Basic Probability and Markov Chains

Prepared by Yoni Nazarathy, (Last Updated: August 24, 2014)

So you didn't study probability nor Markov chains? Or maybe you had a course, but forgot some details. These notes (and the exercises within) summarise the basics. The focus in the probability chapter is on discrete random variables. The focus in the Markov chains chapter is on discrete time and finite state space. If you understand the content of these notes (and are able to do the exercises), then you probably have the needed probability and Markov Chains background for studying optimal control through the Markov decision processes (MDP) viewpoint. The notes contain 75 exercises[1].

These notes don't contain many examples, so in case you feel that you can use some extra reading then you can find dozens of books that cover the Markov Chains material and hundreds of books for the probability material. Out of these, one suggestion (nicely matching the current scope) is the material in Chapter 1 and Appendix A of "*Essentials of Stochastic Processes, Second Edition*" by **Rick Durrett** (2012). This is available (for example) on-line through the UQ Library. If instead you are using the first edition of that book (1999), then the relevant chapters in that edition are Chapter 1 and Chapter 2. Both options are fine.

Thanks to error/typo catchers: **Julia Kuhn, Patrick Laub, Brendan Patch, Daniel Sutherland.**

[1]Exercises decorated with **NFQ** (Not for Quiz) do not need to be mastered for the first Quiz of the course.

Chapter 1

Probability

1.1 The Probability Triple

The basic thing to start with is $\mathbb{P}(A)$. What is this? Read this as the *probability* of the event A. Probability is a number in the interval $[0, 1]$ indicating the chance of the event A occurring. If $\mathbb{P}(A) = 0$ then A will not occur. If $\mathbb{P}(A) = 1$, occurrence is certain. If $\mathbb{P}(A) = 0.78$ then we can read this as a chance of 78% for the event. It can also be read as that if we repeat the experiment that we are talking about many times, the proportion of times of which we will observe the event A occurring is 78%. The higher the probability the more likely the event will occur.

But $\mathbb{P}(A)$ doesn't live by itself. Sometimes people ask me: "You are a researcher in the field of probability, so what is the probability of finding another life form on a different planet?". My response often follows the lines: "Sorry, guys, I need a probability model. For example, you can ask me what is the chance of getting a double when tossing a pair of dice. Then my probability model will tell you this is 1/6. But for finding life forms on a different planet, I don't have a model that I can use. Sorry... But we do have some friendly astrophysicists here at UQ so go ask them!".

So what is a probability model? Well the basic way to handle this is through a *probability triple*, $\Big(\Omega, \mathcal{F}, \mathbb{P}\Big)$. The basic idea is that of an *experiment*. Think of every dynamic situation as an experiment. By this I mean every situation in which there can be one of several possible outcomes. The set of possible outcomes to this experiment is Ω. For example in the case of tossing a pair of dice Ω can be represented by,

$$\Omega = \{(i, j) \ : \ i, j = 1, 2, \ldots, 6\}.$$

I.e. when you roll a pair of dice you can get $(3, 4)$ indicating that the first die was 3 and the second was 4 and you can get any other combination. The set Ω is called the *sample space*. Caution: don't confuse this with "sample" as used by statisticians; In general, you shouldn't confuse the (applied) mathematical field of probability with statistics! Do you know the difference? If not, give it some thought.

Back to the probability triple: How about events? Well an *event* is a subset of Ω and we denote the set of these by $\mathcal{F}$. In complicated experiments not all subsets of Ω are in $\mathcal{F}$, but in

elementary examples such as the rolling of a pair of dice we can take $\mathcal{F}$ to be composed of all possible subsets. Specifically this is the case when Ω is finite. In our specific case there are 2^{36} possible outcomes! Also for our specific example, the event $A \subset \Omega$ which indicates "getting a double" is:

$$A = \{(1,1), (2,2), (3,3), (4,4), (5,5), (6,6)\}.$$

One of the events in $\mathcal{F}$ is $\emptyset$. This is called the null-event. Another event is Ω itself. So basically, events are sets (subsets of Ω and elements of $\mathcal{F}$). The appendix to these notes can help you, if you are not an ace on basic set notation and operations and similarly if you have some gaps of knowledge with respect to basic counting (combinatorics).

Now $\mathbb{P}(\cdot)$ is the *probability measure* (sometimes loosely called the *probability function*). It is a function taking elements of $\mathcal{F}$ (events) and mapping them to $[0,1]$. The basic (and most sensible) model for rolling a pair of dice is to believe that each outcome (i,j) is equally likely. In this case (this is often called a *symmetric probability space*) the probability measure is obvious:

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

So for the event we discussed before, $\mathbb{P}(A) = 6/36 = 1/6$. But in other examples we may have a different type of $\mathbb{P}(\cdot)$ that does not give the same chance for all outcomes.

What properties do we expect Ω, $\mathcal{F}$ and $\mathbb{P}$ to obey? Well, $\mathcal{F}$ needs to be a *sigma-Algebra* (also called *sigma-field*). This is a regularity property on the set (family) of events that ensures that the mathematics end up being well defined. Basically we need:

1. $\emptyset \in \mathcal{F}$.
2. If $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$. The set A^c is the complement with respect to Ω. I.e. $A^c = \Omega \setminus A$.
3. If $A_1, A_2, \ldots \subset \mathcal{F}$ then, $\cup_i A_i \in \mathcal{F}$. The number of sets in the union can be finite or countably infinite.

Some properties follow quite easily:

Exercise 1. *Show that:*

1. $\Omega \in \mathcal{F}$.
2. *If* $A_1, A_2, \ldots \subset \mathcal{F}$ *then,* $\cap_i A_i \in \mathcal{F}$.

A note about exercises: These notes are not complete without the exercises. I.e. the exercises are often used to establish statements that are a key part of the main body of understanding. Also, "show" means "prove", just in case you were wondering.

Having defined the (boring) machinery of $\mathcal{F}$ let's move to the key ingredient of any probability model: $\mathbb{P}(\cdot)$. The probability measure must satisfy:

1. For any $A \in \mathcal{F}$, $\mathbb{P}(A) \geq 0$.
2. $\mathbb{P}(\Omega) = 1$.

3. For any countable sequence of disjoint events, $A_1, A_2, \ldots$:

$$\mathbb{P}(\cup_i A_i) = \sum_i \mathbb{P}(A_i).$$

Key in (3) is the fact that the events are disjoint. I.e. for any A_i and A_j with $i \neq j$ we have $A_i \cap A_j = \emptyset$. The above *probability axioms* imply the following:

Exercise 2. *Show that:*

1. $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.
2. $\mathbb{P}(\emptyset) = 0$.
3. $\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2) - \mathbb{P}(A_1 \cap A_2)$ *(this is called the inclusion-exclusion principle).*

1.2 Independence

Two events A and B are said to independent if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$. A typical example is an experiment where you do two things and they don't affect each other. For the rolling of the dice experiment, this is typically the case: One die does not affect the other. And indeed consider for $i \in \{1, \ldots, 6\}$, the events:

$$A_i := \{(i,1), (i,2), (i,3), (i,4), (i,5), (i,6)\},$$
$$B_i := \{(1,i), (2,i), (3,i), (4,i), (5,i), (6,i)\}.$$

The event A_i implies "The first die yielded i". The event B_i implies "The second die yielded i". What is the event $A_i \cap B_j$? It is read as "The first yield i and the second yielded j." Indeed,

$$A_i \cap B_j = \{(i,j)\},$$

and thus,

$$\mathbb{P}(A_i \cap B_j) = \frac{|A_i \cap B_j|}{|\Omega|} = \frac{1}{36} = \frac{1}{6} \cdot \frac{1}{6} = \frac{|A_i|}{|\Omega|}\frac{|B_j|}{|\Omega|} = \mathbb{P}(A_i)\,\mathbb{P}(B_j).$$

So the events are independent.

This example is almost too trivial to be interesting. But the concept of independence goes a long way in probability. This will become more evident when random variables and conditional probability come into play.

Students starting with probability often get confused between "two events being disjoint" and "two events being independent". After all, both terms specify that the events are non-related in some way. But in fact, these concepts are very different.

Exercise 3. *Consider the experiment of tossing a fair coin (yielding 'H' or 'T') and spinning a wheel divided into three parts (yielding '1', '2' or '3'). Assume the underlying probability space is symmetric. Write Ω, $\mathcal{F} = 2^\Omega$ and specify $\mathbb{P}(A)$ for all $A \in \mathcal{F}$ (you'll have 64 events!). Fish out which events are disjoint and which events are independent. See that if two (non-null) events are disjoint they are not independent. And conversely if two (non-null) events are independent, they are not disjoint.*

Independence goes further than just two events. the events $A_1, \ldots, A_n$ are said to be *pair-wise independent* if for each $i \neq j$, A_i and A_j are independent. This set of events is said to be *independent* (without the "pair-wise prefix") if for any set of indexes, $1 \leq i_1 < i_2 \ldots < i_k \leq n$:

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

1.3 Conditional Probability

Given two events, $A, B \subset \Omega$, with $\mathbb{P}(B) > 0$, the *conditional probability* of A given B, denoted $\mathbb{P}(A \,|\, B)$ is defined as:

$$\mathbb{P}(A \,|\, B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}. \tag{1.1}$$

Exercise 4. *Assume $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$. Show that A and B are independent if and only if $\mathbb{P}(A \,|\, B) = \mathbb{P}(A)$.*

Mmmm... So if A and B are independent then the chance of A happening is not influenced by B. But if there is some dependence, then $\mathbb{P}(A \,|\, B) \neq \mathbb{P}(A)$.

Exercise 5. *Suppose you roll a die. I tell you that the result is an even number. So now what is the chance that the result is 6?*

There are mathematical subtleties in defining conditional probability, but we won't touch these. From our perspective, we can consider the conditional probability $\mathbb{P}(\cdot \,|\, B)$, (1.1), as a new probability measure in a new probability triple, $\big(B, \tilde{\mathcal{F}}, \mathbb{P}(\cdot \,|\, B)\big)$. It is as though the sample space was reduced from Ω to B and all probabilities were simply normalised. This means that all the properties of $\mathbb{P}(\cdot)$ from the previous section carry over. For e.g.,

$$\mathbb{P}(A \,|\, B) = 1 - \mathbb{P}(B \setminus A \,|\, B).$$

Below are three useful basic results that follow immediately from the definition in (1.1). Let $A, B_1, B_2, B_3, \ldots \subset \Omega$ with $\{B_i\}$ mutually disjoint sets such that $\cup_i B_i = \Omega$:

1. *The multiplication rule:* Assume $\mathbb{P}(B) > 0$, then $\mathbb{P}(A \cap B) = \mathbb{P}(B)\,\mathbb{P}(A \,|\, B)$.
2. *The law of total probability:* $\mathbb{P}(A) = \sum_i \mathbb{P}(A \,|\, B_i)\,\mathbb{P}(B_i) = \sum_i \mathbb{P}(A \cap B_i)$.
3. *Bayes' rule:* $\mathbb{P}(B_i \,|\, A) = \frac{\mathbb{P}(A \,|\, B_i)\,\mathbb{P}(B_i)}{\sum_j \mathbb{P}(A \,|\, B_j)\,\mathbb{P}(B_j)}$.

Note that in certain cases, the law of total probability and the celebrated Bayes' rule hold also when there is an non-countable family of events $\{B_t\}$. In that case, replace the summations over i by integrals over t.

Exercise 6. *Prove (1)–(3) above.*

Have you heard of Bayesian statistics? The underlying mechanism is Bayes' rule.

An example that surprises many people is the following: Suppose you are in a television gameshow where you need to choose one of three boxes, one of which has a prize, and the

others are empty. The game-show host knows where the prize is. You point at one of the boxes and say with a hesitant voice: "this is my box". At that point, the flashy gameshow host follows the producer's protocol and reveals another box, showing you that the prize is not in that one. Now you know that either your first choice was the correct box, or perhaps the prize is in the third box. The gameshow continues to follow protocol and says: "So, do you want to stay with your box, or change (to the third box)?". What do you do?

The immediate intuitive answer would be to say: "It doesn't matter, there is a 50% chance for having the prize in either the current box or the other option." But let's look more closely.

Denote the boxes by $1, 2, 3$ and assume without loss of generality that you choose box 1 at first. Denote the event that the prize is in box i by A_i. Clearly,

$$\mathbb{P}(A_i) = \frac{1}{3}, \quad i = 1, 2, 3.$$

Now the host will never reveal a box with a prize. If you initially guessed the correct box, the host will have an option between two boxes to reveal. But if you initially guessed the wrong box, the host only has one option of what to reveal. Denote by B the event that the host reveals box 2 after your choice. I.e. B^c is the event that the host reveals box 3. So:

$$\mathbb{P}(B \,|\, A_1) = \frac{1}{2}, \qquad \mathbb{P}(B^c \,|\, A_1) = \frac{1}{2}.$$

and,

$$\mathbb{P}(B \,|\, A_2) = 0, \qquad \mathbb{P}(B^c \,|\, A_2) = 1,$$

and similarly,

$$\mathbb{P}(B \,|\, A_3) = 1, \qquad \mathbb{P}(B^c \,|\, A_3) = 0.$$

Now using the law of total probability,

$$\mathbb{P}(B) = \mathbb{P}(B \,|\, A_1)\,\mathbb{P}(A_1) + \mathbb{P}(B \,|\, A_2)\,\mathbb{P}(A_2) + \mathbb{P}(B \,|\, A_3)\,\mathbb{P}(A_3) = \frac{1}{2} \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} = \frac{1}{2}.$$

So (not surprisingly) there is a 50% chance that the host reveals box 2.

Now let's put you back in that situation. You are on TV! You just made a choice (box 1), and the gameshow guy (or flashy gal if you wish) just revealed box 2. So you observed the event B. Now you want to compare,

$$\mathbb{P}(A_1 \,|\, B), \qquad \text{v.s.} \qquad \mathbb{P}(A_3 \,|\, B),$$

and choose the box which maximises this probability. Using Bayes' Rule

$$\mathbb{P}(A_1 \,|\, B) = \frac{\mathbb{P}(B \,|\, A_1)\mathbb{P}(A_1)}{\mathbb{P}(B)} = \frac{\frac{1}{2} \cdot \frac{1}{3}}{\frac{1}{2}} = \frac{1}{3},$$

and the complement,

$$\mathbb{P}(A_3 \,|\, B) = \frac{\mathbb{P}(B \,|\, A_3)\mathbb{P}(A_3)}{\mathbb{P}(B)} = \frac{1 \cdot \frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}.$$

So you are better of changing boxes!!! Go for it.

I must admit that this is one of the most confusing elementary examples of conditional probability (and Bayes' rule) that are out there. But it is also one of the more shocking examples; hence it is interesting. I was recently reminded of it by a curious pool-safety-person[1], and I couldn't explain it to him without resorting to formalism. Maybe you can...

Exercise 7. NFQ *Think about this example. Try to describe (in lay-person terms) why you are better off changing boxes.*

1.4 Random Variables and their Probability Distributions

So now you know what a probability triple is and you even know about independence and basic conditional probability. What next? Well typically we work with random quantities. And by "quantity" I mean something that is easier to handle and manipulate in comparison to arbitrary sets (events). By this I mean real numbers, integers, complex numbers, vectors, matrices etc... But let's just think of random quantities that are either real valued (continuous) or integer valued (discrete). Our focus is in fact on discrete (basically integer) quantities..

A *random variable*, X (also referred to sometimes as a *measurable function*), is a mapping from Ω to $\mathbb{R}$ or $\mathbb{N}$ or some other sensible set (vectors, complex numbers etc...). Think for now about integer random variables so, $X : \Omega \to \mathbb{Z}$. Now the idea is that since the $\omega \in \Omega$ is a random outcome of an experiment, then so is $X(\omega)$. Formally, the way to handle this is to define for sensible subsets of $B \subset \mathbb{Z}$, an inverse image set,

$$A = \{\omega \in \Omega \,:\, X(\omega) \in B\}.$$

Think of A as an event; B should not be thought of as an event. It is rather a set of values that the random variable may take.

Now if everything is well defined meaning that $\mathcal{F}$ is rich-enough and that $X(\cdot)$ and B are not too crazy, then $A \in \mathcal{F}$ and hence it is a proper event which we can stick in $\mathbb{P}(\cdot)$. Often instead of the event A we often just write "$X \in B$" instead. So we can calculate probabilities of the form, $\mathbb{P}(X \in B)$. Of course if the set B contains just one point, say b, then we can try and evaluate $\mathbb{P}(X = b)$ or if B is say an interval $[a, b]$ (with possibility one or two of the endpoints being $-\infty$ or ∞, then we can try and evaluate $\mathbb{P}(a \leq X \leq b)$, etc.. etc... The point is that random variables quantify the outcome of the experiment. And for some possible set of outcomes, B, we are asking for the probability of $X \in B$.

Now consider sets B of the form, $(-\infty, b]$. For such sets we have,

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

Such subsets, B are useful because if we know the value of $\mathbb{P}(X \leq b)$ for all b then we can use this to calculate $\mathbb{P}(X \in B)$ for any sensible B. This motivates us to define the *distribution function*:

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

[1]Dan Adelman (Finishing Touch Pool Safety Inspections and Compliance Repairs) – highly recommended for pool safety certificates as well as for a long chat about probability once the job is done.

The subscript X is just part of the notation of the function - it reminds us that this is the distribution of the random variable X. This function is also (less ambiguously) called: the *cumulative distribution function* (CDF). Some prefer to work with the *complementary cumulative distribution function* (CCDF):

$$\overline{F}_X(b) := 1 - F_X(b) = \mathbb{P}(X > b).$$

Some call the above the *survival function* - but these guys are typically wearing suits and don't smile too much because they work in insurance companies or are reliability engineers. The CDF or CCDF are alternative descriptions of the *distribution* of X. There are other descriptions which are sometimes useful also (probability mass function, probability density function, moment generating function, probability generating function, characteristic function, Laplace transform, hazard rate, renewal measure,...). What I'm trying to say is that there are many ways to describe the *distribution of a random variable*, each useful in its own way. But let's get back to CDF:

Exercise 8. *Show that,*

1. $\lim_{x\to-\infty} F_X(x) = 0$.
2. $\lim_{x\to\infty} F_X(x) = 1$.
3. $F_X(\cdot)$ *is non-decreasing.*

The above three properties are often taken to be defining properties of CDFs. For any function satisfying the above, we can actually find a probability triple supporting a random variable X with the desired CDF.

In these notes we focus mostly on random variables whose values fall within a discrete set such as $\{0, \ldots, n\}$ for some finite n or $\mathbb{N}$ or $\mathbb{Z}$ etc. These are sometimes called *discrete random variables*. We call the set of values which the random variable may take, the *support*. If (for e.g.) the support does not have negative values then we say the random variable is *non-negative*.

Exercise 9. *Consider the first example of these notes (tossing of two dice). Let the random variable be the sum of the dice. Illustrate the graph $F_X(x)$. At points of discontinuity, make sure to note open and closed indications on the graph.*

For discrete random variables an alternative (and sometimes easier to handle) representation of the distribution is the *probability mass function* (PMF):

$$p_X(k) := \mathbb{P}(X = k).$$

Assuming that the support is some subset of $\mathbb{Z}$ then,

$$F_X(k) := \sum_{i=-\infty}^{k} p_X(i) \quad \text{and} \quad p_X(k) = F_X(k) - F_X(k - \epsilon),$$

where ϵ is any value in the range $(0, 1]$. For k that are not in the support we simply have $p_X(k) = 0$. Keep this in mind, because when we write things such as,

$$\sum_{k=-\infty}^{\infty} p_X(k),$$

this is equivalent to,

$$\sum_{k \in \text{ support of } X} p_X(k).$$

Exercise 10. *Draw the PMF associated for the previous exercise. Place your illustration under the CDF. Exhibit the relationship between the CDF and the PMF.*

Some people call refer to PMF as "density". I respect these people, some of them are even my friends, but I'm not one of them. I keep the word density for functions $f_X(x)$ that describe the CDF of continuous random variables through:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt.$$

But more on this later (when we briefly touch continuous distributions). Also I should state that in the continuation of these notes, I won't use the notation $p_X(\cdot)$ much, even though PMFs will appear everywhere.

1.5 Expectation, Mean, Variance, Moments

The *mean* of a (discrete) random variable, denoted $\mathbb{E}[X]$ is:

$$\mathbb{E}[X] = \sum_{k=-\infty}^{\infty} k\, p_X(k).$$

An alternative name for the mean is the *expectation* or *expected value*. The expected value describes the "center of mass" of the probability distribution. Another meaning follows from the law of large numbers described in the sequel: If we observe many random variables having this distribution and calculate their average, it will be near the mean. Note that in the summation above, it is enough to sum over the support of the random variable since for other values of k, $p_X(k) = 0$.
Observe that the mean of an integer valued random variable does not have to be an integer.

Exercise 11. *What is the mean value for the sum of two dice? Use the probability model and random variable that appeared in previous exercises.*

Exercise 12. *Show that for a non-negative random variable,*

$$\mathbb{E}[X] = \sum_{k=0}^{\infty} \overline{F}_X(k).$$

Take now $h : \mathbb{R} \to \mathbb{R}$ then $h(X(\omega))$ is some new random variable. We can calculate the mean of this new random variable simply as follows:

$$\mathbb{E}[h(X)] = \sum_{k=-\infty}^{\infty} h(k)\, p_X(k). \tag{1.2}$$

I.e. the expectation functional, $\mathbb{E}[\cdot]$ takes as input a random variable and returns a number. When $h(x) = x^n$ then $\mathbb{E}[h(X)]$ is called the n'th *moment*. I.e. the first moment is the mean. Another important case $h(x) = (x - \mathbb{E}[X])^2$ then $\mathbb{E}[h(X)]$ is called the *variance* and denoted, $\mathbb{V}\mathrm{ar}\,(X)$. Note that it is non-negative. The square root of the variance is called the *standard deviation*. Both the variance and the standard deviation are measures of the spread of the distribution (each one of these is useful in its own way). You can see that:

$$\mathbb{V}\mathrm{ar}\,(X) = \mathbb{E}\big[(X - \mathbb{E}[X])^2\big] = \mathbb{E}\big[X^2 - 2\mathbb{E}[X]\,X + (\mathbb{E}[X])^2\big]. \tag{1.3}$$

Note that inside the expectation operator we are doing algebra involving both the random variable X and the constant values, 2 and $\mathbb{E}[X]$.

Exercise 13. *Show that,*

1. *If c is a constant (non-random quantity), then $\mathbb{E}[c\,X] = c\,\mathbb{E}[X]$.*
2. *For any two random variables, X and Y,*

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

(Illustrate this through the meaning of a random variable – a function of ω).

Now with these basic properties of the expectation, you are ready to proceed with (1.3) to show that,

$$\mathbb{V}\mathrm{ar}\,(X) = \mathbb{E}[X^2] - \big(\mathbb{E}[X]\big)^2.$$

This implies that for "zero-mean" random variables, the variance is simply the second moment.

Exercise 14. *Let, c_1, c_2 be some constants. What is $\mathbb{V}\mathrm{ar}\,(c_1 X + c_2)$ in terms of $\mathbb{V}\mathrm{ar}\,(X)$?*

Exercise 15. *Show that if $\mathbb{V}\mathrm{ar}\,(X) = 0$ then the support of X contains a single value (i.e. there is some k_0 such that $p_X(k_0) = 1$).*

Another very important $h(\cdot)$ is obtained by setting some $B \subset \mathbb{R}$ and then $h(x) = \mathbf{1}_B(x) := \mathbf{1}\{x \in B\}$ (and indicator function returning 1 if $x \in B$ and 0 otherwise). In this case $\mathbb{E}[h(X)] = \mathbb{P}(X \in B)$. Nice, no?

1.6 Bernoulli Trials

We now consider probability spaces where Ω is the set of binary sequences,

$$\Omega = \{(b_1, b_2, b_3, \ldots),\ b_i \in \{0, 1\}\}$$

and where $\mathbb{P}(\cdot)$ is such that the events $\{b_i = 1\}$ are independent. We further assume that $\mathbb{P}(\{b_i = 1\}) = p$ for all i. I.e. this probability space describes experiments involving a sequence of independent "coin flips", each with having the same probability of success: p.

There are now many random variables associated with this probability space. We say X follows a *Bernoulli distribution*, with probability p if,

$$\mathbb{P}(X = 0) = (1 - p), \qquad \text{and} \qquad \mathbb{P}(X = 1) = p.$$

We say that X follows a *binomial distribution* with parameters n and p if,

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, \ldots, n. \tag{1.4}$$

Here n is any integer ≥ 1 and $p \in [0, 1]$.

Exercise 16. *Show that with respect to the Bernoulli Trials probability space,*

$$X(\omega) = \sum_{i=1}^{n} \mathbf{1}\{b_i^\omega = 1\},$$

where b_i^ω is the i'th element of ω. That is, derive the right hand side (1.4)*.*

Exercise 17. *Verify for the binomial distribution of* (1.4)*, that*

$$\sum_{i=0}^{n} \mathbb{P}(X = i) = 1.$$

Exercise 18. *1. Show that the mean of a binomial distribution is $n\,p$.*

2. Let X be binomially distributed with n and p. What is the distribution of $Z = n - X$?

Exercise 19. *Assume you are guessing answers on a multiple choose test that has 20 questions, and each can be answered (a), (b), (c), or (d). What is the chance of getting* 10 *or more answers correct?*

Consider now, $X(\omega) = \inf\{k \in \{1, 2, 3, \ldots\} \mid b_k^\omega = 1\}$. I.e. This is the index of the trial with the first success. Such a random variable is said to follow a *geometric distribution* with success probability p.

Exercise 20. *Show that,*

$$\mathbb{P}(X = k) = (1-p)^{k-1}\, p, \qquad k = 1, 2, \ldots.$$

Verify that (as needed),

$$\sum_{k=1}^{\infty} \mathbb{P}(X = k) = 1.$$

Exercise 21. *The chance of getting a flat-tire on a bicycle ride is* 0.01*. What is the chance of having* 20 *consecutive bicycle rides without a flat tire?*

A related random variable (also referred to as "geometric"), counts the "number of failures until success" as opposed to the "number of trials until success".

Exercise 22. *What is the support and distribution of this version of the geometric?*

A generalisation of the geometric distribution is the *negative binomial distribution.* Here X counts the number of trials till m successes:

$$X(\omega) = \inf\{k \in \{1, 2, 3, \ldots\} \mid \sum_{i=1}^{k} b_i^\omega = m\}.$$

The support of this distribution is $\{m, m+1, m+2, \ldots\}$.

Exercise 23. *Develop the pmf of the negative binomial distribution with parameters $p \in [0,1]$ and $m \geq 1$ from first principles. Do the same for a modification (as was for the geometric) which counts the number of failures till m successes. The support here is $\{0,1,2,\ldots\}$.*

1.7 Other Common Discrete Distributions

You can think of the binomial distribution as follows: You are fishing in a lake where there are M brown fish and N gold fish. You are fishing out n fish, one by one, and whenever you catch a fish you return it to the lake. So assuming your chance of catching a fish of a given type is exactly its proportion, and further assuming that fishing attempts don't interact, the number of gold fish that you get is binomially distributed with n and $p = N/(N+M)$. The thing here is that by catching a fish, you didn't alter the possible future catches.

But what if you (weren't a vegetarian like me), and as you catch a fish, you bop it in the head, fry eat and eat it. Then with every fish you are catching, you are altering the population of fish, and then the binomial description no longer holds. In this case X, the number of gold fish that you catch follows a *hyper-geometric distribution.*

$$\mathbb{P}(X = k) = \frac{\binom{N}{k}\binom{M}{n-k}}{\binom{N+M}{n}}.$$

Exercise 24. *The hyper-geometric distribution is constructed by basic counting arguments on a symmetric probability space. Carry out these arguments. Further, what is the support of this distribution?*

Exercise 25. NFQ *When $N+M \to \infty$ (i.e. big lakes) such that $N/(N+M) \to p$, you would expect that it doesn't matter if you return the fish to the lake or not. This can be formalised by showing the the pmf of the hyper-geometric distribution converges to the binomial distribution. Find this some place in the literature and carry out the computations, describing the steps. Or if you have already had several courses of probability, maybe try to do it without looking elsewhere.*

Another useful discrete distribution is the *Poisson distribution* (incidentally "poisson" means fish in French – but we are now done with fish). The random variable X is distributed Poisson with parameter λ if,

$$\mathbb{P}(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}, \qquad k = 0,1,2,\ldots.$$

Exercise 26. *Show that the mean and variance are both λ.*

The Poisson distribution is useful for describing the number of events in a time-interval. Especially when events occur in a "completely random manner". That is, it may be a good model for the number of shooting stars that you observe while looking at a moon-less desert sky for an hour. To see this, consider the hour and divide it input n intervals, each interval being quite small. Then it is sensible that within each such interval there is a probability of p_n for seeing a shooting star. Here the subscript indicates the dependence on n. The bigger the n the smaller the p. In fact, how about setting $\lambda = n\,p_n$ (this is the mean number of shooting stars

during that hour). Now if we increase $n \to \infty$ then $p_n \to 0$ in such a way that their product remains λ. For any finite n, the number of stars is distributed Binomial(n, p_n). But as $n \to \infty$ this converges to Poisson.

Exercise 27. NFQ *Show that for every k,*

$$\lim_{n\to\infty} \binom{n}{k} (\frac{\lambda}{n})^k \left(1 - \frac{\lambda}{n}\right)^{n-k} = e^{-\lambda} \frac{\lambda^k}{k!}.$$

As a final example of a discrete distribution, consider,

$$\mathbb{P}(X = k) = \frac{1}{k(k+1)}, \quad k = 1, 2, \ldots.$$

Indeed by writing

$$\frac{1}{k(k+1)} = \frac{1}{k} - \frac{1}{k+1},$$

we get a telescopic sum and see that,

$$\sum_{k=1}^{\infty} \mathbb{P}(X = k) = 1,$$

as desired. This distribution is an example of a power-law, since the tails of it decay to 0 like a power law. Such distributions are sometimes called *heavy tailed* and indeed the following distribution does not have a finite mean.

Exercise 28. *Show that the mean is infinite.*

Note that while the mean is infinite it is well defined. I.e. this series diverges to infinity:

$$\sum_{k=1}^{\infty} k\, \mathbb{P}(X = k) = \infty.$$

But in other cases, the mean is not even defined. For e.g. consider this distribution:

$$\mathbb{P}(X = k) = \frac{1}{2|k|(|k|+1)}, \quad k = \ldots, -3, -2, -1, 1, 2, 3 \ldots.$$

1.8 Vector Valued Random Variables

A vector valued random variable doesn't differ much from the scalar (uni-variate) cases described above. We'll present things for a vector of two random variables, X and Y. The generalisation to n random variables is straight forward.

The basic object is the *joint probability mass function*:

$$p_{X,Y}(k, \ell) = \mathbb{P}(X = k,\, Y = \ell).$$

The requirement is that,

$$\sum_{k=-\infty}^{\infty} \sum_{\ell=-\infty}^{\infty} p_{X,Y}(k,\ell) = 1.$$

This is sometimes called the *joint distribution* of X and Y. Knowing this joint distribution, how can we recover the distributions of the individual random variables, X and Y? To get the distribution of X, we sum over all possible Y:

$$p_X(k) = \sum_{\ell=-\infty}^{\infty} p_{X,Y}(k,\ell).$$

Similarly to get the distribution of Y we can sum over all possible X.

Exercise 29. *Derive the above using the law of total probability.*

We know about independence of events, but what is independence of random variables? The random variables X and Y are said to be *independent* if,

$$p_{X,Y}(k,\ell) = p_X(k)\, p_Y(\ell).$$

When the random variables are independent, the knowledge of X yields no information about Y and visa-versa.

Given some function, $h : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ we can compute the expectation of the random variable $h(X,Y)$ as follows:

$$\mathbb{E}[h(X,Y)] = \sum_k \sum_\ell h(k,\ell)\, p_{X,Y}(k,\ell).$$

The *covariance* of X and Y, denoted $\mathbb{C}\mathrm{ov}\,(X,Y)$ is computed in this way using

$$h(x,y) = (x - \mathbb{E}[X])(y - \mathbb{E}[Y]).$$

Exercise 30. *Show that* $\mathbb{C}\mathrm{ov}\,(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$.

Exercise 31. *Show that if X and Y are independent then $\mathbb{E}[X\,Y] = \mathbb{E}[X]\,\mathbb{E}[Y]$ and hence,* $\mathbb{C}\mathrm{ov}\,(X,Y) = 0$.

Exercise 32. *Take a case where the support of X is $\{1,2,3\}$ and the support of Y is $\{1,2\}$.*

1. *Find $p_{X,Y}(x,y)$ such that* $\mathbb{C}\mathrm{ov}\,(X,Y) \neq 0$.
2. *Find $p_{X,Y}(x,y)$ such that X and Y are not independent but* $\mathbb{C}\mathrm{ov}\,(X,Y) = 0$.

1.9 Conditioning and Random Variables

Now that you know about multiple random variables living together in the same probability space, you can start seeing how they interact. Consider first the conditional probability:

$$\mathbb{P}(X = k \,|\, Y = \ell).$$

Since you can read "$X = k$" and "$Y = \ell$" as events then $\mathbb{P}(X = k \,|\, Y = \ell)$ is well defined (well, as long as $Y = \ell$ can occur with a positive probability). Continuing this, define the function, $p_{X|Y=\ell}(\cdot, \cdot)$, as:

$$p_{X|Y=\ell}(k, \ell) := \mathbb{P}(X = k \,|\, Y = \ell) = \frac{\mathbb{P}(X = k,\, Y = \ell)}{\mathbb{P}(Y = \ell)} = \frac{p_{X,Y}(k, \ell)}{p_Y(\ell)}.$$

The function $p_{X|Y=\ell}(\cdot, \ell)$ specifies the *conditional distribution* of X given that $Y = \ell$.

Exercise 33. *Show that $p_{X|Y=\ell}(\cdot, \ell)$ is a valid probability mass function (in the first variable) for any ℓ such that $\mathbb{P}(Y = \ell) > 0$.*

Exercise 34. *Show that if X and Y are independent random variables, then $p_{X|Y=\ell}(\cdot, \ell) = p_X(\cdot)$.*

Exercise 35. *For your example used as solution of Exercise 32 calculate, $p_{X|Y=\ell}(\cdot, \cdot)$ and $p_{Y|X=k}(\cdot, \cdot)$ for all possible values. I.e. specify 6 distributions.*

The geometric distribution is said to be *memoryless* due to this property:

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

Exercise 36. *1. Show that the memoryless holds for geometric random variables.*

2. Comment on why this property makes sense (considering the sequence of Bernoulli trials).

3. Find another discrete distribution which does **not** *satisfy the memoryless property.*

Now that you know about conditional distributions, you can talk about the *conditional expectation*, variance, etc... Simply define:

$$\mathbb{E}[h(X) \,|\, Y = \ell] = \sum_k h(k)\, p_{X|Y=\ell}(k, \ell).$$

Exercise 37. *Calculate the conditional means of the 6 distributions of the previous example. Compare these means to the two (unconditional) means of X and Y.*

Observe that you can think of $\mathbb{E}[h(X) \,|\, Y = \ell]$ as a function of ℓ. So what if you left ℓ unspecified and let it simply be the result of the random variable Y? In this case, you get (also called conditional expectation) the random variable: $\mathbb{E}[h(X) \,|\, Y]$. The conditional expectation is a random variable because it is a function of the random variable on which we are conditioning.

Exercise 38. *Show that,*

$$\mathbb{E}\Big[\mathbb{E}[h(X) \,|\, Y]\Big] = \mathbb{E}[h(X)]. \tag{1.5}$$

Note that the outer expectation is with respect to the random variable Y.

The formula (1.5) is sometimes called the smoothing formula. It is sometimes super-useful because, evaluation of $\mathbb{E}[h(X)]$ in its own may be tough, but if we condition on another random variable Y, things get much easier. This is a classic example: Let $X_1, X_2, \ldots$ be a sequence of

i.i.d. (*independent and identically distributed*) random variables independent of some discrete random variable N. Denote,

$$S := \sum_{i=1}^{N} X_i.$$

The new random variable S is sometimes called a random sum. For example, N may be the number of insurance claims a company has during a month, and each insurance claim is assumed to be distributed as X_1. What is $\mathbb{E}[S]$? Intuition may tell you that, $\mathbb{E}[S] = \mathbb{E}[N]\,\mathbb{E}[X_1]$. This is for example the case if N equals some fixed value with probability 1 (the linearity of expectation). But how can you show (prove) this? Well, condition on N:

$$\begin{aligned}
\mathbb{E}[\sum_{i=1}^{N} X_i] &= \mathbb{E}[\,\mathbb{E}[\sum_{i=1}^{N} X_i \,|N]] \\
&= \mathbb{E}[\sum_{i=1}^{N} \mathbb{E}[X_i \,|\, N]] \\
&= \mathbb{E}[\sum_{i=1}^{N} \mathbb{E}[X_i]] \\
&= \mathbb{E}[N\,\mathbb{E}[X_1]] \\
&= \mathbb{E}[X_1]\,\mathbb{E}[N].
\end{aligned}$$

Exercise 39. *Detail (in words) what is happening in each step of the above derivation.*

1.10 A Bit on Continuous Distributions

The random variables discussed up to now were discrete. Their support is finite or countably infinite. For our purposes, these are indeed the critical cases to master. Nevertheless, we now briefly touch on *continuous random variables*. In the continuous case, the support is some non-countable subset of $\mathbb{R}$: E.g. $[a, b]$ or $[0, \infty)$ or all of $\mathbb{R}$. For such random variables, $\mathbb{P}(X = x) = 0$ for any specific x, but for intervals of strictly positive length, the probability can be non-zero. Such random variables are best described by a *density function*: $f_X(x) : \mathbb{R} \to \mathbb{R}_+$. The best way to think of the density is that it is a function satisfies the following:

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

Exercise 40. *1. What is $\int_{-\infty}^{\infty} f_X(x)\,dx$?*

2. *Given a density, $f_X(\cdot)$, what is the CDF? Is the CDF a continuous function? Or only if the density is continuous?*

3. *Given any integrable, non-negative function $\tilde{f}(x)$, describe how to make a density $f_X(\cdot)$ such that $f_X(x) = K\tilde{f}(x)$ for some constant K.*

For statisticians, the typical way of thinking about a distribution is through the density. If you think about it, indeed a PMF and a density are not so different. You should also know that random variables don't need to be continuous or discrete, you can get mixtures of the two or even more exotic objects. But for an elementary and introductory treatment such as ours, this dichotomy is fine.

The mean, moments and variance of continuous random variables are defined in an analogous way to the discrete case. The basic definitions is:

$$\mathbb{E}[h(X)] = \int_{-\infty}^{\infty} h(x)\, f_X(x)\, dx. \tag{1.6}$$

Once you realise that $\mathbb{P}\big(X \in [x, x+dx)\big) \approx f_X(x)\, dx$, the above should make perfect sense. I.e. compare (1.6) with (1.2). As with discrete random variables, make sure that you know what is the support of the random variable. For x's not in the support, $f_X(x) = 0$. So the region of integration in (1.6) may be limited to the support.

There are many types (parametrised families) of continuous probability distributions and manipulation of these encompasses a good part of a full course of probability. Here we shall outline three key types:

The *uniform distribution* on the range $[a, b]$ has density,

$$f_X(x) = \frac{1}{b-a}, \qquad x \in [a, b].$$

Exercise 41. *Calculate the mean and variance of the uniform distribution. The mean should make "perfect sense" – explain it. The variance: not intuitive.*

Exercise 42. *Write out the CDF of the uniform distribution. Make sure to specify it for the three regions, $x \leq a$, $x \in [a, b]$ and $x > b$.*

But come on! The uniform density is a bit boring. This one is much more exciting:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}.$$

This is the *normal* (also known as Gaussian) density with parameters $\mu \in \mathbb{R}$ and $\sigma > 0$.

Exercise 43. *Show that the mean is μ and that the variance is σ^2.*

Gaussian random variables are everywhere. I said everywhere!!! In the sequel when we discuss the central limit theorem there is some evidence for that.

Exercise 44. NFQ *Do you believe me that for the Gaussian case, $f_X(\cdot)$ is a density? Carry out numerical integration (for some selected μ and σ) to check that,*

$$\int_{-\infty}^{\infty} f_X(x)\, dx = 1.$$

The final example that we briefly describe is the exponential distribution with parameter $\lambda > 0$.

$$f_X(x) = \lambda\, e^{-\lambda x}, \quad x > 0.$$

Exercise 45. *1. Verify that* $\int_0^\infty f_X(x)\,dx = 1$.

2. Calculate the mean.

3. Calculate the variance.

You can get a discrete distribution by transforming a continuous one. Here is one such example:

Exercise 46. *Let* X *be distributed exponential*(λ). *Let* $Y = \lfloor X \rfloor$. *What is the distribution of* Y*?*

Exercise 47. *Show that (as for geometric random variables), exponential random variables also satisfy the memoryless property.*

1.11 Limiting Behaviour of Averages

Much of modern probability deals with limiting results associated with sequences of random variables and stochastic processes. Here we only discuss the two fundamental classic results: The first result states that the sample mean converges to the mean:

Theorem 1 (The Strong Law of Large Numbers (SLLN))**.** *Let* $X_1, X_2, \ldots$ *be and i.i.d. sequence of random variables with finite mean* μ. *Then with probability* 1*:*

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

Exercise 48. *Let* $q = \mathbb{P}(X_i > \alpha)$. *Use the SLLN to show that with probability* 1*:*

$$\lim_{n\to\infty} \frac{\#_n\{X_i > \alpha\}}{n} = q,$$

where $\#_n\{A_i\}$ *is the number of times out of the first* n *during which the event* A_i *occurs.*

The next result is called the *central limit theorem*. It is the reason for the universality of the normal distribution. It shows that normalised sums of random variables converge in distribution to the normal distribution.

Theorem 2 (The Central Limit Theorem (CLT))**.** *Let* $X_1, X_2, \ldots$ *be and i.i.d. sequence of random variables with mean* μ *and finite variance* $\sigma^2 > 0$. *Then,*

$$\lim_{n\to\infty} \mathbb{P}\Big(\frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n\sigma^2}} \le x\Big) = \Phi(x) := \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{\frac{-u^2}{2}}\,du, \quad \forall x \in \mathbb{R}.$$

Exercise 49. *Another version (often more popular with statisticians) of the CLT deals with the asymptotic distribution of the sample mean,* $\frac{1}{n}\sum_{i=1}^n X_i$*:*

$$\lim_{n\to\infty} \mathbb{P}\Big(\frac{\frac{1}{n}\sum_{i=1}^n X_i - a_n}{b_n} \le x\Big) = \Phi(x) \quad \forall x \in \mathbb{R}.$$

Here a_n *is the mean of the sample mean and* b_n *is it's standard deviation. What are* a_n *and* b_n*?*

Exercise 50. NFQ *Let X_1, X_2 and X_3 be i.i.d. uniform$(0,1)$ random variables. Using either a convolution (analytically – if you know how to do that) or via simulation (overviewed in the next section), plot the density of $S_n = \sum_{i=1}^{n} X_i$ for $n = 2$ and 3. What is the relation of this exercise to the CLT?*

1.12 Computer Simulation of Random Variables

When you invoke the **rand**() function in matlab (or similar functions in similar software packages) you get a *pseudo-random* number in the range $[0, 1]$. This number is an element in a deterministic (non-random) sequence initialised by a *seed*. A good pseudorandom sequence has statistical properties similar to an i.i.d. sequence of uniform$(0, 1)$ random variables.

What if you want to use a computer to generate (simulate) random variables from a different distribution? In certain cases, it should be obvious how to do this:

Exercise 51. NFQ *Generate on a computer,* $10,000$ *Bernoulli random variables with success probability $p = 0.25$. Calculate the sample mean and sample variance. How far are these values from the theoretical values?*

So you figured out how to generate Bernoulli random variables. But what about other types of random variables? Below is a general method.

Proposition 3 (Inverse probability transform)**.** *Let $U \sim$ uniform$(0,1)$ and Let $F(\cdot)$ be a CDF with inverse function,*

$$F^{-1}(u) := \inf\{x \mid F(x) = u\}.$$

Then the random variable $X = F^{-1}(U)$ is distributed with CDF $F(\cdot)$.

Proof.

$$\mathbb{P}(X \leq x) = \mathbb{P}(F^{-1}(U) \leq x) = \mathbb{P}(U \leq F(x)) = F(x).$$

□

So if you want to generate from distribution X, you need to find out $F^{-1}(\cdot)$ and apply this function to a pseudorandom uniform. For continuous random variables, this is often very easy.

Exercise 52. NFQ *Generate* $10,000$ *exponential*$(1/2)$ *values. Plot their histogram. Calculate their sample mean and sample variance. Compare this to the theoretical values.*

You will often need to generate from a discrete distribution with probability masses given by some vector $\boldsymbol{p}$. Proposition 3 can be used for that.

Exercise 53. NFQ *Write a function that takes as input $\boldsymbol{p}$ of some arbitrary finite length and generates a random variable distributed according to this vector. Try this on the vector,*

$$\boldsymbol{p} = [0.35,\ 0.25,\ 0.1,\ 0.3].$$

Generate $10,000$ *values distributed according to $\boldsymbol{p}$ and compare their empirical frequencies to $\boldsymbol{p}$.*

Chapter 2

Markov Chains

2.1 Markov Chain Basics

A *stochastic process* is a random function $X(t, \omega)$ where say $t \in \mathbb{R}$ (or $\in \mathbb{Z}$) represents time and $\omega \in \Omega$ is a point in the probability sample space. An alternative view, is to think of a stochastic process as a family (sequence) of random variables $\{X(t, \omega),\, t \in \mathbb{R}\}$ (or $\in \mathbb{Z}$). Stochastic processes get interesting when the random variables are not independent. I.e. when there is some dependence structure between them. In the sequel we omit the fact that $X(\cdot, \omega)$ depends on ω from the notation, but keep in mind it is always there.

When analysing a stochastic process, we sometimes use the term *sample path* or alternatively *realisation* to refer to one instance of the time function $X(\cdot, \omega)$ associated with a single ω.

An elementary, but highly useful stochastic process is the *time homogenous finite state space discrete time Markov chain* (*finite DTMC* for short). This is a sequence of random variables indexed by $t \in \mathbb{Z}_+$ with the following three properties:

1. Lack of memory (Markovian property):

$$\mathbb{P}(X(t+1) = j \mid X(t) = i_t, \ldots, X(0) = i_0) = \mathbb{P}(X(t+1) = j \mid X(t) = i_t).$$

2. Time Homogeneity (this makes the probability law of the the process time-homogenous):

$$\mathbb{P}(X(t+1) = j \mid X(t) = i) = \mathbb{P}(X(1) = j \mid X(0) = i) := p_{i,j}.$$

3. Finite state space: There is some finite set (state space), $\mathcal{S}$, such that,

$$\mathbb{P}(X(t) \notin \mathcal{S}) = 0, \quad \forall t.$$

Since we are considering finite state space Markov chains, we may think of $\mathcal{S} = \{1, \ldots, N\}$ for some fixed integer $N \geq 2$. At the end of section we briefly also discuss infinite (but still countable) state-spaces. If you are reading these notes and have seen Markov chains before, it may be a good idea that you occasionally ask yourself, where (and how) the finite state space assumption is used. If on the other hand you have not encountered Markov chains previously, then don't let this distinction bother you.

Based on properties (1) and (2) above, it can be seen that in order to specify the probability law of the evolution of $\{X(t)\}$ we need to specify, $p_{i,j}$ for $i, j \in \mathcal{S}$ as well as the distribution of $X(0)$ (the *initial distribution*). The convenient way to specify the *transition probabilities* is by an $N \times N$ matrix $P = [p_{i,j}]$ with non-negative elements and with row sums $= 1$. I.e. each row i can be treated as a PMF indicating the distribution of $X(t+1)$ given that $X(t) = i$. A convenient way to is specify the initial distribution is by a row vector, $\mathbf{r}(0)$ of length N having non-negative elements and summing to 1 with i'th entry, $\mathbf{r}_i(0)$ meaning: $\mathbb{P}(X(0) = i) = \mathbf{r}_i(0)$. This is can again be viewed as a PMF.

Note that a non-negative matrix with row sums equal to 1 is called a *stochastic matrix*. Don't let the name confuse you; it isn't a random variable or a random matrix, it is a deterministic object.

Now using basic conditional probability and the law of total probability we can get some very nice properties. First for $t = 0, 1, 2, \ldots$, denote,

$$p_{i,j}^{(t)} = \mathbb{P}(X(t) = j \mid X(0) = i),$$

and the matrix of these probabilities by $P^{(t)} = [p_{i,j}^{(t)}]$. Also denote,

$$r_i(t) = \mathbb{P}(X(t) = i),$$

with $\mathbf{r}(t)$ being the row vector of these probabilities.

Exercise 54. *The basic dynamics of DTMCs is given by the following:*

1. *Show that $P^{(0)}$ is the identity matrix.*
2. *Show (arguing probabilistically) that $P^{(t)}$ is a stochastic matrix for any $t \in \mathbb{Z}_+$.*
3. *Show the* Chapman-Kolmogorov equations *hold:*

$$p_{i,j}^{(m+n)} = \sum_{k=1}^{N} p_{i,k}^{(m)} \, p_{k,j}^{(n)}.$$

4. *Show that $P^{(t)} = P^t$. I.e. $P^t = P \cdot P \cdot \ldots \cdot P$, where the product is of t matrices.*
5. *Show that $\mathbf{r}(t) = \mathbf{r}(0)P^t$ (the right hand side here is a row vector multiplied by a matrix).*

The next exercise, will ensure you got the point. I hope you are in the mood for doing it.

Exercise 55. *Make a model of your feelings. Say $1 \equiv$ "happy", $2 \equiv$ "indifferent", $3 \equiv$ "sad". Assume that you are Markovian (i.e. the way you feel at day $t+1$ is not affected by days prior to day t, if the feelings at day t are known)*[1].

1. *Specify the transition probabilities matrix P which you think matches you best.*
2. *Assume that at day 0 you are sad with probability 1. What is the probability of being happy in day 3.*

[1] This is perhaps a sensible assumption for guys; gals on the other hand may require more complicated models.

3. *Assume that at day 0 you have a (discrete) uniform distribution of feelings, what is the probability of being happy in day 3.*

4. *Assuming again, that the initial distribution is uniform, what is the probability of "happy, happy, sad, sad, happy" (a sequence of 5 values on times $t = 0, 1, \ldots, 4$).*

Markov chains generalised i.i.d. sequences:

Exercise 56. *Assume you are given a PMF $p_X(\cdot)$ with support $\{1, \ldots, N\}$. How can you make a Markov chain such that $\{X(t)\}$ is an i.i.d. sequence of that PMF? I.e. what matrix P will you use? Explain.*

The fact that $\mathbf{r}(t) = \mathbf{r}(0)P^t$ is remarkable and beautiful. But in general it is quite hard to have an explicit analytic expression for P^t. With some effort, you can do this for a two-state Markov chain:

Exercise 57. *Consider the Markov chain over $\mathcal{S} = \{1, 2\}$.*

1. *How many free parameters are in this model (i.e. how many numbers specify $\boldsymbol{r}(0)$ and P)?*

2. *Write an expression for P^t in terms of the parameters (e.g. do this by diagonalising the matrix P so that you can evaluate matrix powers easily).*

3. *Write an expression for $\boldsymbol{r}(t)$.*

4. *What happens to $\boldsymbol{r}(t)$ as $t \to \infty$?*

5. *Do you have any intuition on the previous result?*

2.2 First-Step Analysis

Consider a gambler; one of those hard-core TAB types. She has $X(t)$ dollars at day t. Her goal is to reach L dollars, since this is the amount needed for the new tattoo she wants[2]. She attends the bookies daily and is determined to gamble her one dollar a day, until she reaches either L or goes broke, reaching 0. On each gamble (in each day) she has a chance of p of earning a dollar and a chance of $1 - p$ of loosing a dollar.

This problem is sometimes called the *gambler's ruin* problem. We can view her fortune as the state of a Markov chain on state space, $\mathcal{S} = \{0, 1, 2 \ldots, L-1, L\}$.

Exercise 58. *Specify the transition probabilities $p_{i,j}$ associated with this model.*

At day $t = 0$, our brave gambler begins with $X(0) = x_0$ dollars. As she drives to the bookies, Jimmy texts her: "Hey babe, I was wondering what is the the chance you will eventually reach the desired L dollars?". She thinks while driving, but can't concentrate, so she stops the car by the side of the road and sketches out the following in writing: Define,

$$\tau_0 := \inf\{t \geq 0 \ : \ X(t) = 0\}, \qquad \tau_L := \inf\{t \geq 0 \ : \ X(t) = L\}.$$

[2]The tattoo will feature the name of her boyfriend, "Jimmy" together with a picture of a Holden.

These two objects are random variables which are called *hitting times* (the time it takes till hitting a state for the first time). They are random because different realisations of $X(\cdot, \omega)$ imply different values for τ_0 or τ_L. Note that the infimum of the empty set is defined to be ∞. So if our gambler, for example reaches L, then $\tau_0 = \infty$ and similarly if the other case occurs. In terms of hitting times, Jimmy's question to our gambler, was to evaluate:

$$q_i := \mathbb{P}\big(\tau_L < \tau_0 \mid X(0) = i\big), \qquad \text{with} \qquad i = x_0.$$

We define q_i for all states i, because to evaluate q_{x_0} we will need the other q_i also. It is obvious that $q_0 = 0$ and $q_L = 1$ but what if $i \in \{1, \dots, L-1\}$? Well here we can partition the event $\{\tau_L > \tau_0\}$ based on the *first step*:

$$\begin{aligned}
q_i &= \mathbb{P}\big(\tau_L < \tau_0 \mid X(0) = i\big) \\
&= \mathbb{P}\big(\tau_L < \tau_0 \mid X(0) = i,\, X(1) = i+1\big)\, p_{i,i+1} + \mathbb{P}\big(\tau_L < \tau_0,\, \mid X(0) = i,\, X(1) = i-1\big)\, p_{i,i-1} \\
&= \mathbb{P}\big(\tau_L < \tau_0 \mid X(1) = i+1\big)\, p_{i,i+1} + \mathbb{P}\big(\tau_L < \tau_0,\, \mid X(1) = i-1\big)\, p_{i,i-1} \\
&= \mathbb{P}\big(\tau_L < \tau_0 \mid X(0) = i+1\big)\, p_{i,i+1} + \mathbb{P}\big(\tau_L < \tau_0,\, \mid X(0) = i-1\big)\, p_{i,i-1} \\
&= q_{i+1}\, p + q_{i-1}(1-p).
\end{aligned}$$

So using this *first step analysis* we end up with $L+1$ equations for the $L+1$ unknowns $q_0, q_1, \dots, q_L$:

$$\begin{bmatrix}
1 & 0 & & & & & & 0 \\
(1-p) & -1 & p & & & & & \\
0 & (1-p) & -1 & p & & & & \\
& \ddots & \ddots & \ddots & \ddots & & & \\
& & & (1-p) & -1 & p & 0 & \\
& & & & (1-p) & -1 & p & \\
0 & & & & & 0 & 1 &
\end{bmatrix}
\begin{bmatrix} q_0 \\ q_1 \\ q_2 \\ \vdots \\ \vdots \\ q_{L-2} \\ q_{L-1} \\ q_L \end{bmatrix}
=
\begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ \vdots \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

The unique solution to these equations happens to be,

$$q_i = \begin{cases} i/L & \text{if } p = 1/2, \\ \frac{1-\left(\frac{1-p}{p}\right)^i}{1-\left(\frac{1-p}{p}\right)^L} & \text{if } p \neq 1/2. \end{cases} \tag{2.1}$$

Exercise 59. *Verify that the solution given above is correct.*

1. *(Analytically) – Plug it in the equations and see it satisfies them.*
2. *(Numerically) – Make a* 10×10 *matrix in matlab (or anything else) and see that the vector* q_i *solves the equations above.*
3. *(Simulation) – Simulate this gamblers ruin problem for some given parameters (say with* $L = 9$*) to verify that* q_i *is indeed correct. Basically do this by generating sample paths* $X(\cdot, \omega)$ *for all times* t*, till* $\min\{\tau_0, \tau_L\}$*.*

Exercise 60. NFQ *Assume you didn't know the formula in* (2.1)*. Think of methods in which you can obtain it. Outline your methods. Try to start with* $p = 1/2$ *and then move onto* $p \neq 1/2$*.*

The concept of *first step analysis* goes hand in hand with Markov chains and is useful for a variety of settings. When our gambler finished the calculations above, she texted Jimmy the result (q_{x_0}) and drove off. But then she got another text: "Honey love, for how many more days will you do this? Can't wait babe!". She thinks, and then figures out that Jimmy wants to know,

$$m_i := \mathbb{E}[\min\{\tau_0, \tau_L\} \mid X(0) = i] \qquad \text{with} \qquad i = x_0.$$

By now our gambler knows how to do first step analysis, even while driving. She formulates the following: First,

$$m_0 = 0 \qquad \text{and} \qquad m_L = 0.$$

Even Jimmy can do this part. But further for $i \in \{1, 2, \ldots, L-1\}$:

$$\begin{aligned} m_i &= p_{i,i+1}(1 + m_{i+1}) + p_{i,i-1}(1 + m_{i-1}) \\ &= 1 + p_{i,i+1}m_{i+1} + p_{i,i-1}m_{i-1} \\ &= 1 + p\, m_{i+1} + (1-p)\, m_{i-1} \end{aligned}$$

So again we have $L + 1$ equations with $L + 1$ unknowns.

Exercise 61. *Find the solution when* $p = 1/2$*.*

Exercise 62. NFQ *Find the solution when* $p \neq 1/2$*.*

2.3 Class Structure, Periodicity, Transience and Recurrence

Note: Some of the derivations in this section are heuristic, hence we avoid using the theorem/proof phrasing to things. Nevertheless, the reader should know that without much extra effort, all of the results can be proved in a precise manner.

One way to visualise the transition matrix of a finite DTMC is by drawing the weighted graph associated with P. Edges associated with (i, j) such that $p_{i,j} = 0$ are omitted . If you ignore the weights you simply get a directed graph. What does this graph tell you? Well, by studying it, you can see which paths the process may possibly take, and which paths are never possible. Of course, if $p_{i,j} > 0$ for all state pairs, then there is nothing to do because you have a complete graph. But in applications and theory, we often have $p_{i,j} = 0$ for a significant portion of the tuples (i, j). This allows us to study the *directed graph* that has edge (i, j) only when $p_{i,j} > 0$. This graph obviously doesn't specify all of the information about the DTMC, but it does tell us the *class structure*. We describe this now.

We say that two states, i and j *communicate* if there are two non-negative integers t_1 and t_2 such that $p_{i,j}^{(t_1)} > 0$ and $p_{j,i}^{(t_2)} > 0$. This implies there is a path (in the directed graph) from i to j and a path from j to i. We denote communication of i and j by $i \leftrightarrow j$. The relation of communication is an equivalence relation[3] over the set of states. Namely: $i \leftrightarrow i$ (reflexivity); if $i \leftrightarrow j$ then $j \leftrightarrow i$ (symmetry); and finally if $i \leftrightarrow j$ and $j \leftrightarrow k$ then $i \leftrightarrow k$ (transitivity).

[3]If for some reason you don't know what an *equivalence relation* is, don't stress. You'll understand from the text.

Exercise 63. *Use the Chapman-Kolmogorov equations to prove transitivity.*

The implication of the fact that $\leftrightarrow$ is an equivalence relation is that it induces equivalence classes, $\mathcal{C}_1, \mathcal{C}_2, \ldots$ that are a partition of $\mathcal{S}$. That is, $\mathcal{C}_i$ and $\mathcal{C}_j$ are disjoint for $i \neq j$ and $\cup_i \mathcal{C}_i = \mathcal{S}$. All states within class $\mathcal{C}_i$ communicate with each other, but do not communicate with states that are not in $\mathcal{C}_i$. Obviously for finite state spaces of size N, there can be at most N classes and this upper bound is achieved only when $P = I$, the identity matrix. At the other extreme, we are often interested in Markov chains with only one class. Such Markov chains are said to be *irreducible*.

A state i is said to have a period of d if $p_{i,i}^{(t)} = 0$ for all integers t that are not divisible by d, and further d is the greatest integer with this property. E.g, assume, that $p_{i,i}^{(3)} > 0$, $p_{i,i}^{(6)} > 0$, $p_{i,i}^{(9)} > 0$ etc... and further $p_{i,i}^{(t)} = 0$ for $t \notin \{3, 6, 9, \ldots\}$. So if we start at time 0 in state i we can only expect to be in state i at the times $3, 6, 9, \ldots$. It isn't guaranteed that at those times we visit state i, but we know that if we do visit state i, it is only at those times. It can be shown that all states in the same class have the same period. But we won't ponder on that. In general, we aren't so interested in periodic behaviour, but we need to be aware of it. In particular note that if $p_{i,i} > 0$ for all states i, then the Markov chain is guaranteed to be non-periodic.

Define now, the hitting time[4] (starting at 1): $\tau_i = \inf\{t \geq 1 \mid X(t) = i\}$ and define,

$$f_{i,j}^{(t)} = \begin{cases} \mathbb{P}(\tau_j = t \mid X(0) = i) & \text{if } t \geq 1, \\ 0 & \text{if } t = 0. \end{cases}$$

Further define $f_{i,j} = \sum_{t=1}^{\infty} f_{i,j}^{(t)}$. This is the probability of ever making a transition into state j, when starting at state i:

$$f_{i,j} = \mathbb{P}\Big(\sum_{t=1}^{\infty} \mathbf{1}\{X(t) = i\} \geq 1 \mid X(0) = i \Big).$$

A state i is said to be *recurrent* if $f_{i,i} = 1$. This means that if $X(0) = i$ we will continue visiting the state again and again. A state that is not recurrent is *transient*; i.e. i.e., $f_{i,i} < 1$ then there is a non-zero chance $(1 - f_{i,i})$ that we never return to the state.

Exercise 64. *Assume that $X(0) = i$ and state i is transient. Explain why the distribution of the number of visits to state i after time 0, is geometric with success probability $1 - f_{i,i}$ and mean $1/(1 - f_{i,i})$. I.e.,*

$$\mathbb{P}\big(\sum_{t=1}^{\infty} \mathbb{1}\{X(t) = i\} \;=\; n \mid X(0) = i \big) = (1 - f_{i,i})(f_{i,i})^n, \qquad n = 0, 1, 2, \ldots$$

Further, write an expression (in terms of $f_{i,j}$ values) for,

$$\mathbb{P}\big(\sum_{t=1}^{\infty} \mathbb{1}\{X(t) = j\} \;=\; n \mid X(0) = i \big).$$

[4]Some authors refer to the case starting at time 1 as as a first passage time and to the case starting at time 0 as a *hitting time*. This distinction only matters if the initial state is i itself.

In certain cases, it is obvious to see the values of $f_{i,j}$:

Exercise 65. *Consider the Markov chain with transition matrix,*

$$P = \begin{bmatrix} 0.3 & 0.7 & 0 \\ 0 & 0.5 & 0.5 \\ 0 & 0.5 & 0.5 \end{bmatrix}.$$

1. *What are the classes of the Markov chain.*
2. *Which states are transient, and which are recurrent.*
3. *What are $f_{i,j}$ for all i, j?*

Consider now the following example,

$$P = \begin{bmatrix} 0.1 & 0.7 & 0.2 & 0 \\ 0.4 & 0.3 & 0 & 0.3 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{2.2}$$

The classes of this example are $\mathcal{C}_1 = \{1, 2\}$, $\mathcal{C}_2 = \{3\}$ and $\mathcal{C}_3 = \{4\}$. Here without doing any calculations it is already obvious that $f_{3,3} = 1$ and $f_{4,4} = 1$, since states 3 and 4 are recurrent. They are even called *absorbing*, because once you get to state 3 or state 4, you never leave. So $f_{3,i} = 0$ for $i \neq 3$ and further $f_{4,i} = 0$ for $i \neq 4$. But the values $f_{i,j}$ with $i \in \{1, 2\}$ are not as clear. Starting in state 1, for example, there is a 0.2 chance of absorbing in 3 and and with the complement there is a chance of staying within the class $\mathcal{C}_1$. So how does this affect $f_{1,i}$?
The general mechanism we can use is first step analysis. This is the basic equation:

$$\begin{aligned} f_{i,j} &= \mathbb{P}\big(\sum_{t=1}^{\infty} \mathbb{1}\{X(t) = j\} \geq 1 \mid X(0) = i\big) \\ &= \sum_{k \neq j} \mathbb{P}\big(\sum_{t=1}^{\infty} \mathbb{1}\{X(t) = j\} \geq 1 \mid X(0) = i,\ X(1) = k\big)\, p_{i,k} \\ &+ \mathbb{P}\big(\sum_{t=1}^{\infty} \mathbb{1}\{X(t) = j\} \geq 1 \mid X(0) = i,\ X(1) = j\big)\, p_{i,j} \\ &= \sum_{k \neq j} f_{k,j}\, p_{i,k} + p_{i,j} \\ &= \sum_{k \neq j} p_{i,k} f_{k,j} + p_{i,j}. \end{aligned}$$

Exercise 66. *This exercise relates to the matrix P in* (2.2).

1. *Find $f_{1,3}$ and $f_{1,4}$ (you'll need to find out other $f_{i,j}$ values for this).*
2. *Explain why $f_{1,3} + f_{1,4} = 1$.*
3. *Run a simulation to verify your calculated value of $f_{1,3}$.*

There are many characterisations of recurrent and transient states. One neat characterisation is the following:

$$\text{State } i \text{ is recurrent if and only if } \sum_{t=0}^{\infty} p_{i,i}^{(t)} = \infty. \tag{2.3}$$

The idea of the derivation looks at the expected number of visits to the state:

$$\mathbb{E}\big[\sum_{t=0}^{\infty} \mathbb{1}\{X(t)=i\} \mid X(0)=i\big] = \sum_{t=0}^{\infty} \mathbb{E}\big[\mathbb{1}\{X(t)=i\} \mid X(0)=i\big] = \sum_{t=0}^{\infty} p_{i,i}^{(t)}$$

Now for a recurrent state, we know that $\sum_{t=0}^{\infty} \mathbf{1}\{X(t)=i\} = \infty$ and thus the expectation of this random variable should also be ∞. So this shows the direction $\Longleftarrow$. For the other direction assume that state i is transient (the contrapositive). In this case we saw that $\sum_{t=0}^{\infty} \mathbf{1}\{X(t)=i\}$ is a geometric random variable with finite expectation, so $\sum_{t=0}^{\infty} p_{i,i}^{(t)} < \infty$.

In many cases, we can't explicitly compute $p_{i,i}^{(t)}$ so there isn't much computational use for (2.3). But one classic fascinating example is the *simple random walk*. For this we assume now a state is $\mathcal{S} = \mathbb{Z}$. Take $p \in [0,1]$ and set,

$$p_{i,j} = \begin{cases} p & \text{if } j = i+1, \\ (1-p) & \text{if } j = i-1, \\ 0 & \text{otherwise.} \end{cases}$$

This is the only example in these short notes where we are considering a (countably) infinite state space. A full introductory course on stochastic processes, featuring DTMC would typically put much emphasis on many other countably infinite state space models.

The example is called a random walk because at every time step the walker takes either a step up with probability p or a step down with probability $1-p$. It is called simple, because the change at each time point is a random variable with support $\{-1, 1\}$. In the general random walk, steps would be of arbitrary magnitude.

A nice feature of this model is that we can actually calculate $p_{i,i}^{(t)}$.

Exercise 67. *Verify the following:*

1. *If $p = 0$ or $p = 1$ there is an infinite number of classes, but if $p \in (0,1)$ the model is irreducible.*
   For the rest of the items below, assume $p \in (0,1)$.

2. *The model is periodic with period 2.*
   So now we will consider $p_{i,i}^{(2t)}$, since for $t \in \{1,3,5,7,\ldots\}$, $p_{i,i}^{(t)} = 0$.

3. *Explain why:*
   $$p_{i,i}^{(2t)} = \binom{2t}{t} p^t (1-p)^t.$$

4. **NFQ** *Now use the Stirling approximation for* $t!$ *(see Appendix) to show,*

$$p_{i,i}^{(2t)} \sim \frac{\big(4p(1-p)\big)^t}{\sqrt{\pi t}},$$

where the symbol $\sim$ *implies that as* $t \to \infty$ *the ratio of the left hand side and the right hand side goes to* 1*.*

5. **NFQ** *Verify (using the definition of convergence of a series), that if* $a_t \sim b_t$ *then* $\sum_t a_t < \infty$ *if and only if* $\sum_t b_t < \infty$*.*

6. **NFQ** *Verify that*

$$\sum_{t=0}^{\infty} \frac{\big(4p(1-p)\big)^t}{\sqrt{\pi t}} = \infty,$$

if and only if $p = 1/2$ *(otherwise the series converges).*

With the results of the above exercise we know that state i (for any i) is recurrent if and only if $p = 1/2$. That is if $p \neq 1/2$ then all states are transient. Loosely speaking, the chain will "drift off" towards $+\infty$ if $p > 1/2$ and towards $-\infty$ if $p < 1/2$. States may be revisited, but ultimately, each state i will be revisited only a finite number of times.
In finite Markov chains, we can't have all states transient:

Exercise 68. *Argue why a finite DTMC, must have at least one recurrent state.*

In the infinite state space case, we can sometimes have that,

$$\mathbb{E}[\tau_i \mid X(0) = i] = \infty,$$

even when state i is recurrent. Such is actually the case for the simple random walk in the symmetric case ($p = 1/2$). This cannot happen when the state space is finite. This phenomenon is called *null-recurrence*. The other case,

$$\mathbb{E}[\tau_i \mid X(0) = i] < \infty,$$

is referred to as *positive-recurrence*. In finite state space DTMC all recurrent states are positive-recurrent. Further, in the finite state space case, if the DTMC is irreducible then all states are recurrent and thus all states are positive-recurrent.

2.4 Limiting Probabilities

We are often interested in the behaviour of $\{X(t)\}$ over long time periods. In applied mathematics, infinity, is a good approximation for "long". There is much to say here and we will only cover a small portion of the results and cases heuristically. Specifically, let us now assume that our DTMC has finite state-space, that it is irreducible, and that it is aperiodic (all states have a period of 1). Limiting probability results often hold when these assumptions are partially relaxed, but one needs to take more care in specifying the results.

To illustrate the main idea we return to exercise (55). If your example chain for that exercise had $p_{i,i} \in (0,1)$ then the above conditions are satisfied. Let us assume that this is the case. Now ask[5],

"Over the long range, in what proportion of my days am I happy?"

Remembering that our code for "happy" was 1, the question can be posed as finding

$$\pi_1 := \lim_{t\to\infty} \mathbb{E}\Big[\frac{\sum_{\ell=0}^{t} \mathbb{1}\{X(\ell)=1\}}{t}\Big].$$

The value π_1 is then referred to as the limiting probability of being in state 1. I should hope that for your Markov chain of exercise (55), π_1 is high (close to 1). How can we evaluate it? The key result is that we can solve the system of equations:

$$\begin{aligned}
\pi_1 &= \pi_1\, p_{1,1} + \pi_2\, p_{2,1} + \pi_3\, p_{3,1},\\
\pi_2 &= \pi_1\, p_{1,2} + \pi_2\, p_{2,2} + \pi_3\, p_{3,2},\\
\pi_3 &= \pi_1\, p_{1,3} + \pi_2\, p_{2,3} + \pi_3\, p_{3,3},\\
1 &= \pi_1 + \pi_2 + \pi_3.
\end{aligned}$$

Now the unique solution, $[\pi_1, \pi_2, \pi_3]$ gives the long range proportion during which state i is occupied. Note that we have 4 equations with only 3 unknowns, but we should in fact omit one (any one) of the first 3 equations (this is a consequence of the fact P is a singular matrix). These equations are called the *balance equations*. In matrix form they are compactly written with $\boldsymbol{\pi}$ taken as a row vector and $\mathbf{1}$ a column vector of 1's.

$$\begin{aligned}
\boldsymbol{\pi} &= \boldsymbol{\pi} P,\\
1 &= \boldsymbol{\pi}\mathbf{1}.
\end{aligned} \tag{2.4}$$

Exercise 69. *Consider your matrix P of exercise (55). Use a computer for the following:*

1. *Solve the balance equations for $\boldsymbol{\pi}$.*
2. *Run a single simulation of the DTMC for $T = 10,000$ time points. Choose any initial distribution for $X(0)$. Evaluate for $i \in \{1,2,3\}$,*

$$\hat{\pi}_i := \frac{\sum_{\ell=0}^{T} \mathbb{1}\{X(\ell)=i\}}{T},$$

   compare these values to the answer of item 1.
3. *Compute P^5, P^{10},P^{20} and P^{100}. Compare the rows of these matrices with the answer of item 1.*
4. *The numerical illustration of the previous item, indicates that the rows all converge to $\boldsymbol{\pi}$. If this is indeed true (which it is), argue that for any initial distribution, $\boldsymbol{r}(0)$,*

$$\lim_{t\to\infty} \boldsymbol{r}(t) = \boldsymbol{\pi}.$$

[5]Beware of such questions if your current age is $10 * n \pm \epsilon$ where ϵ is small. Such thoughts can throw you on soul adventures that you may end up regretting – or maybe not.

The numerics of the above exercise, indicate the validity of the following (we omit the proof – note also that there are much more general formulations):

Theorem 4. *Consider a finite DTMC that is irreducible and non-periodic. Then,*

1. *The balance equations* (2.4) *have a unique solution with* $\pi_i \in (0,1)$.
2. *It holds that for any* $i \in \mathcal{S}$,
$$\lim_{t\to\infty} p_{i,j}^{(t)} = \pi_j.$$
3. *It holds that,*
$$\pi_i = \frac{1}{\mathbb{E}[\tau_i \mid X(0) = i]}.$$
4. *For any function,* $f : \mathcal{S} \to \mathbb{R}$, *we have with probability one,*
$$\lim_{t\to\infty} \frac{\sum_{\ell=0}^{t} f\big(X(\ell)\big)}{t} = \sum_{i\in\mathcal{S}} \pi_i\, f(i).$$

So basically, knowing $\boldsymbol{\pi}$ gives us much information about the *long run* or *steady state* behaviour of the system. When talking about long range behaviour it is $\boldsymbol{\pi}$ that matters; the initial distribution, $\boldsymbol{r}(0)$ becomes insignificant. Item 4 (also called the *ergodic property*) shows that long range behaviour can be summarised in terms of $\boldsymbol{\pi}$.

One of the names of the distribution $\boldsymbol{\pi}$ is the *stationary distribution* also known as the *invariant distribution*. A process $\{X(t)\}$ is stationary if for any integer $k \geq 0$ and any integer values, $t_1, \ldots, t_k$, and any integer τ,

$$\mathbb{P}(X(t_1) = i_1, \ldots, X(t_k) = i_k) = \mathbb{P}(X(t_1 + \tau) = i_1, \ldots, X(t_k + \tau) = i_k).$$

Exercise 70. *Use the equations describing* $\boldsymbol{\pi}$ *to show:*

1. *If we start at time* 0 *with* $\boldsymbol{r}(0) = \boldsymbol{\pi}$, *then* $\boldsymbol{r}(1) = \boldsymbol{\pi}$ *and this holds for all* $\boldsymbol{r}(t)$.
2. *More generally, show that if we start at time* 0 *with* $\boldsymbol{r}(0) = \boldsymbol{\pi}$ *then the process is stationary.*

So when we look at a DTMC, we can consider the *stationary version* where we choose $\boldsymbol{r}(0) = \boldsymbol{\pi}$. This means we are looking at the system which is already in "statistical equilibrium". Such systems may not exactly occur in practice, but it is often a very sensible approximation for systems that have been running for a bit of time.

If on the other hand $\boldsymbol{r}(0) \neq \boldsymbol{\pi}$, then the DTMC is not stationary. But still, if we let it run for some time, it can be approximately considered to be stationary. This is due to item 2 of the theorem above.

Appendix A

Basics of Sets and Counting

A.1 Sets

A *set* is a collection of objects, e.g. $\mathcal{A} = \{1, -3, 8, a\}$. Sets are not regarded as ordered and can have a finite or infinite number of objects. $x \in \mathcal{A}$ is read as "x is an element of $\mathcal{A}$". Similarly $x \notin \mathcal{A}$. E.g. for the set above we have $1 \in \mathcal{A}$ and $4 \notin \mathcal{A}$.

We say $\mathcal{A}$ is a *subset* of $\mathcal{B}$ (denoted by $\mathcal{A} \subset \mathcal{B}$) if whenever $x \in \mathcal{A}$ we also have $x \in \mathcal{B}$. We say two sets $\mathcal{A}$ and $\mathcal{B}$ are equal (denoted $\mathcal{A} = \mathcal{B}$) if $\mathcal{A} \subset \mathcal{B}$ and $\mathcal{B} \subset \mathcal{A}$. The empty set, denoted $\emptyset$ has no elements ($\emptyset = \{\}$). It is a subset of any other set.

We often have a *universal set* (in probability theory it is often denoted Ω). Having such a set allows us to define the *complement* of any subset of Ω: $\mathcal{A}^c$. This is the set of all elements that are not in $\mathcal{A}$ but in Ω. This can also be written as,

$$\mathcal{A}^c = \{x \in \Omega : x \notin \mathcal{A}\}.$$

Note that $(\mathcal{A}^c)^c = \mathcal{A}$. Also, $\Omega^c = \emptyset$.

The *union* of two sets $\mathcal{A}$ and $\mathcal{B}$, denoted $\mathcal{A} \cup \mathcal{B}$, is the set that contains all elements that are in either $\mathcal{A}$, $\mathcal{B}$ or both. E.g. $\{-2, 0, 3\} \cup \{0, 1\} = \{0, -2, 3, 1\}$. Note that $\mathcal{A} \cup \mathcal{A}^c = \Omega$. The *intersection* of two sets $\mathcal{A}$ and $\mathcal{B}$, denoted $\mathcal{A} \cap \mathcal{B}$, is the set of all elements that are in both $\mathcal{A}$ and $\mathcal{B}$. E.g. $\{-2, 0, 3\} \cap \{0, 1\} = \{0\}$. Note that $\mathcal{A} \cap \mathcal{A}^c = \emptyset$.

Exercise 71. NFQ *Prove the following:*

1. $\mathcal{A} \cap \mathcal{B} \subset \mathcal{A} \cup \mathcal{B}$.
2. *Commutative properties:* $\mathcal{A} \cup \mathcal{B} = \mathcal{B} \cup \mathcal{A}$ *and* $\mathcal{A} \cap \mathcal{B} = \mathcal{B} \cap \mathcal{A}$.
3. *Associative properties:* $\mathcal{A} \cup (\mathcal{B} \cup \mathcal{C}) = (\mathcal{A} \cup \mathcal{B}) \cup \mathcal{C}$ *and* $\mathcal{A} \cap (\mathcal{B} \cap \mathcal{C}) = (\mathcal{A} \cap \mathcal{B}) \cap \mathcal{C}$.
4. *Distributive properties:* $\mathcal{A} \cup (\mathcal{B} \cap \mathcal{C}) = (\mathcal{A} \cup \mathcal{B}) \cap (\mathcal{A} \cup \mathcal{C})$ *and* $\mathcal{A} \cap (\mathcal{B} \cup \mathcal{C}) = (\mathcal{A} \cap \mathcal{B}) \cup (\mathcal{A} \cap \mathcal{C})$.
5. *DeMorgan's rules:* $(\mathcal{A} \cup \mathcal{B})^c = \mathcal{A}^c \cap \mathcal{B}^c$, $(\mathcal{A} \cap \mathcal{B})^c = \mathcal{A}^c \cup \mathcal{B}^c$.

Two sets $\mathcal{A}$ and $\mathcal{B}$ are said to be *disjoint* if $\mathcal{A} \cap \mathcal{B} = \emptyset$. The *difference* of $\mathcal{A}$ and $\mathcal{B}$, denoted $\mathcal{A} \backslash \mathcal{B}$ is the set of elements that are in $\mathcal{A}$ and not in $\mathcal{B}$. Note that $\mathcal{A} \backslash \mathcal{B} = \mathcal{A} \cap \mathcal{B}^c$.

We can use the following notation for unions: $\bigcup_{\gamma\in\Gamma}\mathcal{A}_\gamma$, or similarly for intersections $\bigcap_{\gamma\in\Gamma}\mathcal{A}_\gamma$. This means taking the union (or intersection) of $\mathcal{A}_\gamma$ for all γ in Γ. E.g. if $\Gamma=\{1,2\}$ it implies $\mathcal{A}_1\cup\mathcal{A}_2$ (or similarly for intersection).

Exercise 72. NFQ *Prove DeMorgan's rules for arbitrary collections:*

$$(\bigcup_{\gamma\in\Gamma}A_\gamma)^c=\bigcap_{\gamma\in\Gamma}A_\gamma^c,\qquad \textit{and}\qquad (\bigcap_{\gamma\in\Gamma}A_\gamma)^c=\bigcup_{\gamma\in\Gamma}A_\gamma^c.$$

The *power set* of a set $\mathcal{A}$, denoted $2^{\mathcal{A}}$ is the set of all subsets of $\mathcal{A}$, e.g.,

$$2^{\{a,b\}}=\{\emptyset,\ \{a\},\ \{b\},\ \{a,b\}\}.$$

A.2 Counting

For a finite set $\mathcal{A}$, $|\mathcal{A}|$ denotes the number of elements in $\mathcal{A}$. E.g. $|\{a,b,c\}|=3$. A *k-tuple* is simply an ordered list with values $(x_1,\ldots,x_k)$. The multiplication principle: The number of distinct ordered k-tuples $(x_1,\ldots,x_k)$ with components $x_i\in\mathcal{A}_i$ is $|\mathcal{A}_1|\cdot|\mathcal{A}_2|\cdot\ldots\cdot|\mathcal{A}_k|$.

Exercise 73. NFQ *Show that for $\mathcal{A}$ finite,*

$$|2^{\mathcal{A}}|=2^{|\mathcal{A}|}.$$

The number of ways to choose k objects from a finite set $\mathcal{A}$ with $|\mathcal{A}|=n$, not requiring the objects to be distinct is: n^k. This is sometimes called sampling with replacement and with ordering. Note that this also corresponds to the number of ways of distributing k distinct balls in n bins where there is no limit on the number of balls that can fit in a bin.

The number of ways to choose k distinct objects from a finite set $\mathcal{A}$ of size n where order matters is

$$n\cdot(n-1)\cdot\ldots\cdot(n-k+1).$$

I.e. this is the number of k-tuples with distinct elements selected from $\mathcal{A}$. This is number also corresponds the number of ways of distributing k distinct balls in n bins where there is a limit of at most one ball per bin. Note that if $k=n$ this number is $n!$ (e.g. $5!=5\cdot4\cdot3\cdot2\cdot1=120$). Each ordering of a finite set of size n is called a *permutation*. Thus the number of permutations is $n!$. Note Stirling's formula:

$$n!\sim\sqrt{2\pi}n^{n+\frac{1}{2}}e^{-n}.$$

The "similar sign" $\sim$ indicates that the ratio of the left hand side and right hand side converges to 1 as $n\to\infty$. Note: We often use $\sim$ to indicate the distribution of a random variable - something completely different.

The number of ways of choosing k distinct objects from a finite set $\mathcal{A}$ where order does not matter is similar to the case where order matters but should be corrected by a factor of $k!$. This number is sometimes called the binomial coefficient:

$$\left(\begin{array}{c}n\\k\end{array}\right):=\frac{n\cdot(n-1)\cdot\ldots\cdot(n-k+1)}{k!}=\frac{n!}{k!(n-k)!}.$$

I.e. this is the number of subsets of size k of a set of size n. It also corresponds to the number of ways of distributing k indistinguishable balls in a n bins with room for at most one ball per bin.

Exercise 74. NFQ *Prove each of these properties both algebraiclly and using counting arguments:*

1. $$\left(\begin{array}{c} n \\ k \end{array} \right) = \left(\begin{array}{c} n \\ n-k \end{array} \right).$$

2. $$\left(\begin{array}{c} n \\ 0 \end{array} \right) = \left(\begin{array}{c} n \\ n \end{array} \right) = 1.$$

3. $$\left(\begin{array}{c} n \\ 1 \end{array} \right) = \left(\begin{array}{c} n \\ n-1 \end{array} \right) = n.$$

4. $$\left(\begin{array}{c} n \\ k-1 \end{array} \right) + \left(\begin{array}{c} n \\ k \end{array} \right) = \left(\begin{array}{c} n+1 \\ k \end{array} \right).$$

5. *The binomial theorem:*
$$(a+b)^n = \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) a^k b^{n-k}.$$

6. $$\sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) = 2^n.$$

A.3 Countable an Not Countable Sets

The set of *natural numbers*, denoted $\mathbb{N}$ is $\{1, 2, 3, \ldots\}$. A set, $\mathcal{S}$ is said to be *countable* if it is either finite, or it is infinite and there exists a one-to-one mapping between $\mathcal{S}$ and $\mathbb{N}$, in the latter case, it is sometimes refered to as *countably infinite*.
The set of *integers*, denoted $\mathbb{Z}$ is $\{\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots\}$. The non-negative integers are denoted $\mathbb{Z}_+ := \{0\} \cup \mathbb{N}$. The set of *rational numbers*, denoted $\mathbb{Q}$ are all numbers that can be represented in the form m/n with $m, n \in \mathbb{Z}$.

Exercise 75. NFQ *Show $\mathbb{Z}$, $\mathbb{Z}_+$ and $\mathbb{Q}$ are countably infinite sets.*

The set of *reals* or *real numbers*, denoted $\mathbb{R}$ contains $\mathbb{Q}$ as well as all limits of sequences of elements in $\mathbb{Q}$. A useful subset of the reals is the interval $[0, 1] := \{x : 0 \leq x \leq 1\}$. Any element of $[0, 1]$ can be represented by an infinite sequence of binary digits such as,

$$0010100111010011111010101011010 1 \ldots,$$

by this representation it can be shown that $[0, 1]$ and hence $\mathbb{R}$ is not a countable set.

Theorem 5. *The set* $\mathbb{R}$ *is not countable.*

The above theorem is proved by assuming that $[0, 1]$ is countable and thus its elements can be ordered. Then showing that the number represented by flipping the i'th digit of the i'th element of the ordered sequence does not equal any of the ordered numbers, yet is an element of $[0, 1]$.