

$$a_i^{-1} = \sum_{j \in S} p_{ij} \eta_{ij}$$

Thus, a_i^{-1} is the mean of the distribution $H_i(t)$, and is assumed to be finite and positive.

Now, as in Section 4.4, define the random variable

$$\xi = \sum_{k=1}^{\infty} \tau_k$$

As for CTMCs, we assume that the Markov renewal sequence is such that $P(\xi = \infty) = 1$, i.e., we assume *regularity*.

[1]Note that such a diagrammatic representation of a Markov renewal sequence is not standard practice.

5.2 Semi-Markov Processes

Now let $(X_n, T_n), n \geq 0$, be a regular Markov renewal sequence, and define the process $X(t), t \geq 0$, by $X(t) = X_n$ for $t \in [T_n, T_{n+1})$. Then $X(t), t \geq 0$, is called a *semi-Markov process (SMP)*. We say that $\{X_n, n \geq 1\}$ is the embedded Markov chain (EMC) of the SMP, and that $\{(X_n, T_n), n \geq 0\}$ is the associated Markov renewal sequence.

For an SMP on the state space $\mathcal{S}$, for each $i \in \mathcal{S}$, we define S_{ij} as in Definition 4.7; i.e., with $X(0) = i$, S_{ij} is the time until $X(t)$ hits j after once leaving i. Note that, unlike CTMCs, where the sojourn time in a state is exponentially distributed and hence memoryless, here we must assert that at $t = 0$, the sojourn in i has "just begun." Let us write $F_{ij}(\cdot)$ as the distribution function of S_{ij}. We say that the state j is *reachable* from i, if, for some $t_{ij} \geq 0$, $F_{ij}(t_{ij}) > 0$, i.e., starting in i at $t = 0$, the SMP hits j with positive probability before the time t_{ij}. The SMP is irreducible if each state is reachable from every other state. As for CTMCs, it can be observed that an SMP is irreducible if and only if its EMC is irreducible. The state j is said to be *recurrent* if $P(S_{jj} < \infty) = 1$, or $\lim_{t\to\infty} F_{jj}(t) = 1$. As in the case of CTMCs (see Theorem 4.8) it is easily seen that an SMP is recurrent if and only if its EMC is recurrent.

For an SMP $X(t)$ on $\mathcal{S}$, with $X(0) = i$, the instants at which the process enters the state $j(\neq i)$ form a delayed renewal process. If at $t = 0$ the sojourn in the state i had just started, then the first lifetime random variable has the distribution of S_{ij}, and the subsequent lifetimes have the distribution of S_{jj}.

Theorem 5.3. *For a recurrent SMP,* $\mathsf{E}(S_{jj}) = \frac{1}{u_j} \sum_{i\in\mathcal{S}} \frac{u_i}{a_i}$, *where* $\mathbf{u}$ *is a positive solution to* $\mathbf{u} = \mathbf{uP}$.

Proof: Write $\sigma_{ij} = \mathsf{E}(S_{ij})$. Then, by conditioning on the first jump, we can see that

$$\sigma_{ij} = \frac{1}{a_i} + \sum_{k\in\mathcal{S}, k\neq j} p_{ik}\sigma_{kj}$$

where notice that we have used the fact that $a_i^{-1} = \sum_{k\in S} p_{ik}\eta_{ik}$. We are given $\mathbf{u}$ positive that solves $\mathbf{u} = \mathbf{uP}$ (recall that such a $\mathbf{u}$ exists since the EMC is recurrent; see Theorem 3.19). Multiply the previous equation on both sides by u_i and sum over $i \in \mathcal{S}$, to obtain

$$\sum_{i\in\mathcal{S}} u_i\sigma_{ij} = \sum_{i\in\mathcal{S}} u_i\frac{1}{a_i} + \sum_{i\in\mathcal{S}} u_i \sum_{k\in\mathcal{S}, k\neq j} p_{ik}\sigma_{kj}$$

Exchanging the summations in the second term on the right, and using the fact that $\mathbf{u} = \mathbf{uP}$, we obtain

$$\sum_{i\in\mathcal{S}} u_i\sigma_{ij} = \sum_{i\in\mathcal{S}} u_i\frac{1}{a_i} + \sum_{k\in\mathcal{S}, k\neq j} u_k\sigma_{kj}$$

which yields

$$u_j \sigma_{jj} = \sum_{i \in \mathcal{S}} u_i \frac{1}{a_i}$$

from which the desired result follows. ■

Remark: Notice that we can write the expression for $\mathsf{E}(S_{jj})$ obtained in the previous theorem as

$$\mathsf{E}(S_{jj}) = \frac{1}{a_j} + \sum_{i \in \mathcal{S}, i \neq j} \frac{u_i}{u_j} \frac{1}{a_i}$$

The first term is the mean time until the process leaves the state j. We recall, from Theorem 3.19, the interpretation of $\frac{u_i}{u_j}$ as the expected number of visits to i between returns to j in the EMC. Each such visit to i incurs a mean sojourn time of $\frac{1}{a_i}$. Hence, the second term is just the sum of mean times spent in each of the other states, before return to j. ■

We now study the pointwise limiting behaviour of an SMP. We assume that the Markov renewal sequence is regular. Let

$$p_{ij}(t) = P(X(t) = j | X(0) = i)$$

where, as stated earlier, at time $t = 0$ we take the process to have just started its sojourn in i.

Theorem 5.4. *$X(t), t \geq 0$, is an irreducible, recurrent SMP such that for each $j \in \mathcal{S}$ the distribution $F_{jj}(\cdot)$ is nonlattice. Let $\mathbf{u}$ be a positive vector that solves $\mathbf{u} = \mathbf{uP}$. Then*

$$\lim_{t \to \infty} p_{ij}(t) = \frac{\frac{u_j}{a_j}}{\sum_{k \in \mathcal{S}} \frac{u_k}{a_k}}$$

Proof: We can write down the following renewal equation

$$p_{jj}(t) = (1 - H_j(t)) + \int_0^t p_{jj}(t-u) dF_{jj}(u)$$

Solving this renewal equation, we obtain

$$p_{jj}(t) = (1 - H_j(t)) + \int_0^t (1 - H_j(t-u)) dm_{jj}(u)$$

where $m_{jj}(\cdot)$ is the renewal function of the renewal process with lifetime distribution $F_{jj}(\cdot)$. Since $F_{jj}(\cdot)$ is nonlattice, using the Key Renewal Theorem (Theorem 3.17) we obtain

$$\lim_{t \to \infty} p_{jj}(t) = \frac{1}{a_j \mathsf{E}(S_{jj})}$$

Similarly, after writing the delayed renewal equation, we obtain

$$\lim_{t\to\infty} p_{ij}(t) = \frac{1}{a_j \mathsf{E}(S_{jj})}$$

Hence, by Theorem 5.3, we conclude that

$$\lim_{t\to\infty} p_{ij}(t) = \frac{\frac{u_j}{a_j}}{\sum_{k\in\mathcal{S}} \frac{u_k}{a_k}}$$

■

Remarks 5.2.

a. We observe that the limit is positive if and only if

$$\sum_{i\in\mathcal{S}} \frac{u_i}{a_i} < \infty$$

thus reiterating what we saw in the case of CTMCs, that the notion of positivity of an SMP and its EMC need not coincide (the latter requiring that $\sum_{i\in\mathcal{S}} u_i < \infty$).

b. We also observe that the limit obtained in the previous theorem depends only on the means of the sojourn time distributions, and not on the distributions themselves. This is an example of *insensitivity* to distributions.

5.3 Markov Regenerative Processes

We now define a class of processes that generalises the class of regenerative processes defined by Definition 3.4.

Definition 5.2. *Consider a process $X(t), t \geq 0$, taking values in the discrete state space $\mathcal{S}$. For each $i \in \mathcal{S}$, let $P_i(\cdot)$ denote the probability law of $X(t), t \geq 0$, given that $X(0) = i$, where we mean that at $t = 0$ the process just began its sojourn in i. $X(t), t \geq 0$, is called a* Markov regenerative process (MRGP) *if, with $X(0) = i$, there exists a stopping time T_1 such that*

a. given $X(T_1) = j$, the probability law of $X(t + T_1),\ t \geq 0$, is $P_j(\cdot)$, and

b. $\{X(t + T_1),\, t \geq 0\}$ II $\{T_1, \text{and } X(u),\ u < T_1\}$, given $X(T_1) = j$.

■

Now, if $X(t), t \geq 0$, is a MRGP on $\mathcal{S}$, let $X_0 = X(0) = i$, with $T_0 = 0$. Then, by definition we have $T_1 \geq T_0$, and define $X_1 = X(T_1)$. Again, with $X_1 = j$ (say), we have T_2 and X_2, similarly defined. It can easily be seen, from the definition of an MRGP, that $(X_n, T_n), n \geq 0$, is a Markov renewal sequence.

For $i \in \mathcal{S}$, let

$$M^{(i)}(t) = |\{n : T_n \leq t, X_n = i\}|$$

i.e., $M^{(i)}(t)$ is the number of times the Markov renewal sequence visited i before time t. Let $Z_n^{(i)}$ denote the n-th Markov renewal instant at which the Markov renewal sequence hits the state i. Evidently, for each i, $Z_n^{(i)}$ constitute renewal instants, and $M^{(i)}(t), t \geq 0$, is the corresponding renewal counting process. Further, by comparing Definition 5.2, and Definition 3.4, we conclude that with respect to the time instants $Z_n^{(i)}$, $X(t), t \geq 0$, is a regenerative process.

In applications we are often interested in the long run fraction of time that an MRGP spends in some state, i.e.,

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t I_{\{X(u)=\ell\}}\, du$$

We can obtain this limit by using the observation that an MRGP has an embedded regenerative processes. But first we need the following notation. Let

$$\alpha_j^{(\ell)} := \mathsf{E}\left(\left(\int_0^{T_1} I_{\{X(u)=\ell\}} du \right) |X_0 = j \right)$$

i.e., the expected time that the process $X(t)$ spends in the state ℓ until the first Markov renewal instant, given that the process starts in the state j. By conditioning on the state at T_1, we can then write

$$\alpha_j^{(\ell)} := \sum_{k\in\mathcal{S}} p_{jk} \mathsf{E}\left(\left(\int_0^{T_1} I_{\{X(u)=\ell\}} du \right) |X_0 = j, X_1 = k \right)$$

Denote

$$\alpha_{jk}^{(\ell)} := \mathsf{E}\left(\left(\int_0^{T_1} I_{\{X(u)=\ell\}} du \right) |X_0 = j, X_1 = k \right)$$

Theorem 5.5. *$X(t), t \geq 0$, is an MRGP with embedded Markov renewal sequence $(X_n, T_n), n \geq 0$,. The embedded DTMC X_n is irreducible, and recurrent, with transition probability matrix $\mathbf{P}$. We are given that, if $\mathbf{u}$ is a positive vector such that $\mathbf{u} = \mathbf{uP}$, then*

$\sum_{k\in\mathcal{S}} u_k \frac{1}{a_k} < \infty$. *Let, for each* $j, \ell \in \mathcal{S}$, $\alpha_j^{(\ell)}$ *be as defined above. Then,*

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t I_{\{X(u)=\ell\}}\, du \stackrel{a.s.}{=} \frac{\sum_{k\in\mathcal{S}} u_k \alpha_k^{(\ell)}}{\sum_{k\in\mathcal{S}} u_k \frac{1}{a_k}}$$

Proof: Fix $j \in \mathcal{S}$, let $X_0 = j$, and consider the regenerative cycles defined by $Z_n^{(j)}$. We observe from the results in Section 3.5.1 that

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t I_{\{X(u)=\ell\}}\, du \stackrel{a.s.}{=} \frac{\mathsf{E}\left(\int_0^{Z_1^{(j)}} I_{\{X(u)=\ell\}}\, du | X_0 = j \right)}{\mathsf{E}\left(Z_1^{(j)}\right)} \tag{5.1}$$

The denominator on the right was earlier shown to be

$$\mathsf{E}\left(Z_1^{(j)}\right) = \sigma_{jj} = \sum_{k\in\mathcal{S}} \frac{u_k}{u_j} \frac{1}{a_k} \tag{5.2}$$

Let

$$r_{ij} = \mathsf{E}\left(\int_0^{Z_1^{(j)}} I_{\{X(u)=\ell\}}\, du | X_0 = i \right)$$

With $X_0 = i$, r_{ij} is the expected time the process spends in state ℓ until $Z_1^{(j)}$. We are interested in r_{jj}. Now, for fixed j, using the notation introduced before the theorem statement,

$$r_{ij} = \sum_{k\in\mathcal{S}} p_{ik} \alpha_{ik}^{(\ell)} + \sum_{k\in\mathcal{S}, k\neq j} p_{ik} r_{kj}$$

i.e.,

$$r_{ij} = \alpha_i^{(\ell)} + \sum_{k\in\mathcal{S}, k\neq j} p_{ik} r_{kj}$$

Multiplying both sides by u_i and summing over $i \in \mathcal{S}$, we obtain

$$\sum_{i\in\mathcal{S}} u_i r_{ij} = \sum_{i\in\mathcal{S}} u_i \alpha_i^{(\ell)} + \sum_{i\in\mathcal{S}} u_i \sum_{k\in\mathcal{S}, k\neq j} p_{ik} r_{kj}$$

Exhanging the summations in the second term on the right, and using the fact that $u_k = \sum_{i\in\mathcal{S}} u_i p_{ik}$

$$\sum_{i\in\mathcal{S}} u_i r_{ij} = \sum_{i\in\mathcal{S}} u_i \alpha_i^{(\ell)} + \sum_{k\in\mathcal{S}, k\neq j} u_k r_{kj}$$

Cancelling terms between the summation on the left and the second summation on the right, we obtain

$$u_j r_{jj} = \sum_{i \in \mathcal{S}} u_i \alpha_i^{(\ell)}$$

or

$$r_{jj} = \sum_{i \in \mathcal{S}} \frac{u_i}{u_j} \alpha_i^{(\ell)} \tag{5.3}$$

Combining Equations 5.1, 5.2, and 5.3

$$\lim_{t \to \infty} \frac{1}{t} \int_0^t I_{\{X(u)=\ell\}} \, du \overset{a.s.}{=} \frac{\sum_{k \in \mathcal{S}} u_k \alpha_k^{(\ell)}}{\sum_{k \in \mathcal{S}} u_k \frac{1}{a_k}}$$

■

Remark: When $\mathbf{u}$ is the stationary probability vector of the embedded DTMC $X_n, n \geq 0$, (if one exists), then there is an obvious interpretation of the result shown in the previous theorem. The fraction of time that the process spends in the state ℓ is the ratio of the expected time that it spends in ℓ in a Markov renewal period (i.e., $\sum_{k \in \mathcal{S}} u_k \alpha_{k\ell}$) to the expected length of a Markov renewal period.

5.4 Notes on the Bibliography

This short treatment of Markov renewal processes was developed from Wolff [17] and Kulkarni [11]. Chapter 10 of Çinlar [5] provides a detailed treatment of Markov renewal theory, with applications to processes arising in queueing models.

5.5 Problems

5.1. Consider the example of a Markov renewal process depicted in Remark 5.1. Consider the semi-Markov process $X(t)$ constructed from this MRP, as in Section 5.2. For $i \in \{0, 1, 2\}$, obtain

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t I_{\{X(u)=i\}}\, du,$$

i.e., the long run fraction of time that $X(t)$ spends in each of the states $\{0, 1, 2\}$.

5.2. Consider the GI/M/1 queue (i.e., renewal arrival process with interarrival time distribution $A(\cdot)$, and i.i.d. exponential service times with rate μ, with the service time sequence being independent of the arrival process). Let $\{X(t),\ t \geq 0\}$, $X(t) = 0$, denote the queue length process.

a. Show that $X(t)$ is not a Markov chain.

b. Let $T_0 = 0$, and $T_n, n \geq 1$, denote the arrival instants. Define $X_n = X(T_n)$ and show that $(X_n, T_n), n \geq 0$, is a Markov renewal sequence.

Bibliography

[1] S. Asmussen. *Applied Probability and Queues*. Springer-Verlag, 2003.

[2] Krishna B. Athreya and Soumendra N. Lahiri. *Probability Theory*. Hindustan Book Agency, 2006.

[3] Pierre Bemaud. *An Introduction to Probabilistic Modelling*. Springer Verlag, 1988.

[4] Pierre Bremaud. *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*. Springer, 1999.

[5] Erhan Çinlar. *Introduction to Stochastic Process*. Prentice-Hall, 1975.

[6] D.R. Cox. *Renewal Theory*. Methuen and Co., 1967.

[7] G. Fayolle, V.A. Malyshev, and M.V. Menshikov. *Topics in the Constructive Theory of Countable Markov Chains*. Cambridge University Press, 1995.

[8] R.G. Gallager. *Discrete Stochastic Processes*. Kluwer, 1996.

[9] M. Kaplan. A sufficient condition for the nonergodicity of a markov chain. *IEEE Transactions on Information Theory*, 25(4):470–471, July 1979.

[10] Samuel Karlin and Howard M. Taylor. *A First Course in Stochastic Processes*. Academic Press, second edition, 1975.

[11] V. G. Kulkarni. *Modeling and Analysis of Stochastic Systems*. Chapman and Hall, London, UK, 1995.

[12] Anurag Kumar, D. Manjunath, and Joy Kuri. *Communication Networking: An Analytical Approach*. Morgan-Kaufmann (an imprint of Elsevier), San Francisco, May 2004.

[13] M. Loeve. *Probability Theory I*. Springer-Verlag, fourth edition, 1977.

[14] M. Loeve. *Prabability Theory II*. Springer-Verlag, fourth edition, 1978.

[15] Athanasios Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw Hill, 1984.

[16] Sheldon M. Ross. *Applied Probability Models with Optimization Applications*. Dover, 1992.

[17] Ronald W. Wolff. *Stochastic Modelling and the Theory of Queues*. Prentice Hall, Englewood Cliffs, New Jersey, 1989.