Introduction to Probability Models

Sivaraman Balakrishnan

Department of Statistics
Carnegie Mellon University,
Pittsburgh, PA 15213.

siva@stat.cmu.edu

May 3, 2018

Abstract

These notes are based on previous versions of this course taught by several faculty in the Department of Statistics at CMU. They also contain lots of material from Sheldon Ross' book. The material on Markov Decision Processes is from Pieter Abbeel and Dan Klein's Intro to AI class at Berkeley.

Contents

Chapter 1

Administrivia

Most of this information can be found in the Syllabus or Calendar. Consult the Calendar for exam dates.

1.1 Administrative things

The course will use Canvas and Piazza. The Piazza URL:

- https://piazza.com/cmu/spring2018/36410/home

Sign up. Please don't sign up as an instructor!

Most important things (class policies, office hours, grading scheme etc.) can be found in the Syllabus document.

1.2 About the class

Broadly, the class is about stochastic processes. We'll be more formal later on but roughly a structured collection of random variables is what we refer to as a stochastic process, and this class will be broadly about various types of stochastic processes and their applications. What will distinguish this course from a typical statistics course is the emphasis on collections of random variables that are *dependent*.

To belabor this point, a typical or canonical problem in a first statistics class is: I have a coin, I want to know if it is a fair coin or not, I will toss the coin a few times and do something with the outcomes I observe. Here each outcome is a random variable (we'll define this more formally in the next Chapter), and importantly the outcomes of the coin tosses are independent random variables, i.e. roughly the outcome of the first coin toss has no effect on the next one. However, many real world phenomena are much better modeled as random variables that have some dependence structure. Typical examples range from the weather to stock prices – in both cases it should be clear that the weather tomorrow is certainly not independent of the weather today, and similarly for stock prices.

So the basic question we would like to understand throughout this course is:

How do we model **dependent** *random variables and what are the main properties of models for dependent random variables?*

The course will roughly be divided into:

1. A quick probability refresher
2. A large portion on Markov Chains
3. Applications of Markov Chains (Pagerank, Reinforcement Learning, Hidden Markov Models, MCMC, ...)
4. Other stochastic processes (Poisson Processes, Brownian Motion, Continuous time Markov Chains)

It is traditionally a course that is heavy in mathematical concepts.

1.3 What is a stochastic (probability) model?

In statistics, machine learning, economics, and all of the basic sciences we use *models* to understand the real-world. Roughly, a model is an idealization (i.e. a simplification) of the real-world that typically is mathematically tractable (i.e. we can study its properties) but still captures some of the real-world phenomena we are interested in. We will see many models in this course. We can broadly characterize models as deterministic and stochastic:

- A *deterministic model* has no random component. A deterministic model will return the same output if restarted from the same *initial conditions*. Physical laws often described by differential equations are a popular example.
- A *stochastic model* or *probability model* has random components. We won't spend much time on this but it is important to ponder where stochastic modeling arises in the "real world". Where does randomness come from?
    1. Physical randomization: Clinical trials, polling (TV ratings), gambling.
    2. Randomization as a modeling device: In many cases, we use stochastic models to describe our uncertainty which can arise as a proxy for our inability to correctly model a complex system. One can argue that few things are "genuinely random".

1.3.1 Models as approximations

Stochastic modeling is a central idea in Statistics, Machine Learning, Data Science etc. Why is this the case?

1. A simplified view of machine learning: Start with our guess of a probability model for the world, observe data, update our guess for the probability model of the world.
2. More pragmatically, one can think of models purely as a way to derive useful algorithms, i.e. *probabilistic reasoning* can lead to good algorithms.

To paraphrase George Box: "All models are wrong, some are useful." Why are models often wrong? Most often the make too many assumptions/simplification. Why do we still use them? Often *interpretable*, easier to work with and good for certain tasks. You should have seen most of this in earlier courses. What are we doing that is different?

One of the most important assumptions we often make in statistics or machine learning is the i.i.d. (independent, identically distributed) assumption. Roughly, every item in our

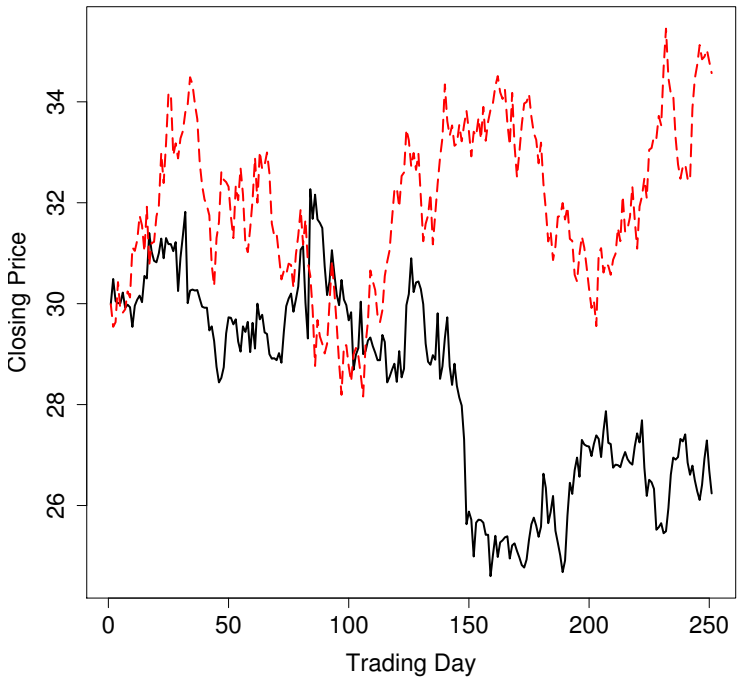


Figure 1.1: Stock prices of the PEET over a period of one year, and the values of a Gaussian process with the same starting point. Which is which? The main point of the example is that even though we don't believe the stochastic model exactly, it can be a useful way to model our ignorance of the world.

dataset is independent. Sometimes reasonable: maybe knowing my height doesn't help me guess yours. Very often not: stock prices, text, speech.

The stochastic models we will see in this class (Markov Chains, Gaussian processes etc.) break the assumption of independence. This seemingly simple idea, will introduce many new models and ideas, and broaden the scope and applicability of the tools at your disposal.

Chapter 2

Introduction to Probability

This chapter introduces some basic ideas in probability. Most of this will be familiar to all of you but there are a few important ideas that you need to pay special attention to. I will point them out as we go along.

References: See Chapter 1 of the Ross book.

2.1 Probability review

1. **Sample space:** Formally, the set of all possible outcomes of an experiment constitute the sample space S.

   Example 2.1. What is the sample space:

   (a) When I toss a coin?

   (b) When I roll two dice?

   (c) When I throw a dart that lands in the unit circle in $\mathbb{R}^2$?

   In the first case the sample space is $S = \{H, T\}$, in the second case there are 36 possible outcomes $S = \{(1,1), (1,2), \ldots, (1,6), (2,1), \ldots, (6,6)\}$. In the third case, the space of possible outcomes is the unit sphere in $\mathbb{R}^2$, i.e.

   $$S = \{(x, y) : x^2 + y^2 \leq 1, (x, y) \in \mathbb{R}^2\}.$$

2. **Event:** Any subset $E \subseteq S$ is an event.

   Example 2.2. When we toss two coins an example of an *event* is that the sum of the two dice equals 6, i.e.

   $$E = \{(1,5), (2,4), (3,3), (4,2), (5,1)\}.$$

   We say that an event occurs if the outcome belongs to E. You can perform usual set operations on events (i.e. take unions, intersection and complements of events).

3. **Probability Distribution/Measure:** A measure is a mapping from events to the reals, i.e. $\mathbb{P} : E \mapsto \mathbb{R}$, that satisfies the so called *axioms* of probability.

   - The probability of the sample space is 1, i.e.

   $$\mathbb{P}(S) = 1.$$

   - For an event E we have that,

   $$0 \leq \mathbb{P}(E) \leq 1.$$

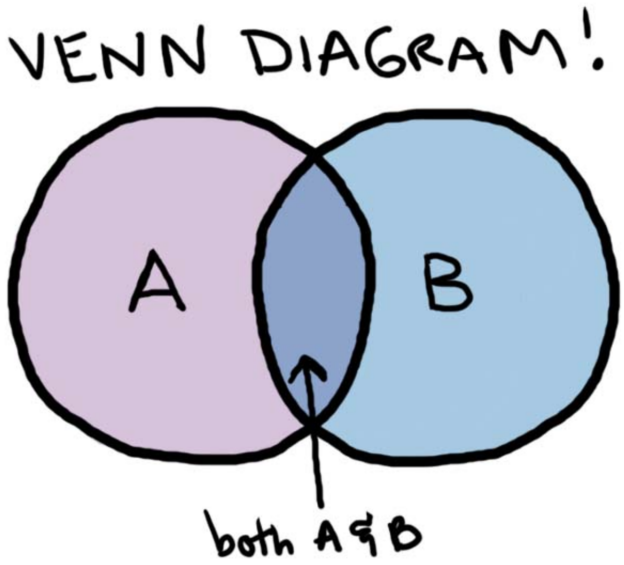
   - Finally, for disjoint events $E_1, E_2, \ldots,$

   $$\mathbb{P}\left(\bigcup E_i\right) = \sum_i \mathbb{P}(E_i).$$

This is the formal way of defining a distribution. One can ask where does this mapping come from? The answer connects back to the experiment we were performing, and gives maybe a "computational" way for defining the mapping, by repeating the experiment many times. If our experiment is repeated over and over again then (with probability 1) the proportion of time that event E occurs will just be $\mathbb{P}(E)$.

2.2 Some basic rules

Always keep in mind a Venn diagram.



Staring at this Venn diagram you can deduce some basic properties satisfied by any probability distribution.

1. **The addition rule:**

   $$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

   In particular, if the events are disjoint then: $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.

2. **The complement rule:**

$$\mathbb{P}(A^c) = 1 - \mathbb{P}(A).$$

Lets do a quick example.

Example 2.3. A drug has the following information:

1. There is a 10% chance of experiencing headache (H).
2. There is a 15% chance of experiencing nausea (N).
3. There is a 5% chance of experiencing both side effects.

What is the probability of experiencing at least one side effect?

We need to compute the probability of $H \cup N$ so we use the addition rule:

$$\mathbb{P}(H \cup N) = \mathbb{P}(H) + \mathbb{P}(N) - \mathbb{P}(H \cap N) = 0.2.$$

2.3 Counting problems

Several probability calculations involve counting combinations. The basic rule is: if there are n distinct objects there are

$$\binom{n}{k} = \frac{n!}{k!(n-k)!},$$

possible combinations of size k. Lets see a couple of examples of using this idea.

Example 2.4. I give you 5 cards (drawn randomly without replacement) from a deck. What is the probability that you get exactly 1 jack?

The probability of getting exactly one J is the ratio of the number of hands with exactly one J to the total number of possible hands, i.e.

$$\mathbb{P}(\text{exactly one J}) = \frac{\#\text{ hands with 1 J}}{\#\text{ hands}} = \frac{\binom{4}{1}\binom{48}{4}}{\binom{52}{5}}.$$

Example 2.5. Poker hands: Which is more likely a four of a kind or a full house?

In poker roughly, you receive 5 cards drawn randomly without replacement (as in the previous example). A four of a kind is a hand that looks like: $\{K, K, K, K, \star\}$ or $\{7, 7, 7, 7, \star\}$. A full house is a hand where you have three of one type and two of another, i.e. hands like $\{6, 6, 6, 2, 2\}$.

We can calculate the probabilities of each of these hands as below:

$$\mathbb{P}(4 \text{ of a kind}) = \frac{13 \times 48}{\binom{52}{5}} \approx 0.00024,$$

and

$$\mathbb{P}(\text{full house}) = \frac{13 \times \binom{4}{3} \times 12 \times \binom{4}{2}}{\binom{52}{5}} \approx 0.0015.$$

So we see that the full house is more likely. Incidentally, this is how hands are ranked in poker, i.e. if you have a less likely hand then you win, so a 4 of a kind beats a full house.

2.4 Conditional Probability

Often we want to calculate the probability that an event A occurs given that an event B occurred. We use the notation: $\mathbb{P}(A|B)$. This is only defined when $\mathbb{P}(B) > 0$. The rule for conditional probability:

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

Example 2.6. We roll a fair dice once:

$$A = \{\text{even number}\},$$
$$B = \{1, 2, 3, 5\}.$$

What is $\mathbb{P}(A|B)$?

We do this by computing $\mathbb{P}(A \cap B) = 1/6$, and $\mathbb{P}(B) = 4/6$, so we obtain that,

$$\mathbb{P}(A|B) = \frac{1}{4}.$$

Example 2.7. I flip two coins, hidden from your view. You ask, "Is at least one coin heads? " I say, "Yes." What is probability (from your perspective) that both are heads?

Maybe a quick guess would be that the probability is 1/2, but lets try to calculate it more carefully. We define the events:

$$A = \{\text{both heads}\},$$
$$B = \{\text{at least one heads}\}.$$

So we can see that, $\mathbb{P}(B) = 3/4$ and $\mathbb{P}(A \cap B) = \mathbb{P}(A) = 1/4$, so we conclude that, $\mathbb{P}(A|B) = 1/3$.

Now that we have the notion of conditional probabilities we can derive a few more rules.

1. **The Multiplication rule:** I call this the *chain rule*. It gives a way to re-write a joint probability as a "chain" of probabilities, i.e.

$$\mathbb{P}(A \cap B) = \begin{cases} \mathbb{P}(A)\mathbb{P}(B|A) \\ \mathbb{P}(B)\mathbb{P}(A|B) \end{cases}$$

2. **The Law of Total Probability:** The law of total probability gives a way to compute a probability by partitioning this probability computation via conditioning, i.e. the law says that for any event B:

$$\mathbb{P}(A) = \mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c).$$

The way to derive this is to observe that:

$$\mathbb{P}(A) = \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^c),$$

and then use the definitions of conditional probabilities.

Example 2.8. A bank is considering extending credit to a new customer and is interested in the probability that the client will default on the loan. Based on historical data, the bank knows that there is a 5% chance that a customer who has overdrawn an account will default, while there is only a 0.5% chance that a customer who has never overdrawn an account will default. Unfortunately, the bank does not know for sure if the customer will overdraw her account. Based on background checks the bank believes there is a 30% chance that the customer will overdraw the account. Calculate the probability that she will default if credit is extended.

So first we need to translate the problem into some events, and then try to compute the desired probability. One way to do this is to say

1. $A = \{\text{customer defaults on the loan}\}$, and
2. $B = \{\text{customer overdraws her account}\}$,

so we want to compute $\mathbb{P}(A)$. We are given:

$$\begin{aligned} \mathbb{P}(A|B) &= 0.05 \\ \mathbb{P}(A|B^c) &= 0.005 \\ \mathbb{P}(B) &= 0.3, \end{aligned}$$

so via the law of total probability we have

$$\begin{aligned} \mathbb{P}(A) &= \mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c) \\ &= 0.05 \times 0.3 + 0.005 \times 0.7 \\ &\approx 0.0185. \end{aligned}$$

We can also generalize the law of total probability. First we define the notion of a partition of the sample space. We say that $B_1, B_2, \ldots, B_n$ is a *partition* of S if

- $\bigcup_{i=1}^n B_i = S$.
- $B_i \cap B_j = \phi, \quad \text{for } \ i \neq j.$

Given a partition of S we have that,

$$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A|B_i)\mathbb{P}(B_i).$$

How does this generalize the previous version? We see that we can apply this version with the partition $S = B \cup B^c$ to obtain the previous version.

Example 2.9. There are 3 drawers, one full of white socks, the second full of black socks, and the third half black and half white. Select one drawer at random, set it aside. Then draw one sock randomly from each of the remaining two drawers. What is the probability of getting a matching pair?

The key yet again is to define a convenient set of events. For instance,

$$\begin{aligned} A &= \{\text{get a matching pair}\}, \\ B_1 &= \{\text{all white drawer left out}\}, \\ B_2 &= \{\text{all black drawer left out}\}, \\ B_3 &= \{\text{50-50 drawer left out}\}. \end{aligned}$$

So we want to calculate $\mathbb{P}(A)$. We have:

$$\begin{aligned} \mathbb{P}(A) &= \mathbb{P}(A|B_1)\mathbb{P}(B_1) + \mathbb{P}(A|B_2)\mathbb{P}(B_2) + \mathbb{P}(A|B_3)\mathbb{P}(B_3) \\ &= \frac{1}{2} \times \frac{1}{3} + \frac{1}{2} \times \frac{1}{3} + 0 \times \frac{1}{3} \\ &= \frac{1}{3}. \end{aligned}$$

Example 2.10. In the game *Phew!* you have a spinner that can land in one of three Phew regions ? "You win," (W) "You lose," (L) and "Spin Again" (A) – with respective probabilities p, q, and r, which sum to one. You continue to spin until you either win or lose. Find the probability that you win the game by conditioning on an appropriate random variable.

There are really many different approaches to solve this problem:

1. Brute force: We list every sequence that results in us winning and sum up their probabilities, i.e. we win if the sequence of outcomes looks like one of these: $\{W, AW, AAW, \ldots\}$. The probabilities of these sequences are: $p, rp, r^2p, \ldots$, so the probability of winning is:

$$\mathbb{P}(\text{win}) = p + rp + r^2p + \ldots = \frac{p}{1-r}.$$

2. Conditioning on the first step: This is the most natural way of solving the problem, and just sets up a recursion. Suppose we let θ be the probability that we win, then by the law of total probability:

$$\begin{aligned} \theta &= \mathbb{P}(\text{win}|\text{first spin} = W)\mathbb{P}(W) + \mathbb{P}(\text{win}|\text{first spin} = L)\mathbb{P}(L) + \mathbb{P}(\text{win}|\text{first spin} = A)\mathbb{P}(A) \\ &= 1 \times p + 0 \times q + \theta \times r, \end{aligned}$$

   so solving for θ we get the same expression as before.

2.4.1 Bayes' rule

This rule is quite central to probability theory (and also statistics, machine learning, etc.). First the simpler version:

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c)} = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)}.$$

You should think about using Bayes' rule when you want to calculate $\mathbb{P}(B|A)$ and $\mathbb{P}(A|B)$ is known or easier to calculate.

The general version of Bayes' rule: Suppose $B_1, B_2, \ldots B_n$ are a partition of S, and A is any event then:

$$\mathbb{P}(B_i|A) = \frac{\mathbb{P}(A|B_i)\mathbb{P}(B_i)}{\mathbb{P}(A)}.$$

Lets consider a few examples.

Example 2.11 (Polygraph tests)**.** If a person is lying, the probability that this is correctly detected by the polygraph is 0.88, whereas if the person is telling the truth, this is correctly detected with probability 0.86. Suppose we are consider a question for which 99% of all subjects tell the truth.

Our polygraph machine says a subject is lying on this question. What is the probability that the polygraph is incorrect?

Once again, we first try to define some events that make sense. One possible solution is to take:

$$\begin{aligned} A &= \{\text{polygraph says the subject is lying}\}, \\ B &= \{\text{subject is actually lying}\}. \end{aligned}$$

Then our goal is to compute $\mathbb{P}(B^c|A)$. We are given:

$$\begin{aligned} \mathbb{P}(A|B) &= 0.88 \\ \mathbb{P}(A^c|B^c) &= 0.86, \\ \mathbb{P}(B) &= 0.01. \end{aligned}$$

Applying Bayes' rule we then get:

$$\mathbb{P}(B^c|A) = \frac{\mathbb{P}(A|B^c)\mathbb{P}(B^c)}{\mathbb{P}(A|B^c)\mathbb{P}(B^c) + \mathbb{P}(A|B)\mathbb{P}(B)} = \frac{0.14 \times 0.99}{0.14 \times 0.99 + 0.88 \times 0.01} \approx 0.94.$$

2.5 Independence

One of the key concepts of this course: recall, Markov chains deviate from the traditional assumption of independence. Events $A_1, A_2, \ldots A_n$ are independent if for every collection of *distinct* indices $\{i_1, i_2, \ldots, i_k\}$,

$$\mathbb{P}(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k}) = \mathbb{P}(A_{i_1})\mathbb{P}(A_{i_2}) \ldots \mathbb{P}(A_{i_k}).$$

Example 2.12. Consider a sequence of n *independent* trials, each of which has a probability $1/n$ of being a "success". What is the probability of zero successes in n trials? What if the number of trials is doubled?

The probability is simply:

$$\begin{aligned}\mathbb{P}(\text{failure on trial } 1, \ldots, \text{failure on trial n}) &= \prod_{i=1}^{n} \mathbb{P}(\text{failure on trial i}) \\ &= \left(1 - \frac{1}{n}\right)^n \approx \exp(-1).\end{aligned}$$

If we double the number of trials then:

$$\begin{aligned}\mathbb{P}(\text{failure on trial } 1, \ldots, \text{failure on trial 2n}) &= \prod_{i=1}^{2n} \mathbb{P}(\text{failure on trial i}) \\ &= \left(1 - \frac{1}{n}\right)^{2n} \approx \exp(-2).\end{aligned}$$

Example 2.13. In 3 out of 45 games that Arlene bowls, she scores at least 280 points. What are her chances of bowling a game over 280 points in her next 18 games?

We can first compute the probability that Arlene bowls under 280 in one game:

$$\mathbb{P}(\text{Arlene scores } \leq \text{ 280 in 1 game}) = \frac{42}{45},$$

so by independence we obtain,

$$\mathbb{P}(\text{Arlene scores } \leq \text{ 280 in all 18 games}) = \left(\frac{42}{45}\right)^{18}.$$

Finally,

$$\mathbb{P}(\text{Arlene scores } \geq \text{ 280 in at least 1 game}) = 1 - \left(\frac{42}{45}\right)^{18} \approx 0.711.$$

Chapter 3

Expectations, Conditional Expectations and Common Distributions

In the last chapter we considered some basic probability from the perspective of events. Particularly, we focused on defining a distribution, methods to calculate the probability of various events using different rules (Bayes' rule, Chain rule, Law of Total Probability and so on). In this chapter we will finish our probability review, and focus on understanding and manipulating random variables.

Reference: See Chapters 2-3 of the Ross book.

3.1 Random Variables

So what is a random variable? Formally, its a map from the outcomes to the reals, i.e. X is a random variable if it can be expressed as a map $X : S \mapsto \mathbb{R}$, where S is the sample space of outcomes, and $\mathbb{R}$ is the real numbers.

This is perhaps not the most intuitive definition. The way I like to think about random variables is that they are a "lens" through which we can "summarize" the outcome of our experiment, i.e. a random variable is a measurement of a stochastic system.

Example 3.1. Suppose we flip a coin twice: $S = \{HH, HT, TH, TT\}$. Describe some random variables.

1. Suppose we take the random variable:

$$X(HH) = 2, \quad X(HT) = X(TH) = 1, \quad X(TT) = 0.$$

   Here the random variable is counting the number of heads and it is a map from the outcomes to $\{0, 1, 2\}$.

2. Another random variable measuring a different stochastic snapshot of the same experiment would be:

$$Y(HH) = 1, \quad Y(HT) = 1, \quad Y(TH) = 0, \quad Y(TT) = 0.$$

   This random variable is simply 1 if the first coin toss outcome was H.

3. A particular type of random variable that we will see often is an *indicator* RV, i.e. for some event E, we define the random variable:

$$\mathbb{I}_E = \begin{cases} 1 & \text{outcome } \in E. \\ 0 & \text{otherwise.} \end{cases}$$

A note on the shorthand notation: Suppose X is a random variable, we often use notation like $\mathbb{P}(X = 1)$. The way you should interpret this is as a shorthand for the probability of all the outcomes for which the random variable takes the value 1, i.e.

$$\mathbb{P}(X = 1) = \mathbb{P}(\{\omega : \omega \in S, X(\omega) = 1\}).$$

Example 3.2 (Indicator Random Variables)**.** Suppose that our experiment consists of measuring how long bulbs last in a household. Further assume we are not interested in precisely how long the bulbs last but just if they last over 5 years or not.

We would use the indicator random variable:

$$\mathbb{I}_{\text{life} \geq 5} = \begin{cases} 1 & \text{if bulb lasts for over 5 years} \\ 0 & \text{otherwise.} \end{cases}$$

This example illustrates an important use of random variables, i.e. we can use them to "coarsen" the sample space. We do not really care about the outcome (i.e. how long the bulb lasted), we only care about a much coarser property (i.e. did it last over 5 years), and so we define a random variable that gives us this information. Another aspect we will return to is that indicator random variables interact nicely with expectations.

Example 3.3. Suppose we consider a discrete random walk on the integers, and are interested in its location after k time steps. What is the appropriate random variable to define and what values can it take?

First, lets define formally a discrete random walk on the integers. The random walk begins at 0, and at each subsequent time step does one of three things (each equally likely): it stays in the same place, it moves right or it moves left. We are actually defining a *sequence* of random variables (this is a stochastic process or a Markov chain but we will define all of this later) where we have: $X_0 = 0$, and

$$X_i = \begin{cases} X_{i-1} + 1 & \text{with probability } 1/3 \\ X_{i-1} - 1 & \text{with probability } 1/3 \\ X_{i-1} & \text{with probability } 1/3. \end{cases}$$

After k steps the random variable X_k takes values in the set $\{-k, -(k-1), \ldots, 0, \ldots, k-1, k\}$.

3.2 Cumulative distribution function

A random variable is *random*. There is no way to determine its value before running the stochastic experiment. In order to *understand* random variables, and make a-priori decisions, we often consider non-random functions of the random variable (cumulative distribution functions, expectations, variances etc.).

The cumulative distribution function or CDF of a random variable X, is defined as,

$$F_X(x) = \mathbb{P}(X \leq x).$$

The CDF F is defined for every x, and *uniquely* determines the distribution of the random variable X.

Example 3.4 (The Kolmogorov-Smirnov test)**.** In statistics, we often make distributional modeling assumptions, i.e. for example that the noise follows a standard Gaussian distribution (i.e. a Gaussian distribution with mean 0 and variance 1). If we are being careful, we would also like to "test" this assumption. Typically this is done via something called the Kolmogorov-Smirnov test.

Given an independent and identically distributed sample the idea of the KS test, is to compare the CDF of the samples to the CDF of the distribution we think that the sample follows.

We compute the so called "Empirical" CDF (i.e. the CDF of the samples)

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_i \leq x),$$

and compare this with the "true" CDF (in this case the Gaussian CDF):

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)dx.$$

The KS test, basically compares these two CDFs and rejects the assumption if the difference between them is large. The key point to take away is that this whole idea is leveraging the fact that the CDF is a unique characterization of the distribution of a random variable.

Example 3.5. Two fair coins are tossed and the outcome is observed. Before the coins are tossed, we are given the following choice of payoffs:

Payoff 1:

- Win \$ 1 for each head.
- Lose \$ 3 for getting two tails.

Payoff 2:

- Win \$ 1 if the coins are different.
- Win \$ 2 if both coins turn up tails.
- Lose \$ 3 if both coins turn up heads.

Which payoff should we choose?

Lets denote the first random payoff by the RV X and the second by Y. We have the following table:

	X	Y
HH	2	-3
HT	1	1
TH	1	1
TT	-3	2

There is no reason to prefer one Payoff to the other – they have exactly the same CDF. Note the important point – this does not mean that the random variables are equal – they just have the same distribution.

3.3 Discrete and Continuous random variables

RVs are sometimes divided into two groups, discrete and continuous. There are RVs that are neither (sometimes called *mixed*).

- **Common discrete distributions:** Bernoulli, Binomial, Poisson, ...
- **Common continuous distributions:** Gaussian, Exponential, ...

We will see some of these repeatedly later in the course but we will describe their properties as we need them.
A discrete RV X typically has an associated probability *mass* function (pmf):

$$f_X(x) = \mathbb{P}(X = x).$$

A continuous RV X has an associated probability *density* function (pdf), also denoted $f_X(x)$, with the property that for $a \leq b$,

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x)dx.$$

For a continuous random variable, we think of

$$\mathbb{P}(X = x) = 0 \neq f_X(x).$$

That is to say, a continuous RV does not take any particular value with positive probability. You can draw infinitely many times from a Gaussian distribution and you will never see a sample that is exactly 0. Roughly, the density $f_X(x)$ is capturing the probability that a continuous random variable takes a value close to x.

3.4 Expectations

Often, things we care about are properties of random variables: is it big or small? Will we win money or not? How much will we win? Again, RVs are *random* and so not tangible before running the experiment. The *expectation* is an important non-random property of a random variable that we can "calculate" before running the experiment.

For a **discrete** random variable,

$$\mathbb{E}(X) = \sum_x x f_X(x).$$

For a **continuous** random variable,

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x) dx.$$

Example 3.6. A point X is chosen uniformly between $[0, 1]$. Let $T = X^2$. Find $\mathbb{E}(T)$.

Since this is a continuous random variable we use integration as above, i.e.:

$$\begin{aligned}\mathbb{E}[T] &= \int_{-\infty}^{\infty} x^2 \mathbb{I}_{x\in[0,1]} dx \\ &= \int_0^1 x^2 dx = \frac{1}{3}.\end{aligned}$$

This example highlights something important. It is often called the rule of the lazy statistician, or sometime the law of the unconscious statistician which has the catchy acronym LOTUS. The rule says that, if we want to calculate: $\mathbb{E}[g(X)]$ we can just use the formula:

$$\mathbb{E}[g(X)] = \int_x g(x) f_X(x) dx,$$

rather than first compute the density of the transformed RV,$Y = g(X)$, and compute the expected value as $\mathbb{E}[Y] = \int y f_Y(y) dy$.

Example 3.7. A bucket contains 100 balls, 30 of which are blue and the rest are red. We draw a random subset of 5 balls (uniformly). Let B be the number of blue balls in the chosen subset, and define the random variable $I = \mathbb{I}_{B=3}$. Find the expectation, $\mathbb{E}(I)$.

This is a discrete random variable, so we use the appropriate formula:

$$\mathbb{E}[I] = \frac{\sum_{\text{subsets}} \mathbb{I}(\text{subset has 3 blue balls})}{\binom{100}{5}} = \frac{\binom{30}{3}\binom{70}{2}}{\binom{100}{5}}.$$

More generally, we always have the relationship between expectations of indicators and probabilities:

$$\mathbb{E}[\mathbb{I}_E] = 1 \times \mathbb{P}(E) + 0 \times \mathbb{P}(E^c) = \mathbb{P}(E),$$

i.e. the expected value of an indicator is a probability.

3.4.1 Expectations as averages

An intuitive way to think about an expectation is as the average value of a random variable. Roughly, if I measured a random variable many times (independently), and took their average, I would obtain the expectation (formally, this is the law of large numbers).

Averages follow certain intuitive properties, and expectations do as well.

1. The **constancy** rule: For any constant c, we have that, $\mathbb{E}[c] = c$.
2. The **scaling** rule: For any constant c, $\mathbb{E}[cX] = c\mathbb{E}[X]$.
3. The **ordering** rule: If $X \leq Y$ always then, $\mathbb{E}[X] \leq \mathbb{E}[Y]$.
4. The **additivity** rule: This one is most important, it is also called the linearity of expectations, i.e.

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y].$$

3.4.2 Conditional Expectation and the Law of Total Expectation

We often want to calculate the expected value of some random variable *conditioned on some event* E, i.e. the conditional expectation is simply the average value of the random variable when we restrict to events in E. Formally, we compute this just like a conditional probability, assuming $\mathbb{P}(E) > 0$:

$$\mathbb{E}[X|E] = \frac{\mathbb{E}[\mathbb{I}_E X]}{\mathbb{P}(E)},$$

assuming for instance that the random variable is continuous you can also write this as:

$$\mathbb{E}[X|E] = \int_x x f_{X|E}(x)dx,$$

where $f_{X|E}$ is the density of X conditioned on the event E.

Much more important is the law of total expectation which like the law of total probability gives a divide and conquer way to compute an expectation. Suppose $B_1, \ldots, B_k$ is a partition of the sample space, then:

$$\mathbb{E}[X] = \sum_{i=1}^{k} \mathbb{E}[X|B_i]\mathbb{P}(B_i).$$

So we can reduce computing an expectation to computing many (hopefully easier) conditional expectations.

Example 3.8. Suppose that two factories supply light bulbs to the market. Factory X's bulbs work for an average of 5000 hours, whereas factory Y's bulbs work for an average of 4000 hours. It is known that factory X supplies 60% of the total bulbs available. What is the expected length of time that a purchased bulb will work for?

Applying the law of total expectation, we have:

$$\mathbb{E}(L) = \mathbb{E}(L \mid X)\mathbb{P}(X) + \mathbb{E}(L \mid Y)\mathbb{P}(Y) = 5000(.6) + 4000(.4) = 4600.$$

Let us look at a couple of nice applications of the law of total expectation:

Example 3.9 (The expectation of sum of a random number of RVs)**.** Suppose that the average number of car accidents in a day in Pittsburgh is μ_1 and that each time there is a car accident on average μ_2 people are injured. What is the average number of people injured in car accidents in a day?

It might initially seem like insufficient information. In particular, it is not completely clear why knowing only the means of the two distributions suffices. Lets try conditioning: let X denote the number of people injured and Y denote the number of accidents then:

$$\begin{aligned}\mathbb{E}[X] &= \sum_{y=0}^{\infty} \mathbb{E}[X|Y=y]\mathbb{P}(Y=y) \\ &= \sum_{y=0}^{\infty} \mu_2 y \mathbb{P}(Y=y) \\ &= \mu_1\mu_2.\end{aligned}$$

This fact is sometimes called Wald's identity.

Example 3.10 (The mean of the Geometric distribution)**.** A coin, with probability of heads p, and we keep tossing it till we see a head. What is the mean number of flips we need?

The distribution of the number of flips is a Geometric distribution, i.e. if we let X be the number of flips we need then:

$$\mathbb{P}(X=k) = (1-p)^{k-1}p.$$

You can now brute force compute the mean. The more elegant way is to condition on the first step, i.e.

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}[X|\text{first toss is heads}]\mathbb{P}(H) + \mathbb{E}[X|\text{first toss is tails}]\mathbb{P}(T) \\ &= 1 \times p + (1 + \mathbb{E}[X]) \times (1-p),\end{aligned}$$

and solving this we obtain that,

$$\mathbb{E}[X] = \frac{1}{p},$$

which should intuitively make sense, i.e. if p is small we expect to need many tosses.

3.5 Variance

For a random variable, the *variance* is a measure of the spread of its distribution. Formally,

$$\mathbb{V}(X) = \mathbb{E}(X - \mathbb{E}(X))^2.$$

The variance captures something like "on average how far is X from its mean $\mathbb{E}[X]$".

3.5.1 Pairs of random variables

Given a pair of random variables we can measure various "cross" quantities, and this leads to various definitions.

3.5.2 Co-variance

The covariance between two random variables X and Y is defined as:

$$\text{cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))],$$

i.e. we say that RVs have high co-variance if they fluctuate around their mean together, i.e. when X is bigger than $\mathbb{E}[X]$ we have that Y is also usually bigger than $\mathbb{E}[Y]$ and vice versa.

3.5.3 Properties of variance and co-variance

Like the expectation, there are some simple rules that help calculate the variance of transformations of a random variable:

1. $$\mathbb{V}(X) = \mathbb{E}(X - \mathbb{E}(X))^2 = \mathbb{E}(X^2 + (\mathbb{E}(X))^2 - 2X\mathbb{E}(X)) = \mathbb{E}X^2 - (\mathbb{E}(X))^2.$$

2. $$\mathbb{V}(X + Y) = \mathbb{V}(X) + \mathbb{V}(Y) + 2\text{cov}(X, Y).$$

   Importantly, the variance is *not* linear unless the covariance is 0.

3. For any constant c, $\mathbb{V}(c) = 0$.

4. $$\mathbb{V}(aX + b) = a^2\mathbb{V}(X).$$

3.6 Independence

We have discussed independence of events and now we turn our attention to independence of random variables. Independence means that you cannot predict one random variable using the other. Roughly, uncorrelated means you cannot predict the other random variable using a linear function.

There are two equivalent ways of defining independence:

1. The more common or intuitive way is that X and Y are independent random variables if their joint distribution factorizes i.e.,

   $$f_{XY}(x, y) = f_X(x)f_Y(y).$$

   This is perhaps easier to interpret for discrete RVs, where we have the condition that:

   $$\mathbb{P}(X = x \cap Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y).$$

   A consequence of independence that is often useful is that:

   $$\mathbb{P}(X = x|Y = y) = \mathbb{P}(X = x),$$

   i.e. knowing Y does not affect the distribution of X if they are independent.

2. Alternatively two random variables X, Y are independent if for every function f, g we have that,

$$\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\mathbb{E}[g(Y)],$$

i.e. every function of the two random variables is uncorrelated.

Chapter 4

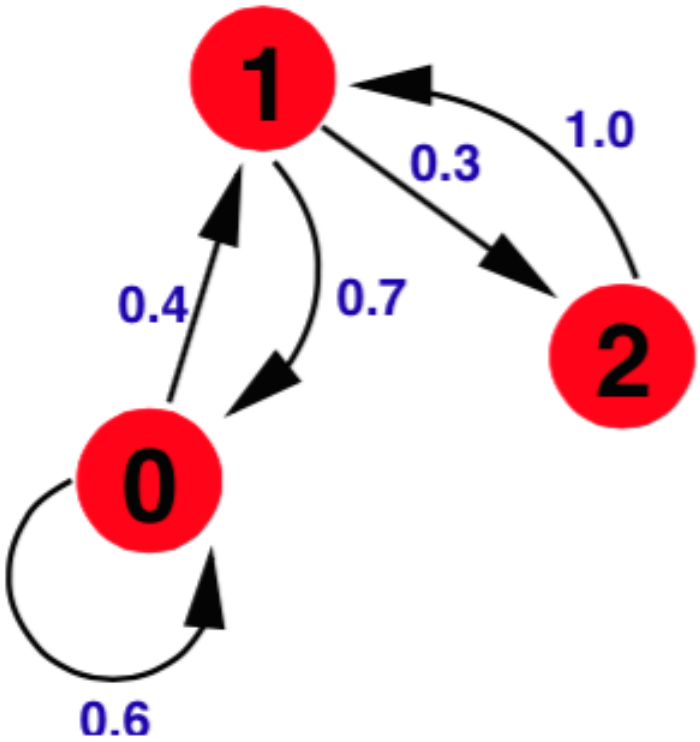
Markov Chain Basics

In this chapter we will begin to discuss the basics of Markov chains. Markov chains are named after Andrey Markov. Here is the story from an online article: "The story starts in the early 20th century with one of those obscure academic squabbles that usually don't amount to much. A mathematician and theologian named Pavel Nekrasov argued that since independent events follow the law of big numbers and social phenomena such as crime statistics do as well, then humans must have free will.

Andrei Markov, one of the great mathematicians of the day, thought Nekrasov's argument was hogwash. After all, he noted, just because independent variables follow a certain mathematical law doesn't mean that directed activity can't do so as well.

To prove his point, he performed a mathematical analysis of Eugene Onegin, Pushkin's famous novel in verse, and showed that the combinations of vowels and consonants followed the law of big numbers as well. A vowel, would most likely be followed by a consonant and vice versa, in proportions that became more stable as you analyzed more text."

We will see more formal definitions later, but for now lets consider a "state diagram".



Some features:

- The state is a random variable and the state evolves over time.
- The nodes represent states, and the edges represent transitions.
- The numbers on the edges are probabilities, they represent conditional probabilities of the next state given the current state. For instance if the process is in State 0, it will transition to State 1 in the next time step with probability 0.4. We often represent these probabilities as a *transition matrix.*
- *For this to be a valid state diagram the sum of weights on outgoing edges from any node should be 1.*

Question 4.1. This one is a bit open ended. What are some questions we could ask (and hopefully answer), about the behavior of this process?

- From a practical standpoint, what can I model using processes like this? Why are they useful?
- From a more conceptual standpoint: suppose I start in some state i, what is the expected number of steps I need to take before I visit every state at least once? This is known as the *cover time.*
- Is there some sense in which a stochastic process converges? In particular, suppose I started in state i and walked "forever", what is the average amount of time I would spend in some state j? This is known as the *stationary distribution.*
- Suppose I start in some state i what is the expected time to hit another state j? This is known as the *hitting time.*
- Suppose I start in a state i what is the probability of being in some state j after k-steps. This is the *k-step transition probability.*

You can see that some of these questions are about the long-term behaviour of the stochastic process while others are trying to characterize its short-term behaviour.

4.1 A few canonical examples

We will see many different Markov chains throughout this class but to set the stage for some of the tools we will develop lets discuss a few examples.

4.1.1 Random walks

There are many different random walks that we will study but here are a few.

- **Simple Random Walk:** Start at zero and then at each time step move up by 1 or down by 1 with probability p and $1-p$ respectively. Here the states are $\{\ldots, -2, -1, 0, 1, 2, \ldots\}$, and the transition matrix is:

$$P = \begin{pmatrix} & \vdots & \vdots & \vdots & \vdots & \vdots & \\ \ldots & 1-p & 0 & p & 0 & 0 & \ldots \\ \ldots & 0 & 1-p & 0 & p & 0 & \ldots \\ \ldots & 0 & 0 & 1-p & 0 & p & \ldots \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \end{pmatrix}.$$

- **Random Walk with One-Sided Barrier:** A variant of this is where we start at 0 and move up or down by 1 at each step with probability p and $1-p$ respectively. However, if we are at state 0 and attempt to move down we hit a barrier and remain at state 0. The states are $\{0, 1, 2, \ldots, \}$, and the transition matrix is:

$$P = \begin{pmatrix} 1-p & p & 0 & 0 & 0 & \ldots \\ 1-p & 0 & p & 0 & 0 & \ldots \\ 0 & 1-p & 0 & p & 0 & \ldots \\ 0 & 0 & 1-p & 0 & p & \ldots \\ & & \vdots & & & \end{pmatrix}.$$

- **Random Walk with Two-Sided Barrier:** Same as above except there are two barriers, one at 0 and one at some fixed positive number N. The upper barrier also reflects the random walk as above, i.e. if we are in state N then with probability p we remain in state N and move to state $N-1$ with probability $1-p$.

4.1.2 Gambler's ruin

A slight variant of the random walk Markov chain introduces so called *absorbing states*. As the name suggests these are states in which the stochastic process gets trapped, i.e. once the process reaches an absorbing state with probability 1 the process remains in this absorbing state.

An important example is the so-called Gambler's ruin problem. Here we have a Gambler who has an initial fortune of $i\$$, and at each step wins a dollar with probability p and loses a dollar with probability $1-p$. For some fixed number N, the Gambler stops if his wealth reaches either $N\$$ or $0\$$ (i.e. the Gambler is ruined).

The state space is $\{0, 1, \ldots, N\}$, and the transition matrix is:

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & \ldots & 0 \\ 1-p & 0 & p & 0 & 0 & \ldots & 0 \\ 0 & 1-p & 0 & p & 0 & \ldots & 0 \\ 0 & 0 & 1-p & 0 & p & \ldots & 0 \\ & & \vdots & & & & \\ 0 & & \ldots & & & & 1 \end{pmatrix}.$$

A natural question, that we will show how to answer is then: what is the probability that the Gambler is ruined in the long-run?

4.2 Classification of Stochastic Processes

Formally, a stochastic process is a family of random variables $\{X_t : t \in \mathcal{T}\}$, where each X_t takes some value in the state space, denoted $\mathbb{S}$. We call $\mathcal{T}$ the index set.

Often we refer to the index set as *time*. We can imagine a broad *classification* of various stochastic processes:

1. **Discrete time, discrete state:** These are Markov chains like the ones above.

2. **Discrete time, continuous state:** These are typically used in tracking problems, where the goal is to track an object moving continuously in space but which we only observe at discrete time-points. The canonical example of such a stochastic process is the Kalman filter.

3. **Continuous time, discrete state:** A canonical example is something called a birth and death process. Imagine you were to model the population of the US over time: the population changes when there are births or deaths, but these need not happen at fixed intervals, i.e. they happen in continuous time.

4. **Continuous time, continuous state:** The canonical example is a Brownian motion, which is just like a random walk, except we take infinitesimal steps, i.e. we move up or down by a tiny amount in a very tiny window.

4.3 The Markov Assumption

All of the examples we have talked about so far have been implicitly making a crucial independence assumption. This assumption is called the Markov assumption or the Markov property. Suppose that the time right now is n, then we often refer to:

1. $X_0, X_1, \ldots, X_{n-1}$ as the *past*.

2. X_n as the *present*.

3. $X_{n+1}, X_{n+2}, \ldots$ as the *future*.

An important question for a discrete time stochastic process:

What do the values of the random variables in the past and present tell us about the distribution of X_{n+1}?

More formally, what can we say about the probability $\mathbb{P}(X_{n+1} = j | X_n = i, X_{n-1} = k, \ldots)$? If we assume $X_0, X_1, \ldots$ are independent, then

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

In words, the past and present give us *no information* about the future. This assumption is easy to work with from a theoretical and computational point of view, but clearly not realistic in many practical situations.

At the other extreme, if we assume that the distribution of X_{n+1} is dependent on **all** of the past $(X_0, X_1, \ldots, X_n)$, then the model will be cumbersome, and will grow more complex as time progresses and will be impractical from both statistical and computational standpoints.

The middle ground is a Markov chain. A Markov Chain of **order** k assumes that

$$\mathbb{P}(X_{n+1} = j | X_0 = i_0, X_1 = i_1, \ldots, X_n = i) = \mathbb{P}(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \ldots, X_{n-k+1} = i_{n-k+1}).$$

So, **conditional on** $X_{n-k+1}, \ldots, X_n$ the state X_{n+1} is *independent* of $X_0, \ldots, X_{n-k}$. In words, the future is independent of the "distant" past conditional on the "near" past.

In fact, all of our examples so far (random walks, Gambler's ruin) have actually been Markov chains of order 1. We focus on these **first order** Markov Chains throughout the course. Why? One justification is that it is possible to make any Markov chain of any order into a first order chain by enlarging the state space appropriately. This idea is best illustrated by an example.

Example 4.1. I want to model the daily weather as a discrete time chain with three states:

$$\mathbb{S} = \{\text{sunny}, \text{cloudy}, \text{rainy}\}.$$

I believe that the distribution over $\mathbb{S}$ for tomorrow's weather can be adequately predicted by the state of the weather yesterday and today. Thus, this is a second order Markov chain. Show that by enlarging the state space, you can create a new *first* order Markov Chain with the same behavior as this second order one.

Lets denote sunny $= S$, cloudy $= C$, rainy $= R$. Imagine a sequence of observations, say $SSCRRSSSRC\star$, i.e. so we have $X_0 = S, X_1 = S, X_2 = C$ and so on.

Suppose we enlarged the state space to include two days of weather (representing yesterday's weather and today's weather), then

$$\mathbb{S}' = \{SS, SC, SR, CS, CC, CR, RS, RC, RR\}.$$

We could then represent the same sequence as $X_0 = SS, X_1 = SC, X_2 = CR, \ldots$. Note that this Markov chain, from the state SS can only transition to one of $\{SR, SR, SS\}$.

The original Markov chain was second order but with the new state space we now have a first-order Markov chain since the state now incorporates both yesterday and today's weather. This is a general principle, i.e. you can convert a k-th order Markov chain into a first-order Markov chain with an enlarged state space.

4.3.1 The Transition Matrix

For a first-order Markov Chain,

$$P(X_{n+1} = j | X_0 = i_0, X_i = i_1, \ldots, X_n = i) = \mathbb{P}(X_{n+1} = j | X_n = i).$$

This quantity is central. We denote

$$P_{ij}(n) = \mathbb{P}(X_{n+1} = j | X_n = i).$$

These are the **transition probabilities** at time n. In words, the transition probability at time n, $P_{ij}(n)$ is the probability of moving from state i at time n to state j at time $n + 1$.

There is another simplification that is common: we call a Markov Chain **time homogeneous** if we assume that $P_{ij}(n)$ does not depend on n. In this case, we just denote the transition probability from state i to state j as P_{ij}. As we have seen already, we can arrange these probabilities into a matrix which we refer to as the transition matrix.

Time homogeneous, first-order Markov chains are prevalent. They are often referred to as simply **Markov chains.**

Example 4.2. $\{X_n : n \geq 0\}$ is a (time homogeneous, first-order) Markov chain. Are the random variables X_1 and X_3 independent?

The answer is that decidedly not independent. The past is independent of the future *conditional on the present*. In particular,

$$\mathbb{P}(X_3 = j | X_1 = k) \neq \mathbb{P}(X_3 = j),$$

but we do have that,

$$\mathbb{P}(X_3 = j | X_1 = k, X_2 = i) = \mathbb{P}(X_3 = j | X_2 = i) = P_{ij}.$$

If you prefer a more intuitive explanation, consider something like a random walk like we considered in the previous chapter, after 3 steps you can be in any location from $\{-3, \ldots, 0, \ldots, 3\}$. However, if I tell you that $X_1 = 1$, then X_3 can only be some location between $\{-1, 0, 1, 2, 3\}$, so the distributions of X_3 and $X_3|X_1$ are clearly different, and so X_3 is not independent of X_1.

4.4 Two-step Transition Probabilities

A natural quantity to attempt to calculate is the two-step transition probability, i.e. we might be interested in:

$$\mathbb{P}(X_{n+2} = j | X_n = i) := P_{ij}^{(2)}.$$

One way to calculate this is to use the law of total probability and condition on the step in between, i.e.

$$\mathbb{P}(X_{n+2} = j | X_n = i) = \sum_k \mathbb{P}(X_{n+2} = j | X_{n+1} = k, X_n = i) \mathbb{P}(X_{n+1} = k | X_n = i).$$

Now by the Markov property we can simplify this as:

$$\begin{aligned} \mathbb{P}(X_{n+2} = j | X_n = i) &= \sum_k \mathbb{P}(X_{n+2} = j | X_{n+1} = k) \mathbb{P}(X_{n+1} = k | X_n = i) \\ &= \sum_k P_{ik} P_{kj}. \end{aligned}$$

In words, this is just telling us that the probability of going from i to j in two-steps is the same as summing over paths of length 2, where we go from i to k in one step and from k to j in the next step.

Suppose we wanted to represent this fact as a matrix of equations (one for each $P_{ij}^{(2)}$). Then observe that we can write this set of equations compactly using the operation of matrix multiplication, i.e. I claim that:

$$P^{(2)} = \begin{pmatrix} P_{11}^{(2)} & P_{12}^{(2)} & \ldots \\ \vdots & & \\ P_{n1}^{(2)} & P_{n2}^{(2)} & \ldots \end{pmatrix} = P \times P.$$

This just follows by observing that the expression we derived above is simply the inner product between the i-th row of the transition matrix and the j-th column of the transition matrix.

Observe that the two-step transition probabilities do not depend on the time index n either. This is a consequence of the time-homogeneity of the Markov chain.

We can generalize this further, suppose that we wanted to calculate m-step transition matrices/probabilities, i.e. we wanted to compute:

$$\mathbb{P}(X_{n+m} = j | X_n = i) := P_{ij}^{(m)}.$$

Then we could again condition on the first-step (or if you want to you could also condition on the last step), to see that,

$$\begin{aligned}
\mathbb{P}(X_{n+m} = j | X_n = i) &= \sum_k \mathbb{P}(X_{n+m} = j | X_n = i, X_{n+1} = k) \mathbb{P}(X_{n+1} = k | X_n = i) \\
&= \sum_k \mathbb{P}(X_{n+m} = j | X_{n+1} = k) P_{ik} \\
&= \sum_k P_{kj}^{(m-1)} P_{ik}.
\end{aligned}$$

We can again express this expression in matrix form as:

$$P^{(m)} = P \times P^{(m-1)}.$$

You can unroll this expression completely or use mathematical induction to see that,

$$P^{(m)} = \underbrace{P \times P \times \ldots \times P}_{m \text{ times}} = P^m.$$

This very neat fact, that the m-step transition probabilities are just the 1-step transition probabilities raised to the m-th power, is known as the *Chapman-Kolmogorov* equation. This fact is at the heart of many "convergence" properties of Markov chains that we will study when we study their long-term behaviour, but for now observe that we have a concrete answer (one that is extremely easy to compute) to the short/intermediate-term question: *what is the probability of being in some state j after m-steps starting from some state i?*

The following example will illustrate this and a few other ideas.

Example 4.3. Suppose that we are distributing balls in 8 urns, and each ball is equally likely to be placed in any of the urns. What is the probability that there will be exactly 3 non-empty urns after 9 balls have been placed?

The first and most critical thing to do is to set up the correct Markov chain. We will see examples like this repeatedly. Intuitively, we would like the state of the Markov chain to be the number of nonempty urns, i.e the state space is $\{0, 1, 2, 3, \ldots, 8\}$. We start in state 0, and the transition probabilities are:

$$P_{i,i+1} = 1 - \frac{i}{8}, \quad \text{and} \quad P_{i,i} = \frac{i}{8}.$$

Now, we are interested in $P_{0,3}^{(9)}$, and to compute this we would simply take the transition matrix and raise it to the 9-th power and look at the corresponding entry of the matrix.

4.4.1 Distribution given an initial state distribution

So far, we have only been calculating conditional distributions, i.e. the conditional probability of being in some state in the future *given* our current location. We can also calculate unconditional probabilities by defining the initial state distribution:

$$\pi_i = \mathbb{P}(X_0 = i).$$

So if we start in some state j (deterministically) then $\pi_j = 1$ and the other entries are 0. We can arrange these values into a vector:

$$\pi = \begin{bmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_n \end{bmatrix}.$$

Now, observe that,

$$\pi^T \times P = \begin{bmatrix} \sum_k P_{k1}\pi_k & \sum_k P_{k2}\pi_k & \dots & \sum_k P_{kn}\pi_k \end{bmatrix}.$$

Now let us take one of these expressions and examine it, i.e.

$$\sum_k P_{k1}\pi_k = \sum_k \mathbb{P}(X_1 = 1 | X_0 = k)\mathbb{P}(X_0 = k) = \mathbb{P}(X_1 = 1),$$

and so on. So we have the relationship that,

$$\begin{bmatrix} \mathbb{P}(X_1 = 1) & \mathbb{P}(X_1 = 2) & \dots \mathbb{P}(X_1 = n) \end{bmatrix} = \pi^T P,$$

and more generally,

$$\begin{bmatrix} \mathbb{P}(X_m = 1) & \mathbb{P}(X_m = 2) & \dots \mathbb{P}(X_m = n) \end{bmatrix} = \pi^T P^m.$$

4.5 Classifying States

So far we have tried to understand the short/medium term properties of a Markov chain, particularly, we have tried to understand the k-step transition probabilities and the distribution over the different states after k-steps. With an eye towards understanding the long-term/limiting behavior of a Markov chain we will try to understand how to group states in a Markov chain. These "groups" will share properties and the groups will be important in understanding the limiting behavior.

The groups in Markov chains are referred to as *classes*. First, we say that state i is accessible from state j if there is some $n \geq 0$ such that, $P_{ij}^{(n)} > 0$. So a state j is accessible from state i if and only if starting in i, it is possible for the stochastic process to enter j. To see this formally, we simply observe that if j were not accessible from i then:

$$\begin{aligned} \mathbb{P}(\text{reaching j}|\text{starting in i}) &= \mathbb{P}\left(\bigcup_{n=0}^{\infty} \{X_n = j\} | X_0 = i\right) \\ &\leq \sum_{n=0}^{\infty} \mathbb{P}(X_n = j | X_0 = i) \\ &= \sum_{n=0}^{\infty} P_{ij}^{(n)} \\ &= 0. \end{aligned}$$

If i is accessible from j and j is accessible from i we say that the states *communicate*. We use the notation $i \leftrightarrow j$. By definition, we say that any state communicates with itself. It is easy to see that if $i \leftrightarrow j$ then $j \leftrightarrow i$ since the relationship is symmetric.

The key property however, is that if $i \leftrightarrow j$ and $j \leftrightarrow k$ then $i \leftrightarrow k$. To see this notice that $i \leftrightarrow j$ means there is some $n \geq 0$ such that $P_{ij}^{(n)} > 0$, and similarly there is some $m \geq 0$ such that $P_{jk}^{(m)} > 0$. Now, we know that,

$$P_{ik}^{(m+n)} \geq P_{ij}^{(n)} P_{jk}^{(m)} > 0.$$

This means that k is accessible from i and similarly we can argue that i is accessible from k. These properties characterize what are known as *equivalence relations*, and equivalence relations induce a partition into classes. What this means is just that we can group states into classes where in each class all the states communicate with each other and between the classes they do not.

Markov chains where all the states communicate with each other, i.e. chains with only one class are known as *irreducible* Markov chains.

Example 4.4. Suppose we have a Markov chain with states $S = \{1, 2, 3\}$ with transition matrix:

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

What are the communicating classes?

We can verify that all states communicate, so the Markov chain is irreducible, and the only communicating class is $\{1, 2, 3\}$.

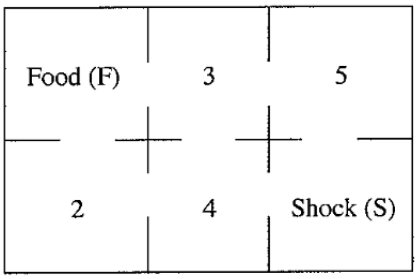
Example 4.5. Suppose we have a Markov chain with states $S = \{1, 2, 3, 4\}$ with transition matrix:

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

What are the communicating classes?

Again we can verify that there are three communicating classes, $\{1, 2\}, \{3\}, \{4\}$. Notice in particular that state 4 is absorbing (i.e. once you enter that state you are stuck there forever) so no other state is accessible from 4. Absorbing states are always their own class.

Example 4.6. Consider the rat maze:



The rat when it reaches either the food or shock states is stuck there (the experiment essentially ends), but otherwise chooses one of the remaining doors at random to go through. What are the communicating classes?

As we discussed above the absorbing states each form their own class, the remaining states all communicate so we have 3 classes: $\{F\}, \{S\}, \{2, 3, 4, 5\}$.

Example 4.7. What are the communicating classes for the random walk with barrier? What about for the Gambler's ruin problem?

4.5.1 Recurrent and Transient States

The next thing we will define are what are called recurrent and transient states. First define,

$$f_i = \mathbb{P}(X_n = i \text{ for some } n > 0 | X_0 = i),$$

which is the *return probability*, i.e. the probability that if you start in some state you eventually return to it. A state i is *recurrent* if $f_i = 1$, and *transient* if $f_i < 1$.

Let us now define, the expected number of visits to a state i, starting at i. Roughly, we expect that if a state is recurrent then we should keep visiting the state so the expected number of visits should be infinite but otherwise if the state is transient we might expect that the expected number of visits is finite, i.e. we should eventually stop returning to the state.

To understand this more formally we define the indicator RVs:

$$\mathbb{I}_n = \begin{cases} 1, \text{if } X_n = i \\ 0 \text{ otherwise.} \end{cases}$$

So our interest is then in:

$$\begin{aligned} \mathbb{E}[\text{number of visits to i} | X_0 = i] &= \mathbb{E}\left[\sum_{n=0}^{\infty} \mathbb{I}_n | X_0 = i\right] \\ &= \sum_{n=0}^{\infty} \mathbb{E}\left[\mathbb{I}_n | X_0 = i\right] \\ &= \sum_{n=0}^{\infty} \mathbb{P}[X_n = i | X_0 = i] \\ &= \sum_{n=0}^{\infty} P_{ii}^{(n)}. \end{aligned}$$

Our next goal is then to show the following characterization of transient and recurrent states:

1. A state i is transient if and only if $\sum_{n=0}^{\infty} P_{ii}^{(n)} < \infty$,

2. A state i is recurrent if and only if $\sum_{n=0}^{\infty} P_{ii}^{(n)} = \infty$.

We will discuss only one direction of these claims (the other direction can be argued similarly). Let us start with the second statement. If a state is recurrent then we know that the return probability is 1, so the chain will certainly return to the state i. The key point is that *whenever the chain returns to state i because of the Markov property, the entire process simply restarts.* This means that we can simply repeat the above argument, to conclude that the Markov chain must return to i *infinitely often.*

Now, the more interesting question is what happens if a state is transient. There is some probability f_i that the Markov chain returns to the state i and the process restarts, and a probability $1 - f_i$ that it never returns, and the game ends.

This is exactly like a geometric random variable. Remember how we defined geometric random variables, we imagined tossing a coin repeatedly until we saw a heads, and the geometric RV simply counted the number of tosses. The expected number of tosses (i.e. the mean of a geometric RV) is just $1/p$.

Returning to the Markov chain at hand, the number of visits to state i, has a geometric distribution with parameter $1 - f_i$, i.e. we stop when we don't return to i which happens with probability $1 - f_i$. So the expected number of visits to state i when we start in state i is just,

$$\sum_{n=0}^{\infty} P_{ii}^{(n)} = \frac{1}{1 - f_i} < \infty,$$

where the last inequality follows since $f_i < 1$ (the state is transient).

Fact 4.1. In a Markov chain with finite number of states (we refer to these as finite Markov chains) all states cannot be transient.

This fact is important to remember. It is intuitively easy to explain, if every state is transient then every state must be visited a finite number of times in the long run so eventually the Markov chain will run out of states to visit which cannot happen.

An even more important fact is that recurrence and transience are examples of what are called *class properties*, i.e. if two states communicate they must be either both transient or both recurrent. Every state in a class must either be transient or every state must be recurrent.

Fact 4.2. Recurrence is a class property, i.e. if state i is recurrent, and $i \leftrightarrow j$ then state j is recurrent.

Lets try to prove this a bit more carefully, since $i \leftrightarrow j$ we know that there are integers u and v such that, $P_{ij}^{(u)} > 0$ and $P_{ji}^{(v)} > 0$. For any integer n we have that,

$$P_{jj}^{(u+n+v)} \geq P_{ji}^{(v)} P_{ii}^{(n)} P_{ij}^{(u)},$$

since the right hand side represents the probability of one possible path and the left hand side is the probability of the sum over all possible paths from j to j in $(u+n+v)$ steps. Summing both sides over n we have that,

$$\sum_{n=0}^{\infty} P_{jj}^{(u+n+v)} \geq P_{ji}^{(v)} P_{ij}^{(u)} \sum_{n=0}^{\infty} P_{ii}^{(n)} = \infty,$$

so we conclude that j is also recurrent. This fact also leads us to two other facts.

Fact 4.3. Transience is a class property, i.e. if state i is transient, and $i \leftrightarrow j$ then state j is transient.

This follows just because from the previous fact because a recurrent state cannot communicate with a transient one.

Fact 4.4. In a finite Markov chain, if the Markov chain is irreducible, then every state must be recurrent.

Since the chain is irreducible, we know that either every state is transient or every state is recurrent. We also have shown that in a finite chain every state cannot be transient, so we conclude that every state must be recurrent.

Lets do a couple of warmup examples before trying a really hard one.

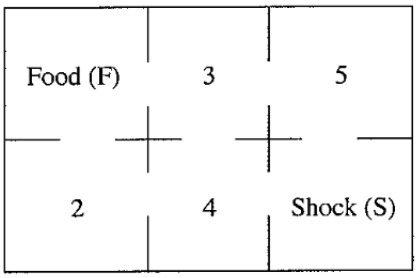
Example 4.8. Suppose we have a Markov chain with states $S = \{1, 2, 3, 4\}$ with transition matrix:

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

Which states are transient and which are recurrent?

We can check that all states communicate and since the chain is finite they must all be recurrent.

Example 4.9. Consider the rat maze:



Which states are transient and which are recurrent?

The states $\{2, 3, 4, 5\}$ are transient while the others are recurrent. Notice that absorbing states are always recurrent.

It is often not too difficult to figure out which states are transient and recurrent in a finite Markov chain without resorting to any real calculations. It is however, considerably more challenging to do this for infinite Markov chains. To give you a flavor of this let us study an example.

Example 4.10. Classify the states of a simple random walk.

Recall, in a simple random walk we start in state 0, and move up with probability p and down with probability $1-p$. We know that all the states communicate, so the chain is irreducible, and either all the states are transient or they are all recurrent.

To figure out which let us attempt to calculate $\sum_{n=1}^{\infty} P_{00}^{(n)}$. We know that if this quantity is infinite then the chain must be recurrent otherwise it is transient.

Now, to compute the sum, we note that it must take an even number of steps to return to state 0, so alternate terms of the sum are 0, and we have:

$$\sum_{n=1}^{\infty} P_{00}^{(n)} = \sum_{n=1}^{\infty} P_{00}^{(2n)}.$$

The latter quantity is computable by noting that to return to state 0 in $2n$ steps we must go up for n steps and return back for n steps, i.e.

$$P_{00}^{(2n)} = \binom{2n}{n} p^n (1-p)^n.$$

Now we will make a quick detour to introduce Stirling's formula. If you have not seen this before, it is a useful tight bound on $n!$. In particular, we have the relationship, for $n \geq 1$,

$$\sqrt{2\pi}\sqrt{n}\left(\frac{n}{e}\right)^n \leq n! \leq e\sqrt{n}\left(\frac{n}{e}\right)^n.$$

So upto constants we can see that,

$$n! \sim \sqrt{n}\left(\frac{n}{e}\right)^n,$$

where we use $\sim$ to indicate that we are ignoring constants. This in turn means that,

$$\begin{aligned}\binom{2n}{n} = \frac{(2n)!}{n! \times n!} &\sim \frac{\sqrt{2n}\left(\frac{2n}{e}\right)^{2n}}{\sqrt{n}\left(\frac{n}{e}\right)^n \times \sqrt{n}\left(\frac{n}{e}\right)^n} \\ &= \sqrt{\frac{2}{n}} 4^n \sim \frac{4^n}{\sqrt{n}}.\end{aligned}$$

Returning to our original calculation, we obtain,

$$\sum_{n=1}^{\infty} P_{00}^{(2n)} \sim \sum_{n=1}^{\infty} \frac{(4p(1-p))^n}{\sqrt{n}}.$$

Now there are two cases:

1. $4p(1-p) = 1$**, i.e.** $p = 1/2$**:** In this case, the sum is

$$\sum_{n=1}^{\infty} P_{00}^{(2n)} \sim \sum_{n=1}^{\infty} \frac{1}{\sqrt{n}} = \infty,$$

so the chain is recurrent.

2. $p \neq 1/2$**:** In this case, denote $r = 4p(1-p)$,

$$\sum_{n=1}^{\infty} P_{00}^{(2n)} \leq \sum_{n=1}^{\infty} r^n = \frac{1}{1-r} < \infty,$$

so the chain is transient.

To summarize if $p = 1/2$ the random walk is recurrent and otherwise it is transient. This calculation also shows some of the difficulties in dealing with infinite Markov chains.

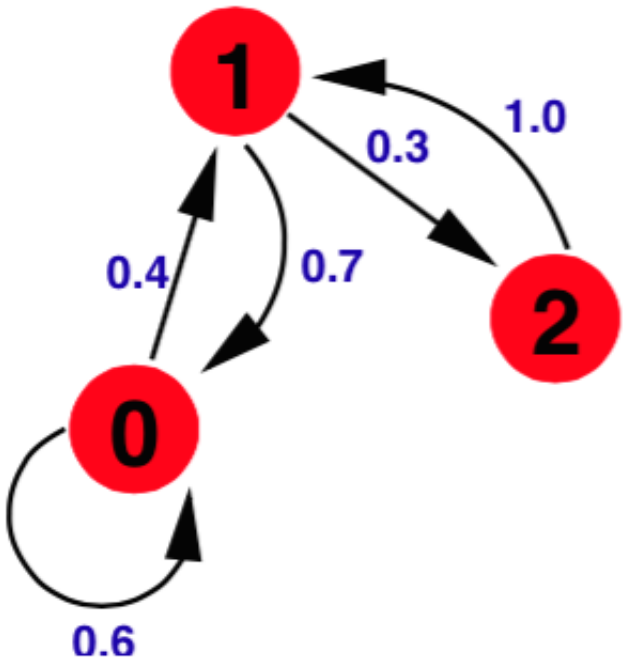
Fact 4.5. Argue that in a Markov chain, every recurrent class is absorbing, i.e. once you enter a recurrent class you can never leave.

Suppose we have a recurrent class A, and some state $i \in A$. Now, for any $j \notin A$, we claim that j is not accessible from i. If it is, then it must be possible to return (otherwise i cannot be recurrent), but if it is possible to return then it must be the case that $i \leftrightarrow j$ so that they both belong to the same class.

4.6 Long Run Behaviour of Markov Chains – Computational Methods

To begin with lets just try to describe certain phenomena, together with some computational methods that will be useful, before trying to analytically characterize the long term behavior.

Recall this Markov Chain we saw previously,



and its transition matrix:

$$\mathbf{P} = \begin{bmatrix} 0.6 & 0.4 & 0 \\ 0.7 & 0 & 0.3 \\ 0 & 1 & 0 \end{bmatrix}$$

Suppose the chain starts in state 1, what is the probability that after n steps it is back in state 1? What happens as n gets big?

Some MATLAB code to help us out:

```
» P = [0.6 0.4 0; 0.7 0 0.3; 0 1 0]
ans =
       0.6000   0.4000        0
       0.7000        0   0.3000
            0   1.0000        0
» P^2
ans =
       0.6400   0.2400   0.1200
       0.4200   0.5800        0
       0.7000        0   0.3000
» P^5
ans =
       0.5616   0.3506   0.0878
       0.6135   0.2554   0.1312
       0.5124   0.4372   0.0504
» P^100
ans =
       0.5738   0.3279   0.0984
       0.5738   0.3279   0.0984
       0.5738   0.3279   0.0984
» P^1000
ans =
       0.5738   0.3279   0.0984
       0.5738   0.3279   0.0984
       0.5738   0.3279   0.0984
```

It appears as though the n-step probabilities are converging, and really we would like to understand when/why this happens. We define $P_{ij}^{\infty} = \lim_{n\to\infty} P_{ij}^{n}$, if the limit exists. If it exists for all pairs (i, j) then we can put these into the matrix $P^{(\infty)}$.

Another key point: Notice that the n-step probabilities are not just converging, they are also *invariant to the initial state* X_0, i.e. it does not matter which state we start in we seem to end up in the same distribution over states.

Example 4.11. Construct a simple transition matrix P for which $P^{(\infty)}$ does not exist.

Suppose we consider the Markov chain with:

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

Then observe that $P^{(2)}$ is given by:

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

and more generally,

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

so the entries are not converging in any sense (they are instead *oscillating*). In this case, the limit $P^{(\infty)}$ does not exist. This is an example of a *periodic* Markov chain.

For now, we will assume that in cases when the limit exists that we can compute it, i.e. for instance we can use MATLAB to approximately obtain the entries of $P^{(\infty)}$. Our next goal will be to try to understand this matrix a bit better, and see how we can use it.

4.6.1 Long run probabilities for absorbing states

Recall that a state j of a Markov chain is called **absorbing** if $P_{jj} = 1$, i.e. if the chain reaches state j, it is stuck there forever.

Example 4.12. Suppose that a Markov chain is such that state j is absorbing. What can you say about P_{ij}^{∞}?

Observe that:

$$\begin{aligned} P_{ij}^{k} &= P(X_k = j | X_0 = i) \\ &= P(X_l = j \text{ for any } l \leq k | X_0 = i). \end{aligned}$$

So that, P_{ij}^{∞} is just the probability of ever reaching the state j, when starting in the state i. So we have,

$$\begin{aligned} P_{ij}^{\infty} &= P(\text{ever reach state j} | X_0 = i) \\ 1 - P_{ij}^{\infty} &= P(\text{never reach state j} | X_0 = i). \end{aligned}$$

Example 4.13 (Random walk on the integers)**.** Recall our usual random walk on integers. What is the probability that you reach "3" before you reach "-2"?

This example illustrates a very useful technique in Markov chain analysis: often we can "mutilate" a Markov chain so as to not change the answer to a particular question while making the desired calculation much easier.

In this case, suppose we made the states 3 and -2 both absorbing? Then observe that P(reach "3" before you reach "-2"$|X_0 = 0$) is identical for both chains.

Furthermore, for the modified chain:

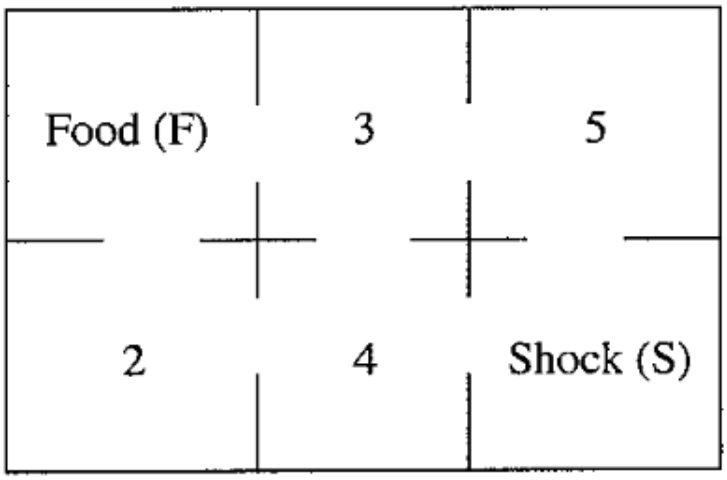
$$P(\text{reach "3" before you reach "-2"} | X_0 = 0) = P(\text{ever reach "3"} | X_0 = 0) = P_{03}^{(\infty)},$$

which we can compute numerically using MATLAB. So we write down the modified transition matrix:

$$P = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0.5 & 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0.5 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0.5 & 0 & 0.5 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

and find that $P_{03}^{(\infty)} \approx P_{03}^{(1000)} = 0.4$.

Example 4.14 (The Rat Maze Example)**.** Suppose a rat is placed into the maze below:



The experiment stops once the rat reaches either the food (F) or the shock (S). If the rat is in one of the other four rooms, it chooses one of the doors at random with equal probability.

What is the limiting transition matrix $P^{(\infty)}$? If the rat starts in room "2" what is the probability it reaches the Food before it is Shocked?

Again, via MATLAB:

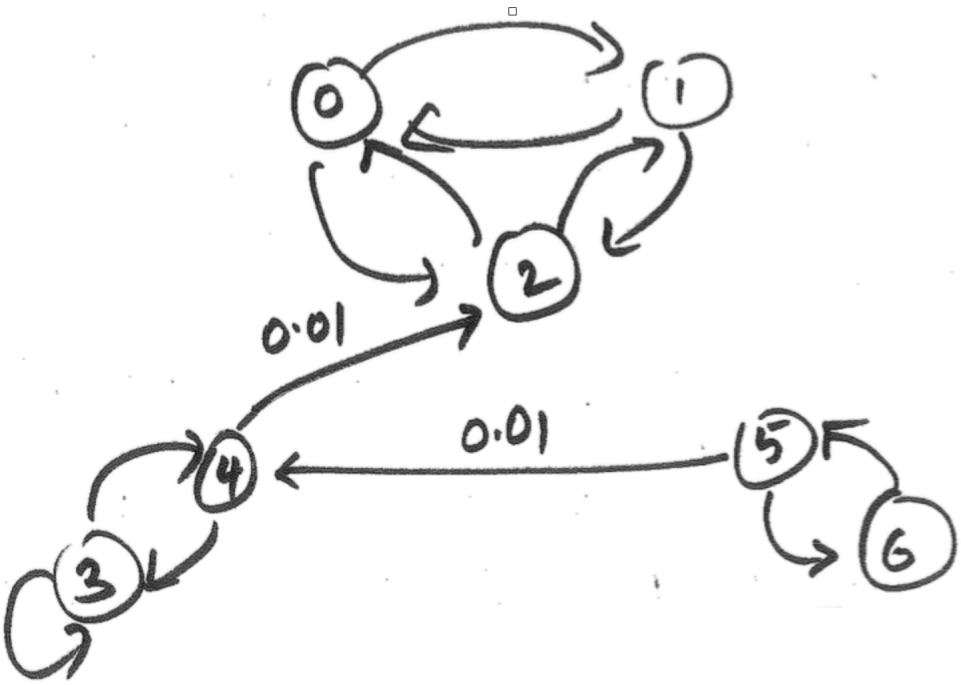
```
» P^1000
ans =

    1.0000    0    0    0    0         0
    0.7143    0    0    0    0    0.2857
    0.5714    0    0    0    0    0.4286
    0.4286    0    0    0    0    0.5714
    0.2857    0    0    0    0    0.7143
         0    0    0    0    0    1.0000
```

So that the desired probability is $P_{21}^{(\infty)} \approx 0.7143$.

Putting together some of the insights from the previous section and this one, we can try to glean some structural information about $P^{(\infty)}$.

Example 4.15. Construct a Markov chain with two transient classes, one recurrent class, and reason about the structure of $P^{(\infty)}$.

Consider the following Markov chain:



Observe that since there is only one recurrent class, in the long run we should spend all of our time there. We can also use a symmetry argument (supposing that all the unmarked transitions are equally likely) to conclude that $P^{(\infty)}$ should have the structure:

$$P^{(\infty)} = \begin{bmatrix} 0.33 & 0.33 & 0.33 & 0 & 0 & 0 & 0 \\ 0.33 & 0.33 & 0.33 & 0 & 0 & 0 & 0 \\ 0.33 & 0.33 & 0.33 & 0 & 0 & 0 & 0 \\ 0.33 & 0.33 & 0.33 & 0 & 0 & 0 & 0 \\ 0.33 & 0.33 & 0.33 & 0 & 0 & 0 & 0 \\ 0.33 & 0.33 & 0.33 & 0 & 0 & 0 & 0 \\ 0.33 & 0.33 & 0.33 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

In particular, the long-run probability of transient states will be 0. We will investigate all of this analytically in the sequel.

4.6.2 Numerically computing return probabilities

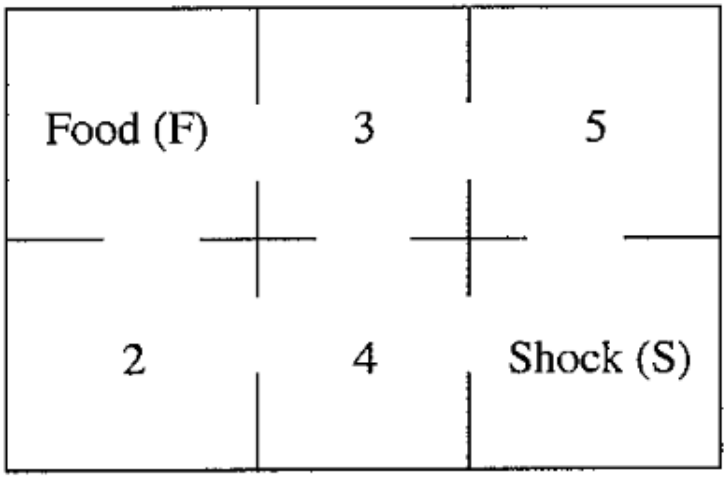
It is also worth noting that we can numerically compute return probabilities. Recall, that for a recurrent state the return probability is 1, and that for transient states we have that,

$$\sum_{k=0}^{\infty} P_{ii}^{(k)} = \frac{1}{1 - f_i}.$$

Also, the expected number of visits to any state j, starting in state i, is given as:

$$\mathbb{E}[\text{number of visits to state j}|X_0 = i] = \sum_{k=0}^{\infty} P_{ij}^{(k)},$$

Example 4.16. Re-visit the rat-maze example.



The rat starts in room 2. What is the expected number of visits to room "3"? What is the probability that it never returns to room "2"? Use MATLAB.

Here is the MATLAB code:

```
» P = [  1    0    0    0    0    0;
        0.5   0    0   0.5   0    0;
        1/3   0    0   1/3  1/3   0;
         0   1/3  1/3   0    0   1/3;
         0    0   1/2   0    0   1/2;
         0    0    0    0    0    1  ]
» M = eye(6);
» for i = 1:1000
M = M + P^i;
end
» M(2:5, 2:5)
ans =
 1.2381  0.2857  0.7143  0.0952
 0.1905  1.4286  0.5714  0.4762
 0.4762  0.5714  1.4286  0.1905
 0.0952  0.7143  0.2857  1.2381
```

Why did I not print out the other numbers? The answer is just that the expected number of visits to the recurrent states is blowing up.
With MATLAB's help we can now answer the questions:

- First we notice that as we derived earlier the expected number of visits to a particular

state can be written as:

$$\mathbb{E}[\text{number of visits to state j}|X_0 = i] = \sum_{k=0}^{\infty} P_{ij}^{(k)},$$

so we have that the expected number of visits to state 3 starting from 2 is 0.2857.

- Similarly, the probability of never returning to state 2 starting in state 2 is $1 - f_2$, and using the previous expression we obtain:

$$1 - f_2 = \frac{1}{\sum_{k=0}^{\infty} P_{22}^{(k)}} = \frac{1}{1.2381}.$$

4.7 Long Run Behaviour of Markov Chains – Analytical Methods

Our goal now will be to try to characterize some of the things we saw in the previous section, using direct analytical methods. Recall that if the initial distribution in a Markov chain is π_0, then we can obtain the distribution after n-steps as:

$$\pi_n^T = \pi_0^T P^{(n)}.$$

We need to define and contrast two things:

- **A Stationary Distribution:** A stationary distribution is one that is invariant, i.e. if we start in a stationary distribution, the distribution after n-steps is the same as the distribution in which we started. Formally, a stationary distribution is any distribution (i.e. π has entries between 0 and 1, and the sum of its entries is 1) such that,

$$\pi^T = \pi^T P.$$

- **The Limiting Distribution:** The limiting distribution, if it exists, is simply the long-term distribution over the different states, i.e.

$$\pi_{\lim}^T = \lim_{n\to\infty} \pi_0^T P^{(n)}.$$

If this limit is not unique (for instance it depends on the initial distribution π_0) then we say that the limiting distribution does not exist.

Fact 4.6. A stationary distribution of a Markov chain need not exist and need not be unique, i.e. a Markov chain can have 0, 1 or infinitely many stationary distributions. On the other hand, if there is a limiting distribution, it is by definition unique.

To substantiate this fact, let us construct Markov chains with different numbers of stationary distributions:

- **No stationary distributions:** It turns out that every *finite* Markov chain has at least one stationary distribution, so our example is necessarily an infinite chain. Consider the Markov chain on the integers, that moves mass to the right (deterministically), i.e.

$$P_{i,i+1} = 1, \quad \text{for all } i.$$

You can convince yourself that no distribution can be stationary here (we'll do so a bit more formally later on) but roughly the mass that π_0 puts on any state moves to its neighbor we need that every state must have the same mass initially. However, there is no distribution on the integers that puts equal mass at every integer (i.e. there is no uniform distribution on an infinite set).

- **One stationary distribution:** This is the typical case. We will see many examples of this case in the future.
- **Multiple stationary distributions:** Consider for example the rat-maze. It is easy to see that both the distributions $\pi^T = [1\ 0\ 0\ 0\ 0\ 0]$ and $\pi^T = [0\ 0\ 0\ 0\ 0\ 1]$ are stationary.

  In general, in a finite Markov chain, you will be able to associate every recurrent class with a stationary distribution.

At a high-level our interest in the rest of this section will be to understand the limiting distribution, when it exists and how to compute it. To compute it, we will try to reason about when the limiting distribution is equal to the (hopefully unique) stationary distribution and then just compute the stationary distribution. The stationary distribution is easy to compute because it just involves solving a linear system.

4.7.1 Computing a stationary distribution

Lets temporarily ignore the limiting distribution, and even ignore the question of when the stationary distribution is unique and just try to compute the stationary distribution (assuming it is unique). Let us consider some examples.

Example 4.17. Suppose we have a Markov chain with transition matrix:

$$P = \begin{bmatrix} \alpha & 1-\alpha \\ \beta & 1-\beta \end{bmatrix},$$

where $0 < \alpha, \beta < 1$. Compute its stationary distribution.

We know that the stationary distribution must satisfy $\pi^T = \pi^T P$, so let us take the entries of $\pi = [\pi_1 \;\; \pi_2]$, and we obtain the equations:

$$\begin{aligned} \pi_1 &= \pi_1\alpha + \pi_2\beta \\ \pi_2 &= \pi_1(1-\alpha) + \pi_2(1-\beta). \end{aligned}$$

Notice that these two equations are redundant so we cannot solve them to obtain π. However, we do have another crucial piece of information, π has to be a distribution, so we know that:

$$\pi_1 + \pi_2 = 1,$$

and using this we obtain from the first equation above that,

$$\pi_1 = \pi_1\alpha + (1-\pi_1)\beta,$$

i.e. that $\pi_1 = \frac{\beta}{1-\alpha+\beta}$, and that $\pi_2 = \frac{1-\alpha}{1-\alpha+\beta}$.

This method of solving for the stationary distribution also works for infinite Markov chains.

Example 4.18. Let us consider the random walk with barrier, i.e. if you are in state 0, you remain in that state with probability $1-p$ and go to state 1 with probability p. For every other positive integer: $X_n = X_{n-1} - 1$ with probability $1-p$, and $X_n = X_{n-1} + 1$ with probability p. Compute the stationary distribution when $p < 1/2$.

Let us first examine the condition $p < 1/2$: just like we argued that for the random walk when $p \neq 1/2$ the chain is transient one can argue that for the RW with barrier the chain is transient if $p > 1/2$. When the chain is transient there is no stationary distribution. When $p = 1/2$ the situation is more complicated: the chain is what is called *null recurrent*. What this means is that even though the chain is recurrent, it can take infinitely long for the chain to return

to any fixed state: in this case also there is no stationary distribution. To summarize, when $p \geq 1/2$ the chain does not have a stationary distribution. You do not need to understand this too deeply.

Let us now focus on the case when $p < 1/2$ and find its stationary distribution. We just follow the usual recipe, i.e. solve $\pi^T = \pi^T P$ (this is now an infinite system of linear equations). Notice that P is given by:

$$P = \begin{bmatrix} 1-p & p & 0 & 0 & 0 & \dots \\ 1-p & 0 & p & 0 & 0 & \dots \\ 0 & 1-p & 0 & p & 0 & \dots \\ 0 & 0 & 1-p & 0 & p & \dots \\ & & \vdots & & & \end{bmatrix}.$$

So we can write out the system of equations as:

$$\begin{aligned} &\pi_0 = (1-p)[\pi_0 + \pi_1] \implies \pi_1 = \frac{p}{1-p}\pi_0 \\ &\pi_1 = \pi_0 p + (1-p)\pi_2 \implies (1-p)\pi_2 = \frac{p}{1-p}\pi_0 - p\pi_0 \implies \pi_2 = \left(\frac{p}{1-p}\right)^2 \pi_0 \\ &\pi_2 = \pi_1 p + (1-p)\pi_3 \implies \pi_2 = \left(\frac{p}{1-p}\right)^3 \pi_0, \\ &\vdots \end{aligned}$$

Notice that we have re-written all of the probabilities in terms of π_0. In order to solve this system we once again need to use the fact that the probabilities sum to 1 to obtain that,

$$\pi_0 \left(1 + \frac{p}{1-p} + \left(\frac{p}{1-p}\right)^2 + \left(\frac{p}{1-p}\right)^3 + \dots\right) = 1,$$

and this yields that,

$$\pi_0 = \frac{1-2p}{1-p}, \quad \pi_1 = \frac{p(1-2p)}{(1-p)^2}, \quad \pi_2 = \frac{p^2(1-2p)}{(1-p)^3}, \dots$$

4.7.2 Interpreting the limiting distribution

There are several ways to interpret the limiting distribution when it exists. There are three things to know that may build intuition:

1. Limiting distributions are always stationary.
2. Limiting distributions capture the long-run proportion of time spent in each state.
3. The limiting probability of any state is inversely proportional to the expected time between visits.

Let us investigate these properties more carefully.

Fact 4.7. The limiting distribution of a Markov chain is always stationary.

Let us denote $\pi_m = [P(X_m = 0) \quad P(X_m = 1) \quad \ldots]$, as the distribution over states after m-steps. We know that it is always the case that for any $m \geq 1$,

$$\pi_m^T = \pi_0^T P^{(m)} = \pi_{m-1}^T P,$$

i.e. to obtain the distribution after m-steps we just multiply the distribution after $m-1$ steps by the transition matrix. If the limiting distribution exists then we can just take the limit as $m \to \infty$ on both sides of this expression to obtain that:

$$\lim_{m\to\infty} \pi_m^T = \lim_{m\to\infty} \pi_{m-1}^T P,$$

and this yields that,

$$\pi_{\lim}^T = \pi_{\lim}^T P.$$

This tells us that the limiting distribution is also a stationary distribution.

Fact 4.8. The entries of the limiting distribution represent the long-run proportion of time spent in each state.

To understand this statement let us recall the usual law of large numbers. One simplified version is that if I observe $X_1, \ldots, X_n$ which are i.i.d and such that $P(X_i = 1) = p, P(X_i = 0) = 1 - p$, then the average "converges" to the expected value, i.e.:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i \to p.$$

One can do the same thing with a Markov chain, suppose we denote the sequence of states $X_1, \ldots, X_n$. Let us fix some state u and suppose that we define an indicator sequence:

$$Y_i = \begin{cases} 1 & \text{if } X_i = u \\ 0 & \text{if } X_i \neq u. \end{cases}$$

Then the average:

$$\frac{1}{n} \sum_{i=1}^{n} Y_i,$$

just represents the average amount of time (i.e. the fraction of time) we spend in the state u. It turns out that there is a similar convergence that happens, i.e. for a Markov chain if the limiting distribution exists:

$$\frac{1}{n} \sum_{i=1}^{n} Y_i \to \pi_{\lim}(u).$$

So the entries of the limiting distribution just tell us what fraction of time we spend in each state.

Example 4.19. Suppose I model my mood each day as a first-order Markov chain $\{X_n : n \geq 0\}$ with states "energetic" (E), "so-so" (S), and "tired" (T) with transition matrix

$$\mathbf{P} = \begin{bmatrix} 0.1 & 0.6 & 0.3 \\ 0.1 & 0.2 & 0.7 \\ 0.4 & 0.3 & 0.3 \end{bmatrix}.$$

For now, take it as given that there is a limiting distribution, and using MATLAB we find that it is

$$\pi_{\text{lim}} = [0.23 \quad 0.34 \quad 0.43].$$

I'm not particularly worried about my mood but I am worried about my coffee consumption. Suppose, that

$$Y_n = \text{number of cups of coffee I drink on day } n.$$

Further, suppose that when I am energetic I drink one cup of coffee, when I am so-so I drink two cups, and if I am tired I drink three cups. In the long-run how many cups of coffee should I anticipate buying per day?

The random variable Y_n is defined as:

$$Y_n = \begin{cases} 1 & \text{if } X_n = 0 \\ 2 & \text{if } X_n = 1 \\ 3 & \text{if } X_n = 2. \end{cases}$$

So we can compute:

$$\begin{aligned} \mathbb{E}[Y_n] &= 1 \times P(X_n = 0) + 2 \times P(X_n = 1) + 3 \times P(X_n = 2) \\ &= 0.23 + 2 \times 0.34 + 3 \times 0.43 \\ &= 2.2. \end{aligned}$$

Example 4.20. More realistically, Y_n is a random variable whose distribution depends on the state I am in on day n, as given in the following table:

State (X_n)	$P(Y_n = 1)$	$P(Y_n = 2)$	$P(Y_n = 3)$
energetic	0.4	0.5	0.1
so-so	0.1	0.6	0.3
tired	0	0.5	0.5

Now, how many cups of coffee should I anticipate buying per day?

This is a basic example of something called a *Hidden Markov Model*, where rather than observe the state X_n directly we observe a "noisy" version of it (in this case the coffee intake).

To compute $\mathbb{E}[Y_n]$ we can use the law of total expectation:

$$\mathbb{E}[Y_n] = \mathbb{E}[Y_n|X_n = E]P(X_n = E) + \mathbb{E}[Y_n|X_n = S]P(X_n = S) + \mathbb{E}[Y_n|X_n = T]P(X_n = T).$$

To compute for instance,

$$\mathbb{E}[Y_n|X_n = E] = 0.4 \times 1 + 0.5 \times 2 + 0.1 \times 3 = 1.7,$$

and similarly,

$$\mathbb{E}[Y_n|X_n = S] = 2.2,$$
$$\mathbb{E}[Y_n|X_n = T] = 2.5,$$

and putting these together we obtain $\mathbb{E}[Y_n] = 2.21$.
Now, suppose we define the *mean first-passage times*: for a pair of states i and j we define,

$$n_{ij} = \mathbb{E}[\#\text{ steps to visit state j}|X_0 = i].$$

The value n_{ii} is the expected time to return to a particular state i, it is sometimes called the mean first-return time.

Fact 4.9. For any finite Markov chain with a (unique) limiting distribution:

$$\pi_{\lim}(i) = \frac{1}{n_{ii}}.$$

Again this fact should intuitively make sense, if a state i has a small expected return time (i.e. on average we return to the state rapidly) then we might imagine that in the limit we are going to spend a good fraction of our time in this state. It is also giving us a way of thinking about the limiting probabilities.

Let us try to prove this fact. The idea is one that we have used many times before, let us try to condition on the first step. So we have:

$$\mathbb{E}[\#\text{ steps to visit state j}|X_0 = i] = \sum_k \mathbb{E}[\#\text{ steps to visit state j}|X_0 = i, X_1 = k]\mathbb{P}(X_1 = k|X_0 = i).$$

Now observe that:

$$\mathbb{E}[\#\text{ steps to visit state j}|X_0 = i, X_1 = k] = \begin{cases} 1 & \text{if } k = j \\ 1 + n_{kj} & \text{if } k \neq j \end{cases}.$$

So we can write the first passage time as:

$$\begin{aligned}\mathbb{E}[\#\text{ steps to visit state j}|X_0 = i] &= \sum_k \mathbb{E}[\#\text{ steps to visit state j}|X_0 = i, X_1 = k]\mathbb{P}(X_1 = k|X_0 = i) \\ &= \sum_k (1 + n_{kj})P_{ik} - n_{jj}P_{ij} \\ &= 1 + \sum_k n_{kj}P_{ik} - n_{jj}P_{ij}.\end{aligned}$$

Now, we can collect these expressions into a matrix set of equations. First we define E to just be a matrix of all 1s (of the same size as the transition matrix), and N_d to be the diagonal matrix (assuming there are s states in the Markov chain):

$$N_d = \begin{bmatrix} n_{11} & 0 & 0 & 0 & \dots & 0 \\ 0 & n_{22} & 0 & 0 & \dots & 0 \\ & & \vdots & & & \\ 0 & 0 & 0 & 0 & \dots & n_{ss} \end{bmatrix}.$$

$$N = E + P \times N - P \times N_d.$$

Now suppose we multiply on the left by π_{lim}^T. Since the sum of entries in the limiting distribution is 1, we have that,

$$\pi_{\text{lim}}^T E = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}.$$

Since the limiting distribution is stationary we have $\pi_{\text{lim}}^T = \pi_{\text{lim}}^T P$, so we obtain:

$$\pi_{\text{lim}}^T N = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + \pi_{\text{lim}}^T N - \pi_{\text{lim}}^T N_d.$$

This says that:

$$\pi_{\text{lim}}^T N_d = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix},$$

and this is simply the fact we wanted to prove, i.e. that for any state j, $\pi_{\text{lim}}(j) n_{jj} = 1$.

4.8 Computing the limiting distribution

So far we have seen that the limiting distribution is a useful and natural object of interest for a Markov chain. We have also seen that the *stationary distribution* is easy to compute (we just solve a system of equations). Now we would like to focus on computing the limiting distribution. The way we'll do it is via what is called the *basic limit theorem*.

4.8.1 The basic limit theorem

I'll state the theorem here and we'll need to figure out what it means in the rest of this section.

Fact 4.10 (The basic limit theorem)**.** Let $X_0, X_1, \ldots$ be an irreducible, aperiodic Markov chain having a stationary distribution π. Let X_0 have the distribution $\pi^{(0)}$, an arbitrary initial distribution. Then $\pi_{\lim}(i) = \pi(i)$ for all states i.

So the basic limit theorem tells us when the limiting distribution and stationary distribution are the same, and hence the limiting distribution is easy to compute. Also, notice that when an irreducible, aperiodic Markov chain has at least one stationary distribution that stationary distribution is unique (because the limiting distribution is always unique).

We need to understand two things: (1) what is an *aperiodic* Markov chain? (2) when does a Markov chain have at least one stationary distribution π? Lets start of with the answer to the second question.

4.8.2 Positive Recurrence and Unique Stationary Distributions

Now that we know what a stationary distribution is, we can ask the question: when is there at least one stationary distribution?

Fact 4.11.
1. Any finite Markov chain always has at least one stationary distribution.
2. An irreducible Markov chain has a stationary distribution if and only if it is *positive recurrent*.

Formally, positive recurrent means that the average time to return to any state if you start from there must be finite.

I do not want to belabor the second point because it is a bit technical. Let me just add a bit of caution: *for infinite Markov chains things can get tricky*. You do not need to memorize/learn the following. Let me also tell you some canonical examples of infinite Markov chains which do and do not have stationary distributions:

1. **Simple Random Walk:**
   - When $p \neq 1/2$ we have seen that all the states are transient. Transient states cannot be part of the stationary distribution so there are no stationary distributions.
   - When $p = 1/2$ the chain is recurrent. However, it turns out that the chain is *not positive recurrent*, i.e. the expected time to return to any state is ∞, so again no stationary distributions.
2. **RW with barrier:**
   - When $p < 1/2$ we computed the stationary distribution and it is unique.
   - When $p \geq 1/2$ it does not have a stationary distribution: again when $p = 1/2$ it is *not positive recurrent*, and when $p > 1/2$ it is transient.

4.8.3 Aperiodic Markov Chains

The last piece of the puzzle is that we need our Markov chain to be aperiodic. Suppose we have a Markov chain with a transition matrix:

$$\mathbf{P} = \begin{bmatrix} 0 & 0.7 & 0 & 0.3 \\ 0.8 & 0 & 0.2 & 0 \\ 0 & 0.15 & 0 & 0.85 \\ 0.7 & 0 & 0.3 & 0 \end{bmatrix}.$$

This transition matrix defines an irreducible Markov chain, but we notice that state 0 (for instance) has the following property: If the chain starts at 0, it can only return on steps which are a multiple of two. In this case, we say that state 0 has **period** two.

Using MATLAB calculate $\mathbf{P}^{(1000)}$ and $\mathbf{P}^{(1001)}$. What do you observe?

You will observe that they have different non-zero entries, i.e.:

$$P^{(1000)} = \begin{bmatrix} \star & 0 & \star & 0 \\ 0 & \star & 0 & \star \\ \star & 0 & \star & 0 \\ 0 & \star & 0 & \star \end{bmatrix},$$

$$P^{(1001)} = \begin{bmatrix} 0 & \star & 0 & \star \\ \star & 0 & \star & 0 \\ 0 & \star & 0 & \star \\ \star & 0 & \star & 0 \end{bmatrix},$$

where all I am indicating are the locations of the non-zeros. Clearly the entries of the matrix are not converging.

Formal definition: State i is said to have period d if $P_{ii}^{(n)} = 0$ whenever n is not divisible by d, d is the largest integer with this property.

Lets just turn this into a sort of "recipe".

1. Find all n such that $P_{ii}^{(n)} > 0$.
2. The period d of the state i is the greatest common divisor of this list.

Some quick consequences:

1. $P_{ii}^{(n)} > 0$ then the period $d \leq n$.
2. If $P_{ii} > 0$, state i is aperiodic.

We define a Markov chain to be aperiodic if every state has period 1.

Example 4.21. Consider a Markov chain with transition matrix

$$\mathbf{P} = \begin{bmatrix} 0 & 1 \\ 0.7 & 0.3 \end{bmatrix}.$$

What is the period of each of the two states?

To compute the period of each state we notice that the second state has a self-loop, i.e. $P_{22} > 0$ so state 2 is aperiodic. Now, we also see that for state 1, $P_{11}^{(n)} > 0$ for $n \in \{2, 3, 4, 5, \ldots\}$, and the GCD of this list is 1 so this state also has period 1, i.e. the chain is aperiodic.

Fact 4.12. The period of a state is a class property, i.e. every state in a class has the same period.

I will prove it here, but the proof is optional (i.e. you do not need to study this unless you are curious). However, you should remember this fact – it gives a simple way to compute the period of each state for a big Markov chain: find the communicating classes and find the period of one state in each class.

Let us now prove the statement: let us take a pair of states in the same class, $i \leftrightarrow j$, and let d_i be the period of i and d_j be the period of j. We want to show that $d_i = d_j$. Instead we will show that $d_i \geq d_j$, and we can then flip the roles of i and j to conclude that $d_j \geq d_i$, i.e. that they are both equal.

First, let us find all $r > 0$, such that $P_{jj}^{(r)} > 0$ and collect them $\{r_1, r_2, \ldots\}$. Then we know that d_j is the GCD of this list. Next, lets find some path from $i \to j$ of length s and $j \to i$ of length t, so we know that d_i must be a common divisor of $\{s+t, s+t+r_1, s+t+r_2, \ldots\}$ since $P_{ii}^{(s+t+r_n)} > 0$. This in turn means that d_i must be a common divisor of $\{r_1, r_2, \ldots\}$. Since d_j is the GCD of this list we have that $d_j \geq d_i$, and reversing the roles of i and j we obtain the desired fact.

4.8.4 Wrapping up

Now we can make sense of the basic limit theorem for Markov chains: if we have an irreducible, aperiodic Markov chain, which has a stationary distribution then (a) the stationary distribution must be unique and (b) the Markov chain has a limiting distribution that is the same as this (unique) stationary distribution.

This might not seem like a very general result, but we will see in the next few sections that this really tells us everything we need to know! Once we have this result we can combine it with a few results about the transient behaviour to figure out almost everything we might need to know.

4.8.5 Computing the Stationary/Limiting Distribution – Conveniently

We have already seen that in the nice cases, all we need to do to compute the limiting distribution is to solve the system of equations $\pi^T = \pi^T P$ with the constraint that the entries of π sum to 1.

One can also write this in a slightly more elegant form. Denoting by e a row vector of 1s, E a matrix of 1s, and I the identity matrix.

Fact 4.13.

$$\pi^T = e(I + E - P)^{-1}.$$

To see this we notice that we have:

$$\pi^T = \pi^T P$$
$$\pi^T E = e.$$

Summing these we obtain,

$$\pi^T (I + E - P) = e,$$

which yields the desired fact.

4.9 Recap: What do we know so far about the long-term behavior?

Lets focus on finite, aperiodic Markov chains.

1. There are a transient classes, recurrent classes (and at least one recurrent class).

2. Transient classes are transient so in the long-term we don't spend any time there.

3. Recurrent classes are absorbing so if we start in one we are stuck in that class forever.

4. If we start in a recurrent class C, because it is finite and aperiodic (by assumption), we know by the basic limit theorem two things (since we can ignore all the other classes it is aperiodic): (1) there is a unique stationary distribution and it is equal to the limiting distribution, (2) we can compute the limiting distribution by solving the system of equations $\pi_C^T = \pi_C^T P_{CC}$.

   By this I mean we just solve for the stationary distribution of the recurrent class (ignoring all the other states).

5. If we start in a transient state we know that we will eventually be absorbed into a recurrent class, and the limiting distribution will be different based on which recurrent class we get absorbed into. This brings us to some questions – can we analytically calculate the probability of being absorbed into each recurrent class? What more can we say about the transient behaviour, i.e. the behaviour before being absorbed into a recurrent class?

4.10 Computing $P^{(\infty)}$ analytically

We have already seen that we can reason a bit about the structure of $P^{(\infty)}$, and we used some numerical computations (i.e. using MATLAB and computing some high power of P) to approximate $P^{(\infty)}$. Now, we will try to compute all the entries analytically.

4.10.1 Finite Absorbing Chains

To answer questions about the transient states and the transient behavior, it will be simpler to think about finite absorbing chains. A finite absorbing chain is one that has only transient states and absorbing states. To reason about the transient behavior we can simply convert any finite Markov chain into a finite absorbing chain by replacing every recurrent class with an absorbing state. This new Markov chain has the same transient behaviour as the old Markov chain.

Notice that I can also permute the states so that the absorbing states come first. Then the transition matrix has the following structure:

$$P = \begin{bmatrix} I & 0 \\ S & T \end{bmatrix}.$$

Supposing that there are a absorbing states and t transient states then I is the $a \times a$ identity matrix, T is a $t \times t$ matrix and S is a $t \times a$ matrix.

4.10.2 Probability of Absorption

In a finite absorbing chain, let us try to compute the probability of being absorbed into a particular absorbing state j when we begin in a transient state i, i.e. we would like to compute:

$$Q_{ij} = \mathbb{P}(\text{ever reaching j}|X_0 = i).$$

We return to our old friend and condition on the first step, i.e. we know that,

$$Q_{ij} = \sum_k \mathbb{P}(\text{ever reaching j}|X_0 = i, X_1 = k)P_{ik}.$$

There are a few possibilities: if k in the above sum is a transient state then $\mathbb{P}(\text{ever reaching j}|X_0 = i, X_1 = k)$ is simply Q_{kj}, if k is the absorbing state j then $\mathbb{P}(\text{ever reaching j}|X_0 = i, X_1 = k) = 1$ and finally if k is some other absorbing state then $\mathbb{P}(\text{ever reaching j}|X_0 = i, X_1 = k) = 0$. So we can put this together and see that:

$$Q_{ij} = P_{ij} + \sum_{k:k \text{ transient}} P_{ik}Q_{kj}.$$

As we have done several times before we can put all of these equations (there is one for (transient,absorbing) pair) and obtain the matrix system of equations:

$$Q = S + T \times Q,$$

where S and T are defined above. So from this we get

Fact 4.14.

$$Q = (I - T)^{-1}S.$$

Let us try to work out a couple of examples:

Example 4.22. We worked out this one numerically before but now we will do it analytically. Recall our usual random walk on integers. What is the probability that you reach "3" before you reach "-2"?

Let me remind you how we solved it before. Suppose we made the states 3 and -2 both absorbing. Then observe that P(reach "3" before you reach "-2"$|X_0 = 0$) is identical for both chains.

Furthermore, for the modified chain:

$$P(\text{reach "3" before you reach "-2"}|X_0 = 0) = P(\text{ever reach "3"}|X_0 = 0) = P_{03}^{(\infty)},$$

which we can compute numerically using MATLAB. So we write down the modified transition matrix:

$$P = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0.5 & 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0.5 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0.5 & 0 & 0.5 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

and find that $P_{03}^{(\infty)} \approx P_{03}^{(1000)} = 0.4$.

Now, suppose we look at the modified chain and try to compute Q. We find that:

$$T = \begin{bmatrix} 0 & 0.5 & 0 & 0 \\ 0.5 & 0 & 0.5 & 0 \\ 0 & 0.5 & 0 & 0.5 \\ 0 & 0 & 0.5 & 0 \end{bmatrix},$$

and

$$S = \begin{bmatrix} 0.5 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0.5 \end{bmatrix},$$

so that we can compute (I did it using MATLAB but one could imagine doing this by hand easily):

$$Q = (I - T)^{-1} S = \begin{bmatrix} 0.8 & 0.2 \\ 0.6 & 0.4 \\ 0.4 & 0.6 \\ 0.2 & 0.8 \end{bmatrix}.$$

This is the matrix of probabilities of being absorbed into the states -2 and 3 for each of the transient states. So now we can see that as we obtained previously: $P_{03}^{(\infty)} = 0.4$ (the $(2, 2)$ entry of the above matrix).

Let us do one more example that is a bit more complicated.

Example 4.23. Suppose we have a Markov chain with transition matrix:

$$P = \begin{bmatrix} 1/3 & 1/3 & 1/3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0 & 0.5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.5 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1/3 & 1/3 & 1/3 \end{bmatrix}.$$

Explain what is going on, and compute the probability of being absorbed into each of the recurrent classes.

When we draw the state diagram we see that there are two recurrent classes and one transient class in between them. Since we want to understand the transient behaviour we convert it to a finite absorbing chain. We can permute so that the two recurrent classes are at the top and we obtain the following transition matrix:

$$P = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0.5 & 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0.5 & 0 & 0.5 \\ 0 & 0.5 & 0 & 0.5 & 0 \end{bmatrix}.$$

As before we can write:

$$T = \begin{bmatrix} 0 & 0.5 & 0 \\ 0.5 & 0 & 0.5 \\ 0 & 0.5 & 0 \end{bmatrix},$$

and

$$S = \begin{bmatrix} 0.5 & 0 \\ 0 & 0 \\ 0 & 0.5 \end{bmatrix},$$

and we can compute:

$$Q = (I - T)^{-1} S = \begin{bmatrix} 0.75 & 0.25 \\ 0.5 & 0.5 \\ 0.25 & 0.75 \end{bmatrix},$$

which gives us the probability of being absorbed into the two recurrent classes from each of the three transient states.

4.10.3 Computing all of $P^{(\infty)}$

Now that we have done the above calculation, we can compute all of $P^{(\infty)}$ analytically. Lets do this in an example and I will explain the general idea.

Example 4.24. Suppose we have a Markov chain with transition matrix:

$$P = \begin{bmatrix} 1/3 & 1/3 & 1/3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0 & 0.5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.5 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1/3 & 1/3 & 1/3 \end{bmatrix}.$$

Compute $P^{(\infty)}$.

Instead of actually computing $P^{(\infty)}$ let me tell you what it is and then we can try to understand it together.

$$P^{(\infty)} = \begin{bmatrix} 0.3333 & 0.3333 & 0.3333 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.3333 & 0.3333 & 0.3333 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.3333 & 0.3333 & 0.3333 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.2500 & 0.2500 & 0.2500 & 0 & 0 & 0 & 0.0833 & 0.0833 & 0.0833 \\ 0.1667 & 0.1667 & 0.1667 & 0 & 0 & 0 & 0.1667 & 0.1667 & 0.1667 \\ 0.0833 & 0.0833 & 0.0833 & 0 & 0 & 0 & 0.2500 & 0.2500 & 0.2500 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.3333 & 0.3333 & 0.3333 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.3333 & 0.3333 & 0.3333 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.3333 & 0.3333 & 0.3333 \end{bmatrix}.$$

All of the zeros should be clear: in the long run you cannot end up in a transient state, and if you start in one recurrent class you cannot leave. Lets focus on the two recurrent classes. They both have transition matrix:

$$P_R = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{bmatrix},$$

and we can solve for the limiting distribution in the usual way or directly see that it must be [1/3 1/3 1/3]. So we can see how to obtain the top-left block and the bottom-right block of the matrix.

The only tricky parts are the rows that correspond to the transient states. I claim that the long run probabilities are actually simple – for any transient state they are just the probability of being absorbed into the recurrent class times the limiting distribution of the recurrent class. We calculated the probability of being absorbed into each class above:

$$Q = (I - T)^{-1}S = \begin{bmatrix} 0.75 & 0.25 \\ 0.5 & 0.5 \\ 0.25 & 0.75 \end{bmatrix},$$

and so if we multiply these numbers by [1/3 1/3 1/3], we will obtain the rest of $P^{(\infty)}$. Putting all these together we see how we can compute the entire matrix $P^{(\infty)}$ by some simple analytical calculations, i.e. we can understand almost all of the long-term behaviour of the Markov chain. There is one small piece left and we'll address that next.

4.11 Mean time spent in transient states

For a Markov chain we can define the expected number of visits to state j given the chain starts in state i. We have seen this quantity before and showed previously that

$$\mathbb{E}[\text{number of visits to state j}|X_0 = i] = \sum_{k=0}^{\infty} P_{ij}^k.$$

Let us denote this quantity v_{ij}, and put the v_{ij}s into a matrix $\mathbf{V}$. For a finite absorbing chain what can we say about the structure of V? Notice that we can write

$$V = \begin{bmatrix} A & \mathbf{0} \\ B & U \end{bmatrix}.$$

What can one say about the form of A, B and U?

The matrix U is known as the **fundamental matrix**, of the finite absorbing chain. There is a very simple way to calculate U. We will prove this in a second but:

Fact 4.15.

$$U = (I - T)^{-1}.$$

Given U we can calculate the expected amount of time spent in transient states given you start in a particular transient state. How?

We simply sum the appropriate row of the U matrix.

Example 4.25. How many times do I need to roll a dice to see n consecutive 6s? Lets do the case when $n = 2$ (it is easy to generalize).

We set up a Markov chain where the number of 6s so far is the state. The transition matrix is:

$$P = \begin{bmatrix} 5/6 & 1/6 & 0 \\ 5/6 & 0 & 1/6 \\ 0 & 0 & 1 \end{bmatrix}.$$

So the matrix T is:

$$T = \begin{bmatrix} 5/6 & 1/6 \\ 5/6 & 0 \end{bmatrix},$$

and we obtain that

$$U = (I - T)^{-1} = \begin{bmatrix} 36 & 6 \\ 30 & 6 \end{bmatrix}.$$

So to calculate the expected number of rolls we simply sum the first row (we started with no 6s) to obtain the answer 42!

Example 4.26. Let $\mathcal{T}$ denote the collection of all the transient states of a finite absorbing chain. Show that for any pair $i, j \in \mathcal{T}$, we have that

$$v_{ij} = \mathbb{I}(i = j) + \sum_{k \in \mathcal{T}} v_{kj} P_{ik}.$$

To show this result we simply do the usual thing, condition on the first step:

$$v_{ij} = \sum_{k} \mathbb{E}[\#\text{ visits to j} | X_0 = i, X_1 = k] P_{ik}.$$

Now we need to notice that if the state i and j are the same then we have already visited the state j once (at time step 0), and further if k is recurrent then the expected number of visits is 0, so we obtain that,

$$v_{ij} = \mathbb{I}(i = j) + \sum_{k \in \mathcal{T}} v_{kj} P_{ik}.$$

We can use this to express the fundamental matrix via a matrix equation in the usual way:

$$U = I + U \times T,$$

which in turn yields the fact.

4.12 Practice problems

In this section we will try to apply the concepts from the previous few sections to solve some problems.

Example 4.27. When flipped a coin comes up heads with probability p and tails with probability $1-p$.

1. Find the expected number of flips until either two consecutive heads or two consecutive tails come up.
2. Find the probability that two consecutive heads come up before two consecutive tails.

Suppose also that you are given the following useful fact:

$$\begin{bmatrix} 1 & -p & -(1-p) \\ 0 & 1 & -(1-p) \\ 0 & -p & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & \frac{(2-p)p}{p^2-p+1} & \frac{1-p^2}{p^2-p+1} \\ 0 & \frac{1}{p^2-p+1} & \frac{1-p}{p^2-p+1} \\ 0 & \frac{p}{p^2-p+1} & \frac{1}{p^2-p+1} \end{bmatrix}.$$

This is one of my favourite questions. It illustrates lots of the ideas we have seen so far. First we need to come up with a state space that is helpful to solve the problem, write down the Markov chain and analyze it.

One candidate state space is to take $\{0, 1h, 1t, 2h, 2t\}$ to denote the starting state, a sequence of 1 head, and so on. It is easy to then verify that this is a finite absorbing chain with the following transition matrix:

$$P = \begin{bmatrix} 0 & p & 1-p & 0 & 0 \\ 0 & 0 & 1-p & p & 0 \\ 0 & p & 0 & 0 & 1-p \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

The two questions are basically, what is the time spent in transient states, and what is the probability of absorption into each of the absorbing states. So we compute the fundamental matrix:

$$U = (I-T)^{-1} = \begin{bmatrix} 1 & \frac{(2-p)p}{p^2-p+1} & \frac{1-p^2}{p^2-p+1} \\ 0 & \frac{1}{p^2-p+1} & \frac{1-p}{p^2-p+1} \\ 0 & \frac{p}{p^2-p+1} & \frac{1}{p^2-p+1} \end{bmatrix}.$$

We also compute the matrix S:

$$S = \begin{bmatrix} 0 & 0 \\ p & 0 \\ 0 & 1-p \end{bmatrix}.$$

So to compute the time spent in transient states we simply sum the first row of U, i.e.

$$t = \frac{2-p^2+p}{p^2-p+1},$$

and we also compute $Q = U \times S$ (but we only need its first row since we start in the state 0):

$$Q_1 = \begin{bmatrix} \frac{(2-p)p^2}{p^2-p+1} & \frac{(1-p^2)(1-p)}{p^2-p+1} \end{bmatrix}.$$

Here is a similar (but slightly more involved) example. We will not solve it in class but it is a great practice problem (you should use MATLAB for the matrix inverses that you need).

Example 4.28. Three people, named A, B, and C, play the following game: In the first round, A plays B while C watches. The winner of that round plays in the second round against C while the loser of the first round watches. In the third round, the winner of the second round plays against the person who sat out the second round. And so on: In the nth round, the winner of round $n-1$ plays against the person that sat out round $n-1$. This continues until a player wins two consecutive rounds; that player is the winner of the game.

Assume that rounds are independent, and if A plays B, A wins the round with probability 0.6; if B plays C, B wins the round with probability 0.5, and if A plays C, A wins the round with probability 0.45.

1. What is the probability of winning for each person?
2. What is the expected number of rounds before the game ends?

Example 4.29. Consider the following system, there is queue of length 2, i.e. there can be 0, 1 or 2 jobs in the queue. Each day one of 3 things happens:

1. either a new job arrives (with probability p),
2. or an existing job (if there is one) is completed (with probability q),
3. or nothing happens.

If a new job arrives when the queue is full, it is simply rejected.

1. Set this up as a Markov chain, where the state space is the number of jobs currently in the queue. What is the transition matrix?
2. What are the classes of this Markov chain. Is this Markov chain, irreducible, aperiodic?
3. Calculate the limiting distribution of the Markov chain. This should be a function of p and q.
4. In the long run, what fraction of jobs will be rejected?

Example 4.30. An important problem in genetics, is to identify frequently occurring patterns in our DNA (these are known as motifs). Suppose that our DNA is a sequence made up of 4 characters: $\{A, G, T, C\}$. Further, for simplicity assume that they are all equally likely and random, i.e. every subsequent character is independently one of the above 4 characters.

Suppose, we are interested in the motif **'GAT'**, i.e. we are interested in how often this sequence occurs, if the sequence was genuinely random.

We let the state space be $X_n = \{0, 1, 2, 3\}$, where $X_n = 0$ if we matched none of the desired motif so far, $X_n = 1$ if we have matched the letter 'G' and so on. Once we match the full sequence, we remain in the state 3.

The transition matrix is:

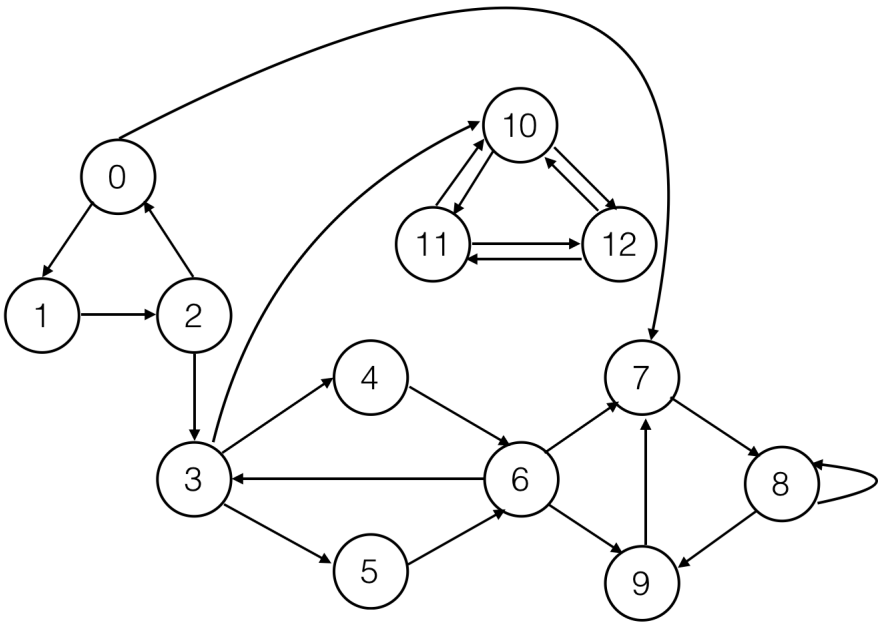
$$P = \begin{bmatrix} 3/4 & 1/4 & 0 & 0 \\ 1/2 & 1/4 & 1/4 & 0 \\ 1/2 & 1/4 & 0 & 1/4 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

This is just the expected time spent in the transient states. We compute the fundamental matrix as:

$$U = \begin{bmatrix} 1/4 & -1/4 & 0 \\ -1/2 & 3/4 & -1/4 \\ -1/2 & -1/4 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 44 & 16 & 4 \\ 40 & 16 & 4 \\ 32 & 12 & 4 \end{bmatrix}.$$

and sum the first row to obtain the answer 64.

Example 4.31. Suppose we have a Markov chain with state diagram:



Describe its classes, whether they are transient/recurrent and what their periods are. Describe the long-term behaviour.

Again, we will not solve this one in lecture but great practice problem (use MATLAB)!

Example 4.32. Smith is in jail and has 3 dollars; he can get out on bail if he has 8 dollars. A guard agrees to make a series of bets with him. If Smith bets A dollars, he wins A dollars with probability .4 and loses A dollars with probability .6. Find the probability that he wins 8 dollars before losing all of his money if

1. He bets 1 dollar each time (timid strategy).
2. He bets, each time, as much as possible but not more than necessary to bring his fortune up to 8 dollars (bold strategy).
3. Which strategy gives Smith the better chance of getting out of jail?

Example 4.33. Mary and John are playing the following game: They have a three-card deck marked with the numbers 1, 2, and 3 and a spinner with the numbers 1, 2, and 3 on it. The game begins by dealing the cards out so that the dealer gets one card and the other person gets two. A move in the game consists of a spin of the spinner. The person having the card with the number that comes up on the spinner hands that card to the other person. The game ends when someone has all the cards.

1. Set up the transition matrix for this absorbing Markov chain, where the states correspond to the number of cards that Mary has.
2. Find the fundamental matrix.
3. On the average, how many moves will the game last?
4. If Mary deals, what is the probability that John will win the game?

The state space has been given to us, and the transition matrix is:

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

We can calculate the fundamental matrix as:

$$U = (I - T)^{-1} = \begin{bmatrix} 1 & -2/3 \\ -2/3 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 1.8 & 1.2 \\ 1.2 & 1.8 \end{bmatrix}.$$

To answer the first question we simply sum (either row) of the fundamental matrix to see that on average the game will last 3 moves.

To answer the second question, we need the matrix S:

$$S = \begin{bmatrix} 1/3 & 0 \\ 0 & 1/3 \end{bmatrix}.$$

If Mary deals, she begins with one card so we start in the first of the two transient states. We want to calculate the probability of absorption into the first of the two absorbing states (where John wins). So we compute $(U \times S)_{11} = 0.6$.

Example 4.34. A certain experiment is believed to be described by a two-state Markov chain with the transition matrix P, where

$$P = \begin{bmatrix} .5 & .5 \\ p & 1-p \end{bmatrix}.$$

and the parameter p is not known. When the experiment is performed many times, the chain ends in state one approximately 20 percent of the time and in state two approximately 80 percent of the time. Compute a sensible estimate for the unknown parameter p and explain how you found it.

We know that the limiting distribution of the chain is: $\pi^T = [0.2 \quad 0.8]$, so we simply solve $\pi^T = \pi^T P$ where the transition matrix has an unknown parameter p. We obtain the equation:

$$0.1 + 0.8p = 0.2 \implies p = \frac{1}{8}.$$

Example 4.35. Consider a chain with transition matrix

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

1. Find the limiting distribution of this chain.
2. Find the expected time between visits to state 1.

It is clear that the chain is irreducible, aperiodic, and finite so we need to solve the system of equations:

$$\pi^T = \pi^T P,$$

to obtain that $\pi^T = [0.0811 \quad 0.6486 \quad 0.2703]$. We use the fact that we can relate the limiting distribution to the expected return times, to see that:

$$n_{11} = \frac{1}{\pi_1} = 12.33.$$

Example 4.36. Consider a sequence $\Xi_1, \Xi_2, \ldots$ of independent Ber(p) random variables, where $0 < p < 1$.

Let $Z_0 = 0$ and define Z_n to be the length of the run of 1s looking back from time n. That is, for sequence 00110111010001111, we get for instance $Z_1 = 0$, $Z_2 = 0$, $Z_3 = 1$, $Z_4 = 2$, $Z_5 = 0$, $Z_8 = 3$, $Z_{13} = 0$, and $Z_{17} = 4$. In general,

$$Z_n = \max\{0 < k \leq n : \Xi_n \cdots \Xi_{n-k+1} = 1\},$$

with $Z_n = 0$ if the set is empty (i.e., $\Xi_n = 0$).

1. Is Z a Markov chain? Justify your answer.

2. What is the state space of Z?

3. Find the initial distribution $\pi^{(0)}$ and transition probabilities P.

4. Find the communicating classes of the chain. Is Z irreducible? Explain your answer.

5. Find the periods of the communicating classes.

6. Is Z recurrent? Justify your answer in the following way: we know that for any recurrent state it must be the case that $\sum_{k=0}^{\infty} P_{ii}^k = \infty$. Argue explicitly that this is the case for state 0, i.e. that $\sum_{k=0}^{\infty} P_{00}^k = \infty$ Hint: Try to lower bound each term in the sum.

7. It turns out that this Markov chain is positive recurrent (you don't need to prove this) and has a unique stationary distribution. Find it. Hint: Try to solve the system of equations $\pi = \pi P$. P is an infinite matrix (and π is an infinite vector) but you should still be able to find a distribution that solves this system of equations.

$$\pi = [q \;\; pq \;\; p^2q \;\; p^3q \;\; p^4q \;\; \dots \;\;].$$

Example 4.37. In each game, a gambler wins the dollars he bets with probability p and loses with probability $q = 1 - p$. If he has less than \$3, he will bet all he has. Otherwise, since his goal is to have \$5, he will only bet the difference between \$5 and what he has. He continues to bet until he has either \$0 or \$5. Let X_n bet the amount he has immediately after the nth bet.

1. Find the transition probabilities for this chain.

2. Verify that the "fundamental matrix" for this chain equals

$$U = \frac{1}{1-p^2q^2}\begin{bmatrix} 1 & p & p^2q & p^2 \\ pq^2 & 1 & pq & p \\ q & pq & 1 & p^2q \\ q^2 & pq^2 & q & 1 \end{bmatrix}$$

   Hint: You shouldn't have to invert a matrix to do this. Why not?

3. Find the expected number of bets before he stops for each of the possible starting states.

So there are 6 states corresponding to the gambler having $\{0, 1, 2, 3, 4, 5\}$ dollars. The transition matrix is:

$$P = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ q & 0 & p & 0 & 0 & 0 \\ q & 0 & 0 & 0 & p & 0 \\ 0 & q & 0 & 0 & 0 & p \\ 0 & 0 & 0 & q & 0 & p \\ 0 & 0 & 0 & 0 & 0 & q \end{bmatrix}.$$

To proceed we need to check that for the given U, $U \times (I - T) = I$.

Example 4.38. Suppose we have a Markov chain with transition matrix:

$$P = \begin{bmatrix} 1/3 & 1/3 & 1/3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0 & 0.5 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0 & 0.5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.5 & 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.5 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/6 & 1/3 & 1/2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/6 & 1/3 & 1/2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/6 & 1/3 & 1/2 \end{bmatrix}.$$

What are the communicating classes, are they transient or recurrent? What is the long-term behavior? Compute all entries of the matrix P^∞ and the matrix V.

There are several steps involved so lets go through them. The communicating classes are $\{1,2,3\},\{4,5,6,7\},\{8,9,10\}$. The first two are transient and the last one is recurrent.

There are three possible settings: if we begin in the first recurrent class then we will converge to its limiting distribution [1/3 1/3 1/3 0 0 0 0 0 0 0] in the long-term. If we begin in the transient class we will end up in one of the recurrent classes in the long-term (we will be more precise about this in a minute), and if we start in the second recurrent class we will have the distribution [0 0 0 0 0 0 0 1/6 1/3 1/2] in the long-term.

Now, we can re-write this as a finite absorbing chain by changing every recurrent class to an absorbing state. The resulting chain has matrices:

$$T = \begin{bmatrix} 0 & 0.5 & 0 & 0 \\ 0.5 & 0 & 0.5 & 0 \\ 0 & 0.5 & 0 & 0.5 \\ 0 & 0 & 0.5 & 0 \end{bmatrix},$$

and

$$S = \begin{bmatrix} 0.5 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0.5 \end{bmatrix},$$

so that we can compute (I did it using MATLAB but one could imagine doing this by hand easily):

$$Q = (I-T)^{-1}S = \begin{bmatrix} 0.8 & 0.2 \\ 0.6 & 0.4 \\ 0.4 & 0.6 \\ 0.2 & 0.8 \end{bmatrix},$$

and

$$U = (I-T)^{1} = \begin{bmatrix} 1.6 & 1.2 & 0.8 & 0.4 \\ 1.2 & 2.4 & 1.6 & 0.8 \\ 0.8 & 1.6 & 2.4 & 1.2 \\ 0.4 & 0.8 & 1.2 & 1.6 \end{bmatrix}.$$

We have computed all of these before (see Example 4.22). With this in place we can compute the matrix V as:

$$V = \begin{bmatrix} \infty & \infty & \infty & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \infty & \infty & \infty & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \infty & \infty & \infty & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \infty & \infty & \infty & 1.6 & 1.2 & 0.8 & 0.4 & \infty & \infty & \infty \\ \infty & \infty & \infty & 1.2 & 2.4 & 1.6 & 0.8 & \infty & \infty & \infty \\ \infty & \infty & \infty & 0.8 & 1.6 & 2.4 & 1.2 & \infty & \infty & \infty \\ \infty & \infty & \infty & 0.4 & 0.8 & 1.2 & 1.6 & \infty & \infty & \infty \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \infty & \infty & \infty \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \infty & \infty & \infty \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \infty & \infty & \infty \end{bmatrix}.$$

By multiplying entries of Q with the limiting distribution we can also compute $P^{(\infty)}$:

$$P^{(\infty)} = \begin{bmatrix} 1/3 & 1/3 & 1/3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 4/15 & 4/15 & 4/15 & 0 & 0 & 0 & 0 & 1/30 & 1/15 & 1/10 \\ 1/5 & 1/5 & 1/5 & 0 & 0 & 0 & 0 & 0.0667 & 0.1333 & 0.2 \\ 2/15 & 2/15 & 2/15 & 0 & 0 & 0 & 0 & 1/10 & 2/10 & 3/10 \\ 1/15 & 1/15 & 1/150 & 0 & 0 & 0 & 0 & 0.1333 & 0.2667 & 0.4 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/6 & 1/3 & 1/2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/6 & 1/3 & 1/2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/6 & 1/3 & 1/2 \end{bmatrix}.$$

Chapter 5

Branching Processes

In this relatively short chapter we will discuss a particular type of infinite Markov chain called a branching process and study some of its properties. The type of branching process we study is sometimes called a "Galton-Watson process". Sir Francis Galton was a very famous statistician who was interested in the probability that a given family name would go extinct (the sort of thing that statisticians of the day cared about). This stochastic process since then has found many applications in genetics and even in particle physics.

5.1 Basic Setup

Consider a population which consists of individuals that can produce offspring. Let X_n denote the number of members in the n^{th} generation. Let Z_{in} equal the number of offspring produced by the i^{th} member of the n^{th} generation. Thus,

$$X_{n+1} = \begin{cases} \sum_{i=1}^{X_n} Z_{in}, & \text{if } X_n \geq 1, \\ 0 & \text{if } X_n = 0. \end{cases}$$

We assume that the Z_{in} are independent, and identically distributed with mean μ and variance $\sigma^2 > 0$. It is often assumed that $X_0 = 1$. Denote the distribution of Z_{in} by:

$$P_j = \mathbb{P}(Z_{in} = j).$$

Fact 5.1. $\{X_n : n \geq 0\}$ is a Markov chain with infinite state space.

In general, if $P_j > 0$ for $j \geq 2$ then it is clear that the Markov chain has an infinite state space. To see that it is Markov chain, let us consider the distribution of $X_{n+1}|X_n = j$. This is the same as the distribution of,

$$\sum_{i=1}^{j} Z_{in},$$

where Z_{in} are all i.i.d. This does not depend in any way on the past $\{X_{n-1}, \ldots, X_0\}$

Question 5.1. Recall, $P_0 = P(Z_{in} = 0)$. Assume $P_0 > 0$. Classify the states of the chain as recurrent or transient.

We know that the state $\{0\}$ is absorbing and so is in its own class. For the rest of the states $\{2, 3, 4, \ldots\}$ we note that they are all communicating and transient, since there is a non-zero probability that we will hit the absorbing state 0 and have a return probability < 1.

5.2 Mean and Variance: Short/Medium term Behavior

A natural question is to characterize the distribution of the population after n time steps, i.e. after n generations. This turns out to be a bit difficult but computing the mean and variance is relatively straightforward.

Fact 5.2.

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

We will prove this via induction.
Base Case: Since we always take $X_0 = 1$, the base case is straightforward.
Induction: Observe that,

$$\begin{aligned}
\mathbb{E}[X_n] &= \sum_j \mathbb{E}[X_n | X_{n-1} = j] \mathbb{P}(X_{n-1} = j) \\
&= \sum_j \mathbb{E}\left[\sum_{i=1}^{j} Z_{in}\right] \mathbb{P}(X_{n-1} = j) \\
&= \sum_j j\mu \mathbb{P}(X_{n-1} = j) = \mu \mathbb{E}[X_{n-1}] = \mu \times \mu^{n-1} = \mu^n.
\end{aligned}$$

We can also calculate the variance:

Fact 5.3.

$$\mathrm{Var}(X_n) = \begin{cases} n\sigma^2 & \text{if } \mu = 1, \\ \sigma^2 \mu^{n-1} \left(\frac{1-\mu^n}{1-\mu}\right) & \text{if } \mu \neq 1. \end{cases}$$

Analogous to the law of total expectation, there is a law of total variance that tells us how to calculate the variance of a random variable by conditioning on another random variable, i.e.

$$\mathrm{Var}(X_n) = \mathbb{E}[\mathrm{Var}(X_n | X_{n-1})] + \mathrm{Var}[\mathbb{E}(X_n | X_{n-1})],$$

Now, conditioning on X_{n-1} the mean of X_n is $X_{n-1}\mu$ and the variance is $X_{n-1}\sigma^2$, so we obtain

$$\begin{aligned}
\mathrm{Var}(X_n) &= \sigma^2 \mathbb{E}[X_{n-1}] + \mu^2 \mathrm{Var}(X_{n-1}) \\
&= \sigma^2 \mu^{n-1} + \mu^2 \mathrm{Var}(X_{n-1}) \\
&= \sigma^2 \left[\mu^{n-1} + \mu^n\right] + \mu^4 \mathrm{Var}(X_{n-2}).
\end{aligned}$$

Continuing to unroll this recursion, and noting $\mathrm{Var}(X_0) = 0$ we obtain that,

$$\mathrm{Var}(X_n) = \sigma^2 \left[\mu^{n-1} + \mu^n + \ldots + \mu^{2n-2}\right],$$

which is precisely the stated fact.

5.3 Probability of Dying Out: Long-term Behavior

Of interest is the probability that the population will eventually die out, if $X_0 = 1$. Denote this probability π_0. As one would expect, π_0 depends on the distribution Z_{in}. Remarkably, however, whether or not $\pi_0 = 1$, depends only on the mean of Z_{in}, denoted μ. In particular, $\pi_0 = 1$ if and only if $\mu \leq 1$.

Fact 5.4. If $\mu < 1$, then $\pi_0 = 1$.

To see this we note that,

$$\begin{aligned}\mathbb{P}(X_n > 0) &= \sum_{j=1}^{\infty} \mathbb{P}(X_n = j) \\ &\leq \sum_{j=1}^{\infty} j\mathbb{P}(X_n = j) \\ &= \mathbb{E}[X_n] = \mu^n.\end{aligned}$$

So the probability that $X_n > 0$ tends to 0 as $n \to \infty$, so we conclude that the $\pi_0 \to 1$.

Fact 5.5. One key fact about a branching process is that if $P_0 > 0$ then either the population will die out or $\to \infty$. Why?

Recall that we showed earlier that the state $\{0\}$ is absorbing while the rest are transient. Let us focus our attention on states $\{1, \ldots, k\}$ for some finite k. Then since the states $\{1, \ldots, k\}$ are transient we must eventually not return to them, i.e. there are two possibilities we get absorbed into 0 (i.e. the population dies out) or the population reaches $k + 1$. Since k was arbitrary this means that for any finite k, the population dies out or grows beyond k (i.e. grows to ∞).

More generally, there are three regimes of interest:

- $\mu < 1$, population will eventually die out, i.e. $P(X_n = 0) \to 1$.
- $\mu = 1$, same as above, but we won't prove this. Intuitively, the mean population size is 1 but the variance will keep growing so reasonable chance at some point the population will die off.
- $\mu > 1$, in this case $P(X_n = 0) < 1$, i.e. there is some chance the population size will converge to infinity.

Fact 5.6. In the case when $\mu > 1$, verify that the limiting probability $\lim_{n\to\infty} P(X_n = 0) = \pi_0$ must satisfy:

$$\pi_0 = \sum_{j=0}^{\infty} \pi_0^j P_j.$$

We rely on the usual trick of conditioning on the first step. So we have that,

$$\begin{aligned}\pi_0 = \mathbb{P}(\text{ever dying out}|X_0 = 1) &= \sum_j \mathbb{P}(\text{ever dying out}|X_0 = 1, X_1 = j)\mathbb{P}(X_1 = j|X_0 = 1) \\ &= \sum_j \mathbb{P}(\text{ever dying out}|X_0 = 1, X_1 = j)P_j \\ &= \sum_j \pi_0^j P_j,\end{aligned}$$

since if there are j different individuals, their branches die out with probability π_0 independently, and so the whole population dies out with probability π_0^j.

Fact 5.7. Of course, $\pi_0 = 1$ always solves this equation. It turns out that when $\mu > 1$ there is always a value $\pi_0 < 1$ that satisfies this equation. The smallest (strictly) positive solution of this equation is the π_0 we are interested in.

You will prove this fact in your HW.

Example 5.1. If $P_0 = \frac{1}{2}$, $P_1 = P_2 = \frac{1}{4}$ then determine π_0.

We go through two steps, first we calculate the mean:

$$\mu = 1 \times P_1 + 2 \times P_2 = \frac{3}{4} < 1,$$

so we know that $\pi_0 = 1$.

Example 5.2. If $P_0 = \frac{1}{4}$, $P_1 = \frac{1}{4}$ and $P_2 = \frac{1}{2}$ then determine π_0.

We again calculate the mean, to obtain:

$$\mu = 1 \times \frac{1}{4} + 2 \times \frac{1}{2} = \frac{5}{4} > 1,$$

so we know that $\pi_0 < 1$, we write down the equation above:

$$\pi_0 = \frac{1}{4} + \frac{1}{4}\pi_0 + \frac{1}{2}\pi_0^2,$$

which has two solutions $\pi_0 = \{1/2, 1\}$. So we take the smallest positive solution: in this case it is $1/2$.

Example 5.3. In each of the previous examples, if instead we had an initial population of size N (where N is some fixed number), what is the probability that the population would die out? i.e. what is the new π_0?

In the first case, we would still have $\pi_0 = 1$, and in the second case we would need each of the N populations to die off, i.e. we obtain that $\pi_0 = \frac{1}{2^N}$.

Chapter 6

Time Reversible Markov Chains and Markov Chain Monte Carlo

We often state the Markov assumption, as "the past is independent of the future given the present". Of course, this statement is symmetrical in time, and it suggests that we could also look at a Markov chain with time running backwards.

This however is a bit simplistic, since not all behaviour of a Markov chain is reversible. For instance, convergence to the limiting distribution is very non-symmetric. If we have a nice Markov chain (one with a unique limiting distribution) you can start in any state, run the chain forwards and end up in the limiting distribution, but this process is not meaningful to reverse[1].

A deeper question, is what happens if we start the Markov chain from the limiting distribution π? Now, fix some large number N, and think about the reversed stochastic process:

$$Y_n = X_{N-n}, \quad \text{for} \quad 0 \leq n \leq N.$$

Question 6.1. Is the stochastic process Y_n a Markov chain? What is its transition matrix? What is its limiting distribution?

As we discussed above by the symmetry of the Markov assumption (i.e. $X_{n-1} \perp\!\!\!\perp X_{n+1} | X_n$ is a symmetric statement in $n+1$ and $n-1$), it is clear that the new stochastic process Y_n is also a Markov chain. The symbol $\perp\!\!\!\perp$ is used to denote independence.

Let Q be the transition matrix for the Markov chain $\{Y_0, \ldots, Y_n, \ldots\}$ then:

$$\begin{aligned} Q_{ij} &= \mathbb{P}(Y_{n+1} = j | Y_n = i) \\ &= \mathbb{P}(X_{N-n-1} = j | X_{N-n} = i) \\ &= \frac{\mathbb{P}(X_{N-n} = i | X_{N-n-1} = j) \times \mathbb{P}(X_{N-n-1} = j)}{\mathbb{P}(X_{N-n} = i)}, \end{aligned}$$

using Bayes' rule. Now we can use the fact that we started in the stationary distribution to observe that,

$$Q_{ij} = \frac{P_{ji} \times \pi_j}{\pi_i},$$

[1] Throughout the rest of the lecture we will focus on Markov chains which have a unique stationary distribution equal to its limiting distribution.

where $\pi_i = \mathbb{P}(X_n = i)$. Notice that it is important that we started the forward (X) Markov chain in its stationary distribution.

It is also not hard to see that any distribution that is stationary for the forward Markov chain is also stationary for the reversed Markov chain (and vice-versa). Since the forward Markov chain has a unique stationary and limiting distribution this tells us that so does the reversed Markov chain. To see the first part, focussing on the i-the entry we see that:

$$[\pi^T Q]_i = \sum_j \pi_j Q_{ji} = \sum_j \pi_i P_{ij} = \pi_i \sum_j P_{ij} = \pi_i,$$

so we have verified that $\pi^T Q = \pi^T$.

6.1 Time-reversible Markov chains and Detailed Balance

A Markov chain is called **time reversible** if the transition probabilities of the reversed chain are the same as the transition probabilities of the forward chain, i.e. if $P_{ij} = Q_{ij}$. Intuitively, a time-reversible Markov chain is one that is indistinguishable when run forwards and backwards (i.e. if I tell you, you are either observing the forward chain X or the backward chain Y you cannot tell which). More formally, if $X_0 \sim \pi$, then

$$(X_0, X_1, \ldots, X_t) \stackrel{d}{=} (X_t, X_{t-1}, \ldots, X_0).$$

Fact 6.1. Detailed balance: The requirement of time-reversibility can be simply expressed as a condition on the original transition matrix (and stationary distribution), i.e.

$$\pi_i P_{ij} = \pi_j P_{ji} \quad \forall \quad (i, j).$$

These conditions are known as the detailed balance conditions. Intuitively, it should be clear why detailed balance implies reversibility.

As we will see in the sequel, detailed balance conditions often give a convenient way to find the limiting distribution of a Markov chain. To understand this we need to first understand the following fact:

Fact 6.2. Suppose that for a transition matrix corresponding to a finite, irreducible and aperiodic Markov chain, we can find a distribution μ that satisfies the detailed balance conditions, then

1. The stationary distribution of the Markov chain is equal to μ.
2. The Markov chain is reversible.

Before we prove this we need to understand its significance. It is essentially telling us that if we could "solve" the detailed balance equations (this is not always possible) then we will have found the stationary distribution (and have established that the chain is time-reversible).

We really only need to show one direction here, i.e. we need to verify that if a distribution μ satisfies detailed balance then it must be stationary. To see this we can simply sum up the

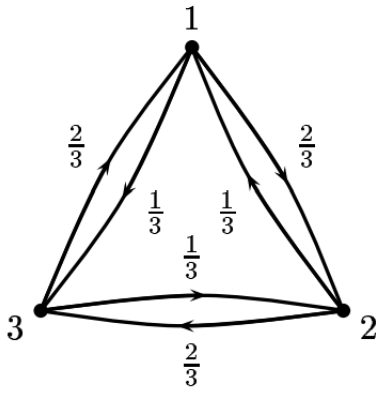
detailed balance conditions for each state, i.e. if a distribution μ satisfied the detailed balance conditions then

$$\sum_i \mu_i P_{ij} = \sum_i \mu_j P_{ji} \implies \mu_j = \sum_i \mu_i P_{ij}, \quad \text{i.e.} \quad \mu^T = \mu^T P,$$

so we conclude that μ is stationary.

Let us consider a few examples.

Example 6.1. Consider the Markov chain depicted in the figure. Is it time-reversible?



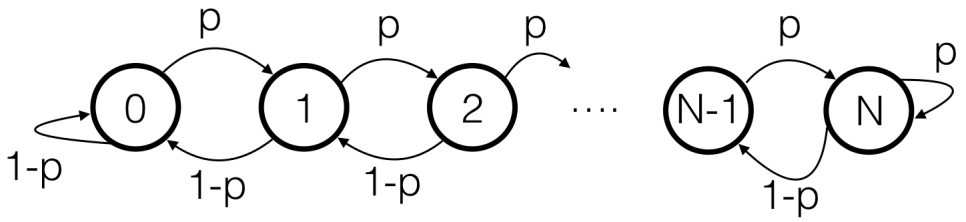
Intuitively, it seems like we are more likely to move clock-wise than counter clock-wise so it seems like the chain is not time-reversible (since in reverse you would see the opposite). This intuition can in some cases be misleading (in this case it is not) but lets try to be a bit more formal.

Formally, we write the transition matrix:

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

and compute that $\pi^T = [1/3 \;\; 1/3 \;\; 1/3]$. Finally, noting that $\pi_1 P_{12} \neq \pi_2 P_{21}$ (for instance) we conclude that the MC is not reversible.

Example 6.2. Consider the Markov chain depicted in the figure. Is it time-reversible?



Again intuitively it seems like if $p \neq 1/2$ then the chain is not time-reversible. Interestingly

this is not correct. Lets do the calculation. We can verify in the usual way that,

$$\pi \propto \begin{bmatrix} 1 \\ \frac{p}{q} \\ \frac{p^2}{q^2} \\ \vdots \\ \frac{p^N}{q^N} \end{bmatrix}.$$

This is telling us that:

$$\pi_{i+1} = \frac{p}{q}\pi_i,$$

for each i which is just the detailed balance condition for this Markov chain. So the chain is indeed time-reversible.

Somewhat more usefully we could have computed the limiting distribution just by solving the detailed balance conditions. In this case, it would have been roughly the same amount of effort as directly solving the equation the $\pi^T = \pi^T P$.

Example 6.3. An important case is the Markov chain corresponding to a (connected) non-negative weighted undirected graph, i.e. consider a particle moving from node to node in the following way: if the particle is at node i, it will move to neighbor j of i with probability

$$P_{ij} = \frac{w_{ij}}{\sum_j w_{ij}}.$$

Furthermore, the graph is symmetric, i.e. $w_{ij} = w_{ji}$. Show that this Markov chain is time reversible, and find its stationary distribution π.

In this case we will try to directly solve the detailed balance equations to obtain π (and show that the MC is time reversible). The detailed balance conditions tell us that for each pair (i, j):

$$\pi_i \frac{w_{ij}}{\sum_u w_{iu}} = \pi_j \frac{w_{ji}}{\sum_v w_{jv}}.$$

Noting that $w_{ij} = w_{ji}$ we can re-arrange this to see that,

$$\pi_i \sum_v w_{jv} = \pi_j \sum_u w_{iu},$$

and we can sum these equations over j to see that,

$$\pi_i \sum_j \sum_v w_{jv} = \sum_u w_{iu} \sum_j \pi_j,$$

so we obtain that,

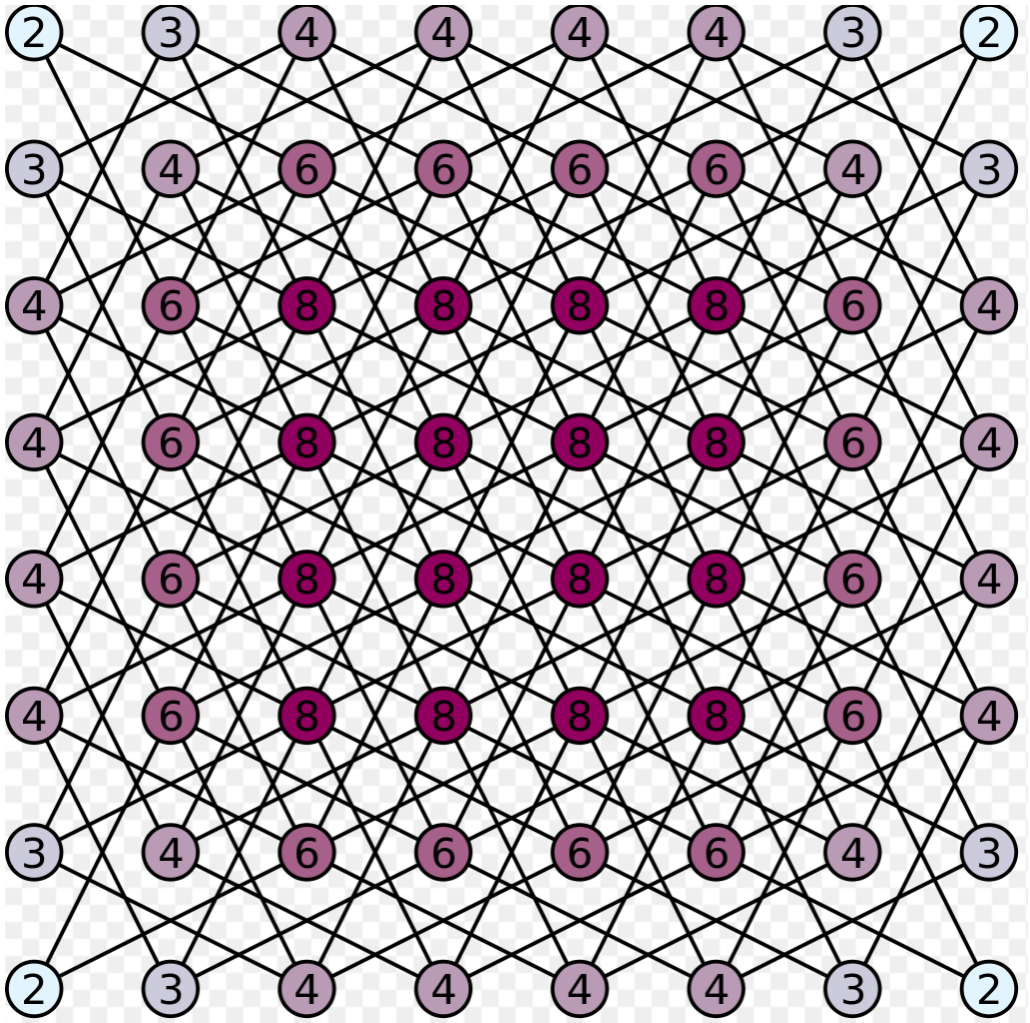
$$\pi_i = \frac{\sum_u w_{iu}}{\sum_j \sum_v w_{jv}}.$$

It is clear that this solution gives us a distribution which satisfied the detailed balance conditions so we conclude that π is stationary and that the Markov chain is time-reversible. The probabilities π_i are simply:

$$\pi_i = \frac{\text{degree of i}}{\text{total degree of graph}},$$

which intuitively makes sense: the time we spend in each state is proportional to its degree.

Example 6.4 (Knight moves)**.** One of the classical problems in probability is the random knight problem. Suppose we have a knight on a chess board that makes random moves, i.e. at each time step it picks a legal move at random and makes it. What is the expected time to return to the starting square?



Notice that this is exactly the same type of problem as the previous example, i.e. it is simply a random walk on an undirected graph. We already know what the stationary distribution is using the previous example. We know that,

$$n_{ii} = \frac{1}{\pi_i}.$$

We simply need to compute the total degree and the degree of each node to compute π_i. The total degree is 336, and so if we select the bottom left corner (which has degree 2) then we will have that the return time is $1/(2/336) = 168$.

6.2 Sampling and integration

One of the main reasons why we introduced time-reversibility was so that we could reason about Markov Chain Monte Carlo (MCMC) methods. MCMC methods are by far the most popular use of Markov Chain theory. At a high-level they give a way to draw samples from distributions that are very difficult to sample otherwise. Before we get to MCMC however we'll need to understand the problems of sampling and integration.

Question 6.2. Suppose that we have a continuous random variable X with pdf f. What is the expected value of $h(X)$, where $h(.)$ is a real-valued function?

The answer is:

$$\mathcal{I} = \mathbb{E}_{X\sim f} h(X) = \int h(x)f(x)dx.$$

Question 6.3. Suppose the random variable has an **unknown** pdf but I have the ability to draw samples from f. How could we estimate the expected value of $h(X)$?

The natural strategy is to draw many random samples $\{X_1, \ldots, X_n\}$ from f, and then try to approximate the expected value as:

$$\widehat{\mathcal{I}} = \frac{1}{n}\sum_{i=1}^{n} h(X_i).$$

It is easy to see that $\mathbb{E}[\widehat{\mathcal{I}}] = \mathcal{I}$, and by the law of large numbers if n is large $\widehat{\mathcal{I}}$ will be close to $\mathcal{I}$ with high-probability.

- This is an example of something called **Monte Carlo integration**. This situation where we cannot explicitly describe a pdf but can sample from it arises frequently in the physical sciences. In these settings, we can simulate a system of interest and then look at the output (this is a sample), but cannot tractably determine the pdf of the output variable.

To simplify things a little bit and to keep it in line with the discrete state Markov chains we have been discussing so far lets consider this problem in the discrete case. If we have a discrete random vector X with p.m.f

$$\pi(j) = \mathbb{P}(X = x_j),$$

and would like to calculate:

$$\mathcal{I} = \mathbb{E}_{X\sim\pi} h(X) = \sum_j \pi(j)h(x_j),$$

we can once again use Monte Carlo integration, i.e. we sample $X_1, \ldots, X_n$ from π and compute:

$$\widehat{\mathcal{I}} = \frac{1}{n} \sum_{i=1}^{n} h(X_i).$$

In many cases, it is difficult to sample directly from the distribution π and in some of these cases MCMC algorithms might be useful.

Here are two examples:

Example 6.5 (Graphical Models)**.** In statistical physics we often come across models like this one. This is called an Ising model. The idea is we have a collection of discrete random variables $(X_1, \ldots, X_n)$ which each take values in $\{-1, 1\}$: each random variable represents an atom and its value represents its direction of spin. Now, in what are called ferromagnetic materials atoms that are adjacent "like" to share the same spin. This is represented by an energy function:

$$E(X_1, \ldots, X_n) = - \sum_{(i,j) \text{ adjacent}} X_i X_j,$$

so the energy is low if adjacent atoms share the same spin. Now, the probabilistic model (i.e. the Ising model) is that the probability of any particular configuration is inversely proportional to its energy, and in particular the probability of a configuration $(x_1, \ldots, x_n)$ has the form:

$$P((x_1, \ldots, x_n)) \propto \exp(-\beta E(x_1, \ldots, x_n)).$$

The parameter β is called a temperature parameter. We'd like to be able to draw samples from this distribution in order to understand what the "likely" configurations are. The key point is that this can be computationally difficult since computing the normalizer:

$$Z = \sum_{(x_1, \ldots, x_n)} \exp(-\beta E(x_1, \ldots, x_n)),$$

is hard because there are 2^n terms in the sum. We will see how we can sample from this distribution without normalizing it – this is the magic of MCMC.

Example 6.6 (Sampling from combinatorial sets)**.** A related example arises when we want to sample distributions with very large support. Suppose that we want to generate a uniformly distributed element $(x_1, \ldots, x_n)$ from the set of all permutations of $\{1, \ldots, n\}$ for which some constraint is satisfied, say:

$$\sum_{j=1}^{n} j x_j > a, \tag{6.1}$$

for some given constant a. Intuitively, these are the subset of permutations that put the higher entries of $\{1, \ldots, n\}$ later on in the permutation, i.e. they are close to the identity permutation $(1, 2, \ldots, n)$. It is difficult to even count how many elements this set has, but

again we know the distribution upto the normalizer, i.e. for any element $(x_1, \ldots, x_n)$

$$P((x_1, \ldots, x_n)) \propto \mathbb{I}\left(\sum_{j=1}^{n} j x_j > a\right).$$

Computing the normalizer requires us to figure out how many permutations there are in the set. As we will see we can use MCMC to construct samplers without computing this normalizer.

The abstract setting: What is common to the above two examples (and many more) is that we know the distribution we want to sample from up to the normalizer, i.e. we know that:

$$\pi(X = x_j) = C b_j, \tag{6.2}$$

where X and x_j may be vectors, and C is an *unknown* constant (i.e. the normalizer). Both b_j for each j and C are assumed to be ≥ 0.

Now, our broad goal is to set up a Markov chain which has limiting distribution π. Furthermore, we want to be able to "construct" a way to draw samples from this Markov chain, knowing only b_j (i.e. not knowing C).

6.3 Metropolis-Hastings Algorithm

Lets first describe the sampling algorithm and then try to figure out what it is doing. We are going to construct a Markov chain with state space given by the values of the discrete random variable we want to draw samples from.

1. In the Ising model example, the states corresponding to the 2^n possible binary vectors.
2. In the permutation example, the states correspond to the different possible permutations that satisfy the condition (6.1).

First we need to describe what are called *proposal distributions*. These are like a transition matrix for a Markov chain: let Q represent the proposal distributions, i.e. for each state i the row $Q(i, \cdot)$ represents a distribution over states that we draw the candidate next state from.

Q is chosen to ensure that the resulting Markov chain is irreducible and aperiodic (this is not difficult). Often we choose Q to make what are called "local-moves". In the permutation example, we might consider the following:

For any permutation, $(x_1, \ldots, x_n)$ we can consider the permutations $(x_1^1, \ldots, x_n^1), (x_2^2, \ldots, x_n^2), \ldots$ which all satisfy the condition in (6.1) that we might obtain by swapping two entries of the original permutation $(x_1, \ldots, x_n)$. Let $N(x_1, \ldots, x_n)$ denote the number of such "neighbors" that satisfy the condition, then we choose one of these neighbors with probability $1/N$ (i.e. we uniformly sample a neighbor). This describes Q.

Q clearly defines the transition matrix of a Markov chain, however its stationary distribution is not going to be π so we need to modify it in some way. This is done by the following algorithm (known as the Metropolis-Hastings (MH) algorithm): we start in some state X_0, and then for each i we repeat the following:

1. Sample a proposal $Y \sim Q(X_i, \cdot)$.

2. Evaluate the ratio:

$$\alpha = \min\left\{1, \frac{\pi(Y)Q(Y, X_i)}{\pi(X_i)Q(X_i, Y)}\right\}.$$

3. We now do the following:

$$X_{i+1} = \begin{cases} X_i & \text{with probability } 1-\alpha, \\ Y & \text{with probability } \alpha. \end{cases}$$

Often you will hear the terminology of accept/reject associated with the last step, i.e. we accept the proposal with probability α and reject it otherwise. Let us denote by T the transition matrix of the Markov chain created by the MH algorithm. To transition from state i to j two things need to happen: we need to propose a move to j and the move needs to be accepted, i.e.:

$$T_{ij} = Q_{ij} \times \min\left\{1, \frac{\pi(j)Q_{ji}}{\pi(i)Q_{ij}}\right\},$$
$$T_{ii} = 1 - \sum_{j \neq i} T_{ij}.$$

Now, there are two fascinating things we need to observe: the algorithm is still constructing a Markov chain, and this Markov chain has limiting distribution π (we'll prove this in a second). More interestingly, we can run this algorithm even if we only know π upto its normalizer (see (6.2)).

Fact 6.3. The Metropolis-Hastings algorithm does not require full-knowledge of π, i.e. it is sufficient if we only know that:

$$\pi(X = x_j) \propto b_j.$$

The key point to observe is that the algorithm only uses ratios, i.e. things of the form $\pi(X)/\pi(Y)$ which we can easily evaluate, i.e.

$$\frac{\pi(X = x_j)}{\pi(X' = x_{j'})} = \frac{b_j}{b_{j'}}.$$

Fact 6.4. The limiting/stationary distribution of the Markov chain constructed by the Metropolis-Hastings algorithm is π.

We see that by our construction of Q we have ensured that the Markov chain is irreducible and aperiodic so the stationary distribution is unique and equal to the limiting distribution. Lets try to verify the detailed balance conditions to obtain a stationary distribution. We need to check that for any pair of states (i, j):

$$T_{ij}\pi(i) = T_{ji}\pi(j).$$

We see that if $i = j$ then the condition is trivially satisfied. Otherwise,

$$T_{ij}\pi(i) = \pi(i)Q_{ij} \times \min\left\{1, \frac{\pi(j)Q_{ji}}{\pi(i)Q_{ij}}\right\}, \quad \text{and,}$$
$$T_{ji}\pi(j) = \pi(j)Q_{ji} \times \min\left\{1, \frac{\pi(i)Q_{ij}}{\pi(j)Q_{ji}}\right\}.$$

Now there are two cases, if $\pi(j)Q_{ji} \geq \pi(i)Q_{ij}$ then we have that both of these expressions are equal to $\pi(i)Q_{ij}$ and if $\pi(j)Q_{ji} \leq \pi(i)Q_{ij}$ then both expressions are equal to $\pi(j)Q_{ji}$.

So we see that the detailed balance conditions are satisfied and we conclude that π is the limiting distribution of the Markov chain constructed by the MH algorithm.

Lets re-visit the permutation example. We described the matrix Q earlier, essentially:

$$Q_{ij} = \frac{1}{N(i)} \quad \text{if } j \text{ is a neighbor of i,}$$

where $N(i)$ is the number of neighbours of the permutation i. So now to complete the description of the MH algorithm we simply note that since our target is the *uniform distribution* over the set of permutations satisfying the condition (6.1) we have that,

$$\alpha = \min\left\{1, \frac{N(j)}{N(i)}\right\},$$

so the algorithm is quite simple, we start at some permutation that satisfies the constraint and propose a neighboring permutation. If that neighbor has more neighbors, then $\alpha = 1$ and we accept the proposal, otherwise we accept it with probability $N(j)/N(i)$. This simple procedure has limiting distribution equal to the uniform distribution on the set of permutations satisfying the constraint!

6.4 Gibbs Sampling

A very popular variant of the MH algorithm arises when we are sampling multi-dimensional random vectors (both the permutation example and the Ising model example are examples of this). Gibbs sampling is just the MH algorithm with a very particular proposal distribution.

The idea in Gibbs sampling is that we start in some state $(x_1, \ldots, x_n)$ and run the following algorithm:

1. Choose a coordinate to update, uniformly at random, i.e. with probability $1/n$ we select co-ordinate j.

2. Update its value by sampling from the conditional distribution keeping all other coordinates fixed, i.e. we sample a new value for x_j:

$$x_j \sim \pi(x_j|(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n)).$$

We repeat these steps. As before there are two miracles: (1) we can often sample from the conditional distribution even when we cannot normalize the distribution (2) the Gibbs sampling algorithm constructs a Markov chain with stationary distribution equal to π.

Gibbs sampling does not make much sense for the permutation example so lets understand how it works for the Ising model.

Fact 6.5. Often sampling from the conditional distribution $\pi(x_j|(x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_n))$ is easy.

This is difficult to make precise but lets see what happens in the Ising model:

$$\begin{aligned}\pi(x_j=1|(x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_n)) &= \frac{\pi((x_1,\ldots,x_{j-1},1,x_{j+1},\ldots,x_n))}{\pi((x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_n))}\\ &= \frac{\pi((x_1,\ldots,x_{j-1},1,x_{j+1},\ldots,x_n))}{\pi((x_1,\ldots,x_{j-1},-1,x_{j+1},\ldots,x_n))+\pi((x_1,\ldots,x_{j-1},+1,x_{j+1},\ldots,x_n))}\\ &= \frac{\exp(-\beta E(x_1,\ldots,x_{j-1},1,x_{j+1},\ldots,x_n))}{\exp(-\beta E(x_1,\ldots,x_{j-1},1,x_{j+1},\ldots,x_n))+\exp(-\beta E(x_1,\ldots,x_{j-1},-1,x_{j+1},\ldots,x_n))},\end{aligned}$$

and this is a very easy distribution to sample from.

In general, a conditional distribution will be a ratio of joint distributions and so the normalizing constant will again cancel (as it did in the MH algorithm).

Fact 6.6. The limiting distribution of the Gibbs Markov chain is π.

We can see that as before the Gibbs algorithm is constructing a Markov chain and we need to figure out what its transition matrix is. It is again easy to check that in most interesting cases the constructed chain will be irreducible and aperiodic (try to do this for the Ising model as an exercise) so we just need to check detailed balance conditions.

We will do this by showing that Gibbs is a special case of the MH algorithm. Notice that there are no accept/reject steps in Gibbs but first what is the proposal distribution? We propose to move from $(x_1,\ldots,x_n)$ to $(x'_1,\ldots,x'_n)$ if they differ in only one co-ordinate. Lets denote this coordinate j (chosen with probability $1/n$) then the proposal happens with probability:

$$Q((x_1,\ldots,x_n),(x'_1,\ldots,x'_n)) = \frac{1}{n}\pi(x'_j|x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_n),$$

and similarly we can calculate:

$$Q((x'_1,\ldots,x'_n),(x_1,\ldots,x_n)) = \frac{1}{n}\pi(x_j|x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_n).$$

Now if this were really an MH algorithm, we should accept the proposal with probability:

$$\begin{aligned}\alpha &= \min\left\{1, \frac{\pi((x'_1,\ldots,x'_n))\times Q((x'_1,\ldots,x'_n),(x_1,\ldots,x_n))}{\pi((x_1,\ldots,x_n))\times Q((x_1,\ldots,x_n),(x'_1,\ldots,x'_n))}\right\}\\ &= \min\left\{1, \frac{\pi((x'_1,\ldots,x'_n))\times \pi(x_j|x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_n)}{\pi((x_1,\ldots,x_n))\times \pi(x'_j|x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_n)}\right\}.\end{aligned}$$

Now notice that,

$$\pi(x_j|x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_n) = \frac{\pi((x_1,\ldots,x_n))}{\sum_x \pi((x_1,\ldots,x_{j-1},x,x_{j+1},\ldots,x_n))},$$

so we see that the ratio α is always equal to 1. So if we use the conditional distribution as a proposal distribution then we should always accept the new sample – which is exactly what Gibbs sampling does. So we conclude that Gibbs sampling has limiting distribution π.

Chapter 7

PageRank Algorithm

In this chapter we will discuss another important application of Markov chains, and some of the theory we have developed so far concerning the limiting behaviour of Markov chains.

7.1 Motivation and history

PageRank is an algorithm that was developed by Sergey Brin, Rajeev Motwani, Larry Page and Terry Winograd. The basic intuition was that one could try to measure the number and quality of links to a webpage, and use that to estimate how important the webpage is.

A good webpage is one that is linked to from many other good webpages.

Similar algorithms were proposed by various people around the same time. Notably, Jon Kleinberg suggested the HITS (Hyperlink-Induced Topic Search) algorithm for the same problem. Today, Google uses many other "signals" in webpage ranking but PageRank remains a very influential idea.

PageRank's first insight is relatively straightforward: we can think of the World-Wide-Web (WWW) as a *directed* graph. The nodes are webpages. Different webpages link to each other: these are the edges. For today we have n webpages numbered $\{1, \ldots, n\}$. We will attempt to assign each webpage a PageRank score. Denote these (suggestively) $\{\pi_1, \pi_2, \ldots, \pi_n\}$.

7.2 Some early attempts

Given our graph on the webpages, a first attempt would be to treat all webpages equally. We then rank each webpage by the number of pages that link to it. So the top webpage is just the webpage with the highest "in-degree", and so on.

Intuitively, this seems like a bad idea. More precisely, you could imagine that I can create a webpage and many sub-pages that all link to this webpage and this will result in the first page having a very high-ranking even though all the pages that linked to it are owned by me. More broadly, we want the ranking to be difficult to manipulate locally. As a side comment ideas like these of trying to artificially boost a webpage's ranking/reputation has led in recent years to many startups attempting "Search Engine Optimization", and on the other hand companies like Google constantly try to keep this in check.

The next attempt is recursive but is very close to the actual PageRank algorithm. We could weight each of the in-links to a webpage. Suppose we are trying to compute the PageRank score for a particular webpage i. Intuitively,

1. Webpages that link to a particular i, and have a high PageRank score, should be given more weight.

2. Webpages that link to a particular i, and have a low PageRank score, should be given less weight.

This is a bit circular, but the key point is that that is not a problem. We have been seeing many "circular" definitions recently. We will re-visit this in a bit.

Lets connect this back to Markov chains. The way to do this is to think of a so-called "random surfer", one who visits webpages by clicking links at random. The random surfer's current location is really just the state in a Markov chain. Define, the adjacency matrix for a directed graph as

$$A = \begin{bmatrix} A_{11} & A_{12} & \dots & A_{1n} \\ A_{21} & A_{22} & \dots & A_{2n} \\ \vdots & & & \\ A_{n1} & A_{n2} & \dots & A_{nn} \end{bmatrix},$$

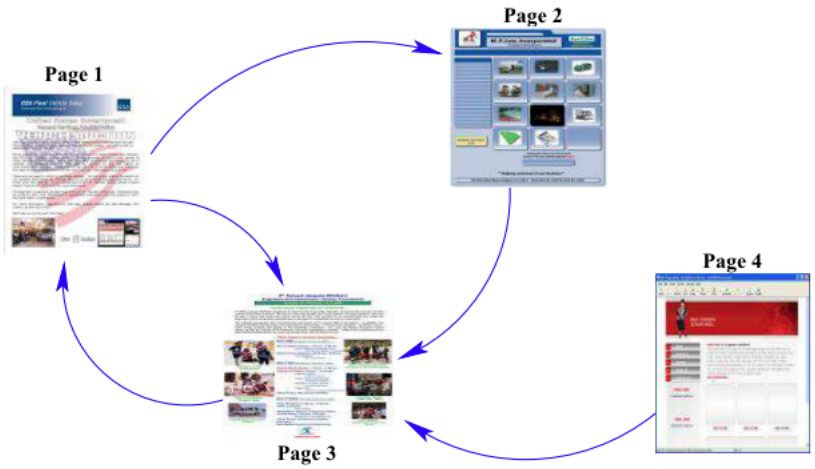
where $A_{ij} = 1$ if there is an edge or link from $i \to j$. Now, we want to turn this into a transition matrix for our random surfer. This is easy to do. We just define:

$$P = \begin{bmatrix} P_{11} & P_{12} & \dots & P_{1n} \\ P_{21} & P_{22} & \dots & P_{2n} \\ \vdots & & & \\ P_{n1} & P_{n2} & \dots & P_{nn} \end{bmatrix},$$

where

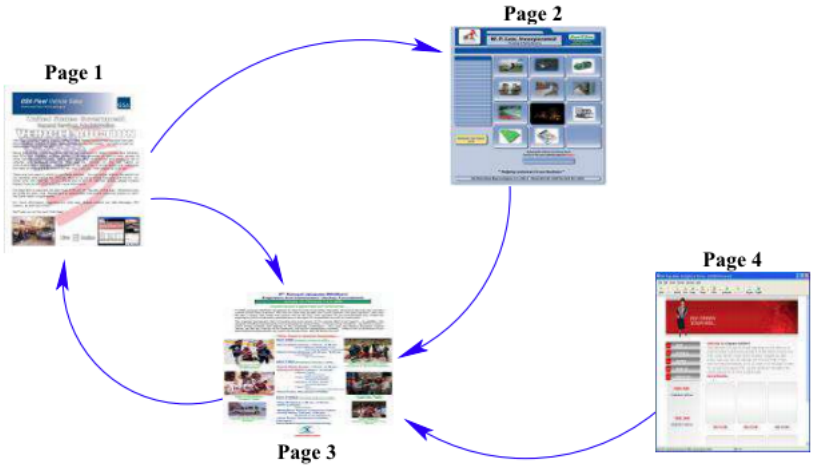
$$P_{ij} = \frac{A_{ij}}{\sum_{j=1}^{n} A_{ij}}.$$

Lets go through a simple example. Consider the following graph:



We can write down the "random surfer" transition matrix.

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



Using MATLAB, we can find the limiting distribution of the random surfer, and we obtain

$$\pi = [0.4, 0.2, 0.4, 0].$$

Is this intuitive? Observe the following things:

1. If there are no links in to a page it has a score of 0.
2. Page 3 has many in-links and gets a high-score.
3. Page 1 has only one in-link but it is from an "influential page" so it also gets a high-score.

So we jumped from some intuition about assigning "high score to webpages which are linked to from other pages with high score" to the random surfer model, and the limiting distribution. Lets take a step back. We know that for a stationary distribution π:

$$\pi^T = \pi^T P.$$

So in particular the PageRank score for a page i satisfies:

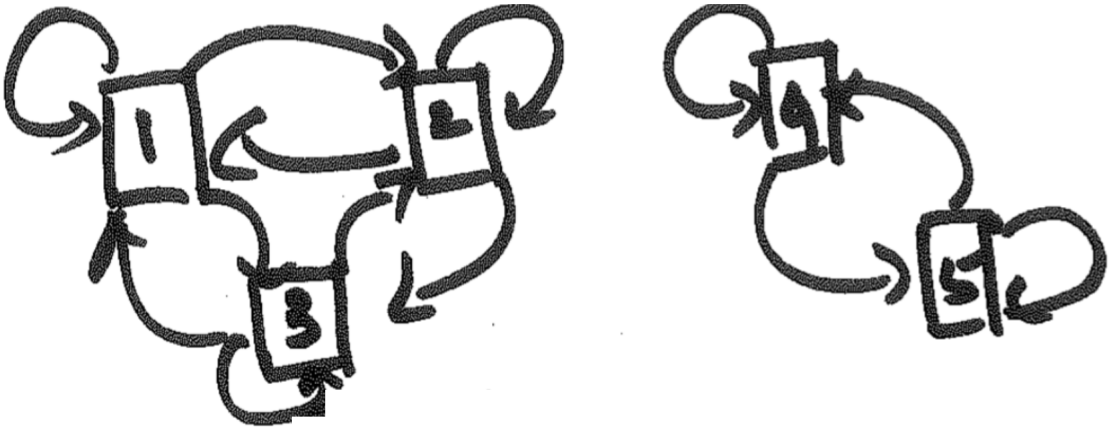
$$\pi_i = \sum_{j=1}^{n} \pi_j P_{ji} = \sum_{j=1}^{n} \frac{\pi_j A_{ji}}{d_j}.$$

Is this intuitive? What about the normalization by degrees? This effectively controls the "total vote" of each page.

So we have made some progress. We have formulated our desired characteristics, into things that are satisfied by the limiting distribution of a particular Markov chain. Now, we just need to make sure that the limiting distribution exists. There are a few problems we might run into.

Well we know that for finite Markov chains there is always at least one stationary distribution. Why not use one of those (it may not be unique)?

Here is a concrete example of what can go wrong. Suppose we have the following web graph:



So the random surfer matrix has two stationary distributions: [1/3 1/3 1/3 0 0] and [0 0 0 1/3 1/3]. If we interpret these entries as PageRank scores then the two ranking are completely at odds with each other. Intuitively, what is happening is that the ranking of the webpages depends on where the random surfer starts.

7.3 The real PageRank algorithm

The real PageRank algorithm makes a slight modification. Suppose that our random surfer occasionally gets bored and clicks on a random webpage.

Our bored random surfer also follows a Markov chain, with transition matrix:

$$Q = (1-\theta)\begin{bmatrix} \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \\ \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \\ \vdots & & & \\ \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \end{bmatrix} + \theta P.$$

We need to set the parameter θ, and need to do so carefully. What can go wrong if we pick a value that is too small? Too large?

Google uses the value $\theta = 0.85$. This simple modification is incredibly important. We have now created a Markov chain with a unique stationary distribution, irrespective of the underlying graph.

We now have all the pieces in place. We can represent the WWW as a (very large) matrix and compute its unique limiting distribution. We'll discuss this in a bit more detail.

7.4 Computing PageRank scores

The real contribution of the PageRank paper was to show that one could compute the PageRank scores on huge graphs. On a small graph we could use MATLAB.

Revisiting our earlier example we would run the following commands:

```
» P = [0 1/2 1/2 0; 0 0 1 0; 1 0 0 0; 0 0 1 0];
» Q = (1-0.85)*ones(4,4)/4 + 0.85*P;
» Q^1000
ans =
0.3725 0.1958 0.3941 0.0375
0.3725 0.1958 0.3941 0.0375
0.3725 0.1958 0.3941 0.0375
0.3725 0.1958 0.3941 0.0375
```

This is somewhat inelegant. The other way of doing this involves observing that the stationary distribution is just the top eigenvector of the transition matrix. Think this through.

So we could use the following instead:

```
» [V,D] = eig(Q');
» V(:,1)/sum(V(:,1))
ans =
0.3725
0.1958
0.3941
0.0375
```

Computing the top eigenvector of an $(n \times n)$ matrix typically takes time $O(n^3)$, which is very slow. One could instead use an iterative method called the power method, to compute the limiting distribution. This is really very similar to what we tried first (i.e. raising the transition matrix to some high power). It is called the "power method" of computing eigenvectors.

Essentially, we could use the following iterative algorithm:

1. Start with some initial distribution π_0.
2. Compute iteratively: $\pi^{t+1} = \pi^t Q$ until the vector stabilizes.

So, in MATLAB we would use:

```
» p = ones(1,4)/4;
» for i = 1:1000
p = p*Q;
end
» p
p =
0.3725 0.1958 0.3941 0.0375
```

The PageRank paper showed that they could compute PageRank scores on web-scale using the power method. An important observation that they made was that one can compute these very fast when the transition matrix (without the "bored" part) is sparse.

7.5 Summary

It is worthwhile to think through how one would combine PageRank scores (which really ignore the content of the webpage) with things like TF-IDF (term frequency-inverse document frequency) that use the content of the webpage (and the search query).

We saw how to cast the problem of ranking webpages as that of finding the limiting distribution of a Markov chain. We saw how to use insights about the existence and uniqueness of limiting distributions to "fix" the Markov chain, and finally we saw how one might compute PageRank scores at web-scale.

In the years since PageRank was proposed, several nice variants have also been proposed to improve the basic algorithm (check out the Wikipedia page for PageRank to find some references).

Chapter 8

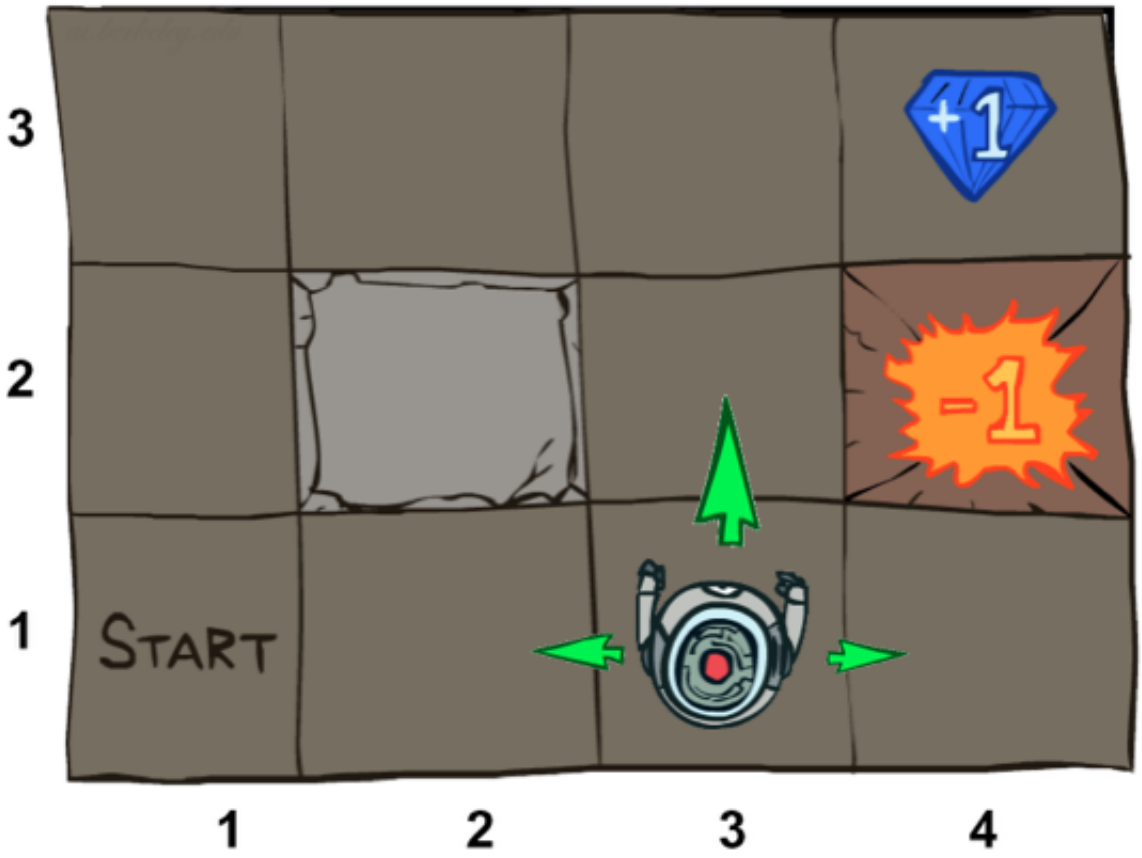
Markov Decision Processes

Today's material contains lots of pictures from Dan Klein and Pieter Abbeel's CS 188 course at Berkeley. You can find videos of their lectures online and I recommend watching them if possible.

8.1 Introduction

Lets begin by considering a simple Robot control problem, that is also known as stochastic search. For today's lecture our running example will be "GridWorld".

We have a robot placed in the following grid world:



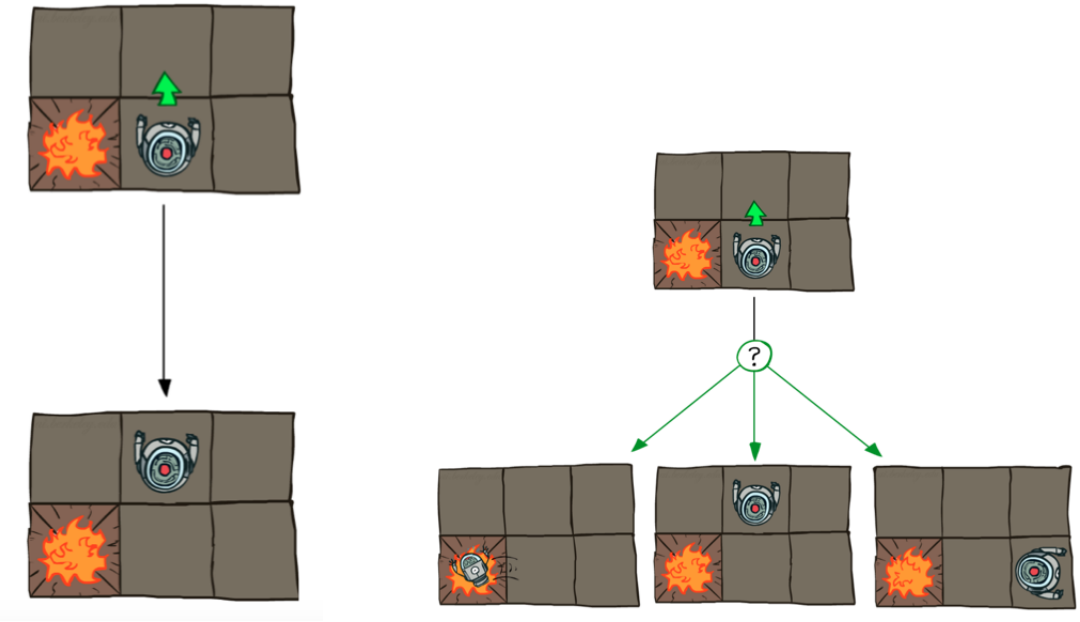
The grid world has the following "rules":

1. 80% of the time, trying to go somewhere takes you there (if there is no wall), and 10% you veer off to the the left and 10% you veer off to the right (if no wall).
2. If you hit the wall you stay put.
3. If you hit the big reward states you exit.
4. "Living" reward at every step, this can be negative.

The broad goal in grid world is that we want to

maximize the sum of rewards.

Why is this in a probability modeling course?



(a) Deterministic grid world (b) Stochastic grid world

Some notation:

1. **States**: $\mathbb{S} \in \{0, 1, \ldots, M\}$.
2. **Actions**: $a \in A$.
3. **Dynamics**: $P_{ij}(a) = P(X_{n+1} = j | X_n = i, A_n = a)$.
4. **Rewards**: $R(i, a, j)$.

If you ignore the rewards then basically you have a Markov chain for every possible action.

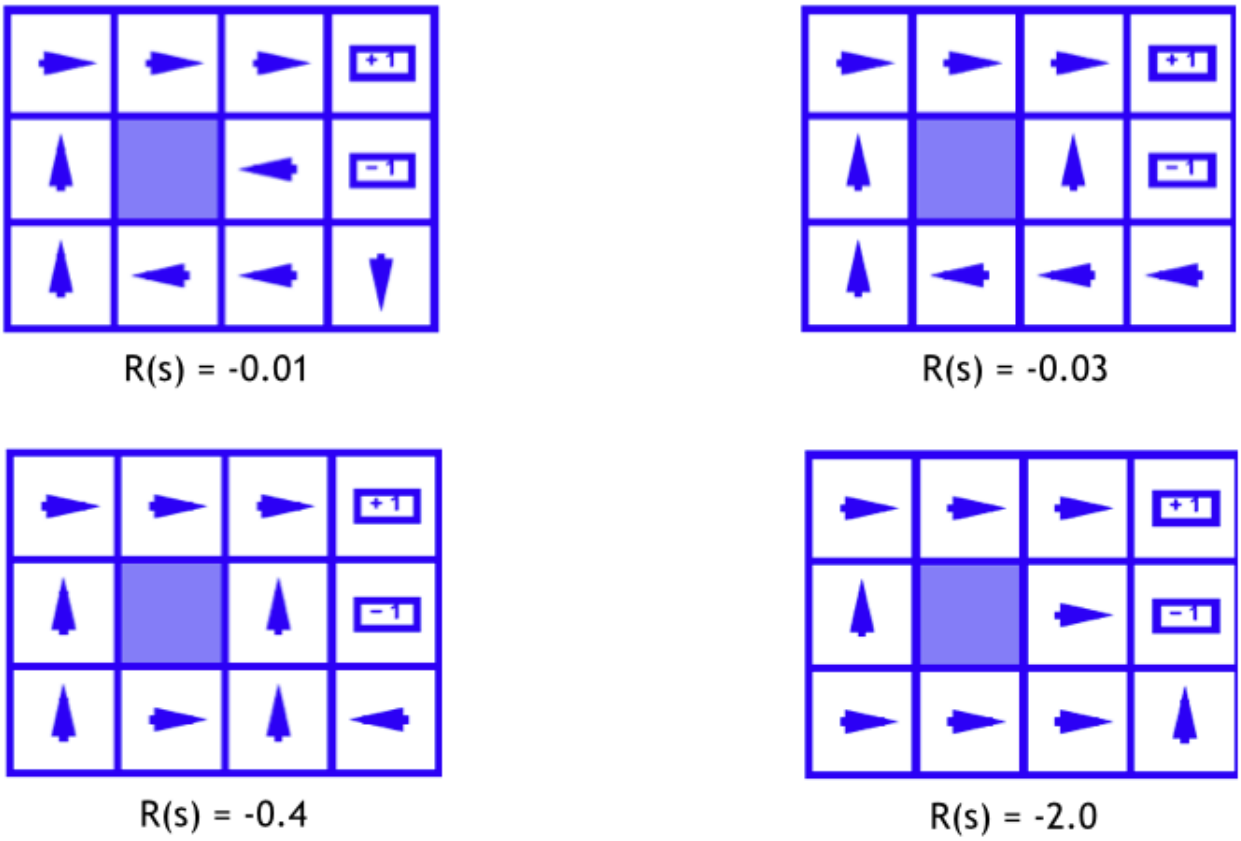
What is Markov about MDPs?

What is the goal?

- The goal is to find a good policy (one that maximizes total reward).
- A policy $\pi : \mathbb{S} \to A$.

8.1.1 Living Rewards and Optimal Policies

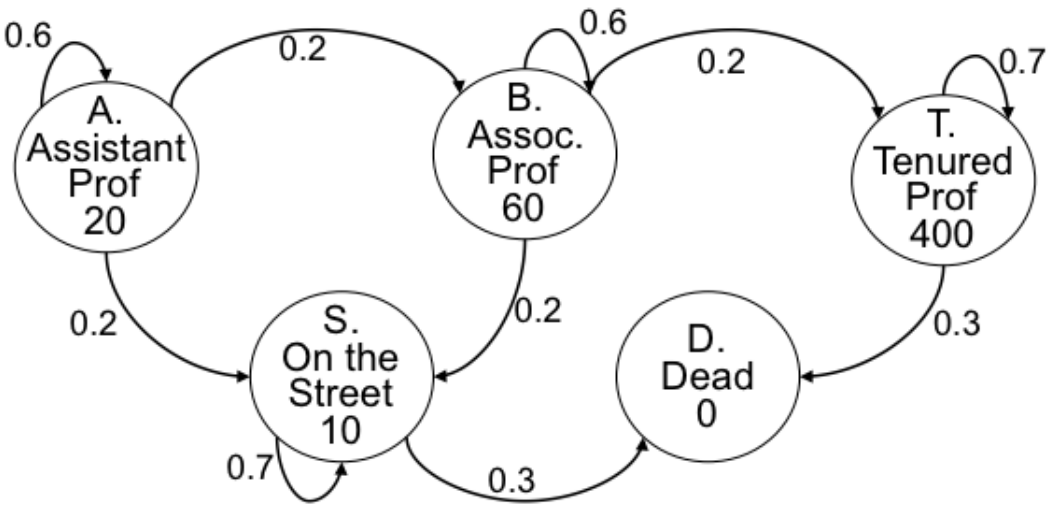
The "living" reward, i.e. the amount of money we get for staying alive can alter the optimal policy in interesting ways.



What you are seeing is that when the living reward is a small negative number the optimal policy is extremely "risk-averse", when the living reward gets larger (at -0.4) the optimal policy is to take the shortest path to the reward (makes sense), and if it gets even more negative (-2) the agent becomes suicidal.

8.2 Discounted Rewards

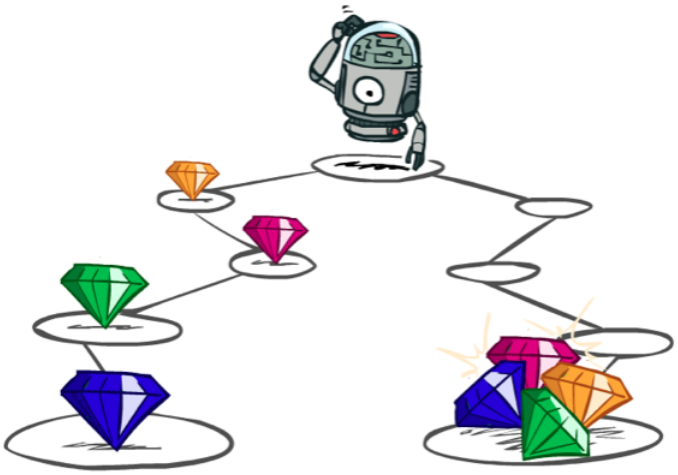
Lets think of another MDP. This is my life. Here $R(i, a, j) = R(j)$.



- There is a somewhat annoying feature in MDPs: if there is an "exit" state then long-term average reward is 0.
- An alternative is to look at *total reward*: but in control settings this can be equally meaningless, we find a state with small positive reward we can stay there forever. This policy and optimal policy will both have infinite total reward.
- Another alternative is to use *finite* time horizons, i.e. declare in advance the game ends after T rounds. This creates non-stationary policies (i.e. π depends on amount of time left) which is a major annoyance.
- The most appealing alternative, and one that we will use for the rest of this lecture, is to use *discounting*, i.e. when evaluating a policy future rewards are discounted by a factor $0 < \gamma < 1$.

  If a policy gets us rewards $\{r_0, r_1, \ldots, r_\infty\}$ our *utility* is

$$U = \sum_{t=0}^{\infty} \gamma^t r_t.$$



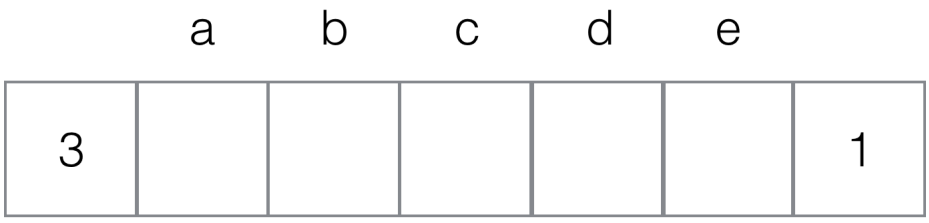
- Investment intuition, i.e. if you imagine every time you collect a reward you store it in a bank and it accumulates interest then you would prefer earlier rewards.

- If you force *stationary preferences* then essentially the only thing that makes sense is multiplicative (or exponential) discounting.

- Discount factor can affect optimal policy.

- Discounting rewards can help algorithms converge faster.

Example: Here is an MDP:



- For $\gamma = 1$ what is the optimal policy?

- For $\gamma = 0.1$ what is the optimal policy?

- For which γ are left and right equally good if you start in state d?

8.3 Value iteration

In the Ross book there is a section on MDPs where he shows that under certain assumptions one can compute an optimal policy via linear programming. This is not a very practical algorithm since solving a (big) linear program can be slow, and it is difficult to incorporate learning into this algorithm. We might return to this and other methods to find an optimal policy in the next lecture. Today we will study a more direct method for computing an optimal policy via dynamic programming.

Define:

1. **Value of a state s:** We will denote this $V^*(s)$. This is the expected (discounted) reward if you start in state s and act optimally.

2. **Value of a q-state (s, a):** We will denote this $Q^*(s, a)$. This is the expected (discounted) reward if you start in s, take action a and then act optimally from then onwards.

3. **The optimal policy:** $\pi^*(s)$ is the optimal action from state s. Note regarding randomized versus pure policies.

These satisfy some important relationships:

$$V^*(s) = \max_a Q^*(s,a)$$

$$Q^*(s,a) = \sum_{s'} \left[R(s,a,s') + \gamma V^*(s')\right] \times P_{ss'}(a) \quad \text{this is just the law of total expectation}$$

$$V^*(s) = \max_a \sum_{s'} \left[R(s,a,s') + \gamma V^*(s')\right] \times P_{ss'}(a)$$

$$\pi^*(s) = \arg\max_a Q^*(s,a).$$

These equations are known as the *Bellman equations*. We are interested in computing (some subset of) these quantities. If these were linear equations we might just re-arrange them and solve them, however the Bellman equations are a system of non-linear equations.

Key idea: Interpreting the Bellman equations as updates, instead of equations. We have seen this before in the context of the power method for PageRank.

This is called *value iteration*. We initialize the value of every state $V_0(s) = 0$, and then update the values of each state:

$$V_{k+1}(s) \leftarrow \max_a \left[\sum_{s'} \left[\left[R(s,a,s') + \gamma V_k(s')\right] \times P_{ss'}(a)\right]\right].$$

We iterate this update until convergence (i.e. until the change in the values for all the states is very small).

- What is the complexity of each iteration? For each state the complexity is $\mathcal{O}(M|A|)$, and so the overall complexity of each iteration is $\mathcal{O}(M^2|A|)$.
- Can prove that this iteration will converge to the (unique) optimal values.
- Empirically at least policies may converge long before the values converge.

8.4 Policy iteration

What is bad about value iteration? The main problem is that it can be very slow to converge, i.e. finding the exact values of each state may take a long time. Often we do not care about the optimal values, we just want to find a good policy.

8.4.1 Policy evaluation

Suppose that I fixed a policy π and I want to know how good it is. First, let us define the utility of a state s, under a policy π:

$$V^\pi(s) = \sum_{s'} \left[R(s,\pi(s),s') + \gamma V^\pi(s)\right] \times P_{ss'}(\pi(s))$$

Again, recursive set of equations, but now a linear system.

- We can solve this via an iterative method: again we start with $V_0(s) = 0$ for each state and update:

$$V^\pi_{k+1}(s) = \sum_{s'} \left[R(s,\pi(s),s') + \gamma V^\pi_k(s)\right] \times P_{ss'}(\pi(s)),$$

  until convergence.

- Or we can just use a linear system solver: we can rewrite the above system of equations in the usual way, i.e. for some (A, b) we have that,

$$V^\pi = AV^\pi + b,$$

  which we can easily solve.

Intuitively, which do you expect will be faster? The answer turns out to be complex but in general, if γ is small the iterative algorithm will converge faster and when γ is close to 1 then the linear system solver will be much faster.

8.4.2 Policy extraction

We have seen how to get values from a policy. How does one get a policy from values?

Suppose we have the optimal values $V^*(s)$ for every state s. We could try to do the following:

$$\pi^*(s) = \arg\max_a \left[\sum_{s'} \left[R(s, a, s') + \gamma V^*(s') \right] \times P_{ss'}(a) \right].$$

This is called policy extraction. It is trying to extract a policy from the values by looking ahead one-step. If you prefer this is just saying:

$$\pi^*(s) = \arg\max_a Q^*(s, a).$$

What happens if we do policy extraction from non-optimal values? This leads to an improved policy, and this is called the policy iteration algorithm.

8.4.3 Policy iteration

As a quick recap: we are still trying to find a good policy. Value iteration has a few drawbacks:

- It can be slow.
- Policy can converge long before values, so max rarely changes.

Policy iteration instead, starts with a policy π_0 and alternates the following two steps:

- **Policy Evaluation:**

$$V_{k+1}^{\pi_i}(s) \leftarrow \sum_{s'} \left[R(s, \pi_i(s), s') + \gamma V_k^{\pi_i}(s') \right] \times P_{ss'}(\pi_i(s)).$$

  We iterate this until convergence.

- **Policy Extraction/Improvement:** Now, we extract a new policy by one-step lookahead, i.e.:

$$\pi_{i+1}(s) = \arg\max_a \left[\sum_{s'} \left[R(s, a, s') + \gamma V^{\pi_i}(s') \right] \times P_{ss'}(a) \right].$$

We know when we are done. It is when the policy stops changing, i.e. $\pi_{i+1} = \pi_i$.

To put them side-by-side:

1. **Value Iteration:**
    - Every iteration updates values (and implicitly the policy).
    - We do not track the policy explicitly.
2. **Policy Iteration:**
    - Several passes to update values with fixed policy (much faster than value iterations): $\mathcal{O}(M^2)$.
    - After evaluation, we choose a new policy which is either better or we are done.

Both have guaranteed convergence to the optimal values and policy, and both are dynamic programs.

Chapter 9

Estimating Markov Chains

In this chapter we will consider the problem of estimating a Markov chain from data, and how to assess the fit of a Markov chain to data.

9.1 The principle of maximum likelihood

Someone comes up to you and says, "I have this coin. If I toss it what is the probability it comes up heads".

In all likelihood (pun intended), you would say "Toss it a few times." She does, and it turns out that the observed sequence of tosses is $HTHTT$. You might then respond: I think P(Heads) = 2/5.

The typical way to justify this response is via the principle of maximum likelihood. First, a definition. The likelihood, is a function of the parameters (what we want to estimate), that is formed after we see the data (i.e. the thing we observe).

Suppose we denote the probability of heads as p, then

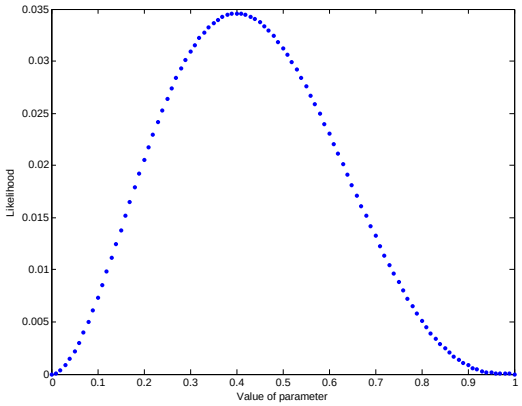
$$\begin{aligned}\mathcal{L}(\text{Sequence} \mid p) &= p \times (1-p) \times p \times (1-p) \times (1-p) \\ &= p^{\#\text{heads}} \times (1-p)^{\#\text{tails}}.\end{aligned}$$

Reiterating: The likelihood function, for every value of the parameters returns a number. We would like to pick a single parameter.

The principle of maximum likelihood is then just:

Choose the parameters that maximize the likelihood.

First, lets look at a plot:



So, how do I maximize this function? One (not ideal) way is to use MATLAB.

```
    p=[0:0.01:1];
for i = 1:length(p)
l(i) = (p(i)^2)*(1-p(i))^3;
end
[ ~,ind]=max(l);
» p(ind)
ans =
0.4000
```

The other way is to use calculus. First, to simplify things observe that the value of p that maximizes the likelihood is the same as the value of p that maximizes the log-likelihood.

We can try to maximize:

$$\mathcal{LL}(\text{Sequence} \mid p) = \#\text{heads} \log p + \#\text{tails} \log(1-p).$$

So we take the derivative and set it equal to 0.

$$\#\text{heads}\frac{1}{p} - \#\text{tails}\frac{1}{1-p} = 0,$$
$$\widehat{p} = \frac{\#\text{heads}}{\#\text{tails} + \#\text{heads}}.$$

To summarize, to estimate the parameter of a Bernoulli, we just take the ratio of # of successes over total # of trials.

A simple generalization of this is that to fit a multinomial, the maximum likelihood estimate would again just take the ratio of the observed count to the total count. For example, if you role a dice 100 times and "3" comes up 16 times you would estimate the probability of seeing a "3" as 16/100.

Intuitively, the principle of maximum likelihood is a natural "rule" to estimate parameters is to pick the ones that are most likely to have generated the data you have already seen. This is quite a general, core statistical concept.

Lets take a look at an example that we will use for the rest of the class:

Example 1: Alofi is a city in Niue (a country near New Zealand). Rainfall was recorded each day from January 1st, 1987 to December 31st, 1989: a total of 1096 measurements. The data was converted into three states: "1" indicates no rainfall, "2" indicates a positive amount of rainfall, but less than 5mm, "3" indicates more than 5mm of rainfall.

Here is the complete data.

```
   [1] 3 2 2 2 2 2 2 3 3 3 2 1 1 1 2 3 3 2 1 1 3 2 2 1 3 3 2 1 2 1 1 3 3 2 2 1 1
  [38] 1 2 1 2 2 2 2 3 1 2 1 1 2 3 2 2 3 1 1 1 1 1 1 1 3 3 3 3 2 2 1 3 1 1 1 2 1
  [75] 2 2 1 3 1 3 1 1 1 1 2 1 2 3 1 1 1 1 3 3 3 2 2 1 1 1 1 2 1 1 1 1 2 1 2 1 1
 [112] 1 3 2 1 1 1 2 1 2 1 2 2 2 1 2 1 3 2 2 2 2 3 3 3 3 1 1 2 3 2 1 1 1 1 1 2 2
 [149] 2 3 1 2 2 2 2 1 2 1 1 2 1 1 1 1 1 2 1 1 1 1 3 3 2 1 1 1 1 1 2 1 1 2 1 1 1
 [186] 2 1 1 1 2 1 1 1 1 3 2 2 1 1 1 1 1 1 1 1 1 1 2 2 2 1 1 1 3 3 2 1 2 1 1 1 1
 [223] 1 1 3 2 1 1 1 1 1 2 1 1 2 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [260] 1 1 1 1 1 1 2 2 1 1 1 2 3 2 1 3 1 1 1 1 1 3 1 1 1 1 1 1 1 1 3 1 1 1 1 1 1
 [297] 1 1 2 3 2 1 1 1 2 1 1 1 1 2 1 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 3 1 1 1 2 2
 [334] 2 2 3 3 3 3 1 1 1 1 1 1 1 2 3 3 2 1 1 2 1 2 2 1 3 3 3 3 3 2 3 1 1 1 1 1 1
 [371] 2 1 1 1 1 1 2 1 1 3 3 2 1 1 2 1 1 1 1 3 2 1 2 2 2 2 3 2 1 3 1 2 2 1 3 2 1
 [408] 1 1 1 2 3 3 2 2 1 2 3 1 1 2 2 3 3 1 1 1 1 3 3 2 1 1 2 1 2 1 1 2 2 1 1 1 1
 [445] 2 1 2 1 1 1 1 2 1 3 3 3 2 3 3 3 3 3 3 3 1 2 1 1 1 1 2 2 2 2 1 1 1 1 1 3 3
 [482] 3 2 1 1 3 3 3 2 2 1 3 1 3 1 1 2 1 1 1 1 1 1 3 2 2 1 2 2 3 3 3 3 3 3 3 2 2
 [519] 3 2 2 3 3 2 1 1 2 2 1 1 1 2 1 1 2 2 1 1 1 1 1 1 1 2 3 1 3 2 2 3 2 2 2 2 2
 [556] 2 2 2 1 1 2 1 3 3 3 3 2 2 1 1 1 2 2 3 1 1 1 1 1 2 3 3 3 2 1 1 1 1 2 1 1 1
 [593] 1 1 2 2 2 3 3 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 2 1 3 3 3 3 1 1 1 1 1 2 1 1
 [630] 1 1 1 3 3 2 3 2 3 1 3 3 3 3 3 2 3 3 3 1 1 1 3 3 1 1 1 1 3 3 2 2 2 3 1 2 2
 [667] 3 1 1 2 2 3 2 1 3 3 2 2 3 3 2 1 3 2 2 2 3 3 2 1 2 3 3 3 2 3 1 3 2 1 3 1 1
 [704] 1 3 3 1 2 2 2 2 3 2 1 1 3 3 2 1 1 3 2 1 1 2 3 2 1 1 2 3 3 3 3 3 3 3 3 3 3
 [741] 3 3 2 3 3 1 3 2 1 1 2 3 2 2 2 1 3 1 2 1 1 1 1 3 3 2 3 3 3 3 2 1 2 3 3 3 3
 [778] 3 3 3 3 2 3 3 2 1 1 2 1 1 2 1 1 1 1 2 1 1 2 2 1 1 1 2 1 2 3 3 2 3 3 2 2 2
 [815] 1 1 2 2 3 3 2 3 2 3 2 3 1 2 3 3 2 2 2 3 3 3 2 1 3 3 2 3 3 1 3 3 3 1 1 2 2
 [852] 2 3 3 2 3 3 2 2 3 3 2 1 1 1 2 1 1 1 1 1 1 1 1 1 1 2 3 1 2 1 1 1 2 3 1 2
 [889] 2 1 2 2 2 3 3 1 2 1 3 3 1 1 1 2 1 1 1 1 2 1 2 1 2 3 2 1 1 1 1 1 1 3 1 1 2
 [926] 1 1 1 2 1 1 1 1 2 1 2 1 1 1 1 2 2 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 2 1 1 1
 [963] 1 1 2 2 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 2 1 1 1 1 2 2 1 1 1 2 1 1 1 1 1 3
[1000] 2 1 2 1 1 1 1 1 2 1 1 2 3 1 1 2 1 1 1 2 1 2 1 3 2 3 2 3 3 3 2 3 3 2 1 2 1
[1037] 1 1 1 1 3 3 1 2 2 3 2 1 1 2 3 3 1 1 2 1 2 3 1 2 1 3 3 3 3 3 3 1 3 3 1 1 1
[1074] 1 1 2 2 1 3 2 1 1 1 1 3 3 1 1 1 1 1 2 1 3 3 2
```

Our first attempt is to fit a model assuming independence. How would we proceed?

We could first compute some statistics from the data:

pattern	count
1	548
2	294
3	254

So our estimates would be:

$$\widehat{p}_1 = \frac{548}{1096}$$
$$\widehat{p}_2 = \frac{294}{1096}$$
$$\widehat{p}_3 = \frac{254}{1096}$$

9.2 Fitting A Markov Chain

Another possibility is to fit a Markov chain. To describe a Markov chain on the state space $\mathbb{S} = \{1, 2, 3\}$ we need to estimate a transition matrix $\mathbf{P}$, and possibly an initial distribution.

Note: It should be clear we cannot really estimate the initial distribution in this setting. We can give it a deterministic value – it will not matter since the chain is irreducible and aperiodic anyway. Let us now focus on estimating the transition matrix.

Now, we need to estimate conditional probabilities. We again do this via maximum likelihood. Let us roughly try to see what the likelihood is. We will continue to ignore $\pi^{(0)}$. The beginning of our sequence was: 322222233321 . . ., so we could write

$$\begin{aligned}\mathcal{L}(\text{Sequence}|\mathbf{P}) &= P_{32}P_{22}P_{22}P_{22}\dots \\ &= \prod_{(i,j)} P_{ij}^{\#i\to j},\end{aligned}$$

i.e. we simply raise each entry of the transition matrix to the power of the number of times we saw that particular transition in our sequence. We can then try to maximize this as a function of the transition matrix $\mathbf{P}$. Some calculus will then show that the maximum likelihood estimate

$$\widehat{P}_{ij} = \frac{\#i \to j}{\#i}.$$

In order to proceed, we could try to count the number of occurrences of the various pairs of states. To make this convenient here is a table

pattern	1	2	3	Total
1	362	126	60	548
2	136	90	68	294
3	50	79	124	253

So in this case we would estimate the entries of the transition matrix as:

$$\widehat{P} = \begin{bmatrix} \frac{362}{548} & \frac{126}{548} & \frac{60}{548} \\ \frac{136}{294} & \frac{90}{294} & \frac{68}{294} \\ \frac{50}{253} & \frac{79}{253} & \frac{124}{253} \end{bmatrix}.$$

9.3 Assessing the fit

So far, we looked at some data, we fit a multinomial model assuming the data were independent, and then we fit a Markov chain assuming the data had a first-order dependence structure.

Intuitively, maybe we believe that the weather has more of a Markov structure than an independence structure (i.e. for instance we expect to see many rainy days in a row etc.).

One simple way to test this is to compare likelihoods:

1. **Likelihood of Independent Model:** $\mathcal{L} = \widehat{p}_1^{\#1} \times \widehat{p}_2^{\#2} \times \widehat{p}_3^{\#3} = \exp(-1138)$.
2. **Likelihood of Markov Chain:** $\mathcal{L} = \prod_{(i,j)} \widehat{P}_{ij}^{\#i \to j} \approx \exp(-1040)$.

The Markov Chain has higher likelihood but it is somewhat unclear how one should interpret a higher likelihood especially since these numbers are not scaled in a meaningful way. It is also not very clear if the Markov chain is a good fit or just a better fit than the independent model. An alternative is to use QQ-plots or goodness of fit tests.

Lets see QQ-plots in action. The basic fact you will need for this lecture is that

Basic Fact: For a multinomial we have that,

$$\frac{\text{Observed count} - \text{Expected count}}{\sqrt{\text{Expected count}}} \approx N(0,1).$$

A QQ-plot is a quantile-quantile plot, i.e. if I have a bunch of samples from a distribution that I believe is a standard Gaussian then the quantiles of the samples must be roughly the same as the quantiles of a standard Gaussian, i.e. if a standard Gaussian puts 2.5% of its mass above 1.96, then so should the samples. We can plot the quantiles of a standard Gaussian versus the samples and if they are close in distribution this should be roughly a plot with slope 1.

How do we use this in our case? Lets re-visit the table:

pattern	1	2	3	Total
1	362	126	60	548
2	136	90	68	294
3	50	79	124	253

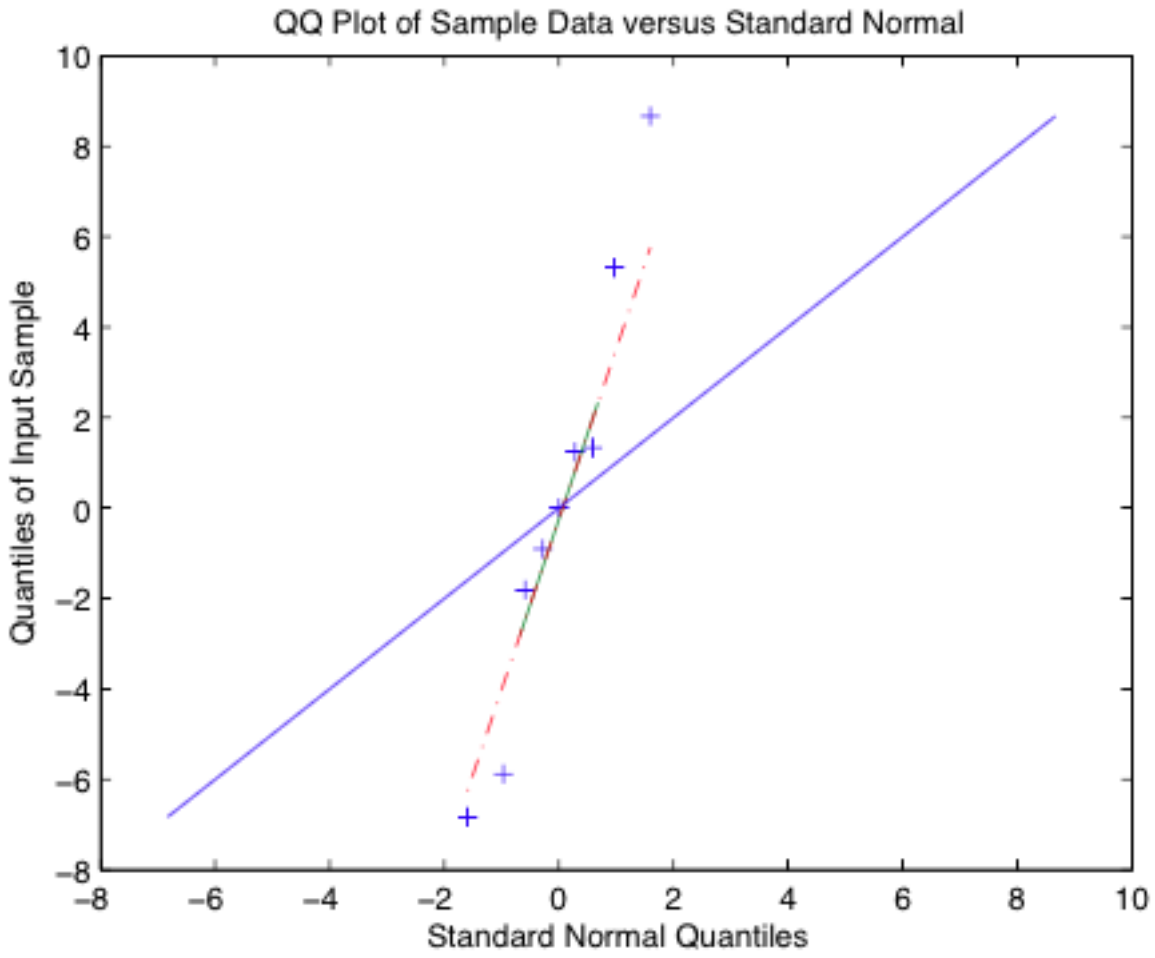
What should the expected counts in each cell here be if the independence assumption was correct? We would expect to see the pattern 11 for instance $1096 \times \widehat{p}_1^2$ times. So the expected counts can be arranged into a table:

$$E = 1096 \times \begin{bmatrix} \widehat{p}_1^2 & \widehat{p}_1\widehat{p}_2 & \widehat{p}_1\widehat{p}_3 \\ \widehat{p}_1\widehat{p}_2 & \widehat{p}_2^2 & \widehat{p}_2\widehat{p}_3 \\ \widehat{p}_1\widehat{p}_3 & \widehat{p}_2\widehat{p}_3 & \widehat{p}_3^2 \end{bmatrix}.$$

So we have the empirical counts (i.e. observed counts) and the expected counts so we can draw the QQ-plot:

In MATLAB you would just take the samples in a vector and use the command: `qqplot(x);`

Here is what happened when I did this to the rainfall data:



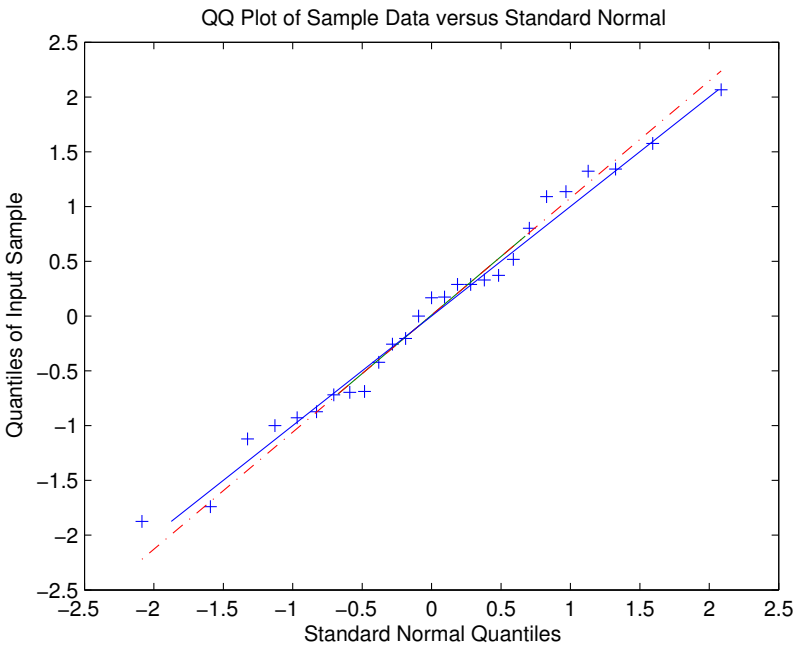
We only have the nine samples but the QQ-plot seems to deviate quite badly from the standard normal indicating that the independence assumption was likely too strong.

In order to repeat this style of analysis for the Markov model, I would need to compute length 3 histories. I have done this for you:

pattern	1	2	3	Total
11	247	86	29	362
21	86	27	23	136
31	29	13	8	50
12	70	32	24	126
22	29	35	26	90
32	37	23	18	78
13	13	16	31	60
23	17	17	34	68
33	20	45	59	124

As before we can calculate the expected counts, and use our **Basic fact** to compute a set of samples that we believe are normally distributed. How?

Again in this setting we can draw the QQ-plot:



This time we see the QQ-plot is quite close to linear and so we might conclude that a first-order Markov model is a good fit to the observed data.

Chapter 10

Poisson Processes

With an eye towards understanding continuous time Markov chains we will in this chapter define and study Poisson processes. Before we get to defining precisely what a Poisson process is we need to review two distributions that will appear repeatedly in the sequel:

1. The Exponential distribution: A *continuous* RV X has an exponential(λ) distribution, if it has pdf

$$f(x) = \lambda \exp(-\lambda x) \quad \text{for all} \quad x \geq 0.$$

2. The Poisson distribution: A *discrete* RV X has a Poisson(λ) distribution if its pmf is given by

$$P(X = k) = \exp(-\lambda)\frac{\lambda^k}{k!} \quad \text{for} \quad k = \{0, 1, 2, \ldots\}.$$

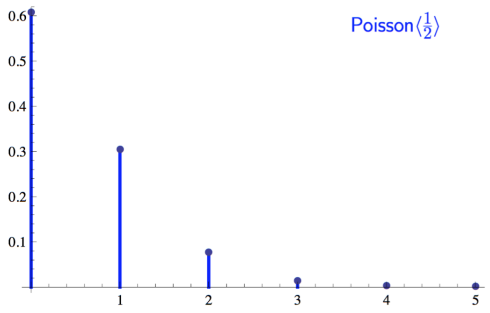
Basic Properties: The mean and variance of an exponential(λ) are:

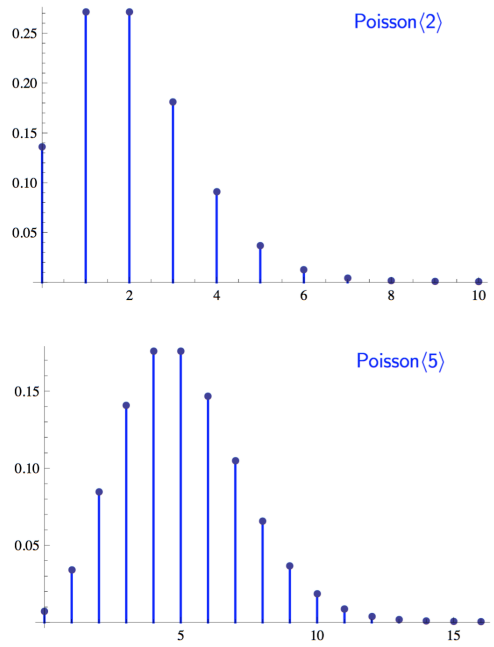
$$\mathbb{E}[X] = \frac{1}{\lambda} \qquad \text{Var}[X] = \frac{1}{\lambda^2}.$$

The mean and variance of a poisson(λ) are:

$$\mathbb{E}[X] = \lambda \qquad \text{Var}[X] = \lambda.$$

10.1 Poisson distribution





Fact 10.1. If X and Y are Poisson(μ) and Poisson(ν) respectively, then $X + Y$ has a Poisson($\mu + \nu$) distribution.

Fact 10.2. The Poisson distribution is often used to approximate Binomials, i.e.

$$\text{Bin}(n, p) \approx \text{Poisson}(np),$$

if p is very small. Some people call this, the law of rare events or the *law of small numbers.*

More concretely, suppose that X_n is Binomial with parameter $p = \lambda/n$, then for any $k \geq 0$,

$$\lim_{n\to\infty} P(X_n = k) = \frac{\exp(-\lambda)\lambda^k}{k!}.$$

To see this, we can write the Binomial probability as:

$$P(X_n = k) = \binom{n}{k} p^k (1-p)^{n-k},$$

and we note by Stirling's formula that

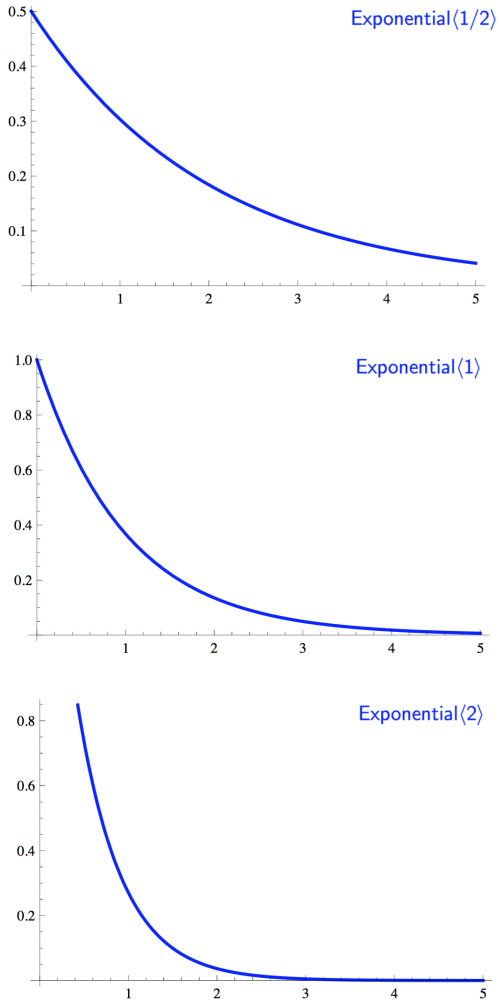
$$\binom{n}{k} \approx \frac{n^k}{k!},$$

so that,

$$\begin{aligned}
\mathbb{P}(X_n = k) &\approx \frac{n^k}{k!} \times \left(\frac{\lambda}{n}\right)^k \times \left(1 - \frac{\lambda}{n}\right)^{n-k} \\
&\approx \frac{\lambda^k}{k!} \exp\left[-\lambda \frac{n-k}{n}\right] \\
&\to \mathbb{P}(Y_n = k),
\end{aligned}$$

where $Y_n \sim \text{Poi}(np)$.

10.2 Exponential distribution



The exponential distribution is often used to model wait times, i.e. I am at a supermarket checkout line, and I want to model the time I have to wait till the next customer arrives.

Fact 10.3. The exponential distribution is (essentially) the only memoryless distribution, i.e.

$$P(X > s + t | X > s) = P(X > t).$$

To verify this fact we can calculate the CDF of the exponential distribution.

$$F(t) = \int_0^t \lambda \exp(-\lambda x) dx = \frac{\lambda \exp(-\lambda x)}{-\lambda}\Big|_0^t = 1 - \exp(-\lambda t).$$

So we see that,

$$\begin{aligned} P(X > t) &= \exp(-\lambda t), \\ P(X > s + t | X > t) &= \frac{P(X > s + t, X > t)}{P(X > t)} = \frac{\exp(-\lambda(s + t))}{\exp(-\lambda t)} \\ &= \exp(-\lambda s) = \mathbb{P}(X > s). \end{aligned}$$

It is worth spending more time digesting memoryless-ness as a property. It is usually a bit counter-intuitive at first sight. The Ross book has several examples.

Example 10.1. Suppose that the time we spend in a bank is exponential with mean 10 minutes. What is the probability that we will spend more than 15 minutes in the bank? What is the probability that we will spend more than 15 minutes in the bank given that we have already spent 10 minutes in the bank?

We are given that $1/\lambda = 10$. The answer to the first question is $P(X > 15) = \exp(-15/10) = \exp(-1.5)$. The answer to the second question is $P(X > 15|X > 10) = \exp(-0.5)$.

Example 10.2. Suppose that the amount of time a lightbulb works is exponential with mean 10 hours. Suppose that I enter the room, and the lightbulb is burning. I would like to work for 5 hours, what is the probability that I can complete my work before the bulb burns out? What is this probability if the distribution was not exponential?

If the distribution is exponential then the quantity is simply $P(X > t + 5|X > t) = P(X > 5) = \exp(-5/10)$.

On the other hand if we did not know that the distribution was exponential then we would say:

$$P(X > t + 5|X > t) = \frac{P(X > t + 5)}{P(X > t)} = \frac{1 - F(t + 5)}{1 - F(t)},$$

where F is the CDF of the distribution. Notice that this quantity depends on t.

The connection between exponential distribution and Markov chains will become clearer as time goes on, but for now it is worth noting that memoryless-ness is like a Markov property, i.e. in some sense the distribution conditioned on the present forgets the past.

Example 10.3. Suppose X_1 and X_2 are independent exponential RVs with parameters λ_1 and λ_2. What is the probability: $\mathbb{P}(X_1 < X_2)$?

$$\begin{aligned}
P(X_1 < X_2) &= \int_x P(X_1 < x|X_2 = x)\mathbb{P}(X_2 = x)dx \\
&= \int_x (1 - \exp(-\lambda_1 x))\lambda_2 \exp(-\lambda_2 x)dx \\
&= \lambda_2 \left[\int_x \exp(-\lambda_2 x)dx - \int_x \exp(-(\lambda_1 + \lambda_2)x)dx \right] \\
&= \lambda_2 \left[\frac{1}{\lambda_2} - \frac{1}{\lambda_1 + \lambda_2} \right] \\
&= \frac{\lambda_1}{\lambda_1 + \lambda_2}.
\end{aligned}$$

Example 10.4. Show that if $X_1, X_2, \ldots, X_n$ are exponentials with rate λ_i, then their minimum is also an exponential. Calculate its rate.

Let us define:

$$Z = \min_i X_i.$$

Then we can compute

$$\begin{aligned} P(Z > t) &= \prod_{i=1}^{n} P(X_i > t) \\ &= \prod_{i=1}^{n} \exp(-\lambda_i t) \\ &= \exp(-(\sum_{i=1}^{n} \lambda_i)t). \end{aligned}$$

So we see that the minimum Z is exponential with rate $\sum_{i=1}^{n} \lambda_i$.

10.3 Counting Processes

Think about the following examples:

1. A store opens its doors and customers arrive at various times through the day.
2. Earthquakes along the San Andreas fault occur from time to time.
3. Copying errors occur at random points along a strand of DNA.
4. Spam arrives in your inbox

All of these situations involve occurrences that are scattered randomly across time or space. For concreteness, we will call the occurrences *arrivals*, thinking of customers entering a store.

We will *count* the number of arrivals till (and including) time t, and denote this N_t. For concreteness, $N_0 = 0$, and now we have a process $(N_t)_{t \geq 0}$, which is just a collection of non-negative discrete RVs. To begin with lets try to understand a basic discrete time counting process.

10.3.1 Bernoulli Processs

Before we move on to general counting processes, lets take a brief minute to review a basic Markov chain counting process known as a Bernoulli process.

Suppose that every day, I toss a coin, and if it comes up heads I count an arrival, i.e. suppose the sequence of tosses was 00101011 then I would record the counting process $N(t)$ as 00112234, which just counts the number of heads I have seen so far.

This is just an infinite Markov chain with $\mathbb{S} = \{0, 1, \ldots\}$ and transition matrix, $P_{i,i+1} = p$ and $P_{i,i} = 1 - p$.

This Markov chain has no limiting distribution, but we might ask other questions:

1. What is the time to the k-th head?
2. What is the distribution of time between the k-th head and $k+1$-st head?

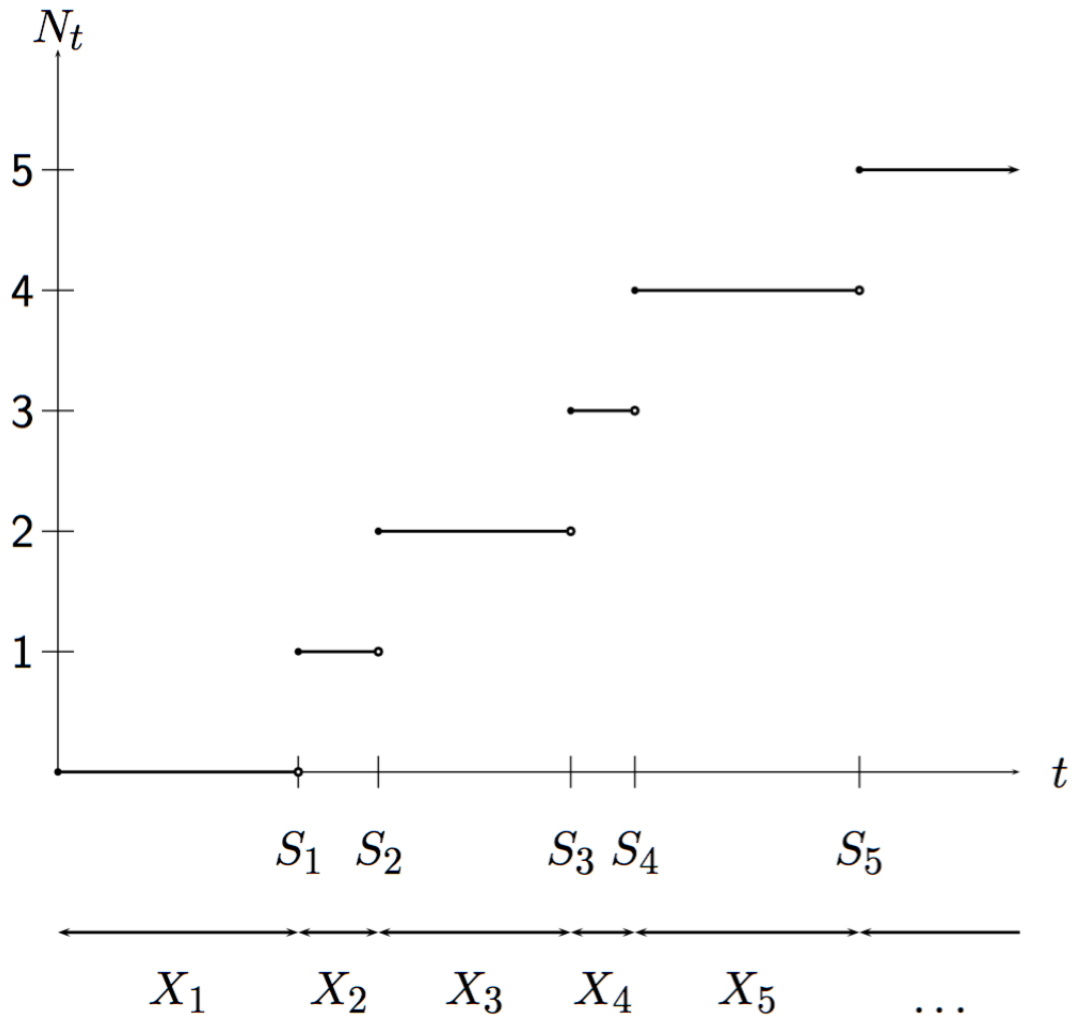
10.3.2 More General Counting Processes

We want to often model things like arrivals at a store. Customers really can arrive at any time at a store: we could discretize more finely.

Or we could just keep track of the arrival times. The key point for us is that the number of arrivals is now a continuous time stochastic process, i.e. unlike in Markov chains/Bernoulli processes where time was discrete (and events/transitions happened at equally spaced intervals). Here are a few basic observations we can make about counting processes:

- The counting process $N(t)$ jumps at each arrival.
- It is monotonically increasing.
- $N(t) - N(s)$ gives us the number of arrivals in $(s, t]$.

Here is a picture to keep in mind:



Here $S_1, S_2, \ldots$ are known as arrival times, and $X_1, X_2, \ldots$ are known as inter-arrival times. In order to describe the behavior of a counting process we typically need to describe N_0 and $N_t - N_s$ for every $t, s \geq 0$.

A basic set of choices in this context:

1. $N_0 = 0$.
2. $N_{t_1} - N_{s_1}, N_{t_2} - N_{s_2}, \ldots, N_{t_k} - N_{s_k}$ are independent for every positive integer k and every $0 \leq s_1 < t_1 < s_2 < \ldots < t_{k-1} < s_k < t_k$.

   Note that this is a *disjoint* set of intervals. This property is called **independent increments**.
3. The marginal distribution of $N_t - N_s$, for $t > s$, has a distribution that depends only on the length of the interval $t - s$. This property is called **stationary increments.**

We are now ready to define a Poisson process which is a particular kind of counting process.

Definition 1:
A counting process $N = (N_t)_{t \geq 0}$ is said to be a (homogeneous) Poisson Process with rate λ if:

1. $N_0 = 0$
2. N has independent increments.
3. If $t > s$, $N_t - N_s$ has a Poisson($\lambda(t - s)$) distribution. Note that N consequently has stationary increments.

We have that $\mathbb{E}(N_t - N_s) = \lambda(t - s)$, revealing that λ is the expected number of arrivals per unit time.

At this point it should be clear that Poisson is one of the few choices that would give us stationary increments. Particularly, we need the sum property to work out correctly.

Example 10.5. Suppose that calls arrive in a call center as a Poisson process with rate $\lambda = 30$/hour. What is the probability that in the next 5 minutes no calls come in? What is the expected number of calls during a ten minute period? What is the probability of receiving over 40 calls in the next hour?

So we let X denote the number of calls in the next 5 minutes. Then $X \sim \text{Poi}(30/12)$.

Then we have that,

$$P(X = 0) = \exp(-30/12) \approx 0.082.$$

Suppose that Y denotes the number of calls in a ten minute period, then $Y \sim \text{Poi}(30/6)$, and then $\mathbb{E}[Y] = 5$.

Now, let Z denote the number of calls in a 60 minute period, then $Z \sim \text{Poi}(30)$, so

$$P(Z > 40) = 1 - \text{poisson cdf}(40; \lambda = 30) \approx 0.0323.$$

Example 10.6. The Poisson process, as we just described it may seem like the result of just assuming an arbitrary distribution for the number of events in an interval. While this is somewhat true, one can also arrive at the Poisson process via a more natural generative process. The key point is: the Poisson process can be viewed as a limit of a Bernoulli

process.

Suppose we took a time window say $[0, 1]$ and discretized it very finely into bins of size δ (where δ is very small). Now, in each window we toss a coin with probability of heads $p = \lambda\delta$ (for some parameter $\lambda > 0$), and if it comes up heads we mark an arrival at the center of the window. This is now a Bernoulli process since the arrivals can only happen in discrete time, but it behaves like a Poisson process. In particular, for any $0 < t < 1$,

$$\#\text{number of arrivals} \sim \text{Bin}(\frac{t}{\delta}, \lambda\delta) \to \text{Poi}(\lambda t),$$

as $\delta \to 0$ using the law of small numbers.

Definition 2: The second definition, connects Poisson processes to exponential distributions. Suppose we have a random walk: with $S_0 = 0$, and $S_n = \sum_{i=1}^{n} X_i$ for $n \geq 1$ where the X_i s are exponential(λ) RVs.

Define,

$$N_t = \max\{n : S_n \leq t\}$$

Then $(N_t)_{t\geq 0}$ is a (homogenous) Poisson process with rate λ. This gives a way to *simulate* a Poisson process.

Example 10.7. Suppose we have a Poisson process with rate λ. Show that X_1 has an exponential(λ) distribution.

Notice that,

$$\begin{aligned}\mathbb{P}(X_1 > t) &= \mathbb{P}(N(t) = 0) \\ &= \mathbb{P}(\text{Poi}(\lambda t) = 0) \\ &= \exp(-\lambda t).\end{aligned}$$

This is just 1 minus the CDF of the exponential distribution, which confirms that X_1 is indeed exponentially distributed.

One can also use the independent, stationary increments property to similarly argue that $X_2, X_3, \ldots$ are all exponentially distributed with parameter λ.

Example 10.8. Suppose that buses arrive at a bus stop as a Poisson process with rate λ. I arrive at a randomly chosen time. What is the expected time I have to wait for the next bus?

Example 10.9. What is the pdf of the time of the k^{th} arrival in a Poisson process? This is called the *Erlang* distribution.

We're going to use the fact that

$$\{N(t) \geq k\} \quad \text{if and only if} \quad \{S_k \leq t\}.$$

First, calculate the cdf and then take a derivative to obtain the pdf.

$$F_{S_k}(t) \;=\; \mathbb{P}\left(S_k \le t\right) \;=\; \mathbb{P}(N(t) \ge k) \;=\; \sum_{j=k}^{\infty} \exp\left(-\lambda t\right) \frac{(\lambda t)^j}{j!},$$

which, upon differentiation, yields

$$\begin{aligned} f_{S_k}(t) \;&=\; \sum_{j=k}^{\infty} \Big\{ -\lambda \exp(-\lambda t) \frac{(\lambda t)^j}{j!} + \lambda \exp(-\lambda t) \frac{(\lambda t)^{j-1}}{(j-1)!} \Big\} \\ &=\; \lambda \exp(-\lambda t) \frac{(\lambda t)^{k-1}}{(k-1)!} + \sum_{j=k+1}^{\infty} \lambda \exp(-\lambda t) \frac{(\lambda t)^{j-1}}{(j-1)!} - \sum_{j=k}^{\infty} \lambda \exp(-\lambda t) \frac{(\lambda t)^j}{j!} \\ &=\; \lambda \exp(-\lambda t) \frac{(\lambda t)^{k-1}}{(k-1)!}. \end{aligned}$$

This is called the *Erlang distribution* with the parameters (k, λ). Note that if $k = 1$, it becomes the exponential distribution with λ.

Definition 3: A counting process $N = (N_t)_{t \ge 0}$ is said to be a (homogeneous) Poisson Process with rate λ if

1. $N_0 = 0$
2. N has independent increments.
3. $\mathbb{P}(N_{t+h} - N_t > 1) = o(h)$ as $h \to 0$.
4. $\mathbb{P}(N_{t+h} - N_t = 1) = \lambda h + o(h)$ as $h \to 0$.

You should think about the $o(h)$ as something very small (treat it as 0). More formally, $f(h) = o(h)$ if $\lim_{h \to 0} \frac{f(h)}{h} = 0$. So for instance $f(h) = h^2$ would be $o(h)$.

The third condition says that we are very unlikely to see more than one arrival in a small interval. The fourth condition says that the chance of one arrival in a small interval is roughly proportional to λ.

We won't explore this definition too much, but this is closely related to the Binomial intuition we tried to develop before.

10.4 Multiple Independent Poisson Processes

Intuitively we can imagine that the total number of arrivals is also a Poisson process, whose arrival rate is just the sum of the arrival rates of the individual processes.

Example 10.10. Suppose that buses from two routes, named A and B, stop at a particular bus stop. Buses from route A, arrive as a Poisson process with rate λ_A and buses from route B, arrive as a Poisson process with rate λ_B. Arrivals of buses from the two routes are independent. Suppose I arrive at the bus stop at a random time:

1. Argue that the bus arrival rate is $\lambda_A + \lambda_B$.
2. What is the expected waiting time until the first bus (of either route)?
3. What is the probability that the next bus that arrives is from route A?
4. What is the expected number of buses from route A that will pass before the first bus from route B?

We answer each of these questions in turn.

1. Choose an interval of length h of the form $(t, t+h]$ and let,

$$X = \text{number of A buses in } (t, t+h]$$
$$Y = \text{number of B buses in } (t, t+h],$$

   then $X \sim \text{Poi}(\lambda_A h)$ and $Y \sim \text{Poi}(\lambda_B h)$ so that $X + Y \sim \text{Poi}((\lambda_A + \lambda_B)h)$, as desired.

2. $\mathbb{E}(\text{Exp}(\lambda_A + \lambda_B)) = \frac{1}{\lambda_A + \lambda_B}$.
3. We know that $T_1 \sim \text{Exp}(\lambda_A)$ and $T_2 \sim \text{Exp}(\lambda_B)$ so we have that,

$$\mathbb{P}(T_1 < T_2) = \frac{\lambda_A}{\lambda_A + \lambda_B}.$$

4.

Example 10.11. We have k lightbulbs, each with a lifetime which is an Exponential(λ). Find the expected time until the last light bulb dies out.

We want to find the expected value of the maximum

$$\mathbb{E}\left[\max\{X_1, \cdots X_k\}\right]$$

where $X_i \overset{i.i.d.}{\sim} \text{Exp}(\lambda)$.

Approach 1: directly calculate the cdf or the pdf of the maximum value and take the expectation.

Let $Y_{(k)} = \max\{X_1, \cdots, X_k\}$. The cdf of $Y_{(k)}$ is givey by

$$\begin{aligned} F_{Y_{(k)}}(t) &= \mathbb{P}(Y_{(k)} \leq t) \\ &= \mathbb{P}(X_1 \leq t, \cdots, X_k \leq t) \\ &= \prod_{i=1}^{k} \mathbb{P}(X_i \leq t) \\ &= (1 - \exp(-\lambda t))^k. \end{aligned}$$

The expected value of the nonnegative random variable can be obtained by using the following fact:

Useful fact:
If $X \geq 0$,

$$\mathbb{E}(X) = \int_0^\infty 1 - F_X(t)dt.$$

Using the above, we have

$$\mathbb{E}\left[Y_{(k)}\right] \;=\; \int_0^\infty 1 - (1 - \exp(-\lambda t))^k dt \;=\; \sum_{j=1}^{k} \frac{1}{\lambda j}.$$

Approach 2: Use the connection between Poisson processes and exponential distributions.

Consider the order statistics $Y_{(1)}, \cdots, Y_{(k)}$ of $X_1, \cdots, X_k$. The maximum value can be decomposed into

$$Y_{(k)} = Y_{(1)} + (Y_{(2)} - Y_{(1)}) + \cdots + (Y_{(k)} - Y_{(k-1)})$$

so that

$$\mathbb{E}\left[Y_{(k)}\right] \;=\; \mathbb{E}\left[Y_{(1)} - 0\right] + \mathbb{E}\left[Y_{(2)} - Y_{(1)}\right] + \cdots + \mathbb{E}\left[Y_{(k)} - Y_{(k-1)}\right].$$

Next, consider k independent Poisson processes with the same rate λ. In this case, $Y_{(1)}$ is the waiting time of the first event of k independent Poisson processes. We can merge the k Poisson processes to obtain $\mathbb{E}\left[Y_{(1)}\right]$. So, we have

$$Y_{(1)} \sim \text{Exp}(k\lambda) \quad \text{and} \quad \mathbb{E}\left[Y_{(1)} - 0\right] \;=\; \frac{1}{k\lambda}.$$

Similarly, consider $(k-1)$ independent Poisson processes starting from $Y_{(1)}$. Then you can argue that $Y_{(2)} - Y_{(1)}$ is the waiting time of the first event from the Poisson process with rate $(k-1)\lambda$.

Therefore,

$$Y_{(2)} - Y_{(1)} \sim \text{Exp}((k-1)\lambda) \quad \text{and} \quad \mathbb{E}\left[Y_{(2)} - Y_{(1)}\right] \;=\; \frac{1}{(k-1)\lambda}.$$

Based on the same argument, we can deduce

$$\mathbb{E}\left[Y_{(k)}\right] \;=\; \sum_{j=1}^{k} \frac{1}{\lambda j}.$$

10.5 Non-homogeneous Poisson processes

Up to now, we have looked at Poisson processes where the rate λ is a fixed constant $\lambda > 0$. However, in many application areas, it makes sense to allow the rate paramter to change over time

For instance, we can expect that there are more customers vising a coffee shop in the morning than at night.

In this case, we can use a generalization of the Poisson process where the rate parameter changes over time, and we call $\lambda(t)$ as the **intensity function** of a non-homogeneous Poisson process.

Formally, a counting process is said to be a **non-homogeneous Poisson process** if it satisfies

1. $N_0 = 0$.
2. N has independent increments.
3. If $t > s$, $N_t - N_s$ has a Poisson distribution with rate

$$\lambda_{st} = \int_s^t \lambda(u)du.$$

Example 10.12. Siegbert runs a hot dog stand that opens at 8 AM.

1. From 8 AM until 11 AM, customers seem to arrive, on the average, at a steadily increasing rate that starts with an initial rate of 5 customers per hour at 8 AM and reaches a maximum of 20 customers per hour at 11 AM.
2. From 11 AM until 1 PM the (average) rate seems to remain constant at 20 customers per hour.
3. However, the (average) arrival rate then drops steadily from 1 PM until closing time at 5 PM at which time it has the value of 12 customers per hour.

If we assume that the numbers of customers arriving at Siegbert's stand during disjoint time periods are independent, then what is a good probability model for the preceding? What is the probability that no customers arrive between 8:30 AM and 9:30 AM on Monday morning? What is the expected number of arrivals in this period?

Q1: What is the probability that no customers arrive between 8:30 AM and 9:30 AM on Monday morning?

Answer: Assume that arrivals constitute a non-homogeneous Poisson process with the intensity function

$$\lambda(t) = \begin{cases} 5 + 5t, & 0 \le t \le 3 \\ 20, & 3 \le t \le 5 \\ 20 - 2(t-5), & 5 \le t \le 9 \end{cases}$$

Then we're interested in

$$\mathbb{P}(\text{no customers arrive between 8:30 AM and 9:30 AM}) \;=\; \mathbb{P}(X = 0)$$

where X has a Poisson distribution with rate

$$\lambda = \int_{0.5}^{1.5} (5 + 5t)dt = 10.$$

So the answer is $\exp(-10)$.

Q2: What is the expected number of arrivals in this period?

Answer: Since we know $X \sim \text{Poi}(10)$, $\mathbb{E}[X] = 10$.

Example 10.13. What is the distribution of X_1 in a non-homogeneous Poisson process with intensity $\lambda(t)$?

Remember X_1 is the waiting time of the first event from a non-homogeneous Poisson process with intensity $\lambda(t)$. Thus, the distribution of X_1 is

$$\begin{aligned}\mathbb{P}(X_1 \leq t) &= 1 - \mathbb{P}(X_1 > t) \\ &= 1 - \mathbb{P}(\text{zero arrivals in } [0,t)) \\ &= 1 - \exp\left(-\int_0^t \lambda(u)du\right).\end{aligned}$$

Note that if $\lambda(t)$ is a constant in terms of t, then X_1 has an exponential distribution.

10.6 Compound Poisson processes

Suppose that $Y_1, Y_2, \ldots$ are independent, identically distributed random variables. Assume $\{N(t), t \geq 0\}$ is a Poisson process which is independent of the Y_i. Define

$$X(t) = \begin{cases} \sum_{i=1}^{N(t)} Y_i, & \text{if } N(t) \geq 1 \\ 0, & \text{if } N(t) = 0. \end{cases}$$

Then $\{X(t) : t \geq 0\}$ is a **compound Poisson process**.

Examples:

- If $Y_i = 1$, then $X(t) = N(t)$, and we have the usual Poisson process.
- Assume that customers check out at a store as a Poisson process with rate λ (or with intensity function $\lambda(t)$). Then let Y_i denote the amount of time spend by the ith customer. $X(t)$ would be the total amount of time spent by all customers up to time t.
- Assume that an event is either a birth ($Y_i = +1$) or a death ($Y_i = -1$) within a population. If these events occur as a Poisson process, then $X(t)$ is the change in population size over the period $(0, t]$. This is a type of **birth and death process**, which will be part of our next topic.

Example 10.14. What is the expected value of $X(t)$?

We're going to use the law of total expectation, which says

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]].$$

Using this, we have

$$\mathbb{E}[X(t)] = \lambda t \times \mu_Y.$$

10.7 More practice problems

In this section we cover some more practice problems to build familiarity with calculations involving Poisson processes.

Example 10.15. Suppose we have a PP with rate λ.

1. $\mathbb{E}$[time of the 10th arrival]
2. $\mathbb{P}$(10th arrival occurs 2 or more time units after the 9th arrival)
3. $\mathbb{P}$(10th arrival occurs later than time 20)
4. $\mathbb{P}$(2 arrivals in [1,4] and 3 arrivals in [3,5]).

The answers are:

1. $10/\lambda$.

2. $$\begin{aligned}&\mathbb{P}(\text{10th arrival occurs 2 or more time units after the 9th arrival})\\ &= \int_t \mathbb{P}(\text{no arrival in (t,t+2]}|S_9 = t)\mathbb{P}(S_9 = t)dt\\ &= \int_t \mathbb{P}(N(2) = 0)\mathbb{P}(S_9 = t)dt\\ &= \exp(-2\lambda)\int_t \mathbb{P}(S_9 = t)dt = \exp(-2\lambda).\end{aligned}$$

3. $$\begin{aligned}\mathbb{P}(\text{10th arrival occurs later than time 20}) &= \mathbb{P}(N(20) \leq 9)\\ &= \sum_{j=0}^{9} \exp(-20\lambda)\frac{(20\lambda)^j}{j!}.\end{aligned}$$

   One can also write this using the CDF of the Erlang distribution (how?).

4. Notice that there are two overlapping windows that we need to reason about so we cannot just multiply the probabilities. However, suppose we defined:

   $$\begin{aligned}X &= \text{number of arrivals in [1,3]}\\ Y &= \text{number of arrivals in (3,4]}\\ Z &= \text{number of arrivals in (4,5]}.\end{aligned}$$

   We can see that these are all independent random variables and further that we need to

calculate

$$\begin{aligned}\mathbb{P}(X+Y=2, Y+Z=3) &= \sum_{k=0}^{2} \mathbb{P}(X=2-k, Z=3-k|Y=k)\mathbb{P}(Y=k) \\ &= \sum_{k=0}^{2} \exp(-\lambda)\frac{\lambda^k}{k!} \times \exp(-2\lambda)\frac{(2\lambda)^{2-k}}{(2-k)!} \times \exp(-\lambda)\frac{\lambda^{3-k}}{(3-k)!} \\ &= \exp(-4\lambda)\left[\frac{\lambda^5}{3} + \lambda^4 + \frac{\lambda^3}{2}\right].\end{aligned}$$

Example 10.16. Assume that cars arrive at rate 10 per hour. Assume each car will pick up a hitchhiker with probability 1/10. You are second in line. What is the probability that you will have to wait for more than 2 hours?

The key point is just that the cars that pick up the hitchhikers follow a PP with rate 1. So we would like to compute for this process:

$$\begin{aligned}\mathbb{P}(T_1 + T_2 > 2) = \mathbb{P}(N(2) \leq 1) &= \mathbb{P}(N(2)=0) + \mathbb{P}(N(2)=1) \\ &= \exp(-2)\,[1+2] = 3\exp(-2).\end{aligned}$$

Example 10.17. Suppose that we have arrivals according to two independent Poisson processes, one with rate $\lambda_1 = 5$ and the other with rate $\lambda_2 = 1$. What is the probability that we see at least 4 events from the first process before we see any events from the second process?

The key idea is that we can merge the two PPs into a single PP with rate $\lambda_1 + \lambda_2$. Now in the merged process, each arrival is independently from the first process with probability $\lambda_1/(\lambda_1+\lambda_2)$ and from the second process with probability $\lambda_2/(\lambda_1+\lambda_2)$.

Now the probability of at least 4 events from the first process before any events from the second process is just given by $(\lambda_1/(\lambda_1+\lambda_2))^4$.

Observe that you can follow exactly the same reasoning to reason about different patterns, i.e. for example: what is the probability that we see 3 events from the first process before we see 4 events from the second process?

Longer way: One can also solve this problem in a more brute-force fashion. Suppose we denote the first arrival time of the second process by X_{12}, then we can write the desired probability as:

$$\begin{aligned}p &= \int_t \mathbb{P}(\text{Poi}(\lambda t) \geq 4 | X_{12} = t)\mathbb{P}(X_{12}=t)dt \\ &= \int_t \sum_{k=4}^{\infty} \frac{\exp(-\lambda_1 t)(\lambda_1 t)^k}{k!}\lambda_2 \exp(-\lambda_2 t)dt \\ &= \sum_{k=4}^{\infty} \frac{\lambda_1^k \lambda_2}{k!}\int_t \exp(-(\lambda_1+\lambda_2)t)t^k dt.\end{aligned}$$

Now you will have to look up the moments of the exponential distribution to see that if $X \sim \text{Exp}(\lambda)$ then,

$$\mathbb{E}[X^n] = \frac{n!}{\lambda^n},$$

so we have that,

$$\begin{aligned} p &= \sum_{k=4}^{\infty} \frac{\lambda_2 \lambda_1^k}{k!} \frac{k!}{(\lambda_1+\lambda_2)^{k+1}} \\ &= \frac{\lambda_2}{\lambda_1+\lambda_2} \sum_{k=4}^{\infty} \left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^k \\ &= \left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^4. \end{aligned}$$

Example 10.18. Two people, Alice and Bob, are hitchhiking. Cars that would pick up a hitchhiker arrive as a Poisson process with rate λ_C. Alice is first in line for a ride. Moreover, after $\text{Exp}(\lambda_A)$ time, Alice quits, and after $\text{Exp}(\lambda_B)$ time, Bob quits. Compute the probability that Alice is picked up before she quits, and the same for Bob.

For Alice we can see that the probability is simply $\lambda_C/(\lambda_A+\lambda_C)$.

For Bob the calculation is more difficult, but the easiest way to do it is to see that:

$$\begin{aligned} \mathbb{P}(\text{Bob is picked up}) &= \mathbb{P}(\text{first arrival is either A or C, and second arrival amongst B or C arrivals is C}) \\ &= \frac{\lambda_A+\lambda_C}{(\lambda_A+\lambda_B+\lambda_C)} \times \frac{\lambda_C}{\lambda_B+\lambda_C}. \end{aligned}$$

Example 10.19. Suppose that $N(t)$ is a Poisson process with rate λ. Calculate the covariance function, i.e. for two times $t_1, t_2 > 0$ calculate $\text{Cov}(N(t_1), N(t_2))$.

The way to solve this problem is to try to convert it into a problem involving independent RVs. Assume that $t_1 \leq t_2$, then we can write:

$$\begin{aligned} \text{Cov}(N(t_1), N(t_2)) &= \text{Cov}(N(t_1), N(t_2) - N(t_1) + N(t_1)) \\ &= \text{Cov}(N(t_1), N(t_2) - N(t_1) + \text{Var}(N(t_1)) \\ &= 0 + \lambda t_1, \end{aligned}$$

where the first term is 0 because those RVs are independent. More generally, $\text{Cov}(N(t_1), N(t_2)) = \min\{t_1, t_2\}$.

Example 10.20. Suppose that we have a PP with rate λ, show that conditional on $N(t) = 1$ the first arrival time T_1 has a uniform distribution on $(0, t]$.

To show this we can simply try to calculate:

$$\begin{aligned}
\mathbb{P}(T_1 \leq x | N(t) = 1) &= \frac{\mathbb{P}(T_1 \leq x, N(t) = 1)}{\mathbb{P}(N(t) = 1)} \\
&= \frac{\mathbb{P}(T_1 \leq x, N(t) = 1)}{\lambda t \exp(-\lambda t)} \\
&= \frac{\mathbb{P}(\text{one arrival in (0,x] and no arrivals in (x,t])}}{\lambda t \exp(-\lambda t)} \\
&= \frac{(\lambda x) \exp(-\lambda x) \exp(-\lambda(t-x))}{\lambda t \exp(-\lambda t)} \\
&= \frac{x}{t}.
\end{aligned}$$

Chapter 11

Continuous Time Markov Chains

In the last chapter we discussed Poisson Processes, where the goal was to model a stochastic process which counts the number of arrivals we have seen so far. The key point was that the arrivals happened in continuous time so we could not model this using a Markov chain model.

A CTMC is a generalization of the Poisson process, where we have discrete state space $\mathbb{S}$, but allow for transitions in continuous time.

11.1 The Markov Property and Exponential Distributions

We write a CTMC as $\{X(t), t \geq 0\}$ where $X : t \mapsto \mathbb{S}$ is a map from time to states, i.e. $X(t)$ just represents the state of the stochastic process at time t.

Now we can say what the Markov property of CTMCs is. We require that,

$$\mathbb{P}(X(t+s) = j | X(s) = i, X(u) = x(u) \text{ for } 0 \leq u < s) = \mathbb{P}(X(t+s) = j | X(s) = i),$$

so that the past is independent of the future conditioned on the present. We will also insist on time-homogeneity or stationarity of the transition probabilities, i.e. that,

$$P_{ij}(t) = \mathbb{P}(X(t+s) = j | X(s) = i) = \mathbb{P}(X(t) = j | X(0) = i).$$

Here $P_{ij}(t)$ denotes the transition probability function.

An important consequence of these two assumptions (Markov + Time-Homogeneity) is that the holding time of each state, i.e. the time spent in a state before transitioning out has an *exponential distribution*.

Intuitively, this is because of the nature of the Markov assumption, i.e. once I condition on the state at time s it does not matter how long I have already spent in the state (which is part of the past history of the process): it is as though the process restarts itself. This is only possible if the distribution of the time-spent in each state is memoryless. To see this more formally:

The key take-home from this section is that to specify a CTMC, we need to specify two things: a transition probability matrix (as in the discrete-time case) and a vector of rates which parameterize the exponential distributions of the holding times. We will denote the rates by ν_i for $i \in \mathbb{S}$.

11.2 Examples

There are many examples of stochastic processes that are naturally modeled as CTMCs. We have already discussed counting processes which are an example. Counting processes are sometimes referred to as *pure birth* processes in order to distinguish them from a generalization which allows for both arrivals and departures – these are called *birth and death processes*. Finally, most models that arise in queuing theory are also CTMCs and we will see examples of this in this section.

The basic goal for this section is to illustrate how to convert various descriptions of CTMCs into a standard transition matrix + rates description so that we can then study CTMCs in this standard form.

Example 11.1. A Poisson Process with rate parameter λ is a CTMC.

1. The states of the CTMC are $\{0, 1, 2, \dots, \}$.
2. The transition matrix of the CTMC is given by: for any state i, $P_{i,i+1} = 1$ and all other transitions have 0 probability.
3. Finally, every state has a rate parameter $\nu_i = \lambda$.

A more complex example is a birth and death process (BADP). In a BADP each state $i > 0$ is associated with a birth rate λ_i and a death rate μ_i. The state corresponding to $i = 0$ only allows births, i.e. $\mu_0 = 0$.

Example 11.2. The BADP is a CTMC.

1. The states of the CTMC are $\{0, 1, 2, \dots, \}$.
2. The transition matrix of the CTMC is given by: for any state i,

$$P_{i,i+1} = \frac{\lambda_i}{\lambda_i + \mu_i},$$
$$P_{i,i-1} = \frac{\mu_i}{\lambda_i + \mu_i},$$

   and all other transitions have 0 probability.
3. Finally, every state has a rate parameter $\nu_i = \lambda_i + \mu_i$.

The best way to understand this example is to visualize the BADP as a competition between two Poisson Processes (we studied examples like this in the previous chapter) – the birth process and the death process. If the first arrival comes from the birth process then we transition from i to $i + 1$. The probability of transitioning to $i + 1$ is therefore:

$$\mathbb{P}(\text{Exp}(\lambda_i) < \text{Exp}(\mu_i)) = \frac{\lambda_i}{\lambda_i + \mu_i},$$

as desired.

Example 11.3. A *Yule* process is a pure-birth process (like a Poisson process), where the birth-rate is dependent on the number of individuals in the population, i.e. for a parameter λ the birth-rates are given as:

$$\lambda_i = i\lambda.$$

Example 11.4. Another variant of the BADP is known as a *linear BADP with immigration*. In this case we have that,

$$\mu_i = i\mu$$
$$\lambda_i = i\lambda + \theta.$$

Our next example comes from queuing theory.

Example 11.5. An M/M/s queue is a queue is one in which customers arrive according to Poisson process with parameter λ. Upon arrival they find s-servers and each server has a service time that has an Exp(μ) distribution. If all the servers are busy the customers join the queue, and wait to be served.

Let $X(t)$ denote the number of people currently being served and waiting to be served. Then $X(t)$ is a BADP with birth-rates:

$$\lambda_i = \lambda,$$

and death-rates:

$$\mu_i = \min\{i, s\}\mu.$$

11.3 The short and intermediate-term behavior of CTMCs

Recall that for discrete-time Markov chains, we computed k-step transition probabilities using the Chapman-Kolmogorov equations.

For CTMCs we would like to similarly compute $P_{ij}(t)$ i.e. the probability of going from i to j in t units of time. This transition probability function (for each pair of states) characterizes the intermediate-term behavior of CTMCs.

Fact 11.1. One way to understand the transition probability function is via discretizing a CTMC. Suppose we have a CTMC, but we only observe it at discrete times, i.e. for some fixed t, we observe the process $X'(k) = X(kt)$ for $k = \{0, 1, \ldots\}$, where $X(t)$ is a CTMC.

The process X' is now a discrete time Markov chain and its transition probability matrix is given by $P_{ij}(t)$.

In order to actually compute $P_{ij}(t)$ we need to do more work. We first define the *instantaneous rates*, i.e. for a pair of states (i, j) we define:

$$q_{ij} = \nu_i \times P_{ij}.$$

One simple fact is that we can recover both the transition probabilities and the rates from the instantaneous rates, i.e.:

$$\nu_i = \sum_j q_{ij},$$
$$P_{ij} = \frac{q_{ij}}{\sum_j q_{ij}}.$$

The significance of the instantaneous rates is that they give us a way to calculate the transition probability function on a very small time-scale.

Fact 11.2. For $i \neq j$,

$$\lim_{h \to 0} \frac{P_{ij}(h)}{h} = q_{ij},$$

and furthermore,

$$\lim_{h \to 0} \frac{1 - P_{ii}(h)}{h} = \nu_i.$$

So intuitively this is saying that for two states (i, j) if t is very small then $P_{ij}(t) \approx q_{ij}t$.

There is a slightly more formal proof in the Ross book, but all this amounts to is the observation that if h is very small then we are unlikely to make two (or more) transitions in a window of size h, and the probability of making one transition is

$$\mathbb{P}(\text{Exp}(\nu_i) \leq h) = 1 - \exp(-\nu_i h) \to \nu_i h.$$

So we make a transition with probability $\nu_i h$ and this transition takes us to state j with probability P_{ij}, i.e.

$$P_{ij}(h) \approx \nu_i h \times P_{ij} = q_{ij} h.$$

It is also straightforward to conclude that the probability that we make no transition in a small window of size h is given by:

$$P_{ii}(h) = 1 - \sum_{j \neq i} P_{ij}(h) \approx 1 - \nu_i h.$$

In essence the instantaneous rates capture the behavior of a CTMC on a very short time-scale. It will be convenient to arrange the q_{ij} into a matrix known as the *fundamental matrix of a CTMC*, i.e.

$$Q = \begin{bmatrix} -\nu_1 & q_{12} & q_{13} & \dots \\ q_{21} & -\nu_2 & q_{23} & \dots \\ \vdots & & & \end{bmatrix}.$$

Observe that each row of the fundamental matrix sums to 0. The next goal will be to try to understand CTMCs on a longer time-scale.

11.3.1 Intermediate term behavior

The intermediate term behavior of the CTMC is characterized by Kolmogorov's Forward and Backward equations. These equations are differential equations that characterize the transition probability function[1].

As a preliminary let us note that the Chapman-Kolmogorov equations continue to make sense for CTMCs, i.e. we have that for any pair of states (i, j) and $s, t \geq 0$:

$$P_{ij}(s+t) = \sum_k P_{ik}(s)P_{kj}(t).$$

Kolmogorov's Backward Equations: Kolmogorov's Backward equations say that,

$$P'(t) = QP(t).$$

Where $P'(t)$ is a matrix of time-derivatives of the transition probability function. Roughly, the way to interpret these equations is that if we know the transition function at time t, the instantaneous rates (in the matrix Q) can be used to derive what the transition function at time $t - \epsilon$ for some small $\epsilon > 0$ is.

More formally,

$$P'_{ij}(t) = \lim_{h\to 0}\left[\frac{P_{ij}(t+h) - P_{ij}(t)}{h}\right] = \lim_{h\to 0}\left[\frac{\sum_k P_{ik}(h)P_{kj}(t) - P_{ij}(t)}{h}\right],$$

using the C-K equations above. This can be further written as:

$$\lim_{h\to 0}\left[\frac{\sum_k P_{ik}(h)P_{kj}(t) - P_{ij}(t)}{h}\right] = \lim_{h\to 0}\left[\frac{\sum_{k\neq i} P_{ik}(h)P_{kj}(t) - (1 - P_{ii}(h))P_{ij}(t)}{h}\right],$$

which using the definitions of the instantaneous rates can be written as:

$$\lim_{h\to 0}\left[\frac{\sum_{k\neq i} P_{ik}(h)P_{kj}(t) - (1 - P_{ii}(h))P_{ij}(t)}{h}\right] = \sum_{k\neq i} q_{ik}P_{kj}(t) - \nu_i P_{ij}(t).$$

Putting all of this together we obtain that,

$$[P'(t)]_{ij} = [QP(t)]_{ij},$$

which are the Kolmogorov Backward equations.

Kolmogorov's Forward Equations: Kolmogorov's Forward equations say that,

$$P^{'}(t) = P(t)Q.$$

They can be derived in exactly the same way as the Backward equations but we do the first step differently, i.e. we write:

$$P'_{ij}(t) = \lim_{h\to 0}\left[\frac{P_{ij}(t+h) - P_{ij}(t)}{h}\right] = \lim_{h\to 0}\left[\frac{\sum_k P_{ik}(t)P_{kj}(h) - P_{ij}(t)}{h}\right],$$

[1] As a technical note (ignore this comment unless you are especially curious) deriving these equations requires interchanging certain limits and sums and as a result Kolmogorov's Backward equations are always valid but Kolmogorov's Forward equations are only valid when there is a valid limiting distribution for the CTMC. We will just assume that the conditions needed hold

and proceed exactly as before.

Solving these equations: In order to obtain the transition probability function for any value $t > 0$ we simply need to solve this system of differential equations, with the appropriate boundary conditions (at $t = 0$ we know that $P(t) = I$). It turns out that the solution mirrors the solution of the scalar differential equation:

$$f'(t) = qf(t) \implies f(t) = \exp(qt) + C.$$

In the matrix case we obtain the solution that:

$$P(t) = \exp(Qt),$$

where

$$\exp(Qt) = I + Qt + \frac{Q^2t^2}{2!} + \dots$$

11.3.2 Long term behavior

Fact 11.3. The limiting distribution π for a CTMC, when it exists, can be found by finding a distribution that solves the system of equations:

$$\pi^T Q = 0.$$

Before we prove this fact lets see an example.

Example 11.6. Molecules arrive to the surface of a bacterium according to a Poisson process with rate λ. An α fraction of these molecules are acceptable and the rest are unacceptable. An unacceptable molecule spends a time on the surface that has an $\text{Exp}(\mu_1)$ distribution and an acceptable molecule spends a time on the surface that has an $\text{Exp}(\mu_2)$ distribution. A new molecule only attaches to the surface if it is currently unoccupied. What fraction of time is the surface of the bacterium unoccupied?

Let us denote the state space as $\mathbb{S} = \{\text{unoccupied, acceptable, unacceptable}\}$. Then we have that the transition matrix is given as:

$$P = \begin{bmatrix} 0 & \alpha & 1-\alpha \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix},$$

and the rates are $\nu_1 = \lambda, \nu_2 = \mu_2, \nu_3 = \mu_1$. Putting these together we obtain the Q matrix:

$$Q = \begin{bmatrix} -\lambda & \alpha\lambda & (1-\alpha)\lambda \\ \mu_2 & -\mu_2 & 0 \\ \mu_1 & 0 & -\mu_1 \end{bmatrix}.$$

Now, solving the system $\pi^T Q = 0$ for the value of π_0 we obtain that,

$$\pi_0 = \frac{1}{1 + \frac{\alpha\lambda}{\mu_2} + \frac{(1-\alpha)\lambda}{\mu_2}}.$$

We now return to trying to verify that the limiting distribution must in fact satisfy the expression that $\pi^T Q = 0$. Let us return to the Kolmogorov forward equations which tell us that:

$$P'_{ij}(t) = \sum_{k \neq j} P_{ik}(t) q_{kj} - \nu_j P_{ij}(t).$$

Now suppose we take the limit as $t \to \infty$. If the limit of the derivative exists, it must be 0 – this is because P_{ij} is bounded between $[0, 1]$. If the derivative is not 0 then $P_{ij}(t)$ will eventually exit this bounded interval.

On the other hand $\lim_{t\to\infty} P_{ik}(t) = \pi_k$ (again when the limit exists). Using this we obtain that,

$$0 = \sum_{k \neq j} \pi_k q_{kj} - \nu_j \pi_j.$$

Putting these equations together we obtain the desired system of equations $\pi^T Q = 0$.

Intuitively, we can interpret the equation:

$$\sum_{k \neq j} \pi_k q_{kj} = \nu_j \pi_j,$$

as follows. The RHS is the rate at which we transition *out of* the state j (i.e. it is the probability of being in state j times the rate at which we make transitions when in state j). On the other hand, the LHS is the rate at which we transition *in to* the state j (i.e. it is the probability of being in state k times the rate at which we make transitions from k to j). So the system $\pi^T Q = 0$ is simply telling us that in the limit, if we have a well-defined limiting distribution, it must be balanced in this sense, the rate at which we transition in to and out of any state must be equal.

11.4 Practice Problems

Example 11.7. After being repaired, a machine functions for an exponential time with rate λ and then fails. Upon failure, a repair process begins. The repair process proceeds sequentially through k distinct phases. First a phase 1 repair must be performed, then a phase 2, and so on. The times to complete these phases are independent, with phase i taking an exponential time with rate $\mu_i, i = 1, \ldots, k$.

- What proportion of time is the machine undergoing a phase i repair?
- What proportion of time is the machine working?

We first define the state space $\mathbb{S} = \{0, 1, 2, \ldots, k\}$ denoting that the machine is working (state 0) or is in stage i repair for $i \in \{1, \ldots, k\}$. The rates are given as $\nu_0 = \lambda$, $\nu_1 = \mu_1, \ldots, \nu_k = \mu_k$. The transition matrix is:

$$P = \begin{bmatrix} 0 & 1 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 1 & 0 & \ldots & 0 \\ \vdots & & & & & \\ 1 & 0 & 0 & 0 & \ldots & 0 \end{bmatrix},$$

and the fundamental matrix Q is given by:

$$Q = \begin{bmatrix} -\lambda & \lambda & 0 & 0 & \dots & 0 \\ 0 & -\mu_1 & \mu_1 & 0 & \dots & 0 \\ 0 & 0 & -\mu_2 & \mu_2 & \dots & 0 \\ \vdots & & & & & \\ \mu_k & 0 & 0 & 0 & \dots & -\mu_k \end{bmatrix}.$$

Writing out the system of equations we see that:

$$-\lambda\pi_0 + \mu_k\pi_k = 0 \implies \pi_k = \frac{\lambda}{\mu_k}\pi_0$$
$$\lambda\pi_0 - \mu_1\pi_1 = 0 \implies \pi_1 = \frac{\lambda}{\mu_1}\pi_0, \dots$$

so using that the sum must be 1 we obtain that,

$$\pi_0 = \frac{1}{1 + \lambda\sum_{i=1}^{k}\frac{1}{\mu_i}}$$
$$\vdots$$
$$\pi_j = \frac{\lambda/\mu_j}{1 + \lambda\sum_{i=1}^{k}\frac{1}{\mu_i}}.$$

Example 11.8. Consider two copiers that are maintained by a single repairman. Machine i functions for a time that is exponentially distributed with rate λ_i, and the repair times for the i-th machine are exponentially distributed with rate μ_i. Assume that the machines are repaired in the order in which they fail. Suppose that we wish to construct a CTMC model of this system, with the goal of finding the long-run proportions of time that each copier is working and the repairman is busy. How can we proceed? Calculate these long-run probabilities.

First we need to decide on the state space. Notice that since we need to know which machine failed first (when the machines fail) so our state space needs to account for this:

$\mathbb{S} = \{\text{both working}, \text{machine 1 broken}, \text{machine 2 broken}, \text{both broken 1 first}, \text{both broken 2 first}\}$.

With this in place we can write down the rates,

$$\nu_1 = \lambda_1 + \lambda_2$$
$$\nu_2 = \mu_1 + \lambda_2$$
$$\nu_3 = \mu_2 + \lambda_1$$
$$\nu_4 = \mu_1$$
$$\nu_5 = \mu_2$$

and the transition matrix,

$$P = \begin{bmatrix} 0 & \frac{\lambda_1}{\lambda_1+\lambda_2} & \frac{\lambda_2}{\lambda_1+\lambda_2} & 0 & 0 \\ \frac{\mu_1}{\mu_1+\lambda_2} & 0 & 0 & \frac{\lambda_2}{\mu_1+\lambda_2} & 0 \\ \frac{\mu_2}{\mu_2+\lambda_1} & 0 & 0 & 0 & \frac{\lambda_1}{\lambda_1+\mu_2} \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}.$$

We can also find the fundamental matrix:

$$Q = \begin{bmatrix} -(\lambda_1+\lambda_2) & \lambda_1 & \lambda_2 & 0 & 0 \\ \mu_1 & -(\mu_1+\lambda_2) & 0 & \lambda_2 & 0 \\ \mu_2 & 0 & -(\mu_2+\lambda_1) & 0 & \lambda_1 \\ 0 & 0 & \mu_1 & -\mu_1 & 0 \\ 0 & \mu_2 & 0 & 0 & -\mu_2 \end{bmatrix}.$$

Solving the system of equations $\pi^T Q = 0$ and $\sum_i \pi_i = 1$ will give the desired result.

Example 11.9. What is the limiting distribution for a BADP?

We have already seen that for a BADP we can calculate the fundamental matrix as:

$$Q = \begin{bmatrix} -\lambda_0 & \lambda_0 & 0 & 0 & \dots & 0 & \\ \mu_1 & -(\lambda_1+\mu_1) & \lambda_1 & 0 & 0 & \dots & 0 \\ 0 & \mu_2 & -(\mu_2+\lambda_2) & \lambda_2 & 0 & \dots & 0 \\ \vdots & & & & & & \end{bmatrix}.$$

Writing out the usual system of equations we see that,

$$\begin{aligned} \pi_0\lambda_0 &= \pi_1\mu_1 \\ \pi_0\lambda_0 &= \pi_1(\lambda_1+\mu_1) - \pi_2\mu_2 \implies \pi_1\lambda_1 = \pi_2\mu_2 \\ &\vdots \end{aligned}$$

so we obtain that,

$$\begin{aligned} \pi_1 &= \frac{\lambda_0}{\mu_1}\pi_0 \\ \pi_2 &= \frac{\lambda_1}{\mu_2}\pi_1 = \frac{\lambda_1\lambda_0}{\mu_1\mu_2}\pi_0 \\ &\vdots \end{aligned}$$

and using the fact that the sum must be 1 we obtain that,

$$\pi_0 = \frac{1}{1+\sum_{n=1}^{\infty}\frac{\lambda_0\dots\lambda_{n-1}}{\mu_1\dots\mu_n}}$$

and denoting

$$\theta_n = \frac{\lambda_0 \dots \lambda_{n-1}}{\mu_1 \dots \mu_n},$$

we have that,

$$\pi_n = \frac{\theta_n}{1 + \sum_{n=1}^{\infty} \theta_n}.$$

So for a BADP to determine the limiting distribution we simply need to compute these θ_n.

Example 11.10. Consider a small barbershop, where there are only two barbers, each with his own barber chair. Suppose that there is only room for at most 5 customers, with 2 in service and 3 waiting. Assume that potential customers arrive according to a Poisson process at rate $\lambda = 6$ per hour. Customers arriving when the system is full are blocked and lost, leaving without receiving service and without affecting future arrivals. Assume that the duration of each haircut is an independent exponential random variable with a mean parameter $1/\mu = 15$ minutes. Customers are served in a first-come first-served manner by the first available barber.

1. What is the long-run proportion of time there are two customers in service plus two customers waiting?

We first need to decide on the state space. We can simply use the number of customers in the shop $\mathbb{S} = \{0, 1, 2, 3, 4, 5\}$.

Now the rates are given as:

$$\begin{aligned}
\nu_0 &= 6 \\
\nu_1 &= 10 \\
\nu_2 &= 14 \\
\nu_3 &= 14 \\
\nu_4 &= 14 \\
\nu_5 &= 8,
\end{aligned}$$

and the fundamental matrix:

$$Q = \begin{bmatrix}
-6 & 6 & 0 & 0 & 0 & 0 \\
4 & -10 & 6 & 0 & 0 & 0 \\
0 & 8 & -14 & 6 & 0 & 0 \\
0 & 0 & 8 & -14 & 6 & 0 \\
0 & 0 & 0 & 8 & -14 & 6 \\
0 & 0 & 0 & 0 & 8 & -8
\end{bmatrix}.$$

From this we can calculate the limiting distribution in the usual way.

Chapter 12

Markov Chain Mixing

In this chapter we will try to understand some partial answers to the question:

How long does a Markov chain take to reach its limiting distribution?

Before we start talking about this it is worth pondering why one should care about the answer. As you think about it you will notice that in every single application of Markov chains we discussed this was actually a core question:

1. PageRank Algorithm: Basically, run a Markov chain till it converges and hope that it converges quickly.
2. MCMC: I know eventually I will be drawing samples from the stationary distribution, but how long is "eventually"?
3. MDP: How fast do policy and value iterations converge?
4. Estimating a Markov chain: One can ask, is the MLE a good estimate of transition probabilities? The answer turns out to be "only when the Markov chain converges fast".

Throughout the chapter we will assume we are dealing with finite, irreducible, aperiodic Markov Chains, i.e. Markov Chains with unique limiting distributions. We will denote by π_0 some initial distribution and by π^* the unique limiting distribution. We know that we can write the distribution after t time-steps as:

$$\pi_t^T = \pi_0^T P^t.$$

We would like to understand how far apart π_t and π^* are.

12.1 How far are two distributions?

Before we can make much sense of this question we need some way to measure the distance between distributions. One way to do this is via the *total variation distance*, defined as:

$$\mathrm{TV}(p, q) = \frac{1}{2} \sum_{i=1}^{k} |p(i) - q(i)|,$$

where in our setting k denotes the number of states in the Markov chain.

It turns out that the Total Variation distance has many nice properties. For instance, you will show on your HW that the total variation distance can also be written as:

$$\mathrm{TV}(p,q) = \max_A \left[p(A) - q(A)\right],$$

where A denotes any event in the sample-space, i.e. the total variation distance is the most the two distributions can differ in terms of the probability they assign to events. Now that we have defined the TV distance, we might ask the question: how large is $\mathrm{TV}(\pi_t, \pi_0)$?

12.2 A basic example

To build some intuition we can try to work this out in some (very simple) examples.

Example 12.1. Suppose we have a frog that can jump between two lily pads, and has transition matrix:

$$P = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}.$$

Calculate the TV distance between π_t and π^* when π_0 is $(1,0)^T$, i.e. the frog starts on the first lily pad.

In this example we can easily calculate the limiting distribution

$$\pi^* = \begin{pmatrix} \frac{q}{p+q} \\ \frac{p}{p+q} \end{pmatrix},$$

so we can define,

$$\Delta_t = \pi_t(1) - \frac{q}{p+q},$$

as the distance of the first coordinate from its limiting value. Now, we can see that,

$$\begin{aligned}
\Delta_{t+1} &= \pi_{t+1}(1) - \frac{q}{p+q} \\
&= \pi_t(1)(1-p) + \pi_t(2)q - \frac{q}{p+q} \\
&= \pi_t(1)(1-p) + (1-\pi_t(1))q - \frac{q}{p+q} \\
&= (1-p-q)\left[\pi_t(1) - \frac{q}{p+q}\right] \\
&= (1-p-q)\Delta_t.
\end{aligned}$$

Since both π_t and π^* sum to 1, it is easy to verify that,

$$\pi_t(2) - \frac{p}{p+q} = -\Delta_t,$$

so now we can conclude that the total variation distance:

$$\mathrm{TV}(\pi_t, \pi^*) = \frac{1}{2}(2|\Delta_t|) = |\Delta_t| = |1-p-q|^t|\Delta_0|.$$

Since we started in the first state

$$\Delta_0 = 1 - \frac{q}{p+q} = \frac{p}{p+q},$$

so we see that,

$$\mathrm{TV}(\pi_t, \pi^*) = |1-p-q|^t \frac{p}{p+q}.$$

If $0 < p < 1, 0 < q < 1$ then $|1-p-q| < 1$ and we see that the total variation distance converges exponentially fast to 0. Markov chains of this type are referred to as *rapidly mixing*.

12.3 Mixing Time

A related parameter is often used to measure the speed of convergence to the limiting distribution. This is called the *mixing time* for a Markov chain:

$$\tau_{\mathrm{mix}}(\epsilon) = \min\{t : \mathrm{TV}(\pi_t, \pi^*) \leq \epsilon\}.$$

So the mixing time measures the amount of time we need to run the Markov chain to reach a distribution that is ϵ-close to the limiting distribution.

In particular, for the frog example above, we need that:

$$|1-p-q|^{\tau_{\mathrm{mix}}(\epsilon)} \frac{p}{p+q} \leq \epsilon,$$

i.e. that,

$$\tau_{\mathrm{mix}}(\epsilon) \leq \log_{1/|1-p-q|} \left[\frac{p}{(p+q)\epsilon} \right],$$

and if we treat p, q as constants then,

$$\tau_{\mathrm{mix}}(\epsilon) = C \log \frac{1}{\epsilon}.$$

12.4 Basic Convergence Theorem for Markov Chains

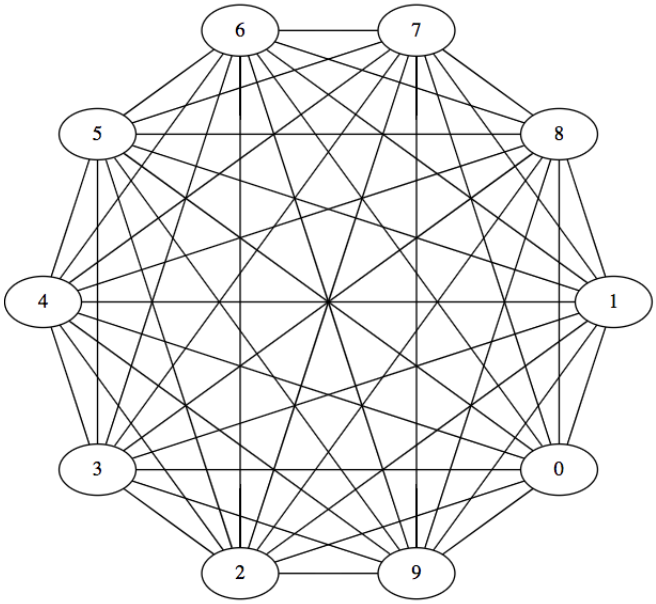
It turns out that the above result is true for every finite, irreducible, aperiodic Markov chain, i.e.

Fact 12.1. For any finite, irreducible, aperiodic Markov chain we have that there are constants $\alpha \in [0, 1)$ and $C > 0$ such that for $t > 0$,

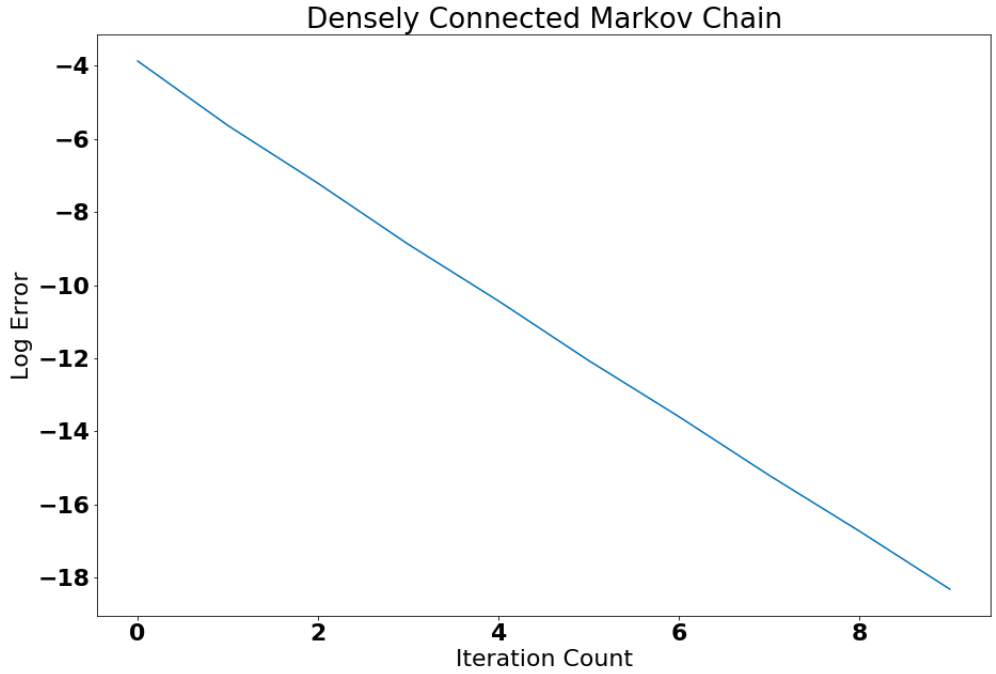
$$\mathrm{TV}(\pi_t, \pi^*) \leq C\alpha^t.$$

This is a fairly general result (and in fact implies our earlier Basic Limit Theorem, at least for finite Markov chains). However, there are some drawbacks to this generality. We only know of the existence of constants C, α but do not really have good quantitative bounds. In order, to see why this might be a problem lets consider two extreme cases:

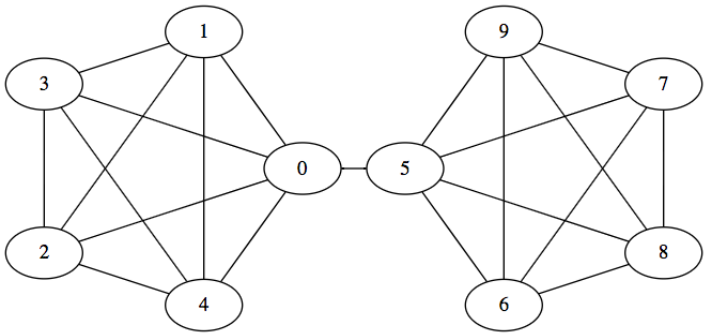
1. A dense, "well-connected" Markov Chain: Here we can go from each state to every other state with some non-trivial probability.



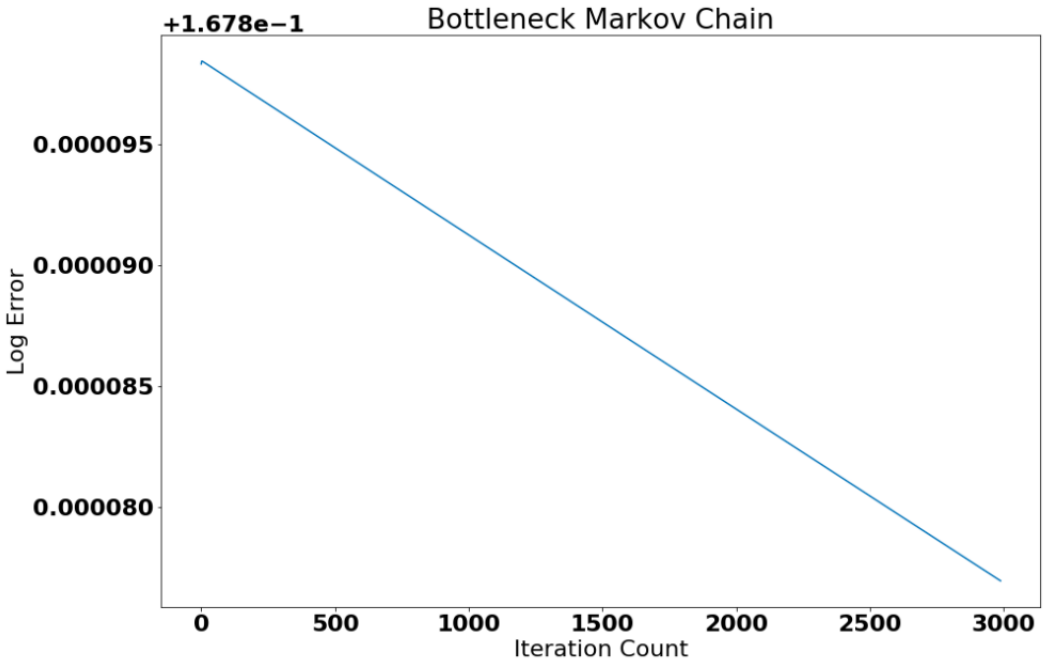
In this case, we can plot the total-variation distance versus time on a log-log scale. Notice that the basic convergence theorem assures us that this plot will be linear with slope $\log(\alpha)$.



2. A "dumbell" Markov Chain with a bottleneck: Intuitively, we see that



In this case, again we have a linear plot but notice that its slope is much closer to 0, and the error goes down very slowly.



What all of this means is that even though the basic convergence theorem is useful, we still need ways to understand what exactly the constant α is to be really sure that the stochastic process converges to its limiting distribution quickly.

The rest of this section is strictly optional, and gives a proof of the Basic Convergence Theorem.

Proof: Since the Markov Chain is finite, and irreducible we know that there must exist an $r > 0$ such that P^r has all positive entries, i.e. eventually we must have a non-zero probability of being in each state of the Markov Chain.

Let Π denote a matrix which has identical rows, each equal to the unique limiting distribution π^*. For some sufficiently small $\delta > 0$ it must be the case that, $P^r_{ij} \geq \delta \pi^*_j$ or equivalently,

$$P^r \geq \delta \Pi^*.$$

Let $\theta := 1 - \delta$, where since $\delta > 0$, we have that $\theta < 1$. Noting that both P^r and Π^* have rows that sum to 1 (and are all positive) we obtain that there must be a matrix $Q \geq 0$ whose rows

sum to 1, such that,

$$P^r = (1-\theta)\Pi^* + \theta Q.$$

Now by induction we can see that, for any integer $k \geq 1$ we claim that,

$$P^{rk} = (1-\theta^k)\Pi^* + \theta^k Q^k.$$

The base case, $k = 1$ is straightforward. Assuming it to be true for $k = n$ we simply need to verify this for $k = n+1$, i.e.

$$\begin{aligned} P^{r(n+1)} &= \left[(1-\theta^n)\Pi^* + \theta^n Q^n\right] P^r \\ &= \theta^{n+1}Q^{n+1} + (1-\theta^n)\Pi^* P^r + \theta^n(1-\theta^n)Q^n\Pi^*. \end{aligned}$$

Since the rows of Q sum to 1 we have that $Q^n\Pi^* = \Pi^*$. On the other hand since $(\pi^*)^T P = (\pi^*)^T$ we obtain that, $\Pi^* P^r = \Pi^*$. From this we see that,

$$\begin{aligned} P^{r(n+1)} &= \theta^{n+1}Q^{n+1} + (1-\theta^n)\Pi^* + \theta^n(1-\theta^n)\Pi^* \\ &= (1-\theta^{n+1})\Pi^* + \theta^{n+1}Q^{n+1}, \end{aligned}$$

proving the inductive claim.

Now, we can see that,

$$P^{rk+j} = (1-\theta^k)\Pi^* P^j + \theta^k Q^k P^j,$$

i.e. that

$$P^{rk+j} - \Pi^* = \theta^k \left[Q^k P^j - \Pi^*\right].$$

Now lets suppose that $X_0 = i$, i.e. we begin in state i for an arbitrary i, then summing the i-th row of this equation in absolute value and dividing by 2 gives us what we need. Let us first note that on the right hand side, $Q^k P^j$ is the product of two matrices with positive entries whose rows sum to 1, so it is also a matrix with positive entries whose rows sum to 1. So any row of $Q^k P^j - \Pi^*$ corresponds to the difference between two distributions, and so if we sum the absolute value and divide by 2 we get a number less than or equal to 1, i.e. for any i,

$$\mathrm{TV}([Q^k P^j]_i, \pi^*) \leq 1,$$

since $Q^k P^j$ is a distribution. Now we obtain that,

$$\mathrm{TV}(\pi_{rk+j}, \pi^*) \leq \theta^k,$$

from which we obtain the Basic Convergence Theorem in a simple way.

12.5 Spectral Conditions

Returning to the frog example, let us note two facts:

1. The Markov Chain is reversible (it is easy to verify that it satisfies detailed balance).

2. The transition matrix P has the following left eigenvectors:

$$\left(\frac{q}{p+q}, \frac{p}{p+q}\right)^T, (-1/2, 1/2)^T,$$

with corresponding eigenvalues of 1 and $1-p-q$.

Some of these facts are always true, i.e. the leading left eigenvector of any finite, irreducible, aperiodic Markov chain will be the limiting distribution, and it will always have an eigenvalue of 1. The other eigenvalues are always strictly smaller than 1 and trapped between $(1, -1]$.

Suppose we take the absolute value of the eigenvalues of P, then the second absolute eigenvalue (the one closest to 1 in magnitude) is special – it is known as the Fiedler eigenvalue or the Cheeger eigenvalue, i.e.

$$\lambda_\star = \max\{|\lambda| : \lambda \text{ is an eigenvalue of } P, \lambda \neq 1\}.$$

So for the frog Markov Chain $\lambda_\star = p + q$. We also define $\pi_{\min} = \min_{i=1}^k \pi^*(i)$ to be the smallest value of the limiting distribution. We have the following general fact:

Fact 12.2. For any reversible, finite, irreducible, aperiodic Markov chain:

$$\mathrm{TV}(\pi_t, \pi^*) \leq \frac{1}{\pi_{\min}}(1-\lambda_\star)^t.$$

Again for the frog Markov chain we almost recover the result we obtained by direct calculation. On the other hand in general it tells us something quite nice in general. If we have a (reversible) Markov chain where there is a small spectral gap (roughly, it is difficult to partition the Markov chain into two "separate" Markov chains) then the convergence is quite fast, and we have some quantitative upper bound on α.

12.6 Coupling

This section is also completely optional, and will not be on any exam. The last way that we will study involves an idea known as coupling. Before we get there let us try to understand the solution to a simple probability puzzle:

Suppose we consider a random walk with barrier on $\{0, 1, \ldots, n\}$ where at each time step we either move up or down by 1 (with equal probability). If we try to reach either -1 or $n+1$ we just stay put. It seems intuitive that for two states $x \leq y$, the following must be true:

$$P^t(x, n) \leq P^t(y, n),$$

i.e. the probability of reaching n after t time-steps must be higher if we start in a higher state. How might we try to prove this? One way is to try to count paths that go from x to n in t time-steps and then try to argue that there are more paths from y to n in t time-steps. This is quite tedious to do. Another way involves *coupling*.

Suppose we imagined two random walks, one that starts at x and another that starts at y but they both *make the same moves*, i.e. the either both attempt to move up or they both attempt to move down at any time step.

Let us denote the first random walk Markov chain by $X_0, \ldots, X_t$ and the second by $Y_0, \ldots, Y_t$. Our question then just boils down to asking: is it the case that:

$$P(X_t = n) \leq P(Y_t = n),$$

and we can see that this is clearly true.

Now, why might this be useful in reasoning about convergence to stationarity? Imagine the following thought experiment: we start two copies of the Markov chain – one from the stationary distribution π^* and the other from some arbitrary initial state.

These two Markov chains evolve independently until the first time they meet, i.e. they land in the same state, and then from that time on they evolve together (i.e. they make the same moves). It is easy to see that for one of these Markov chains the distribution after t time-steps is precisely π_t and for the other Markov chain it is π^*.

Fact 12.3. Basic coupling theorem:

$$\mathrm{TV}(\pi_t, \pi^*) \leq \mathbb{P}(\tau_{\text{couple}} > t)$$

Intuitively, this is saying that suppose we started two different copies of the Markov Chain (from different initial states) and they quickly converged to the same state then the Markov Chain is rapidly mixing.

There are many examples of this type of argument, but bounding the coupling time usually requires some more advanced probability theory. If you are really curious see the book by (Levin, Peres and Wilmer) which is entirely devoted to using coupling to prove bounds on the total-variation distance for many different Markov chains.

While the whole section is optional, the proof is even more optional.

Proof: The first thing we need to know is that the Total Variation distance has many probabilistic interpretations. One of them is in terms of couplings. In particular, suppose that we want to understand the total variation distance between two distribution μ and ν.

Suppose we have a pair of random variables X and Y with a joint distribution ω such that,

$$\mu(x) = \sum_y \omega(x, y) \quad \forall \ x,$$
$$\nu(y) = \sum_x \omega(x, y) \quad \forall \ y,$$

i.e. the joint distribution has X and Y marginals μ and ν respectively. These joint distributions, with the correct marginals, are known as couplings. Now, one can show that the total variation distance between ν and μ is given by:

$$\mathrm{TV}(\mu, \nu) = \inf_\omega \mathbb{P}_{(X,Y)\sim\omega}(X \neq Y).$$

Intuitively, this is just saying that if we can find a joint distribution with the "right" marginals so that under this joint distribution X and Y are usually equal then we can conclude that the Total Variation between the two marginals is small.

On the other hand this also means that for any coupling θ of μ and ν we have that,

$$\mathrm{TV}(\mu, \nu) \leq \mathbb{P}_{(X,Y)\sim\theta}(X \neq Y).$$

This is interesting because in order to bound the Total Variation we just need to construct a "good coupling", i.e. one for which the probability that the coupled random variables disagree is small.

Now, returning to Markov chains, for any two starting distributions, π_0 and π_0' lets suppose we run the corresponding Markov chains $\{X_t\}$ and $\{Y_t\}$ in parallel independently until the first time that they meet at time τ_{couple} and once that happens the two chains evolve together. Notice that this is a coupling, in the sense that if you just watched the X_t process it would look like a chain that started in π_0 and if you just watched the Y_t process it would look like a chain started in π_0', i.e. the marginal distributions are correct. Now we have that, after t steps:

$$\text{TV}(\pi_t, \pi_t') \leq \mathbb{P}(X_t \neq Y_t) = \mathbb{P}(\tau_{\text{couple}} > t).$$

Now taking one of these two distributions (say π_0') to be π^* we recover the coupling theorem.

Chapter 13

Martingales

In this chapter we will study martingales, which roughly attempt to model what are called *fair wagers*. Before we dive into the formal definitions lets just think about this intuitively: what does it mean to be fair? One reasonable criterion is that on average no matter how we bet we should come out even. In mathematical terms, we are just saying that our expected winnings should be 0. It turns out that this is necessary but we still need to go a bit deeper. In particular, we need that the expected winnings should be 0 at the *time of the bet*.

To understand the distinction between these two things consider the following scenario:

- We toss a fair coin. If it comes up heads, we toss a second coin which is unfair and has probability of heads $2/3$. If it instead comes up tails, we toss a second coin which is also unfair and has probability of tails $2/3$.

Say we were going to wager on the second outcome.

If we do this, *before the first toss*, then it is clear that it is a fair wager and we cannot make any money. However, if we do this *after the first toss* then it is clear that the odds are in our favor and it is no longer a fair wager.

To maintain fairness, we need some way of saying that the expected winnings should be 0, given *all the information* we have until the time of the bet. This is precisely a martingale, and we will define it more formally soon.

13.1 More Formal Definition

A completely formal definition (really, even to define conditional expectations) requires a lot of definitions and work. We will do it semi-formally.

Suppose we want to say that some stochastic process $\{X_0, X_1, \ldots\}$ is a *martingale*. At time t, to define the information we have until the time of the bet, we will either use the X stochastic process, i.e. $\{X_0, \ldots, X_{t-1}\}$ or sometimes by the X stochastic process and another auxiliary stochastic process $\{Y_0, \ldots, Y_{t-1}\}$. We will denote this information available at time t by $\mathcal{F}_{t-1}$.

Now that we have defined the *information available* the definition of the martingale is simply that:

$$\mathbb{E}[X_t - X_{t-1} | \mathcal{F}_{t-1}] = 0,$$

i.e. our expected winnings at time t conditional on all the information we have at time t is 0. Noticing that X_{t-1} is part of the information available at time $t-1$ we have that,

$$\mathbb{E}[X_{t-1}|\mathcal{F}_{t-1}] = X_{t-1},$$

and so the martingale definition can be written as:

$$\mathbb{E}[X_t|\mathcal{F}_{t-1}] = X_{t-1}.$$

This captures our notion of conditional fairness that we described earlier.

Martingales are important for various reasons in finance (roughly, in an efficient market all decisions should be fair so martingales form the basis for modelling efficient markets), and in probability theory or statistics (roughly, martingale-winnings, i.e. $X_t - X_{t-1}$ are stochastic processes that behave very much like independent zero-mean random variables).

13.2 The optional stopping theorem

The optional stopping theorem formalizes the intuition that if every wager is fair then there should be no strategy that allows a gambler to make money.

Formally, suppose we have a gambler playing a martingale game, i.e. the martingale X_t at time t models the amount of wealth the gambler has at time t. The gambler plays for some time τ, and then based on what he has already seen, i.e. $\{X_0, \ldots, X_\tau\}$ decides to stop at time τ (following some type of stopping rule).

Fact 13.1. If the stopping time τ is *finite* with probability 1, then it must be the case that,

$$\mathbb{E}[X_\tau] = X_0.$$

This fact is known as the optional stopping theorem.

To see what this means, think of the simplest martingale game. A gambler starts with 0 dollars (in winnings), and at each time bets 1\$. He can choose to stop at any point:

1. Say after 100 rounds.

2. If he wins 1\$ or after 1 million rounds have passed,

and so on. The optional stopping theorem says that no matter what rule the gambler follows (as long as he plays for a finite number of rounds), at the end of the day in expectation he will have 0\$.

It is worth pondering over why the following rule: stop when you have 1\$ (which guarantees a profit of 1\$) is not valid. In particular, you should observe that this stopping rule can lead to the gambler never stopping, because there is actually a non-zero probability that he never wins a dollar (overall). In Markov Chain terminology, the state 1 in this Markov chain, is recurrent but not positive recurrent (so the expected return time to state 1 can be infinite).

The optional stopping theorem can give us slick ways to prove some things about certain Markov chains. In particular, the gambler's ruin Markov chain or various balanced random walk Markov chains.

Example 13.1. Suppose we consider the gambler's ruin problem described above, and want to calculate the probability that the gambler wins a$ before losing b$ (i.e. reaches the state a before $-b$).

We have solved similar problems in simple cases using the idea of making the Markov chain finite absorbing. However, in this case it will be difficult to directly apply that technique. Instead we can use the optional stopping theorem.

Suppose the gambler stops when he either wins a$ or loses b$. This is a valid stopping rule so in expectation it must yield 0 winning. So let us denote the probability of winning a$ by p then we have that:

$$0 = \mathbb{E}[\text{winning}] = p \times a + (1-p) \times (-b),$$

i.e. we see that

$$p = \frac{b}{a+b}.$$

13.3 Examples

To see the definition of a martingale in action let us consider a few examples.

Example 13.2. Random Walk Martingale. Suppose we consider $X_1, X_2, \ldots$ such that for each i,

$$X_i = \begin{cases} +1 & \text{with probability } 1/2 \\ -1 & \text{with probability } 1/2. \end{cases}$$

Consider the associated random walk, i.e. $S_0 = 0$ and

$$S_i = \sum_{j=1}^{i} X_i.$$

Show that the sequence S_t is a martingale with respect to the sequence $X_1, X_2, \ldots$.

We check the definition,

$$\begin{aligned} \mathbb{E}[S_t|\mathcal{F}_{t-1}] &= \mathbb{E}[S_t|X_1, \ldots, X_{t-1}] = \mathbb{E}[X_t + \sum_{j=1}^{t-1} X_j | X_1, \ldots, X_{t-1}] \\ &= \mathbb{E}[X_t] + \mathbb{E}[S_{t-1}|X_1, \ldots, X_{t-1}] = 0 + S_{t-1}, \end{aligned}$$

as desired.

Example 13.3. Multiplicative Martingale. Suppose we consider a sequence of random

variables X_t such that,

$$\begin{aligned} X_0 &= 1, \\ X_1 &= Y_1 \\ X_2 &= X_1 \times Y_2 \\ X_3 &= X_2 \times Y_3, \\ &\vdots \end{aligned}$$

where each Y_i is independent and has mean 1. Show that the stochastic process X_t is a martingale.

We simply check the definition:

$$\begin{aligned} \mathbb{E}[X_t|\mathcal{F}_{t-1}] &= \mathbb{E}[X_t|X_0, \ldots, X_{t-1}] \\ &= \mathbb{E}[X_{t-1} \times Y_t|X_0, \ldots, X_{t-1}] \\ &= X_{t-1}\mathbb{E}[Y_t] = X_{t-1}. \end{aligned}$$

Example 13.4. Branching Process Martingale

Example 13.5. Doob Martingale

Example 13.6. Polya's Urn

Chapter 14

Brownian Motion

In this chapter we will study the most famous example of a stochastic process in continuous time with a continuous state space.

In some ways it is the most natural, but also the most mysterious. It has several extremely strange-at-first-sight properties, but hopefully we will see why they are all very natural. Surprisingly, the least scary thing about Brownian motions is its formal definition:

Fact 14.1. A standard Brownian motion (SBM) $\{W(t) : t \geq 0\}$ is a stochastic process having

1. continuous paths, (with probability 1)
2. stationary, independent increments, and
3. $W(t) \sim N(0, t)$ for all $t \geq 0$.

To elaborate on the second property: it requires that for any window $[s, t]$ the distribution of $W(t) - W(s)$ is just a function of the window length $t - s$ (stationarity). Furthermore, if we consider two disjoint intervals $[s_1, t_1]$ and $[s_2, t_2]$ then the two increments $W(t_1) - W(s_1)$ is independent of $W(t_2) - W(s_2)$.

An important point: In general, when you specify a stochastic process in this indirect way, it is not obvious if the specification makes sense. David Freedman famously said: "One of the leading results on Brownian motion is that it exists."

Another historical remark: Brownian motions, were originally described by a botanist Robert Brown, although maybe really by Einstein. We often use the letter W in honor of Norbert Wiener.

Just intuitively, wanting continuous paths *and* independent increments is the cause for most of the difficulty, i.e. consider the following continuous time, continuous state stochastic process:

1. Sample $T_1, T_2, \ldots$ i.i.d. Exp(λ), and $Z_1, Z_2, \ldots$ i.i.d. N(0,1). Define, $S_1 = T_1, S_2 = T_1 + T_2, \ldots$, and $U_1 = Z_1, U_2 = Z_1 + Z_2, \ldots$.

2. Set:

$$W(t) = \begin{cases} 0 & \text{for} \quad 0 \leq t < T_1 \\ U_1 & \text{for} \quad T_1 \leq t < T_2 \\ U_2 & \text{for} \quad T_2 \leq t < T_3 \\ \vdots \end{cases}$$

Intuitively, this is just like a Poisson process, except each arrival brings a Gaussian random variable. This stochastic process satisfies all the conditions of the Brownian motion (it is a continuous time/space process, with stationary and independent increments) except that its sample paths are not continuous. Intuitively, requiring continuity is at odds with requiring independence and this is what causes lots of conceptual difficulties in dealing with Brownian motions. On the other hand one can reasonably expect many natural stochastic processes to have continuous sample paths and the Brownian motion is a very important addition to the toolbox.

14.1 A Brownian motion is the limit of simple random walks

The first thing to observe is that I could sample a Brownian motion at time $0, 1, 2, \ldots$, and look at the values $\{W(0), W(1), W(2), \ldots\}$ this is a familiar stochastic process. It is a particular kind of simple random walk.

We have that

$$\begin{aligned} W(0) &= 0, \\ W(1) &\sim N(0,1) \\ W(2) - W(1) &\sim N(0,1) \\ &\vdots \end{aligned}$$

So the discretized Brownian motion is a Gaussian random walk.

Suppose I just wanted to get a clearer picture of the first second, and sampled the process in intervals of length 0.01 instead of length 1. This is also a simple random walk.

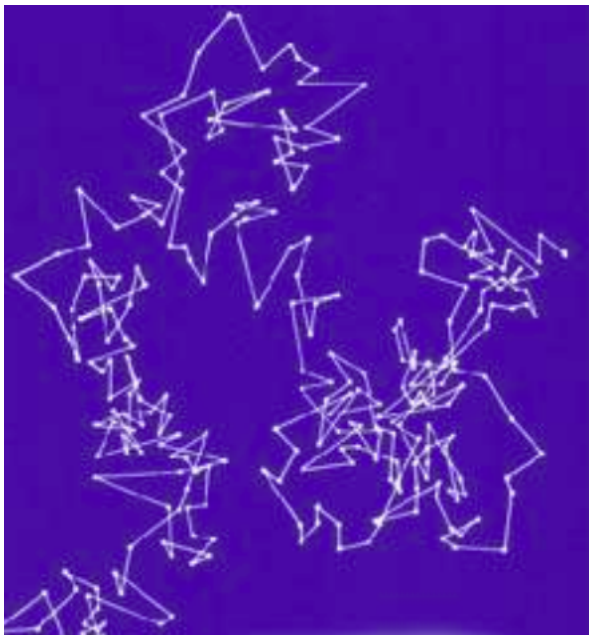
$$\begin{aligned} W(0) &= 0, \\ W(0.01) &\sim N(0, 0.01) \\ W(0.02) - W(0.01) &\sim N(0, 0.01) \\ &\vdots \end{aligned}$$

We could extend the sampling (linearly), and take finer and finer sampling to somehow get a more complete picture of a Brownian motion. Alternatively, you could define a Brownian motion in this way: it is the "limit" of the linear extension of appropriately scaled simple random walks. This is similar to the spirit in which a Poisson process was the limit of a Bernoulli process.

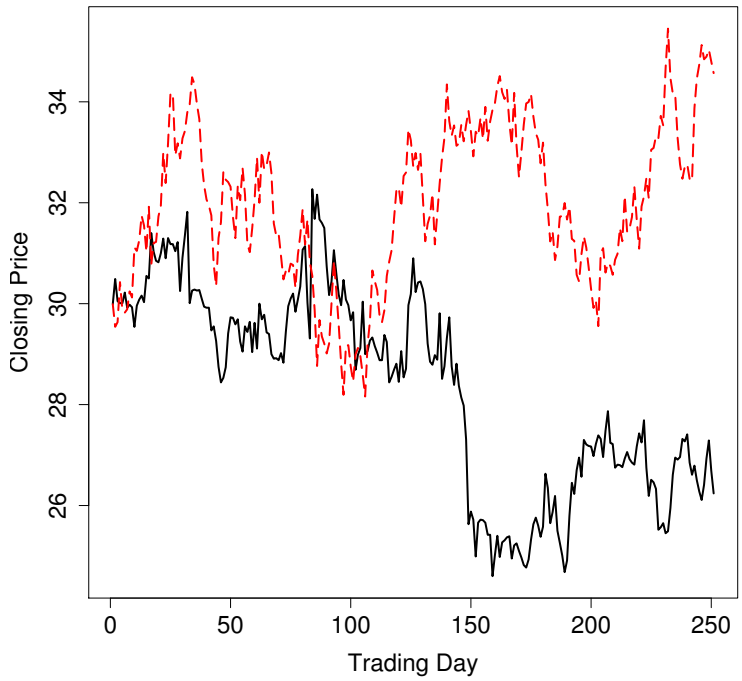
14.2 Examples

There are two canonical examples of Brownian motions. In general, when one hypothesizes a Brownian motion one also tries to give a physical interpretation of why this might be reasonable.

Example: Pollen particles: This was the motivating example for Brown (the botanist who lent his name to Brownian motion). He observed that pollen particles in water/air had a zig-zag path in 2D but could not quite explain why they were in constant motion. Einstein gave a formal explanation and argued that being constantly bombarded by environmental molecules caused the erratic motion, and formally described Brownian motion.



Example: Stock prices: Variants of Brownian motion are often used as natural models for stock prices. Roughly, the reasoning is that beyond some natural drift, stock prices go up and down erratically as trades are made.



14.3 Exercises

These are meant to give you some basic familiarity with the properties defining a Brownian motion.

Example 14.1. For a standard Brownian motion what is the distribution of $W(s)+W(t)$ for $s \leq t$?

While marginally $W(s)$ and $W(t)$ have a Gaussian distribution (with variance s and t) they are not independent. So we rewrite them in terms of independent RVs, i.e.,

$$W(s) + W(t) = 2W(s) + (W(t) - W(s)),$$

and note that the two terms are in fact independent. This leads to the conclusion that,

$$2W(s) + (W(t) - W(s)) \sim N(0, 4s) + N(0, t - s) \sim N(0, 3s + t).$$

Example 14.2. Calculate the covariance between $W(s)$ and $W(t)$.

Let us suppose that $s \leq t$. Noting that $\mathbb{E}[W(s)] = 0$ we have that,

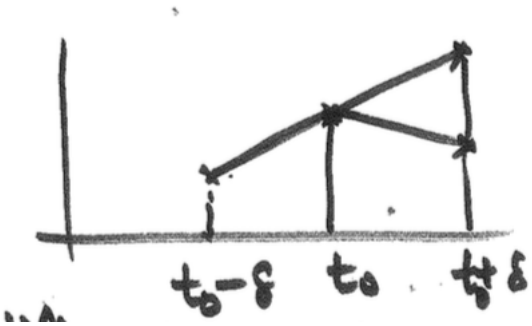
$$\begin{aligned}
\operatorname{cov}(W(s), W(t)) &= \mathbb{E}[W(s)W(t)] \\
&= \mathbb{E}[W(s)(W(s) + W(t) - W(s)] \\
&= \mathbb{E}[W^2(s)] + \mathbb{E}[W(s)(W(t) - W(s))] \\
&= s + \mathbb{E}[W(s)]\mathbb{E}[W(t) - W(s)] \\
&= s.
\end{aligned}$$

14.4 Properties of Brownian motions

Brownian motions have many curious properties. All of them roughly translate to: "unlike normal functions Brownian motions are extremely wiggly".

Fact 14.2. A Brownian motion, with probability 1 is nowhere differentiable.

It used to be widely believed that for continuous functions, points of non-differentiability (i.e. "kinks") were rare (i.e. countably infinite) and that continuous functions were differentiable almost everywhere. Surprisingly, Weierstrass showed this was not the case by constructing a continuous, nowhere-differentiable function. Brownian motions are an example of this type of function. We will try to see this intuitively.



In essence for a continuous function to be differentiable at some point t, we need that for some small δ the two increments: $W(t) - W(t-\delta) \approx W(t+\delta) - W(t)$ (otherwise there would be a kink at t). For a Brownian motion however these two increments are in fact two independent $N(0,\delta)$ random variables so the probability with which they exactly line up is 0.

Fact 14.3. A Brownian motion, with probability 1, exhibits quadratic variation, i.e.

$$\lim_{n\to\infty} \sum_{i=0}^{n} \left[W\left(\frac{iT}{n}\right) - W\left(\frac{(i-1)T}{n}\right) \right]^2 = T.$$

Before we prove this fact, we should try to understand why it is surprising. In order to do this let us first consider a smooth function f defined on $[0,T]$, i.e. one for which its derivative $|f'(x)| \leq M$ where $M > 0$ is some big constant.

Now, by Taylor's theorem we know that for this function,

$$\left| f\left(\frac{iT}{n}\right) - f\left(\frac{(i-1)T}{n}\right) \right| \leq \left[\max_x |f'(x)| \right] \times \frac{T}{n}.$$

Using this we can see that,

$$\begin{aligned} \lim_{n\to\infty} \sum_{i=0}^{n} \left[f\left(\frac{iT}{n}\right) - f\left(\frac{(i-1)T}{n}\right) \right]^2 &\leq \left[\max_x |f'(x)| \right]^2 \times \frac{T^2}{n^2} \\ &\leq \lim_{n\to\infty} \frac{M^2T^2}{n} = 0. \end{aligned}$$

So for any smooth function the quadratic variation is 0. Brownian motions are very *non-smooth*.

For a Brownian motion we see that,

$$W\left(\frac{iT}{n}\right) - W\left(\frac{(i-1)T}{n}\right) \sim N(0, T/n),$$

and these random variables are independent for each i. So we obtain that,

$$\lim_{n\to\infty} \sum_{i=0}^{n} \left[W\left(\frac{iT}{n}\right) - W\left(\frac{(i-1)T}{n}\right)\right]^2 = \lim_{n\to\infty} \sum_{i=0}^{n} Z_i^2,$$

where each Z_i is independent and identically distributed. By the law of large numbers we can see that,

$$\lim_{n\to\infty} \sum_{i=0}^{n} Z_i^2 \to n \times \mathbb{E}[Z_i^2] = n \times \frac{T}{n},$$

so we conclude that unlike smooth functions Brownian motions exhibit (non-zero) quadratic variation.

Fact 14.4. Fix any $\epsilon > 0$, a Brownian motion has infinitely many zeros on the interval $(0, \epsilon)$.

Fact 14.5. Extrema of Brownian motions, a.k.a. the reflection principle. For a Brownian motion,

For things like stock prices, you might care about what is the maximum and minimum values the stock price will take over a day. We will return to these facts in a subsequent lecture.

Fact 14.6. A Brownian motion is a special case of a Gaussian Process. Fix some number n, and suppose we take a SBM and sample it at n time points: $(t_1, \ldots, t_n)$, then the random vector $(W(t_1), W(t_2), \ldots, W(t_n))$ has a jointly Gaussian distribution, with mean 0, and covariance$(W(s), W(t)) = \min(s, t)$. These are more generally known as *Gaussian Processes*.

This fact follows from the expression for the covariance we derived earlier. This leads to the alternate definition of a Brownian motion:

A Brownian motion is:

1. A Gaussian process with,
2. continuous paths,
3. mean 0, and
4. covariance function $r(s, t) = \min(s, t)$ is a standard Brownian motion.

14.5 Understanding conditioning in a Brownian motion

14.5.1 Conditioning on a point in the future

Suppose we want to understand the behavior of a Brownian motion in between two sampled points. Concretely, we want to understand the distribution of $W(t)|W(u)$ for $0 < t < u$.

To begin, let us consider an exercise:

Example 14.3. For a SBM with $0 < t < u$, show that $W(t) - \frac{t}{u}W(u)$ is independent of $W(u)$.

Since these are (jointly) Gaussian RVs to show independence we simply need to check that the two random variables have zero correlation.

To see this notice that,

$$\begin{aligned}\mathbb{E}\left[\left(W(t) - \frac{t}{u}W(u)\right) \times W(u)\right] &= \mathbb{E}[W(t)W(u)] - \frac{t}{u}\mathbb{E}[W(u)^2] \\ &= \min\{t, u\} - \frac{t}{u} \times u \\ &= 0.\end{aligned}$$

Example 14.4. Using the above calculation find the mean $\mathbb{E}(W(t)|W(u))$.

To calculate this we note that,

$$\begin{aligned}0 = \mathbb{E}(W(t) - \frac{t}{u}W(u)) &= \mathbb{E}(W(t) - \frac{t}{u}W(u)|W(u)) \\ &= \mathbb{E}(W(t)|W(u)) - \frac{t}{u}W(u).\end{aligned}$$

Re-arranging this we obtain that:

$$\mathbb{E}(W(t)|W(u)) = \frac{t}{u}W(u).$$

So the mean falls on the line that joins 0 at time 0, to the point $W(u)$ at time u.

Example 14.5. Find the variance$(W(t)|W(u))$, and use this to conclude what its distribution is. Does the variance you calculated make sense?

To calculate the variance we see that,

$$\begin{aligned}\text{Var}(W(t)|W(u)) &= \mathbb{E}\left[(W(t) - \mathbb{E}(W(t)|W(u)))^2|W(u)\right] \\ &= \mathbb{E}\left[(W(t) - \frac{t}{u}W(u))^2|W(u)\right], \\ &= \mathbb{E}\left[(W(t) - \frac{t}{u}W(u))^2\right],\end{aligned}$$

using the independence property we derived above. This in turn yields that,

$$\begin{aligned}\mathrm{Var}(W(t)|W(u)) &= \mathbb{E}[W^2(t) + \frac{t^2}{u^2}W^2(u) - \frac{2t}{u}\mathbb{E}[W(t)W(u)] \\ &= t + \frac{t^2}{u^2} \times u - \frac{2t}{u} \times t \\ &= \frac{t(u-t)}{u}.\end{aligned}$$

Notice that this expression for the variance makes sense: it is 0 at 0, and 0 at the point u (since we know the values of the Brownian motion at those two points).

14.5.2 Conditioning on a point in the past

When conditioning on a point in the past the Markov property of the Brownian motion is relevant.

Fact 14.7. Brownian motions are both martingale and Markov.

Suppose that W is a standard Brownian motion, and let $c > 0$. Define, the new stochastic process: $X(t) = W(c+t) - W(c)$. Then $\{X(t) : t \geq 0\}$ is a standard Brownian motion that is independent of $\{W(t) : 0 \leq t \leq c\}$. This is a way of formally saying that a Brownian motion is *Markov process.* Alternatively, we can say that for $t > c$:

$$W(t)|W(u), 0 \leq u \leq c \sim W(t)|W(c) \sim N\left(W(c), t-c\right).$$

The Brownian motion is also a martingale with respect to itself.

14.6 The important variants

1. Brownian motion with drift: Just like we can generate the family of Gaussian random variables by re-centering and scaling a standard Gaussian, we can generate the full family of Brownian motions by re-centering and scaling a standard Brownian motion appropriately. Formally,

   A process X is called a (μ, σ^2) Brownian motion if it can be written in the form

   $$X(t) = X(0) + \mu t + \sigma W(t),$$

   where W is a standard Brownian motion.

   A fact to keep in mind is that this family of stochastic processes is essentially the only family that satisfies the properties of our earlier definition, i.e.

   If a stochastic process X has continuous paths and stationary, independent increments, then X is a Brownian motion (with drift).

2. Brownian bridge: The standard Brownian bridge is a SBM, conditioned to have $W(1) = 0$. Intuitively, it is tied down at 0 and 1. We can use our earlier calculations to find the mean and covariance of the Brownian bridge. Formally,

   A **standard Brownian bridge** is a Gaussian process X with continuous paths, mean 0, and covariance function, $\mathrm{cov}(X(s), X(t)) = s(1-t)$ for $0 \leq s \leq t \leq 1$.

3. Geometric Brownian motion: The geometric Brownian motion is the stochastic process most commonly associated with stock prices. Intuitively, it is attempting to model a stochastic process where the *percentage change* is independent and identically distributed.

   A key quantity of interest for a Geometric Brownian motion is its expected value given the past, i.e. we are interested in calculating $\mathbb{E}(Y(t)|Y(u), 0 \leq u \leq s)$.

4. Integrated Brownian motion: If $\{W(t), t \geq 0\}$ is Brownian motion, then the process $\{Z(t), t \geq 0\}$ defined by

$$Z(t) = \int_0^t W(s)ds$$

   is called an *integrated Brownian motion.*

14.7 Some more Brownian motion problems

Example 14.6. Use the formula for extrema of a Brownian motion to calculate an expression for

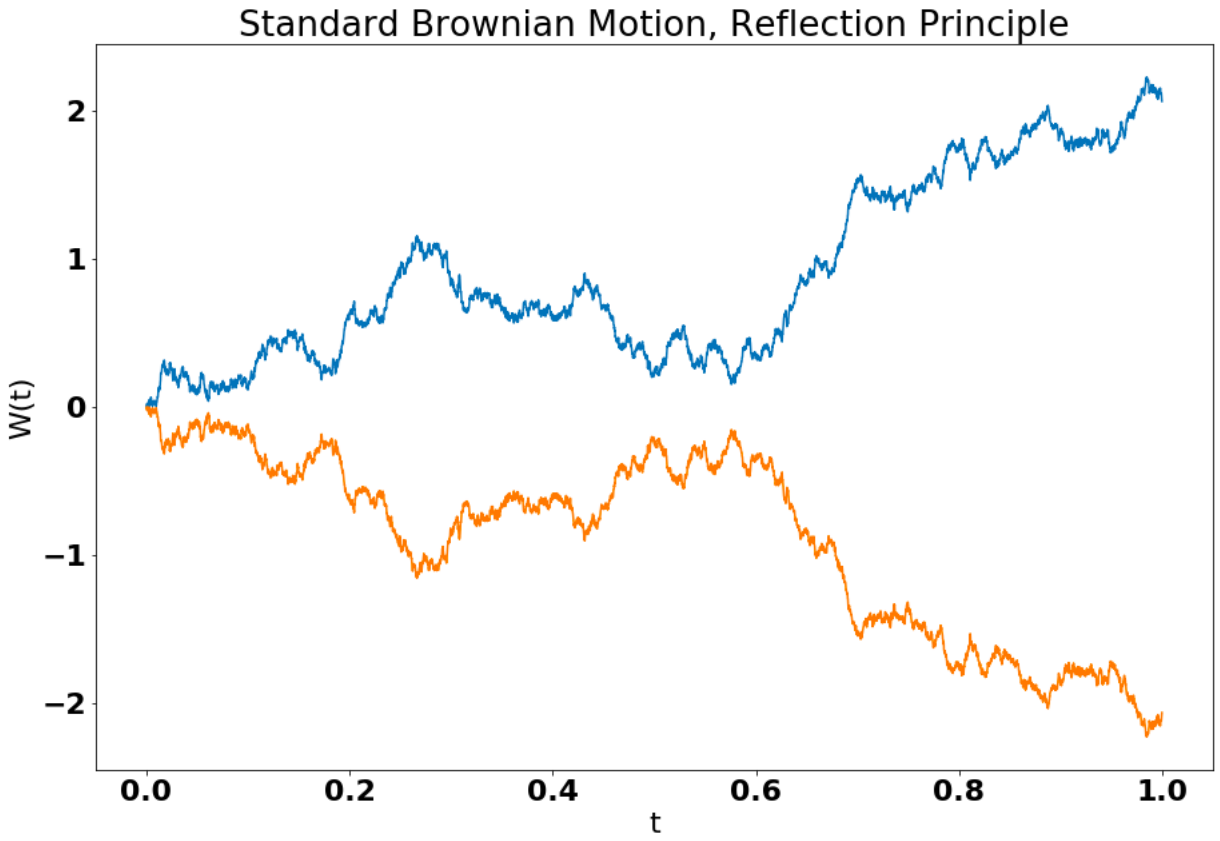
$$\mathbb{P}(\max_{t_1 \leq s \leq t_2} W(s) > x).$$

Example 14.7. One way to reason about a Brownian motion with drift is to think about a biased random walk. Argue informally, that a random walk with an infinitesimal bias, i.e. a random walk that in a time window of width δ goes up with probability $\frac{1}{2}(1+\mu\sqrt{\delta})$ by an amount of $\sqrt{\delta}$, converges to a Brownian motion with drift μ.

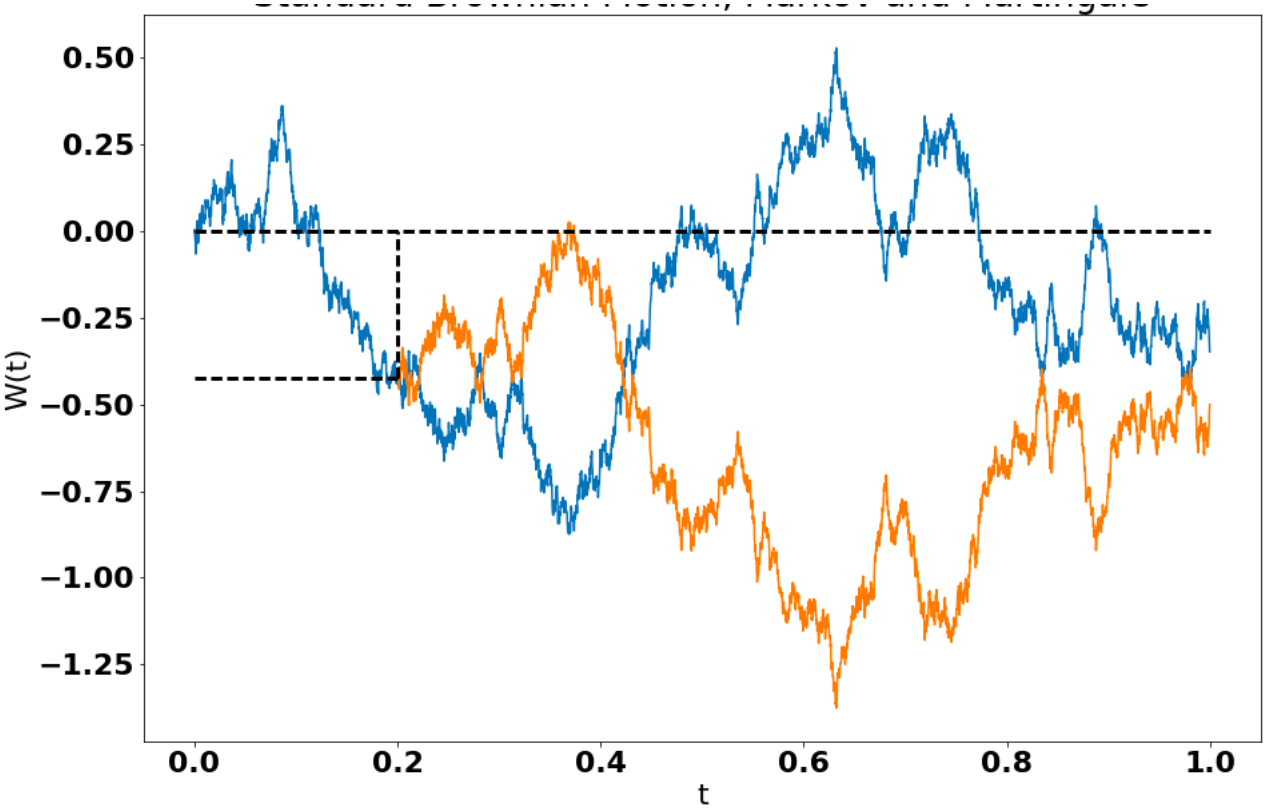
We begin today by using an important property of Brownian motions called the *reflection principle* to derive an important fact about the extrema of Brownian motions.

What does the reflection principle say? It says that if I condition on the value of a Brownian motion at some time t then any path the Brownian motion can take after time t has the same probability as its reflection.

Here are some pictures illustrating the *reflection principle*. In this figure, the reflection principle tells us the blue and orange paths are equally likely.



In this figure, the reflection principle tells us the blue and orange paths are equally likely. Here we additionally condition on the value of the Brownian motion at time 0.2.



Using the reflection principle we can prove the following fact:

Fact 14.8. For a Brownian motion we have the following,

$$\mathbb{P}\left(\max_{0\leq u\leq t} W(u) \geq a\right) = 2 \times \mathbb{P}(W(t) \geq a).$$

First, let us notice that this is an extremely useful expression. The right hand side here is very simple. We know that

$$W(t) \sim N(0, t),$$

so that

$$\mathbb{P}(W(t) \geq a) = \mathbb{P}(N(0,t) \geq a) = \mathbb{P}(N(0,1) \geq a/\sqrt{t}) = 1 - \Phi(a/\sqrt{t}).$$

Here Φ is the standard Normal CDF and is a quantity we understand very well.

To prove the stated fact let us define,

$$Z(t) = \max_{0\leq u\leq t} W(u),$$

and let us also define,

$$\tau_a = \min\{u : W(u) = a\},$$

as the first time the Brownian motion crosses the value a. Now observe that,

$$\mathbb{P}\left(\max_{0\le u\le t} W(u) \ge a\right) = \mathbb{P}(\tau_a \le t).$$

Using this we obtain that,

$$\begin{aligned}\mathbb{P}(\tau_a \le t) &= \mathbb{P}(W(t) < a, \tau_a \le t) + \mathbb{P}(W(t) \ge a, \tau_a \le t)\\ &= \mathbb{P}(W(t) < a|\tau_a \le t)\mathbb{P}(\tau_a \le t) + \mathbb{P}(W(t) \ge a),\end{aligned}$$

so if we can show that $\mathbb{P}(W(t) < a|\tau_a \le t) = 1/2$ then a simple re-arrangement yields the desired fact. This is a simple consequence of the reflection principle which tells us that,

$$\mathbb{P}(W(t) < a|\tau_a \le t) = \mathbb{P}(W(t) > a|\tau_a \le t),$$

and since they sum to 1 they must each $= 1/2$.

Example 14.8. Suppose I have several stocks $\{1, \dots, n\}$ each of which are modeled by a Brownian motion with different variances $\{\sigma_1^2, \dots, \sigma_n^2\}$, i.e. for the i-th stock the price $P_i(t) = \sigma_i \times W_i(t)$ where each W_i is an independent standard Brownian motion.

Compute the probability that over the time window $[0, t]$ *all* of the stock prices cross the value a.

Since each stock price is independent we can simply multiply the individual probabilities. Now for the i-th stock we have that,

$$\mathbb{P}\left(\max_{0\le u\le t} P_i(u) \ge a\right) = \mathbb{P}\left(\max_{0\le u\le t} W_i(u) \ge a/\sigma_i\right) = 2\left(1 - \Phi(a/(\sigma_i\sqrt{t}))\right).$$

Multiplying these together we obtain the desired result.

14.8 The important variants

1. Brownian motion with drift: Just like we can generate the family of Gaussian random variables by re-centering and scaling a standard Gaussian, we can generate the full family of Brownian motions by re-centering and scaling a standard Brownian motion appropriately. Formally,

   A process X is called a (μ, σ^2) Brownian motion if it can be written in the form

   $$X(t) = X(0) + \mu t + \sigma W(t),$$

   where W is a standard Brownian motion.

   A fact to keep in mind is that this family of stochastic processes is essentially the only family that satisfies the properties of our earlier definition, i.e.

   If a stochastic process X has continuous paths and stationary, independent increments, then X is a Brownian motion (with drift).

   Many properties of Brownian motions with drift (for instance, the conditional distribution of the process) can be computed in the same way we followed for the Standard Brownian motion.

Example 14.9. Suppose that X is a Brownian motion with drift and $s < t$, what is the distribution of $X(t) - X(s)$?

It is easy to check that $X(t) - X(s) \sim N(\mu(t-s), \sigma^2(t-s))$, since:

$$\begin{aligned} X(t) - X(s) &= X(0) + \mu t + \sigma W(t) - X(0) - \mu s - \sigma W(s) \\ &= \mu(t-s) + \sigma(W(t) - W(s)), \end{aligned}$$

and now the claim follows using the properties of W.

2. Brownian bridge: The standard Brownian bridge is a SBM, conditioned to have $W(1) = 0$. Intuitively, it is tied down at 0 and 1. We can use our earlier calculations to find the mean and covariance of the Brownian bridge. Formally,

Example 14.10. A **standard Brownian bridge** is a Gaussian process X with continuous paths, mean 0, and covariance function, $\text{cov}(X(s), X(t)) = s(1-t)$ for $0 \leq s \leq t \leq 1$.

This just follows by properties we have seen already, i.e. the standard Brownian bridge is simply a standard Brownian motion conditioned on $W(1) = 0$.

3. Geometric Brownian motion: The geometric Brownian motion is the stochastic process most commonly associated with stock prices. Intuitively, it is attempting to model a stochastic process where the *percentage change* is independent and identically distributed.

   Formally, if $Y(t)$ is a Brownian motion (possibly with drift μ and variance σ^2) then the process

$$X(t) = \exp(Y(t)),$$

   is a *Geometric Brownian Motion*.

Example 14.11. Suppose we discretize a GBM, i.e. we observe $\{X(0), X(1), \ldots\}$ then argue that the resulting stochastic process has independent *ratios*.

Formally, by independent ratios we mean that for disjoint windows $[i, j]$ and $[k, l]$ we have that,

$$\frac{X(j)}{X(i)} \perp\!\!\!\perp \frac{X(l)}{X(k)}.$$

This follows directly from the independent increments property of Y, i.e. since

$$Y(j) - Y(i) \perp\!\!\!\perp Y(l) - Y(k),$$

it follows that their exponentials are also independent.

14.9 Gaussian Process Regression

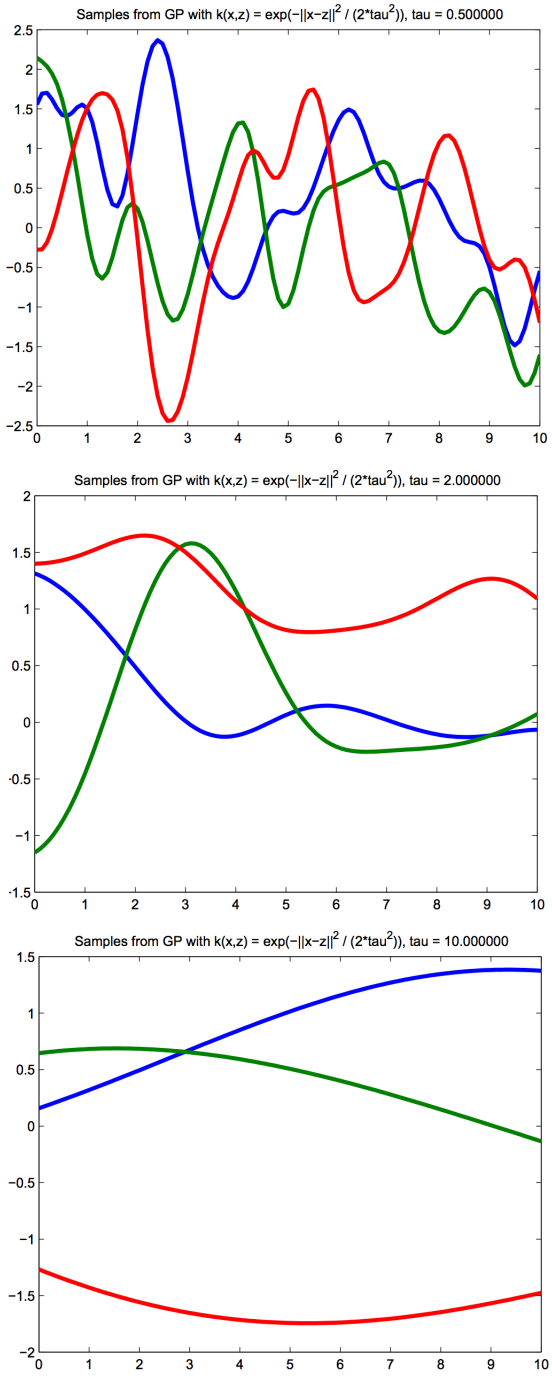
Brownian motions are somewhat strange beasts. One way to think about them is as a distribution over continuous functions, but it is a distribution that puts all of its mass on continuous functions that are extremely wiggly.

Brownian motions are excellent ways to model *noise* or purely stochastic processes. However, in machine learning/statistics we often have a different goal. We want to model some signal so we want a stochastic process that puts most of its mass on "nice" functions, and we want to be able to model our data using this stochastic process (i.e. not have it model just the noise). It will turn out that Gaussian processes, which generalize Brownian motions, are a great tool for this.

Let me skip to the punchline: in a Gaussian process, if we assume the covariance function, is smooth (think differentiable) then the Gaussian process will put its mass on smooth functions. We can then try to use these smooth functions in machine learning problems.

First, lets try to recall what we need to specify for a Gaussian process.

I made a claim, that if we use smooth covariance functions, then the Gaussian process will produce sample paths that are smooth functions. So you have a picture in mind:



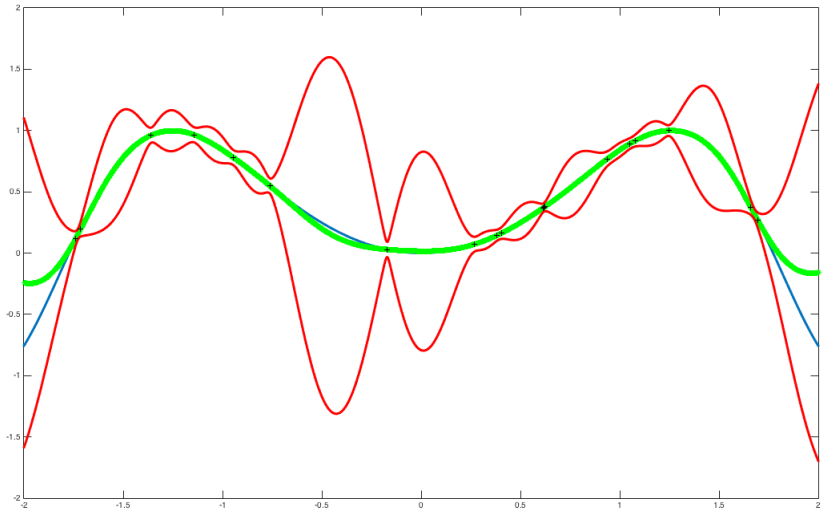
There is a very critical caveat here: Gaussian processes in general do not satisfy the Markov property, do not have stationary/independent increments etc.

We are interested in solving a canonical problem in statistics/machine learning: the problem of regression:

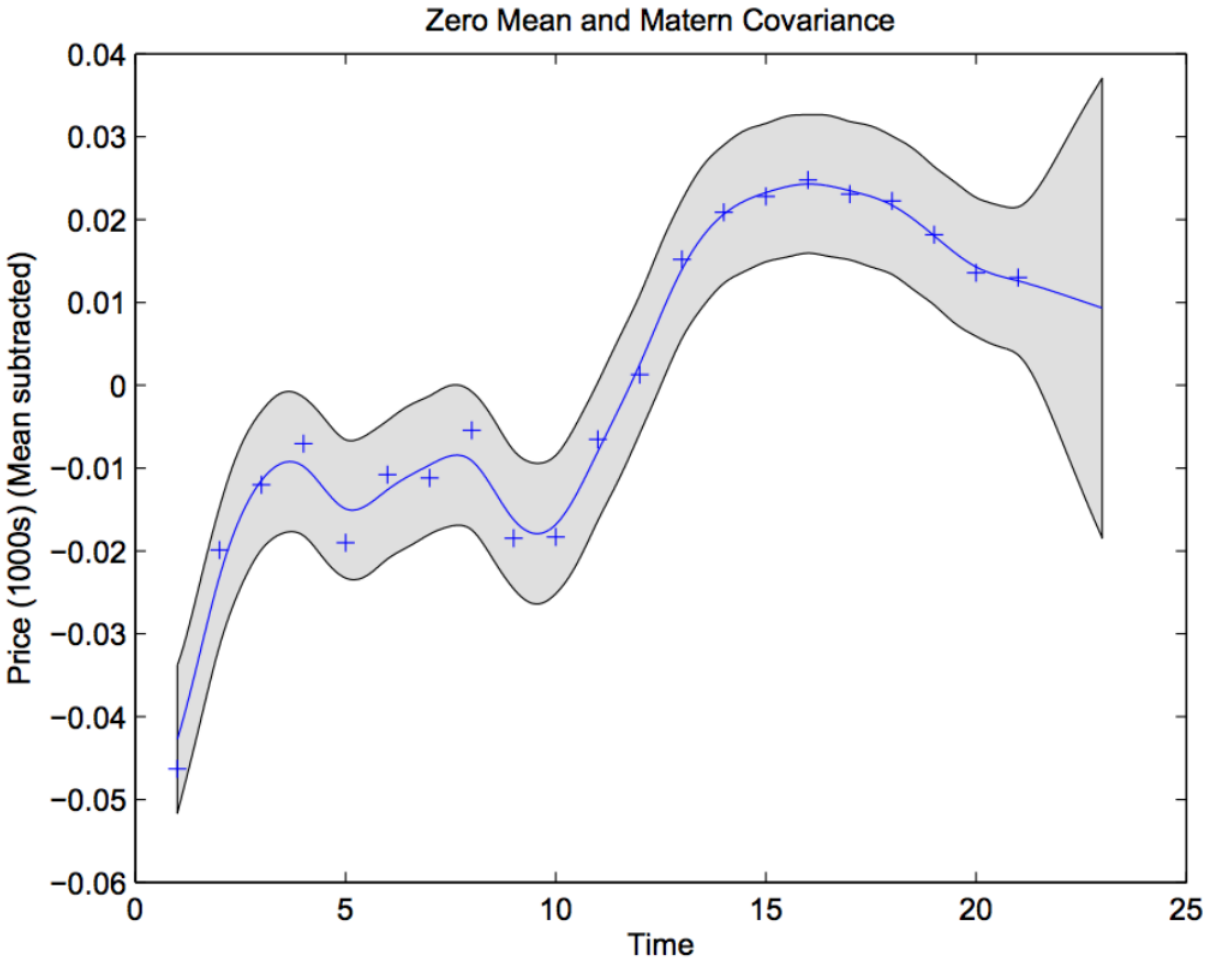
A brief description of non-parametric regression:

In GP regression, we model the data as a GP which we observe at some points and then want to predict at some new points. We will not do this in detail but you can imagine that just like we conditioned in a Brownian motion, we can determine the conditional distribution in a general GP, and use this conditional distribution to estimate the function value at new points (interpolation/extrapolation).

Notice that this gives us both means and standard errors. Here is an illustration:



A simple finance example: Here is an illustration of GP regression. The data is Google stock closing prices in November 2010.



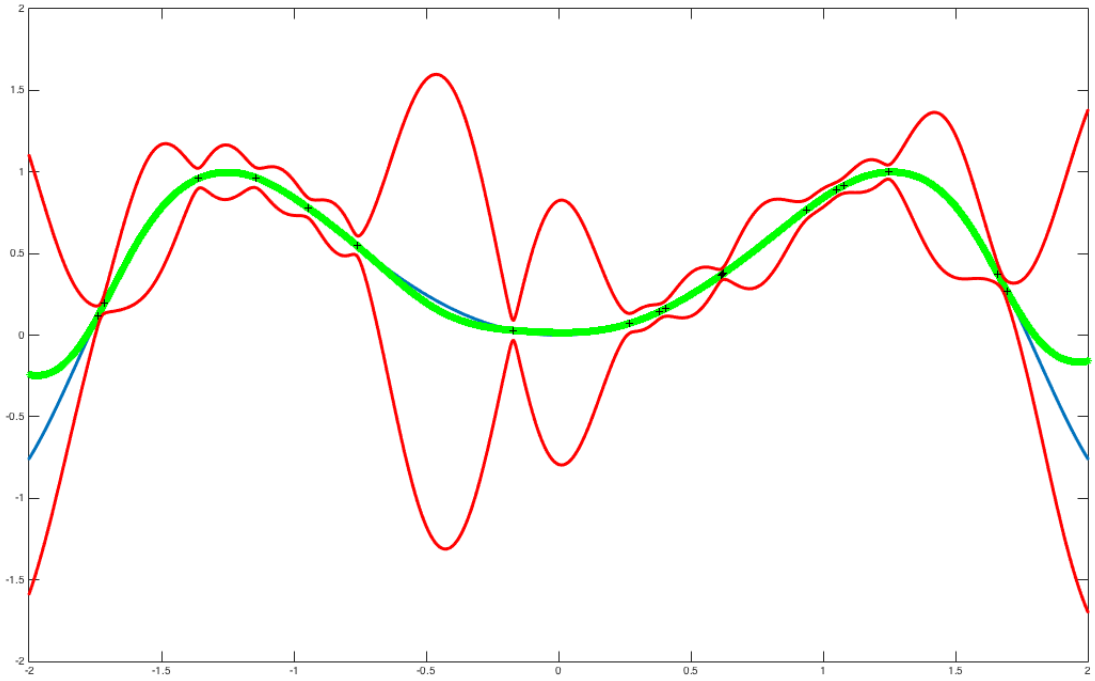
You could imagine trying to predict the closing price on some day in the near future, and could use GP regression for this task.

Maybe importantly, over these wide time-scales there is no particular reason to think that a Brownian motion/Geometric Brownian motion is a good model. The same is true for a linear model.

14.10 Applications

It is actually very nice that we have both an estimate of the function, and some notion of uncertainty. As an illustration let us consider two things you could do with GP regression: (1) Active Learning, (2) Zero-order optimization.

Looking at the same picture again:


2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2

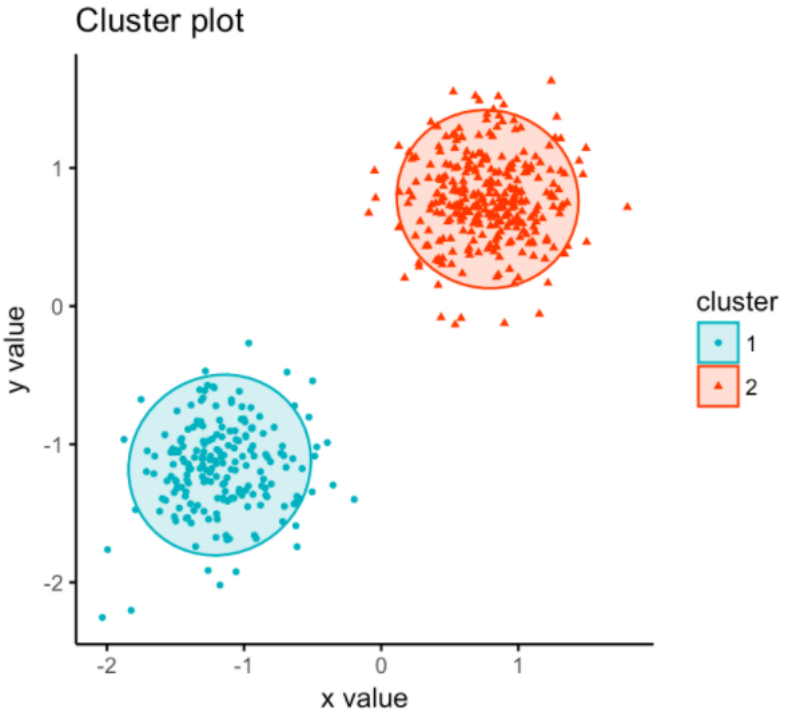
Chapter 15

Hidden Markov Models

For this part we will at least roughly follow the Ross book. The Hidden Markov Model (HMM) is another modification of a Markov Chain that is very widely used in practical applications.

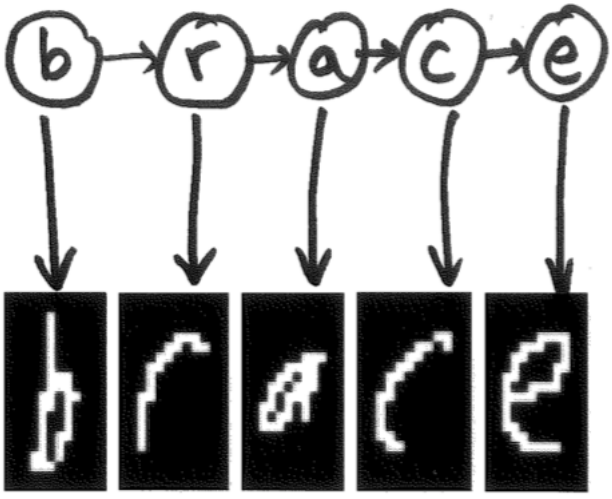
In a very broad sense, a Hidden Markov Model is an example of something known as a latent variable model in statistics. The idea of latent variable models is very simple, I have some complex and messy data that I want to understand. One way to do this, is to model it as some "simpler" data corrupted by noise, and then try to recover this "simpler" representation.

A canonical example of this is *clustering*. In clustering, we observe data (clouds of points), and we imagine that these points are *noisy observations* of a few points (cluster centers) and we just want to recover this noiseless representation, i.e. the cluster memberships.



In an HMM the "simpler" data is a Markov chain. So we observe some noisy version of the states of the Markov chain and want to decipher the true state.

A canonical application is in handwriting recognition:



15.1 What precisely is an HMM?

To define a Markov chain we needed to specify two things: the distribution of the initial state π_0, and the transition matrix P_{ij}.

To define an HMM we define three things:

- π_0
- The transition matrix $P_{ij} = P(X_{n+1} = j | X_n = i)$
- The observation distributions: $P(E_n | X_n = i)$ for each i

The Markov assumption in a HMM is that:

- E_{i-1} is independent of E_{i+1} conditional on X_i,
- the usual X_{i-1} is independent of X_{i+1} conditional on X_i

What are the questions of interest in an HMM? There are two questions that we typically try to answer:

1. *Decoding:* Given a sequence of noisy observations $\{E_1, \ldots, E_n\}$ we want to know what are the *most likely* $\{X_1, \ldots, X_n\}$.
2. *Estimation:* Just as we estimated parameters of a Markov chain from data, it is common to *train* a HMM, i.e. to estimate the transition matrix and observation distributions from data.

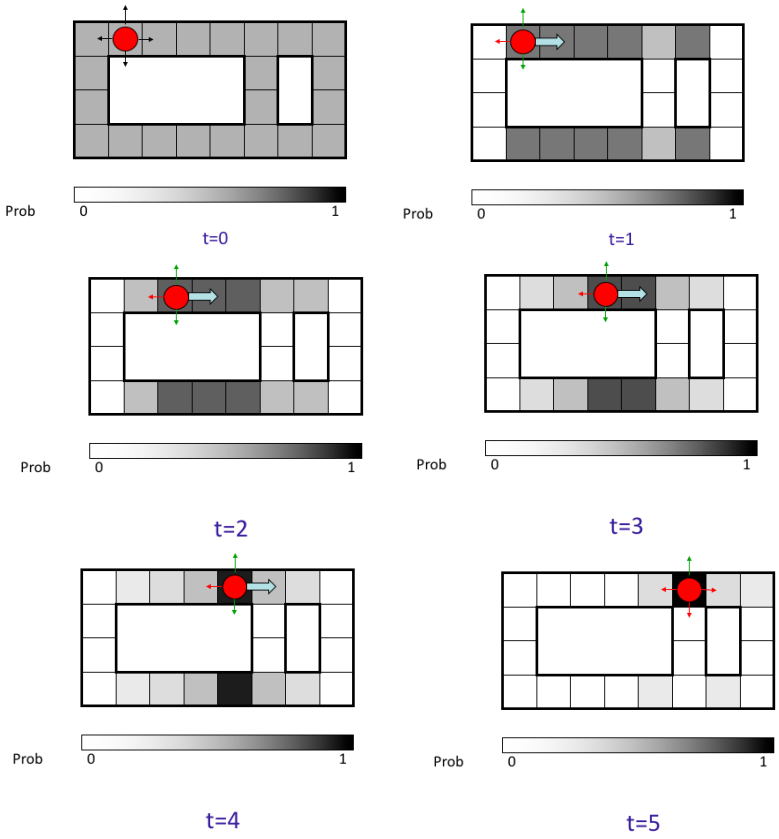
We will only focus on the first one.

Each of these tasks could be absurdly computationally intractable: there are exponentially many possible sequences that we need to search over to find the most likely one. The key point that we will explore in the next lecture is that by exploiting the Markov property of the Markov chains we will be able to design a fast dynamic programming algorithm for decoding. This algorithm is known as the *Viterbi* algorithm.

Today we are going to develop a deeper understanding of Hidden Markov Models by trying to understand two basic algorithms: the forward algorithm and the Viterbi algorithm – in the context of their most canonical applications robot localization and speech/handwriting recognition.

15.2 Robot Localization and the Forward Algorithm

Here is an example of robot localization. A robot walks around a world for which we have a map and the sensors on the robot tell us in which directions there is a wall, with never more than 1 mistake (and all possibilities being equally likely). Initially, our belief of where the robot is, is flat. As we receive more and more sensor readings fewer and fewer squares are consistent with the possible location, and after a while we are certain of where the robot is.



Lets try to understand mathematically what is going on. Recall, that to specify a HMM, we specify a few different things:

- π_0
- The transition matrix $P_{ij} = P(X_{n+1} = j | X_n = i)$
- The observation/emission distributions: $P(E_n | X_n = i)$ for each i

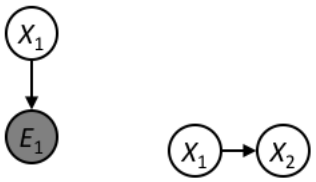
In the robot localization problem, X_t is the location of the robot that we would like to infer, the Es are the sensor readings, and we are given $P(E_t|X_t)$ in the specification above. The initial distribution π_0 is uniform over the space. If we made no observations, then we just

have a Markov chain and we know that eventually the distribution $P(X_n)$ would converge to the limiting distribution of the Markov chain (assuming it existed). However, the observations affect our belief of the state, and this is one of the central ideas in an HMM.

In robot localization we would like to compute (in an online fashion),

$$P(X_t|e_1, \ldots, e_t),$$

which is our estimate of the robot's location at time t after observing the sensor readings $e_1, \ldots, e_t$. Lets first try to understand the basic operations (emissions and transitions), and then we will build this into an algorithm:



To understand the effect of observing e_1 we simply use Bayes' rule:

$$P(X_1|e_1) = \frac{P(e_1|X_1) \times P(X_1)}{P(e_1)} \propto P(e_1|X_1) \times P(X_1).$$

If we want the distribution over X_1 we can simply re-normalize these values to form a distribution. On the other hand we have seen many times how to compute the effect of transitions, we see that,

$$P(X_2) = \sum_{x_1} P(x_1)P(X_2|x_1).$$

Let us now try to understand the steps of the forward algorithm. Suppose we have computed our current "beliefs": $B(X_t) = P(X_t|e_1, \ldots, e_t)$ and would like to compute $B(X_{t+1})$. There are two steps:

1. We first calculate $\widetilde{B}(X_{t+1}) = P(X_{t+1}|e_1, \ldots, e_t)$. To do this we note that,

$$\widetilde{B}(X_{t+1}) = \sum_{x_t} P(X_{t+1}, x_t|e_1, \ldots, e_t) = \sum_{x_t} P(X_{t+1}|x_t) \times P(x_t|e_1, \ldots, e_t),$$

   where the second step uses the fact that conditioned on the hidden state x_t the past observations do not affect the state X_{t+1} (one of the Markov properties of the HMM). Now, we see that,

$$\widetilde{B}(X_{t+1}) = \sum_{x_t} P(X_{t+1}|x_t) \times B(x_t).$$

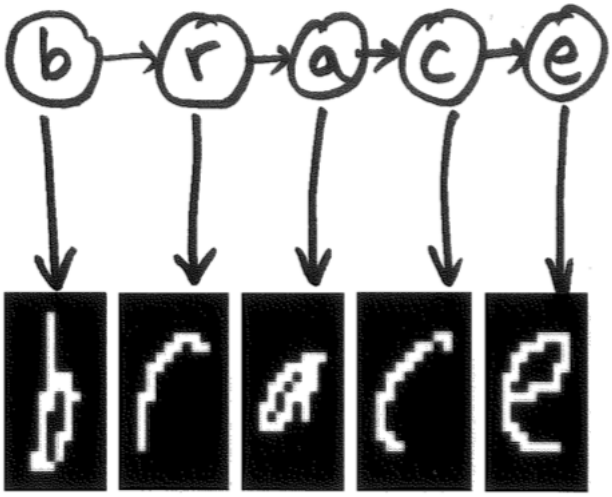
2. Now that, we have $\widetilde{B}(X_{t+1})$ we still need to incorporate the observation e_{t+1}, i.e. using Bayes' rule as above we have that,

$$\begin{aligned} B(X_{t+1}) &= P(X_{t+1}|e_1, \ldots, e_{t+1}) \propto P(e_{t+1}|X_{t+1}) \times P(X_{t+1}|e_1, \ldots, e_t) \\ &= P(e_{t+1}|X_{t+1}) \times \widetilde{B}(X_{t+1}). \end{aligned}$$

These two steps together, incorporating the effects of transitions and observations, are known as the forward algorithm. The forward algorithm forms the basis of robot localization.

15.3 Handwriting/Speech Recognition and the Viterbi Algorithm

The other canonical application of the HMM is in Handwriting/Speech Recognition which we illustrated earlier.



In handwriting recognition, the states X_t correspond to the alphabet, the observations correspond to images (i.e. they are in this case distributions over 64×64 binary images).

Mathematically, the task of decoding is to compute the most likely sequence of states to have generated the observations, i.e. we want to find the sequence of states $x_1, \ldots, x_t$ that maximizes the probability:

$$P(x_1, \ldots, x_t | e_1, \ldots, e_t).$$

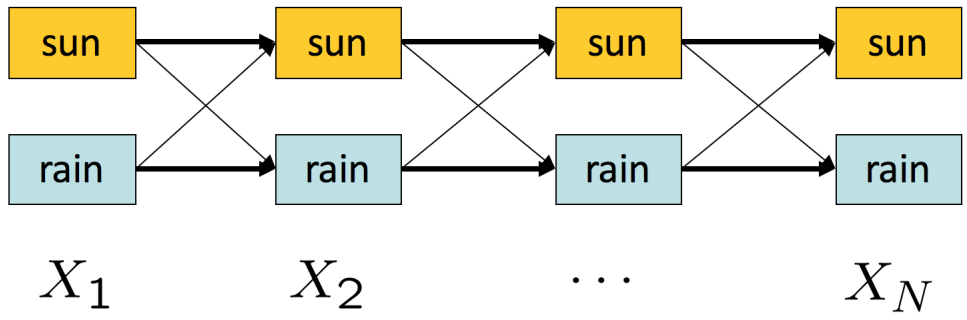
Let us first consider a simple example of an HMM. This is the "sad graduate student" HMM from the Russell-Norvig book. We have a two-state weather Markov chain, where each day we have Sun/Rain with some transition matrix between these states. Each day, there is a hard-working graduate student who does not go out (so does not know if it is Sun/Rain) but instead observes people heading out either carrying an Umbrella or Not.

To specify this HMM fully we need a 2×2 transition matrix as usual and we need to specify:

$$P(\text{Umbrella}|\text{Rain}) = 0.9, \qquad P(\text{Umbrella}|\text{Sun}) = 0.1.$$

We also need some initial distribution over Sun/Rain say [0.5 0.5].

Now, there is a slightly different way of thinking about a Markov chain by un-wrapping it over time. This is sometimes called the *trellis* representation.



Each edge here represents a transition. So to compute the probability of a sequence of states $x_1, \ldots, x_n$ we simply multiply the transition probabilities along the path.

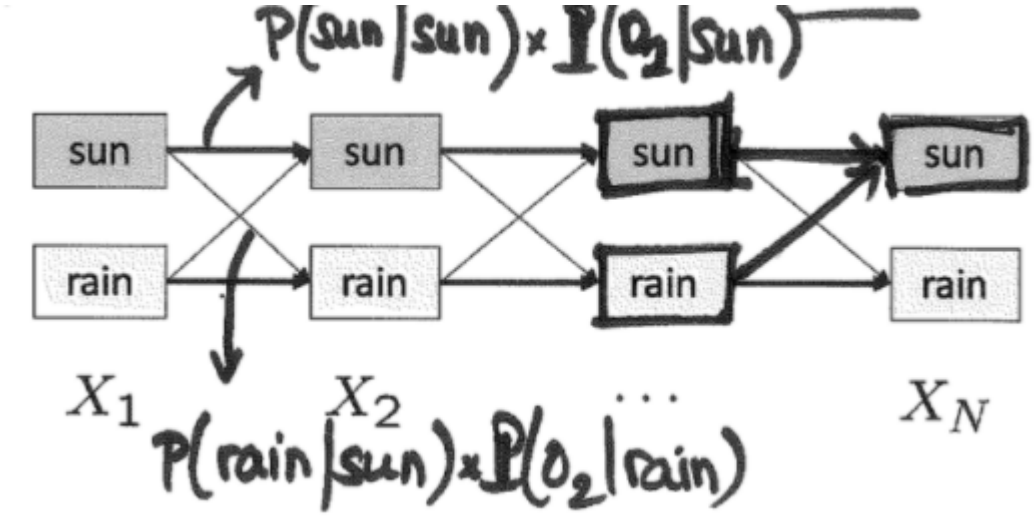
The trellis representation can also be used to derive the Viterbi algorithm. Let us stay with the Sun/Rain HMM. Recall the goal of decoding: we want to find $x_1, \ldots, x_t$ that maximizes:

$$P(x_1, \ldots, x_t | e_1, \ldots, e_t).$$

This is the same as finding $x_1, \ldots, x_t$ that maximizes the joint probability:

$$P(x_1, \ldots, x_t, e_1, \ldots, e_t) = [P(x_1) \times P(e_1|x_1)] \times [P(x_2|x_1) \times P(e_2|x_2)] \times \ldots \times [P(x_t|x_{t-1}) \times P(e_t|x_t)].$$

Now, we can label the trellis with each of these terms. This is illustrated in the figure below (change the *o*s to *e*s):



The goal of decoding is simple – we have many possible paths through this trellis and the probability of each path is obtained by multiplying the numbers along the edges. We want to find the most probable path. Since there are exponentially many paths we still need to do this carefully.

Just as we derived a recursion on beliefs in the forward algorithm we will derive a recursion here as well. Let us define:

$$M(X_t) = \max_{x_1, \ldots, x_{t-1}} P(x_1, \ldots, x_{t-1}, X_t, e_1, \ldots, e_t).$$

In words, $M(X_t)$ is the most likely path that ends in the state X_t at time t.

Now,

$$\begin{aligned}
M(x_t) &= \max_{x_1,\ldots,x_{t-1}} P(x_1,\ldots,x_{t-1},x_t,e_1,\ldots,e_t) \\
&= \max_{x_1,\ldots,x_{t-1}} P(x_1,\ldots,x_{t-1},e_1,\ldots,e_{t-1}) \times P(e_t|x_t) \times P(x_t|x_{t-1}) \\
&= P(e_t|x_t) \times \max_{x_1,\ldots,x_{t-1}} P(x_1,\ldots,x_{t-1},e_1,\ldots,e_{t-1}) \times P(x_t|x_{t-1}) \\
&= P(e_t|x_t) \times \max_{x_{t-1}} P(x_t|x_{t-1}) \max_{x_1,\ldots,x_{t-2}} P(x_1,\ldots,x_{t-1},e_1,\ldots,e_{t-1}) \\
&= P(e_t|x_t) \times \max_{x_{t-1}} P(x_t|x_{t-1}) M(x_{t-1}).
\end{aligned}$$

So we can compute the $M(x_t)$ for each state at each time t. Notice the similarity to the forward algorithm. It is the same recursion, except we replace the sum over x_{t-1} by the maximum over x_{t-1}.

Now, once we have the $M(x_t)$ it is very easy to find the sequence of states that has highest probability. This is known as *backtracking*. For this you need to imagine writing out the $M(X_t)$s in a table.

The last state is easy, we simply say:

$$x_N^* = \arg\max M(x_N).$$

Now, we want to know x_{N-1}^*. We know that in the next step we need to end in x_N^* so we would compute,

$$x_{N-1}^* = \arg\max_{x_{N-1}} \left[M(x_{N-1}) \times P(x^*N|x_{N-1}) \times P(e_N|x_N^*) \right].$$

We can keep doing this to find x_{N-2}^* and so on.

These two steps define the Viterbi algorithm. The first step is to compute $m(x_t)$ for each x_t and each time t. Then we backtrack and find the most likely sequence. At a high-level, all this is, is a dynamic program to find the most likely path through the trellis we drew above.