## Contents

1 Introduction ........................................... 4
   1.1 Why study stochastic models in biology? ............ 5
   1.2 Stochastic Versus Deterministic Models ................. 5
   1.3 MATLAB ..................................................... 9

2 Discrete Time Markov Chains .......................... 11
   2.1 Discrete time Markov chains: the basic model ........ 12
       2.1.1 Examples .................................................. 15
   2.2 Simulating paths of a Discrete Time Markov Chain ... 18
   2.3 Higher Order Transition Probabilities .................. 21
   2.4 Reducibility, periodicity, and recurrence of discrete time Markov chains 24
       2.4.1 Reducibility ................................................. 24
       2.4.2 Periodicity ................................................. 27
       2.4.3 Recurrence and Transience ................................ 29
   2.5 Stationary distributions and the long time behavior of the chain . . . 34
       2.5.1 Markov chains with a finite state space . . . . . . . 37
       2.5.2 Countable Markov chains ................................. 43
   2.6 Transient behavior ........................................... 49
   2.7 Exercises ....................................................... 56

3 Discrete Time Markov Chain Models in the Biosciences .... 61
   3.1 Genetic Models .............................................. 61
       3.1.1 Mendelian Genetics ....................................... 61
       3.1.2 The Wright-Fischer Model .............................. 65
       3.1.3 Phylogenetic Distance and the Jukes-Cantor Model . . . 66
   3.2 Discrete Time Birth and Death models ................. 68
   3.3 Branching Processes ......................................... 78
       3.3.1 Terminology and notation ................................. 78
       3.3.2 Behavior of the mean ...................................... 79
       3.3.3 Probability of extinction ................................ 81
   3.4 Exercises ....................................................... 89

4 Renewal and Point processes ............................ 93
   4.1 Renewal Processes ........................................... 93
       4.1.1 The behavior of $N(t)/t$, as $t \to \infty$ .................. 95
4.1.2 The renewal reward process ........................................... 99
4.2 Point Process ................................................................. 101
  4.2.1 Preliminaries ............................................................. 101
  4.2.2 The Poisson process .................................................... 103
  4.2.3 Transformations of Poisson processes ............................... 110
4.3 Exercises ................................................................. 122

5 Continuous Time Markov Chains ........................................... 125
  5.1 Construction and Basic Definitions ................................... 125
  5.2 Explosions ..................................................................... 132
  5.3 Forward Equation, Backward Equation, and the Generator Matrix .. 135
  5.4 Stationary Distributions ..................................................... 143
    5.4.1 Classification of states .................................................. 143
    5.4.2 Invariant measures ..................................................... 143
    5.4.3 Limiting distributions and convergence ............................ 149
  5.5 The Generator, Revisited ................................................. 150
  5.6 Continuous Time Birth and Death Processes ......................... 154
  5.7 Exercises ................................................................. 159

6 Continuous Time Markov Chain Models for Biochemical Reaction
Networks ................................................................. 161
  6.1 Chemical reaction networks: basic construction ....................... 161
    6.1.1 The stochastic equations and basic terminology ............... 161
    6.1.2 Rates for the law of mass action .................................. 164
    6.1.3 Example: Gene transcription and translation .................... 165
    6.1.4 Generator of the process and the forward equations ........... 168
    6.1.5 General continuous time Markov chains built using random
time changes ................................................................. 170
  6.2 Simulation ................................................................. 173
    6.2.1 The stochastic simulation algorithm ............................... 173
    6.2.2 The next reaction method ............................................ 174
    6.2.3 Euler’s method ......................................................... 175
  6.3 Advanced topics in computing ........................................... 176
    6.3.1 Numerically approximating expectations ........................ 176
  6.4 First order reaction networks ........................................... 184
  6.5 Relationship with Deterministic Model: the Law of Large Numbers . 186
    6.5.1 Conversion of rate constants ....................................... 187
    6.5.2 The classical scaling ................................................ 188
  6.6 Brownian Motions ........................................................ 190
    6.6.1 Markov property and generator of a Brownian motion .......... 192
  6.7 Integration with Respect to Brownian Motion ........................ 193
  6.8 Diffusion and Linear Noise Approximations ............................ 198
    6.8.1 Diffusion approximation .............................................. 198
    6.8.2 Linear noise approximation ....................................... 200
  6.9 Solving stochastic differential equations numerically ............... 201
A A quick review of linear differential equations and linear difference equations

A.1 Linear differential equations

A.2 Linear Difference Equations

A.3 Exercises

B A review of probability theory

B.1 The Probability Space

B.1.1 The sample space \( \Omega \)

B.1.2 The collection of events \( \mathcal{F} \)

B.1.3 The probability measure \( P \)

B.2 Conditional Probability and Independence

B.3 Random Variables

B.3.1 Expectations of random variables

B.3.2 Variance of a random variable

B.3.3 Some common discrete random variables

B.3.4 Some common continuous random variables

B.3.5 Transformations of random variables

B.3.6 More than one random variable

B.3.7 Variance of linear combinations

B.4 Inequalities and Limit Theorems

B.4.1 Important inequalities

B.4.2 Limit theorems

B.5 Estimation via simulation: Monte Carlo

B.6 Exercises
Chapter 1

Introduction

This course is an introduction to stochastic processes, with an added focus on computational techniques and applications arising from biology. A stochastic process, denoted $X(t)$ or $X_t$, is a collection of random variables indexed by time, $t$. Most often, the time parameter $t$ will be a subset of the nonnegative integers $\{0, 1, 2, \ldots\}$, in which case it will often be denoted by $n$, or a subset of $[0, \infty)$, the nonnegative real numbers. When time is indexed by the nonnegative integers, we say it is a discrete time process, whereas when time is indexed by the nonnegative reals, we say it is a continuous time process. The process $X_t$ will take values in a state space, which can itself be either discrete (finite or countably infinite) or continuous (for example, the real line or $\mathbb{R}^d$).

The main mathematical models of study in these notes are discrete and continuous time Markov chains, branching processes, point processes, and diffusion processes (those incorporating Brownian motion). Examples of how each such model arises in the biosciences will be provided throughout. We will also spend time focusing on different simulation and computational methods for the different mathematical models. The software package of choice will be Matlab.

We will pay special attention to the continuous time Markov chain models that arise in the study of population processes, such as cellular level biochemical models. Our perspective will differ from that of other texts on the subject in that we will primarily attempt to understand these processes via the random time change representation of Kurtz [15, 28, 30, 31], which will be used to derive the relevant approximations (including the standard mass-action kinetics model for deterministic dynamics) and simulation strategies.

The study of stochastic processes requires different mathematical machinery than does the study of deterministic processes, which are often modeled via ordinary or partial differential equations. In these notes, the student is required to possess an understanding of: basic probability at the undergraduate level, basic linear algebra (for example, eigenvalues and eigenvectors), linear differential equations, and linear difference equations. With the exception of linear algebra, these topics are reviewed in the appendix. No knowledge of measure theory will be assumed.

Copyright © 2013 by David F. Anderson.
1.1 Why study stochastic models in biology?

Stochastic models have a long history in biology; however, over the past decade their use has increased dramatically. One reason for this great increase is that recent advances in experimental methods in biology, such as fluorescent proteins and laser traps, have enabled quantitative measurements at the single cell, and even single molecule, level. Such experiments show time and time again that the dynamics of the constituent species at this level behave in an intrinsically stochastic manner. The (mathematical) implication of this observation is that the standard deterministic models for dynamics need to sometimes be replaced with analogous stochastic models. However, the stochastic models should not be constructed by simply “adding noise” to the deterministic system. In fact, the opposite is true; the stochastic models should be developed via first principle arguments, and the deterministic models should arise as some limit of the stochastic system. Later in the course, we will discuss how such limiting arguments can be carried out. Cellular level biochemical processes and other population processes (such as predator-prey models) will be the main biological models studied in this course, though other models, such as branching processes and models from genetics will also be considered.

My hope. It is my hope that the students coming to the class with a mathematics and/or physical science background will gain an appreciation for some of the current mathematical challenges in biology and that the students coming to the class with more of a biological background will gain the basic mathematical tools necessary to study the models that are currently arising in their field. It is also one of my strong desires that both groups of students gain an appreciation for the other’s field, and the potential benefits of interdisciplinary collaborations. That being said, my background is one of mathematics and I make no claims to being an expert biologist. I therefore strongly encourage my biology students to correct me when I (inevitably) make an error pertaining to any of the biological content.

Finally, these notes are a work in progress and are undoubtedly littered with errors and typos. Please let me know of any such errors, or of any suggestions you have for improvements.

1.2 Stochastic Versus Deterministic Models

Before proceeding too far, it is important to gain a good understanding of the terms stochastic and deterministic. Loosely, a process is deterministic if its future is completely determined by its present and past. Examples of deterministic processes include solutions to differential and difference equations, which are reviewed in Section A.1 and A.2 of the appendix.

Example 1.2.1. The initial value problem
\[ \dot{x}(t) = 7x(t), \quad x(0) = 2, \]
has the solution \( x(t) = 2e^{7t} \). This is a linear differential equation for which knowledge of the value at the single time \( t = 0 \), determines the value of the process at all times \( t \in \mathbb{R} \).

**Example 1.2.2.** Consider the difference equation

\[
F_n = F_{n-1} + F_{n-2}, \quad \text{for } n > 2.
\]

Then, if \( F_1 = F_2 = 1 \), this is the well known Fibonacci sequence and \( \{F_n\}_{n=1}^{\infty} = \{1, 1, 2, 3, 5, 8, \ldots \} \).

On the other hand, a stochastic process is a random process evolving in time. Informally, this means that even if you have full knowledge of the state of the system and its entire past, you can not be sure of its value at future times with certainty. We will present two examples of stochastic processes, one discrete in time and the other continuous in time.

**Example 1.2.3.** Consider rolling a fair, six-sided die many times, and for \( k \in \{1, 2, \ldots \} \), let \( Z_k \) be the outcome of the \( k \)th roll. Let

\[
X_n = \sum_{k=1}^{n} Z_k.
\]

Thus, \( X_n \) is the accumulated total of the first \( n \) rolls. It should be clear that knowing \( X_1 = Z_1 = 3 \) only tells you that \( X_2 \in \{4, 5, 6, 7, 8, 9\} \), each with equal probability. Note that time, indexed here by \( n \), is discrete in that we only update the system after each roll of the die.

**Example 1.2.4.** Consider a frog sitting in a pond with \( k \) lily pads, labeled 1 through \( k \). The frog starts the day by sitting on a randomly chosen pad (for example, they could be chosen with equal probability). However, after a random amount of time (perhaps modeled by an exponential random variable), the frog will jump to another pad, also randomly chosen. We also assume that the frog is very fast and the jump happens instantaneously. Letting \( t = 0 \) denote the start of the day, we let \( X(t) \in \{1, \ldots, k\} \) denote the lily pad occupied by the frog at time \( t \). In this example, time is naturally continuous.

However, if we are only interested in which lily pad the frog is on after a given number of jumps, then we may let \( Z_n \) denote the lily pad occupied by the frog after the \( n \)th jump, with \( Z_0 \) defined to be the starting lily pad. The process \( Z_n \) is discrete in time. The processes \( Z_n \) and \( X(t) \) are clearly related, and \( Z_n \) is usually called the embedded discrete process associated with \( X(t) \). Other authors refer to \( Z_n \) as the skeletal chain of the process.

We pause to make the philosophical point that it could be argued that neither of the processes described in Examples 1.2.3 or 1.2.4 are actually random. For example, if in the context of Example 1.2.3 you know the physical state of every molecule in the die and the room you are within, as the die left your hand, you could (theoretically)
predict with certainty which number was going to appear. Similarly, the frog in Example 1.2.4 could have some decision mechanism that we have simply not been able to parse. However, the physical correctness of these statements aside, we are making a modeling choice and, for example, are effectively giving up on trying to understand the underlying dynamics of the die. Depending upon the questions one wants to ask, this may be a good or a bad, choice, but is a question clearly left to the modeler.

To varying degrees, all models used in the sciences have been devised by making such choices, with the tradeoff usually being between realistic modeling of the underlying process and analytic tractability. The famous quote from George Box comes to mind: “all models are wrong, but some are useful.” To demonstrate this point, we turn to a processes that can be modeled either stochastically or deterministically in an effort to understand why stochastic models, which are often less analytically tractable, may oftentimes be preferable to the standard deterministic models. The point is that the stochastic model may capture the actual behavior of the underlying true process much more accurately than a deterministic model could. Of course, the stochastic model itself is still an approximation to the “truth”, just a better one than the deterministic model.

Example 1.2.5 (Bacterial Growth). Consider two competing models for bacterial growth (by growth here, we mean the growth of the size of the colony, not of an individual bacteria): one deterministic and one stochastic. We suppose there are 10 bacteria at time zero. We further suppose that bacteria divide after approximately three hours into two identical daughter cells. If one were to model the growth of the colony deterministically, a reasonable model would be

\[
\dot{x}(t) = \frac{1}{3} x(t) \quad x(0) = 10,
\]  

(1.1)

with solution \( x(t) = 10e^{t/3} \), where the units of \( t \) are hours.

Without going into the finer details yet (though we will later), a reasonable choice for a stochastic model would be one in which each bacteria divides after a random amount of time with an average wait of 3 hours (more technically, I am taking this waiting time to be an exponential random variable with parameter \( 1/3 \)). It is further assumed that the amount of time that one bacteria takes to divide is independent of the amount of time it takes the other bacteria to divide. Similar to equation (1.1) for the deterministic model, it is possible to write down a system of equations describing the time evolution of such a model; however, I will postpone doing so until later in the course. Figure 1.2.1 provides a plot of the solution of the deterministic system versus three different realizations of the stochastic system. The stochastic realizations appear to follow the deterministic system in a noisy way. It is clear, however, that the behavior of a single realization of the stochastic system can not be predicted with absolute accuracy. □

Example 1.2.6 (Bacterial Growth and Death). Now suppose that we change the model described in Example 1.2.5 in that we allow bacteria to die as well as divide,
and we suppose we begin with only two bacteria. We suppose that they die (either through natural causes or via some predator or a combination of both) after about five hours. Our new deterministic model is then

$$\dot{x}(t) = \frac{1}{3} x(t) - \frac{1}{5} x(t) = \frac{2}{15} x(t), \quad x(0) = 2,$$

with solution $x(t) = 2e^{2t/15}$. Note that this is again simply exponential growth, albeit with a slower growth rate.

For the stochastic model, we now model the two possible changes to the size of the colony separately. That is, either the next event is a growth event (via a division) or a decrease event (via a death). As before, we assume an exponential waiting time between such events (technically, we make the modeling choice that the amount of time that must pass before a bacteria dies is an exponential random variable with a parameter of $1/5$). We will again wait to write down the equations governing such a process until later in the class. Figure 1.2.2 provides a plot of the solution of the deterministic system versus three different realizations of the stochastic system. We now see that the solutions of the deterministic and stochastic models behave qualitatively differently: one of the realizations of the stochastic model (i.e. one of the colonies under observation) has been completely wiped out because the two bacteria died before they could reproduce, something not possible in the deterministic modeling context!

We will discuss many examples throughout this course, however it is worth pausing and noting the importance of Example 1.2.6. Even though any reasonable person
Figure 1.2.2: Deterministic model of bacterial division and death (black curve) versus three realizations of the analogous stochastic model. As described in Example 1.2.6 would agree that a deterministic and continuous model is probably not appropriate for a model of the growth of a colony consisting of two bacteria, it was the randomness (and not the discreteness) that allowed the size of the colony to actually hit zero. We also start to see some of the different types of questions that become interesting in the stochastic context, for example:

1. For a given birth and death rate in Example 1.2.6, what is the probability that the colony will eventually die out?

2. For models in which extinction is eventually guaranteed: what is the expected amount of time before extinction?

3. If we know a stochastic processes $X_t$ neither dies out nor diverges to infinity, and if $a < b$ are real numbers, then what is the probability that the value of the process is between $a$ and $b$ for very large $t$? That is, what is

$$
\lim_{t \to \infty} \text{Prob}\{a \leq X_t \leq b\}?
$$

4. How did we make those plots in Figures 1.2.1 and 1.2.2?

### 1.3 MATLAB

In this course, we will use MATLAB as our primary software package. It is my goal to assign problems that require MATLAB on every assignment. I recognize that this
will be intimidating to some people who have never used such software packages in
the past. However, the use of simulation has become an integral part of the study of
stochastic processes that arise in the life sciences, and it is important for all students
to at least be exposed to it. To ease the transition, and hopefully flatten the learning
curve, I will provide substantial help in terms of MATLAB programming by providing
helpful hints in the assignments, even going so far as providing some full codes to use
as solutions, and templates to further problems. While I will do my best to provide
such help, it is the student’s responsibility to learn how to complete the assignments.
Thankfully, you live in an age when you can simply type any question you have into
a search engine, and an answer will be provided. For example, if you wish to learn
how to generate a Poisson random variable using MATLAB, simply type “Poisson
random variable Matlab”. As is course policy, you must of course complete your
own assignment. However, I strongly recommend that students work in groups while
completing the assignments so as to benefit from each other’s skills.
Chapter 2

Discrete Time Markov Chains

In this chapter we introduce discrete time Markov chains (DTMCs). For these models both time and space are discrete. We will begin in Section 2.1 by introducing the basic model and providing some examples. Next, in Section 2.2 we will provide an explicit probability space on which a Markov chain is defined that is particularly useful for simulation purposes. In Sections 2.3 and beyond we develop some of the basic theory of discrete time Markov chains.

Before proceeding, it is worthwhile to review more technically than in the introduction what, precisely, a stochastic process is. First, recall that for some probability space \((\Omega, \mathcal{F}, P)\), a random variable \(X\) is a real-valued function defined on the sample space \(\Omega\). That is, \(X : \Omega \to \mathbb{R}\) and \(X(\omega) \in \mathbb{R}\) for each \(\omega \in \Omega\) (see Appendix B if you need a refresher on any of this). The values that a random variable can take (i.e. its range) is called its state space.

We can now ask for probabilities associated with the random variable \(X\). For example, we may be interested in the probability that a certain random variable defined on a certain probability space takes a value between 1 and 6, inclusive. That is, we may be interested in

\[
P\{\omega : 1 \leq X(\omega) \leq 6\}.
\]  

Note that we are still asking for a probability of a subset of \(\Omega\), that is, an element of \(\mathcal{F}\). It is often the case that we are not particularly interested in the probability space itself, and are only interested in the probabilities of random variables. In this case it is typical to ignore the probability space and simply write

\[
P\{1 \leq X \leq 6\},
\]

where it is understood that the equation above is shorthand for (2.1). We take this perspective through most of these notes, with a few exceptions such as Section 2.2, in which the precise probability space is of importance.

Having reviewed what we mean by a random variable, the formal definition of a stochastic process is quite simple. A stochastic process \(\{X_t\}, t \in I\), is a collection of random variables defined on a common probability space, and where \(I\) is some index.
set. For example, if time is continuous and non-negative then it is often the case that \( I = \mathbb{R}_{\geq 0} \). When time is discrete it is often natural to take \( I = \mathbb{Z}_{\geq 0} \), the non-negative integers; in this case we often denote time by \( n \), as opposed to \( t \). As in the case of a random variable, it is technically correct to write things like \( X_t(\omega) \), where \( \omega \in \Omega \), the common probability space, and

\[
P\{\omega : X_{t_1}(\omega) \in A_1, X_{t_2}(\omega) \in A_2\},
\]

giving the probability that the random variables \( X_{t_1} \) and \( X_{t_2} \) takes values in the sets \( A_1 \in \mathcal{F} \) and \( A_2 \in \mathcal{F} \), respectively. However, we will most often instead simply write \( X_t \) and \( P\{X_{t_1} \in A_1, X_{t_2} \in A_2\} \) and largely ignore the probability space upon which the process is defined.

It is a reasonable question to ask: why ignore the probability space? The reason is that we often truly do not care about where the seeming randomness is coming from. For just one example, consider the modeling of the gambling game of roulette, in which there are 38 distinct numbers, each of which has an equal chance of being selected by a bouncing ball on a spinning wheel. In this case, the actual probability space would be nearly impossible to explicitly write down as it should involve detailed physical knowledge pertaining to the bouncing and spinning ball, the spinning wheel, the molecules in the room, etc. Instead, we only care about the fact that there is some probability space and random variable, \( X \), for which \( P\{X = i\} = 1/38 \) for each \( i \in \{1, \ldots, 38\} \). That is, we take the existence of \((\Omega, \mathcal{F}, P)\) for granted, and now work with the actual model of interest, captured by the probability mass function of the random variable \( X \).

### 2.1 Discrete time Markov chains: the basic model

Let \( X_n, n = 0, 1, 2 \ldots \), be a discrete time stochastic process with a discrete state space \( S \). For the time being, we ignore the probability space upon which the process is defined. Recall that \( S \) is said to be discrete if it is either finite or countably infinite. Without loss of generality, we will nearly always assume that \( S \) is either \( \{1, \ldots, N\} \) or \( \{0, \ldots, N-1\} \) in the finite case, and either \( \{0,1,\ldots\} \) or \( \{1,2,\ldots\} \) in the infinite setting.

For any process, it is typically of interest to know the values of

\[
P\{X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n\},
\]

where \( n \in \{0,1,2,\ldots\} \) and \( \{i_0, \ldots, i_n\} \) is some finite sequence of states in \( S \). Note that having such finite dimensional distributions allows for the calculation of any path probability. For example, by the axioms of probability

\[
P\{X_0 = i_0, X_3 = i_3\} = P\{X_0 = i_0, X_1 \in S, X_2 \in S, X_3 = i_3\}
= \sum_{j_1 \in S} \sum_{j_2 \in S} P\{X_0 = i_0, X_1 = j_1, X_2 = j_2, X_3 = i_3\},
\]

where the second equality holds as the events are mutually exclusive.
Example 2.1.1. Recall Example 1.2.3, where we let $Z_k$ be the outcome of the $k$th roll of a fair die and we let

$$X_n = \sum_{k=1}^{n} Z_k.$$ 

Then, assuming the rolls, and hence the random variables $\{Z_k\}$, are independent, we can calculate any probability of the form (2.2) for the process $X_n$. For example,

$$P\{X_1 = 2, X_2 = 4, X_3 = 6\} = P\{X_3 = 6|X_2 = 4\} P\{X_2 = 4|X_1 = 2\} P\{X_1 = 2\}$$

$$= P\{Z_3 = 2\} P\{Z_2 = 2\} P\{Z_1 = 2\}$$

$$= \left(\frac{1}{6}\right)^3.$$ 

Example 2.1.2. Suppose a frog can jump between three lily pads, labeled 1, 2, and 3. We suppose that if the frog is on lily pad number 1, it will jump next to lily pad number 2 with a probability of one. Similarly, if the frog is on lily pad number 3, it will next jump to lily pad number 2 with a probability of one. However, when the frog is on lily pad number 2, it will next jump to lily pad 1 with probability 1/4, and to lily pad three with probability 3/4. We can depict this process graphically via

$$1 \xrightarrow{1/4} 2 \xrightarrow{1} 3.$$ 

We let $X_n$ denote the position of the frog after the $n$th jump, and assume that $X_0 = 1$. We then have by the basic properties of conditional probabilities

$$P\{X_0 = 1, X_1 = 2, X_2 = 3\} = P\{X_2 = 3|X_0 = 1, X_1 = 2\} P\{X_0 = 1, X_1 = 2\}$$

$$= P\{X_2 = 3|X_1 = 2\} P\{X_1 = 2|X_0 = 1\} P\{X_0 = 1\}$$

$$= \frac{3}{4} \times 1 \times 1$$

$$= \frac{3}{4},$$

whereas

$$P\{X_0 = 1, X_1 = 3\} = 0.$$ 

Note that in Example 2.1.2, we made the intuitively straightforward assumption that $P\{X_2 = 3|X_0 = 1, X_1 = 2\} = P\{X_2 = 3|X_1 = 2\}$. That is, we noted that the frog starting in state 1 has no bearing on the probability that it would jump from state 2 (at time 1) to state 3 (at time 2). This assumption is the essence of the Markov assumption.

13
Returning to (2.2), we have

\[ P\{X_n = i_n \mid X_0 = i_0, \ldots, X_{n-1} = i_{n-1}\} = P\{X_n = i_n \mid X_{n-1} = i_{n-1}\}. \]  

(2.4)

Definition 2.1.3 says that the probabilities associated with future states only depend upon the current state, and not on the history of the process. Note that the processes described in Examples 2.1.1 and 2.1.2 are both discrete time Markov chains.

Having the definition of the object of study in hand, we turn to developing the mathematical machinery, including definitions and theoretical results, necessary for the analysis of such models. However, we pause to stress that in many, though not all, situations the fact that we are dealing with a Markov chain is an assumption being made, and not a consequence of previous analysis. For example, the mathematical step \( P\{X_2 = 3 \mid X_0 = 1, X_1 = 2\} = P\{X_2 = 3 \mid X_1 = 2\} \) we made in Example 2.1.2 was a result of the explanation of the problem itself.

Definition 2.1.4. The one-step transition probability of a Markov chain from state \( i \) to state \( j \), denoted by \( p_{ij}(n) \), is

\[ p_{ij}(n) \overset{\text{def}}{=} P\{X_{n+1} = j \mid X_n = i\}. \]

If the transition probabilities do not depend upon \( n \), in which case \( p_{ij}(n) = P\{X_1 = j \mid X_0 = i\} \), then the processes is said to be time homogeneous, or simply homogeneous, and we will use the notation \( p_{ij} \) as opposed to \( p_{ij}(n) \).

Unless explicitly stated otherwise, all discrete time Markov chain models considered in these notes will be time homogeneous.

It is a straightforward use of conditional probabilities to show that any process satisfying the Markov property (2.4) satisfies the more general condition

\[ P\{X_{n+m} = i_{n+m}, \ldots, X_n = i_n \mid X_0 = i_0, \ldots, X_{n-1} = i_{n-1}\} = P\{X_{n+m} = i_{n+m}, \ldots, X_n = i_n \mid X_{n-1} = i_{n-1}\}, \]

(2.5)

for any choice of \( n, m \geq 1 \), and states \( i_j \in S \), with \( j \in 0, \ldots, n + m \). Similarly, any Markov chain satisfies intuitively pleasing identities such as

\[ P\{X_n = i_n \mid X_{n-1} = i_{n-1}, X_0 = i_0\} = P\{X_n = i_n \mid X_{n-1} = i_{n-1}\}. \]  

(2.6)

We will denote the initial probability distribution of the process by \( \alpha \) (which we think of as a column vector):

\[ \alpha(j) = P\{X_0 = j\}, \quad j \in S. \]

Returning to (2.2), we have

\[
\begin{align*}
P\{X_0 = i_0, \ldots, X_n = i_n\} & = P\{X_n = i_n \mid X_0 = i_0, \ldots, X_{n-1} = i_{n-1}\}P\{X_0 = i_0, \ldots, X_{n-1} = i_{n-1}\} \\
& = p_{i_{n-1}i_n}P\{X_0 = i_0, \ldots, X_{n-1} = i_{n-1}\} \\
& \quad \vdots \\
& = \alpha_0 p_{i_0i_1} \cdots p_{i_{n-1}i_n}. 
\end{align*}
\]  

(2.7)
and under the Markov assumption the problem of computing probabilities has been converted to one of simple multiplication. For example, returning to Example 2.1.2, we have

\[ P\{X_0 = 1, X_1 = 2, X_2 = 3\} = \alpha_1 p_{12} p_{23} = 1 \times 1 \times \frac{3}{4} = 3/4. \]

The one-step transition probabilities are most conveniently expressed in matrix form.

**Definition 2.1.5.** The transition matrix \( P \) for a Markov chain with state space \( S = \{1, 2, \ldots, N\} \) and one-step transition probabilities \( p_{ij} \) is the \( N \times N \) matrix

\[
P = \begin{pmatrix}
p_{11} & p_{12} & \cdots & p_{1N} \\
p_{21} & p_{22} & \cdots & p_{2N} \\
\vdots & \vdots & \ddots & \vdots \\
p_{N1} & p_{N2} & \cdots & p_{NN}
\end{pmatrix}.
\]

If the state space \( S \) is infinite, then \( P \) is formally defined to be the infinite matrix with \( i, j \)th component \( p_{ij} \).

Note that the matrix \( P \) satisfies

\[
0 \leq P_{ij} \leq 1, \quad 1 \leq i, j, \leq N, \quad (2.8)
\]

\[
\sum_{j=1}^{N} P_{ij} = 1, \quad 1 \leq i \leq N. \quad (2.9)
\]

Any matrix satisfying the two conditions (2.8) and (2.9) is called a Markov or stochastic matrix, and can be the transition matrix for a Markov chain. If \( P \) also satisfies the condition

\[
\sum_{i=1}^{N} P_{ij} = 1, \quad 1 \leq j \leq N,
\]

so that the column sums are also equal to 1, then \( P \) is termed doubly stochastic.

### 2.1.1 Examples

We list examples that will be returned to throughout these notes.

**Example 2.1.6.** This example, termed the deterministically monotone Markov chain, is quite simple but will serve as a building block for more important models in the continuous time setting, in particular the Poisson process.

Consider \( X_n \) with state space \( \{1, 2, \ldots\} \), and with transition probabilities \( p_{i,i+1} = 1 \), and all others are zero. Thus, if \( \alpha \) is the initial distribution and \( \alpha_1 = 1 \), then the process simply starts at 1 and proceeds deterministically up the integers towards positive infinity. \( \square \)
Example 2.1.7. This example formalizes a computation made in Example 2.1.1. Suppose that $X_n$ are independent and identically distributed with
\[ P\{X_0 = k\} = a_k, \quad k = 0, 1, \ldots, N, \]
where $a_k \geq 0$ and $\sum_k a_k = 1$. Then,
\[ P\{X_{n+1} = i_{n+1} \mid X_0 = i_0, \ldots, X_n = i_n\} = P\{X_{n+1} = i_{n+1}\} = a_{i_{n+1}} = P\{X_{n+1} = i_{n+1} \mid X_n = i_n\}, \]
and the process is Markovian. Here
\[ P = \begin{pmatrix} a_0 & a_1 & \cdots & a_N \\ \vdots & \ddots & \vdots \\ a_0 & a_1 & \cdots & a_N \end{pmatrix}. \]

Example 2.1.8. Consider a gene that can be repressed by a protein. By $X_n = 0$, we mean the gene is free at time $n$, and by $X_n = 1$ we mean that the gene is repressed. We make the following assumptions:

1. If the gene is free at time $n$, there will be a probability of $p \geq 0$ that it is repressed at time $n + 1$.
2. If the gene is represses at time $n$, there will be a probability of $q \geq 0$ that it is free at time $n + 1$.

In this setting $X_n$ can be modeled as a discrete time Markov chain with finite state space $S = \{0, 1\}$. The transition matrix is
\[ P = \begin{bmatrix} 1 - p & p \\ q & 1 - q \end{bmatrix}, \quad (2.10) \]
where the first row/column is associated with state 0. Note that any two state discrete time Markov chain has a transition matrix of the form (2.10).

Example 2.1.9 (Random walk with finite state space). A random walk is a model used to describe the motion of an entity, the walker, on some discrete space. Taking our state space to be $\{0, \ldots, N\}$, for some $N > 0$, we think of the walker flipping a coin to decide whether or not to move to the right or left during the next move. That is, at each time-step the walker moves one step to the right with probability $p$ (she flipped a heads) and to the left with probability $1 - p$ (she flipped a tails). If $p = 1/2$, the walk is termed symmetric or unbiased, whereas if $p \neq 1/2$, the walk is biased. The one step transition intensities for $i \in \{1, \ldots, N - 1\}$ are,
\[ p_{i,i+1} = p, \quad p_{i,i-1} = 1 - p, \quad 0 < i < N, \]
though we must still give the transition intensities at the boundaries. One choice for the boundary conditions would be to assume that with probability one, the walker transitions away from the boundary during the next time step. That is, we could have

\[ p_{01} = 1, \quad p_{N,N-1} = 1. \]

We say such a process has *reflecting boundaries*. Note that Example 2.1.2 was a model of a random walk on \{1, 2, 3\} with reflecting boundaries. Another option for the boundary conditions is to assume there is *absorption*, yielding the boundary conditions

\[ p_{00} = 1, \quad p_{NN} = 1, \]

in which case the chain is often called the *Gambler’s Ruin*, which can be understood by assuming \( p < 1/2 \). Finally, we could have a partial type of reflection

\[ p_{00} = 1 - p, \quad p_{01} = p, \quad p_{N,N-1} = 1 - p, \quad p_{NN} = p. \]

Of course, we could also have any combination of the above conditions at the different boundary points. We could also generalize the model to allow for the possibility of the walker choosing to stay at a given site \( i \in \{1, \ldots, N - 1\} \) during a time interval, and not moving either left or right for at least one turn.

In the most general case, we could let \( q_i, p_i \) and \( r_i \) be the probabilities that the walker moves to the left, right, and stays put given that she is in state \( i \). Assuming absorbing boundary conditions, the transition matrix for this model is

\[
P = \begin{pmatrix}
1 & 0 & 0 & 0 & \cdots & 0 & 0 \\
q_1 & r_1 & p_1 & 0 & \cdots & 0 & 0 \\
0 & q_2 & r_2 & p_2 & \cdots & 0 & 0 \\
\vdots & \ddots & \ddots & \ddots & \ddots & \vdots & \vdots \\
0 & \cdots & 0 & q_{N-1} & r_{N-1} & p_{N-1} & 0 \\
0 & 0 & 0 & 0 & \cdots & 0 & 1
\end{pmatrix},
\]

where it is understood that \( q_i, p_i, r_i \geq 0 \), and \( q_i + p_i + r_i = 1 \) for all \( i \in \{1, \ldots, N - 1\} \).

**Example 2.1.10** (Axonal transport). One method of transport used in living cells is axonal transport in which certain (motor) proteins carry cargo such as mitochondria, other proteins, and other cell parts, on long microtubules. These microtubule can be thought of as the “tracks” of the transportation mechanism, with the motor protein as the random walker. One natural, and simple, model for such transport would begin by breaking the microtubule into \( N \) equally sized intervals, and then letting \( X_n \) be the position of the motor protein on the state space \( \{1, \ldots, N\} \). We could then let the transition probabilities satisfy

\[ p_{i,i+1} = p_i, \quad p_{i,i-1} = q_i, \quad p_{i,i} = r_i, \quad i \in \{2, \ldots, N-1\}, \]

\[ p_{1,0} = 1, \quad p_{0,1} = 1, \quad p_{N,N} = 1. \]
where \( p_i + q_i + r_i = 1 \) with \( p_i, q_i, r_i \geq 0 \), and with boundary conditions

\[
p_{1,1} = p_1 + q_1, \quad p_{1,2} = r_1, \quad p_{N,N} = 1,
\]

where we think of the end of the microtubule associated with state \( N \) as the destination of the cargo. In this case, it would be natural to expect \( p_i > q_i \). □

Example 2.1.11 (Random walk on the integers). This Markov chain is like that of Example 2.1.9, except now we assume that the state space is all the integers \( S = \mathbb{Z} = \{\ldots, -1, 0, 1, \ldots\} \). That is, \( X_n \) is the position of the walker at time \( n \), where for some \( 0 \leq p \leq 1 \) the transition probabilities are given by

\[
p_{i,i+1} = p, \quad p_{i,i-1} = 1 - p,
\]

for all \( i \in S \). This model is one of the most studied stochastic processes and will be returned to frequently as a canonical example. □

Example 2.1.12 (Random walk on \( \mathbb{Z}^d \)). We let \( \mathbb{Z}^d \) be the \( d \)-dimensional integer lattice:

\[
\mathbb{Z}^d = \{(x_1, \ldots, x_d) : x_i \in \mathbb{Z}\}.
\]

Note that for each \( x \in \mathbb{Z}^d \) there are exactly \( 2d \) values \( y \) with \( |x - y| = 1 \) (as there are precisely \( d \) components that can be changed by a value of \( \pm 1 \)). We may let

\[
p_{xy} = \begin{cases} 
\frac{1}{2d} & \text{if } |x - y| = 1 \\
0 & \text{else}
\end{cases}.
\]

□

2.2 Simulating paths of a Discrete Time Markov Chain

We turn to the problem of simulating a discrete time Markov chain with a given initial distribution, \( \alpha \), and transition matrix, \( P \). More explicitly, for the discrete set \( S = \{1, 2, \ldots, N\} \), where \( N \) could be infinity, we assume the existence of:

(i) An initial distribution \( \alpha = \{\alpha_k\} \) giving the associated probabilities for the random variable \( X_0 \). That is, for \( k \in S \),

\[
\alpha_k = P\{X_0 = k\}.
\]

(ii) Transition probabilities, \( p_{ij} \), giving the desired probability of transitioning from state \( i \in S \) to state \( j \in S \):

\[
p_{ij} = P\{X_{n+1} = j \mid X_n = i\}.
\]
Note that we will require that \( \alpha \) is a probability vector in that \( \alpha_i \geq 0 \) for each \( i \) and
\[
\sum_{i \in S} \alpha_i = 1.
\]
We further require that for all \( i \in S \)
\[
\sum_{j \in S} p_{ij} = 1,
\]
which simply says that the chain will transition somewhere from state \( i \) (including the possibility that the chain transitions back to state \( i \)).

Given \( \alpha \) and the transition probabilities \( \{ p_{ij} \} \), it will be useful for the time being to be explicit about the probability space we define our stochastic process on. We let \( \Omega = \{ (\omega_0, \omega_1, \ldots) : \omega_i \geq 0 \} \) consist of sequences of independent uniform \([0, 1]\) random variables. That is, each element \( \omega \in \Omega \) consists of an infinite stream of independent uniform\([0, 1]\) random variables. We further assume that the \( \omega \in \Omega \) themselves are independent. We let \( (\Omega, \mathcal{F}, P) \) denote this probability space, and simply mention that it is an advanced topic in probability theory that such a probability space even exists (Math 733 would be the appropriate course for students at the University of Wisconsin who are interested in this topic).

We will now define our DTMC with initial probability vector \( \alpha \) and transition rates \( p_{ij} \) on this specific probability space, and confirm it has the desired Markov property. Will this give us a simulation strategy? The answer is yes because modern software, such as MATLAB, R, Octave, etc., all have built in random number generators that can efficiently produce sequences of independent uniform random variables, which we have simply denoted here by \( \omega \). Hence, if we describe below how to turn a realization, \( \omega \), into a realization of a stochastic process \( X_n \), then we have a method for simulating realizations of the DTMC on a computer. It is worth explicitly noting that we have reduced the question of how to simulate a discrete time Markov chain using a computer to the mathematical question of how to produce a function \( f \) with \( \omega \) as input and \( X_n \), a path of a DTMC, as output.

**The construction.** We let \( \omega = \{ \omega_0, \omega_1, \ldots \} \) be independent random variables that are uniformly distributed on the interval \((0, 1)\). The plan is the following. We will use the initial distribution, \( \alpha \), to produce \( X_0 \) from \( \omega_0 \), and then for \( n \geq 1 \), we will use the transition matrix to produce \( X_n \) from the pair \((X_{n-1}, \omega_n)\). Note, therefore, that each choice of sequence of uniform random variables \( \{ \omega_0, \omega_1, \ldots \} \) will correspond with a single path of the process \( X_n, n \geq 0 \).

Specifically, we generate \( X_0 \) from \( \omega_0 \) using the transformation method detailed in Theorem B.3.22. Next, we note that,
\[
P\{X_1 = j \mid X_0 = i\} = p_{ij}.
\]
Therefore, conditioned upon \( X_0 = i \), the random variable \( X_1 \) has probability mass function determined by the \( i \)th row of the transition matrix \( P \). We may then use
Theorem B.3.22 again to generate \( X_1 \) using only \( \omega_1 \). Continuing in this manner constructs the Markov chain \( X_n \).

It is straightforward to verify that the constructed process is a discrete time Markov chain with the desired transition probabilities. Using that \( X_{n+1} \) is simply a function of \( X_n \) and \( \omega_{n+1} \), that is \( X_{n+1} = f(X_n, \omega_{n+1}) \), and that by construction \( X_0, \ldots, X_n \) are independent of \( \omega_{n+1} \), we have

\[
P\{X_{n+1} = j \mid X_0 = i_0, \ldots, X_{n-1} = i_{n-1}, X_n = i\}
\]

\[
= P\{f(X_n, \omega_{n+1}) = j \mid X_0 = i_0, \ldots, X_{n-1} = i_{n-1}, X_n = i\} \tag{definition of \( f \)}
\]

\[
= P\{f(i, \omega_{n+1}) = j \mid X_0 = i_0, \ldots, X_{n-1} = i_{n-1}, X_n = i\} \tag{using conditioning}
\]

\[
= P\{f(i, \omega_{n+1}) = j \} \tag{independence}
\]

\[
= p_{ij}.
\]

We present the formal algorithm for the generation of the desired DTMC below. The state space is assumed to be \( S = \{0, 1, 2, \ldots, N\} \), where, again, \( N \) could be infinity.

**Algorithm 1** (Simulating a discrete time Markov chain). Take as input a probability vector \( \alpha \) and transition rates \( p_{ij} \).

1. Generate an independent uniform\([0, 1]\) random variable \( \omega_0 \). Find the smallest \( m \in \{0, 1, \ldots, N\} \) satisfying

\[
\sum_{j=0}^{m} \alpha_j \geq \omega_0,
\]

and set \( X_0 = m \).

2. Let \( k = 1 \) and perform the following steps repeatedly (i.e. until a decision to stop is made):

   a. Generate a uniform\([0, 1]\) random variable \( \omega_k \) that is independent of all previous random variables generated \( \{\omega_0, \ldots, \omega_{k-1}\} \).

   b. Find the smallest \( m \in \{0, 1, \ldots, N\} \) satisfying

\[
\sum_{j=0}^{m} p_{X_{k-1}j} \geq \omega_k,
\]

   where \( p_{X_{k-1}j} \) is the probability of transitioning to state \( j \) given the chain is in state \( X_{k-1} \).

   c. Set \( X_k = m \).

   d. Set \( k \leftarrow k + 1 \).

**Example 2.2.1.** Suppose that the initial distribution is \( \alpha = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}) \), and the transition matrix for the DTMC on the state space \( S = \{0, 1, 2\} \) is

\[
P = \begin{pmatrix}
0 & 1 & 2 \\
3 & 3 & 3 \\
2 & 1 & 1 & 1 & 0 \\
\end{pmatrix}
\]

20
Then the random sequence \( \omega = (0.814, 0.906, 0.127, 0.632, \ldots) \) produces the realization
\[
X_0 = 2, \quad X_1 = 2, \quad X_2 = 0, \quad X_3 = 2, \ldots.
\]
whereas the sequence \( \omega = (0.096, 0.278, 0.547, 0.958, \ldots) \) produces the realization
\[
X_0 = 0, \quad X_1 = 1, \quad X_2 = 1, \quad X_3 = 1, \ldots.
\]

\[\square\]

### 2.3 Higher Order Transition Probabilities

Having in hand a basic concept of a discrete time Markov chain, together with a simulation strategy, we turn to the question of: what can we learn from a Markov model? For example, consider again the frog in Example 2.1.2. Perhaps we would like to have some sense of where the frog will be at time \( n = 2 \), or even \( n = 200 \). We could simulate the behavior of the frog multiple times using the methods of the previous section, but this seems uninspired and would not allow us to conclude anything with surety. Something more is desired.

Leaving the example of the frog, we ask one of the most basic questions possible of a discrete time Markov chain: given an initial distribution \( \alpha \), and a transition matrix \( P \), what is the probability that the Markov chain will be in state \( i \in S \) at time \( n \geq 0 \)?

Note that the question only has content. To begin answering this question we have the following definition.

**Definition 2.3.1.** The \( n \)-step transition probability, denoted \( p_{ij}^{(n)} \), is the probability of moving from state \( i \) to state \( j \) in \( n \) steps,
\[
p_{ij}^{(n)} := P\{X_n = j \mid X_0 = i\} = P\{X_{n+k} = j \mid X_k = i\},
\]
where the final equality is a consequence of time homogeneity.

The \( n \)-step transition probabilities are intimately linked with powers of the transition matrix. Hence, we let \( P_{ij}^n \) denote the \( i, j \)th entry of the matrix \( P^n \). We note that if the state space is infinite, then we formally have that
\[
P_{ij}^{2} = \sum_{k \in S} p_{ik}p_{kj},
\]
which converges since \( \sum_k p_{ik}p_{kj} \leq \sum_k p_{ik} = 1 \), with similar expressions for \( P_{ij}^n \). The following is one of the most useful results in the study of discrete time Markov chains, and is the reason much of their study reduces to linear algebra. (This points out that if you feel weak in your linear algebra, you should brush up on it.)

**Proposition 2.3.2.** For all \( n \geq 0 \) and \( i, j \in S \),
\[
p_{ij}^{(n)} = P_{ij}^n.
\]

21
Proof. We will show the result by induction on \( n \). First, note that the cases \( n = 0 \) and \( n = 1 \) follow by definition. Next, assuming the result is true for a given \( n \geq 1 \), we have

\[
P_{ij}^{(n+1)} = P\{X_{n+1} = j \mid X_0 = i\} \quad \text{(definition)}
\]

\[
= P\{X_{n+1} = j, X_n \in S \mid X_0 = i\} \quad \text{(} X_n \text{ is somewhere)}
\]

\[
= \sum_{k \in S} P\{X_{n+1} = j, X_n = k \mid X_0 = i\} \quad \text{(prop. of probabilities)}
\]

\[
= \sum_{k \in S} P^{(n)}_{ik} P_{kj} \quad \text{(prop. of cond. prob.)}
\]

\[
= \sum_{k \in S} P^n_{ik} P_{kj},
\]

where the final equality is our inductive hypothesis. The last term is the \( i,j \)th entry of \( P^{n+1} \), and the proof is complete. \( \square \)

We note that a slight generalization of the above computation yields

\[
P_{ij}^{m+n} = \sum_{k \in S} P^{(m)}_{ik} P^{(n)}_{kj}, \quad \text{(2.11)}
\]

for all \( i, j \in S \), and \( m, n \geq 0 \). These are usually called the Chapman-Kolmogorov equations, and they have an intuitive interpretation: the chain must be somewhere after \( m \) steps, and we are simply summing over the associated probabilities. Note that the Chapman-Kolmogorov equations is the probabilistic version of the well known matrix identity

\[
P^{m+n} = P^m P^n.
\]

We may now answer our original question pertaining to the probability that the Markov chain will be in state \( i \in S \) at time \( n \geq 0 \) for a given initial distribution \( \alpha \). Writing \( P_\alpha \) for the probabilities associated with a chain with initial distribution \( \alpha \), we have

\[
P_\alpha\{X_n = i\} = \sum_{k \in S} P_\alpha\{X_n = i \mid X_0 = k\} P_\alpha\{X_0 = k\}
\]

\[
= \sum_{k \in S} \alpha_k P^n_{ki} \quad \text{(2.12)}
\]

Thus, calculating higher order transition probabilities is computationally equivalent to computing powers of the transition matrix.
Example 2.3.3. Consider again Example 2.1.8 pertaining to the gene that can be repressed. Suppose that \( p = \frac{1}{3} \) and \( q = \frac{1}{8} \) and we know that the gene is unbound at time 0, and so

\[
\alpha = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.
\]

Suppose we want to know the probability that the gene is unbound at time \( n = 4 \). We have

\[
P = \begin{bmatrix} 2/3 & 1/3 \\ 1/8 & 7/8 \end{bmatrix},
\]

and so

\[
P^4 = \begin{bmatrix} .33533 & .66467 \\ .24925 & .75075 \end{bmatrix},
\]

and

\[
\alpha^T P^4 = [.33533, .66467].
\]

Thus, the desired probability is .33533.

A natural question, and the focus of Section 2.5, is the following: for large \( n \), what are the values \( P\{X_n = i\} \), for \( i \in S \). That is, after a very long time what are the probabilities of being in different states. By Proposition 2.3.2, we see that this question, at least in the case of a finite state space, can be understood simply through matrix multiplication.

For example, suppose that \( X_n \) is a two-state Markov chain with transition matrix

\[
P = \begin{bmatrix} 2/3 & 1/3 \\ 1/8 & 7/8 \end{bmatrix}.
\]

It is easy to check with a computer, or linear algebra, that for very large \( n \),

\[
P^n \approx \begin{bmatrix} 3/11 & 8/11 \\ 3/11 & 8/11 \end{bmatrix} \overset{\text{def}}{=} \Pi.
\]

Note that the rows of \( \Pi \) are identical and equal to \( \pi^T = [3/11, 8/11] \). Therefore, if \( v \) is a probability vector (that is, a column vector with non-negative elements that sum to one, think of it as an initial distribution), we see that

\[
\lim_{n \to \infty} v^T P^n = v^T \Pi = \pi^T.
\]

Therefore, for this example we may conclude that

\[
\lim_{n \to \infty} P\{X_n = 1\} = \frac{3}{11}, \quad \text{and} \quad \lim_{n \to \infty} P\{X_n = 2\} = \frac{8}{11},
\]

no matter the initial distribution.

Such a vector \( \pi \) will eventually be termed a stationary, or invariant, distribution of the process, and is usually of great interest to anyone wishing to understand the underlying model. Natural questions now include: does every discrete time Markov
chain have such a stationary distribution? If so, is it unique? Can we quantify how long it takes to converge to a stationary distribution? To answer these questions we need more terminology and mathematical machinery that will be developed in the next section. We will return to these questions in Section 2.5.

2.4 Reducibility, periodicity, and recurrence of discrete time Markov chains

Our eventual goal is to be able to answer the questions posed at the end of the last section. We will do so in Section 2.5. In this section, we discuss three different aspects of the Markov chain: reducibility, periodicity, and recurrence. These aspects determine the answers to the above questions. Each aspect is quite intuitive: (i) reducibility captures whether or not the Markov chain consists of one or more pieces, (ii) periodicity captures whether or not the chain is bound to be in certain states at certain times (for example, the even numbers at even times and the odd numbers at odd times), and (iii) recurrence captures whether or not the chain will return to a given state over and over again.

2.4.1 Reducibility

Suppose that $X_n$ is a Markov chain with state space $S = \{1, 2, 3, 4\}$ and transition matrix

$$
P = \begin{bmatrix}
1/2 & 1/2 & 0 & 0 \\
1/3 & 2/3 & 0 & 0 \\
0 & 0 & 1/3 & 2/3 \\
0 & 0 & 3/4 & 1/4 \\
\end{bmatrix}.
$$

Note that if the chain starts in either state 1 or 2, then it will remain in $\{1, 2\}$ for all time, whereas if the chain starts in state 3 or 4, it will remain in $\{3, 4\}$ for all time. It seems natural to study this chain by analyzing the reduced chains consisting of state spaces $S_1 = \{1, 2\}$ and $S_2 = \{3, 4\}$ and transition matrices

$$
P_1 = \begin{bmatrix}
1/2 & 1/2 \\
1/3 & 2/3 \\
\end{bmatrix}, \quad P_2 = \begin{bmatrix}
1/3 & 2/3 \\
3/4 & 1/4 \\
\end{bmatrix}.
$$

separately.

If instead the original transition matrix was

$$
P = \begin{bmatrix}
1/2 & 1/4 & 1/4 & 0 \\
1/3 & 2/3 & 0 & 0 \\
0 & 0 & 1/3 & 2/3 \\
0 & 0 & 3/4 & 1/4 \\
\end{bmatrix},
$$

1The answers are: no, sometimes, yes.
then it should be at least intuitively clear that even if \( X_0 \in \{1, 2\} \), the chain will eventually move to the states \( \{3, 4\} \) as every time the chain enters state 1, it has a probability of 0.25 of next transitioning to state 3. Once such a transition occurs, the chain remains in the states \( \{3, 4\} \) for all time. This intuition will be shown to be true later in the notes. For this example, if only the probabilities associated with very large \( n \) are desired, then it seems natural to only consider the reduced chain consisting of states \( \{3, 4\} \).

The following definitions describe when chains can be so reduced.

**Definition 2.4.1.** The state \( j \in S \) is *accessible* from \( i \in S \), and we write \( i \rightarrow j \), if there is an \( n \geq 0 \) such that

\[
p^{(n)}_{ij} > 0.
\]

That is, \( j \) is accessible from \( i \) if there is a positive probability of the chain hitting \( j \) if it starts in \( i \).

For example, for the chain with transition matrix (2.13) we have the relations \( 1 \rightarrow 2, 2 \rightarrow 1, 3 \rightarrow 4, \) and \( 4 \rightarrow 3 \), together with all the relations \( i \rightarrow i \) (since \( n = 0 \) is allowed in Definition 2.4.1). However, for the chain with transition matrix (2.14), we have all the relations \( i \rightarrow i \) and

- \( 1 \rightarrow 2, 1 \rightarrow 3, 1 \rightarrow 4, \)
- \( 2 \rightarrow 1, 2 \rightarrow 3, 2 \rightarrow 4, \)
- \( 3 \rightarrow 4, \)
- \( 4 \rightarrow 3, \)

which can be seen from the fact that

\[
P^4 = \begin{bmatrix}
19/72 & 5/18 & 5/18 & 13/72 \\
10/27 & 97/216 & 1/8 & 1/18 \\
0 & 0 & 107/216 & 109/216 \\
0 & 0 & 109/192 & 83/192
\end{bmatrix},
\]

combined with the fact that the bottom left \( 2 \times 2 \) sub-matrix of \( P^n \) will always consist entirely of zeros. (Why? You should be able to figure this out!)

**Definition 2.4.2.** States \( i, j \in S \) of a Markov chain *communicate* with each other, and we write \( i \leftrightarrow j \), if \( i \rightarrow j \) and \( j \rightarrow i \).

It is straightforward to verify that the relation \( \leftrightarrow \) is

1. Reflexive: \( i \leftrightarrow i \).
2. Symmetric: \( i \leftrightarrow j \) implies \( j \leftrightarrow i \).
3. Transitive: \( i \leftrightarrow j \) and \( j \leftrightarrow k \) implies \( i \leftrightarrow k \).
Only the third condition need be checked, and it follows from the Chapman-Kolmogorov equations (2.11): Since \( i \rightarrow j \), there is an \( n \geq 0 \) such that \( p^{(n)}_{ij} > 0 \). Since \( j \rightarrow k \), there is an \( m \geq 0 \) such that \( p^{(m)}_{jk} > 0 \). Therefore, by (2.11)
\[
p^{n+m}_{ik} = \sum_{\ell} p^{(n)}_{i\ell} p^{(m)}_{\ell k} \geq p^{(n)}_{ij} p^{(m)}_{jk} > 0,
\]
and \( i \rightarrow k \).

We may now decompose the state space using the relation \( \leftrightarrow \) into disjoint equivalence classes called communication classes. For example, the Markov chain with transition matrix (2.13) has two communication classes \( \{1, 2\} \) and \( \{3, 4\} \). Also, the Markov chain with transition matrix (2.14) has the same communication classes: \( \{1, 2\} \) and \( \{3, 4\} \). For the deterministically monotone process of Example 2.1.6, each singleton \( \{i\} \), \( i \geq 0 \), is its own communication class (remember that \( n = 0 \) is allowed in Definition 2.4.1). For the symmetric random walk of Example 2.1.9 with absorbing boundaries (the Gambler’s Ruin problem) the communication classes are \( \{0\}, \{N\} \), and \( \{1, \ldots, N-1\} \), whereas for the symmetric random walk with reflecting boundaries the only communication class is the entire state space \( \{0, \ldots, N\} \). For the random walk on the integer lattice \( \mathbb{Z}^d \) described in Example 2.1.12, the only communication class is all of \( \mathbb{Z}^d \).

**Definition 2.4.3.** A Markov chain is **irreducible** if there is only one communication class. That is, if \( i \leftrightarrow j \) for all \( i, j \in S \). Otherwise, it is called **reducible**.

Having the notion of irreducibility in hand, we turn to a related concept, that of a closed subset of the state space. Consider again the Markov chains with transition matrices (2.13) and (2.14), each of which have communication classes \( \{1, 2\} \) and \( \{3, 4\} \). For both chains, the set of states \( \{1, 2\} \) is a communication class. However, it should be clear that the behavior of the chains on \( \{1, 2\} \) are quite different as the chain with transition matrix (2.14) will eventually leave those states (assuming it starts there), and never return.

**Definition 2.4.4.** A subset of the state space \( C \subset S \), is said to be **closed** if it is impossible to reach any state outside of \( C \) from any state in \( C \) via one-step transitions. That is, \( C \) is closed if \( p_{ij} = 0 \) for all \( i \in C \) and \( j \notin C \). We say that the state \( j \) is absorbing if \( \{j\} \) is closed.

The set \( \{1, 2\} \) is closed for the chain with transition matrix (2.13), whereas it is not for that with transition matrix (2.14). However, the set \( \{3, 4\} \) is closed for both. For the deterministically monotone system, the subset \( \{n, n+1, n+2, \ldots\} \) is closed for any \( n \geq 0 \). For the Gambler’s ruin problem of random walk on \( \{0, \ldots, N\} \) with absorbing boundary conditions, only \( \{0\} \) and \( \{N\} \) are closed, and hence the states 0 and \( N \) are absorbing.

It is important to note that if \( C \subset S \) is closed, then the matrix with elements \( p_{ij} \) for \( i, j \in C \), that is the transition matrix restricted to \( C \), is also a stochastic matrix.
This fact follows because for any \( i \in C \),
\[
\sum_{j \in C} p_{ij} = 1, \quad \text{and} \quad \sum_{j \in C^c} p_{ij} = 0.
\]

Therefore, if we restrict our attention to any closed subset of the state space, we can treat the resulting model as a discrete time Markov chain itself. The most interesting subsets will be those that are both closed and irreducible: for example the subset \{3, 4\} of the Markov chain with transition matrix (2.13) or (2.14), which for either model is a two-state Markov chain with transition matrix
\[
P_2 = \begin{bmatrix} 1/3 & 2/3 \\ 3/4 & 1/4 \end{bmatrix}.
\]

### 2.4.2 Periodicity

Periodicity helps us understand the possible motion of a discrete time Markov chain. As a canonical example, consider the random walker of Example 2.1.9 with state space \( S = \{0, 1, 2, 3, 4\} \) and reflecting boundary conditions. Note that if this chain starts in state \( i \in S \), it can only return to state \( i \) on even times.

For another example, consider the Markov chain on \( \{0, 1, 2\} \) with
\[
p_{01} = p_{12} = p_{20} = 1.
\]
Thus, the chain deterministically moves from state 0 to state 1, then to state 2, then back to 0, etc. Here, if the chain starts in state \( i \), it can (and will) only return to state \( i \) at times that are multiples of 3.

On the other hand, consider the random walk on \( S = \{0, 1, 2, 3, 4\} \) with boundary conditions
\[
p_{00} = 1/2, \quad p_{01} = 1/2, \quad \text{and} \quad p_{43} = 1.
\]
In this case, if the chain starts at state 0, there is no condition similar to those above on the times that the chain can return to state 0.

**Definition 2.4.5.** The period of state \( i \in S \) is
\[
d(i) = \gcd\{n \geq 1 : p_i^{(n)} > 0\},
\]
where \( \gcd \) stands for greatest common divisor. If \( \{n \geq 1 : p_i^{(n)} > 0\} = \emptyset \), we take \( d(i) = 1 \). If \( d(i) = 1 \), we say that \( i \) is *aperiodic*, and if \( d(i) > 1 \), we say that \( i \) is periodic with a period of \( d(i) \).

The proof of the following theorem can be found in either [32, Chapter 1] or [37, Chapter 2].

**Theorem 2.4.6.** Let \( X_n \) be a Markov chain with state space \( S \). If \( i, j \in S \) are in the same communication class, then \( d(i) = d(j) \). That is, they have the same period.

---

\(^2\)This happens, for example, for the deterministically monotone chain of Example 2.1.6.
Therefore, we may speak of the period of a communication class, and if the chain is irreducible, we may speak of the period of the Markov chain itself. Any property which necessarily holds for all states in a communication class is called a class property. Periodicity is, therefore, the first class property we have seen, though recurrence and transience, which are discussed in the next section, are also important class properties.

Periodicity is often obvious when powers of the transition matrix are taken.

**Example 2.4.7.** Consider a random walk on \( \{0, 1, 2, 3\} \) with reflecting boundary conditions. This chain is periodic with a period of two. Further, we have

\[
P = \begin{bmatrix}
0 & 1 & 0 & 0 \\
1/2 & 0 & 1/2 & 0 \\
0 & 1/2 & 0 & 1/2 \\
0 & 0 & 1 & 0 \\
\end{bmatrix},
\]

and for any \( n \geq 1 \),

\[
P^{2n} = \begin{bmatrix}
* & 0 & * & 0 \\
0 & * & 0 & * \\
* & 0 & * & 0 \\
0 & * & 0 & * \\
\end{bmatrix}, \quad \text{and} \quad P^{2n+1} = \begin{bmatrix}
0 & * & 0 & * \\
* & 0 & * & 0 \\
0 & * & 0 & * \\
* & 0 & * & 0 \\
\end{bmatrix},
\]

where * is a generic placeholder for a positive number.

**Example 2.4.8.** Consider the random walk on \( S = \{0, 1, 2, 3, 4\} \) with boundary conditions

\[
p_{0,0} = 1/2, \quad p_{0,1} = 1/2, \quad \text{and} \quad p_{4,3} = 1.
\]

The transition matrix is

\[
P = \begin{bmatrix}
1/2 & 1/2 & 0 & 0 & 0 \\
1/2 & 0 & 1/2 & 0 & 0 \\
0 & 1/2 & 0 & 1/2 & 0 \\
0 & 0 & 1/2 & 0 & 1/2 \\
0 & 0 & 0 & 1 & 0 \\
\end{bmatrix},
\]

and

\[
P^8 = \begin{bmatrix}
71 & 57 & 1 & 9 & 7 \\
256 & 256 & 4 & 64 & 64 \\
57 & 39 & 29 & 21 & 1 \\
256 & 128 & 256 & 64 & 32 \\
7 & 21 & 9 & 63 & 1 \\
64 & 64 & 256 & 128 & 256 \\
7 & 32 & 1 & 7 & 35 \\
32 & 16 & 16 & 128 & 128 \\
\end{bmatrix},
\]

showing that \( d(i) = 1 \) for each \( i \in S \).
In the previous example, we used the basic fact that if each element of \( P^n \) is positive for some \( n \geq 1 \), then \( P^{n+k} \) has strictly positive elements for all \( k \geq 0 \). This follows because (i) each element of \( P \) is nonnegative, (ii) the rows of \( P \) sum to one, and (iii) \( P^{n+k} = PP^{n+k-1} \).

### 2.4.3 Recurrence and Transience

A state \( i \in S \) of a Markov chain will be called recurrent if after every visit to state \( i \), the chain will eventually return for another visit with a probability of one. Otherwise, we will call the state transient. More formally, we begin by fixing a state \( i \in S \) and then defining the probability measure \( P_i \) by

\[
P_i\{A\} \overset{\text{def}}{=} P\{A|X_0 = i\}, \quad A \in \mathcal{F}.
\]

(Note that we have abused notation slightly in that we should technically write \( P_{e_i} \) for the above probability measure, where \( e_i \) is the probability vector with a one in the \( i \)th component. For ease of notation, we instead choose to write \( P_i \).) We let \( \mathbb{E}_i \) be the expected value associated with the probability measure \( P_i \). Let \( \tau_i \) denote the first return time to state \( i \),

\[
\tau_i \overset{\text{def}}{=} \min\{n \geq 1 : X_n = i\},
\]

where we take \( \tau_i = \infty \) if the chain never returns.

**Definition 2.4.9.** The state \( i \in S \) is **recurrent** if

\[
P_i\{\tau_i < \infty\} = 1,
\]

and **transient** if \( P_i\{\tau_i < \infty\} < 1 \). Note that an equivalent formulation for transience is that \( P_i\{\tau_i = \infty\} > 0 \).

It should be at least intuitive at this point that there is a large difference between recurrent and transient states. In essence, recurrent states will be those that the chain spends most of its time in, whereas transient states will, in the long run, be left and never returned to. We make these intuitive ideas mathematically precise below, and we develop mathematical machinery which will allow us to discover which states are recurrent, and which are transient.

To begin our study of the difference between a recurrent and transient state we let

\[
R = \sum_{n=0}^{\infty} 1_{\{X_n = i\}}
\]

denote the random variable giving the total number of times the chain returns to state \( i \). Computing the expectation of \( R \) we see that

\[
\mathbb{E}_i R = \sum_{n=0}^{\infty} P_i\{X_n = i\} = \sum_{n=0}^{\infty} p^{(n)}_{ii},
\]

29
where we made use of the fact that the expectation of an indicator function is the probability of the event being indicated upon. Suppose that the chain is transient and let

\[ p \overset{\text{def}}{=} P_i \{ \tau_i < \infty \} < 1. \]

The random variable \( R \) is geometric with parameter \( 1 - p > 0 \). That is, for \( k \geq 1 \)

\[ P_i \{ R = 1 \} = 1 - p, \quad P_i \{ R = 2 \} = p(1 - p), \quad \ldots \quad P_i \{ R = k \} = p^{k-1}(1 - p). \]

Therefore,

\[
\mathbb{E}_i R = \sum_{k} k P_i \{ R = k \} = \sum_{k=1}^{\infty} kp^{k-1}(1 - p) = \frac{1}{1 - p} < \infty.
\]

Note that equation (2.15) also shows that if the chain is transient, then

\[ P_i \{ R = \infty \} = 0, \]

for otherwise we would have \( \mathbb{E}_i R = \infty \). Hence, we may conclude that there is, with a probability of one, a last time the chain visits the site \( i \). Similarly, if state \( i \) is recurrent, then \( P_i \{ R = \infty \} = 1 \) and \( \mathbb{E}_i R = \infty \). Combining the above yields the following.

**Theorem 2.4.10.** A state \( i \) is transient if and only if the expected number of returns is finite, which occurs if and only if

\[ \sum_{n=0}^{\infty} p^{(n)}_{ii} < \infty. \]

Further, if \( i \) is recurrent, then with a probability of one, \( X_n \) returns to \( i \) infinitely often, whereas if \( i \) is transient, there is a last time a visit occurs.

The set of recurrent states can be subdivided further. We say that the state \( i \) is **positive recurrent** if we also have

\[ \mathbb{E}_i [\tau_i] < \infty. \]

Otherwise, we say that the state \( i \) is **null recurrent**. The different types of recurrence will be explored further in Section 2.5, where we will show why positive recurrence is a much stronger form of recurrence than null recurrence. In fact, in many important ways positive recurrent chains with an infinite state space behave like finite state space chains.

The following theorem shows that recurrence, and hence transience, is a class property. Thus, when the chain is irreducible and each state is recurrent, we may say that the chain itself is recurrent.
Theorem 2.4.11. Suppose that \( i \leftrightarrow j \). Then state \( i \) is recurrent if and only if state \( j \) is recurrent.

Proof. The following argument is the intuition needed to understand the result (which is also the basis of the proof): because state \( i \) is recurrent, we return to it an infinite number of times with a probability of one. We also know that there is an \( n > 0 \) for which \( p_{ij}^{(n)} > 0 \). Thus, every time we are in state \( i \), which happens an infinite number of times, there is a positive probability that we get to state \( j \) in \( n \) steps. Thus, we will enter state \( j \) an infinite number of times. The formal proof, based upon this intuition, is below.

Suppose that \( i \) is recurrent. We must show that \( j \) is recurrent. Because \( i \leftrightarrow j \), there are nonnegative integers \( n \) and \( m \) that satisfy \( p_{ij}^{(n)}, p_{ji}^{(m)} > 0 \). Let \( k \) be a nonnegative integer. It is an exercise in the use of conditional probabilities to show that

\[
p_{jj}^{(m+n+k)} \geq p_{ji}^{(m)} p_{ii}^{(k)} p_{ij}^{(n)},
\]

which says that one way to get from state \( j \) to state \( j \) in precisely \( m + n + k \) steps is to first go to \( i \) in \( m \) steps, then return to \( i \) in \( k \) steps, then return to \( j \) in \( n \) steps. Using this observation, we have

\[
\sum_{k=0}^{\infty} p_{jj}^{(k)} \geq \sum_{k=0}^{\infty} p_{jj}^{(m+n+k)} \geq \sum_{k=0}^{\infty} p_{ji}^{(m)} p_{ii}^{(k)} p_{ij}^{(n)} = p_{ji}^{(m)} p_{ij}^{(n)} \sum_{k=0}^{\infty} p_{ii}^{(k)}.
\]

Because \( i \) is recurrent, Theorem 2.4.10 shows that the sum is infinite, and hence that state \( j \) is recurrent. \( \square \)

Note that Theorems 2.4.10 and 2.4.11 together guarantee the following:

Fact: All states of an irreducible, finite state space Markov chain are recurrent.

The above fact holds by the following logic: if the states were not recurrent, they are each transient. Hence, there is a last time, call it \( T_i \), that a particular realization of the chain visits state \( i \). Therefore, \( \max_i \{ T_i \} \) is the last time the realization visits any state, which can not be (where is the chain after time \( T \)?). Things are significantly less clear in the infinite state space setting as the next few examples demonstrate.

Example 2.4.12. Consider a one dimensional random walk on the integer lattice \( S = \mathbb{Z} = \{ \ldots, -2, -1, 0, 1, 2, \ldots \} \) where for some \( 0 < p < 1 \) we have

\[
p_{i,i+1} = p, \quad p_{i,i-1} = q, \quad \text{with} \quad q \overset{\text{def}}{=} 1 - p.
\]

This chain is irreducible and has a period of 2. We will show that it is recurrent if \( p = 1/2 \), and transient otherwise. To do so, we will verify the result at the origin using Theorem 2.4.10, and then use Theorem 2.4.11 to extend the result to the entire state space.
Notice that because of the periodicity of the system, we have
\[ p_{00}^{(2n+1)} = 0, \]
for all \( n \geq 0 \). Therefore,
\[ \sum_{n=0}^{\infty} p_{00}^{(n)} = \sum_{n=0}^{\infty} p_{00}^{(2n)}. \]

Given that \( X_0 = 0 \), if \( X_{2n} = 0 \) the chain must have moved to the right \( n \) times and to the left \( n \) times. Each such sequence of steps has a probability of \( p^n q^n \) of occurring. Because there are exactly \( \binom{2n}{n} \) such paths, we see
\[ p_{00}^{(2n)} = \binom{2n}{n} (pq)^n = \frac{(2n)!}{n!n!} (pq)^n. \]

Therefore,
\[ \sum_{n=0}^{\infty} p_{00}^{(2n)} = \sum_{n=0}^{\infty} \frac{(2n)!}{n!n!} (pq)^n. \]

Recall that Stirling’s formula states that for \( m \gg 1 \),
\[ m! \sim m^m e^{-m} \sqrt{2\pi m}, \]
where by \( f(m) \sim g(m) \) we mean
\[ \lim_{m \to \infty} \frac{f(m)}{g(m)} = 1. \]

Verification of Stirling’s formula can be found in a number of places, for example in [16]. Stirling’s formula yields
\[ p_{00}^{(2n)} = \binom{2n}{n} (pq)^n \sim \frac{\sqrt{4\pi n (2n)^2 e^{-2n}}}{2 \pi n n^2 e^{-2n}} (pq)^n = \frac{1}{\sqrt{\pi n}} (4pq)^n. \]

Therefore, there is an \( N > 0 \) such that \( n \geq N \) implies
\[ \frac{1}{2\sqrt{\pi n}} (4pq)^n < p_{00}^{(2n)} < \frac{2}{\sqrt{\pi n}} (4pq)^n. \]

The function \( 4pq = 4p(1-p) \) is strictly less than one for all \( p \in [0, 1] \) with \( p \neq 1/2 \). However, when \( p = 1/2 \), we have that \( 4p(1-p) = 1 \). Therefore, in the case that \( p = 1/2 \) we have
\[ \sum_{n=0}^{\infty} p_{00}^{(2n)} > \sum_{n=N}^{\infty} p_{00}^{(2n)} > \sum_{n=N}^{\infty} \frac{1}{2\sqrt{\pi n}} = \infty, \]
and by Theorem 2.4.10, the chain is recurrent. When \( p \neq 1/2 \), let \( \rho = 4pq < 1 \). We have
\[ \sum_{n=0}^{\infty} p_{00}^{(2n)} < N + \sum_{n=N}^{\infty} \frac{2}{\sqrt{\pi n}} \rho^n < \infty, \]
and by Theorem 2.4.10, the chain is transient. \( \square \)
Example 2.4.13. We consider now the symmetric random walk on the integer lattice $\mathbb{Z}^d$ introduced in Example 2.1.12. Recall that for this example,

$$p_{ij} = \begin{cases} 1/2d & \text{if } |i - j| = 1 \\ 0 & \text{else} \end{cases}.$$ 

We again consider starting the walk at the origin $\vec{0} = (0, 0, \ldots, 0)$. The chain has a period of 2, and so $p_{\vec{0},\vec{0}}^{(2n+1)} = 0$ for all $n \geq 0$. Thus, to apply Theorem 2.4.10 we only need an expression for $p_{\vec{0},\vec{0}}^{(2n)}$. We will not give a rigorous derivation of the main results here as the combinatorics for this example are substantially more cumbersome than the last. Instead, we will make use of the following facts, which are intuitive:

(i) For large value of $n$, approximately $2n/d$ of these steps will be taken in each of the $d$ dimensions.

(ii) In each of the $d$ dimensions, the analysis of the previous example implies that the probability that that component is at zero at time $2n/d$ is asymptotic to $1/\sqrt{\pi(n/d)}$.

Therefore, as there are $d$ dimensions, we have

$$p_{\vec{0},\vec{0}}^{(2n)} \sim C \left( \frac{d}{n\pi} \right)^{d/2},$$

for some $C > 0$ (that depends upon $d$, of course). Recalling that $\sum_{n=1}^{\infty} n^{-a} < \infty$ if and only if $a > 1$, we see that

$$\sum_{n=1}^{\infty} p_{\vec{0},\vec{0}}^{(2n)} \begin{cases} = \infty, & d = 1, 2 \\ < \infty, & d \geq 3 \end{cases}$$

Thus, simple random walk in $\mathbb{Z}^d$ is recurrent if $d = 1$ or 2 and is transient if $d \geq 3$. This points out the general phenomenon that dynamics, in general, are quite different in dimensions greater than or equal to three than in dimensions one and two. Essentially, a path restricted to a line or a plane is much more restricted than one in space.\footnote{The video game “Tron” points this out well. Imagine how the game would play in three dimensions.}  

The following should, at this point, be intuitive.

Theorem 2.4.14. Every recurrent class of a Markov chain is a closed set.

Proof. Suppose $C$ is a recurrent class that is not closed. Then, there exists $i \in C$ and $j \notin C$ such that $p_{ij} > 0$, but it is impossible to return to state $i$ (otherwise, $i \leftrightarrow j$). Therefore, the probability of starting in $i$ and never returning is at least $p_{ij} > 0$, a contradiction with the class, and hence state $i$, being recurrent. \qed
Note that the converse of the above theorem is, in general, false. For example, for
the deterministic monotone chain, each set \( \{n, n+1, \ldots \} \) is closed, though no state
is recurrent.

Suppose that \( P \) is a transition matrix for a Markov chain and that \( R_1, \ldots, R_r \) are
the recurrent communication classes and \( T_1, \ldots, T_s \) are the transient classes. Then, after potentially reordering the indices of the state, we can write \( P \) in the following form:

\[
P = \begin{bmatrix}
P_1 & 0 & \cdots & 0 \\
P_2 & P_3 & \cdots & 0 \\
0 & \ddots & \ddots & \vdots \\
S & & & P_r \\
\end{bmatrix}
\]

where \( P_k \) is the transition matrix for the Markov chain restricted to \( R_k \). Raising \( P \)
to powers of \( n \geq 1 \) yields

\[
P^n = \begin{bmatrix}
P_1^n & 0 & \cdots & 0 \\
P_2^n & P_3^n & \cdots & 0 \\
0 & \ddots & \ddots & \vdots \\
S_n & & & P_r^n \\
\end{bmatrix}
\]

and to understand the behavior of the chain on \( R_k \), we need only study \( P_k \). The
Matrix \( Q \) is sub-stochastic in that the row sums are all less than or equal to one,
and at least one of the row sums is strictly less than one. In this case each of the
eigenvalues has an absolute value that is strictly less than one, and it can be shown
that \( Q^n \to 0 \), as \( n \to \infty \).

### 2.5 Stationary distributions and the long time behavior of the chain

Just as stable fixed points characterize the long time behavior of solutions to differential equations, stationary distributions characterize the long time behavior of Markov chains.

**Definition 2.5.1.** Consider a Markov chain with transition matrix \( P \). A non-negative
vector \( \pi \) is said to be an invariant measure if

\[
\pi^T P = \pi^T,
\]

which in component form is

\[
\pi_i = \sum_j \pi_j p_{ji}, \quad \text{for all } i \in S.
\]
If $\pi$ also satisfies $\sum_k \pi_k = 1$, then $\pi$ is called a stationary, equilibrium or steady state probability distribution.

Thus, a stationary distribution is a left eigenvector of the transition matrix with associated eigenvalue equal to one. Note that if one views $p_{ji}$ as a “flow rate” of probability from state $j$ to state $i$, then (2.18) can be interpreted in the following manner: for each state $i$, the probability of being in state $i$ is equal to the sum of the probability of being in state $j$ times the “flow rate” from state $j$ to $i$.

A stationary distribution can be interpreted as a fixed point for the Markov chain because if the initial distribution of the chain is given by $\pi$, then the distribution at all times $n \geq 0$ is also $\pi$,

$$\pi^T P^n = \pi^T P P^{n-1} = \pi^T P^{n-1} = \cdots = \pi^T,$$

where we are using equation (2.12). Of course, in the theory of dynamical systems it is well known that simply knowing a fixed point exists does not guarantee that the system will converge to it, or that it is unique. Similar questions exist in the Markov chain setting:

1. Under what conditions on a Markov chain will a stationary distribution exist?
2. When a stationary distribution exists, when is it unique?
3. Under what conditions can we guarantee convergence to a unique stationary distribution?

We note that we have already seen an example in which all of the above questions were answered. Recall that in Section 2.3, we showed that if the two-state Markov chain has transition matrix

$$P = \begin{bmatrix} 2/3 & 1/3 \\ 1/8 & 7/8 \end{bmatrix},$$

then for very large $n$,

$$P^n \approx \begin{bmatrix} 3/11 & 8/11 \\ 3/11 & 8/11 \end{bmatrix} = \Pi.$$

The important point was that the rows of $\Pi$ are identical and equal to $\pi^T = [3/11, 8/11]$, and therefore, if $v$ is an arbitrary probability vector,

$$\lim_{n \to \infty} v^T P^n = v^T \Pi = \pi^T,$$

and so no matter the initial distribution we have

$$\lim_{n \to \infty} P\{X_n = 1\} = \frac{3}{11}, \quad \text{and} \quad \lim_{n \to \infty} P\{X_n = 2\} = \frac{8}{11}.$$

It is straightforward to check that $[3/11, 8/11]$ is the unique left eigenvector of $P$ with an eigenvalue of 1.

Let us consider at least one more example.
Example 2.5.2. Suppose that $X_n$ is a three state Markov chain with transition matrix

$$P = \begin{bmatrix} \frac{2}{3} & \frac{1}{3} & 0 \\ \frac{1}{12} & \frac{5}{8} & \frac{7}{24} \\ 0 & \frac{1}{8} & \frac{7}{8} \end{bmatrix}. \quad (2.20)$$

Then, for large $n$

$$P^n \approx \begin{bmatrix} \frac{3}{43} & \frac{12}{43} & \frac{28}{43} \\ \frac{3}{43} & \frac{12}{43} & \frac{28}{43} \\ \frac{3}{43} & \frac{12}{43} & \frac{28}{43} \end{bmatrix} = \Pi,$$

where we again note that each row of $\Pi$ is identical. Therefore, regardless of the initial distribution, we have

$$\lim_{n \to \infty} P\{X_n = 1\} = \frac{3}{43}, \quad \lim_{n \to \infty} P\{X_n = 2\} = \frac{12}{43}, \quad \text{and} \quad \lim_{n \to \infty} P\{X_n = 3\} = \frac{28}{43}.$$

We again note that it is straightforward to check that $[\frac{3}{43}, \frac{12}{43}, \frac{28}{43}]$ is the unique left eigenvalue of $P$ with an eigenvalue of 1. \hfill \Box

Interestingly, we were able to find stationary distributions for the above transition matrices without actually computing the left eigenvectors. Instead, we just found the large $n$ probabilities. Question 3 above asks when such a link between stationary distributions and large $n$ probabilities holds (similar to convergence to a fixed point for a dynamical system). This question will be explored in detail in the current section, however we begin my making the observation that if

$$\pi^T = \lim_{n \to \infty} v^T P^n,$$

for all probability vectors $v$ (which should be interpreted as an initial distribution), then

$$\pi^T = \lim_{n \to \infty} v^T P^{n+1} = \left( \lim_{n \to \infty} v^T P^n \right) P = \pi^T P.$$

Therefore, if $P^n$ converges to a matrix with a common row, $\pi$, then that common row is, in fact, a stationary distribution.

The logic of the preceding paragraph is actually backwards in how one typically studies Markov chains. Most often, the modeler has a Markov chain describing something of interest to him or her. If this person would like to study the behavior of their process for very large $n$, then it would be reasonable to consider the limiting probabilities, assuming they exist. To get at these probabilities, they would need to compute $\pi$ as the left-eigenvector of their transition matrix and verify that this is the unique stationary distribution, and hence all probabilities converge to it, see Theorems 2.5.6 and 2.5.16 below.

We will answer the three questions posed above first in the finite state space setting, where many of the technical details reduce to linear algebra. We then extend all the results to the infinite state space setting.
2.5.1 Markov chains with a finite state space

Irreducible, aperiodic chains

For a finite Markov chain with transition matrix $P$, we wish to understand the long term behavior of $P^n$ and, relatedly, to find conditions that guarantee a unique stationary distribution exists. However, we first provide a few examples showing when such a unique limiting distribution does not exist.

Example 2.5.3. Consider simple random walk on $\{0, 1, 2\}$ with reflecting boundaries. In this case we have

$$P = \begin{bmatrix} 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 1 & 0 \end{bmatrix}.$$  

Note that for $n \geq 1$,  

$$P^{2n} = \begin{bmatrix} 1/2 & 0 & 1/2 \\ 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \end{bmatrix},$$

and,  

$$P^{2n+1} = \begin{bmatrix} 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 1 & 0 \end{bmatrix}.$$  

It is easy to see why this happens. If the walker starts at 1, then she must be at one after an even number of steps, etc. This chain is therefore periodic. Clearly, $P^n$ does not converge in this example. $\Box$

Example 2.5.4. Consider simple random walk on $\{0, 1, 2, 3\}$ with absorbing boundaries. That is

$$P = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 \\ 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$  

For $n$ large we have

$$P^n \approx \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2/3 & 0 & 0 & 1/3 \\ 1/3 & 0 & 0 & 2/3 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$  

Again, this is believable, as you are assured that you will end up at 0 or 3 after enough time has passed. We see the problem here is that the states $\{1, 2\}$ are transient. $\Box$

Example 2.5.5. Suppose that $S = \{1, 2, 3, 4, 5\}$ and

$$P = \begin{bmatrix} 1/3 & 2/3 & 0 & 0 & 0 \\ 3/4 & 1/4 & 0 & 0 & 0 \\ 0 & 0 & 1/8 & 1/4 & 5/8 \\ 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 1/3 & 0 & 2/3 \end{bmatrix}.$$
For \( n \gg 1 \), we have

\[
P^n \approx \begin{bmatrix}
\frac{9}{17} & \frac{8}{17} & 0 & 0 & 0 \\
\frac{9}{17} & \frac{8}{17} & 0 & 0 & 0 \\
0 & 0 & \frac{8}{33} & \frac{4}{33} & \frac{7}{11} \\
0 & 0 & \frac{8}{33} & \frac{4}{33} & \frac{7}{11} \\
0 & 0 & \frac{8}{33} & \frac{4}{33} & \frac{7}{11}
\end{bmatrix}.
\]

In this case, the Markov chain really consists of two smaller, noninteracting chains: one on \( \{1, 2\} \) and another on \( \{3, 4, 5\} \). Each subchain will converge to its equilibrium distribution, but there is no way to move from one subchain to the other. Here the problem is that the chain is reducible.

These examples actually demonstrate everything that can go wrong. The following theorem is the main result of this section.

**Theorem 2.5.6.** Suppose that \( P \) is the transition matrix for a finite Markov chain that is irreducible and aperiodic. Then, there is a unique stationary distribution \( \pi \),

\[
\pi^T P = \pi^T,
\]

for which \( \pi_i > 0 \) for each \( i \). Further, if \( v \) is any probability vector, then

\[
\lim_{n \to \infty} v^T P^n = \pi^T.
\]

The remainder of this sub-section consists of verifying Theorem 2.5.6. However, before proceeding with the general theory, we attempt to better understand why the processes at the beginning of this section converged to a limiting distribution \( \pi \). Note that the eigenvalues of the matrix (2.19) are

\[
\lambda_1 = 1 \quad \text{and} \quad \lambda_2 = \frac{13}{24} < 1.
\]

If we let \( \pi_1 \) and \( \pi_2 \) denote the respective left eigenvectors, and let \( v \) be an arbitrary probability vector, then because \( \pi_1 \) and \( \pi_2 \) are necessarily linearly independent

\[
v^T P^n = (c_1 \pi_1^T P^n + c_2 \pi_2^T P^n) = c_1 \pi_1^T + c_2 (13/24)^n \pi_2^T \to c_1 \pi_1^T, \quad \text{as } n \to \infty.
\]

The normalization constant \( c_1 \) is then chosen so that \( c_1 \pi_1 \) is a probability vector.

Similarly, the eigenvalues of the transition matrix for the Markov chain of Example 2.5.2 are \( \lambda_1 = 1 \) and \( \lambda_2, \lambda_3 = (14 \pm \sqrt{14})/24 \). Thus, \( |\lambda_i| < 1 \) for \( i \in \{1, 2\} \), and \( \lambda_1 = 1 \) is again the dominant eigenvalue. Therefore, by the same reasoning as in the \( 2 \times 2 \) case, we again have that for any probability vector \( v \),

\[
v^T P^n \to c_1 \pi_1^T, \quad \text{as } n \to \infty,
\]

where \( c_1 \) is chosen so that \( c_1 \pi_1 \) is a probability vector.

The above considerations suggest the following plan of attack for proving Theorem 2.5.6, which we write in terms of a claim.

**Claim:** Suppose that a stochastic matrix, \( P \), satisfies the following three conditions:
(i) $P$ has an eigenvalue of 1, which is simple (has a multiplicity of one).

(ii) All other eigenvalues have absolute value less than 1.

(iii) The left eigenvector associated with the eigenvalue 1 has strictly positive entries.

Then
\[ v^T P^n \to \pi^T, \quad \text{as } n \to \infty, \] (2.21)
for any probability vector $v$, where $\pi$ is the unique left eigenvector normalized to sum to one, and $\pi_i > 0$ for each $i$.

Note that condition (iii) above is not strictly necessary as it just guarantees convergence to a vector giving non-zero probability to each state. However, it is included for completeness (since this will be the case for irreducible chains), and we will consider the possibility of $\pi_i = 0$ for some $i$ (which will occur if there are transient states) later.

It turns out the above claim follows from a straightforward use of Jordan canonical forms. We point the interested reader to [32, Chapter 1] for full details. However, it is probably more instructive to show the result in a slightly less general setting by also assuming that there is a full set of distinct eigenvalues for $P$ (though we stress that the claim holds even without this added assumption). Thus, let $\lambda_1, \lambda_2, \ldots, \lambda_N$, be the eigenvalues of $P$ with $\lambda_1 = 1$ and $|\lambda_i| < 1$, for $i > 1$. Let the corresponding left eigenvectors be denoted by $\pi_i$, where $\pi_1$ is normalized to sum to one (that is, it is a probability vector). The eigenvectors are necessarily linearly independent and so we can write our initial distribution as
\[ v = c_1 \pi_1 + c_2 \pi_2 + \cdots + c_N \pi_N, \]
for some choice of $c_i$, which depend upon our choice of $v$. Thus, letting $\pi^{(n)}$ denote the distribution at time $n$ we see
\[ \pi^{(n)} = v^T P^n = (c_1 \pi_1^T + c_2 \pi_2^T + \cdots + c_N \pi_N^T) P^n = c_1 \lambda_1^n \pi_1^T + c_2 \lambda_2^n \pi_2^T + \cdots + c_N \lambda_N^n \pi_N^T \to c_1 \pi_1, \]
as $n \to \infty$. Note that as both $\pi^{(n)}$ and $\pi_1$ are probability vectors, we see that $c_1 = 1$, which agrees with our examples above. We further note the useful fact that the rate of convergence to the stationary distribution is dictated by the size of the second largest (in absolute value) eigenvalue.

Returning to Theorem 2.5.6, we see that the theorem will be proved if we can verify that the transition matrix of an aperiodic, irreducible chain satisfies the three conditions above. By the Perron-Frobenius theorem, any stochastic matrix, $Q$, that has all strictly positive entries satisfies the following:

(i) 1 is a simple eigenvalue of $Q$,
(ii) the left eigenvector associated with 1 can be chosen to have strictly positive entries,

(iii) all other eigenvalues have absolute value less than 1.

Therefore, the Perron-Frobenius theorem almost gives us what we want. However, the transition matrix for aperiodic, irreducible chains do not necessarily have strictly positive entries, see (2.20) of Example 2.5.2, and so the above can not be applied directly.

However, suppose instead that $P^n$ has strictly positive entries for some $n \geq 1$. Then the Perron-Frobenius theorem can be applied to $P^n$, and conditions (i), (ii), and (iii) directly above hold for $P^n$. However, by the spectral mapping theorem the eigenvalues of $P^n$ are simply the $n$th powers of the eigenvalues of $P$, and the eigenvectors of $P$ are the eigenvectors of $P^n$. We can now conclude that $P$ also satisfies the conclusions of the Perron-Frobenius theorem by the following arguments:

1. The vector consisting of all ones is a right eigenvector of $P$ with eigenvalue 1, showing $P$ always has such an eigenvalue.

2. If $\lambda \neq 1$ were an eigenvalue of $P$ with $|\lambda| = 1$ and eigenvector $v$, then $v^T P^n = \lambda^n v^T$, showing $v$ is a left eigenvector of $P^n$ with eigenvalue of absolute value equal to one. This is impossible as the eigenvalue 1 is simple for $P^n$. Thus, 1 is a simple eigenvalue of $P$ and all others have absolute value less than one.

3. The left eigenvector of $P$ associated with eigenvalue 1 has strictly positive components since this is the eigenvector with eigenvalue 1 for $P^n$.

Therefore, Theorem 2.5.6 will be shown if the following claim holds:

**Claim:** Suppose that $P$ is the transition matrix for an aperiodic, irreducible Markov chain. Then, there is an $n \geq 1$ for which $P^n$ has strictly positive entries.

**Proof.** The proof of the claim is relatively straightforward, and the following is taken from [32, Chapter 1]. We take the following fact for granted, which follows from a result in number theory: if the chain is aperiodic, then for each state $i$, there is an $M(i)$ for which $p_{ii}^{(n)} > 0$ for all $n \geq M(i)$.

Returning to the proof of the claim, we need to show that there is an $M > 0$ so that if $n \geq M$, then we have that $P^n$ has strictly positive entries. Let $i, j \in S$. By the irreducibility of the chain, there is an $m(i,j)$ for which $p_{ij}^{(m(i,j))} > 0$.

Thus, for all $n \geq M(i)$,

$$p_{ij}^{(n+m(i,j))} \geq p_{ii}^{(n)} p_{ij}^{(m(i,j))} > 0.$$ 

Now, simply let $M$ be the maximum over $M(i) + m(i,j)$, which exists since the state space is finite. Thus, $p_{ij}^{(n)} > 0$ for all $n \geq M$ and all $i,j \in S$. \qed
We pause to reflect upon what we have shown. We have concluded that for an irreducible, aperiodic Markov chain, if we wish to understand the large time probabilities associated with the chain, then it is sufficient to calculate the unique left eigenvector of the transition matrix with eigenvalue equal to one. Such computations can be carried out by hand for small examples, though are usually performed with software (such as Maple, MATLAB, or Mathematica) for larger systems. In the next sub-section we consider what changes when we drop the irreducible assumption. We will consider the periodic case when we turn to infinite state space Markov chains in Section 2.5.2.

**Example 2.5.7.** Consider a Markov chain with state space \( \{0, 1, 2, 3\} \) and transition matrix

\[
P = \begin{bmatrix}
0 & 1/5 & 3/5 & 1/5 \\
1/4 & 1/4 & 1/4 & 1/4 \\
1 & 0 & 0 & 0 \\
0 & 1/2 & 1/2 & 0 \\
\end{bmatrix}.
\]

Find \( \lim_{n \to \infty} P\{X_n = 2\} \).

**Solution.** It is easy to verify that

\[
P^3 = \begin{bmatrix}
3/16 & 77/400 & 18/400 & 67/400 \\
127/320 & 57/320 & 97/320 & 39/320 \\
13/20 & 3/20 & 3/20 & 1/20 \\
5/32 & 7/32 & 15/32 & 5/32 \\
\end{bmatrix},
\]

showing that Theorem 2.5.6 applies. The eigenvector of \( P \) (normalized to be a probability distribution) associated with eigenvalue 1 is

\[
\pi = \begin{bmatrix}
25/67 \\
12/67 \\
22/67 \\
8/67
\end{bmatrix}.
\]

Therefore,

\[
\lim_{n \to \infty} P\{X_n = 2\} = \frac{12}{67}.
\]

**Reducible chains**

We turn to the case of a reducible chain and begin with some examples.

**Example 2.5.8.** Consider the gambler’s ruin problem on the state space \( \{0, 1, 2, \ldots, N\} \).

Setting \( \pi_\alpha = (\alpha, 0, 0, \ldots, 1 - \alpha) \),
for any \( 0 \leq \alpha \leq 1 \), it is straightforward to show that \( \pi_\alpha^T P = \pi_\alpha^T \). Thus, there are uncountably many stationary distributions for this example, though it is important to note that they are all linear combinations of \((1, 0, \ldots, 0)\) and \((0, \ldots, 0, 1)\), which are the stationary distributions on the recurrent classes \(\{0\}\) and \(\{N\}\).

\[
\square
\]
Example 2.5.9. Consider the Markov chain on \( \{1, 2, 3, 4\} \) with

\[
P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix},
\]

with

\[
P_i = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}, \quad \text{for } i \in \{1, 2\}.
\]

Then, the communication classes \( \{1, 2\} \) and \( \{3, 4\} \) are each irreducible and aperiodic, and have stationary distribution \((1/2, 1/2)\). Also for any 0 ≤ \( \alpha \) ≤ 1,

\[
\alpha(1/2, 1/2, 0, 0) + (1 - \alpha)(0, 0, 1/2, 1/2) = (\alpha/2, \alpha/2, (1 - \alpha)/2, (1 - \alpha)/2)
\]

is a stationary distribution for the transition matrix \( P \). □

The above examples essentially show what happens in this case of a reducible Markov chain with a finite state space. All of the mass of a limiting distribution will end up on the recurrent classes, and the form of the stationary distribution on the recurrent classes can be found by the results in the previous section.

Consider now a general finite state space Markov chain with reducible state space, \( S \), that is restricted to any recurrent communication class \( R_k \subset S \). If the Markov chain is aperiodic on \( R_1 \), then by Theorem 2.5.6 a unique stationary distribution, \( \pi^{(1)} \), exists with support only on \( R_1 \). Clearly, the previous argument works for each recurrent communication class \( R_k \subset S \). Therefore, we have the existence of a family of stationary distributions, \( \pi^{(k)} \), which are limiting stationary distributions for the Markov chain restricted to the different \( R_k \). We note the following (some of which are left as homework exercises to verify):

1. Each such \( \pi^{(k)} \) is a stationary distribution for the original, unrestricted Markov chain.

2. Assuming there are \( m \) recurrent communication classes, each linear combination

\[
a_1 \pi^{(1)} + \cdots + a_m \pi^{(m)}
\]

with \( a_i \geq 0 \) and \( \sum_i a_i = 1 \), is a stationary distribution for the unrestricted Markov chain, \( X_n \).

3. All stationary distributions of the Markov chain \( X_n \) can be written as a linear combination of the form (2.22).

Thus, in the case that the Markov chain is reducible, the limiting probabilities will depend on the initial condition. That is, if \( \alpha_k(i) \) is the probability that the chain ends up in recurrent class \( R_k \) given it starts in state \( i \), then for \( j \in R_k \),

\[
\lim_{n \to \infty} p^{(n)}_{ij} = \alpha_k(i) \pi^{(k)}_j
\]

(2.23)

where we will discuss how to calculate \( \alpha_k(i) \) in later sections. Note, however, that \( \alpha_k(i) \) will be one if \( i, j \) are in the same recurrent class, zero if they are in different recurrent classes, and between zero and one if \( i \) is transient and \( i \to j \). We conclude that if \( v \) is an initial distribution for a reducible finite state space Markov chain, then the limit \( \lim_{n \to \infty} v^T P^n \) will always exists, though it will depend upon \( v \).
2.5.2 Countable Markov chains

We now extend the results of the previous section to the setting of a countably infinite state space. We note that every result stated in this section also holds for the finite state space case, and these are the most general results. We begin with an example demonstrating a major difference between the finite and countable state space setting.

Example 2.5.10. Consider symmetric random walk on the integers. That is, the state space is $S = \mathbb{Z}$ and $p_{i,i+1} = p_{i,i-1} = 1/2$ for all $i$. We know from Example 2.4.12 that this chain is recurrent, and we search for a stationary distribution $\pi$ satisfying $\pi^T = \pi^T P$, where $P$ is the transition matrix. We find that $\pi^T = \pi^T P$ is equivalent to

$$\pi_j = \sum_k \pi_k p_{kj} = \pi_{j-1} p_{j-1,j} + \pi_{j+1} p_{j+1,j} = \pi_{j-1} (1/2) + \pi_{j+1} (1/2) = \frac{1}{2} (\pi_{j-1} + \pi_{j+1}),$$

for all $j \in \mathbb{Z}$. These can be solved by taking $\pi_j \equiv 1$. Note, however, that in this case we cannot scale the solution to get a stationary distribution, and so such a $\pi$ is an invariant measure, though not a stationary distribution. □

Recalling that the Markov chain of the previous example is recurrent (see Example 2.4.12), and therefore one might expect a stationary distribution to exist, it turns out the chain “is not recurrent enough.” We recall that we define $\tau_i$ to be the first return time to state $i$,

$$\tau_i \overset{\text{def}}{=} \min\{n \geq 1 : X_n = i\},$$

where we take $\tau_i = \infty$ if the chain never returns. We further recall that the state $i$ is called recurrent if $P_i(\tau_i < \infty) = 1$ and transient otherwise. In the infinite state space setting it is useful to subdivide the set of recurrent states even further.

Definition 2.5.11. The value

$$\mu_i \overset{\text{def}}{=} \mathbb{E}_i \tau_i = \sum_{n=1}^{\infty} n P_i\{\tau_i = n\}$$

is called the mean recurrence time or mean first return time for state $i$. We say that a recurrent state $i$ is positive recurrent if $\mathbb{E}_i \tau_i < \infty$, and null recurrent otherwise, i.e. if $\mathbb{E}_i \tau_i = \infty$.

Note that we have $\mu_i = \infty$ for a transient state as in this case $P_i\{\tau_i = \infty\} > 0$.

The following is stated without proof. However, for those that are interested, the result follows directly from basic renewal theory, see [37, Chapter 3]. Theorem 2.5.12 captures the main difference between positive recurrent and other (null recurrent and transient) chains.

Theorem 2.5.12. Consider a recurrent, irreducible, aperiodic Markov chain. Then, for any $i,j \in S$

$$\lim_{n \to \infty} p_{ji}^{(n)} = \frac{1}{\mu_i},$$

where if $\mu_i = \infty$ (null recurrence), we interpret the right hand side as zero.
The similar theorem for periodic chains is the following.

**Theorem 2.5.13.** Let $X_n$ be a recurrent, irreducible, $d$-periodic Markov chain. Then, for any $i \in S$

$$\lim_{n \to \infty} p_{ii}^{(nd)} = \frac{d}{\mu_i},$$

where if $\mu_i = \infty$ (null recurrence), then we interpret the right hand side as zero.

Recurrence has already been shown to be a class property. The following theorem shows that positive recurrence is also a class property.

**Theorem 2.5.14.** Suppose that $i \leftrightarrow j$ belong to the same class and that state $i$ is positive recurrent. Then state $j$ is positive recurrent.

**Proof.** We will prove the result in the aperiodic case so that we may make use of Theorem 2.5.12. We know from Theorem 2.5.12 that

$$\lim_{n \to \infty} p_{kj}^{(n)} = \frac{1}{\mu_j},$$

for any $k$ in the same class as $j$. Because $j$ is positive recurrent if and only if $\mu_j < \infty$, we see it is sufficient to show that

$$\lim_{n \to \infty} p_{ij}^{(n)} > 0.$$

Because $i \leftrightarrow j$, there is an $m > 0$ for which $p_{ij}^{(m)} > 0$. Therefore,

$$\lim_{n \to \infty} p_{ij}^{(n)} = \lim_{n \to \infty} p_{ij}^{(n+m)} \geq \lim_{n \to \infty} p_{ii}^{(n)} p_{ij}^{(m)} = p_{ij}^{(m)} \lim_{n \to \infty} p_{ii}^{(n)} = p_{ij}^{(m)} \frac{1}{\mu_i} > 0,$$

where the final equality holds from Theorem 2.5.12 applied to state $i$. \(\square\)

Therefore, we can speak of positive recurrent chains or null recurrent chains.

**Example 2.5.15.** Consider again the symmetric \((p = 1/2)\) random walk on the integer lattice. We previously showed that

$$P_{00}^{(2n)} \sim \frac{1}{\sqrt{\pi n}}.$$

Therefore, $\lim_{n \to \infty} P_{00}^{(2n)} \sim \frac{1}{\sqrt{\pi n}} = 0$, and by Theorem 2.5.13 we have that $\mu_0 = \infty$, and the chain is null recurrent. Thus, when $p = 1/2$, the chain is periodic and null recurrent, and when $p \neq 1/2$, the chain is periodic and transient. \(\square\)

Theorem 2.5.12 also gives a strong candidate for a limiting stationary distribution for a positive recurrent, irreducible, aperiodic Markov chain.
**Theorem 2.5.16.** If a Markov chain is irreducible and recurrent, then there is an invariant measure \( \pi \), unique up to multiplicative constants, that satisfies \( 0 < \pi_j < \infty \) for all \( j \in S \). Further, if the Markov chain is positive recurrent then

\[
\pi_i = \frac{1}{\mu_i},
\]

where \( \mu_i \) is the mean recurrence time of state \( i \), \( \sum_i \pi_i = 1 \), and \( \pi \) is a stationary distribution of the Markov chain. If the Markov chain is also aperiodic, then \( p_j^{(n)} \rightarrow \pi_i \), as \( n \rightarrow \infty \), for all \( i, j \in S \).

**Proof.** We will verify the result in the positive recurrent case only, and direct the reader to [37, Chapter 2.12] for the full details. We first show that \( \sum_{i \in S} \pi_i = 1 \).

Choosing any \( k \in S \), we see

\[
1 = \lim_{n \rightarrow \infty} \sum_{j \in S} p_j^{(n)} = \sum_{j \in S} \frac{1}{\mu_j},
\]

where the final equality follows from Theorem 2.5.12. Next, for any \( k \in S \),

\[
\frac{1}{\mu_i} = \lim_{n \rightarrow \infty} p_{ki}^{(n+1)} = \sum_{j \in S} \lim_{n \rightarrow \infty} P\{X_{n+1} = i \mid X_n = j\} P\{X_n = j \mid X_0 = k\} = \sum_{j \in S} p_{ji} \lim_{n \rightarrow \infty} p_{kj}^{(n)} = \sum_{j \in S} p_{ji} \frac{1}{\mu_j}.
\]

Thus, the result is shown. \( \square \)

Note that Theorem 2.5.16 guarantees the existence of a stationary distribution even in the chain is periodic.

**Example 2.5.17.** Consider reflecting random walk on \( \{1, 2, 3, 4\} \). That is, the Markov chain with transition matrix

\[
P = \begin{bmatrix}
0 & 1 & 0 & 0 \\
1/2 & 0 & 1/2 & 0 \\
0 & 1/2 & 0 & 1/2 \\
0 & 0 & 1 & 0
\end{bmatrix}.
\]

This chain has period two, and for large \( n \) we have

\[
p^{2n} \approx \begin{bmatrix}
1/3 & 0 & 2/3 & 0 \\
0 & 2/3 & 0 & 1/3 \\
1/3 & 0 & 2/3 & 0 \\
0 & 2/3 & 0 & 1/3
\end{bmatrix}, \quad p^{2n+1} \approx \begin{bmatrix}
0 & 2/3 & 0 & 1/3 \\
1/3 & 0 & 2/3 & 0 \\
0 & 2/3 & 0 & 1/3 \\
1/3 & 0 & 2/3 & 0
\end{bmatrix}.
\]

The unique stationary distribution of the chain can be calculated, however, and is \( \pi = [1/6, 1/3, 1/3, 1/6] \). While \( \pi \) does not, in this case, give the long run probabilities of the associated chain, we will see in Theorem 2.5.22 a useful interpretation of \( \pi \) as giving the average amount of time spent in each state. \( \square \)
A question still remains: can the invariant measure of a null recurrent chain be normalized to give a stationary distribution? The answer, given in the following theorem, is no.

**Theorem 2.5.18.** Suppose a Markov chain is irreducible and that a stationary distribution \( \pi \) exists:

\[
\pi' = \pi' P, \quad \sum_{j \in S} \pi_j = 1, \quad \pi_j > 0.
\]

Then, the Markov chain is positive recurrent.

Thus, a necessary and sufficient condition for determining positive recurrence is simply demonstrating the existence or non-existence of a stationary distribution. Note also that the above results provides an effective algorithm for computing the mean return times: compute the invariant distribution using

\[
\pi^T = \pi^T P,
\]

and invert the component of interest.

**Example 2.5.19** (Random walk with partially reflecting boundaries, [32]). Consider again a random walker on \( S = \{0, 1, 2, \ldots\} \). Suppose that for \( j \in S \) the transition probabilities are given by

\[
p_{j,j+1} = p, \quad p_{j,j-1} = 1 - p, \quad \text{if } j \geq 1,
p_{01} = p, \quad p_{00} = 1 - p.
\]

This Markov chain is irreducible and aperiodic. We want to determine when this model will have a limiting stationary distribution, and, hence, when it is positive recurrent.

A stationary distribution for this system must satisfy

\[
\pi_{j+1}(1 - p) + \pi_{j-1} p = \pi_j, \quad j > 0 \tag{2.24}
\]

\[
\pi_1(1 - p) + \pi_0(1 - p) = \pi_0 \tag{2.25}
\]

with the condition that \( \pi_j \geq 0 \) and \( \sum_{j=0}^{\infty} \pi_j = 1 \). Solving the difference equations, the general solution to equation (2.24) is

\[
\pi_j = \begin{cases} 
  c_1 + c_2 \left(\frac{p}{1-p}\right)^j, & p \neq 1/2 \\
  c_1 + c_2 j, & p = 1/2
\end{cases}
\]

However, equation (2.25) shows

\[
\pi_0 = \frac{1 - p}{p} \pi_1.
\]
Plugging this into the above equation shows $c_1 = 0$ in the $p \neq 1/2$ case, and that $c_2 = 0$ in the $p = 1/2$ case. Therefore,

$$
\pi_j = \begin{cases} 
  c_2 \left( \frac{p}{1-p} \right)^j, & p \neq 1/2 \\
  c_1, & p = 1/2
\end{cases}
$$

Because we need $\sum_{j=0}^{\infty} \pi_j = 1$ for a distribution to exist, we see that if $p = 1/2$, no choice of $c_1$ could satisfy this condition.

Now just consider the case $p \neq 1/2$. We obviously require that $c_2 > 0$. If $p > 1/2$, then $p/(1-p) > 1$ and the sum

$$
\sum_{j=0}^{\infty} c_2 \left( \frac{p}{1-p} \right)^j = \infty.
$$

If, on the other hand, $p < 1/2$, then

$$
\sum_{j=0}^{\infty} c_2 \left( \frac{p}{1-p} \right)^j = c_2 \frac{1-p}{1-2p}.
$$

Therefore, taking $c_2 = (1-2p)/(1-p)$ gives us a stationary distribution of

$$
\pi_j = \frac{1-2p}{1-p} \left( \frac{p}{1-p} \right)^j.
$$

Thus, the chain is positive recurrent when $p < 1/2$, which is believable. We also know that the chain is either null recurrent or transient if $p \geq 1/2$. □

Suppose that we want to figure out when the chain of the previous example is either null recurrent or transient. We will make use of the following non-trivial fact, which is stated without proof. We will make use of this fact again in later sections.

**Theorem 2.5.20.** Let $X_n$ be an irreducible Markov chain with state space $S$, and let $i \in S$ be arbitrary. Then $X_n$ is transient if and only if there is a unique solution, $\alpha : S \to \mathbb{R}$, to the following set of equations

$$
0 \leq \alpha_j \leq 1, \quad \alpha_i = 1, \quad \inf \{ \alpha_j : j \in S \} = 0 \quad (2.26)
$$

$$
\alpha_j = \sum_{k \in S} p_{jk} \alpha_k, \quad i \neq j. \quad (2.27)
$$

It is reasonable to ask why these conditions are at least believable. Suppose we define

$$
\alpha_j = P\{X_n = i \text{ for some } n \geq 0 \mid X_0 = j\},
$$

47
and we assume our chain is transient. Then, $\alpha_i = 1$ by construction and we should have $\inf \{ \alpha_j : j \in S \} = 0$ by transience (though we are not going to prove this fact). Finally, for $j \neq i$, we have

$$
\alpha_j = P\{ X_n = i \text{ for some } n \geq 0 \mid X_0 = j \} 
= P\{ X_n = i \text{ for some } n \geq 1 \mid X_0 = j \} 
= \sum_k P\{ X_n = i \text{ for some } n \geq 1 \mid X_1 = k \} P\{ X_1 = k \mid X_0 = j \} 
= \sum_k p_{jk} \alpha_k.
$$

In the recurrent case, we know $\alpha_j \equiv 1$, and so there should be no solution satisfying (2.27).

**Example 2.5.21.** We return to the previous example and try to figure out when the chain is transient. Take $i = 0$. We will try to find a solution to the above equations. Equation (2.28) states that we must have

$$
\alpha_j = (1 - p)\alpha_{j-1} + p \alpha_{j+1}, \quad j > 0.
$$

The solution to this difference equation is

$$
\alpha_j = \begin{cases} 
  c_1 + c_2 \left( \frac{1-p}{p} \right)^j, & \text{if } p \neq 1/2 \\
  c_1 + c_2 j, & \text{if } p = 1/2
\end{cases}.
$$

We must have that $\alpha_0 = 1$. Therefore, we have

$$
\alpha_j = \begin{cases} 
  (1 - c_2) + c_2 \left( \frac{1-p}{p} \right)^j, & \text{if } p \neq 1/2 \\
  1 + c_2 j, & \text{if } p = 1/2
\end{cases}.
$$

If $c_2 = 0$ in either, then $\alpha_j \equiv 1$, and we can not satisfy our decay condition. Also, if $p = 1/2$ and $c_2 \neq 0$, then the solution is not bounded. Thus, there can be no solution in the case $p = 1/2$, and the chain is recurrent in this case. If $p < 1/2$, we see that the solution will explode if $c_2 \neq 0$. Thus, there is no solution for $p < 1/2$. Of course we knew this already because we already showed it was positive recurrent in this case! For the case $p > 1/2$, we have that $1 - p < p$, we see we can take $c_2 = 1$ and find that

$$
\alpha_j = \left( \frac{1-p}{p} \right)^j,
$$

is a solution. Thus, when $p > 1/2$, the chain is transient. \(\square\)

We end this section with a theorem that shows that the time averages of a single path of an irreducible and positive recurrent Markov chain is equal to the chains space average. This is incredibly useful and shows that one way to compute statistics of the stationary distribution is to compute one very long path and average over that path. For a proof of the Theorem below, we point the interested reader to [37, Chapter 2.12].

48
Theorem 2.5.22. Consider an irreducible, positive recurrent Markov chain with unique stationary distribution \( \pi \). If we let

\[
N_i(n) = \sum_{k=0}^{n-1} 1\{X_k=i\},
\]

denote the number of visits to state \( i \) before time \( n \), then

\[
P \left( \frac{N_i(n)}{n} \to \pi_i, \text{ as } n \to \infty \right) = 1.
\]

Moreover, for any bounded function \( f : S \to \mathbb{R} \),

\[
P \left( \frac{1}{n} \sum_{k=0}^{n-1} f(X_k) \to \sum_{i \in S} f(i)\pi_i, \text{ as } n \to \infty \right) = 1.
\]

The final result says that the time average of a single realization of the Markov chain converges (with probability one) to the “space averages” obtained by simply taking expectations with respect to the distribution \( \pi \). More explicitly, think of a random variable \( X_\infty \) having probability mass function \( P\{X_\infty=i\}=\pi_i \). Then, by definition,

\[
\sum_{i \in S} f(i)\pi_i = \mathbb{E}_\pi f(X_\infty).
\]

Therefore, another, more suggestive, way to write the last result is

\[
P \left( \frac{1}{n} \sum_{k=0}^{n-1} f(X_k) \to \mathbb{E}_\pi f(X_\infty), \text{ as } n \to \infty \right) = 1.
\]

Example 2.5.23. Consider the Markov chain with state space \( \{1, 2, 3\} \) and transition matrix

\[
P = \begin{bmatrix}
1/3 & 2/3 & 0 \\
1/4 & 1/2 & 1/4 \\
1 & 0 & 0
\end{bmatrix}.
\]  

(2.29)

It is simply to check that the unique stationary distribution of this chain is \( \pi = [3/8, 1/2, 1/8] \). Therefore, for example, \( \lim_{n \to \infty} P\{X_n = 3\} = 1/8 \). However, we can also approximate this value using Theorem 2.5.22. Figure 2.5.1 plots \( (1/n) \sum_{k=0}^{n-1} 1\{X_k=3\} \) versus \( n \) for one realization of the chain. We see it appears to converge to 1/8. \( \square \)

2.6 Transient behavior

In this section we ask the following questions for Markov chains with finite state spaces.

1. How many steps do we expect the chain to make before being absorbed by a recurrent class if \( X_0 = i \) is a transient state?
Figure 2.5.1: \( (1/n) \sum_{k=0}^{n-1} 1_{\{X_k = 3\}} \) versus \( n \) for one realization of the Markov chain with transition matrix \( (2.29) \). A line of height \( 0.125 = 1/8 \) has been added for reference.

2. For given states \( i, j \in S \) of an irreducible chain, what is the expected number of needed steps to go from state \( i \) to state \( j \)?

3. If \( X_0 = j \) is a transient state, and the recurrent classes are denoted \( R_1, R_2, \ldots \), what is the probability that the chain eventually ends up in recurrent class \( R_k \)?

We answer these questions sequentially and note that much of the treatment presented here follows Section 1.5 in Greg Lawler’s book [32].

**Question 1.** We let \( P \) be the transition matrix for some finite Markov chain \( X_n \). We recall that after a possible reordering of the indices, we can write \( P \) as

\[
P = \begin{bmatrix} \tilde{P} & 0 \\ S & Q \end{bmatrix},
\]

where \( \tilde{P} \) is the transition matrix for only those states associated with recurrent states, \( Q \) is the submatrix of \( P \) giving the transition probabilities from the transient states to the transient states, and \( S \) is the submatrix of \( P \) giving the transition probabilities from the transient states to the recurrent states. Raising powers of \( P \) in the form \( (2.30) \) yields,

\[
P^n = \begin{bmatrix} \tilde{P}^n & 0 \\ S^n & Q^n \end{bmatrix}.
\]
For example, consider the the Markov chain with state space \(\{1, 2, 3, 4\}\) and transition matrix given by (2.14),

\[
P = \begin{bmatrix}
\frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 \\
\frac{1}{3} & \frac{2}{3} & 0 & 0 \\
0 & 0 & \frac{1}{3} & \frac{2}{3} \\
0 & 0 & \frac{3}{4} & \frac{1}{4}
\end{bmatrix}.
\]

After reordering the elements of the state space as \(\{3, 4, 1, 2\}\) the new transition matrix is

\[
\begin{bmatrix}
\frac{1}{3} & \frac{2}{3} & 0 & 0 \\
\frac{3}{4} & \frac{1}{4} & 0 & 0 \\
\frac{1}{4} & 0 & \frac{1}{2} & \frac{1}{4} \\
0 & 0 & \frac{1}{3} & \frac{2}{3}
\end{bmatrix},
\]

and for this example

\[
\tilde{P} = \begin{bmatrix}
\frac{1}{3} & \frac{2}{3} \\
\frac{3}{4} & \frac{1}{4}
\end{bmatrix}, \quad Q = \begin{bmatrix}
\frac{1}{2} & \frac{1}{4} \\
\frac{1}{3} & \frac{2}{3}
\end{bmatrix}, \quad \text{and} \quad S = \begin{bmatrix}
\frac{1}{4} & 0 \\
0 & 0
\end{bmatrix}.
\]

Note that, in general, \(S\) will not be a square matrix.

The matrix \(Q\) will always be a substochastic matrix, meaning the row sums are less than or equal to one, with at least one row summing to a value that is strictly less than one.

**Proposition 2.6.1.** Let \(Q\) be a substochastic matrix. Then the eigenvalues of \(Q\) all have absolute values strictly less than one.

The above proposition can be proved in a number of ways using basic linear algebra techniques. However, for our purposes it may be best to understand it in the following probabilistic way. Because each of the states represented by \(Q\) are transient, we know that \(Q^n\), which gives the \(n\) step transition probabilities between the transient states, converges to zero, implying the result.

Because the eigenvalues of \(Q\) have absolute value strictly less than one, the equation \((I_d - Q)\nu = 0\), where \(I_d\) is the identity matrix with the same dimensions as \(Q\), has no solutions. Thus, \(I_d - Q\) is invertible and we define

\[
M \overset{\text{def}}{=} (I_d - Q)^{-1} = I_d + Q + Q^2 + \cdots,
\]

where the second equality follows from the identity

\[
(I_d + Q + Q^2 + \cdots)(I_d - Q) = I_d.
\]

Now consider a transient state \(j\). We let \(R_j\) denote the total number of visits to \(j\),

\[
R_j = \sum_{n=0}^{\infty} 1_{\{X_n = j\}},
\]
where we explicitly note that if the chain starts in state \( j \), then we count that as one visit. Note that \( R_j < \infty \) with a probability of one no matter the initial condition since \( j \) is transient.

Suppose that \( X_0 = i \), where \( i \) is also transient. Then,

\[
\mathbb{E}[R_j \mid X_0 = i] = \sum_{n=0}^{\infty} P\{X_n = j \mid X_0 = i\} = \sum_{n=0}^{\infty} p_{ij}^{(n)}.
\]

Therefore, we have shown that \( \mathbb{E}[R_j \mid X_0 = i] \) is the \( i, j \)th entry of

\[
I_d + P + P^2 + \cdots,
\]

which, because both \( i \) and \( j \) are transient, is the same as the \( i, j \)th entry of

\[
I_d + Q + Q^2 + \cdots = (I - Q)^{-1}.
\]

Therefore, we can conclude that the expected number of visits to state \( j \), given that the chain starts in state \( i \), is \( M_{ij} \), defined in (2.32).

For example, consider the Markov chain with transition matrix (2.31). For this example, the matrix \( M \) for the transient states \( \{1, 2\} \) is

\[
M = (I_d - Q)^{-1} = \begin{bmatrix} 4 & 3 \\ 4 & 6 \end{bmatrix}.
\]

We see that starting in state 1, for example, the expected number of visits to state 2 before being absorbed to the recurrent states is equal to \( M_{12} = 3 \). Starting in state 2, the expected number of visits to state 2 (including the first) is \( M_{22} = 6 \). Now suppose we want to know the total number of visits to any recurrent state given that \( X_0 = 2 \). This value is given by

\[
\mathbb{E}_2 R_1 + \mathbb{E}_2 R_2 = M_{21} + M_{22} = 10,
\]

and we see that we simply need to sum the second row of \( M \). We also see that the expected total number of steps needed to transition from state 2 to a recurrent state is 10.

More generally, we have shown the following.

**Proposition 2.6.2.** Consider a Markov chain with transition matrix \( P \) given by (2.31). Then, with \( M \) defined via (2.32), and states \( i, j \) both transient, \( M_{ij} \) gives the expected number of visits to the transient state \( j \) given that \( X_0 = i \). Further, if we define 1 to be the vector consisting of all ones, then \( M1 \) is a vector whose \( i \)th component gives the total expected number of visits to transient states, given that \( X_0 = i \), before the chain is absorbed by the recurrent states.

**Question 2.** We now turn to the second question posed at the beginning of this section: for given states \( i, j \in S \) of an irreducible chain, what is the expected number of needed steps to go from state \( i \) to state \( j \)?
With the machinery just developed, this problem is actually quite simple now. We begin by reordering the state space so that \( j \) is the first element. Hence, the transition matrix can be written as

\[
P = \left[ \frac{p_{jj}}{S} U \right] Q,
\]

where \( Q \) is a substochastic matrix and the row vector \( U \) has the transition probabilities from \( j \) to the other states. Next, simply note that the answer to the question of how many steps are required to move from \( i \) to \( j \) would be unchanged if we made \( j \) an absorbing state. Thus, we can consider the problem on the system with transition matrix

\[
\tilde{P} = \left[ \frac{1}{S} Q \right],
\]

where all notation is as before. However, this is now exactly the same problem solved above and we see the answer is \( M_{1i} \), where all notation is as before.

**Example 2.6.3** (Taken from Lawler, [32]). Suppose that \( P \) is the transition matrix for random walk on \( \{0, 1, 2, 3, 4\} \) with reflecting boundary:

\[
P = \begin{bmatrix}
0 & 0 & 1 & 0 & 0 \\
1 \frac{1}{2} & 0 & 1 \frac{1}{2} & 0 & 0 \\
2 & 0 & 1 \frac{1}{2} & 0 & 1 \frac{1}{2} & 0 \\
3 & 0 & 0 & 1 \frac{1}{2} & 0 & 1 \frac{1}{2} \\
4 & 0 & 0 & 0 & 1 & 0
\end{bmatrix}.
\]

If we let \( i = 0 \), then

\[
Q = \begin{bmatrix}
0 & 1 \frac{1}{2} & 0 & 0 \\
1 \frac{1}{2} & 0 & 1 \frac{1}{2} & 0 \\
0 & 1 \frac{1}{2} & 0 & 1 \frac{1}{2} \\
0 & 0 & 1 & 0
\end{bmatrix}, \quad M = (I - Q)^{-1} = \begin{bmatrix}
2 & 2 & 2 & 1 \\
2 & 4 & 4 & 2 \\
2 & 4 & 6 & 3 \\
2 & 4 & 6 & 4
\end{bmatrix}.
\]

Thus,

\[
M_{1} = (7, 12, 15, 16),
\]

Therefore, the expected number of steps needed to get from state 3 to state 0 15.

**Example 2.6.4.** Consider the Jukes-Cantor model of DNA mutation. The transition matrix for this model is

\[
P = \begin{bmatrix}
1 - \rho & \rho/3 & \rho/3 & \rho/3 \\
\rho/3 & 1 - \rho & \rho/3 & \rho/3 \\
\rho/3 & \rho/3 & 1 - \rho & \rho/3 \\
\rho/3 & \rho/3 & \rho/3 & 1 - \rho
\end{bmatrix}.
\]

If at time zero the nucleotide is in state 1, how many steps do we expect to take place before it enters states 3 or 4? Recalling that the different states are A, G, C, and T,
we note that A (adenine) and G (guanine) are purines and that C (cytosine) and T (thymine) are pyrimidines. Thus, this question is asking for the expected time until a given purine converts to a pyrimidine.

We make \{3, 4\} absorbing states, reorder the state space as \{3, 4, 1, 2\} and note that the new transition matrix is

\[
\begin{bmatrix}
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
\frac{\rho}{3} & \frac{\rho}{3} & 1 - \rho & \frac{\rho}{3} \\
\frac{\rho}{3} & \frac{\rho}{3} & \rho/3 & 1 - \rho
\end{bmatrix},
\]

with \( Q \) and \( M = (I_d - Q)^{-1} \) given via

\[
Q = \begin{bmatrix}
1 - \rho & \rho/3 \\
\rho/3 & 1 - \rho
\end{bmatrix}, \quad \text{and} \quad M = \begin{bmatrix}
\frac{9}{8\rho} & \frac{3}{8\rho} \\
\frac{3}{8\rho} & \frac{9}{8\rho}
\end{bmatrix}.
\]

Therefore, the expected number of transitions needed to go from state 1 (A) to states 3 or 4 (C or T) is

\[
M_{11} + M_{12} = \frac{9}{8\rho} + \frac{3}{8\rho} = \frac{3}{2\rho}.
\]

Note that this value goes to \( \infty \) as \( \rho \to 0 \), which is reasonable.

\[\square\]

**Question 3.** We turn now to the third question laid out at the beginning of this section: if \( X_0 = j \) is a transient state, and the recurrent classes are denoted \( R_1, R_2, \ldots \), what is the probability that the chain eventually ends up in recurrent class \( R_k \)? Note that this question was asked first in and around equation (2.23).

We begin by noting that we can assume that each recurrent class consists of a single point (just group all the states of a class together). Therefore, we denote the recurrent classes as \( r_1, r_2, \ldots \), with \( p_{r_i, r_i} = 1 \). Next, we let \( t_1, t_2, \ldots \) denote the transient states. We may now write the transition matrix as

\[
P = \begin{bmatrix}
I & 0 \\
S & Q
\end{bmatrix},
\]

where we put the recurrent states first. For any transient state \( t_i \) and recurrent class \( k \), we define

\[
\alpha_k(t_i) \overset{\text{def}}{=} P\{X_n = r_k \text{ for some } n \geq 0 \mid X_0 = t_i\}.
\]
For a recurrent states $r_k, r_i$ we set $\alpha_k(r_k) \equiv 1$, and $\alpha_k(r_i) = 0$, if $i \neq k$. Then, for any transient state $t_i$ we have
\[
\alpha_k(t_i) = P\{X_n = r_k \text{ for some } n \geq 0 \mid X_0 = t_i\} = \sum_{j \in S} p_{t_i,j} \alpha_k(j) = \sum_{r_j} p_{t_i,r_j} \alpha_k(r_j) + \sum_{t_j} p_{t_i,t_j} \alpha_k(t_j) = p_{t_i,r_k} + \sum_{t_j} p_{t_i,t_j} \alpha_k(t_j),
\]
where the first sum was over the recurrent states and the second (and remaining) sum is over the transient states. If $A$ is the matrix whose $i,k$th entry is $\alpha_k(t_i)$, then the above can be written in matrix form:
\[
A = S + QA.
\]
Again letting $M = (I - Q)^{-1}$ we have
\[
A = (I - Q)^{-1}S = MS.
\]

**Example 2.6.5.** Consider again the Markov chain with state space $\{3, 4, 1, 2\}$ and transition matrix
\[
\begin{pmatrix}
1/3 & 2/3 & 0 & 0 \\
3/4 & 1/4 & 0 & 0 \\
1/4 & 0 & 1/2 & 1/4 \\
0 & 0 & 1/3 & 2/3
\end{pmatrix}
\]
Note that for this example, we know that we must enter state 3 before state 4, so it is a good reality check on our analysis above. We again have
\[
M = \begin{bmatrix} 4 & 3 \\ 4 & 6 \end{bmatrix}, \quad \text{and} \quad S = \begin{bmatrix} 1/4 & 0 \\ 0 & 0 \end{bmatrix},
\]
and so
\[
MS = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix},
\]
as expected. \hfill \Box

**Example 2.6.6** (Taken from Lawler, [32]). As an example, consider random walk with absorbing boundaries. We order the states $S = \{0, 4, 1, 2, 3, \}$ and have
\[
P = \begin{bmatrix}
1 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 \\
1/2 & 0 & 0 & 1/2 & 0 \\
0 & 0 & 1/2 & 0 & 1/2 \\
0 & 1/2 & 0 & 1/2 & 0
\end{bmatrix}
\]
Then,

\[
S = \begin{bmatrix}
1/2 & 0 \\
0 & 0 \\
0 & 1/2
\end{bmatrix}, \quad
M = \begin{bmatrix}
3/2 & 1 & 1/2 \\
1 & 2 & 1 \\
1/2 & 1 & 3/2
\end{bmatrix}, \quad
MS = \begin{bmatrix}
3/4 & 1/4 \\
1/2 & 1/2 \\
1/4 & 3/4
\end{bmatrix}
\]

Thus, starting at state 1, the probability that the walk is eventually absorbed at state 0 is 3/4.

\[\square\]

### 2.7 Exercises

1. Suppose there are three white and three black balls in two urns distributed so that each urn contains three balls. We say the system is in state \( i \), \( i = 0, 1, 2, 3 \), if there are \( i \) white balls in urn one. At each stage one ball is drawn at random from each urn and interchanged. Let \( X_n \) denote the state of the system after the \( n \)th draw. What is the transition matrix for the Markov chain \( \{X_n : n \geq 0\} \).

2. (Success run chain.) Suppose that Jake is shooting baskets in the school gym and is very interested in the number of baskets he is able to make in a row. Suppose that every shot will go in with a probability of \( p \in (0, 1) \), and the success or failure of each shot is independent of all other shots. Let \( X_n \) be the number of shots he has currently made in a row after \( n \) shots (so, for example, \( X_0 = 0 \) and \( X_1 \in \{0, 1\} \), depending upon whether or not he hit the first shot). Is it reasonable to model \( X_n \) as a Markov chain? What is the state space? What is the transition matrix?

3. (Jukes-Cantor model of DNA mutations) Consider a single nucleotide on a strand of DNA. We are interested in modeling possible mutations to this single spot on the DNA. We say that \( X_n \) is in state 1, 2, 3, or 4, if the nucleotide is the base A, G, C, or T, respectively. We assume that there is a probability, \( \rho \in (0, 1) \), that between one time period and the next, we will observe a change in this base. If it does change, we make the simple assumption that each of the other three bases are equally likely.

(a) What is the transition matrix for this Markov chain?

(b) If \( \rho = 0.01 \), what are (approximately): \( p_{13}^{(10)} \), \( p_{13}^{(100)} \), \( p_{13}^{(1,000)} \), \( p_{13}^{(10,000)} \)? (You should use MATLAB for this problem.)

4. Suppose that whether or not it rains tomorrow depends on previous weather conditions only through whether or not it is raining today. Assume that the probability it will rain tomorrow given it rains today is \( \alpha \) and the probability it will rain tomorrow given it is not raining today is \( \beta \). Supposing that the state space is \( S = \{0, 1\} \), where state 0 means it rains and state 1 means it does not rain on a given day, what is the transition matrix when we model this situation with a Markov chain. If we assume there is a 40\% chance of rain today, what is the probability it will rain three days from now if \( \alpha = 6/10 \) and \( \beta = 3/10 \).
5. Verify the condition (2.5). Hint, use an argument like equation (2.7).

6. (a) Show that the product of two stochastic matrices is stochastic.

(b) Show that for stochastic matrix $P$, and any row vector $\pi$, we have $\|\pi P\|_1 \leq \|\pi\|_1$, where $\|v\|_1 = \sum_i |v_i|$. Deduce that all eigenvalues, $\lambda$, of $P$ must satisfy $|\lambda| \leq 1$.

Hint: For the first part of (b) just write out each side and manipulate the expressions as needed. For the second part, just use the definition of a left-eigenvector and follow your nose.

7. Let $X_n$ denote a discrete time Markov chain with state space $S = \{1, 2, 3, 4\}$, transition Matrix

\[
P = \begin{bmatrix}
1/4 & 0 & 1/5 & 11/20 \\
0 & 0 & 0 & 1 \\
1/6 & 1/7 & 0 & 29/42 \\
1/4 & 1/4 & 1/2 & 0
\end{bmatrix},
\]

and initial distribution is $\alpha = (1/4, 0, 1/4, 1/2)^T$.

(a) Suppose that

\[
(\omega_0, \omega_1, \omega_2, \ldots, \omega_{10}) = (0.421, 0.794, 0.311, 0.528, 0.165, 0.602, 0.263, 0.654, 0.689, 0.748, 0.450)
\]

is a sequence of 11 independent uniform(0,1) random variables. Using these random variables (in the order presented above) and the construction of Section 2.2, what are $X_n$, $n \in \{0, 1, \ldots, 10\}$? Note, you are supposed to do this problem by hand (without the aid of a computer).

(b) Using Matlab, simulate a path of $X_n$ up to time $n = 100$ using the construction of Section 2.2. A helpful sample Matlab code has been provided on the course website that is intended to get you started (make sure you read my comments in the code). Play around with your script. Try different values of $n$ and see the behavior of the chain. Turn in your MATLAB code along with a plot of a realization of your chain.

8. Consider a chain with state space $\{0, 1, 2, 3, 4, 5\}$ and transition matrix

\[
P = \begin{bmatrix}
1/2 & 0 & 0 & 0 & 1/2 & 0 \\
0 & 3/4 & 1/4 & 0 & 0 & 0 \\
0 & 1/8 & 7/8 & 0 & 0 & 0 \\
1/2 & 1/4 & 1/4 & 0 & 0 & 0 \\
1/3 & 0 & 0 & 0 & 2/3 & 0 \\
0 & 0 & 0 & 1/2 & 0 & 1/2
\end{bmatrix}
\]

What are the communication classes? Which classes are closed? Which classes are recurrent and which are transient?
9. Consider a finite state space Markov chain, $X_n$. Suppose that the recurrent communication classes are $R_1, R_2, \ldots, R_m$. Suppose that restricted to $R_k$, the Markov chain is irreducible and aperiodic, and let $\tilde{\pi}^{(k)}$ be the unique limiting stationary distribution for the Markov chain restricted to $R_k$. Now, for each $R_k$, let $\pi_k$ (note the lack of a tilde) be the vector with components equal to those of $\tilde{\pi}^{(k)}$ for those states in $R_k$, and zero otherwise. For example, assuming there are three states in $R_1$,

$$\pi^{(1)} = (\tilde{\pi}_1^{(1)}, \tilde{\pi}_2^{(1)}, \tilde{\pi}_3^{(1)}, 0, 0, \ldots, 0),$$

and if there are two states in $R_2$, then

$$\pi^{(2)} = (0, 0, 0, \tilde{\pi}_1^{(2)}, \tilde{\pi}_2^{(2)}, 0, \ldots, 0),$$

Prove both the following:

(a) Each linear combination

$$a_1 \pi^{(1)} + \cdots + a_m \pi^{(m)}$$

with $a_i \geq 0$ and $\sum_i a_i = 1$, is a stationary distribution for the unrestricted Markov chain, $X_n$.

(b) All stationary distributions of the Markov chain $X_n$ can be written as such a linear combination. (Hint: use the general form of the transition matrix given by equation (2.16). Now, break up an arbitrary stationary distribution, $\pi$, into the different components associated with each communication class. What can be concluded about each piece of $\pi$?)

10. Consider the Markov chain described in Problem 3 above. What is the stationary distribution for this Markov chain? Interpret this result in terms of the probabilities of the nucleotide being the different possible values for large times. Does this result make sense intuitively?

11. Show that the success run chain of Problem 2 above is positive recurrent. What is the stationary distribution of this chain? Using the stationary distribution, what is the expected number of shots Jake will hit in a row.

12. Let $X_n$ be the number of customers in line for some service at time $n$. During each time interval, we assume that there is a probability of $p$ that a new customer arrives. Also, with probability $q$, the service for the first customer is completed and that customer leaves the queue. Assuming at most one arrival and at most one departure can happen per time interval, the transition probabilities are

$$p_{i,i-1} = q(1-p), \quad p_{i,i+1} = p(1-q)$$

$$p_{ii} = 1 - q(1-p) - p(1-q), \quad i > 0$$

$$p_{00} = 1 - p, \quad p_{01} = p.$$
(a) Argue why the above transition probabilities are the correct ones for this model.

(b) For which values of $p$ and $q$ is the chain null recurrent, positive recurrent, transient?

(c) For the positive recurrent case, give the limiting probability distribution $\pi$. (Hint: note that the equation for $\pi_0$ and $\pi_1$ are both different than the general $n$th term.)

(d) Again in the positive recurrent case, using the stationary distribution you just calculated, what is the expected length of the queue in equilibrium? What happens to this average length as $p \to q$. Does this make sense?

13. This problem has you redo the computation of Example 2.5.23, though with a different Markov chain. Suppose our state space is $\{1, 2, 3, 4\}$ and the transition matrix is

\[
P = \begin{bmatrix}
\frac{1}{4} & 0 & \frac{1}{5} & \frac{11}{20} \\
0 & 0 & 0 & 1 \\
\frac{1}{6} & \frac{1}{7} & 0 & \frac{29}{42} \\
\frac{1}{4} & \frac{1}{4} & \frac{1}{2} & 0
\end{bmatrix},
\]

which was the transition matrix of problem 7 above. Using Theorem 2.5.22, and MATLAB, estimate $\lim_{n \to \infty} P\{X_n = 2\}$. Make sure you choose a long enough path, and that you plot your output (to turn in). Compare your solution with the actual answer computed via the left eigenvector with an eigenvalue of one.

14. (Taken from Lawler, [32]) You will need software for this problem to deal with the matrix manipulations. Suppose that we flip a fair coin repeatedly until we flip four consecutive heads. What is the expected number of flips that are needed? (Hint: consider a Markov chain with state space $\{0, 1, 2, 3, 4\}$.)

15. You will need software for this problem to deal with the matrix manipulations. Consider a Markov chain $X_n$ with state space $\{0, 1, 2, 3, 4, 5\}$ and transition matrix

\[
P = \begin{bmatrix}
\frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} & 0 \\
0 & \frac{3}{4} & \frac{1}{4} & 0 & 0 & 0 \\
0 & \frac{1}{8} & \frac{7}{8} & 0 & 0 & 0 \\
\frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 & 0 & 0 \\
\frac{1}{3} & 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} \\
0 & 0 & \frac{1}{4} & \frac{1}{4} & 0 & \frac{1}{2}
\end{bmatrix}
\]

Here the only recurrent class is $\{1, 2\}$. Suppose that $X_0 = 0$ and let

\[T = \inf\{n : X_n \in \{1, 2\}\}.\]

(a) What is $ET$?
(b) What is $P\{X_T = 1\}$? $P\{X_T = 2\}$? (Note that this is asking for the probabilities that when the chain enters the recurrent class, it enters into state 1 or 2.)

16. (Taken from Lawler, [32]) You will need software for this problem to deal with the matrix manipulations. Let $X_n$ and $Y_n$ be independent Markov chains with state space $\{0, 1, 2\}$ and transition matrix

$$
P = \begin{bmatrix}
\frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\
\frac{1}{4} & \frac{1}{4} & \frac{1}{2} \\
0 & \frac{1}{2} & \frac{1}{2}
\end{bmatrix}.
$$

Suppose that $X_0 = 0$ and $Y_0 = 2$ and let

$$
T = \inf\{n : X_n = Y_n\}.
$$

A hint for all parts of this problem: consider the nine-state Markov chain $Z_n = (X_n, Y_n)$.

(a) Find $\mathbb{E}(T)$.

(b) What is $P\{X_T = 2\}$?

(c) In the long run, what percentage of the time are both chains in the same state?
Chapter 3

Discrete Time Markov Chain Models in the Biosciences

In this Chapter, we review some basic discrete time Markov chain models used in the biosciences. In Section 3.1 we discuss models in genetics, in Section 3.2 we discuss discrete time birth and death models, and in Section 3.3 we discuss branching processes.

3.1 Genetic Models

3.1.1 Mendelian Genetics

One of the greatest scientific achievements of the past 200 years was due to Gregor Mendel who, in his work on the inheritance traits of pea plants, founded the science of genetics. Mendel predicted that each gene (although there was not even a notion that we had something called genes in his day) could come in multiple different alleles. The different observable traits would then depend upon which different types of alleles were possessed. For example, in his work on pea plants Mendel predicted that there were alleles for, among others: tall and dwarf plants, round and wrinkled seeds, and yellow and green seeds.

To come to his conclusions, Mendel made the following series of experiments and observations. First, if he took tall plants that had been bred from a line of only tall plants, and cross-bred them with dwarf plants that had been bred from a line of only dwarf plants, then the resulting plants were all tall. Next, he took these plants and produced a second generation of plants. In this second generation, roughly a quarter of the plants were now dwarfs, even though none of the plants from the first generation were dwarfs! Mendel observed the above behavior of one trait dominating another in the first generation, followed by a 3:1 ratio in the second generation for all traits he had under consideration.

From these observations Mendel made the following conclusions, which still form the basic understanding we have today. Each plant has two genes, one from each...
parent. Each gene can come in one of two alleles, call them $A$ and $a$, with one being dominant, such as “tall” in the example above. Thus, when the gene sequence is either $AA$ or $Aa$ (which is equivalent to $aA$, order does not matter), the dominant trait is observed, and only when the gene sequence is $aa$ does the recessive trait emerge. The combination of alleles, $AA$, $Aa$, and $aa$ are called genotypes. The observable traits are often called phenotypes.

We now discuss a model of genetic inbreeding that has been formulated in a few places, including [1] and [16]. Consider two randomly chosen individuals, who we will call the first generation, that are mated to yield a second generation. Next, two offspring of opposite sex from this second generation are mated to give rise to a third generation. A fourth generation is then made by choosing two of the offspring of opposite sex from the third generation and mating them. This process then continues indefinitely. We formulate a discrete time Markov chain describing this process in the following way. We let $X_n$ characterize the genotypes of both parents in the above formulation. Therefore, the different possible values for $X_n$ are:

<table>
<thead>
<tr>
<th>$X_n$</th>
<th>Genotypes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$AA \times AA$</td>
</tr>
<tr>
<td>2</td>
<td>$AA \times Aa$</td>
</tr>
<tr>
<td>3</td>
<td>$Aa \times Aa$</td>
</tr>
<tr>
<td>4</td>
<td>$Aa \times aa$</td>
</tr>
<tr>
<td>5</td>
<td>$AA \times aa$</td>
</tr>
<tr>
<td>6</td>
<td>$aa \times aa$</td>
</tr>
</tbody>
</table>

where we note that order of the genotypes of the individuals does not matter. We must now find the associated transition probabilities.

First, note that if the genotype pair of a couple is $AA \times AA$, that is if both parents have genotype $AA$, then all of their offspring necessarily have genotype $AA$, and so $p_{11} = 1$. Similarly, we have $p_{66} = 1$. Now consider the possibility that the genotype pairing is of type 2, that is $AA \times Aa$. Consider a specific offspring of this pairing and let $Z_1 \in \{A, a\}$ be the allele passed down to the offspring from the parent with genotype $AA$, and let $Z_2 \in \{A, a\}$ be the allele passed down to the offspring from the parent with genotype $Aa$. Finally, let $Y$ be the genotype of a randomly selected offspring. We then have

\[
P\{Y = AA\} = P\{Z_1 = A, Z_2 = A\} = P\{Z_1 = A\}P\{Z_2 = A\} = 1 \times \frac{1}{2} = \frac{1}{2}
\]

\[
P\{Y = Aa\} = P\{Z_1 = A, Z_2 = a\} + P\{Z_1 = a, Z_2 = A\}
\]

\[
= P\{Z_1 = A\}P\{Z_2 = a\} + P\{Z_1 = a\}P\{Z_2 = A\}
\]

\[
= \frac{1}{2} + 0
\]

\[
= \frac{1}{2}
\]

\[
P\{Y = aa\} = P\{Z_1 = a, Z_2 = a\} = P\{Z_1 = a\}P\{Z_2 = a\} = 0.
\]

Thus, letting $Y_1$ and $Y_2$ be the genotypes of two randomly chosen offspring, we have
that

\[
P\{Y_1 \times Y_2 = AA \times AA\} = P\{Y_1 = AA\}P\{Y_2 = AA\} = \frac{1}{4}
\]

\[
P\{Y_1 \times Y_2 = AA \times Aa\} = P\{Y_1 = AA\}P\{Y_2 = Aa\} = \frac{1}{2}
\]

\[
P\{Y_1 \times Y_2 = Aa \times AA\} = P\{Y_1 = Aa\}P\{Y_2 = AA\} = \frac{1}{4}
\]

\[
P\{Y_1 \times Y_2 = Aa \times Aa\} = P\{Y_1 = Aa\}P\{Y_2 = Aa\} = \frac{1}{4}.
\]

Thus, we have that

\[
p_{21} = \frac{1}{4}, \quad p_{22} = \frac{1}{2}, \quad \text{and} \quad p_{23} = \frac{1}{4},
\]

where the second equality holds because \(AA \times Aa = Aa \times AA\). Continuing in this manner, the transition matrix for this six state Markov chain is

\[
P = \begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 \\
1/4 & 1/2 & 1/4 & 0 & 0 & 0 \\
1/16 & 1/4 & 1/4 & 1/8 & 1/16 \\
0 & 0 & 1/4 & 1/2 & 0 & 1/4 \\
0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1
\end{pmatrix},
\]

where the states have been listed to the left of the transition matrix for reference. Note that the communication classes are \(\{1\}, \{6\}\), and \(\{2, 3, 4, 5\}\) with only the first two classes being recurrent (as states 1 and 6 are absorbing).

We wish to calculate the relevant transition probabilities discussed in Section 2.6. Therefore, we reorder the states to be 1, 6, 2, 3, 4, 5, yielding the transition matrix

\[
\tilde{P} = \begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1/2 & 1/4 & 0 & 0 \\
0 & 1/4 & 1/16 & 1/4 & 1/4 & 1/8 \\
0 & 0 & 1/4 & 0 & 1/2 & 0 \\
0 & 0 & 0 & 0 & 1 & 0
\end{pmatrix}.
\]
Therefore, using the notation of Section 2.6,

\[
Q = \begin{pmatrix}
  1/2 & 1/4 & 0 & 0 \\
  1/4 & 1/4 & 1/4 & 1/8 \\
  0 & 1/4 & 1/2 & 0 \\
  0 & 1 & 0 & 0
\end{pmatrix}
\]

\[
M = (I_d - Q)^{-1} = \begin{pmatrix}
  8/3 & 4/3 & 2/3 & 1/6 \\
  4/3 & 8/3 & 4/3 & 1/3 \\
  2/3 & 4/3 & 8/3 & 1/6 \\
  4/3 & 8/3 & 4/3 & 4/3
\end{pmatrix}
\]

\[
S = \begin{pmatrix}
  1/4 & 0 \\
  1/16 & 1/16 \\
  0 & 1/4 \\
  0 & 0
\end{pmatrix},
\]

and

\[
MS = \begin{pmatrix}
  3/4 & 1/4 \\
  1/2 & 1/2 \\
  1/4 & 3/4 \\
  1/2 & 1/2
\end{pmatrix}
\]

Note also that

\[
M1 = \begin{pmatrix}
  29/6 \\
  17/3 \\
  29/6 \\
  20/3
\end{pmatrix}
\]

We may now make a few conclusions.

1. If \( X_0 = 2 \), i.e. genotype \( AA \times Aa \), then the probability of eventual absorption into state 1, \( AA \times AA \), is 3/4 and the probability of absorption into state 2, \( aa \times aa \), is 1/4. This is seen in the first row of the matrix \( MS \).

2. If \( X_0 \in \{3, 5\} \), that is if the initial genotype pairing is of type \( Aa \times Aa \) or \( AA \times aa \), then there are equal probabilities of ending up in states 1 and 6. This should be intuitive via symmetry of the alleles \( A \) and \( a \).

However, the expected number of steps to achieve such absorption is not the same. It is 5.666... if \( X_0 = 3 \), and is 6.666... if \( X_0 = 5 \). This can be understood by observing that if the initial genotype is \( AA \times aa \), then all offspring of the original pairing will have genotype \( A \times a \).

Of course, more conclusions than the above are readily available, and we leave them to the interested reader.
3.1.2 The Wright-Fischer Model

We consider another model from population genetics, which was first developed by Fischer, and later extended by Wright. In this model we assume the existence of $N$ diploid (two copies of each gene) individuals. Thus, there are a total of $2N$ genes in the gene pool. We make the following assumptions:

1. The number of individuals remains constant at $N$ from generation to generation.

2. The genes for any individual in the $(n+1)$st generation are randomly selected (with replacement) from the pool of genes in the $n$th generation.

Note that the last assumption allows us to disregard the individuals, and only consider the gene pool itself. The model is by now quite classical and should be regarded as the starting building block for such genetic models. The above assumptions can be weakened substantially if so desired, although it is usually most instructive to start with such a simple model.

We suppose we have two alleles of the gene in question, which we denote by $A$ and $a$. We let $X_n \in \{0, 1, \ldots, 2N\}$ denote the number of alleles of type $A$ in the entire gene pool. Oftentimes $A$ is assumed to be a mutant that has entered the population. We are interested in the probabilities associated with fixation, meaning when the system becomes homogeneous in $A$, which occurs when $X_n = 2N$ and $A$ has overtaken the population, or in $a$, which occurs when $X_n = 0$.

We build our Markov model by finding the transition probabilities. Supposing that $X_n = i$, for some $i \geq 0$, what is the probability that $X_{n+1} = j$? Because of our simplifying assumptions, we see that, conditioned on $X_n$, the value of $X_{n+1}$ is a binomial random variable with parameters $2N$ and $X_n / (2N)$. Therefore,

$$p_{ij} = \binom{2N}{j} \left( \frac{i}{2N} \right)^j \left( 1 - \frac{i}{2N} \right)^{2N-j} \text{ for } i, j \in \{0, \ldots, 2N\}. \tag{8.5}$$

We now let $\alpha_0(j)$ denote the probability that the chain eventually ends up in state 0, given that it started in state $j$. That is, it is the probability that the allele $A$ eventually dies out completely from the population. We recall that we have methods, from Section 2.6, for the calculation of such probabilities. However, it is not clear how to use those methods here. Trying a first step analysis, we are led to the system of equations,

$$\alpha_0(j) = 1 \cdot p_{j0} + \alpha_0(1)p_{j1} + \cdots + \alpha_0(2N-1)p_{j2N-1}, \quad j \in \{1, 2, \ldots, 2N-1\}. \tag{8.6}$$

This looks quite scary and is not very promising. A different method is required to solve this problem elegantly.

The method is essentially a martingale method. We note that

$$\mathbb{E} \left[ \frac{X_{n+1}}{2N} \middle| X_n \right] = \frac{X_n}{2N}. \tag{8.7}$$
Therefore, if $X_0 = j$, say, then

$$\lim_{n \to \infty} E[X_n] = j.$$ 

However, assuming I can pass the limit inside (and we can because of something called the optional stopping theorem),

$$j = \lim_{n \to \infty} E[X_n] = E \lim_{n \to \infty} X_n = 0\alpha_0(j) + 2N(1 - \alpha_0(j)).$$

Solving yields

$$\alpha_0(j) = \frac{2N - j}{2N}.$$ 

Note also that $\alpha_{2N}(j) = j/(2N)$.

### 3.1.3 Phylogenetic Distance and the Jukes-Cantor Model

We present a model of phylogenetic distance called the Jukes-Cantor distance. See Section 4.5 of [2] for a more thorough formulation of this model, and see Problem 3 of Chapter 2 for the first appearance of this model in these notes.

Consider a single nucleotide on a strand of DNA. It can take any one of the four values $A, G, C, T$. We say that $X_n$ is in state 1, 2, 3, or 4, if the nucleotide is the base $A$, $G$, $C$, or $T$, respectively. We assume that there is a probability, $\rho \in (0, 1)$, that between one time period and the next (you can think of a generation as a time period), we will observe a change in this base. If it does change, we make the simple assumption that each of the other three bases are equally likely. Weakening of this last assumption leads to other models, such as the Kimura model (which, in turn, leads to “Kimura distances.”) The transition matrix for this Markov chain is

$$P = \begin{pmatrix}
1 - \rho & \rho/3 & \rho/3 & \rho/3 \\
\rho/3 & 1 - \rho & \rho/3 & \rho/3 \\
\rho/3 & \rho/3 & 1 - \rho & \rho/3 \\
\rho/3 & \rho/3 & \rho/3 & 1 - \rho
\end{pmatrix},$$

and we assume that $\rho$ is small.

A very natural question is now the following: given two strands of DNA, one in the present and one in the fossil record, how long have the two species spent evolving from each other? That is, can we use the DNA to date the original fossil? Also, if we know how old the fossil is, can we recover the mutation rate, $\rho$?

We begin by asking for the probability that after $n$ generations, a given, single, nucleotide is the same value as it was at time zero. Note that, by symmetry, this probability will be the same regardless of the initial state. We therefore consider $e_1^T P^n$, where $e_1 = (1, 0, 0, 0)$, which will yield the first (and hence all) diagonal element(s). Note that the left eigenvector of $P$ with associated eigenvalue 1 is

$$\begin{pmatrix}
1/4 \\
1/4 \\
1/4 \\
1/4
\end{pmatrix},$$

66
and that the eigenvalue $1 - (4/3)\rho$ has the three linearly independent eigenvectors

$$
\begin{pmatrix}
\frac{1}{4} \\
0 \\
0 \\
-\frac{1}{4}
\end{pmatrix}, \quad
\begin{pmatrix}
\frac{1}{4} \\
0 \\
-\frac{1}{4} \\
0
\end{pmatrix}, \quad
\begin{pmatrix}
\frac{1}{4} \\
0 \\
0 \\
0
\end{pmatrix}.
$$

Therefore, we may write

$$
\begin{pmatrix}
1 \\
0 \\
0 \\
0
\end{pmatrix} = \begin{pmatrix}
\frac{1}{4} \\
\frac{1}{4} \\
\frac{1}{4} \\
\frac{1}{4}
\end{pmatrix} + \begin{pmatrix}
\frac{1}{4} \\
0 \\
0 \\
-\frac{1}{4}
\end{pmatrix} + \begin{pmatrix}
\frac{1}{4} \\
0 \\
-\frac{1}{4} \\
0
\end{pmatrix} + \begin{pmatrix}
\frac{1}{4} \\
0 \\
0 \\
0
\end{pmatrix},
$$

and conclude

$$e_1^T P^n = (1/4)(1, 1, 1, 1) + \left(1 - \frac{4}{3}\rho\right)^n (3/4, -1/4, -1/4, -1/4).$$

Therefore, the first term of the above vector is the general diagonal term of $P^n$ and is given by

$$d(n) \overset{\text{def}}{=} \frac{1}{4} + \frac{3}{4} \left(1 - \frac{4}{3}\rho\right)^n.$$

Note the fact that $d(0) = 1$, and $\lim_{n \to \infty} d(n) = 1/4$, which should agree with your intuition.

Now we reconsider the two strands of DNA. Suppose we observe that the fraction of nucleotides that are different is $m$. As the probability of a single nucleotide being the same after $n$ time steps is given by $d(n)$ above, we see the expected fraction of nucleotides that should be different is $1 - d(n)$. Note that for very long strands of DNA, the law of large numbers states that $m$ should be very close to $1 - d(n)$. Therefore, we set

$$m = 1 - d(n) = \frac{3}{4} - \frac{3}{4} \left(1 - \frac{4}{3}\rho\right)^n.$$

Solving for $n$ yields

$$n = \frac{\ln(1 - (4/3)m)}{\ln(1 - (4/3)\rho)},$$

which yields $n$ as a function of $m$, which is observable, and $\rho$. If we somehow know $\rho$, perhaps through other fossil records, then we can figure out $n$. However, if $\rho$ is unknown, then we obviously can not find $n$.

However, we define the Jukes-Cantor phylogenetic distance as $d_{JC} \overset{\text{def}}{=} n\rho$, which gives the expected number of mutations, per site, over the elapsed time period. Assuming that $\rho$ is small, which is typically the case, we may make use of the fact that

$$\ln(1 - (4/3)\rho) = -(4/3)\rho + O(\rho^2),$$

which is
to conclude
\[ d_{JC} = \frac{\ln(1 - (4/3)m)}{\ln(1 - (4/3)\rho)} \approx \frac{\ln(1 - (4/3)m)}{-(4/3)\rho} \rho = -\frac{3}{4} \ln \left(1 - \frac{4}{3}m\right), \]
yielding a measure of the difference between the strands of DNA. We again note that if \( \rho \) is unknown, then we were not able to find \( n \). Instead we were able to find a value for the product of \( n \) and \( \rho \). If the age of the older strand of DNA is somehow known or can be estimated (perhaps through a geological record), then the mutation rate \( \rho \) can be recovered.

**Example 3.1.1.** Suppose that two strands of DNA have been found, one from the distant past and one in the present. Suppose further they differ in 22.3% of the nucleotide sites. Then, the Jukes-Cantor distance for the two strands is

\[ d_{JC} = -\frac{3}{4} \ln \left(1 - \frac{4}{3} \cdot 0.223\right) = 0.26465. \]

Note that this value, giving the expected number of substitutions per site over the time period of interest, is higher than the observed 22.3% differences, as \( d_{JC} \) takes into account mutations that have taken place, and then changed back. That is, when \( A \to G \to A \), and we have not observed the (two or more) changes that have taken place. If we also somehow know that the mutation rate is \( \rho = .002/\text{generation} \), then the difference, in generations, between the strands can be estimated to be

\[ n = d_{JC} \frac{1}{\rho} \approx \frac{0.26465}{.002} \approx 132 \text{ generations}. \]

\[ \square \]

### 3.2 Discrete Time Birth and Death models

We will consider models that, intuitively, can be thought of as modeling the size of some population. The possible changes include an increase or a decrease of one (a birth or a death), or no change. Such models are called birth and death models for obvious reasons, though they are not only useful in the study of population processes. For example they are used to model such disparate things as the accumulated winnings of a gambler, the length of a line, and the state of a chemical system. We will use this section as an introduction to the basic idea of a birth and death process and revisit them again in the continuous time setting.

We consider a discrete time Markov chain with the following transition probabilities:

\[
p_{ij} = \begin{cases} 
    p_i, & \text{if } j = i + 1 \\
    q_i, & \text{if } j = i - 1 \\
    1 - p_i - q_i, & \text{if } j = i, \\
    0, & \text{else}
\end{cases}
\]
where \( p_i, q_i \geq 0 \), and \( p_i + q_i \leq 1 \), and the usual state space is \( S = \{0, 1, 2, \ldots\} \), or a finite subset \( \{0, 1, \ldots, N\} \). The values \( p_i \) and \( q_i \) are often called the birth rate and death rate, respectively. Note that to ensure the zero boundary condition, we usually have \( q_0 = 0 \). The key fact to note is that the probabilities associated with an increase and a decrease, denoted \( p_i \) and \( q_i \) respectively, are functions of the state \( i \). That is, they are state dependent functions. Note that a defining characteristic of a birth and death process is that the associated transition matrix is tridiagonal. We start with some examples

**Example 3.2.1.** The deterministically monotone chain of Example 2.1.6 is a birth and death process with \( p_i \equiv 1 \), and hence \( q_i \equiv 0 \), for all \( i \in \{0, 1, \ldots\} \).

**Example 3.2.2.** The Gambler’s ruin problem, which is random walk on \( \{0, 1, \ldots, N\} \) with absorbing boundaries, is a birth and death process with \( p_i \equiv p \) and \( q_i \equiv 1 - p \) for all \( 1 \leq i \leq N - 1 \), and \( p_0 = q_0 = p_N = q_N = 0 \) (and hence \( p_{00} = p_{NN} = 1 \)).

**Example 3.2.3 (A linear reversible chemical reaction).** We consider the chemical system

\[ A \rightleftharpoons B, \]

meaning that there are two types of molecules, \( A \) and \( B \), and they are converting back and forth. We assume that there are a total of \( N \) molecules in the system. That is, if we let \( X_n(1) \) and \( X_n(2) \) denote the number of molecules of chemical species \( A \) and \( B \), respectively, at time \( n \), then \( X_n(1) + X_n(2) = N \) for all \( n \geq 0 \).

We make the following assumptions on the model. We suppose that within a given (small) period of time, the probability that a particular \( A \) molecule converts to a \( B \) molecule is \( c_1 \), and that the probability that a particular \( B \) molecule converts to an \( A \) molecule is \( c_2 \). The constants \( c_1 \) and \( c_2 \) are very small as they scale with the size of the time interval. Therefore, the probability that some \( A \) molecule converts to a \( B \) molecule is \( c_1X_n(1) \) and the probability that some \( B \) molecule converts to an \( A \) molecule is \( c_2X_n(2) \). Finally, we suppose that our time interval is so small that at most one reaction can take place per time interval. Clearly, the probability that no reaction takes place in the time interval is \( 1 - c_1X_n(1) - c_2X_n(2) \). Because we have the relation \( X_n(1) + X_n(2) = N \), we can consider the system as a 1-dimensional birth and death process with state \( X_n(1) \in \{0, 1, \ldots, N\} \) and birth and death rates

\[
\begin{align*}
\text{State 0} : & \quad p_0 = c_2N, & q_0 = 0, \\
\text{State } i \in \{1, \ldots, N - 1\} : & \quad p_i = c_2(N - i), & q_i = c_1i \\
\text{State } N : & \quad p_N = 0, & q_N = c_1N.
\end{align*}
\]

The transition matrix for this Markov chain is therefore

\[
\begin{pmatrix}
1 - c_2N & c_2N & 0 & 0 & 0 & \cdots & 0 \\
1 - c_1 - (N - 1)c_2 & (N - 1)c_2 & 0 & 0 & \cdots & 0 \\
0 & 2c_1 & 1 - 2c_1 - (N - 2)c_2 & (N - 2)c_2 & 0 & \cdots & 0 \\
\vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\
0 & 0 & \cdots & 0 & (N - 1)c_1 & 1 - (N - 1)c_1 - c_2 & c_2 \\
0 & 0 & \cdots & 0 & 0 & c_1N & 1 - c_1N
\end{pmatrix}
\]

\( \square \)
Example 3.2.4 (Population size). Suppose that $X_n$ is the total population on an island and that the growth rate is proportional, with constant of proportionality $\lambda > 0$, to the size of the population so long as the population remains below $N$. Once the population reaches $N$, the growth rate is zero. Further, suppose that the death rate is always proportional to the size of the population, with constant of proportionality $\mu > 0$. Thus, the transition probabilities are

$$p_i = \lambda i, \quad q_i = \mu i, \quad \text{if} \quad 0 < i < N,$$

with boundary conditions $q_0 = p_0 = 0$, $p_N = 0$, and $q_N = \mu N$. Note that having $p_N = 0$ ensures that the state space is $\{0, 1, \ldots, N\}$ so that for $\lambda$ and $\mu$ small enough we can always guarantee that $p_i + q_i \leq 1$. In the continuous time setting we will drop the need for the bounded growth rate.

Example 3.2.5 (Queueing Models). We suppose that $X_n$ represents the number of people in line for some service at time $n$. We suppose that people are arriving at a rate of $\lambda$, i.e. each $p_i = \lambda$, for $i \in S = \{0, 1, 2, \ldots\}$. However, customers may also leave the line because they have been served. We can choose how to model the service times in a number of ways.

(a) (Single server) If there is a single server, and that person always serves the first person in line, then we take $q_i = \mu > 0$ if $i \geq 1$ and $q_0 = 0$ (as there is no one to serve).

(b) ($k$ servers) If there are $k \geq 1$ servers, and the first $k$ people in line are always being served, then for some $\mu > 0$ we take

$$q_i = \begin{cases} 
i\mu, & \text{if } i \leq k \\ k\mu, & \text{if } i \geq k \end{cases}.$$

There are other options for choosing how the people in line can be served, but we delay until later. Note that $\lambda$ and $\mu$ must satisfy $\lambda + k\mu \leq 1$ for this model to make sense.

Recurrence, transience, and stationary distributions

We begin by searching for a condition on the values $p_i$ and $q_i$ that separates recurrent from transient chains. We note that this analysis will also help us understand the behavior of continuous time birth and death models that will appear later in the notes. Before beginning our analysis, however, we restrict our attention to chains that are irreducible. It is easy to see that a birth and death chain is irreducible if and only if $p_i > 0$, for all $i \geq 0$ (except maybe at the upper boundary), and $q_i > 0$, for all $i \geq 1$ (as we need $q_0 = 0$). We will use the conditions for transience used for more general Markov chains in Section 2.5.

Let $\alpha_n$ denote the probability that the chain, starting at state $n \in \{0, 1, 2, \ldots\}$, ever returns to state 0. We clearly have that $\alpha_0 = 1$, and using a first step analysis
and the Markov property we have that for \( n \geq 0 \) \( \alpha_n \) satisfies

\[
\alpha_n = P\{X_i = 0 \text{ for some } i \geq 1 \mid X_0 = n\} \\
= \sum_k P\{X_i = 0 \text{ for some } i \geq 1, X_1 = k \mid X_0 = n\} \\
= \sum_k P\{X_i = 0 \mid X_1 = k\}P\{X_1 = k \mid X_0 = n\} \\
= p_n\alpha_{n+1} + q_n\alpha_{n-1} + (1 - p_n - q_n)\alpha_n,
\]

which yields the relation

\[
(p_n + q_n)\alpha_n = p_n\alpha_{n+1} + q_n\alpha_{n-1},
\]

for \( n \geq 1 \). By Theorem 2.5.20, there will be a solution to these equations with \( \alpha_n \to 0 \) as \( n \to \infty \) if and only if the chain is transient.

Equation 3.3 implies

\[
\alpha_n - \alpha_{n+1} = \frac{q_n}{p_n}[\alpha_{n-1} - \alpha_n],
\]

which is valid for \( n \geq 1 \). Iterating the above yields

\[
\alpha_n - \alpha_{n+1} = \frac{q_1 \cdots q_n}{p_1 \cdots p_n}[\alpha_0 - \alpha_1] = \frac{q_1 \cdots q_n}{p_1 \cdots p_n}[1 - \alpha_1]
\]

Therefore,

\[
\alpha_{n+1} = [\alpha_{n+1} - \alpha_0] + \alpha_0 \\
= \sum_{k=0}^{n} [\alpha_{k+1} - \alpha_k] + 1 \\
= [\alpha_1 - 1] \sum_{k=0}^{n} \frac{q_1 \cdots q_k}{p_1 \cdots p_k} + 1,
\]

where the \( k = 0 \) term in the sum is taken to be equal to 1. We can find a “well-behaved” solution to these equations if and only if the sum converges. Note, therefore, that we can already conclude that if the above sum does not converge, then the chain is recurrent. In the case that the sum does converge, we define

\[
Z \overset{\text{def}}{=} \sum_{k=0}^{\infty} \frac{q_1 \cdots q_k}{p_1 \cdots p_k} < \infty.
\]

Noting that we need \( \alpha_{n+1} \to 0 \), as \( n \to \infty \), to be able to conclude that the chain is transient, we set

\[
\alpha_1 = 1 - \frac{1}{Z} = \frac{Z - 1}{Z} = \left( \sum_{k=1}^{\infty} \frac{q_1 \cdots q_k}{p_1 \cdots p_k} \right) / \left( \sum_{k=0}^{\infty} \frac{q_1 \cdots q_k}{p_1 \cdots p_k} \right).
\]
so that
\[ \alpha_{n+1} = 1 - \frac{1}{Z} \sum_{k=0}^{n} \frac{q_1 \cdots q_k}{p_1 \cdots p_k} = \left( \frac{1}{\sum_{k=0}^{\infty} \frac{q_1 \cdots q_k}{p_1 \cdots p_k}} \right) \sum_{k=n+1}^{\infty} \frac{q_1 \cdots q_k}{p_1 \cdots p_k}, \rightarrow 0, \text{ as } n \rightarrow \infty. \]

Hence, when \( Z < \infty \) we can solve the relevant equations and we see \( Z < \infty \) is a necessary and sufficient condition for transience. We pull this fact out as a proposition.

**Proposition 3.2.6.** A birth and death chain is transient if and only if
\[ \sum_{k=1}^{\infty} \frac{q_1 \cdots q_k}{p_1 \cdots p_k} < \infty. \]

Note that in the above proposition we have pulled out the term \( k = 0 \), as this is always equal to one.

For example, consider the single server queue. Here \( p_i = \lambda \) and \( q_i = \mu \) for \( i \geq 1 \). In this case we have
\[ \sum_{k=1}^{\infty} \frac{q_1 \cdots q_k}{p_1 \cdots p_k} = \sum_{k=1}^{\infty} \left( \frac{\mu}{\lambda} \right)^k, \]
which converges if and only if \( \mu < \lambda \). This says that the single server queue is transient if and only if the arrival rate is strictly greater than the rate at which the server can work, and recurrent otherwise. Generalizing this result a bit, consider the multi-server queue with \( k \) servers. In this case, for \( n \geq k \) we have
\[ \frac{q_1 \cdots q_n}{p_1 \cdots p_n} = \frac{q_1 \cdots q_k q_{k+1} \cdots q_n}{p_1 \cdots p_n} = \frac{k!}{k^k} \left( \frac{k\mu}{\lambda} \right)^n. \]

Therefore, the sum converges, and the chain is transient if and only if the total possible serving rate \( k\mu \) is strictly less than the arrival rate \( \lambda \), which is intuitive.

We now turn our attention to trying to differentiate between positive recurrent and null recurrent birth and death chain. We solve this problem by solving for a stationary distribution and applying the results of Section 2.5.2. We again note that these results will also help us understand birth and death models in the continuous time setting.

Recall that for any discrete time Markov chain, the vector \( \pi \) is a stationary distribution if it is a probability vector and if \( \pi^T = \pi^T P \). For the birth-death model this equation is
\begin{align*}
\pi_k &= \pi_{k-1}p_{k-1} + \pi_k(1 - p_k - q_k) + \pi_{k+1}q_{k+1}, \quad \text{if } k \geq 1 \\
\pi_0 &= \pi_0(1 - p_0) + \pi_1q_1. \tag{3.5} \tag{3.6}
\end{align*}

Rearranging equation (3.6) shows
\[ q_1\pi_1 - p_0\pi_0 = 0, \]
while rearranging terms in equation (3.5) gives that for \( k \geq 1 \)
\[
q_{k+1} \pi_{k+1} - p_k \pi_k = q_k \pi_k - p_{k-1} \pi_{k-1}.
\]
Iterating the above equations shows that for all \( k \geq 0 \)
\[
q_{k+1} \pi_{k+1} - p_k \pi_k = q_1 \pi_1 - p_0 \pi_0 = 0.
\]
Therefore, for all \( k \geq 0 \)
\[
\pi_{k+1} = \frac{p_k}{q_{k+1}} \pi_k,
\]
and we conclude that for \( k \geq 1 \)
\[
\pi_k = \frac{p_k - 1}{q_k} \pi_{k-1} = \cdots = \frac{p_0 p_1 \cdots p_{k-1}}{q_1 q_2 \cdots q_k} \pi_0.
\]
We can make this a probability vector if and only if
\[
\sum_{k=1}^{\infty} \frac{p_0 p_1 \cdots p_{k-1}}{q_1 q_2 \cdots q_k} < \infty. \tag{3.7}
\]
In this case, we set
\[
W \overset{\text{def}}{=} \sum_{k=0}^{\infty} \frac{p_0 p_1 \cdots p_{k-1}}{q_1 q_2 \cdots q_k},
\]
where the \( k = 0 \) term in the sum is taken to be equal to one, and note that
\[
\sum_{k=0}^{\infty} \pi_k = \pi_0 W = 1.
\]
Therefore, we see that \( \pi_0 = 1/W \) and
\[
\pi_k = \frac{p_0 p_1 \cdots p_{k-1}}{q_1 q_2 \cdots q_k} W^{-1}, \tag{3.8}
\]
for \( k \geq 1 \). We collect thoughts with the following proposition.

**Proposition 3.2.7.** A birth and death chain is positive recurrent if and only if
\[
\sum_{k=1}^{\infty} \frac{p_0 p_1 \cdots p_{k-1}}{q_1 q_2 \cdots q_k} < \infty.
\]
In this case, \( \pi_0 = \left( \sum_{k=0}^{\infty} \frac{p_0 p_1 \cdots p_{k-1}}{q_1 q_2 \cdots q_k} \right)^{-1} \), and \( \pi_k \) satisfies (3.8).

For example, consider again the queuing network with one server. In this case
\[
\sum_{k=0}^{\infty} \frac{p_0 p_1 \cdots p_{k-1}}{q_1 q_2 \cdots q_k} = \sum_{k=0}^{\infty} \left( \frac{\lambda}{\mu} \right)^k = \frac{1}{1 - (\lambda/\mu)},
\]
where the final equation holds provided that $\lambda < \mu$. The equilibrium distribution in this case of $\lambda < \mu$ is now given as

$$\pi_k = \left(1 - \frac{\lambda}{\mu}\right) \left(\frac{\lambda}{\mu}\right)^k.$$  \hfill (3.9)

We have also shown that the case $\lambda = \mu$ is null recurrent, as expected.

We still have not solved for the stationary distribution when the state space is finite. Therefore, now suppose that $S = \{0, 1, \ldots, N\}$. By the same logic as above, we still have that (3.5) and (3.6) hold for $k \in \{0, 1, \ldots, N - 1\}$. The equation for $\pi_N$ is

$$\pi_N = \pi_N(1 - q_N) + \pi_{N-1}p_{N-1},$$

implying

$$q_N\pi_N = p_{N-1}\pi_{N-1} = p_{N-1}\frac{p_0 \cdots p_{N-2}}{q_1 q_2 \cdots q_{N-1}} \pi_0$$

and so

$$\pi_N = \frac{p_0 \cdots p_{N-1}}{q_1 q_2 \cdots q_N} \pi_0,$$

satisfying the same general form as $\pi_k$.

Let us return to Example 3.2.3 pertaining to the linearly reversible system $A \rightleftharpoons B$.

The transition rates for $X_n$, giving the number of $A$ molecules at time $n$ are given in (3.2). We use these rates to compute the first few terms of $\pi$ to see if a pattern emerges:

$$\pi_1 = \frac{c_2 N}{c_1} \pi_0,$$

$$\pi_2 = \frac{c_2 N c_2(N - 1)}{c_1 \times 2c_1} \pi_0 = \frac{N}{2} \left(\frac{c_2}{c_1}\right)^2 \pi_0,$$

$$\pi_3 = \frac{c_2 N c_2(N - 1)c_2(N - 2)}{c_1 \times 2c_1 \times 3c_1} \pi_0 = \frac{N}{3} \left(\frac{c_2}{c_1}\right)^3 \pi_0,$$

$$\vdots$$

$$\pi_k = \frac{c_2 N c_2(N - 1)c_2(N - 2) \cdots c_2(N - (k - 1))}{c_1 \times 2c_1 \times 3c_1 \cdots k c_1} \pi_0 = \frac{N}{k} \left(\frac{c_2}{c_1}\right)^k \pi_0.$$

Also note that

$$1 = \sum_{k=0}^{N} \pi_k = \sum_{k=0}^{N} \binom{N}{k} \left(\frac{c_2}{c_1}\right)^k \pi_0 = \pi_0 \left(1 + \frac{c_2}{c_1}\right)^N,$$
which implies
\[ \pi_0 = \left( \frac{c_1}{c_1 + c_2} \right)^N. \]

Therefore,
\[ \pi_k = \binom{N}{k} \left( \frac{c_2}{c_1} \right)^k \left( \frac{c_1}{c_1 + c_2} \right)^N = \binom{N}{k} \left( \frac{c_2}{c_1 + c_2} \right)^k \left( \frac{c_1}{c_1 + c_2} \right)^{N-k}. \]  (3.10)

So, the stationary distribution is binomial with parameters \( N \) and \( c_2/(c_1 + c_2) \). Note that, for example, this implies that if \( c_2 \gg c_1 \) (i.e. the rate of \( A \) production is much higher than the rate of \( B \) production), then most molecules will be \( A \) after a very long time.

Also note that if we let \( Z_n \in \{A, B\} \) be the state of a particular molecule at time \( n \), then the transition matrix for \( Z_n \) is

\[ P = \begin{bmatrix} 1 - c_1 & c_1 \\ c_2 & 1 - c_2 \end{bmatrix}, \]

and the stationary distribution, denoted \( \pi_Z \), is

\[ \pi_Z = \left[ \frac{c_2}{c_1 + c_2}, \frac{c_1}{c_1 + c_2} \right]. \]

Therefore, equation (3.10) can be understood by assuming all of the molecules are behaving independently, with the number of molecules being in state \( A \) simply being a binomial random variable with parameters \( N \) and \( p \) determined by the first component of \( \pi_Z \) above.

### Expected time until extinction

Consider a birth and death process such that 0 is a recurrent absorbing state \( (p_0 = 0) \). In this situation, we know that the population will die out at some point (with probability one). A very reasonable question one can then ask is, how long do we expect to wait before the population dies out?

Thus far, first step analysis has been useful, so we continue with it in the obvious manner. We let \( \tau_k \) denote the expected amount of time before the process hits zero, conditioned on an initial population of size \( k \). We clearly have that \( \tau_0 = 0 \) and for \( k \geq 1 \)

\[ \tau_k = p_k(1 + \tau_{k+1}) + q_k(1 + \tau_{k-1}) + (1 - p_k - q_k)(1 + \tau_k). \]

Therefore, after some rearranging, for \( k \geq 1 \)

\[ \tau_{k+1} = \tau_k + \frac{q_k}{p_k} \left( \tau_k - \tau_{k-1} - \frac{1}{q_k} \right). \]

For example, because \( \tau_0 = 0 \),

\[ \tau_2 = \tau_1 + \frac{q_1}{p_1} \left( \tau_1 - \frac{1}{q_1} \right). \]
Similar to the analyses for transience and positive recurrence, we search for a usable pattern. For \( k = 2 \) in the above we have

\[
\tau_3 = \tau_2 + \frac{q_2}{p_2} (\tau_2 - \tau_1 - \frac{1}{q_2}) \\
= \tau_2 + \frac{q_2}{p_2} \left( \frac{q_1}{p_1} \left( \frac{1}{q_1} - \frac{1}{q_2} \right) \right) \\
= \tau_1 + \frac{q_1}{p_1} \left( \frac{1}{q_1} \right) + \frac{q_1 q_2}{p_1 p_2} \left( \tau_1 - \frac{1}{q_1} - \frac{p_1}{q_1 q_2} \right),
\]

where we have used our result for \( \tau_2 \) multiple times. Let us think about the next step.

\[
\tau_4 = \tau_3 + \frac{q_3}{p_3} (\tau_3 - \tau_2 - \frac{1}{q_3}).
\]

We see that the first term, \( \tau_3 \), was just calculated. We begin to see that a summation is forming for the general term. The new term in the sum comes from

\[
\frac{q_3}{p_3} (\tau_3 - \tau_2 - \frac{1}{q_3}) = \frac{q_3}{p_3} \left( \frac{q_1 q_2}{p_1 p_2} \left( \tau_1 - \frac{1}{q_1} - \frac{p_1}{q_1 q_2} \right) - \frac{1}{q_3} \right) \\
= \frac{q_1 q_2 q_3}{p_1 p_2 p_3} \left( \tau_1 - \frac{1}{q_1} - \frac{p_1}{q_1 q_2} - \frac{p_1 p_2}{q_1 q_2 q_3} \right),
\]

and the pattern is emerging. In general, we have for \( m \geq 1 \)

\[
\tau_m = \tau_1 + \sum_{k=1}^{m-1} \frac{q_1 \cdots q_k}{p_1 \cdots p_k} \left[ \tau_1 - \frac{1}{q_1} - \sum_{i=2}^{k} \frac{p_1 \cdots p_{i-1}}{q_1 \cdots q_i} \right], \quad (3.11)
\]

where we interpret the second summation as zero when \( k < 2 \). Therefore, if we can determine \( \tau_1 \), we are done.

To determine \( \tau_1 \), the expected amount of time needed to hit state zero conditioned upon an initial population of one, we change our model slightly. We suppose that instead of being an absorbing state, our new model has \( p_0 = 1 \). That is, if the population enters state zero at any time, then with probability one it leaves and goes to state 1 during the next time interval. Now we denote by \( T_0 \) the first return time of this new, modified chain and note that \( E T_0 = \tau_1 + 1 \), where \( \tau_1 \) is what we want. However, computing \( E T_0 \) is easy because we know that it is equal to \( 1/\bar{\pi}_0 \), where \( \bar{\pi}_0 \) is the stationary distribution of the modified chain. We pause here and point out that if the original chain is null recurrent, then so is the modified chain, however the logic above still holds. This proves that \( \tau_1 = \infty \), and hence, \( \tau_m = \infty \) for all \( m \geq 1 \), which agrees with our analysis of the gambler’s ruin problem.

Now, in the positive recurrent case, we use our machinery from before and find that

\[
\bar{\pi}_0 = \frac{1}{1 + \sum_{k=1}^{\infty} \frac{p_0 p_1 \cdots p_{k-1}}{q_1 q_2 \cdots q_k}}.
\]
Therefore,
\[ \tau_1 = \mathbb{E}T_0 - 1 = \pi_0^{-1} - 1 \]
\[ = \sum_{k=1}^{\infty} \frac{p_0 p_1 \cdots p_{k-1}}{q_1 q_2 \cdots q_k} = \frac{1}{q_1} + \sum_{k=2}^{\infty} \frac{p_1 \cdots p_{k-1}}{q_1 q_2 \cdots q_k}. \]  
(3.12)

Plugging (3.12) into (3.11) yields
\[ \tau_m = \tau_1 + \sum_{k=1}^{m-1} \left[ \frac{q_1 \cdots q_k}{p_1 \cdots p_k} \sum_{i=k+1}^{\infty} \frac{p_1 \cdots p_{i-1}}{q_1 \cdots q_i} \right]. \]

This formula is legitimately scary, however such is life. In the finite state space case, \( S = \{0, 1, 2, \ldots, N\} \), the only things that change in the formula for \( \tau_1 \) and \( \tau_m \) is that the infinities become \( N \)'s.

**Example 3.2.8.** Consider the model of the single server queue in Example 3.2.5. Here \( q_i = \mu \) and \( p_i = \lambda \). Assuming \( \lambda < \mu \), so that the chain is positive recurrent, we have
\[ \tau_1 = \frac{1}{\mu} + \sum_{k=2}^{\infty} \frac{\lambda^{k-1}}{\mu^k} = \frac{1}{\lambda} \sum_{k=1}^{\infty} \frac{\lambda^k}{\mu^k} = \frac{1}{\lambda} \left( \sum_{k=0}^{\infty} \frac{\lambda^k}{\mu^k} - 1 \right) \]
\[ = \frac{1}{\lambda} \left( \frac{1}{1 - (\lambda/\mu)} - 1 \right) \]
\[ = \frac{1}{\mu - \lambda}. \]

Note that as \( \lambda \to \mu \), the expected time needed to hit state zero conditioned on the fact that the chain starts at state 1 tends to infinity! \( \square \)
3.3 Branching Processes

Branching processes are a subclass of discrete time Markov chains. However, the methods used in their study are typically quite different than what we have seen so far.

Historically, these models are traced back to Francis Galton and Henry William Watson. In 1873, Galton posed the following problem: suppose that $N$ adult males in a population each have different surnames. Suppose in each generation, $a_0$ percent of the adult males have no male children who survive to adulthood; $a_1$ have one such child; $a_2$ have two, and so on up to $a_5$, who have five. Then, what proportion of the surnames become extinct after $r$ generations? Watson helped solve the problem by using probability generating functions (which we’ll see in a bit). It was not until the 1930’s, however, that complete solutions to Galton’s problems were completed. It is for the above reason that these models are commonly referred to as Galton-Watson processes.

3.3.1 Terminology and notation

We denote the size of a certain population at time $n$ by $X_n$, where $n \in \{0, 1, 2, \ldots \}$. So our state space is $\{0, 1, 2, \ldots \}$. We make the following assumptions on the dynamics of the population:

1. Each individual produces a random number of offspring, each independently.
2. The distribution of the number of offspring is the same for each individual.
3. After reproducing, each individual dies.

We will denote by $Y$ (or $Y_i$ if we want to specify an individual) the random variable determining the number of offspring for an individual and suppose

$$P\{Y = n\} = p_n,$$

for $n \in \{0, 1, 2, 3, \ldots \}$ subject to the following constraints (to avoid trivialities)

$$p_0 > 0, \quad p_0 + p_1 < 1.$$

Therefore,

$$\sum_{n=0}^{\infty} p_n = 1, \quad p_n = P\{\text{individual produces exactly } n \text{ offspring}\}.$$ 

Unfortunately, it is not easy to write down transition probabilities for this model. Note that if $X_n = k$, then $k$ individuals produces offspring for the $(k+1)$st generation. If $Y_1, Y_2, \ldots, Y_k$ are independent random variables each with distribution

$$P\{Y_i = j\} = p_j,$$
then
\[ p_{k,j} = P\{X_{n+1} = j \mid X_n = k\} = P\{Y_1 + Y_2 + \cdots + Y_k = j\}. \] (3.13)

Note the important fact that once the process dies out, it never returns:
\[ p_{0,j} = \begin{cases} 1 & \text{if } j = 0 \\ 0 & \text{else} \end{cases}. \]

Equation (3.13) can be solved for using convolutions, but we do not choose that route. Instead, we begin by simply computing how the mean of the process evolves in time.

### 3.3.2 Behavior of the mean

Let \( \mu \) denote the mean number of offspring produced per individual. That is,
\[ \mu = \mathbb{E}(Y) = \sum_{k=0}^{\infty} kp_k < \infty, \]
where the last inequality is an assumption in our model. Then,
\[
\mathbb{E}(X_{n+1} \mid X_n = k) = \sum_{j=0}^{\infty} j P\{X_{n+1} = j \mid X_n = k\}
= \sum_{j=0}^{\infty} j P\{Y_1 + Y_2 + \cdots + Y_k = j\}
= \mathbb{E}(Y_1 + Y_2 + \cdots + Y_k)
= k \mu.
\]

Further, we also have that
\[
\mathbb{E}(X_{n+1} \mid X_0) = \sum_{j=0}^{\infty} j P\{X_{n+1} = j \mid X_0\} = \sum_{j=0}^{\infty} j \sum_{k=0}^{\infty} P\{X_{n+1} = j, X_n = k \mid X_0\}
= \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} P\{X_n = k \mid X_0\} P\{X_{n+1} = j \mid X_n = k\}
= \sum_{k=0}^{\infty} P\{X_n = k \mid X_0\} \sum_{j=0}^{\infty} j P\{X_{n+1} = j \mid X_n = k\}
= \sum_{k=0}^{\infty} P\{X_n = k \mid X_0\} \mathbb{E}(X_{n+1} \mid X_n = k)
= \sum_{k=0}^{\infty} P\{X_n = k \mid X_0\} k \mu
= \mu \mathbb{E}(X_n \mid X_0).
\]
Therefore,

\[ \mathbb{E}(X_n \mid X_0) = \mu \mathbb{E}(X_{n-1} \mid X_0) = \mu^2 \mathbb{E}(X_{n-2} \mid X_0) = \cdots = \mu^n \mathbb{E}(X_0 \mid X_0) = \mu^n X_0. \]

Taking expectations yields

\[ \mathbb{E}X_n = \mu^n \mathbb{E}X_0. \]

We can already draw some interesting conclusions from this calculation. Clearly, if \( \mu < 1 \), then the mean number of offspring goes to zero as \( n \) gets large (so long as \( \mathbb{E}X_0 < \infty \), of course). We have the crude estimate

\[ \mathbb{E}(X_n) = \sum_{j=0}^{\infty} j P\{X_n = j\} \geq \sum_{j=1}^{\infty} P\{X_n = j\} = P\{X_n \geq 1\}. \]

Therefore, if \( \mu < 1 \), we know that

\[ P\{X_n \geq 1\} \leq \mathbb{E}(X_n) = \mu^n \mathbb{E}X_0 \to 0, \]

as \( n \to \infty \), which is the same as saying

\[ \lim_{n \to \infty} P\{X_n = 0\} = 1. \quad (3.14) \]

Equation (3.14) says that the probability of the process having died out by time \( n \) goes to 1, as \( n \to \infty \). However, this is not the same as saying “with probability one, the process eventually dies out,” which we would write mathematically as

\[ P\left\{ \lim_{n \to \infty} X_n = 0 \right\} = 1. \quad (3.15) \]

We will prove equation (3.15) also holds. Let \( E_n \) be the event that the process has died out by time \( n \), then \( E_n \subset E_{n+1} \), and \( \{E_n\} \) is an increasing sequence of sets. Further,

\[ \left\{ \lim_{n \to \infty} X_n = 0 \right\} = \bigcup_{i=1}^{\infty} E_i = \lim_{n \to \infty} \bigcup_{i=1}^{n} E_i = \lim_{n \to \infty} E_n \]

is the event that the process dies out at some time, and so by the continuity of probability functions we have that

\[ P\left\{ \lim_{n \to \infty} X_n = 0 \right\} = P\left\{ \lim_{n \to \infty} E_n \right\} = \lim_{n \to \infty} P\{E_n\} = \lim_{n \to \infty} P\{X_n = 0\} = 1. \]

Therefore, having (3.14) allows us to conclude that, with probability equal to one, the population eventually dies out.

It is not entirely clear what happens when \( \mu = 1 \) or \( \mu > 1 \). In the case \( \mu = 1 \) we get the interesting case that \( \mathbb{E}(X_n) = \mathbb{E}(X_0) \) for all \( n \), so the expected size of the population stays constant. Also, when \( \mu > 1 \) the mean grows exponentially. However, it is possible for \( X_n \) to be zero with probability very near one, yet for \( \mathbb{E}(X_n) \) to not be small, so it is not clear what the probability of extinction will be in this case. Note that there is \textit{always} a non-zero probability that the population will go extinct in the next generation, regardless of how large the population is, so even the case \( \mu > 1 \) will have a probability of eventual extinction.
3.3.3 Probability of extinction

Recall that we are assuming that \( p_0 > 0 \) and \( p_0 + p_1 < 1 \). We now let

\[
a_n(k) = P\{X_n = 0 \mid X_0 = k\},
\]

and \( a(k) \) be the probability that the population eventually dies out given an initial population of \( k \)

\[
a(k) = \lim_{n \to \infty} a_n(k).
\]

Note that if there are originally \( k \) people in the population, then for the entire population to die out, each of the \( k \) branches must to die out. Because the branches are all acting independently we have

\[
a(k) = [a(1)]^k,
\]

and so it suffices to determine \( a(1) \). We will denote the extinction probability \( a(1) \) by \( a \).

As a first calculation we find that

\[
a = P\{\text{population dies out} \mid X_0 = 1\}
\]

\[
= \sum_{k=0}^{\infty} P\{X_1 = k \mid X_0 = 1\} P\{\text{population dies out} \mid X_1 = k\}
\]

\[
= \sum_{k=0}^{\infty} p_k a^k.
\]

We pause here to discuss the term on the right. We will also generalize a bit. For any random variable \( X \) taking values in \( \{0, 1, 2, \ldots\} \), the probability generating function of \( X \) is the function

\[
\phi(s) = \phi_X(s) = \mathbb{E}(s^X) = \sum_{k=0}^{\infty} s^k P\{X = k\}.
\]

A few properties are immediate.

1. Because each \( P\{X = k\} \geq 0 \) for each \( k \), \( \phi(s) \) is an increasing function of \( s \) for \( s \geq 0 \).

2. We have \( \phi(0) = P\{X = 0\} \) and \( \phi(1) = 1 \).

3. The domain of convergence for the infinite sum includes \( |s| \leq 1 \), thus the sum can be differentiated termwise. We find

\[
\phi'(s) = \sum_{k=1}^{\infty} k s^{k-1} P\{X = k\}
\]

\[
\phi''(s) = \sum_{k=2}^{\infty} k(k-1) s^{k-2} P\{X = k\}
\]
Therefore, so long as $P\{X \geq 1\} > 0$, we have
\[ \phi'(s) > 0, \]
and so long as $P\{X \geq 2\} > 0$ we have
\[ \phi''(s) > 0. \]

Note that in our branching process case, we have $p_0 > 0$ and $p_0 + p_1 < 1$, so both of these conditions are satisfied. Therefore, the function is strictly convex.

4. Possibly defining the derivative as a limit from the left, we see
\[ \phi'(1) = \sum_{k=1}^{\infty} kP\{X = k\} = \mathbb{E}(X). \]
\[ \phi''(1) = \sum_{k=2}^{\infty} k(k - 1)P\{X = k\} = \mathbb{E}(X^2) - \mathbb{E}(X). \]

5. Suppose that $X_1, X_2, \ldots, X_n$ are independent random variables taking values in \{0, 1, 2, \ldots\}, then
\[ \phi_{X_1+\ldots+X_n}(s) = \phi_{X_1}(s)\phi_{X_2}(s)\cdots\phi_{X_n}(s). \]

This follows from a simple computation
\[ \phi_{X_1+\ldots+X_n}(s) = \mathbb{E}[s^{X_1+\ldots+X_n}] \]
\[ = \mathbb{E}[s^{X_1}]\cdots\mathbb{E}[s^{X_n}] \]
\[ = \phi_{X_1}(s)\phi_{X_2}(s)\cdots\phi_{X_n}(s). \]

Returning to our study of branching processes. We see that the extinction probability $a$ satisfies
\[ a = \phi(a). \]

Obviously, $a = 1$ satisfies the above equation. The question is whether or not there are more such solutions. Also, we see that there can be at most two solutions to the equation on the interval $[0, 1]$. This follows from the convexity of $\phi(a)$. See Figure 3.3.1.

Recalling that we are taking $X_0 = 1$, we will now show that the generating function of $X_n$ is
\[ \phi(\phi(\cdots(\phi(s))\cdots)) = \phi \circ \phi \circ \cdots \circ \phi(s), \text{ (n compositions)}. \]

This clearly holds for $X_0$ as in this case $P\{X_0 = 1\} = 1$. We now show the general
Figure 3.3.1: The different ways that the generating function can intersect the line $y = s$ on the interval $[0, 1]$. There are two possibilities: (a) one intersection at $s = 1$ or (b) one intersection at $s < 1$ and one intersection at $s = 1$. In the case that there is an intersection at a point strictly before 1, the point of intersection is the extinction probability, $a$.

result by induction. We denote the generating function for $X_n$ by $\phi_n$. We have

$$
\phi_n(s) = \sum_{k=0}^{\infty} P\{X_n = k\} s^k
$$

$$
= \sum_{k=0}^{\infty} \left[ \sum_{j=0}^{\infty} P\{X_1 = j\} P\{X_n = k \mid X_1 = j\} \right] s^k
$$

$$
= \sum_{j=0}^{\infty} P\{X_1 = j\} \left[ \sum_{k=0}^{\infty} P\{X_{n-1} = k \mid X_0 = j\} s^k \right].
$$

The last term in brackets is the probability generating function of $X_{n-1}$ given an initial condition of $X_0 = j$. However, if $Y_{n}^i$, $i \in \{1, \ldots, j\}$ denotes the number of offspring of the $j$th individual at time $n$, then we see that the $Y_{n}^i$ are independent, and, conditioned on an initial population of $j$ individuals, $X_{n-1} = Y_{n-1}^1 + \cdots + Y_{n-1}^j$. We know (our inductive hypothesis) that the generating function for each $Y_{n-1}^i$ is

$$
\phi \circ \phi \circ \cdots \circ \phi(s), \quad (n-1 \text{ compositions}).
$$

We also know that by their independence, the generating function for $X_{n-1}$, which is exactly the last term in brackets, is the product of the generating functions of the
Therefore, writing $\phi^{(i)}(s)$ as $i$ convolutions of $\phi$, we have

$$\phi_n(s) = \sum_{j=0}^{\infty} P\{X_1 = j\} (\phi_{n-1}(s))^j$$

$$= \sum_{j=0}^{\infty} P\{X_1 = j\} (\phi^{(n-1)}(s))^j$$

$$= \phi(\phi^{(n-1)}(s))$$

$$= \phi^{(n)}(s).$$

Now we note that

$$a_n(1) = P\{X_n = 0 \mid X_0 = 1\} = \phi_n(0) = \phi^{(n)}(0).$$

Also,

$$a = \lim_{n \to \infty} a_n(1) = \phi(a),$$

and so $\lim_{n \to \infty} a_n(1)$ is a root of the equation $a = \phi(a)$. We now prove the following.

**Lemma 3.3.1.** The extinction probability of the branching process is the smallest positive root of the equation $s = \phi(s)$.

**Proof.** We already know that $a$ must satisfy the equation. We let $\bar{a}$ denote the smallest root of the equation. Now we will show that for every $n$, $a_n(1) = P\{X_n = 0 \mid X_0 = 1\} \leq \bar{a}$. This will imply that

$$a = \lim_{n \to \infty} a_n(1) \leq \bar{a}.$$

However, $\lim_{n \to \infty} a_n(1)$ is also a root of the equation $s = \phi(s)$ and so we are forced to conclude that $\lim_{n \to \infty} a_n(1) = \bar{a}$. To prove the desired inequality, we first note that it is trivially true for $n = 0$ because $a_0(1) = 0$. We now proceed by induction on $n$. Thus, assuming $a_{n-1} \leq \bar{a}$, we have

$$P\{X_n = 0\} = \phi^{(n)}(0) = \phi(\phi^{(n-1)}(0)) = \phi(a_{n-1}) \leq \phi(\bar{a}) = \bar{a},$$

where the inequality holds because $\phi$ is an increasing function. \qed

**Example 3.3.2.** Suppose that $p_0 = .3, p_1 = .6, p_2 = .05, p_3 = .05$. Then $\mu = .85$ and

$$\phi(a) = .3 + .6a + .05a^2 + .05a^3.$$

Solving the equation $a = \phi(a)$ yields $a = 1, 1.64575$. Thus, the extinction probability is 1. \qed

**Example 3.3.3.** Suppose that $p_0 = .2, p_1 = .2, p_2 = .3, p_3 = .3$. Then $\mu = 1.7$ and

$$\phi(a) = .2 + .2a + .3a^2 + .3a^3.$$

Solving the equation $a = \phi(a)$ yields $a = 1, .291$. Thus, the extinction probability is .291. \qed
Example 3.3.4. Suppose that \( p_0 = 1/4, p_1 = 1/2, p_2 = 1/4 \). Then \( \mu = 1 \) and
\[
\phi(a) = (1/4) + (1/2)a + (1/4)a^2.
\]
Solving the equation \( a = \phi(a) \) yields \( a = 1, 1 \). Thus, the extinction probability is 1.\( \square \)

We now establish the criteria for when \( a < 1 \).

Theorem 3.3.5. For a branching process with \( p_0 + p_1 < 1 \) and \( p_0 > 0 \), the extinction probability \( a \) satisfies \( a = 1 \) if and only if \( \mu \leq 1 \).

Proof. Note that we have already proven that if \( \mu < 1 \), then \( a = 1 \). We now suppose that \( \mu = 1 \). Then, we know that
\[
\phi'(1) = \mathbb{E}(Y) = \mu = 1.
\]
Therefore, by the concavity of \( \phi \), we must have that \( \phi'(s) < 1 \) for \( s < 1 \). Therefore, for any \( s < 1 \) we have
\[
1 - \phi(s) = \int_s^1 \phi'(t)dt < 1 - s.
\]
That is, \( \phi(s) > s \). Therefore, there can be no root less than one and we have shown that in this case the extinction probability is one.

Now we consider the case of \( \mu > 1 \). Then, by the same reasoning,
\[
\phi'(1) = \mu > 1.
\]
However, we also know that \( \phi(1) = 1 \). Therefore, there is an \( s < 1 \) with \( \phi(s) < s \). However, \( \phi(0) = p_0 > 0 \), and so by continuity we have that there must be some \( a \in (0, s) \) for which \( \phi(a) = a \). Because \( \phi''(s) > 0 \), the curve is convex and there is at most one such solution. By previous lemma, this root of \( a = \phi(a) \) is the extinction probability. \( \square \)

Note that the result that \( \mu = 1 \) implies \( a = 1 \) is quite interesting. For example we see that
\[
\mathbb{E}(X_n \mid X_n > 0) = \sum_{k=0}^{\infty} kP\{X_n = k \mid X_n > 0\}
\]
\[
= \sum_{k=0}^{\infty} kP\{X_n = k, X_n > 0\}/P\{X_n > 0\}
\]
\[
= \frac{1}{P\{X_n > 0\}} \sum_{k=0}^{\infty} k[P\{X_n = k\} - P\{X_n = k, X_n = 0\}]
\]
\[
= \frac{1}{P\{X_n > 0\}} \mathbb{E}[X_n]
\]
\[
= \frac{1}{P\{X_n > 0\}} \rightarrow \infty, \quad \text{as} \quad n \rightarrow \infty.
\]
Therefore, the expected value of the size of the population, conditioned on the population having survived to time \( n \), goes to infinity.
Definition 3.3.6. The mean of the offspring distribution, $\mu$, is known as the criticality parameter.

- If $\mu < 1$, extinction is definite, $a = 1$, and the process is called subcritical.
- If $\mu = 1$, extinction is definite (so long as $p_1 \neq 1$), $a = 1$, and the process is called critical.
- If $\mu > 1$, extinction is not definite, $a < 1$, and the process is called supercritical.

Note that we can also now give the solution to Galton’s original question, where there were $N$ males worried about the longevity of their surnames. After $n$ generations, the probability that a given surname has gone extinct is $a_n(1)$ and the probability that any $j$ surnames have gone extinct is

$$\binom{N}{j} a_n(1)^j (1 - a_n(1))^{N-j}.$$

Because $a_n(1) = \phi(n)(0)$, these values can be solved for iteratively. If $a < 1$ we see that as $n \to \infty$ the probability that exactly $j$ surnames eventually go extinct is

$$\binom{N}{j} \alpha^j (1 - \alpha)^{N-j},$$

that is, the probability distribution is binomial($N, \alpha$).

We now show that for every branching process the states $\{1, 2, \ldots\}$ are transient.

**Theorem 3.3.7.** For every branching process, the states $\{1, 2, \ldots\}$ are transient.

**Proof.** The cases $p_0 = 0$ or $p_0 = 1$ are clear. Therefore, supposing that $p_0 \in (0,1)$, we consider state $k \geq 1$. Because $p_0 > 0$ by assumption, we have that

$$p_{k,0} = p_0^k > 0,$$

by the independence of the $k$ individuals. Thus, $k \to 0$ for all $k \geq 0$, and since zero is absorbing, $k$ is transient. \(\square\)

We now compute the variance of a branching process under the assumption that $X_0 = 1$. The proof proceeds similarly to the computation of the mean. We let
\( \sigma^2 = \text{Var}(Y) \), and have

\[
\mathbb{E}(X_{n+1}^2) = \sum_{j=0}^{\infty} j^2 P\{X_{n+1} = j | X_0 = 1\}
\]

\[
= \sum_{j=0}^{\infty} j^2 \sum_{k=0}^{\infty} P\{X_{n+1} = j, X_n = k | X_0 = 1\}
\]

\[
= \sum_{j=0}^{\infty} P\{X_n = k | X_0 = 1\} \sum_{j=0}^{\infty} j^2 P\{X_{n+1} = j | X_n = k\}
\]

\[
= \sum_{j=0}^{\infty} P\{X_n = k | X_0 = 1\} \sum_{j=0}^{\infty} j^2 P\{Y_1^n + \cdots + Y_k^n = j\}
\]

\[
= \sum_{k=0}^{\infty} P\{X_n = k | X_0 = 1\} (\text{Var}(Y_1 + \cdots + Y_k) + (k\mu)^2)
\]

\[
= \sum_{k=0}^{\infty} P\{X_n = k | X_0 = 1\} (k\sigma^2 + (k\mu)^2)
\]

\[
= \sigma^2 \mathbb{E}(X_n) + \mu^2 \mathbb{E}(X_n^2).
\]

Doing the obvious things yields

\[
\text{Var}(X_{n+1}) = \sigma^2 \mu^n + \mu^2 \text{Var}(X_n).
\]

You can iterate this and eventually find (note we are shifting indices from \( n + 1 \) to \( n \))

\[
\text{Var}(X_n) = \begin{cases} 
\mu^{n-1} \left( \frac{\mu^n - 1}{\mu - 1} \right) \sigma^2, & \text{if } \mu \neq 1 \\
\frac{n \sigma^2}{\mu}, & \text{if } \mu = 1
\end{cases}
\]

We close with an example taken from Allen, 2003.

**Example 3.3.8.** Suppose that a certain population is very large. Suppose also that a mutant gene appears in \( N \) individuals of the population simultaneously. Both individuals with and without the mutant gene reproduce according to a branching process. We suppose that the mean number of individuals produced by those with the mutant gene is \( \mu \). Suppose that the mean number of individuals produced by those without the mutant gene is 1. Suppose that

\[
\mu = 1 + \epsilon,
\]

for a small \( \epsilon > 0 \). What is the probability that the mutant gene will become extinct?
Let $a$ denote the extinction probability. We note that $a \approx 1$. Why must this be? We know that $\phi'(1) = \mu = 1 + \epsilon$. Also, $\phi''(s)$ is bounded from below by $2p_2$ because

$$\phi''(s) = \sum_{k=2}^{\infty} p_k k(k-1)s^{k-2} = 2p_2 + \text{non-negative term.}$$

If $p_2 = 0$, just go out to $p_n$ where $p_n \neq 0$ and $n \geq 2$. In this case the lower bound is $n(n-1)p_n(1/2)^n - 2$, valid for $s \geq 1/2$.

We will now approximate the value of $a$. We change variables by defining $\theta$ via the equation $a = e^\theta$. Because $a \approx 1$, we know $\theta \approx 0$. Now we define $M(s) = \phi(e^s)$, and note that $M(\theta) = \phi(e^\theta) = \phi(a) = a$. Next define

$$K(s) = \ln M(s).$$

Note that

$$K(\theta) = \ln(M(\theta)) = \ln(a) = \ln(e^\theta) = \theta.$$  \hfill (3.16)

We will now find the first few terms in the Taylor expansion of $K$ around $\theta \approx 0$, and use these, in conjunction with (3.16), to solve, at least approximately, for $\theta$, and hence $a$.

It is easy to check that $K(0) = 0$, $K'(0) = \mu$, and $K''(0) = \sigma^2$. For example, we have

$$K(0) = \ln M(0) = \ln \phi(e^0) = \ln \phi(1) = \ln 1 = 0.$$

Next,

$$K'(s) = \frac{d}{ds} \ln M(s) = \frac{M'(s)}{M(s)},$$

$$= \frac{1}{M(s)} \phi'(e^s) e^s$$

$$\implies K'(0) = \frac{1}{M(0)} \phi'(1) = \mu.$$  \hfill (3.17)

Finally, using Equation (3.17), we have

$$K''(s) = -\frac{1}{M(s)^2} M'(s)^2 + \frac{M''(s)}{M(s)}.$$

Also,

$$M'(s) = \phi'(e^s) e^s$$

$$M''(s) = \phi''(e^s) e^{2s} + \phi'(e^s) e^s.$$
Thus, because \( \phi''(1) = \mathbb{E}[Y^2] - \mathbb{E}[Y] = \sigma^2 + \mu^2 - \mu, \)

\[
K''(0) = \frac{1}{1} \mu^2 + \phi''(1) + \phi'(1) \\
= -\mu^2 + (\sigma^2 + \mu^2 - \mu) + \mu \\
= \sigma^2.
\]

Therefore, expanding the left hand side of (3.16) using a Taylor series, we have

\[
\theta \approx K(0) + K'(0)\theta + \frac{1}{2} K''(0)\theta^2 = \mu \theta + \sigma^2 \frac{\theta^2}{2}.
\]

Using this equation to solve for \( \theta \) yields

\[
\theta \approx \frac{2}{\sigma^2} (1 - \mu) = -\frac{2}{\sigma^2} \epsilon.
\]

That is,

\[
a \approx e^{-2 \epsilon / \sigma^2}.
\]

Also, for an initial size of \( N \) mutants,

\[
a^N \approx e^{-2N \epsilon / \sigma^2}.
\]

For example, in the case that the birth probabilities are Poisson, \( m = \sigma^2 = 1 + \epsilon. \)

Taking \( \epsilon = .01 \), we have

\[
a^N \approx e^{-2N(.01)/1.01} \approx 0.98039^N.
\]

The probability that the gene becomes established is \( 1 - a^N \). See Table 3.1.

### 3.4 Exercises

1. Consider the genetics inbreeding problem with transition matrix (3.1). Compute the third row of \( P \) similarly to how we computed the second row. That is, formally find all of \( p_{3i} \) for \( i \in \{1, 2, \ldots, 6\} \).
2. Consider the single server queue. We found that when $\lambda < \mu$, the stationary distribution is given by (3.9). What is the expected length of the queue in equilibrium. What happens as $\lambda \to \mu$?

3. Consider a birth-death process with $q_i = 1/4$, for all $i \geq 1$ with $q_0 = 0$. Suppose

$$p_i = \frac{1}{4} \left( \frac{i + 1}{i + 2} \right)^2, \quad i \geq 0.$$ 

Note that $p_i < q_i$ for all $i$, and $p_i \to q_i$, as $i \to \infty$. Is this chain transient, positive recurrent, or null recurrent?

4. Consider a birth-death process with $q_i = 1/4$, for all $i \geq 1$ (with, as always $q_0 = 0$). Suppose that

$$p_i = \frac{1}{4} \left( \frac{i + 2}{i + 1} \right)^2, \quad i \geq 0.$$ 

Note that $p_i > q_i$ for all $i$, and $p_i \to q_i$, as $i \to \infty$. Is this chain transient, positive recurrent, or null recurrent?

5. Consider a birth-death process with $q_i = 1/4$, for all $i \geq 1$ with $q_0 = 0$. Suppose

$$p_i = \frac{1}{4} \left( \frac{i + 1}{i + 2} \right)^2, \quad i \geq 0.$$ 

Note that $p_i < q_i$ for all $i$, and $p_i \to q_i$, as $i \to \infty$. Is this chain transient, positive recurrent, or null recurrent?

6. Consider a birth-death process with $q_i = 1/4$, for all $i \geq 1$ with $q_0 = 0$. Suppose

$$p_i = \frac{1}{4} \left( \frac{i + 2}{i + 1} \right)^2, \quad i \geq 0.$$ 

Note that $p_i > q_i$ for all $i$, and $p_i \to q_i$, as $i \to \infty$. Is this chain transient, positive recurrent, or null recurrent?

7. Consider a birth and death process with $q_i = 1/4$ if $i \geq 1$, and $q_0 = 0$, and

$$p_i = \frac{1}{4} \left( \frac{i + 1}{i + 2} \right)^2, \quad i \geq 0.$$ 

Note that this is the same chain as in Problem 5 above. We will again use Theorem 2.5.22 to estimate the stationary distribution. Simulate this process, with $X_0 = 0$, and average over the path to estimate $\pi_i = \lim_{n \to \infty} P\{X_n = i\}$, for $i \in \{0, 1, 2, 3, 4, 5\}$. Note that this problem is similar to (and in some ways easier, even though the state space is infinite) that of 13 of Chapter 3.

8. (Lawler, 2006) Given a branching process with the following offspring distributions, determine the extinction probability $a$. 

90
9. Consider again the branching process with \( p_0 = .5, p_1 = .1, p_3 = .4 \), and suppose that \( X_0 = 1 \). What is the probability that the population is extinct in the second generation \( (X_2 = 0) \), given that it did not die out in the first generation \( (X_1 > 0) \)!

10. (Lawler, 2006) Consider a branching process with \( p_0 = 1/3, p_1 = 1/3, p_2 = 1/3 \). Find, with the aid of a computer, the probability that the population dies out in the first \( n \) steps for \( n = 20, 100, 200, 1000, 1500, 2000, 5000 \). Do the same with the values \( p_0 = .35, p_1 = .33, p_2 = .32 \), and then do it for \( p_0 = .32, p_1 = .33 \), and \( p_2 = .35 \).

11. Suppose a branching process has a Poisson offspring distribution

\[
p_k = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots
\]

(a) Supposing that \( X_0 = 1 \), find the mean and variance of \( X_n \), the size of the population at time \( n \).

(b) For \( \lambda = 1.5 \) and \( \lambda = 2 \), find the probability that the process eventually goes extinct.

12. We consider a model of replication for polymer chains. We consider a polymer chain consisting of \( m \) nucleotides that will itself be copied, and whose offspring chains will be copied, etc. We assume a fixed probability of \( p \) that a single nucleotide is correctly copied during replication. Thus, the probability that the entire chain is copied correctly is \( p^m \). We assume that the chain replicates at discrete times. We also suppose that during each time window, any polymer chain is destroyed with a probability of \( 1 - q \), and survives with a probability of \( q \). Therefore, a given polymer can yield zero (if it is destroyed), one (if it survives but is copied incorrectly), or two (survives and is replicated perfectly) exact replica chains in the next time step with respective probabilities

\[
p_0 = 1 - q, \quad p_1 = q(1 - p^m), \quad p_2 = qp^m.
\]

For a given \( p \) and \( q \), determine the threshold for \( m \), the size of the polymer, for which extinction of the exact chain is not assured. That is, give a condition on \( m \) for which the probability of extinction is less than one.

For a survival rate of \( q = .8 \), determine how long the chain has to be so that extinction is not guaranteed for the cases \( p = .2, .4, .5, .9, .99 \). Finally, using any formulas you have computed, what happens when \( q < 1/2 \)? Please interpret this result. Does it make sense?
Consider a branching process with

\[ p_0 = \frac{1}{2}, \quad p_1 = \frac{1}{4}, \quad \text{and} \quad p_2 = \frac{1}{4}. \]

For this chain, we have

\[ \mu = p_1 + 2p_2 = \frac{3}{4}, \]

and so extinction is guaranteed. Let \( X_0 = 1 \), and denote the time at which the chain goes extinct by \( T_{\text{extinct}} \). Write a Matlab script that simulates individual realizations of this branching process. Use your script to estimate \( \mathbb{E}[T_{\text{extinct}}] \) via Monte Carlo.
Chapter 4
Renewal and Point processes

Markov processes are only one class of stochastic process. In this chapter we will study a class of processes called point processes. This chapter will seem quite dry and disconnected from the previous material at first pass. However, it will play a critical role in the study of stochastic models of population processes and, in particular, biochemical reaction networks in later chapters. We start with a special class of point processes called renewal processes.

4.1 Renewal Processes

A renewal process is used to model occurrences of events happening at random times, where the times between the occurrences are modeled using independent and identically distributed random variables. Many of even the most complicated models have within them an embedded renewal process.

The formal model is as follows. We let $Y_n, n \geq 1$, be a sequence of independent and identically distributed random variables which take only non-negative values. We also let $Y_0$ be a non-negative random variable, independent from $Y_n, n \geq 1$, though not necessarily of the same distribution. The range of these random variables could be discrete, perhaps \{0, 1, 2, \ldots\}, or continuous, perhaps [0, $\infty$). The random variables $Y_n$ will be the inter-event times of the occurrences. We assume throughout that for all $n \geq 1$

$$P\{Y_n = 0\} < 1.$$  

Next, for all $n \geq 0$, we define the process $S_n$ as

$$S_n = \sum_{i=0}^{n} Y_i.$$  

For example,

$$S_0 = Y_0, \quad S_1 = Y_0 + Y_1, \quad \text{and} \quad S_2 = Y_0 + Y_1 + Y_2.$$  

Copyright © 2013 by David F. Anderson.
Note that the random variable $S_n$ gives the amount of “time” that passes until the $n$th occurrence (we think of $Y_0$ as the “zeroth” occurrence). The sequence $\{S_n, n \geq 0\}$ is called a renewal sequence. The times $S_n$ are called renewal times. The occurrences are usually called the renewals. Note that if we define $\mu \triangleq EY_i, i \geq 1$, then for $n \geq 1$,

$$E S_n = EY_0 + \cdots + EY_n = n\mu + EY_0.$$ 

The random variable $Y_0$ gives the time until the zeroth occurrence. If $P\{Y_0 > 0\} > 0$, then the process is called delayed. If, on the other hand, we have that $P\{Y_0 = 0\} = 1$, in which case $S_0 = Y_0 = 0$ with a probability of one, then the process is called pure.

Recalling that the indicator function, $1_A : \mathbb{R} \to \{0, 1\}$, satisfies

$$1_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases},$$

we define the counting process

$$N(t) = \sum_{n=0}^{\infty} 1_{[0,t]}(S_n), \quad (4.1)$$

which yields the number of renewals in the time interval $[0, t]$. Note that if the process is pure, then $Y_0 \equiv 0$ and $N(0) = 1$ with probability one. The counting function $N$ will be our main object of study in this section.

A few more definitions are in required. If $P\{Y_n < \infty\} = 1$ for $n \geq 1$, then the renewal process is called proper. However, if $P\{Y_n < \infty\} < 1$, then the process is called defective. Note that in the defective case there will be a final renewal. In this case, $N(t)$ remains bounded with a probability of one, though the bound is a random variable. In fact, the bound is a geometric random variable with parameter $P\{Y_n < \infty\}$.

Our focus will be on trying to understand the large time behavior of the process $N(t)$. We will derive both a law of large numbers, and a central limit theorem for $N(t)$. However, we point the interested reader to either [37] or [32] for a more complete treatment on renewal processes, including a study of the renewal function

$$U(t) \triangleq E N(t) = E \sum_{n=0}^{\infty} 1_{[0,t]}(S_n).$$

We begin by considering a few concrete examples.

**Example 4.1.1.** Consider the Poisson process with rate $\lambda > 0$. The usual model is to take $N(0) = 0$, and define the waiting times between events, $Y_i$, to be independent, exponential random variables with parameter $\lambda$. In this case, the process is delayed, as $Y_0 \sim Y_i, i \geq 1$, is also an exponential random variable with a parameter of $\lambda$. The renewal function, that is the expectation of $N(t)$, is

$$U(t) = EN(t) = E \text{ Poiss}(\lambda t) = \lambda t,$$

where Poiss($x$) refers to a Poisson random variable with a parameter of $x$.  \[ \Box \]
Example 4.1.2. Let \( \{X_n, n \geq 0\} \) be an irreducible, positive recurrent, discrete time Markov chain. For \( i \in S \), let

\[
S_0 = \inf\{n \geq 0 : X_n = i\},
\]

and for \( n \geq 1 \)

\[
S_n = \inf\{j > S_{n-1} : X_j = i\}.
\]

Then, \( \{S_n, n \geq 0\} \) is a sequence of renewal times [13]. Note that if \( X_0 = i \), then the process is pure, whereas if \( X_0 \neq i \) then the process is delayed. Letting

\[
Y_n = S_n - S_{n-1},
\]

for \( n \geq 1 \), we note that in the delayed case the distribution of \( Y_0 \) is different than the distribution of \( Y_n, n \geq 1 \). As in many examples, the distribution of the inter-event times, \( Y_n, n \geq 1 \), can be difficult to find. However, we know at least that

\[
\mathbb{E} Y_n = \frac{1}{\pi_i}, \quad \text{for } n \geq 0,
\]

where \( \pi \) is the stationary distribution.

If instead the Markov chain were irreducible and transient, then the associated renewal process defined above is defective, and there is a last time the process returns to state \( i \). \( \square \)

4.1.1 The behavior of \( N(t)/t \), as \( t \to \infty \)

Our goal in this section is to characterize the limiting behavior of \( N(t)/t \), as \( t \to \infty \), where we preserve the notation from the previous section. We begin with a law or large result.

Theorem 4.1.3. Let \( N(t) \) be the counting process (4.1) associated with the renewal sequence \( \{S_n, n \geq 0\} \). Letting \( \mu \overset{\text{def}}{=} \mathbb{E} Y_1 \),

\[
\lim_{t \to \infty} \frac{N(t)}{t} = \frac{1}{\mu},
\]

with a probability of one, where the right-hand side is interpreted as zero if \( \mu = \infty \).

Proof. Note that if the process is defective, then (i) \( \mu = \infty \), and (ii) \( N(t) \) is uniformly bounded in \( t \). Using (ii) we have that \( \lim_{t \to \infty} N(t)/t = 0 \) with probability one, and the result holds. Hence, we need only consider the case that the process is proper, in which case \( N(t) \to \infty \) as \( t \to \infty \) with a probability of one.

We begin by recalling the strong law of large numbers [13]. Suppose that \( \{Z_n\} \) are i.i.d. non-negative random variables with \( \mathbb{E} Z_1 \leq \infty \). Then

\[
\frac{1}{n}(Z_1 + \cdots + Z_n) \to \mathbb{E} Z_1, \quad \text{as } n \to \infty
\]

almost surely. Note that the above allows for \( \mathbb{E} Z_1 = \infty \).
Returning to our processes, the strong law of large numbers states that with a probability of one
\[ \lim_{n \to \infty} \frac{Y_1 + \cdots + Y_n}{n} = \mu. \]
Thus, we may conclude that with probability one,
\[ \frac{S_n}{n} = \frac{Y_0}{n} + \frac{1}{n} \sum_{i=1}^{n} Y_i \to \mu, \text{ as } n \to \infty. \]
Noting that \( N(t) \to \infty, \text{ as } t \to \infty, \) we can therefore conclude that
\[ \frac{1}{N(t)} S_{N(t)} \to \mu, \text{ as } t \to \infty, \tag{4.2} \]
almost surely. By construction, for \( N(t) \geq 1 \) we have that
\[ S_{N(t)-1} \leq t < S_{N(t)}. \]
Therefore, so long as \( N(t) \geq 2, \)
\[ \frac{S_{N(t)-1}}{N(t)} \leq \frac{t}{N(t)} < \frac{S_{N(t)}}{N(t)} \implies \frac{S_{N(t)-1}}{N(t) - 1} \times \frac{N(t) - 1}{N(t)} \leq \frac{t}{N(t)} < \frac{S_{N(t)}}{N(t)}. \]
Applying (4.2), we see that as \( t \to \infty \)
\[ \frac{t}{N(t)} \to \mu \implies \frac{N(t)}{t} \to \frac{1}{\mu}. \]

Note that the above result is intuitively pleasing as it says that the shorter the wait between events, as characterized by \( \mu, \) the more events you expect to see in a given time interval. Note that it also says that
\[ N(t) = \frac{t}{\mu} + o(t), \text{ as } t \to \infty. \tag{4.3} \]
We now ask the question, how good of an approximation is (4.3), and, more precisely, what is the next term in the expansion (which is currently incorporated in the \( o(t) \) term). This is a central limit theorem type question.

We begin by defining
\[ \sigma^2 \overset{\text{def}}{=} \text{Var}(Y_i), \]
for \( i \geq 1, \) and assume that \( \sigma^2 < \infty. \)

Lemma 4.1.4. As \( t \to \infty, \) the distribution of
\[ \frac{N(t) - \mu^{-1}t}{\sigma \mu^{-3/2} \sqrt{t}}, \]
approaches that of a standard normal. That is,
\[ \lim_{t \to \infty} \text{P} \left\{ \frac{N(t) - t\mu^{-1}}{\sqrt{t} \sigma \mu^{-3/2}} \leq x \right\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt. \tag{4.4} \]
Note that the above lemma loosely states that
\[ \frac{N(t) - \mu^{-1}t}{\sigma \mu^{-3/2} \sqrt{t}} \approx B, \] (4.5)
where \( B \) is a Gaussian random variable with mean zero and variance one. Rearranging terms in (4.5) yields,
\[ \frac{N(t)}{t} \approx \frac{1}{\mu} + \frac{\sigma}{\mu^{3/2}} B, \]
showing that we are indeed uncovering the next term in the expansion of \( N(t)/t \).

**Proof of Lemma 4.1.4.** We begin by stating the usual central limit theorem (CLT) for the \( S_n \) process,
\[ \lim_{n \to \infty} P \left\{ \frac{S_n - n\mu}{\sigma \sqrt{n}} \leq x \right\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt. \] (4.6)
which we will need in the derivation of the CLT for the process \( N \).

Instead of just proving that (4.4) holds, which would probably not give much intuition, we will show why the process
\[ \frac{N(t) - t\mu^{-1}}{\sqrt{t\sigma^2 \mu^{-3}}} \]
is properly scaled to give the desired convergence. That is, we will suppose that we believe a CLT should hold for the process \( N(t) \) in that
\[ Z(t) \overset{d}{=} \frac{N(t) - h(t)}{g(t)} \overset{d}{\to} \eta, \quad \text{as } t \to \infty, \]
where \( \eta \) is a standard Gaussian and the convergence is in the sense of distributions, but that we do not know either \( h(t) \) or \( g(t) \).

The numerator is, of course, straightforward. In order for \( Z(t) \) to converge to a mean zero process, we see that Theorem 4.1.3 implies that we should choose \( h(t) = t\mu^{-1} \). Hence, we turn our attention to the scaling factor \( g(t) \) in the denominator.

Noting that we have the CLT for the process \( S_n \), we would like to convert statements about \( N \) to equivalent statements pertaining to \( S_n \). Thus, we note that for any \( n \geq 0 \) we have
\[ \{N(t) \leq n\} = \{S_n > t\}, \]
both of which state that the number of jumps by time \( t \) is less than or equal to \( n \). Therefore, denoting \( \lceil y \rceil \) to be the greatest integer less than or equal to \( y \in \mathbb{R} \), we have
\[ P \left\{ \frac{N(t) - t\mu^{-1}}{g(t)} \leq x \right\} = P \left\{ N(t) \leq \lceil t\mu^{-1} + g(t)x \rceil \right\} \]
\[ = P \left\{ S_{\lceil t\mu^{-1} + g(t)x \rceil} > t \right\} \]
\[ = P \left\{ \frac{S_{\lceil t\mu^{-1} + g(t)x \rceil} - \mu\lceil t\mu^{-1} + g(t)x \rceil}{\sigma \sqrt{\lceil t\mu^{-1} + g(t)x \rceil}} > \frac{t - \mu\lceil t\mu^{-1} + g(t)x \rceil}{\sigma \sqrt{\lceil t\mu^{-1} + g(t)x \rceil}} \right\}, \] (4.7)
where in the final equality we simply put the statement into the framework of \((S_n - \mu n)/(\sigma \sqrt{n})\) of the CLT for \(S_n\) found in (4.6). Because of (4.6) and (4.7), we see that we may conclude

\[
P \left\{ \frac{N(t) - t\mu^{-1}}{g(t)} \leq x \right\} \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt, \quad \text{as } t \to \infty,
\]
so long as two conditions are met:

(i) \([t\mu^{-1} + g(t)x] \to \infty \text{ as } t \to \infty\), and

(ii)

\[
\frac{t - \mu[t\mu^{-1} + g(t)x]}{\sigma \sqrt{t\mu^{-1} + g(t)x}} \to -x,
\]

as \(t \to \infty\).

Considering condition (ii), and dropping the floor function \(\lfloor \cdot \rfloor\) as the terms will be large, we see that we require

\[
\frac{-\mu g(t)x}{\sigma \sqrt{t\mu^{-1} + g(t)x}} \to -x,
\]

which tells us that \(g(t)\) must scale like \(O(\sqrt{t})\). Therefore, writing \(g(t) \equiv \sqrt{t}K\), we see that we must have

\[
\frac{\mu \sqrt{t}Kx}{\sigma \sqrt{t\mu^{-1} + \sqrt{t}Kx}} = \frac{-\mu Kx}{\sigma \sqrt{\mu^{-1} + t^{-1/2}Kx}} \to -x.
\]

As \(t^{-1/2}Kx \to 0\) as \(t \to \infty\), we arrive simply at the requirement

\[
\frac{-\mu Kx}{\sigma u^{-3/2}} = -x \quad \Rightarrow \quad K = \sigma u^{-3/2}.
\]

Therefore condition (ii) is satisfied if

\[
g(t) = \sqrt{t} \sigma^2 \mu^{-3}.
\]

Note that in this case, condition (i) is also satisfied and the result is proved.

Another result from renewal theory, and one which should not be surprising at this point, is called the elementary renewal theorem, and is stated below without proof.

**Theorem 4.1.5.** Let \(\mu = \mathbb{E}Y_i \leq \infty\), \(i \geq 1\). Then,

\[
\lim_{t \to \infty} t^{-1} \mathbb{E}N(t) = \frac{1}{\mu}.
\]
4.1.2 The renewal reward process

A renewal reward process is only a slight generalization of the standard renewal process. We now suppose that at each renewal time, $S_n$, we are given a random reward, $R_n$, where $R_n$ can be positive, negative, or zero. We assume that the random variables \( \{R_n, n \geq 0\} \) are independent, and that the sequence \( \{R_n, n \geq 1\} \) are also identically distributed. However, we do not assume that the $R_n$ are necessarily independent of $S_n$. For example, they could be functions of the inter-event times $Y_n$. We then define the renewal reward process to be

\[
R(t) = \sum_{i=0}^{N(t)-1} R_i = \sum_{i=0}^{\infty} R_i \mathbf{1}_{[0,t]}(S_i).
\]

That is, $R(t)$ gives the total accumulated reward up to time $t$.

**Example 4.1.6.** Consider an insurance company with claims coming in at times $S_n$. The sizes of the claims are $R_n$ and the total amount of claims against the company at time $t$ is $R(t)$. \(\square\)

**Example 4.1.7.** Consider a metabolic network that requires the amino acid methionine to run properly. Methionine is ingested when the organism eats food. Letting $S_n$ denote the times the organism eats, and $R_n$ the amount of methionine ingested, $R(t)$ is the total amount of methionine ingested by time $t$. \(\square\)

The following theorem should not be surprising at this point.

**Theorem 4.1.8.** If $\mathbb{E}|R_j| < \infty$ and $\mathbb{E}Y_j = \mu < \infty$, for $j \geq 1$, then

\[
\lim_{t \to \infty} \frac{R(t)}{t} = \frac{\mathbb{E}R_1}{\mu}.
\]

**Proof.** We have

\[
\lim_{t \to \infty} \frac{R(t)}{t} = \lim_{t \to \infty} \frac{\sum_{i=0}^{N(t)-1} R_i}{t} = \lim_{t \to \infty} \frac{\sum_{i=0}^{N(t)-1} R_i}{N(t)} \cdot \frac{N(t)}{t}.
\]

From Theorem 4.1.3 we know that

\[
\lim_{t \to \infty} \frac{N(t)}{t} = \frac{1}{\mu}.
\]

From the usual law of large numbers applied to the random variables $R_i$ we have

\[
\lim_{t \to \infty} \frac{1}{N(t)} \sum_{i=0}^{N(t)-1} R_i = \mathbb{E}R_1.
\]
Hence,
\[
\lim_{t \to \infty} \frac{R(t)}{t} = \lim_{t \to \infty} \frac{\sum_{i=0}^{N(t)-1} R_i}{N(t)} \cdot \frac{N(t)}{t} = [\mathbb{E}R_1] \frac{1}{\mu},
\]
and the result is shown.

\[\square\]

**Example 4.1.9.** This example is a generalization of an industrial example from [37]. Consider a species, termed \(Sp_1\), that must occasionally fight another species, which we will call \(Sp_2\). We suppose that \(Sp_1\) only comes within range of \(Sp_2\) sporadically, and that the number of days between the events of \(Sp_1\) and \(Sp_2\) being close is random, with an average of \(d\). We also suppose that \(Sp_1\) can expend energy each day in an effort to avoid detection in the event that an \(Sp_2\) is lurking nearby. Specifically, by expending an energy amount per day that is equivalent to a caloric intake of \(e\) calories, the probability of actually being caught by \(Sp_2\) if it is nearby is \(p(e)\). That is, the probability is a function of \(e\). Finally, suppose that if \(Sp_1\) is seen, then it will have to fight and expend an amount of energy which is random, but has a mean which is equivalent to requiring a caloric intake of \(E\). The question is, what is the amount of energy, quantified by \(e\), which minimizes the long term total energy expenditure of species \(Sp_1\)?

To begin answering our question, we start building a renewal process. Let \(Y_i\) denote the number of days between the times the species \(Sp_1\) is within range of \(Sp_2\). We know only that \(\mathbb{E}Y_i = d\). Next, we let \(N\) be be the number of times that \(Sp_2\) is within range of \(Sp_1\) before actually seeing, and fighting, it. Then, \(N\) is a geometric random variable with range \(\{1, 2, \ldots\}\) and for \(k \in \{1, 2, \ldots\}\) we have
\[
P\{N = k\} = q(e)^{k-1}p(e),\]
where \(q(e) \overset{\text{def}}{=} 1 - p(e)\). Thus, \(\mathbb{E}N = 1/p(e)\). Therefore, if \(S_n\) are the days in which \(Sp_2\) catches \(Sp_1\), then we see
\[
S_1 = \sum_{i=1}^{N} Y_i,
\]
and \(\mathbb{E}S_1 = \mathbb{E}Y_i\mathbb{E}N = d/p(e)\). Note that this implies that \(\mathbb{E}[S_n - S_{n-1}] = d/p(e)\) for all \(n \geq 1\). Let \(C_i\) be the cost, in terms of energy used, of being caught the \(i\)th time. We have that \(\mathbb{E}C_i = E\), \(i \geq 1\).

Now let \(N(t)\) count the number of times the prey has been caught by time \(t\). Note that it is the times \(S_n\), constructed above, that \(N(t)\) is counting and not the process with inter-event times \(Y_i\). The total cost is then
\[
C(t) = et + \sum_{i=1}^{N(t)-1} C_i,
\]
where we are taking \(S_0 = 1\) to keep with the general construction. The long run average of the cost, which is what the organism would like to minimize, is then
\[
\frac{C(t)}{t} = e + \frac{1}{t} \sum_{i=1}^{N(t)-1} C_i \to e + \mathbb{E}C_1 \mathbb{E}S_1 = e + \frac{E}{d}p(e).
\]
For example, suppose that \( p(e) = \max\{1 - e, 0\} \) for \( e \geq 0 \). Then,

\[
e + \frac{E}{d} p(e)
\]

is minimized at \( E/d \) when \( 1 \geq E/d \) (take \( e = 0 \)) and minimized at 1 when \( 1 \leq E/d \) (take \( e = 1 \)).

For another example, suppose that \( d = 20, E = 15 \), and that we model

\[
p(e) = \max\{1 - \sqrt{e}, 0\},
\]

for \( e \geq 0 \). Differentiating \( Q(e) \equiv e + (15/20)(1 - \sqrt{e}) \) and setting to zero yields

\[
0 = Q'(e) = 1 - \frac{1}{2}e^{-1/2}(15/20) \implies e = \frac{9}{64},
\]

yielding a minimum value of \( Q(9/64) = 39/64 \). \( \Box \)

### 4.2 Point Process

As the name suggests, the basic idea of a point process is to allow us to model a random distribution of points in a space, usually a subset of Euclidean space such as \( \mathbb{R}, [0, \infty) \), or \( \mathbb{R}^d \), for \( d \geq 1 \). Here are a few examples.

**Example 4.2.1.** Renewal processes distribute points on \([0, \infty)\) so that gaps between points are i.i.d. random variables. \( \Box \)

**Example 4.2.2.** The Poisson process, which will be the main object of our focus, is a renewal process which distributes points so gaps are i.i.d. exponential random variables. \( \Box \)

Some other examples could include the following:

1. the breakdown times of a certain part of a car,
2. the position of proteins on a cell membrane,
3. the positions and times of earthquakes in the next 100 years,
4. the locations of diseased deer in a given region of Wisconsin.

### 4.2.1 Preliminaries

We begin with an important mathematical notion, that of a measure.

**Definition 4.2.3.** Let \( E \) be a subset of Euclidean space, and let \( \mathcal{F} \) be a \( \sigma \)-algebra of \( E \) (think of this as the subsets of \( E \), recall Chapter B). Then, \( \mu : \mathcal{F} \to \mathbb{R} \) is a measure if the following three conditions hold.
1. For $A \in \mathcal{F}$, i.e. for $A$ a subset of $E$, $\mu(A) \geq 0$.

2. If $\{A_i\}$ are disjoint sets of $\mathcal{F}$, then

$$\mu \left( \bigcup_i A_i \right) = \sum_i \mu(A_i).$$

3. $\mu(\emptyset) = 0$.

The concept of a measure generalizes the idea of length in one dimension, area in two, etc. In fact, the most important measure is that of Lebesgue measure, which is precisely length, area, volume, etc. For example, if $\mu$ is Lebesgue measure on $\mathbb{R}$, then for any $a < b$ we have $\mu([a, b]) = b - a$, and if $[a, b]$ and $[c, d]$ are disjoint, then

$$\mu \left( [a, b] \cup [c, d] \right) = (b - a) + (d - c).$$

If $\mu$ is Lebesgue measure on $\mathbb{R}^2$, then for any region of space $A \subset \mathbb{R}^2$, we have $\mu(A) = \text{Area}(A)$. The interested reader can verify that Lebesgue measure does satisfy the three assumptions above.

We now turn to point processes and start with some notation, terminology, and definitions. We suppose that $E$ is a subset of Euclidian space, $\mathbb{R}^d$ (or $[0, \infty)$, $\mathbb{R}^2$, etc.). Similarly to our study of the renewal process in the previous section, we want to be able to distribute the points throughout $E$, and have a compact notation that counts the number of points that fall in a given subset $A \subset E$. For renewal processes, we distributed the points by assuming independent gaps between them, and let $N(t)$ denote the counting process giving the number of points up to time $t$.

We assume that $\{X_n, n \geq 0\}$ are random elements of $E$, which represent points in the state space $E$. Next, we define the discrete (random, as it depends upon the point $X_n$) measure by

$$1_{X_n}(A) = \begin{cases} 1, & \text{if } X_n \in A, \\ 0, & \text{if } X_n \notin A. \end{cases}$$

(4.8)

Note, therefore, that $1_{X_n}$ is a function whose domain is $\mathcal{F}$, i.e. the subsets of $E$, and whose range is $\{0, 1\}$, and that it takes the value one whenever $X_n$ is in the subset of interest. Next, we note that by taking the sum over $n$, we find the total number of the points $\{X_n\}$ contained in the set $A$. Therefore, we define the counting measure $N$ by

$$N \overset{\text{def}}{=} \sum_n 1_{X_n},$$

so that for $A \subset E$,

$$N(A) = \sum_n 1_{X_n}(A),$$

simply gives the total number of points in $A \subset E$.

**Definition 4.2.4.** The function $N$ is called a point process, and $\{X_n\}$ are called the points.
We note that as $N$ depends explicitly on the values of the points, $X_n$, it is natural to call such an object a *random measure* since the points themselves are random.

We will make the (technical) running assumption that bounded regions of $A$ must always contain a finite number of points with a probability of one. That is, for any bounded set $A$,

$$P\{N(A) < \infty\} = 1.$$ 

**Example 4.2.5.** For a renewal process, we have $E = [0, \infty)$, and the points are the renewal times $\{S_n\}_{n=0}^{\infty}$. The point process is

$$N = \sum_{n=0}^{\infty} 1_{S_n}.$$ 

Note that the notation for the counting process has changed from $N(t)$ to $N([0,t])$.

**Example 4.2.6.** Consider modeling the positions, and illnesses, of sick deer in a given region. A good choice for a state space would be

$$E = \mathbb{R}^2 \times \{1, 2, \ldots, M\},$$

where the first component of $E$ determines the deer’s location and the second lists its ailment. The point process would then be

$$N = \sum_{n} 1_{(L_{n1}, L_{n2}), m},$$

where $(L_{n1}, L_{n2})$ represents the latitude and longitude of the deer, and $m$ its ailment.

An important statistic of a point process is the *mean measure*, or *intensity*, of the process, which is defined to be

$$\mu(A) = \mathbb{E}N(A),$$

giving the expected number of points found in the region $A$. We note that the intensity is commonly referred to as the *propensity* in the biosciences.

### 4.2.2 The Poisson process

Poisson processes will play a critical role in our modeling and understanding of continuous time Markov chains. We will eventually develop a general notion of a Poisson process, though we begin with the formulation of the one-dimensional model that most people see in their first introduction to probability course. We will refer to this as the *first* formulation.

We suppose that starting at some time, usually taken to be zero, we start counting events (earthquakes, arrivals at a post-office, number of meteors, number of mRNA molecules transcribed, etc.). For each $t$, we obtain a number, $N(t)$, giving the number
of events that has occurred up to time $t$. Note that we have reverted, for the time being, to our notation from renewal processes.

Recalling that a function $f$ is said to be $o(h)$, and written $f \in o(h)$ or $f = o(h)$, if $f(h)/h \to 0$ as $h \to 0$, we make the following modeling assumptions on the process $N(t) (= N([0, t]))$.

**Assumption 4.2.7.** 1. For some $\lambda > 0$, the probability of exactly one event occurring in a given time interval of length $h$ is equal to $\lambda h + o(h)$. Mathematically, this assumption states that for any $t \geq 0$

$$P\{N(t + h) - N(t) = 1\} = \lambda h + o(h), \text{ as } h \to 0.$$ 

2. The probability that 2 or more events occur in an interval of length $h$ is $o(h)$:

$$P\{N(t + h) - N(t) \geq 2\} = o(h), \text{ as } h \to 0.$$ 

3. The random variables $N(t_1) - N(s_1)$ and $N(t_2) - N(s_2)$ are independent for any choice of $s_1 \leq t_1 \leq s_2 \leq t_2$. This is usually termed an independent interval assumption.

We then say that $N(t)$ is a homogeneous Poisson process with *intensity*, *propensity*, or *rate*, $\lambda$. The following proposition describes the distribution associated to the random variables $N(t) - N(s)$, and makes clear why the process $N(t)$ is termed a Poisson process.

**Proposition 4.2.8.** Let $N(t)$ be a Poisson process satisfying the three assumptions above. Then, for any $t \geq s \geq 0$ and $k \in \{0, 1, 2, \ldots \}$,

$$P\{N(t) - N(s) = k\} = e^{-\lambda(t-s)}(\lambda(t-s))^k/k!.$$ 

Proof. We will prove the proposition in the case $s = 0$, with the more general case proved similarly. We must show that under the above assumptions, the number of events happening in any length of time $t$ has a Poisson distribution with parameter $\lambda t$. That is, we will show that

$$P\{N(t) = k\} = e^{-\lambda t}(\lambda t)^k/k!.$$ 

We begin by breaking the interval $[0, t]$ up in to $n$ subintervals of length $t/n$, where $n$ should be thought of as large. We will eventually let $n \to \infty$. We define the two events

$$A_n \overset{\text{def}}{=} \{k \text{ of the subintervals contain exactly 1 event and other } n-k \text{ contain zero}\}$$

$$B_n \overset{\text{def}}{=} \{N(t) = k \text{ and at least one of the subintervals contains 2 or more events}\}$$

The events are mutually exclusive and

$$P\{N(t) = k\} = P\{A_n\} + P\{B_n\}.$$
Note that the left hand side does not depend upon $n$. We will show that $P\{B_n\} \to 0$ as $n \to \infty$, hence proving that events happen one at a time. This is called orderliness. Using Boole’s inequality, which states that
\[
P\left\{\bigcup_i C_i\right\} \leq \sum_i P\{C_i\},
\]
for any set of events $\{C_i\}$, we have
\[
P\{B_n\} \leq P\{\text{at least one subinterval has 2 or more events}\}
\]
\[
= P\left\{\bigcup_{i=1}^n \{ \text{ith subinterval contains 2 or more} \} \right\}
\]
\[
\leq \sum_{i=1}^n P\{ \text{ith subinterval contains 2 or more} \}
\]
\[
= \sum_{i=1}^n o(t/n)
\]
\[
= no(t/n)
\]
\[
= t \left[ \frac{o(t/n)}{t/n} \right].
\]
Thus, $P\{B_n\} \to 0$, as $n \to \infty$. Now we just need to understand the limiting behavior of $P\{A_n\}$. We see from assumptions 1 and 2 that
\[
P\{0 \text{ events occur in a given interval of length } h\} = 1 - \lambda h - o(h).
\]
Now using assumption 3, independence of intervals, we see that
\[
P\{A_n\} = \binom{n}{k} \left[ \frac{\lambda t}{n} + o(t/n) \right]^k \left[ 1 - \left( \frac{\lambda t}{n} - o(t/n) \right) \right]^{n-k}
\]
\[
\approx \binom{n}{k} \left( \frac{\lambda t}{n} \right)^k \left[ 1 - \left( \frac{\lambda t}{n} \right) \right]^{n-k}
\]
\[
= (\lambda t)^k \frac{n!}{(n-k)! k!} \left( \frac{1}{n} \right)^k \left( 1 - \lambda t \left( \frac{1}{n} \right) \right)^n \left( 1 - \lambda t \left( \frac{1}{n} \right) \right)^{-k}.
\]
Noting that
\[
\frac{n!}{(n-k)!} \frac{1}{n^k} = \frac{n \cdot (n-1) \cdots (n-k+1)}{n^k}
\]
\[
= 1 \cdot \left( 1 - \frac{1}{n} \right) \cdots \left( 1 - \frac{k}{n} + \frac{1}{n} \right) \xrightarrow{n \to \infty} 1
\]
\[
\left( 1 - \lambda t \left( \frac{1}{n} \right) \right)^n \xrightarrow{n \to \infty} e^{-\lambda t},
\]
\[
\left( 1 - \frac{\lambda t}{n} \right)^{-k} \xrightarrow{n \to \infty} 1,
\]
we find that

\[ \lim_{n \to \infty} P\{A_n\} = e^{-\lambda t} \frac{(\lambda t)^k}{k!}. \]

Therefore, we see that

\[ P\{N(t) = k\} = \lim_{n \to \infty} P\{A_n\} + \lim_{n \to \infty} P\{B_n\} = e^{-\lambda t} \frac{(\lambda t)^k}{k!}, \]

proving the result.

We may use Proposition 4.2.8 to make the following (non-rigorous) argument. First, letting \( S_0 \) denote the time of the first increase in the process,

\[ P\{S_0 > t\} = P\{N(t) = 0\} = e^{-\lambda t}. \]

Therefore, the distribution of the first event time is exponentially distributed with a parameter of \( \lambda \). Next, it should be at least believable that because of the independent increments assumption, if we are given \( S_0 \) then the distribution of \( S_1 - S_0 \) is also exponentially distributed with parameter \( \lambda \). (To make this point precise, we need the Strong Markov Property, which is beyond the scope of these notes.) Continuing, the distribution of the inter-event times are all seen to be exponentially distributed with a parameter of \( \lambda \). For a rigorous derivation of this result, see [37]. Thus, we see that \( N(t) \) is simply the counting process of a renewal process with inter-event times determined by exponential random variables, which we refer to here as the second formulation of a Poisson process.

We now present a third formulation of a one dimensional Poisson process. We say that \( N \) is a Poisson process with intensity \( \lambda \) if

(i) for any \( A \subset \mathbb{R}_{\geq 0} \) and \( k \geq 0 \), we have that

\[ P\{N(A) = k\} = e^{-|A|} \frac{\lambda^k |A|^k}{k!}, \]

where \( |A| \) is the Lebesgue measure of \( A \), and

(ii) If \( A_1, \ldots, A_k \) are disjoint subsets of \( E \), then \( N(A_1), \ldots, N(A_k) \) are independent random variables.

It is straightforward to show, by taking \( |A| = o(h) \), that any process satisfying this third formulation (or the second) also satisfies the assumptions of the first. Also, as this formulation assumes Proposition 4.2.8, the logic of the previous paragraph applies, and any process satisfying this third formulation also satisfies the second. Therefore, we have the implications \( 1 \implies 2, 3 \implies 1, 1 \implies 3 \) (from a slight generalization of Proposition 4.2.8), \( 3 \implies 2 \), and \( 2 \implies 1 \). Therefore, the formulations are all equivalent.

We will now generalize the notion of a Poisson process in multiple ways, and do so by generalizing the third formulation.
Definition 4.2.9. Let $N$ be a point process with state space $E \in \mathbb{R}^d$, and let $\mu$ be a measure on $\mathbb{R}^d$. We say that $N$ is a \textit{Poisson process with mean measure $\mu$}, or a \textit{Poisson random measure}, if the following two conditions hold:

(i) For $A \subseteq E$ and $k \geq 0$, we have that

$$P\{N(A) = k\} = \begin{cases} e^{-\mu(A)}(\mu(A))^k k!, & \text{if } \mu(A) < \infty, \\ 0, & \text{if } \mu(A) = \infty. \end{cases}$$

(ii) If $A_1, \ldots, A_k$ are disjoint subsets of $E$, then $N(A_1), \ldots, N(A_k)$ are independent random variables.

Therefore, $N$ is a Poisson process if the random number of points in $A$ has a Poisson distribution with parameter $\mu(A)$, and the number of points in disjoint regions are independent. Note that the mean measure of a Poisson process, $\mu(\cdot)$, completely determines the process.

One choice for the mean measure would be a multiple of Lebesgue measure, which gives length in $\mathbb{R}$, area in $\mathbb{R}^2$, volume in $\mathbb{R}^3$, etc. That is, if $\mu((a, b]) = \lambda(b - a)$, for $a, b \in \mathbb{R}$, then $\mu$ is said to be Lebesgue measure with rate, or intensity, $\lambda$. If $\lambda = 1$, then the measure is said to be \textit{unit-rate}. For another example, a Poisson process with Lebesgue measure in $\mathbb{R}^2$ satisfies $\mu(A) = \text{Area}(A)$. When the mean measure is a multiple of Lebesgue measure, we call the process \textit{homogeneous}. In the homogeneous case, a more compact notation for the mean measure is $EN(A) = \lambda |A|$, where $|A|$ is the Lebesgue measure of $A$. Note that the original formulation of a Poisson process, given in the three assumptions at the beginning of this section, was for a homogeneous Poisson process in one dimension with rate $\lambda$.

Of course, with our new notion of a Poisson process, we can generalize easily away from the homogeneous case. As an important example of such a generalization, suppose that for open intervals $(a, b)$, the mean measure $\mu$ for a Poisson process is

$$\mu((a, b)) = \Lambda(b) - \Lambda(a),$$

for some non-decreasing, absolutely continuous function $\Lambda$. If $\Lambda$ has density $\lambda$ (i.e. if $\Lambda$ is differentiable), then

$$\mu((a, b)) = \Lambda(b) - \Lambda(a) = \int_a^b \lambda(s)ds,$$

or, more generally,

$$\mu(A) = \int_A \lambda(s)ds.$$ 

Note that, by construction, $\lambda$ takes only non-negative values as $\Lambda$ is non-decreasing. Further, we have that

$$P\{N((a, b)) = k\} = e^{-(\Lambda(b) - \Lambda(a))}(\Lambda(b) - \Lambda(a))^k k! = e^{-\int_a^b \lambda(s)ds} \left(\int_a^b \lambda(s)ds\right)^k k!,$$
and more generally,

\[ P\{N(A) = k\} = e^{-\int_A \lambda(s)ds} \frac{\left(\int_A \lambda(s)ds\right)^k}{k!}, \]

for any set \( A \subset \mathbb{R} \). The function \( \lambda \) is usually termed the rate, intensity, or propensity function of the Poisson process, depending upon the specific scientific field in which the model is being considered.

We note that if the first assumption in Assumption 4.2.7 were changed to

\[ P\{N(t + h) - N(t) = 1\} = \lambda(t)h + o(h), \quad \text{as} \quad h \to 0, \tag{4.9} \]

then it can be argued that the resulting process is equivalent to the non-homogeneous Poisson process just described. This is important from a modeling perspective, as it is usually an assumption of the form (4.9) that is the starting point of a mathematical model.

**Example 4.2.10.** Suppose we believe that the arrival times of frogs to a pond can be reasonably modeled by a Poisson process. We suppose that frogs are arriving at a rate of 3 per hour. What is the probability that no frogs will arrive in the next hour? What is the probability that 12 or less frogs arrive in the next five hours?

**Solution:** Let \( N([0,t]) \) be the Poisson process of rate 3 determining the number of frogs to arrive in the next \( t \) hours. Then,

\[ P\{N([0,1]) = 0\} = e^{-3\lambda(3 \cdot 1)} \frac{0!}{0!} = e^{-3} = 0.04978. \]

Further,

\[ P\{N([0,5]) \leq 12\} = \sum_{k=0}^{12} P\{N([0,5]) = k\} = \sum_{k=0}^{12} e^{-3 \cdot 5} \frac{(3 \cdot 5)^k}{k!} \approx 0.2676. \]

**Example 4.2.11.** We change the previous example by recognizing that it is unlikely that frogs would arrive anywhere according to a homogenous process. Instead, the rate of arrival should fluctuate throughout the day. Therefore, we change our model and suppose that the arrival of the frogs is modeled by a non-homogeneous Poisson process with intensity function

\[ \lambda(t) = 3 + \sin(t/4), \]

where \( t = 0 \) is taken to be 8AM. Assuming it is exactly 8AM now, what is the probability that no frogs will arrive in the next hour? What is the probability that 12 or less frogs arrive in the next five hours?

**Solution:** We let \( N([0,t]) \) be the Poisson process with intensity \( \lambda(t) = 3 + \sin(t/4) \). Then,

\[ P\{N([0,1]) = 0\} = e^{-\int_0^1 (3 + \sin(t/4))dt} \frac{\int_0^1 (3 + \sin(t/4))dt}{0!} = e^{-7 - 4 \cos(1/4)} \approx 0.044. \]
Further,
\[
P\{N([0,5]) \leq 12\} = k = \sum_{k=0}^{12} P\{N([0,5]) = k\}
\]
\[
= \sum_{k=0}^{12} e^{-\int_{0}^{5} (3+\sin(t/4))dt} \frac{\left(\int_{0}^{5} (3+\sin(t/4))dt\right)^{k}}{k!}
\]
\[
\approx 0.1017.
\]

**Example 4.2.12.** Suppose that we believe that bears in upper Wisconsin are distributed as a spatial Poisson process with rate \(\lambda = 3\) per square mile. Suppose you are camping in northern Wisconsin tonight. What is the expected distance between you and the nearest bear?

**Solution:** Let \(R\) be the distance of the nearest bear from the campground and let \(d(r)\) be a disc of radius \(r\) centered at the campground. Then,
\[
P\{R > r\} = P\{N(d(r)) = 0\} = e^{-\lambda|d(r)|}.
\]
Note that we have
\[
|d(r)| = \pi r^2,
\]
and so
\[
P\{R > r\} = e^{-3\pi r^2}.
\]
Therefore, using the substitution \(u/\sqrt{2} = \sqrt{3\pi r}\), the expected distance is
\[
\mathbb{E}(R) = \int_{0}^{\infty} P\{R > r\}dr \\
= \int_{0}^{\infty} e^{-3\pi r^2}dr \\
= \frac{1}{\sqrt{2\sqrt{3\pi}}} \int_{0}^{\infty} e^{-u^2/2}du \\
= \frac{1}{\sqrt{2\sqrt{3\pi}}} \sqrt{2\pi} \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} e^{-u^2/2}du \\
= \frac{1}{\sqrt{3}} \cdot \frac{1}{2} \\
= .2887 \text{ miles}.
\]

**Simulating non-homogeneous Poisson processes: a first try**

Suppose we believe that a physical process of interest to us can be modeled as a non-homogeneous Poisson process with intensity \(\lambda(t)\). We would like to be able to
simulate this process. Recall that in the homogeneous case this was simple: generate exponential random variables to give the gaps between points. We try a similar approach here by calculating the distribution of the waiting times between points. We let \( S_1 \) be the time of the first point of the process, which we denote by \( N \). Then,

\[
P\{S_1 > t\} = P\{N(t) = 0\} = e^{-\int_0^t \lambda(s)ds}.
\]

Therefore, for \( t \geq 0 \), the distribution function of the first point is

\[
F_{S_1}(t) = 1 - e^{-\int_0^t \lambda(s)ds},
\]

which has density

\[
f_{S_1}(t) = \lambda(t)e^{-\int_0^t \lambda(s)ds}.
\]

Supposing we can generate a random variable according to this distribution, and we will see how to do this in later sections, the distribution of the wait time between \( S_1 \) and \( S_2 \) appears to be

\[
F_{S_2-S_1}(t) = 1 - e^{-\int_{S_1}^{S_1+S_2} \lambda(s)ds},
\]

which depends explicitly upon the value of \( S_1 \). At this point, it does not appear we have the necessary mathematical machinery to efficiently simulate this process. We will see that it will be helpful for us to consider transformations of Poisson processes in order to (efficiently) solve this problem.

### 4.2.3 Transformations of Poisson processes

**Understanding the homogeneous Poisson processes: our first transformation**

Suppose that we let \( E_n^\lambda, n \geq 0 \), be i.i.d. exponential random variables with parameter \( \lambda > 0 \). Define

\[
S_n^\lambda \overset{\text{def}}{=} \sum_{i=1}^n E_i^\lambda,
\]

to be the points associated with a counting process \( N_\lambda \). That is, we let

\[
N_\lambda([0, t]) \overset{\text{def}}{=} \sum_{n=1}^\infty 1_{[0, t]}(S_n^\lambda) = \sum_{n=1}^\infty 1\{S_n^\lambda \leq t\} = \sum_{n=1}^\infty 1\{\sum_{i=1}^n E_i^\lambda \leq t\}.
\]

We know that \( N_\lambda \) is a one-dimensional homogeneous Poisson process with intensity \( \lambda > 0 \). However, by the basic properties of exponential random variables (see Section B.3.4 if you need a refresher), we know that \( E_i^\lambda \overset{\mathcal{D}}{=} E_i/\lambda \), where \( E_i \) are unit exponential random variables. Therefore,

\[
N_\lambda([0, t]) = \sum_{n=1}^\infty 1\{\sum_{i=1}^n E_i \leq t\} \overset{\mathcal{D}}{=} \sum_{n=1}^\infty 1\{\sum_{i=1}^n E_i \leq \lambda t\} = N([0, \lambda t]), \quad (4.10)
\]
where $N$ is a *unit-rate* Poisson process. The importance of the relation (4.10) can not be overstated, and we will interpret it in two different ways. First, it can be viewed as a *time-change*. That is, it shows that if a homogenous Poisson process with rate $\lambda > 0$ is desired, then it is sufficient to start with a *unit-rate* process and simply “move” along its time-frame at rate $\lambda$. If $\lambda > 1$ we move faster than unit speed, whereas if $\lambda < 1$, we move slower.

Second, (4.10) can be viewed as a spacial shifting of points. It shows if the position of the points of a homogeneous process of rate $\lambda > 0$ are multiplied by $\lambda$, then the resulting point process is also Poisson, and it is, in fact, a homogeneous process of rate 1. Likewise, we could start with a unit-rate process and divide the position of each point by $\lambda$ to get a homogeneous process with rate $\lambda$. This phenomenon will be explored further in the next subsection.

Both interpretations are important and will be returned to repeatedly. This is our first example of a transformation of a Poisson process, via time or space, yielding another Poisson process. In the next section we greatly expand our understanding of such transformations.

**General transformations of Poisson processes**

The material in this section is, to a large extent, similar to that of Section 4.3 in Resnick [37].

We return to the example at the end of the last subsection in which we transformed one Poisson process into another, though we now take a slightly different perspective. Let $E_i$ denote independent, unit-exponential random variables, and let $S_n = \sum_{i=1}^n E_i$. Letting

$$N([0, t]) = \sum_{n=1}^\infty 1_{\{S_n \leq t\}},$$

we know that $N$ is a unit-rate Poisson process. For $\lambda > 0$, let $T : \mathbb{R} \to \mathbb{R}$ be defined via

$$T(x) = \frac{x}{\lambda}.$$

We then have that

$$N_{\lambda}([0, t]) \overset{\text{def}}{=} \sum_{n=1}^\infty 1_{\{T(S_n) \leq t\}} = \sum_{n=1}^\infty 1_{\{S_n/\lambda \leq t\}} = \sum_{n=1}^\infty 1_{\{\sum_{i=1}^n E_i/\lambda \leq t\}}$$

is a homogeneous Poisson process with rate $\lambda$ because the $E_i/\lambda$ are independent exponential random variables with a parameter of $\lambda$. Also, the mean measure of the process has changed during the transformation from

$$\mu((a, b)) = b - a,$$

to

$$\mu'((a, b)) = \lambda(b - a),$$
where $b > a$. Note that
\[
\mu'((a,b)) = \lambda(b-a) = (\lambda b - \lambda a) = \mu((\lambda a, \lambda b)) = \mu(T^{-1}(a,b)).
\]
Collecting thoughts, we see that moving the points around via a function, or transformation, resulted in another Poisson process, and the new mean measure can be understood via the inverse of $T$ and the original measure $\mu$. That is, the new measure is induced by the transformation. We will see below that this is a general result, however we begin by building up our terminology.

For two euclidean spaces $E$ and $E'$, we assume the existence of a some one-to-one function $T : E \to E'$. Note that the function $T^{-1}$ induces a set mapping from the subsets of $E'$ to those of $E$. That is, for $A' \subset E'$ we have
\[
T^{-1}(A') = \{x \in E : T(x) \in A'\}.
\]
Therefore, $T^{-1}(A')$ is simply the pre-image of $A'$ under $T$.

We want to take the points of a general Poisson process, $N$, defined on $E$, apply $T$ to them, and consider the resulting point process in $E'$. Note that because $T$ is one-to-one, for any $A \in E$ we have that $N(A) = N(T(A))$. That is, points can not be combined.

We will denote the mean measure of $N$ by $\mu$ and the points associated with $N$ as $X_n$. The goal is to be able to count the number of points, $T(X_n) \in E'$, in a given region $A' \subset E'$. Letting $N'$ denote that counting process, we see that
\[
N'(A') = \sum_n 1_{T(X_n)(A')} = \sum_n 1_{\{T(X_n) \in A'\}} = \sum_n 1_{\{X_n \in T^{-1}(A')\}}
= \sum_n 1_{X_n}(T^{-1}(A'))
= N(T^{-1}(A'))
= N \circ T^{-1}(A').
\]
Further, we see the expected number of points can also be calculated via
\[
\mu'(A') \overset{\text{def}}{=} \mu(T^{-1}(A')) = \mu \circ T^{-1}(A').
\]
That is, once again, the mean measure is determined by $T^{-1}$ and the original mean measure $\mu$. The following proposition is incredibly useful and, even though the proof is rather straightforward, nearly miraculous.

**Proposition 4.2.13.** Suppose that $T : E \to E'$ is a one-to-one (injective) mapping between Euclidean spaces such that if $B' \subset E'$ is bounded, then so is $T^{-1}(B') \subset E$. If $N$ is Poisson process on $E$ with mean measure $\mu$ and points $\{X_n\}$, then
\[
N' \overset{\text{def}}{=} N \circ T^{-1}
\]
is a Poisson process on $E'$ with points $\{T(X_n)\}$ and mean measure
\[
\mu' = \mu \circ T^{-1}.
\]
Proof. We need to show that the two properties of a Poisson process as given in Definition 4.2.9 are satisfied. Firstly, we have for any $B' \subset E'$ and $k \geq 0$,

$$\begin{align*}
P\{N'(B') = k\} &= P\{N(T^{-1}(B')) = k\} \\
&= e^{-\mu(T^{-1}(B'))} \left(\frac{\mu(T^{-1}(B'))^k}{k!}\right) \\
&= e^{-\mu'(B')} \left(\frac{\mu'(B')^k}{k!}\right),
\end{align*}$$

where the second equality follows since $N$ is a Poisson process with mean measure $\mu$. Next, if $B'_1, \ldots, B'_m$ are disjoint, then so are $T^{-1}(B'_1), \ldots, T^{-1}(B'_m)$. Therefore, the random variables

$$\{N'(B'_1), \ldots, N'(B'_m)\} = \{N(T^{-1}(B'_1)), \ldots, N(T^{-1}(B'_m))\},$$

are independent. \qed

Example 4.2.14. Let $N = \sum_n 1_{X_n}$ be the usual homogeneous Poisson process with rate $\lambda = 1$ and state space $E = [0, \infty)$. Let $T(x) = x^2$. Then

$$N' = \sum_n 1_{X_n^2}$$

is a Poisson process on $[0, \infty)$ with mean measure

$$\mu'([0, t]) = \mu(T^{-1}([0, t])) = \mu\{x: x^2 \leq t\} = \mu([0, \sqrt{t}]) = \sqrt{t}. \qed$$

We now return to the non-homogeneous Poisson process. Specifically, we suppose we have a Poisson process with state space $[0, \infty)$ and mean measure $\mu$, which we assume is absolutely continuous with density $\lambda(t)$. That is,

$$\mu(A) = \int_A \lambda(s)ds.$$ 

As already discussed, we call $N$ a non-homogeneous Poisson process with local intensity $\lambda(t)$. We assume that $\int_0^\infty \lambda(s)ds = \infty$.

We will now show how to obtain this process as a transformation of a homogeneous Poisson process, which, aside from being interesting in its own right, will show how to simulate a non-homogeneous Poisson process in an efficient manner. We begin by defining the non-decreasing function $m$ via

$$m(t) = \int_0^t \lambda(s)ds.$$ 

Next, we define an “inverse function” for $m$, denoted $I$, via

$$I(x) = \inf\{t: m(t) \geq x\}, \quad x \geq 0.$$
Note, for example, that
\[ m(I(x)) = \int_0^{I(x)} \lambda(s) ds = x, \]

as \( I(x) \) is defined to be the first such value, \( u \) say, for which
\[ \int_0^u \lambda(s) ds = x. \]  \hspace{1cm} (4.11)

Note, however, that if there is an \( \epsilon > 0 \) for which \( \lambda(s) = 0 \) for \( s \in (I(x), I(x)+\epsilon) \), then the solution, \( u \), to (4.11) need not be unique. Thus, we may not have \( I(m(x)) = x \).

For example suppose that \( u < x \) with \( \lambda(s) = 0 \) for all \( s \in (u,x) \). Then
\[ I(m(x)) = \inf\{ t : m(t) \geq m(x) \} \leq u < x, \]

since
\[ \int_0^u \lambda(s) ds = m(u) = m(x) = \int_0^x \lambda(s) ds. \]

We do note, however, that we must have \( I(m(x)) \leq x \), as \( x \) will certainly always satisfy \( m(x) \geq m(x) \), it just may not be the smallest such value.

We have the following lemma.

**Lemma 4.2.15.** The function \( I \) is strictly increasing.

**Proof.** Suppose that \( x < y \). Then,
\[ \int_0^{I(x)} \lambda(s) ds = x < y = \int_0^{I(y)} \lambda(s) ds \implies I(x) < I(y). \]

Thus, the function \( I \) is a valid transformation for our purposes, and we will apply it to the points of a homogeneous Poisson process and see what we get (hint: it will be the correct non-homogeneous process).

Therefore, we let \( N = \sum_n 1_{X_n} \) be a unit-rate homogeneous Poisson process, with corresponding points \( X_n \). We then let
\[ N' = \sum_n 1_{I(X_n)}, \] \hspace{1cm} (4.12)

which by Proposition 4.2.13, and the fact that \( I \) is strictly increasing and hence one to one, is a Poisson process on \([0, \infty)\).

What is the mean measure? To find out, the following lemma is useful.

**Lemma 4.2.16.** For all \( x, t \geq 0 \),
\[ I(x) \leq t \iff x \leq m(t). \]
Proof. First suppose that $I(x) \leq t$. Applying $m$, which is non-decreasing, to both sides yields

$$m(I(x)) \leq m(t) \implies x \leq m(t),$$

where we used that $m(I(x)) = x$ for any $x$.

Next, we suppose that $x \leq m(t)$ and apply $I$, which is also non-decreasing. We see

$$I(x) \leq I(m(t)).$$

Using that $I(m(t)) \leq t$ gives us the desired implication.

Returning to the mean measure for our new process, we have that for any $t \geq 0$,

$$\{x : I(x) \leq t\} = \{x : x \leq m(t)\},$$

and letting $\mu$ be (unit-rate) Lebesgue measure, we have

$$\mu'([0,t]) = \mu(I^{-1}([0,t])) = \mu(\{x : I(x) \leq t\}) = \mu(\{x : x \leq m(t)\})$$

$$= m(t)$$

$$= \int_0^t \lambda(s)ds,$$

certifying the resulting process is, indeed, the non-homogeneous process with intensity $\lambda$.

Returning to $N'$ defined in (4.12), we also see that

$$N'([0,t]) = \sum_n 1_{I(X_n)}([0,t]) = \sum_n 1_{I(X_n) \leq t} = \sum_n 1_{X_n \leq m(t)}$$

$$= \sum_n 1_{X_n([0,m(t])] = N([0, \int_0^t \lambda(s)ds]).$$

Condensing notation, we see that if $N'$ is a Poisson process with intensity $\lambda$, then

$$N'(t) = N\left(\int_0^t \lambda(s)ds\right). \quad (4.13)$$

Thus, to get a non-homogeneous Poisson process it is enough to correctly modulate the speed at which the “clock” runs on a Poisson process with points determined by unit exponentials. This is called a time-changed representation for the non-homogeneous Poisson process and will play a critical role in our understanding of continuous time Markov chains in later chapters.
Simulating non-homogeneous Poisson processes

We may now return to the question of efficiently simulating a non-homogeneous Poisson process. We see that simulating such a process is equivalent to simulating the right hand side of (4.13), and the following strategy does just that.

1. Let $E_1$ be an exponential random variable with parameter one. This is the first point of the homogeneous Poisson point process.

2. Solve for the smallest $t_1$ that satisfies
   \[ \int_0^{t_1} \lambda(s)ds = E_1. \]
   Using the notation from above, we see this is equivalent to setting $t_1 = I(E_1)$. Note that if the anti-derivative of $\lambda$ is of a nice form, then solving this equation will be simple. The value $t_1$ is the first point of the non-homogeneous process.

3. Repeat. Let $E_2$ be an exponential random variable with parameter one. This is the second point of the homogeneous Poisson point process.

4. Solve for the smallest $t_2$ that satisfies
   \[ \int_{t_1}^{t_2} \lambda(s)ds = E_2. \]
   The value $t_2$ is the second point of the non-homogeneous process.

5. Repeat until a desired number of points for the non-homogeneous process have been generated.

The following is the general algorithm for generating the points of a non-homogeneous Poisson process with intensity $\lambda$. The points of the non-homogeneous process will be $t_n$ for $n \geq 1$.

Algorithm. Set $t_0 = 0$. Set $n = 1$.

1. Let $E_n$ be an exponential random variable with parameter one, which is independent from all other random variables already generated.

2. Find the smallest $u \geq 0$ for which
   \[ \int_{t_{n-1}}^{u} \lambda(s)ds = E_n. \]
   Set $t_n = u$. Note this is equivalent to solving
   \[ \int_0^{u} \lambda(s)ds = E_1 + \cdots + E_n, \]
   for $u$.  

116
3. Set $n \leftarrow n + 1$.

4. Return to step 1 or break.

The above algorithm for simulating a non-homogeneous Poisson process is the core of future methods for the simulation of continuous time Markov chains. You should truly try to understand the algorithm before moving on.

**Example 4.2.17.** Suppose that $\lambda(t) = t$ for all $t \geq 0$. We consider the problem of simulating the non-homogeneous Poisson process with intensity $\lambda(t)$:

$$N \left( \int_0^t \lambda(s) ds \right) = N \left( \int_0^t s ds \right) = N \left( \frac{1}{2} t^2 \right),$$

where $N$ is a unit-rate Poisson process. Supposing our stream of exponential random variables begins with $E = [1.8190, 0.2303, 1.1673, 0.6376, 1.7979]$, we have the following.

1. To find the first jump of our counting process we solve

$$\int_0^{t_1} \lambda(s) ds = \frac{1}{2} t_1^2 = 1.8190 \implies t_1 \approx 1.907.$$

2. To find the second jump, we solve

$$\int_{1.907}^{t_2} s ds = 0.2303 \implies \frac{1}{2} t_2^2 - \frac{1}{2} (1.907)^2 = 0.2303 \implies t_2 = 2.024.$$
Similarly, we simply could have solved for $t_2$ via

$$\int_0^{t_2} \lambda(s)\,ds = \frac{1}{2}t_2^2 = (1.8190 + 0.2303) \implies t_2 = 2.024.$$ 

3. Solving for $t_3$:

$$\frac{1}{2}t_3^2 = (1.8190 + 0.2303 + 1.1673) \implies t_3 = 2.536.$$ 

4. Solving for $t_4$:

$$\frac{1}{2}t_4^2 = (1.8190 + 0.2303 + 1.1673 + 0.6376) \implies t_4 = 2.776.$$ 

5. Solving for $t_5$:

$$\frac{1}{2}t_5^2 = (1.8190 + 0.2303 + 1.1673 + 0.6376 + 1.7979) \implies t_5 = 3.362.$$ 

Note that the actual time between events, $t_n - t_{n-1}$, is (non-monotonically) reducing. Plots of (a) the counting process, and (b) the deterministic function $t^2/2$, are found in Figure 4.2.1. □

**Differing time frames**

Consider the non-homogeneous Poisson process with intensity $\lambda(t)$,

$$\tilde{N}(t) \overset{\text{def}}{=} N\left(\int_0^t \lambda(s)\,ds\right),$$

where $N$ is a unit-rate Poisson process. Note that the variable $t$ represents time. However, there is another time frame in the problem: that of the Poisson process $N$. Let

$$\tau(t) \overset{\text{def}}{=} \int_0^t \lambda(s)\,ds,$$

and note that the process $\tilde{N}$ can be written as

$$\tilde{N}(t) = N\left(\int_0^t \lambda(s)\,ds\right) = N(\tau(t)).$$

We see that $\tau(t)$ give the current time of the Poisson process $N$. We can use this notation to simplify our calculations. For example, letting $\mathcal{F}_{s}$ denote all the information pertaining to the entire history of $N$ up through time $s$, we have that

$$P\{N(\tau(t) + h) - N(\tau(t)) = 1 \mid \mathcal{F}_{\tau(t)}\} = h + o(h), \quad \text{as } h \to 0,$$
which implies that
\[
P\{\tilde{N}(t+h) - \tilde{N}(t) = 1 \mid \mathcal{F}_{\tau(t)}\} = P\{N \left( \int_0^{t+h} \lambda(s)ds \right) - N \left( \int_0^t \lambda(s)ds \right) = 1 \mid \mathcal{F}_{\tau(t)}\}
\]
\[
= P\{N \left( \int_t^{t+h} \lambda(s)ds + \tau(t) \right) - N(\tau(t)) = 1 \mid \mathcal{F}_{\tau(t)}\}
\]
\[
= P\{N(\lambda(t)h + o(h) + \tau(t)) - N(\tau(t)) = 1 \mid \mathcal{F}_{\tau(t)}\}
\]
\[
= \lambda(t)h + o(h), \quad \text{as } h \to 0,
\]
where the second to last equality follows from calculus.

**Processes with random intensity**

Suppose that \( X(t) \) is some stochastic process defined for all time \( t \geq 0 \). For example, we could have that \( X_n, n \geq 0 \), is a discrete time Markov chain and then extend \( X_n \) to all of \( \mathbb{R} \geq 0 \) by defining
\[
X(t) \overset{\text{def}}{=} X_{\lfloor t \rfloor},
\]
where \( \lfloor x \rfloor \) is the largest integer less than or equal to \( x \). Note that in this case \( X(t) \) is a step function, though in general this need not be the case.

We now consider the process which, conditioned upon the value \( X(t) \), behaves locally like a non-homogeneous Poisson process with intensity function \( \lambda(X(t)) \). That is, letting our desired process be denoted by \( \tilde{N} \) and letting all of the information pertaining to the history of \( X \) up through time \( t \) be denoted as \( \mathcal{F}_{X(t)} \), we want \( \tilde{N} \) to satisfy the relations
\[
P\{\tilde{N}(t+h) - \tilde{N}(t) = 1 \mid \mathcal{F}_{X(t)}\} = \lambda(X(t))h + o(h), \quad \text{as } h \to 0
\]
\[
P\{\tilde{N}(t+h) - \tilde{N}(t) \geq 2 \mid \mathcal{F}_{X(t)}\} = o(h), \quad \text{as } h \to 0.
\]

From the results of the previous section, we believe we can model such a process via
\[
\tilde{N}(t) = N \left( \int_0^t \lambda(X(s))ds \right),
\]
where \( N \) is a unit-rate Poisson process that is independent from \( X \). Later we will discuss how the independence assumption can be weakened a bit. Note that conditioned upon \( \mathcal{F}_{X(t)} \), the history of \( X \) up through time \( t \), the expected number of points in the interval \([0, t] \) is
\[
\mathbb{E}\left[ \tilde{N}(t) \mid \mathcal{F}_{X_t} \right] = \mathbb{E}\left[ N \left( \int_0^t \lambda(X(s))ds \right) \mid \mathcal{F}_{X_t} \right] = \int_0^t \lambda(X(s))ds.
\]
Taking expectations again shows that
\[
\mathbb{E}\tilde{N}(t) = \mathbb{E} \int_0^t \lambda(X(s))ds = \int_0^t \mathbb{E}\lambda(X(s))ds.
\]
If $\lambda$ is a nonlinear function, then it is not permissible to switch the expectation with $\lambda$, and so we do not have that the expected number of points is $\int_0^t \lambda(\mathbb{E}(X(s)))ds$! This mistake has been made repeatedly in the science literature and will be returned to again when we discuss stochastic models of biochemical reaction networks.

We will show that the process (4.15) satisfies one of the modeling assumptions (4.14), leaving the second for a homework exercise. Denoting

$$\tau(t) = \int_0^t \lambda(X(s))ds,$$

which itself is a stochastic process, we have

$$P\{\tilde{N}(t+h) - \tilde{N}(t) = 1 \mid \mathcal{F}_{X(t)}\}$$

$$= P\left\{N\left(\int_0^{t+h} \lambda(X(s))ds\right) - N\left(\int_0^t \lambda(X(s))ds\right) = 1 \mid \mathcal{F}_{X(t)}\right\}$$

$$= P\left\{N\left(\int_t^{t+h} \lambda(X(s))ds + \tau(t)\right) - N(\tau(t)) = 1 \mid \mathcal{F}_{X(t)}\right\}$$

$$= P\left\{N(\lambda(X(t))h + o(h) + \tau(t)) - N(\tau(t)) = 1 \mid \mathcal{F}_{X(t)}\right\}$$

$$= \lambda(X(t))h + o(h),$$

valid as $h \to 0$, where the last equality follows from the independence of $X$ (and hence $\mathcal{F}_{X(t)}$) and $N$. The remaining condition is left as a homework exercise.

Note that we did not strictly require that $N$ and $X$ be independent. Instead, we only require that for all $t \geq 0$, the increments $N(s + \tau(t)) - N(\tau(t))$ are independent from $\mathcal{F}_{X(t)}$. That is, loosely, we require that $X$ can not look into the future behavior of $N$. This seems like a minor point, but it will have large modeling consequences with the first such example found in Example 4.2.19. In Example 4.2.20 we demonstrate what can go wrong if we do not have such a condition.

Example 4.2.18. Suppose $X_n$ gives the number of people living in a valley at the beginning of year $n$ that are of child bearing age. We suppose $X_n$ is a random process dependent upon multiple environmental factors, such as recent weather, and the population from the previous year. We then suppose that the time of births in this valley can be modeled via a Poisson process with local intensity $\lambda(X(n))$ for all of year $n$. Letting $X(t) \overset{\text{def}}{=} X_{\lfloor t \rfloor}$ (which is the process attained by extending $X_n$ to all of $\mathbb{R}_{\geq 0}$), we see that the time of births can be modeled via

$$N\left(\int_0^t \lambda(X(s))ds\right),$$

where $N$ is a unit-rate Poisson process. Note that it would be reasonable to assume that $\lambda(\cdot)$ is a non-negative function with $\lambda(0) = 0$. 

120
Example 4.2.19. We give a general model for arrivals and departures. This formulation could be used to model a queue, the transcription and degradation of mRNA, or something else.

We suppose that arrivals are taking place at a constant rate of $\lambda > 0$. Therefore, letting $N_1$ denote a unit-rate Poisson process, we define

$$A(t) \overset{\text{def}}{=} N_1 \left( \int_0^t \lambda ds \right) = N_1(\lambda t),$$

as our arrival process. We let $N_2$ be another Poisson process that is independent of $N_1$. Let $X(t)$ be the number of people in the queue at time $t$ and assume that departures are happening at a rate of $\mu(X(t))$. That is,

$$D(t) \overset{\text{def}}{=} N_2 \left( \int_0^t \mu(X(s))ds \right) \quad (4.16)$$
is the departure process. Another way to formulate the departure process would be to simply require that

$$P\{D(t + h) - D(t) = 1 \mid \mathcal{F}_X(t)\} = \lambda(X(t))h + o(h), \quad \text{as } h \to 0$$

and

$$P\{D(t + h) - D(t) \geq 2 \mid \mathcal{F}_X(t)\} = o(h), \quad \text{as } h \to 0, \quad (4.17)$$

and then note that $D(t)$, as defined in (4.16), satisfies the conditions. Noting that we must have $X(t) = X(0) + A(t) - D(t)$, we see that $X(t)$ is the solution to the stochastic equation

$$X(t) = X(0) + N_1(\lambda t) - N_2 \left( \int_0^t \mu(X(s))ds \right).$$

Such an equation is an example of a random time change representation. Existence of a unique solution to the above equation can be shown by a “jump by jump” argument, which we will discuss in detail later in the course. We note, however, that in the above formulation $N_2$ is not independent from $X$. However, letting

$$\tau(t) = \int_0^t \mu(X(s))ds,$$

we see that $N_2(\tau(t) + t) - N_2(\tau(t)), \ t \geq 0,$ is independent of $\mathcal{F}_X(t)$, all information pertaining to $X$ up until time $t$. This allows us to conclude that conditions (4.17) are still valid.

□

Example 4.2.20. We now demonstrate what can go wrong when the processes $X$ and $N$ are too dependent. Let $E_i$ denote independent, exponential random variables with a parameter of one. Let $N$ be the Poisson point process with points

$$S_n = \sum_{i=1}^n E_i,$$
where we take \( S_0 = 0 \). Now define \( X(t) = S_{\lfloor t \rfloor + 1} - S_{\lfloor t \rfloor} \). That is, \( X(t) \) gives the value of the waiting time of the current “gap” in the points. For example,

\[
X(0) = S_1 - S_0 = E_1.
\]

Finally, define the counting process

\[
\tilde{N}(t) \overset{\text{def}}{=} N \left( \int_0^t X(s) ds \right).
\]

Note that \( X(t) \) looks into the future of \( N \) by giving the time between points; for example, \( X(0) \) determines the behavior of \( N \) up until the first jump. Thus, for example,

\[
P\{ \tilde{N}(h) - \tilde{N}(0) \geq 1 \mid X(0) \} = \begin{cases} 1 & \text{if } h > 1 \\ 0 & \text{else} \end{cases},
\]

showing this counting process does not satisfy the equations (4.14).

\[
\square
\]

### 4.3 Exercises

1. Recall that for a renewal process if \( P\{Y_n < \infty\} < 1 \), then the process is called **defective**. Suppose that \( P\{Y_0 < \infty\} = 1 \) and argue why in the defective case \( \tilde{N}(t) \) is bounded, with the bound given by a geometric random variable with a parameter of \( \defP = 1 - P\{Y_n < \infty\} \). Give the distribution precisely in the event that the process is a pure renewal process.

2. Show that for a renewal process

\[
S_{N(t)-1} \leq t < S_{N(t)}
\]

so long as \( N(t) \geq 1 \). Hint: draw a picture.

3. Let \( E \) be an exponential random variable with parameter 1. For \( \lambda > 0 \), show that \( E/\lambda \) is an exponential random variable with parameter \( \lambda \). That is, if \( E^\lambda \) is an exponential random variable with parameter \( \lambda > 0 \), show that \( E^\lambda \overset{D}{=} E/\lambda \), where \( E \) is a unit exponential random variable.

4. Verify that \( 1_{X_n} \), defined in (4.8), satisfies the three properties that make it a measure.

5. Let \( N(t) \) be a one-dimensional homogeneous Poisson process with rate \( \lambda > 0 \) and points \( S_n \) (i.e. jumps happen at the times \( S_n \)). We are assuming \( N(0) = 0 \). Suppose I tell you that \( \tilde{N}(t) = 1 \) for some \( t > 0 \). Find the conditional distribution of the time \( S_0 \). That is, find the distribution of the time of the first jump, \( S_0 \), conditioned upon knowing \( \tilde{N}(t) = 1 \). Hint: the distribution has a name, give it. Note how the answer depends upon \( \lambda \).
6. Let $N(t)$ be a one-dimensional homogeneous Poisson process with rate $\lambda > 0$ and points $S_n$ (i.e. jumps happen at the times $S_n$). We are assuming $N(0) = 0$. Suppose that I tell you $N(t) = n$ for some $t > 0$ and $n \geq 1$. Find the conditional distribution of the time $S_0$. That is, find the distribution of the time of the *first* jump conditioned upon knowing $N(t) = n$ for some $n \geq 1$. Note how the answer depends upon $\lambda$.

7. Let $N$ be a unit-rate Poisson process with associated points $S_n, n \geq 0$. That is, $N([0,t]) \overset{\text{def}}{=} \sum_{n=0}^{\infty} 1_{\{S_n \leq t\}}$, and $S_n - S_{n-1}$ are independent unit exponentials. Let $T : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}_{\geq 0}^2$ be defined via $T(x) = (x, x^2)$. Describe the resulting counting process $\tilde{N}$ when the points $S_n$ are transformed by $T$.

8. Let $N$ be a non-homogeneous Poisson process with local intensity $\lambda(t) = t^2$. Write a Matlab code that simulates this process until 500 jumps have taken place. Next, write a script that simulates this process up to a time of 15. You are required to turn in your script, and three plots from each.

9. Students at a boarding school can be in one of three states: sad, neutral, or happy. If they are sad, they do not want to make many phone calls to each other. If they are neutral, they make some phone calls to each other, and if they are happy, they tend to make lots of phone calls to each other. Suppose that the state of the student body changes each day according to a discrete time Markov chain with state 1 being sad, state 2 being neutral, and state 3 being happy, and that the transition matrix is given by

$$P = \begin{bmatrix} .1 & .8 & .1 \\ .3 & .1 & .6 \\ .1 & .4 & .5 \end{bmatrix}.$$

Let $X(t) \in \{1, 2, 3\}$ denote the state of this Markov chain at time $t$, noting that it is a step function and constant each day. Now suppose we believe that the number of calls to the local cell phone tower can be modeled as a time non-homogeneous Poisson process with local intensity

$$\lambda(t) = \lambda(X(t)) = \begin{cases} 10, & \text{if } X(t) = 1 \\ 33, & \text{if } X(t) = 2 \\ 56, & \text{if } X(t) = 3 \end{cases},$$

where the units of $t$ are days. Assuming that the Markov chain starts day 1 in state 1, approximate the probability that the cell tower receives more calls on day two than day one and that it receives more calls on day 3 than day 2. Solve this problem by simulating the model $n = 10^2, 10^3,$ and $10^4$ times and averaging (that is, give three different answer based upon the different choices.
of \( n \). Note that you will be applying to law of large numbers to conclude that these values constitute an estimate for the desired probability.

Next, answer the same question for 4 days (that is, probability of increasing number of calls for first 4 days), and then answer the same question for 5 days.

10. Suppose that \( X(t) \) is a stochastic process and that \( \tilde{N} \) satisfies (4.15), where \( N \) is a unit-rate Poisson process independent from \( X \). Prove that

\[
P\{\tilde{N}(t + h) - \tilde{N}(t) \geq 2 \mid \mathcal{F}_{X(t)}\} = o(h), \quad \text{as } h \to 0.
\]

where \( \mathcal{F}_{X(t)} \) represents all the information pertaining to the history of \( X \) up through time \( t \).
Chapter 5

Continuous Time Markov Chains

In Chapter 2, we considered stochastic processes that were discrete in both time and space, and that satisfied the Markov property: the distribution of the future of the process depends only upon the current state of the process. Here we generalize such models by allowing for time to be continuous. As before, we will always take our state space to be either finite or countably infinite.

A good mental image to have when first encountering continuous time Markov chains is simply a discrete time Markov chain in which transitions can happen at any time. We will see in the next section that this image is a very good one, and that the Markov property will imply that the jump times, as opposed to simply being integers as in the discrete time setting, will be exponentially distributed.

5.1 Construction and Basic Definitions

We wish to construct a continuous time process on some countable state space $S$ that satisfies the Markov property. That is, letting $\mathcal{F}_{X(s)}$ denote all the information pertaining to the history of $X$ up to time $s$, and letting $j \in S$ and $s \leq t$, we want

$$P\{X(t) = j \mid \mathcal{F}_{X(s)}\} = P\{X(t) = j \mid X(s)\}. \tag{5.1}$$

We also want the process to be time-homogeneous so that

$$P\{X(t) = j \mid X(s) = k\} = P\{X(t - s) = j \mid X(0) = k\}. \tag{5.2}$$

We will call any process satisfying (5.1) and (5.2) a time-homogeneous continuous time Markov chain, though a more useful equivalent definition in terms of transition rates will be given in Definition 5.1.3 below. Property (5.1) should be compared with the discrete time analog (2.4). As we did for the Poisson process, which we shall see is the simplest (and most important) continuous time Markov chain, we will attempt to understand such processes in more than one way.

Before proceeding, we make the technical assumption that the processes under consideration are right-continuous. This implies that if a transition occurs “at time $t$”, then we take $X(t)$ to be the new state and note that $X(t) \neq X(t-)$. 
Example 5.1.1. Consider a two state continuous time Markov chain. We denote the states by 1 and 2, and assume there can only be transitions between the two states (i.e. we do not allow $1 \rightarrow 1$). Graphically, we have

$$1 \iff 2.$$  

Note that if we were to model the dynamics via a discrete time Markov chain, the transition matrix would simply be

$$P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$  

and the dynamics are quite trivial: the process begins in state 1 or 2, depending upon the initial distribution, and then deterministically transitions between the two states. At this point, we do not know how to understand the dynamics in the continuous time setting. All we know is that the distribution of the process should only depend upon the current state, and not the history. This does not yet tell us when the firings will occur. □

Motivated by Example 5.1.1, our first question pertaining to continuous time Markov chains, and one whose answer will eventually lead to a general construction/simulation method, is: how long will this process remain in a given state, say $x \in S$? Explicitly, suppose $X(0) = x$ and let $T_x$ denote the time we transition away from state $x$. To find the distribution of $T_x$, we let $s, t \geq 0$ and consider

$$P\{T_x > s + t \mid T_x > s\}$$

$$= P\{X(r) = x \text{ for } r \in [0, s + t] \mid X(r) = x \text{ for } r \in [0, s]\}$$

$$= \frac{P\{X(r) = x \text{ for } r \in [0, s + t], X(r) = x \text{ for } r \in [0, s]\}}{P\{X(r) = x \text{ for } r \in [0, s]\}}$$

(conditional probability)

$$= \frac{P\{X(r) = x \text{ for } r \in [s, s + t], X(r) = x \text{ for } r \in [0, s]\}}{P\{X(r) = x \text{ for } r \in [0, s]\}}$$

(set equivalence)

$$= P\{X(r) = x \text{ for } r \in [s, s + t] \mid X(r) = x \text{ for } r \in [0, s]\}$$

(conditional probability)

$$= P\{X(r) = x \text{ for } r \in [0, t] \mid X(0) = x\}$$

(Markov property)

$$= P\{T_x > t\}.$$

Therefore, $T_x$ satisfies the loss of memory property, and is therefore exponentially distributed (since the exponential random variable is the only continuous random variable with this property). We denote the parameter of the exponential holding time for state $x$ as $\lambda(x)$. We make the useful observation that

$$\mathbb{E}T_x = \frac{1}{\lambda(x)}.$$  

126
Thus, the higher the rate $\lambda(x)$, representing the rate out of state $x$, the smaller the expected time for the transition to occur, which is intuitively pleasing.

**Example 5.1.2.** We return to Example 5.1.1, though now we assume the rate from state 1 to state 2 is $\lambda(1) > 0$, and the rate from state 2 to state 1 is $\lambda(2) > 0$. We commonly incorporate these parameters into the model by placing them next to the transition arrow in the graph:

$$
\begin{align*}
1 & \xrightarrow{\lambda(1)} 2, \\
\end{align*}
$$

The dynamics of the model are now clear. Assuming $X(0) = 1$, the process will remain in state 1 for an exponentially distributed amount of time, with parameter $\lambda(1)$, at which point it will transition to state 2, where it will remain for an exponentially distributed amount of time, with parameter $\lambda(2)$. This process then continuous indefinitely.

Example 5.1.2 is deceptively simple as it is clear that when the process transitions out of state 1, it must go to state 2, and vice versa. However, consider the process with states 1, 2, and 3 satisfying

$$
1 \xrightarrow{\lambda(1)} 2 \xrightarrow{\lambda(2)} 3.
$$

Even if you are told the holding time parameter for state 2, without further information you can not figure out which you transition to state 1 or state 3 after you leave state 2. Therefore, we see we want to study the transition probabilities associated with the process, which we do now.

Still letting $T_x$ denote the amount of time the process stays in state $x$ after entering state $x$, and which we now know is exponentially distributed with a parameter of $\lambda(x)$, we define for $y \neq x$

$$
p_{xy} \overset{\text{def}}{=} P\{X(T_x) = y \mid X(0) = x\},
$$

to be the probability that the process transitions to state $y$ after leaving state $x$. It can be shown that the time of the transition, $T_x$, and the value of the new state, $y$, are independent random variables. Loosely, this follows since if the amount of time the chain stays in state $x$ affects the transition probabilities, then the Markov property (5.1) is not satisfied as we would require to know both the current state and the amount of time the chain has been there to know the probabilities associated with ending up in the different states.

We next define

$$
\lambda(x, y) \overset{\text{def}}{=} \lambda(x)p_{xy}.
$$

Since $T_x$ is exponential with parameter $\lambda(x)$, we have that

$$
P\{T_x < h\} = 1 - e^{-\lambda(x)h} = \lambda(x)h + o(h), \text{ as } h \to 0.
$$
Combining the above, for \( y \neq x \) and mild assumptions on the function \( \lambda \), we have

\[
P\{X(h) = y \mid X(0) = x\} = P\{T_x < h, X(T_x) = y \mid X(0) = x\} + o(h)
\]

\[
= \lambda(x)hp_{xy} + o(h)
\]

\[
= \lambda(x,y)h + o(h),
\]

as \( h \to 0 \), where the \( o(h) \) in the first equality represents the probability of seeing two or more jumps (each with an exponential distribution) in the time window \([0, h]\).

Therefore, \( \lambda(x,y) \) yields the \textit{local rate}, or intensity, of transitioning from state \( x \) to state \( y \). It is worth explicitly pointing out that for \( x \in S \)

\[
\sum_{y \neq x} \lambda(x,y) = \sum_{y \neq x} \lambda(x)p_{xy} = \lambda(x).
\]

Note that we also have

\[
P\{X(h) = x \mid X(0) = x\} = 1 - \sum_{y \neq x} P\{X(h) = y \mid X(0) = x\}
\]

\[
= 1 - \sum_{y \neq x} \lambda(x,y)h + o(h)
\]

\[
= 1 - \lambda(x)h \sum_{y \neq x} p_{xy} + o(h)
\]

\[
= 1 - \lambda(x)h + o(h).
\]

Similarly to our consideration of the Poisson process, it can be argued that any time homogeneous process satisfying the local conditions (5.3) and (5.4) also satisfies the Markov property (5.1). This is not surprising as the conditions (5.3)-(5.4) only make use of the current state of the system and ignore the entire past. This leads to a formal definition of a continuous time Markov chain that incorporates all the relevant parameters of the model and is probably the most common definition in the literature.

**Definition 5.1.3.** A time-homogeneous \textit{continuous time Markov chain} with transition rates \( \lambda(x,y) \) is a stochastic process \( X(t) \) taking values in a finite or countably infinite state space \( S \) satisfying

\[
P\{X(t+h) = x \mid X(t) = x\} = 1 - \lambda(x)h + o(h)
\]

\[
P\{X(t+h) = y \mid X(t) = x\} = \lambda(x,y)h + o(h),
\]

where \( y \neq x \), and \( \lambda(x) = \sum_{y \neq x} \lambda(x,y) \).

When only the local rates \( \lambda(x,y) \) are given in the construction of the chain, then it is important to recognize that the transition probabilities of the chain can be recovered via the identity

\[
p_{xy} = \frac{\lambda(x,y)}{\lambda(x)} = \frac{\lambda(x,y)}{\sum_{y \neq x} \lambda(x,y)}.
\]

\[\footnote{For example, we do not want to let \( \lambda(z) = \infty \) for any \( z \in E \)}\]
Example 5.1.4. Let $N$ be a Poisson process with intensity $\lambda > 0$. As $N$ satisfies
\[
P\{N(t + h) = j + 1 \mid N(t) = j\} = \lambda h + o(h)
\]
\[
P\{N(t + h) = j \mid N(t) = j\} = 1 - \lambda h + o(h),
\]
we see that it is a continuous time Markov chain. Note also that the Poisson process is the continuous time version of the deterministically monotone chain from Chapter 2.

\[\square\]

Example 5.1.5. Consider again the three state Markov chain
\[
\begin{array}{ccc}
1 & \xrightleftharpoons{\lambda(1,2)} & 2 \\
\xrightleftharpoons{\lambda(2,1)} & \lambda(2,3) & \xrightleftharpoons{\lambda(3,2)} & 3,
\end{array}
\]
where the local transition rates have been placed next to their respective arrows. Note that the holding time in state two is an exponential random variable with a parameter of
\[
\lambda(2) \overset{\text{def}}{=} \lambda(2, 1) + \lambda(2, 3),
\]
and the probability that the chain enters state 1 after leaving state 2 is
\[
p_{21} \overset{\text{def}}{=} \frac{\lambda(2, 1)}{\lambda(2, 1) + \lambda(2, 3)},
\]
whereas the probability that the chain enters state 3 after leaving state 2 is
\[
p_{23} \overset{\text{def}}{=} \frac{\lambda(2, 3)}{\lambda(2, 1) + \lambda(2, 3)}.
\]
This chain could then be simulated by sequentially computing holding times and transitions.

\[\square\]

An algorithmic construction of a general continuous time Markov chain should now be apparent, and will involve two building blocks. The first will be a stream of unit exponential random variables used to construct our holding times, and the second will be a discrete time Markov chain, denoted $X_n$, with transition probabilities $p_{xy}$ that will be used to determine the sequence of states. Note that for this discrete time chain we necessarily have that $p_{xx} = 0$ for each $x$. We also explicitly note that the discrete time chain, $X_n$, is different than the continuous time Markov chain, $X(t)$, and the reader should be certain to clarify this distinction. The discrete time chain is often called the \textit{embedded} chain associated with the process $X(t)$.

Algorithm 2. (Algorithmic construction of continuous time Markov chain)
Input:

- Let $X_n$, $n \geq 0$, be a discrete time Markov chain with transition matrix $Q$. Let the initial distribution of this chain be denoted by $\alpha$ so that $P\{X_0 = k\} = \alpha_k$.
- Let $E_n$, $n \geq 0$, be a sequence of independent unit exponential random variables.
Algorithmic construction:

1. Select $X(0) = X_0$ according to the initial distribution $\alpha$.

2. Let $T_0 = 0$ and define $W(0) = E_0/\lambda(X(0))$, which is exponential with parameter $\lambda(X(0))$, to be the waiting time in state $X(0)$.

3. Let $T_1 = T_0 + W(0)$, and define $X(t) = X(0)$ for all $t \in [T_0, T_1)$.

4. Let $X_1$ be chosen according to the transition matrix $Q$, and define $W(1) = E_1/\lambda(X_1)$.

5. Let $T_2 = T_1 + W(1)$ and define $X(t) = X_1$ for all $t \in [T_1, T_2)$.

6. Continue process.

Note that two random variables will be needed at each iteration of Algorithm 2, one to compute the holding time, and one to compute the next state of the discrete time Markov chain. In the biology/chemistry context, the algorithm implicit in the above construction is typically called the Gillespie algorithm, after Dan Gillespie. However, it (and its natural variants) is also called, depending on the field, the stochastic simulation algorithm, kinetic Monte Carlo, dynamic Monte Carlo, the residence-time algorithm, the n-fold way, or the Bortz-Kalos-Liebowitz algorithm; needless to say, this algorithm has been discovered many times and plays a critical role in many branches of science.

As the future of the process constructed in Algorithm 2 only depends upon the current state of the system, and the current holding time is exponentially distributed, it satisfies the Markov property (5.1). Further, for $y \neq x$ we have

$$P\{X(h) = y \mid X(0) = x\} = P\{X(T_1) = y, T_1 \leq h \mid X(0) = h\} + o(h)$$

$$= \lambda(x)hp_{xy} + o(h)$$

$$= \lambda(x,y)h,$$

showing we also get the correct local intensities. Therefore, the above construction via a stream of exponentials and an embedded discrete time Markov chain could be taken to be an alternative definition of a continuous time Markov chain.

One useful way to think about the construction in Algorithm 2 is in terms of alarm clocks:

1. When the chain enters state $x$, independent “alarm clocks” are placed at each state $y$, and the $y$th is programed to go off after an exponentially distributed amount of time with parameter $\lambda(x,y)$.

2. When the first alarm goes off, the chain moves to that state, all alarm clock are discarded, and we repeat the process.
Note that to prove that this algorithm is, in fact, equivalent to the algorithmic construction above, you need to recall that the minimum of exponential random variables with parameters \( \lambda(x, y) \) is itself exponentially distributed with parameter

\[
\lambda(x) = \sum_y \lambda(x, y),
\]

and that it is the \( y \)th that went off with probability

\[
\frac{\lambda(x, y)}{\sum_{j \neq x} \lambda(x, j)} = \frac{\lambda(x, y)}{\lambda(x)}.
\]

See Propositions B.3.18 and B.3.19.

We close this section with three examples.

**Example 5.1.6.** We consider again a random walker on \( S = \{0, 1, \ldots \} \). We suppose the transition intensities are

\[
\lambda(i, i + 1) = \lambda \\
\lambda(i, i - 1) = \mu, \quad \text{if } i > 0,
\]

and \( \lambda(0, -1) = 0 \). Therefore, the probability of the embedded discrete time Markov chain transitioning up if the current state is \( i \neq 0 \), is \( \lambda/(\lambda + \mu) \), whereas the probability of transitioning down is \( \mu/(\lambda + \mu) \). When \( i \neq 0 \), the holding times will always be exponentially distributed with a parameter of \( \lambda + \mu \).

**Example 5.1.7.** We generalize Example 5.1.6 by allowing the transition rates to depend upon the current state of the system. As in the discrete time setting this leads to a birth and death process. More explicitly, for \( i \in \{0, 1, \ldots \} \) we let

\[
\lambda(i, i + 1) = B(i) \\
\lambda(i, i - 1) = D(i),
\]

where we take \( D(0) = 0 \). Note that the transition rates are now state dependent, and may even be unbounded as \( i \to \infty \). Common choices for the rates include

\[
B(i) = \lambda i \\
D(i) = \mu i,
\]

for some scalar \( \lambda, \mu > 0 \). Another common model would be to assume a population satisfies a logistical growth model,

\[
B(i) = ri \\
D(i) = \frac{r}{K} i^2.
\]

where \( K \) is the carrying capacity.
Analogously to Example 4.2.19, if we let $X(t)$ denote the state of the system at time $t$, we have that $X(t)$ solves the stochastic equation

$$X(t) = X(0) + Y_1 \left( \int_0^t B(X(s))ds \right) - Y_2 \left( \int_0^t D(X(s))ds \right), \quad (5.5)$$

where $Y_1$ and $Y_2$ are independent unit-rate Poisson processes. As in Example 4.2.19, it is now an exercise to show that the solution to (5.5) satisfies the correct local intensity relations of Definition 5.1.3. For example, denoting

$$A(t) \overset{\text{def}}{=} Y_1 \left( \int_0^t B(X(s))ds \right), \quad D(t) \overset{\text{def}}{=} Y_2 \left( \int_0^t D(X(s))ds \right),$$

we see that

$$P\{X(t + h) = x + 1 \mid X(t) = x\}$$

$$= P\{A(t + h) - A(t) = 1, D(t + h) - D(t) = 0 \mid X(t) = x\} + o(h)$$

$$= B(x)h(1 - D(x)h) + o(h)$$

$$= B(x)h + o(h).$$

\[\square\]

5.2 Explosions

Now that we have a good idea of what a continuous time Markov chain is, we demonstrate a behavior that is not possible in the discrete time setting: explosions. Recall that in Algorithm 2, which constructs a continuous time Markov chain, the value $T_n$ represents the time of the $n$th transition of the chain. Therefore, the chain so constructed is only defined up until the (random) time

$$T_\infty \overset{\text{def}}{=} \lim_{n \to \infty} T_n.$$

If $T_\infty < \infty$, then we say that an explosion has happened.

**Definition 5.2.1.** If

$$P_i\{T_\infty = \infty\} \overset{\text{def}}{=} P\{T_\infty = \infty \mid X(0) = i\} = 1, \quad \text{for all } i \in S,$$

then we will say the process is non-explosive. Otherwise we will say the process is explosive.

Note that a process could be explosive even if

$$P_i\{T_\infty = \infty\} = 1,$$

for some $i \in S$; see Example 5.2.4. It is not too difficult to construct an explosive process. To do so, we will first need the following result pertaining to exponential random variables.
Proposition 5.2.2. Suppose that \( \{E_n\} \), \( n \geq 1 \), are independent exponential random variables with respective parameters \( \lambda_n \). Then,

\[
P \left\{ \sum_{n=1}^{\infty} E_n < \infty \right\} = 1 \iff \sum_{n=1}^{\infty} \frac{1}{\lambda_n} < \infty.
\]

Proof. We will prove one direction of the implication (the one we will use). For the other direction, see [37, Section 5.1]. We suppose that \( \sum_n \frac{1}{\lambda_n} < \infty \). Because \( \sum_n E_n \geq 0 \) and

\[
\mathbb{E} \left( \sum_n E_n \right) = \sum_n \mathbb{E} E_n = \sum_n \frac{1}{\lambda_n} < \infty,
\]

we may conclude that \( \sum_n E_n < \infty \) with probability one. \( \square \)

Thus, we see that we can construct an explosive birth process by requiring that the holding times satisfy \( \sum_n 1/\lambda(X_n) < \infty \).

Example 5.2.3. Consider a pure birth process in which the embedded discrete time Markov chain is the deterministically monotone chain of Example 2.1.6. Suppose that the holding time parameter in state \( i \) is \( \lambda(i) \). Finally, let \( X(t) \) denote the state of the continuous time process at time \( t \). Note that the stochastic equation satisfied by \( X \) is

\[
X(t) = X(0) + N \left( \int_0^t \lambda(X(s))ds \right).
\]

Suppose that \( \lambda(n) = \lambda n^2 \) for some \( \lambda > 0 \) and that \( X(0) = 1 \). Then the \( n \)th holding time is determined by an exponential random variable with parameter \( \lambda n^2 \), which we denote by \( E_n \). Since

\[
\sum_n \frac{1}{\lambda n^2} < \infty,
\]

we may conclude by Proposition 5.2.2 that

\[
P \left\{ \sum_n E_n < \infty \right\} = 1,
\]

and the process is explosive. The stochastic equation for this model is

\[
X(t) = X(0) + N \left( \lambda \int_0^t X(s)^2 ds \right),
\]

and should be compared with the deterministic ordinary differential equation

\[
x'(t) = \lambda x^2(t) \iff x(t) = x(0) + \lambda \int_0^t x(s)^2 ds,
\]

which also explodes in finite time. \( \square \)
Example 5.2.4. Consider a continuous time Markov chain with state space \{-2, -1, 0, 1, 2, \ldots \}. We suppose that the graph of the model is 

\[ -2 \xrightarrow{1} -1 \xleftarrow{2} 0 \xrightarrow{1} 1 \xrightarrow{2^2} 3 \xrightarrow{3^2} \cdots , \]

where, in general, the intensity of \( n \to n + 1 \), for \( n \geq 1 \) is \( \lambda(n) = n^2 \). From the previous example, we know this process is explosive. However, if \( X(0) \in \{-2, -1\} \), then the probability of explosion is zero\(^2\), whereas if \( X(0) = 0 \), the probability of explosion is 1/3.

The following proposition characterizes the most common ways in which a process is non-explosive. A full proof can be found in \([37]\).

**Proposition 5.2.5.** For any \( i \in S \),

\[
P_i\{T_\infty < \infty\} = P_i\left\{ \sum_n \frac{1}{\lambda(X_n)} < \infty \right\},
\]

and therefore, the continuous time Markov chain is non-explosive iff

\[
\sum_n \frac{1}{\lambda(X_n)} = \infty
\]

\( P_i \)- almost surely for every \( i \in S \). In particular,

1. If \( \lambda(i) \leq c \) for all \( i \in S \) for some \( c > 0 \), then the chain is non-explosive.
2. If \( S \) is a finite set, then the chain is non-explosive.
3. If \( T \subset S \) are the transient states of \( \{X_n\} \) and if

\[
P_i\{X_n \in T, \forall n\} = 0,
\]

for every \( i \in S \), then the chain is non-explosive.

**Proof.** The equivalence of the probabilities is shown in \([37, Section 5.2]\). Will prove the results 1,2,3. For (1), simply note that

\[
\sum_n \frac{1}{\lambda(X(n))} \geq \sum_n \frac{1}{c} = \infty.
\]

To show (2), we note that if the state space is finite, we may simply take \( c = \max\{\lambda_i\} \), and apply (1).

We will now show (3). If \( P_i\{X_n \in T, \forall n\} = 0 \), then entry into \( T^c \) is assured. There must, therefore, be a state \( i \in T^c \), which is hit infinitely often (note that this value can be different for different realizations of the process). Let the infinite sequence of times when \( X_n = i \) be denoted by \( \{n_j\} \). Then,

\[
\sum_n \frac{1}{\lambda(X_n)} \geq \sum_j \frac{1}{\lambda(X_{n_j})} = \sum_j \frac{1}{\lambda(i)} = \infty.
\]

\( \square \)

\(^2\)This is proven by the next proposition, but it should be clear
We will henceforth have a running assumption that unless otherwise explicitly stated, all processes considered are non-explosive. However, we will return to explosiveness later and prove another useful condition that implies a process is non-explosive. This condition will essentially be a linearity condition on the intensities, and is sufficient to prove the non-explosiveness of most processes in the queueing literature. Unfortunately, the world of biology is not so easy and most processes of interest are highly non-linear and it is, in general, quite a difficult (and open) problem to characterize which systems are non-explosive.

5.3 Forward Equation, Backward Equation, and the Generator Matrix

We note that in each of the constructions of a continuous time Markov chain, we are given only the local behavior of the model, such as transition probabilities and holding times. Similarly to when we studied the Poisson process, the question now becomes: how do these local behaviors determine the global behavior? In particular, how can we find terms of the form

\[ P_{ij}(t) = P\{X(t) = j \mid X(0) = i\}, \]

for \( i, j \in S \), the state space, and \( t \geq 0 \)?

We begin to answer this question by first deriving the Kolmogorov forward equations, which are a system of ordinary differential equations governing the behaviors of the probabilities \( P_{ij}(t) \). We note that the forward equations are only valid if the process is non-explosive as we will derive them by conditioning on the state of the system “directly before” our time of interest. If that time is \( T_\infty < \infty \), then this question does not really make sense, for what is the last jump before \( T_\infty \)?

Proceeding, we have

\[
P'_{ij}(t) = \lim_{h \to 0} \frac{P_{ij}(t + h) - P_{ij}(t)}{h} \\
= \lim_{h \to 0} \frac{1}{h} \left( P\{X(t + h) = j \mid X(0) = i\} - P\{X(t) = j \mid X(0) = i\} \right) \\
= \lim_{h \to 0} \frac{1}{h} \left( \sum_{y \in S} P\{X(t + h) = j \mid X(t) = y, X(0) = i\} P\{X(t) = y \mid X(0) = i\} \\
- P\{X(t) = j \mid X(0) = i\} \right).
\]
However,

\[
\sum_{y \in S} P\{X(t + h) = j \mid X(t) = y, X(0) = i\} P\{X(t) = y \mid X(0) = i\} = P\{X(t) = j \mid X(0) = i\}
\]

\[
+ \sum_{y \neq j} P\{X(t + h) = j \mid X(t) = y, X(0) = i\} P\{X(t) = y \mid X(0) = i\} \quad (5.6)
\]

\[
= (1 - \lambda(j)h)P_{ij}(t) + \sum_{y \neq j} \lambda(y, j)hP_{iy}(t) + o(h), \quad (5.7)
\]

and so

\[
P'_{ij}(t) = \lim_{h \to 0} \frac{1}{h} \left( (1 - \lambda(j)h - 1)P_{ij}(t) + \sum_{y \neq j} \lambda(y, j)P_{iy}(t)h + o(h) \right)
\]

\[
= -\lambda(j)P_{ij}(t) + \sum_{y \neq j} \lambda(y, j)P_{iy}(t). \quad (5.8)
\]

Thus,

\[
P'_{ij}(t) = -\lambda(j)P_{ij}(t) + \sum_{y \neq j} P_{iy}(t)\lambda(y, j). \quad (5.8)
\]

These are the Kolmogorov forward equations for the process. In the biology literature this system of equations is termed the chemical master equation.

We point out that there was a small mathematical “slight of hand” in the above calculation. To move from (5.6) to (5.7), we had to assume that

\[
\sum_{y} P_{iy}(t)\sigma_{y}(h) = o(h),
\]

where we write \(\sigma_{y}(h)\) to show that the size of the error can depend upon the state \(y\). This condition is satisfied for all systems we will consider.

**Definition 5.3.1.** Let \(X(t)\) be a continuous time Markov chain on some state space \(S\) with transition intensities \(\lambda(i, j) \geq 0\). Recalling that

\[
\lambda(i) = \sum_{j \neq i} \lambda(i, j),
\]

The matrix

\[
A_{ij} = \begin{cases} 
-\lambda(i), & \text{if } i = j \\
\lambda(i, j), & \text{if } i \neq j
\end{cases} = \begin{cases} 
-\sum_{j} \lambda(i, j), & \text{if } i = j \\
\lambda(i, j), & \text{if } i \neq j
\end{cases}
\]

is called the generator, or infinitesimal generator, or generator matrix of the Markov chain.
We see that the Kolmogorov forward equations (5.8) can be written as the matrix differential equation
\[ P'(t) = P(t)A, \]
since
\[
(P(t)A)_{ij} = \sum_y P_{iy}(t)A_{yj} = P_{ij}A_{jj} + \sum_{y \neq j} P_{iy}A_{yj} \\
= -\lambda(j)P_{ij}(t) + \sum_{y \neq j} P_{iy}\lambda(y,j).
\]

At least formally, this system can be solved
\[ P(t) = P(0)e^{tA} = e^{tA}, \]
where \( e^{tA} \) is the matrix exponential and we used that \( P(0) = I \), the identity matrix. Recall that the matrix exponential is defined by
\[ e^{At} \overset{\text{def}}{=} \sum_{k=0}^{\infty} \frac{t^n A^n}{n!}. \]

This solution is always valid in the case that the state space is finite.

We make the following observations pertaining to the generator \( A \):

1. The elements on the main diagonal are all strictly negative.
2. The elements off the main diagonal are non-negative.
3. Each row sums to zero.

We also point out that given a state space \( S \), the infinitesimal generator \( A \) completely determines the Markov chain as it contains all the local information pertaining to the transitions: \( \lambda(i,j) \). Thus, it is sufficient to characterize a chain by simply providing a state space, \( S \), and generator, \( A \).

**Example 5.3.2.** A molecule transitions between states 0 and 1. The transition rates are \( \lambda(0,1) = 3 \) and \( \lambda(1,0) = 1 \). The generator matrix is
\[ A = \begin{bmatrix} -3 & 3 \\ 1 & -1 \end{bmatrix}. \]

**Example 5.3.3.** Consider a mathematician wandering between three coffee shops with graphical structure
\[ A \overset{\mu_1}{\underset{\lambda_1}{\leftrightarrow}} B \overset{\mu_2}{\underset{\lambda_2}{\leftrightarrow}} C. \]
The infinitesimal generator of this process is

\[ A = \begin{bmatrix} -\mu_1 & \mu_1 & 0 \\ \lambda_1 & -(\lambda_1 + \mu_2) & \mu_2 \\ 0 & \lambda_2 & -\lambda_2 \end{bmatrix}, \]

and the transition matrix for the embedded Markov chain is

\[ P = \begin{bmatrix} 0 & 1 & 0 \\ \frac{\lambda_1}{\lambda_1 + \mu_2} & 0 & \frac{\mu_2}{\lambda_1 + \mu_2} \\ 0 & 1 & 0 \end{bmatrix}. \]

Example 5.3.4. For a unit-rate Poisson process, we have

\[ A = \begin{bmatrix} -1 & 1 & 0 & \ldots \\ 0 & -1 & 1 & 0 & \ldots \\ 0 & 0 & -1 & 1 & \ddots \\ \vdots & \vdots & \ddots & \ddots \end{bmatrix}. \]

If we are given an initial distribution of \( \alpha \) and want the probability of being in state \( j \) at time \( t \), then we simply note

\[ P_\alpha\{X(t) = j\} = \sum_i P\{X(t) = j \mid X(0) = i\}P\{X(0) = i\} = \sum_i \alpha_i P_{ij} = (\alpha P(t))_j. \]

Thus,

\[ \alpha P(t) = P_\alpha(t) = \alpha e^{tA}. \quad (5.9) \]

Backward equation

Before attempting to solve a system using Kolmogorov’s forward equations, we introduce another set of equations, called Kolmogorov’s backward equations, which are valid for all continuous time Markov chains. The derivation below follows that of [37].

We begin by finding an integral equation satisfied by \( P_{ij}(t) \). We will then differentiate it to get the backward equations.

**Proposition 5.3.5.** For all \( i, j \in S \) and \( t \geq 0 \), we have

\[ P_{ij}(t) = \delta_{ij}e^{-\lambda(i)t} + \int_0^t \lambda(i)e^{-\lambda(i)s} \sum_{k \neq i} Q_{ik} P_{kj}(t-s) ds, \]

where, as usual,

\[ \delta_{ij} = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{if } i \neq j \end{cases} \]

is the Kronecker delta function, and \( Q \) is the transition matrix of the embedded discrete time Markov chain.
Proof. Conditioning on the first jump time of the chain, $T_1$, we have

$$P\{X(t) = j \mid X(0) = i\} = P\{X(t) = j, T_1 > t \mid X(0) = i\} + P\{X(t) = j, T_1 \leq t \mid X(0) = i\}. $$

We handle these terms separately. For the first term on the right hand side of the above equation, a first transition has not been made. Thus, $X(t) = j$ iff $j = i$ and does so with a probability of one. That is,

$$P\{X(t) = j, T_1 > t \mid X(0) = i\} = \delta_{ij} P\{T_1 > t\} = \delta_{ij} e^{-\lambda(i)t}.$$

For the second term, we will condition on the time of the first jump happening in $(s, s + \Delta)$, for small $\Delta$ (we will eventually take $\Delta \to 0$). As the holding time is exponential with parameter $\lambda(i)$, this event has probability

$$\int_s^{s+\Delta} \lambda(i) e^{-\lambda(i)r} dr = \lambda(i) e^{-\lambda(i)s} \Delta + O(\Delta^2).$$

We let $s_n = nt/N$ for some large $N$, denote $\Delta = t/N$, and see

$$P\{X(t) = j, T_1 \leq t \mid X(0) = i\} = \sum_{n=0}^{N-1} P\{X(t) = j, T_1 \in (s_n, s_{n+1}) \mid X(0) = i\}$$

$$= \sum_{n=0}^{N-1} P\{X(t) = j \mid X(0) = i, T_1 \in (s_n, s_{n+1})\} P\{T_1 \in (s_n, s_{n+1}) \mid X(0) = i\}$$

$$= \sum_{n=0}^{N-1} \lambda(i) e^{-\lambda(i)s_n} \sum_{k \neq i} P\{X(t) = j, X_1 = k \mid X(0) = i, T_1 \in (s_n, s_{n+1})\} \Delta + O(\Delta)$$

$$= \sum_{n=0}^{N-1} \lambda(i) e^{-\lambda(i)s_n} \sum_{k \neq i} \left[ P\{X(t) = j \mid X_1 = k, X(0) = i, T_1 \in (s_n, s_{n+1})\} \right] \Delta + O(\Delta)$$

$$\approx \sum_{n=0}^{N-1} \lambda(i) e^{-\lambda(i)s_n} \sum_{k \neq i} Q_{ik} P_{kj}(t - s_n) \Delta + O(\Delta)$$

$$\to \int_0^t \lambda(i) e^{-\lambda(i)s} \sum_{k \neq i} Q_{ik} P_{kj}(t - s) ds,$$

as $\Delta \to 0$. Combining the above shows the result. \qed
Proposition 5.3.6. For all \(i, j \in S\), we have that \(P_{ij}(t)\) is continuously differentiable and
\[
P'(t) = AP(t), \tag{5.10}
\]
which in component form is
\[
P'_{ij}(t) = \sum_k A_{ik} P_{kj}(t).
\]

The system of equations (5.10) is called the Kolmogorov backwards equations. Note that the difference with the forward equations is the order of the multiplication of \(P(t)\) and \(A\). However, the solution of the backwards equation is once again seen to be
\[
P(t) = e^{tA},
\]
agreeing with previous results.

Proof. Use the substitution \(u = t - s\) in the integral equation to find that
\[
P_{ij}(t) = \delta_{ij} e^{-\lambda(i)t} + \int_0^t \lambda(i) e^{-\lambda(i)s} \sum_{k \neq i} Q_{ik} P_{kj}(t - s) ds
\]
\[
= \delta_{ij} e^{-\lambda(i)t} + \int_0^t \lambda(i) e^{-\lambda(i)(t-u)} \sum_{k \neq i} Q_{ik} P_{kj}(u) ds
\]
\[
= e^{-\lambda(i)t} \left[ \delta_{ij} + \int_0^t \lambda(i) e^{\lambda(i)u} \sum_{k \neq i} Q_{ik} P_{kj}(u) ds \right].
\]
Differentiating yields
\[
P'_{ij}(t) = -\lambda(i) e^{-\lambda(i)t} \left[ \delta_{ij} + \int_0^t \lambda(i) e^{\lambda(i)u} \sum_{k \neq i} Q_{ik} P_{kj}(u) ds \right]
\]
\[
+ e^{-\lambda(i)t} \cdot \lambda(i) e^{\lambda(i)t} \sum_{k \neq i} Q_{ik} P_{kj}(t)
\]
\[
= -\lambda(i) P_{ij}(t) + \lambda(i) \sum_{k \neq i} Q_{ik} P_{kj}(t)
\]
\[
= \sum_k (-\lambda(i) \delta_{ik} P_{kj}(t)) + \sum_k \lambda(i) Q_{ik} P_{kj}(t)
\]
\[
= \sum_k (-\lambda(i) \delta_{ik} + \lambda(i) Q_{ik}) P_{kj}(t)
\]
\[
= \sum_k A_{ik} P_{kj}(t).
\]

Both the forward and backward equations can be used to solve for the associated probabilities as the next example demonstrates.
Example 5.3.7. We consider a two state, \(\{0, 1\}\), continuous time Markov chain with generator matrix

\[
A = \begin{bmatrix}
-\lambda & \lambda \\
\mu & -\mu
\end{bmatrix}.
\]

We will use both the forwards and backwards equations to solve for \(P(t)\).

**Approach 1: Backward equation.** While we want to compute \(P_{ij}(t)\) for each pair \(i, j \in \{0, 1\}\), we know that \(P_{00}(t) + P_{01}(t) = P_{10}(t) + P_{11}(t) = 1\), for all \(t \geq 0\), and so it is sufficient to solve just for \(P_{00}(t)\) and \(P_{10}(t)\).

The backwards equation is \(P'(t) = AP(t)\), yielding the equations

\[
\begin{align*}
P'_{00}(t) &= \lambda [P_{10}(t) - P_{00}(t)] \\
P'_{10}(t) &= \mu [P_{00}(t) - P_{10}(t)].
\end{align*}
\]

We see that

\[
\mu P'_{00}(t) + \lambda P'_{10}(t) = 0 \implies \mu P_{00}(t) + \lambda P_{10}(t) = c.
\]

We know that \(P(0) = I\), so we see that

\[
\mu P_{00}(0) + \lambda P_{10}(0) = c \iff \mu = c.
\]

Thus,

\[
\mu P_{00}(t) + \lambda P_{10}(t) = \mu \implies \lambda P_{10}(t) = \mu - \mu P_{00}(t).
\]

Putting this back into our differential equations above we have that

\[
P'_{00}(t) = \mu - \mu P_{00}(t) - \lambda P_{00}(t) = \mu - (\mu + \lambda)P_{00}(t).
\]

Solving, with \(P_{00}(t) = 1\) yields

\[
P_{00}(t) = \frac{\mu}{\mu + \lambda} + \frac{\lambda}{\mu + \lambda} e^{-(\mu + \lambda)t}.
\]

Of course, we also have that

\[
P_{01}(t) = 1 - P_{00}(t)
\]

\[
P_{10}(t) = \frac{\mu}{\lambda} - \frac{\mu}{\lambda} \left( \frac{\mu}{\mu + \lambda} + \frac{\lambda}{\mu + \lambda} e^{-(\mu + \lambda)t} \right) = \frac{\mu}{\mu + \lambda} - \frac{\mu}{\mu + \lambda} e^{-(\mu + \lambda)t}.
\]

**Approach 1: Forward equation.** This is easier. We want to solve

\[
P'(t) = P(t)A.
\]
We now get

\[ P_0'(t) = -P_0(t) \lambda + P_0(t) \mu = -P_0(t) \lambda + (1 - P_0(t)) \mu = \mu - (\lambda + \mu) P_0(t) \]

\[ P_1'(t) = -\lambda P_1(t) + \mu P_1(t) = -\lambda P_1(t) + \mu (1 - P_1(t)) = \mu - (\lambda + \mu) P_0(t), \]

and the solutions above follow easily.

Note that, as in the discrete time setting, we have that

\[ \lim_{t \to \infty} P(t) = \frac{1}{\lambda + \mu} \begin{bmatrix} \mu & \lambda \\ \mu & \lambda \end{bmatrix}, \]

yielding a common row vector which can be interpreted as a limiting distribution. □

There is a more straightforward way to make the above computations: simply solve the matrix exponential.

**Example 5.3.8** (Computing matrix exponentials). Suppose that \( A \) is an \( n \times n \) matrix with \( n \) distinct eigenvectors. Then, letting \( D \) be a diagonal matrix consisting of the eigenvalues of \( A \), we can decompose \( A \) into

\[ A = QDQ^{-1}, \]

where \( Q \) consists of the eigenvectors of \( A \) (ordered similarly to the order of the eigenvalues in \( D \)). In this case, we get the very nice identity

\[ e^{At} = \sum_{n=0}^{\infty} \frac{t^n (QDQ^{-1})^n}{n!} = Q \left( \sum_{n=0}^{\infty} \frac{t^n D^n}{n!} \right) Q^{-1} = Q e^{Dt} Q^{-1}, \]

where \( e^{Dt} \), because \( D \) is diagonal, is a diagonal matrix with diagonal elements \( e^{\lambda_i t} \) where \( \lambda_i \) is the \( i \)th eigenvalue.

**Example 5.3.9.** We now solve the above problem using the matrix exponential. Supposing, for concreteness, that \( \lambda = 3 \) and \( \mu = 1 \), we have that the generator matrix is

\[ A = \begin{bmatrix} -3 & 3 \\ 1 & -1 \end{bmatrix} \]

It is easy to check that the eigenvalues are 0, -4 and the associated eigenvalues are \([1, 1]^t\) and \([-3, 1]^t\). Therefore,

\[ Q = \begin{bmatrix} 1 & -3 \\ 1 & 1 \end{bmatrix}, \quad Q^{-1} = \begin{bmatrix} 1/4 & 3/4 \\ -1/4 & 1/4 \end{bmatrix}, \]

and

\[ e^{tA} = \begin{bmatrix} 1/4 + (3/4)e^{-4t} & 3/4 - (3/4)e^{-4t} \\ 1/4 - (1/4)e^{-4t} & 3/4 + (1/4)e^{-4t} \end{bmatrix}. \]

You should note that

\[ \lim_{t \to \infty} e^{tA} = \begin{bmatrix} 1/4 & 3/4 \\ 1/4 & 3/4 \end{bmatrix}, \]

which has a common row. Thus, for example, in the long run, the chain will be in state zero with a probability of 1/4. □
5.4 Stationary Distributions

In this section we will parallel our treatment of stationary distributions for discrete time Markov chains. We will aim for intuition, as opposed to attempting to prove everything, and point the interested reader to [37] and [34] for the full details of the proofs.

5.4.1 Classification of states

We start by again classifying the states of our process. Viewing a continuous time Markov chain as an embedded discrete time Markov chain with exponential holding times makes the classification of states, analogous to Section 2.4 in the discrete time setting, easy. We will again denote our state space as \( S \).

**Definition 5.4.1.** The communication classes of the continuous time Markov chain \( X(t) \) are the communication classes of the embedded Markov chain \( X_n \). If there is only one communication class, we say the chain is *irreducible*; otherwise it is said to be *reducible*.

Noting that \( X(t) \) will return to a state \( i \) infinitely often if and only if the embedded discrete time chain does (even in the case of an explosion!) motivates the following.

**Definition 5.4.2.** State \( i \in S \) is called *recurrent* for \( X(t) \) if \( i \) is recurrent for the embedded discrete time chain \( X_n \). Otherwise, \( i \) is *transient*.

**Definition 5.4.3.** Let \( T_1 \) denote the first jump time of the continuous time chain. We define
\[
\tau_i \overset{\text{def}}{=} \inf\{t \geq T_1 : X(t) = i\},
\]
and set \( m_i = \mathbb{E}_i \tau_i \). We say that state \( i \) is *positive recurrent* if \( m_i < \infty \).

Note that, perhaps surprisingly, we do *not* define \( i \) to be positive recurrent if \( i \) is positive recurrent for the discrete time chain. In Example 5.4.10 we will demonstrate that \( i \) may be positive recurrent for \( X_n \), while not for \( X(t) \).

As in the discrete time setting, recurrence, transience, and positive recurrence are class properties.

Note that the concept of periodicity no longer plays a role, or even makes sense to define, as time is no longer discrete. In fact, if \( P(t) \) is the matrix with entries \( P_{ij}(t) = P\{X(t) = j \mid X(0) = i\} \) for an irreducible continuous time chain, then for every \( t > 0 \), \( P_{ij}(t) \) has strictly positive entries because there is necessarily a path between \( i \) and \( j \), and a non-zero probability of moving along that path in time \( t > 0 \).

5.4.2 Invariant measures

Recall that equation (5.9) states that if the initial distribution of the process is \( \alpha \), then \( \alpha P(t) \) is the vector whose \( i \)th component gives the probability that \( X(t) = i \). We therefore define an invariant measure in the following manner.
**Definition 5.4.4.** A measure $\eta = \{\eta_j, j \in S\}$ on $S$ is called *invariant* if for all $t > 0$

$$\eta P(t) = \eta.$$ 

If this measure is a probability distribution (i.e. sums to one), then it is called a *stationary distribution*.

Note, therefore, that if the initial distribution is $\eta$, then $P_t\{X(t) = i\} = \eta_i$, for all $t \geq 0$, demonstrating why such a measure is called *invariant*.

The following theorem gives us a nice way to find stationary distributions of continuous time Markov chains.

**Theorem 5.4.5.** Let $X(t)$ be an irreducible and recurrent continuous time Markov chain with generator matrix $A$. Then the following statements are equivalent:

1. $\eta A = 0$;
2. $\eta P(t) = \eta$, for all $t \geq 0$.

**Proof.** The proof of this fact is easy in the case of a finite state space, which is what we will assume here. Recall Kolmogorov’s backward equation

$$P'(t) = AP(t).$$

Assume that $\eta A = 0$. Multiplying the backwards equation on the left by $\eta$ shows

$$0 = \eta AP(t) = \eta P'(t) = \frac{d}{dt} \eta P(t).$$

Thus,

$$\eta P(t) = \eta P(0) = \eta,$$

for all $t \geq 0$.

Now assume that $\eta P(t) = \eta$ for al $t \geq 0$. Then, for all $h > 0$, we have

$$\eta P(h) = \eta \implies \eta(P(h) - I) = 0 \implies \frac{\eta}{h}(P(h) - I) = 0.$$

Taking $h \to 0$ now shows that

$$0 = \eta P'(0) = \eta A,$$

where we have used that $P'(0) = A$, which follows from either the forward or backward equations.

The interchange above of differentiation with summation can not in general be justified in the infinite dimensional setting, and different proof is needed and we refer the reader to [34, Section 3.5].
Theorem 5.4.6. Suppose that $X(t)$ is irreducible and recurrent. Then $X(t)$ has an invariant measure $\eta$, which is unique up to multiplicative factors. Moreover, for each $k \in S$, we have

$$\eta_k = \pi_k / \lambda(k),$$

where $\pi$ is the unique invariant measure of the embedded discrete time Markov chain $X_n$. Finally, $\eta$ satisfies

$$0 < \eta_j < \infty, \quad \forall j \in S,$$

and if $\sum_i \eta_i < \infty$ then $\eta$ can normalize by $1 / \sum_k \eta_k$ to give a stationary distribution.

Proof. By Theorem 5.4.5, we must only show that there is a solution to $\eta A = 0$, satisfying all the desired results, if and only if there is an invariant measure to the discrete time chain. We first recall that $\pi$ was an invariant measure for a discrete time Markov chain if and only if $\pi Q = \pi$, where $Q$ is the transition matrix. By Theorem 2.5.16, such a $\pi$ exists, and is unique up to multiplicative constants, if $X_n$ is irreducible and recurrent.

Recall that if $j \neq k$, then $A_{jk} = \lambda(j) Q_{jk}$ and that $A_{jj} = -\lambda(j)$. We now simply note that

$$\eta' A = 0 \iff \sum_j \eta_j A_{jk} = 0, \quad \forall k \iff \sum_{j \neq k} \eta_j \lambda(j) Q_{jk} - \eta_k \lambda(k) = 0.$$ 

However, this holds if and only if

$$\sum_{j \neq k} \eta_j \lambda(j) Q_{jk} = \eta_k \lambda(k) \iff \pi Q = \pi,$$

where $\pi_k \overset{\text{def}}{=} \lambda(k) \eta_k$.

That is, the final equation (and hence all the others) holds if and only if $\pi$ is invariant for the Markov matrix $Q$. Such a $\pi$ exists, and satisfies all the desired properties, by Theorem 2.5.16. Further, we see the invariant measure of the continuous time Process satisfies $\eta_k = \pi_k / \lambda(k)$, as desired. \qed

Example 5.4.7. Consider the continuous time Markov chain with generator matrix

$$A = \begin{bmatrix}
-5 & 3 & 1 & 1 \\
1 & -1 & 0 & 0 \\
2 & 1 & -4 & 1 \\
0 & 2 & 2 & -4
\end{bmatrix}.$$ 

The unique left eigenvector of $A$ with eigenvalue 0, i.e. the solution to $\eta A = 0$, normalized to sum to one is

$$\eta = \begin{bmatrix}
14 \\
58 \\
6 \\
5
\end{bmatrix}.$$

145
Further, note that the transition matrix for the embedded discrete time Markov chain is

\[
P = \begin{bmatrix}
0 & 3/5 & 1/5 & 1/5 \\
1 & 0 & 0 & 0 \\
1/2 & 1/4 & 0 & 1/4 \\
0 & 1/2 & 1/2 & 0
\end{bmatrix}.
\]

Solving for the stationary distribution of the embedded chain, i.e. solving \( \pi P = \pi \), yields

\[
\pi = \begin{bmatrix}
35/86 & 29/86 & 6/43 & 5/43
\end{bmatrix}.
\]

Finally, note that

\[
[\eta_1 \lambda(1), \eta_2 \lambda(2), \eta_3 \lambda(3), \eta_4 \lambda(4)] = \begin{bmatrix}
5 \cdot 14/83 & 58/83 & 6 \cdot 4/83 & 5 \cdot 5/83
\end{bmatrix} = \begin{bmatrix}
70/83 & 24/83 & 20/83
\end{bmatrix} = \frac{172}{83} \begin{bmatrix}
35/86 & 29/86 & 6/43 & 5/43
\end{bmatrix} = \frac{172}{83} \pi,
\]

as predicted by the theory. \( \square \)

We now consider the positive recurrent case. We recall that \( m_i = E_i \tau_i \), the expected first return time to state \( i \). The following result should not be surprising at this point. See [34] for a proof.

**Theorem 5.4.8.** Let \( A \) be the generator matrix for an irreducible continuous time Markov chain. Then the following are equivalent

1. Every state is positive recurrent.
2. Some state is positive recurrent.
3. \( A \) is non-explosive and has an invariant distribution \( \eta \).

**Definition 5.4.9.** We call the non-explosive continuous time Markov chain \( \{X(t)\} \) **ergodic** if \( \{X_n\} \) is recurrent and irreducible and a stationary distribution exists.

Note, therefore, that \( X(t) \) is ergodic if and only if the chain is irreducible and positive recurrent.

The following example shows that positive recurrence of \( X_n \) does not guarantee existence of stationary distribution for \( X(t) \). That is, \( X(t) \) may not be positive recurrent.

146
Example 5.4.10. We consider a continuous time Markov chain whose embedded
discrete time Markov chain has state space \( S = \{0, 1, 2, \ldots\} \) and transition matrix

\[
Q = \begin{pmatrix}
  0 & 1 & 0 & 0 & \cdots \\
  q & 0 & p & 0 & \cdots \\
  q & 0 & 0 & p & \cdots \\
  \vdots & & \ddots & \ddots & \ddots
\end{pmatrix},
\]

where \( p + q = 1 \). This is the “success run chain” and we showed in Problem 2.11 that
the discrete time chain is positive recurrent. Let \( \lambda(i) \) be the holding time parameter
for state \( i \) of the associated continuous time Markov chain, and let \( E_m, m \geq 0, \)
denote a sequence of independent unit exponential random variables, which are also
independent of the embedded discrete time Markov chain. Finally, assuming that
\( X_0 = 0 \), let \( T_1 \) denote the first return time to state 0 of the _embedded chain_. For
example, if \( T_1 = 3 \), then \( X_0 = 0, X_1 = 1, X_2 = 2, \) and \( X_3 = 0 \). More generally,
we have \( X_0 = 0, X_1 = 1, \ldots, X_{T_1-1} = T_1 - 1, \) and \( X_T = 0 \). For \( m < T_1 \), we let
\( W(m) = E_m/\lambda(m) \) be the holding time in state \( m \). We have

\[
m_0 = \mathbb{E}_0 T_0 = \mathbb{E}_0 \sum_{m=0}^{T_1-1} W(m) \\
= \mathbb{E} \sum_{m=0}^{\infty} W(m) 1_{\{m<T_1\}} \\
= \sum_{m=0}^{\infty} \mathbb{E}[W(m) 1_{\{m<T_1\}}].
\]

However, we know that the holding times and the embedded chain are independent.
Thus, as \( 1_{\{m<T_1\}} \) is simply a statement pertaining to the embedded chain,

\[
\mathbb{E}[W(m) 1_{\{m<T_1\}}] = [\mathbb{E} W(m)][\mathbb{E} 1_{\{m<T_1\}}] = \frac{1}{\lambda(m)} P_0 \{ m < T_1 \}.
\]

Combining the above,

\[
m_0 = \sum_{m=0}^{\infty} \frac{1}{\lambda(m)} P_0 \{ m < T_1 \} \\
= \frac{1}{\lambda(0)} + \sum_{m=1}^{\infty} \frac{1}{\lambda(m)} P_0 \{ m < T_1 \}.
\]

For \( m \geq 1, \)

\[
P\{m < T_1\} = \sum_{n=m+1}^{\infty} P\{T_1 = n\} = \sum_{n=m+1}^{\infty} p^{n-2} q = qp^{m-1} \sum_{n=0}^{\infty} p^n = p^{m-1}.
\]

147
Thus,

\[ m_0 = \frac{1}{\lambda(0)} + \sum_{m=1}^{\infty} \frac{1}{\lambda(m)p^{m-1}}. \]

Of course, we have not chosen \( \lambda(m) \) yet. Taking \( \lambda(m) = p^m \), we see

\[ m_0 = \frac{1}{\lambda(0)} + \sum_{m=1}^{\infty} \frac{1}{p^mp^{m-1}} = 1 + \sum_{m=1}^{\infty} \frac{1}{p} = \infty. \]

So, \( \{X_n\} \) is positive recurrent, but \( X(t) \) is not. \( \square \)

The following example, taken from [34], shows two things. First, it demonstrates that a transient chain can have an invariant measure. Further, it even shows stranger behavior is possible: a transient chain can have an invariant distribution! Of course, the previous theorems seem to suggest that this is not possible. However, there is a catch: the chain could be explosive. In fact, if a transient chain is shown to have a stationary distribution, then the chain must be explosive for otherwise Theorem 5.4.8 is violated.

**Example 5.4.11.** Consider a discrete time random walker on \( S = \{0,1,2,\ldots\} \). Suppose that the probability of moving to the right is \( p > 0 \) and to the left is \( q = 1 - p \). To convert this into a continuous time chain, we suppose that \( \lambda(i) \) is the holding time parameter in state \( i \). More specifically, we assume \( X(t) \) is a continuous time Markov chain with generator matrix \( A \) satisfying

\[
A = \begin{pmatrix}
-\lambda(0)p & \lambda(0)p & 0 & 0 & 0 & \cdots \\
q\lambda(1) & -\lambda(1) & p\lambda(1) & 0 & 0 & \cdots \\
0 & q\lambda(2) & -\lambda(2) & p\lambda(2) & 0 & \cdots \\
0 & 0 & q\lambda(3) & -\lambda(3) & p\lambda(3) & \cdots \\
\vdots & \ddots & \ddots & \ddots & \ddots
\end{pmatrix}
\]

We know that this chain is transient if \( p > q \) since the discrete time chain is. We now search for an invariant measure satisfying

\[ \eta A = 0, \]

which in component form is

\[
-\lambda(0)p\eta_0 + q\lambda(1)\eta_1 = 0 \\
\lambda(i-1)p\eta_{i-1} - \lambda(i)\eta_i + \lambda(i+1)q\eta_{i+1} = 0 \quad i > 0.
\]

We will confirm that \( \eta \) satisfying

\[ \eta(i) = \frac{1}{\lambda(i)} \left( \frac{p}{q} \right)^i, \]

148
is a solution. The case $i = 0$ is easy to verify

$$\lambda(0)p\eta_0 = \lambda(0)p \frac{1}{\lambda(0)} = p = q\lambda(1)\frac{1}{\lambda(1)} \frac{p}{q} = q\lambda(1)\eta_1.$$ 

The $i > 0$ case follows similarly.

Therefore, there is always an invariant distribution, regardless of the values $p$ and $q$. Taking $p > q$ and $\lambda(i) = 1$ for all $i$, we see that the resulting continuous time Markov chain is transient, and has an invariant distribution

$$\eta(i) = \left(\frac{p}{q}\right)^i,$$

which can not be normalized to provide an invariant distribution.

Now, consider the case when $p > q$, with $1 < p/q < 2$, and take $\lambda(i) = 2^i$. Define $\alpha \overset{\text{def}}{=} p/q < 2$. Then,

$$\sum_{i=0}^\infty \eta(i) = \sum_{i=0}^\infty \left(\frac{\alpha}{2}\right)^i = \frac{1}{1-\alpha/2} = \frac{2}{2-\alpha} < \infty,$$

Therefore, we can normalize to get a stationary distribution. Since we already know this chain is transient, we have shown that it must, in fact, explode. \hfill \square

### 5.4.3 Limiting distributions and convergence

We have found conditions for the existence of a unique stationary distribution to a continuous time Markov chain: irreducibility and positive recurrence (i.e. ergodicity). As in the discrete time case, there is still the question of convergence. The following is proven in [34].

**Theorem 5.4.12.** Let $X(t)$ be an ergodic continuous time Markov chain with unique invariant distribution $\eta$. Then, for all $i, j \in S$,

$$\lim_{t \to \infty} P_{ij}(t) = \eta_j.$$

**Example 5.4.13.** Let $S = \{0, 1\}$ with transition rates $\lambda(0,1) = 3$ and $\lambda(1,0) = 1$. Then the generator matrix is

$$A = \begin{bmatrix} -3 & 3 \\ 1 & -1 \end{bmatrix}.$$ 

Solving directly for the left eigenvector of $A$ with eigenvalue 0 yields

$$\pi = [1/4, 3/4],$$

which agrees with the result found in Example 5.3.9. \hfill \square

As in the discrete time setting, we have an ergodic theorem, which we simply state. For a proof, see [37, Section 5.5].

149
Theorem 5.4.14. Let $X(t)$ be an irreducible, positive recurrent continuous time Markov chain with unique stationary distribution $\eta$. Then, for any initial condition, and any $i \in S$, 

$$P\left( \frac{1}{t} \int_0^t 1_{\{X(s)=i\}} ds \to \eta_i, \ as \ t \to \infty \right) = 1.$$ 

Moreover, for any bounded function $f : S \to \mathbb{R}$ we have 

$$P\left( \frac{1}{t} \int_0^t f(X(s)) ds \to \bar{f}, \ as \ t \to \infty \right) = 1,$$

where 

$$\bar{f} = \sum_{j \in S} \eta_j f(j) = E_{\eta} f(X_\infty),$$

where $X_\infty$ has distribution $\eta$.

Thus, as in the discrete time setting, we see that $\eta_i$ gives the proportion of time spent in state $i$ over long periods of time. This gives us an algorithmic way to sample from the stationary distribution: simulate a single long trajectory and average over it.

5.5 The Generator, Revisited

Consider a function $f : S \to \mathbb{R}$. Noting that $f$ is simply a mapping from $S$ to $\mathbb{R}$, and that $S$ is discrete, we can view $f$ as a column vector whose $i$th component is equal to $f(i)$. For example, if $S = \{1, 2, 3\}$ and $f(1) = -2, f(2) = \pi,$ and $f(3) = 100$, then we take

$$f = \begin{bmatrix} -2 \\ \pi \\ 100 \end{bmatrix}.$$ 

As $A$ is a matrix, it therefore makes sense to discuss the well defined object $Af$, which is itself a column vector, and hence a function from $S$ to $\mathbb{R}$.

Next, we note that if the initial distribution for our Markov chain is $\alpha$, then for any $f$ we have that 

$$E_{\alpha} f(X(t)) = \sum_{j \in S} P_{\alpha} \{X(t) = j\} f(j)$$

$$= \sum_{j \in S} \left( \sum_{i \in S} P\{X(t) = j \mid X(0) = i\} P\{X(0) = i\} \right) f(j)$$

$$= \sum_{i \in S} \alpha_i \left( \sum_{j \in S} P_{ij}(t) f(j) \right)$$

$$= \sum_{i \in S} \alpha_i (P(t)f)_i$$

$$= \alpha P(t)f.$$
Now recall that the forward equation stated that $P'(t) = P(t)A$. Integrating this equation yields

$$P(t) = I + \int_0^t P(s)A ds,$$

and multiplication on the right by $f$ gives

$$P(t)f = f + \int_0^t P(s)Af ds.$$ (5.12)

Multiplying (5.12) on the left by $\alpha$ yields

$$\alpha P(t) = \alpha f + \int_0^t \alpha P(s)(Af) ds,$$

which combined with (5.11) gives

$$E_{\alpha}f(X(t)) = E_{\alpha}f(X(0)) + \int_0^t E_{\alpha}(Af)(X(s)) \, ds$$

This is a version of Dynkin’s formula. For a more formal derivation in the Markov process setting, see [15, Section 1.1]. In the next section, we will use this formulation to calculate the mean and variance of a linear birth and death model.

**Example 5.5.1.** We will re-derive the mean and variance of a Poisson process using Dynkin’s formula. Let $X(t)$ be a Poisson process with intensity $\lambda > 0$ defined on $S = \{0, 1, 2, \ldots \}$. Then, for any function $f : S \to \mathbb{R}$

$$(Af) = \begin{bmatrix} -\lambda & \lambda & 0 & 0 & \cdots \\ 0 & -\lambda & \lambda & 0 & \cdots \\ 0 & 0 & -\lambda & \lambda & \cdots \\ \vdots & \ddots & \ddots & \ddots & \ddots \end{bmatrix} \begin{bmatrix} f(0) \\ f(1) \\ f(2) \\ \vdots \end{bmatrix} = \begin{bmatrix} -\lambda f(0) + \lambda f(1) \\ -\lambda f(1) + \lambda f(2) \\ -\lambda f(2) + \lambda f(3) \\ \vdots \end{bmatrix},$$

and so, for any $i \geq 0$,

$$(Af)(i) = \lambda(f(i + 1) - f(i)).$$

Letting $f(i) = i$, and taking $X(0) = 0$ with a probability of one, we therefore see that

$$\mathbb{E}f(X(t)) = \mathbb{E}X(t) = 0 + \int_0^t \mathbb{E}(Af)(X(s)) ds$$

$$= \int_0^t \mathbb{E}\lambda(f(X(s) + 1) - f(X(s))) ds$$

$$= \lambda \int_0^t ds$$

$$= \lambda t.$$
Next, letting \( g(i) = i^2 \) (in order to find the second moment), we have

\[
E_g(X(t)) = E X(t)^2 = 0 + \int_0^t E (Af)(X(s))ds \\
= \int_0^t E \lambda (g(X(s) + 1) - g(X(s)))ds \\
= \lambda \int_0^t E (X(s)^2 + 2X(s) + 1 - X(s)^2)ds \\
= \lambda \int_0^t E (2X(s) + 1)ds \\
= \lambda \int_0^t (2\lambda s + 1)ds \\
= \lambda^2 t^2 + \lambda t.
\]

Therefore, the variance is

\[
\text{Var}(X(t)) = E X(t)^2 - (E X(t))^2 = \lambda t.
\]

Of course, both the mean and variance of a Poisson process are well known. However, the above method is quite general and is useful in a myriad of applications.

**Example 5.5.2.** Consider a pure birth process with growth rate \( \lambda(i) = bi \) for some \( b > 0 \). That is, the embedded chain is the deterministic monotone chain and the holding time in state \( i \) is \( bi \). For an arbitrary function \( f \), we have that

\[
(Af)(i) = bi (f(i + 1) - f(i)), \tag{5.13}
\]

for all \( i \geq 0 \), where \( A \) is the generator for the continuous time chain. Assuming \( X(0) = 1 \), we will derive the mean of the process.

For \( f(i) = i \), we have that

\[
E f(X(t)) = E X(t) = 1 + \int_0^t E (Af)(X(s))ds \\
= 1 + \int_0^t E b X(s)(f(X(s) + 1) - f(X(s)))ds \\
= 1 + b \int_0^t E X(s)ds.
\]

Therefore, defining \( g(t) = E X(t) \), we see that

\[
g'(t) = bg(t), \quad g(0) = 1.
\]

Thus,

\[
g(t) = E X(t) = e^{bt}.
\]

This result should be compared with the solution to the ODE linear growth model \( x'(t) = bx(t) \), which yields a similar solution. You will derive the variance for a homework exercise.
Now consider the (row) vector $e_i$, with a one in the $i$th component, and zeros everywhere else. Taking $e_i$ as an initial distribution, we see from (5.11) that for all $t \geq 0$

$$e_iP(t)f = \mathbb{E}_i f(X(t)).$$

In words, the $i$th component of the vector $P(t)f$ gives $\mathbb{E}_i f(X(t))$. Next, note that

$$(Af)(i) = e_i(Af) = e_i(P'(0)f) = e_i \lim_{h \to 0} \frac{1}{h} (P(h)f - P(0)f)$$

$$= \lim_{h \to 0} \frac{1}{h} (e_iP(h)f - e_if)$$

$$= \lim_{h \to 0} \frac{\mathbb{E}_i f(X(h)) - f(i)}{h}.$$

Further, taking $f(i) = 1_{i = j}$ for some $j$, we see that

$$(Af)(i) = \lim_{h \to 0} \frac{\mathbb{E}_i f(X(h)) - f(i)}{h},$$

(5.14) gives

$$A_{ij} = \lim_{h \to 0} \frac{1}{h} (P\{X(h) = j \mid X(0) = i\}) = \lambda(i, j),$$

when $i \neq j$, and

$$A_{jj} = \lim_{h \to 0} \frac{1}{h} (P\{X(h) = j \mid X(0) = j\} - 1) = -\lambda(j),$$

for the diagonal elements. Therefore, (5.14) could be taken as an alternative definition of the generator for a Markov process, though one which views the generator as an operator and not simply as a matrix that stores the transition intensities. In fact, in many ways this definition is much more useful than that of simply the matrix with transition rates.

**Example 5.5.3.** Consider a process with arrivals coming in at rate $\lambda > 0$ and departures taking place at rate $\mu X(t)$, where $X(t)$ is the number of items at time $t$. Then, for $i \geq 0$ we have

$$(Af)(i) = \lim_{h \to 0} \frac{\mathbb{E}_i f(X(h)) - f(i)}{h}$$

$$= \lim_{h \to 0} \frac{1}{h} \left[ f(i + 1)P_i\{X(h) = i + 1\} + f(i - 1)P_i\{X(h) = i - 1\} \right.$$

$$\left. + f(i)P_i\{X(h) = i\} - f(i) + o(h) \right]$$

$$= \lim_{h \to 0} \frac{1}{h} \left[ f(i + 1)\lambda h + f(i - 1)\mu i h + f(i)(1 - \lambda h - \mu i h) - f(i) + o(h) \right]$$

$$= \lambda(f(i + 1) - f(i)) + \mu i(f(i - 1) - f(i)).$$
So, for example, taking \( f(y) = y \) to be the identity, and \( X(0) = x \), we have that

\[
\mathbb{E}X(t) = \mathbb{E}f(X(t)) = \mathbb{E}X(0) + \mathbb{E} \int_0^t (Af)(X(s))\,ds
\]

\[
= x + \mathbb{E} \int_0^t (\lambda(X(s) + 1 - X(s))) + \mu X(s)(X(s) - 1 - X(s))\,ds
\]

\[
= x + \int_0^t (\lambda - \mu \mathbb{E}X(s))\,ds.
\]

Setting \( g(t) = \mathbb{E}X(t) \), we see that \( g(0) = x \) and \( g'(t) = \lambda - \mu g(t) \). Solving this initial value problem yields the solution

\[
\mathbb{E}X(t) = xe^{-\mu t} + \frac{\lambda}{\mu}(1 - e^{-\mu t}).
\]

The second moment, and hence the variance, of the process can be calculated in a similar manner.

5.6 Continuous Time Birth and Death Processes

We again consider a Markovian birth and death process, though now in the continuous time setting. As in Section 3.2, our state space is \( S = \{0, 1, 2, \ldots \} \). The transition rates are

\[
\lambda(n, n + 1) = b_n,
\]

\[
\lambda(n, n - 1) = d_n,
\]

\[
\lambda(n, j) = 0, \quad \text{if } |j - n| \geq 2,
\]

for some values \( b_n, d_n \geq 0 \), and \( d_0 = 0 \), yielding a tridiagonal generator matrix

\[
A = \begin{bmatrix}
-b_0 & b_0 & 0 & 0 & 0 & \cdots \\
d_1 & -(b_1 + d_1) & b_1 & 0 & 0 & \cdots \\
0 & d_2 & -(b_2 + d_2) & b_2 & 0 & \cdots \\
0 & 0 & d_3 & -(b_3 + d_3) & b_3 & \cdots \\
\vdots & \vdots & \ddots & \ddots & \ddots & \ddots
\end{bmatrix}.
\]

We begin with examples, many of which are analogous to those in the discrete time setting.

Example 5.6.1. The Poisson process is a birth-death process with \( b_n \equiv \lambda \), for some \( \lambda > 0 \), and \( d_n \equiv 0 \).

Example 5.6.2. A pure birth process with \( b_n \geq 0 \), and \( d_n \equiv 0 \) is an example of a birth and death process.
Example 5.6.3 (Queueing Models). We suppose that arrivals of customers are occurring at a constant rate of $\lambda > 0$. That is, we assume that $b_n \equiv \lambda$. However, departures occur when a customer has been served. There are a number of natural choices for the model of the service times.

(a) (Single server) If there is a single server, and that person always serves the first person in line, then we take $d_n = \mu > 0$, if $n \geq 1$, and $d_0 = 0$ (as there is no one to serve).

(b) ($k$ servers) If there are $k \geq 1$ servers, and the first $k$ people in line are always being served, then for some $\mu > 0$ we take

$$d_n = \begin{cases} n\mu, & \text{if } n \leq k \\ k\mu, & \text{if } n \geq k \end{cases}.$$ 

(c) ($\infty$ servers) If we suppose that there are an infinite number of servers, then $d_n = n\mu$ for some $\mu > 0$.

Example 5.6.4 (Population Models). Suppose that $X(t)$ represents the number of individuals in a certain population at time $t$. Assuming the rates of both reproduction and death are proportional to population size we have

$$b_n = \lambda n,$$
$$d_n = \mu n,$$

for some $\lambda, \mu > 0$.

Example 5.6.5 (Population with immigration). Consider the previous system except $b_n = \lambda n + \nu$ for some $\nu > 0$, representing an inflow due to immigration. Now 0 is no longer an absorbing state.

Example 5.6.6 (Fast growing population). Consider a population that grows at a rate equal to the square of the number of individuals. Assuming no deaths, we have for some $\lambda > 0$ that

$$b_n = \lambda n^2,$$
$$\mu_n = 0.$$

We have seen that this population grows so fast that it reaches an infinite population in finite time with a probability of one.

Example 5.6.7 (Chemistry 1). Consider the chemical system $A \overset{k_1}{\underset{k_2}{\rightleftharpoons}} B$ with $A(0) + B(0) = N$ and mass action kinetics. Then, $A(t)$, giving the number of $A$ molecules at time $t$, is a birth and death process with state space $\{0, 1, \ldots, N\}$ and transition rates

$$b_n = k_2(N - n), \quad d_n = k_1n.$$ 

□
Example 5.6.8 (Chemistry 2). Consider the chemical system
\[ \emptyset \xrightleftharpoons[\mu]{\lambda} A \]
and suppose \( X(t) \) tracks the number of \( A \) molecules. Then this model is a birth and death process with the exact same transition rates as the infinite server queue of Example 5.6.3. \( \square \)

Returning to a general system, consider the embedded discrete time Markov chain of a general continuous time birth and death process. The transition probabilities of this chain are

\[ p_{n,n+1} = p_n \overset{\text{def}}{=} \frac{b_n}{b_n + d_n}, \]

\[ q_{n,n-1} = q_n \overset{\text{def}}{=} \frac{d_n}{b_n + d_n}. \]

Note that in this case we have \( p_n + q_n = 1 \) for all \( n \geq 0 \). We will first consider when these processes are recurrent and transient, and then consider positive recurrence. The following proposition follows directly from Proposition 3.2.6.

Proposition 5.6.9. A continuous time birth and death process is transient if and only if

\[ \sum_{k=1}^{\infty} \frac{d_1 \cdots d_k}{b_1 \cdots b_k} < \infty. \]

Proof. From Proposition 3.2.6, the embedded chain, and hence the continuous time chain, is transient if and only if

\[ \sum_{k=1}^{\infty} \frac{q_1 \cdots q_n}{p_1 \cdots p_k} < \infty. \]

Noting that

\[ \sum_{k=1}^{\infty} \frac{q_1 \cdots q_n}{p_1 \cdots p_k} = \sum_{k=1}^{\infty} \frac{d_1 \cdots d_k}{b_1 \cdots b_k}, \]

completes the proof. \( \square \)

Similarly to the discrete time case, we can now conclude that the single server queue is transient if and only if \( \mu < \lambda \), and that the \( k \) server queue is transient if and only if \( k\mu < \lambda \). For the infinite server queue, and the analogous chemistry example in Example 5.6.8, we have

\[ \sum_{k=1}^{\infty} \frac{d_1 \cdots d_k}{p_1 \cdots p_k} = \sum_{k=1}^{\infty} k! \left( \frac{\mu}{\lambda} \right)^k = \infty. \]
Thus, the infinite server queue is always recurrent.

We now turn to the question of positive recurrence and stationary distributions. We know that a stationary distribution $\eta$ must satisfy $\eta A = 0$, which in component form is

$$\eta_0 b_0 = \eta_1 d_1$$

$$(b_k + d_k)\eta_k = b_{k-1}\eta_{k-1} + d_{k+1}\eta_{k+1}, \quad \text{for } k \geq 1.$$  

Noting that these are the same equations as (3.5) and (3.6), we can conclude that such an $\eta$ exists and can be made into a probability vector if and only if

$$\sum_{k=1}^{\infty} \frac{b_0 \cdots b_{k-1}}{d_1 \cdots d_k} < \infty.$$  

The following is analogous to Proposition 3.2.7.

**Proposition 5.6.10.** There exists a stationary distribution for a continuous time birth and death chain if and only if

$$\sum_{k=1}^{\infty} \frac{b_0 b_1 \cdots b_{k-1}}{d_1 \cdots d_k} < \infty.$$  

In this case,

$$\eta_0 = \left(\sum_{k=0}^{\infty} \frac{b_0 b_1 \cdots b_{k-1}}{d_1 \cdots d_k}\right)^{-1},$$  

where the $k = 0$ term in the above sum is taken to be equal to one, and for $k \geq 1$,

$$\eta_k = \frac{b_0 \cdots b_{k-1}}{d_1 \cdots d_k} \eta_0.$$  

For example, for the single server queue we have

$$\sum_{k=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^k = \left(1 - \frac{\lambda}{\mu}\right)^{-1},$$  

provided $\lambda < \mu$, and in this case

$$\eta_k = \left(1 - \frac{\lambda}{\mu}\right) \left(\frac{\lambda}{\mu}\right)^k.$$  

The expected length of the queue is

$$\sum_{k=0}^{\infty} k\eta_k = k \left(1 - \frac{\lambda}{\mu}\right) \left(\frac{\lambda}{\mu}\right)^k = \frac{\lambda}{\mu} \left(1 - \frac{\lambda}{\mu}\right)^{-1} = \frac{\lambda}{\mu - \lambda},$$  

where the $k = 0$ term in the above sum is taken to be equal to one, and for $k \geq 1$. 

157
which grows to infinity as $\lambda$ approaches $\mu$.

For the infinite server queue and the chemistry model of Example 5.6.8, we have

$$\sum_{k=0}^{\infty} b_0 \cdots b_{k-1} d_1 \cdots d_k = \sum_{k=0}^{\infty} \frac{1}{k!} \left( \frac{\lambda}{\mu} \right)^k = e^{\lambda/\mu}.$$ 

Therefore, a stationary distribution exists, and since we already know the chain is recurrent we may conclude it is positive recurrent. Note that the stationary distribution is Poisson($\lambda/\mu$), and

$$\eta_k = e^{-\lambda/\mu} \frac{(\lambda/\mu)^k}{k!}, \quad \text{for } k \geq 0.$$ 

In the next chapter, we will see why many models from chemistry and biology have stationary distributions that are Poisson.

We close by considering the generator for a continuous time birth and death process. It is straightforward to show that it satisfies

$$(Af)(i) = b_i (f(i + 1) - f(i)) + d_i (f(i - 1) - f(i)),$$

for all $i \geq 0$. This fact can be used in the case of linear intensities to easily calculate the time-dependent moments.

**Example 5.6.11.** Consider linear birth and death process with transition rates

$$b_i = \lambda i, \quad d_i = \mu i,$$ 

where $\lambda, \mu > 0$. The generator of the process satisfies

$$(Af)(i) = \lambda i (f(i + 1) - f(i)) + \mu i (f(i - 1) - f(i)),$$

for all $i \geq 0$. Taking $f(i) = i$ to be the identity, and $X(0) = x$, we have that

$$\mathbb{E}f(X(t)) = \mathbb{E}X(t) = \mathbb{E}X(0) + \mathbb{E} \int_0^t (Af)(X(s))ds$$

$$= x + \mathbb{E} \int_0^t \left[ \lambda X(s)(X(s) + 1 - X(s)) + \mu X(s)(X(s) - 1 - X(s)) \right] ds$$

$$= x + (\lambda - \mu) \int_0^t \mathbb{E}X(s)ds.$$ 

Solving yields

$$\mathbb{E}X(t) = xe^{(\lambda-\mu)t}, \quad (5.15)$$

which, it is worth noting, is the solution to the ordinary differential equation $x'(t) = (\lambda - \mu)x(t)$ that is the standard deterministic model for this system. \[\square\]
5.7 Exercises

1. Consider a continuous time Markov chain with state space \{1, 2, 3, 4\} and generator matrix

\[
A = \begin{bmatrix}
-3 & 2 & 0 & 1 \\
0 & -2 & 1/2 & 3/2 \\
1 & 1 & -4 & 2 \\
1 & 0 & 0 & -1
\end{bmatrix}.
\]

Write a Matlab code that simulates a path of this chain. To do so, use the construction provided in the notes (i.e. simulate the embedded chain and holding times sequentially). Using this code and assuming that \(X(0) = 1\), estimate \(E X(3)\) by averaging over 10,000 such paths. Note that you will need to make sure you break your “for” or “while” loop after you see that the time will go beyond \(T = 3\), without updating the state for that step.

2. In Example 5.2.4, it was stated that if \(X(0) = 0\), then the probability of an explosion was 1/3. Why is that?

3. Suppose that \(\alpha > 1\). Prove that any birth process for which \(\lambda(n) = n^\alpha\) is explosive.

4. Consider a continuous time Markov chain with \(S = \{1, 2\}\) and

\[\lambda(1, 2) = 2, \quad \lambda(2, 1) = 3.\]

Find \(P(t)\), the matrix whose \(i, j\)th entry gives \(P\{X(t) = j \mid X(0) = i\}\).

5. Let \(A\) be the generator matrix for an irreducible continuous time Markov chain with finite state space \(S\). Let \(\lambda_0 > 0\) be such that \(|A_{ij}| < \lambda_0\) for all \(i, j \in S\). Now define

\[P = \frac{1}{\lambda_0} A + I,\]

where \(I\) is the identity matrix.

(a) Show that \(P\) is a transition matrix for an irreducible, aperiodic discrete time Markov chain on the state space \(S\).

(b) Using part (a), show that \(A\) has a unique left eigenvector with eigenvalue 0 that is a probability vector and that all other eigenvalues have negative real part.

6. (Taken from Durrett, [12]) Consider a hemoglobin molecule that can carry either one oxygen molecule or one carbon monoxide molecule. Suppose that the oxygen and carbon dioxide molecules are arriving one at a time according to rates \(\lambda_1\) and \(\lambda_2\) respectively. Suppose further that if the hemoglobin molecule is free when one of the molecules arrives it attaches to the hemoglobin molecule for an exponential amount of time with parameter \(\mu_1\) (for oxygen) and \(\mu_2\) (for carbon...
dioxide). Formulate a continuous time Markov chain for this system with state space \( \{O, E, C\} \), where \( O \) represents an attached oxygen molecule, \( C \) represents an attached carbon dioxide molecule, and \( E \) represents an empty hemoglobin molecule. Next, find the long run fraction of time the hemoglobin molecule is in each of its states.

7. For Example 5.5.2, verify that the generator of the process satisfies equation (5.13).

8. Using Dynkin’s formula, calculate \( \text{Var}(X(t)) \) of the linear birth process of Example 5.5.2.

9. Using Dynkin’s formula, calculate \( \text{Var}(X(t)) \) of the linear birth and death process of Example 5.6.11.
Chapter 6

Continuous Time Markov Chain Models for Biochemical Reaction Networks

6.1 Chemical reaction networks: basic construction

Some of this material is adapted from the survey article [7]. We will begin in Section 6.1.1 by developing the basic stochastic equations for chemical systems. Next, in Section 6.1.2 we describe the most common choice for intensity (or propensity) function: mass-action kinetics. In Section 6.1.3 we give a series of models for gene transcription and translation. In Section 6.1.4 we derive both the generator and the forward equations, which is also called the chemical master equation in this context. Finally, in Section 6.1.5 we point out that the model developed here is, in fact, quite general and not confined to the chemical setting.

6.1.1 The stochastic equations and basic terminology

We begin our study of chemical systems by considering a system consisting of three “species,” denoted $A$, $B$, and $C$, and one reaction,

$$A + B \rightarrow C,$$

in which one molecule of $A$ and one molecule of $B$ are consumed to produce one molecule of $C$. The intuition for the model for the reaction is that the probability of the reaction occurring in a small time interval $(t, t + \Delta t]$ should be proportional to the product of the numbers of molecules of each of the reactants and to the length of the time interval. In other words, since for the reaction to occur a molecule of $A$ and a molecule of $B$ must be close to each other, the probability should be proportional to the number of pairs of molecules that can react.

\textsuperscript{0}Copyright © 2013 by David F. Anderson.
Assuming that at time \( t \) there are \( X_A(t) \) molecules of \( A \) and \( X_B(t) \) molecules of \( B \) in our system, we express our assumption about the probability of the reaction occurring by

\[
P\{\text{reaction occurs in } (t, t + \Delta t) | \mathcal{F}_t\} \approx \kappa X_A(t) X_B(t) \Delta t \tag{6.1}
\]

where \( \mathcal{F}_t \) represents the information about the system that is available at time \( t \) and \( \kappa \) is a positive constant, the reaction rate constant. Mathematically, \( \mathcal{F}_t \) is a \( \sigma \)-algebra, but readers unfamiliar with this terminology should just keep the idea of information in mind when we write expressions like this, that is, \( \mathcal{F}_t \) just represents the information available at time \( t \).

As in Chapter 5, we can use assumption (6.1) to build a continuous time Markov chain model. Let \( X(t) = (X_A(t), X_B(t), X_C(t)) \) give the state of the process at time \( t \). Simple bookkeeping implies

\[
X(t) = X(0) + R(t) \begin{pmatrix} -1 \\ -1 \\ 1 \end{pmatrix}, \tag{6.2}
\]

where \( R(t) \) is the number of times the reaction has occurred by time \( t \) and \( X(0) \) is the vector giving the numbers of molecules of each of the chemical species in the system at time zero. We will assume that two reactions cannot occur at exactly the same time, so \( R \) is a counting process, that is, \( R(0) = 0 \) and \( R \) is constant except for jumps of plus one.

Based on our work in Chapters 4 and 5, we can model \( R(t) \) via

\[
R(t) = Y \left( \int_0^t \kappa X_A(s) X_B(s) ds \right), \tag{6.3}
\]

where \( Y \) is a unit rate Poisson process. Hence

\[
\begin{pmatrix} X_A(t) \\ X_B(t) \\ X_C(t) \end{pmatrix} \equiv X(t) = X(0) + \begin{pmatrix} -1 \\ -1 \\ 1 \end{pmatrix} Y \left( \int_0^t \kappa X_A(s) X_B(s) ds \right). \tag{6.4}
\]

Given \( Y \) and the initial state \( X(0) \) (which we assume is independent of \( Y \)), (6.4) is an equation that uniquely determines \( X \) for all \( t > 0 \). To see that this assertion is correct, let \( \tau_k \) be the \( k \)th jump time of \( R \) determined by (6.3). Then letting

\[
\zeta = \begin{pmatrix} -1 \\ -1 \\ 1 \end{pmatrix},
\]

(6.4) implies \( X(t) = X(0) \) for \( 0 \leq t < \tau_1 \), \( X(t) = X(0) + \zeta \) for \( \tau_1 \leq t < \tau_2 \), and so forth. To see that the solution of this equation has the properties suggested by (6.1), let \( \lambda(X(t)) = \kappa X_A(t) X_B(t) \) and observe that occurrence of the reaction in \( (t, t + \Delta t] \)
is equivalent to \( R(t + \Delta t) > R(t) \), so the left side of (6.1) becomes

\[
P\{R(t + \Delta t) > R(t)|\mathcal{F}_t\} = 1 - P\{R(t + \Delta t) = R(t)|\mathcal{F}_t\} = 1 - P \left\{ Y \left( \int_0^t \lambda(X(s))ds + \lambda(X(t))\Delta t \right) = Y \left( \int_0^t \lambda(X(s))ds \right) | \mathcal{F}_t \right\} = 1 - e^{-\lambda(X(t))\Delta t} \approx \lambda(X(t))\Delta t,
\]

where the third equality follows from the fact that \( Y(\int_0^t \lambda(X(s))ds) \) and \( X(t) \) are part of the information in \( \mathcal{F}_t \) (are \( \mathcal{F}_t \)-measurable in the mathematical terminology) and the independence properties of \( Y \).

More generally, we now consider a network of \( M \) chemical reactions involving \( N \) chemical species, \( S_1, \ldots, S_N \),

\[
\sum_{i=1}^{N} \nu_{ik} S_i \rightarrow \sum_{i=1}^{N} \nu'_{ik} S_i, \quad k = 1, \ldots, M,
\]

where \( \nu_{ik} \) and \( \nu'_{ik} \) are nonnegative integers. We let the components of \( X(t) \in \mathbb{Z}_{\geq 0}^N \) give the numbers of molecules of each species in the system at time \( t \). Let \( \nu_k \) be the vector whose \( i \)th component is \( \nu_{ik} \), the number of molecules of the \( i \)th chemical species consumed in the \( k \)th reaction, and let \( \nu'_k \) be the vector whose \( i \)th component is \( \nu'_{ik} \), the number of molecules of the \( i \)th species produced by the \( k \)th reaction. Let \( \lambda_k(x) \) be the rate at which the \( k \)th reaction occurs, that is, it gives the propensity/intensity of the \( k \)th reaction as a function of the numbers of molecules of the chemical species.

Letting \( R_k \) denote the number of times that the \( k \)th reaction occurs by time \( t \), we have

\[
P\{R_k(t + \Delta t) = R_k(t)|\mathcal{F}_t\} = \lambda_k(X(t))\Delta t + o(\Delta t), \quad \text{as } \Delta t \to 0.
\]

If the \( k \)th reaction occurs at time \( t \), the new state becomes

\[
X(t) = X(t-) + \nu'_k - \nu_k.
\]

The number of times that the \( k \)th reaction occurs by time \( t \) is given by the counting process satisfying

\[
R_k(t) = Y_k \left( \int_0^t \lambda_k(X(s))ds \right),
\]

where the \( Y_k \) are independent unit Poisson processes. The state of the system then satisfies

\[
X(t) = X(0) + \sum_k R_k(t)(\nu'_k - \nu_k) = X(0) + \sum_k Y_k \left( \int_0^t \lambda_k(X(s))ds \right) (\nu'_k - \nu_k).
\]
To simplify notation, we will write

\[ \zeta_k = \nu'_k - \nu_k, \]

so that the general form of the stochastic equation is

\[ X(t) = X(0) + \sum_k Y_k \left( \int_0^t \lambda(X(s))ds \right) \zeta_k. \quad (6.5) \]

The vectors \( \zeta_k \) are termed the reaction vectors of the model. The representation (6.5) is termed a random time change representation and is mainly due to work by Thomas Kurtz [15, Chapter 6] [31].

**Example 6.1.1.** In the original example we had \( M = 1, N = 3 \) and

\[
\begin{bmatrix}
1 \\
1 \\
0
\end{bmatrix}, \quad \begin{bmatrix}
0 \\
0 \\
1
\end{bmatrix}, \quad \text{and} \quad \begin{bmatrix}
-1 \\
-1 \\
1
\end{bmatrix},
\]

with \( X(t) \) satisfying (6.4). \( \square \)

**Example 6.1.2.** Consider the chemical system from Example 5.6.8,

\[ \emptyset \xrightleftharpoons{\lambda} \mu A. \]

Here, there is only one species and the stochastic equation is

\[ X(t) = X(0) + Y_1(\lambda t) + Y_k \left( \int_0^t \mu X(s)ds \right), \]

where \( Y_1, Y_2 \) are independent unit Poisson processes. \( \square \)

### 6.1.2 Rates for the law of mass action

The stochastic form of the law of mass action says that the rate at which a reaction occurs should be proportional to the number of distinct subsets of the molecules present that can form the inputs for the reaction. That is,

\[ \lambda_k = \kappa_k \left( \prod_{i=1}^N \nu_{ik}! \right) \prod_{i=1}^N \left( \frac{x_i}{\nu_{ik}} \right) \mathbf{1}_{\{x_i \geq \nu_{ik}\}}. \]

Intuitively, the mass action assumption reflects the idea that the system is well-stirred in the sense that all molecules are equally likely to be at any location at any time. For example, for a binary reaction

\[ S_1 + S_2 \rightarrow S_3 \quad \text{or} \quad S_1 + S_2 \rightarrow S_3 + S_4, \]

164
we take

\[ \lambda_k(x) = \kappa_k x_1 x_2, \]

where \( \kappa_k \) is a rate constant. For a unary reaction

\[ S_1 \rightarrow S_2 \quad \text{or} \quad S_1 \rightarrow S_2 + S_3, \]

we take

\[ \lambda_k(x) = \kappa_k x_1. \]

For a reaction with source \( \emptyset \), for example \( \emptyset \rightarrow S_1 \), which would represent an input to the system, the rate is simply

\[ \lambda_k(x) \equiv \kappa_k, \]

for some \( \kappa_k > 0 \). For a binary reaction with source complex of the form \( 2S_1 \), for example

\[ 2S_1 \rightarrow S_2, \]

we have

\[ \lambda_k(x) = \kappa_k x_1 (x_1 - 1), \]

which differs from deterministic mass action kinetics by the presence of the minus one.

### 6.1.3 Example: Gene transcription and translation

We give a series of models for the dynamical behavior of gene transcription and translation. Transcription is the process by which the information contained in a section of DNA is transferred to a newly assembled piece of messenger RNA (mRNA). Later this mRNA is translated by a ribosome, yielding proteins. We will give a series of three examples, with the first, Example 6.1.3, being the most basic and just include basic transcription, translation, and degradation. Next, in Example 6.1.4, we allow for the developed proteins to dimerize. Finally, in Example 6.1.5, we allow the resulting dimer to inhibit the production of the mRNA, and hence the protein and dimers themselves. This is an example of a negative feedback loop. We note that each of our models leave out many components, such as the RNA polymerase that is necessary for transcription, and the ribosomes that are critical for translation. Instead, we will assume that the abundances of such species are fixed and have been incorporated into the rate constants. More complicated, and realistic, models can incorporate both ribosomes and RNA polymerase.

We will model using a continuous time Markov chain. It is an entirely reasonable question to ask whether it makes sense to model the reaction times of such cellular processes via exponential random variables. The answer is almost undoubtably no. However, the model should be interpreted as an approximation to reality, and of course not reality itself, and has been quite successful in elucidating cellular dynamics. It is also a much more realistic model than a classical ODE approach, which is itself a crude approximation to the continuous time Markov chain model (we will discuss this fact later).
Example 6.1.3. Consider a single gene that is producing mRNA (this process is called *transcription*) at a constant rate of $\kappa_1$, where the units of time are hours, say. Further, we suppose the mRNA molecules are producing proteins (this process is called *translation*) at a rate of $\kappa_2 \cdot (\# \text{mRNA})$, for some $\kappa_2 > 0$. Next, we assume that the mRNA molecules are being degraded at a rate of $d_M \cdot (\# \text{mRNA})$, and proteins are being degraded at a rate of $d_P \cdot (\# \text{proteins})$. Graphically, we may represent this system via the four reactions

\begin{align*}
R1) & \quad G \xrightarrow{\kappa_1} G + M \\
R2) & \quad M \xrightarrow{\kappa_2} M + P \\
R3) & \quad M \xrightarrow{d_M} \emptyset \\
R4) & \quad P \xrightarrow{d_P} \emptyset
\end{align*}

We note that this is not the only way to write down these reactions. For example, many in the biological communities would write $M \rightarrow P$, as opposed to $M \rightarrow M + P$. However, we feel it is important to stress, through the notation $M \rightarrow M + P$, that the mRNA molecule is not lost during the course of the reaction.

As the number of genes in the model is assumed to be constant in time, the state space is $X(t) \in \mathbb{Z}_{\geq 0}^2$, where the first component gives the number of mRNA molecules and the second gives the number of proteins. That is, the species set is $S = \{M, P\}$. Note, therefore, that the first reaction is now viewed as $\emptyset \xrightarrow{\kappa_1} M$. The reaction vectors are

\begin{align*}
\zeta_1 &= \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \zeta_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \zeta_3 = \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \quad \zeta_4 = \begin{bmatrix} 0 \\ -1 \end{bmatrix},
\end{align*}

with respective rates

\begin{align*}
\kappa_1, \quad \kappa_2 X_1(t), \quad d_M X_1(t), \quad d_P X_2(t).
\end{align*}

Note that for this model, the reaction rates are linear functions of abundances. The stochastic equation governing $X(t)$ is

\begin{align*}
X(t) &= X(0) + Y_1 (\kappa_1 t) \zeta_1 + Y_2 \left( \kappa_2 \int_0^t X_1(s)ds \right) \zeta_2 \\
&\quad + Y_3 \left( d_M \int_0^t X_1(s)ds \right) \zeta_3 + Y_4 \left( d_P \int_0^t X_2(s)ds \right) \zeta_4,
\end{align*}

where $Y_i, i \in \{1, 2, 3, 4\}$ are independent unit-rate Poisson processes. Note that the rate of reaction 3, respectively 4, will be zero when $X_1(t) = 0$, respectively $X_2(t) = 0$. Therefore, non-negativity of the molecules is assured.

Example 6.1.4. We consider the previous example, but also allow for the possibility that the protein produces a dimer via the reaction $P + P \rightarrow D$, with rate constant
κ₅. Note that the set of species is now \{M, P, D\}. Letting \(X_3(t)\) denote the number of dimers at time \(t\), with \(X_1\) and \(X_2\) as before, the stochastic equation is now

\[
X(t) = X(0) + Y_1(\kappa_1 t) \begin{bmatrix} 1 \\ 0 \end{bmatrix} + Y_2 \left( \kappa_2 \int_0^t X_1(s) ds \right) \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\
+ Y_3 \left( \kappa_3 \int_0^t X_1(s) ds \right) \begin{bmatrix} -1 \\ 0 \end{bmatrix} + Y_4 \left( \kappa_4 \int_0^t X_2(s) ds \right) \begin{bmatrix} 0 \\ -1 \end{bmatrix} \\
+ Y_5 \left( \kappa_5 \int_0^t X_2(s)(X_2(s) - 1) ds \right) \begin{bmatrix} 0 \\ -2 \\ 1 \end{bmatrix},
\]

where \(Y_i, i \in \{1, \ldots, 5\}\), are independent unit-rate Poisson processes. □

**Example 6.1.5.** We consider the previous examples, but now allow for the dimer to interfere, or *inhibit*, with the production of the mRNA. That is, we assume the dimer can bind to the DNA at which point no mRNA can be produced. Because the resulting dimers inhibit their own production (through the mRNA), this is an example of a *negative feedback loop*. We must now explicitly model the DNA, in both a bound and unbound version. Therefore, we let \(X_4\) and \(X_5\) denote the total number of unbound, and bound, strands of DNA. We take \(X_4(t) + X_5(t) \equiv 1\). Note that the set of species is now \{M, P, D, G, B\}. We add the reactions corresponding to binding and unbinding to our model,

\[
G + D \xrightarrow{\kappa_{\text{on}}} B, \quad B \xrightarrow{\kappa_{\text{off}}} G + D,
\]

where \(\kappa_{\text{on}}, \kappa_{\text{off}} > 0\), and \(B\) represents the bound gene, and we recall that \(D\) represents
the dimer. The stochastic equations are now

\[
X(t) = X(0) + Y_1 \left( \kappa_1 \int_0^t X_4(s) \, ds \right) \begin{bmatrix}
1 \\
0 \\
0 \\
0 
\end{bmatrix} + Y_2 \left( \kappa_2 \int_0^t X_1(s) \, ds \right) \begin{bmatrix}
0 \\
1 \\
0 \\
0 
\end{bmatrix}
\]

\[
+ Y_3 \left( d_M \int_0^t X_1(s) \, ds \right) \begin{bmatrix}
-1 \\
0 \\
0 \\
0 
\end{bmatrix} + Y_4 \left( d_P \int_0^t X_2(s) \, ds \right) \begin{bmatrix}
0 \\
0 \\
-1 \\
0 
\end{bmatrix}
\]

\[
+ Y_5 \left( \kappa_5 \int_0^t X_2(s)(X_2(s) - 1) \, ds \right) \begin{bmatrix}
0 \\
-2 \\
1 \\
0 
\end{bmatrix}
\]

\[
+ Y_6 \left( \kappa_{on} \int_0^t X_4(s)X_3(s) \, ds \right) \begin{bmatrix}
0 \\
0 \\
-1 \\
-1 
\end{bmatrix} + Y_7 \left( \kappa_{on} \int_0^t X_5(s) \, ds \right) \begin{bmatrix}
0 \\
0 \\
0 \\
1 
\end{bmatrix} + Y_8 \left( \kappa_{on} \int_0^t X_6(s) \, ds \right) \begin{bmatrix}
0 \\
0 \\
0 \\
1 
\end{bmatrix}.
\]

Note that the rate of the first reaction has changed to incorporate the fact that mRNA molecules will only be produced when the gene is free. More general models include the possibility of more binding sites on the DNA molecule for the dimers. We note that this example can be easily modified to have the dimer only slow the rate of production, or even raise rate of production. If the rate of production is raised, then this would be an example of a positive feedback loop.

\[\square\]

6.1.4 Generator of the process and the forward equations

Throughout, we make the assumption that the processes under consideration are non-explosive. We derive the generator, \(A\), for the process (6.5). For \(h\) small we have

\[
\mathbb{E}_i f(X(h)) = \left( \sum_{k=1}^M f(i + \zeta_k) P_i \{ X(h) = i + \zeta_k \} \right) + f(i) P_i \{ X(h) = i \} + o(h)
\]

\[
= \left( \sum_{k=1}^M f(i + \zeta_k) \lambda_k(i) h + o(h) \right) + f(i) (1 - \lambda(i) h + o(h)) + o(h),
\]

where the first \(o(h)\) term incorporates the probability of having more than one reaction in the time interval of size \(h\). Since

\[
\lambda(i) = \sum_k \lambda_k(i),
\]

168
we have
\begin{align*}
\mathbb{E}_i f(X(h)) &= \left( \sum_{k=1}^{M} f(i + \zeta_k) \lambda_k(i) h + o(h) \right) + f(i) \left( 1 - \sum_{k=1}^{M} \lambda_k(i) h + o(h) \right) + o(h) \\
&= \sum_{k} \lambda_k(i) (f(i + \zeta_k) - f(i)) h + o(h),
\end{align*}

where we have collected the $o(h)$ terms. Therefore, for all $f$ we have that the generator satisfies
\begin{equation}
(Af)(i) = \sum_{k} \lambda_k(i) (f(i + \zeta_k) - f(i)). \tag{6.6}
\end{equation}

The generator can be viewed as giving us the correct chain-rule for our stochastic equations. For example, consider the related ODE system
\begin{equation}
\dot{x}(t) = \sum_{k} \lambda_k(x(t)) \zeta_k. \tag{6.7}
\end{equation}

Then, for any $f : \mathbb{R}^N \to \mathbb{R}$, we have that
\begin{equation*}
\frac{d}{dt} f(x(t)) = \sum_{k} \lambda_k(x(t)) \zeta_k \cdot \nabla f(x(t)),
\end{equation*}

Further,
\begin{equation*}
f(x(t)) = f(x(0)) + \int_0^t \left( \frac{d}{ds} f(x(s)) \right) ds.
\end{equation*}

For the stochastic equation Dynkin’s formula yields
\begin{equation*}
\mathbb{E} f(X(t)) = \mathbb{E} f(X(0)) + \mathbb{E} \int_0^t (Af)(X(s)) ds.
\end{equation*}

Noting that
\begin{equation*}
\lim_{h \to 0} \frac{f(x + h \zeta_k) - f(x)}{h} = \nabla f(x) \cdot \zeta_k
\end{equation*}
completes the analogy. Note that letting $x_i(t)$ be the solution to the ode (6.7) with initial condition $i$, it would be fair to call the operator
\begin{equation*}
(Bf)(i) = \lim_{h \to 0} \frac{f(x_i(h)) - f(i)}{h} = \sum_{k} \lambda_k(i) \zeta_k \cdot \nabla f(i)
\end{equation*}
the generator of the deterministic process.

We turn to the forward equations. Noting that $\lambda(y, j) \neq 0$ if and only if $y = j - \zeta_k$ for some $k$, we see that the forward equations are
\begin{align*}
\frac{d}{dt} P_i \{ X(t) = j \} &= \sum_{y \neq j} \lambda(y, j) P_i \{ X(t) = y \} - \lambda(j) P_i \{ X(t) = j \} \\
&= \sum_{k=1}^{M} \lambda_k(j - \zeta_k) P_i \{ X(t) = j - \zeta_k \} - \sum_{k=1}^{M} \lambda_k(j) P_i \{ X(t) = j \},
\end{align*}
or, more generally,
\[
\frac{d}{dt} P\{X(t) = j|\alpha\} = \sum_k \lambda_k (j - \zeta_k) P\{X(t) = j - \zeta_k|\alpha\} - \sum_k \lambda_k(x) P\{X(t) = j|\alpha\},
\]

(6.8)

where \(\alpha\) is the initial distribution. Equation (6.8) is often called the chemical master equation in the biological literature and is probably the most well known equation in those settings.

We can use the forward equation (6.8) to find the system of equations that must be satisfied by any stationary distribution. Setting the left hand side of (6.8) to zero, we see a stationary distribution \(\nu\) satisfies
\[
\sum_k \lambda_k(x - \zeta_k) \nu(x - \zeta_k) = \sum_k \lambda_k(x) \nu(x).
\]

Solving for such a \(\nu\) is, in general, non-trivial. In fact, it is an open problem to even classify those networks for which a solution exists. Later, we will see a large class of systems for which the equations can be solved.

### 6.1.5 General continuous time Markov chains built using random time changes

We point out that systems of the form (6.5) are, in fact, quite general. A continuous time Markov chain \(X\) taking values in \(\mathbb{R}^N\) is specified by giving its transition rates \(\lambda_l\) that determine
\[
P\{X(t + \Delta t) - X(t) = \zeta_l | \mathcal{F}_t^X\} \approx \lambda_l(X(t)) \Delta t,
\]

(6.9)

for the different possible jumps \(\zeta_l \in \mathbb{R}^N\), where \(\mathcal{F}_t^X\) is the \(\sigma\)-algebra generated by \(X\) (all the information available from the observation of the process up to time \(t\)). If we write
\[
X(t) = X(0) + \sum_l \zeta_l R_l(t)
\]

where \(R_l(t)\) is the number of jumps of \(\zeta_l\) at or before time \(t\), then (6.9) implies
\[
P\{R_l(t + \Delta t) - R_l(t) = 1 | \mathcal{F}_t^X\} \approx \lambda_l(X(t)) \Delta t, \quad l \in \mathbb{R}^N.
\]

\(R_l\) is a counting process with intensity \(\lambda_l(X(t))\), and by analogy with (6.4), we write
\[
X(t) = X(0) + \sum \zeta_l Y_l \left( \int_0^t \lambda_l(X(s)) ds \right),
\]

(6.10)

where the \(Y_l\) are independent unit-rate Poisson processes. This equation, also a random time change, has a unique solution by the same jump by jump argument used in Section 6.1.1, provided \(\sum_l \lambda_l(x) < \infty\) for all \(x\). Of course, as we know from Section 5.2, unless we add additional assumptions, we cannot rule out the possibility
that the solution only exists up to some finite time. For example, if $d = 1$ and $\lambda_1(k) = (1 + k)^2$, the solution of

$$X(t) = Y_1 \left( \int_0^t (1 + X(s))^2 ds \right)$$

hits infinity in finite time.

We present two examples that are not chemical in nature, but instead are population processes.

**Example 6.1.6.** We attempt to build a model for the behavior of predator-prey interactions. We denote the predator by $F$ (foxes) and the prey by $R$ (rabbits). We now consider what would make a reasonable model. We first note that because rabbits will reproduce, we have a transition of the general form

$$R \xrightarrow{\kappa_1} 2R.$$

This yields an intensity function of the form $\lambda_1(X_R) = \kappa_1 X_R$, and simply assumes the growth rate is proportional to the population size. We also have reproduction of the foxes. However, note that the rate of reproduction should be a function of the number of rabbits. Thus, we have

$$F \xrightarrow{\kappa_2 g(R)} 2F.$$

where $g(R)$ is some function of the number of rabbits. Said differently, we are claiming that the intensity of this transition should be of the form $\kappa_2 X_F g(X_R)$, where $X$ is the state of the system giving the numbers of each animal. It seems plausible that $g$ should be non-decreasing and $g(0) = 0$. For ease, we take $g(R) = R$, though do not try to provide a good argument for why. That is, we are just choosing something. If data is provided for an actual model, more could be said about the function $g$. Next, it should be that interactions between rabbits and foxes decrease the rabbit population. That is, we have a transition of the form

$$R + F \xrightarrow{\kappa_3} F.$$

Finally, we have death by natural causes

$$R \xrightarrow{\kappa_4} \emptyset, \quad F \xrightarrow{\kappa_5} \emptyset.$$

This example is famous for producing oscillations. For example, when we choose

$$\kappa_1 = 10, \quad \kappa_2 = 0.01, \quad \kappa_3 = 0.01, \quad \kappa_4 = 0.01, \quad \kappa_5 = 10,$$

and an initial condition of $X_R(0) = X_F(0) = 1,000$, we get dynamics as exemplified in Figure 6.1.1.

**Example 6.1.7.** We consider a basic stochastic model for the transmission of a disease in a population. We suppose that there are three types of people in the population:
Figure 6.1.1: The numbers of Rabbits and Foxes as a function of time for the predator-prey model of Example 6.1.6. The oscillatory behavior of this model is apparent.

1. Those that are susceptible to infection, denoted $S$.

2. Those that are infected, denote $I$.

3. Those that are recovered, and no longer susceptible, denote $R$.

Such models are commonly referred to as SIR models.

Letting $X = (X_S, X_I, X_R)$, natural transitions, including rates, for the model include

<table>
<thead>
<tr>
<th>Transition</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\emptyset \rightarrow S$</td>
<td>$\lambda_1(X) = \kappa_1(X_S + X_I + X_R)$</td>
</tr>
<tr>
<td>$S + I \rightarrow 2I$</td>
<td>$\lambda_2(X) = \kappa_2 X_S X_I$</td>
</tr>
<tr>
<td>$I \rightarrow R$</td>
<td>$\kappa_3 X_I$</td>
</tr>
<tr>
<td>$S \rightarrow \emptyset$</td>
<td>$\kappa_4 X_S$</td>
</tr>
<tr>
<td>$I \rightarrow \emptyset$</td>
<td>$\kappa_5 X_I$</td>
</tr>
<tr>
<td>$R \rightarrow \emptyset$</td>
<td>$\kappa_6 X_R$</td>
</tr>
</tbody>
</table>

Note the the first rate says that the rate of growth of the susceptible population is proportional to the size of the entire population (i.e. only healthy children are born). Such models are extensively studied in population health where the goal is to find the best treatment so that the disease goes extinct with some high probability. Obviously, for each disease, the parameters will be different and must be estimated. Natural conditions would be that $\kappa_5 \geq \kappa_4$ and $\kappa_5 \geq \kappa_6$. Depending on the time-scale

172
of the problem, it may be natural to assume that $\kappa_1 = \kappa_4 = \kappa_5 = \kappa_6 = 0$, leaving only the system
\[
S + I \xrightarrow{\kappa_3} 2I, \quad I \xrightarrow{\kappa_3} R.
\]
For example, if the infection being modeled is a cold then the time-frame may only be a few weeks. Obviously, this is not the case for a longer term infection such as HIV.

There are many variants to this model, including sophisticated models that take space (i.e. countries, rates of air travel, etc.) into account.

### 6.2 Simulation

Three algorithms for the numerical simulation of sample paths of stochastically modeled chemical reaction systems are provided below. In Section 6.2.1, the “Gillespie algorithm” or stochastic simulation algorithm is presented. This is simply Algorithm 2 from Section 5.1 for simulating the embedded discrete time Markov chain. It is redundant to give the algorithm here, but we present it for completeness. Next, in Section 6.2.2, the algorithm for simulating the random time change representation (6.5) is given. This method often goes by the name “the next reaction method.” Finally, in Section 6.2.3 a natural approximate algorithm is presented. The method is known as $\tau$-leaping, and is simply Euler’s method applied to the random time change representation (6.5). Throughout this section all random variables generated are assumed to be independent of each other and all previous random variables.

#### 6.2.1 The stochastic simulation algorithm

The stochastic simulation algorithm, or Gillespie’s algorithm, is simply Algorithm 2 of Section 5.1 applied the chemical context. See [20, 21] for historical references.

**Algorithm 3. (Gillespie Algorithm)**

**Initialize:** Set the initial number of molecules of each species and set $t = 0$.

1. Calculate the intensity function, $\lambda_k$, for each reaction.
2. Set $\lambda_0 = \sum_{k=1}^{M} \lambda_k$.
3. Generate two independent uniform(0,1) random numbers $r_1$ and $r_2$.
4. Set $\Delta = \ln(1/r_1)/\lambda_0$ (equivalent to drawing an exponential random variable with parameter $\lambda_0$).
5. Find $\mu \in [1, \ldots, M]$ such that
   \[
   (1/\lambda_0) \sum_{k=1}^{\mu-1} \lambda_k < r_2 \leq (1/\lambda_0) \sum_{k=1}^{\mu} \lambda_k,
   \]
   which is equivalent to choosing from reactions 1, $\ldots$, $M$, with the $k$th reaction having probability $\lambda_k/\lambda_0$. 173
6. Set $X(t + \Delta) = X(t) + \zeta_u$.

7. Set $t = t + \Delta$.

8. Return to step 1 or quit.

Note that Algorithm 3 uses two random numbers per step. The first is used to find when the next reaction occurs and the second is used to determine which reaction occurs at that time. That is, the second random variable (in step 5) simulates the next step of the embedded chain.

6.2.2 The next reaction method

The algorithm below simulates the random time change representation (6.5). The method is usually termed the next reaction method. See [18, 3]. The formal algorithm we present follows that of [3].

In Algorithm 4 below, the variable $T_k$ will represent the value of the integrated intensity function, $\int_0^t \lambda_k(X(s))ds$, where $t$ is the current time. Further, the variable $P_k$ will represent the first jump time of the process $Y_k$ after time $T_k$. That is,

$$P_k = \inf\{s > T_k \ni Y_k(s) > Y_k(T_k)\}.$$

Note that if $T_k$ happens to be equal to a jump time of $Y_k$, then $P_k - T_k$ is an exponential random variable with a parameter of one. Continuing, for each $k \in \{1, \ldots, M\}$ we will set $\Delta t_k$ to be the solution to the equation

$$\int_t^{t+\Delta t_k} \lambda(X(s))ds = P_k - T_k.$$

Under the assumption that no other reaction takes place before $t + \Delta t_k$, we see

$$\int_t^{t+\Delta t_k} \lambda(X(s))ds = \Delta t_k \lambda_k(X(t)) \implies \Delta t_k = (P_k - T_k)/\lambda_k(X(t)).$$

Finding the minimum of these $\Delta t_k$ then yields both the reaction (given by the index of the minimum) and the time until the next reaction takes place.

Algorithm 4. (Next Reaction Method)

Initialize: Set the initial number of molecules of each species. Set $t = 0$. For each $k \in \{1, \ldots, M\}$, set $P_k = \ln(1/r_k)$, where $r_k$ are independent uniform(0,1) random variables, and set $T_k = 0$.

1. Calculate the intensity function, $\lambda_k$, for each reaction.

2. For each $k$, set

$$\Delta t_k = \begin{cases} (P_k - T_k)/\lambda_k, & \text{if } \lambda_k \neq 0 \\ \infty, & \text{if } \lambda_k = 0 \end{cases}.$$
3. Set $\Delta = \min_k \{\Delta t_k\}$, and let $\mu$ be the index where the minimum is achieved.
4. Set $X(t + \Delta) = X(t) + \zeta_k$.
5. For each $k \in \{1, \ldots, M\}$, set $T_k = T_k + \lambda_k \cdot \Delta$.
6. Set $P_\mu = P_\mu + \ln(1/r)$, where $r$ is uniform(0,1).
7. Set $t = t + \Delta$.
8. Return to step 1 or quit.

Note that after initialization the Next Reaction Method only demands one random number to be generated per step.

**Time dependent intensity functions**

Due to changes in temperature and/or volume, the rate constants of a chemical system may change in time. Therefore, the intensity functions will no longer be constant between reactions. That is, we may have $\lambda_k(t) = \lambda_k(X(t), t)$, and the random time change representation is

$$X(t) = X(0) + \sum_k Y_k \left( \int_0^t \lambda_k(X(s), s) ds \right) \zeta_k,$$  

(6.11)

where the $Y_k$ are independent, unit rate Poisson processes. The next reaction method as presented in Algorithm 4 is easily modified to incorporate this time dependence. The only step that would change is step 2., which becomes:

2. For each $k$, find $\Delta t_k$ satisfying

$$\int_t^{t+\Delta t_k} \lambda_k(X(s), s) ds = P_k - T_k.$$ 

Note, in particular, that the integral ranges from $t$ to $t + \Delta t_k$. The remainder of the algorithm stays the same.

**6.2.3 Euler’s method**

We briefly review Euler’s method, termed tau-leaping in the chemical kinetic literature [22], as applied to the models (6.5). The basic idea of tau-leaping is to hold the intensity functions fixed over a time interval $[t_n, t_n + h]$ at the values $\lambda_k(X(t_n))$, where $X(t_n)$ is the current state of the system, and, under this assumption, compute the number of times each reaction takes place over this period. Analogously to (6.5), a path-wise representation of Euler tau-leaping defined for all $t \geq 0$ can be given through a random time change of Poisson processes:

$$Z_h(t) = Z_h(0) + \sum_k Y_k \left( \int_0^t \lambda_k(Z_h \circ \eta(s)) ds \right) \zeta_k,$$  

(6.12)
where the $Y_k$ are as before, and $\eta(s) \overset{\text{def}}{=} \left\lfloor \frac{s}{h} \right\rfloor h$. Thus, $Z_h \circ \eta(s) = Z_h(t_n)$ if $t_n \leq s < t_{n+1}$. Noting that

$$\int_0^{t_{n+1}} \lambda_k(Z_h \circ \eta(s))ds = \sum_{i=0}^{n} \lambda_k(Z_h(t_i))(t_{i+1} - t_i)$$

explains why this method is called Euler tau-leaping.

The following algorithm simulates (6.12) up to a time of $T > 0$. Below and in the sequel, for $x \geq 0$ we will write $\text{Poisson}(x)$ to denote a sample from the Poisson distribution with parameter $x$, with all such samples being independent of each other and of all other sources of randomness used.

**Algorithm 5. (Euler tau-leaping)**

**Initialize:** Fix $h > 0$. Set $Z_h(0) = x_0$, $t_0 = 0$, $n = 0$ and repeat the following until $t_n = T$:

1. Set $t_{n+1} = t_n + h$. If $t_{n+1} \geq T$, set $t_{n+1} = T$ and $h = T - t_n$.
2. For each $k$, let $\Lambda_k = \text{Poisson}(\lambda_k(Z_h(t_n))h)$.
3. Set $Z_h(t_{n+1}) = Z_h(t_n) + \sum_k \Lambda_k \zeta_k$.
4. Set $n \leftarrow n + 1$.

Several improvements and modifications have been made to the basic algorithm described above over the years. Some concern adaptive step-size selection along a path [10, 23]. Others focus on ensuring non-negative population values [3, 9, 11, 40]. We also note that it is straightforward to incorporate time dependence of the intensity functions $\lambda_k$ in the above algorithm; simply change step 2 to read

$$\Lambda_k = \text{Poisson} \left( \int_t^{t+h} \lambda_k(Z(t_n), s)ds \right),$$

or

$$\Lambda_k = \text{Poisson} \left( \lambda_k(Z(t_n), t_n)h \right),$$

with the choice depending upon the specific problem.

Historically, the time discretization parameter for Euler’s method has been $\tau$, leading to the name “$\tau$-leaping methods.” We break from this tradition so as not to confuse $\tau$ with a stopping time, and we denote our time-step by $h$ to be consistent with much of the numerical analysis literature.

### 6.3 Advanced topics in computing

#### 6.3.1 Numerically approximating expectations

This section is a condensed version of the research article [5], to which we point the interested reader for a more thorough treatment.
A common task in the study of stochastic models is to approximate \( \mathbb{E}f(X(T)) \), where \( f \) is a scalar-valued function of the state of the system which gives a measurement of interest. For example, the function \( f \) could be:

1. \( f(X(T)) = X_i(T) \), yielding estimates for mean values, or
2. \( f(X(T)) = X_i(T)X_j(t) \), which can be used with estimates for the mean values to provide estimates of variances (when \( i = j \)) and covariances (when \( i \neq j \)), or
3. \( f(X(T)) = 1_{\{X(T) \in B\}} \), the indicator function giving 1 if the state is in some specified set. Such functions could also be used to construct histograms, for example, since \( \mathbb{E}f(X(T)) = P\{X(T) \in B\} \).

We will consider different methods for solving this problem.

Suppose we use an exact simulation algorithm, such as Gillespie’s algorithm or the next reaction method, to approximate \( \mathbb{E}f(X(T)) \) to \( O(\epsilon) \) accuracy in the sense of confidence intervals (recall Section B.5). To do so, we need to generate \( n = O(\epsilon^{-2}) \) paths so that the standard deviation of the usual Monte Carlo estimator,

\[
\mu_n = \frac{1}{n} \sum_{j=1}^{n} f(X_{[j]}(T)),
\]

where \( X_{[j]} \) are independent realizations generated via an exact algorithm, is \( O(\epsilon) \). If we let \( N > 0 \) be the order of magnitude of the number of computations needed to produce a single sample path using an exact algorithm, then the total computational complexity becomes \( O(N\epsilon^{-2}) \).

When \( N \gg 1 \), which is the norm as opposed to the exception in many settings, it may be desirable to make use of an approximate algorithm, such as Euler’s method. Suppose

\[
\mathbb{E}f(X(T)) - \mathbb{E}f(Z_h(T)) = O(h), \tag{6.13}
\]

where \( Z_h \) is an approximate path generated with a time discretization parameter \( h \). That is, we assume we have a “weakly order one method” in that the left hand side of (6.13) (the bias) goes to zero at rate \( h \) as \( h \to 0 \). More generally, we would say we have a method that has a weak order of \( p > 0 \) if the bias goes to zero at a rate of \( h^p \), as \( h \to 0 \). We first make the trivial observation that the estimator

\[
\mu_n = \frac{1}{n} \sum_{j=1}^{n} f(Z_{h,[j]}(T)), \tag{6.14}
\]

where \( Z_{h,[j]} \) are independent paths generated via the approximate algorithm with a step size of \( h \), is an unbiased estimator of \( \mathbb{E}f(Z_h(T)) \), and not \( \mathbb{E}f(X(T)) \). However, noting that

\[
\mathbb{E}f(X(T)) - \mu_n = \left[ \mathbb{E}f(X(T)) - \mathbb{E}f(Z_h(T)) \right] + \left[ \mathbb{E}f(Z_h(T)) - \mu_n \right], \tag{6.15}
\]

we see that choosing \( h = O(\epsilon) \), so that the first term on the right of (6.15) is \( O(\epsilon) \), and \( n = O(\epsilon^{-2}) \), so that the standard deviation of the second term is \( O(\epsilon) \), delivers
the desired accuracy. With a fixed cost per time step, the computational complexity of generating a single such path is $O(\epsilon^{-1})$ and we find that the total computational complexity is $O(\epsilon^{-3})$. Second order methods lower the computational complexity to $O(\epsilon^{-2.5})$, as $h$ may be chosen to be $O(\epsilon^{1/2})$.

The discussion above suggests that the choice between exact or approximate path computation should depend upon whether $\epsilon^{-1}$ or $N$ is the larger value, with an exact algorithm being beneficial when $N < \epsilon^{-1}$. It is again worth noting, however, that the estimators built from approximate methods are biased, and while analytic bounds can be provided for that bias [4, 6, 33] these are typically neither sharp nor computable, and hence of limited practical value. The exact algorithm, on the other hand, trivially produces an unbiased estimator, so it may be argued that $\epsilon^{-1} \ll N$ is necessary before it is worthwhile to switch to an approximate method.

We will now introduce a more sophisticated method for the approximation of expected values. Called multi-level Monte Carlo, the method was originally developed by Mike Giles [19], with related earlier work by Heinrich [24], in the diffusive setting (i.e. stochastic equations incorporating Brownian motions). In that setting, multi-level Monte Carlo has the remarkable property of lowering the standard $O(\epsilon^{-3})$ cost of computing an $O(\epsilon)$ accurate Monte Carlo estimate of $E_f(X(T))$ down to $O(\epsilon^{-2} \log(\epsilon)^2)$ [19]. Here, we are assuming that a weak order one and strong order $1/2$ discretization method, such as Euler–Maruyama, is used. (See, for example [25], which is the classical text on numerical methods for stochastic differential equations incorporating Brownian motions, for an explanation Euler-Maruyama in the diffusive setting.)

**Multi-level Monte Carlo**

At its heart, multi-level Monte Carlo is a control variate method to lower the variance of the estimator for $E_f(X(t))$, and the basic idea is as follows. Supposing we want to estimate $E_f(X(t))$, we could use the estimator

$$E_f(X(t)) \approx \frac{1}{n} \sum_{i=1}^n f(X_i(t)), \quad (6.16)$$

where $X_i(t)$ are independent copies of $X(t)$, though we assume that realizations of $X$ are (relatively) expensive. However, suppose that

$$f(X(t)) \approx f(Z_L(t)),$$

for some approximate process $Z_L$, and $Z_L$ is cheap to generate (the subscript $L$ will be explained below). Further, suppose $X, Z_L$ can be generated simultaneously so that

$$\text{Var}(f(X(t)) - f(Z_L(t)))$$

is small. Then simply note that

$$E_f(X(t)) = E[f(X(t)) - f(Z_L(t))] + E_f(Z_L(t))$$

$$\approx \frac{1}{n_1} \sum_{i=1}^{n_1} (f(X_i(t)) - f(Z_{L,i}(t))) + \frac{1}{n_2} \sum_{i=1}^{n_2} f(Z_{L,i}(t)), \quad (6.17)$$

178
where the right hand side of (6.17) has two estimators, one for \( \mathbb{E}[f(X(t)) - f(Z_L(t))] \) and the other for \( \mathbb{E}f(Z_L(t)) \). However, that the variance of the first estimator is
\[
\frac{1}{n_1} \text{Var}(f(X(t)) - f(Z_L(t)));
\]
which, because \( \text{Var}(f(X(t)) - f(Z_L(t))) \) itself is assumed small, will not require a large \( n_1 \) to be very small. Therefore, while each path may be expensive, we will not need very many of them. On the other hand, the second estimator on the right hand side of (6.17) may still require a very large number of paths, \( n_2 \). However these will be cheap to generate, and so, again, not computationally expensive. Combined, we see that it is reasonable to hope that the full computational complexity required to use the estimator (6.17) could be substantially lower than that needed using crude Monte Carlo (6.16).

The above is a basic control variate idea. The essence of multi-level Monte Carlo is that you now keep going by using another control variate for \( Z_L \), call it \( Z_{L-1} \), and then another for \( Z_{L-1} \), etc. That is, we use
\[
\mathbb{E}f(X(t)) = \mathbb{E}(f(X(t)) - f(Z_L(t))) + \mathbb{E}f(Z_L(t))
\]
\[
= \mathbb{E}(f(X(t)) - f(Z_L(t))) + \mathbb{E}(f(Z_L(t)) - f(Z_{L-1}(t))) + \mathbb{E}f(Z_{L-1}(t))
\]
\[
\vdots
\]
\[
= \mathbb{E}(f(X(t)) - f(Z_L(t))) + \sum_{\ell=0+1}^{L} \mathbb{E}[f(Z_{\ell}(t)) - f(Z_{\ell-1}(t))] + \mathbb{E}[f(Z_{0}(t))].
\]
(6.18)

For now we assume that we can couple (i.e. simultaneously generate) the processes \( X, Z_L \) and \( Z_\ell, Z_{\ell-1} \) in such a way that the variance of their respective differences are small (we will show how to do this below). For some choices of \( n_0, \{n_\ell\}_{\ell=0+1}^{L} \), and \( n_E \), we define the estimators for the terms in (6.18):
\[
\hat{Q}_E \overset{\text{def}}{=} \frac{1}{n_E} \sum_{i=1}^{n_E} (f(X_{[i]}(t)) - f(Z_{L,[i]}(t))),
\]
\[
\hat{Q}_{\ell} \overset{\text{def}}{=} \frac{1}{n_{\ell}} \sum_{i=1}^{n_{\ell}} (f(Z_{\ell,[i]}(t)) - f(Z_{\ell-1,[i]}(t))), \quad \text{for } \ell \in \{\ell_0 + 1, \ldots, L\}
\]
\[
\hat{Q}_0 \overset{\text{def}}{=} \frac{1}{n_0} \sum_{i=1}^{n_0} f(Z_{0,[i]}(t)),
\]
and note that
\[
\hat{Q} \overset{\text{def}}{=} \hat{Q}_E + \sum_{\ell=0+1}^{L} \hat{Q}_{\ell} + \hat{Q}_0,
\]
is an unbiased estimator for \( \mathbb{E}f(X(t)) \).

We have not yet said what the approximate process \( Z_\ell \) should be. In fact, it can be nearly anything. The only criteria it must satisfy is that (i) we must be able to
couple it with $X$, and (ii) it must be easy to generate. We choose to let $Z_\ell$ be the \tau-leap process (6.12) with step size $h_\ell = 1/M^\ell$, for some reasonable choice of $M \geq 2$. Typically, $M \in \{2, 3, 4\}$.

The most important question remains: how should we actually generate the coupled processes $(X, Z_L)$ and $(Z_\ell, Z_{\ell-1})$. We motivate our choice of coupling by first treating two simpler tasks. First, consider the problem of trying to understand the difference between $Z_1(t)$ and $Z_2(t)$, where $Z_1, Z_2$ are homogeneous Poisson processes with rates $13.1$ and $13$, respectively. A simple approach is to let $Y_1$ and $Y_2$ be independent, unit-rate Poisson processes, set

$$Z_1(t) = Y_1(13.1t) \quad \text{and} \quad Z_2(t) = Y_2(13t),$$

and consider $Z_1(t) - Z_2(t)$. Using this representation, these processes are independent and, hence, not coupled. Further, the variance of their difference is the sum of their variances, and so

$$\text{Var}(Z_1(t) - Z_2(t)) = \text{Var}(Z_1(t)) + \text{Var}(Z_2(t)) = 26.1t.$$  

Another choice is to let $Y_1$ and $Y_2$ be independent unit-rate Poisson processes, and set

$$Z_1(t) = Y_1(13t) + Y_2(0.1t) \quad \text{and} \quad Z_2(t) = Y_2(13t),$$

where we have used the additivity property of Poisson processes. The important point to note is that both $Z_1$ and $Z_2$ are using the process $Y_1(13t)$ to generate simultaneous jumps. The process $Z_1$ then uses the auxiliary process $Y_2(0.1t)$ to jump the extra times that $Z_2$ does not. The processes $Z_1$ and $Z_2$ will jump together the vast majority of times, and hence are tightly coupled; by construction $\text{Var}(Z_1(t) - Z_2(t)) = \text{Var}(Y_2(0.1t)) = 0.1t$. More generally, if $Z_1$ and $Z_2$ are instead non-homogeneous Poisson processes with intensities $f(t)$ and $g(t)$, respectively, then we could let $Y_1, Y_2,$ and $Y_3$ be independent, unit-rate Poisson processes and define

$$Z_1(t) = Y_1 \left( \int_0^t f(s) \land g(s) ds \right) + Y_2 \left( \int_0^t f(s) - (f(s) \land g(s)) ds \right),$$

$$Z_2(t) = Y_1 \left( \int_0^t f(s) \land g(s) ds \right) + Y_3 \left( \int_0^t g(s) - (f(s) \land g(s)) ds \right),$$

where we are using that, for example,

$$Y_1 \left( \int_0^t f(s) \land g(s) ds \right) + Y_2 \left( \int_0^t f(s) - (f(s) \land g(s)) ds \right) \overset{D}{=} Y \left( \int_0^t f(s) ds \right),$$

where $Y$ is a unit rate Poisson process and we define $a \land b \overset{\text{def}}{=} \min\{a, b\}$.

We now return to the main problem of coupling the processes $X$ and $Z_L$, and the processes $Z_\ell$ and $Z_{\ell-1}$, each satisfying (6.12) with their respective step-sizes. We
couple the processes $Z_\ell$ and $Z_{\ell-1}$ by generating them in the following manner:

$$Z_\ell(t) = Z_\ell(0) + \sum_k Y_{k,1} \left( \int_0^t \lambda_k(Z_\ell \circ \eta_\ell(s)) \wedge \lambda_k(Z_{\ell-1} \circ \eta_{\ell-1}(s)) \, ds \right) \zeta_k$$

$$+ \sum_k Y_{k,2} \left( \int_0^t \lambda_k(Z_\ell \circ \eta_\ell(s)) - \lambda_k(Z_{\ell-1} \circ \eta_{\ell-1}(s)) \, ds \right) \zeta_k,$$

$$Z_{\ell-1}(t) = Z_{\ell-1}(0) + \sum_k Y_{k,1} \left( \int_0^t \lambda_k(Z_{\ell-1} \circ \eta_{\ell-1}(s)) \wedge \lambda_k(Z_{\ell-1} \circ \eta_{\ell-1}(s)) \, ds \right) \zeta_k$$

$$+ \sum_k Y_{k,3} \left( \int_0^t \lambda_k(Z_{\ell-1} \circ \eta_{\ell-1}(s)) - \lambda_k(Z_{\ell-1} \circ \eta_{\ell-1}(s)) \, ds \right) \zeta_k,$$

(6.19)

where the $Y_{k,i}$, $i \in \{1, 2, 3\}$, are independent, unit-rate Poisson processes, and for each \( \ell \), we define $\eta_\ell(s) \overset{\text{def}}{=} [s/h_\ell]h_\ell$. Note that we essentially used the coupling of the simpler examples above (pertaining to $Z_1$ and $Z_2$) for each of the reaction channels. The paths of the coupled processes can easily be computed simultaneously using Algorithm 6 below, and the distributions of the marginal processes $Z_\ell$ and $Z_{\ell-1}$ are the same as the usual Euler approximate paths (6.12) with similar step-sizes. (This can be seen by thinking about the holding times in each state for each process. This is an exercise worth doing.)

**Algorithm 6** (Simulation of the representation (6.19)). Fix an integer $M \geq 2$. Fix $h_\ell > 0$ and set $h_{\ell-1} = M \times h_\ell$. Set $Z_\ell(0) = Z_{\ell-1}(0) = x_0$, $t_0 = 0$, $n = 0$. Repeat the following steps until $n \geq T$:

(i) For $j = 0, \ldots, M - 1$,

(a) Set

- $A_{k,1} = \lambda_k(Z_\ell(t_n + j \times h_\ell)) \wedge \lambda_k(Z_{\ell-1}(t_n))$.
- $A_{k,2} = \lambda_k(Z_\ell(t_n + j \times h_\ell)) - A_{k,1}$.
- $A_{k,3} = \lambda_k(Z_{\ell-1}(t_n)) - A_{k,1}$.

(b) For each $k$, let

- $\Lambda_{k,1} = \text{Poisson}(A_{k,1}h_\ell)$.
- $\Lambda_{k,2} = \text{Poisson}(A_{k,2}h_\ell)$.
- $\Lambda_{k,3} = \text{Poisson}(A_{k,3}h_\ell)$.

(c) Set

- $Z_\ell(t_n + (j + 1) \times h_\ell) = Z_\ell(t_n + j \times h_\ell) + \sum_k (\Lambda_{k,1} + \Lambda_{k,2}) \zeta_k$.
- $Z_{\ell-1}(t_n + (j + 1) \times h_\ell) = Z_{\ell-1}(t_n + j \times h_\ell) + \sum_k (\Lambda_{k,1} + \Lambda_{k,3}) \zeta_k$.

(ii) Set $t_{n+1} = t_n + h_{\ell-1}$.

(iii) Set $n \leftarrow n + 1$. 

181
Similarly, we generate $X$ and $Z_\ell$ simultaneously via

\[ X(t) = X(0) + \sum_k Y_{k,1} \left( \int_0^t \lambda_k(X(s)) \wedge \lambda_k(Z_\ell \circ \eta_\ell(s)) ds \right) \zeta_k \]

\[ + \sum_k Y_{k,2} \left( \int_0^t \lambda_k(X(s)) - \lambda_k(X(s)) \wedge \lambda_k(Z_\ell \circ \eta_\ell(s)) ds \right) \zeta_k, \]

\[ Z_\ell(t) = Z_\ell(0) + \sum_k Y_{k,1} \left( \int_0^t \lambda_k(X(s)) \wedge \lambda_k(Z_\ell \circ \eta_\ell(s)) ds \right) \zeta_k \]

\[ + \sum_k Y_{k,3} \left( \int_0^t \lambda_k(Z_\ell \circ \eta_\ell(s)) - \lambda_k(X(s)) \wedge \lambda_k(Z_\ell \circ \eta_\ell(s)) ds \right) \zeta_k, \]

where all notation is as before. Note that the distributions of the marginal processes $X$ and $Z_\ell$ are equal to those of (6.5) and (6.12). The natural algorithm to simulate (6.20) is the next reaction method, where the system is viewed as having dimension $2N$ with state $(X, Z_\ell)$, and each of the “next reactions” must be calculated over the Poisson processes $Y_{k,1}, Y_{k,2}, Y_{k,3}$. This version of the algorithm is formally given below.

**Algorithm 7. (Simulation of the representation (6.20))**

**Initialize.** Fix $h_\ell > 0$. Set $X(0) = Z_\ell(0) = x_0$ and $t = 0$. Set $\tilde{Z}_\ell = Z_\ell(0)$. Set $T_{\text{tau}} = h_\ell$. For each $k \in \{1, \ldots, R\}$ and $i \in \{1, 2, 3\}$, set $P_{k,i} = \ln(1/r_{k,i})$, where $r_{k,i}$ is rand$(0, 1)$, and $T_{k,i} = 0$.

(i) For each $k$, set

- $A_{k,1} = \lambda_k(X(t)) \wedge \lambda_k(\tilde{Z}_\ell)$.
- $A_{k,2} = \lambda_k(X(t)) - A_{k,1}$.
- $A_{k,3} = \lambda_k(\tilde{Z}_\ell) - A_{k,1}$.

(ii) For each $k \in \{1, \ldots, R\}$ and $i \in \{1, 2, 3\}$, set

\[ \Delta t_{k,i} = \begin{cases} \frac{(P_{k,i} - T_{k,i})}{A_{k,i}}, & \text{if } A_{k,i} \neq 0 \\ \infty, & \text{if } A_{k,i} = 0 \end{cases} \]

(iii) Set $\Delta = \min_{k,i} \{\Delta t_{k,i}\}$, and let $\mu \equiv \{k, i\}$ be the indices where the minimum is achieved.

(iv) If $t + \Delta \geq T_{\text{tau}}$,

(a) Set $\tilde{Z}_\ell = Z_\ell(t)$.

(b) For each $k \in \{1, \ldots, R\}$ and $i \in \{1, 2, 3\}$, set $T_{k,i} = T_{k,i} + A_{k,i} \times (T_{\text{tau}} - t)$.

(c) Set $t = T_{\text{tau}}$.

(d) Set $T_{\text{tau}} = T_{\text{tau}} + h_\ell$.

(e) Return to step (i) or quit.
Table 6.1: Performance of exact algorithm with crude Monte Carlo for Example 6.3.1.

The mean number of dimers at time 1 is reported with 95% a confidence interval. (v) Else,

(a) Update. For \( \{k, i\} = \mu \), where \( \mu \) is from (iii),

- If \( i = 1 \), set \( X(t + \Delta) = X(t) + \zeta_k \) and \( Z_{\ell}(t + \Delta) = Z_{\ell}(t) + \zeta_k \).
- If \( i = 2 \), set \( X(t + \Delta) = X(t) + \zeta_k \).
- If \( i = 3 \), set \( Z_{\ell}(t + \Delta) = Z_{\ell}(t) + \zeta_k \).

(b) For each \( k \in \{1, \ldots, R\} \) and \( i \in \{1, 2, 3\} \), set \( T_{k,i} = T_{k,i} + A_{k,i} \times \Delta \).

(c) Set \( P_\mu = P_\mu + \ln(1/r) \), where \( r \) is rand(0, 1), and \( \mu \) is from (iii).

(d) Set \( t = t + \Delta \).

(e) Return to step (i) or quit.

Returning to the main problem, we see we have all the necessary pieces to build our estimators. In [5], this unbiased MLMC method is theoretically analyzed, and many more implementation issues are discussed. The theory in [5], together with examples, shows that MLMC will decrease the computational complexity required substantially for almost all examples. Of course, implementation of this multi-level Monte Carlo method is more difficult than using the crude Monte Carlo estimator, so there is a natural tradeoff in terms of implementation time versus CPU time. This issue will only be resolved conclusively after good open source software becomes available for the multi-level Monte Carlo method.

We close with an example demonstrating the performance of the method. This example, with even more detail, can be found in [5].

**Example 6.3.1.** Consider the following model of gene transcription and translation:

\[
G \xrightarrow{25} G + M, \quad M \xrightarrow{1000} M + P, \quad P + P \xrightarrow{0.001} D, \quad M \xrightarrow{0.1} \emptyset, \quad P \xrightarrow{1} \emptyset.
\]

Here, a single gene is being transcribed into mRNA, which is then being translated into proteins, and finally the proteins produce stable dimers. The final two reactions represent degradation of mRNA and proteins, respectively. Suppose the system starts with one gene and no other molecules, so \( X(0) = (1, 0, 0) \) where \( X_1, X_2, X_3 \) give the molecular counts of the mRNA, proteins, and dimers, respectively. Finally, suppose that we want to estimate the expected number of dimers at time \( T = 1 \) to an accuracy of ±1 with 95% confidence. Thus, we want the variance of our estimator to be smaller than \( (1/1.96)^2 \approx .2603 \).

Solving this problem using the next reaction method (Algorithm 4) took nearly 41 hours and required nearly five million sample paths, see Table 6.1. The results of

<table>
<thead>
<tr>
<th>Mean</th>
<th># paths</th>
<th>CPU Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>3714.2 ± 1.0</td>
<td>4,740,000</td>
<td>1.49 × 10^6 CPU S</td>
</tr>
</tbody>
</table>


Table 6.2: Performance of Euler tau-leaping with crude Monte Carlo for Example 6.3.1. The bias of the method is apparent.

<table>
<thead>
<tr>
<th>Step-size</th>
<th>Mean</th>
<th># paths</th>
<th>CPU Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h = 3^{-7}$</td>
<td>3,712.3 ± 1.0</td>
<td>4,750,000</td>
<td>13,374.6 S</td>
</tr>
<tr>
<td>$h = 3^{-6}$</td>
<td>3,707.5 ± 1.0</td>
<td>4,750,000</td>
<td>6,207.9 S</td>
</tr>
<tr>
<td>$h = 3^{-5}$</td>
<td>3,693.4 ± 1.0</td>
<td>4,700,000</td>
<td>2,803.9 S</td>
</tr>
<tr>
<td>$h = 3^{-4}$</td>
<td>3,655.2 ± 1.0</td>
<td>4,650,000</td>
<td>1,219.0 S</td>
</tr>
</tbody>
</table>

Table 6.3: Performance of unbiased MLMC with $\ell_0 = 2$, and $M$ and $L$ detailed above for Example 6.3.1.

<table>
<thead>
<tr>
<th>Step-size parameters</th>
<th>Mean</th>
<th>CPU Time</th>
<th>Var. of estimator</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M = 3, L = 6$</td>
<td>3,713.9 ± 1.0</td>
<td>1,063.3 S</td>
<td>0.2535</td>
</tr>
<tr>
<td>$M = 3, L = 5$</td>
<td>3,714.7 ± 1.0</td>
<td>1,114.9 S</td>
<td>0.2565</td>
</tr>
<tr>
<td>$M = 3, L = 4$</td>
<td>3,714.2 ± 1.0</td>
<td>1,656.6 S</td>
<td>0.2595</td>
</tr>
<tr>
<td>$M = 4, L = 4$</td>
<td>3,714.2 ± 1.0</td>
<td>1,334.8 S</td>
<td>0.2580</td>
</tr>
<tr>
<td>$M = 4, L = 5$</td>
<td>3,713.8 ± 1.0</td>
<td>1,014.9 S</td>
<td>0.2561</td>
</tr>
</tbody>
</table>

It is instructive to give more details for at least one choice of $M$ and $L$ for the unbiased MLMC estimator. For the case with $M = 3, L = 5$, and $\ell_0 = 2$, Table 6.4 provides the relevant data for the different levels. Note that most of the CPU time was taken up at the coarsest level, as is common with MLMC. Also, while the exact algorithm with crude Monte Carlo demanded the generation of almost five million exact sample paths, only 3,900 such paths were required at the finest level of the MLMC estimator. This difference is the main reason for the dramatic reduction in CPU time. Of course, the MLMC method required more than eight million paths at the coarsest level, but these paths are very cheap to generate.

6.4 First order reaction networks

We briefly discuss first order reaction networks. A system is said to be a first order reaction network if each intensity function $\lambda_k$ is linear. In the chemical context with mass action kinetics, the system is linear if and only if each reaction is either unary,
<table>
<thead>
<tr>
<th>Level</th>
<th># paths</th>
<th>Mean</th>
<th>Var. estimator</th>
<th>CPU Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>((X, Z_{3-5}))</td>
<td>3,900</td>
<td>20.1</td>
<td>0.0658</td>
<td>279.6 S</td>
</tr>
<tr>
<td>((Z_{3-5}, Z_{3-4}))</td>
<td>30,000</td>
<td>39.2</td>
<td>0.0217</td>
<td>49.0 S</td>
</tr>
<tr>
<td>((Z_{3-4}, Z_{3-3}))</td>
<td>150,000</td>
<td>117.6</td>
<td>0.0179</td>
<td>71.7 S</td>
</tr>
<tr>
<td>((Z_{3-3}, Z_{3-2}))</td>
<td>510,000</td>
<td>350.4</td>
<td>0.0319</td>
<td>112.3 S</td>
</tr>
<tr>
<td>Euler, (h = 3^{-2})</td>
<td>8,630,000</td>
<td>3,187.4</td>
<td>0.1192</td>
<td>518.4 S</td>
</tr>
<tr>
<td>Totals</td>
<td>N.A.</td>
<td>3,714.7</td>
<td>0.2565</td>
<td>1031.0 S</td>
</tr>
</tbody>
</table>

Table 6.4: Details of the different levels for the implementation of the unbiased MLMC method with \(M = 3\), \(L = 5\), and \(\ell_0 = 2\) for Example 6.3.1. By \((X, Z_{3-5})\) we mean the level in which the exact process is coupled to the approximate process with \(h = 3^{-5}\), and by \((Z_{3-\ell}, Z_{3-\ell+1})\) we mean the level with \(Z_{3-\ell}\) coupled to \(Z_{3-\ell+1}\). The value Mean gives the mean difference for all but the Euler step.

\(S_i \rightarrow \ast\), where “\(\ast\)” could refer to any linear combination of the species, or of the form \(\emptyset \rightarrow \ast\). Note that in this case we have the identity

\[
\mathbb{E}\lambda_k(X(s)) = \lambda_k(\mathbb{E}X(s)),
\]

for each \(k\). Therefore,

\[
\mathbb{E}X(t) = \mathbb{E}X(0) + \sum_k \mathbb{E}Y_k \left( \int_0^t \lambda_k(X(s))ds \right) \zeta_k
\]

\[
= \mathbb{E}X(0) + \sum_k \zeta_k \int_0^t \mathbb{E}\lambda_k(X(s))ds
\]

\[
= \mathbb{E}X(0) + \sum_k \zeta_k \int_0^t \lambda_k(\mathbb{E}X(s))ds,
\]

where we used that \(\mathbb{E}Y_k(u) = u\) for each process \(Y_k\).\(^1\) This gives a very easy method for solving for the means: just solve the associated ordinary differential equation implied by the above integral equation.

**Example 6.4.1.** We consider a model of gene transcription and translation

\[G \xrightarrow{2} G + M, \quad M \xrightarrow{10} M + P, \quad M \xrightarrow{k} \emptyset, \quad P \xrightarrow{1} \emptyset,\]

(6.21)

where a single gene is being translated into mRNA, which is then being transcribed into proteins. The final two reactions represent degradation of the mRNA and protein molecules, respectively. Assuming that there is a single gene copy, the stochastic equation for this model is

\[
X^k(t) = X^k(0) + Y_1(2t) \left( \begin{array}{c} 1 \\ 0 \end{array} \right) + Y_2 \left( \int_0^t 10X_1^k(s)ds \right) \left( \begin{array}{c} 0 \\ 1 \end{array} \right) \\
+ Y_3 \left( \int_0^t kX_1^k(s)ds \right) \left( \begin{array}{c} -1 \\ 0 \end{array} \right) + Y_4 \left( \int_0^t X_2^k(s)ds \right) \left( \begin{array}{c} 0 \\ -1 \end{array} \right),
\]

(6.22)

\(^1\)We are actually using the fact that \(\mathbb{E}Y_k(\tau) = \mathbb{E}\tau\) for any stopping time \(\tau\).
where \( X_1^k(t) \) and \( X_2^k(t) \) give the number of mRNA and protein molecules at time \( t \), respectively, and the \( Y_i, i \in \{1, \ldots, 4\} \), are independent unit-rate Poisson processes. The differential equations governing the mean values are

\[
\begin{align*}
\dot{x}_1(t) &= 2 - k x_1(t) \\
\dot{x}_2(t) &= 10 x_1(t) - x_2(t).
\end{align*}
\]

Solving this system yields

\[
\begin{align*}
x_1(t) &= x_1(0) e^{-kt} + \frac{2}{k} (1 - e^{-kt}) \\
x_2(t) &= e^{-t} \left( x_2(0) - 10 \frac{k x_1(0) - 2k}{k (1 - k)} \right) + 10 \frac{x_1(0) e^{-kt} k + 2 - 2k - 2 e^{-kt}}{k (1 - k)}.
\end{align*}
\]

We note that if \( \lambda_k \) is non-linear, then \( \mathbb{E} \lambda_k(X(s)) \neq \lambda_k(\mathbb{E} X(s)) \), and the mean value of the stochastic process does not satisfy the ordinary differential equation

\[
\dot{x}(t) = \sum_k \lambda_k(x(t)) \zeta_k. \quad (6.23)
\]

In the next section we will show, however, when a scaled version of the stochastic process can be shown to be well approximated by the solution to the differential equation (6.23).

For more information pertaining to linear systems, including the computation of the second moments and covariances, see [7] or [17].

### 6.5 Relationship with Deterministic Model: the Law of Large Numbers

We want to explore the relationship between the stochastic and deterministic models for biochemical systems. Recall that the associated ordinary differential equation for the stochastic process (6.5) is

\[
\dot{x}(t) = \sum_k \hat{\lambda}_k(x(t)) \zeta_k,
\]

where \( \hat{\lambda}_k \) is deterministic mass action kinetics. That is,

\[
\hat{\lambda}_k(x) = \hat{\kappa}_k \prod_i x_i^{\nu_{ik}},
\]

where \( \hat{\kappa}_k \) is the rate constant, and we recall that \( \nu_{ik} \) is the number of molecules required of species \( S_i \) for the \( k \)th reaction.
6.5.1 Conversion of rate constants

To understand the relationship between the stochastic and deterministic models, we must first note that they differ in how they are representing the abundance of each species. In the stochastic model, the abundance is an integer representing the number of molecules present. However, in the usual deterministic model it is the concentration of the species that are being modeled, for example in moles per liter. Therefore, we begin trying to understand the relationship between the two models by explicitly taking the volume into account in the stochastic system, and will do so by introducing a scaling parameter $V$, which is defined to be the volume multiplied by Avogadro’s number.

Consider the stochastic rate constants, $\kappa_k$, used in mass-action kinetics. Explicitly thinking of $V$ as proportional to the volume of the system, we see that for a binary reaction the rate constant $\kappa_k$ should satisfy

$$\kappa_k = \frac{1}{V} \tilde{\kappa}_k,$$

for some constant $\tilde{\kappa}_k$. This follows since $\kappa_k$ is assumed to be proportional to the probability that a particular pair of molecules interact within a small time window, and this probability should intuitively scale inversely with the volume. For unary reactions the volume should not affect the associated probability of a reaction, and so $\kappa_k = \tilde{\kappa}_k$. In general, we have

$$\kappa_k = V^{-(\sum_i \nu_{ik}-1)} \tilde{\kappa}_k,$$

(6.24)

where $\nu_{ik}$ is the number of molecules of species $S_i$ required for the $k$th reaction channel. See, for example, [27].

We will show the relation (6.24) in a second way that follows [42]. Note that if $x$ gives the concentration of a particular species in moles per unit volume, then $Vx$ gives the total number of molecules present. We first consider the zeroth order reaction

$$\emptyset \rightarrow S.$$

The rate of change induced by this reaction for the deterministic model is $\tilde{\kappa} Ms^{-1}$. Thus, if $X$ represents the number of molecules of species $S$, this reaction increases $X$ at a rate of

$$\tilde{\kappa} V$$

molecules per second. Because the stochastic rate law is $\kappa$ molecules per second, we have that

$$\kappa = V \tilde{\kappa}.$$

Now consider the first order reaction

$$S \rightarrow \ast,$$

where “$\ast$” can represent any linear combination of the species. The deterministic rate is $\tilde{\kappa}x Ms^{-1}$. Multiplying the rate $\tilde{\kappa}x$ by $V$, we see that $X$, the number of molecules of $S$, is changing at a rate of

$$\tilde{\kappa}xV,$$
molecules per second. However, since $X = xV$, we have that $X$ is changing at a rate $\hat{\kappa}X$. Since the stochastic rate law for this reaction is $\kappa X$ molecules per second, we see $\kappa = \hat{\kappa}$.

Finally, consider the binary reaction

$$S_1 + S_2 \rightarrow \ast.$$ 

The deterministic rate is $cx_1x_2 \, M \, s^{-1}$, implying the rate of change is $\hat{\kappa}x_1x_2V$ molecules per second. Since $X_1 = x_1V$ and $X_2 = x_2V$ give the number of molecules of $S_1$ and $S_2$, respectively, we see that the rate of change of the $X_i$ is

$$\hat{\kappa}x_1x_2V = \hat{\kappa} \frac{1}{V} X_1 X_2,$$

implying $\kappa = V^{-1}\hat{\kappa}$.

The above arguments again confirm the relation (6.24) for the most common types of reactions. Similar arguments work for all higher order reactions.

### 6.5.2 The classical scaling

The following scaling argument is commonly referred to as the classical scaling. It shows how to understand the relationship between the stochastic and deterministic models. See [27] or [30] for technical details and full statements of the requisite theorems.

Again let $V$ be the volume of the system multiplied by Avogadro’s numbers. If $X$ is the solution to the stochastic system, which counts the numbers of molecules, then

$$X^V(t) \overset{\text{def}}{=} \frac{X(t)}{V},$$

gives the concentration of the different species in moles per unit volume. Therefore, the stochastic equation governing the concentrations is

$$X^V(t) = X^V(0) + \sum_k \frac{1}{V} Y_k \left( \int_0^t \lambda_k(VX^V(s))ds \right) \zeta_k,$$

where we divided each instance of $X$ by $V$.

Letting $\lambda_k$ be stochastic mass-action kinetics, and using the relation (6.24), we have

$$\lambda_k(X) = V^{-\left(\sum_i \nu_i k - 1\right)} \hat{\kappa}_k \left( \prod_i \nu_i ! \right) \prod_i \left( \frac{X_i}{\nu_i k} \right)$$

$$= \hat{\kappa}_k V \prod_i V^{-\nu_i k} X_i \cdots (X_i - \nu_i k + 1)$$

$$\approx V\hat{\kappa}_k \prod_i (X_i^V)^{\nu_i k},$$
where the approximation is valid for large $V$. For vectors $u, v$, define

$$u^v \overset{\text{def}}{=} \prod_i u_i^v,$$

where we take $0^0 = 1$. Now, the stochastic equation governing $X^V$ is

$$X^V(t) = X^V(0) + \sum_k \frac{1}{V} Y_k \left( \int_0^t \lambda_k (V X^V(s)) ds \right) \zeta_k \approx X^V(0) + \sum_k \frac{1}{V} Y_k \left( V \int_0^t \hat{\kappa}_k X^V(s)^{\nu_k} ds \right) \zeta_k. \quad (6.25)$$

Recalling that

$$\lim_{V \to \infty} \frac{1}{V} Y(V u) = u,$$

for any unit-rate Poisson process $Y$, we see that in the limit as $V \to \infty$, $X^V$ satisfies the integral equation

$$x(t) = x(0) + \sum_k \zeta_k \int_0^t \hat{\kappa}_k x(s)^{\nu_k} ds, \quad (6.26)$$

which in differential form is

$$x'(t) = \sum_k \hat{\lambda}_k (x(t)) \zeta_k,$$

where $\hat{\lambda}_k$ is deterministic mass action kinetics with rate parameter $\hat{\kappa}_k$. That is, for large $V$, we have that $X^V \approx x$, and $X^V$ is well approximated by the solution to the deterministic model.

We were quite loose with the above scaling. The correct technical result, which can be found in [27] states that if the rate constants for the stochastic model satisfy the above scaling assumptions, and if $X^V(0) = O(1)$, then for any $\epsilon > 0$ and any $t > 0$

$$\lim_{V \to \infty} P \left\{ \sup_{s \leq t} |X^V(s) - x(s)| > \epsilon \right\} = 0.$$

See also [15, 28, 30]. For practical purposes, this result say that if $X^V(0) = O(1)$ for some large $V$, and if the systems satisfies mass-action kinetics, then it is plausible to use the deterministic model, with deterministic mass-action kinetics, as an approximation for the stochastic model that governs the concentrations of the molecular abundances. This result has been used for decades to justify the use of ordinary differential equation models in chemistry and, more generally, population processes (though such models were used well before such a rigorous justification was available).

**Example 6.5.1.** Consider the dimerization of a certain protein

$$2P \overset{\kappa_1}{\rightleftharpoons} \overset{\kappa_2}{D}.$$
Note that we have the conservation relation $2X_D + X_P = M$, for some $M > 0$, where $X_D$ and $X_P$ give the numbers of dimers and proteins, respectively. The stochastic model for this system is

$$X_P(t) = X_P(0) - 2Y_1 \left( \int_0^t \kappa_1 X_P(s)(X_P(s) - 1)\,ds \right) + 2Y_2 \left( \int_0^t \kappa_2 \frac{1}{2}(M - X_P(s))\,ds \right),$$

where we made use of the conservation relation. The corresponding deterministic model is

$$x_p(t) = x_p(0) - 2 \int_0^t \hat{\kappa}_1 p(s)^2\,ds + 2 \int_0^t \kappa_2 \frac{1}{2}(\hat{M} - x_p(s))\,ds,$$

where $\hat{\kappa}_2$ and $\hat{M}$ are the normalized rate constant and conservation relation, respectively. Taking expectations of the stochastic version yields

$$\mathbb{E} X_P(t) = \mathbb{E} X_P(0) - 2 \int_0^t \kappa_1 \mathbb{E} X_P(s)(X_P(s) - 1)\,ds + 2 \int_0^t \kappa_2 \frac{1}{2}(M - \mathbb{E} X_P(s))\,ds$$

$$= \mathbb{E} X_P(0) - 2 \int_0^t \kappa_1 (\mathbb{E} X_P(s))^2\,ds + 2 \int_0^t \kappa_2 \frac{1}{2}(M - \mathbb{E} X_P(s))\,ds$$

$$- 2 \int_0^t \kappa_1 (\text{Var}(P(s)) - \mathbb{E} X_P(s))\,ds. \quad (6.28)$$

Note that the portion of the above integral equation given by (6.28) is precisely of the same form as (6.27). Further, the part given by (6.29) is, in general, non-zero. Therefore, we see that the equation for the mean of the stochastic process is not the same as the equation for the associated deterministic model. In general, only linear systems have the property that the mean of the stochastic process satisfies the equations of the associated deterministic model.

### 6.6 Brownian Motions

We wish to construct a process, $W$, satisfying the following four properties.

1. $W(0) = 0$;

2. For any $s_1 \leq t_1 \leq s_2 \leq t_2 \leq \cdots \leq s_n \leq t_n$, the random variables $W(t_1) - W(s_1), \ldots, W(t_n) - W(s_n)$ are independent;

3. For any $s < t$, the random variable $W(t) - W(s)$ is normal with mean zero and variance $\lambda(t - s)$.

4. The function $t \to W(t)$ is a continuous function of $t$.

---

2To see this, take $X_P(0) = 10$ with a probability of one.
Any process satisfying the four conditions above is termed a Brownian motion or Wiener process, with variance parameter \( \lambda > 0 \). A standard Brownian motion is one in which \( \lambda = 1 \).

There are a number of ways to construct such a process, with the most common method using symmetric random walks. We choose a different method. Consider a homogeneous Poisson process, \( Y_\lambda \), with rate \( \lambda > 0 \). Viewed as a counting process, the holding time is exponential with mean \( \mu = 1/\lambda \) and variance \( \sigma^2 = 1/\lambda^2 \). Therefore, from Lemma 4.1.4

\[
\frac{Y_\lambda(Vt) - \lambda Vt}{\lambda^{-1}\lambda^{3/2}\sqrt{Vt}} = \frac{Y_\lambda(Vt) - \lambda Vt}{\lambda^{1/2}\sqrt{Vt}} \approx N(0,1),
\]

implying

\[
V^{-1/2}[Y_\lambda(Vt) - V\lambda t] \approx N(0,\lambda t).
\]

Letting \( Y \) be a unit rate Poisson process, the above is equivalent to

\[
V^{-1/2}[Y(V\lambda t) - \lambda Vt] \approx N(0,\lambda t).
\]

Define

\[
W^{(V)}(t) \overset{\text{def}}{=} V^{-1/2}[Y(V\lambda t) - \lambda Vt].
\]

The following four properties of the process \( W^{(V)} \) all follow from the corresponding properties of the Poisson process.

1. \( W^{(V)}(0) = 0 \);
2. For any \( s_1 \leq t_1 \leq s_2 \leq t_2 \leq \cdots \leq s_n \leq t_n \), the random variables \( W^{(V)}(t_1) - W^{(V)}(s_1), \ldots, W^{(V)}(t_n) - W^{(V)}(s_n) \) are independent;
3. For any \( s < t \), the random variable \( W^{(V)}(t) - W^{(V)}(s) \) is approximately normal with mean zero and variance \( \lambda(t-s) \).
4. \( W^{(V)} \) is constant except for jumps of size \( 1/\sqrt{V} \).

It can be shown that as \( V \to \infty \), the above process converges to a process, \( W \),\(^3\) written \( W^{(V)} \Rightarrow W \), satisfying the conditions of a Brownian motion. Note that in this scaling/limit above, a standard Brownian motion arises from scaling a unit-rate Poisson process.

Note that if we wish the process to start at \( x \in \mathbb{R} \) as opposed to zero, then item 1 becomes \( W(0) = x \) and all other items remain the same. Further, note the important fact that for any \( s < t \),

\[
W(t) - W(s)
\]

has a normal distribution with mean zero and variance \( \lambda(t-s) \), and so, for example,

\[
\mathbb{E}[W(t) - W(s)] = 0, \quad \text{and} \quad \mathbb{E}[(W(t) - W(s))^2] = \lambda(t-s).
\]

We will not rigorously study Brownian motions in these notes.

\(^3\)The type of convergence is beyond the scope of this book. Technically, it is convergence in distribution and we write \( W^{(V)} \Rightarrow W \).
6.6.1 Markov property and generator of a Brownian motion

The Brownian motion $W$ has the Markov property by condition (ii) in its definition. That is, the future behavior of the process only depends upon its current value. We may therefore ask: is there a generator for the process?

Let $W$ be a Brownian motion with variance parameter $\sigma^2$. We attempt to find a generator in the sense of (5.14):

$$(Af)(x) \overset{\text{def}}{=} \lim_{h \to 0} \frac{\mathbb{E}_x f(W(h)) - f(x)}{h}.$$  

Letting $\rho = N(0, \sigma^2 h)$, we note

- $\mathbb{E}\rho = 0$
- $\mathbb{E}\rho^2 = \sigma^2 h$
- $\mathbb{E}\rho^3 = 0$
- $\mathbb{E}\rho^4 = O(h^2)$.

Assuming $W(0) = x$, we have that $W(h) = W(0) + W(h) - W(0) = x + \rho$, and so

$$(Af)(x) = \lim_{h \to 0} \frac{\mathbb{E}_x f(W(h)) - f(x)}{h} = \lim_{h \to 0} \frac{\mathbb{E}_x f(x + \rho) - f(x)}{h}.$$  

Taking a Taylor expansion of $f$ (we are assuming that $f$ is smooth enough for all the derivatives needed to exist) yields

$$(Af)(x) = \lim_{h \to 0} \frac{\mathbb{E}_x f'(x)\rho + (1/2)f''(x)\rho^2 + (1/3!)f'''(x)\rho^3 + \cdots}{h} = \frac{1}{2} \sigma^2 f''(x).$$

Said differently, the operator $A$ is the second derivative times $\sigma^2/2$:

$$A = \frac{1}{2} \sigma^2 \frac{\partial^2}{\partial x^2}. \quad (6.30)$$

We return to our construction of the Brownian motion as a limit of a properly scaled Poisson processes to see if we can understand this generator in a different way. Let $Y$ be a unit rate Poisson process. Then, $Y(\sigma^2 \cdot)$ is a Poisson process with rate $\sigma^2$. We have that for large $V$,

$$\frac{1}{\sqrt{V}} \left[ Y(V\sigma^2 t) - \sigma^2 Vt \right] \approx W(\sigma^2 t),$$

where $W$ is a standard Brownian motion. Consider now the generator of the process

$$Z^V(t) = Z^V(0) + \frac{1}{\sqrt{V}} \left[ Y(V\sigma^2 t) - \sigma^2 Vt \right].$$
Denoting the generator of $Z^V$ by $A_Z$, we have

$$(A_Z f)(x) \overset{\text{def}}{=} \lim_{h \to 0} \frac{1}{h} \left[ \mathbb{E}_x f(Z(h)) - f(x) \right].$$

Further,

$$\mathbb{E}_x f(Z(h)) = f(x + \frac{1}{\sqrt{V}} \left[ 1 - \sigma^2 V h \right]) V \sigma^2 h + f(x - \sigma^2 \sqrt{V} h) (1 - V \sigma^2 h) + o(h)$$

$$= f(x + 1/\sqrt{V}) V \sigma^2 h + \left( f(x) - f'(x) \sqrt{V} \sigma^2 h + o(h) \right) (1 - V \sigma^2 h) + o(h)$$

$$= f(x + 1/\sqrt{V}) V \sigma^2 h + f(x)(1 - V \sigma^2 h) - f'(x) \sqrt{V} \sigma^2 h + o(h).$$

Therefore,

$$(A_Z f)(x) = \lim_{h \to 0} \frac{1}{h} \left[ \mathbb{E}_x f(Z(h)) - f(x) \right]$$

$$= f(x + 1/\sqrt{V}) V \sigma^2 - f(x) V \sigma^2 - f'(x) \sqrt{V} \sigma^2$$

$$= V \sigma^2 \left( f(x + 1/\sqrt{V}) - f(x) \right) - \sqrt{V} \sigma^2 f'(x).$$

Note that this generator can be understood by first considering the jump portion,

$$V \sigma^2 \left( f(x + 1/\sqrt{V}) - f(x) \right),$$

followed by the deterministic portion $-\sqrt{V} \sigma^2 f'(x)$ (recall Section 6.1.4). Taking a Taylor approximation of $A_Z f$ now yields:

$$(A_Z f)(x) = V \sigma^2 \left( f'(x)/\sqrt{V} + (1/2) f''(x)/V + (1/3!) f'''(x)/V^{3/2} + O(V^{-2}) \right)$$

$$- \sqrt{V} \sigma^2 f'(x)$$

$$= \frac{1}{2} \sigma^2 f''(x) + O(V^{-1/2})$$

$$\approx (Af)(x),$$

where $A$ is the generator for the Brownian motion as given by (6.30). Thus, not unexpectedly, the generator of the Brownian motion can be obtained via the generator of the scaled Poisson process simply by truncating the Taylor expansion of $A_Z f$.

### 6.7 Integration with Respect to Brownian Motion

Before discussing how to integrate with respect to Brownian motion, we consider how to integrate with respect to a more standard function. Consider two functions, $g$ and $U$. We will discuss what we mean by “integration of $g$ with respect to $U$.” That is, we will define

$$\int_0^t g(x) dU(x).$$
Letting $t_i = it/n$, this integral basically means
\[ \int_0^t g(x) dU(x) = \sum_{i=0}^{n-1} \int_{t_i}^{t_{i+1}} g(x) dU(x) \approx \sum_{i=0}^{n-1} g(t_i)(U(t_i + \Delta t) - U(t_i)), \]

(6.31)
where $\Delta t = t_{i+1} - t_i$, where the definition of the integral is in the limit $\Delta t \to 0$. The most common type of such integration is when $U$ is absolutely continuous. If you are not sure what this means, just think of $U$ as being differentiable, with derivative (or density), $u$. In this case, (6.31) yields
\[ \int_0^t g(x) dU(x) \approx \sum_{i=0}^{n-1} g(t_i)u(t_i)\Delta t \approx \int_0^t g(t_i)u(t_i)dt. \]

As discussed in the previous section, a Brownian motion is not differentiable, and it can be shown to not be absolutely continuous. However, you can still make sense of
\[ \int_0^t g(s)dW(s) = \lim_{\Delta t \to \infty} \sum_{i=0}^{n-1} g(t_i)(W(t_i + \Delta t) - W(t_i)), \]

(6.32)
where $g$ is some function and $W$ is a (standard) Brownian motion, in a similar manner. This integral is termed an Itô integral. The reason that constructing such an integral is actually difficult is because of (??), which implies $W(t_i + \Delta t) - W(t_i) \approx O(\sqrt{\Delta})$, and so
\[ \int_0^t g(s)dW(s) \approx \sum_{i=0}^{n-1} g(t_i)\sqrt{\Delta t} \frac{(W(t_i + \Delta t) - W(t_i))}{\sqrt{\Delta t}}, \]

appears to blow up as $\Delta t \to 0$ since $n = 1/\Delta t$. However, using the properties of a Brownian motion,
\[ \mathbb{E} \left( \sum_{i=0}^{n-1} g(t_i)(W(t_i + \Delta t) - W(t_i)) \right)^2 = \sum_{i=0}^{n-1} g(t_i)^2 \mathbb{E}(W(t_i + \Delta t) - W(t_i))^2 \]
\[ = \sum_{i=0}^{n-1} g(t_i)^2 \Delta t \approx \int_0^t g(s)^2 ds, \]

(6.33)
where the first equality holds because all the cross terms are zero for $j > i$
\[ \mathbb{E}(W(t_{j+1}) - W(t_j))(W(t_{i+1}) - W(t_i)) = \mathbb{E}(W(t_{j+1}) - W(t_j))\mathbb{E}(W(t_{i+1}) - W(t_i)) = 0. \]
Thus, the right hand side of (6.32) at least does not blow up as $\Delta t \to 0$. To see that such an integral is actually well defined, with all technical details provided, see, for example, [35]. However, the basic construction implied above is correct and, as implied by equation 6.33 above,

$$\mathbb{E}\left(\int_0^t g(s)dW(s)\right)^2 = \int_0^t g(s)^2 ds,$$

which is called the Itô isometry. Even more generally, if $g(s,\omega)$ depends upon $W$ only up through time $s$ (that is, $g(s,\omega)$ is contained in $\mathcal{F}_s$, or is $\mathcal{F}_s$ measurable), then the Itô isometry still holds,

$$\mathbb{E}\left(\int_0^t g(s,\omega)dW(s)\right)^2 = \int_0^t \mathbb{E}g(s,\omega)^2 ds.$$

This follows from only a slight reworking of (6.33)

$$\mathbb{E}\left(\sum_{i=0}^{n-1} g(t_i,\omega)(W(t_i+\Delta t) - W(t_i))\right)^2 = \sum_{i=0}^{n-1} \mathbb{E}g(t_i,\omega)^2 \mathbb{E}(W(t_i+\Delta t) - W(t_i))^2$$

$$= \sum_{i=0}^{n-1} \mathbb{E}g(t_i,\omega)^2 \Delta t$$

$$\approx \int_0^t \mathbb{E}g(s,\omega)^2 ds,$$

where the first equality holds since $g(t_i,\omega)$ and $W(t_i+\Delta) - W(t_i)$ are independent. For example,

$$\mathbb{E}\left(\int_0^t W(s)dW(s)\right)^2 = \int_0^t \mathbb{E}[W(s)^2] ds = \int_0^t s ds = \frac{t^2}{2}.$$

We will not give a comprehensive introduction to stochastic integration. The interested reader is instead pointed towards [32]. However, we will point out a few things, and solve two integrals explicitly. First, we note that

$$\mathbb{E}\int_0^t g(s,\omega)dW(s) \approx \mathbb{E}\sum_{i} \mathbb{E}g(t_i,\omega)\mathbb{E}(W(t_{i+1}) - W(t_i)) = 0,$$

and so all such integrals have a mean of zero. This is in stark departure from standard Riemannian integration where $g(s) > 0$ implies $\int_0^t g(s) ds > 0$.

We now solve two examples explicitly. First, for constant $\sigma > 0$, we have that

$$\int_0^t \sigma dW(s) = \sigma \sum_{i} (W(t_{i+1}) - W(t_i)) = \sigma W(t).$$

This agrees with our intuition that comes from integrating deterministic functions:

$$\int_0^t \sigma df(s) = \sigma (f(t) - f(0)).$$
Second, we consider the integral above: \( \int_0^t W(s) dW(s) \). A first (incorrect) guess would be to argue as follows: since for differentiable \( f \) we have
\[
\int_0^t f(s) df(s) = \int_0^t f(s) f'(s) ds = \frac{1}{2} (f(t)^2 - f(0)^2),
\]
it must be that
\[
\int_0^t W(s) dW(s) = \frac{1}{2} W(t)^2.
\]
However, we can instantly see this is incorrect by simply checking the moments of \( (1/2)W(t)^2 \):
\[
\mathbb{E} \frac{1}{2} W(t)^2 = \frac{1}{2} t \neq 0
\]
\[
\mathbb{E} \left( \frac{1}{2} W(t)^2 \right)^2 = \frac{1}{4} \mathbb{E} W(t)^4 = \frac{1}{4} 3 t^2 \neq \frac{1}{2} t^2.
\]
Thus, it has incorrect first and second moments, and so we must be more careful. Let
\[
Z_{\Delta t}(t) = \sum_i W(t_i)(W(t_{i+1}) - W(t_i)),
\]
where \( \Delta t = t_{i+1} - t_i \). Then,
\[
Z_{\Delta t}(t) = \sum_i \frac{1}{2} (W(t_{i+1}) + W(t_i))(W(t_{i+1}) - W(t_i))
- \sum_i \frac{1}{2} (W(t_{i+1}) - W(t_i))(W(t_{i+1}) - W(t_i))
= \sum_i \frac{1}{2} (W(t_{i+1})^2 - W(t_i)^2) - \sum_i \frac{1}{2} (W(t_{i+1}) - W(t_i))^2
= \frac{1}{2} W(t)^2 - \frac{1}{2} \sum_i (W(t_{i+1}) - W(t_i))^2. \quad (6.34)
\]
Note that we have recovered our first guess of: \( W(t)^2/2 \), though there is now a correction in the form of the sum on the far right hand side of (6.34)? Let
\[
Q_i = (W(t_{i+1}) - W(t_i))^2.
\]
\( Q_i \) has a mean of \( \Delta t \) and variance of \( O(\Delta t^2) \). Therefore, the random variable \( \sum_i Q_i \) has a mean of \( t \) and a variance of \( O(\Delta) \). Thus, for an appropriate constant \( C \),
\[
\frac{\sum_i Q_i - t}{C \sqrt{\Delta}} \approx N(0, 1).
\]
In particular,
\[
\sum_i Q_i - t = O(\sqrt{\Delta t}) \to 0, \text{ as } \Delta t \to 0,
\]
196
implying
\[ \sum_i Q_i \to t, \text{ as } \Delta t \to 0. \]

Hence,
\[ \frac{1}{2} \sum_i Q_i \to \frac{1}{2} t. \]

Collecting the above shows that
\[ \int_0^t W(s) dW(s) \approx Z_{\Delta t}(t) \to \frac{1}{2} W(t)^2 - \frac{1}{2} t, \text{ as } \Delta t \to 0. \]

Hence, we conclude
\[ \int_0^t W(s) dW(s) = \frac{1}{2} W(t)^2 - \frac{1}{2} t. \]

Note that, as expected, we have
\[ \mathbb{E} \left[ \frac{1}{2} W(t)^2 - \frac{1}{2} t \right] = 0, \]
and
\[ \mathbb{E} \left( \int_0^t W(s) dW(s) \right)^2 = \mathbb{E} \left( \frac{1}{2} W(t)^2 - \frac{1}{2} t \right)^2 = \frac{1}{4} \mathbb{E} W(t)^4 - \frac{1}{2} t \mathbb{E} W(t)^2 + \frac{1}{4} t^2 \]
\[ = \frac{3}{4} t^2 - \frac{1}{2} t^2 \]
\[ = \frac{1}{2} t^2. \]

We now construct another process from a Brownian motion that is in many ways equivalent to the one constructed above, and will be of use to us. Let \( W \) be a standard Brownian motion and consider the process
\[ Z(t) \overset{\text{def}}{=} W \left( \int_0^t g(s, \omega)^2 ds \right), \]
where, again, \( g(s, \omega) \) may depend upon \( W \), but only up until time \( \tau = \int_0^t g(s, \omega) ds \). That is, it is contained within \( \mathcal{F}_\tau \). The above is a time-changed Brownian motion. We have
\[ Z(t + h) - Z(t) = W \left( \int_t^{t+h} g(s, \omega)^2 ds + \int_0^t g(s, \omega)^2 ds \right) - W \left( \int_0^t g(s, \omega)^2 ds \right), \]
which by the independent increments of \( W \) is approximately normal with mean zero and variance \( g(t, \omega)^2 h \), which is exactly the same distribution as the infinitesimal increment
\[ \int_0^{t+h} g(s, \omega) dW(s) - \int_0^t g(s, \omega) dW(s) \approx g(t, \omega)(W(t + h) - W(t)). \]
This implies that the two processes

\[ \int_0^t g(s, \omega) dW(s) \quad \text{and} \quad W \left( \int_0^t g(s, \omega)^2 ds \right) \tag{6.35} \]

are distributionally equivalent (thought not equal for a given Brownian path). The representation on the left of (6.35) is termed an Itô integral, whereas the process on the right is the time changed process, and can be traced back to Wolfgang Doeblin.

### 6.8 Diffusion and Linear Noise Approximations

We are in position to give two approximations to the process (6.5) which use Brownian motions.

#### 6.8.1 Diffusion approximation

Define the function \( F \) via

\[ F(x) = \sum_k \hat{\kappa}_k x^{\nu_k} \zeta_k, \tag{6.36} \]

which is deterministic mass-action kinetics. Returning to (6.25), the scaled model satisfies

\[ X^V(t) = X^V(0) + \sum_k \frac{1}{V} Y_k \left( V \int_0^t \hat{\kappa}_k X(s)^{\nu_k} ds \right) \zeta_k, \]

which, after centering the counting processes, yields

\[ X^V(t) = X^V(0) + \sum_k \frac{1}{V} \left( Y_k \left( V \int_0^t \hat{\kappa}_k X^V(s)^{\nu_k} ds \right) - V \int_0^t \hat{\kappa}_k X^V(s)^{\nu_k} ds \right) \zeta_k \]

\[ + \int_0^t F(X^V(s)) ds. \]

Using that

\[ \frac{1}{\sqrt{V}} [Y_k(Vu) - Vu] \approx W_k(u), \tag{6.37} \]

where \( W \) is a standard Brownian motion, we then have that

\[ X^V(t) \approx X^V(0) + \int_0^t F(X^V(s)) ds + \sum_k \frac{1}{\sqrt{V}} W_k \left( \int_0^t \hat{\kappa}_k X^V(s)^{\nu_k} ds \right) \zeta_k, \]

where the \( W_k \) are independent standard Brownian motions. This implies that a good approximation to \( X^V \) would be the process \( Z^V \) satisfying

\[ Z^V(t) = X^V(0) + \int_0^t F(Z^V(s)) ds + \sum_k \frac{1}{\sqrt{V}} W_k \left( \int_0^t \hat{\kappa}_k Z^V(s)^{\nu_k} ds \right) \zeta_k. \]

\[ ^4 \text{The centered version of } Y(u) \text{ is } Y(u) - u. \text{ That is, it arises simply by subtracting off the mean.} \]
Considering (6.35), an equivalent way to represent $Z^V$ is via the Itô representation
\[
Z^V(t) = Z^V(0) + \int_0^t F(Z^V(s))ds + \sum_k \frac{1}{\sqrt{V}} \zeta_k \int_0^t \sqrt{\kappa_k Z^V(s)^{\nu_k}}dW_k(s).
\]
This equation is often represented in differential form
\[
dZ^V(t) = F(Z^V(t))dt + \sum_k \frac{1}{\sqrt{V}} \zeta_k \sqrt{\kappa_k Z^V(t)^{\nu_k}}dW_k(s). \quad (6.38)
\]
Note that the portion of the system that is stochastic, often termed the “noise” in the system, is $O(1/\sqrt{V})$, and hence assumed small. This equation is known as the Langevin approximation in the biology literature, and as the diffusion approximation in probability. There is actually an issue related to this approximation that is still not completely resolved in the chemical setting pertaining to the non-negativity of the system. Therefore, perhaps a more reasonable representation would be
\[
dZ^V(t) = F(Z^V(t))dt + \sum_k \frac{1}{\sqrt{V}} \zeta_k \sqrt{\kappa_k [Z^V(t)^{\nu_k}]}dW_k(s),
\]
where $[x]^+ = \max\{x, 0\}$.

Note that there is no limit taking place in the derivation of the diffusion approximation. In fact, the system satisfying (6.38) converges to the solution of the deterministic process with mass action kinetics in the limit $V \to \infty$.

**Example 6.8.1.** Consider the system

\[
A^{\kappa_1} \rightleftharpoons B^{\kappa_2}.
\]

Letting $X_A^V, X_B^V$ denote the normalized abundances of the species $A$ and $B$, respectively, we have that $X_A^V(t) + X_B^V(t) = M$, for some $M > 0$, and

\[
X_A^V(t) = X_A^V(0) + \frac{1}{V} Y_1 \left( V \int_0^t \kappa_2 (M - X_A^V(s))ds \right) - \frac{1}{V} Y_2 \left( V \int_0^t \kappa_1 X_A^V(s)ds \right).
\]

Therefore, the diffusion approximation is the solution to

\[
Z^V(t) = Z^V(0) + \kappa_2 \int_0^t (M - Z^V(s))ds - \kappa_1 \int_0^t Z^V(s)ds
\]

\[
+ \frac{1}{\sqrt{V}} W_1 \left( \int_0^t \kappa_2 [M - Z^V(s)]^+ ds \right) - \frac{1}{\sqrt{V}} W_2 \left( \int_0^t \kappa_1 [Z^V(s)]^+ ds \right),
\]

or, equivalently, the solution to the stochastic differential equation

\[
dZ^V(t) = \kappa_2 (M - Z^V(t))dt - \kappa_1 \int_0^t Z^V(s)ds
\]

\[
+ \frac{1}{\sqrt{V}} \sqrt{\kappa_2 [M - Z^V(t)]^+} dW_1(t) - \frac{1}{\sqrt{V}} \sqrt{\kappa_1 [Z^V(t)]^+} dW_2(t).
\]

\[\square\]
6.8.2 Linear noise approximation

Let \( x(t) \) be the solution to the limiting deterministic system (6.26), and recall that \( F \) is the deterministic kinetics defined in (6.36). Since by (6.37), \( V^{-1/2} [Y_k(Vu) - Vu] \) is approximately a Brownian motion,

\[
L^V(t) \overset{\text{def}}{=} \sqrt{V} (X^V(t) - x(t)) = L^V(0) + \sqrt{V} \left( \sum_k \frac{1}{V} Y_k \left( V \int_0^t \lambda_k(X^V(s)) ds \right) \zeta_k - \int_0^t F(x(s)) ds \right)
\]

\[
= L^V(0) + \sum_k \frac{1}{\sqrt{V}} \left[ Y_k \left( V \int_0^t \lambda_k(X^V(s)) ds \right) - V \int_0^t \lambda_k(X^V(s)) ds \right] \zeta_k
+ \int_0^t \sqrt{V} (F(X^V(s)) - F(x(s))) ds
\]

\[
\approx L^V(0) + \sum_k W_k \left( \int_0^t \lambda_k(x(s)) ds \right) \zeta_k + \int_0^t \nabla F(x(s)) \cdot L^V(s) ds.
\]

The limit as \( V \) goes to infinity gives \( L^V \Rightarrow L \) where

\[
L(t) = L(0) + \sum_k W_k \left( \int_0^t \lambda_k(x(s)) ds \right) \zeta_k + \int_0^t \nabla F(x(s)) \cdot L(s) ds.
\] (6.39)

For more details, see [26, 29, 41] and Chapter 11 of [15]. Note that an alternative representation of (6.39) is

\[
L(t) = L(0) + \sum_k \zeta_k \int_0^t \sqrt{\lambda_k(x(s))} dW_k(s) + \int_0^t \nabla F(x(s)) \cdot L(s) ds,
\]

where now positivity of the term in the square root is guaranteed as \( x \) is the solution to the deterministic model and stays positive for all time (so long as each \( x_i(t) > 0 \)). The above limit suggests the approximation

\[
X^V(t) \approx \tilde{X}_V(t) \overset{\text{def}}{=} x(t) + \frac{1}{\sqrt{V}} L(t),
\]

which is often called the linear noise approximation to \( X^V \), and is used quite extensively. Note that, once again, the “noise” scales like \( 1/\sqrt{V} \).

**Example 6.8.2.** Consider the system

\[
A \xrightleftharpoons{\kappa_1}{\kappa_2} B.
\]

The ordinary differential equation governing \( x_A(t) \), the concentration of \( A \), is

\[
\dot{x}_A(t) = F(x(t)) \overset{\text{def}}{=} \kappa_2(M - x_A(t)) - \kappa_1 x_A(t),
\] (6.40)
where \( M = x_A(t) + x_B(t) \). Therefore,

\[
F'(x) = -\kappa_2 - \kappa_1.
\]

Assuming \( X_A'(0) = x_A(0) \), the equation for \( L \) is then

\[
L(t) = \int_0^t \sqrt{\kappa_2(M - x_A(s))} \, dW_1(s) - \int_0^t \sqrt{\kappa_1 x_A(s)} \, dW_2(s) - (\kappa_1 + \kappa_2) \int_0^t L(s) \, ds,
\]

or

\[
dL(t) = \sqrt{\kappa_2(M - x_A(t))} \, dW_1(t) - \sqrt{\kappa_1 x_A(t)} \, dW_2(t) - (\kappa_1 + \kappa_2)L(t) \, dt.
\]

Solving this equation yields,

\[
L(t) = \int_0^t e^{-(\kappa_1+\kappa_2)(t-s)} \sqrt{\kappa_2(M - x_A(s))} \, dW_1(s) - \int_0^t e^{-(\kappa_1+\kappa_2)(t-s)} \sqrt{\kappa_1 x_A(s)} \, dW_2(s).
\]

Finally, we set

\[
\hat{X}_V(t) = x_A(t) + \frac{1}{\sqrt{V}} L(t).
\]

\[\square\]

### 6.9 Solving stochastic differential equations numerically

Consider the stochastic equation

\[
X(t) = X(0) + \int_0^t b(X(s)) \, ds + \int_0^t \sigma(X(s)) \, dW(s),
\]

where \( X \in \mathbb{R} \), \( b: \mathbb{R} \to \mathbb{R} \), and \( \sigma: \mathbb{R} \to \mathbb{R}_{\geq 0} \). The differential form of the above equation is

\[
dX(t) = b(X(t)) \, dt + \sigma(X(t)) \, dW(t),
\]

where we do not use the notation \( dX(t)/dt \) since \( W \) is not differentiable, and hence \( dW(t)/dt \) is problematic; therefore, it is better to think in terms of the differentials \( dX \) or \( dW \). Note, however, that this is all just notation and no matter what we write, we mean that \( X \) satisfies the integral equation above.

The most common numerical method, by far, to solve for the solution to (6.41) is Euler’s method. That is, we use

\[
X(t + h) = X(t) + \int_t^{t+h} b(X(s)) \, ds + \int_t^{t+h} \sigma(X(s)) \, dW(s)
\]

\[
\approx X(t) + b(X(t))h + \sigma(X(t))(W(t+h) - W(t))
\]

\[
dist = X(t) + b(X(t))h + \sigma(X(t))\sqrt{h}\rho,
\]

where \( \rho \sim N(0,1) \). This observation leads to the following algorithm. In the algorithm below, all random variables generated are assumed to be independent of all previous random variables. The constructed process will be denoted by \( Z \) so as to differentiate it and the exact process \( X \).
Algorithm 8. Fix $Z(0)$ and $h > 0$. Set $n = 0$. Repeat the following steps.

1. Let $\rho \sim N(0, 1)$.

2. Set
   
   $$Z((n+1)h) = Z(nh) + b(Z(nh))h + \sigma(Z(nh))\sqrt{h}\rho$$

   Note that one way to represent $Z$ (at least at the points $nh$), is as the solution to
   
   $$Z(t) = Z(0) + \int_0^t b(Z \circ \eta(s))ds + \int_0^t \sigma(Z \circ \eta(s))dW(s),$$
   
   where $\eta(s) = nh$ for $nh \leq s < (n+1)h$. This follows since if $Z$ is so generated then
   
   $$Z((n+1)h) = Z(nh) + \int_{nh}^{(n+1)h} b(Z(nh))ds + \int_{nh}^{(n+1)h} \sigma(Z(nh))dW(s)$$
   
   $$= Z(nh) + b(Z(nh))h + \sigma(Z(nh))(W((n+1)h) - W(nh)),$$
   
   and $W((n+1)h) - W(nh) \sim N(0, h)$.

   Note also that (6.41) is distributionally equivalent to the time changed representation
   
   $$X(t) = X(0) + \int_0^t b(X(s))ds + W\left(\int_0^t \sigma^2(X(s))ds\right). \quad (6.42)$$
   
   This can be seen by noting that for (6.42), Euler’s method reduces to
   
   $$Z(t) = Z(0) + \int_0^t b(Z \circ \eta(s))ds + W\left(\int_0^t \sigma^2(Z \circ \eta(s))ds\right),$$
   
   yielding
   
   $$Z((n+1)h) = Z(nh) + \int_{nh}^{(n+1)h} b(Z(nh))ds$$
   
   $$+ W\left(\int_{nh}^{(n+1)h} \sigma^2(Z \circ \eta(s))ds + \int_0^{nh} \sigma^2(Z \circ \eta(s))ds\right)$$
   
   $$- W\left(\int_0^{nh} \sigma^2(Z \circ \eta(s))ds\right)$$
   
   $$= Z(nh) + b(Z(nh))hds + W\left(\sigma^2(Z(nh))h + \int_0^{nh} \sigma^2(Z \circ \eta(s))ds\right)$$
   
   $$- W\left(\int_0^{nh} \sigma^2(Z \circ \eta(s))ds\right),$$
   
   where
   
   $$W\left(\sigma^2(Z(nh))h + \int_0^{nh} \sigma^2(Z \circ \eta(s))ds\right) - W\left(\int_0^{nh} \sigma^2(Z \circ \eta(s))ds\right),$$
   
   is normally distributed with mean zero and variance $\sigma^2(Z(nh))h$.

202
6.10 Exercises

1. Simulate the model of foxes and rabbits in Example 6.1.6 in order to output a plot similar to that in Figure 6.1.1. Do this two ways, using Gillespie’s algorithm and the next reaction method.

2. Repeat number 1 above for the SIR model of Example 6.1.7. Choose rate constants of
Appendix A

A quick review of linear differential equations and linear difference equations

In Section A.1 we review linear differential equations. In Section A.2 we review linear difference equations. It is assumed throughout that the student has a basic knowledge of linear algebra. If this is not the case, please review linear algebra from any number of introductory texts, and/or talk with the instructor. If after reading these sections you feel you need more help on this subject, please consult any reference book on differential equations such as *Elementary differential equations with boundary value problems*, by Boyce & DiPrima (2012) [8].

A.1 Linear differential equations

Consider the first order linear differential equation

\[
\begin{align*}
y_1'(t) &= a_{11}y_1(t) + a_{12}y_2(t) + \cdots + a_{1n}y_n(t) \\
y_2'(t) &= a_{21}y_1(t) + a_{22}y_2(t) + \cdots + a_{2n}y_n(t) \\
&\vdots \nonumber \\
y_n'(t) &= a_{n1}y_1(t) + a_{n2}y_2(t) + \cdots + a_{nn}y_n(t),
\end{align*}
\]  

(A.1)

where the \(a_{ij}\) are constants and \(y_1, \ldots, y_n\) are the unknown functions. The system (A.1) can be written succinctly via the vector equation

\[
y'(t) = Ay(t),
\]

where

\[
y(t) = \begin{bmatrix} y_1(t) \\ y_2(t) \\ \vdots \\ y_n(t) \end{bmatrix}, \quad \text{and} \quad A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}.
\]

0Copyright © 2013 by David F. Anderson.
If \( y(0) = v \in \mathbb{R}^n \), the above system has solution
\[
y(t) = e^{At}v,
\]
where
\[
e^{At} = \sum_{k=0}^{\infty} \frac{(At)^k}{k!}.
\]

Equation (A.2) is almost never used itself to compute \( e^{At} \). Instead, we usually try to diagonalize the matrix \( A \). Suppose that \( A \) has \( n \) distinct eigenvalues, \( \lambda_1, \ldots, \lambda_n \), with corresponding eigenvectors \( v_1, \ldots, v_n \). Let \( Q \) be the matrix whose \( k \)th column is \( v_k \). It is a standard fact from linear algebra that \( Q \) is invertible under these assumptions (since the \( v_k \) are necessarily linearly independent).

We are interested in the form of the matrix \( Q^{-1}AQ \), which we can discover column by column by letting \( e_i \in \mathbb{R}^n \) be the canonical basis with a one in the \( i \)th row and zeros elsewhere, and noting,
\[
Q^{-1}AQe_i = Q^{-1}Av_i = \lambda_i Q^{-1}v_i = \lambda_i e_i.
\]

Hence, the \( i \)th column of \( Q^{-1}AQ \) is simply the diagonal matrix with \( \lambda_i \), the \( i \)th eigenvalue, as the \( i \)th diagonal element. That is, \( D := Q^{-1}AQ \) is the diagonal matrix
\[
D = \begin{bmatrix}
\lambda_1 & 0 & \cdots & 0 \\
0 & \lambda_2 & \cdots & 0 \\
\vdots & \ddots & \ddots & \vdots \\
0 & 0 & \cdots & \lambda_n
\end{bmatrix},
\]
and we have
\[
e^{At} = Qe^{Dt}Q^{-1} = Q \begin{bmatrix}
e^{\lambda_1 t} & 0 & \cdots & 0 \\
0 & e^{\lambda_2 t} & \cdots & 0 \\
\vdots & \ddots & \ddots & \vdots \\
0 & 0 & \cdots & e^{\lambda_n t}
\end{bmatrix} Q^{-1}.
\]

If \( A \) is not diagonalizable, then \( A = QJQ^{-1} \), where \( J \) is in Jordan canonical form. One may then find the matrix exponential \( e^{Jt} \), and solve the system similarly. See, for example, [36] for more details.

\section*{A.2 Linear Difference Equations}

Difference equations come up naturally in the study of discrete time Markov chains, and are briefly reviewed. This subsection closely follows that of Lawler [32].

Consider the following second-order linear difference equation
\[
f(n) = af(n - 1) + bf(n + 1), \quad K < n < N,
\]

(A.4)
where $f(n)$ is a function defined on the integers $K \leq n \leq N$, the value $N$ can be chosen to be infinity, and $a$ and $b$ are nonzero real numbers. Note that if $f$ satisfies (A.4) and if the values $f(K)$, $f(K + 1)$ are known then $f(n)$ can be determined for all $K \leq n \leq N$ recursively via the formula

$$f(n + 1) = \frac{1}{b} [f(n) - af(n - 1)].$$

Note also that if $f_1$ and $f_2$ are two solutions of (A.4), then $c_1 f_1 + c_2 f_2$ is a solution for any real numbers $c_1, c_2$. Therefore, the solution space of (A.4) is a two-dimensional vector space and one basis for the space is $\{f_1, f_2\}$ with $f_1(K) = 1, f_1(K + 1) = 0$ and $f_2(K) = 0, f_2(K + 1) = 1$.

We will solve (A.4) by looking for solutions of the form $f(n) = \alpha^n$, for some $\alpha \neq 0$. Plugging $\alpha^n$ into equation (A.4) yields

$$\alpha^n = a\alpha^{n-1} + b\alpha^{n+1}, \quad K < n < N,$$

or

$$\alpha = a + b\alpha^2 \iff b\alpha^2 - \alpha + a = 0.$$

Solving this quadratic gives

$$\alpha = \frac{1 \pm \sqrt{1 - 4ba}}{2b}. \quad (A.5)$$

There are two cases that need handling based upon whether or not the discriminant, $1 - 4ba$, is zero.

**Case 1:** If $1 - 4ba \neq 0$, we find two solutions, $\alpha_1$ and $\alpha_2$, and see that the general solution to the difference equation (A.4) is

$$c_1\alpha_1^n + c_2\alpha_2^n,$$

with $c_1, c_2$ found depending upon the boundary conditions. If $1 - 4ba < 0$, then the roots are complex and the general solution is found by switching to polar coordinates. That is, we let $\alpha = re^{i\theta}$, and find

$$f(n) = r^n e^{in\theta} = r^n \cos(n\theta) \pm ir^n \sin(n\theta),$$

are solutions, implying both the real and imaginary parts are solutions. Therefore, the general solution is

$$c_1r^n \cos(n\theta) + c_2r^n \sin(n\theta),$$

with $c_1, c_2$ found depending upon the boundary conditions.

**Case 2:** If $1 - 4ba = 0$, we only find the one solution, $f_1(n) = (1/2b)^n$ by solving the
quadratic. However, let \( f_2(n) = n(2b)^{-n} \). We have that

\[
af_2(n - 1) + bf_2(n + 1) = a(n - 1)(2b)^{-(n-1)} + b(n + 1)(2b)^{-(n+1)}
\]

\[
= \left( \frac{1}{2b} \right)^n \left[ a(n - 1)2b + b(n + 1) \right] 2b^{-1}
\]

\[
= \left( \frac{1}{2b} \right)^n \left[ a(n - 1) \frac{1}{2} + b(n + 1) \frac{1}{2} \right]
\]

(remember, \( 4ab = 1 \))

\[
= \left( \frac{1}{2b} \right)^n n
\]

\[
= f_2(n).
\]

Note that \( f_2 \) is trivially linearly independent from \( f_1 \). Thus, when \( 4ab = 1 \), the general form of the solution is

\[
f(n) = c_1 \left( \frac{1}{2b} \right)^n + c_2 n \left( \frac{1}{2b} \right)^n.
\]

with \( c_1, c_2 \) found depending upon the boundary conditions.

**Example A.2.1.** Find a function \( f(n) \) satisfying

\[
f(n) = 2f(n - 1) + \frac{1}{10} f(n + 1), \quad 0 < n < \infty,
\]

with \( f(0) = 8, f(1) = 2 \).

**Solution.** Here, \( a = 2 \) and \( b = 1/10 \). Therefore, plugging into (A.5) gives

\[
\alpha = 5 \pm \sqrt{5},
\]

and the general solution is

\[
f(n) = c_1 \left( 5 + \sqrt{5} \right)^n + c_2 \left( 5 - \sqrt{5} \right)^n.
\]

Using the boundary conditions yields

\[
8 = f(0) = c_1 + c_2
\]

\[
2 = f(1) = c_1(5 + \sqrt{5}) + c_2(5 - \sqrt{5}),
\]

which has solution \( c_1 = 4 - 19\sqrt{5}/5, c_2 = 4 + 19\sqrt{5}/5 \). Thus, the solution to the problem is

\[
f(n) = \left( 4 - \frac{19\sqrt{5}}{5} \right) \left( 5 + \sqrt{5} \right)^n + \left( 4 + \frac{19\sqrt{5}}{5} \right) \left( 5 - \sqrt{5} \right)^n.
\]

\( \square \)
Some of the most important difference equations we will see in this course are those of the form

\[ f(n) = pf(n - 1) + qf(n + 1), \quad \text{with } p + q = 1, \quad p, q \geq 0. \]

These will arise when studying random walks with \( p \) and \( q \) interpreted as the associated probabilities of moving right or left. Supposing that \( p \neq q \), the roots of the quadratic formula (A.5) can be found:

\[ \frac{1 \pm \sqrt{1 - 4(1 - p)p}}{2q} = \frac{1 \pm \sqrt{(1 - 2p)^2}}{2q} = \frac{1 \pm |q - p|}{2q} = \left\{ 1, \frac{p}{q} \right\}. \]

Thus, the general solution when \( p \neq 1/2 \) is

\[ f(n) = c_1 + c_2 \left( \frac{p}{q} \right)^n. \]

For the case that \( p = q = 1/2 \), the only root is 1, hence the general solution is

\[ f(n) = c_1 + c_2 n. \]

We analyzed only second-order linear difference equations above. However, and similar to the study of differential equations, higher order difference equations can be studied in the same manner. Consider the general \( k \)th order, homogeneous linear difference equation:

\[ f(n + k) = a_0 f(n) + a_1 f(n + 1) + \cdots + a_{k-1} f(n + k - 1), \quad \text{(A.6)} \]

where we are given \( f(0), f(1), \ldots, f(k - 1) \). Again, we may solve for the general \( f(n) \) recursively using (A.6). We look for solutions of the form \( f(n) = \alpha^n \), which is a solution if and only if

\[ \alpha^k = a_0 + a_1 \alpha + \cdots + a_{k-1} \alpha^{k-1}. \]

If there are \( k \) distinct roots of the above equation, then we automatically get \( k \) linearly independent solutions to (A.6). However, if a given root \( \alpha \) is a root with a multiplicity of \( j \), then

\[ \alpha^n, n\alpha^n, \ldots, n^{j-1} \alpha^n, \]

are linearly independent solutions. We can then use the given initial conditions to find the desired particular solution.

### A.3 Exercises

Problems 1 through 4 require basic linear algebra, which is not explicitly covered in these notes. They are provided to ensure that you have the minimum required knowledge of linear algebra. If the concepts being presented are foreign to you, I strongly encouraged you to consult a text on basic linear algebra, for example, [14], and get in touch with your instructor.

Some of the problems require you to use MATLAB. You must always turn in your MATLAB code when a problem is solved using MATLAB.
1. Let
\[
A = \begin{bmatrix}
-2 & 1 \\
2 & -1
\end{bmatrix}, \quad B = \begin{bmatrix}
0 & -1 \\
1 & 0
\end{bmatrix}, \quad v = \begin{bmatrix}
\frac{1}{3} \\
\frac{2}{3}
\end{bmatrix}
\]
Calculate (by hand) the following:

(a) \(A^2\).
(b) \(AB\).
(c) \(BA\).
(d) \(Av\).

2. Redo Problem 1 using MATLAB. Sample code showing how to multiply two matrices is provided on the course website.

3. Find the eigenvalues and eigenvectors of the matrix \(A\) in Problem 1. Do this in two ways: by hand and by using MATLAB.

4. Find \(e^{At}\) using the relation (A.3).

5. Find all solutions to the following system of differential equations:
\[
x_1'(t) = -2x_1(t) + x_2(t) \\
x_2'(t) = 2x_1(t) - x_2(t).
\]
Next, find the particular solution with initial condition \(x_1(0) = 1/4, x_2(0) = 3/4\).

6. Let \(x(t) = [x_1(t), x_2(t)]^T\), where \(x_1\) and \(x_2\) are the functions found in Problem 5, and \(w^T\) is the transpose of \(w\). What is \(\lim_{t \to \infty} x(t)\). Does this value depend upon the choice of initial condition \(x(0)\)? How does the solution relate to the eigenvalues and eigenvectors of \(A\), be clear in your explanation.

7. Suppose that \(A\) is an \(n \times n\) matrix with \(n\) distinct eigenvalues, \(\lambda_1, \ldots, \lambda_n\). Let \(x(t) = e^{At}x(0)\). What conditions on the eigenvalues guarantee the following (justify your answers):
   (a) \(\lim_{t \to \infty} |x(t)| = 0\) for all choices of \(x(0)\).
   (b) \(\lim_{t \to \infty} |x(t)| = \infty\) for at least some choices of \(x(0)\).

8. Solve the following difference equation by finding a formula for \(f(n)\), valid on \(n = 0, \ldots, 100\).
\[
f(n) = \frac{2}{5}f(n-1) + \frac{3}{5}f(n+1), \quad n = 1, \ldots, 99,
\]
and \(f(0) = 1, f(1) = 0\).
Appendix B

A review of probability theory

These notes are not intended to be a comprehensive introduction to the theory of probability. Instead, they constitute a brief introduction that should be sufficient to allow a student to understand the stochastic models they will encounter in later chapters. These notes were heavily influenced by Sheldon Ross’s text [38], and Timo Seppäläinen’s notes on probability theory that serve a similar purpose [39]. Any student who finds this material difficult should review an introductory probability book such as Sheldon Ross’s A first course in probability [38], which is on reserve in the Mathematics library.

B.1 The Probability Space

Probability theory is used to model experiments (defined loosely) whose outcome can not be predicted with certainty beforehand. For any such experiment, there is a triple \((\Omega, \mathcal{F}, P)\), called a probability space, where

- \(\Omega\) is the sample space,
- \(\mathcal{F}\) is a collection of events,
- \(P\) is a probability measure.

We will consider each in turn.

B.1.1 The sample space \(\Omega\)

The sample space of an experiment is the set of all possible outcomes. Elements of \(\Omega\) are called sample points and are often denoted by \(\omega\). Subsets of \(\Omega\) are referred to as events.

Example B.1.1. Consider the experiment of rolling a six-sided die. Then the natural sample space is \(\Omega = \{1, 2, 3, 4, 5, 6\}\).  

\(^0\)Copyright © 2011 by David F. Anderson.
Example B.1.2. Consider the experiment of tossing a coin three times. Let us write 1 for heads and 0 for tails. Then the sample space consists of all sequences of length three consisting only of zeros and ones. Each of the following representations is valid

\[ \Omega = \{0, 1\}^3 \]
\[ = \{0, 1\} \times \{0, 1\} \times \{0, 1\} \]
\[ = \{(x_1, x_2, x_3) : x_i \in \{0, 1\} \text{ for } i = 1, 2, 3\} \]
\[ = \{(0, 0, 0), (0, 0, 1), (0, 1, 0), (1, 0, 0), (0, 1, 1), (1, 0, 1), (1, 1, 0), (1, 1, 1)\}. \]

Example B.1.3. Consider the experiment of counting the number of mRNA molecules transcribed by a given gene in some interval of time. Here it is most natural to let \( \Omega = \{0, 1, 2, \ldots\} \).

Example B.1.4. Consider the experiment of waiting for a bacteria to divide. In this case, it is natural to take as our sample space all values greater than or equal to zero. That is, \( \Omega = \{t : t \geq 0\} \), where the units of \( t \) are specified as hours, for example.

Note that the above sample spaces are quite different in nature. Those of Examples B.1.1 and B.1.2 are finite, while those of B.1.3 and B.1.4 are infinite. The sample space of Example B.1.3 is countably infinite while that of Example B.1.4 is uncountably infinite. A set that is finite or countably infinite is called discrete. Most, though not all, of the sample spaces encountered in this course will be discrete, in which case probability theory requires no mathematics beyond calculus and linear algebra.

### B.1.2 The collection of events \( \mathcal{F} \)

Events are simply subsets of the state space \( \Omega \). They are often denoted by \( A, B, C \); etc., and they are usually the objects we wish to know the probability of. They can be described in words, or using mathematical notation. Examples of events of the experiments described above are the following:

Example B.1.1, continued. Let \( A \) be the event that a 2 or a 4 is rolled. That is, \( A = \{2, 4\} \).

Example B.1.2, continued. Let \( A \) be the event that the final two tosses of the coin are tails. Thus,

\[ A = \{(1, 0, 0), (0, 0, 0)\}. \]

Example B.1.3, continued. Let \( A \) be the event that no more than 10 mRNA molecules have appeared. Thus,

\[ A = \{0, 1, 2, \ldots, 10\}. \]
Example B.1.4, continued. Let $A$ be the event that it took longer than 2 hours for the bacteria to divide. Then,

$$A = \{ t : t > 2 \}.$$  

\[ \square \]

We will often have need to consider the unions and intersections of events. We write $A \cup B$ for the union of $A$ and $B$, and either $A \cap B$ or $AB$ for the intersection.

For discrete sample spaces, $\mathcal{F}$ will contain all subsets of $\Omega$, and will play very little role. This is the case for nearly all of the models in this course. When the state space is more complicated, $\mathcal{F}$ is assumed to be a $\sigma$-algebra. That is, it satisfies the following three axioms:

1. $\Omega \in \mathcal{F}$.
2. If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$, where $A^c$ is the complement of $A$.
3. If $A_1, A_2, \ldots \in \mathcal{F}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

B.1.3 The probability measure $P$

Definition B.1.5. The real valued function $P$, with domain $\mathcal{F}$, is a probability measure if it satisfies the following three axioms

1. $P(\Omega) = 1$.
2. If $A \in \mathcal{F}$ (or equivalently for discrete spaces if $A \subset \Omega$), then $P(A) \geq 0$.
3. If for a sequence of events $A_1, A_2, \ldots$, we have that $A_i \cap A_j = \emptyset$ for all $i \neq j$ (i.e. the sets are mutually exclusive) then

$$P\left( \bigcup_{i=1}^{\infty} A_i \right) = \sum_{i=1}^{\infty} P(A_i).$$

The following is a listing of some of the basic properties of any probability measure, which are stated without proof.

Lemma B.1.6. Let $P(\cdot)$ be a probability measure. Then

1. If $A_1, \ldots, A_n$ is a finite sequence of mutually exclusive events, then

$$P\left( \bigcup_{i=1}^{n} A_i \right) = \sum_{i=1}^{n} P(A_i).$$

2. $P(A^c) = 1 - P(A)$.
3. $P(\emptyset) = 0.$
4. If \( A \subset B \), then \( P(A) \leq P(B) \).

5. \( P(A \cup B) = P(A) + P(B) - P(A \cap B) \).

Note that for discrete spaces, we can (at least theoretically) find \( P(A) \) for any \( A \in \mathcal{F} \) so long as we know \( P(\omega) \) for every \( \omega \in \Omega \).

**Example B.1.7.** Suppose we roll an unfair die that yields a 1, 2, or 3 each with a probability of 1/10, that yields a 4 with a probability of 1/5, and yields a 5 or 6 each with a probability of 1/4. Then, the probability we roll an even number is

\[
P\{2, 4, 6\} = P\{2\} + P\{4\} + P\{6\} = \frac{1}{10} + \frac{1}{5} + \frac{1}{4} = \frac{11}{20}.
\]

\( \square \)

### B.2 Conditional Probability and Independence

Suppose we are interested in the probability that some event \( A \) took place, though we have some extra information in that we know some other event \( B \) took place. For example, suppose that we want to know the probability that a fair die rolled a 4 given that we know an even number came up. Most people would answer this as 1/3, as there are three possibilities for an even number, \( \{2, 4, 6\} \), and as the die was fair, each of the options should be equally probable. The following definition generalizes this intuition.

**Definition B.2.1.** For two events \( A, B \subset \Omega \), the conditional probability of \( A \) given \( B \) is

\[
P(A|B) = \frac{P(AB)}{P(B)}.
\]

provided that \( P(B) > 0 \).

**Example B.2.2.** The probability that it takes a bacteria over 2 hours to divide is 0.64, and the probability it takes over three hours is 0.51. What is the probability that it will take over three hours to divide, given that two hours have already passed?

**Solution:** Let \( A \) be the event that the bacteria takes over three hours to split and let \( B \) be the event that it takes over two hours to split. Then, because \( A \subset B \),

\[
P(A|B) = \frac{P(AB)}{P(B)} = \frac{P(A)}{P(B)} = \frac{.51}{.64} \approx 0.797.
\]

\( \square \)

We intuitively think of \( A \) being independent from \( B \) if \( P(A|B) = P(A) \), and \( P(B|A) = P(B) \). More generally, we have the following definition.
Definition B.2.3. The events $A, B \in \mathcal{F}$ are called independent if

$$P(AB) = P(A)P(B).$$

It is easy to check that the definition of independence implies both $P(A|B) = P(A)$ and $P(B|A) = P(B)$, and vice versa when $P(A) > 0$ and $P(B) > 0$. The concept of independence will play a key role in our study of Markov chains.

Theorem B.2.4. Let $\Omega$ be a sample space with $B \in \mathcal{F}$, and $P(B) > 0$. Then

(a) $P(A | B) \geq 0$ for any event $A \in \mathcal{F}$.
(b) $P(\Omega | B) = 1$.
(c) If $A_1, A_2, \cdots \in \mathcal{F}$ is a sequence of mutually exclusive events, then

$$P \left( \bigcup_{i=1}^{\infty} A_i \mid B \right) = \sum_{i=1}^{\infty} P(A_i | B).$$

Therefore, conditional probability measures are themselves probability measures, and we may write $Q(\cdot) = P(\cdot | B)$.

By definition

$$P(A|B) = \frac{P(AB)}{P(B)}, \quad \text{and} \quad P(B|A) = \frac{P(AB)}{P(A)}.$$  

Rearranging terms yields

$$P(AB) = P(A|B)P(B), \quad \text{and} \quad P(AB) = P(B|A)P(A).$$

This can be generalized further to the following.

Theorem B.2.5. If $P(A_1A_2\cdots A_n) > 0$, then

$$P(A_1A_2\cdots A_n) = P(A_1)P(A_2 | A_1)P(A_3 | A_1A_2)P(A_4 | A_1A_2A_3)\cdots P(A_n | A_1A_2\cdots A_{n-1}).$$

Definition B.2.6. Let $\{B_1, \ldots, B_n\}$ be a set of nonempty subsets of $\mathcal{F}$. If the sets $B_i$ are mutually exclusive and $\bigcup B_i = \Omega$, then the set $\{B_1, \ldots, B_n\}$ is called a partition of $\Omega$.

Theorem B.2.7 (Law of total probability). Let $\{B_1, \ldots, B_n\}$ be a partition of $\Omega$ with $P(B_i) > 0$. Then for any $A \in \mathcal{F}$

$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i).$$

214
Theorem B.2.8 (Bayes’ Theorem). For all events $A, B \in \mathcal{F}$ such that $P(B) > 0$ we have

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$  

This allows us to “turn conditional probabilities around.”

Proof. We have

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}.$$

\[\square\]

B.3 Random Variables

Henceforth, we assume the existence of some probability space $(\Omega, \mathcal{F}, P)$. All random variables are assumed to be defined on this space in the following manner.

Definition B.3.1. A random variable $X$ is a real-valued function defined on the sample space $\Omega$. That is, $X : \Omega \to \mathbb{R}$.

If the range of $X$ is finite or countably infinite, then $X$ is said to be a discrete random variable, whereas if the range is an interval of the real line (or some other uncountably infinite set), then $X$ is said to be a continuous random variable.

Example B.3.2. Suppose we roll two die and take $\Omega = \{(i, j) \mid i, j \in \{1, \ldots, 6\}\}$. We let $X(i, j) = i + j$ be the discrete random variable giving the sum of the rolls. The range is $\{2, \ldots, 12\}$. \[\square\]

Example B.3.3. Consider two bacteria, labeled 1 and 2. Let $T_1$ and $T_2$ denote the times they will divide to give birth to daughter cells, respectively. Then, $\Omega = \{(T_1, T_2) \mid T_1, T_2 \geq 0\}$. Let $X$ be the continuous random variable giving the time of the first division: $X(T_1, T_2) = \min\{T_1, T_2\}$. The range of $X$ is $t \in \mathbb{R}_{\geq 0}$. \[\square\]

Notation: As is traditional, we will write $\{X \in I\}$ as opposed to the cumbersome $\{\omega \in \Omega \mid X(\omega) \in I\}$.

Definition B.3.4. If $X$ is a random variable, then the function $F_X$, or simply $F$, defined on $(-\infty, \infty)$ by

$$F_X(t) = P\{X \leq t\}$$

is called the distribution function, or cumulative distribution function, of $X$.

Theorem B.3.5 (Properties of the distribution function). Let $X$ be a random variable defined on some probability space $(\Omega, \mathcal{F}, P)$, with distribution function $F$. Then,
1. $F$ is nondecreasing. Thus, if $s \leq t$, then $F(s) = P\{X \leq s\} \leq P\{X \leq t\} = F(t)$.

2. $\lim_{t \to \infty} F(t) = 1$.

3. $\lim_{t \to -\infty} F(t) = 0$.

4. $F$ is right continuous. So, $\lim_{h \to 0^+} f(t + h) = f(t)$ for all $t \in \mathbb{R}$.

For discrete random variables, it is natural to consider a function giving the probability of each possible event, whereas for continuous random variables we need the concept of a density function.

**Definition B.3.6.** Let $X$ be a discrete random variable. Then for $x \in \mathbb{R}$, the function

$$p_X(x) = P\{X = x\}$$

is called the *probability mass function* of $X$.

By the axioms of probability, a probability mass function $p_X$ satisfies

$$P\{X \in A\} = \sum_{x \in A} p_X(x).$$

**Definition B.3.7.** Let $X$ be a continuous random variable with distribution function $F(t) = P\{X \leq t\}$. Suppose that there exists a nonnegative, integrable function $f : \mathbb{R} \to [0, \infty)$, or sometimes $f_X$, such that

$$F(x) = \int_{-\infty}^x f(y)dy.$$

Then the function $f$ is called the *probability density function* of $X$.

We now have that for any $A \subset \mathbb{R}$ (or, more precisely, for any $A \in \mathcal{F}$, but we are going to ignore this point),

$$P\{X \in A\} = \int_A f_X(x)dx.$$

**B.3.1 Expectations of random variables**

Let $X$ be a random variable. Then, the *expected value* of $X$ is

$$\mathbb{E}[X] = \sum_{x \in \mathcal{R}(X)} xp_X(x)$$

in the case of discrete $X$, and

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} xf_X(x)dx,$$

in the case of continuous $X$. The functions $p_X$ and $f_X(x)$ above are the probability mass function and density function, respectively. The expected value of a random variable is also called its *mean* or *expectation* and is often denoted $\mu$ or $\mu_X$. 

216
Example B.3.8. Consider a random variable taking values in \{1, \ldots, n\} with
\[ P\{X = i\} = \frac{1}{n}, \quad i \in \{1, \ldots, n\}. \]
We say that \(X\) is distributed uniformly over \{1, \ldots, n\}. What is the expectation?

Solution. We have
\[ \mathbb{E}[X] = \sum_{i=1}^{n} i P\{X = i\} = \sum_{i=1}^{n} \frac{1}{n} = \frac{1}{n} \frac{n(n+1)}{2} = \frac{n+1}{2}. \]

Example B.3.9. Consider rolling a die and letting \(X\) be the outcome. Then \(X\) is uniformly distributed on \{1, \ldots, 6\}. Thus, \(\mathbb{E}[X] = 7/2 = 3.5\).

Example B.3.10. Consider the weighted die from Example B.1.7. The expectation of the outcome is
\[
1 \times \frac{1}{10} + 2 \times \frac{1}{10} + 3 \times \frac{1}{10} + 4 \times \frac{1}{5} + 5 \times \frac{1}{4} + 6 \times \frac{1}{4} = \frac{83}{20} = 4.15.
\]

Example B.3.11. Suppose that \(X\) is exponentially distributed with density function
\[ f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & \text{else} \end{cases} \]
where \(\lambda > 0\) is a constant. In this case,
\[ \mathbb{E}[X] = \int_{0}^{\infty} x\lambda e^{-\lambda x} dx = \frac{1}{\lambda}. \quad \text{(B.1)} \]

Suppose we instead want the expectation of a function of a random variable: \(g(X) = g \circ X\), which is itself a random variable. That is, \(g \circ X : \Omega \to \mathbb{R}\). The following is proved in any introduction to probability book.

Theorem B.3.12. Let \(X\) be a random variable and let \(g : \mathbb{R} \to \mathbb{R}\) be a function. Then,
\[ \mathbb{E}[g(X)] = \sum_{x \in \mathbb{R}(X)} g(x)p_{X}(x), \]
in the case of discrete \(X\), and
\[ \mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x)f_{X}(x)dx, \]
in the case of continuous \(X\). The functions \(p_{X}\) and \(f_{X}(x)\) are the probability mass function and density function, respectively.

An important property of expectations is that for any random variable \(X\), real numbers \(\alpha_1, \ldots, \alpha_n\), and functions \(g_1, \ldots, g_n : \mathbb{R} \to \mathbb{R}\),
\[ \mathbb{E}[\alpha_1 g_1(X) + \cdots + \alpha_n g_n(X)] = \alpha_1 \mathbb{E}[g_1(X)] + \cdots + \alpha_n \mathbb{E}[g_n(X)]. \]
B.3.2 Variance of a random variable

The variance gives a measure on the “spread” of a random variable around its mean.

Definition B.3.13. Let $\mu$ denote the mean of a random variable $X$. The variance and standard deviation of $X$ are

\[ \text{Var}(X) = \mathbb{E}[(X - \mu)^2] \]
\[ \sigma_X = \sqrt{\text{Var}(X)}, \]

respectively. Note that the units of the variance are the square of the units of $X$, whereas the units of standard deviation are the units of $X$ and $\mathbb{E}[X]$.

A sometimes more convenient formula for the variance can be computed straight-away:

\[ \text{Var}(X) = \mathbb{E}[(X - \mu)^2] = \mathbb{E}[X^2 - 2\mu X + \mu^2] = \mathbb{E}[X^2] - 2\mu \mathbb{E}[X] + \mu^2 = \mathbb{E}[X^2] - \mu^2. \]

A useful fact that follows directly from the definition of the variance is that for any constants $a$ and $b$,

\[ \text{Var}(aX + b) = a^2 \text{Var}(X) \]
\[ \sigma_{aX+b} = |a| \sigma_X. \]

B.3.3 Some common discrete random variables

Bernoulli random variables: $X$ is a Bernoulli random variable with parameter $p \in (0, 1)$ if

\[ P\{X = 1\} = p, \]
\[ P\{X = 0\} = 1 - p. \]

Bernoulli random variables are quite useful because they are the building blocks for more complicated random variables. For a Bernoulli random variable with a parameter of $p$,

\[ \mathbb{E}[X] = p \quad \text{and} \quad \text{Var}(X) = p(1 - p). \]

For any event $A \in \mathcal{F}$, we define the indicator function $1_A$, or $I_A$, to be equal to one if $A$ occurs, and zero otherwise. That is,

\[ 1_A(\omega) = \begin{cases} 1, & \text{if } \omega \in A \\ 0, & \text{if } \omega \notin A \end{cases} \]

$1_A$ is a Bernoulli random variable with parameter $P(A)$. 218
Binomial random variables: Consider \( n \) independent repeated trials of a Bernoulli random variable. Let \( X \) be the number of “successes” (i.e. 1’s) in the \( n \) trials. The range of \( X \) is \( \{0, 1, \ldots, n\} \) and it can be shown that the probability mass function is

\[
P\{X = k\} = \begin{cases} \binom{n}{k} p^k (1 - p)^{n-k}, & \text{if } k \in \{0, 1, 2, \ldots, n\} \\ 0, & \text{else} \end{cases}
\]

Any random variable with probability mass function (B.2) is a binomial random variable with parameters \( n \) and \( p \).

Example B.3.14. From the interval \((0, 1)\), 10 points are selected at random. What is the probability that at least 5 of them are less than 1/3?

Solution: A success is defined by \( x_i < 1/3 \) for \( i \in \{1, 2, \ldots, 10\} \). Thus, \( p = 1/3 \). Let \( X \) be the number of successes. \( X \) is a binomial(10,1/3) random variable. The probability of at least 5 successes in 10 tries is then

\[
P\{X \geq 5\} = \sum_{k=5}^{10} \binom{10}{k} \left(\frac{1}{3}\right)^k \left(\frac{2}{3}\right)^{10-k} = 0.21312.
\]

For a binomial random variable with parameters \( n \) and \( p \),

\[
\mathbb{E}[X] = np \quad \text{and} \quad \text{Var}(X) = np(1 - p).
\]

Geometric random variables: Consider repeating a Bernoulli trial until a success happens. In this case, the sample space is

\[
\Omega = \{s, fs, ffs, fffs, \ldots\},
\]

where \( s \) denotes a success and \( f \) denotes a failure. Suppose that the probability of success is \( p \). Let \( X \) be the number of trials until a success happens. The range of \( X \) is \( \mathcal{R}(X) = \{1, 2, 3, \ldots\} \). The probability mass function is given by

\[
P\{X = n\} = \begin{cases} (1 - p)^{n-1} p, & n \in \{1, 2, 3, \ldots\} \\ 0, & \text{else} \end{cases}
\]

Any random variable with this probability mass function is called a geometric random variable with a parameter of \( p \). For a geometric random variable with a parameter of \( p \),

\[
\mathbb{E}[X] = \frac{1}{p} \quad \text{and} \quad \text{Var}(X) = \frac{1-p}{p^2}.
\]
Geometric random variables, along with exponential random variables, have the memoryless property. Let \( X \) be a Geometric(\( p \)) random variable. Then for all \( n, m \geq 1 \)

\[
P\{X > n + m \mid X > m\} = \frac{P\{X > n + m\}}{P\{X > m\}} = \frac{n + m \text{ failures to start}}{m \text{ failures to start}}
\]

\[
= \frac{(1 - p)^{n+m}}{(1 - p)^m} = (1 - p)^n = P\{X > n\}.
\]

In words, this says that the probability that the next \( n \) trials will be failures, given that the first \( m \) trials were failures, is the same as the probability that first \( n \) are failures.

**Poisson random variables:** This is one of the most important random variables in the study of stochastic models of reaction networks and will arise time and time again in this class.

A random variable with range \( \{0, 1, 2, \ldots \} \) is a Poisson random variable with parameter \( \lambda > 0 \) if

\[
P\{X = k\} = \begin{cases} 
\frac{\lambda^k e^{-\lambda}}{k!}, & k = 0, 1, 2, \ldots \\
0, & \text{else}
\end{cases}
\]

For a Poisson random variable with a parameter of \( \lambda \),

\[
\mathbb{E}[X] = \lambda \quad \text{and} \quad \text{Var}(X) = \lambda.
\]

The Poisson random variable, together with the Poisson process, will play a central role in this class, especially when we discuss continuous time Markov chains. The following can be shown by, for example, generating functions.

**Theorem B.3.15.** If \( X \sim \text{Poisson}(\lambda) \) and \( Y \sim \text{Poisson}(\mu) \) then \( Z = X + Y \sim \text{Poisson}(\lambda + \mu) \).

**The Poisson process.** An increasing process on the integers \( Y(t) \) is said to be a Poisson process with intensity (or rate or propensity) \( \lambda \) if

1. \( Y(0) = 0 \).
2. The number of events in disjoint time intervals are independent.
3. \( P\{Y(s + t) - Y(t) = i\} = e^{-\lambda_s} \frac{(\lambda_s)^i}{i!}, \; i = 0, 1, 2, \ldots \) for any \( s, t \geq 0 \)

The Poisson process will be the main tool in the development of (pathwise) stochastic models of biochemical reaction systems in this class and will be derived more rigorously later.
B.3.4 Some common continuous random variables

Uniform random variables. Uniform random variables will play a central role in efficiently generating different types of random variables for simulation methods. Consider an interval \((a, b)\), where we will often have \(a = 0\) and \(b = 1\). The random variable is said to be uniformly distributed over \((a, b)\) if

\[
F(t) = \begin{cases} 
0 & t < 1 \\
(t - a)/(b - a) & a \leq t < b \\
1 & t \geq b 
\end{cases},
\]

\[
f(t) = F'(t) = \begin{cases} 
1/(b - a) & a < t < b \\
0 & \text{else} 
\end{cases}.
\]

For a uniform random variable over the interval \((a, b)\),

\[
\mathbb{E}[X] = \frac{a + b}{2} \quad \text{and} \quad \text{Var}(X) = \frac{(b - a)^2}{12}.
\]

Normal random variables. Also called Gaussian random variables, normal random variables play a central role in the theory of probability due to their connection to the central limit theorem and Brownian motions. These connections will arise in this class when we consider diffusion approximations, called Langevin approximations in some of the sciences, to the continuous time Markov chain models of chemically reacting species.

A random variable \(X\) is called a normal with mean \(\mu\) and variance \(\sigma^2\), and we write \(X \sim N(\mu, \sigma^2)\), if its density is

\[
f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ \frac{-(x - \mu)^2}{2\sigma^2} \right\}, \quad x \in \mathbb{R}.
\]

A standard normal random variable is a normal random variable with \(\mu = 0\) and \(\sigma = 1\). For a normal random variable with parameters \(\mu\) and \(\sigma^2\),

\[
\mathbb{E}[X] = \mu \quad \text{and} \quad \text{Var}(X) = \sigma^2.
\]

Exponential random variables. For reasons to be demonstrated later, the exponential random variable will be the most important continuous random variable in the study of continuous time Markov chains. It will turn out to be linked to the Poisson process in that the inter-event times of a Poisson process will be given by an exponential random variable. This will lead to the important fact that many simulation methods will consist of generating a sequence of correctly chosen exponential random variables.

A random variable \(X\) has an exponential distribution with parameter \(\lambda > 0\) if it has a probability density function

\[
f(x) = \begin{cases} 
\lambda e^{-\lambda x} & x \geq 0 \\
0 & \text{else} 
\end{cases}.
\]
For an exponential random variable with a parameter of $\lambda > 0$,

$$E[X] = \frac{1}{\lambda} \quad \text{and} \quad \text{Var}(X) = \frac{1}{\lambda^2}.$$  

Similar to the geometric random variable, the exponential random variable has the memoryless property.

**Proposition B.3.16** (Memoryless property). Let $X \sim \text{Exp}(\lambda)$, then for any $s, t \geq 0$,

$$P\{X > (s + t) \mid X > t\} = P\{X > s\}. \quad (B.3)$$

Probably the most important role the exponential random variable will play in these notes is as the *inter-event time* of Poisson random variables.

**Proposition B.3.17.** Consider a Poisson process with rate $\lambda > 0$. Let $T_i$ be the time between the $i$th and $i + 1$st events. Then $T_i \sim \text{Exp}(\lambda)$.

The following propositions are relatively straightforward to prove and form the heart of the usual methods used to simulate continuous time Markov chains. This method is often termed the *Gillespie algorithm* in the biochemical community, and will be discussed later in the notes.

**Proposition B.3.18.** If for $i = 1, \ldots, n$, the random variables $X_i \sim \text{Exp}(\lambda_i)$ are independent, then

$$X_0 \equiv \min_i \{X_i\} \sim \text{Exp}(\lambda_0), \quad \text{where} \quad \lambda_0 = \sum_{i=1}^{n} \lambda_i.$$  

*Proof.* Let $X_0 = \min_i \{X_i\}$. Set $\lambda_0 = \sum_i \lambda_i$. Then,

$$P\{X_0 > t\} = P\{X_1 > t, \ldots, X_n > t\} = \prod_{i=1}^{n} P\{X_i > t\} = \prod_{i=1}^{n} e^{-\lambda_i t} = e^{-\lambda_0 t}.$$  

**Proposition B.3.19.** For $i = 1, \ldots, n$, let the random variables $X_i \sim \text{Exp}(\lambda_i)$ be independent. Let $j$ be the index of the smallest of the $X_i$. Then $j$ is a discrete random variable with probability mass function

$$P\{j = i\} = \frac{\lambda_i}{\lambda_0}, \quad \text{where} \quad \lambda_0 = \sum_{i=1}^{n} \lambda_i.$$  

222
Proof. We first consider the case of \( n = 2 \). Let \( X \sim \text{Exp}(\lambda) \) and \( Y \sim \text{Exp}(\mu) \) be independent. Then,

\[
P\{X < Y\} = \int \int_{x<y} \lambda e^{-\lambda x} \mu e^{-\mu y} \, dx \, dy = \int_0^\infty \int_0^y \lambda e^{-\lambda x} \mu e^{-\mu y} \, dx \, dy = \frac{\lambda}{\mu + \lambda}.
\]

Now, returning to the general case of arbitrary \( n \), we let \( Y_i = \min_{j \neq i} \{ S_j \} \), so that \( Y_i \) is exponential with rate \( \sum_{j \neq i} \lambda_j \) by Proposition B.3.18. Using the case \( n = 2 \) proved above then yields

\[
P\{X_i < \min_{j \neq i} \{ X_j \}\} = P\{X_i < Y_i\} = \frac{\lambda_i}{\lambda_i + \sum_{j \neq i} \lambda_j} = \frac{\lambda_i}{\sum_{j=1}^n \lambda_j}.
\]

One interpretation of the above two propositions is the following. If you have \( n \) alarm clocks, with the \( i \)th set to go off after an \( \text{Exp}(\lambda_i) \) amount of time, then Proposition B.3.18 tells you when the first will go off, and Proposition B.3.19 tells you which one will go off at that time.

B.3.5 Transformations of random variables

Most software packages have very good and efficient methods for the generation of pseudo-random numbers that are uniformly distributed on the interval \((0, 1)\). These pseudo random numbers are so good that we will take the perspective throughout these notes that they are, in fact, truly uniformly distributed over \((0, 1)\). We would then like to be able to construct all other random variables as transformations, or functions, of these uniform random variables. The method for doing so will depend upon whether or not the desired random variable is continuous or discrete. In the continuous case, Theorem B.3.20 will often be used, whereas in the discrete case Theorem B.3.22 will be used.

Theorem B.3.20. Let \( U \) be uniformly distributed on the interval \((0, 1)\) and let \( F \) be an invertible distribution function. Then \( X = F^{-1}(U) \) has distribution function \( F \).

Before proving the theorem, we show how it may be used in practice.

Example B.3.21. Suppose that we want to be able to generate an exponential random variable with parameter \( \lambda > 0 \). Such a random variable has distribution function \( F : \mathbb{R}_{\geq 0} \rightarrow [0, 1) \)

\[
F(t) = 1 - e^{-\lambda t}, \quad t \geq 0.
\]

Therefore, \( F^{-1} : [0, 1) \rightarrow \mathbb{R}_{\geq 0} \) is given by

\[
F^{-1}(u) = -\frac{1}{\lambda} \ln(1 - u), \quad 0 \leq u \leq 1.
\]

If \( U \) is \text{uniform}(0, 1), then so is \( 1 - U \). Thus, to simulate a realization of \( X \sim \text{Exp}(\lambda) \), you first simulate \( U \) from \text{uniform}(0, 1), and then set

\[
x = -\frac{1}{\lambda} \ln(U) = \ln(1/U)/\lambda.
\]
Proof. (of Theorem B.3.20) Letting $X = F^{-1}(U)$ where $U$ is uniform$(0, 1)$, we have

$$P\{X \leq t\} = P\{F^{-1}(U) \leq t\} = P\{U \leq F(t)\} = F(t).$$

\[\square\]

Theorem B.3.22. Let $U$ be uniformly distributed on the interval $(0, 1)$. Suppose that $p_k \geq 0$ for each $k \in \{0, 1, \ldots\}$, and that $\sum_k p_k = 1$. Define

$$q_k = P\{X \leq k\} = \sum_{i=0}^{k} p_i.$$

Let

$$X = \min\{k \mid q_k \geq U\}.$$

Then,

$$P\{X = k\} = p_k.$$

Proof. Taking $q_{-1} = 0$, we have for any $k \in \{0, 1, \ldots\}$,

$$P\{X = k\} = P\{q_k - 1 < U \leq q_k\} = q_k - q_{k-1} = p_k.$$

\[\square\]

In practice, the above theorem is typically used by repeatedly checking whether or not $U \leq \sum_{i=0}^{k} p_i$, and stopping the first time the inequality holds. We note that the theorem is stated in the setting of an infinite state space, though the analogous theorem holds in the finite state space case.

B.3.6 More than one random variable

To discuss more than one random variable defined on the same probability space $(\Omega, \mathcal{F}, P)$, we need joint distributions.

Definition B.3.23. Let $X_1, \ldots, X_n$ be discrete random variables with domain $\Omega$. Then

$$p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = P\{X_1 = x_1, \ldots, X_n = x_n\}$$

is called the joint probability mass function of $X_1, \ldots, X_n$.

Definition B.3.24. We say that $X_1, \ldots, X_n$ are jointly continuous if there exists a function $f(x_1, \ldots, x_n)$, defined for all reals, such that for all $A \subset \mathbb{R}^n$

$$P\{(X_1, \ldots, X_n) \in A\} = \int \cdots \int_{(x_1, \ldots, x_n \in A)} f(x_1, \ldots, x_n)dx_1 \ldots dx_n.$$

The function $f(x_1, \ldots, x_n)$ is called the joint probability density function.
Expectations are found in the obvious way.

**Theorem B.3.25.** If \( h : \mathbb{R}^n \rightarrow \mathbb{R} \) then
\[
\mathbb{E}[h(X_1, \ldots, X_n)] = \sum_{x_1 \in \mathcal{R}(X_1)} \cdots \sum_{x_n \in \mathcal{R}(X_n)} h(x_1, \ldots, x_n)p_{X_1, \ldots, X_n}(x_1, \ldots, x_n).
\]

**Corollary B.3.26.** For random variables \( X \) and \( Y \) on the same probability space
\[
\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y).
\]

**Definition B.3.27.** The random variables \( X \) and \( Y \) are **independent** if for any sets of real numbers \( A \) and \( B \)
\[
P\{X \in A, Y \in B\} = P\{X \in A\}P\{Y \in B\}.
\]
This implies that \( X \) and \( Y \) are independent if and only if
\[
p(x, y) = p_X(x)p_Y(y),
\]
\[
f(x, y) = f_X(x)f_Y(y),
\]
for discrete and continuous random variables, respectively.

**Theorem B.3.28.** Let \( X \) and \( Y \) be independent random variables and \( g : \mathbb{R} \rightarrow \mathbb{R} \) and \( h : \mathbb{R} \rightarrow \mathbb{R} \) be real valued functions; then \( g(X) \) and \( h(Y) \) are also independent random variables.

**Theorem B.3.29.** Let \( X \) and \( Y \) be independent random variables. Then for all real valued functions \( g \) and \( h \),
\[
\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)].
\]

One important application of the above theorem is the relation
\[
\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]
\]
if \( X \) and \( Y \) are independent. However, the converse is, in general, false.

**Example B.3.30.** Let \( R(X) = \{-1, 0, 1\} \) with \( p(-1) = p(0) = p(1) = 1/3 \). Let \( Y = X^2 \). We have
\[
\mathbb{E}[X] = 0, \quad \mathbb{E}[Y] = 2/3, \quad \text{and} \quad \mathbb{E}[XY] = 0.
\]
However,
\[
P\{X = 1, Y = 1\} = P\{Y = 1\}P\{X = 1\} = 1/3
\]
\[
P\{X = 1\}P\{Y = 1\} = (1/3) \times (2/3) = 2/9,
\]
demonstrating these are *not* independent random variables. \(\square\)

More generally, if \( X_1, X_2, \ldots, X_n \) are independent random variables, then
\[
\mathbb{E}[X_1 \cdots X_n] = \mathbb{E}[X_1] \cdots \mathbb{E}[X_n].
\]
B.3.7 Variance of linear combinations.

Suppose that

$$X = X_1 + X_2 + \cdots + X_n.$$  

We already know that for any $X_i$ defined on the same probability space

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X_i].$$

For the variance of a linear combination, a direct calculation shows that for $a_i \in \mathbb{R},$

$$\text{Var} \left( \sum_{i=1}^{n} a_iX_i \right) = \sum_{i=1}^{n} a_i^2 \text{Var}(X_i) + 2 \sum_{i=1}^{n} \sum_{i<j} a_i a_j \text{Cov}(X_i, X_j),$$

where

$$\text{Cov}(X_i, X_j) = \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)] = \mathbb{E}[X_iX_j] - 2\mu_i\mu_j + \mu_i\mu_j = \mathbb{E}[X_iX_j] - \mu_i\mu_j.$$ 

Therefore, if the $X_i$ are pairwise independent,

$$\text{Var} \left( \sum_{i=1}^{n} a_iX_i \right) = \sum_{i=1}^{n} a_i^2 \text{Var}(X_i).$$

**Example B.3.31.** Let $X$ be a binomial random variable with parameters $n$ and $p$. Since $X$ is the number of successes in $n$ independent trials, we can write

$$X = X_1 + \cdots + X_n,$$

where $X_i$ is 1 if $i$th trial was success, and zero otherwise. Therefore, the $X_i$’s are independent Bernoulli random variables and $\mathbb{E}[X_i] = P\{X_i = 1\} = p.$ Thus,

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \sum_{i=1}^{n} p = np.$$ 

Because each of the $X_i$’s are independent

$$\text{Var}(X) = \sum_{i=1}^{n} \text{Var}(X_i) = \sum_{i=1}^{n} p(1-p) = np(1-p).$$

\[\square\]

**Proposition B.3.32.** Let $X_1, \ldots, X_n$ be $n$ independent random variables with mean $\mu$ and variance $\sigma^2$. Let $\bar{X} = (1/n)(X_1 + \cdots + X_n)$ be the average of the sample. Then

$$\mathbb{E}(\bar{X}) = \mu, \quad \text{Var}(\bar{X}) = \frac{\sigma^2}{n}.$$ 

**Proof.** Calculating shows

$$\mathbb{E}[\bar{X}] = \mathbb{E} \left( \frac{X_1 + \cdots + X_n}{n} \right) = \frac{1}{n} n\mu = \mu$$

$$\text{Var}(\bar{X}) = \text{Var} \left( \frac{1}{n} (X_1 + \cdots + X_n) \right) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}.$$ 

\[\square\]
B.4 Inequalities and Limit Theorems

B.4.1 Important inequalities

Oftentimes we do not have explicit representations for the distributions, probability mass functions, or densities of the random variables of interest. Instead, we may have means, variances, or some other information. We may use this information to garner some bounds on probabilities of events.

**Theorem B.4.1** (Markov’s Inequality). Let $X$ be a non-negative random variable; then for any $t > 0$

$$P\{X \geq t\} \leq \frac{E[X]}{t}.$$  

*Proof.* The proof essentially uses the indicator functions $1_{\{X \geq t\}}$ and $1_{\{X < t\}}$ to break up $X$ into two pieces:

$$E[X] = E[X1_{\{X \geq t\}}] + E[X(1 - 1_{\{X \geq t\}})] \geq E[X1_{\{X \geq t\}}] \geq E[t1_{\{X \geq t\}}] = tP\{X \geq t\}.$$  

\[ \square \]

**Theorem B.4.2** (Chebyshev’s Inequality). If $X$ is a random variable with mean $\mu$ and variance $\sigma^2 < \infty$, then for any $t > 0$,

$$P\{|X - \mu| \geq t\} \leq \frac{\sigma^2}{t^2}.$$  

*Proof.* We have that $(X - \mu)^2 \geq 0$, so we may use the Markov inequality:

$$P\{(X - \mu)^2 \geq t^2\} \leq \frac{E[(X - \mu)^2]}{t^2} = \frac{\sigma^2}{t^2}.$$  

But, $P\{(X - \mu)^2 \geq t^2\} = P\{|X - \mu| \geq t\}$, and so the result is shown.  

\[ \square \]

B.4.2 Limit theorems

We now present three limit theorems that will be used later in the course.

**Theorem B.4.3** (Weak Law of Large Numbers). Let $X_1, X_2, X_3, \ldots$ be a sequence of independent and identically distributed random variables with $\mu = E[X_i]$ and $\sigma^2 = \text{Var}(X_i) < \infty, i = 1, 2, \ldots$. Then for all $\epsilon > 0$

$$\lim_{n \to \infty} P\left\{\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| > \epsilon\right\} = 0.$$  

*Proof.* Let $\bar{X} = (1/n)\sum_i X_i$ be the sample average. We know that $E\bar{X} = \mu$ and $\text{Var}(\bar{X}) = \sigma^2/n$. Thus, by Chebyshev’s inequality we get

$$P\{|\bar{X} - \mu| > \epsilon\} \leq \frac{\sigma^2/n}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0 \text{ as } n \to \infty.$$  

\[ \square \]
Note that the proof gives a rate of convergence of $O(1/n)$. The corresponding strong law of large numbers is now stated.

**Theorem B.4.4** (Strong law of large numbers). Let $X_1, X_2, X_3, \ldots$ be a sequence of independent and identically distributed random variables with mean $\mu$. then

$$P \left\{ \lim_{n \to \infty} \frac{X_1 + \cdots + X_n}{n} = \mu \right\} = 1.$$ 

So $\bar{X} = (X_1 + \cdots X_n)/n$ converges to $\mu$ almost surely, or with a probability of one.

We now state the central limit theorem.

**Theorem B.4.5** (Central Limit Theorem). Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables, each with expectation $\mu$ and variance $\sigma^2$. Then the distribution of

$$Z_n = \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$$

converges to the distribution of a standard normal random variable. That is, for any $t \in (-\infty, \infty)$

$$\lim_{n \to \infty} P\{Z_n \leq t\} = \lim_{n \to \infty} P \left\{ \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \leq t \right\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-x^2/2} dx.$$

**B.5 Estimation via simulation: Monte Carlo**

Consider the following question: given a random variable $X$ with unknown distribution function $F(x)$, how can we estimate $\mu = \mathbb{E}[X]$?

Assuming that we can generate realizations of $X$ via a computer, the simulation approach to solving this problem is to estimate $\mu = \mathbb{E}[X]$ by running $n$ independent and identical experiments, thereby obtaining $n$ i.i.d. random variables $X_1, X_2, \ldots, X_n$, with each having the distribution $F(x)$. Then, take the estimate as

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$ 

We call $\hat{\mu}_n$ an estimator. In this case, we have an unbiased estimator as

$$\mathbb{E}\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[X_i] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[X_i] = \mu.$$ 

Further, by the strong law of large numbers we know that

$$\hat{\mu}_n \to \mu,$$
as \( n \to \infty \), with a probability of one.

Of course, knowing that \( \hat{\mu}_n \to \mu \), as \( n \to \infty \), does not actually tell us how large of an \( n \) we need in practice. This brings us to the next logical question: how good is the estimate for a given, finite \( n \). To answer this question, we will apply the central limit theorem.

We know from the central limit theorem that

\[
\frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma \sqrt{n}} \xrightarrow{D} N(0,1),
\]

or

\[
\frac{\sqrt{n}}{\sigma} (\hat{\mu}_n - \mu) \xrightarrow{D} N(0,1).
\]

Specifically, for any \( z \in \mathbb{R} \)

\[
P \{-z \leq N(0,1) \leq z\} \approx P \left\{ -\frac{\sqrt{n}}{\sigma} (\hat{\mu}_n - \mu) \leq z \right\}
= P \left\{ -\frac{\sigma z}{\sqrt{n}} \leq (\hat{\mu}_n - \mu) \leq \frac{\sigma z}{\sqrt{n}} \right\}
= P \left\{ \hat{\mu}_n - \frac{\sigma z}{\sqrt{n}} \leq \mu \leq \hat{\mu}_n + \frac{\sigma z}{\sqrt{n}} \right\}.
\]

In words, the above says that the probability that the true value, \( \mu \), is within \( \pm \sigma z/\sqrt{n} \) of the estimator \( \hat{\mu}_n \) is \( P \{-z \leq N(0,1) \leq z\} \), which can be found for any \( z \). More importantly, a value \( z \) can be found for any desired level of confidence. The interval \( (\mu - \sigma z/\sqrt{n}, \mu + \sigma z/\sqrt{n}) \) is called our confidence interval and the probability \( P \{-z \leq N(0,1) \leq z\} \) is our confidence. Note that both our confidence and the size of the confidence interval increase as \( z \) is increased.

We now turn to finding the value \( z \) for a desired confidence level. Letting

\[
\Phi(z) = P\{N(0,1) \leq z\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2} dt,
\]

we have

\[
P \{-z \leq N(0,1) \leq z\} = P\{N(0,1) \leq z\} - P\{N(0,1) \leq -z\}
= \Phi(z) - \Phi(-z)
= \Phi(z) - (1 - \Phi(z))
= 2\Phi(z) - 1.
\]

Therefore, if for some \( \delta > 0 \) we want to have a probability of \( 1 - \delta \) that the true value is in the constructed confidence interval, then we must choose \( z \) so that

\[
P \{-z \leq N(0,1) \leq z\} = 1 - \delta.
\]

That is, we need to find a \( z \) so that

\[
2\Phi(z) - 1 = 1 - \delta,
\]

229
For example, if $\delta = .1$, so that a 90% confidence interval is required, then

$$\Phi(z) = 1 - .05 = .95,$$

and $z = 1.65$. If, on the other hand, we want $\delta = .05$, so that a 95% confidence interval is desired, then

$$\Phi(z) = 1 - .025 = .975$$

and $z = 1.96$.

Summarizing, we see that for a given $\delta$, we can find a $z$ so that the probability that the parameter $\mu$, which is what we are after, lies in the interval

$$\left[\hat{\mu}_n - \frac{\sigma z}{\sqrt{n}}, \hat{\mu}_n + \frac{\sigma z}{\sqrt{n}}\right]$$

is approximately $1 - \delta$. Further, as $n$ gets larger, the confidence interval shrinks. It is worth pointing out that the confidence interval shrinks at a rate of $1/\sqrt{n}$. Therefore, to get a 10-fold increase in accuracy, we need a 100-fold increase in work.

There is a major problem with the preceding arguments: if we don’t know $\mu$, which is what we are after, we most likely do not know $\sigma$ either. Therefore, we will also need to estimate it from our independent samples $X_1, X_2, \ldots, X_n$.

**Theorem B.5.1.** Let $X_1, \ldots, X_n$ be independent and identical samples with mean $\mu$ and variance $\sigma^2$, and let

$$\hat{\sigma}_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \hat{\mu}_n)^2,$$

where $\hat{\mu}_n$ is the sample mean

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Then,

$$\mathbb{E} \hat{\sigma}_n^2 = \sigma^2.$$

**Proof.** We have

$$\mathbb{E} \hat{\sigma}_n^2 = \mathbb{E} \left[ \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \hat{\mu}_n)^2 \right],$$

which yields

$$(n - 1)\mathbb{E} \hat{\sigma}_n^2 = \mathbb{E} \left[ \sum_{i=1}^{n} (X_i - \hat{\mu}_n)^2 \right] = \sum_{i=1}^{n} \mathbb{E} [X_i^2] - 2\mathbb{E} \left[ \hat{\mu}_n \sum_{i=1}^{n} X_i \right] + n\mathbb{E} [\hat{\mu}_n^2]

= \sum_{i=1}^{n} \mathbb{E} [X_i^2] - 2\mathbb{E} [\hat{\mu}_n n\hat{\mu}_n] + n\mathbb{E} [\hat{\mu}_n^2]

= n\mathbb{E} [X^2] - n\mathbb{E} [\hat{\mu}_n^2].$$
However,

\[ \mathbb{E} \left[ \hat{\mu}_n^2 \right] = \text{Var}(\hat{\mu}_n) + \mathbb{E} [\mu_n]^2 = \text{Var} \left( \frac{1}{n} \sum_{i=1}^{n} X_i \right) + \mu^2 = \frac{1}{n} \sigma^2 + \mu^2. \]

Therefore,

\[ \frac{n-1}{n} \mathbb{E} \hat{\sigma}_n^2 = \mathbb{E} [X^2] - \mathbb{E} [\hat{\mu}_n^2] = (\sigma^2 + \mu^2) - \left( \frac{1}{n} \sigma^2 + \mu \right) = \frac{n-1}{n} \sigma^2, \]

completing the proof.

Therefore, we can use

\[ \hat{\sigma}_n = \sqrt{\hat{\sigma}_n^2} \]

as an estimate of the standard deviation in the confidence interval and

\[ \left[ \hat{\mu}_n - \frac{\hat{\sigma}_n z}{\sqrt{n}}, \hat{\mu}_n + \frac{\hat{\sigma}_n z}{\sqrt{n}} \right] \]

is an approximate \((1 - \delta)100\%\) confidence interval for \(\mu = \mathbb{E}[X]\).

We note that there are two sources of error in the development of the above confidence interval that we will not explore here. First, there is the question of how good an approximation the central limit theorem is giving us. For reasonably sized \(n\), this should not give too much of an error. The second source of error is in using \(\sigma_n\) as opposed to \(\sigma\). Again, for large \(n\), this error will be relatively small.

We have the following algorithm for producing a confidence interval for an expectation given a number of realizations.

Algorithm for producing confidence intervals for a given \(n\).

1. Select \(n\), the number of experiments to be run, and \(\delta > 0\).

2. Perform \(n\) independent replications of the experiment, obtaining the observations \(X_1, X_2, \ldots, X_n\) of the random variable \(X\).

3. Compute the sample mean and sample variance

\[ \hat{\mu}_n = \frac{1}{n} (X_1 + \cdots + X_n) \]

\[ \hat{\sigma}_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \hat{\mu}_n)^2. \]

4. Select \(z\) such that \(\Phi(z) = 1 - \delta/2\). Then an approximate \((1 - \delta)100\%\) confidence interval for \(\mu = \mathbb{E}[X]\) is

\[ \left[ \hat{\mu}_n - \frac{\hat{\sigma}_n z}{\sqrt{n}}, \hat{\mu}_n + \frac{\hat{\sigma}_n z}{\sqrt{n}} \right]. \]
If a level of precision is desired, and \( n \) is allowed to depend upon \( \delta \) and a tolerance \( \epsilon \), then the following algorithm is most useful.

**Algorithm for producing confidence intervals to a given tolerance.**

1. Select \( \delta > 0 \), determining the desired confidence, and \( \epsilon > 0 \) giving the desired precision. Select \( z \) such that \( \Phi(z) = 1 - \delta/2 \).

2. Perform independent replications of the experiment, obtaining the observations \( X_1, X_2, \ldots, X_n \) of the random variable \( X \), until

\[
\frac{\hat{\sigma}_n z}{\sqrt{n}} < \epsilon.
\]

3. Report

\[
\hat{\mu}_n = \frac{1}{n}(X_1 + \cdots + X_n)
\]

and the \((1 - \delta)100\%\) confidence interval for \( \hat{\mu} = \mathbb{E}[X] \),

\[
\left[ \hat{\mu}_n - \frac{\hat{\sigma}_n z}{\sqrt{n}}, \hat{\mu}_n + \frac{\hat{\sigma}_n z}{\sqrt{n}} \right] \approx [\mu - \epsilon, \mu + \epsilon].
\]

There are two conditions usually added to the above algorithm. First, there is normally some minimal number of samples generated, \( n_0 \) say, before one checks whether or not \( \hat{\sigma}_n z/\sqrt{n} < \epsilon \). Second, it can be time consuming to compute the standard deviation after every generation of a new iterate. Therefore, one normally only does so after every multiple of \( M \) iterates, for some positive integer \( M > 0 \). We will not explore the question of what “good” values of \( n_0 \) and \( M \) are, though taking \( n_0 = M \approx 100 \) is usually sufficient.

### B.6 Exercises

Some of the problems require you to use MATLAB. You must always turn in your MATLAB code when a problem is solved using MATLAB.

1. Verify, through a direct calculation, Equation (B.1).

2. Verify the memoryless property, Equation (B.3), for exponential random variables.

3. MATLAB exercise. Perform the following tasks using Matlab.

   (a) Using a FOR LOOP, use the etime command to time how long it takes Matlab to generate 100,000 exponential random variables with a parameter of \( 1/10 \) using the built-in exponential random number generator. Sample code for this procedure is provided on the course website.
(b) Again using a FOR LOOP, use the etime command to time how long it takes Matlab to generate 100,000 exponential random variables with parameter $1/10$ using the transformation method given in Theorem B.3.20.

4. MATLAB exercise. Let $X$ be a random variable with range $\{-10, 0, 1, 4, 12\}$ and probability mass function

$$P\{X = -10\} = \frac{1}{5}, \quad P\{X = 0\} = \frac{1}{8}, \quad P\{X = 1\} = \frac{1}{4},$$

$$P\{X = 4\} = \frac{1}{3}, \quad P\{X = 12\} = \frac{11}{120}.$$ Using Theorem B.3.22, generate $N$ independent copies of $X$ and use them to estimate $\mathbb{E}X$ via

$$\mathbb{E}X \approx \frac{1}{N} \sum_{i=1}^{N} X[i],$$

where $X[i]$ is the $i$th independent copy of $X$ and $N \in \{100, 10^3, 10^4, 10^5\}$. Compare the result for each $N$ to the actual expected value. A helpful sample MATLAB code has been provided on the course website.


σ-algebra, 212
n-step transition probability, 21
absorbing state, 26
accessible, 25
aperiodic, 27
axonal transport, 17
Bias, 177
birth and death, discrete time, 68
branching process, 78
Brownian motion, 191
chemical master equation, 136
class property, 28
closed set, 26
communicate, 25
communication class, 26
conditional probability, 213
continuous time Markov chain, 125, 128
cumulative distribution function, 215
defective renewal process, 94
deterministically monotone Markov chain, 15
discrete time Markov chain, 13
distribution function, 215
doubly stochastic matrix, 15
Euler’s method, 176
events, 210
explosive process, 132
finite dimensional distribution, 12
First order reaction networks, 184
generator matrix of a Markov chain, 136
genotype, 62
Gillespie algorithm, 130
Gillespie’s Algorithm, 173
homogeneous Poisson process, 107
homogeneous process, 14
independent sets, 213
infinitesimal generator, 136
intensity function, 108
invariant measure, 34
irreducible DTMC, 26
Itô integral, 198
Itô isometry, 195
Jukes-Cantor model, 66
kinetic Monte Carlo, 130
Kolmogorov forward equations, 136
Lebesgue measure, 102
Linear noise approximation, 200
Markov matrix, 15
Markov property, 13
mass action kinetics - stochastic form, 164
mean first return time, 43
mean recurrence time, 43
Mendelian genetics, 61
Next reaction method, 174
null recurrent, 43
one-step transition probability, 14
partition, 214
period, 27
Perron-Frobenius theorem, 39
phenotype, 62
phylogenetic distance, 66
point process, 102
Poisson process as counting process, 106
Poisson process via infinitesimal assumptions, 104
Poisson process with mean measure, 106
positive recurrent, 43
probability measure, 212
probability space, 210
propensity function, 108
proper renewal process, 94
pure renewal process, 94
queuing model, 70
random measure, 103
Random time change, 170
random time change representation, 121
random variable, 11, 215
random walk, 16, 18
reaction rate constant, 162
recurrence, 29
reducible DTMC, 26
renewal function, 94
renewal process, 93
renewal reward process, 99
renewal times, 94
sample space, 210
simulate a DTMC, 20
simulation algorithm for continuous time
Markov chain, 129
state space, 11
stationary distribution, 34
stochastic matrix, 15
stochastic process, 11
Stochastic simulation algorithm, 173
stochastic simulation algorithm, 130
tau-leaping, 176
Time dependent intensity functions, 175
time homogeneous process, 14
transience, 29
transition matrix, 15

Weak error, 177
Wiener process, 191
Wright-Fischer model, 65