MATH 3170 - Elementary Stochastic Processes - Spring 2014

Class notes

(Last updated: May 5, 2014)

These notes are roughly based on the book *Essentials of Stochastic Processes* (2nd ed.) by Rick Durrett, the official text for MATH 3170. You may find, as well as contribute to, the list of typos in the book via this Wiki site, which was started by Lionel Levine and his stochastic processes class at Cornell.

To the UConn students: You may notice that Durrett makes plenty of references to Cornell (Ithaca, NY) and Duke (Durham, NC) in the textbook. As a proud Ph.D. recipient from Cornell, I can totally relate to the Ithaca references, and will happily explain them upon request. As for the Duke references, please simply replace them with the appropriate UConn terms, *e.g.* ~~Great Hall~~ → Student Union; ~~Duke Blue Devils basketball~~ → UConn Huskies basketball[1].

Acknowledgements. The author would like to thank Rich Bass, Tom Laetsch, and Nate Eldredge for graciously offering their lecture materials when they taught this class at, respectively, UConn (the first two) and Cornell.

Comments and corrections are welcome! E-mail: joe.p.chen@uconn.edu.

Contents

[1] Even if you are a closeted Jabari Parker fan, just remember that we have Shabazz to save the day.

1 Discrete-time Markov chains

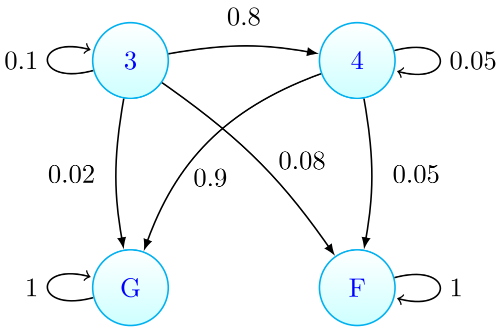
1.1 Definition and examples

Example 1.1 (Graduation). Consider a population of UConn juniors and seniors in Year 2014. We are interested in the fate of these students in Year 2015 and beyond. A junior making satisfactory progress will advance to senior standing in one year. A senior making satisfactory progress will graduate in one year. However, the following scenarios are also possible:

- A student "repeats the same grade," say, a junior stays a junior a year later.
- An ambitious junior jumps to "graduated" status after one year. (See: Emeka Okafor '04)
- A student flunks out of UConn (or transfers to a different college).

Note that once a student graduated, (s)he stays graduated. Similarly once a student flunks out, (s)he stays flunked out. We also rule out the possibility that a student "downgrades" to a lower standing, *e.g.* senior status to junior status.

Let's summarize the above scenarios by the following chain diagram:



Here the states are **3** =junior, **4** =senior, **G**=graduated, and **F**=flunked out. The numbers attached to the arrows indicate the probabilities of moving from one state to another state after one year.[2]

Question: Given the population of juniors and seniors in Year 2014, what will be the proportion of graduates amongst this group in Year 2015? Year 2016? And beyond?

Example 1.2 (Gambler's ruin). Consider the following simple gambling game. Say you start with \$10. At every turn, with probability p you win \$1, and with probability $(1-p)$ you lose \$1. You keep playing until you have $\$N$ $(N > 10)$, or until you run out of money (at which point the casino forces you to stop).

Question: What is the probability that you win $\$N$ before running out of money?

Throughout the entire course, $(\Omega, \mathcal{F}, \mathbb{P})$ denotes a probability space, where Ω is the *sample space* ("space of outcomes"), $\mathcal{F}$ is a *σ-field* on Ω ("collection of events"), and $\mathbb{P} : \mathcal{F} \to [0,1]$ is a *probability* (more precisely, a probability measure) on $(\Omega, \mathcal{F})$ which satisfies Kolmogorov's three axioms of probability. A *random variable* is a function $X : \Omega \to S$, where S is called the *state space*. In this course S may be a subset of $\mathbb{R}^d$, $\mathbb{Z}^d$, $\mathbb{N}^d$, or $\mathbb{N}_0^d$, depending on the context. (d is the dimension.) The number of elements in S is always denoted by $|S|$.

Recall the definition of **conditional probability**: for any events A and B,

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \quad \text{if } \mathbb{P}(B) > 0.$$

[2] I made up these probabilities. Whether these accurately reflect the general UConn population, or a subset thereof, is unclear.

Definition 1.1 (Discrete-time Markov chain)**.** Let $\{X_n\}_{n=0}^\infty$ be a sequence of random variables indexed by time n, with $X_n : \Omega \to S$ for each n. We say that $\{X_n\}_n$ forms a **Markov chain** if for any $n \in \mathbb{N}_0$ and any $j, i, i_{n-1}, \cdots, i_0 \in S$,

$$\mathbb{P}(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \cdots, X_0 = i_0) = \mathbb{P}(X_{n+1} = j | X_n = i). \tag{1}$$

Equation (1) is called the **Markov property**, which says that the state at time $(n+1)$ depends only on the state at time n, regardless of the information of the chain prior to time n (*i.e.,* how the chain got to state i at time n). This may remind you of the memoryless property (which you learned in MATH 3160); the connection will be made more precise later on.

In this course we will almost always assume that the chain is **temporally homogeneous**, that is, for all $n \in \mathbb{N}_0$ and $i, j \in S$,

$$\mathbb{P}(X_{n+1} = j | X_n = i) = \mathbb{P}(X_1 = j | X_0 = i) =: p(i,j),$$

where $p(i,j)$ encodes the transition probability from state i to state j, assumed to be the same for all times. Note that $p(i,j) \geqslant 0$ for all i, j, and $\sum_j p(i,j) = 1$ for all i, by the axioms of probability. [End lecture Tu 1/21]

For efficient computation it helps to adopt the language of linear algebra. Let us introduce the **transition matrix** $\mathbf{P} = \{p(i,j)\}_{i,j \in S}$, an $|S|$-by-$|S|$ matrix whose entries are the $p(i,j)$ (i indexes the row, j indexes the column). From the previous paragraph, infer that all entries of $\mathbf{P}$ are nonnegative, and that the entries along each row of $\mathbf{P}$ must sum to 1.

In the graduation example (Example 1.1), the associated transition matrix is

$$\mathbf{P} = \begin{array}{c} \\ \mathbf{3} \\ \mathbf{4} \\ \mathbf{G} \\ \mathbf{F} \end{array} \begin{array}{c} \begin{array}{cccc} \mathbf{3} & \mathbf{4} & \mathbf{G} & \mathbf{F} \end{array} \\ \begin{pmatrix} 0.1 & 0.8 & 0.02 & 0.08 \\ 0 & 0.05 & 0.9 & 0.05 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{array}.$$

In the gambler's ruin (Example 1.2), the state space is $S = \{0, 1, \cdots, N\}$ (denoting all the possible dollar values), and the transition probability for the Markov chain is

$$p(i, i+1) = p \quad \text{and} \quad p(i+1, i) = 1 - p \quad \text{for all } 1 \leqslant i \leqslant N-1,$$
$$p(0,0) = 1, \quad p(N,N) = 1,$$

where the last condition represents when the gambler stops playing. You could try to express this in matrix form if you wish. But do notice that as N gets large, the transition matrix will contain lots of 0's, *i.e.,* the matrix becomes **sparse** (a terminology often used in CS and numerical programming).

Example 1.3 (Weather chain). Let $S = \{\mathbf{S}, \mathbf{R}\}$, where $\mathbf{S}$ = sunny and $\mathbf{R}$ = rainy. Let X_n stands for the weather on day n. In an (overly) simplistic model, suppose that tomorrow's weather depends only on today's weather. Then $\{X_n\}_n$ forms a Markov chain with transition matrix, say,

$$\mathbf{P} = \begin{array}{c} \\ \mathbf{S} \\ \mathbf{R} \end{array} \begin{array}{c} \begin{array}{cc} \mathbf{S} & \mathbf{R} \end{array} \\ \begin{pmatrix} 0.6 & 0.4 \\ 0.3 & 0.7 \end{pmatrix} \end{array}.$$

Example 1.4 (Two-stage weather chain). One may argue that the example above is not realistic enough, say, that tomorrow's weather should depend on not only today's weather, **but also** yesterday's weather. In other words, X_{n+1} depends on the random vector (X_{n-1}, X_n). Based on the Markov property (1) above you may argue that $\{X_n\}_n$ is no longer a Markov chain, and you'd be correct.

However we can make it into a two-stage Markov chain. Let $Y_n = (X_n, X_{n+1})$. Then it's clear that Y_{n+1} depends solely on Y_n, and $\{Y_n\}_n$ forms a Markov chain on the state space $S^2 = \{\mathbf{SS}, \mathbf{SR}, \mathbf{RS}, \mathbf{RR}\}$.

Example 1.5 (Ehrenfest chain). This urn model is due to the physicist Paul Ehrenfest and his wife Tatianna Ehrenfest, a mathematician, as a toy model for the exchange of gas molecules taking place within two adjoining chambers. Suppose we have two urns ("left urn" and "right urn"), containing a total of N balls. At each turn, you pick up one of the N balls at random and move it to the other urn. Let X_n be the number of balls in the left urn after the nth draw. A moment's thought will tell you that $\{X_n\}_n$ is a Markov chain: indeed,

$$\mathbb{P}[X_{n+1} = i+1 | X_n = i, X_{n-1} = i_{n-1}, \cdots, X_0 = i_0] = \mathbb{P}[X_{n+1} = i+1 | X_n = i] = \frac{N-i}{N}.$$

Similarly,

$$\mathbb{P}[X_{n+1} = i-1 | X_n = i, X_{n-1} = i_{n-1}, \cdots, X_0 = i_0] = \mathbb{P}[X_{n+1} = i-1 | X_n = i] = \frac{i}{N}.$$

To write this more compactly, we recognize that the state space is $S = \{0, 1, \cdots, N\}$. Then the transition probabilities are

$$p(i, i+1) = \frac{N-i}{N} \quad \text{and} \quad p(i, i-1) = \frac{i}{N} \quad \text{for all } i \in S.$$

(What's implicit above, of course, is that $p(i,j) = 0$ whenever $j \notin \{i-1, i+1\}$.)

Example 1.6 (Repair chain). Suppose a machine has 3 critical parts that may fail with positive probability, but the machine will function so long as 2 of the parts are working. When 2 parts are broken, they are replaced and put back in working order the next day.

To set up the Markov chain, let X_n be the state of the machine on day n, which we represent by the parts that are broken: $S = \{0, 1, 2, 3, 12, 13, 23\}$. Suppose that Part 1, 2, 3 fail with respective probability .01, .02, and .04, but that no two parts fail on the same day. Then $\{X_n\}_n$ is a Markov chain with transition matrix

$$\mathbf{P} = \begin{array}{c|ccccccc} & \mathbf{0} & \mathbf{1} & \mathbf{2} & \mathbf{3} & \mathbf{12} & \mathbf{13} & \mathbf{23} \\ \hline \mathbf{0} & .93 & .01 & .02 & .04 & & & \\ \mathbf{1} & & .94 & & & .02 & .04 & \\ \mathbf{2} & & & .95 & & .01 & & .04 \\ \mathbf{3} & & & & .97 & & .01 & .02 \\ \mathbf{12} & 1 & & & & & & \\ \mathbf{13} & 1 & & & & & & \\ \mathbf{23} & 1 & & & & & & \end{array}.$$

(For clarity I have replaced the 0 entries by blanks.)

Question: If we operate this machine for 4 years, how many different parts of types 1, 2, or 3 would we need to replace?

Example 1.7 (Simple random walk on $\mathbb{Z}^d$). As you probably saw in an earlier probability course, a simple random walk on $\mathbb{Z}$ involves adding n i.i.d. random variables Y_i, where $\mathbb{P}[Y_i = +1] = p$ and $\mathbb{P}[Y_i = -1] = 1-p$. Then the displacement of the random walker after n steps is $X_n = \sum_{i=1}^n Y_i$.

Alternatively, we can say that $\{X_n\}_n$ forms a Markov chain on the state space $\mathbb{Z}$ with transition probability

$$p(x, x+1) = p \quad \text{and} \quad p(x, x-1) = 1-p \quad \text{for all } x \in \mathbb{Z}.$$

This is rather like gambler's ruin, except that now you can win an arbitrarily large sum or go in great debt. Nobody will stop you from playing the game indefinitely.

As we will see later the case $p = \frac{1}{2}$, corresponding to the **symmetric simple random walk** on $\mathbb{Z}$, is most interesting. We can also consider symmetric simple random walk on the d-dimensional integer lattice $\mathbb{Z}^d$, where in each step, the walker can move to any of the $2d$ nearest neighbors with equal probability.

Question: As time goes on, will the random walker hit the origin (the place where he started from) infinitely often, or will he go off to infinity with positive probability?

Understanding the (asymptotic) behavior of random walks on $\mathbb{Z}^d$ (or on arbitrary graphs) is one of the most fundamental topics in probability theory, and has all sorts of ramifications to other areas of math, science, engineering, and economics. So keep this example in mind; we will come back to analyze it at various points during this course.

Example 1.8 ((Galton-Watson) branching processes)*.* Consider a population in which each individual in generation n independently gives birth to k children (who belong to generation $(n+1)$) with probability p_k. To formulate this in terms of a Markov chain, let X_n be the number of individuals in generation n (so $S = \mathbb{N}_0$), and let $Y_1, Y_2, \cdots$ be independent and identically distributed (i.i.d.) random variables with $\mathbb{P}[Y_m = k] = p_k$. Then the transition probability from having i individuals in one generation to having j individuals in the next generation is

$$p(i,j) = \mathbb{P}[Y_1 + \cdots + Y_i = j] \quad \text{for all } i > 0 \text{ and } j \geqslant 0.$$

When $i = 0$ then there cannot be any descendant: $p(0,0) = 1$.

Terminology. A state i for which $p(i,i) = 1$ is called an **absorbing state**.

Question: What is the probability that the lineage of a certain individual becomes extinct?

This problem was posed by Francis Galton, who was investigating the extinction of family names. Rev. Henry William Watson gave an answer to Galton's problem. As a result the branching process is often named after the two gentlemen. We will spend about 1 lecture discussing the analysis of branching processes.

Example 1.9 (Wright-Fisher model)*.* Fix a large integer N. Suppose you have a gene population (much larger than N) with proportion i/N of type A and proportion $(N-i)/N$ of type a. Take a sample of size N, which forms the next generation of gene population. Let X_n be the number of type A genes in generation n. Then $\{X_n\}_n$ forms a Markov chain with transition probability

$$p(i,j) = \binom{N}{j} \left(\frac{i}{N}\right)^j \left(1 - \frac{i}{N}\right)^{N-j}.$$

This is the probability mass function associated to sampling with replacement.

1.2 Multistep transition probabilities

The next question we want to address is: given that the chain started at state i, what is the probability that it will be at state j after n time steps? In other words, we want to find

$$\mathbb{P}(X_n = j | X_0 = i).$$

Example calculation. Suppose $n = 2$. To move from $\{X_0 = i\}$ to $\{X_2 = j\}$ the chain needs to go through X_1, but X_1 can take on any state in S. Formalizing this idea, we have

$$\begin{aligned}
\mathbb{P}(X_2 = j | X_0 = i) &= \frac{\mathbb{P}(X_2 = j, X_0 = i)}{\mathbb{P}(X_0 = i)} = \frac{\displaystyle\sum_k \mathbb{P}(X_2 = j, X_1 = k, X_0 = i)}{\mathbb{P}(X_0 = i)} \\
&= \frac{\displaystyle\sum_k \mathbb{P}(X_2 = j | X_1 = k, X_0 = i)\mathbb{P}(X_1 = k, X_0 = i)}{\mathbb{P}(X_0 = i)} \\
&= \sum_k \mathbb{P}(X_2 = j | X_1 = k, X_0 = i) \frac{\mathbb{P}(X_1 = k, X_0 = i)}{\mathbb{P}(X_0 = i)} \\
&= \sum_k \mathbb{P}(X_2 = j | X_1 = k)\mathbb{P}(X_1 = k | X_0 = i),
\end{aligned}$$

where we repeatedly used the definition of conditional probability, and in the last line we used the Markov property (1).

Under the temporally homogeneous condition,

$$\mathbb{P}(X_2 = j | X_1 = k) = \mathbb{P}(X_1 = j | X_0 = k) = p(k, j),$$

so we can simplify the result further as

$$p^{(2)}(i, j) := \mathbb{P}(X_2 = j | X_0 = i) = \sum_k p(i, k)p(k, j).$$

The sharp-eyed reader will recognize the right-hand side as the (i, j)-entry of the matrix $\mathbf{PP} = \mathbf{P}^2$. Indeed this explains why we use matrices to study Markov chains. If we introduce the two-step transition matrix $\mathbf{P}^{(2)} = \{p^{(2)}(i, j)\}_{i,j \in S}$, then the above equality says that

$$\mathbf{P}^{(2)} = \mathbf{P}^2.$$

This generalizes to higher n: if $p^{(n)}(i, j) := \mathbb{P}(X_n = j | X_0 = i)$ and $\mathbf{P}^{(n)} = \{p^{(n)}(i, j)\}_{i,j \in S}$, then $\mathbf{P}^{(n)} = \mathbf{P}^n$. [End lecture Th 1/23] To see why this is so, we need

Theorem 1.1 (The Chapman-Kolmogorov equations)**.** *For all $n, m \in \mathbb{N}$ and all $i, j \in S$,*

$$p^{(m+n)}(i, j) = \sum_k p^{(m)}(i, k)p^{(n)}(k, j).$$

(In matrix notation, $\mathbf{P}^{(m+n)} = \mathbf{P}^{(m)}\mathbf{P}^{(n)}$.)

Proof. Replace X_1 and X_2 in the above example calculation by X_n and X_{n+m}, respectively, and use the temporally homogeneous condition at the end. □

Since $\mathbf{P}^{(1)} = \mathbf{P}$ trivially, it follows from iterating Theorem 1.1 n times that $\mathbf{P}^{(n)} = \mathbf{P}^n$.

A word about notation. From now on, we introduce the probability $\mathbb{P}^x$, defined by $\mathbb{P}^x(A) = \mathbb{P}(A | X_0 = x)$ for any event A, to be the probability given that the chain is initiated at state x.[3] We use $\mathbb{E}^x$ to denote the corresponding expectation.

Sometimes the initial state X_0 may be random. We can capture this randomness by a probability distribution μ on the state space S, so that $\mu(x) = \mathbb{P}(X_0 = x)$. If $\mathbb{P}^\mu(A)$ denotes the probability of an event A under the initial distribution μ, then by the "conditioning trick"[4]

$$\mathbb{P}^\mu(A) = \sum_x \mathbb{P}(A | X_0 = x)\mathbb{P}(X_0 = x) = \sum_x \mathbb{P}^x(A)\mu(x). \tag{2}$$

As a concrete example, take A to be the event $\{X_n = j\}$. The probability that the chain reaches state j at time n given the initial distribution μ is

$$\mathbb{P}^\mu(X_n = j) = \sum_x \mathbb{P}^x(X_n = j)\mu(x) = \sum_x p^{(n)}(x, j)\mu(x).$$

For a linear algebra interpretation, suppose that we form a row vector $\boldsymbol{\mu}$ whose entries are the $\mu(x)$. Then $\mathbb{P}^\mu(X_n = j)$ is the jth entry of the row vector $\boldsymbol{\mu}\mathbf{P}^n$. ($\boldsymbol{\mu}\mathbf{P}^n$ is the product of a 1-by-$|S|$ matrix and a $|S|$-by-$|S|$ matrix, whence a 1-by-$|S|$ matrix). In other words, if the initial distribution of the chain is $\boldsymbol{\mu}$, then at time n the distribution is $\boldsymbol{\mu}\mathbf{P}^n$.

[3] Pardon the switch of the dummy index from i to x. But as far as I can tell, most people write $\mathbb{P}^x$ instead of $\mathbb{P}^i$, for reasons that will become clear later (see: random walk).

[4] which is nothing but a reincarnation of the **multiplication rule** $\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B)$.

1.3 Classification of states

In this section we introduce the notion of a **recurrent** state vs. a **transient** state for a Markov chain on a finite state space S. We will explain where this notion comes from, and give some criteria for determining whether a state is recurrent or transient. At the end of the section, we will prove a "decomposition theorem," which says that S can be divided into a set of transient states and a number of mutually disjoint sets of recurrent states.

1.3.1 Recurrent state vs. transient state

We define the **first return time** of state y by

$$T_y = \min\{n \geqslant 1 : X_n = y\}.$$

According to this definition, $X_0 = y$ does not imply that $T_y = 0$, since the minimum is taken over $n \geqslant 1$.[5] Later in this chapter we will introduce the **first hitting time** of y by

$$U_y = \min\{n \geqslant 0 : X_n = y\},$$

whereby $X_0 = y$ does imply that $U_y = 0$. Don't be overly concerned about the word usage "hit" vs. "return." We will use both terms interchangeably so long as there is no confusion.

We will be interested in $r_{xy} := \mathbb{P}^x(T_y < \infty)$, the probability that a chain started at state x returns to state y after some finite time. In particular, $r_{yy} := \mathbb{P}^y(T_y < \infty)$ represents the probability that the chain started at y *returns* to y (after some finite time).

Let $T_y^{(1)} = T_y$ and define inductively the time of kth return to y by

$$T_y^{(k)} = \min\left\{n > T_y^{(k-1)} : X_n = y\right\} \quad \text{for } k \geqslant 2.$$

Then $\mathbb{P}^y(T_y^{(k)} < \infty)$ represents the probability that starting from y, the chain makes k returns to y. Recalling the Markov property, we may intuit that

$$\mathbb{P}^y(T_y^{(2)} < \infty) = \mathbb{P}^y(T_y^{(1)} < \infty)\mathbb{P}^y(T_y^{(1)} < \infty),$$

since upon the first visit to y, the chain could be started anew as if nothing had happened before, and the probability of revisiting y is again $\mathbb{P}^y(T_y^{(1)} < \infty)$. Extending this heuristic we may argue that

$$\mathbb{P}^y(T_y^{(k)} < \infty) = [\mathbb{P}^y(T_y < \infty)]^k.$$

Interesting things happen when we take $k \to \infty$. There is a dichotomy:

Either $\mathbb{P}^y(T_y < \infty) = 1$, in which case $\mathbb{P}^y(T_y^{(k)} < \infty) = 1$ for all $k \geqslant 1$ and $\lim_{k\to\infty} \mathbb{P}^y(T_y^{(k)} < \infty) = 1$. Almost surely, the chain returns to y a 2nd time, a 3rd time, etc. In fact, it returns to y infinitely many times;

or $\mathbb{P}^y(T_y < \infty) < 1$, in which case $\mathbb{P}^y(T_y^{(k)} < \infty) = (\rho_{yy})^k$ and $\lim_{k\to\infty} \mathbb{P}^y(T_y^{(k)} < \infty) = 0$. It is less and less likely that the chain will make an additional return to y. As such there is no chance that the chain makes an infinite number of returns to y.

Definition 1.2 (Recurrent state vs. transient state)**.** A state y is said to be **recurrent** if $\mathbb{P}^y(T_y < \infty) = 1$. Otherwise, it is **transient**.

[5] We follow the convention from analysis that $\min \varnothing = +\infty$. Indeed, if the chain never returns to y for any finite time, then $\{n \geqslant 1 : X_n = y\} = \varnothing$, and we have $T_y = \min \varnothing = +\infty$.

Everything we just said is correct except for the "heuristic" involving the Markov property, where we cheated a little. This is because the Markov property as stated in (1) applies to *nonrandom*, *deterministic* times only:

$$\mathbb{P}(X_{n+m} = j | X_n = i) = \mathbb{P}(X_m = j | X_0 = i).$$

However, $T_y^{(k)}$ is a *random* time. To make the entire argument rigorous, we need to prove that

$$\mathbb{P}\left(X_{T_y^{(k)}+m} = j \middle| X_{T_y^{(k)}} = i\right) = \mathbb{P}(X_m = j | X_0 = i). \tag{3}$$

The next subsection ties up this loose end. [End lecture Tu 1/28]

1.3.2 Stopping times & the strong Markov property

[*Note:* This subsection contains some technicalities and extra terminology which may seem overwhelming at first. If you feel that these go above your head upon first reading, take (3) on faith and skip this subsection; this will not affect your comprehension of what follows after. But do come back to this subsection in the near future, as the concepts here will reappear at several points later in the course.]

For those who did not learn or forgot the notion of a σ-field, here is the definition.[6]

Definition 1.3. Let Ω be a sample space. A collection $\mathcal{F}$ of subsets of Ω ("events") is called a **σ-field** on Ω if the following conditions are met:

(S1) $\varnothing \in \mathcal{F}$ and $\Omega \in \mathcal{F}$.

(S2) If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$.

(S3) If $A_1, A_2, \cdots, A_n \in \mathcal{F}$, then $\bigcup_{i=1}^n A_i \in \mathcal{F}$ and $\bigcap_{i=1}^n A_i \in \mathcal{F}$.

Example. The collection $\{\varnothing, \Omega\}$, being the simplest kind of a σ-field, is called the *trivial σ-field*.

Example. The power set 2^Ω, the collection of all sets composed of any number of elements in Ω, is a σ-field. Note that the number of elements in 2^Ω is $2^{|\Omega|}$, where $|\Omega|$ is the number of elements in Ω.

The σ-field which will be most useful in our context is the one generated by a set of random variables. Intuitively, this σ-field contains all information pertaining to these random variables.

Definition 1.4. Let $X_0, X_1, \cdots, X_n : \Omega \to \mathbb{R}$ be real-valued random variables. The σ-field generated by $X_0, X_1, \cdots, X_n$, denoted by $\sigma(X_0, X_1, \cdots, X_n)$, is the collection of all events of the form

$$\{X_0 \leqslant a_0, X_1 \leqslant a_1, \cdots, X_n \leqslant a_n\} \quad \text{for } a_0, a_1, \cdots, a_n \in \mathbb{R},$$

as well as the complements, countable unions, and countable intersections thereof.

If the random variables are discrete, then one can equivalently state the above condition using the events

$$\{X_0 = k_0, X_1 = k_1, \cdots, X_n = k_n\} \quad \text{for } k_0, k_1, \cdots, k_n \in \mathbb{R}.$$

This is what we will use in formulating

Definition 1.5 (Stopping time)**.** We say that a random variable $T : \Omega \to \mathbb{N}_0$ is a **stopping time** with respect to a Markov chain $\{X_n\}_n$ if, for every $n \in \mathbb{N}_0$, the (non)occurrence of the event $\{T = n\}$ ("stop at time n") is determined by $X_0, X_1, \cdots, X_n$ ("information of the chain up to time n"). In other words, both $\{T = n\}$ and $\{T \neq n\}$ belong to the σ-field $\sigma(X_0, X_1, \cdots, X_n)$.

[6] When I taught MATH 3160 last fall, I discussed σ-fields when introducing the axiomatic definition of probability during Week 2. If you were in Prof. Bass's MATH 3160 section you should have seen this as well.

Example 1.10 (Intuition behind stopping times). This example is courtesy of Prof. Richard Bass. Suppose you leave the UConn campus and drive north on Route 195 toward I-84. (This is the Markov process in question.) If you stop at the second traffic light after the intersection with Route 44, that is a stopping time. On the other hand, if you stop at the second traffic light before I-84, that is NOT a stopping time, because this time relies on information "after the fact."

We list a few more examples of stopping times.

- A nonrandom (deterministic) time n is a stopping time.
- If T is a stopping time, so is $T+n$.
- If T_1 and T_2 are stopping times, so are $T_1 \wedge T_2$ and $T_1 \vee T_2$. (Here $a \wedge b = \min(a,b)$ and $a \vee b = \max(a,b)$. This notation is standard in the probability literature, and will appear from time to time in this course.)
- Last but not least, both $T_y^{(k)}$ and $U_y^{(k)}$ are stopping times. Take $T_y^{(1)}$ for instance. For every $n \in \mathbb{N}$, $\{T_y^{(1)} = n\}$ coincides with the event
$$\{X_1 \neq y, X_2 \neq y, \cdots, X_{n-1} \neq y, X_n = y\},$$
while $\{T_y^{(1)} \neq n\}$ is the union of
$$\{X_i = y \text{ for some } 1 \leqslant i \leqslant n-1\} \quad \text{and} \quad \{X_i \neq y \text{ for all } 1 \leqslant i \leqslant n\}.$$
Since both sets are in $\sigma(X_0, X_1, \cdots, X_n)$, $T_y^{(1)}$ is a stopping time.

We now formulate the Markov property for stopping times, which allows us to close the gap from the previous subsection. Since this is a stronger condition than the Markov property stated in (1), we call this the **strong Markov property**.

Theorem 1.2 (The strong Markov property)**.** *Let $\{X_n\}_n$ be a Markov chain, and T be a stopping time with respect to $\{X_n\}_n$. Then for all $m \in \mathbb{N}$ and all $i, j \in S$,*
$$\mathbb{P}(X_{T+m} = j | X_T = i) = \mathbb{P}(X_m = j | X_0 = i).$$

Proof. We show this for $m=1$; the case $m \geqslant 2$ can be handled similarly.

$$\begin{aligned}
\mathbb{P}(X_T = i, X_{T+1} = j) &= \sum_{n=0}^{\infty} \mathbb{P}(X_T = i, X_{T+1} = j, T = n) \\
&= \sum_{n=0}^{\infty} \mathbb{P}(X_n = i, X_{n+1} = j, T = n) \\
&= \sum_{n=0}^{\infty} \mathbb{P}(X_{n+1} = j | X_n = i, T = n)\mathbb{P}(X_n = i, T = n) \\
&\overset{(\star)}{=} \sum_{n=0}^{\infty} \mathbb{P}(X_{n+1} = j | X_n = i)\mathbb{P}(X_n = i, T = n) \\
&= \sum_{n=0}^{\infty} \mathbb{P}(X_1 = j | X_0 = i)\mathbb{P}(X_n = i, T = n) \\
&= \sum_{n=0}^{\infty} p(i,j)\mathbb{P}(X_n = i, T = n) \ = \ p(i,j)\mathbb{P}(X_T = i).
\end{aligned}$$

In the second from last equality we applied the Markov property (1). Dividing both sides by $\mathbb{P}(X_T = i)$ yields the desired result. □

Remark. You might wonder where in the above proof did we use the property that T is a stopping time. It is in the equality ($\star$), where we used that fact that the event $\{X_n = i, T = n\}$ belongs to $\sigma(X_0, X_1, \cdots, X_n)$, and thus there is no loss or gain of information in replacing it with $\{X_n = i\}$.

1.3.3 Criteria for a state to be recurrent or transient

Definition 1.6. We say that state x **communicates with** state y, denoted $x \to y$, if $r_{xy} := \mathbb{P}^x(T_y < \infty) > 0$. (This means that there exists some $m \in \mathbb{N}$ such that $p^{(m)}(x, y) > 0$.)

The next result says that the communication relation is transitive. You should find this quite intuitive.

Proposition 1.3. *If $x \to y$ and $y \to z$, then $x \to z$.*

Proof. $x \to y$ implies that $p^{(m)}(x, y) > 0$ for some $m \in \mathbb{N}$. Similarly $y \to z$ implies that $p^{(n)}(y, z) > 0$ for some $n \in \mathbb{N}$. By the Chapman-Kolmogorov equations (Theorem 1.1),

$$p^{(m+n)}(x, z) = \sum_w p^{(m)}(x, w)p^{(n)}(w, z) \geqslant p^{(m)}(x, y)p^{(n)}(y, z) > 0,$$

so $x \to z$. □

The next two lemmas play a crucial role in the proof of the decomposition theorem, Theorem 1.12. In particular, Lemma 1.4 allows us to classify all transient states.

Lemma 1.4. *If $x \to y$ but $y \not\to x$, then x is transient.*

Proof. There is positive probability that the chain moves from x to y, and once at y, it never comes back to x. So $r_{xx} < 1$, which implies that x is transient. □

Lemma 1.5. *If x is recurrent and $x \to y$, then y is recurrent.*

Proof. The probability of starting at x and not hitting y before returning to x is

$$\mathbb{P}^x(T_x < T_y) = \mathbb{P}^x(T_x < \infty) - \mathbb{P}^x(T_y < \infty) = r_{xx} - r_{xy} = 1 - r_{xy},$$

where at the end we used the assumption that x is recurrent. Iterating we get that the probability of twice hitting x without hitting y is $(1 - r_{xy})^2$, and doing this k times is $(1 - r_{xy})^k$. Since $r_{xy} > 0$, $(1 - r_{xy}) < 1$, so the probability of hitting x infinitely often without *ever* hitting y is 0. It follows that the chain will hit y with probability 1. Then we can use the recurrence of x to deduce that y will be hit infinitely often with probability 1. Hence y is recurrent. □

We now discuss a useful criterion for a state to be recurrent (resp. transient), based on the expected number of return visits. This is given in Theorem 1.9, which says that y is recurrent if and only if the expected number of returns to y is infinite. This fundamental result will be invoked later in the course.

Proposition 1.6. $\mathbb{P}^x(T_y^{(k)} < \infty) = r_{xy}(r_{yy})^{k-1}$.

Proof. To make k visits to y from x, we first go from x to y, then make $(k - 1)$ returns to y. Now use the strong Markov property. □

[End lecture Th 1/30]

Let

$$N(y) = \sum_{n=1}^{\infty} \mathbb{1}_{\{X_n = y\}} \tag{4}$$

be the number of visits to y.

Proposition 1.7. $\mathbb{E}^x N(y) = \dfrac{r_{xy}}{1-r_{yy}}$.

Proof. Observe that

$$N(y) = \sum_{k=1}^{\infty} \mathbb{1}_{\{N(y)\geqslant k\}}.$$

Taking expectation $\mathbb{E}^x$ on both sides gives

$$\mathbb{E}^x N(y) = \mathbb{E}^x \left[\sum_{k=1}^{\infty} \mathbb{1}_{\{N(y)\geqslant k\}}\right] = \sum_{k=1}^{\infty} \mathbb{E}^x \left[\mathbb{1}_{\{N(y)\geqslant k\}}\right] = \sum_{k=1}^{\infty} \mathbb{P}^x[N(y)\geqslant k].$$

For those who worry about interchanging the order of the two summations $\mathbb{E}^x$ and $\sum_{k=1}^{\infty}$, remember that the state space S is finite, so $\mathbb{E}^x$ is a finite weighted sum. Therefore the interchange is legitimate.[7]

Now $\{N(y)\geqslant k\}$ is the same event as $\{T_y^{(k)} < \infty\}$, so by Proposition 1.6,

$$\mathbb{E}^x N(y) = \sum_{k=1}^{\infty} \mathbb{P}^x[T_y^{(k)} < \infty] = \sum_{k=1}^{\infty} r_{xy}(r_{yy})^{k-1} = \frac{r_{xy}}{1-r_{yy}}.$$

□

Proposition 1.8. $\mathbb{E}^x N(y) = \sum_{n=1}^{\infty} p^{(n)}(x,y)$.

Proof. Taking expectation $\mathbb{E}^x$ on both sides of (4) yields

$$\mathbb{E}^x N(y) = \mathbb{E}^x \left[\sum_{n=1}^{\infty} \mathbb{1}_{\{X_n=y\}}\right] = \sum_{n=1}^{\infty} \mathbb{E}^x \left[\mathbb{1}_{\{X_n=y\}}\right] = \sum_{n=1}^{\infty} \mathbb{P}^x[X_n=y] = \sum_{n=1}^{\infty} p^{(n)}(x,y).$$

□

Theorem 1.9. *y is recurrent if and only if*

$$\mathbb{E}^y N(y) = \sum_{n=1}^{\infty} p^{(n)}(y,y) = \infty.$$

Proof. By Proposition 1.7, $\mathbb{E}^y N(y) = \dfrac{r_{yy}}{1-r_{yy}}$. It's not hard to see that this quantity is ∞ if and only if $r_{yy}=1$, which is the definition for y being recurrent. □

1.3.4 The decomposition theorem for finite state space Markov chains

We introduce two fundamental notions pertaining to any subset of the state space S with respect to the given Markov chain.

Definition 1.7. A subset A of S is said to be **closed** if starting from A, the chain never hits $A^c = S - A$. Put it another way, whenever $x \in A$ and $y \in A^c$, $p(x,y)=0$.

Definition 1.8. A subset A of S is said to be **irreducible** (or **ergodic**) if whenever $x,y\in A$, $x\to y$.

Proposition 1.10. *In a finite closed set, there is at least one recurrent state.*

[7] If S is infinite, we have to be more careful to make sure that the interchange adheres to **Fubini's theorem** (assuming that the summand/integrand is summable/integrable) or **Tonelli's theorem** (assuming that the summand/integrand is nonnegative). For this particular instance, the interchange of summations is valid even when $|S|=\infty$.

Proof. Suppose a finite closed set A contains all transient states. By definition, a transient state is visited only finitely many times, with probability 1. Due to the finiteness of A, it follows that A is visited finitely many times, which implies that the chain must leave A, with probability 1. This contradicts the condition that A is closed. □

We are closing in on the first major milestone of this chapter.

Theorem 1.11. *If C is a finite, closed, and irreducible set, then all states in C are recurrent.*

Proof. Since C is finite and closed, Proposition 1.10 implies that there exists a recurrent state $x \in C$. Moreover, since C is irreducible, $x \to y$ for every $y \in C - \{x\}$, so Lemma 1.5 implies that every $y \in C - \{x\}$ is recurrent. □

Theorem 1.12 (The decomposition theorem)**.** *If the state space S is finite, then S equals the disjoint union*

$$\mathcal{T} \cup \mathcal{R}_1 \cup \mathcal{R}_2 \cup \cdots \cup \mathcal{R}_k,$$

where $\mathcal{T}$ is the set of transient states, and each $\mathcal{R}_i$ is a closed irreducible set of recurrent states.

Proof. Let $\mathcal{T} = \{x \in S : x \to y \text{ but } y \not\to x \text{ for some } y \in S\}$. By Lemma 1.4, all states in $\mathcal{T}$ are transient.

Take an $x_1 \in S - \mathcal{T}$ and let $C_{x_1} = \{y \in S : x_1 \to y\}$. Since $x_1 \notin \mathcal{T}$, $y \to x_1$ for every $y \in C_{x_1}$. If $y \in C_{x_1}$ and $y \to z$, then $x_1 \to z$ by Proposition 1.3 and $z \in C_{x_1}$, so C_{x_1} is closed. If $y, z \in C_{x_1}$, $y \to x_1 \to z$ and $z \to x_1 \to y$, so C_{x_1} is irreducible. Hence C_{x_1} is finite, closed, and irreducible, and Theorem 1.11 says that all states in C_{x_1} are recurrent.

If $S - \mathcal{T} - C_{x_1} = \varnothing$, we're done. If not, pick an $x_2 \in S - \mathcal{T} - C_{x_1}$ and let $C_{x_2} = \{y \in S : x_2 \to y\}$. Repeating the above argument we find that C_{x_2} is another closed irreducible set of recurrent states which is disjoint from C_{x_1}. Continue this procedure until we have exhausted the whole state space S. Conclude that S is the disjoint union of $\mathcal{T}$ and finitely many disjoint $\mathcal{R}_i = C_{x_i}$. □

Proposition 1.3, Lemmas 1.4 and 1.5, together with Theorems 1.11 and 1.12, provide us with an algorithm for classifying all recurrent and transient states of a finite state space Markov chain.

. . . . BEGIN OPTIONAL READING . . . .

Remark. In many stochastic processes books (but not Durrett) the authors bring up the notion of a **communication class**. First, we say that states i and j *intercommunicate*, denoted $i \leftrightarrow j$, if $i \to j$ and $j \to i$. It is easy to verify that the the intercommunication relation is an *equivalence relation* on S, *i.e.*,

$i \leftrightarrow i$ (reflexivity); $i \leftrightarrow j$ if and only if $j \leftrightarrow i$ (symmetry); If $i \leftrightarrow j$ and $j \leftrightarrow k$, then $i \leftrightarrow k$ (transitivity).

As such, S can be partitioned into different *equivalence classes* based on the equivalence relation $\leftrightarrow$:

$$S = S_1 \cup S_2 \cup \cdots \cup S_j,$$

where $S_i \cap S_j = \varnothing$ for all $i \neq j$, and $x \leftrightarrow y$ whenever x, y are in the same S_i. Each S_i is called a *communication class.*

Theorem 1.12 can thus be rephrased as follows: the set $\mathcal{T}$, as well as each of the $\mathcal{R}_i$, is a communication class. It's also not hard to show that within a communication class, if a state is recurrent (resp. transient), then all other states are recurrent (resp. transient).

. . . . END OPTIONAL READING . . . .

If you want to practice with some explicit examples, study Example 1.14 (a seven-state chain) in Durrett, as well as the four examples in Exercise 1.8 (two of which will be assigned as homework problems).

1.4 Stationary distributions

We say that a probability distribution μ on S is a **stationary** (or invariant) **distribution** if

$$\mathbb{P}^\mu(X_1 = y) = \mathbb{P}^\mu(X_0 = y) = \mu(y) \quad \text{for all } y \in S,$$

that is, advancing the chain by one time step does not change the distribution. For such a distribution we denote by the symbol π from now on. By (2) this equality can be rewritten as

$$\pi(y) = \sum_x \pi(x)p(x, y),$$

or in matrix notation,

$$\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}. \tag{5}$$

Recall from the end of §1.2 that if the Markov chain has initial distribution $\boldsymbol{\pi}$, then $\boldsymbol{\pi}\mathbf{P}^n$ gives the distribution at time n. Iterating (5) repeatedly we find

$$\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P} = (\boldsymbol{\pi}\mathbf{P})\mathbf{P} = \boldsymbol{\pi}\mathbf{P}^2 = (\boldsymbol{\pi}\mathbf{P})\mathbf{P}^{\mathbf{2}} = \boldsymbol{\pi}\mathbf{P}^3 = \cdots = \boldsymbol{\pi}\mathbf{P}^n = \cdots,$$

that is, a stationary distribution is invariant under the Markov chain.

You might ask: given this invariance, is it plausible that $\pi(j)$ coincides with the limiting distribution of the chain, $\lim_{n\to\infty} p^{(n)}(i, j)$, *for all* i? If this holds, we can conclude that the limiting distribution is π *regardless of the initial distribution* μ:

$$\lim_{n\to\infty} \mathbb{P}^\mu(X_n = y) = \lim_{n\to\infty} \sum_x \mu(x)p^{(n)}(x, y) = \sum_x \mu(x) \left(\lim_{n\to\infty} p^{(n)}(x, y)\right) = \sum_x \mu(x)\pi(y) = \pi(y).$$

To answer this question, we will break it down into two smaller questions:

- Under what conditions does the stationary distribution π exist?
- Assume that π exists. Under what conditions does $\lim_{n\to\infty} p^{(n)}(i, j) = \pi(j)$ hold?

A sufficient condition will be established in §1.6.

[End lecture Tu 2/4]

1.4.1 The transition rate (or infinitesimal generator) matrix Q

A matrix which will make frequent appearances throughout the course is the **transition rate** (or **infinitesimal generator**) **matrix**[8]

$$\mathbf{Q} = \mathbf{P} - \mathbf{I},$$

where **I** is the identity matrix. The **Q**-matrix plays a fundamental role in the study of Markov chain dynamics, and segues nicely to branches of analysis (differential equations) and combinatorics (graph theory).

Example 1.11 (Simple random walk). Consider the symmetric simple random walk on $\mathbb{Z}$ (or a finite subset thereof, say, $I \cap \mathbb{Z}$ where I is an interval in $\mathbb{R}$). The transition probability is given by

$$p(x, y) = \begin{cases} 1/2, & \text{if } y = x + 1 \text{ or } y = x - 1, \\ 0, & \text{otherwise.} \end{cases}$$

For any bounded function $h : \mathbb{Z} \to \mathbb{R}$ and interior point x, we find

$$(Ph)(x) = \sum_{y\in S} p(x, y)h(y) = p(x, x + 1)h(x + 1) + p(x, x - 1)h(x - 1) = \frac{1}{2}\left[h(x + 1) + h(x - 1)\right].$$

[8] Here a "transition rate" matrix is to be distinguished from a "transition" matrix. Also I have not explained where the terms "transition rate" and "infinitesimal generator" come from. They will be explained when we discuss continuous-time Markov chains, Chapter 4.

Therefore

$$(Qh)(x) = (Ph)(x) - h(x) = \frac{1}{2}\left[h(x+1) + h(x-1) - 2h(x)\right].$$

The RHS is the 2nd-order centered difference of the function h at x, which is a discrete analog of the second derivative $h''(x)$, also written as the **Laplacian** $(\Delta h)(x)$. Indeed, $\mathbf{Q}$ in this context is often called the (probabilistic) **discrete Laplacian** on $\mathbb{Z}$.

This idea can be extended to symmetric simple random walk on *any* irreducible graph $G = (V, E)$ of bounded degree (that is, every vertex $x \in V$ is connected to finitely many vertices.) We define the transition probability to be

$$p(x,y) = \begin{cases} 1/\deg(x), & \text{if } y \sim x, \\ 0, & \text{otherwise,} \end{cases}$$

where $y \sim x$ means that vertex y is connected by an edge to x. Then for any bounded function $h : V \to \mathbb{R}$ and all $x \in V$,

$$(Ph)(x) = \sum_{y \in V} p(x,y)h(y) = \sum_{y \sim x} p(x,y)h(y) = \frac{1}{\deg(x)} \sum_{y \sim x} h(y).$$

Therefore

$$(Qh)(x) = (Ph)(x) - h(x) = \frac{1}{\deg(x)} \sum_{y \sim x} h(y) - h(x) = \frac{1}{\deg(x)} \sum_{y \sim x} [h(y) - h(x)].$$

In the language of spectral graph theory, $\mathbf{P}$ is the (normalized) **adjacency matrix** of G, and $\mathbf{Q}$ the (probabilistic) **discrete Laplacian** on G. There is an elegant theory which links together random walk (probability), the discrete Laplacian (analysis), and electrical resistance (physics/EE) on graphs, which we will spell out later in the chapter.

To see the $\mathbf{Q}$-matrix in the context of stationary distribution, we have the following. If a stationary distribution π exists, then

$$\boldsymbol{\pi}\mathbf{Q} = \boldsymbol{\pi}(\mathbf{P} - \mathbf{I}) = \boldsymbol{\pi}\mathbf{P} - \boldsymbol{\pi} = 0,$$

that is, $\boldsymbol{\pi}$ lies in the left nullspace of $\mathbf{Q}$.

.............................BEGIN LINEAR ALGEBRA DISCUSSION..............................

Throughout this discussion we assume that $\dim \mathbf{Q} = |S|$ is finite. The **rank** of $\mathbf{Q}$ is the dimension of either the column space or the row space of $\mathbf{Q}$. The nullspace (resp. left nullspace) of $\mathbf{Q}$ is the space spanned by the column vectors (resp. row vectors) $\mathbf{v}$ such that $\mathbf{Qv} = 0$ (resp. $\mathbf{vQ} = 0$); its dimension is called the **nullity** (resp. **left nullity**). An important result from linear algebra is the **rank-nullity theorem**:

$$\dim \mathbf{Q} = \text{rank of } \mathbf{Q} + \text{(left) nullity of } \mathbf{Q}.$$

In principle, if the left nullity of $\mathbf{Q}$ is less than (resp. equal to, greater than) 1, then there are 0 (resp. 1, multiple) stationary distributions $\boldsymbol{\pi}$. It turns out that the left nullity of $\mathbf{Q}$ can never be 0.

Proposition 1.13. *The (left) nullity of* $\mathbf{Q}$ *is at least* 1.

Proof. Since all rows of $\mathbf{P}$ add up to 1, one can easily deduce that all rows of $\mathbf{Q}$ add up to 0: for all $1 \leqslant i \leqslant |S|$,

$$\sum_{j=1}^{|S|} Q(i,j) = 0, \quad \text{or} \quad Q(i,|S|) = -\sum_{j=1}^{|S|-1} Q(i,j).$$

The latter condition says that the last row vector of $\mathbf{Q}$ can be written as a linear combination of the first $(|S|-1)$ row vectors of $\mathbf{Q}$, that is, the last row vector is dependent upon the first $(|S|-1)$ row vectors. Therefore the rank of $\mathbf{Q}$ is at most $(|S|-1)$. By the rank-nullity theorem, the (left) nullity of $\mathbf{Q}$ is at least 1. □

[Loose end: check that all components of π have the same sign.]

.... END LINEAR ALGEBRA DISCUSSION ....

Upshot: **a stationary distribution always exists when the state space is finite.** This is however not the case if the state space is infinite, as we will see later.

That leaves us with two possibilities, a unique stationary distribution or multiple stationary distributions. It's not hard to find examples from the latter case.

Example 1.12 (A chain that sits still). Take $\mathbf{P}$ to be the identity matrix $\mathbf{I}$.[9] Then $\mathbf{Q} = \mathbf{P} - \mathbf{I}$ is the zero matrix (all entries are 0). You can take any stationary distribution you like: $(1, 0, 0, \cdots, 0)$, $(0, 1, 0, 0, \cdots, 0)$, etc., and any linear combination of these. There are infinitely many stationary distributions.

So under what conditions is a stationary distribution unique? Here's one scenario.

Proposition 1.14. *If the state space S is finite and irreducible, there exists a unique stationary measure π such that $\pi(x) \geqslant 0$ for all x, $\sum_x \pi(x) = 1$, and $\boldsymbol{\pi}\mathbf{Q} = 0$. In particular, $\pi(x) = 1/\mathbb{E}^x[T_x]$.*

For a proof heavily based on linear algebra, see Theorem 1.14 in Durrett. We will give a probabilistic proof later in §1.6.

1.5 Periodicity

We begin with a simple but instructive example.

Example 1.13 (A two-state switching chain). Let $|S| = 2$ and $\mathbf{P} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$. A straightforward matrix multiplication shows that

$$\mathbf{P}^n = \begin{cases} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, & \text{if } n \text{ is even,} \\ \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, & \text{if } n \text{ is odd.} \end{cases}$$

Therefore $p^{(n)}(x, x) > 0$ only if n is a multiple of 2: $n = 2, 4, 6, \cdots$. Note that in this case $\lim\limits_{n\to\infty} p^n(x, y)$ does NOT exist, since the sequence $\{p^n(x, y)\}_n$ oscillates between 0 and 1.

Exercise: Construct a three-state switching chain whereby $p^{(n)}(x, x) > 0$ only if n is a multiple of 3.

Here's a slightly more sophisticated example.

Example 1.14 (Random walk on a bipartite graph). A bipartite graph is a graph whose vertices can be colored either black ($\bullet$) or white ($\circ$) in such a way that every vertex is connected by an edge to a vertex of the opposite color, but never to a vertex of the same color. The integer lattice $\mathbb{Z}^d$ in any integer dimension d is a bipartite graph. The triangular lattice in the plane is not.

Let's consider a simple random walk on a bipartite graph: for all vertices x and y,

$$p(x, y) = \begin{cases} \frac{1}{\deg(x)}, & \text{if } x \text{ and } y \text{ are connected by an edge,} \\ 0, & \text{otherwise.} \end{cases}$$

It's not hard to see that all random walk trajectories are of the form $\bullet - \circ - \bullet - \circ - \cdots$ or $\circ - \bullet - \circ - \bullet - \cdots$. In order to return to where you started, you must take an even number of steps: $p^{(n)}(x, x) = 0$ whenever n is odd. As in Example 1.13, the limit $\lim\limits_{n\to\infty} p^{(n)}(i, j)$ does not exist generally. In the probability literature you will see that authors define $\lim\limits_{n\to\infty} p^{(2n)}(x, y)$ to be the limiting distribution, which thankfully does exist provided that the graph is irreducible.

[9] Observe that this is a chain where each state forms its own irreducible closed set, and hence every state is recurrent.

For each state $x \in S$ we let $I_x := \{n \in \mathbb{N} : p^{(n)}(x,x) > 0\}$. Note that if $m, n \in I_x$, then by Chapman-Kolmogorov (Theorem 1.1),

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

so $(m+n) \in I_x$.[10] (In higher math terminology we say that I_x is *closed* under addition of natural numbers.)

Definition 1.9 (Period)**.** We define the **period** of a state $x \in S$ as the *g*reatest *c*ommon *d*ivisor (gcd) of the set of integers I_x. A state is said to be **aperiodic** if its period is 1. A subset A of S is said to be aperiodic if all states in A are aperiodic.

Here's an immediate consequence of this definition: If $p(x,x) > 0$, then x is aperiodic. Indeed it's not difficult to verify that in this case, $p^{(n)}(x,x) > 0$ for every $n \geqslant 1$, that is, $I_x = \mathbb{N}$.

Example 1.15. Suppose $p^{(5)}(x,x) > 0$ and $p^{(6)}(x,x) > 0$. By Definition 1.9 x has period 1. Since $5, 6 \in I_x$, it follows that $5^2 = 25 = 5+5+5+5+5 \in I_x$. Changing each 5 into 6 we get that $26, 27, 28, 29, 30 \in I_x$. Then adding 5 to the previous numbers generate 31 through 35 in I_x, etc. From this we conclude that $p^{(n)}(x,x) > 0$ for all $n \geqslant 25$.

An easy generalization of the above example gives

Proposition 1.15. *If $k, (k+1) \in I_x$, then $p^{(n)}(x,x) > 0$ for all $n \geqslant k^2$.*

The proof is straightforward and is left as an exercise. [End lecture Th 2/6; we also covered doubly stochastic chains.]

[UPDATE W 2/12: As pointed out by Zhiguo Wang on Piazza, one can actually improve the lower bound on n in Proposition 1.15 from k^2 to $k(k-1)$. This improvement is not hard to show and will be left as an exercise for the curious student.]

The next two lemmas are the highlights of this section.

Lemma 1.16. *If $x \in S$ is aperiodic, there exists an $n_0 \in \mathbb{N}$ such that $p^{(n)}(x,x) > 0$ for all $n \geqslant n_0$.*

Proof. We need a fact from elementary number theory: given a set of positive integers $n_1, \cdots, n_p$ whose gcd is 1, there exist (not necessarily positive) integers $a_1, \cdots, a_p$ such that

$$a_1 n_1 + \cdots + a_p n_p = 1.$$

The proof of this fact can be found in any introductory number theory book, or by Googling "Diophantine equations." To give you a taste of this fact, let $p = 2$, and suppose $a_1 = 5$ and $a_2 = 8$. We want to find the integer solutions (n_1, n_2) of $5n_1 + 8n_2 = 1$. One possibility is $(n_1, n_2) = (-3, 1)$. Another possibility is $(n_1, n_2) = (5, -4)$. In fact, all possible integer solutions are of the form

$$(n_1, n_2) = (-3+8m, 1-5m), \ \ m \in \mathbb{Z}.$$

On the other hand, take $a_1 = 2$ and $a_2 = 4$. The equation $2n_1 + 4n_2 = 1$ has no integer solution (n_1, n_2) because the left-hand side and the right-hand side has different odd-even parity. Also observe that $\gcd(a_1, a_2) = \gcd(2,4) = 2 \neq 1$. A theorem from modular arithmetic says that the integer-coefficient equation $ax + by = c$ has integer solutions (x, y) *if and only if* $\gcd(a,b)$ divides c.

Back to our example, we shall take the n_j from I_x. Moreover, write $a_j = (a_j)_+ - (a_j)_-$, where $(a_j)_+ = a_j \vee 0$ and $(a_j)_- = (-a_j) \vee 0$ are, respectively, the positive and negative parts of a_j. Then we can write the above equation as

$$(a_1)_+ n_1 + \cdots + (a_p)_+ n_p = [(a_1)_- n_1 + \cdots + (a_p)_- n_p] + 1.$$

This equality implies that shows that k and $(k+1)$, where $k = [(a_1)_- n_1 + \cdots + (a_p)_- n_p]$, both belong to I_x. Now use Proposition 1.15. □

[10] Where have you seen a proof via Chapman-Kolmogorov previously? See Proposition 1.3. In fact the two proofs are utterly similar.

Lemma 1.17. *If $x \to y$ and $y \to x$, then x and y have the same period.*

Proof. Let d_x (resp. d_y) be the period of x (resp. y). If $n \in I_x$, then d_x divides n by definition. Now by assumption there exist $k, \ell \in \mathbb{N}$ such that $p^{(k)}(x,y) > 0$ and $p^{(\ell)}(y,x) > 0$. By Chapman-Kolmogorov,

$$p^{(k+\ell)}(x,x) = \sum_z p^{(k)}(x,z)p^{(\ell)}(z,x) \geqslant p^{(k)}(x,y)p^{(\ell)}(y,x) > 0,$$

so $(k+\ell) \in I_x$. Therefore d_x divides $(k+\ell)$.

Now suppose $m \in I_y$. A two-step Chapman-Kolmogorov shows that $p^{(k+m+\ell)}(x,x) > 0$, so $(k+m+\ell) \in I_x$ and d_x divides $(k+m+\ell)$. Infer then that d_x divides m. Since this is true for all $m \in I_y$, we deduce that d_x divides the gcd of I_y, or d_y.

Repeat the previous argument, but reverse the roles of x and y to deduce that d_y divides d_x. Whence $d_x = d_y$. □

1.6 Limit theorems on Markov chains

We've come to the major theorems concerning Markov chains on a finite state space S. As mentioned before, a stationary distribution π always exists in this setting.

Not all of the results below apply to the setting of infinite state space. For the precise modifications and statements see §1.9.

[In lecture I will first state these theorems without proof. After finishing the special examples (§1.7) and the one-step calculations (§1.8), I will come back to give the proofs.]

1.6.1 The ergodic theorem: irreducibility + aperiodicity ⇔ "mixing"

The **ergodic theorem** is by far the most important limit theorem for Markov chains. It says that any initial distribution will be "mixed" under the evolving Markov chain and tends to the stationary distribution as time goes to infinity. To guarantee this mixing we require two ingredients:

- Irreducibility, so that the probability mass can be redistributed across all states; and
- Aperiodicity, so that the dynamics of the mass redistribution does not get trapped in a limit cycle.

Theorem 1.18 (Ergodic theorem)**.** *Suppose S is finite, irreducible, and aperiodic. Then $\lim_{n\to\infty} p^{(n)}(x,y) = \pi(y)$ for all $x, y \in S$.*

In particular this implies the uniqueness of the stationary distribution π. See also Theorem 1.20.

Proof. The key technique used in this proof goes by the name of **coupling**. Instead of looking at one single Markov chain, we consider two independent copies X_n and Y_n of the same Markov chain, and put $Z_n := (X_n, Y_n)$. Then for all $n \in \mathbb{N}_0$ and all $i, j, k, l \in S$,

$$\begin{aligned} & \mathbb{P}(Z_{n+1} = (j,l) | Z_n = (i,k), Z_{n-1}, \cdots, Z_0) \\ = \quad & \mathbb{P}(Z_{n+1} = (j,l)|Z_n = (i,k)) = \mathbb{P}(X_{n+1} = j|X_n = i)\mathbb{P}(Y_{n+1} = l|Y_n = k) = p(i,j)p(k,l), \end{aligned}$$

that is, $\{Z_n\}_n$ forms a Markov chain on $S \times S$. Moreover, since S is irreducible and *aperiodic* with respect to the single chain, there exists $N = N(i,j,k,l)$ such that $p^{(n)}(i,j)p^{(n)}(k,l) > 0$ for all $n \geqslant N$. This guarantees that $S \times S$ is irreducible with respect to $\{Z_n\}_n$. (Note that this is the only place where aperiodicity is used in the proof. *Exercise:* Find an explicit example in which not all states in S are aperiodic with respect to the single chain, which leads to $S \times S$ being reducible with respect to $\{Z_n\}_n$.)

We will initiate Z_n as follows: start X_n from state $x_1 \in S$, and start Y_n from the stationary distribution π on S. We will let $\mathbb{P}$ denote the probability law of this chain subject to these initial conditions. As S is finite and irreducible with respect to X_n (and Y_n), it is easy to deduce that $S \times S$ is finite and irreducible with respect to Z_n, hence all states in $S \times S$ are recurrent by Theorems 1.11 and 1.12. Therefore with probability 1, the two chains $\{X_n\}_n$ and $\{Y_n\}_n$ will collide at some finite time (or put in another way, $\{Z_n\}_n$ will "hit the diagonal"): there exist some $n \in \mathbb{N}$ such that $X_n = Y_n$.

With this in mind let us introduce the following coupling. Run $\{X_n\}$ and $\{Y_n\}$ independently until the time of their first collision U. After the collision the two chains are merged, that is, set $X_n = Y_n$ for all $n \geqslant U$. By this construction, for all $n \in \mathbb{N}$ and all $y \in S$,

$$\mathbb{P}(X_n = y) = \mathbb{P}(X_n = y, U \leqslant n) + \mathbb{P}(X_n = y, U > n) = \mathbb{P}(Y_n = y, U \leqslant n) + \mathbb{P}(X_n = y, U > n),$$

and

$$\mathbb{P}(Y_n = y) = \mathbb{P}(Y_n = y, U \leqslant n) + \mathbb{P}(Y_n = y, U > n).$$

Subtracting the second equation from the first we find

$$\mathbb{P}(X_n = y) - \mathbb{P}(Y_n = y) = \mathbb{P}(X_n = y, U > n) - \mathbb{P}(Y_n = y, U > n) \leqslant \mathbb{P}(U > n).$$

Reversing the role of x and y and repeat the argument. Then we get

$$|\mathbb{P}(X_n = y) - \mathbb{P}(Y_n = y)| \leqslant \mathbb{P}(U > n) \underset{n\to\infty}{\longrightarrow} 0.$$

But $\mathbb{P}(Y_n = y) = \pi(y)$. This shows that $\lim_{n\to\infty} |\mathbb{P}(X_n = y) - \pi(y)| = 0$, or $\lim_{n\to\infty} p^{(n)}(x_1, y) = \pi(y)$. □

1.6.2 Frequency of visits tends to the stationary distribution

This is the Markov chain version of the law of large numbers (LLN).

Theorem 1.19. *Suppose S is finite, irreducible, and recurrent. If $N_n(y)$ denotes the number of visits to $y \in S$ up to time n, then with probability 1,* $\lim_{n\to\infty} \frac{N_n(y)}{n} = \frac{1}{\mathbb{E}^y[T_y]}$.

Proof. Realize that each successive time T_y of return to y is independent, so by the **strong law of large numbers**,

$$\frac{T_y^{(k)}}{k} \underset{k\to\infty}{\longrightarrow} \mathbb{E}^y[T_y] \quad \text{with probability 1.}$$

An equivalent way of saying this is that we make, on average, a return trip to y in a period $\mathbb{E}^y[T_y]$, that is,

$$\frac{N_n(y)}{n} \underset{n\to\infty}{\longrightarrow} \frac{1}{\mathbb{E}^y[T_y]} \quad \text{with probability 1.}$$

If this does not look convincing to you, here is a rigorous argument. Let $R(k) = \min\{n \geqslant 1 : N_n(y) = k\}$. Then $R(N_n(y)) \leqslant n < R(N_n(y) + 1)$. Dividing all sides of this inequality by $N_n(y)$ yields

$$\frac{R(N_n(y))}{N_n(y)} \leqslant \frac{n}{N_n(y)} \leqslant \frac{R(N_n(y)+1)}{N_n(y)} < \frac{R(N_n(y)+1)}{N_n(y)+1} \cdot \frac{N_n(y)+1}{N_n(y)}.$$

Taking $n \to \infty$ you'd find that the left and right sides of the inequality both converge to $\mathbb{E}^y[T_y]$, so it must be that $n/N_n(y) \to \mathbb{E}^y[T_y]$. □

Theorem 1.20. *Suppose S is finite and irreducible. Then* $\pi(y) = \frac{1}{\mathbb{E}^y[T_y]}$.

This theorem can interpreted as follows: If S is irreducible, then there exists at most 1 stationary distribution. Note that we do not require aperiodicity in this theorem.

Proof. By Theorem 1.19 we have

$$\frac{N_n(y)}{n} \to \frac{1}{\mathbb{E}^y[T_y]}.$$

Let π be a stationary distribution (which exists by the given assumptions). Then

$$\mathbb{E}^\pi[N_n(y)] = \mathbb{E}^\pi\left[\sum_{i=1}^n \mathbb{1}_{\{X_i=y\}}\right] = \sum_{i=1}^n \mathbb{E}^\pi[\mathbb{1}_{\{X_i=y\}}] = \sum_{i=1}^n \mathbb{P}^\pi[X_i=y] = \sum_{i=1}^n \pi(y) = n\pi(y).$$

Therefore $\dfrac{\mathbb{E}^\pi[N_n(y)]}{n} = \pi(y)$. In light of the limit result above we conclude that $\pi(y) = 1/\mathbb{E}^y[T_y]$. □

1.7 Special examples

1.7.1 Doubly stochastic chains

Remember that for every Markov chain, the entries along each *row* of the transition matrix $\mathbf{P}$ must add up to 1: $\sum_j p(i,j) = 1$ for all i. If furthermore the entries along each *column* of $\mathbf{P}$ add up to 1, $\sum_i p(i,j) = 1$ for all j, then we say that the transition matrix or the Markov chain is **doubly stochastic**.

For a finite state space ($|S| < \infty$) doubly stochastic chain, it turns out that the uniform distribution, $\pi(x) = \frac{1}{|S|}$ for all $x \in S$, is a stationary distribution. This can be verified via an explicit calculation:

$$\sum_x \pi(x)p(x,y) = \sum_x \frac{1}{|S|}p(x,y) = \frac{1}{|S|}\sum_x p(x,y) = \frac{1}{|S|} = \pi(y).$$

Remark. Here the assumption of finite state space is important. If S is infinite, $|S| = \infty$ and $\frac{1}{|S|} = 0$, so it doesn't make sense to call $\pi(x) = \frac{1}{|S|}$ a probability distribution. Nevertheless, we can still define a **stationary measure** m on S which satisfies $\mathbf{m} = \mathbf{mP}$, whether the total mass of m is finite (so m can be normalized into a probability) or infinite (so m cannot be normalized). In the case of an infinite state space doubly stochastic chain, $\mathbf{m}$ can be taken to be the infinite row vector consisting of all 1's.

1.7.2 Chains satisfying detailed balance

The term "detailed balance" is more closely associated with physics and chemistry, especially in statistical mechanics, the study of systems consisting of a large number of particles (such as fluids and gases). In the pure probability literature the term "symmetric" is more often used.

Definition 1.10. We say that a Markov chain is **symmetric** with respect to a probability distribution π on S if the condition of **detailed balance**,

$$\pi(x)p(x,y) = \pi(y)p(y,x),$$

holds for all $x, y \in S$.

For instance, suppose x and y represent two compounds participating in a chemical reaction, and let $\pi(x)$ and $\pi(y)$ represent the amount of x and y in the flask. Let $p(x,y)$ be the rate per mole at which x is turned into y, and vice versa for $p(y,x)$. To say that this chemical system obeys detailed balance means that at any given time, the rate of production of y from x equals the rate of production of x from y. You can infer that the reaction has reached an equilibrium.[11]

The reason we used the symbol π in the definition above is because it is a stationary distribution. (In the chemical reaction example, π would be the equilibrum concentration of the two chemicals.)

[11] To the chemistry-minded reader, I hope this paragraph makes perfect sense to you. (I was quite a chemist before freshman year in college...)

Proposition 1.21. *Given the conditions in Definition 1.10, π is a stationary distribution.*

Proof. Using the condition of detailed balance we have

$$\sum_x \pi(x)p(x,y) = \sum_x \pi(y)p(y,x) = \pi(y)\sum_x p(y,x) = \pi(y).$$

□

While every distribution which satisfies the condition of detailed balance is a stationary distribution, the converse is NOT true. There are many stationary distributions which do NOT satisfy the condition of detailed balance.

Example 1.16 (A stationary distribution which does not satisfy detailed balance)*.* Take $S = \{1,2,3\}$ and $\mathbf{P} = \begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.3 & 0.5 \\ 0.4 & 0.4 & 0.2 \end{bmatrix}$. This is a doubly stochastic chain, so the stationary distribution is the uniform distribution on S: $\pi(x) = \frac{1}{3}$ for $x \in \{1,2,3\}$. On the other hand, $\pi(1)p(1,2) = \frac{1}{3}\cdot 0.3$ and $\pi(2)p(2,1) = \frac{1}{3}\cdot 0.2$, so the detailed balance condition is not satisfied.

[End lecture Tu 2/11]

1.7.3 Time-reversed chains

Imagine a Markov chain as a movie "moving forward in time." Now suppose we rewind the movie "backward in time" at speed 1x. Is the time-reversed chain a Markov chain as well?

To fix ideas, we assume that the Markov chain $\{X_n\}_n$ is started from a stationary distribution π.

Example 1.17*.* Consider a three-state Markov chain, where $S = \{a,b,c\}$ and

$$\mathbf{P} = \begin{array}{c|ccc} & a & b & c \\ \hline a & 0 & 1/2 & 1/2 \\ b & 1/3 & 0 & 2/3 \\ c & 1/3 & 2/3 & 0 \end{array}.$$

The stationary distribution is $(\pi_a, \pi_b, \pi_c) = (1/4, 3/8, 3/8)$, and in fact this Markov chain is symmetric with respect to π.[12]

Consider the following sequence of transitions: $a \to b \to c \to b$. Starting from π_a, the probability of this transition happening is

$$\pi_a p(a,b)p(b,c)p(c,b) = \frac{1}{4}\cdot\frac{1}{2}\cdot\frac{2}{3}\cdot\frac{2}{3} = \frac{1}{18}.$$

Now reverse the transition: $b \to c \to b \to a$. Starting from π_b, the probability of the reverse transition happening is

$$\pi_b p(b,c)p(c,b)p(b,a) = \frac{3}{8}\cdot\frac{2}{3}\cdot\frac{2}{3}\cdot\frac{1}{3} = \frac{1}{18}.$$

So the probability of the forward transition and the probablity of the time-reversed transition are equal.

Exercise: Pick your favorite sequence of transitions and verify that for this example, the probability of the forward transition equals the probability of the reverse transition.

So given the probabilities of transitions, you would not be able to whether the chain is evolving *forward* in time or *backward* in time. The chain is now **reversible** with respect to π: you cannot tell the "arrow of time" anymore.

[12] You might ask how I knew this. This example comes from a particular electrical network. See below.

Reversibility is an important concept in the study of stochastic processes.[13] To put this in greater generality, we state a proposition and a definition. [I will state these in class, but will leave the proof of the proposition for your reading.]

Proposition 1.22. *Fix $n \in \mathbb{N}$, and let $Y_m = X_{n-m}$ for $0 \leqslant m \leqslant n$. Then the time-reversed chain $\{Y_m\}_{m=0}^n$ is a Markov chain with transition probability*

$$\hat{p}(i,j) = \frac{\pi(j)p(j,i)}{\pi(i)}.$$

Proof. We need to show that for all $0 \leqslant m \leqslant n$ and all $i_0, \cdots, i_{m-1}, i, j \in S$,

$$\mathbb{P}[Y_{m+1} = j | Y_m = i] = \mathbb{P}[Y_{m+1} = j | Y_m = i, Y_{m-1} = i_{m-1}, \cdots, Y_0 = i_0] = \frac{\pi(j)p(j,i)}{\pi(i)}.$$

First of all,

$$\begin{aligned} & \mathbb{P}[Y_{m+1} = j, Y_m = i, Y_{m-1} = i_{m-1}, \cdots, Y_0 = i_0] \\ = \;& \mathbb{P}[X_{n-m-1} = j, X_{n-m} = i, X_{n-m+1} = i_{m-1}, \cdots, X_n = i_0] \\ = \;& \mathbb{P}[X_{n-m-1} = j] \cdot \mathbb{P}[X_{n-m} = i | X_{n-m-i} = j] \cdot \mathbb{P}[X_{n-m+1} = i_{m-1}, \cdots, X_n = i_0 | X_{n-m-i} = j, X_{n-m} = i] \\ = \;& \pi(j) \cdot p(j,i) \cdot \mathbb{P}[X_{n-m+1} = i_{m-1}, \cdots, X_n = i_0 | X_{n-m} = i]. \qquad (6) \end{aligned}$$

The decisive step is in the second equality, where we used the multiplication rule. The three terms in the third line were then respectively transformed according to the following facts: that a stationary distribution is invariant under the Markov chain; the definition of a transition probability; and the Markov property. By a similar and slightly easier reasoning,

$$\begin{aligned} & \mathbb{P}[Y_m = i, Y_{m-1} = i_{m-1}, \cdots, Y_0 = i_0] \\ = \;& \mathbb{P}[X_{n-m} = i, X_{n-m+1} = i_{m-1}, \cdots, X_n = i_0] \\ = \;& \mathbb{P}[X_{n-m} = i] \cdot \mathbb{P}[X_{n-m+1} = i_{m-1}, \cdots, X_n = i_0 | X_{n-m} = i] \\ = \;& \pi(i) \cdot \mathbb{P}[X_{n-m+1} = i_{m-1}, \cdots, X_n = i_0 | X_{n-m} = i]. \qquad (7) \end{aligned}$$

The claim follows from dividing (6) by (7). □

Now we want the condition that the forward transition probability equals the backward (time-reversed) transition probability.

Definition 1.11. We say that a Markov chain started from an initial stationary distribution π is **reversible** (with respect to π) if for all $i, j \in S$,

$$\hat{p}(i,j) = \frac{\pi(j)p(j,i)}{\pi(i)} = p(i,j).$$

This is nothing but the symmetry (or detailed balance) condition, *cf.* Definition 1.10. Therefore we have

Proposition 1.23. *The following are equivalent for a Markov chain and a probability distribution π on S:*

(a) The Markov chain is ***symmetric*** *with respect to π (*i.e., *satisfies the condition of* ***detailed balance****).*

(b) The Markov chain is ***reversible*** *with respect to π.*

Now that you know that symmetry, detailed balance, and reversibility are equivalent conditions for Markov chains, we turn to some important examples of reversible chains.

[13] The study of *equilibrium statistical mechanics* often involve deriving physical laws (for fluids and gases) which are invariant under time-reversal. Understanding irreversible processes is one of the main challenges in *non-equilibrium statistical mechanics*.

1.7.4 Some examples of reversible chains

Example 1.18 (Birth-and-death chains). Here $S = \{\ell, \ell+1, \cdots, r-1, r\} \subset \mathbb{Z}$, and $p(x,y) = 0$ whenever $|x-y| > 1$. In other words, we allow only "birth" by 1, "death" by 1, or "stay put." Let's introduce the shorthands $p_x := p(x, x+1)$, $q_x := p(x, x-1)$, so $p(x,x) = 1 - p_x - q_x$. (Insert figure.)

To find the stationary distribution π, we start with $\pi(\ell)$ and find the rest of π by solving the equations $\pi(y) = \sum_x \pi(x)p(x,y)$ iteratively. Starting from the very left we have

$$\pi(\ell) = \pi(\ell)p(\ell,\ell) + \pi(\ell+1)p(\ell+1,\ell) \Longrightarrow \pi(\ell)p_\ell = \pi(\ell+1)q_{\ell+1},$$

while

$$\begin{aligned}
& \pi(\ell+1) = \pi(\ell)p(\ell,\ell+1) + \pi(\ell+1)p(\ell+1,\ell+1) + \pi(\ell+2)p(\ell+1,\ell+2) \\
\Longrightarrow \quad & \pi(\ell+1)(p_{\ell+1} + q_{\ell+1}) = \pi(\ell)p_\ell + \pi(\ell+2)q_{\ell+2} \\
\Longrightarrow \quad & \pi(\ell+1)(p_{\ell+1} + q_{\ell+1}) = \pi(\ell+1)q_{\ell+1} + \pi(\ell+2)q_{\ell+2} \\
\Longrightarrow \quad & \pi(\ell+1)p_{\ell+1} = \pi(\ell+2)q_{\ell+2}.
\end{aligned}$$

In the end we find

$$\pi(x)p_x = \pi(x+1)q_{x+1} \quad \text{for all } x \in \{\ell, \ell+1, \cdots, r-1\},$$

which is nothing but the symmetry/reversible condition. As long as the q_i don't vanish, π will be well defined. (Of course, we choose $\pi(\ell)$ suitably to make π a probability distribution: $\sum_{x \in S} \pi(x) = 1$.)

Some examples of birth-and-death chains include: the Ehrenfest chain (Example 1.5); the Bernoulli-Laplace model of diffusion (Exercise 1.46 in Durrett, assigned in HW2); gambler's ruin; and simple random walk on the line.

Example 1.19 (Random walk on graphs). Let $G = (V, E)$ be a finite, connected graph of bounded degree. (This means that $1 \leqslant \deg(x) < \infty$ for all $x \in V$.) A symmetric simple random walk $\{X_n\}_n$ on G is defined as the Markov chain on the set V of vertices with transition probability $p(x,y) = A(x,y)/\deg(x)$, where

$$A(x,y) = \begin{cases} 1, & \text{if } \langle xy \rangle \in E, \\ 0, & \text{otherwise,} \end{cases}$$

is the **adjacency matrix** of G, which is clearly a symmetric matrix. For simplicity we write $x \sim y$ whenever $\langle xy \rangle \in E$. You can check from the assumptions that this Markov chain is irreducible. Therefore a unique stationary distribution π exists.

To find π, we exploit the symmetry of the adjacency matrix and obtain the following identity:

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

Therefore $c\deg(\cdot)$ satisfies the condition of detailed balance, and becomes the desired stationary distribution provided that we normalize it into a probability distribution. The correct normalization is $c = \left[\sum_{y \in S} \deg(y)\right]^{-1}$. So the symmetric simple random walk on G has stationary distribution

$$\pi(x) = \frac{\deg(x)}{\sum_{y \in S} \deg(y)}.$$

This whole machinery can be generalized to weighted graphs, *i.e.,* electrical networks. Consider a resistor network $G = (V, E)$ whereby each resistor $e = \langle xy \rangle \in E$ has conductance C_{xy} $(= C_{yx})$, which is one over the resistance. A random walk on this resistor network is defined with transition probability

$$p(x,y) = \frac{C_{xy}}{\sum_{z \sim x} C_{xz}}$$

whenever $y \sim x$, and 0 otherwise. Then the result of the previous paragraph can be carried over to this setting by replacing the degree $\deg(x)$ by the "capacitance" $C_x := \sum_{z \sim x} C_{xz}$. [JPC: Clarify the nature of the capacitance C_x. Doyle & Snell claims that there is a way to understand C_x as the charge divided by voltage at x. Make this physics connections clear.]

[Add the triangular network example with resistances 1, 1 and 1/2.]

1.8 One-step calculations

The guiding example for this section is gambler's ruin.

Example 1.20 (Gambler's ruin, Part I). Suppose you carry \$$x$ ($x \in \mathbb{N}_0$, $x \leqslant 10$) to play a turn-based game at a casino. In every turn, with probability p you win \$1, and with probability $(1-p)$ you lose \$1. The game stops when either you win \$10 or you lose all money. What is the probability that you will win \$10 before you lose all money?

Let X_n be the money you have at time n. Then $\{X_n\}_n$ forms a Markov chain on the state space $S = \{0, 1, \cdots, 10\}$ with transition probability

$$p(i, i+1) = p \quad \text{if } 1 \leqslant i \leqslant 9; \quad p(i, i-1) = 1-p \quad \text{if } 1 \leqslant i \leqslant 9; \quad p(0,0) = 1; \quad p(10,10) = 1.$$

For each $y \in S$, let $U_y := \min\{n \geqslant 0 : X_n = y\}$ be the **first hitting time of** y. Unlike in the definition of T_y, we now take the minimum over all $n \geqslant 0$, *inclusive of* 0, in defining U_y. More generally, for any subset $A \subset S$, we set the first hitting time of A to be

$$U_A := \min\{n \geqslant 0 : X_n \in A\}.$$

Remember that we use the convention $\min \varnothing = +\infty$.

So if you translate the problem into symbols, what we are asked is to find

$$h(x) := \mathbb{P}^x(U_{10} < U_0),$$

where we regard h as a function on S.

Obviously $h(0) = 0$ and $h(10) = 1$. When $1 \leqslant x \leqslant 9$, we can use the Markov property applied to one-step transitions from x to find that for any event A,

$$\begin{aligned} \mathbb{P}^x(A) &= \mathbb{P}(A|X_0 = x) = \sum_y \mathbb{P}(X_1 = y|X_0 = x)\mathbb{P}(A|X_1 = y, X_0 = x) \\ &= \sum_y p(x,y)\mathbb{P}(A|X_1 = y) = \sum_y p(x,y)\mathbb{P}^y(A). \end{aligned}$$

This is the basis of the so-called "one-step calculations." Put $A = \{U_{10} < U_0\}$ in the previous equation to get that for $1 \leqslant x \leqslant 9$,

$$h(x) = \sum_y p(x,y)h(y) = p(x, x-1)h(x-1) + p(x+1, x)h(x+1) = (1-p)h(x-1) + p \cdot h(x+1).$$

To summarize, we need to solve the following difference equation:

$$\begin{cases} h(x) = (1-p)h(x-1) + p \cdot h(x+1), & \text{if } 1 \leqslant x \leqslant 9, \\ h(x) = 1, & \text{if } x = 10, \\ h(x) = 0, & \text{if } x = 0. \end{cases} \tag{8}$$

How to solve this equation? First let's look at the case $p = 1/2$. Then the first equation boils down to $h(x) = \frac{1}{2}[h(x-1) + h(x+1)]$: the value of h at x is the *average/mean* of the values of h at its neighbors. This is true as long as x is an "interior point" of the space S, that is, any integer but 0 or 10. A moment's thought will tell you that the graph of h must be a straight line, which passes through $(0,0)$ and $(10,1)$. (Draw it!) From this infer that $h(x) = x/10$.

What you just found is the **harmonic function** in one dimension subject to the boundary conditions $h(0) = 0$ and $h(10) = 1$. [In fact the letter h stands for harmonic.] You may have encountered the notion of harmonic function in analysis or differential equations, in the sense that h satisfies the second-order equation

$$\begin{cases} \Delta h(x) = 0, & \text{if } 1 \leqslant x \leqslant 9, \\ h(x) = 1, & \text{if } x = 10, \\ h(x) = 0, & \text{if } x = 0. \end{cases} \tag{9}$$

where Δ is the **Laplacian** in 1 dimension. To be more precise, this Laplacian is nothing but the **Q**-matrix, or the 2nd-order centered difference operator: $(\Delta h)(x) = \frac{1}{2}[h(x+1) + h(x-1) - 2h(x)] = (Qh)(x)$.

Definition 1.12. A function h on a subset $A \subset S$ is **harmonic** with respect to a Markov chain $\{X_n\}_n$ on S if the **mean value property**

$$h(x) = (Ph)(x) \text{ for all } x \in A$$

holds. (Equivalently, if $(Qh)(x) = 0$ for all $x \in A$.)

In matrix form you should represent h as a column vector $\mathbf{h}$, and solve for the matrix equation $\mathbf{Ph} = \mathbf{h}$. (Don't confuse this with $\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}$!) Note that h almost always comes with **boundary conditions**, that is, the values of h are already specified on the "boundary" $S - A$. To solve for h on A, one way is to write down the linear algebra equation in block matrix form

$$\begin{bmatrix} \mathbf{P}_{A,A} & \mathbf{P}_{A,S-A} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{h}_A \\ \mathbf{h}_{S-A} \end{bmatrix} = \begin{bmatrix} \mathbf{h}_A \\ \mathbf{h}_{S-A} \end{bmatrix},$$

where we've arranged the ordering of the states so that those in the interior A come before those on the boundary $S - A$. This equation can be stated equivalently using the **Q**-matrix as

$$\begin{bmatrix} \mathbf{Q}_{A,A} & \mathbf{P}_{A,S-A} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{h}_A \\ \mathbf{h}_{S-A} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix},$$

as you can verify by using $\mathbf{Q} = \mathbf{P} - \mathbf{I}$.

Going back to the gambler's ruin example, we now consider the case $p \neq \frac{1}{2}$. The graph of the harmonic function in this case will not be a straight line, so a bit of calculation is needed. But the idea remains the same: we extend from the boundary values $h(0)$ and $h(10)$ "inward" to get the interior values $h(x)$, $1 \leqslant x \leqslant 9$, using the (weighted) mean value property. (This procedure is called "harmonic extension.") The answer is

$$h(x) = \frac{\left(\frac{1-p}{p}\right)^x - 1}{\left(\frac{1-p}{p}\right)^{10} - 1}.$$

We just demonstrated a fairly nice example of the so-called **Dirichlet problem**, whose solution is known as a harmonic function. This may beg the following questions for general Dirichlet problems:

- Does a harmonic function always exist?
- If it exists, is it unique?
- In 1 dimension we saw that the harmonic function attains its maximal (and minimal) value on the boundary (points) of S. Does this **maximum** (and **minimum**) **principle** hold in more generality?

For an irreducible Markov chain on a finite state space, the answers to all three questions are affirmative. We'll explain this shortly.

. . . .

Let's turn to a different question involving gambler's ruin.

Example 1.21 (Gambler's ruin, Part II). Consider the same setup as in Part I. What is the expected first time that you will end the game, that is, when you hit either \$10 or \$0?

This time we are looking at $g(x) := \mathbb{E}^x[U_B]$, where $B = \{0, 10\}$. Note that $g(0) = 0$ and $g(10) = 0$. For $1 \leqslant x \leqslant 9$, we again apply the "one-step" reasoning to find

$$g(x) = 1 + \sum_y p(x,y)g(y).$$

Here the $+1$ comes from the fact that 1 time step has elapsed upon the jump from x to y. Plugging in the numbers for $p(x,y)$ we find

$$g(x) = 1 + p \cdot g(x+1) + (1-p)g(x-1) \quad \text{for } 1 \leqslant x \leqslant 9.$$

This leads to the difference equation

$$\begin{cases} g(x) = 1 + p \cdot g(x+1) + (1-p)g(x-1), & \text{if } 1 \leqslant x \leqslant 9, \\ g(x) = 0, & \text{if } x \in \{0, 10\}. \end{cases} \tag{10}$$

If $p = 1/2$ then the solution turns out to be nice: $g(x) = x(10-x)$. Notice that the value of g is maximal at $x = 5$, the point which is furthest from either boundary point, which makes intuitive sense. It's also possible to solve for g when $p \neq 1/2$. [End lecture Th 2/13]

At this moment you may be curious about the use of the letter g to denote this expected exit time. The g stands for Green's, as in Green's function. The connection is as follows.

Recall that $\mathbb{1}_x$ is the indicator function for x, that is, $\mathbb{1}_x(y) = 1$ if $y = x$, and 0 if $y \neq x$.

Definition 1.13. Fix a subset $A \subset S$. A function $g : A \times A \to \mathbb{R}$ is called a **Green's function** with respect to a Markov chain $\{X_n\}_n$ on S if for all $x_1, x_2 \in A$,

$$g(x_1, x_2) = \mathbb{1}_{x_1}(x_2) + \sum_y p(x_1, y)g(y, x_2)$$

(Equivalently, for each $x \in A$, $(Qg)(x, \cdot) = -\mathbb{1}_x$ on A.)

Once again, we should specify the boundary values of g. In all the examples we will discuss in this course, $g(x_1, x_2) = 0$ whenever either x_1 or x_2 is in $S - A$. Then we say that g is a **Green's function with Dirichlet boundary condition** on $S - A$.[14]

If you learned Green's function from a course in ODE/PDE[15], this is the pretty much the same Green's function we're talking about. On the other hand, our $g(x)$ above is not a Green's function, but the summation/integral of a Green's function over the second argument: $g(x) = \sum_y g(x, y)$. Unfortunately there is no short name for the integral of a Green's function, as far as I'm aware of.

. . . .

Having looked at gambler's ruin, we now turn to a more interesting example involving random walk on graphs.

Example 1.22 (Random walk on resistor networks). [To be filled in after class Tu 2/18]

. . . .

Now that you have seen some explicit examples, we discuss some general and rigorous statements concerning one-step calculations for Markov chains. Modulo the probabilistic interpretations, these statements can also be found in any differential equations book. (In fact, for those who are in the know, one of the most elementary questions concerning a differential equation is, "Does the solution *exist*? Is the solution *unique*?")

[14] For those who already know DiffEq, you may ask what is the Green's function with *Neumann* boundary conditions, or even *Robin* boundary conditions. If you are interested I can tell you in person.

[15] and not from physics. A physicist's Green's function may mean several different things, not all which a mathematician recognizes as the Green's function (s)he knows. From my experience this has created some "language barrier" between physicists and mathematicians.

1.8.1 The maximum principle for harmonic functions

We will restrict our attention to resistor networks.

[To be filled in after class Tu 2/18]

1.8.2 Exit (or hitting) distributions (h)

Theorem 1.24. *Consider a Markov chain on a finite state space S. Let $A, B \subset S$, $A \cap B = \varnothing$, and assume that $\mathbb{P}^x(U_A \wedge U_B < \infty) > 0$ for all $x \in S - (A \cup B)$. Then the* ***Dirichlet problem***

$$\begin{cases} h(x) = \sum_{y \in S} p(x,y)h(y), & \text{for all } x \in S - (A \cup B), \\ h(x) = 1, & \text{for all } x \in A, \\ h(x) = 0, & \text{for all } x \in B, \end{cases} \tag{11}$$

has a unique solution $h(x) = \mathbb{P}^x(U_A < U_B)$.

Recall that the first equation encodes the "mean value property" for harmonic functions.

Proof. We break down the proof into two parts: show the existence of a solution, then prove the uniqueness of the solution. The existence part is where probability comes in; the uniqueness part involves no probability at all.

Existence. Let $h(x) = \mathbb{P}^x(U_A < U_B)$. It is clear that $h = 1$ on A and $h = 0$ on B, so it suffices to check that this h satisfies the first equation. By the Markov property (the basis for the one-step calculation),

$$\mathbb{P}^x(U_A < U_B) = \sum_{y \in S} p(x,y)\mathbb{P}^y(U_A < U_B) \quad \text{for all } x \in S - (A \cup B),$$

as long as $\mathbb{P}^x(U_A \wedge U_B < \infty) > 0$. This proves the existence.

Uniqueness. Suppose h_1 and h_2 are two different solutions to the Dirichlet problem. Then by the maximum principle, $h_1 = h_2$ on all of S. This is a contradiction.

□

1.8.3 Expected hitting (or exit, or first passage) times (g)

Theorem 1.25. *Consider a Markov chain on a finite state space S. Let $A \subset S$, and suppose that $\mathbb{P}^x(U_A < \infty) > 0$ for all $x \in S - A$. Then the equation*

$$\begin{cases} g(x) = 1 + \sum_{y \in S} p(x,y)g(y), & \text{for all } x \in S - A, \\ g(x) = 0, & \text{for all } x \in A, \end{cases} \tag{12}$$

has a unique solution $g(x) = \mathbb{E}^x[U_A]$.

1.9 Infinite state space Markov chains

1.9.1 Positive recurrent, null recurrent, and transient

Definition 1.14 (Positive recurrent vs. null recurrent)**.** An irreducible, recurrent Markov chain on an infinite state space is said to be **null recurrent** if $\mathbb{E}^x[T_x] = \infty$ for some $x \in S$ (and hence for all $x \in S$ by irreducibility). Otherwise, it is **positive recurrent**.

[Justify the "for all $x \in S$" part.]

Theorem 1.26. *An irreducible, recurrent Markov chain on an infinite state space is positive recurrent if and only if a stationary distribution π exists. In this case, π is unique and*

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

Proof. JPC: Need to give proofs of both directions, see Nate Eldredge's notes. □

1.9.2 Reflected simple random walk on $\mathbb{Z}_+$

1.9.3 Simple random walk on $\mathbb{Z}$

Fix $p \in (0,1)$ and let $q = 1-p$.[16] Consider the Markov chain on $\mathbb{Z}$ with transition probability

$$p(x,y) = \begin{cases} p, & \text{if } y = x+1, \\ q, & \text{if } y = x-1, \\ 0, & \text{otherwise,} \end{cases}$$

for all $x, y \in \mathbb{Z}$. This describes a simple random walk on $\mathbb{Z}$: if $p = 1/2$ (resp. $p \neq 1/2$) the walk is *symmetric* (resp. *asymmetric*) in both directions. If you wish, think of it as a gambler's ruin with no holds barred; you can be as rich as you want, or be as much in debt as you want.

In order we will answer the following questions:

(Q1) What is the probability that a random walker returns to the origin after $2n$ steps? (Remember that $\mathbb{Z}$ is a bipartite graph, so to return to where you started, you must take an even number of steps.)

(Q2) What is the probability that the random walker's first return to the origin occurs at time $2n$?

(Q3) Will the random walker return to the origin infinitely often? (Recurrent or transient?) And if so, is the expected time of return finite or infinite? (Positive recurrent or null recurrent?)

(Q4) Maximum of the random walk. (Use the reflection principle)

Answer to (Q1). This is an exercise at the level of MATH 3160. To return to the origin after $2n$ steps, you must take n steps to the right and n steps to the left. The number of possible trajectories, *i.e.*, ways to arrange the left and right steps, is $\binom{2n}{n}$. And since each trajectory happens with probability $p^n q^n$, we conclude that

$$p_{2n} := \mathbb{P}^0(X_{2n} = 0) = \binom{2n}{n} p^n q^n.$$

Exercise: For any $n \in \mathbb{N}$ and $x \in \mathbb{Z}$, find an explicit formula for $\mathbb{P}^0(X_n = x)$.

Answer to (Q2). We are after $r_{2n} := \mathbb{P}^0[T_0 = 2n]$, where as before $T_y = \min\{n \geqslant 1 : X_n = y\}$. Observe that whenever $n \neq m$, the events $\{T_0 = 2n\}$ and $\{T_0 = 2m\}$ are disjoint, and $\Omega = \bigcup_{n=1}^{\infty}\{T_0 = 2n\}$. Applying the conditioning trick followed by the strong Markov property, we have that for all $n \in \mathbb{N}$,

$$\mathbb{P}^0(X_{2n} = 0) = \sum_{m=1}^{n} \mathbb{P}^0[X_{2n} = 0 | T_0 = 2m]\mathbb{P}^0[T_0 = 2m] = \sum_{m=1}^{n} \mathbb{P}^0[X_{2(n-m)} = 0]\mathbb{P}^0[T_0 = 2m],$$

or

$$p_{2n} = \sum_{m=1}^{n} p_{2(n-m)} r_{2m} \quad (n \geqslant 1). \tag{13}$$

Also

$$p_0 = \mathbb{P}^0(X_0 = 0) = 1. \tag{14}$$

[16] Looking at the transition probability, you should be able to see why the cases $p = 0$ and $p = 1$ are not interesting.

If we introduce two functions $P, R : \mathbb{N}_0 \to \mathbb{R}$ defined by $P(n) = p_{2n}$ and $R(n) = r_{2n}$, then (13) and (14) imply that

$$P(n) = \mathbb{1}_{\{0\}}(n) + (P \star R)(n),$$

where we used the convolution introduced in Appendix B.3.

To solve for R, it is best to invoke the generating functions

$$\widehat{P}(x) = \sum_{n=0}^{\infty} p_{2n} x^n \quad \text{and} \quad \widehat{R}(x) = \sum_{n=0}^{\infty} r_{2n} x^n.$$

Utilizing Proposition B.4, we find that

$$\widehat{P}(x) = 1 + \widehat{P}(x)\widehat{R}(x),$$

or

$$\widehat{R}(x) = \frac{\widehat{P}(x) - 1}{\widehat{P}(x)}.$$

It suffices to find

$$\widehat{P}(x) = \sum_{n=0}^{\infty} \binom{2n}{n} p^n q^n x^n.$$

Using the rather curious generating function identity[17]

$$\frac{1}{\sqrt{1-4x}} = \sum_{n=0}^{\infty} \binom{2n}{n} x^n,$$

we deduce that

$$\widehat{P}(x) = \frac{1}{\sqrt{1-4pqx}} \quad \text{and} \quad \widehat{R}(x) = 1 - \sqrt{1-4pqx}.$$

To find r_{2n}, we use the fact that

$$\widehat{R}'(x) = \frac{1}{2\sqrt{1-4pqx}} = \frac{1}{2}\widehat{P}(x).$$

Writing this out in series,

$$\sum_{n=1}^{\infty} n r_{2n} x^{n-1} = \sum_{m=0}^{\infty} (m+1) r_{2(m+1)} x^m = \frac{1}{2} \sum_{m=0}^{\infty} p_{2m} x^m.$$

Matching the series term by term yields

$$r_{2n} = \frac{p_{2(n-1)}}{2n} = \frac{1}{2n} \binom{2(n-1)}{n-1} (pq)^{n-1} = \cdots.$$

Answer to (Q3). To determine whether this Markov chain is recurrent or not, we look at

$$\mathbb{P}^0(T_0 < \infty) = \sum_{n=1}^{\infty} \mathbb{P}^0(T_0 = 2n) = \sum_{n=1}^{\infty} r_{2n}.$$

The RHS is none other than $\widehat{R}(1)$, which equals $1 - \sqrt{1-4pq}$. There are two possibilities:

[17] which I will either justify in class, or leave it as a homework problem.

- $p \neq 1/2$: Then $1-4pq > 0$ and $\widehat{R}(1) = 1-\sqrt{1-4pq} < 1$. **An asymmetric simple random walk on $\mathbb{Z}$ is transient.** It will eventually escape to $+\infty$ (if $p > 1/2$) or $-\infty$ (if $p < 1/2$), which is not a surprising conclusion.
- $p = 1/2$: Then $1-4pq = 0$ and $\widehat{R}(1) = 1$. **A symmetric simple random walk on $\mathbb{Z}$ is recurrent.**

Moreover, when $p = 1/2$,

$$\mathbb{E}^0[T_0] = \sum_{n=1}^{\infty} 2n \cdot \mathbb{P}^0(T_0 = 2n) = 2\sum_{n=1}^{\infty} nr_{2n} = 2\widehat{R}'(1) = \widehat{P}(1) = \infty,$$

so **a symmetric simple random walk on $\mathbb{Z}$ is null recurrent**.

[JPC: Give an alternative proof of null recurrence using Wald's identity. Then leave another proof, using the expected number of returns, as a homework exercise.]

Answer to (Q4).

1.9.4 Symmetric simple random walk on $\mathbb{Z}^d$, $d \geqslant 2$

Moral of the story: *One should not get drunk in* 3 *dimensions (or higher)!*

1.9.5 Galton-Watson branching processes

This is a Markov chain $\{X_n\}_{n=0}^{\infty}$ on $\mathbb{N}_0 = \{0,1,2,\cdots\}$. The rules for this branching process are:

- Fix an offspring distribution $\{p_k\}_{k=0}^{\infty}$, where p_k represents the probability that an individual gives birth to k children. Clearly $\sum_{k=0}^{\infty} p_k = 1$.
- Let X_n be the number of individuals in generation n. Almost always we start with $X_0 = 1$. Then each individual in generation n independently gives birth according to the offspring distribution. The children borne out of generation n form the individuals in generation $(n+1)$.

When X_n hits 0 for some n, we say that the family line goes extinct. The key question to this section is: for which offspring distributions does the branching process have a (non)zero probability of surviving (and not going extinct)?

Theorem 1.27. *Let $\mu := \sum_{k=0}^{\infty} kp_k$ be the mean number of offsprings per parent. Then the following trichotomy holds:*

- *If $\mu < 1$, then with probability* 1 *the branching process goes extinct, and the mean extinction time is finite.*
- *If $\mu = 1$, excluding the trivial case $p_1 = 1$, then with probability* 1 *the branching process goes extinct, and the mean extinction time is infinite.*
- *If $\mu > 1$, then there is a positive probability that the branching process survives.*

There is an easy argument for the case $\mu < 1$; see Durrett, p.56. To resolve the other two cases we need some machinery.

Let ρ be the extinction probability starting from the beginning "ancestor." Using one-step calculation, we may argue that

$$\rho = \sum_{k=0}^{\infty} p_k \rho^k,$$

where in the RHS we made a one-step transition to the various possibilities in generation 1, then ask what the probability that *all* individuals in generation 1 have their family lines go extinct. If there are k individuals, the probability that *all* their lines become extinct is k powers of ρ by independence.

We reserve a special notation for the right-hand sum:

$$\phi(\rho) = \sum_{k=0}^{\infty} p_k \rho^k.$$

This is called the **generating function** of the probability distribution $\{p_k\}_{k=0}^{\infty}$. You have seen the *moment generating function* in MATH 3160, which encodes all the moments of a given probability distribution. Generating functions are useful devices in the study of combinatorics and probability. Here are some basic properties, which are left as an exercise for you.

Proposition 1.28. *The following holds for a generating function $\phi(\rho)$ of $\{p_k\}_{k=0}^{\infty}$:*

(a) $\phi(0) = p_0$, $\phi(1) = 1$.

(b) ϕ is a (strictly) monotone increasing function on $[0,1]$.

(c) ϕ' is also a (strictly) monotone increasing function on $[0,1]$. In particular, $\phi'(1) = \mu$.

Another property which we will use, but is harder to prove, is that ϕ is continuous on $[0,1]$.

We now state the main lemma to proving Theorem 1.27.

Lemma 1.29. *The extinction probability ρ is the smallest root of the equation $x = \phi(x)$ for $0 \leqslant x \leqslant 1$.*

Proof. To begin let us make another one-step calculation. Let $\rho_n = \mathbb{P}^1[X_n = 0]$, the probability that one's family line goes extinct at generation n. Then by the same argument as before,

$$\mathbb{P}^1[X_n = 0] = \sum_{k=0}^{\infty} p_k \left(\mathbb{P}^1[X_{n-1} = 0]\right)^k,$$

or $\rho_n = \phi(\rho_{n-1})$.

Now since 0 is an absorbing state, $\rho_n \geqslant \rho_{n-1}$ for each n. Therefore the sequence $\{\rho_n\}_{n=0}^{\infty}$ is monotone nondecreasing. A fundamental result from undergraduate analysis is that *a monotone sequence always has a limit, be it finite or infinite.* In our case, since $\rho_n \leqslant 1$ for each n, the sequence is bounded and therefore the limit must be finite (bounded by 1). Let's call $\rho_* = \lim_{n\to\infty} \rho_n$; this is the probability of eventually going extinct. Then by the monotonicity of ϕ, we can take the following limit

$$\lim_{n\to\infty} \rho_n = \lim_{n\to\infty} \phi(\rho_{n-1}) = \phi\left(\lim_{n\to\infty} \rho_{n-1}\right) \quad \Rightarrow \quad \rho_* = \phi(\rho_*).$$

Thus far we have shown that ρ_* is a root of the equation $x = \phi(x)$. We want to show that it is in fact the smallest root. So let ρ be this smallest root. By the fact that $\rho \geqslant \rho_0 = 0$ and $\rho_n = \phi(\rho_{n-1})$, we utilize the monotonicity of the function ϕ to deduce that

$$\rho = \phi(\rho) \geqslant \phi(\rho_0) = \rho_1.$$

Similarly

$$\begin{aligned} \rho \geqslant \rho_1 \quad &\Rightarrow \quad \rho = \phi(\rho) \geqslant \phi(\rho_1) = \rho_2 \\ \rho \geqslant \rho_2 \quad &\Rightarrow \quad \rho = \phi(\rho) \geqslant \phi(\rho_2) = \rho_3, \quad \text{etc.} \end{aligned}$$

From this conclude that $\rho \geqslant \rho_*$. But since ρ is already the smallest root, it must be that $\rho = \rho_*$. □

Proof of Theorem 1.27. With Lemma 1.29 in place, it suffices to ask if the equation $x = \phi(x)$ has a root for $x \in [0,1)$. (Note that 1 is always a (trivial) root.) Believe it or not, this is essentially a calculus exercise.

Let $g(x) := \phi(x) - x$. Then $g'(x) = \phi'(x) - 1$, and $g'(1) = \mu - 1$. The fundamental theorem of calculus says that for all $x \in [0,1)$,

$$g(1) - g(x) = \int_x^1 g'(w)dw \quad \Rightarrow \quad g(x) = g(1) - \int_x^1 g'(w)dw = -\int_x^1 g'(w)dw.$$

Since ϕ' is strictly monotone increasing, so is $g' = \phi' - 1$. Therefore $g'(w) \lneq g'(1)$ for all $w \in [0,1)$. It follows that

$$g(x) \gneq -\int_x^1 g'(1)dw = -g'(1) \cdot (1-x) = -(\mu-1)(1-x).$$

So if $\mu < 1$ or $\mu = 1$ (with the exception of the trivial case $p_1 = 1$), this implies that $g(x) > 0$ for all $x \in [0,1)$, that is, $\phi(x) \gneq x$ for all $x \in [0,1)$. The equation $x = \phi(x)$ has only one root for $x \in [0,1]$, namely, 1. Extinction probability is 1.

If $\mu > 1$, we again apply the fundamental theorem of calculus, but make the opposite bound. For any small $\epsilon > 0$,

$$g(1-\epsilon) = g(1) - \int_{1-\epsilon}^1 g'(w)dw = -\int_{1-\epsilon}^1 g'(w)dw \lneq -\epsilon \cdot g'(1-\epsilon)(\lneq 0)$$

Meanwhile we have

$$g(0) = \phi(0) - 0 = p_0(\geqslant 0).$$

This means that the function g has a positive value at 0 and a negative value at $1-\epsilon$. Since g is continuous, we can use the **intermediate value theorem** to deduce that $g(y) = 0$, or $y = \phi(y)$, for some $y \in [0, 1-\epsilon)$.

There is only a loose end to tie up, namely, to show that the mean extinction time for the case $\mu = 1$ (again excluding the trivial case) is infinite. As mentioned in Durrett, this relies upon an advanced result which can be found in the monograph "Branching Processes" by Athreya and Ney, Section 1.9: If the offspring distribution has mean 1 and variance $\sigma^2 > 0$, then asymptotically

$$\mathbb{P}^1[X_n > 0] \sim \frac{2}{n\sigma^2}.$$

Therefore

$$\mathbb{E}^1[T_0] = \sum_n \mathbb{P}^1[X_n > 0] \sim \frac{2}{\sigma^2}\sum_n \frac{1}{n} = \infty.$$

□

2 Poisson processes

2.1 What you did not learn in MATH 3160

Everything that you learned about exponential and Poisson random variables from MATH 3160 can be found in Appendix A.2. I will not waste much time going over these in lecture.

Let's highlight a few things which MATH 3160 did not cover. Recall the notation $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$.

Proposition 2.1. *Let $S \sim \text{Exp}(\lambda)$ and $T \sim \text{Exp}(\mu)$ be independent exponential random variables. Then*

(a) $(S \wedge T) \sim \text{Exp}(\lambda + \mu)$.

(b) $\mathbb{P}(S < T) = \dfrac{\lambda}{\lambda + \mu}$.

(c) Let

$$I = \begin{cases} 1, & \textit{if } S < T, \\ 2, & \textit{if } S > T. \end{cases}$$

Then I and $S \wedge T$ are independent.

Proof. (a): The event $\{(S \wedge T) > t\}$ is the same as $\{S > t, T > t\}$. By independence,

$$\mathbb{P}((S \wedge T) > t) = \mathbb{P}(S > t, T > t) = \mathbb{P}(S > t)\mathbb{P}(T > t) = e^{-\lambda t} e^{-\mu t} = e^{-(\lambda+\mu)t},$$

which is what one expects for a random variable with distribution $\text{Exp}(\lambda + \mu)$.

(b): Let $f_{S,T}(s, t)$ be the joint density function for S and T. By independence,

$$f_{S,T}(s, t) = f_S(s) f_T(t) = \left(\lambda e^{-\lambda s}\right)\left(\mu e^{-\mu t}\right) \mathbb{1}_{\{s \geqslant 0, t \geqslant 0\}}.$$

Therefore

$$\begin{aligned} \mathbb{P}(S < T) &= \int_{-\infty}^{\infty} \int_{-\infty}^{t} f_{S,T}(s, t)\, ds\, dt = \int_0^\infty \int_0^t \left(\lambda e^{-\lambda s}\right)\left(\mu e^{-\mu t}\right) ds\, dt \\ &= \int_0^\infty \mu e^{-\mu t}\left(1 - e^{-\lambda t}\right) dt = 1 - \frac{\mu}{\mu + \lambda} = \frac{\lambda}{\mu + \lambda}. \end{aligned}$$

(c): We'd like to check that the joint probability of I and $S \wedge T$ factorizes into the product of individual probabilities. For each $r > 0$,

$$\begin{aligned} \mathbb{P}(I = 1, (S \wedge T) > r) &= \mathbb{P}(r < S < T) = \int_r^\infty \int_r^t f_{S,T}(s, t)\, ds\, dt = \int_r^\infty \int_r^t \left(\lambda e^{-\lambda s}\right)\left(\mu e^{-\mu t}\right) ds\, dt \\ &= \int_r^\infty \mu e^{-\mu t}(e^{-\lambda r} - e^{-\lambda t})\, dt = e^{-(\lambda+\mu)r} - \frac{\mu}{\mu + \lambda} e^{-(\lambda+\mu)r} \\ &= \frac{\lambda}{\mu + \lambda} e^{-(\mu+\lambda)r} = \mathbb{P}(I = 1)\mathbb{P}((S \wedge T) > r), \\ \mathbb{P}(I = 2, (S \wedge T) > r) &= \mathbb{P}(r < T < S) = \cdots \text{(details omitted)} \cdots \\ &= \frac{\mu}{\mu + \lambda} e^{-(\mu+\lambda)r} = \mathbb{P}(I = 2)\mathbb{P}((S \wedge T) > r). \end{aligned}$$

This shows the independence. □

One can generalize Proposition 2.1 to n independent exponentials. The proof is mostly by induction and hence omitted.

Proposition 2.2. *Let $X_i \sim \text{Exp}(\lambda_i)$, $1 \leqslant i \leqslant n$, be independent exponential random variables. Then*

(a) $\min(X_1, \cdots, X_n) \sim \text{Exp}(\lambda_1 + \cdots + \lambda_n)$.

(b) $\mathbb{P}(X_i = \min(X_1, \cdots, X_n)) = \dfrac{\lambda_i}{\sum_{j=1}^n \lambda_j}$.

(c) Let I be the random integer i for which $X_i = \min(X_1, \cdots, X_n)$. Then the random variables I and $\min(X_1, \cdots, X_n)$ are independent.

2.2 Simple Poisson process

Definition 2.1 (Construction #1 of Poisson process)**.** Let $\tau_1, \cdots, \tau_n, \cdots$ be i.i.d. exponential random variables of parameter λ, and for each $n \in \mathbb{N}$, let $T_n = \tau_1 + \cdots \tau_n$. (By default put $T_0 = 0$.) We can construct a Poisson process of rate λ $\{N(t) : t \geqslant 0\}$ by setting $N(t) = \max\{n : t \geqslant T_n\}$.

Definition 2.2 (Construction #2 of Poisson process)**.** Let $\{N(t) : t \geqslant 0\}$ be a process governed by the following properties:

(a) $N(0) = 0$.

(b) For each $t > 0$, $N(t)$ is a Poisson random variable of parameter λt.

(c) $N(t)$ has **independent increments**, that is, $N(s+t) - N(s)$ is independent of $\{N(r) : r \leqslant s\}$.

Then $N(t)$ is a Poisson process of rate λ.

Theorem 2.3. *The two constructions of Poisson process, Definition 2.1 and Definition 2.2, are equivalent.*

To check this theorem there are a few facts one needs to verify. They can all be found in Durrett.

- For each $s \geqslant 0$, $N(s+t) - N(s)$ is a Poisson process of rate λ.
- $N(T_1)$, $N(T_2) - N(T_1)$, $\cdots$, $N(T_{k+1}) - N(T_k)$, etc. are all independent.

2.3 Compound Poisson process

Let Y_1, $\cdots$, Y_n be i.i.d. random variables. Given a Poisson process, define

$$S(t) = Y_1 + \cdots + Y_{N(t)}.$$

The process $\{S(t) : t \geqslant 0\}$ is called a **compound Poisson process**.

Proposition 2.4. *Let $N(t)$ be a Poisson process of rate λ, and $S(t)$ be as above. Then*

- $\mathbb{E}[S(t)] = \mathbb{E}[N(t)]\mathbb{E}[Y_i] = (\lambda t)\mathbb{E}[Y_i]$.
- $\text{Var}(S(t)) = (\lambda t)\mathbb{E}[Y_i^2]$.

The first item is often called **Wald's identity**.

2.4 Superposition, thinning, and conditioning

Proposition 2.5 (Superposition)**.** *If $N_1(t)$ and $N_2(t)$ are independent Poisson processes with respective rates λ_1 and λ_2, then $N(t) := N_1(t) + N_2(t)$ is a Poisson process with rate $(\lambda_1 + \lambda_2)$.*

Proposition 2.6 (Thinning)**.** *Let $N(t)$ be a simple Poisson process with rate λ, and let us embellish $N(t)$ by iid discrete random variables Y_i to form a compound Poisson process $S(t)$. Let $N_k(t)$ be the number of $i \leqslant N(t)$ for which $Y_m = k$. Then $N_k(t)$ is a simple Poisson process with rate $\lambda\mathbb{P}[Y_i = k]$. Moreover, the different $N_k(t)$ are all independent Poisson processes.*

Theorem 2.7 (Conditioning)**.** *Let $T_1, T_2, \cdots$ be the times of jumps of a simple Poisson process $N(t)$. Then the distribution of $(T_1, \cdots, T_n)$ given $N(t) = n$ is the same as the distribution of $(V_1, \cdots, V_n)$, the order statistics of $(U_1, \cdots, U_n)$ where the U_i are iid uniform from $[0, t]$.*

Corollary 2.8. *Let $N(t)$ be a simple Poisson process. For any $m \leqslant n$ and $s \leqslant t$,*

$$\mathbb{P}[N(s) = m | N(t) = n] = \binom{n}{m} \left(\frac{s}{t}\right)^m \left(1 - \frac{s}{t}\right)^{n-m}.$$

[End lecture Th 3/12]

3 Renewal processes

3.1 Definition and properties

3.2 Applications to queueing theory

3.3 Age and residual life

4 Continuous-time Markov chains

4.1 Definition and properties

4.2 Constructing a continuous-time chain from a discrete-time chain

4.3 Computing transition probabilities

4.3.1 The backward & forward Kolmogorov equation

4.4 Stationary distributions & limit behavior

4.5 Markovian chains

5 Martingales

5.1 Conditional expectations

Below X and Y are arbitrary random variables, unless otherwise stated.

Fact 1. If g is any function, then $\mathbb{E}[g(X)Y|X] = g(X)\mathbb{E}[Y|X]$.

Fact 2. If X and Y are independent, then $\mathbb{E}[X|Y] = \mathbb{E}[X]$.

Fact 3. $\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X]$.

5.2 Definition and examples of martingales

Definition 5.1. A sequence of random variables $\{M_n\}_{n=0}^{\infty}$ is called a **martingale** (resp. supermartingale, submartingale) if for all $n \geqslant 0$,

$$\mathbb{E}[M_{n+1}|M_n, M_{n-1}, \cdots, M_0] = M_n, \tag{15}$$

(resp., if the equality is replaced by $\leqslant$ and $\geqslant$).

Martingales originate from the game of betting. In that context, M_n stands for the amount of money you have at time n. To say that M_n is a martingale means that the expected amount of money you will hold next, conditional upon your betting history, is equal to the money you have now. In other words the game is 'fair.' If M_n is a supermartingale, then the game is unfavorable to you, and you will lose money in the long run.

Here are the examples of martingales discussed in lecture Tu 4/15.

Example 5.1. Let $S_n = X_1 + \cdots X_n$, where the X_i are i.i.d. with mean zero ($\mathbb{E}[X_1] = 0$). Then S_n is a martingale. (Slight generalization: If $\mu = \mathbb{E}[X_1]$ is not necessarily 0, then we get that $S_n - n\mu$ is a martingale. This should make intuitive sense since we're always 'recentering' the sum to have mean 0.)

Example 5.2. Same setup as Example 5.1, except that in addition we put $c = \mathbb{E}[(X_1)^2]$. Then $(S_n)^2 - cn$ is a martingale.

Example 5.3. Same setup as Example 5.1, except that we do not require X_1 to have mean zero. Fix a number a, and put $\varphi(a) = \mathbb{E}[e^{aX_1}]$. Then $e^{aS_n}/[\varphi(a)]^n$ is a martingale.

Example 5.4 (Galton-Watson branching process). Let Z_n be the number of individuals in generation n, and let μ be the mean offspring number. It should be clear that

$$\mathbb{E}[Z_{n+1}|Z_n, \cdots, Z_0] = \mu Z_n.$$

Use this to deduce that Z_n/μ^n is a martingale.

Example 5.5 (Stock prices). Let $S_n = S_0 X_1 X_2 \cdots X_n$, where the X_i are i.i.d. Think S_n as the stock price on Day n, and X_n is the multiplier of the stock's price from Day $(n-1)$ to Day n. Then S_n is a martingale (resp. supermartingale, submartingale) if $\mathbb{E}[X_1] = 1$ (resp. $\leqslant 1$, $\geqslant 1$).

Example 5.6 (Harmonic functions). Let $(X_n)_n$ be a discrete-time Markov chain on S with transition probability $p(\cdot,\cdot)$, and h be a function on S such that

$$h(x) = \sum_{y \in S} p(x,y)h(y) \quad \text{for all } x \in S.$$

[Recall where you have seen this before.] Then $M_n = h(X_n)$ is a martingale.

5.3 Optional stopping

Applying Fact 3 to equation (15), we find that for any martingale M_n,

$$\mathbb{E}[M_{n+1}] = \mathbb{E}[\mathbb{E}[M_{n+1}|M_n, M_{n-1}, \cdots, M_0]] = \mathbb{E}[M_n], \tag{16}$$

or

$$\mathbb{E}[M_n] = \mathbb{E}[M_0] \tag{17}$$

for any n. This should not be surprising; after all the game is 'fair.' Similarly, if M_n is a supermartingale (resp. a submartingale), $\mathbb{E}[M_{n+1}] \leqslant \mathbb{E}[M_n]$ (resp. $\mathbb{E}[M_{n+1}] \geqslant \mathbb{E}[M_n]$).

Recall the notion of a **stopping time**: we say that T is a stopping time with respect to $(M_n)_{n=0}^{\infty}$ if for every n, the (non)occurrence of the event $\{T = n\}$ can be written as a union of events of the form $\{M_n = x_n, M_{n-1} = x_{n-1}, \cdots, M_0 = x_0\}$. It is then not difficult to see that the events $\{T > n\}$ and $\{T \leqslant n\}$ can also be written in such fashion.

The optional stopping theorem says that (17) still holds if we replace the deterministic time n by a stopping time.

Theorem 5.1 (The optional stopping theorem)**.** *Let M_n be a martingale, and T be a stopping time with respect to M_n. Then $\mathbb{E}[M_{T\wedge n}] = \mathbb{E}[M_0]$ for all n.*

(Recall that $a \wedge b = \min(a, b)$.) There are also corresponding theorems for super- and sub-martingales.

Proof. We want to show that $\mathbb{E}[M_{T\wedge(n+1)} - M_{T\wedge n}] = 0$. Let's break this down into two cases, $T \leqslant n$ and $T > n$. If $T \leqslant n$, then $M_{T\wedge n} = M_T$ and $M_{T\wedge(n+1)} = M_T$, so their difference is zero. Thus it suffices to show that

$$\mathbb{E}[M_{T\wedge(n+1)} - M_{T\wedge n}; T > n] = 0.$$

This is the same as

$$\mathbb{E}[M_{n+1} - M_n; T > n] = 0.$$

This holds because we know that

$$\mathbb{E}[M_{n+1} - M_n; M_n = x_n, M_{n-1} = x_{n-1}, \cdots, M_0 = x_0] = 0$$

for every choice of $x_0, \cdots, x_n$, and that $\{T > n\}$ is a union of events of the form $\{M_n = x_n, \cdots, M_0 = x_0\}$. □

Remark (May be skipped). The reason why we used $T \wedge n$ instead of T in the statement of Theorem 5.1 is because 'bad' things may happen if the stopping time T is infinite.

Let's turn to some applications of optional stopping.

Example 5.7 (Gambling). As we mentioned, the amount of money held by a gambler playing a fair casino game is a martingale. Optional stopping says that at any (finite) time when you decide to stop the game, the expected money you have at the end is the same as the expected money you have at the beginning. In this sense there is no successful betting strategy. Can you generalize this interpretation to supermartingales and submartingales?

Example 5.8. Consider the sum of i.i.d. r.v.'s, $S_n = X_1 + \cdots + X_n$. We shall assume in the following that with probability 1, the stopping time T is finite, so $T \wedge n \underset{n\to\infty}{\longrightarrow} T$.

- As discussed in Example 5.1, $S_n - n\mu$ is a martingale, so applying optional stopping to this we get $\mathbb{E}[S_T - T\mu] = 0$, or $\mathbb{E}[S_T] = T\mu = T\mathbb{E}[X_1]$. If you interpret T as a 'random number of terms in the sum,' then this is nothing but Wald's identity (or Wald's 1st equation) first encountered in compound Poisson processes.

- Under the further assumption that $\mu = 0$, Example 5.2 says that $(S_n)^2 - n\mathbb{E}[(X_1)^2]$ is a martingale. So applying optional stopping we find $\mathbb{E}[(S_T)^2] = \mathbb{E}[T]\mathbb{E}[(X_1)^2]$, also known as Wald's 2nd equation. (This was also alluded to during the discussions on compound Poisson process, *viz.* the variance thereof.)

For a special but very useful example, consider symmetric simple random walk, where $\mathbb{P}[X_1 = +1] = \mathbb{P}[X_1 = -1] = \frac{1}{2}$. Let T be the first time that S_n is either $\geqslant b$ or $\leqslant -a$. Then by the first bullet point,

$$0 = \mathbb{E}[S_T] = b\mathbb{P}[S_T = b] + (-a)\mathbb{P}[S_T = -a]$$

Since $\mathbb{P}[S_T = b] = 1 - \mathbb{P}[S_T = -a]$, we deduce that

$$\mathbb{P}[S_T = b] = \frac{a}{a+b}, \quad \mathbb{P}[S_T = -a] = \frac{b}{a+b}.$$

On the other hand, by the second bullet point, $\mathbb{E}[T] = \mathbb{E}[(S_T)^2]$. Meanwhile

$$\mathbb{E}[(S_T)^2] = b^2\mathbb{P}[S_T = b] + (-a)^2\mathbb{P}[S_T = a] = ab.$$

So the expected first time to hit either b or $-a$ is ab, a result which we got from the gambler's ruin calculation earlier.

....

A martingale $(M_t)_{t\geqslant 0}$ in continuous time is one which satisfies

$$\mathbb{E}[M_t | M_r, r \leqslant s] = M_s.$$

There is an analogous optional stopping theorem for continuous-time martingales, which will be stated in the next section.

6 Brownian motion

Definition 6.1. A **standard Brownian motion**[18](also known as a **Wiener process**[19]) $\{B_t : t \geqslant 0\}$ is a $\mathbb{R}$-valued stochastic process satisfying

(BM1) $B_0 = 0$.

(BM2) For all $t > s > 0$, $B_t - B_s$ is independent of $\{B_r : r \leqslant s\}$. [Independent increments.]

(BM3) For all $t > s > 0$, $B_t - B_s$ is a normal random variable with mean 0 and variance $(t - s)$. [Stationary increments.]

(BM4) The function $t \mapsto B_t$ is continuous.

A sample Brownian motion is shown in Figure 1. In contrast to a pure jump process like the Poisson process, a Brownian motion has continuous path and no jumps. However the path is *nowhere differentiable* (it is a fractal). Do not ever ask a probabilist what (dB_t/dt) is! It does not exist![20]

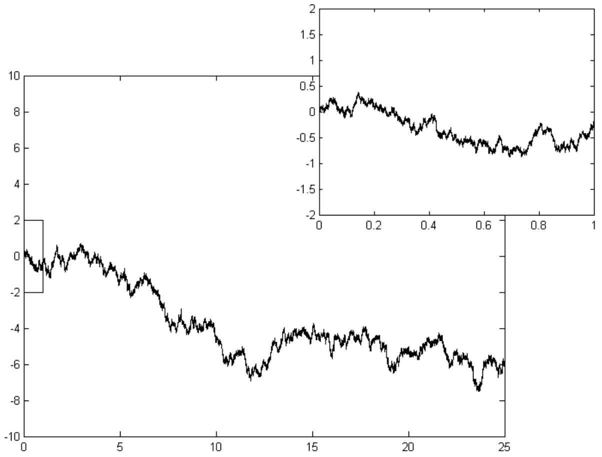


Figure 1: A realization of a Brownian motion. Courtesy of Wikipedia.

Here are two straightforward generalizations of Definition 6.1.

- A (standard) Brownian motion started from point x (instead of 0) is $x + B_t$.
- A standard d-dimensional Brownian motion is a $\mathbb{R}^d$-valued stochastic process $(B_t^1, B_t^2, \cdots, B_t^d)$, where each B_t^i $(i = 1, 2, \cdots, d)$ is an independent one-dimensional standard Brownian motion.

We will not discuss at all the formal construction of Brownian motion[21]. Instead let us focus on the important features of Brownian motion by appealing to Definition 6.1.

The (strong) Markov property. This follows from (BM2). In fact, one can make a stronger statement that $\{B_{u+t} - B_u : t \geqslant 0\}$ is another Brownian motion, independent of $\{B_r : r \leqslant u\}$. This is the Markov property. The strong Markov property says that for any stopping time T, $\{B_{T+t} - B_T : t \geqslant 0\}$ is itself a Brownian motion, independent of $\{B_r : r \leqslant T\}$.

[18] named after the English botanist Robert Brown.

[19] named after the American mathematician Norbert Wiener.

[20] However dB_t shows up frequently in stochastic differential equations.

[21] which is way beyond the level of this course; take graduate-level probability to learn about it.

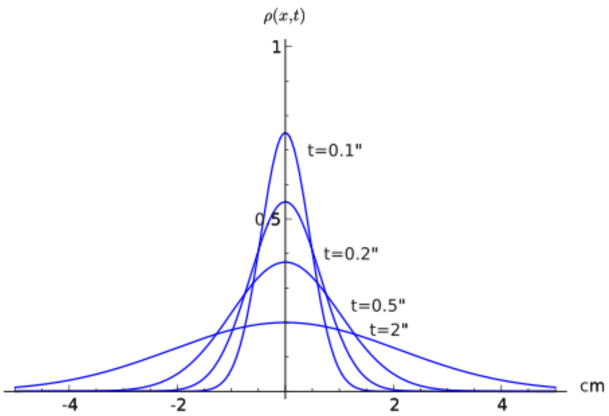


Figure 2: The heat kernel $p_t(0,x) = \frac{1}{\sqrt{2\pi t}} e^{-x^2/2t}$ for several values of t. Courtesy of Wikipedia.

Distribution of B_t. From (BM3) we deduce that for each $t > 0$, B_t is distributed as a normal distribution with mean 0 and variance t. In other words, the density function of B_t reads

$$f_{B_t}(x) = \frac{1}{\sqrt{2\pi t}} e^{-x^2/2t}, \quad x \in \mathbb{R}.$$

Since by (BM1) the process starts from the origin ($B_0 = 0$), you can interpret the above density as the *transition probability density* from point 0 to x over a time period t, denoted by $p_t(0,x)$. See Figure 2. Furthermore, using the translational invariance of this process in space (a B.M. started at x is simply $x + B_t$), one can show that the transition probability density from x to y over time t is

$$p_t(x,y) = \frac{1}{\sqrt{2\pi t}} e^{-(x-y)^2/2t}, \quad x, y \in \mathbb{R}.$$

This is known as the **heat kernel** associated with (one-dimensional) Brownian motion. The name 'heat' comes from the fact that $p_t(x,y)$ is the fundamental solution of the one-dimensional **heat equation**

$$\begin{cases} \partial_t u(x,t) = \frac{1}{2}\partial_{xx} u(x,t) & \text{for all } x \in \mathbb{R},\ t > 0, \\ u(x,0) = \delta_y(x) & \text{for all } x \in \mathbb{R}, \end{cases}$$

where δ_y is the 'δ-function' at y. This may seem familiar if you have had a course in partial differential equations. The point I want to make here is that once again, we are connecting probability (stochastic processes) with analysis (DiffEq)!

[JPC: Add an optional section on finding the 'Kolmogorov equation' for Brownian motion, just to demonstrate that the 'Q-matrix' is the Laplacian $\Delta = \frac{1}{2}\partial_{xx}$ on $\mathbb{R}$.]

Covariance of B_t and B_s. If $t > s$ then (BM3) says that $\text{Var}(B_t - B_s) = (t - s)$. Meanwhile by definition

$$\text{Var}(B_t - B_s) = \text{Var}(B_t) - 2\text{Cov}(B_t, B_s) + \text{Var}(B_s) = t - 2\text{Cov}(B_t, B_s) + s.$$

So $\text{Cov}(B_t, B_s) = s$. If $s > t$, reverse the role of s and t in the preceding argument to get $\text{Cov}(B_t, B_s) = t$. To summarize, $\text{Cov}(B_t, B_s) = t \wedge s$.

Scaling. We claim that for any number $c > 0$, cB_{t/c^2} is another Brownian motion. [Mnemonic: If scale the size of the B.M. by c, then also speed up the motion by c^2 to recover the 'same' B.M.] To see this we check the definition. (BM1), (BM2), and (BM4) are more or less clear, so the nontrivial item left to check is (BM3). If we put $t' = t/c^2$ and $s' = s/c^2$, then

$$\left[cB_{t/c^2} - cB_{s/c^2}\right] \overset{d}{=} c\left[B_{t'} - B_{s'}\right] \overset{d}{=} c\mathcal{N}\left(0, t' - s'\right) \overset{d}{=} c\mathcal{N}\left(0, \frac{t-s}{c^2}\right) \overset{d}{=} \mathcal{N}(0, t - s).$$

(In succession we used simple substitution; (BM3) applied to $(B_{t'} - B_{s'})$; simple substitution; and the scaling property of normal distribution $\mathcal{N}(0, \sigma^2) \overset{d}{=} \sigma\mathcal{N}(0,1)$. The notation '$\overset{d}{=}$' means 'equal in distribution/law.')

Martingale. B_t is a continuous-time martingale. To see this we verify the definition:

$$\mathbb{E}[B_t | B_r, r \leqslant s] = \mathbb{E}[(B_t - B_s) | B_r, r \leqslant s] + \mathbb{E}[B_s | B_r, r \leqslant s] = \cancel{\mathbb{E}[B_t - B_s]} + B_s = B_s.$$

The optional stopping theorem in this context says that if T is a stopping time with $\mathbb{E}[T] < \infty$, then $\mathbb{E}[B_T] = \mathbb{E}[B_0] = 0$.

Let T be the first time B_t hits either b or $-a$. Then using exactly the same reasoning as the symmetric random walk example in the last section, we have

$$\mathbb{P}[B_T = b] = \frac{a}{a+b}, \quad \mathbb{P}[B_T = -a] = \frac{b}{a+b}.$$

Moreover $(B_t)^2 - t$ is a martingale (compare this with the discrete-time analog $(S_n)^2 - n$). The same reasoning from the discrete case carries over to yield $\mathbb{E}[T] = ab$.

You may ask why we are able to borrow the reasoning from symmetric simple random walk and apply it to Brownian motion. Is there a deeper connection between the two? Intuitively there should be: if we shrink down the size of each random walk step to something very tiny, and at the same time adjust the proper time unit, then the random walk path should look close to being continuous, and in fact would approximate a Brownian motion. Precisely we have the following

Theorem 6.1 (Invariance principle for random walks)**.** *Let $(S_n)_{n=0}^{\infty}$ be a symmetric simple random walk on $\mathbb{Z}$. Then for each $t > 0$,*

$$\frac{S_{\lfloor nt \rfloor}}{\sqrt{n}} \xrightarrow[n\to\infty]{d} B_t,$$

where $\lfloor x \rfloor$ means the integer part of x, and $\xrightarrow{d}$ means convergence in distribution.

Once again notice how the scaling works: if you scale the size of the walk by $\sqrt{n}$, then time has to be sped up by n.

An almost complete (but really incomplete) proof. Recall the central limit theorem from MATH 3160: if $S_n = X_1 + \cdots + X_n$ where the X_i are i.i.d. with finite second moment, then

$$\frac{S_n - n\mu}{\sqrt{n\sigma^2}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1).$$

To apply the CLT in our current context, replace n by $\lfloor nt \rfloor$ (defining the integer part is just a technicality), μ by 0, and σ^2 by 1. We thus have

$$\frac{S_{\lfloor nt \rfloor}}{\sqrt{\lfloor nt \rfloor}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1),$$

or

$$\frac{S_{\lfloor nt \rfloor}}{\sqrt{n}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,t).$$

This should make you believe that (BM3) holds, at least when $s = 0$. A suitable modification gives

$$\frac{S_{\lfloor nt \rfloor} - S_{\lfloor ns \rfloor}}{\sqrt{n}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0, t-s),$$

which would validate (BM3) for all $t > s > 0$. (BM1) and (BM2) are clear. So what remains to check is (BM4), the continuity of the limit process in time t, which is quite crucial. Unfortunately the verification would require a different type of CLT which is beyond the scope of our course[22], which is why I said this is "an almost complete (but really incomplete) proof." □

[22] If you're curious look up 'Donsker's invariance principle.'

Finally we shall use the fact that $M(t) = e^{\theta B_t - \theta^2 t/2}$ is a martingale. (Will not be proved in class. This will be one of the homework problems.) This $M(t)$ is a **geometric Brownian motion** and will be used in financial math.

7 Mathematical finance

Plan: Do the binomial model in-depth. Why ‘there is no free lunch’ (absence of arbitrage is essential for an efficient market). Fundamental theorem of finance. Derive the analog of the B-S equation in the binomial model. Time remaining, discuss the continuous-time model (the actual B-S).

An excellent readable reference: S. E. Shreve, *Stochastic Calculus for Finance, Vol. I: The Binomial Asset Pricing Model.* Springer-Verlag, 2004. Some of the notes below are drawn from the first 2 chapters of the text.

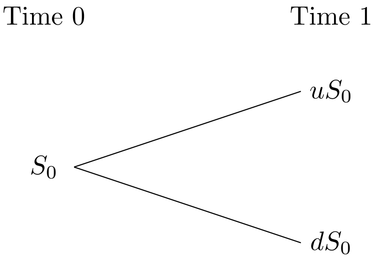
7.1 One-step binomial model

7.1.1 The setting

There is a certain company stock whose price is S_0 at time 0. At time 1, the stock price takes value

$$S_1 = \begin{cases} uS_0, & \text{with probability } p, \\ dS_0, & \text{with probability } (1-p). \end{cases}$$

See the tree diagram below. Without loss of generality we shall assume $0 < d < 1 < u$ and $0 < p < 1$.



An investor, holding an amount of cash (or **wealth**) W_0, can choose to buy Δ_0 shares of the stock, then invest the remaining money in a risk-free savings account, which has an interest rate of r per unit time. Under this scheme, his wealth at time 1 is

$$W_1 = \Delta_0 S_1 + (1+r)(W_0 - \Delta_0 S_0). \tag{18}$$

Note that since S_1 is a random variable, so is W_1.

Two things of note: First, $(W_0 - \Delta_0 S_0)$ is allowed to be negative, which means that the investor gets a loan from the bank to purchase the Δ_0 shares of stock. As a result he carries a debt, which is being assessed an interest rate r. We shall assume throughout that the savings rate and the borrowing rate are the same r. Second, Δ_0 is allowed to be negative, that is, he sells the stocks *short*.

An important concept in financial markets is that of *arbitrage*. Roughly speaking, for an efficient market to run, any trading strategy which grows money must also run the risk of losing money.

Definition 7.1 (Arbitrage)**.** A trading strategy is called an *arbitrage* if it starts with no money, has zero probability of losing money, and has a positive probability of making money.

[Why arbitrage free is essential to analysis.]

Consider the scenario described above. Suppose $(1+r) > u$. Then an investor can come in with no money, sells the stock short at time 0, and invest his earnings into a savings account. Since he is guaranteed to make more money from the savings than from the stock, this constitutes an arbitrage. Mathematically we have (note $\Delta_0 < 0$)

$$W_1 > -|\Delta_0|S_1 + u(0 + |\Delta_0|S_0) \geqslant -|\Delta_0|(uS_0) + |\Delta_0|(uS_0) = 0,$$

so wealth is generated. Similarly, if $d > (1+r)$, then the investor should buy in as many stocks as possible. Then (note $\Delta_0 > 0$)

$$W_1 > \Delta_0 S_1 + d(0 - \Delta_0 S_0) \geqslant \Delta_0(dS_0) - \Delta_0(dS_0) = 0,$$

again generating wealth. Thus we have found the following condition for an arbitrage-free trading:

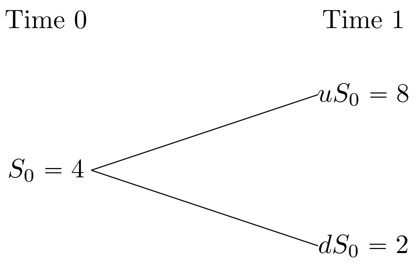
Proposition 7.1. *In the binomial model (sans options), one avoids arbitrage if and only if* $0 < d < 1+r < u$.

7.1.2 Call options with expiry 1

Now suppose the investor is given one unit of a stock **option**. We first consider the **European call option**, where he is allowed to, though not obligated to, exercise the option at time 1. Accompanied with the option is a **strike price**, denoted by K. If the stock price at time 1, S_1, exceeds K, then he would choose to exercise the option and receive the profit, which is the difference between S_1 and K. However, if S_1 falls short of K, there is no point of exercising the option, so the option is worthless. Thus the **value** of the option, denoted by V_1, is given by $(S_1 - K)^+$, where $x^+ = \max(x, 0)$. Other options will be discussed in the next section.

The relevant question is: at time 0, how much is the option worth? In other words, what is a fair value of V_0? To answer this question, we will try to **replicate** the option using a combination of stocks and cash. It turns out this is doable, and in fact can be done for any option other than the European option. We say that the binomial model is **complete** (more on this notion in the next section).

To be concrete we start with some numbers. Suppose $S_0 = 4$, $u = 2$, $d = 1/2$, and $r = 0.25$. The European option expires at time 1 with a strike price $K = 5$. How can we replicate this option using (W_0, Δ_0), the initial wealth and the amount of stocks bought?



It turns out that the right combination is $(W_0, \Delta_0) = (1.2, 1/2)$. Let's verify this via a concrete calculation. Remember that the wealth at time 1 is given by (18). Let's denote by $W_1(u)$ and $W_1(d)$, respectively, the value of W_1 when the stock goes up or down. Then

$$\begin{aligned} W_1(u) &= \Delta_0(uS_0) + (1+r)(W_0 - \Delta_0 S_0); \\ W_1(d) &= \Delta_0(dS_0) + (1+r)(W_0 - \Delta_0 S_0). \end{aligned}$$

Plugging in the numbers for the first, we find

$$W_1(u) = \frac{1}{2} \cdot 8 + (1+0.25)(1.2 - \frac{1}{2} \cdot 4) = 4 - 1 = 3.$$

Similarly

$$W_1(d) = \frac{1}{2} \cdot 2 + (1+0.25)(1.2 - \frac{1}{2} \cdot 4) = 1 - 1 = 0.$$

These agree with the values of the given option, 3 when up and 0 when down.

What's so special about this particular value of (W_0, Δ_0)? **If the option price V_0 deviates any bit from W_0, we can trade the Δ_0 shares of stock in a way to make a riskless profit, whence arbitrage.** To see this let's adopt the point of view of the option trader. Suppose the option sells for 1.21. Then he can sell the option to someone else, use the 1.20 to replicate the option, and invest the remaining

0.01 into the savings bank. At time 1 his wealth will be 0.0125 plus whatever the outcome of the stock option. Since the seller needs no money initially, the creation of wealth from the savings leads to arbitrage.

On the other hand suppose the option sells for 1.19. Then the option trader should do the opposite by buying one option. Sell short 1/2 share of the stock at time 0, and use the money earned (2) to buy one unit of option at 1.19. As for the remaining money, invest 0.80 into savings account A and 0.01 into savings account B. If the stock goes up at time 1, then his wealth would be 4 were he to hold onto the $^1\!/_2$ share of stock. That's OK: the option generates 3, while the money in savings account A increased to 1 thanks to interest, so these made up for the 4—plus the extra 0.0125 from savings account B. Similarly, if the stock goes down at time 1, then his wealth would be 1 were he to hold onto the stock. While the option generates nothing, the money in savings account A exactly makes up for the 1, and something extra comes from savings account B. To conclude, the trader will receive an extra 0.0125 no matter the outcome of the stock price at time 1. Since he didn't need the money initially, we have an arbitrage.

Now that you get the hang of the idea, let's discuss how exactly to find the **arbitrage-free value** of any option in general.

Given: S_0, u, d, r, K, W_1. *Unknown variables to be solved:* W_0, Δ_0. Remember that we want to set $V_0 = W_0$ to get the fair value for the option at time 0.

We have the following system of two equations

$$\begin{aligned} W_1(u) &= \Delta_0(uS_0) + (1+r)(W_0 - \Delta_0 S_0); && (19)\\ W_1(d) &= \Delta_0(dS_0) + (1+r)(W_0 - \Delta_0 S_0). && (20)\end{aligned}$$

Here $W_1(u)$ and $W_1(d)$ can be the value of *any* option at time 1, depending on whether the stock is up (u) or down (d). For concreteness, you can take the European call option, $W_1 = (S_1 - K)^+$.

Solving for (W_0, Δ_0) is now a matter of algebra. To make things easier we introduce the following parameters:

$$\bar{p} = \frac{(1+r)-d}{u-d}, \quad \bar{q} = \frac{u-(1+r)}{u-d}.$$

Note that $\bar{p}, \bar{q} > 0$ and $\bar{p}+\bar{q} = 1$, so one may be tempted to interpret this as a probability. However (and this is a crucial subtlety) **so far we have not used any probability!** In particular, $\bar{p}$ has nothing to do with p, the original probability that the stock moves up at time 1. So for now just think of $\bar{p}$ and $\bar{q}$ as convenient parameters. Their roles will be clarified shortly.

Back to solving the equations. Let's multiply (19) by $\bar{p}$ and (20) by $\bar{q}$, then add the two together:

$$\bar{p}W_1(u) + \bar{q}W_1(d) = (\bar{p}u + \bar{q}d)\Delta_0 S_0 + (1+r)(W_0 - \Delta_0 S_0) = (1+r)W_0,$$

the last equality coming from some nifty (!) algebra, as you should verify on your own. In other words we find

$$W_0 = \frac{1}{1+r}\left[\bar{p}W_1(u) + \bar{q}W_1(d)\right]. \qquad (21)$$

We then solve for Δ_0 to find

$$\Delta_0 = \frac{W_1(u) - W_1(d)}{uS_0 - dS_0}. \qquad (22)$$

These are deceptively simple-looking formulae for W_0 and Δ_0 which you can use on the HW/exam. But let's pause for a moment and interpret what they mean. Let $\bar{\mathbb{P}}$ be the probability which charges the event 'stock goes up' (u) with probability $\bar{p}$, and the event 'stock goes down' (d) with probability $\bar{q}$. Further let $\bar{\mathbb{E}}$ denote the corresponding expectation. Since we're going to match V_0 with W_0 and V_1 with W_1 (otherwise there is arbitrage), (21) boils down to

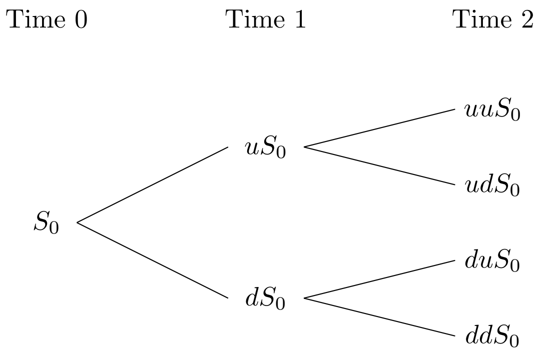
$$V_0 = \frac{1}{1+r}\bar{\mathbb{E}}[V_1].$$

In other words, under $\bar{\mathbb{P}}$, the expected value of the option stays the same upon adjusting for inflation (r can be seen as the rate of inflation per unit time). Therefore $\bar{\mathbb{P}}$ can be viewed as the **risk-neutral probability** for investing under the said option. Again, we emphasize that $\bar{\mathbb{P}}$ has nothing to do with the original probability (p, q). But what this says is that for the given option, there exists a risk-neutral probability under which the inflation-adjusted expected option value 'stays constant' (read: $\{(1+r)^{-n}V_n\}_n$ forms a martingale; see next section). Such a result goes under the heading of the **first fundamental theorem of finance** (or asset pricing, or arbitrage).

7.2 Multi-step binomial model

7.2.1 Call options with expiry 2

Let's now turn to the binomial model with 2 time steps. The stock price evolves according to the following tree diagram.



Suppose we are given one unit of stock option with expiry 2 and value $V(= V_2)$. What is a fair price V_0 for the option at time 0?

Once again we will attempt to replicate the option using a combination of stocks and cash at *both* time 0 *and* time 1. Starting with wealth W_0 at time 0, we buy (a possibly negative) number Δ_0 of stocks, so that at time 1 our wealth is

$$W_1 = \Delta_0 S_1 + (1+r)(W_0 - \Delta_0 S_0).$$

This boils down to two equations

$$W_1(u) = \Delta_0(uS_0) + (1+r)(W_0 - \Delta_0 S_0); \tag{23}$$
$$W_1(d) = \Delta_0(dS_0) + (1+r)(W_0 - \Delta_0 S_0), \tag{24}$$

depending on whether the stock goes up (u) or down (d). To complete the replication process, we want to hold Δ_1 stocks at time 1 **depending on the stock price at that time**, so the wealth W_2 at time 2 matches V_2. Nominally this means $V_2 = W_2$ where

$$W_2 = \Delta_1 S_2 + (1+r)(W_1 - \Delta_1 S_1),$$

but NOTE that both W_1 and Δ_1 ought to be **adapted** to the value of S_1. This boils down to four equations, depending on the outcome of the stock (uu, ud, du, dd):

$$W_2(uu) = \Delta_1(u)\cdot(uuS_0) + (1+r)(W_1(u) - \Delta_1(u)\cdot(uS_0)); \tag{25}$$
$$W_2(ud) = \Delta_1(u)\cdot(udS_0) + (1+r)(W_1(u) - \Delta_1(u)\cdot(uS_0)); \tag{26}$$
$$W_2(du) = \Delta_1(d)\cdot(duS_0) + (1+r)(W_1(d) - \Delta_1(d)\cdot(dS_0)); \tag{27}$$
$$W_2(dd) = \Delta_1(d)\cdot(ddS_0) + (1+r)(W_1(d) - \Delta_1(d)\cdot(dS_0)). \tag{28}$$

Equations (23) through (28) form a system of six linear equations with six unknowns

$$(W_0, W_1(u), W_1(d), \Delta_0, \Delta_1(u), \Delta_1(d)),$$

so a unique solution exists. We will not go through the calculational details, but the key idea is to adapt the arguments from the one-step model and infer the solutions recursively from W_2 to W_1 to W_0. To begin we have

$$W_1(u) = \frac{1}{1+r}\left[\bar{p}W_2(uu) + \bar{q}W_2(ud)\right] \quad \text{and} \quad W_1(d) = \frac{1}{1+r}\left[\bar{p}W_2(du) + \bar{q}W_2(dd)\right],$$

therefore

$$W_0 = \frac{1}{1+r}\left[\bar{p}W_1(u) + \bar{q}W_1(d)\right] = \frac{1}{(1+r)^2}\left[\bar{p}^2W_2(uu) + \bar{p}\bar{q}W_2(ud) + \bar{q}\bar{p}W_2(du) + \bar{q}^2W_2(dd)\right].$$

Thus $V_0 = W_0$ is the fair price of the option. If V_0 deviates any bit from W_0 then arbitrage is created. Similarly, V_1 must match W_1 to avoid arbitrage. Symbolically we can write the above as

$$V_1 = \frac{1}{1+r}\bar{\mathbb{E}}[V_2|S_1] \quad \text{and} \quad V_0 = \frac{1}{1+r}\bar{\mathbb{E}}[V_1] = \frac{1}{(1+r)^2}\bar{\mathbb{E}}[\bar{\mathbb{E}}[V_2|S_1]] = \frac{1}{(1+r)^2}\bar{\mathbb{E}}[V_2],$$

where in the last equality we used Fact 3 of conditional expectation.

The calculations for Δ_0 and Δ_1 also follow recursively:

$$\Delta_1(u) = \frac{V_2(uu) - V_2(ud)}{uuS_0 - udS_0}, \quad \Delta_1(d) = \frac{V_2(du) - V_2(dd)}{duS_0 - ddS_0}, \quad \Delta_0 = \frac{V_1(u) - V_1(d)}{uS_0 - dS_0}.$$

For a concrete example with numbers, consider the European call option with expiry 2, that is, $V = (S_2 - K)^+$. We use the parameters

$$S_0 = 4, \ u = 2, \ d = 1/2, \ r = 0.25, \ K = 3.$$

Then

$$\bar{p} = \frac{(1+r)-d}{u-d} = \frac{1.25 - 0.5}{2 - 0.5} = 0.5, \quad \bar{q} = \frac{u-(1+r)}{u-d} = \frac{2 - 1.25}{2 - 0.5} = 0.5.$$

We tabulate the 4 possible scenarios for the stock price at time 2, as well as the value of the option, below. Note that to compute the fair price V_0 we need to use the risk-neutral probability $\bar{\mathbb{P}}$.

ω	S_1	S_2	$V = (S_2 - K)^+$	Prob.
uu	8	16	13	$\bar{p}^2 = (1/2)^2$
ud	8	4	1	$\bar{p}\bar{q} = (1/2)^2$
du	2	4	1	$\bar{q}\bar{p} = (1/2)^2$
dd	2	1	0	$\bar{q}^2 = (1/2)^2$

So

$$V_0 = \frac{1}{(1+r)^2}\bar{\mathbb{E}}[V] = \frac{1}{(1+0.25)^2}\left(13\cdot(1/2)^2 + 1\cdot(1/2)^2 + 1\cdot(1/2)^2 + 0\cdot(1/2)^2\right) = 2.4.$$

On the homework assignment you will get to work through a three-step binomial model. An example similar to the homework problem will appear on the final exam.

7.2.2 General analysis

The purpose of this subsection is to set everything up systematically, so that you see why the binomial options pricing model is tractable (even though in general it does not admit closed-form solutions).

The question. Consider the binomial options pricing model (with parameters S_0, u, d, r) where the option (need not be an European call option) expires at time n. Let $V = V_n$ be the value of the option (at time n). What is the arbitrage-free price V_0 of this option at time 0?

Short answer. $V_0 = (1+r)^{-n}\bar{\mathbb{E}}[V]$, where $\bar{\mathbb{E}}$ is the expectation corresponding to the risk-neutral measure $\bar{\mathbb{P}}$.

While this result can be readily used on the HW/exam, it is instructive to go through the derivation once just to see all the mechanics involved. Believe it or not, you are about to see a discretized version of an Itô integral used in stochastic calculus (even though I am not supposed to tell you what it is).

To carry out the general analysis, we will ALWAYS make the following assumptions, which are in general NOT satisfied by real financial markets.

Assumption 1. Throughout the derivation of the options pricing, we assume that:

(a) There is unlimited short selling of stock.

(b) There is unlimited borrowing of cash from the bank.

(c) There are no transaction costs/fees whatsoever.

(d) Our transactions take place on a small enough scale that they will not affect the entire market.

We also bring up some assumptions alluded previously.

Assumption 2. We also assume that:

(a) Shares of stocks can be subdivided for sale or purchase.

(b) The interest rate for saving/investing is equal to the interest rate for borrowing (denoted by r).

(c) The purchase price of a stock is the same as the selling price ('no bid-ask spread').

We can now describe the stochastic processes involved.

- $(S_0, S_1, \cdots, S_n)$ represents the sequence of *stock prices* from time 0 to time n. In the binomial model, S_{k+1} takes only two possible values, uS_k or dS_k.
- $(\Delta_0, \Delta_1, \cdots, \Delta_n)$ stands for the amount of *stock* held by the investor. In financial math terminology this sequence of random variables is called the **portfolio process**.
- $(W_0, W_1, \cdots, W_n)$ stands for the *wealth* held by the investor. We will refer to this sequence of random variables as the **wealth process**.

Both the portfolio process and the wealth process should be **adapted** to the stock prices, that is, both Δ_k and W_k should be functions of $(S_0, S_1, \cdots, S_k)$ and no more. This makes sense in practice, since the investor makes his stock transactions (and hence his wealth decision) based on the performance of the stock price up to that point. Depending on whether the stock moves up or down, he may choose to readjust his investment strategy.

With this in mind, we now write down an equation linking the wealth process at time $(k+1)$ and time k. It is

$$W_{k+1} = \Delta_k S_{k+1} + (1+r)(W_k - \Delta_k S_k). \tag{29}$$

Recall that we have seen this equation when $k=0$ and $k=1$. At time k, the wealth W_k consists of the value of the stock, $\Delta_k S_k$, as well as the cash value, $(W_k - \Delta_k S_k)$. When one unit time elapses, the stock value changes to $\Delta_k S_{k+1}$, while the cash value increases by rate r. This explains why W_{k+1} is equal to the RHS.

Perhaps to your lack of surprise, this equation possesses a nice recursive structure. Let's verify this in the case $r=0$. Using the identity $W_k = \Delta_{k-1}S_k + (W_{k-1} - \Delta_{k-1}S_{k-1})$, we replace the W_k on the RHS of (29) above to get

$$W_{k+1} = \Delta_k(S_{k+1} - S_k) + \Delta_{k-1}(S_k - S_{k-1}) + W_{k-1} = W_{k-1} + \sum_{j=k-1}^{k} \Delta_j(S_{j+1} - S_j).$$

Hopefully you're sensing a pattern emerging. Continue the recursion to W_0 yields

$$W_{k+1} = W_0 + \sum_{j=0}^{k} \Delta_j (S_{j+1} - S_j).$$

This is in fact a 'baby' discretized version of an Itô integral, the sum being the discrete version of $\int_0^T \Delta_t dS_t$. Note that dS_t is the differential of a stochastic process, and in general it CANNOT be understood as $\frac{dS_t}{dt} dt$, since $\frac{dS_t}{dt}$ doesn't even exist! Making sense of this integral requires some leg work (which you can learn in a course on stochastic calculus), but one important condition is that the portfolio process be **adapted** to the movement of the stock price, as previously mentioned. After all, we can only use the prior history of the stock price, NOT the future, to update our portfolio.

A similar calculation for $r > 0$ gives (details omitted)

$$\widetilde{W}_{k+1} = \widetilde{W}_0 + \sum_{j=0}^{k} \Delta_j (\widetilde{S}_{j+1} - \widetilde{S}_j),$$

where $\widetilde{S}_k = (1+r)^{-k} S_k$ and $\widetilde{W}_k = (1+r)^{-k} W_k$ are, respectively, the discounted (or inflation-adjusted) stock price and wealth processes at time k.

Shortly we shall uncover the underlying **martingale** structure of both the stock and the wealth processes. It turns out that this is key to pinning down a *unique* fair price for any option in the binomial model.

The roadmap for the rest of the subsection consists of four propositions. We will establish them one by one in their logical order, culminating with the final and the most useful result. Keep in mind that Assumptions 1 and 2 are in full force.

Proposition 7.2. *Under the risk-neutral probability $\bar{\mathbb{P}}$, the discounted (or inflation-adjusted) stock price $(1+r)^{-k} S_k$ forms a martingale.*

Proposition 7.3. *Under the risk-neutral probability $\bar{\mathbb{P}}$, the discounted (or inflation-adjusted) wealth process $(1+r)^{-k} W_k$ forms a martingale.*

Propositions 7.2 and 7.3 together form the **(first) fundamental theorem of finance** in the context of the binomial model.

Proposition 7.4. *The binomial asset pricing model is* ***complete****: that is, any option (with expiry n) can be replicated using a fixed constant W_0 and a portfolio process $\{\Delta_0, \Delta_1, \cdots, \Delta_n\}$.*

(*Translation:* Starting from W_0 dollars, we can trade shares of stock to exactly duplicate the value of any option.)

Proposition 7.5. *The value of the option V at time 0 is $V_0 = (1+r)^{-n} \bar{\mathbb{E}}[V]$.*

Proof of Proposition 7.2. We want to show that for each k,

$$\bar{\mathbb{E}}[(1+r)^{-(k+1)} S_{k+1} | S_k, \cdots, S_0] = (1+r)^{-k} S_k.$$

Method 1: We split S_{k+1} into the sum of $(S_{k+1} - S_k)$ and S_k. Then by Fact 1 and Fact 2 of conditional expectation, respectively,

$$\bar{\mathbb{E}}[S_k | S_k, \cdots, S_0] = S_k \quad \text{and} \quad \bar{\mathbb{E}}[(S_{k+1} - S_k) | S_k, \cdots, S_0] = \bar{\mathbb{E}}[S_{k+1} - S_k].$$

But $\bar{\mathbb{E}}[S_{k+1} - S_k] = (u-1) S_k \cdot \bar{p} + (d-1) S_k \cdot \bar{q} =$ (after some nifty algebra) $= r S_k$. Thus

$$\begin{aligned} \bar{\mathbb{E}}[(1+r)^{-(k+1)} S_{k+1} | S_k, \cdots, S_0] &= (1+r)^{-(k+1)} \left(\bar{\mathbb{E}}[S_k | S_k, \cdots, S_0] + \bar{\mathbb{E}}[(S_{k+1} - S_k) | S_k, \cdots, S_0] \right) \\ &= (1+r)^{-(k+1)} (S_k + r S_k) = (1+r)^{-k} S_k, \end{aligned}$$

which is what we want.

Method 2: Alternatively, we may split S_{k+1} into the product of $\frac{S_{k+1}}{S_k}$ and S_k. Since the ratio $\frac{S_{k+1}}{S_k}$ is either u or d, irrespective of $(S_k, \cdots, S_0)$, we can use Fact 2 of conditional expectation to get

$$\bar{\mathbb{E}}\left[\frac{S_{k+1}}{S_k}\middle| S_k, \cdots, S_0\right] = \bar{\mathbb{E}}\left[\frac{S_{k+1}}{S_k}\right] = \bar{p}u + \bar{q}d.$$

Thus

$$\begin{aligned}\bar{\mathbb{E}}[(1+r)^{-(k+1)}S_{k+1}|S_k, \cdots, S_0] &= (1+r)^{-(k+1)}\bar{\mathbb{E}}\left[\frac{S_{k+1}}{S_k}S_k\middle| S_k, \cdots, S_0\right] \\ &= (1+r)^{-(k+1)}S_k\bar{\mathbb{E}}\left[\frac{S_{k+1}}{S_k}\middle| S_k, \cdots, S_0\right] \\ &= (1+r)^{-(k+1)}S_k \cdot (\bar{p}u + \bar{q}d) = (1+r)^{-k}S_k.\end{aligned}$$

In the second line we used Fact 1 of conditional expectation. □

Proof of Proposition 7.3. We want to show that for each k,

$$\bar{\mathbb{E}}[(1+r)^{-(k+1)}W_{k+1}|S_k, \cdots, S_0] = (1+r)^{-k}W_k.$$

Here we use the generalized notion of a martingale, since the information generated by $(W_k, \cdots, W_0)$ is governed by the information about $(S_k, \cdots, S_0)$. Actually it is more convenient to establish the form

$$\bar{\mathbb{E}}[(1+r)^{-(k+1)}W_{k+1} - (1+r)^{-k}W_k|S_k, \cdots, S_0] = 0.$$

The reason is as follows. By (29),

$$W_{k+1} - (1+r)W_k = \Delta_k S_{k+1} - (1+r)\Delta_k S_k = \Delta_k(S_{k+1} - (1+r)S_k),$$

and multiplying both sides by $(1+r)^{-(k+1)}$ gives

$$(1+r)^{-(k+1)}W_{k+1} - (1+r)^{-k}W_k = \Delta_k\left((1+r)^{-(k+1)}S_{k+1} - (1+r)^{-k}S_k\right).$$

Therefore using Fact 1 of conditional expectation,

$$\bar{\mathbb{E}}[(1+r)^{-(k+1)}W_{k+1} - (1+r)^{-k}W_k)|S_k, \cdots, S_0] = \Delta_k\bar{\mathbb{E}}[(1+r)^{-(k+1)}S_{k+1} - (1+r)^{-k}S_k)|S_k, \cdots, S_0] = 0,$$

where in the end we used that $(1+r)^{-k}S_k$ is a martingale, Proposition 7.2. □

This proof actually extends to the continuous-time Itô integral as well. If for all $T \geqslant 0$

$$W_T = W_0 + \int_0^T \Delta_t dS_t,$$

and $\{S_t : t \geqslant 0\}$ is a continuous-time martingale, then $\{W_t : t \geqslant 0\}$ is a continuous-time martingale by essentially the same argument (key is that Δ_t is adapted to S_t).

Proof of Proposition 7.4. We want to show that if V is any random variable which is a function of $S_0, \cdots, S_n$, there exists a constant W_0 (initial wealth) and a portfolio process $(\Delta_0, \Delta_1, \cdots, \Delta_n)$ so that the wealth process W_n at time n matches V.

Let $V_k = (1+r)^k\bar{\mathbb{E}}[(1+r)^{-n}V|S_k, \cdots, S_0]$. Then $(1+r)^{-k}V_k$ is a martingale, because

$$\begin{aligned}\bar{\mathbb{E}}[(1+r)^{-(k+1)}V_{k+1}|S_{k+1}, \cdots, S_0] &= \bar{\mathbb{E}}[\bar{\mathbb{E}}[(1+r)^{-n}V|S_k, \cdots, S_0]|S_{k+1}, \cdots, S_0] \\ &= \bar{\mathbb{E}}[(1+r)^{-n}V|S_k, \cdots, S_0] \\ &= (1+r)^{-k}V_k.\end{aligned}$$

(In the second line we used the identity $\bar{\mathbb{E}}[\bar{\mathbb{E}}[\cdot|S_k,\cdots,S_0]|S_{k+1},S_k,\cdots,S_0] = \bar{\mathbb{E}}[\cdot|S_k,\cdots,S_0]$, which is a slight generalization of Fact 3 of conditional expectation.)

Let $\omega = (t_1, t_2, \cdots, t_n)$, where each t_i can be either u or d, denote the outcome of the stock at time n. Set

$$\Delta_k(\omega) = \frac{V_{k+1}(t_1,\cdots,t_k,u,t_{k+2},\cdots,t_n) - V_{k+1}(t_1,\cdots,t_k,d,t_{k+2},\cdots,t_n)}{S_{k+1}(t_1,\cdots,t_k,u,t_{k+2},\cdots,t_n) - S_{k+1}(t_1,\cdots,t_k,d,t_{k+2},\cdots,t_n)}. \tag{30}$$

Since V_{k+1} and S_{k+1} both depend on $(t_1,\cdots,t_k)$, but NOT on $(t_{k+2},\cdots,t_n)$, it follows that Δ_k is a function of $(t_1,\cdots,t_k)$ ONLY. This shows the adaptability of Δ_k to $(S_0,\cdots,S_k)$ as alluded to previously. In the rest of the proof, we will fix the $(t_1,\cdots,t_k)$, so it is ok to drop the t's and write $V_{k+1}(u)$ as a shorthand for $V_{k+1}(t_1,\cdots,t_k,u,t_{k+2},\cdots,t_n)$. Similar shorthands apply to $V_{k+1}(d)$, $S_{k+1}(u)$ $(= uS_k)$, and $S_{k+1}(d)$ $(= dS_k)$.

Using this shorthand, we can then deduce the following: Since $(1+r)^{-k}V_k$ is a martingale (see above),

$$V_k = \bar{\mathbb{E}}[(1+r)^{-1}V_{k+1}|S_{k+1},\cdots,S_0] = (1+r)^{-1}\left[\bar{p}V_{k+1}(u) + \bar{q}V_{k+1}(d)\right]. \tag{31}$$

Now we are ready to prove the claim set out at the beginning: if we set $W_0 = V_0 = \bar{\mathbb{E}}[(1+r)^{-n}V]$ and $(\Delta_0,\cdots,\Delta_n)$ as in (30), then we will match W_k with V_k for every k, and hence $W_n = V$. This will be proved by induction on k.

When $k = 0$, this boils down to the one-step model which was computed in the previous section. There we found indeed that $W_1 = V_1$.

Now suppose $W_k = V_k$. If we can show that $W_{k+1} = V_{k+1}$ then we're done. Let's use (29), (30), and (31):

$$\begin{aligned}
W_{k+1}(u) &= \Delta_k S_{k+1}(u) + (1+r)(W_k - \Delta_k S_k) \\
&= \Delta_k\left[uS_k - (1+r)S_k\right] + (1+r)V_k \\
&= \frac{V_{k+1}(u) - V_{k+1}(d)}{uS_k - dS_k}\left[uS_k - (1+r)S_k\right] + \left[\bar{p}V_{k+1}(u) + \bar{q}V_{k+1}(d)\right] \\
&= \left(\frac{u-(1+r)}{u-d} + \bar{p}\right)V_{k+1}(u) + \left(-\frac{u-(1+r)}{u-d} + \bar{q}\right)V_{k+1}(d) = V_{k+1}(u).
\end{aligned}$$

Similarly we can show that $W_{k+1}(d) = V_{k+1}(d)$. That does it. □

Proof of Proposition 7.5. By Proposition 7.4, we can build a portfolio process $(\Delta_0,\cdots,\Delta_n)$ so that if we start with $W_0 = (1+r)^{-n}\bar{\mathbb{E}}[V]$, then at time n we would have $W_n = V$, no matter what the market does in between. We now show that if the option is priced at anything but W_0, then riskless profit can occur, which would violate the 'no-arbitrage' rule. [This argument mimics what was done for the one-step model in the last section.]

Suppose the option trades at W_0 plus a positive amount β_0. Then the trader can sell the option to someone else, use the W_0 and the portfolio process $(\Delta_0,\cdots,\Delta_n)$ to replicate the option, and invest the extra β_0 into savings. The total amount generated from the savings constitutes a riskless profit.

Now suppose the option trades at W_0 minus β_0. The trader should then do the opposite by holding $-\Delta_k$ shares of stock at time k and buying an option (and putting the rest in savings). This again generates a riskless profit. □

This last proof is actually instructive because it tells you what hedging strategy (or portfolio process) to adopt when an arbitrage opportunity arises.

7.2.3 The 'Black-Scholes formula' in the binomial model

Using Proposition 7.5 we can now derive explicit expressions for the value of certain options. For example, the European call option $V = (S_n - K)^+$ with expiry n has a fair price of

$$V_0 = \frac{1}{(1+r)^n} \underbrace{\sum_{k=0}^{n} \left(u^k d^{n-k} S_0 - K\right)^+ \binom{n}{k} \bar{p}^k \bar{q}^{n-k}}_{=\bar{\mathbb{E}}[V]},$$

where we remember that to reach stock price $u^k d^{n-k} S_0$ at time n, there are $\binom{n}{k}$ ways, each having probability $\bar{p}^k \bar{q}^{n-k}$ under $\bar{\mathbb{P}}$.

To put this in parallel with the continuous-time Black-Scholes formula, we imagine that each time step is measured in unit of (δt), where (δt) is small enough. Assume that the European option expires at time $T = n(\delta t)$. Then

$$V_0 = \frac{1}{(1+r)^{n(\delta t)}} \sum_{k=0}^{n} \left(u^k d^{n-k} S_0 - K\right)^+ \binom{n}{k} \bar{p}^k \bar{q}^{n-k}.$$

Assuming that $T = n(\delta t)$ is small, we can approximate $(1+r)^{-n(\delta t)}$ by $e^{-rn(\delta t)}$ to first order in $(n(\delta t))$, and get

$$V_0 \approx e^{-rn(\delta t)} \sum_{k=0}^{n} \left(u^k d^{n-k} S_0 - K\right)^+ \binom{n}{k} \bar{p}^k \bar{q}^{n-k}.$$

This is basically the discrete analog of the continuous-time Black-Scholes formula (see next section). We have actually derived something very nontrivial (in fact close to being worthy of a Nobel Prize)! Also this solution is 'closed-form' because the option value at expiry is *independent* of the evolution path of the stock price: uud, udu, and duu gives rise to the same option price. Other options may not enjoy the same 'path-independence' property.

7.2.4 Other options

Besides the European call option, one may also explore other options.

American options.

Look-back options.

Asian options.

7.3 The Black-Scholes model

A Review items from MATH 3160

A.1 Conditional probability & conditional expectation

Conditional distribution. Let X and Y be two jointly distributed discrete random variables. From our prior discussions on conditional probability, it makes sense to ask, "What is the conditional probability of $\{X = x\}$ given that $\{Y = y\}$?" This motivates the **conditional probability mass function** $p_{X|Y}(\cdot|y)$, defined by

$$p_{X|Y}(x|y) := \mathbb{P}[X = x|Y = y] = \frac{\mathbb{P}[X = x, Y = y]}{\mathbb{P}[Y = y]} = \frac{p_{X,Y}(x, y)}{p_Y(y)}$$

whenever $p_Y(y) > 0$. To see why this is a kosher probability mass function, check on your own that $p_{X|Y}(\cdot|y) \geqslant 0$ and $\sum_x p_{X|Y}(x|y) = 1$.

[Another approach to understanding conditional distribution: Think of $m(\cdot) = \mathbb{P}[X = \cdot|Y = y]$ as a mass distribution on X. (Visualize it!) By construction, the total mass, $\sum_x m(x)$, is finite. In order to make this into a probability mass distribution, one needs to divide m by the total mass $\sum_x m(x)$, that is, by $\sum_x \mathbb{P}[X = x|Y = y] = \mathbb{P}[Y = y]$.]

Similarly, if X and Y are jointly distributed continuous random variables, then by analogy (and with a little bit of cheating—read Ross and you'll understand the nature of the cheating) we can define the **conditional probability density function** $f_{X|Y}(\cdot|y)$ via

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

whenever $f_Y(y) > 0$. [Recall also that the marginal density of Y is given by $f_Y(y) = \int f_{X,Y}(x, y)dx$.] Again, I claim that $f_{X|Y}(\cdot|y)$ is a legit probability density function, because (check on your own) $f_{X|Y}(x|y) \geqslant 0$ and $\int f_{X|Y}(x|y)dx = 1$).

Important exercise: Show that if X and Y are jointly discrete (resp. jointly continuous) and independent, then $p_{X|Y}(x|y) = p_X(x)$ (resp. $f_{X|Y}(x|y) = f_X(x)$).

Conditional expectation. Recall the conditional distribution of X given $\{Y = y\}$ from §6.4-6.5. If X and Y are jointly discrete, then we have the conditional probability mass function

$$p_{X|Y}(\cdot|y) = \frac{p_{X,Y}(\cdot, y)}{p_Y(y)}.$$

What is the expectation of X given $\{Y = y\}$? Naturally, it is

$$\sum_x x p_{X|Y}(x|y) =: \mathbb{E}[X|Y = y],$$

whereby a standard notation for conditional expectation has been introduced.

By total (!) analogy, if X and Y are jointly continuous, them the conditional expectation of X given $\{Y = y\}$ is

$$\mathbb{E}[X|Y = y] = \int x f_{X|Y}(x|y)dx = \int x \frac{f_{X,Y}(x, y)}{f_Y(y)}dx.$$

In what follows we write $\mathbb{E}[X|Y]$ instead of $\mathbb{E}[X|Y = y]$. The conditional variance of X given Y is

$$\mathrm{Var}[X|Y] = \mathbb{E}[(X - \mathbb{E}[X|Y])^2|Y].$$

Extremely important remark: $\mathbb{E}[X|Y]$ and $\mathrm{Var}[X|Y]$ are both still random variables, since Y does not have a definite value! Another way to see this is that conditioning on Y means only knowing "partial information" about the joint distribution of (X, Y). Therefore, the conditional expectation (resp. conditional variance) of X remains random. That being said...

Proposition A.1 (Law of total expectation)**.** $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$.

Proof. We only give the proof for when X and Y are jointly discrete. (The proof for when X and Y are jointly continuous is completely analogous.) Notice that since the events $\{Y = y\}$ are disjoint for different y's, and that $\bigcup_y \{Y = y\}$ is the entire sample space, the RHS can be expressed as

$$\begin{aligned}\sum_y \mathbb{E}[X|Y=y]\mathbb{P}[Y=y] &= \sum_y \left(\sum_x x \cdot p_{X|Y}(x|y)\right) p_Y(y) = \sum_y \sum_x x \cdot p_{X,Y}(x,y) \\ &= \sum_x x \sum_y p_{X,Y}(x,y) = \sum_x x p_X(x) = \mathbb{E}[X].\end{aligned}$$

In the intervening steps the identity $p_{X|Y}(x|y)p_Y(y) = p_{X,Y}(x,y)$ was used, as well as an interchange of the two summations. □

Important exercise: Show that if X and Y are independent random variables, then $\mathbb{E}[X|Y] = \mathbb{E}[X]$.

This concludes a recap of conditional probability and expectation at the level of MATH 3160. We will give a slightly more sophisticated treatment of conditional expectation in MATH 3170, in preparation for our discussions on martingales (Chapter 5).

A.2 Poisson and exponential random variables

A.2.1 Properties

A **Poisson** random variable with parameter $\lambda > 0$ (denoted by $X \sim \text{Pois}(\lambda)$) is a discrete random variable taking values in the nonnegative integers $\mathbb{N}_0$, with probability mass function

$$\mathbb{P}[X = k] = e^{-\lambda}\frac{\lambda^k}{k!} \quad \text{for each } k \in \mathbb{N}_0.$$

As you learned in MATH 3160, a Poisson random variable is often used to model the number of "rare" events, such as accidents on I-84, typos in a manuscript, number of street crimes, etc. The parameter λ represents the *mean* number of such events. Indeed, $\mathbb{E}[X] = \lambda$ and $\text{Var}(X) = \lambda$.

To be more quantitative, a binomial distribution with small/rare probability of "success" is well approximated by a Poisson distribution. The following "Poisson approximation to the binomial" is also known as the "Poisson paradigm" or the "law of rare events."

Theorem A.2 (Poisson approximation to the binomial)**.** *Let $\{X_n\}_{n=1}^\infty$ be a sequence of binomial random variables, $X_n \sim \text{Bin}(n, p_n)$, with $\lim_{n\to\infty}(np_n) = \lambda$. Then as $n \to \infty$, X_n converges in distribution to a Poisson random variable $X \sim \text{Pois}(\lambda)$, that is:*

$$\lim_{n\to\infty} \mathbb{P}[X_n \in A] = \mathbb{P}[X \in A] \quad \textit{for every subset } A \subset \mathbb{R}.$$

Proof. Since the binomial and Poisson are both discrete random variables taking nonnegative integer values, it suffices to show that

$$\lim_{n\to\infty} \mathbb{P}[X_n = k] = \mathbb{P}[X = k] \quad \text{for all } k \in \mathbb{N}_0.$$

Recall that for each $k \in \mathbb{N}_0$,

$$\begin{aligned}\mathbb{P}[X_n = k] &= \binom{n}{k}(p_n)^k(1-p_n)^{n-k} = \frac{n!}{k!(n-k)!}(p_n)^k(1-p_n)^{n-k} \\ &= \frac{1}{k!}\cdot\frac{n!}{n^k(n-k)!}\cdot\left(\frac{np_n}{1-p_n}\right)^k\cdot(1-p_n)^n,\end{aligned}$$

where the expression has been rewritten in a way which facilitates taking limits. Since

$$\lim_{n\to\infty} \frac{n!}{n^k(n-k)!} = 1, \quad \lim_{n\to\infty} \left(\frac{np_n}{1-p_n}\right)^k = \lambda^k, \quad \text{and } \lim_{n\to\infty}(1-p_n)^n = \lim_{n\to\infty} e^{-np_n} = e^{-\lambda},$$

we conclude that

$$\lim_{n\to\infty} \mathbb{P}[X_n = k] = e^{-\lambda}\frac{\lambda^k}{k!}.$$

□

Theorem A.2 does not tell us about the error of the Poisson approximation when n is finite. It turns out that whenever $X_n \sim \text{Bin}(n,p)$ and $X \sim \text{Pois}(np)$,

$$|\mathbb{P}[X_n \in A] - \mathbb{P}[X \in A]| \leqslant np^2 \quad \text{for every subset } A \subset \mathbb{R},$$

so the Poisson approximation is "good" when np^2 is "small." This will be discussed in lecture: see Theorem ? of the main text for the statement and the proof.[23]

An **exponential** random variable with parameter $\lambda > 0$ (denoted by $T \sim \text{Exp}(\lambda)$) is a continuous random variable taking values in the positive reals.[24] Its defining property is

$$\mathbb{P}[T > t] = e^{-\lambda t} \quad \text{for every } t > 0. \tag{32}$$

An exponential random variable models the time for a single event to occur, such as the time to the next accident on I-84, time to failure of the power generator, etc. As seen from (32), the probability that this time exceeds t decays at an exponential rate λ in t, whence the name.

An equivalent formulation to (32) is

$$F_T(t) = \mathbb{P}[T \leqslant t] = 1 - e^{-\lambda t} \quad \text{for every } t > 0,$$

which is the cumulative distribution function (cdf) of T. Recall that the probability density function (pdf, or density) f_T of a continuous random variable T is the derivative of the cdf F_T, by the fundamental theorem of calculus. Thus the density of T reads

$$f_T(t) = \begin{cases} \lambda e^{-\lambda t}, & t > 0, \\ 0, & t \leqslant 0. \end{cases}$$

An elementary calculation shows that $\mathbb{E}[T] = \lambda^{-1}$ and $\text{Var}(T) = \lambda^{-2}$.

We say that a random variable X is **memoryless** if

$$\mathbb{P}[X > s + t | X > s] = \mathbb{P}[X > t]$$

for all $s, t > 0$. For example, suppose that X represents the time for an air conditioner to fail. If X is memoryless, then the probability that the A/C lasts more than $(s+t)$ years, conditional upon it being used for at least s years, is the same as the probability that the A/C lasts more than t years from the beginning. In other words, the A/C does not "remember" that it has been used for s years.

Proposition A.3. *The only continuous random variable having the memoryless property is the* ***exponential*** *random variable.*[25]

[23] Had I had an extra lecture in MATH 3160 last fall, I would have discussed the error estimate of the Poisson approximation to the binomial.

[24] Note that some authors use λ^{-1} as the parameter in defining the exponential distribution. The convention appearing here is consistent with Ross, and is in my opinion the better convention because of the connection to the Poisson process.

[25] The only discrete, integer-valued random variable having the memoryless property is the **geometric** random variable.

Proof. First we show that the exponential random variable is memoryless, which presumably you've seen in MATH 3160. Let $T \sim \text{Exp}(\lambda)$. By the definition of conditional probability and (32),

$$\mathbb{P}[T > s+t | T > s] = \frac{\mathbb{P}[T > s+t]}{\mathbb{P}[T > s]} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} = \mathbb{P}[T > t]$$

for all $s, t > 0$.

Next we show that a memoryless continuous random variable X must be an exponential random variable. The proof involves a simple exercise in DiffEq. Let $F_X = \mathbb{P}[X \leqslant \cdot]$ be the cdf of X, and $h_X = 1 - F_X = \mathbb{P}[X > \cdot]$. Using the memoryless property one verifies that

$$h_X(s+t) = h_X(s) h_X(t) \quad \text{for all } s, t > 0. \tag{33}$$

Since X is a continuous random variable, F_X, and in turn h_X, is an (almost everywhere) differentiable function. So we can differentiate (33) with respect to either s or t, and find

$$h'_X(s+t) = h'_X(s) h_X(t) = h_X(s) h'_X(t),$$

or

$$\frac{h'_X(s)}{h_X(s)} = \frac{h'_X(t)}{h_X(t)}.$$

Since each side of the equality depends on a separate variable, the only way to uphold the equality is if both sides are equal to a constant, say, β.[26] Thus we deduce the ordinary differential equation

$$h'_X(t) = \beta h_X(t) \quad \text{for all } t > 0,$$

which has solution $h_X(t) = Ce^{\beta t}$ for an appropriate constant C. Since h_X is a probability, β cannot be positive (or $h_X(t)$ blows up as $t \to \infty$) or zero (which implies that the density $f_X(t) = 0$ for all $t > 0$, a trivial distribution). So $h_X(t) = Ce^{-\lambda t}$ for some $\lambda > 0$, and in particular, $h_X(t)$ increases monotonically toward C as $t \downarrow 0$. As a probability cannot exceed 1, $C = 1$. This proves that X is an exponential random variable by (32). □

A.2.2 Sums of independent Poissons or of i.i.d. exponentials

Key message: The sum of independent Poissons is another Poisson. But the sum of i.i.d. exponentials is a Gamma, not an exponential.

Recall that if X and Y are independent discrete random variables, then their sum $(X+Y)$ has probability mass function

$$p_{X+Y}(k) = \sum_j p_X(j) p_Y(k-j) = \sum_j p_X(k-j) p_Y(j),$$

where the sum runs over all values of X (or Y). Likewise, if X and Y are independent continuous random variables, then $(X+Y)$ has density

$$f_{X+Y}(x) = \int_{-\infty}^{\infty} f_X(y) f_Y(x-y)\,dy = \int_{-\infty}^{\infty} f_X(x-y) f_Y(y)\,dy.$$

In advanced/engineering math terminology, f_{X+Y} is the *convolution* of f_X and f_Y, denoted by $f_X \star f_Y$.

Proposition A.4 (Sum of independent Poissons). *If $X_1 \sim \text{Pois}(\lambda_1)$ and $X_2 \sim \text{Pois}(\lambda_2)$ are independent, then $(X_1 + X_2) \sim \text{Pois}(\lambda_1 + \lambda_2)$. By induction, if $\{X_i\}_{i=1}^n$ is a sequence of independent Poisson random variables with $X_i \sim \text{Pois}(\lambda_i)$, then $(\sum_{i=1}^n X_i) \sim \text{Pois}(\sum_{i=1}^n \lambda_i)$.*

[26] For those of you who had DiffEq, this "separation of variables" trick hopefully rings a bell.

Proof. Proving the first statement is a direct computation: for every $k \in \mathbb{N}_0$,

$$\begin{aligned} p_{X_1+X_2}(k) &= \sum_{j:\ p_{X_1}(j)\neq 0,\ p_{X_2}(k-j)\neq 0} p_{X_1}(j)p_{X_2}(k-j) = \sum_{j=0}^{k}\left(e^{-\lambda_1}\frac{\lambda_1^j}{j!}\right)\left(e^{-\lambda_2}\frac{\lambda_2^{k-j}}{(k-j)!}\right) \\ &= e^{-(\lambda_1+\lambda_2)}\cdot\frac{1}{k!}\sum_{j=0}^{k}\frac{k!}{j!(k-j)!}\lambda_1^j\lambda_2^{k-j} = e^{-(\lambda_1+\lambda_2)}\frac{(\lambda_1+\lambda_2)^k}{k!}, \end{aligned}$$

where in the last equality we used the binomial theorem. Meanwhile $p_{X_1+X_2}(k) = 0$ whenever $k \notin \mathbb{N}_0$.

The induction procedure is straightforward and is left as an exercise for you. □

Proposition A.5 (Sum of i.i.d. exponentials). *If $\{T_i\}_{i=1}^n$ is a sequence of i.i.d. exponential random variables with $T_i \sim \mathrm{Exp}(\lambda)$, then $S_n := (\sum_{i=1}^n T_i)$ has density*

$$f_{S_n}(t) = \begin{cases} e^{-\lambda t}\dfrac{\lambda^n t^{n-1}}{(n-1)!}, & t > 0, \\ 0, & t \leqslant 0. \end{cases}$$

In other words, $S_n \sim \mathrm{Gamma}(n,\lambda)$.

Proof. Let's start with $S_2 = T_1 + T_2$: for every $t > 0$,

$$\begin{aligned} f_{T_1+T_2}(t) &= \int_{-\infty}^{\infty} f_{T_1}(s)f_{T_2}(t-s)ds = \int_0^{\infty}(\lambda e^{-\lambda s})(\lambda e^{-\lambda(t-s)}\mathbb{1}_{\{t>s\}})ds \\ &= \int_0^t \lambda^2 e^{-\lambda t}ds = \lambda^2 t e^{-\lambda t}, \end{aligned}$$

which matches the density of the Gamma distribution $\mathrm{Gamma}(2,\lambda)$.

For the induction step, suppose that S_k follows the $\mathrm{Gamma}(k,\lambda)$ distribution. Then S_{k+1}, being the independent sum of S_k and T_{k+1}, has density

$$\begin{aligned} f_{S_k+T_{k+1}}(t) &= \int_{-\infty}^{\infty} f_{S_k}(s)f_{T_{k+1}}(t-s)ds = \int_0^{\infty}\left(e^{-\lambda s}\frac{\lambda^k s^{k-1}}{(k-1)!}\right)\left(\lambda e^{-\lambda(t-s)}\mathbb{1}_{\{t>s\}}\right)ds \\ &= \int_0^t \frac{\lambda^k s^{k-1}}{(k-1)!}\lambda e^{-\lambda t}ds = e^{-\lambda t}\frac{\lambda^{k+1}t^k}{k!} \end{aligned}$$

if $t > 0$, and 0 if $t \leqslant 0$. This matches the density of the $\mathrm{Gamma}(k+1,\lambda)$ distribution. □

This concludes a short review of Poisson and exponential random variables at the level of MATH 3160. You can now go back to Chapter 2 of the main text to see how they are used in action.

A.3 Useful inequalities

A.3.1 Chebyshev's inequality

Theorem A.6 (Markov's inequality). *Let X be any nonnegative r.v. For every $a > 0$,*

$$\mathbb{P}[X \geqslant a] \leqslant \frac{\mathbb{E}[X]}{a}.$$

Proof. First observe the following:

$$X = X\mathbb{1}_{\{X<a\}} + X\mathbb{1}_{\{X\geqslant a\}} \geqslant a\mathbb{1}_{\{X\geqslant a\}}.$$

The equality comes from splitting up the constant function 1 into the sum of two indicator functions, $\mathbb{1}_{\{X<a\}}$ and $\mathbb{1}_{\{X\geqslant a\}}$. As for the inequality, notice that $X\mathbb{1}_{\{X<a\}} \geqslant 0$ since X is nonnegative by assumption, and that $X\mathbb{1}_{\{X\geqslant a\}} \geqslant a\mathbb{1}_{\{X\geqslant a\}}$.

Taking expectation on both sides preserves the inequality. (Why?)

$$\mathbb{E}[X] \geqslant a\mathbb{E}[\mathbb{1}_{\{X\geqslant a\}}] = a\mathbb{P}[X \geqslant a],$$

where in the last equality we used the identity that for any event A, $\mathbb{E}[\mathbb{1}_A] = \mathbb{P}(A)$. □

Theorem A.7 (Chebyshev's inequality)**.** *Let X be any r.v. with finite first moment (mean) μ and finite variance σ^2. Then for any $k > 0$,*

$$\mathbb{P}[|X-\mu| \geqslant k] \leqslant \frac{\sigma^2}{k^2}.$$

Proof. Realize that $|X-\mu|^2$ is a nonnegative r.v. So by a trivial identity followed by Markov's inequality (Theorem A.6), we have

$$\mathbb{P}[|X-\mu| \geqslant k] = \mathbb{P}[|X-\mu|^2 \geqslant k^2] \leqslant \frac{\mathbb{E}[|X-\mu|^2]}{k^2} = \frac{\sigma^2}{k^2}.$$

□

In some textbooks Markov's inequality is also called Chebyshev's inequality.

A.3.2 Jensen's inequality

Theorem A.8 (Jensen's inequality)**.** *For any real-valued r.v. X and any convex function $\varphi : \mathbb{R} \to \mathbb{R}$,*

$$\varphi(\mathbb{E}[X]) \leqslant \mathbb{E}[\varphi(X)].$$

A.4 Limit theorems

Throughout this section, $\{X_i\}_{i=1}^\infty$ is a sequence of i.i.d. r.v.'s drawn from a probability distribution with mean μ and variance σ^2. We denote by

$$\bar{X}_N = \frac{1}{N}\sum_{i=1}^N X_i$$

the *sample mean* of the distribution.

A.4.1 The law of large numbers (LLN)

The LLN refers to the fact that as the sample size goes to infinity, the (sequence of) sample means, as random variables, will converge to the mean of the distribution μ *almost surely*. We begin with a weaker version of this statement, which has a much simpler proof.

Theorem A.9 (Weak LLN for i.i.d. r.v.'s)**.** *Let $\{X_i\}_{i=1}^\infty$ be any sequence of i.i.d. r.v.'s having finite mean $\mathbb{E}[X_i] = \mu$. Then for every $\epsilon > 0$,*

$$\mathbb{P}\left[\left|\bar{X}_N - \mu\right| \geqslant \epsilon\right] \xrightarrow[N\to\infty]{} 0.$$

Proof, assuming that the variance $\sigma^2 = \mathrm{Var}(X_i)$ is finite. Since

$$\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N X_i\right] = \mu \quad \text{and} \quad \mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^N X_i\right) = \frac{1}{N^2}\sum_{i=1}^N \mathrm{Var}(X_i) = \frac{1}{N^2}\cdot N\sigma^2 = \frac{\sigma^2}{N},$$

the latter of which follows from the independence of the X_i, we can apply Chebyshev's inequality (Theorem A.7) to find

$$0 \leqslant \mathbb{P}\left[\left|\frac{1}{N}\sum_{i=1}^{N} X_i - \mu\right| \geqslant \epsilon\right] \leqslant \frac{1}{\epsilon^2}\cdot\frac{\sigma^2}{N} \xrightarrow[N\to\infty]{} 0.$$

□

A different proof of Theorem A.9 using characteristic functions is slightly outside the scope of MATH 3160. However, it has the advantage of dispensing with the finite variance assumption. See Wikipedia.

Theorem A.10 (Strong LLN for i.i.d. r.v.'s)**.** *Let $\{X_i\}_{i=1}^{\infty}$ be any sequence of i.i.d. r.v.'s having finite mean $\mathbb{E}[X_i] = \mu$. Then with probability* 1 *(that is, almost surely),*

$$\bar{X}_N \xrightarrow[N\to\infty]{} \mu.$$

[For those who recall Bolzano's definition of convergence[27] from Calculus (whatever version), the strong LLN can be restated as follows. *For every $\epsilon > 0$, there exists a positive integer N_0 such that for all $N \geqslant N_0$, $|\bar{X}_N - \mu| < \epsilon$ with probability* 1.]

Due to the complexity of the proof and the lack of class time, the strong LLN was not proved in MATH 3160 (when I taught it in Fall 2013). Read Ross §8.4 if you are interested.

What is the difference between the weak LLN and the strong LLN? Even though both versions of the LLN say that the sample means $\{\bar{X}_N\}_N$ converge to μ as $N \to \infty$, the two *modes of convergence* differ subtly[28]. The weak LLN says that for a fixed large N, the sample mean $\bar{X}_N$ is near μ with a bit of wiggle room. It lends to the possibility that the event $\{|\bar{X}_N - \mu| > \epsilon\}$ happens infinitely many times, though very infrequently (when considered against the far more infinite space of all outcomes). The strong LLN, however, rules out such a possibility: Given any $\epsilon > 0$, $|\bar{X}_N - \mu| < \epsilon$ for *all* sufficiently large N *with probability 1.* A moment's thought will tell you that the strong LLN implies the weak LLN, but not the other way around.

At any rate, the LLN justifies the frequentist approach to probability. If one flips a biased coin (with probability p of landing a head) independently many times, the frequency of showing heads (relative to the number of flips) will converge to p, as the number of flips tends to infinity.

A.4.2 The central limit theorem (CLT)

Let

$$Z_N = \frac{S_N - N\mu}{\sqrt{N\sigma^2}} = \sqrt{N}\left(\frac{\bar{X}_N - \mu}{\sqrt{\sigma^2}}\right).$$

Note that each Z_N has mean 0 and variance 1. Recall also that Φ denotes the c.d.f. of a standard normal distribution.

Theorem A.11 (Classical CLT)**.** *Let $\{X_i\}_{i=1}^{\infty}$ be $\underline{\text{any}}$ sequence of i.i.d. r.v.'s with finite mean μ and finite variance σ^2. Then*

$$Z_N \xrightarrow{d} \mathcal{N}(0,1),$$

where "$\xrightarrow{d}$" means "converges in distribution." In other words, for every $a \in \mathbb{R}$,

$$\mathbb{P}\left[Z_N \leqslant a\right] \xrightarrow[N\to\infty]{} \Phi(a).$$

[27] an extremely fundamental mathematical concept which a great majority of calculus students forget or do not learn properly.

[28] For the technically minded: The weak LLN is the *convergence in probability* of the sample means; the strong LLN, *almost sure convergence* or *strong convergence* of the sample means.

Informally, the CLT tells us that the sample mean $\bar{X}_N$ has the following asymptotic expansion:

$$\bar{X}_N \sim \left[\mu + \frac{\sigma}{\sqrt{N}}\mathcal{N}(0,1) + (\text{lower order terms in } N)\right] \quad \text{as } N \to \infty.$$

Though the precise mode of convergence must be clarified, see the caveat below.

Caveat: It is NOT correct to say that every instance/sample of the r.v. $\lim_{N\to\infty} Z_N$ is a standard normal r.v. (It's false BTW. If $N_1 \ll N_2$, then the difference between Z_{N_1} and Z_{N_2} are almost independent from each other, since the bulk of Z_{N_2} is determined by the r.v.'s X_i for $i > N_1$. As a result, $|Z_{N_2} - Z_{N_1}|$ need not be close to 0 for all sufficiently large N_1 and N_2 in the almost sure sense, or even in probability.) The correct statement is: the standard normal distribution (whose density is "the mother of all bell curves") is the *asymptotic distribution* of the r.v. Z_N as $N \to \infty$. The CLT convergence refers to the c.d.f. (distribution), not to the the r.v. itself (unlike the weak/strong LLN).

Before proving the classical CLT, we need a fact/lemma, which will be stated without proof[29].

Lemma A.12 (Convergence of the moment-generating functions implies convergence in distribution)**.** *Let $\{X_n\}_n$ be a sequence of r.v.'s and Y be another r.v. Suppose that the moment-generating functions of the X_n converge to the moment-generating function of Y:*

$$m_{X_n}(t) \underset{n\to\infty}{\longrightarrow} m_Y(t) \quad \textit{for every } t \in \mathbb{R}.$$

Then $X_n \xrightarrow{d} Y$, that is,

$$\lim_{n\to\infty} \mathbb{P}[X_n \leqslant a] = \mathbb{P}[Y \leqslant a] \quad \textit{for all } a \in \mathbb{R}^{30}.$$

Proof of Theorem A.11. Let us assume that the i.i.d. sequence $\{X_i\}_{i=1}^\infty$ is drawn from a distribution with mean μ and variance σ^2. For each i, let $Y_i = \dfrac{X_i - \mu}{\sqrt{\sigma^2}}$. Then an elementary calculation shows that sequence $\{Y_i\}_{i=1}^\infty$ is i.i.d. with mean zero and variance 1. It follows then that

$$Z_N = \frac{S_N - N\mu}{\sqrt{N\sigma^2}} = \frac{\sum_{i=1}^N [X_i - \mu]}{\sqrt{N\sigma^2}} = \frac{\sum_{i=1}^N Y_i}{\sqrt{N}}.$$

In other words, it is enough to prove the CLT in the setting where the distribution is "standardized" with mean zero and variance 1.

Consider the sequence of r.v.'s $\left\{\dfrac{1}{\sqrt{N}}\displaystyle\sum_{i=1}^N Y_i\right\}_N$. Their moment-generating functions can be expressed as follows:

$$m_{\frac{1}{\sqrt{N}}\sum_{i=1}^N Y_i}(t) = \mathbb{E}\left[e^{t\left(\frac{1}{\sqrt{N}}\sum_{i=1}^N Y_i\right)}\right] = \mathbb{E}\left[\prod_{i=1}^N e^{tY_i/\sqrt{N}}\right] = \prod_{i=1}^N \mathbb{E}\left[e^{tY_i/\sqrt{N}}\right] = \left(\mathbb{E}\left[e^{tY_1/\sqrt{N}}\right]\right)^N,$$

where in the penultimate equality we used the identity that $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ whenever X and Y are independent.

By Taylor expansion (about $t = 0$),

$$\mathbb{E}\left[e^{tY_1/\sqrt{N}}\right] = 1 + \frac{1}{\sqrt{N}}t\cancel{\mathbb{E}[Y_1]} + \frac{1}{N}\frac{t^2}{2}\mathbb{E}[Y_1^2] + R_N = 1 + \frac{t^2}{2N} + R_N, \tag{34}$$

where R_N denotes the error term which decays faster than $\frac{1}{N}$. We claim that

$$\lim_{N\to\infty}\left(\mathbb{E}\left[e^{tY_1/\sqrt{N}}\right]\right)^N = \lim_{N\to\infty}\left(1 + \frac{t^2}{2N} + R_N\right)^N = e^{t^2/2}.$$

[29] Recall that a moment-generating function uniquely defines a probability distribution, and vice versa. The content of Lemma A.12 is that this one-to-one relationship remains intact upon taking limits.

[30] More precisely, for all continuity points a of the c.d.f. F_Y.

To see why this is true, take the logarithm on both sides of (34), multiply by N, and then take the limit $N \to \infty$:

$$\lim_{N\to\infty} N \log \mathbb{E}\left[e^{tX_1/\sqrt{N}}\right] = \lim_{N\to\infty} N \log\left(1 + \frac{t^2}{2N} + R_N\right) = \lim_{N\to\infty} N \left[\frac{t^2}{2N} + R_N + R'_N\right] = \frac{t^2}{2}.$$

Here recall the Taylor expansion $\log(1+x) = x + \frac{x^2}{2} + \frac{x^3}{3} + \cdots$, so it's not hard to convince yourself that the remainder term R'_N decays faster than $\frac{1}{N}$. Next, exponentiate both sides, and realize that the monotonicity of the exponential function allows you to interchange the limit with the exponentiation:

$$\lim_{N\to\infty} \left(\mathbb{E}\left[e^{tY_1/\sqrt{N}}\right]\right)^N = \exp\left(\lim_{N\to\infty} N \log \mathbb{E}\left[e^{tY_1/\sqrt{N}}\right]\right) = e^{t^2/2}.$$

Thus we have demonstrated that

$$\lim_{N\to\infty} m_{\frac{1}{\sqrt{N}}\sum_{i=1}^N Y_i}(t) = e^{t^2/2} = m_Z(t),$$

the moment-generating function of a standard normal r.v. By Lemma A.12, $Z_N = \frac{1}{\sqrt{N}}\sum_{i=1}^N Y_i \xrightarrow{d} Z$. □

B Other Potpourri

B.1 Some linear algebra facts

B.1.1 The Perron-Frobenius theorem

B.2 Some analysis (advanced calculus) facts

This section contains things that aren't necessarily taught in three semesters of calculus, but are most certainly covered in the first semester of undergraduate analysis. I suspect some of these will be useful for understanding the subtler notions in MATH 3170.

B.2.1 Supremum and infimum

An important property of the real number system is the *least upper bound property*. Let S be any set of real numbers. If $b \geqslant x$ for all $x \in S$, we say that b is an upper bound on S. Let $U = \{b \in \mathbb{R} : b \text{ is an upper bound on } S\}$ be the set of upper bounds on S.

Proposition B.1 (The least upper bound property)**.** *There is a smallest element $c \in U$ such that $c \leqslant b$ for all $b \in U$.*

We call c the **least upper bound** on, or the **supremum** of, S, denoted by $\sup S$. Likewise, there is a **greatest lower bound** on, or the **infimum** of, S, denoted by $\inf S$. Note that since the empty set $\varnothing$ contains no real numbers, any real number is considered an upper bound *and* a lower bound on $\varnothing$. By this convention, $\sup \varnothing = -\infty$ and $\inf \varnothing = +\infty$.

Exercise: Formulate the greatest lower bound property of the real numbers, and show that it is equivalent to the least upper bound property.

The supremum (resp. infimum) of a set should be distinguished from the maximum (resp. minimum), the largest (resp. smallest) element of the set. The former always exists, but the latter need not.

Example. Let $S = \{1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}\}$. Then $\sup S = \max S = 1$ and $\inf S = \min S = \frac{1}{4}$. In fact, if S is a finite set, it always has a maximum and a minimum.

Example. Let $S = \{1, \frac{1}{2}, \frac{1}{3}, \cdots, \frac{1}{n}, \cdots\}$. Here $\sup S = \max S = 1$ still, but S has no smallest element. However $\inf S$ exists and is equal to 0.

B.2.2 Limsup and liminf

As you surely know, a given sequence of real numbers $\{a_n\}_n$ need not have a limit. But that does not prevent us from putting the least upper asymptotic bound, and the greatest lower asymptotic bound, on $\{a_n\}_n$. In fact we will show that these two numbers *always* exist; when they coincide, we say that the limit of $\{a_n\}_n$ exists.

We begin with a very elementary but useful fact. Recall that a sequence $\{a_n\}_n$ is *monotone* if either $a_n \leqslant a_{n+1}$ for all n (monotone increasing) or $a_n \geqslant a_{n+1}$ for all n (monotone decreasing).

Proposition B.2. *Let $\{a_n\}_n$ be a monotone sequence. Then $\lim\limits_{n\to\infty} a_n$ always exist as an extended real number (that is, it may be $\pm\infty$). In particular, if $\{a_n\}_n$ is a bounded monotone sequence, then $\lim\limits_{n\to\infty} a_n$ is finite.*

Proof. □

Let us now define, for any sequence $\{a_n\}_n$, its **limit superior** and **limit inferior**:

$$\limsup_{n\to\infty} a_n = \lim_{n\to\infty} \sup_{m \geqslant n} a_m \quad \text{and} \quad \liminf_{n\to\infty} a_n = \lim_{n\to\infty} \inf_{m \geqslant n} a_m.$$

Proposition B.3. *For any sequence $\{a_n\}_n$ of real numbers, $\limsup a_n$ and $\liminf a_n$ always exist as extended real numbers (that is, they may be $\pm\infty$).*

Proof. Let $B_n = \sup_{m \geqslant n} a_m$; observe that $\{B_n\}_n$ is a monotone decreasing sequence. By Proposition B.2, $\lim_{n\to\infty} B_n$, or $\limsup_{n\to\infty} a_n$, exists. The argument for liminf is utterly similar. □

Example. Let $\{a_n\}_n$ be a sequence where $a_n = 0$ if n is even, and $a_n = 1$ if n is odd. This sequence does not have a limit, but $\limsup a_n = 1$ and $\liminf a_n = 0$.

B.3 Generating functions

Let $\{p_n\}_{n=0}^\infty$ be a sequence of real numbers. In most of our examples, $p_n \geqslant 0$ for all n, and sometimes we may have $\sum_n p_n = 1$ (that is, $\{p_n\}$ is a probability distribution on $\mathbb{N}_0$). The **generating function** associated with $\{p_n\}$ is defined by

$$P(x) = \sum_{n=0}^\infty p_n x^n$$

for all $x \in \mathbb{R}$ (or $x \in \mathbb{C}$) where this series converges. As you learned from Calc II, the *radius of convergence* for the series P is the number $r \in \mathbb{R}_+ \cup \{+\infty\}$ such that the series converges if $|x| < r$, and diverges if $|x| > r$. It is proved in undergraduate analysis that

$$r^{-1} = \limsup_{n\to\infty} |p_n|^{1/n}.$$

Example B.1 (Moment-generating function). Suppose $\{p_n\}$ is a probability distribution on $\mathbb{N}_0$. For each $t \in \mathbb{R}$, put

$$P(e^t) = \sum_{n=0}^\infty p_n e^{tn} = \mathbb{E}[e^{tX}],$$

where X is a discrete random variable with probability mass function $\mathbb{P}[X = n] = p_n$. This is known as the **moment-generating function (MGF)** of X, which you presumably saw in MATH 3160. (Engineers will recognize this as a Laplace transform on the p_n.) The MGF need not be defined for all $t \in \mathbb{R}$. When it is defined on a neighborhood of 0, we can take its t-derivatives at $t = 0$ to find the various moments of X:

$$\mathbb{E}[X^n] = \frac{d^n}{dt^n}\mathbb{E}[e^{tx}]\bigg|_{t=0}.$$

This explains the origin of "moment-generating."

Example B.2 (Characteristic function). Again we assume that $\{p_n\}$ is a probability distribution on $\mathbb{N}_0$. Suppose we replace the $t \in \mathbb{R}$ in the moment-generating function by it, where $i = \sqrt{-1}$. The function $\varphi_X : \mathbb{R} \to \mathbb{C}$ defined by

$$\varphi_X(t) = P(e^{it}) \left(= \mathbb{E}[e^{itX}]\right)$$

is called the **characteristic function** of the random variable X. (It is the Fourier transform of p_n.) The nice thing about the characteristic function is that it is defined on all of $\mathbb{R}$ (*i.e.,* the radius of convergence is ∞), since for any $t \in \mathbb{R}$,

$$|P(e^{it})| = \left|\sum_{n=0}^\infty p_n e^{itn}\right| \leqslant \sum_{n=0}^\infty p_n |e^{itn}| = \sum_{n=0}^\infty p_n = 1 < \infty.$$

In MATH 3160 we established several results using the moment-generating function, assuming that the MGF converges. If one uses the corresponding characteristic function, then no convergence assumption is necessary. That being said, we will not discuss the use of characteristic function because it involves notions from complex analysis which would take us too far afield.

Let $f, g : \mathbb{N}_0 \to \mathbb{R}$ be two functions on the nonnegative integers. We write $f_n = f(n)$, and set $f_{-n} = 0$ for every $n \in \mathbb{N}$ by default.

Definition B.1. The function $(f \star g) : \mathbb{N}_0 \to \mathbb{R}$ defined by

$$(f \star g)_n = \sum_{m=0}^{\infty} f_m g_{n-m}$$

is called the **convolution** of f and g.

Convolution shows up in many contexts. In electrical engineering, when you take two time signals and add them, you convolve them. (*Good to know:* the sum of two identical square waves is NOT a square wave, but a triangular (sawtooth) wave.) In probability, when you add two independent random variables X and Y, the distribution of $(X+Y)$ is the convolution of the distribution of X and the distribution of Y.

More often than not, convolutions can be difficult to compute by hand. This is where generating functions become extremely handy. Let f, g, and $f \star g$ be as above, and let $\widehat{f}$, $\widehat{g}$, and $\widehat{(f \star g)}$ be their respective generating functions. The next result says that the generating function of the convolution is the *product* of the individual generating functions.

Proposition B.4. *$\widehat{(f \star g)}(x) = \widehat{f}(x)\widehat{g}(x)$ for all $x \in \mathbb{N}_0$ whereby all three functions make sense.*

Proof. Using Definition B.1 we find

$$\widehat{(f \star g)}(x) = \sum_{n=0}^{\infty} (f \star g)_n x^n = \sum_{n=0}^{\infty} \left(\sum_{m=0}^{\infty} f_m g_{n-m} \right) x^n.$$

Meanwhile

$$\widehat{f}(x)\widehat{g}(x) = \left(\sum_{j=0}^{\infty} f_j x^j \right) \left(\sum_{k=0}^{\infty} g_k x^k \right) = \sum_{k=0}^{\infty} \sum_{j=0}^{\infty} f_j g_k x^{j+k} = \sum_{n=0}^{\infty} \sum_{j=0}^{\infty} f_j g_{n-j} x^n,$$

where we made a change of variables from k to $n = j+k$ at the end (and swept any convergence issues under the rug). Matching the two series term by term proves the identity. □

It is straightforward to extend this result to n-fold convolutions:

$$(f_1 \widehat{\star \cdots \star} f_n)(x) = \prod_{i=1}^{n} \widehat{f}_i(x).$$

This identity is useful for various computations involving the sum of independent random variables. Applications include the proof of the central limit theorem, and the derivation of certain quantities associated with random walk on $\mathbb{Z}^d$.

B.4 Stirling's approximation

The big-$\mathcal{O}$ notation. Let $f, g : \mathbb{N} \to \mathbb{R}$ be two functions. We say that $f(n) = \mathcal{O}(g(n))$ as $n \to \infty$ if there exist a positive constant M and $n_0 \in \mathbb{N}$ such that for all $n > n_0$,

$$\frac{|f(n|}{|g(n)|} \leqslant M.$$

That is to say, $|f(n)|$ grows asymptotically at the same rate as $|g(n)|$.

One of the hallmark results of analysis, which has numerous applications in all areas of mathematics and the physical sciences is

Theorem B.5 (Stirling's approximation)**.** *As $n \to \infty$,*

$$n! = n^n e^{-n} \sqrt{2\pi n} \cdot \mathcal{O}(1).$$

Proof. There are multiple proofs, which you can look up on Wikipedia. Prof. Keith Conrad has also written a blurb on this subject, which I invite you to read at your leisure. □

We will use Stirling's approximation several times in this course to find asymptotic quantities.