Stochastic Calculus

Jonathan Goodman

last revised September 20, 2011

Preface

This book expands on the material in the one semester Masters level class with the same name I taught several times at the Courant Institute of NYU. A majority of the students were from our masters program in financial mathematics but there always was a healthy minority from disciplines ranging from physics and chemistry to economics and statistics. Over the years I have experimented with various forms of the material and presentation styles. This version have proven more effective than my earlier attempts. It benefits from years of mostly friendly feedback from many students, for which I am very grateful.

The class and the book present a succinct introduction to random processes as viewed by this applied mathematician. There are four main goals: some important properties of random processes, an appreciation of one mathematical formalism for describing and studying random processes, the techniques for creating a mathematical model of a random system, and methods for making quantitative calculations about the model. I always assume that the student/reader wants to apply these ideas to real problems.

This book is not completely rigorous in the mathematical sense. Many of the hardest theorems are not proven and serious mathematical issues go unmentioned. I have included quantitative arguments, such as the convergence of approximations to the Ito integral. This is similar to many Calculus textbooks that explain the convergence of Riemann sums without proving the completeness of the real number system. The industrious reader with time to kill will have no trouble finding rigorous versions. But many readers do not have the time nor the inclination for a deep and extended study of the measure theoretic foundations of probability theory.

The prerequisites for the book are as linear algebra, multivariate calculus, and basic probability. The book uses the level of mathematical maturity I found was appropriate for most students in the class. The book has many computational exercises that require the student to use some programming system. Students used systems ranging from direct C/C++ programming to packages such as Matlab and R, to spreadsheets and Visual Basic.

Contents

Chapter 1

Discrete Probability

This section covers the basic ideas of discrete probability and conditional expectation. The material might seem dry now, but examples are coming. We develop the *modern* view of stochastic processes, partially revealed information, and conditional expectation. These are are easy to understand in the discrete setting because they are simple restatements of *classical* conditional expectation. The modern formalism is not always helpful for simple problems, but it can be just the thing for understanding more subtle stochastic processes and decision problems under incomplete information.

A central role in this discussion is played by algebras (systems) of sets that convey a state of partial information. It may seem artificial at first, but sets and algebras of sets are a convenient framework that makes general abstract probability simple. The review here is not so much to remind the reader of basic discrete probability, but to use it as a simple context in which to explain the role of algebras of sets.

Computational methods are an essential part of all fields of applied mathematics today. For Stochastic Calculus, much of this computation is stochastic simulation and Monte Carlo – the distinction is explained below.

1.1 Discrete probability

In abstract probability, we imagine that there is some *experiment* or *trial* that produce a randoms *outcome*. This outcome is called ω. The set of all possible outcomes is Ω, which is the *probability space*. The probability space Ω is *discrete* if it is finite or countable.[1] We discuss only discrete probability here. The outcome ω is often called a *random variable*. I avoid that term because I (and most other people) want to call functions $X(\omega)$ random variables, see below.

The probability of a specific outcome is $P(\omega)$. It is a number between 0 and 1, with $P = 0$ for events that never happen and $P = 1$ is for events that we know (before doing the experiment) will happen. A probability cannot be negative. We always assume that $\sum_{\omega\in\Omega} P(\omega) = 1$. This is the statement that Ω is a complete list of all possible outcomes. If the experiment that produces ω has multiple steps, then the description of ω has multiple corresponding components. For example, if the experiment calls for measuring three numbers, then ω has three components and may be considered to be an element of a three dimensional space.

The interpretation of probability is a matter for philosophers, but we might say that $P(\omega)$ is the probability of outcome ω happening, or the fraction of times event ω would happen in a large number of independent trials. The philosophical problem is that it may be impossible actually to perform a large number of independent trials. People also sometimes say that probabilities

[1] A set is *countable* if it is possible to make an infinite list of all elements in the set. For example, the rational numbers between 0 and 1 are countable because of the list: (1) $\frac{1}{2}$, (2) $\frac{1}{3}$, (3) $\frac{2}{3}$, (4) $\frac{1}{4}$, (5) $\frac{2}{4}$, (6) $\frac{3}{4}$, (7) $\frac{1}{5}$, etc. The set of all real numbers between 0 and 1 is not countable in this sense.

represent our often subjective (lack of) knowledge of future events. Probability 1 means something that is certain to happen, while probability 0 is for something that cannot happen. "Probability zero $\Rightarrow$ impossible" is strictly true only for discrete probability.

An *event* is a set of outcomes, which is the same as a subset of Ω. The probability of an event is the sum of the probabilities of the outcomes that make up the event

$$P(A) = \sum_{\omega \in A} P(\omega) \ . \tag{1.1}$$

Usually, we specify an event in some way other than listing all the outcomes in it (see below). We do not distinguish between the outcome ω and the event that that outcome occurred $A = \{\omega\}$. That is, we write $P(\omega)$ for $P(\{\omega\})$ or vice versa. This is called "abuse of notation": we use notation in a way that is not absolutely correct but whose meaning is clear.

The formula (1.1) defines the probabilities of events from the probabilities of the outcomes that make up the event. This is possible only when the probability space is discrete. In this book, a probability space that is not discrete is called *continuous*. Roughly speaking, the difference between continuous and discrete probability is the difference between integrals and sums. The probabilities of events in continuous probability are found by integrating using a probability density or probability measure.

Here is a simple example of the above definitions. Suppose you toss a coin four times and that each output is H (heads) or T (tails).[2] The outcome "first H, then T, then H, then H" is written $\omega = \text{HTHH}$. The possible outcomes are $\Omega = \{\text{HHHH}, \text{HHHT}, \text{HTHH}, \ldots, \text{TTTT}\}$. The number of outcomes is $\#(\Omega) = |\Omega| = 16$. If each outcome is equally likely, $P(\omega) = \frac{1}{16}$ for each $\omega \in \Omega$. If A is the event that the first two tosses are H, then

$$A = \{\text{HHHH, HHHT, HHTH, HHTT}\} \ .$$

There are 4 elements (outcomes) in A, each having probability $\frac{1}{16}$ Therefore

$$P(\text{first two H}) = P(A) = \sum_{\omega \in A} P(\omega) = \sum_{\omega \in A} \frac{1}{16} = 4 \cdot \frac{1}{16} = \frac{1}{4} \ .$$

Set operations apply to events since events are sets of outcomes. If A and B are events, then "A and B" means that both A and B happened. That means that the outcome ω is in both A and B. Thus, "A and B" is the intersection $A \cap B$. Similarly, "A or B" is the event consisting of outcomes ω with $\omega \in A$ or $\omega \in B$ or both.[3] Thus "A or B" is the union $A \cup B$. The *complement* of A, A^c, is the event "not A", the set of outcomes not in A. The empty event is the empty set, the set with no elements, $\emptyset$. The probability of $\emptyset$ is zero because the sum (1.1) that defines it has no terms. The complement of $\emptyset$ is Ω. Events A

[2] Most coins have a face, or head, on one side. A U.S. nickel last made in 1938 has a buffalo with a tail on the other side.

[3] There is an *exclusive or*, which means A or B, but not both. The "or" here is not exclusive.

and B are disjoint if $A \cap B = \emptyset$. Event A is contained in event B, $A \subseteq B$, if every outcome in A is also in B. For example, if the event A is as above and B is the event that the last toss in T, then $|B| = 8$, and the event "A and B" is $A \cap B = \{\text{HHHT}, \text{HHTT}\}$. If C is the event the first toss is H, then $A \subseteq C$.

The representation (1.1) implies some of the basic facts about probabilities of events. First $P(A) \leq P(B)$ if $A \subseteq B$. Also, $P(A) + P(B) \geq P(A \cup B)$, with $P(A) + P(B) = P(A \cup B)$ if $P(A \cap B) = \emptyset$, but usually not otherwise. Clearly, $P(A) + P(A^c) = P(\Omega) = 1$.

1.2 Conditional probability

The probability of outcome A given that B has occurred is the *conditional probability* of A given B,

$$P(A \,|\, B) = \frac{P(A \cap B)}{P(B)} \ . \tag{1.2}$$

This is the fraction of B outcomes that are also A outcomes. The formula (1.2) is called *Bayes' rule*. It is often used to calculate $P(A \cap B)$ once we know $P(B)$ and $P(A \,|\, B)$. The formula for that is $P(A \cap B) = P(A \,|\, B) P(B)$.

Events A and B are *independent* if $P(A \mid B) = P(A)$. That is, knowing whether or not B occurred does not change the probability of A. In view of Bayes' rule, this is expressed as

$$P(A \cap B) \ = \ P(A) \cdot P(B) \ . \tag{1.3}$$

For example, suppose A is the event that two of the four tosses are H, and B is the event that the first toss is H. Then A has 6 elements, B has 8, and $A \cap B$ has 3 (check this by listing them). If each outcome has probability $\frac{1}{16}$, this gives $P(A \cap B) = \frac{3}{16}$ while $P(A) = \frac{6}{16}$ and $P(B) = \frac{8}{16} = \frac{1}{2}$, and (1.3) is satisfied. Knowing that half the tosses are H (event A) does not change the probability that the first toss was H. Note that the definition (1.3) is symmetric. If A is independent of B then B is independent of A. Knowing that the first toss was H (event B) does not change the probability that exactly half the tosses are H.

On the other hand, if C is the event that 3 of the 4 tosses are H, then $P(C) = \frac{4}{16} = \frac{1}{4}$ and $P(B \cap C) = \frac{3}{16}$, because

$$B \cap C = \{\text{HHHT, HHTH, HTHH}\}$$

has three elements. Bayes' rule (1.2) gives $P(B \,|\, C) = \frac{3}{16} / \frac{3}{4} = \frac{3}{4}$. Knowing that there are 3 heads in all raises the probability that the first toss is H from $\frac{1}{2}$ to $\frac{3}{4}$.

These two examples show that the question of independence is a question about the equality (1.3). In many cases events A and B are obviously independent because whatever factors determine whether $\omega \in A$ have nothing to do with the what determines whether $\omega \in B$. This might be the case, for example,

for two independent coin tosses. In other cases the events may be intertwined and yet technically independent.

Let us fix the event B with $P(B) > 0$, and discuss the conditional probability $P_B(\omega) = P(\omega \mid B)$. There are two slightly different ways to discuss P_B. One way is to take B to be the probability space and define

$$P_B(\omega) = \frac{P(\omega)}{P(B)}$$

for all $\omega \in B$. Since B is the probability space for P_B, we do not define P_B for $\omega \notin B$. This P_B is a probability because $P_B(\omega) \geq 0$ for all $\omega \in B$ and $\sum_{\omega \in B} P_B(\omega) = 1$. A slightly different way to think of P_B is to keep Ω as the probability space and set the conditional probabilities to zero for $\omega \notin B$. If we know the event B happened, then the probability of an outcome not in B is zero. This version of conditional probability is:

$$P_B(\omega) = P(\{\omega\} \mid B) = \begin{cases} \frac{P(\omega)}{P(B)} & \text{for } \omega \in B, \\ 0 & \text{for } \omega \notin B. \end{cases} \tag{1.4}$$

Either way, we restrict to outcomes in B and *renormalize* the probabilities by dividing by $P(B)$ so that they again sum to one.

We can condition a second time by conditioning P_B on another event, C. Then $P_B(\omega \mid C)$ is the conditional probability of ω given that C occurred, given that B occurred. It seems natural that this should be be the P conditional probability of ω given that both B and C occurred. Bayes' rule verifies this intuition. If $\omega \in B$ and $\omega \in C$, then

$$\begin{aligned} P_B(\omega \mid C) &= \frac{P_B(\omega)}{P_B(C)} \\ &= \frac{P(\omega \mid B)}{P(C \mid B)} \\ &= \frac{P(\omega)}{P(B)} \cdot \frac{1}{\frac{P(C \cap B)}{P(B)}} \\ &= \frac{P(\omega)}{P(B \cap C)} \\ &= P_{B \cap C}(\omega) \, . \end{aligned}$$

You can check that these definitions give $P_B(\omega \mid C) = 0$ if $\omega \notin B \cap C$. The conclusion is the *tower property*, that conditioning on B and then on C is the same as conditioning just once on $B \cap C$. Recurrence relations based on the tower property are very important in stochastic calculus.

1.3 Algebras of sets, partial information

The study of random processes relies on the idea of states of partial information. At a given time one may have some information about ω but not yet know

everything about it. If you toss coins sequentially, there is a time when you know the results of the first two tosses but not the third or fourth. The concept of an *algebra* of sets helps formalize the idea of partial information.

An algebra of sets, $\mathcal{F}$, is a a collection of subsets of Ω. We interpret $A \in \mathcal{F}$ to mean that in our state of partial information we know whether $\omega \in A$ or not. For example, if the partial information is the outcomes of the first two tosses, then the event $A = \{\text{HHHH}, \text{HHHT}, \text{HHTH}, \text{HHTT}\}$ is in $\mathcal{F}$ because it is the event "first two tosses are H". The event $B =$"exactly one H", which is $\{\text{HTTT}, \text{THTT}, \text{TTHT}, \text{TTTH}\}$ is not in $\mathcal{F}$ because we may not know after just two tosses whether $\omega \in B$. It is possible that we do know (e.g. if $\omega \in A$), but it is possible that we do not.

The mathematical definition is as follows. The set of events $\mathcal{F}$ is an *algebra* if there is at least one $A \in \mathcal{F}$ and

i. $A \in \mathcal{F}$ implies that $A^c \in \mathcal{F}$.

ii. $A \in \mathcal{F}$ and $B \in \mathcal{F}$ implies that $A \cup B \in \mathcal{F}$ and $A \cap B \in \mathcal{F}$.

These are natural from the point of view of partial information. If we know whether A occurred then we also know whether "not A"$(= A^c)$ occurred. If we know whether A happened and whether B happened, then we can tell whether "A and B" happened. We definitely know whether $\emptyset = A \cap A^c$ happened (it did not) and whether $\Omega = A \cup A^c$ happened (it did). For historical reasons (see below), events in $\mathcal{F}$ are called *measurable* with respect to $\mathcal{F}$.

As another example, suppose we know only the results of the tosses but not the order, which might happen if we toss 4 identical coins at the same time. Some measurable sets are (with an abuse of notation)

$$\begin{aligned} \{4\} &= \{\text{HHHH}\} \\ \{3\} &= \{\text{HHHT, HHTH, HTHH, THHH}\} \\ &\vdots \\ \{0\} &= \{\text{TTTT}\} \end{aligned}$$

The event $\{2\}$ has 6 outcomes (list them), so its probability is $6 \cdot \dfrac{1}{16} = \dfrac{3}{8}$. There are other events measurable in this algebra, such as "fewer than 3 H" $= \{2\} \cup \{1\} \cup \{0\}$. The events $\{4\}, \ldots, \{0\}$ *generate* the algebra in the sense that any other event in the algebra is determined by some combination of these.

An algebra of sets is a $\sigma-$algebra (pronounced "sigma algebra") if it is *closed under countable unions*. The distinction between algebra and $\sigma-$algebra makes sense only if Ω is infinite. Suppose $A_n \in \mathcal{F}$ is an infinite sequence of events measurable in $\mathcal{F}$. Let $A = \cup_n A_n$ be the set of outcomes in any of the A_n. If $\mathcal{F}$ is a $\sigma-$algebra, then $A \in \mathcal{F}$, too. The reader can check that an algebra closed under countable unions is also closed under countable intersections, and conversely. An algebra is automatically a $\sigma-$algebra if Ω is finite. If Ω is infinite,

an algebra might or might not be a $\sigma-$algebra.[4] In a $\sigma-$algebra, it is possible to take limits of infinite sequences of events, just as it is possible to take limits of sequences of real numbers. We will never (again) refer to an algebra of events that is not a $\sigma-$algebra.

A *random variable* is a quantity whose value is random. If X is a random variable, a *probability model* of X says how X is determined by the outcome of some experiment. This is the same as specifying a probability space Ω, a probability function $P(\omega)$, and a function $X = X(\omega)$. For example, if Ω is the 16 element coin tossing space then $X(\omega)$ could be the number of heads in ω (so $X(HHHT) = 3$, $X(TTTT) = 0$, etc.). We sometimes call ω itself the random variable and $X(\omega)$ a function of a random variable. It is common to omit the argument ω in formulas involving X. It also is traditional to write random variables with capital letters and use the corresponding lower case letter for a value the random variable might have. For example, we write $\Pr(X = x) = \sum_{\omega | X(\omega)=x} P(\omega)$. The *cumulative distribution function* (or *CDF* or just *distribution function*) of X is $F(x) = \Pr(X \leq x)$.

We often describe events in words. For example, we might write $P(X \leq x)$ where, strictly, we might be supposed to say $A_x = \{\omega \mid X(\omega) \leq x\}$ then $P(X \leq x) = P(A_x)$. For example, if there are two random variables on the same probability space, X_1 and X_2, we might try to calculate the probability that they are equal, $P(X_1 = X_2)$. Strictly speaking, this is the probability of the set of ω so that $X_1(\omega) = X_2(\omega)$.

A random variable $X(\omega)$ is *measurable* with respect to the algebra $\mathcal{F}$ if the information in $\mathcal{F}$ is enough to determine the value of X. More precisely, for any number, x, we can consider the event that $X = x$, which is $B_x = \{\omega : X(\omega) = x\}$. In discrete probability, B_x will be the empty set for almost all x values and will not be empty only for those values of x actually taken by $X(\omega)$ for one of the outcomes ω. The function $X(\omega)$ is measurable with respect to $\mathcal{F}$ if the sets B_x are all measurable. It is common to write $X \in \mathcal{F}$ (an abuse of notation) to indicate that X is measurable with respect to $\mathcal{F}$. For example, suppose $\mathcal{F}$ is the algebra generated by the sets $\{0\}$, $\{1\}$, etc. above. The random variable $X =$ number of H minus number of T is measurable with respect to this algebra, but the function $X =$ number of T before the first H is not (find an x and $B_x \notin \mathcal{F}$ to show this).

There are several ways to describe a $\sigma-$algebra. One way is to give a family of generating events. Suppose there are events $A_1, \ldots, A_n$ that you know. The algebra *generated* by these sets is the one that expresses the information you have by virtue of knowing these events. For example, if you know whether A_1 happened and whether A_2 happened, then you also know A_1^c, and $A_1 \cap A_2$, etc. We denote by $\mathcal{F}$ the $\sigma-$algebra generated by the A_k. One definition of $\mathcal{F}$ is that it consists of all events A that are formed by finitely many or countably many set operations (intersection, union, complement) from the sets A_k.

[4] Let Ω be the set of integers and $A \in \mathcal{F}$ if A is finite or A^c is finite. This $\mathcal{F}$ is an algebra (check), but not a $\sigma-$algebra. For example, if A_n is the first n even integers (a finite set), then $A = \cup_n A_n$ is the set of all even integers. However A is not in $\mathcal{F}$ because neither A not A^c is finite.

An equivalent definition is that $\mathcal{F}$ is the smallest $\sigma-$algebra that contains all the events A_k. There is a smallest one because if $\mathcal{F}_1$ and $\mathcal{F}_2$ are two algebras, then $\mathcal{F}_1 \cap \mathcal{F}_2$ also is a $\sigma-$algebra, and is contained in both $\mathcal{F}_1$ and $\mathcal{F}_2$. The intersection of all $\sigma-$algebras that contain the A_k is the smallest one that contains the A_k. This is particularly convenient when defining the $\sigma-$algebra generated by an infinite family of events. It is the intersection of all $\sigma-$algebras that contain all the events.

A function $X(\omega)$ defines a $\sigma-$algebra of sets generated by the sets B_x defined above. This is the smallest $\sigma-$algebra, $\mathcal{F}_X$, so that X is measurable with respect to $\mathcal{F}$. Suppose $Y(\omega)$ is another random variable, then $Y \in \mathcal{F}_X$ if the value of Y is known once the value of X is known. That is the same as saying there is a function $y = u(x)$ so that $Y(\omega) = u(X(\omega))$ or all $\omega \in \Omega$.

As an example, let $\mathcal{F}$ be generated by the events $A_k = \{k\}$ above. Let $X(\omega) = k$ for all $\omega \in A_k$. Then $X(\omega)$ is the number of heads in ω and $\mathcal{F}$ is the algebra generated by N. Suppose $Y(\omega)$ is the number of heads minus the number of tails. Then $Y \in \mathcal{F}$ and $Y = 2X(\omega) - 4$. This means that $u(x) = 2x - 4$.

We can consider the algebra generated by a family of functions $X_1, X_2, \ldots$. This is the algebra generated by the algebras $\mathcal{F}_{X_1}$, $\mathcal{F}_{X_2}$, etc. It is possible to show that $Y \in \mathcal{F}$ if there is a function $y = u(x_1, x_2, \ldots)$ so that $Y(\omega) = u(X_1, X_2, \ldots)$. For example, let $X_k(\omega)$ be the result of the k-th toss. If Y is a random variable that is measurable with respect to the algebra generated by $\mathcal{F}_{X_1}$ and $\mathcal{F}_{X_2}$, then the value of Y is determined by the first two tosses.

Suppose we learn the values of random variables X_k one at a time, first $X_1(\omega)$, then $X_2(\omega)$, etc. After learning X_k, our state of knowledge is modeled by the $\sigma-$algebra $\mathcal{F}_k$ generated by $X_1, \ldots, X_k$. Now suppose Y_k is a decision that must be made using the the knowledge of $X_1, \ldots, X_k$ only. Then there is a function $u_k(x_1, \ldots x_k)$ so that $Y_k(\omega) = u_k(X_1(\omega), \ldots, X_k(\omega))$. This is what it means to say that $Y_k \in \mathcal{F}_k$. A sequence of $\sigma-$algebras $\mathcal{F}_k$ with $\mathcal{F}_k \subseteq \mathcal{F}_{k+1}$ is a *filtration*. A filtration may be *generated* by a sequence of random variables, but there are other ways to create filtrations. A family of random variables $Y_k \in \mathcal{F}_k$ is called *progressively measurable*, or *adapted*, or *non-anticipating* with respect to this filtration, though these terms have slightly different meanings in more technical situations.

A $\sigma-$algebra determines an *equivalence relation* and *partitions* the probability space into *equivalence classes*. Outcomes ω_1 and ω_2 are *distinguishable* in $\mathcal{F}$ if there is some event in $\mathcal{F}$ that contains ω_1 but not ω_2. If ω_1 and ω_2 are not distinguishable, they are called *equivalent*. We write $\omega_1 \sim \omega_2$ in that case. Put slightly differently, $\omega_1 \sim \omega_2$ if $\omega_1 \in A \Rightarrow \omega_2 \in A$ for every $A \in \mathcal{F}$. For example, in algebra that counts the number of heads, THTT $\sim$ TTTH. Since $\mathcal{F}$ is an algebra, $\omega_1 \sim \omega_2$ also implies that $\omega_1 \notin A \Rightarrow \omega_2 \notin A$ (think this through). It is easy to check that any ω has $\omega \sim \omega$ (similar to itself), and, for any three outcomes, $\omega_1 \sim \omega_2$ and $\omega_2 \sim \omega_3$ implies that $\omega_1 \sim \omega_3$ (transitivity). A relation that has these properties is an equivalence relation.

Any equivalence relation defines *equivalence classes*. The equivalence class of

ω, written[5] A_ω, is the set of outcomes ω' that are not distinguishable from ω in $\mathcal{F}$. This definition may be written as $A_\omega = \{\omega' \text{ with } \omega' \sim \omega\}$. In discrete probability, equivalence classes are measurable. (Proof: for any ω' not equivalent to ω in $\mathcal{F}$, there is at least one $B \in \mathcal{F}$ with $\omega \in B$ but $\omega' \notin B$. Choose any one of these and call it $B_{\omega'}$. The sets $B_{\omega'}$ form a finite or countable family because Ω is discrete. Moreover, $A_\omega = \cap_{\omega'} B_{\omega'}$, because any ω'' that is indistinguishable from ω in $\mathcal{F}$ is in all the events on the right, and any ω' that is distinguishable from ω is not in $B_{\omega'}$.) In the example above, the equivalence class of THTT is the event $A_{THTT} = \{\text{HTTT, THTT, TTHT, TTTH}\}$.

A *partition* of Ω is a collection of events, $\mathcal{P} = \{B_1, B_2, \ldots\}$ so that every outcome $\omega \in \Omega$ is in exactly one of the events B_k. For example, in the coin tossing example, knowing the first two tosses partitions the space of all outcomes into the four events $B_1 = \{\text{HH}\cdots\}$, $B_2 = \{\text{HT}\cdots\}$, $B_3 = \{\text{TH}\cdots\}$, and $B_4 = \{\text{TT}\cdots\}$. Every $\omega \in \Omega$ is in one of the B_k. For example, $\omega = \text{THHT} \in B_3$. No outcome is in more than one of the B_k. For another example, the events $A_k = \{k\}$ above are a partition of Ω into five events. More generally, $X(\omega)$ a random variable partitions Ω into the sets $B_x = \{\omega \mid X(\omega) = x\}$. This is the partition corresponding to the $\sigma-$ algebra generated by X.

The equivalence classes of an equivalence relation form a partition of Ω. Two equivalence classes either are identical or disjoint. To see this, suppose A_ω and $A_{\omega'}$ are two equivalence classes. If A_ω and $A_{\omega'}$ are not disjoint, then there is some ω'' with $\omega'' \in A_\omega$ and $\omega'' \in A_{\omega'}$, which is to say that $\omega'' \sim \omega$ and $\omega'' \sim \omega'$. If the equivalence relation is transitive and symmetric, this implies that $\omega \sim \omega'$. Now, if $\omega''' \in A_{\omega'}$ then $\omega''' \sim \omega' \sim \omega$, so $\omega''' \in A_\omega$. This shows that if A_ω and $A_{\omega'}$ have any elements in common, then every element in $A_{\omega'}$ also is in A_ω, and vice versa. Thus, if two equivalence classes are not disjoint, they are the same.

Partitions and $\sigma-$algebras are equivalent in discrete probability. The $\sigma-$algebra generated by $\mathcal{P}$, which we call $\mathcal{F}_\mathcal{P}$, consists of events that are unions of events in $\mathcal{P}$ (Why are complements and intersections not needed?). A $\sigma-$algebra, $\mathcal{F}$, determines a partition, $\mathcal{P}_\mathcal{F}$, described above. If you start with a partition, $\mathcal{P}$, then form the $\sigma-$algebra $\mathcal{F}_\mathcal{P}$, then the partition for $\mathcal{F}_\mathcal{P}$ is just $\mathcal{P}$. Similarly, if you start with $\mathcal{F}$ and form $\mathcal{P}_\mathcal{F}$, then the algebra corresponding to $\mathcal{P}_\mathcal{F}$ is $\mathcal{F}_{\mathcal{P}_\mathcal{F}} = \mathcal{F}$. We use both points of view because partitions are easier to work with in discrete probability while $\sigma-$algebras are much better in continuous probability.

1.4 Conditional expectation with partial information

A random variable $X(\omega)$ has expected value

$$E[X] = \sum_{\omega \in \Omega} X(\omega) P(\omega) \ .$$

[5] Other notations for the equivalence class of ω are $\overline{\omega}$ and (ω).

Note that we do not write ω on the left. We think of X as simply a random number and ω as a story telling how X was generated. This is the average value in the sense that if you could perform the experiment of sampling X many times then average the resulting numbers, you would get roughly $E[X]$. This is because $P(\omega)$ is the fraction of the time you would get ω and $X(\omega)$ is the number you get for ω. The operation of taking the expected value (or expectation) is linear. That means that if $X_1(\omega)$ and $X_2(\omega)$ are two random variables and c_1 and c_2 are constants, then $E[c_1X_1 + c_2X_2] = c_1E[X_1] + c_2E[X_2]$.

Conditional expectation is the expected value in a state of partial information. The classical conditional expectation with respect to an event B. The information is that $\omega \in B$. This conditional expectation is defined from conditional probability (1.4) in the natural way

$$E[X|B] = \sum_{\omega \in B} X(\omega)P(\omega|B) \; . \tag{1.5}$$

For example, we can calculate

$$E[\#\text{ of H in 4 tosses } \mid \text{ at least one H}] \; .$$

To get $\Pr(B)$ for the event $B = \{$ at least one H$\}$, note that only $\omega =$TTTT in not in B. Therefore $|B| = 15$ and $\Pr(B) = \frac{15}{16}$. That implies that

$$P(\omega \mid B) \;=\; \frac{\frac{1}{16}}{\frac{15}{16}} \;=\; \frac{1}{15}$$

for any $\omega \in B$. Let $X(\omega)$ be the number of H in ω. Unconditionally, $E[X] = 2$, which means

$$\frac{1}{16} \sum_{x \in \Omega} X(\omega) = 2 \; .$$

Note that $X(\omega) = 0$ for all $\omega \notin B$ (only TTTT), so

$$\sum_{\omega \in \Omega} X(\omega)P(\omega) = \sum_{\omega \in B} X(\omega)P(\omega) \; ,$$

and therefore

$$\begin{aligned}
\frac{1}{16} \sum_{\omega \in B} X(\omega)P(\omega) &= 2 \\
\frac{15}{16} \cdot \frac{1}{15} \sum_{\omega \in B} X(\omega)P(\omega) &= 2 \\
\frac{1}{15} \sum_{\omega \in B} X(\omega)P(\omega) &= \frac{2 \cdot 16}{15} \\
E[X \mid B] &= \frac{32}{15} \;=\; 2 + .133\ldots \; .
\end{aligned}$$

Knowing that there was at least one H increases the expected number of H by .133 . . ..

Suppose $\mathcal{P} = \{B_1, B_2, \ldots\}$ is a partition of Ω. The *law of total probability* is the formula

$$E[X] = \sum_k E[X \mid B_k] P(B_k) \; . \tag{1.6}$$

This is easy to understand: exactly one of the events B_k happens. The expected value of X is the sum over each of the events B_k of the expected value of X given that B_k happened, multiplied by the probability that B_k did happen. The mathematical statement of this argument is a combination of the definitions of conditional expectation (1.5) and conditional probability (1.4):

$$\begin{aligned} E[X] &= \sum_{\omega \in \Omega} X(\omega) P(\omega) \\ &= \sum_k \left(\sum_{\omega \in B_k} X(\omega) P(\omega) \right) \\ &= \sum_k \left(\sum_{\omega \in B_k} X(\omega) \frac{P(\omega)}{P(B_k)} \right) P(B_k) \\ &= \sum_k E[X \mid B_k] P(B_k) \; . \end{aligned}$$

This fact underlies the recurrence relations that are among the primary tools of stochastic calculus. It will be reformulated below as the *tower property* when we discuss the modern view of conditional probability.

Decision trees illustrate one of the uses of conditional probability. A *decision tree* is a model for a sequence of random choices, each dependent on the earlier ones. See Figure 1.1. Let X_1 be the first decision and X_2 the second. Then $X_1 = \text{H}$ or $X_1 = \text{T}$, and similarly for X_2. The numbers on the top row are probabilities, so $\Pr(X_1 = \text{H}) = \frac{2}{3}$ and $\Pr(X_1 = \text{T}) = \frac{1}{3}$. The numbers on the second row are conditional probabilities, conditioned on the value of X_1, so $\Pr(X_2 = \text{H} \mid X_1 = \text{H}) = \frac{3}{4}$, and $\Pr(X_2 = \text{H} \mid X_1 = \text{T}) = \frac{1}{2}$, etc. The probability on the first row and the conditional probability on the second row are combined using Bayes' rule (1.2) as the following example. Define events $B_{1\text{T}} = \{X_1 = \text{T}\}$ and $B_{2\text{H}} = \{X_2 = \text{H}\}$. Then $B_{1\text{T}} \cap B_{2\text{H}}$ is the single outcome $\omega = \text{TH}$. Since $\Pr(B_{1\text{T}}) = \frac{1}{3}$ and $\Pr(B_{2\text{H}} \text{ and } B_{1\text{T}}) = \frac{1}{2}$, we have

$$\Pr(\text{TH}) \;=\; \Pr(B_{2\text{H}} \cap B_{1\text{T}}) \;=\; \Pr(B_{2\text{H}} \mid B_{1\text{T}}) \cdot \Pr(B_{1\text{T}}) \;=\; \frac{1}{3} \cdot \frac{1}{2} \;=\; \frac{1}{6} \; .$$

To summarize, a decision tree is a model given by specifying conditional probabilities at each stage (after the first stage). Bayes' rule allows you to compute probabilities of sequences by multiplying conditional probabilities.

The decision tree also offers a good example of conditional expectation under partial information. The numbers at the bottom of Figure 1.1 are "payouts" for the corresponding sequences of decisions. For example, if $\omega = TH$, then

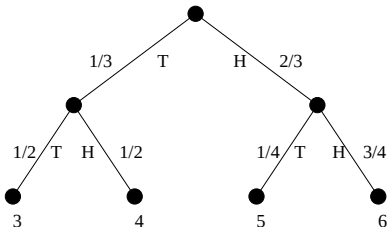

Figure 1.1: *A two stage decision tree. The probability that the first decision is H is* $\frac{2}{3}$*. If the first decision is H, the probability that the second one is T is* $\frac{1}{4}$*. The payout for decisions TH is* $V = 4$*. The payout for HH is* $V = 6$*, etc.*

$V(\omega) = 4$. For reasons that will become clear, we call the expected payout f_0. Its value is

$$\begin{aligned} f_0 \;=\; E[V] \;&=\; \sum_{\omega\in\Omega} V(\omega)P(\omega) \\ &=\; 3\cdot P(TT) + 4\cdot P(TH) + 5\cdot P(HT) + 6\cdot P(HH) \\ &=\; 3\cdot\frac{1}{3}\cdot\frac{1}{2} \;+\; 4\cdot\frac{1}{3}\cdot\frac{1}{2} \;+\; 5\cdot\frac{2}{3}\cdot\frac{1}{4} \;+\; 6\cdot\frac{2}{3}\cdot\frac{3}{4} \\ &=\; 5\,. \end{aligned}$$

Let $\mathcal{F}$ be the algebra generated by X_1, which is the first decision. This algebra is generated by the partition of Ω into two pieces, $B_{1\mathrm{T}} = \{TT, TH\}$, and $B_{1\mathrm{H}} = \{HT, HH\}$. The conditional expectations are

$$\begin{aligned} f_1(T) \;=\; E[V \mid B_{1\mathrm{T}}] \;&=\; 3\cdot P(TT \mid B_{1\mathrm{T}}) + 4\cdot P(TH \mid B_{1\mathrm{T}}) \\ &=\; 3\cdot\frac{1}{2} \;+\; 4\cdot\frac{1}{2} \\ &=\; 3.5\,, \\ f_1(H) \;=\; E[V \mid B_{1\mathrm{H}}] \;&=\; 5\cdot P(HT \mid B_{1\mathrm{H}}) + 6\cdot P(HH \mid B_{1\mathrm{H}}) \\ &=\; 5\cdot\frac{1}{4} \;+\; 6\cdot\frac{3}{4} \\ &=\; 5.75\,. \end{aligned}$$

The probabilities used in the f_1 calculations are the conditional probabilities given in in the second row of Figure 1.1. The law of total probability (1.6) gives the overall mean, f_0, as the mean of the conditional means:

$$\begin{aligned} f_0 \;&=\; f_1(T)\cdot P(B_{1\mathrm{H}}) + f_1(H)\cdot P(B_{1\mathrm{H}}) \\ 5 \;&=\; 3.5\cdot\frac{1}{3} + 5.75\cdot\frac{2}{3}\;. \end{aligned}$$

The previous paragraph gives two ways to calculate f_0. The direct method first calculates the probabilities of all four *paths* through the decision tree. The indirect method computes the conditional expectations f_1, then combines these

to get f_0. The indirect method has advantages over the direct method. For one thing, it uses less arithmetic. In this two level tree it was three rather than four multiplications, but for an n level tree it is $n(n+1)/2$ instead of 2^n, which is a big difference. For $n = 10$ it is 55 instead of 1024. Another advantage is the extra information in the *value function*, f.

A fancier *modern* conditional expectation handles the situation where the information is about more than a single event. If X is a random variable and $\mathcal{F}$ is a $\sigma-$algebra, then the conditional expectation with respect to $\mathcal{F}$ is written $Y = E[X|\mathcal{F}]$. It is a random variable whose value is determined by the information in $\mathcal{F}$. If $\mathcal{P}_\mathcal{F} = \{B_1, \ldots\}$ is the partition corresponding partition, then the information in $\mathcal{F}$ is which B_j ω falls into. The value of Y is the classical conditional expectation (1.5) of X for this B_j:

$$Y(\omega) = E[X \mid B_j] \text{ if } \omega \in B_j \ . \tag{1.7}$$

The modern definition (1.7) is consistent with the classical definition (1.5). A single set B defines a partition: $B_1 = B$, $B_2 = B^c$. In this case (1.7) gives $Y(\omega) = E[X \mid B]$ if $\omega \in B$, and $Y(\omega) = E[X \mid B^c]$ if $\omega \notin B$.

The modern point of view – conditional expectation as a function – makes particular sense when the partial information is the value of some function $Z(\omega)$. Since the partial information is the value of Z, it makes sense that the conditional expectation is a function of Z. If $f(z) = E[X \mid Z(\omega) = z]$, then $Y(\omega) = f(Z(\omega))$. In the decision tree example above, let Z be the first decision, so $Z(TT) = T$, $Z(HT) = H$, etc. Then the possible values are $z = T$ and $z = H$. The conditional expectation with respect to this information is the value function f_1 above.

Another example is given in Table 1.1. Take Ω to be sequences of 4 coin tosses. Take $Z(\omega)$ to be the number of H tosses and let $\mathcal{F}$ to be the $\sigma-$algebra generated by Z. This is generated by the partition $\mathcal{P} = \{\{0\}, \{1\}, \ldots\}$ corresponding to the values $z = 0$, $z = 1$, etc. Let $X(\omega)$ be the number of H tosses before the first T (e.g. $X(\text{HHTH}) = 2$, $X(\text{TTTT}) = 0$, $X(\text{HHHH}) = 4$, etc.). We calculate (table below): $f(0) = Y(\{0\}) = 0$, $f(1) = Y(\{1\}) = 1/4$, $f(2) = Y(\{2\}) = 2/3$, $f(3) = Y(\{3\}) = 3/2$, and $f(4) = Y(\{4\}) = 4$. Note, for example, that HHTT and HTHT are equivalent in $\mathcal{F}$, both in the equivalence class B_2. This is because the information in $\mathcal{F}$ does not distinguish HHTT from HTHT. Therefore $Y(\text{HHTT}) = Y(\text{HTHT})$, even though $X(\text{HHTT}) \neq X(\text{HTHT})$. The common value $Y(\text{HHTT}) = Y(\text{HTHT}) = \frac{1}{4}$) is the average value of X over the outcomes in the equivalence class B_2.

Partitions and $\sigma-$algebras give a way to express the idea of adding new information. We say that $\sigma-$algebra $\mathcal{F}$ is an *extension* of $\sigma-$algebra $\mathcal{G}$ if $\mathcal{G} \subseteq \mathcal{F}$, which is to say that $B \in \mathcal{G} \Longrightarrow B \in \mathcal{F}$. This means that if the information in in $\mathcal{G}$ determines whether B occurred, then the information in $\mathcal{F}$ also determines B. Let $\mathcal{P}_\mathcal{F}$ and $\mathcal{P}_\mathcal{G}$ be the corresponding partitions of Ω. Then $\mathcal{P}_\mathcal{F}$ is a *refinement* of $\mathcal{P}_\mathcal{G}$ if every $C \in \mathcal{P}_\mathcal{G}$ is a union of elements of $\mathcal{P}_\mathcal{F}$, written

$$C = B_{1_C} \cup \cdots \cup B_{m_C} \ , \ \text{with } B_{j_C} \in \mathcal{P}_\mathcal{F}.$$

$z = 0, B_0 = \{0\}$	ω: TTTT; $X(\omega)$: 0; expected value: 0
$z = 1, B_1 = \{1\}$	ω: HTTT, THTT, TTHT, TTTH; $X(\omega)$: 1, 0, 0, 0; expected value: $(1+0+0+0)/4 = 1/4$
$z = 2, B_2 = \{2\}$	ω: HHTT, HTHT, HTTH, THHT, THTH, TTHH; $X(\omega)$: 2, 1, 1, 0, 0, 0; expected value: $(2+1+1+0+0+0)/6 = 2/3$
$z = 3, B_3 = \{3\}$	ω: HHHT, HHTH, HTHH, THHH; $X(\omega)$: 3, 2, 1, 0; expected value: $(3+2+1+0)/4 = 3/2$
$z = 4, B_4 = \{4\}$	ω: HHHH; $X(\omega)$: 4; expected value: 4

Table 1.1: Computing the conditional expectation of X = the number of H before a T, conditioned on the number of H tosses in all.

This is the same as saying that if $C \in \mathcal{P}_\mathcal{G}$ and $B \in \mathcal{P}_\mathcal{F}$, then either $B \subseteq C$ or B is disjoint from C. If $\mathcal{P}_\mathcal{F}$ refines $\mathcal{P}_\mathcal{G}$, and $C \in \mathcal{P}_\mathcal{G}$, then the $B_{jC} \in \mathcal{P}_\mathcal{F}$ with $B_{jC} \subseteq C$ form a partition of C. The information in $\mathcal{G}$ divides ω into the sets $C \in \mathcal{P}_\mathcal{G}$. The extra information in $\mathcal{F}$ subdivides each $C \in \mathcal{P}_\mathcal{G}$ into the B_{jC}. The reader should take the time to verify that these notions are equivalent: $\mathcal{F}$ extends $\mathcal{G}$ if and only if $\mathcal{P}_\mathcal{F}$ refines $\mathcal{P}_\mathcal{G}$.

The *tower property* is the fact that that if $\mathcal{G} \subseteq \mathcal{F}$, and $Y_\mathcal{F} = E[X|\mathcal{F}]$, and $Y_\mathcal{G} = E[X|\mathcal{G}]$, then $Y_\mathcal{G} = E[Y_\mathcal{F}|\mathcal{G}]$. To show this, let $Z = E[Y_\mathcal{F}|\mathcal{G}]$. We must show that for any ω, $Z(\omega) = Y_\mathcal{G}(\omega)$. Let $C \in \mathcal{P}_\mathcal{G}$ be the $\mathcal{G}$ partition member that ω is a member of. The tower property is the law of total probability (1.6) applied to the conditional probability P_C. Indeed, the events B_{jC} above are a partition of C. Moreover, for $\omega \in C$,

$$\begin{aligned} Z &= E_{P_C}[Y_\mathcal{F}] \\ &= E_{P_C}[X \mid B_{1C}]\, P_C(B_{1C}) + \cdots . \end{aligned}$$

Exercise 2 asks the reader to complete the argument.

Recall that a random variable $X(\omega)$ is one way to specify a $\sigma-$algebra. Let $\mathcal{F}$ be this $\sigma-$algebra. The corresponding partition consists of the sets $B_x = \{\omega \mid X(\omega) = x\}$. Let Y be another random variable. The conditional expectation $Y_\mathcal{F} = E[X \mid \mathcal{F}]$ is determined by the function (described in several ways, the last being most common and most informal)

$$f(x) \;=\; E[\,Y \mid X = x\,] \;=\; E_{B_x}[\,Y\,] \;=\; E_x[\,Y\,] \;. \tag{1.8}$$

The formula is $Y_\mathcal{F} = f(X)$. This is convenient, but it may take some unwinding to understand fully. Both sides are random variables, which makes them functions of ω. The random variables are equal in the sense that for every ω, $Y_\mathcal{F}(\omega)) = f(X(\omega))$. It is easy to see that the definition (1.8) is just this statement, if x is the value of $X(\omega)$. In the example of Table 1.1, the values are $f(0) = 0$, $f(1) = \frac{1}{4}$, $f(2) = \frac{2}{3}$, $f(3) = \frac{3}{2}$, and $f(4) = 4$.

1.5 Simulation and Monte Carlo

Simulating a process means making samples that follow the probability distribution of the process. *Monte Carlo* is a computational technique that uses random samples to evaluate probabilities or expected values or to solve other problems. The most direct Monte Carlo method is to simulate a large number of samples and then take the empirical mean from the simulation. There are much more powerful Monte Carlo methods, but these are not the primary focus here.

The basis of most simulation and Monte Carlo is a *random number generator*. The ideal random number generator would produce a sequence of random variables U_k uniformly distributed in $[0, 1]$ and independent. In Java, the line of code `U = rand.nextDouble();` executed n times would produce n independent uniforms. The actual procedure ("method" in Java) is a *pseudo random number* generator. It is deterministic algorithm that returns numbers that look random

```
import java.util.Random;  // Put this at the top of the file before
                          // the executable code.


//     Put this at the top of the code, execute it only once.

Random rand;          // Declare the symbol "rand" to be a random
                      // number generator.
rand = new Random();  // Initialize this random number generator
double   U;           // A double precision variable, to be random.

        .

//      Use random numbers in the body of the code.

  for ( . . . ) {      // Start of the Monte Carlo loop
        ...
     U = rand.nextDouble();  //  These are the independent, uniform
                             //  [0,1] samples.
        ...
    }                 // End of the Monte Carlo loop
```

Figure 1.2: How to use the native uniform pseudo random number generator in Java.

in many respects. Good (pseudo) random number generators produce numbers that work as random numbers for most simulation and Monte Carlo purposes.

As we just said, a pseudo random number generator is a deterministic algorithm. It has two parts. One updates an internal variable called the *seed*. The other uses the seed to produce a number in the interval $[0, 1]$. Executing `U = rand.nextDouble();` is essentially equivalent to `seed = seedUpdate(seed);` then `U = getUnif(seed);`. Successive calls to `rand.nextDouble()` produce different results because the seen has changed. The actual `seed` typically is a data structure consisting of one or more integers. In the Java implementation of Figure 1.2, the seed is stored in (is a *member of*) the `rand` instance of the class `Random`. The line `rand = new Random();`, among other things, gives the seed an initial value.

Simulation or Monte Carlo computations can be repeatable or not depending on how the seed is set. The version in Figure 1.2 uses the current time as the initial seed. This means that successive runs have different initial seeds and therefore different results. If the initialization is changed to

```
rand = new Random(1234L);  // Initialize  with a fixed seed
```

then the results will be the same each time. This can be helpful, for example, in looking for bugs.

In simulation and Monte Carlo, all random events are built from standard uniforms (independent random variables uniformly distributed in the unit interval) in some way. For example, suppose X is the result of a single coin toss, with $P(X = \mathrm{H}) = p$. If U is a standard uniform, then $p = P(U < p)$, so we can make X like this:

```
U = rand.nextDouble();
if ( U < p )  X = H;        // This has probability p
else          X = T;        // This has probability 1-p
```

While simulation just means producing samples of some kind of random object, Monte Carlo means at the end of the day giving approximations to numbers that themselves are not random. The simplest is the expected value of a random variable, $A = E[X]$. If we can make independent samples of X, an estimator of the mean is

$$\widehat{A}_n \;=\; \frac{1}{n}\sum_{k=1}^{n} X_k \;. \tag{1.9}$$

Statisticians often use a hat over a quantity to denote a statistical estimate of it. So $\widehat{A}$ is an estimate of A. The estimation error is $\widehat{A}_n - A$. The *weak law* of large numbers states that the probability of making an error of size ϵ goes to zero as the simple size goes to infinity. More precisely, if $E\left[|X|\right] < \infty$, then for any $\epsilon > 0$

$$\lim_{n\to\infty} P\left(\left|\widehat{A}_n - A\right| > \epsilon\right) \;\to\; 0 \quad \text{as} \;\; n \to \infty \;. \tag{1.10}$$

What is random in (1.10) is $\widehat{A}_n$; A is just a number.

The convergence theorem (1.10) never guarantees that $\left|\widehat{A}_n - A\right| \leq \epsilon$, even though the probability of this is very small. It always is possible that Monte Carlo estimates have large errors, even if large errors are very unlikely. If you do a large number of Monte Carlo estimates, even if most of them are accurate, it becomes likely that some are inaccurate.

A practical person hopes for a convergence statement more quantitative than (1.10). Is n large enough so that the error is likely to be below a desired level? Approximately what is $P\left(\left|\widehat{A}_n - A\right| > \epsilon\right)$. If $\mathrm{var}(X) < \infty$ the *central limit theorem* applies and says that $\widehat{A}_n$ is approximately Gaussian with mean A and standard deviation

$$\sigma_{\widehat{A}} \;=\; \frac{1}{\sqrt{n}}\sigma_X \;. \tag{1.11}$$

It is standard Monte Carlo practice – something every Monte Carlo practitioner and every student should do – to estimate σ_X using

$$\widehat{\sigma_X^2} \;=\; \frac{1}{n}\sum_{k=1}^{n}\left(X_k - \widehat{A}_n\right)^2 \;. \tag{1.12}$$

The estimator of the standard deviation is

$$\widehat{\sigma_X} \;=\; \sqrt{\frac{1}{n}\sum_{k=1}^{n}\left(X_k - \widehat{A}_n\right)^2}\;.$$

Some people prefer to use $n-1$ instead of n in (1.12) because it makes the result unbiased. If this matters, you don't have enough data. Moreover, the relation between σ^2 and σ is nonlinear. The estimator of σ will have a small bias from this nonlinearity even if the estimate of σ^2 is unbiased.

The *cumulative normal* is $N(x) = P(Z < x)$ when Z is a standard normal random variable:

$$N(x) \;=\; \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-z^2/2}\,dz\;.$$

Since we have to include two *tails*, the central limit theorem implies that if $Y \sim \mathcal{N}(\mu, \sigma_2)$ (i.e. Y is Gaussian with mean μ and variance σ^2) then

$$P\left(|Y-\mu| \;>\; k\sigma\right) \;=\; 1 \;-\; 2N(k)\;.$$

The values of N are well known. The probability of a *one sigma* error ($k = 1$) is about 30%. A two sigma event ($k = 2$) has about 5% probability, and three sigma is very unlikely. The central limit theorem makes these statements approximately true of the Monte Carlo estimation error

$$P\left(\left|\widehat{A}_n - A\right| > k\frac{\widehat{\sigma_X}}{\sqrt{n}}\right) \;\approx\; 1 - 2N(k)\;.$$

The quantity $\frac{\widehat{\sigma_X}}{\sqrt{n}}$ often is called the Monte Carlo *error bar*, because it is indicated as a bar around $\widehat{A}_n$ on a graph. If $\widehat{A}_n = 3.15$ and $\frac{\widehat{\sigma_X}}{\sqrt{n}} = .01$, we might write $A = 3.15 \pm .01$. This means there is about a 66% chance that $3.14 < A < 3.16$ and about a 95% chance that $3.13 < A < 3.17$, etc.

You report error bars in different ways depending on who is looking at them. You report one sigma error bars (as above) when the person has the scientific training to understand what they mean. You can report two or even three sigma error bars if the consumer appreciates the need for error bars but does not know about gaussians.

This is not to say that you believe strongly that the error is within the error bar – it has about at 30% chance of being outside. Rather you trust the consumer of your information to understand the meaning of a one standard deviation error bar and to double it (for example) for 95% conficence. You are free to report a two or even three standard deviation error bar if you do not trust the consumer of your information to understand the central limit theorem. If you are in a business setting and cannot report error bars to a customer, you should make sure to look at the error bars first to make sure they are small enough for your purpose.

1.6 Exercises and examples

1. Verify the law of total probability (1.6) in the example of the number algebra. Letting X be the number of tails before a head, calculate $E[X]$ directly using

$$\begin{aligned} P(X=0) &= P\left(1^{st} \text{ toss} = \text{H}\right) = \frac{1}{2} \\ P(X=1) &= P\left(1^{st} \text{ toss} = \text{T and } 2^{nd} \text{ toss} = \text{H}\right) = \frac{1}{2}\cdot\frac{1}{2} \\ &\text{etc.} \end{aligned}$$

   Check that this gives the same answer as $E[X|B_0]\cdot P(B_0) + \cdots$.

2. Do the algebra to show that the tower property is a consequence of the law of total probability in the conditional probability distribution P_C.

Chapter 2

Markov Chains

Markov chains are a simple class of models of dynamics with randomness. Many specific useful models take the form of Markov chains. They also provide a simple setting in which to understand one of the most useful ideas in stochastic calculus – recursive calculation of probabilities and expectation values.

2.1 Discrete time stochastic processes

There is an abstract view of random dynamical systems in terms of successive revealing of information. Let us suppose there is a time variable t that takes discrete values $t = 0, 1, \ldots$. The information available at time t is described by the $\sigma-$algebra $\mathcal{F}_t$. It is natural to suppose that $\mathcal{F}_{t+1}$ is an extension of $\mathcal{F}_t$. This just says that all the information available at time t is still available at time $t+1$. Such an expanding family of $\sigma-$algebras, $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots$, is a *filtration*.

Let $\mathcal{S}$, the *state space*, be a discrete (finite or countable) set. This set is a list of the possible states of some system. The actual state of the system changes with time in some random way. Let X_t be the state of the system at time t. The random sequence X_t is a *random process*, or *stochastic process*. It is natural to suppose that X_t is known at time t. This means that X_t is measurable in $\mathcal{F}_t$, and is written $X_t \in \mathcal{F}_t$, though that is not literally true as X_t is not an event. If $X_t \in \mathcal{F}_t$ for all t, then the random process X_t is *adapted* to, or *non-anticipating*, or *progressively measurable*[1] with respect to the filtration $\mathcal{F}_t$.

This definition of stochastic process, a sequence measurable with respect to a filtration, may seem hopelessly abstract. But there is a benefit to abstraction – it strips away all unnecessary detail and leaves only what is absolutely necessary to understand the situation. In this respect is is something like abstract art. Figure 2.1 shows an abstract painting *Night Creatures*. All the irrelevant details are gone and the important feelings are left clearly exposed.

A sequence of states X_0, X_1, ..., is a *path* in $\mathcal{S}$. If $a < T$, we write $X_{[a,T]}$ for the sequence between times $t = a$ and $t = T$. That is, $X_{[a:T]} = (X_a, X_{a+1}, \cdots, X_T)$. The set of all such paths is the *path space*, denoted $\mathcal{P}$. If it is important to emphasize the start and end times, we write $\mathcal{P}_{[a:T]}$. The state at time t is the value of the path $X_{[0:T]}$ at time t.

There is a *minimal* probability space and filtration for the stochastic process X_t up to time T. The probability space Ω is the path space $\mathcal{P}_{[0:T]}$. An element $\omega \in \Omega$ is a path $X_{[0:T]} \in \mathcal{P}_{[0:T]}$. The state X_t is a function of the path in the obvious but possibly confusing way: $X_t\left(X_{[0:T]}\right) = X_t$. The filtration is the one in which $\mathcal{F}_t$ is generated by $X_0, \ldots, X_t$. At time t, we know the values X_s for $s \leq t$ but not the future states X_{t+1}, etc. If you are only interested in the process X_t it probably is wise to work with this minimal space and filtration. But other situations arise. For example, Ω might describe two stochastic processes in two state spaces $X_t \in \mathcal{S}$ and $Y_t \in \mathcal{S}'$. Or Ω might describe a stochastic process depending on a random and unknown parameter so that even after the whole path $X_{[0:T]}$ is known, the parameter still has some uncertainty. This is

[1] These three terms mean the same thing here. In more sophisticated settings, continuous time and state space, their meanings may differ from each other.

Figure 2.1: *An abstract painting,* NightCreatures *by Lee Krasner. This painting was made at about the same time the abstract definitions of filtration and stochastic process were given.*

a *Bayesian* model. Finally, it is conceivable that at time t you already know X_{t+1}. If $\mathcal{F}_t$ is generated by $X_{[0:t+1]}$, then X_t is measurable in $\mathcal{F}_t$.

A one dimensional random walk illustrates these definitions. The state space is the positive and negative integers: $\mathcal{S} = \mathbb{Z}$. An outcome $\omega \in \Omega$ is a sequence of integers.

The writer Carl Jung reports a story that fits with this view of dynamics. That view (supposedly held by a tribe somewhere in east Africa) compares a person moving through time to a person riding backwards on the back of a cart. The future is that part of the landscape in front of the cart but behind the rider's back that she cannot see. As the cart rolls forward, more of the world comes into view. A person going through life is the same. What is behind the cart is the part she can see. This is the past. She cannot see the future. In the same way, you can think of the path $X_{[0:T]}$ as given from the beginning, but you learn about it one step at a time.

2.2 Markov chains

A *Markov chain* is a probability distribution P, on a path space $\Omega = \mathcal{P}_{[0:T]}$, that has the *Markov property.* Informally, P has the Markov property if the present is all the information about the past that is relevant for predicting the future. To say this mathematically, suppose $x_{[0:T]} = (x_0, \ldots, x_T)$ is a fixed path of length T. Let $t \leq T$ be some earlier time and consider the initial part of the

path $x_{[0:t]} = (x_0, \ldots, x_t)$. If P satisfies the Markov property, then

$$\Pr\left(X_{t+1} = x_{t+1} \mid X_{[0:t]} = x_{[0:t]}\right) \;=\; \Pr\left(X_{t+1} = x_{t+1} \mid X_t = x_t\right) \;. \tag{2.1}$$

We say that P has the Markov property if this is true for every path $x_{[0:T]} \in \mathcal{P}_{[0:T]}$ and every t with $0 \le t \le T$.

A Markov chain is *stationary* if the transition probabilities

$$P_{xy} \;=\; \Pr\left(X_{t+1} = y \mid X_t = x\right) \tag{2.2}$$

are independent of t. One usually assumes a Markov chain is stationary unless someone explicitly says otherwise. The order in (2.2) is important: P_{xy} is the probability of an $x \to y$ transition, not a $y \to x$ transition.

To appreciate the difference between Markovian and non-Markovian, consider first a simple linear process with state space $\mathcal{S} = \mathbb{R}$ (the state at time t is a real number). This state space is not discrete, but it will illustrate the ideas. Suppose that $X_0 = 0$ and that for $t > 0$,

$$X_{t+1} \;=\; aX_t \,+\, bZ_t \;, \tag{2.3}$$

where the $Z_t \sim \mathcal{N}(0,1)$ are independent standard normal random variables. Linear Gaussian models like this are common in time series modeling. If $X_t = x_t$ is known, then the remaining uncertainty in X_{t+1} comes from bZ_t, which has mean zero and variance b^2. Therefore both sides of (2.1) are given by the same expression

$$X_{t+t} \;\sim\; \mathcal{N}(ax_t, b^2) \;.$$

Knowing the value of X_{t-1} does not help us forecast X_{t+1} if X_t is known.

Contrast this with a more general linear model

$$Y_{t+1} \;=\; a_0 Y_t + a_1 Y_{t-1} \,+\, bZ_t \;. \tag{2.4}$$

In this case, the conditional distribution of Y_{t+1} is different depending on whether we know only Y_t or both Y_t and Y_{t-1}. If we define the "state" at time t to be all the information about the past that is helpful for forecasting the future, this would be[2] $X_t = (Y_t, Y_{t-1})'$. Now the state space is $\mathcal{S} = \mathbb{R}^2$ for the two components of X_t. The recurrence relation (2.4) may be reformulated as

$$\begin{pmatrix} Y_{t+1} \\ Y_t \end{pmatrix} \;=\; \begin{pmatrix} a_0 & a_1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} Y_t \\ Y_{t-1} \end{pmatrix} \,+\, \begin{pmatrix} b \\ 0 \end{pmatrix} Z_t \;.$$

This is the same as the matrix equation

$$X_{t+1} \;=\; AX_t \,+\, BZ_t \;, \tag{2.5}$$

where

$$A \;=\; \begin{pmatrix} a_0 & a_1 \\ 1 & 0 \end{pmatrix} \quad , \quad B \;=\; \begin{pmatrix} b \\ 0 \end{pmatrix} \;.$$

[2] We write $(u, v)'$ for the column vector with components u and v. It is convenient to use column vectors when we are about to multiply by a matrix. Do not be confused by the conflict of notation – the superscript $'$ is for "transpose" while the subscript t is for "time".

This example illustrates a general point of view. You formulate a general random process as a Markov chain by enlarging the state space to include everything about the past that matters for predicting the future.

2.3 Transition probabilities

A Markov chain is largely described by the probabilities on the right side of (2.1). Suppose the state space if finite: $\mathcal{S} = \{s_1, \ldots, s_n\}$. Then the right side of (2.1) consists of the n^2 numbers[3]

$$p_{ij} \;=\; \Pr\left(s_i \to s_j\right) \;=\; \Pr\left(X_{t+1} = s_j \mid X_t = s_i\right) \;. \tag{2.6}$$

We often "abuse notation" by conflating i with s_i and writing, for example, $p_{ij} = \Pr(i \to j)$ instead of the slightly more correct (2.6). The p_{ij} are the *transition probabilities* corresponding to the Markov chain. The *transition matrix*, P, is the $n \times n$ matrix whose (i, j) entry is p_{ij}. The transition matrix may be called the *generator* of the Markov chain.

The transition matrix, together with the starting probabilities

$$u_{0,i} \;=\; \Pr\left(X_0 = i\right) \;, \tag{2.7}$$

determines the probability of any path $x_{[0:T]} \in \mathcal{P}_{[0:T]}$. It is just (2.6), the Markov property, and Bayes' rule. For example,

$$\begin{aligned} \Pr\left(x_{[0:1]}\right) &= \Pr\left(X_0 = x_0 \text{ and } X_1 = x_1\right) \\ &= \Pr\left(X_0 = x_0\right) \cdot \Pr\left(X_1 = x_1 \mid X_0 = x_0\right) \\ &= u_{0,x_0} p_{x_0 x_1} \;. \end{aligned}$$

You have to use the Markov property to get the formula for two or more transitions. For example:

$$\begin{aligned} \Pr\left(x_{[0:2]}\right) &= \Pr\left(X_0 = x_0 \text{ and } X_1 = x_1 \text{ and } X_2 = x_2\right) \\ &= \Pr\left(\ (X_0 = x_0 \text{ and } X_1 = x_1)\ \text{ and } X_2 = x_2\right) \\ &= \Pr\left(X_0 = x_0 \text{ and } X_1 = x_1\right) \cdot \Pr\left(X_2 = x_2 \mid X_0 = x_0 \text{ and } X_1 = x_1\right) \\ &= \Pr\left(X_0 = x_0 \text{ and } X_1 = x_1\right) \cdot \Pr\left(X_2 = x_2 \mid X_1 = x_1\right) \\ &= u_{0,x_0} p_{x_0 x_1} p_{x_1 x_2} \;. \end{aligned}$$

Continuing in this way, the general formula clearly involves the starting probabilities and the product of the transition probabilities:

$$\Pr\left(x_{[0:T]}\right) \;=\; u_{0,x_0} \prod_{t=0}^{T-1} p_{x_t x_{t+1}} \;. \tag{2.8}$$

[3]The letter P is overworked here. To reduce confusion I write generic probabilities as $\Pr(\cdot)$ leaving P to be a transition matrix.

This is how the overall probability distribution on $\mathcal{P}_{[0:T]}$ is determined by the starting probabilities and the transition matrix.

The transition matrix has two simple mathematical properties. First, all the elements are non-negative: $p_{ij} \geq 0$ for all i and j. Second,

$$\sum_{j=1}^{n} p_{ij} \;=\; \sum_{s \in \mathcal{S}} \Pr\left(s_i \to s\right) \;=\; 1 \; . \tag{2.9}$$

The sum is equal to one because it is over all the states that s_i possibly could transition to. Of course, the term corresponding to no transition, $s_i \to s_i$, must be included in the sum. The sum in (2.9) is over all the elements in row i of P. If P is a transition matrix of a Markov chain, its elements are non-negative and all its row sums are equal to one. A square matrix with these properties is a *stochastic* matrix. The formula (2.8) shows that any stochastic matrix defines a Markov chain.

Matrices are meant to be multiplied, a statement that applies particularly to the transition matrix of a Markov chain. Let p_{ij}^k be the elements of P^k. These have the probabilistic interpretation

$$p_{ij}^k \;=\; \Pr\left(X_{t+k} = j \mid X_t = i\right) \;=\; \Pr\left(i \to j \text{ in } k \text{ steps }\right) \; . \tag{2.10}$$

We start by proving this for the case of $k = 2$ steps. Conditional on $X_t = i$, the $i \to j$ transition event $A_j = \{X_2 = j\}$ is partitioned into n events depending on the intermediate state, l: $B_{lj} = \{X_{t+1} = l \text{ and } X_{t+2} = j\}$. Therefore

$$\begin{aligned}
\Pr\left(X_{t+2} = j \mid X_t = i\right) &= \sum_{l=1}^{n} \Pr\left(X_{t+1} = l \text{ and } X_{t+2} = j \mid X_t = i\right) \\
&= \sum_{l=1}^{n} \Pr\left(X_{t+1} = l \mid X_t = i\right) \Pr\left(X_{t+2} = j \mid X_{t+1} = l\right) \\
&= \sum_{l=1}^{n} p_{il} p_{lj}
\end{aligned}$$

You recognize the last line as the (i, j) entry of P^2. Continuing like this you can see that

$$\begin{aligned}
\Pr\left(i \to j \text{ in three steps }\right) &= \sum_{l=1}^{n} \Pr\left(i \to l \text{ in two steps }\right) \cdot \Pr\left(l \to j \text{ in one step }\right) \\
&= \sum_{l=1}^{n} p_{il}^2 \cdot p_{lj} \;=\; p_{ij}^3 \; .
\end{aligned}$$

The formula for larger k follows by induction.

Urn processes are a family of simple Markov chains used as examples. The simplest urn process involves a single urn (container) with m balls in it. The balls are either red or blue but are otherwise indistinguishable. At each stage,

you choose a ball at random (all balls equally likely to be chosen) from the urn and replace it with a new ball that has probability p to be blue and probability $1-p$ to be red. All choices are independent. The state at time t depends on the number of blue balls, which we call x. The size of the state space is $n = m+1$ corresponding to the possible states $x = 0, 1, \ldots, m$.

Exercise 1 asks you to calculate the transition matrix elements p_{ij}. Most of the p_{ij} are zero since only transitions $i \to i-1$, $i \to i$ and $i \to i+1$ have non-zero probability. The matrix P is *tridiagonal* because the non-zeros are on the diagonal (p_{ii}, the *subdiagonal* ($p_{i,i-1}$), and the *superdiagonal* ($p_{i,i+1}$). Higher powers of P have more non-zeros. For example, $p^2_{i,i+2} > 0$ because it is possible to add two blue balls in two steps. You can check that $p^m_{ij} > 0$ for all i and j because it is possible to go from any number of blue balls to any other in m steps. Note however that some of the probabilities may be quite small. It may be extremely unlikely, for example, to go from $i = 0$ blue balls to $i = m$ by m, which requires m successive choices to add a blue ball.

A simpler example for calculation is *simple random walk*. In this process, X_t is any integer, positive or negative: $\mathcal{S} = \mathbb{Z}$. At each step, the walker goes to the left with probability α, stays in place with probability β, and goes to the right with probability γ. As with the urn process, the walker can go at most one unit in one time step, so $\alpha + \beta + \gamma = 1$. The parameters defining the random walk are determined by the *one step drift* and the *one step variance*. Both of these depend on the step at time t, which is $\Delta X_t = X_{t+1} - X_t$. The one step drift is just the expected value of the step:

$$a \;=\; E\left[\Delta X_t\right] \;. \tag{2.11}$$

The one step variance is the variance of the step

$$b^2 \;=\; \mathrm{var}\left(\Delta X_t\right) \;=\; E\left[\Delta X_t^2\right] \;-\; E\left[\Delta X_t\right]^2 \;. \tag{2.12}$$

The notation a, b, and ΔX_t will be used later in discussing diffusions and stochastic differential equations. The random walk is "simple" in that the statistics of ΔX_t do not depend on t (a given for time homogeneous Markov chains) or on X_t.

2.4 Dynamics of probabilities

The transition matrix describes how the probabilities of states change in time. Let $u_{i,t} = \Pr\left(X_t = i\right)$. There is a simple formula for the $u_{i,t+1}$ in terms of the $u_{i,t}$ and the transition probabilities.

$$\begin{aligned} u_{i,t+1} &= \Pr\left(X_{t+1} = i\right) \\ &= \sum_{j=1}^{n} \Pr\left(X_t = j\right) \cdot \Pr\left(X_{t+1} = i \mid X_t = j\right) \\ u_{i,t+1} &= \sum_{j=1}^{n} u_{j,t} p_{ji} \;, \end{aligned} \tag{2.13}$$

where $n = |\mathcal{S}|$ is the size of the state space. This is the *forward equation* for the evolution of the probabilities $u_{i,t}$. It also is called the*Kolmogorov* forward equation or the *Chapman Kolmogorov* equation. It is called "forward" because it carries the probabilities forward through time. We will come to the "backward" equation soon.

The forward equation is conveniently expressed in matrix/vector notation. The $u_{i,t}$ form the components of a *row vector* $u_t = (u_{1,t}, \ldots, u_{n,t})$. The equations (2.13) are equivalent to the matrix equation

$$u_{t+1} \;=\; u_t P \;. \tag{2.14}$$

The right side of (2.14) makes sense as the product of a $1 \times n$ matrix (i.e. a row vector) with the $n \times n$ transition marix. According to the rules of matrix multiplication, the result is another $1 \times n$ vector. It may take time to get used to putting the vector on the left of the matrix, but that is what people have come to do in this situation.

If you use two step transition probabilities, the argument that led to (2.13) gives

$$u_{i,t+2} \;=\; \sum_{j=1}^{n} u_{j,t} \cdot p_{ji}^2 \;.$$

In matrix form, this is $u_{t+2} = u_t P^2$. This is consistent with (2.14), as

$$u_{t+2} \;=\; u_{t+1} P \;=\; (u_t P)\, P \;=\; u_t\, (PP) \;=\; u_t P^2 \;.$$

The same applies for $k \geq 2$ step transition probabilities:

$$u_{t+k} \;=\; u_t P^k \;. \tag{2.15}$$

Indeed, the argument given for (2.13) is essentially the same as the argument for (2.10) above.

2.5 The value function

A *value function* is a kind of conditional expectation related to a Markov chain. Value functions are important for many reasons. One is that many important quantities of interest can be formulated as value functions. Another is that value functions satisfy recurrence relations that make them easy to calculate, if the state space is not too large.

The simplest value function is defined as follows. Suppose X_t is the state at time t of a Markov chain with transition matrix P. Let $V(x)$ be a *payout function*, which at this point is any real valued function of $x \in \mathcal{S}$. Define $f(x,t)$ as the expected payout at time T starting at state x at time $t \leq T$:

$$f(x,t) \;=\; E\left[V(X_T) \mid X_t = x\right] \;. \tag{2.16}$$

Consider, for example, simple random walk with $V(x) = x$. Then $f(x,t) = x + a(T-t)$ because the one step drift is a and $T - t$ is the number of steps.

The *backward equation* is a relation that determines the numbers $f(x,t)$ from the $f(x,t+1)$. The Markov property also plays a role because

$$E\left[V(X_T) \mid X_t = x \text{ and } X_{t+1} = y\right] \;=\; E\left[V(X_T) \mid X_{t+1} = y\right] \;=\; f(y,t+1)\;.$$

Then the tower property from Chapter 1 gives:

$$\begin{aligned} f(x,t) &= E\left[V(X_T) \mid X_t = x\right] \\ &= \sum_{y\in\mathcal{S}} E\left[V(X_T) \mid X_t = x \text{ and } X_{t+1} = y\right] \cdot \Pr\left(X_{t+1} = y \mid X_t = x\right) \\ &= \sum_{y\in\mathcal{S}} E\left[V(X_T) \mid X_{t+1} = y\right] \cdot \Pr\left(X_{t+1} = y \mid X_t = x\right) \\ f(x,t) &= \sum_{y\in\mathcal{S}} p_{xy} \cdot f(y,t+1)\;. \end{aligned} \tag{2.17}$$

As for the forward equation, (2.17) may be written in matrix/vector notation. The vector this time is the n component column vector $f_t = (f(1,t),\ldots,f(n,t))'$. The matrix form of (2.17) is

$$f_t \;=\; P f_{t+1}\;. \tag{2.18}$$

This determines the f_t for all $t \le T$, given the extra obvious *final condition*

$$f_T \;=\; V\;, \tag{2.19}$$

where V is the column vector $V_j = V(j)$, which is the known values

$$f(j,T) \;=\; E\left[V(X_T) \mid X_T = j\right] \;=\; V(j)\;.$$

The value function/backward equation method is very useful tool. Suppose, for example, a Markov chain X_t is known to be in state x_0 at time $t-0$ and you just want $E[V(X_T)]$. One way to get this number, sometimes the only practical way, is to calculate all the other intermediate quantities $f(j,t)$. You would compute the whole value function to get the single desired number.

As an example, consider a simple random walk on a finite state space $\mathcal{S} = \{1,\ldots,n\}$. Suppose that the for $i \neq 1$ and $i \neq n$ the transition probabilities are the usual $p_{ii} = \beta$, $p_{i,i-1} = \alpha$, $p_{i,i+1} = \gamma$, and $p_{ij} = 0$ otherwise. Suppose that on the ends, transitions out of $\mathcal{S}$ are just not taken. That means that $p_{12} = \Pr(1 \to 2) = \gamma$ and $p_{11} = \Pr(1 \to 1) = \alpha + \beta$. Similarly, $p_{n,n-1} = \alpha$ and $p_{nn} = \beta + \gamma$. The transition matrix is

$$P \;=\; \begin{pmatrix} \alpha+\beta & \gamma & 0 & \cdots & & 0 \\ \alpha & \beta & \gamma & 0 & & \vdots \\ 0 & \alpha & \ddots & \ddots & 0 & \vdots \\ \vdots & 0 & \ddots & \ddots & \gamma & 0 \\ & & & \alpha & \beta & \gamma \\ 0 & \cdots & & 0 & \alpha & \beta+\gamma \end{pmatrix}\;.$$

Each of the row sums of this matrix is equal to one because $\alpha + \beta + \gamma = 1$, including the top and bottom rows. The backward equation in matrix form is[4]

$$\begin{pmatrix} f(1,t) \\ f(2,t) \\ \vdots \\ \\ \\ f(n,t) \end{pmatrix} = \begin{pmatrix} \alpha+\beta & \gamma & 0 & \cdots & & 0 \\ \alpha & \beta & \gamma & 0 & & \vdots \\ 0 & \alpha & \ddots & \ddots & 0 & \vdots \\ \vdots & 0 & \ddots & \ddots & \gamma & 0 \\ & & & \alpha & \beta & \gamma \\ 0 & \cdots & & 0 & \alpha & \beta+\gamma \end{pmatrix} \begin{pmatrix} f(1,t+1) \\ f(2,t+1) \\ \vdots \\ \\ \\ f(n,t+1) \end{pmatrix} .$$

In components, these equations are

$$\left.\begin{array}{rcl} f(1,t) &=& (\alpha+\beta)f(1,t+1) + \gamma f(2,t+1) \\ f(2,t) &=& \alpha f(1,t+1) + \beta f(2,t+1) + \gamma f(3,t+1) \\ &\vdots& \\ f(i,t) &=& \alpha f(i-1,t+1) + \beta f(i,t+1) + \gamma f(i+1,t+1) \\ &\vdots& \\ f(n,t) &=& \alpha f(n-1,t+1) + (\beta+\gamma) f(n,t+1) \, . \end{array}\right\} \tag{2.20}$$

These equations give the value function at time t from the value function at time $t+1$.

There are not many cases where the value function can be evaluated explicitly. You can find some in the exercises. More commonly, one would use a computer. The algorithm starts with $f(i,T) = V(i)$ and then does matrix vector multiplies to produce successively the values $f(i,T-1)$, then $f(i,T-2)$, and so on down to $f(i,0)$. Of course, you need several values at time $t+1$ to compute values at time t. If T is not small, you probably need all the values $f(i,T)$ to get any of the numbers $f(i,0)$.

Unfortunately, there are many practical Markov chains where the state space is so large that it is impractical to compute and store a value function. As an example, consider a d dimensional simple random walk. In this process, the state consists of state consists of d integers: $X = (X_1, \ldots, X_d)$, with each $X_k \in \{1, \ldots, l\}$. Mathematically, we write $X \in \mathcal{S} = \{1, \ldots, l\}^d$. The size of the state space is $n = |\mathcal{S}| = l^d$. Suppose, for example, that $l = 10$ and $d = 30$. Of the computer uses one word (= 44 bytes) per number $f(x,t)$ and we have only one t value, the number of words is $10^{30} = 10^{21}$ Giga-words. No computer now being planned has anything like this much memory or the processing power to compute that many numbers in a practical length of time. On the other hand, the code in Figure 2.2 shows that the system is easy to simulate.

[4]Maybe a LaTeX wizzard can figure out how to make the vertical spacing in the matrix and the vectors the same.

```
import java.util.Random;  // Put this at the top of the file before
                          // the executable code.

//     Put this at the top of the code, execute it only once.

Random rand;          // Declare the symbol "rand" to be a random
                      // number generator.
rand = new Random();  // Initialize this random number generator
double   U;           // A double precision variable, to be random.

//     Variables for the multi dimensional random walk simulation

int      d;           // The dimension of the random walk.
int      l;           // The number of positions in each direction
double   alpha;       // The parameters for the random walk ...
double   beta;        // ... They must be positive ...
double   gamma;       // ... and add up to one.
int[]    X;           // The state of the system at time t
X       = new int[d]; // Allocate memory for d integers

//       Code for setup etc. .....

//   Take a step in the multi-dimensional random walk.

  for ( int k = 0; k < d; k++ ) {     // Loop over the coordinates

     U = rand.nextDouble();  //  Draw a uniform U for this step
     if ( U < alpha ) {      //  With probability alpha . . .
        X[k]--;              //  Move to the left in the i direction
        if ( X[k] = 0 ) {    //  Except that moves outside the  ...
           X[k] = 1;         //  ... allowed range are rejected
         }
      }
     if ( U > 1.- gamma ) {  //  With probability gamma . . .
        X[k]++;              //  Move to the right in the i direction
        if ( X[k] > d ) {    //  Except that moves outside the  ...
           X[k] = d;         //  ... allowed range are rejected
         }                   //  If alpha < U < 1-gamma, do not move.
      }                      //  ... This event has probability beta
    }                        //  Done with this time step
```

Figure 2.2: How to take one step of a multi-dimensional random walk. This code could take a million time steps per second with $l = 10$ and $d = 30$, while computing and storing the value function would be impossible.

2.6 Continuous time Markov chains

Many stochastic dynamical systems are modeled as taking place in continuous time rather than discrete time. In this case the time variable t is continuous, taking values in $\mathbb{R}$. Rather than speaking about transition probabilities, we talk about *transition rates*. In a discrete space, $r_{ij} = \text{rate}(i \to j)$ is defined, if $i \neq j$, by

$$r_{ij}\, dt \;=\; \Pr\left(X_{t+dt} = j \mid X_t = i\right) \;. \tag{2.21}$$

The backward and forward equations (2.18) and (2.14) are replaced by differential equations derived below. The matrix powers above are replaced by the matrix exponential.

We start with the single transition process that consists of a single exponential random variable. By definition, $T > 0$ is an exponential random variable whose *rate constant* is λ if the probability density of T is $u(t) = \lambda e^{-\lambda t}$ (and $u(t) = 0$ for $t < 0$). A simple calculation shows that such a transition is memoryless in the sense that the probability of having to wait a time t does not depend on how long you have waited already:

$$\Pr(T > t_0 + t \mid T > t_0) \;=\; e^{-\lambda t} \;.$$

Now replace t with a small time dt and apply ordinary calculus to $e^{-\lambda\, dt}$ and you get

$$\Pr(t_0 < T < t_0 + t \mid T > t_0) \;=\; \lambda\, dt \;.$$

This makes λ a transition rate in the sense of (2.21).

The *Poisson process* is the stochastic process that counts how many exponential times have happened within a given time t. Suppose T_1, T_2, etc. are independent exponential random variables with common rate constant λ. These are the *inter-arrival* times, also called *waiting times* between events. The time of event k is $T_1 + \cdots + T_k$. The *counting process*, $N(t)$ records the number of arrivals up to time t

$$N(t) \;=\; k \;\text{ if } T_1 + \cdots + T_{k-1} < t \text{ and } T_1 + \cdots + T_{k-1} > t. \tag{2.22}$$

The arrival rate is λ in the sense that

$$\Pr(\text{ arrival in } (t, t+dt) \;) \;=\; \Pr(N(t+dt) > N(t)) \;=\; \lambda dt \;.$$

This is a Markov process because the inter-arrival times are independent. The events {arrival in $(t, t+dt)$} are independent of each other.

2.7 Exercises

1. Find formulas for the three non-zero probabilities in row i of the transition matrix for the simple urn process. For example, to have an $i \to i+1$ transition, an addition of a blue ball, you first have to select a red ball to remove and then choose to replace it with a blue ball. The probability

of choosing red is $(m-i)/m$, which is the fraction of balls that are red. The probability of replacing it with blue is just p. This gives $p_{i,i+1} = p(m-i)/m$. Note that this is zero if $i = m$; if all the balls already are blue, the number of blue balls cannot increase.

2. Find formulas for the random walk one step drift (2.11) and variance (2.12) in terms of α, β, and γ. Show, conversely, how to determine α, β, and γ to achieve a desired one step drift and variance. What values of a and b may be achieved by simple random walks?

3. Describe the "matrix" P that corresponds to simple random walk. More precisely,

   (a) what are the diagonals p_{ii}, the sub-diagonals $p_{i,i-1}$, and the super-diagonals $p_{i,i+1}$, and the other p_{ij}?

   (b) Show that $p^k_{ij} = 0$ if $j > i + k$ and that $p^k_{i,i+k} = \gamma^k$.

   (c) Show that next entries are given by $p^k_{i,i+k-1} = (k-1)\beta\gamma^{k-1}$. Hint: Describe all the paths that go from i to $i + k - 1$, figure out their probability, and how many there are.

   (d) Verify the formula from part (c) by induction on k using the result of part (b) and the recurrence relation (2.14). Hint: in terms of matrix elements, the recurrence relation has the form $p^{k+1}_{ij} = *p^k_{i,j-1} + * * p^k_{ij} + * * * p^k_{i,j+1}$. Figure it out completely then use it.

4. This exercise gives several approaches to the value function for $V(x) = x^2$ in simple random walk.

   (a) Express the backward equation

   (b) Assume that the solution has the form $f(x,t) = A(t)x^2 + B(t)x + C(t)$. This is the *ansatz* method we will use later to find solutions of many partial differential equations. Find a formula for $A(t)$ in terms of $A(t+1)$, and use it (together with $A(T) = 1$) to give a formula for $A(t)$ for $t \leq T$.

Chapter 3

Brownian motion limits

This chapter describes together some general background about measures and limits of measures. It also describes Gaussian and Brownian motion measures and the limits that produce them. For Gaussians, this is the classical central limit theorem. For Brownian motion, this is the more recent (1950's) theorem about the convergence of random walk to Brownian motion.

The rigorous mathematical treatment of these limits is technical, but most people are able to work with the limits confidently without mastering the proofs in detail. This is like the way most people use ordinary calculus. The understand that limits are behind the definitions of derivative and integral but may not have studied the mathematical theory of these limits in detail.

3.1 Central limit theorem

The central limit theorem gives an approximation for the distribution of sums of large numbers of random variables. Suppose Y is a random variable and Y_k is a sequence of independent *samples* having the same distribution as Y. We describe this by saying that $Y_k \sim Y$ are *i.i.d.* (independent and identically distributed). Let $\mu = E[Y]$ and $\sigma^2 = \text{var}(Y)$. We are interested in the sums $S_n = Y_1 + \cdots + Y_n$.

The simplest description is

$$S_n \approx n\mu \, . \tag{3.1}$$

This is the *law of large numbers*, which gives an approximate value for the sum that is not random. For example, suppose that random variables A_k are i.i.d. exponentials with mean $E[A_k] = \mu = 1/\lambda$, where λ is the rate constant. In the previous chapter we interpreted the A_k as the inter-arrival times of a Poisson process. The n^{th} event happens at time $T_n = A_1 + \cdots + A_n$. The law of large numbers says that if n is large, then $T_n \approx n/\lambda$. The average rate of arrivals during that time is the number of arrivals divided by the time, which is n/T_n. For large n, this is approximately $n/(n/\lambda) = \lambda$.

One aspect of the law of large numbers is that the behavior of S_n for large n does not depend much on the distribution of the Y_k. It depends only on the mean. At this level of precision, S_n does not even depend on the variance of the Y_k. This is one of the things that make the limits of stochastic calculus so useful.

The value of S_n is not known exactly, even if it is known approximately. The first issue is *scaling*. The *fluctuation* is $S_n - n\mu$? How big is this likely to be when n is large? The answer comes from finding something you can calculate, in this case the variance

$$E\left[(S_n - n\mu)^2\right] \;=\; \text{var}(S_n) \;=\; n\,\text{var}(Y) \;=\; n\sigma^2 \, .$$

Since typical values of $(S_n - n\mu)^2$ are on the order of n, we may expect that typical values of $S_n - n\mu$ are on the order of $n^{1/2}$. We put this hypothesis into

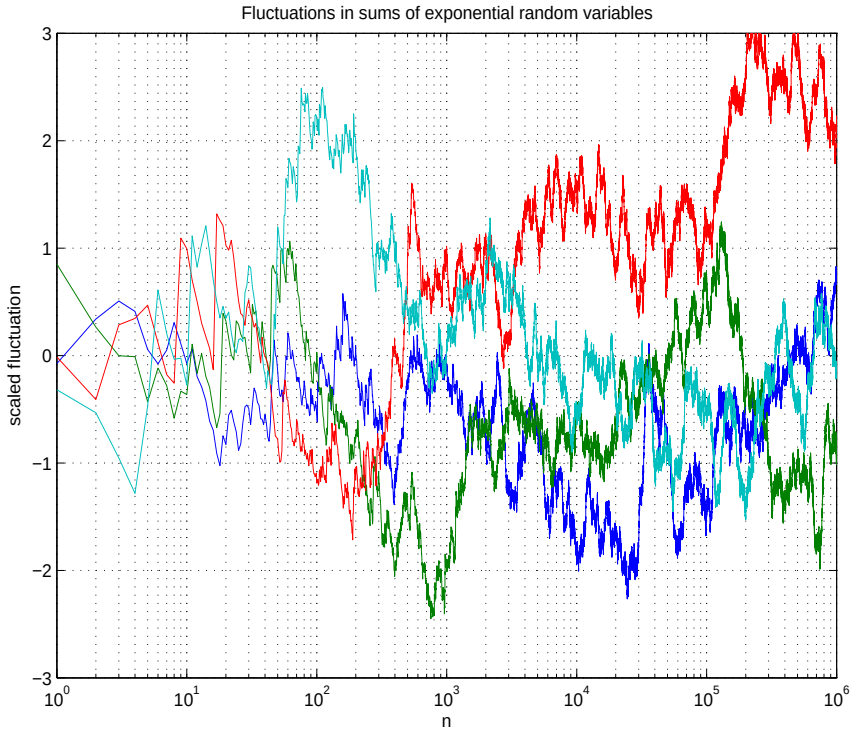

Figure 3.1: Four sample paths of X_n from (3.2). The Y_k are exponential random variables with rate one, which gives them probability density $f(y) = e^{-y}$ for $y > 0$. These trajectories do not seem to have limits or approach each other as $n \to \infty$, at least not up to $n = 10^6$. A later chapter will explain that these paths are well approximated by an Ornstein Uhlenbeck process.

mathematical form with a *scaling ansatz* of the form

$$S_n - n\mu \;=\; n^{1/2} X_n \;. \tag{3.2}$$

You might view this just as the definition of X_n. But if the scaling is done right, the probability distribution of X_n should approach a limit as $n \to \infty$. The *central limit theorem* states that this limiting distribution exists and is Gaussian. As with the law of large numbers, the central limit theorem gives a limit that does not depend much on the distribution of Y, only its variance.

It is important to understand what convergence in distribution does and does not say about X_n. It does not say, for example, that the limit $\lim_{n\to\infty} X_n$ exists. Indeed, figure 3.1 shows that the limit probably does not exist. What is supposed to converge is the probability distribution of X_n rather than X_n itself.

But there also are different senses of convergence of a probability distribu-

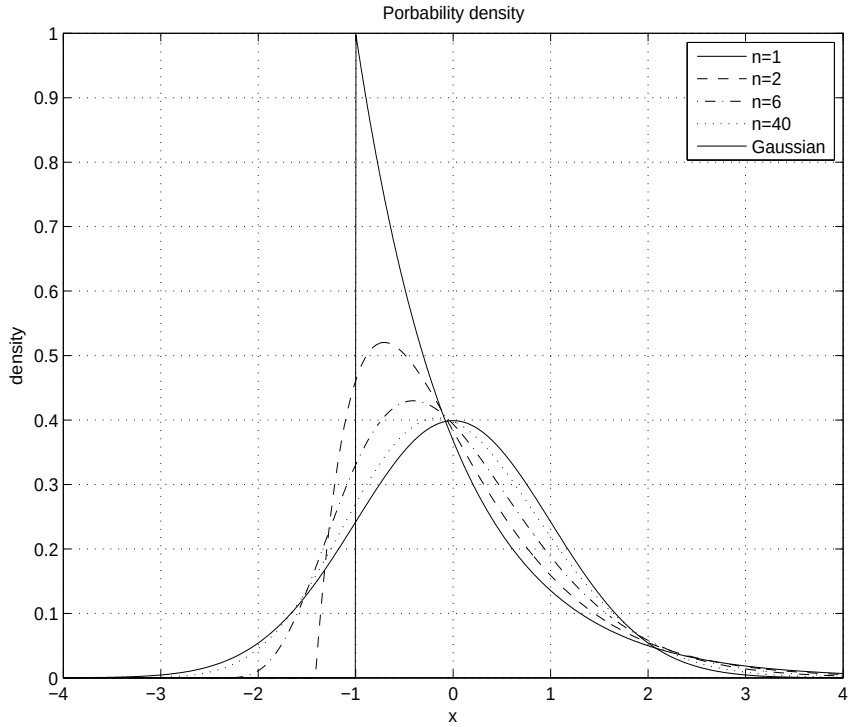

Figure 3.2: The probability density functions for X_n defined by (3.2) for various values of n. The Y_k are i.i.d standard exponentials. It is not clear at first from the figure that each of the curves has the same integral.

tion. The simplest would be convergence of the probability densities. Let $u_n(x)$ be the probability density of X_n. Figure 3.2 shows the $u_n(x)$ converging to a Gaussian density as $n \to \infty$. But there are random variables that do not have probability densities in this sense. For example, suppose $Y = \pm 1$ with probability $\frac{1}{2}$ for each possibility. Then the probability "density" for X_n is equal to zero for all x except for numbers of the form $integer/\sqrt{n}$. These "densities" do not converge in the sense of figure 3.2.

The basic central limit theorem is about convergence *in distribution*, which means that, for suitable *test functions* F,

$$E\left[F(X_n)\right] \;\to\; E\left[F(X)\right] \qquad \text{as } n \to \infty. \tag{3.3}$$

If F is continuous, this can be true even when X_n is a sequence of random variables. Suppose $x_{n,k}$ is a list of the possible values of X_n and $u_{n,k} = \Pr(X_{n,k} = x_{n,k}$. In the specific example above, $x_{n,k} = k/\sqrt{n}$. The spacing between these points, $\Delta x = x_{n,k+1} - x_{n,k} = n^{-1/2}$, goes to zero as $n \to \infty$. Therefore, the

limit (3.3) is

$$\lim_{n\to\infty} \sum_{k=-\infty}^{\infty} F(k/\sqrt{n})\, u_{k,n} \;\to\; \int_{-\infty}^{\infty} F(x)u(x)\, dx \;.$$

This is something like the convergence of Riemann sums to an integral.

The formula (3.3) suggests a strategy for proving the central limit theorem or other limit theorems in probability. The strategy is to find specific functions F for which you can calculate the limit directly. If you can calculate the limit, then you know $\int F(x)u(x)dx$, which is a piece of information about u. One useful family of test functions is powers of x. This tells us moments of u. For a Gaussian, $u \sim \mathcal{N}(0, \sigma^2)$, a calculation shows that the even moments are given by (the sequence is $3\sigma^4$, $15\sigma^6$, $105\sigma^8$, etc.)

$$E\left[X^{2p}\right] \;=\; \sigma^{2p}(2p-1)(2p-3)\cdots(3) \;,$$

while the odd moments all are zero. If $E\left[Y^{2p}\right] < \infty$, it is not so hard to show that $E\left[X_n^{2p}\right]$ has the same limits as $n \to \infty$. Another convenient family of test functions is $F(x) = e^{i\xi x}$. These have (the last step uses the independence of the random variables $W_k = e^{i\xi Y_k/\sqrt{n}}$ and the fact that $Y_k \sim Y$ for each k.)

$$\begin{aligned}
E\left[F(X_n)\right] &= E\left[\exp\left(i\xi \frac{1}{\sqrt{n}} \sum_k Y_k\right)\right] \\
&= E\left[\exp\left(\sum_k \left\{\frac{i\xi Y_k}{\sqrt{n}}\right\}\right)\right] \\
&= E\left[\prod_k e^{i\xi Y_k/\sqrt{n}}\right] \\
&= \left(E\left[e^{i\xi Y/\sqrt{n}}\right]\right)^n
\end{aligned}$$

The exponent is small when n is large and we have the approximation $e^\epsilon \approx 1 + \epsilon + \frac{1}{2}\epsilon^2$, which gives

$$E\left[e^{i\xi Y/\sqrt{n}}\right] \;\approx\; 1 - \frac{\sigma^2 \xi}{n} \;.$$

This implies that

$$E\left[e^{i\xi X_n}\right] \;\to\; e^{-\xi^2/2\sigma^2} \qquad \text{as } n \to \infty,$$

which is the correct value for $X \sim \mathcal{N}(0, \sigma^2)$.

Chapter 4

Continuous Probability

Chapter 5

Gaussian Random Variables

Index