Notes on Probability Theory

Christopher King

Department of Mathematics
Northeastern University

July 31, 2009

Abstract

These notes are intended to give a solid introduction to Probability Theory with a reasonable level of mathematical rigor. Results are carefully stated, and many are proved. Numerous examples and exercises are included to illustrate the applications of the ideas. Many important concepts are already evident in simple situations and the notes include a review of elementary probability theory.

Contents

1 Elementary probability

Many important concepts are already evident in simple situations so we start with a review of elementary probability theory.

1.1 Why all the fuss? – can't we just figure it out??

Questions in probability can be tricky, and we benefit from a clear understanding of how to set up the solution to a problem (even if we can't solve it!). Here is an example where intuition may need to be helped along a bit:

"Remove all cards except aces and kings from a deck, so that only eight cards remain, of which four are aces and four are kings. From this abbreviated deck, deal two cards to a friend. If he looks at his card and announces (truthfully) that his hand contains an ace, what is the probability that both his cards are aces? If he announces instead that one of his cards is the ace of spades, what is the probability then that both his cards are aces? Are these two probabilities the same?"

Probability theory provides the tools to organize our thinking about how to set up calculations like this. It does this by separating out the two important ingredients, namely events (which are collections of possible outcomes) and probabilities (which are numbers assigned to events). This separation into two logically distinct camps is the key which lets us think clearly about such problems. For example, in the first case above, we ask "which outcomes make such an event possible?". Once this has been done we then figure out how to assign a probability to the event (for this example it is just a ratio of integers, but often it is more complicated).

First case: there are 28 possible 'hands' that can be dealt (choose 2 cards out of 8). Out of these 28 hands, exactly 6 contain no aces (choose 2 cards out of 4). Hence 28-6=22 contain at least one ace. Our friend tells us he has an ace, hence he has been dealt one of these 22 hands. Out of these exactly 6 contain two aces (again choose 2 out of 4). Therefore he has a probability of 6/22=3/11 of having two aces.

Second case: one of his cards is the ace of spades. There are 7 possibilities for the other card, out of which 3 will yield a hand with 2 aces. Thus the probability is 3/7.

Any implicit assumptions?? Yes: we assume all hands are equally likely.

1.2 Sample space

The basic setting for a probability model is the *random experiment* or *random trial*. This is your mental model of what is going on. In our previous example this would be the dealer passing over two cards to your friend.

Definition 1 *The sample space S is the set of all possible outcomes of the random experiment.*

Depending on the random experiment, S may be finite, countably infinite or uncountably infinite. For a random coin toss, $S = \{H, T\}$, so $|S| = 2$. For our card example, $|S| = 28$, and consists of all possible unordered pairs of cards, eg (Ace of Hearts, King of Spades) etc. But note that you have some choice here: you could decide to include the order in which two cards are dealt. Your sample space would then be twice as large, and would include both (Ace of Hearts, King of Spades) and (King of Spades, Ace of Hearts). Both of these are valid sample spaces for the experiment. So you get the first hint that there is some *artistry* in probability theory! namely how to choose the 'best' sample space.

Other examples:

(1) Roll a die: the outcome is the number on the upturned face, so $S = \{1, 2, 3, 4, 5, 6\}$, $|S| = 6$.

(2) Toss a coin until Heads appears: the outcome is the number of tosses required, so $S = \{1, 2, 3, \dots\}$, $|S| = \infty$.

(3) Choose a random number between 0 and 1: $S = [0, 1]$. (This is the first example of an uncountable sample space).

(4) Throw a dart at a circular dartboard:

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

For this review of elementary probability we will restrict ourselves to finite and countably infinite sample spaces.

1.3 Events

An event is a collection of possible outcomes of a random experiment. Usually write $A, B, \ldots$ to denote events. So an event A is a subset of S, the sample space, that is $A \subset S$. Usually an event contains the set of outcomes which make the answer to a question 'Yes'. Saying 'the outcome is in A' is the same as saying 'the event A is true'. For the first question in our card example, one event of interest is that both cards are aces. This event is the collection of all outcomes which make it true, namely the 6 hands with two aces.

There are two special events: the whole sample space S is called the certain or the sure event. The empty set $\emptyset$ is the null event.

1.4 Combining events

We often want to combine events in various ways. For example given events E, F, G, might want to investigate the event that at least 2 out of these 3 events are true. There are 3 basic operations for combining events.

Complement

$$E^c = \text{“not } E\text{”} = \text{collection of outcomes not in } E \tag{1}$$

Intersection

$$E \cap F = \text{“}E \text{ and } F\text{”} = \text{collection of outcomes in both } E \text{ and } F \tag{2}$$

Union

$$E \cup F = \text{“}E \text{ or } F\text{”} = \text{collection of outcomes in either } E \text{ or } F \text{ or both} \tag{3}$$

By combining operations can build up more and more complicated events.

Exercise 1 Given three events E, F, G, write formulas for the following events: only E is true; both E and F but not G; at least two of the events are true.

The union and intersection operations distribute like addition and multiplication respectively: for example

$$(E \cup F) \cap G = (E \cap G) \cup (F \cap G) \tag{4}$$

The complement squares to the identity: $(E^c)^c = E$. De Morgan's Laws are

$$(E \cap F)^c = E^c \cup F^c, \quad (E \cup F)^c = E^c \cap F^c \tag{5}$$

Exercise 2 Circuit with switches in parallel or in series. Describe event that circuit is open in terms of events that each switch is open or closed.

1.5 Assigning probabilities

The second ingredient in our setup is the assignment of a probability to an event. These probabilities can often be calculated from 'first principles'. In our card example we did this by counting and dividing. In other cases the probabilities may be given as part of the description of the problem; for example if you are told that a coin is biased and comes up Heads twice as often as Tails. We next analyze the requirements for a satisfactory assignment.

The basic step is that every event E is assigned a probability $P(E)$. This is a number satisfying

$$0 \leq P(E) \leq 1 \tag{6}$$

The meaning is "$P(E)$ is the probability that event E is true". The operational meaning (which will follow from the mathematical setup) is that if the random experiment (our mental image of the process) is repeated many times under identical conditions, then in the long-run the *fraction of times* when E is true will approach $P(E)$ as the number of trials becomes arbitrarily large. Since this can never be checked in practice, it remains an article of faith about how the universe works. Nevertheless it can be formulated as a mathematical statement in probability theory, and then it can be shown to be a consequence of the axioms of the theory. This result is called the Law of Large Numbers and will be studied in detail later in the course.

There are lots of possible events, so there are consistency relations that must be satisfied. Here are some:

(1) $P(E^c) = 1 - P(E)$

(2) $P(S) = 1$

(3) if $E \subset F$ then $P(E) \leq P(F)$

(4) if $E \cap F = \emptyset$ (aka E and F are disjoint, or mutually exclusive), then

$$P(E \cup F) = P(E) + P(F) \tag{7}$$

(5) for any events E, F,

$$P(E \cup F) = P(E) + P(F) - P(E \cap F) \tag{8}$$

(6) if $E_1, E_2, \ldots, E_n, \ldots$ is a sequence of pairwise disjoint events, so that $E_i \cap E_j = \emptyset$ for all $i \neq j$, then

$$P(\bigcup_{n=1}^{\infty} E_n) = \sum_{n=1}^{\infty} P(E_n) \tag{9}$$

The last property (6) is crucial, and it cannot be deduced from the previous relations which involve only finitely many sets. This property is called countable additivity and we will have much more to say about it later.

Other relations then follow from these. However it can be shown that there are no other independent relations; if conditions (1) – (6) hold for all events then P is a consistent assignment of probabilities on S. In this case the assignment P is called a *probability model* or *probability law* on S.

Some work has gone into finding a minimal set of relations which generate all others: one such minimal set is the two relations (2) and (6) above.

Exercise 3 Derive (1), (3), (5) from (2) and (4).

Exercise 4 Two events E and F; the probability that neither is true is 0.6, the probability that both are true is 0.2; find the probability that exactly one of E or F is true.

In elementary probability theory where S is either finite or countably infinite, every possible outcome $s \in S$ is assigned a probability $p(s)$, and then the probability of any event A can be calculated by the sum

$$P(A) = \sum_{s \in A} p(s) \tag{10}$$

This relation follows from (6) above, since $A = \cup_{s \in A}\{s\}$ is a countable union of disjoint sets. The sum always converges, even if S is (countably) infinite. Furthermore, if $p : S \to [0,1],\ s \mapsto p(s)$ is any map that satisfies the condition

$$\sum_{s \in S} p(s) = 1 \tag{11}$$

then it defines a probability law on S.

Exercise 5 For any sequence of events $\{A_n\}$, use Property (6) to show that

$$P(\bigcup_{n=1}^{\infty} A_n) \leq \sum_{n=1}^{\infty} P(A_n) \tag{12}$$

[Hint: rewrite $\bigcup_{n=1}^{\infty} A_n$ as a union of pairwise disjoint sets]

1.6 Drawing breath

To summarize: we have laid down the mathematical foundations of probability theory. The key step is to recognize the two separate pillars of the subject, namely on the one hand the sample space of outcomes, and on the other hand the numerical probabilities which are assigned to events. Next we use this basic setup to define the familiar notions of probability, such as independence, random variables etc..

1.7 Conditional probability

$P(B|A) =$ conditional probability that B is true given that A is true

Imagine the following 2-step thought experiment: you toss a coin; if it comes up Heads, you draw one card at random from a standard deck; if it comes up Tails you draw two cards at random (without replacement). Let A be the event that you get Heads on the coin toss, and let B be the event that you draw at least one Ace from the deck. Then $P(B|A)$ is clearly $4/52 = 1/13$. What about $P(A \cap B)$? Imagine lining up all your many repeated experiments, then for approximately one-half of them the event A will be

true. Out of these approximately 1/13 will have B also true. So we expect that $P(A \cap B) = (1/2)(1/13) = P(A)P(B|A)$. This line of reasoning leads to the following definition:

$$P(B|A) = \frac{P(B \cap A)}{P(A)} \tag{13}$$

It is important to note that $P(B|A)$ is defined only if $P(A) \neq 0$.

Exercise 6 Suppose that $P(B|A) > P(B)$. What does this imply about the relation between $P(A|B)$ and $P(A)$?

Exercise 7 Show that

$$\begin{aligned} &P(A_1 \cap A_2 \cap \cdots \cap A_n) \\ &\quad = P(A_1)\, P(A_2|A_1)\, P(A_3|A_1 \cap A_2)\, \ldots P(A_n|A_1 \cap A_2 \cap \cdots \cap A_{n-1}) \end{aligned} \tag{14}$$

Exercise 8 A standard deck of 52 playing cards is randomly divided into 4 piles of 13 cards each. Find the probability that each pile has exactly one Ace.
[Hint: define events $A_1, \ldots, A_4$ by

$$A_k = \{\text{the } k^{th} \text{ pile has exactly one Ace}\}, \quad k = 1, 2, 3, 4 \tag{15}$$

and use the previous Exercise]

One useful application is the formula for total probability: suppose that there is a collection of events $A_1, A_2, \ldots, A_n$ which are mutually disjoint, so $A_i \cap A_j = \emptyset$ for all $i \neq j$, and also exhaustive, meaning they include every outcome so that $A_1 \cup A_2 \cup \cdots \cup A_n = S$. Then for any event B,

$$\begin{aligned} P(B) &= P(B \cap A_1) + P(B \cap A_2) + \cdots + P(B \cap A_n) \\ &= P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \cdots + P(B|A_n)P(A_n) \end{aligned} \tag{16}$$

Note that the first equality follows from Property (4) of the probability law.

Exercise 9 Derive Bayes formula: for mutually exclusive and exhaustive events $A_1, \ldots, A_n$,

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \cdots + P(B|A_n)P(A_n)} \quad (17)$$

Exercise 10 A patient walks in who has a fever and chills. The doctor wonders, "what is the chance that this patient has tuberculosis given the symptoms I am seeing?" Let A be the event that the patient has TB, let B be the event that the patient has fever and chills. Assume that TB is present in 0.01% of the population, whereas 3% of the population exhibits fever and chills. Assume that $P(B|A) = 0.5$. What is the answer to the doctor's question?

Exercise 11 Rework our old card problem using conditional probabilities.

1.8 Independence

Two events A, B are *independent* if

$$P(A|B) = P(A) \Longleftrightarrow P(B|A) = P(B) \Longleftrightarrow P(A \cap B) = P(A)\,P(B) \quad (18)$$

In other words these three conditions are equivalent.

The collection of events $A_1, \ldots, A_n, \ldots$ is independent if for every finite subset $A_{i_1}, \ldots, A_{i_k}$,

$$P(A_{i_1} \cap \cdots \cap A_{i_k}) = P(A_{i_1}) \cdots P(A_{i_k}) \quad (19)$$

Independence is very important in probability theory because it occurs naturally in many applications, and also because it provides very useful tools for solving problems.

Exercise 12 Successive coin tosses are independent. A biased coin has probability p of coming up Heads. The coin is tossed 10 times. Find the probability that it comes up Heads at least twice.

Exercise 13 Two dice are rolled many times, and each time the sum of the numbers on the dice is recorded. Find the probability that the value 8 will occur before the value 7.

1.9 Random variables

A random variable is a 'random number', meaning a number which is determined by the outcome of a random experiment. Usually denoted $X, Y, \ldots$. The *range* of X is the set of possible values for X. Mathematically, X is a real-valued map on the sample space S:

$$X : S \to \mathbb{R}, \quad s \mapsto X(s) \tag{20}$$

Another way to say this is that X is the result of a measurement of interest on S.

In elementary probability we consider only discrete random variables whose range is either finite or countably infinite. If the range of X is finite then we say that X is a *simple random variable*. The event $\{X = x\}$ is the set of outcomes in S for which the value x is assigned to X. Mathematically,

$$\{X = x\} = \{s \in S \,|\, X(s) = x\} = X^{-1}(\{x\}) \tag{21}$$

The probability of this event is written $P(X = x)$. At this point the sample space S recedes into the background, and we can concentrate just on the range of possible values of X and their probabilities. This list is called the probability mass function or pmf of X:

$$(x_1, p_1),\ (x_2, p_2),\ \ldots \tag{22}$$

where $\mathrm{Ran}(X) = \{x_1, x_2, \ldots\}$ and $p_k = P(X = x_k)$.

Exercise 14 Roll two fair dice, Y is the maximum of the two numbers on their faces, find the pmf of Y.

Given just the pmf of X, is there a unique underlying sample space S with its probability assignments? The answer is no. There are many sample spaces which would yield the same pmf for X. But there is a minimal sample space which does the job. Just take S to be the set of points in the range of X, and assign probabilities to these points according to the pmf of X. So $S = \{x_1, x_2, \ldots\}$ and $P(\{x_k\}) = p_k$. In this case the map which defines X is particularly simple, it is just the identity function: $X(x_k) = x_k$. This exercise also shows that there is a random variable defined for every pmf: given a countable set of real numbers $\{x_k\}$ and a set of probabilities $\{p_k\}$ satisfying $\sum_k p_k = 1$, there is a random variable X whose pmf is $\{(x_k, p_k)\}$.

To see an example, suppose we roll two fair dice and define

$$X = \begin{cases} 0 & \text{if the dice are different} \\ 1 & \text{if the dice are the same} \end{cases}$$

The obvious sample space has 36 elements, namely all pairs of outcomes for the dice. The map X assigns value either 0 or 1 to every outcome, e.g. $X(1,3) = 0$, $X(4,4) = 1$, etc. The pmf of X is

$$P(X = 0) = \frac{1}{6}, \qquad P(X = 1) = \frac{5}{6}$$

We could instead take our sample space to consist of just two elements, namely $S = \{0, 1\}$ with probabilities $(1/6, 5/6)$, then define $X(0) = 0$ and $X(1) = 1$, and we end up with the same pmf for X. So all that really matters for X is that it is *possible* to construct a sample space on which X can be defined with this pmf, we don't actually care about the details of the sample space.

There are several special discrete random variables which are especially important because they arise in many situations.

Bernoulli
$\text{Ran}(X) = \{0, 1\}$, $p = P(X = 1)$, $1 - p = P(X = 0)$. For example, a biased coin has probability p of coming up Heads. Toss coin, X is number of Heads.

Binomial
$\text{Ran}(X) = \{0, 1, 2, \ldots, n\}$, $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$. Now X is number of Heads for n tosses of a biased coin. As a shorthand write

$$X \sim \text{Bin}(n, p) \tag{23}$$

Poisson
$\text{Ran}(X) = \{0, 1, 2, \ldots\}$, $P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}$. For example, X counts number of occurrences of rare events, like radioactive decays from a sample.

There is an important relation between the Binomial and Poisson formulas.

Lemma 2 *Fix $\lambda = np$, then*

$$\lim_{n \to \infty, p \to 0} \binom{n}{k} p^k (1-p)^{n-k} = e^{-\lambda} \frac{\lambda^k}{k!}$$

Exercise 15 Biased coin, p is probability of Heads. Toss coin until Heads appears. Let N be number of tosses, find the pmf of N. [This is the *geometric* distribution].

1.10 Joint distributions

In many circumstances we encounter a collection of random variables which are all related to each other. For example, X and Y could be the minimum and maximum respectively of two rolls of the dice. Often we want to consider these related random variables together.

Given a collection of discrete random variables $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, let R_i be the range of X_i. Then the range of $\mathbf{X}$ is the Cartesian product $R_1 \times \cdots \times R_n$. Their joint pmf is the collection of probabilities $P(X_1 = x_1, \ldots, X_n = x_n)$ for every point $(x_1, \ldots, x_n)$ in the range of $X_1, X_2, \ldots, X_n$. It is also convenient to view $\mathbf{X}$ as a random vector in $\mathbb{R}^n$.

Exercise 16 Let X and Y be the minimum and maximum respectively of two rolls of the dice. Find the joint pmf of X and Y.

The random variables $X_1, X_2, \ldots, X_n$ are defined on the same sample space S. Just as for a single discrete random variable, if S is not known a priori we can always construct a sample space for $X_1, X_2, \ldots, X_n$ by taking S to be the range $R_1 \times \cdots \times R_n$, and defining the probability of a point using the pmf. Then X_i is the projection onto the i^{th} coordinate.

We can recover the pmf of X_1 by itself from the joint pmf:

$$P(X_1 = x_1) = \sum_{x_2, \ldots, x_n} P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) \tag{24}$$

This procedure is called finding the marginal pmf of X_1. The same procedure works for $X_2, \ldots, X_n$.

The random variables $X_1, X_2, \ldots, X_n$ are independent if for every point $(x_1, \ldots, x_n)$ the events $\{X_1 = x_1\}, \{X_2 = x_2\}, \ldots, \{X_n = x_n\}$ are independent. Equivalently, $X_1, X_2, \ldots, X_n$ are independent if and only if the joint pmf is the product of the marginals, that is for every point $(x_1, \ldots, x_n)$,

$$P(X_1 = x_1, \ldots, X_n = x_n) = P(X_1 = x_1) \ldots P(X_n = x_n) \tag{25}$$

Exercise 17 You have two coins, one is unbiased, the other is biased with probability of Heads equal to 2/3. You toss both coins twice, X is the number of Heads for the fair coin, Y is the number of Heads for the biased coin. Find $P(X > Y)$. [Hint: X and Y are independent].

Exercise 18 Two biased coins have the same probability p of coming up Heads. The first coin is tossed until Heads appears for the first time, let N be the number of tosses. The second coin is then tossed N times. Let X be the number of times the second coin comes up Heads. Find the pmf of X (express the pmf by writing $P(X = k)$ as an infinite series).

1.11 Expected value or expectation

Let X be a discrete random variable with pmf $(x_1, p_1), (x_2, p_2), \ldots$. If the range of X is finite the expected value or expectation of X is defined to be

$$\mathbb{E}X = \sum_n p_n x_n \tag{26}$$

As an example can compute expected value of roll of die ($= 7/2$).

If the range of X is infinite, the sum is defined as follows: first divide X into its positive and negative parts X^+ and X^-,

$$X^+ = \max\{X, 0\}, \quad X^- = X - X^+ \tag{27}$$

Define

$$\mathbb{E}X^+ = \sum_{n\,:\,x_n \geq 0} p_n x_n, \quad \mathbb{E}X^- = \sum_{n\,:\,x_n < 0} p_n |x_n| \tag{28}$$

Both are sums of positive terms, hence each either converges or is $+\infty$. Unless both $\mathbb{E}X^+ = \mathbb{E}X^- = +\infty$ we say that $\mathbb{E}X$ exists and define it to be

$$\mathbb{E}X = \mathbb{E}X^+ - \mathbb{E}X^- \tag{29}$$

The value of $\mathbb{E}X$ may be finite, or $\pm\infty$. If both $\mathbb{E}X^+ = \mathbb{E}X^- = +\infty$ then $\mathbb{E}X$ does not exist. Note that $|X| = X^+ + X^-$. Hence $\mathbb{E}X$ exists and is finite if and only if $\mathbb{E}|X|$ exists and is finite.

$\mathbb{E}X$ has a nice operational meaning. Repeat the underlying random experiment many times, and measure the value of X each time. Let $\mathrm{Av}(X; n)$

be the average of these values for n successive measurements. This average value depends on n and is itself a random variable. However our experience with the universe shows that $\mathrm{Av}(X;n)$ converges as $n \to \infty$, and this limiting value is $\mathbb{E}X$. Again this can never be verified by experiment but it can be derived mathematically from the axioms.

Exercise 19 Find $\mathbb{E}X$ when: (a) X is maximum of two dice rolls ($= 161/36$), (b)X is number of tosses of biased coin until Heads first appears ($= 1/p$).

1.12 Draw breath again

We have now met all the ingredients of elementary probability theory. With these in hand you can tackle any problem involving finite sample spaces and discrete random variables. We will shortly look in detail at one such class of models, namely finite state Markov chains. First we consider some further technical questions.

1.13 Function of a random variable

Let $X : S \to \mathbb{R}$ be a discrete random variable, and $g : \mathbb{R} \to \mathbb{R}$ a real-valued function. Then $Y = g \circ X : S \to \mathbb{R}$ is also a random variable. Its range is

$$\mathrm{Ran}(Y) = g(\mathrm{Ran}(X)) = \{g(x_k) \,|\, x_k \in \mathrm{Ran}(X)\} \tag{30}$$

and its pmf is

$$\begin{aligned} P(Y = y) &= P(\{s \,:\, g(X(s)) = y\}) \\ &= \sum_{s\,:\,g(X(s))=y} p(s) \\ &= \sum_{k:g(x_k)=y} \sum_{s\,:\,X(s)=x_k} p(s) \\ &= \sum_{k:g(x_k)=y} P(X = x_k) \end{aligned} \tag{31}$$

Write $Y = g(X)$. To compute $\mathbb{E}Y$, first define the positive and negative

parts Y^+ and Y^- as before. Then

$$\begin{aligned}
\mathbb{E}Y^+ &= \sum_{y_j \geq 0} y_j P(Y = y_j) \\
&= \sum_{y_j \geq 0} y_j \sum_{k:g(x_k)=y_j} P(X = x_k) \\
&= \sum_{y_j \geq 0} \sum_{k:g(x_k)=y_j} g(x_k)\, P(X = x_k) \\
&= \sum_{y_j \geq 0} \sum_{k} 1_{\{g(x_k)=y_j\}}\, g(x_k)\, P(X = x_k) \qquad (32)
\end{aligned}$$

where 1_A is the indicator function of the event A: it equals 1 if A is true, and 0 if false. All terms in the double summation are positive, so we can change the order of summation without changing the value of the sum. Hence

$$\begin{aligned}
\mathbb{E}Y^+ &= \sum_{k} \sum_{y_j \geq 0} 1_{\{g(x_k)=y_j\}}\, g(x_k)\, P(X = x_k) \\
&= \sum_{k:g(x_k)\geq 0} g(x_k)\, P(X = x_k) \qquad (33)
\end{aligned}$$

The same calculation shows that

$$\mathbb{E}Y^- = \sum_{k:g(x_k)<0} g(x_k)\, P(X = x_k) \qquad (34)$$

Assuming $\mathbb{E}Y$ is defined, so at least one of $\mathbb{E}Y^+$ and $\mathbb{E}Y^-$ is finite, we conclude that

$$\mathbb{E}Y = \sum_{k} g(x_k)\, P(X = x_k) \qquad (35)$$

or more casually

$$\mathbb{E}g(X) = \sum_{x} g(x)\, P(X = x) \qquad (36)$$

This is a change of variables formula which allows us to compute expectations of functions of X directly from the pmf of X itself.

Exercise 20 Suppose $a_{i,j} \geq 0$ for all $i, j \geq 1$, show that

$$\sum_{i=1}^{\infty} \Big(\sum_{j=1}^{\infty} a_{i,j} \Big) = \sum_{j=1}^{\infty} \Big(\sum_{i=1}^{\infty} a_{i,j} \Big) \tag{37}$$

where $+\infty$ is a possible value for both sums.

Exercise 21 Show that $\mathbb{E}[\cdot]$ is a linear operator.

Exercise 22 If $\mathrm{Ran}(N) = \{1, 2, \dots\}$ show that

$$\mathbb{E}N = \sum_{n=1}^{\infty} P(N \geq n) \tag{38}$$

Exercise 23 Compute $\mathbb{E}X^2$ where X is the number of tosses of a biased coin until Heads first appears ($= (2-p)/p^2$).

1.14 Moments of X

The k^{th} moment of X is defined to be $\mathbb{E}X^k$ (if it exists). The first moment is the expected value of X, also called the mean of X.

Exercise 24 If the k^{th} moment of X exists and is finite, show that the j^{th} moment exists and is finite for all $1 \leq j \leq k$. [Hint: if $j \leq k$ show that $|X|^j \leq 1 + |X|^k$]

The variance of X is defined to be

$$\mathrm{VAR}(X) = \mathbb{E}\Big(X - \mathbb{E}X \Big)^2 \tag{39}$$

If the second moment of X exists then

$$\mathrm{VAR}(X) = \mathbb{E}[X^2 - 2X\mathbb{E}X + (\mathbb{E}X)^2] = \mathbb{E}X^2 - (\mathbb{E}X)^2 \tag{40}$$

The standard deviation of X is defined as

$$\mathrm{STD}(X) = \sqrt{\mathrm{VAR}(X)} \tag{41}$$

Exercise 25 Suppose $\mathrm{Ran}(X) = \{1, 2, \dots\}$ and $P(X = k) = c\,k^{-t}$ where $t > 1$ and $c > 0$ is an irrelevant constant. Find which moments of X are finite (the answer depends on t).

The moment generating function (mgf) of X is defined for $t \in \mathbb{R}$ by

$$M_X(t) = \mathbb{E}e^{tX} = \sum_x e^{tx} P(X = x) \tag{42}$$

If X is finite then M_X always exists. If X is infinite it may or may not exist for a given value t. Since $e^{tx} > 0$ for all t, x, the mgf is either finite or $+\infty$. Clearly $M_X(0) = 1$.

1.15 Function of a random vector

Suppose $X_1, \dots, X_n$ are random variables with joint pmf $p(x_1, \dots, x_n)$. Let $g : \mathbb{R}^n \mapsto \mathbb{R}$, then

$$\mathbb{E}g(X_1, \dots, X_n) = \sum_{x_1,\dots,x_n} g(x_1, \dots, x_n)\, p(x_1, \dots, x_n) \tag{43}$$

In particular if $g(x_1, \dots, x_n) = x_k$ is the projection onto the k^{th} coordinate then

$$\mathbb{E}g(X_1, \dots, X_n) = \mathbb{E}X_k = \sum_{x_1,\dots,x_n} x_k\, p(x_1, \dots, x_n) = \sum_{x_k} x_k\, p(x_k) \tag{44}$$

where $p(x_k) = P(X = x_k)$ is the marginal pmf of X_k.

Commonly encountered applications of this formula include:

$$\mathbb{E}(aX + bY) = a\mathbb{E}X + b\mathbb{E}Y \tag{45}$$

$$\mathrm{COV}(X, Y) = \mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y) = \mathbb{E}(XY) - (\mathbb{E}X)(\mathbb{E}Y) \tag{46}$$

The last number is the covariance if X and Y and it measures the degree of dependence between the two random variables.

Exercise 26 If X and Y are independent show that $\mathrm{COV}(X, Y) = 0$.

Exercise 27 Calculate $\mathrm{COV}(X, Y)$ when X, Y are respectively the max and min of two dice rolls.

As noted above, if X and Y are independent then $\mathrm{COV}(X, Y) = 0$, that is $\mathbb{E}(XY) = (\mathbb{E}X)(\mathbb{E}Y)$. Application of this and a little algebra shows that if $X_1, X_2, \ldots, X_n$ are all independent, then

$$\mathrm{VAR}[X_1 + X_2 + \cdots + X_n] = \mathrm{VAR}[X_1] + \mathrm{VAR}[X_2] + \cdots + \mathrm{VAR}[X_n] \quad (47)$$

Exercise 28 Using the linearity of expected value and the above property of variance of a sum of independent random variables, calculate the mean and variance of the binomial random variable. [Hint: write $X = X_1 + \cdots + X_n$ where X_k counts the number of Heads on the k^{th} toss].

Exercise 29 Derive the formula

$$\mathrm{VAR}[X_1 + X_2 + \cdots + X_n] = \sum_{k=1}^{n} \mathrm{VAR}[X_k] + 2 \sum_{i<j} \mathrm{COV}(X_i, X_j) \quad (48)$$

1.16 Conditional distribution and expectation

If two random variables are independent then knowing the value of one of them tells you nothing about the other. However if they are dependent, then knowledge of one gives you information about the likely value of the other. Let X, Y be discrete random variables. The conditional distribution of X conditioned on the event $\{Y = y\}$ is (assuming $P(Y = y) \neq 0$)

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

For a non-negative random variable X and an event A with $P(A) \neq 0$, define the conditional expectation of X with respect to A as

$$\mathbb{E}[X|A] = \sum_{x} x\, P(X = x|A) \quad (49)$$

Since all terms in the sum are positive, either the sum converges or else it is $+\infty$. For a general random variable X, write $X = X^+ - X^-$ and then define

$$\mathbb{E}[X|A] = \mathbb{E}[X^+|A] - \mathbb{E}[X^-|A] \tag{50}$$

assuming as usual that both terms are not infinite.

An important special case is where $A = \{Y = y\}$ for some random variable Y, with $P(Y = y) \neq 0$. Since $\mathbb{E}[X|Y = y]$ is defined for every $y \in \mathrm{Ran}(Y)$, it defines a real-valued function on S, and hence is itself a random variable. It is denoted $\mathbb{E}[X|Y]$ and is defined by

$$\mathbb{E}[X|Y] : S \to \mathbb{R}, \quad s \mapsto \mathbb{E}[X|Y = Y(s)] \tag{51}$$

Since the value of $\mathbb{E}[X|Y](s)$ depends only on $Y(s)$, it follows that $\mathbb{E}[X|Y]$ is a function of Y. Hence its expected value can be computed using our formula for expectation of a function of a random variable:

$$\mathbb{E}[\mathbb{E}[X|Y]] = \sum_y \mathbb{E}[X|Y = y]P(Y = y) \tag{52}$$

Exercise 30 Assuming that $\mathbb{E}X$ exists, show that

$$\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}X \tag{53}$$

Exercise 31 Let $N, X_1, X_2, \ldots$ be independent random variables on a discrete sample space S. Assume the X_k are identically distributed with finite mean $\mathbb{E}X_k = \mu$. Also assume that $\mathrm{Ran}(N) = \{1, 2, 3, \ldots\} = \mathbb{N}$, and that $\mathbb{E}N < \infty$. Define

$$Y = \sum_{n=1}^{N} X_n \tag{54}$$

Prove that $\mathbb{E}Y = \mu\,\mathbb{E}N$.

Exercise 32 Same setup as in previous exercise, assume in addition that both $\mathrm{VAR}[X_k] < \infty$ and $\mathrm{VAR}[N] < \infty$. Prove that

$$\mathrm{VAR}[Y] = \mathbb{E}N\,\mathrm{VAR}[X] + \mu^2\,\mathrm{VAR}[N] \tag{55}$$

Exercise 33 A rat is trapped in a maze with three doors. Door #1 leads to the exit after 1 minute. Door #2 returns to the maze after three minutes. Door #3 returns to the maze after five minutes. Assuming that the rat is at all times equally likely to choose any one of the doors, what is the expected length of time until the rat reaches the exit?

2 Discrete-time finite state Markov chains

Now we can 'get moving' with the Markov chain. This is the simplest non-trivial example of a stochastic process. It has an enormous range of applications, including:

- statistical physics
- queueing theory
- communication networks
- voice recognition
- bioinformatics
- Google's pagerank algorithm
- computer learning and inference
- economics
- gambling
- data compression

2.1 Definition of the chain

Let $S = \{1, \dots, N\}$. A collection of S-valued random variables $\{X_0, X_1, X_2, \dots\}$ is called a discrete-time Markov chain on S if it satisfies the *Markov condition*:

$$P(X_n = j \,|\, X_0 = i_0, \dots, X_{n-1} = i_{n-1}) = P(X_n = j \,|\, X_{n-1} = i_{n-1}) \qquad (56)$$

for all $n \geq 1$ and all states $j, i_0, \dots, i_{n-1} \in S$.

Regarding the index of X_n as a discrete time the Markov condition can be summarized by saying that the conditional distribution of the present state X_n conditioned on the past states $X_0, \dots, X_{n-1}$ is equal to the conditional distribution of X_n conditioned on the most recent past state X_{n-1}. In other words, the future (random) behavior of the chain only depends on where the chain sits right now, and not on how it got to its present position.

We will mostly consider homogeneous chains, meaning that for all n and $i, j \in S$

$$P(X_n = j \,|\, X_{n-1} = i) = P(X_1 = j \,|\, X_0 = i) = p_{ij} \tag{57}$$

This defines the $N \times N$ transition matrix P with entries p_{ij}.

A transition matrix must satisfy these properties:

(P1) $p_{ij} \geq 0$ for all $i, j \in S$

(P2) $\sum_{j \in S} p_{ij} = 1$ for all $i \in S$

Such a matrix is also called row-stochastic. So a square matrix is a transition matrix if and only if it is row-stochastic.

Once the initial probability distribution of X_0 is specified, the joint distribution of the X_i is determined. So let $\alpha_i = P(X_0 = i)$ for all $i \in S$, then for any sequence of states $i_0, i_1, \ldots, i_m$ we have (recall Exercise **)

$$P(X_0 = i_0, X_1 = i_1, \ldots, X_m = i_m) = \alpha_{i_0} p_{i_0,i_1} p_{i_1,i_2} \ldots p_{i_{m-1},i_m} \tag{58}$$

The transition matrix contains the information about how the chain evolves over successive transitions. For example,

$$\begin{aligned}
P(X_2 = j | X_0 = i) &= \sum_k P(X_2 = j, X_1 = k | X_0 = i) \\
&= \sum_k P(X_2 = j | X_1 = k, X_0 = i)\, P(X_1 = k | X_0 = i) \\
&= \sum_k P(X_2 = j | X_1 = k)\, P(X_1 = k | X_0 = i) \\
&= \sum_k p_{kj}\, p_{ik} \\
&= \sum_k (P)_{ik}\, (P)_{kj} \\
&= (P^2)_{ij}
\end{aligned} \tag{59}$$

So the matrix P^2 provides the transition rule for two consecutive steps of the chain. It is easy to check that P^2 is also row-stochastic, and hence is the transition matrix for a Markov chain, namely the two-step chain $X_0, X_2, X_4, \ldots$, or $X_1, X_3, \ldots$. A similar calculation shows that for any $k \geq 1$

$$P(X_k = j | X_0 = i) = (P^k)_{ij} \tag{60}$$

and hence P^k is the k-step transition matrix. We write

$$p_{ij}(n) = (P^n)_{ij} = P(X_n = j | X_0 = i) \tag{61}$$

Note that $p_{ij} = 0$ means that the chain cannot move from state i to state j in one step. However it is possible in this situation that there is an integer n such that $p_{ij}(n) > 0$, meaning that it is possible to move from i to j in n steps. In this case we say that state j is *accessible* from state i.

Example 1

Consider the following model. There are four balls, two White and two Black, and two boxes. Two balls are placed in each box. The transition mechanism is that at each time unit one ball is randomly selected from each box, these balls are exchanged, and then placed into the boxes. Let X_n be the number of White balls in the first box after n steps. The state space is $S = \{0, 1, 2\}$. The transition matrix is

$$P = \begin{pmatrix} 0 & 1 & 0 \\ 1/4 & 1/2 & 1/4 \\ 0 & 1 & 0 \end{pmatrix} \tag{62}$$

Why is it a Markov chain? The transition mechanism only depends on the current state of the system. Once you know the current state (= number of balls in first box) you can calculate the probabilities of the next state.

Example 2

The drunkard's walk. The state space is $S = \{0, 1, 2, 3, 4\}$, and X_n is the drunkard's position after n steps. At each step he goes left or right with probability $1/2$ until he reaches an endpoint 0 or 4, where he stays forever. The transition matrix is

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \tag{63}$$

Again the transition mechanism depends only on the current state, which means that this is a Markov chain.

Exercise 34 Decide if the following are Markov chains. A deck of cards is randomly shuffled. (1) The top card is selected, X is the value of this card.

The card is replaced in the deck at a random position. The top card is again drawn and so on. (2) The top card is selected, X is the value of this card. The card is *not* replaced in the deck. The top card is again drawn and so on.

Exercise 35 Suppose that $S_n = \sum_{i=1}^n X_i$, where $\{X_i\}$ are IID random variables which assume a finite number of values. Assume that the distribution of X_i is known. In each case, either show that the given sequence is a Markov chain, or give an example to show that it is not.
i). $\{S_n\}$
ii). $\{S_{\gamma_n}\}$ where $\gamma_n = \min\{k \leq n : \quad S_k = \max\{S_1, \ldots, S_n\}\}$
iii). The ordered pair (S_n, S_{γ_n}).
[Hint: for (ii) take $\text{Ran}(X) = \{0, -1, 1\}$]

A finite state Markov chain can be usefully represented by a directed graph where the vertices are the states, and edges are the allowed one-step transitions.

2.2 Absorbing chains

This is one special type of chain, exemplified by Example 2 above.

Definition 3 *A state i is* absorbing *if $p_{ii} = 1$. A chain is* absorbing *if for every state i there is an absorbing state which is accessible from i. A non-absorbing state in an absorbing chain is called a* transient *state.*

Consider an absorbing chain with r absorbing states and t transient states. Denote by R the set of absorbing states and by T the set of transient states. Re-order the states so that the transient states come first, then the absorbing states. The transition matrix then has the form

$$P = \begin{pmatrix} Q & R \\ 0 & I \end{pmatrix} \tag{64}$$

where I is the $r \times r$ identity matrix.

Exercise 36 For the drunkard's walk, show that

$$Q = \begin{pmatrix} 0 & 1/2 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 1/2 & 0 \end{pmatrix}, \quad R = \begin{pmatrix} 1/2 & 0 \\ 0 & 0 \\ 0 & 1/2 \end{pmatrix}, \quad I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{65}$$

Simple calculations show that for all $n \geq 1$

$$P^n = \begin{pmatrix} Q^n & R_n \\ 0 & I \end{pmatrix} \tag{66}$$

where R_n is a complicated matrix depending on Q and R.

Lemma 4 *As $n \to \infty$,*

$$(Q^n)_{ij} \to 0$$

for all absorbing states i, j.

Proof: for a transient state i, there is an absorbing state k, an integer n_i and $\delta_i > 0$ such that

$$p_{ik}(n_i) = \delta_i > 0 \tag{67}$$

Let $n = \max n_i$, and $\delta = \min \delta_i$, then for any $i \in T$, there is a state $k \in R$ such that

$$p_{ik}(n) \geq \delta \tag{68}$$

Hence for any $i \in T$,

$$\sum_{j \in T} Q_{ij}^n = 1 - \sum_{k \in R} P_{ik}^n = 1 - \sum_{k \in R} p_{ik}(n) \leq 1 - \delta \tag{69}$$

In particular this means that $Q_{ij}^n \leq 1 - \delta$ for all $i, j \in T$. So for all $i \in T$ we get

$$\sum_{j \in T} Q_{ij}^{2n} = \sum_{k \in T} Q_{ik}^n \sum_{j \in T} Q_{kj}^n \leq (1 - \delta) \sum_{k \in T} Q_{ik}^n \leq (1 - \delta)^2 \tag{70}$$

This iterates to give

$$\sum_{j \in T} Q_{ij}^{kn} \to 0 \quad \text{as } k \to \infty \tag{71}$$

for all $i \in T$. It remains to notice that

$$\sum_{j \in T} Q_{ij}^{m+1} = \sum_{k \in T} Q_{ik}^m \sum_{j \in T} Q_{kj} \leq \sum_{k \in T} Q_{ik}^m \tag{72}$$

and hence the sequence $\{\sum_{k\in T} Q_{ik}^m\}$ in monotone decreasing in m. Therefore

$$\sum_{j\in T} Q_{ij}^k \to 0 \quad \text{as } k \to \infty \tag{73}$$

for all $i \in T$, which proves the result.

QED

Notice what the result says: the probability of remaining in the transient states goes to zero, so eventually the chain must transition to the absorbing states. So the quantities of interest are related to the time (=number of steps) needed until the chain exits the transient states and enters the absorbing states, and the number of visits to other transient states.

Consider the equation

$$x = Qx \tag{74}$$

Applying Q to both sides we deduce that

$$x = Q^2 x \tag{75}$$

and iterating this leads to

$$x = Q^n x \tag{76}$$

for all n. Since $Q^n \to 0$ it follows that $x = 0$. Hence there is no nonzero solution of the equation $x = Qx$ and therefore the matrix $I-Q$ is non-singular and so invertible. Define the fundamental matrix

$$N = (I - Q)^{-1} \tag{77}$$

Note that

$$(I + Q + Q^2 + \cdots + Q^n)(I - Q) = I - Q^{n+1} \tag{78}$$

and letting $n \to \infty$ we deduce that

$$N = I + Q + Q^2 + \cdots \tag{79}$$

Theorem 5 *Let i, j be transient states. Then*

(1) N_{ij} *is the expected number of visits to state* j *starting from state* i *(counting initial state if* $i = j$*).*

(2) $\sum_j N_{ij}$ *is the expected number of steps of the chain, starting in state* i*, until it is absorbed.*

(3) define the $t \times r$ *matrix* $B = NR$*. Then* B_{ik} *is the probability that the chain is absorbed in state* k*, given that it started in state* i*.*

Proof: the chain starts at $X_0 = i$. Given a state $j \in T$, for $k \geq 0$ define indicator random variables as follows:

$$Y^{(k)} = \begin{cases} 1 & \text{if } X_k = j \\ 0 & \text{else} \end{cases} \tag{80}$$

Then for $k \geq 1$

$$\mathbb{E}Y^{(k)} = P(Y^{(k)} = 1) = P(X_k = j) = p_{ij}(k) = (Q^k)_{ij} \tag{81}$$

and for $k = 0$ we get $\mathbb{E}Y^{(0)} = \delta_{ij}$. Now the number of visits to the state j in the first n steps is $Y^{(0)} + Y^{(1)} + \cdots + Y^{(n)}$. Taking the expected value yields the sum

$$\delta_{ij} + Q_{ij} + (Q^2)_{ij} + \cdots + (Q^n)_{ij} = (I + Q + Q^2 + \cdots + Q^n)_{ij} \tag{82}$$

which converges to N_{ij} as $n \to \infty$. This proves (1). For (2), note that the sum of visits to all transient states is the total number of steps of the chain before it leaves the transient states. For (3), use $N = \sum Q^n$ to write

$$\begin{aligned} (NR)_{ik} &= \sum_{j \in T} N_{ij} R_{jk} \\ &= \sum_{j \in T} \sum_{n=0}^{\infty} (Q^n)_{ij} R_{jk} \\ &= \sum_{n=0}^{\infty} \sum_{j \in T} (Q^n)_{ij} R_{jk} \end{aligned} \tag{83}$$

and note that $\sum_{j \in T} (Q^n)_{ij} R_{jk}$ is the probability that the chain takes n steps to transient states before exiting to the absorbing state k. Since this is the only way that the chain can transition to k in $n+1$ steps, the result follows.

QED

Exercise 37 For the drunkard's walk,

$$Q^{2n+1} = 2^{-n}Q, \quad Q^{2n+2} = 2^{-n}Q^2 \tag{84}$$

and

$$N = \begin{pmatrix} 3/2 & 1 & 1/2 \\ 1 & 2 & 1 \\ 1/2 & 1 & 3/2 \end{pmatrix} \tag{85}$$

Also

$$B = NR = \begin{pmatrix} 3/4 & 1/4 \\ 1/2 & 1/2 \\ 1/4 & 3/4 \end{pmatrix} \tag{86}$$

Exercise 38 Rework the drunkard's walk, assuming that a step to the right has probability 1/3 and a step to the left has probability 2/3.

Exercise 39 [Snell and Grinstead] A city is divided into three areas $1, 2, 3$. It is estimated that amounts u_1, u_2, u_3 of pollution are emitted each day from these three areas. A fraction q_{ij} of the pollution from region i ends up the next day at region j. A fraction $q_i = 1 - \sum_j q_{ij} > 0$ escapes into the atmosphere. Let $w_i^{(n)}$ be the amount of pollution in area i after n days.
(a) Show that $w^{(n)} = u + uQ + \cdots + uQ^{n-1}$.
(b) Show that $w^{(n)} \to w$.
(c) Show how to determine the levels of pollution u which would result in a prescribed level w.

Exercise 40 [The gambler's ruin] At each play a gambler has probability p of winning one unit and probability $q = 1 - p$ of losing one unit. Assuming that successive plays of the game are independent, what is the probability that, starting with i units, the gambler's fortune will reach N before reaching 0? [Hint: define P_i to be the probability that the gambler's fortune reaches N before reaching 0 conditioned on starting in state i. By conditioning on the first step derive a recursion relation between P_i, P_{i+1} and P_{i-1}.]

2.3 Ergodic Markov chains

These are a kind of opposite to absorbing chains: the state never settles down to a fixed value but continues making jumps forever. As before the case is characterized by the transition matrix. Notation: for a matrix T write $T \geq 0$ if $T_{ij} \geq 0$ for all i, j and $T > 0$ if $T_{ij} > 0$ for all i, j.

Definition 6 *Let P be the transition matrix of a Markov chain.*

(1) The Markov chain is primitive *if there is an integer n such that $P^n > 0$.*

(2) The Markov chain is irreducible *if for all states i, j there is an integer $n(i, j)$ such that $p_{ij}(n(i, j)) > 0$.*

Exercise 41 Recall the balls in boxes model:

$$P = \begin{pmatrix} 0 & 1 & 0 \\ 1/4 & 1/2 & 1/4 \\ 0 & 1 & 0 \end{pmatrix} \tag{87}$$

Since

$$P^2 = \begin{pmatrix} 1/4 & 1/2 & 1/4 \\ 1/8 & 3/4 & 1/8 \\ 1/4 & 1/2 & 1/4 \end{pmatrix} \tag{88}$$

it follows that P is primitive.

Exercise 42 Define the two-state swapping chain:

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \tag{89}$$

Then $P^2 = I$ is the identity, hence for all $n \geq 1$

$$P^{2n} = I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad P^{2n+1} = P \tag{90}$$

So P is irreducible but not primitive.

Let e denote the vector in $\mathbb{R}^n$ with all entries 1, so

$$e = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \tag{91}$$

Theorem 7 [Perron-Frobenius] *Suppose P is a primitive $n \times n$ transition matrix. Then there is a unique strictly positive vector $w \in \mathbb{R}^n$ such that*

$$w^T P = w^T \tag{92}$$

and such that

$$P^k \to e\, w^T \quad as\ k \to \infty \tag{93}$$

Proof: we show that for all vectors $y \in \mathbb{R}^n$,

$$P^k y \to e\, w^T y \tag{94}$$

which is a positive multiple of the constant vector e. This implies the result.

Suppose first that $P > 0$ so that $p_{ij} > 0$ for all $i, j \in S$. Let $d > 0$ be the smallest entry in P (so $d \leq 1/2$). For any $y \in \mathbb{R}^n$ define

$$m_0 = \min_j\, y_j, \quad M_0 = \max_j\, y_j \tag{95}$$

and

$$m_1 = \min_j\, (Py)_j, \quad M_1 = \max_j\, (Py)_j \tag{96}$$

Consider $(Py)_i = \sum_j p_{ij} y_j$. This is maximized by pairing the smallest entry m_0 of y with the smallest entry d of p_{ij}, and then taking all other entries of y to be M_0. In other words,

$$\begin{aligned} M_1 &= \max_i (Py)_i \\ &= \max_i \sum_j p_{ij} y_j \\ &\leq (1-d) M_0 + d m_0 \end{aligned} \tag{97}$$

By similar reasoning,

$$m_1 = \min_i (Py)_i \geq (1-d)m_0 + dM_0 \tag{98}$$

Subtracting these bounds gives

$$M_1 - m_1 \leq (1-2d)(M_0 - m_0) \tag{99}$$

Now we iterate to give

$$M_k - m_k \leq (1-2d)^k (M_0 - m_0) \tag{100}$$

where again

$$M_k = \max_i (P^k y)_i, \quad m_k = \min_i (P^k y)_i \tag{101}$$

Furthermore the sequence $\{M_k\}$ is decreasing since

$$M_{k+1} = \max_i (PP^k y)_i = \max_i \sum_j p_{ij} (P^k y)_j \leq M_k \tag{102}$$

and the sequence $\{m_k\}$ is increasing for similar reasons. Therefore both sequences converge as $k \to \infty$, and the difference between them also converges to zero. Hence we conclude that the components of the vector $P^k y$ converge to a constant value, meaning that

$$P^k y \to m\, e \tag{103}$$

for some m. We can pick out the value of m with the inner product

$$m(e^T e) = e^T \lim_{k\to\infty} P^k y = \lim_{k\to\infty} e^T\, P^k y \tag{104}$$

Note that for $k \geq 1$,

$$e^T\, P^k y \geq m_k (e^T e) \geq m_1 (e^T e) = \min_i (Py)_i (e^T e)$$

Since P is assumed positive, if $y_i \geq 0$ for all i it follows that $(Py)_i > 0$ for all i, and hence $m > 0$.

Now define

$$w_j = \lim_{k\to\infty} P^k e_j / (e^T e) \tag{105}$$

where e_j is the vector with entry 1 in the j^{th} component, and zero elsewhere. It follows that $w_j > 0$ so w is strictly positive, and

$$P^k \to ew^T \tag{106}$$

By continuity this implies

$$\lim_{k\to\infty} P^k P = ew^T P \tag{107}$$

and hence $w^T P = w^T$. This proves the result in the case where $P > 0$.

Now turn to the case where P is primitive. Since P is primitive, there exists integer N such that

$$P^N > 0 \tag{108}$$

Hence by the previous result there is a strictly positive $w \in \mathbb{R}^n$ such that

$$P^{kN} \to ew^T \tag{109}$$

as $k \to \infty$, satisfying $w^T P^N = w^T$. It follows that $P^{N+1} > 0$, and hence there is also a vector v such that

$$P^{k(N+1)} \to ev^T \tag{110}$$

as $k \to \infty$, and $v^T P^{N+1} = v^T$. Considering convergence along the subsequence $kN(N+1)$ it follows that $w = v$, and hence

$$w^T P^{N+1} = v^T P^{N+1} = v^T = w^T = w^T P^N \tag{111}$$

and so

$$w^T P = w^T \tag{112}$$

The subsequence $P^{kN} y$ converges to $ew^T y$ for every y, and we want to show that the full sequence $P^m y$ does the same. For any $\epsilon > 0$ there is $K < \infty$ such that for all $k \geq K$ and all probability vectors y

$$\|(P^{kN} - ew^T)y\| \leq \epsilon \tag{113}$$

Let $m = kN + j$ where $j < N$, then for any probability vector y

$$\|(P^m - ew^T)y\| = \|(P^{kN+j} - ew^T)y\| = \|(P^{kN} - ew^T)P^j y\| \leq \epsilon \tag{114}$$

which proves convergence along the full sequence.

QED

Note that as a corollary of the Theorem we deduce that the vector w is the unique (up to scalar multiples) solution of the equation

$$w^T P = w^T \tag{115}$$

Also since $v^T e = \sum v_i = 1$ for a probability vector v, it follows that

$$v^T P^n \to w^T \tag{116}$$

for any probability vector v.

Exercise 43 Recall the balls in boxes model:

$$P = \begin{pmatrix} 0 & 1 & 0 \\ 1/4 & 1/2 & 1/4 \\ 0 & 1 & 0 \end{pmatrix} \tag{117}$$

We saw that P is primitive. Solving the equation $w^T P = w^T$ yields the solution

$$w^T = (1/6, 2/3, 1/6) \tag{118}$$

Furthermore we can compute

$$P^{10} = \begin{pmatrix} 0.167 & 0.666 & 0.167 \\ 0.1665 & 0.667 & 0.1665 \\ 0.167 & 0.666 & 0.167 \end{pmatrix} \tag{119}$$

showing the rate of convergence.

Aside on convergence [Seneta]: another way to express the Perron-Frobenius result is to say that for the matrix P, 1 is the largest eigenvalue (in absolute value) and w is the unique eigenvector (up to scalar multiples). Let λ_2 be the second largest eigenvalue of P so that $1 > |\lambda_2| \geq |\lambda_i|$. Let m_2 be the multiplicity of λ_2. Then the following estimate holds: there is $C < \infty$ such that for all $n \geq 1$

$$\|P^n - ew^T\| \leq C\, n^{m_2 - 1}\, |\lambda_2|^n \tag{120}$$

So the convergence $P^n \to ew^T$ is exponential with rate determined by the first spectral gap.

Concerning the interpretation of the result. Suppose that the distribution of X_0 is

$$P(X_0 = i) = \alpha_i \tag{121}$$

for all $i \in S$. Then

$$P(X_k = j) = \sum_i P(X_k = j | X_0 = i)\, P(X_0 = i) = \sum_i (P^k)_{ij}\, \alpha_i = (\alpha^T P^k)_j \tag{122}$$

where α is the vector with entries α_i. Using our Theorem we deduce that

$$P(X_k = j) \to w_j \tag{123}$$

as $k \to \infty$ for any initial distribution α. Furthermore if $\alpha = w$ then $\alpha^T P^k = w^T P^k = w^T$ and therefore

$$P(X_k = j) = w_j \tag{124}$$

for all k. So w is called the *equilibrium* or *stationary* distribution of the chain. The Theorem says that the state of the chain rapidly forgets its initial distribution and converges to the stationary value.

Now suppose the chain is irreducible but not primitive. Then we get a similar but weaker result.

Theorem 8 *Let P be the transition matrix of an irreducible Markov chain. Then there is a unique strictly positive probability vector w such that*

$$w^T P = w^T \tag{125}$$

Furthermore

$$\frac{1}{n+1}\Big(I + P + P^2 + \cdots + P^n\Big) \to ew^T \tag{126}$$

as $n \to \infty$.

Proof: define

$$Q = \frac{1}{2}I + \frac{1}{2}P \tag{127}$$

Then Q is a transition matrix. Also

$$2^n Q^n = \sum_{k=0}^{n} \binom{n}{k} P^k \tag{128}$$

Because the chain is irreducible, for all pairs of states i, j there is an integer $n(i,j)$ such that $(P^{n(i,j)})_{ij} > 0$. Let $n = \max n(i,j)$, then for all i, j we have

$$2^n (Q^n)_{ij} = \sum_{k=0}^{n} \binom{n}{k} (P^k)_{ij} \geq \binom{n}{n(i,j)} (P^{n(i,j)})_{ij} > 0 \tag{129}$$

and hence Q is primitive. Let w be the unique stationary vector for Q then

$$w^T Q = w^T \leftrightarrow w^T P = w^T \tag{130}$$

which shows existence and uniqueness for P.

Let $W = ew^T$ then a calculation shows that for all n

$$\Big(I + P + P^2 + \cdots + P^{n-1}\Big)(I - P + W) = I - P^n + nW \tag{131}$$

Note that $I - P + W$ is invertible: indeed if $y^T(I - P + W) = 0$ then

$$y^T - y^T P + (y^T e)w = 0 \tag{132}$$

Multiply by e on the right and use $Pe = e$ to deduce

$$y^T e - y^T Pe + (y^T e)(w^T e) = (y^T e)(w^T e) = 0 \tag{133}$$

Since $w^T e = 1 > 0$ it follows that $y^T e = 0$ and so $y^T - y^T P = 0$. By uniqueness this means that y is a multiple of w, but then $y^T e = 0$ means that $y = 0$. Therefore $I - P + W$ is invertible, and so

$$I + P + P^2 + \cdots + P^{n-1} = (I - P^n + nW)(I - P + W)^{-1} \tag{134}$$

Now $WP = W = W^2$ hence

$$W(I - P + W) = W \implies W = W(I - P + W)^{-1} \tag{135}$$

therefore

$$I + P + P^2 + \cdots + P^{n-1} = (I - P^n)(I - P + W)^{-1} + nW \tag{136}$$

and so

$$\frac{1}{n}\Big(I + P + P^2 + \cdots + P^{n-1}\Big) = W + \frac{1}{n}(I - P^n)(I - P + W)^{-1} \tag{137}$$

It remains to show that the norm of the matrix $(I - P^n)(I - P + W)^{-1}$ is bounded as $n \to \infty$, or equivalently that $\|(I - P^n)\|$ is uniformly bounded. This follows from the bound

$$\|P^n z\| \leq \sum_{ij} (P^n)_{ij}|z_j| = \sum_j |z_j| \tag{138}$$

Therefore $\frac{1}{n}(I - P^n)(I - P + W)^{-1} \to 0$ and the result follows,

QED

This Theorem allows the following interpretation: for an irreducible chain, w_j is the long-run fraction of time the chain spends in state j.

Exercise 44 A transition matrix is doubly stochastic if each column sum is 1. Find the stationary distribution for a doubly stochastic chain with M states.

Exercise 45 [Ross] Trials are performed in sequence. If the last two trials were successes, then the next trial is a success with probability 0.8; otherwise the next trial is a success with probability 0.5. In the long run, what proportion of trials are successes?

Exercise 46 Let $\{X_n\}$ be a primitive finite state Markov chain with transition matrix P and stationary distribution w. Define the process $\{Y_n\}$ by $Y_n = (X_{n-1}, X_n)$. Show that $\{Y_n\}$ is a Markov chain, and compute

$$\lim_{n\to\infty} P(Y_n = (i, j)) \tag{139}$$

Definition 9 *Consider an irreducible Markov chain.*

(1) starting in state i, m_{ij} is the expected number of steps to visit state j for the first time (by convention $m_{ii} = 0$)

(2) starting in state i, r_i is the expected number of steps for the first return to state i

(3) the fundamental matrix is $Z = (I - P + W)^{-1}$

Theorem 10 *Let w be the stationary distribution of an irreducible Markov chain. Then for all states $i, j \in S$,*

$$r_i = \frac{1}{w_i}, \quad m_{ij} = \frac{z_{jj} - z_{ij}}{w_j} \tag{140}$$

where z_{ij} is the (i, j) entry of the fundamental matrix Z.

Proof: let M be the matrix with entries $M_{ij} = m_{ij}$, let E be the matrix with entries $E_{ij} = 1$, and let D be the diagonal matrix with diagonal entries $D_{ii} = r_i$. For all $i \neq j$,

$$m_{ij} = p_{ij} + \sum_{k \neq j} p_{ik}(m_{kj} + 1) = 1 + \sum_{k \neq j} p_{ik} m_{kj} \tag{141}$$

For all i,

$$r_i = \sum_k p_{ik}(m_{ki} + 1) = 1 + \sum_k p_{ik} m_{ki} \tag{142}$$

Thus for all i, j,

$$M_{ij} = 1 + \sum_{k \neq j} p_{ik} M_{kj} - D_{ij} \tag{143}$$

which can be written as the matrix equation

$$M = E + PM - D \tag{144}$$

Multiplying on the left by w^T and noting that $w^T = w^T P$ gives

$$0 = w^T E - w^T D \tag{145}$$

The i^{th} component of the right side is $1 - w_i r_i$, hence this implies that for all i

$$r_i = \frac{1}{w_i} \tag{146}$$

Recall the definition of the matrix $Z = (I - P + W)^{-1}$, and vector $e = (1, 1, \ldots, 1)^T$. Since $Pe = We = e$ it follows that $(I - P + W)e = e$ and hence $Ze = e$ and $ZE = E = ee^T$. Furthermore $w^T P = w^T W = w^T$ and so similarly $w^T Z = w^T$ and $W = WZ$. Therefore from (144),

$$Z(I - P)M = ZE - ZD = E - ZD \tag{147}$$

Since $Z(I - P) = I - ZW = I - W$ this yields

$$M = E - ZD + WM \tag{148}$$

The (i, j) component of this equation is

$$m_{ij} = 1 - z_{ij} r_j + (w^T M)_j \tag{149}$$

Setting $i = j$ gives $0 = 1 - z_{jj} r_j + (w^T M)_j$, hence

$$m_{ij} = (z_{jj} - z_{ij}) r_j = \frac{z_{jj} - z_{ij}}{w_j} \tag{150}$$

QED

2.4 Classification of finite-state Markov chains

Say that states i and j *intercommunicate* if there are integers n, m such that $p_{ij}(n) > 0$ and $p_{ji}(m) > 0$. In other words it is possible to go from each state to the other. A class of states C in S is called *closed* if $p_{ij} = 0$ whenever $i \in C$ and $j \notin C$. The class is called *irreducible* if all states in C intercommunicate.

Note that if the chain is irreducible then all states intercommunicate and hence for all i, j there is an integer n such that $p_{ij}(n) > 0$.

Theorem 11 *The state space S can be partitioned uniquely as*

$$S = T \cup C_1 \cup C_2 \cup \cdots \tag{151}$$

where T is the set of all transient states, and each class C_i is closed and irreducible.

Proof: (later)

If the chain starts with $X_0 \in C_i$ then it stays in C_i forever. If it starts with $X_0 \in T$ then eventually it enters one of the classes C_i and stays there forever.

Exercise 47 Determine the classes of the chain:

$$P = \begin{pmatrix} 1/2 & 1/2 & 0 & 0 & 0 & 0 \\ 1/4 & 3/4 & 0 & 0 & 0 & 0 \\ 1/4 & 1/4 & 1/4 & 1/4 & 0 & 0 \\ 1/4 & 0 & 1/4 & 1/4 & 0 & 1/4 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \end{pmatrix} \tag{152}$$

3 Existence of Markov Chains

We have been making an implicit assumption about the Markov chains, and now it is time to address this openly. Namely, we have been assuming that the statement "let $X_0, X_1, \ldots$ be random variables ..." makes sense. Because we use the joint distribution for the X_i we need to know that they all exist as random variables on the *same* underlying sample space, with the same underlying probability function. If we use a finite number of X_i then there is no problem, we are still working with simple random variables. But we want to consider an infinite number at the same time: for example what does it mean to say (as we did) "$\lim_{n\to\infty} P(X_n = j)$"? To make sense of this at the very least we need all X_n to be defined as random variables on the same sample space. So let's do that. We will handle the case of a countable state space at the same time.

3.1 Sample space for Markov chain

Let S be the countable state space of a Markov chain. Let α_i and p_{ij} be respectively the initial distribution of X_0 and the transition matrix of the chain. This means that

$$P(X_0 = i) = \alpha_i \quad \text{for all } i \in S \tag{153}$$

and also that

$$P(X_{n+1} = j \,|\, X_n = i) = p_{ij} \quad \text{for all } i, j \in S, \text{ all } n \geq 1 \tag{154}$$

The joint distribution of the X_i follows from this: for example

$$\begin{aligned} P(X_2 = k, X_1 = j, X_0 = i) &= P(X_2 = k \,|\, X_1 = j)\, P(X_1 = j \,|\, X_0 = i)\, P(X_0 = i) \\ &= p_{jk}\, p_{ij}\, \alpha_i \end{aligned}$$

In general we have

$$P(X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n) = \alpha_{i_0}\, p_{i_0 i_1}\, \ldots\, p_{i_{n-1} i_n} \tag{155}$$

We want a sample space Ω with a probability function P defined on it so that for all n

$$X_n : \Omega \to S \tag{156}$$

and so that the joint pmf is given as above.

We take $\Omega = [0,1]$, that is the closed interval of real numbers. We take the probability function P to be the 'usual' length function (more about this shortly). So an event in Ω is a subset of $[0,1]$, say $A \subset [0,1]$, and the probability of A is the 'length' of A. For example if $A = (a,b]$ then

$$P(A) = \text{length}(a,b] = |b-a| \tag{157}$$

Of course we need to extend the notion of length to more general subsets. But assuming that there is such an extension then we define the random variables as follows.

a) Partition $[0,1]$ into S subintervals, call them

$$\{I_i^{(0)}\} \quad i \in S$$

with $\text{length}(I_i^{(0)}) = \alpha_i$ for each i. Since $\sum_i \alpha_i = 1$ this covers the whole interval.

b) Partition each such interval $I_i^{(0)}$ into further subintervals

$$\{I_{i,j}^{(1)}\}_{j \in S}$$

so that for every $i,j \in S$ we have $\text{length}(I_{i,j}^{(1)}) = \alpha_i\, p_{ij}$. Again note that $\sum_j p_{ij} = 1$ so we are covering the whole interval $I_i^{(0)}$.

c) Inductively partition $[0,1]$ into intervals $\{I_{i_0,i_1,\ldots,i_n}^{(n)}\}$ such that they are nested according to $I_{i_0,i_1,\ldots,i_n}^{(n)} \subset I_{i_0,i_1,\ldots,i_{n-1}}^{(n-1)}$, and so that their lengths are given by $\text{length}(I_{i_0,i_1,\ldots,i_n}^{(n)}) = \alpha_{i_0}\, p_{i_0 i_1} \ldots p_{i_{n-1} i_n}$.

d) Define X_n by $X_n(x) = i_n$ if $x \in I_{i_0,i_1,\ldots,i_n}^{(n)}$ for some choice of $i_0, i_1, \ldots, i_{n-1}$.

It remains to verify that X_n have the required joint distribution. First note that

$$P(X_0 = i) = P(x \in I_i^{(0)}) = \text{length}\, I_i^{(0)} = \alpha_i \tag{158}$$

Then note that

$$P(X_1 = j, X_0 = i) = P(x \in I_{i,j}^{(1)}) = \text{length}\, I_{i,j}^{(1)} = p_{ij}\, \alpha_i \tag{159}$$

The general case follows just as easily.

To summarize: we have exhibited the random variables $X_0, X_1, \ldots$ as functions on the same probability space namely $\Omega = [0,1]$, equipped with the probability function defined by the usual length.

3.2 Lebesgue measure

To complete the demonstration we need to extend the notion of length to include more general subsets of $[0, 1]$. Why is this? Suppose we want to calculate

$$P(X_n = i \text{ infinitely often})$$

This event can be written as

$$\{X_n = i \text{ infinitely often}\} = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} \{X_k = i\} \tag{160}$$

On the right side we have an infinite union of an infinite intersection of intervals. What is the length of such a set? Clearly we need to extend the notion of length.

The correct extension is called the Lebesgue measure, or just the measure. We will delve into this shortly but here we just note that it is possible to do it in such a way that we can consistently assign a probability to the right side above, and in fact make sense of any such complicated event that might arise in the study of our Markov chains.

4 Discrete-time Markov chains with countable state space

Moving from a finite state space to an infinite but countable state space leads to novel effects and a broader class of applications. The basic setup is the same as before: a finite or countably infinite state space S, a sequence of S-valued random variables $\{X_0, X_1, \dots\}$, and a set of transition probabilities $\{p_{ij}\}$ for each pair of states $i, j \in S$. The Markov property is the same:

$$P(X_n = y \,|\, X_0 = x_0, \dots, X_{n-1} = x_{n-1}) = P(X_n = y \,|\, X_{n-1} = x_{n-1}) \quad (161)$$

for all $n \geq 1$ and all states $y, x_0, \dots, x_{n-1} \in S$. Also the transition 'matrix' satisfies

$$\sum_{j \in S} p_{ij} = 1 \quad \text{for all } i \in S \quad (162)$$

4.1 Some motivating examples

The one-dimensional random walk has state space $\mathbb{Z} = \{\dots, -1, 0, 1, \dots\}$, and transition probabilities

$$p_{ij} = \begin{cases} p & \text{if } j = i + 1 \\ q & \text{if } j = i - 1 \\ 0 & \text{else} \end{cases} \quad (163)$$

So at each time unit the chain takes one step either to the left or the right, with probabilities q and p respectively. The chain has no absorbing states so it keeps moving forever. Interesting questions are whether it wanders off to infinity or stays around its starting position, and also rates of various long-run behavior.

A second important example is the branching process. This describes the growth of a population. The state is the number of individuals in successive generations. This changes because of the random number of births and deaths. In the simplest case each individual produces a random number of individuals B in the next generation:

$$X_{n+1} = \sum_{i=1}^{X_n} B_i$$

There is one absorbing state corresponding to extinction. So the interesting question is whether the chain reaches extinction or keeps growing forever – or more precisely, the probability that it ever reaches extinction.

4.2 Classification of states

Define

$$f_{ij}(n) = P(X_1 \neq j, X_2 \neq j, \ldots, X_{n-1} \neq j, X_n = j \,|\, X_0 = i) \tag{164}$$

to be the probability that starting in state i the chain first visits state j after n steps. Define

$$f_{ij} = \sum_{n=1}^{\infty} f_{ij}(n) \tag{165}$$

This is the probability that the chain eventually visits state j starting in state i.

Definition 12 *The state j is* persistent *if $f_{jj} = 1$. The state j is* transient *if $f_{jj} < 1$.*

There is a further separation of persistent states which occurs for infinite state space.

Definition 13 *The* mean return time μ_j *of state j is*

$$\mu_j = \begin{cases} \sum_{n=1}^{\infty} n f_{jj}(n) & \text{if } j \text{ is persistent} \\ \infty & \text{if } j \text{ is transient} \end{cases} \tag{166}$$

Note that μ_j may be finite or infinite for a persistent state (this is what we called r_j for the finite state space).

Definition 14 *The persistent state j is* null-persistent *if $\mu_j = \infty$, and it is* non-null persistent *if $\mu_j < \infty$.*

So there are three types of states in a Markov chain: transient, null persistent and non-null persistent. This is the classification of states.

Exercise 48 Define generating functions

$$P_{ij}(s) = \sum_{n=0}^{\infty} s^n p_{ij}(n), \quad F_{ij}(s) = \sum_{n=0}^{\infty} s^n f_{ij}(n) \tag{167}$$

with the conventions $p_{ij}(0) = \delta_{ij}$ and $f_{ij}(0) = 0$. Show that

$$P_{ii}(s) = 1 + F_{ii}(s)\, P_{ii}(s) \tag{168}$$

Show that state i is persistent if and only if $\sum_n p_{ii}(n) = \infty$.
[Hint: recall Abel's theorem: if $a_n \geq 0$ for all n and $\sum_n a_n s^n$ is finite for all $|s| < 1$, then

$$\lim_{s \uparrow 1} \sum_{n=0}^{\infty} a_n s^n = \sum_{n=0}^{\infty} a_n \tag{169}$$

4.3 Classification of Markov chains

Say that states i and j *intercommunicate* if there are integers n, m such that $p_{ij}(n) > 0$ and $p_{ji}(m) > 0$. In other words it is possible to go from each state to the other.

Theorem 15 *Let i, j intercommunicate, then they are either both transient, both null persistent or both non-null persistent.*

Proof: Since i, j intercommunicate there are integers n, m such that

$$h = p_{ij}(n) p_{ji}(m) > 0 \tag{170}$$

Hence for any r,

$$p_{ii}(n + m + r) \geq p_{ij}(n) p_{jj}(r) p_{ji}(m) = h\, p_{jj}(r) \tag{171}$$

Sum over r to deduce

$$\sum_k p_{ii}(k) \geq \sum_r p_{ii}(n + m + r) \geq h \sum_r p_{jj}(r) \tag{172}$$

Therefore either both sums are finite or both are infinite, hence either both states are transient or both are persistent. [Omit the proof about null persistent and non-null persistent].

QED

Exercise 49 Suppose that state i is transient, and that state i is accessible from state j. Show that $p_{ij}(n) \to 0$ as $n \to \infty$.

A class of states C in S is called *closed* if $p_{ij} = 0$ whenever $i \in C$ and $j \notin C$. The class is called *irreducible* if all states in C intercommunicate.

This usage is consistent with the finite state space case – if the chain is an irreducible class then all states intercommunicate and hence for all i, j there is an integer n such that $p_{ij}(n) > 0$.

Theorem 16 *The state space S can be partitioned uniquely as*

$$S = T \cup C_1 \cup C_2 \cup \cdots \tag{173}$$

where T is the set of all transient states. Each class C_i is closed and irreducible, and contains persistent states. Either all states in C_i are null persistent, or all states in C_i are non-null persistent.

Proof: mostly clear except maybe that C_i is closed. So suppose indeed that there are states $i \in C$ and $j \notin C$ with $p_{ij} > 0$. Since i is not accessible from j, it follows that $p_{ji}(n) = 0$ for all $n \geq 1$. Hence

$$1 - f_{ii} = P(X_n \neq i \text{ for all } n \geq 1 | X_0 = i) \geq P(X_1 = j | X_0 = i) = p_{ij} \tag{174}$$

which means that $f_{ii} < 1$, but this contradicts the persistence of state i.

QED

If the chain starts with $X_0 \in C_i$ then it stays in C_i forever. If it starts with $X_0 \in T$ then eventually it enters one of the classes C_i and stays there forever. We will restrict attention to irreducible chains now. The first issue is to determine which of the three types of chains it may be. Recall the

definition of a stationary distribution of the chain: this is a distribution π such that $\pi_i \geq 0$ and $\sum_i \pi_i = 1$, and for all $j \in S$,

$$\pi_j = \sum_i \pi_i \, p_{ij} \tag{175}$$

(it is conventional to use π for discrete chains, we do so from now on).

Theorem 17 *Consider an irreducible chain with transition probabilities p_{ij}.*

(1) The chain is transient if and only if $\sum_n p_{jj}(n) < \infty$ for any (and hence all) states $j \in S$.

(2) The chain is persistent if and only if $\sum_n p_{jj}(n) = \infty$ for any (and hence all) states $j \in S$.

(3) There is a positive vector x satisfying $x^T = x^T P$, that is

$$x_j = \sum_{i \in S} x_i p_{ij} \tag{176}$$

The chain is non-null persistent if and only if $\sum_i x_i < \infty$.

(4) If the chain has a stationary distribution then it is non-null persistent.

In case (3) we can normalize x by dividing by $\sum_i x_i$ and hence recover the stationary distribution π. Thus as a Corollary we see that a chain has a stationary distribution if and only if it is non-null persistent.
Proof: items (1), (2) were shown in the exercises. For item (4), suppose that π is a stationary distribution and note that if the chain is transient then $p_{ij}(n) \to 0$ for all states i, j and hence

$$\pi_j = \sum_i \pi_i p_{ij}(n) \to 0 \tag{177}$$

(this needs a little care when the sum is infinite – see Comment after the proof of Theorem 20).

Turn to item (3). Fix a state k, and let T_k be the time (number of steps) until the first return to state k. Let $N_i(k)$ be the time spent in state i during this sojourn, or more precisely,

$$N_i(k) = \sum_{n=1}^{\infty} 1_{\{X_n = i\} \cap \{T_k \geq n\}} \tag{178}$$

It follows that $N_k(k) = 1$. Hence

$$T_k = \sum_{i \in S} N_i(k) \tag{179}$$

By definition

$$\mu_k = \mathbb{E}[T_k \,|\, X_0 = k] \tag{180}$$

Define $\rho_i(k) = \mathbb{E}[N_i(k) \,|\, X_0 = k]$ then

$$\mu_k = \sum_{i \in S} \rho_i(k) \tag{181}$$

It turns out that $\rho_i(k)$ will yield the components of the stationary distribution.

First we claim that $\rho_i(k) < \infty$. To see this, write

$$L_{ki}(n) = \mathbb{E}[1_{\{X_n = i\} \cap \{T_k \geq n\}}] = P(\{X_n = i\} \cap \{T_k \geq n\}) \tag{182}$$

so that $\mathbb{E}[N_i(k)] = \sum_{n=1}^{\infty} L_{ki}(n)$. Now

$$f_{kk}(m+n) \geq L_{ki}(n)\, f_{ik}(m) \tag{183}$$

Choose m so that $f_{ik}(m) > 0$ (chain is irreducible) then

$$L_{ki}(n) \leq \frac{f_{kk}(m+n)}{f_{ik}(m)} \tag{184}$$

Hence

$$\begin{aligned} \rho_i(k) &= \mathbb{E}[N_i(k) \,|\, X_0 = k] \\ &= \sum_{n=1}^{\infty} L_{ki}(n) \\ &\leq \sum_{n=1}^{\infty} \frac{f_{kk}(m+n)}{f_{ik}(m)} \\ &\leq \frac{1}{f_{ik}(m)} < \infty \end{aligned} \tag{185}$$

Next we claim that ρ_i is stationary. Note that for $n \geq 2$,

$$L_{ki}(n) = \sum_{j \neq k} L_{kj}(n-1) p_{ji} \tag{186}$$

Hence

$$\begin{aligned} \rho_i(k) &= L_{ki}(1) + \sum_{n=2}^{\infty} L_{ki}(n) \\ &= p_{ki} + \sum_{j \neq k} \sum_{n=2}^{\infty} L_{kj}(n-1) p_{ji} \\ &= p_{ki} + \sum_{j \neq k} \rho_j(k) p_{ji} \\ &= \sum_{j \in S} \rho_j(k) p_{ji} \end{aligned} \tag{187}$$

where in the last equality we used $\rho_k(k) = 1$ (true because $N_k(k) = 1$). Hence $\rho_i(k)$ is stationary.

So for every $k \in S$ we have a stationary vector. The chain is non-null persistent if and only if $\mu_k < \infty$, in which case we can normalize to get a probability distribution. It remains to show that this distribution is unique and positive. For positivity, suppose that $\pi_j = 0$ for some j, then

$$0 = \sum_i \pi_i p_{ij}(n) \geq \pi_i p_{ij}(n) \tag{188}$$

for all i and n. Hence if i and j communicate then $\pi_i = 0$ also. But the chain is irreducible, hence $\pi_i = 0$ for all $i \in S$. For uniqueness, use the following Theorem 20.

QED

Definition 18 *The state i is aperiodic if*

$$1 = \gcd\{n \,|\, p_{ii}(n) > 0\} \tag{189}$$

If a chain is irreducible then either all states are aperiodic or none are.

Exercise 50 Construct the coupled chain $Z = (X, Y)$ consisting of the ordered pair of independent chains with the same transition matrix P. If

X and Y are irreducible and aperiodic, show that Z is also irreducible and aperiodic. [Hint: use the following theorem: "An infinite set of integers which is closed under addition contains all but a finite number of positive multiples of its greatest common divisor" [Seneta]].

Definition 19 *An irreducible, aperiodic, non-null persistent Markov chain is called* ergodic.

Theorem 20 *For an ergodic chain,*

$$p_{ij}(n) \to \pi_j = \frac{1}{\mu_j} \tag{190}$$

as $n \to \infty$, for all $i, j \in S$.

Proof: Use the coupled chain described above. It follows that Z is also ergodic. Suppose that $X_0 = i$ and $Y_0 = j$, so $Z_0 = (i, j)$. Choose $s \in S$ and define

$$T = \min\{n \geq 1 \,|\, Z_n = (s, s)\} \tag{191}$$

This is the 'first passage time' to state (s, s). Hence

$$\begin{aligned} p_{ik}(n) &= P(X_n = k) \\ &= P(X_n = k, T \leq n) + P(X_n = k, T > n) \\ &= P(Y_n = k, T \leq n) + P(X_n = k, T > n) \\ &\leq P(Y_n = k) + P(T > n) \\ &= p_{jk}(n) + P(T > n) \end{aligned} \tag{192}$$

where we used the fact that if $T \leq n$ then X_n and Y_n have the same distribution. This and related inequality with i and j switched gives

$$|p_{ik}(n) - p_{jk}(n)| \leq P(T > n) \tag{193}$$

But since Z is persistent, $P(T < \infty) = 1$ and hence

$$|p_{ik}(n) - p_{jk}(n)| \to 0 \tag{194}$$

as $n \to \infty$. Furthermore, let π be a stationary distribution for X, then

$$\pi_k - p_{jk}(n) = \sum_i \pi_i \left(p_{ik}(n) - p_{jk}(n) \right) \to 0 \tag{195}$$

as $n \to \infty$. Together (194) and (195) show that $p_{jk}(n)$ converges as $n \to \infty$ to a limit which does not depend on j or on the choice of stationary distribution for X. Hence there is a unique stationary distribution for X. Finally from the previous Theorem we had $\rho_k(k) = 1$ and so

$$\pi_k = \frac{\rho_k(k)}{\sum_j \rho_j(k)} = \frac{1}{\mu_k} \tag{196}$$

QED

Comment: the limit in (195) needs to be justified. Let F be a finite subset of S then

$$\begin{aligned} \sum_i \pi_i \left| p_{ik}(n) - p_{jk}(n) \right| &\leq \sum_{i \in F} \pi_i \left| p_{ik}(n) - p_{jk}(n) \right| + 2 \sum_{i \notin F} \pi_i \\ &\to 2 \sum_{i \notin F} \pi_i \end{aligned} \tag{197}$$

as $n \to \infty$. Now take an increasing sequence of finite subsets F_a converging to S, and use $\sum_{i \in S} \pi_i = 1$ to conclude that $\sum_{i \notin F_a} \pi_i \to 0$.

Exercise 51 Show that the one-dimensional random walk is transient if $p \neq 1/2$. If $p = 1/2$ (called the symmetric random walk) show that the chain is null persistent. [Hint: use Stirling's formula for the asymptotics of $n!$:

$$n! \sim n^n \, e^{-n} \sqrt{2\pi n} \tag{198}$$

Exercise 52 Consider a Markov chain on the set $S = \{0, 1, 2, \dots\}$ with transition probabilities

$$p_{i,i+1} = a_i, \quad p_{i,0} = 1 - a_i$$

for $i \geq 0$, where $\{a_i \,|\, i \geq 0\}$ is a sequence of constants which satisfy $0 < a_i < 1$ for all i. Let $b_0 = 1$, $b_i = a_0 a_1 \cdots a_{i-1}$ for $i \geq 1$. Show that the chain is

(a) persistent if and only if $b_i \to 0$ as $i \to \infty$
(b) non-null persistent if and only if $\sum_i b_i < \infty$,

and write down the stationary distribution if the latter condition holds.

Let A and β be positive constants and suppose that $a_i = 1 - Ai^{-\beta}$ for all large values of i. Show that the chain is

(c) transient if $\beta > 1$
(d) non-null persistent if $\beta < 1$. Finally, if $\beta = 1$ show that the chain is

(e) non-null persistent if $A > 1$
(f) null persistent if $A \leq 1$.

Exercise 53 For a branching process the population after n steps can be written as

$$X_n = \sum_{i=1}^{X_{n-1}} Z_i \tag{199}$$

where $X_0 = 1$, and where Z_i is the number of offspring of the i^{th} individual of the $(n-1)^{st}$ generation. It is assumed that all the variables Z_i are IID. Let π_0 be the probability that the population dies out,

$$\pi_0 = \lim_{n \to \infty} P(X_n = 0 \,|\, X_0 = 1) \tag{200}$$

Show that π_0 is the smallest positive number satisfying the equation

$$\pi_0 = \sum_{j=0}^{\infty} \pi_0^j \, P(Z = j) \tag{201}$$

[Hint: define the generating functions $\phi(s) = \mathbb{E}s^Z$ and $\phi_n(s) = \mathbb{E}s^{X_n}$ for $s > 0$. Show that $\phi_{n+1}(s) = \phi(\phi_n(s))$ and deduce that π_0 is a fixed point of ϕ.]

4.4 Time reversible Markov chains

Consider an ergodic chain $\{\dots, X_{n-1}, X_n, \dots\}$ with transition probabilities p_{ij} and stationary distribution π_j. We have

$$p_{ij} = P(X_n = j \,|\, X_{n-1} = i) \tag{202}$$

Now consider the reversed chain, where we run the sequence backwards: $\{\dots, X_n, X_{n-1}, \dots\}$. The transition matrix is

$$\begin{aligned} q_{ij} &= P(X_{n-1} = j \,|\, X_n = i) \\ &= \frac{P(X_{n-1} = j, X_n = i)}{P(X_n = i)} \\ &= P(X_n = i \,|\, X_{n-1} = j)\frac{P(X_{n-1} = j)}{P(X_n = i)} \\ &= p_{ji}\,\frac{P(X_{n-1} = j)}{P(X_n = i)} \end{aligned} \tag{203}$$

Assume that the original chain is in its stationary distribution so that $P(X_n = i) = \pi_i$ for all i, then this is

$$q_{ij} = p_{ji}\,\frac{\pi_j}{\pi_i} \tag{204}$$

Definition 21 *The Markov chain is* reversible *if $q_{ij} = p_{ij}$ for all $i, j \in S$.*

The meaning of this equation is that the chain "looks the same" when it is run backwards in time (in its stationary distribution). So you cannot tell whether a movie of the chain is running backwards or forwards in time. Equivalently, for all $i, j \in S$

$$\pi_i p_{ij} = \pi_j p_{ji} \tag{205}$$

The main advantage of this result is that these equations are much easier to solve than the original defining equations for π. There is a nice result which helps here.

Lemma 22 *Consider a non-null persistent Markov chain with transition probabilities p_{ij}. Suppose there is a positive vector $x_j > 0$ with $\sum_j x_j < \infty$, such that for all $i, j \in S$*

$$x_i p_{ij} = x_j p_{ji} \tag{206}$$

Then the chain is time reversible and x_j is a multiple of the stationary distribution.

So this result says that if you can find a positive solution of the simpler equation then you have solved for the stationary distribution.

Exercise 54 A total of m white and m black balls are distributed among two boxes, with m balls in each box. At each step, a ball is randomly selected fro each box and the two selected balls are exchanged and put back in the boxes. Let X_n be the number of white balls in the first box. Show that the chain is time reversible and find the stationary distribution.

The quantity $\pi_i p_{ij}$ has another interpretation: it is the rate of jumps of the chain from state i to state j. More precisely, it is the long-run average rate at which the chain makes the transition between these states:

$$\lim_{n\to\infty} P(X_n = i\,,\, X_{n+1} = j) = \pi_i p_{ij} \tag{207}$$

This often helps to figure out if a chain is reversible.

Exercise 55 Argue that any Markov chain on $\mathbb{Z}$ which makes jumps only between nearest neighbor sites is reversible.

Exercise 56 Consider a Markov chain on a finite graph. The states are the vertices, and jumps are made along edges connecting vertices. If the chain is at a vertex with n edges, then at the next step it jumps along an edge with probability $1/n$. Argue that the chain is reversible, and find the stationary distribution.

5 Probability triples

5.1 Rules of the road

Why do we need to go beyond elementary probability theory? We already saw the need even for finite state Markov chains. Here is an analogy. When you drive along the road you operate under the assumption that the road will continue after the next bend (even though you cannot see it before getting there). So you can rely on an *existence theorem*, namely that the road system of the country is constructed in such a way that roads do not simply 'end' without warning, and that you will not drive over a cliff if you follow the road. Another way of saying this is that you trust that by following the rules of the road you can proceed toward your destination. Of course whether you reach your destination may depend on your map-reading ability and skill in following directions, but that is another matter!

So let's apply this analogy to probability theory. Our 'rules of the road' are the basic definitions and operations introduced before, including events, probabilities, random variables and so on. Our destination may be quite complicated, for example: "what is the probability that the pattern HHT-THH will appear infinitely often if a coin is repeatedly tossed?". In order to answer this question we will take our simple rules and push them beyond the limits of elementary probability. So we will not be able to use mental models (imagine tossing infinitely many coins!) or rely on our intuition about what is or is not reasonable. We want to know that our methods will make sense out there, and that there will be a sample space and an event for which we can compute this probability. We don't want to 'fall off a cliff' and end up proving that $0 = 1$! So we need an existence theorem that gives conditions ('rules of the road') which will guarantee that we will not end up with mathematical nonsense. This is the business addressed in this section. It turns out to be difficult because there really are dangerous cliffs out there.

5.2 Uncountable sample spaces

We resolved the Markov chain existence question by pushing it onto the problem of defining a measure on the interval $[0, 1]$. The resolution is a general theory that applies to any probability model. So we introduce the general theory first and then apply it to the Lebesgue measure. As a bonus we will see how to define continuous random variables also.

For uncountable sample spaces a new kind of difficulty arises when trying to define a probability law. We want to maintain the consistency relations as before, in particular countable additivity: if $E_1, E_2, \ldots, E_n, \ldots$ is a sequence of pairwise disjoint events, then we want

$$P(\bigcup_{n=1}^{\infty} E_n) = \sum_{n=1}^{\infty} P(E_n) \tag{208}$$

But there are so many subsets of an infinite sample space S that it turns out to be impossible to satisfy countable additivity for all sequences of disjoint sets. Something has to give. The resolution is to restrict the class of events by excluding some subsets of S. The class of events should still be large enough to include everything we encounter in practice, and it should also include everything we can get by combining events in the usual way. The correct formulation is called a σ-algebra.

5.3 σ-algebras

Definition 23 *Let S be a nonempty set. A σ-algebra in S is a collection of subsets $\mathcal{A}$ satisfying the following conditions:*

(1) $S \in \mathcal{A}$

(2) if $A \in \mathcal{A}$ *then* $A^c \in \mathcal{A}$

(3) if $A_1, A_2, \ldots$ *is a countable collection of sets in* $\mathcal{A}$*, then their union* $\bigcup_n A_n$ *also belongs to* $\mathcal{A}$

These properties are expressed by saying $\mathcal{A}$ is closed under complements and countable unions.

Exercise 57 Show that $\mathcal{A}$ is also closed under countable intersections.

We summarize this by saying that $\mathcal{A}$ is closed under the operations of complement, countable union and countable intersection. Note however that in general an *uncountable* union or intersection of sets in $\mathcal{A}$ will not be contained in $\mathcal{A}$. For this reason there may be subsets of S which are not in the collection $\mathcal{A}$.

Clearly the collection of all subsets of S is a σ-algebra.

Exercise 58 Let $\mathcal{A}$ be the subsets of S which are either countable or whose complement is countable. Show that $\mathcal{A}$ is a σ-algebra.

So in our new way of thinking a sample space S will be equipped with a σ-algebra $\mathcal{A}$, and only the sets in $\mathcal{A}$ will be considered as events. So the probability law needs to be defined only on $\mathcal{A}$.

Definition 24 *A probability triple $(S, \mathcal{A}, P)$ consists of a nonempty set S, a σ-algebra $\mathcal{A}$ in S, and a map $P : \mathcal{A} \to [0, 1]$ satisfying*
(i) $0 \leq P(A)$ *for all* $A \in \mathcal{A}$,
(ii) $P(S) = 1$,
(iii) *if $A_1, A_2, \ldots$ is a pairwise disjoint sequence of sets in $\mathcal{A}$ then*

$$P(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n) \tag{209}$$

The axioms given are tightly compressed, and imply a host of other relations. For example, a finite sequence $A_1, \ldots, A_n$ can be augmented to $A_1, \ldots, A_n, \emptyset, \ldots$ and then (iii) provides finite additivity. Since $A \cap A^c = \emptyset$ and $S = A \cup A^c$, it follows from (ii) and (iii) that

$$1 = P(S) = P(A) + P(A^c) \tag{210}$$

This also implies monotonicity: if $A \subset B$ then $B = A \cup (B - A)$ is a disjoint union, so $P(B) = P(A) + P(B - A) \geq P(A)$.

Exercise 59 Derive the inclusion-exclusion formula:

$$\begin{aligned} P(\bigcup_{k=1}^{n} A_k) &= \sum_i P(A_i) - \sum_{i<j} P(A_i \cap A_j) \\ &+ \sum_{i<j<k} P(A_i \cap A_j \cap A_k) + \cdots + (-1)^{n+1} P(A_1 \cap \cdots \cap A_n) \end{aligned} \tag{211}$$

There is one important special case, namely where $\mathcal{A} = 2^S$ is the σ-algebra of all subsets of S. When S is finite or countable and P is any map on S satisfying $\sum_{s \in S} p(s) = 1$, then the probability triple can always be taken as $(S, 2^S, P)$, and in this case all subsets are events. But if S is uncountable then it may not be possible to extend P to a map on the σ-algebra 2^S.

The important thing to recognize is that we are putting conditions on $(S, \mathcal{A}, P)$ by demanding that these properties are satisfied. As we will see these conditions are enough to guarantee that we will 'stay on the road' if we follow the rules.

5.4 Continuity

The axioms guarantee that P has a nice continuity property. A sequence A_n of sets is increasing if $A_n \subset A_{n+1}$ for all n. The limit of this sequence is defined to be $\bigcup_{n=1}^{\infty} A_n$. Similarly a sequence A_n is decreasing if A_n^c is increasing, and the limit is then $\bigcap_{n=1}^{\infty} A_n$.

Lemma 25 *If A_n are increasing then $P(A_n)$ is increasing and $\lim P(A_n) = P(\bigcup_{n=1}^{\infty} A_n)$. If A_n are decreasing then $P(A_n)$ is decreasing and $\lim P(A_n) = P(\bigcap_{n=1}^{\infty} A_n)$.*

Proof: suppose A_n are increasing. For each $n \geq 1$ define

$$B_{n+1} = A_{n+1} - A_n, \quad B_1 = A_1 \tag{212}$$

Then B_n are disjoint, and for every $N \geq 1$

$$\bigcup_{n=1}^{N} A_n = \bigcup_{n=1}^{N} B_n \tag{213}$$

as well as

$$\bigcup_{n=1}^{\infty} A_n = \bigcup_{n=1}^{\infty} B_n \tag{214}$$

Hence

$$\begin{aligned} P(\bigcup_{n=1}^{\infty} A_n) &= P(\bigcup_{n=1}^{\infty} B_n) \\ &= \sum_{n=1}^{\infty} P(B_n) \\ &= \lim_{N\to\infty} \sum_{n=1}^{N} P(B_n) \\ &= \lim_{N\to\infty} P(\bigcup_{n=1}^{N} B_n) \\ &= \lim_{N\to\infty} P(\bigcup_{n=1}^{N} A_n) \\ &= \lim_{N\to\infty} P(A_N) \end{aligned} \tag{215}$$

QED

Exercise 60 Complete the proof for a decreasing sequence.

Recall that Exercise 5 derived countable subadditivity:

$$P(\bigcup_{n=1}^{\infty} A_n) \leq \sum_{n=1}^{\infty} P(A_n) \tag{216}$$

5.5 Draw breath

It is worth returning to the original reason for introducing σ-algebras, namely the impossibility of satisfying the consistency relations (208) for all subsets of S. At this point it is not clear that P can be defined even on the smaller collection of sets $\mathcal{A}$. In fact this is possible, and leads to powerful models in probability theory.

The issue of how to do this is tackled in probability theory by a general strategy. First there is a small class of sets where it is 'obvious' how to define the probabilities. The probabilities defined on this small class are then used to construct a function called outer measure that assigns a value $P^*(E)$ to every subset $E \subset S$. The value $P^*(E)$ agrees with the original probability value on the small class of 'obvious' sets, but cannot be interpreted as a probability for all sets. Finally a special σ-algebra $\mathcal{A}$ is identified where the function P^* satisfies the properties (i), (ii), (iii) required for a probability law. Then $(S, \mathcal{A}, P^*)$ is the probability triple. When done in the right way this leads to a sufficiently large σ-algebra that includes the events of interest for the problem. Of course, once the probability law has been defined you can start trying to compute probabilities of interesting events, which is where the real hard work starts!

A theory is only as good as its useful examples. We will shortly look at how Lebesgue measure is constructed. For the moment we note that property (iii) of the triple does not lead to inconsistencies.

Lemma 26 *Suppose* $\{A_n\}$ *and* $\{B_n\}$ *are each pairwise disjoint sequences of sets in* $\mathcal{A}$*, and also*

$$\bigcup_{n=1}^{\infty} A_n = \bigcup_{n=1}^{\infty} B_n \tag{217}$$

Then

$$\sum_{n=1}^{\infty} P(A_n) = \sum_{n=1}^{\infty} P(B_n) \tag{218}$$

Proof: Let $E = \bigcup_{n=1}^{\infty} A_n = \bigcup_{n=1}^{\infty} B_n$. For all n, m define the set $C_{n,m} = A_n \cap B_m$. Then the sets $C_{n,m}$ are disjoint and belong to $\mathcal{A}$, and

$$\bigcup_{n=1}^{\infty} C_{n,m} = E \cap B_m = B_m, \quad \bigcup_{m=1}^{\infty} C_{n,m} = A_n \cap E = A_n \tag{219}$$

Hence

$$\begin{aligned}
\sum_{n=1}^{\infty} P(A_n) &= \sum_{n=1}^{\infty} P(\bigcup_{m=1}^{\infty} C_{n,m}) \\
&= \sum_{n=1}^{\infty}\sum_{m=1}^{\infty} P(C_{n,m}) \\
&= \sum_{m=1}^{\infty}\sum_{n=1}^{\infty} P(C_{n,m}) \\
&= \sum_{m=1}^{\infty} P(\bigcup_{n=1}^{\infty} C_{n,m}) \\
&= \sum_{m=1}^{\infty} P(B_m) \qquad (220)
\end{aligned}$$

QED

So it follows that if a set $A \in \mathcal{A}$ can be decomposed in several different ways as a countable union of disjoint sets then the relation (iii) is satisfied in every case. This is an important consistency check for the definition of a probability law.

Exercise 61 Let $\mathcal{A}$ and $\mathcal{B}$ be σ-algebras in S. Show that $\mathcal{A} \cap \mathcal{B}$ is also a σ-algebra (note that $C \in \mathcal{A} \cap \mathcal{B}$ if and only if $C \in \mathcal{A}$ and $C \in \mathcal{B}$).

5.6 σ-algebra generated by a class

For a finite collection of sets, you can enumerate all the sets obtained by taking complements, unions and intersections of these sets. This larger collection is called the σ-algebra generated by the original set. This procedure does not work if you start with an infinite collection of sets, hence another method of construction is needed.

Let $\mathcal{C}$ be a collection of subsets of S. Define $\sigma(\mathcal{C})$ to be the smallest σ-algebra in S containing $\mathcal{C}$. More precisely, $\mathcal{C} \subset \sigma(\mathcal{C})$ and if $\mathcal{A}$ is any σ-algebra containing $\mathcal{C}$ then $\sigma(\mathcal{C}) \subset \mathcal{A}$. This is called the σ-algebra generated by $\mathcal{C}$. The construction of $\sigma(\mathcal{C})$ is quite strange but it gives a flavor of how things are done in the world of measure theory. First we note the following.

Lemma 27 *Let $\mathcal{T}$ be a collection of σ-algebras, then $\mathcal{B} = \bigcap_{\mathcal{A}\in\mathcal{T}} \mathcal{A}$ is a σ-algebra.*

Proof: let B_k be a sequence in $\mathcal{B}$, then $B_k \in \mathcal{A}$ for every $\mathcal{A} \in \mathcal{T}$, hence $\bigcup B_k \in \mathcal{A}$ for every $\mathcal{A} \in \mathcal{T}$, hence $\bigcup B_k \in \mathcal{B}$. Similarly $\mathcal{B}$ is closed under complement.

QED

Now define $\mathcal{T}$ to be the collection of all σ-algebras in S which contain $\mathcal{C}$. Then

$$\sigma(\mathcal{C}) = \bigcap_{\mathcal{A}\in\mathcal{T}} \mathcal{A} \tag{221}$$

To see why this is true, note that $\bigcap_{\mathcal{A}\in\mathcal{T}} \mathcal{A}$ is a σ-algebra, it contains $\mathcal{C}$, and it is the smallest σ-algebra which does so.

Exercise 62 Let $\mathcal{C}$ denote the collection of all half-open intervals $(a, b] \subset \mathbb{R}$ where $a < b$. Show that $\sigma(\mathcal{C})$ contains all intervals of the form (a, b), $[a, b]$ and $[a, b)$ with $a < b$.

5.7 Borel sets

The Borel sets constitute an important σ-algebra in $\mathbb{R}$. They are built up by starting with the half-open intervals $(a, b]$ where $a < b$. Let $\mathcal{C}$ be the collection of all such intervals. Then the σ-algebra of Borel sets is defined to be $\mathcal{B} = \sigma(\mathcal{C})$, that is the smallest σ-algebra containing all of these intervals.

The Borel σ-algebra $\mathcal{B}$ plays an important role because it is large enough that we can construct continuous random variables on $(\mathbb{R}, \mathcal{B}, P)$. Of course we have not specified P yet but there are plenty of ways to do this. For the moment we note some properties of $\mathcal{B}$. Recall that we showed in Exercise 35 that $\mathcal{B}$ contains all intervals of the form (a, b), $[a, b]$ and $[a, b)$ with $a < b$.

Exercise 63 Show that $\mathcal{B}$ contains all open and closed sets in $\mathbb{R}$.
[Hint: use the fact that every open set in $\mathbb{R}$ is a countable union of pairwise disjoint open intervals].

In fact (though we will not prove it here) $\mathcal{B}$ is also the σ-algebra generated by the open sets in $\mathbb{R}$.

For the next exercise, recall that a function $f : \mathbb{R} \to \mathbb{R}$ is continuous if and only if $f^{-1}(A)$ is open for every open set $A \subset \mathbb{R}$.

Exercise 64 Let $f : \mathbb{R} \to \mathbb{R}$ be continuous. Define

$$\mathcal{C} = \{E \subset \mathbb{R} : f^{-1}(E) \in \mathcal{B}\} \tag{222}$$

Show that $\mathcal{C}$ is a σ-algebra. Show that $\mathcal{C}$ contains all open sets. Deduce that $\mathcal{B} \subset \mathcal{C}$.

5.8 Lebesgue measure

This is the prototype for probability functions on continuous spaces. The subsets of $\mathbb{R}$ with an obvious length are the intervals:

$$l(a, b] = l(a, b) = |a - b| \tag{223}$$

We want to extend this to a measure on the Borel sets. First define outer measure for all subsets:

$$m^*(A) = \inf \Big\{ \sum_{n=1}^{\infty} l(I_n) : A \subset \bigcup_{n=1}^{\infty} I_n \Big\} \tag{224}$$

where the infimum is taken over all countable collections of intervals whose union contains A. There is some work to do now. Must check that $m^*(I) = l(I)$ for every interval, so that m^* really is an extension of the length function. This is quite non-trivial, and requires using compactness properties of $\mathbb{R}$.

The next step is to select a good collection of sets where countable additivity will hold.

Definition 28 *A set $E \subset \mathbb{R}$ is* measurable *if for every set $A \subset \mathbb{R}$ we have*

$$m^*(A) = m^*(A \cap E) + m^*(A \cap E^c) \tag{225}$$

So whenever a measurable set E divides a set A into two disjoint pieces $A \cap E$ and $A \cap E^c$, the sum of the measures must equal the measure of the whole. Let $\mathcal{M}$ be the collection of all measurable sets. The key result is the following.

Lemma 29 *$\mathcal{M}$ is a σ-algebra, and $\mathcal{M}$ contains the Borel sets. If E_n are pairwise disjoint sets in $\mathcal{M}$, then*

$$m^*(\bigcup E_n) = \sum m^*(E_n) \tag{226}$$

It follows that m^* defines a measure on $\mathcal{M}$, and this is called the Lebesgue measure. Can check that it is translation invariant. The σ-algebra $\mathcal{M}$ is strictly larger than the Borel sets $\mathcal{B}$, but for most purposes the distinction is irrelevant, and we restrict the measure to $\mathcal{B}$. We will denote Lebesgue measure by λ henceforth.

Exercise 65 Show $\mathbb{Q}$ has measure zero. Same for any countable set. Same for the Cantor set.

Despite the complexity of their definition, the Borel sets are not too much different from open and closed sets, as the following result shows.

Lemma 30 *Let $B \in \mathcal{B}$ be a Borel set. Then for every $\epsilon > 0$, there is a closed set F and an open set G such that $F \subset B \subset G$, and $\lambda(B - F) < \epsilon$ and $\lambda(G - B) < \epsilon$.*

Proof: first we construct the open set G which contains B. By definition of outer measure, for every $\epsilon > 0$ there is a countable union of open intervals $\{I_n\}$ such that $B \subset \bigcup I_n$ and

$$\lambda(B) = m^*(B) > \sum_{n=1}^{\infty} l(I_n) - \epsilon$$

Let $G = \bigcup I_n$, then G is open, $B \subset G$ and

$$\lambda(G - B) = \lambda(G) - \lambda(B) \leq \sum_{n=1}^{\infty} l(I_n) - \lambda(B) < \epsilon$$

For the closed set inside F, take F^c to be the open set containing B^c as above.

QED

Exercise 66 Show that outer measure m^* is translation invariant, that is $m^*(A + \{x\}) = m^*(A)$ for every $A \subset \mathbb{R}$ and every $x \in \mathbb{R}$.

Exercise 67 Show that Lebesgue measure is complete: if B is measurable and $\lambda(B) = 0$, then every subset $A \subset B$ is also measurable and $\lambda(A) = 0$.

5.9 Lebesgue-Stieltjes measure

The study of continuous random variables will lead us to a generalization of Lebesgue measure. Suppose that $F : \mathbb{R} \to \mathbb{R}$ is a non-decreasing function which is continuous on the right. So if $x < y$ then $F(x) \leq F(y)$, and for all x

$$F(x) = \lim_{h \to 0+} F(x + h) \tag{227}$$

Then we can assign a new measure to half-open intervals as follows:

$$\mu(a, b] = F(b) - F(a) \tag{228}$$

The construction of the Lebesgue measure can now be repeated with the measure μ used instead of l for the intervals. Everything goes through and we end up with a new measure μ_F defined on $\mathcal{B}$.

Lemma 31 *Let F be a non-decreasing function which is continuous on the right, and satisfies $\lim_{x \to -\infty} F(x) = 0$. Then there is a unique measure μ_F on $\mathcal{B}$ such that for all $a < b$,*

$$\mu_F(a, b] = F(b) - F(a) \tag{229}$$

Exercise 68 Define

$$F(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ x^2 & \text{for } 0 < x < \frac{1}{2} \\ \frac{1}{2} & \text{for } \frac{1}{2} \leq x < 1 \\ 1 & \text{for } x \geq 1 \end{cases} \tag{230}$$

Calculate $\mu_F(0, 1/2)$, $\mu_F(0, 1/2]$, $\mu_F(\{1/2\})$, $\mu_F[1/2, 1)$, $\mu_F[1/2, 1]$.

5.10 Lebesgue-Stieltjes measure on $\mathbb{R}^n$

The Borel sets on $\mathbb{R}^n$ are denoted $\mathcal{B}(\mathbb{R}^n)$. This is the σ-algebra generated by the open sets in $\mathbb{R}^n$, and is also the σ-algebra generated by the rectangles $(a_1, b_1] \times \cdots \times (a_n, b_n]$. Let $F : \mathbb{R}^n \to \mathbb{R}$ be increasing and right continuous in each component, then there is a unique measure μ_F on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ satisfying

$$F(x_1, \ldots, x_n) = \mu_F((-\infty, x_1] \times \cdots \times (-\infty, x_n]) \tag{231}$$

for all $(x_1, \ldots, x_n) \in \mathbb{R}^n$. This is the Lebesgue-Stieltjes measure defined by F. One special case arises when F is a product, that is $F(x_1, \ldots, x_n) = F_1(x_1) \ldots F_n(x_n)$. In this case

$$\mu_F = \mu_{F_1} \times \cdots \times \mu_{F_n} \tag{232}$$

is a product measure on $\mathbb{R}^n$.

5.11 Random variables

Let S, T be sets, and let $\mathcal{A}, \mathcal{C}$ be σ-algebras of subsets of S, T respectively. A map $f : S \to T$ is called *measurable* if $f^{-1}(C) \in \mathcal{A}$ for every $C \in \mathcal{C}$.

Definition 32 *Consider a probability triple $(S, \mathcal{A}, P)$. A random variable on S is a measurable function from $(S, \mathcal{A})$ to $(\mathbb{R}, \mathcal{B})$.*

So the preimage of every Borel set must be a measurable set. By σ-algebra properties, it is sufficient to check this for the sets that generate $\mathcal{B}$, namely the half-open intervals. Even this can be simplified to the following statement: X is a random variable if and only if for every $a \in \mathbb{R}$, the set $X^{-1}(-\infty, a]$ is in $\mathcal{A}$.

Exercise 69 Let $A \in \mathcal{A}$, and let 1_A be the indicator function of A. Show that 1_A is a random variable.

Exercise 70 Let $f, g : \mathbb{R} \to \mathbb{R}$ be measurable. Show that $f + g$ and fg are also measurable.

In the previous section we studied the case where $\mathrm{Ran}(X)$ is countable, that is where X is discrete, and S is countable. Measurability does not arise

in this case because all subsets of S are measurable. Furthermore the pmf contains all information about probabilities involving X; this is just the list of probabilities of the (countably) many different values for X.

In general for uncountable S the pmf makes no sense. What takes its place is the cdf (cumulative generating function). This is the real-valued function $F_X : \mathbb{R} \to [0,1]$ defined by

$$F_X(x) = P(X \leq x) = P(\{\omega : X(\omega) \leq x\}) = P(X^{-1}(-\infty, x]) \tag{233}$$

Notice it is well-defined because $X^{-1}(-\infty, x] \in \mathcal{A}$ for all x. It is convenient to drop the subscript X unless we need it to distinguish between cdf's.

Important properties are:

(a) $0 \leq F(x) \leq 1$ for all $x \in \mathbb{R}$

(b) if $x < y$ then $F(x) \leq F(y)$

(c) $\lim_{x\to\infty} F(x) = 1$, $\lim_{x\to-\infty} F(x) = 0$

(d) F is right continuous: if x_n is a decreasing sequence and $\lim x_n = x$ then $\lim F(x_n) = F(x)$

Exercise 71 Prove (a)–(d).

Exercise 72 Prove that

$$P(X = x) = F(x) - \lim_{h \downarrow 0} F(x - h) \tag{234}$$

As far as the random variable X is concerned, everything that can be known about P is contained in the cdf F. More precisely, for any Borel set B, the probability $P(X \in B) = P(X^{-1}(B))$ can be computed from F. This is because F is a non-decreasing function on $\mathbb{R}$ which is continuous on the right, and hence there is a unique Lebesgue-Stieltjes measure μ_F on $\mathbb{R}$ which satisfies

$$\mu_F(a, b] = F(b) - F(a) \tag{235}$$

for every $a < b$. Looking at this we find that

$$P(a < X \leq b) = \mu_F(a, b] \tag{236}$$

So the probability of any half-open interval $(a, b]$ is uniquely determined by F in this way. By our Lebesgue-Stieltjes theorem, we know that μ_F is the unique measure on $\mathcal{B}$ which satisfies this. Therefore $\mu_F(B)$ is uniquely determined by F.

This is very nice because it means that we can concentrate on the cdf F_X and forget about the underlying probability triple. All the information about X is contained in this one function.

Another way to express this is to note that a measurable function "pushes forward" a measure. Since X is a measurable function from $(S, \mathcal{A}, P)$ to $(\mathbb{R}, \mathcal{B})$, it pushes forward the measure P to the measure μ_F on $(\mathbb{R}, \mathcal{B})$, namely

$$\mu_F(B) = P(X \in B), \quad B \in \mathcal{B} \tag{237}$$

Exercise 73 Let $f : S \to T$ be a measurable function from $(S, \mathcal{A}, \mu)$ to $(T, \mathcal{C})$. For $C \in \mathcal{C}$ define

$$\nu(C) = \mu(f^{-1}(C)) \tag{238}$$

Prove that ν is a measure on $(T, \mathcal{C})$.

5.12 Continuous random variables

Although the measure μ_F is always defined it may be quite difficult to work with. In many cases X satisfies an additional condition which greatly simplifies the measure. Recall that a map $g : \mathbb{R} \to \mathbb{R}$ is *continuous* if given any $\epsilon > 0$, there is a $\delta > 0$ such that

$$|g(y) - g(x)| < \epsilon \tag{239}$$

for every interval (x, y) with

$$|y - x| < \delta \tag{240}$$

There is also a slightly stronger condition: a map $g : \mathbb{R} \to \mathbb{R}$ is *absolutely continuous* if given any $\epsilon > 0$, there is a $\delta > 0$ such that

$$\sum_{i=1}^{n} |g(y_i) - g(x_i)| < \epsilon \tag{241}$$

for every finite collection $\{(x_i, y_i)\}$ of nonoverlapping intervals with

$$\sum_{i=1}^{n} |y_i - x_i| < \delta \tag{242}$$

This rather formidable definition is important because of the following Theorem.

Theorem 33 *A function F is an indefinite integral if and only if it is absolutely continuous.*

In other words, the function F is absolutely continuous if and only if there is an integrable function f such that for all $a, x \in \mathbb{R}$,

$$F(x) = \int_a^x f(t)\ dt + F(a) \tag{243}$$

Comment: we have not defined the Lebesgue integral yet! this will be done shortly. In the meantime we will work with examples where $f(t)$ is continuous and so the Riemann integral is sufficient.

Definition 34 *The random variable X is continuous if the function F_X is absolutely continuous.*

Comment: strictly we should define X in this case to be absolutely continuous. But everyone uses this notation so we follow suit.

If X is continuous then its cdf is completely determined by the pdf f_X, which satisfies the following:

(1) f_X is measurable and non-negative

(2) for all $a \in \mathbb{R}$, $P(X \leq a) = F_X(a) = \int_{-\infty}^{a} f_X(x)\ dx$

It follows as a consequence that for a continuous random variable X,

$$P(a < X \leq b) = \int_a^b f_X(x)\ dx \tag{244}$$

and therefore that $P(X = x) = 0$ for every $x \in \mathbb{R}$. Thus for continuous random variables the events $\{X < a\}$ and $\{X \leq a\}$ have the same probability, and so on. The value of f_X at any particular point is irrelevant, as it does not affect the value of the integral. Notice the normalization condition $\int_{-\infty}^{\infty} f_X(x)dx = 1$.

Many special cases are important, we list a few here.

Uniform The pdf is

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases} \tag{245}$$

where $a < b$. Loosely, X is 'equally likely' to be anywhere in the interval $[a, b]$.

Exponential The pdf is

$$f(x) = \begin{cases} ke^{-kx} & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases} \tag{246}$$

where $k > 0$. This is often the model for the time until failure of a device.

Normal The pdf is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{247}$$

where $\mu \in \mathbb{R}$ is the mean and $\sigma^2 > 0$ is the variance. The special case $\mu = 0$, $\sigma = 1$ is called the standard normal. This is the best known and most widely used random variable, we will see why later.

Exercise 74 Compute the cdf's of the uniform and exponential.

Exercise 75 For the exponential, show that

$$P(X > s + t | X > s) = P(X > t)$$

for all $s, t > 0$. This is the famous 'memoryless' property of the exponential.

Exercise 76 Verify that the normal is correctly normalized using the following integration formula: for $a > 0$ and all b,

$$\int_{-\infty}^{\infty} e^{-ax^2+bx} dx = \sqrt{\frac{\pi}{a}}\, e^{b^2/4a} \tag{248}$$

Exercise 77 Here is another continuous random variable. Imagine dropping a coin onto a tiled floor. The tiles are squares of unit side length, the coin has radius $r < \frac{1}{2}$. Let R be the distance from the coin's center to the nearest square center. Find the pdf of R.

5.13 Several random variables

Often have to consider several random variables together. This presents no problems. By assumption the random variables $X_1, \ldots, X_n$ are each defined on the same probability triple $(S, \mathcal{A}, P)$. Define the map $\mathbf{X} = (X_1, \ldots, X_n) : S \to \mathbb{R}^n$. Then $\mathbf{X}$ is a vector-valued random variable. We must check that for every Borel set $B \subset \mathbb{R}^n$, the set $\mathbf{X}^{-1}(B)$ is measurable. But this is guaranteed by the condition that each component function X_k is a random variable.

The joint cdf of $X_1, \ldots, X_n$ is defined by

$$F_{\mathbf{X}}(x_1, \ldots, x_n) = P(X_1 \leq x_1, \ldots, X_n \leq x_n) \tag{249}$$

Once again all the information about $\mathbf{X}$ is contained in this function, and it defines a measure μ_F on $\mathbb{R}^n$ which determines the probabilities for all events $\{\mathbf{X} \in B\}$. The two most commonly encountered cases are where each X_k is discrete and where each X_k is continuous. In the latter case the random variable $\mathbf{X}$ has a joint pdf $f_{\mathbf{X}}$ which determines probabilities according to

$$P(\mathbf{X} \in B) = \int_B f_{\mathbf{X}}(x_1, \ldots, x_n) dx_1 \ldots dx_n \tag{250}$$

5.14 Independence

Recall that a collection of events $A_1, \ldots, A_n$ is independent if

$$P(A_{i_1} \cap \cdots \cap A_{i_k}) = P(A_{i_1}) \cdots P(A_{i_k}) \tag{251}$$

for all subsets $A_{i_1}, \ldots, A_{i_k}$.

For each $i = 1, \ldots, n$ let $\mathcal{A}_i$ be a collection of events. Then the sequence of *collections* $\mathcal{A}_1, \ldots, \mathcal{A}_n$ is independent if for each choice of $A_i \in \mathcal{A}_i$ for $i = 1, \ldots, n$, the events $A_1, \ldots, A_n$ are independent.

Finally, for a random variable X define $\sigma(X)$ to be the σ-algebra generated by X, namely the smallest σ-algebra with respect to which X is measurable. In other words, $\sigma(X)$ is the smallest σ-algebra which contains all the events $\{X^{-1}(B)\}$ for all Borel sets B in $\mathbb{R}$.

As a special case, if X is discrete then $\mathrm{Ran}(X) = \{x_1, x_2, \ldots\}$ is countable. Let $A_i = X^{-1}(x_i)$, then $\sigma(X)$ is the σ-algebra generated by the events $A_1, A_2, \ldots$.

Definition 35 *The random variables $X_1, X_2, \ldots, X_n$ are independent if $\sigma(X_1), \sigma(X_2), \ldots, \sigma(X_n)$ are independent.*

Explicitly, $X_1, X_2, \ldots, X_n$ are independent if for all Borel sets $B_1, B_2, \ldots, B_n$,

$$P(X_1 \in B_1, \ldots, X_n \in B_n) = P(X_1 \in B_1) \ldots P(X_n \in B_n) \tag{252}$$

(we allow $B_i = \mathbb{R}$ so this checks all subsets of the X_i). The Borel sets are generated by intervals, so it is enough to check this for Borel sets of the form $B = (-\infty, a]$, and thus independence is equivalent to

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = F_{X_1}(x_1) \ldots F_{X_n}(x_n) \tag{253}$$

for all $(x_1, \ldots, x_n) \in \mathbb{R}^n$.

If all X_i are continuous, then independence is equivalent to factorization of the joint pdf, that is

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = f_{X_1}(x_1) \ldots f_{X_n}(x_n) \tag{254}$$

Exercise 78 A dart is randomly thrown at a square dartboard with unit side length. It lands at the point (X, Y). Find the probability that $|X - Y| \leq 1/4$.

5.15 Expectations

Let X be a random variable with cdf F, then μ_F is the Lebesgue-Stieltjes measure on $\mathbb{R}$ induced by F. The expected value of X is computed using this

measure. This is done by defining an integral using this measure in exact analogy to how the usual Lebesgue integral is defined starting from Lebesgue measure. We outline the steps below. For convenience we drop the subscript X on F.

First, recall the indicator function 1_A for a set A. For any Borel set B define

$$\int 1_B \, dF = \mu_F(B) = P(X \in B) \tag{255}$$

By linearity this extends to simple functions of the form $\phi = \sum_{i=1}^n c_i \, 1_{A_i}$:

$$\int \phi \, dF = \sum_{i=1}^n c_i \int 1_{A_i} \, dF \tag{256}$$

Exercise 79 Suppose a simple function ϕ is written in two ways as a sum of indicator functions:

$$\phi = \sum_{i=1}^n c_i \, 1_{A_i} = \sum_{j=1}^m d_j \, 1_{B_j} \tag{257}$$

Show that $\int \phi \, dF$ is the same when calculated with either expression. [Hint: first show that a simple function has a unique representation of this form with disjoint sets $\{B_j\}$, then show that the statement holds in this case].

Most of the work goes into showing that any measurable function g can be approximated by a sequence of simple functions ϕ_n, and that the integrals of the simple functions converge as $n \to \infty$. We will assume these results here and jump to the conclusion, which is that the integral of a bounded non-negative measurable function g is defined to be

$$\int g \, dF = \sup_{\phi \leq g} \int \phi \, dF \tag{258}$$

where the sup runs over simple functions which are upper bounded by g. The following properties of the integral can then be deduced.

Lemma 36 *The integral $\int \cdot dF$ is defined for all non-negative measurable functions on $\mathbb{R}$, and satisfies*

(i) $\int cg\ dF = c\int g\ dF$ *for all* $c \in \mathbb{R}$
(ii) $\int (g+h)\ dF = \int g\ dF + \int h\ dF$
(iii) $\int g\ dF \geq 0$ *for* $g \geq 0$
(iv) *if* $g_n \uparrow g$ *then* $\int g_n\ dF \uparrow \int g\ dF$

The last property (4) is called the monotone convergence theorem and plays an important role in the theory.

Exercise 80 Let g be bounded and measurable, say $|g(x)| \leq M$. For all $n \geq 1$ define the sets

$$E_k = \Big\{ x\ :\ \frac{kM}{n} \geq g(x) > \frac{(k-1)M}{n} \Big\}, \quad -n \leq k \leq n \tag{259}$$

Define the simple functions

$$\psi_n(x) = \frac{M}{n} \sum_{k=-n}^{n} k\, 1_{E_k}(x), \quad \phi_n(x) = \frac{M}{n} \sum_{k=-n}^{n} (k-1)\, 1_{E_k}(x) \tag{260}$$

Show that $\phi_n(x) \leq g(x) \leq \psi_n(x)$ for all x and all n. Deduce that

$$\inf_{g \leq \psi} \int \psi\, dF = \sup_{g \geq \phi} \int \phi\, dF \tag{261}$$

where the infimum and supremum are taken over simple functions.

There is one other important convergence result. First, if g is measurable write $g = g^+ - g^-$. If both $\int g^{\pm}\ dF < \infty$ then say g is integrable and define

$$\int g\ dF = \int g^+\ dF - \int g^-\ dF \tag{262}$$

Lemma 37 (Dominated Convergence Theorem) *Suppose* $\{g_n\}$ *are integrable, and* $g_n \to g$ *as* $n \to \infty$. *Suppose also that there is an integrable function* h *such that* $|g_n| \leq h$, *then*

$$\int g_n\ dF \to \int g\ dF \tag{263}$$

Exercise 81 Use the Dominated Convergence Theorem to show that the following limit exists, and compute it:

$$\lim_{n\to\infty} \int_0^1 n \sin\Big(\frac{1}{n\sqrt{x}}\Big)\, dx \tag{264}$$

Now that we have the integral defined, we can define the expectation.

Definition 38 *Let X be a random variable with cdf F. For any real-valued measurable function $g : \mathbb{R} \to \mathbb{R}$,*

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x)\, dF \tag{265}$$

If X is discrete and $\mathrm{Ran}(X) = \{x_1, x_2, \dots\}$ then

$$\mathbb{E}[g(X)] = \sum_i g(x_i)\, P(X = x_i) \tag{266}$$

If X is continuous with pdf f then

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f(x)\, dx \tag{267}$$

5.16 Calculations with continuous random variables

There are various useful formulas for calculations which deserve a special mention.

Change of variables Let X be continuous and $Y = g(X)$ for some measurable function g. The cdf of Y is obtained by using

$$P(Y \le y) = P(g(X) \le y) = \int_{x\,:\,g(x)\le y} f_X(x) dx \tag{268}$$

Exercise 82 Let Z be a standard normal, show the pdf of $Y = Z^2$ is

$$f_Y(y) = \begin{cases} 0 & \text{for } y \le 0 \\ \frac{1}{\sqrt{2\pi}}\, y^{-1/2}\, e^{-y/2} & \text{for } y > 0 \end{cases} \tag{269}$$

If g is invertible there is a formula. Suppose that $\mathbf{X} = (X_1, \ldots, X_n)$ are continuous with joint pdf $f_{\mathbf{X}}$. Suppose that g is a one-to-one, continuously differentiable map on $\mathbb{R}^n$. Let T be the inverse of g, and suppose that its Jacobian J is nonzero everywhere. Define $\mathbf{Y} = (Y_1, \ldots, Y_n) = g(X_1, \ldots, X_n)$. Then Y is continuous and its pdf is

$$f_{\mathbf{Y}}(x) = f_{\mathbf{X}}(T(x))\,|J(x)| \tag{270}$$

Exercise 83 Suppose X_1, X_2 are normal with the joint pdf

$$f(x, y) = \frac{3}{2\pi}\, e^{-\frac{1}{2}(2x^2+2xy+5y^2)} \tag{271}$$

Define $U = X_1 - X_2$ and $V = X_1 + 2X_2$, show that the joint pdf of (U, V) is

$$f_{U,V}(u, v) = \frac{1}{2\pi}\, e^{-\frac{1}{2}(u^2+v^2)} \tag{272}$$

Events involving independent random variables The probability of an event involving two independent random variables X, Y can be computed using an iterated integral. More precisely, for any event B,

$$\begin{aligned} P((X,Y) \in B) &= \int_{-\infty}^{\infty} P((x,Y) \in B)\, dF_X \\ &= \int_{-\infty}^{\infty} \left(\int_{y\,:\,(x,y)\in B} dF_Y \right) dF_X \end{aligned} \tag{273}$$

Exercise 84 Suppose X is exponential and Y is uniform on $[0, 1]$, and X, Y are independent. Show that

$$P(X + Y \leq 1) = e^{-1} \tag{274}$$

Although we have not yet defined conditioning with respect to a continuous random variable, it is often useful to rewrite this result using the conditioning notation. So we write

$$P((x,Y) \in B) = P((X,Y) \in B \,|\, X = x) \tag{275}$$

then our formula becomes

$$\begin{aligned} P((X,Y) \in B) &= \int_{-\infty}^{\infty} P((X,Y) \in B \,|\, X = x)\, dF_X \\ &= \mathbb{E}[P((X,Y) \in B \,|\, X)] \end{aligned} \tag{276}$$

As an illustration, suppose that X, Y are independent exponentials with mean 1 and we want $P(X + Y \geq z)$ where $z \geq 0$. Now

$$P(X + Y \geq z \,|\, X = x) = P(Y \geq z - x \,|\, X = x) = P(Y \geq z - x) \tag{277}$$

because they are independent. Thus

$$P(Y \geq z - x) = \begin{cases} e^{-(z-x)} & \text{for } z - x \geq 0 \\ 1 & \text{for } z - x < 0 \end{cases} \tag{278}$$

and hence

$$\begin{aligned} P(X + Y \geq z) &= \int_0^{\infty} P(X + Y \geq z \,|\, X = x) e^{-x} dx \\ &= \int_0^{\infty} P(Y \geq z - x) e^{-x} dx \\ &= \int_0^{z} e^{-z} dx + \int_z^{\infty} e^{-x} dx \\ &= z e^{-z} + e^{-z} \end{aligned} \tag{279}$$

Similar reasoning applies to several independent random variables. The same technique can be applied even when the random variables are dependent.

Exercise 85 Suppose X is uniform on $[0,1]$ and Y is uniform on $[0,X]$. Calculate $\mathbb{E}Y$.

Comment: for a continuous random variable X the event $\{X = x\}$ has probability zero, so our earlier definition of conditional probability does not give meaning to the expression $P(A \,|\, X = x)$. We will return later to this problem.

5.17 Stochastic processes

In subsequent sections we will often want to work with an infinite sequence of random variables $X_1, X_2, \ldots$ with some prescribed joint distributions. For Markov chains with countable state space we did this by defining them on $[0,1]$ with Lebesgue measure. But for continuous r.v.'s we need a better way. We consider the basic case where the X_k are all independent. In this case all the information is contained in the individual cdf's $F_1, F_2, \ldots$.

Theorem 39 *Let $\{F_k\}$ be a sequence of cdf's on $\mathbb{R}$. There exists on some probability space $(S, \mathcal{A}, P)$ an independent sequence of random variables $\{X_k\}$ such that X_k has cdf F_k.*

A few words about the proof. The general strategy runs as before: first define probabilities for a small class of sets, then extend to a larger σ-algebra. The process is constructed on the infinite product space $\mathbb{R}^\infty = \mathbb{R} \times \mathbb{R} \times \cdots$. A point in $\mathbb{R}^\infty$ is a sequence $\mathbf{s} = (s_1, s_2, \ldots)$. A set $A \subset \mathbb{R}^\infty$ is called a *cylinder set* if there are integers $(i_1, \ldots, i_k)$ and measurable sets $B_{i_1}, \ldots, B_{i_k}$ such that

$$A = \{\mathbf{s} \,|\, s_{i_1} \in B_{i_1}, \ldots, s_{i_k} \in B_{i_k}\} \tag{280}$$

The probability of this cylinder set is defined to be

$$\begin{aligned} P(A) &= \mu_{F_{i_1}}(B_{i_1}) \ldots \mu_{F_{i_k}}(B_{i_k}) \\ &= P(X_{i_1} \in B_{i_1}) \ldots P(X_{i_k} \in B_{i_k}) \end{aligned} \tag{281}$$

It is not hard to show that P is finitely additive on the cylinder sets.

Exercise 86 Let A, B be disjoint cylinder sets such that $A \cup B$ is also a cylinder set. Show that $P(A \cup B) = P(A) + P(B)$.

The hard work comes in showing countable additivity for P on the cylinder sets. As for the Lebesgue measure this needs a compactness argument.

Exercise 87 Suppose that $\mathcal{T}$ is a collection of sets and P is a probability with the following properties: if $T_1, \ldots, T_n \in \mathcal{T}$ are pairwise disjoint, such that $\bigcup_{i=1}^n T_i \in \mathcal{T}$ then

$$P(\bigcup_{i=1}^n T_i) = \sum_{i=1}^n P(T_i) \tag{282}$$

Suppose also that whenever $T_1, T_2, \cdots \in \mathcal{T}$ such that $T_{n+1} \subset T_n$ and $\bigcap_{n=1}^{\infty} T_n = \emptyset$, it follows that

$$P(T_n) \to 0 \quad \text{as } n \to \infty \tag{283}$$

Prove that P is countably additive on $\mathcal{T}$: if $T_1, \cdots \in \mathcal{T}$ are pairwise disjoint, such that $\bigcup_{i=1}^{\infty} T_i \in \mathcal{T}$ then

$$P(\bigcup_{i=1}^{\infty} T_i) = \sum_{i=1}^{\infty} P(T_i) \tag{284}$$

Once these properties are established there is a general machinery for extending P to a measure on the σ-algebra generated by the cylinder sets. In the case where the random variables are discrete the construction is somewhat simplified although the compactness property is still needed. In this case the state space $\mathbb{R}^{\infty}$ can be replaced by S^{∞} where S is discrete. The process can then be constructed on $([0,1], \mathcal{B})$ where $\mathcal{B}$ is the Borel σ-algebra on $[0,1]$ (this is clear when $S = \{0,1\}$).

6 Limit Theorems for stochastic sequences

This section is concerned with sums and averages of random variables. We already noted that $\mathrm{E}(X)$ has an operational meaning, namely that if an experiment is repeated many times under identical conditions and X is measured each time, then $\mathrm{E}(X)$ is the long-run average value of the measurements of X. To analyze this we need to look at the average of many trials. This leads to the Law of Large Numbers (LLN) and also to the Central Limit Theorem (CLT).

6.1 Basics about means and variances

The Limit Theorems mostly concern sequences of random variables. For a sequence of r.v.'s $X_1, X_2, \ldots, X_n$ the sample mean is defined to be

$$\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n) \tag{285}$$

The expected value is a linear operator so the mean is easily found:

$$\mathbb{E}\overline{X} = \frac{1}{n}(\mathbb{E}X_1 + \cdots + \mathbb{E}X_n) \tag{286}$$

The variance is not linear. However in most applications the variables X_i are independent, and in this case the variance distributes also:

$$\mathrm{VAR}[\overline{X}] = \frac{1}{n^2}(\mathrm{VAR}[X_1] + \cdots + \mathrm{VAR}[X_n]) \tag{287}$$

Recall the notation for mean and variance:

$$\mu = \mathbb{E}X, \quad \sigma^2 = \mathrm{VAR}[X] = \mathbb{E}X^2 - \mu^2 \tag{288}$$

6.2 Review of sequences: numbers and events

Before considering convergence of random variables we first recall how convergence is defined for ordinary sequences of numbers. Consider a sequence of real numbers $a_1, a_2, \ldots$. The sequence converges to a if for every $\epsilon > 0$ there is an integer $N < \infty$ such that

$$|a_n - a| < \epsilon \quad \text{for all } n \geq N \tag{289}$$

Exercise 88 Prove the sequence $a_n = n\sin(x/n)$ converges to x.

There is another way to formulate convergence that suits our needs better. Recall some definitions: the number b is an upper bound for the set $A \subset \mathbb{R}$ if $x \leq b$ for all $x \in A$. The number c is the least upper bound for A if c is an upper bound, and if $c \leq b$ for every upper bound b. A basic ingredient of real analysis is the fact that every bounded set has a least upper bound. We will write sup (supremum) for the least upper bound. The inf (infimum) is defined in a similar way as the greatest lower bound.

The sup of the sequence $\{a_n\}$ is the least upper bound, written as $\sup a_n$. Similarly for $\inf a_n$. The lim sup of a_n is defined as

$$\limsup a_n = \overline{\lim}\, a_n = \inf_{n\geq 1} \sup_{k\geq n} a_k \tag{290}$$

The meaning is: eventually the sequence is bounded above by $\overline{\lim} a_n + \epsilon$ for any $\epsilon > 0$. The lim inf is defined similarly:

$$\liminf a_n = \underline{\lim}\, a_n = \sup_{n\geq 1} \inf_{k\geq n} a_k \tag{291}$$

So loosely speaking this means that the sequence eventually ends up in the interval $[\underline{\lim} a_n, \overline{\lim} a_n]$. This gives a way to define convergence: the sequence converges if and only if $\underline{\lim} a_n = \overline{\lim} a_n$, in which case we define $\lim a_n$ to be this common value.

Exercise 89 Show that this definition of convergence agrees with the previous one.

Exercise 90 Compute $\underline{\lim} a_n$ and $\overline{\lim} a_n$ for $a_n = (n\cos(n\pi))/(n+1)$.

Now we turn to sequences of events $A_1, A_2, \ldots$. It is assumed that all sets A_n are subsets of the same state space S. By analogy with real sequences define

$$\overline{\lim} A_n = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k, \quad \underline{\lim} A_n = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k \tag{292}$$

What does this mean? Suppose first that $\omega \in \overline{\lim} A_n$, then for every $n \geq 1$, $\omega \in \bigcup_{k=n}^{\infty} A_k$, meaning that ω belongs to at least one of the sets $A_n, A_{n+1}, \ldots$.

Thus ω appears in the sets A_n infinitely often; no matter how far along in the sequence you go, ω will belong to a set further along. Thus

$$\overline{\lim} A_n = \{\omega \in S \,|\, \omega \in A_n \text{ i.o.}\} \tag{293}$$

where i.o. stands for infinitely often. Similarly, if $\omega \in \underline{\lim} A_n$, then for some n, it must be true that $x \in A_k$ for all $k \geq n$, meaning that ω belongs to every set A_k beyond some point in the sequence. So

$$\underline{\lim} A_n = \{\omega \in S \,|\, \omega \in A_n \text{ eventually}\} \tag{294}$$

From this it is clear that

$$\underline{\lim} A_n \subset \overline{\lim} A_n \tag{295}$$

And this leads to the definition of convergence: the sequence $\{A_n\}$ converges if and only if $\underline{\lim} A_n = \overline{\lim} A_n$, in which case $\lim A_n$ is defined to be this common event.

Note that the operations used to construct $\underline{\lim} A_n$ and $\overline{\lim} A_n$ are all operations that preserve the σ-algebra structure. So if $\{A_n\}$ are measurable (i.e. events) then so are $\underline{\lim} A_n$ and $\overline{\lim} A_n$.

Consider now a sequence of random variables $X_1, X_2, \ldots$. Each of these is a measurable function on a probability triple $(S, \mathcal{A}, P)$.

Exercise 91 Show that

$$\begin{aligned} \{s \,|\, \sup X_n(s) \leq x\} &= \bigcap_{n=1}^{\infty} \{s \,|\, X_n(s) \leq x\} \\ \{s \,|\, \inf X_n(s) \geq x\} &= \bigcap_{n=1}^{\infty} \{s \,|\, X_n(s) \geq x\} \end{aligned} \tag{296}$$

It follows that $\sup X_n$ and $\inf X_n$ are also random variables. Therefore so are

$$\overline{\lim} X_n = \inf_{n \geq 1} \sup_{k \geq n} X_k, \quad \underline{\lim} X_n = \sup_{n \geq 1} \inf_{k \geq n} X_k \tag{297}$$

So convergence of X_n concerns the properties of these random variables. If these are equal at some point $s \in S$ then we define the common value to

$\lim X_n(s)$. This is generally not a definition of the random variable $\lim X_n$ since it may not exist on the whole space. This will not matter as we will be concerned only with the set where it is defined.

Definition 40 *The sequence $\{X_n\}$ converges to X almost surely (a.s.) if*

$$P(\{s \in S \mid \lim X_n(s) = X(s)\}) = 1 \tag{298}$$

Note: this is saying both that X_n converges on a set of measure 1, and also that the limiting value equals X on a set of measure 1. The following result gives a useful way to prove convergence.

Lemma 41 *$\{X_n\}$ converges to X a.s. if and only if for every $\epsilon > 0$,*

$$P(|X_n - X| \geq \epsilon \text{ } i.o.) = 0 \tag{299}$$

Proof: let $s \in S$. The sequence $X_n(s)$ fails to converge to $X(s)$ if and only if there is some $\epsilon > 0$ such that $|X_n(s) - X(s)| > \epsilon$ infinitely often. Hence

$$\{s \mid \lim_n X_n(s) \neq X(s)\} = \bigcup_\epsilon \{s \mid |X_n(s) - X(s)| \geq \epsilon \,\text{i.o.}\} \tag{300}$$

Thus for all $\epsilon > 0$,

$$P(\lim_n X_n \neq X) \geq P(|X_n - X| \geq \epsilon \,\text{i.o.}) \tag{301}$$

Thus if $\{X_n\}$ converges to X a.s. then $P(\lim_n X_n \neq X) = 0$, and hence (301) implies (299). Conversely, the union over all ϵ in (300) can be restricted to rational values, because the event $\{s \mid |X_n(s) - X(s)| \geq \epsilon \,\text{i.o.}\}$ is increasing as ϵ decreases. Hence if (299) holds for every $\epsilon > 0$, then by countable additivity the union over rational values in (300) is zero. Thus by continuity the right side of (300) has probability zero also, and hence so does the left side, and so $\{X_n\}$ converges to X a.s..

QED

6.3 The Borel-Cantelli Lemmas and the $0-1$ Law

Lemma 42 (First Borel-Cantelli Lemma) *Let A_n be a sequence of events, and suppose that $\sum_n P(A_n) < \infty$. Then*

$$P(\limsup A_n) = P(A_n \ i.o.) = 0 \tag{302}$$

Proof: By definition

$$\limsup A_n = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k \subset \bigcup_{k=n}^{\infty} A_k \tag{303}$$

holds for all n. Hence

$$P(\limsup A_n) \leq P(\bigcup_{k=n}^{\infty} A_k) \leq \sum_{k=n}^{\infty} P(A_k) \tag{304}$$

which goes to zero as $n \to \infty$ because the infinite sum converges.

QED

Lemma 43 (Second Borel-Cantelli Lemma) *Let A_n be a sequence of* independent *events, and suppose that $\sum_n P(A_n) = \infty$. Then*

$$P(\limsup A_n) = P(A_n \ i.o.) = 1 \tag{305}$$

Proof: Sufficient to show that $P(\limsup A_n)^c = 0$. Now

$$(\limsup A_n)^c = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k^c \tag{306}$$

so sufficient to show that for all $n \geq 1$,

$$P(\bigcap_{k=n}^{\infty} A_k^c) = 0 \tag{307}$$

Using independence and the inequality $1 - x \leq e^{-x}$ valid for all real x gives

$$\begin{aligned} P(\bigcap_{k=n}^{\infty} A_k^c) &\leq P(\bigcap_{k=n}^{n+m} A_k^c) \\ &= \prod_{k=n}^{n+m} P(A_k^c) \\ &= \prod_{k=n}^{n+m} (1 - P(A_k)) \\ &\leq \exp[-\sum_{k=n}^{n+m} P(A_k)] \end{aligned} \tag{308}$$

By assumption $\sum_{k=n}^{n+m} P(A_k) \to \infty$ as $m \to \infty$, hence $\exp[-\sum_{k=n}^{n+m} P(A_k)] \to 0$.

QED

These Lemmas have a surprising consequence, namely that for an independent sequence A_n, the event $\limsup A_n$, i.e. the event that A_n is true infinitely often, either has probability zero or probability one – it can never have probability 1/2 or 3/4 etc. This is the Borel $0-1$ Law. For example, toss a coin and consider the event "does the pattern HHTTHH appear infinitely often?". By lumping together every 6 successive tosses you get an independent sequence, and then it follows that the event is either certain or impossible (in this case it is certain).

The $0-1$ law was generalized by Kolmogorov to include all events which are determined by the tail of the sequence $A_1, A_2, \ldots$. More precisely, the *tail field* is defined to be

$$\tau = \bigcap_{n=1}^{\infty} \sigma(A_n, A_{n+1}, \ldots) \tag{309}$$

where $\sigma(A, B, \ldots)$ is the σ-algebra generated by the events $A, B, \ldots$. So the meaning of the tail field is that it contains events which for any n do not depend on the first n events $A_1, \ldots, A_n$.

Lemma 44 (Kolmogorov's $0-1$ Law) *If the events $A_1, A_2, \ldots$ are independent, with tail field τ, and A is any event in τ, then $P(A) = 0$ or $P(A) = 1$.*

Corollary 45 *If X is measurable with respect to τ, then X is constant almost surely.*

The meaning of these $0-1$ laws is that every event in the tail field is either certain or impossible, and any measurement which depends only on the tail field is constant. For example, consider the long-run average

$$\overline{X} = \lim_{n\to\infty} \frac{1}{n}\Big(X_1 + \cdots + X_n\Big)$$

We don't know yet that the right side converges. But supposing it does, then the left side is a random variable that depends only on the tail σ-field – for example if you throw away the first 1,000,000 measurements of X it will not change this limit. Hence its value must be a constant. If the X_i all have mean μ, then so does $\overline{X}$. But a constant is equal to its mean, therefore if it exists we deduce that $\overline{X} = \mu$. The next few sections concern the proof that $\overline{X}$ exists.

6.4 Some inequalities

First recall *Markov's inequality*: for any random variable X and for any numbers $a > 0$ and $k > 0$,

$$P(|X| \geq a) \leq \frac{1}{a^k}\,\mathbb{E}[|X|^k] \tag{310}$$

The proof is easy:

$$\begin{aligned}
\mathbb{E}[|X|^k] &= \int |x|^k dF \\
&\geq \int_{|x|\geq a} |x|^k dF \\
&\geq a^k \int_{|x|\geq a} dF \\
&= a^k P(|X| \geq a)
\end{aligned} \tag{311}$$

An important special case of Markov's inequality is called *Chebyshev's inequality*: take $X = Y - \mathbb{E}Y$ and $k = 2$ to get

$$P(|Y - \mathbb{E}Y| \geq a) \leq \frac{1}{a^2} \operatorname{Var}(Y) \tag{312}$$

Exercise 92 Take Y exponential, find a large enough so that the right side of Chebyshev's inequality is less than 0.1.

6.5 Modes of convergence

The Limit Theorems concern the behavior of sums of random variables as the number of summands grows without bound. We need to know how to determine convergence and limiting behavior. There are several ways to do this. One is the notion of strong convergence introduced above which is repeated below in item (1).

Consider a sequence of random variables $\{X_1, X_2, \dots\}$.

(1) the sequence converges to X almost surely (a.s.) if

$$P(\lim_{n\to\infty} X_n = X) = 1 \tag{313}$$

or more precisely, the event $\{\omega \mid \lim_{n\to\infty} X_n(\omega) = X(\omega)\}$ has probability one.

(2) the sequence converges to X in L^2 if

$$\mathbb{E}|X_n - X|^2 \to 0 \tag{314}$$

as $n \to \infty$.

(3) the sequence converges to X in probability if for every $\epsilon > 0$

$$P(|X_n - X| > \epsilon) \to 0 \tag{315}$$

as $n \to \infty$.

(4) the sequence converges to X weakly if

$$\lim_{n\to\infty} P(X_n \leq t) = P(X \leq t) \tag{316}$$

for all t where the cdf $F_X(t) = P(X \leq t)$ is continuous.

These notions of convergence are related as the following lemma shows. Notice that only (1) is a statement about a limiting event.

Lemma 46

$$(1), (2) \implies (3) \implies (4) \tag{317}$$

Proof: (2) $\implies$ (3) follows from the Chebyshev inequality: for any $\epsilon > 0$

$$P(|X_n - X| \geq \epsilon) \leq \frac{\mathbb{E}(X_n - X)^2}{\epsilon^2} \tag{318}$$

The right side converges to zero as $n \to \infty$, therefore so does the left side.

(1) $\implies$ (3): for any $\epsilon > 0$ and any n, a.s. convergence guarantees that

$$P(\bigcup_{k \geq n} \{\omega \,|\, |X_k - X| \geq \epsilon\}) \to 0 \tag{319}$$

Hence

$$P(\{\omega \,|\, |X_k - X| \geq \epsilon\}) \leq P(\bigcup_{k \geq n} \{\omega \,|\, |X_k - X| \geq \epsilon\}) \to 0 \tag{320}$$

(3) $\implies$ (4): let $F(t) = P(X \leq t)$, $F_n(t) = P(X_n \leq t)$ then

$$\begin{aligned} F(t - \epsilon) &= P(X \leq t - \epsilon) \\ &= P(X \leq t - \epsilon, X_n \leq t) + P(X \leq t - \epsilon, X_n > t) \\ &\leq P(X_n \leq t) + P(X_n - X \geq \epsilon) \end{aligned} \tag{321}$$

By assumption the second term goes to zero as $n \to \infty$, and so

$$F(t - \epsilon) \leq \liminf F_n(t) \tag{322}$$

A similar argument works to lower bound $F(t + \epsilon)$ by $\limsup F_n(t)$. Hence for all $\epsilon > 0$,

$$F(t - \epsilon) \leq \liminf F_n(t) \leq \limsup F_n(t) \leq F(t + \epsilon) \tag{323}$$

If t is a continuity point of F then taking $\epsilon \downarrow 0$ shows that $F_n(t) \to F(t)$.

QED

Exercise 93 Suppose that X_n converges to X in probability, and that f is uniformly continuous. Show that $f(X_n)$ converges to $f(X)$ in probability.

Exercise 94 Let X_n be uniform (discrete) on the set $\{1, 2, \ldots, n\}$. Show that for $0 \leq x \leq 1$,

$$\lim_{n\to\infty} P(n^{-1}X_n \leq x) = x \tag{324}$$

6.6 Weak law of large numbers

The first version is easy to prove and the most widely used, so we start here. We have a sequence of random variables $X_1, X_2, \ldots$ which are all independent and identically distributed (IID). We should think of these as successive independent measurements of the same random variable. They have a common mean μ and variance σ^2. We assume that both of these are finite. Recall that two variables X, Y are *uncorrelated* if $\mathbb{E}XY = \mathbb{E}X\,\mathbb{E}Y$, or equivalently $\mathrm{COV}[X, Y] = 0$. This is weaker than independence.

Theorem 47 *Let $S_n = X_1 + \cdots + X_n$ where X_i have a common mean μ and variance σ^2 (both assumed finite), and where the variables are all uncorrelated. Then as $n \to \infty$,*

$$\frac{S_n}{n} \to \mu \quad \textit{in } L^2 \textit{ and in probability} \tag{325}$$

Proof: by Lemma 46 it is sufficient to prove convergence in L^2. Since $\mathbb{E}S_n = n\mu$ we have

$$\mathbb{E}\Big(\frac{S_n}{n} - \mu\Big)^2 = \mathrm{VAR}\frac{S_n}{n} = \frac{1}{n^2}\,\mathrm{VAR}[X_1 + \cdots + X_n] = \frac{1}{n^2}\,n\sigma^2 = \frac{\sigma^2}{n} \tag{326}$$

Hence $S_n/n \to \mu$ in L^2.

QED

The weak law can be used to justify the 'obvious' meaning of probabilities as limiting frequencies of occurrence.

Exercise 95 [Shannon's Theorem] Suppose that X_i are IID discrete taking values $\{1, \ldots, r\}$ with positive probabilities $p_1, \ldots, p_r$. For a sequence

$i_1, \ldots, i_n$ define

$$p_n(i_1, \ldots, i_n) = p_{i_1} \ldots p_{i_n} \tag{327}$$

Define $Y_n = p_n(X_1, \ldots, X_n)$. Show that

$$-\frac{1}{n} \log Y_n \to -\sum_{i=1}^{r} p_i \log p_i \tag{328}$$

with probability 1.

The second version is more general and applies to a broader range of situations, for example when the variance may not exist.

Theorem 48 *Let $S_n = X_1 + \cdots + X_n$ where X_i are IID, and assume that $\mathbb{E}|X_i| < \infty$, so that $\mu = \mathbb{E}X_i$ exists. Then*

$$\frac{S_n}{n} \to \mu \quad \textit{in probability} \tag{329}$$

We will not prove the Theorem here, but note that the conclusion is weaker than in Theorem 47, because it only guarantees convergence in probability.

6.7 Strong law of large numbers

The Strong Law gives conditions for a.s. convergence of the sample mean of a sequence of random variables. We state a first version of the SLLN which can be proved without too much work. Then we state a stronger version that needs more work.

Theorem 49 *Let $X_1, X_2, \ldots$ be IID random variables with $\mu = \mathbb{E}X$. Assume that $\mathbb{E}|X|^p < \infty$ for $p = 1, 2, 3, 4$. Let $S_n = X_1 + \cdots + X_n$, then*

$$\frac{S_n}{n} \to \mu \quad \textit{a.s. as } n \to \infty \tag{330}$$

Proof: without loss of generality we assume $\mu = 0$, as this can be achieved by replacing X by $X - \mu$. Then we wish to show that $n^{-1}S_n \to 0$ a.s.. or more precisely that the event

$$\{\omega \mid \lim_{n\to\infty} n^{-1} S_n(\omega) = 0\} \tag{331}$$

has probability one. Lemma 41 established that this is equivalent to the following statement: for every $\epsilon > 0$,

$$P(|n^{-1}S_n| \geq \epsilon \text{ i.o.}) = 0 \tag{332}$$

Let

$$A_n = \{\omega \,|\, |n^{-1}S_n(\omega)| \geq \epsilon\} \tag{333}$$

The event $\{\omega \,|\, |n^{-1}S_n(\omega)| \geq \epsilon \text{ i.o.}\}$ is just $\limsup A_n$. So we want to show that $P(\limsup A_n) = 0$. By Borel-Cantelli we just need to show that $\sum_k P(A_k)$ is convergent and we are done.

Obviously the convergence of the sum depends on how quickly the terms $P(A_k)$ go to zero as $k \to \infty$. As long as they go faster than k^{-1} we are fine. So that is what we will show. We use Markov's inequality, but now with exponent 4:

$$P(A_k) = P(|S_k| \geq k\epsilon) \leq \frac{\mathbb{E}S_k^4}{k^4\epsilon^4} \tag{334}$$

We have

$$\mathbb{E}S_k^4 = \sum_{a,b,c,d} \mathbb{E}X_aX_bX_cX_d \tag{335}$$

where each index runs from 1 to k. Since $\mathbb{E}X_k = 0$ and the variables are independent, if $a \neq b \neq c \neq d$ then

$$\mathbb{E}X_aX_bX_cX_d = \mathbb{E}X_a\mathbb{E}X_b\mathbb{E}X_c\mathbb{E}X_d = 0 \tag{336}$$

Similarly if three indices are different, say $a = b \neq c \neq d$ then

$$\mathbb{E}X_aX_bX_cX_d = \mathbb{E}X_a^2\mathbb{E}X_c\mathbb{E}X_d = 0 \tag{337}$$

So the only nonzero contributions arise when there are either two or one distinct indices. This gives

$$\mathbb{E}S_k^4 = \sum_a \mathbb{E}X_a^4 + 3\sum_{a\neq b} \mathbb{E}X_a^2\mathbb{E}X_b^2 = k(\mathbb{E}X^4) + 3k(k-1)(\mathbb{E}X^2) \tag{338}$$

As a function of k this expression grows proportionately to k^2, so we have a constant C such that

$$\mathbb{E}S_k^4 \leq Ck^2 \tag{339}$$

Inserting in our bound gives

$$P(A_k) \leq \frac{C}{k^2\epsilon^4} \tag{340}$$

and since the series $\sum k^{-2}$ is finite this proves the result.

QED

So this proves the Strong Law. It can be extended by weakening the conditions as follows.

Theorem 50 *The Strong Law holds for an IID sequence $X_1, X_2, \ldots$ if $\mathbb{E}|X_i| < \infty$.*

Exercise 96 Define the sequence X_n inductively by setting $X_0 = 1$, and selecting X_{n+1} randomly and uniformly from the interval $[0, X_n]$. Prove that $\frac{1}{n} \log X_n$ converges a.s. to a constant, and evaluate the limit.

6.8 Applications of the Strong Law

Renewal theory offers a nice application. Suppose the X_i are positive and IID, then think of $T_k = X_1 + \cdots + X_k$ as the time of the k^{th} occurrence of some event. For example, X_i could be the lifetime of a component (battery, lightbulb etc) which gets replaced as soon as it breaks. Then T_k is the time when the k^{th} component fails, and

$$N_t = \sup\{n \,|\, T_n \leq t\} \tag{341}$$

is the number of breakdowns up to time t.

Lemma 51 *Assuming that $\mathbb{E}X_i = \mu < \infty$, then as $t \to \infty$ we have*

$$\frac{N_t}{t} \to \frac{1}{\mu} \quad a.s. \tag{342}$$

Comment: The result says that the rate of breakdowns converges to the inverse of the lifetime of the components – a result which agrees with our intuition.

Proof: the proof is an application of the SLLN. We know from SLLN that

$$\frac{T_n}{n} \to \mu \quad \text{a.s.} \tag{343}$$

as $n \to \infty$. Also we have

$$T_{N_t} \leq t < T_{N_t+1} \tag{344}$$

so therefore

$$\frac{T_{N_t}}{N_t} \leq \frac{t}{N_t} \leq \frac{T_{N_t+1}}{N_t+1} \, \frac{N_t+1}{N_t} \tag{345}$$

Now $T_n < \infty$ for all n and hence $N_t \to \infty$ as $t \to \infty$. By SLLN there is an event B_1 with $P(B_1) = 1$ such that

$$N_t \to \infty \quad \text{as } t \to \infty \quad \text{and} \quad \frac{T_n}{n} \to \mu \quad \text{as } n \to \infty \tag{346}$$

Therefore on this event we also have

$$\frac{T_{N_t}}{N_t} \to \mu \quad \text{and} \quad \frac{N_t+1}{N_t} \to 1 \quad \text{as } t \to \infty \tag{347}$$

Hence by the pinching inequalities we get

$$\frac{t}{N_t} \to \frac{1}{\mu} \quad \text{as } t \to \infty \tag{348}$$

QED

7 Moment Generating Function

One useful way to present the information about a random variable is through its moment generating function (mgf). We will use this representation to derive the Central Limit Theorem and also large deviation bounds.

7.1 Moments of X

For $k \geq 0$ the quantity $\mathbb{E}[X^k]$ is called the k^{th} moment of the r.v. X. These are defined by integrals which may or may not exist:

$$\mathbb{E}X^k = \int x^k \, dF(x) \tag{349}$$

To see an example where moments do not exist, consider the Cauchy distribution with pdf

$$f(x) = \frac{a}{\pi} \frac{1}{x^2 + a^2} \tag{350}$$

with $a > 0$. The slow decay of f as $x \to \infty$ means that all integrals $\int |x|^p f(x) dx$ with $p \geq 1$ do not exist. So the moments of X do not exist for all $k \geq 1$.

The first and second moments are the most important features of a random variable and usually exist. They produce the mean and variance which are the most widely used statistics of X.

Exercise 97 Compute the mean and variance for the uniform and the exponential pdf's.

7.2 Moment Generating Functions

The moment generating function (mgf) $M(t)$, if it exists, is defined for $t \in \mathbb{R}$ by

$$M(t) = \mathbb{E}[e^{tX}] = \int_{\mathbb{R}} e^{tx} \, dF(x) \tag{351}$$

Let us assume for the moment that $M(t)$ exists for all t in an interval containing 0. Then ignoring technical issues for a moment we may differentiate

and find

$$\begin{aligned}\frac{d}{dt}M(t) &= \frac{d}{dt}\int_{\mathbb{R}} e^{tx}dF(x) \\ &= \int_{\mathbb{R}} \frac{d}{dt}\, e^{tx}dF(x) \\ &= \int_{\mathbb{R}} x\, e^{tx}dF(x) \\ &= \mathbb{E}[Xe^{tX}] \end{aligned} \tag{352}$$

Now setting $t = 0$ we find

$$\frac{d}{dt}M(t)|_{t=0} = \mathbb{E}X \tag{353}$$

Thus the first derivative of the mgf gives the mean value. By taking the second derivative we find

$$\frac{d^2}{dt^2}M(t)|_{t=0} = \mathbb{E}X^2 \tag{354}$$

and similarly for all $k \geq 1$ (again assuming that this formal operation can be justified)

$$\frac{d^k}{dt^k}M(t)|_{t=0} = \mathbb{E}X^k \tag{355}$$

This explains the name mgf, and also gives a procedure for recovering the moments from the mgf. As an example, consider the normal distribution. The pdf is

$$f(x) = \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2} \tag{356}$$

The mgf is

$$M(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx}e^{-x^2/2}dx \tag{357}$$

Complete the square to get

$$M(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{t^2/2}e^{-(x-t)^2/2}dx = e^{t^2/2} \tag{358}$$

This is defined for all $t \in \mathbb{R}$.

Exercise 98 For the exponential with pdf $f(x) = ke^{-kx}$ show that $M(t)$ is defined for all $t < k$. Deduce that the n^{th} moment is $n!k^{-n}$.

Back to technicalities. Assume that $M(t)$ is defined throughout an interval $(-t_0, t_0)$ containing 0. So the function e^{tx} is integrable for all $t \in (-t_0, t_0)$. Recall the Taylor series expansion

$$\left|e^{tx}\right| = \left|\sum_{k=0}^{\infty} \frac{(tx)^k}{k!}\right| \leq \sum_{k=0}^{\infty} \frac{|tx|^k}{k!} = e^{|tx|} \tag{359}$$

By assumption the right side is integrable on $(-t_0, t_0)$ and uniformly bounds the partial sums, so the Dominated Convergence Theorem implies

$$\sum_{k=0}^{\infty} \frac{t^k}{k!} \int x^k \, dF = \sum_{k=0}^{\infty} \int \frac{(tx)^k}{k!} \, dF = \int \sum_{k=0}^{\infty} \frac{(tx)^k}{k!} \, dF = \int e^{tx} \, dF = M(t) \tag{360}$$

This is the Taylor series expansion for $M(t)$, and the Taylor coefficients are $\int x^k dF = \mathbb{E}[X^k]$. Since this is valid in a nonzero interval around 0 it follows that

$$\frac{d^k}{dt^k} M(t)|_{t=0} = \mathbb{E}X^k \tag{361}$$

as desired. So the key requirement is existence of the mgf in some open interval containing 0.

Exercise 99 Suppose X is continuous and has a power-law tail, meaning that there are $C, K < \infty$ and $s > 1$ such that

$$f(x) \geq C \, |x|^{-s} \quad \text{for } |x| \geq K \tag{362}$$

Show that $M(t)$ exists only at $t = 0$. Do the moments exist?

Exercise 100 Suppose $M(t_1) < \infty$ and $M(t_2) < \infty$, where $t_1 < t_2$. Show that $M(s) < \infty$ for all $t_1 \leq s \leq t_2$.

The next property states that the mgf of a sum is the product of the mgf's. This is one of the most useful properties of the mgf.

Suppose $X_1, \ldots, X_n$ are independent and their mgf's $M_1, \ldots, M_n$ all exist in some interval containing 0. Then the mgf of $X_1 + \cdots + X_n$ exists in the same interval and is equal to

$$\mathbb{E}[e^{t(X_1+\cdots+X_n)}] = M_1(t)\, M_2(t) \ldots M_n(t) \tag{363}$$

Exercise 101 Find the mgf of a sum of n IID random variables uniform on $[0, L]$.

8 The Central Limit Theorem

The Law of Large Numbers says that the sample mean converges to the true mean, but does not say anything about how the convergence occurs. The Central Limit Theorem gives more precise information, and in particular shows a universal behavior for the mode of convergence. Recall that Z is a standard normal random variable if

$$P(a \leq Z < b) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\, dx \tag{364}$$

Recall also that mgf of the standard normal is

$$M_Z(t) = e^{t^2/2} \tag{365}$$

We will say that the sequence of random variables X_n converges to X *in distribution* if for all $a < b$

$$P(a \leq X_n < b) \to P(a \leq X < b) \tag{366}$$

as $n \to \infty$. This is also called *weak convergence.*

Theorem 52 *Let $X_1, X_2, \ldots$ be IID with finite mean $\mathbb{E}X_i = \mu$ and finite variance $\mathrm{VAR}[X_i] = \sigma^2$. Let $S_n = X_1 + \cdots + X_n$. Then the sequence*

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \tag{367}$$

converges to the standard normal in distribution.

Recall that the LLN says that $\frac{S_n - n\mu}{n}$ converges to zero a.s.. So the CLT tells us about the rate of convergence, namely that if we scale up by the factor $\sqrt{n}$ then the sequence settles down to a nonzero limit (the factor of σ is pulled out for convenience).
Proof: (actually just a sketch) First define

$$Y_n = \frac{X_n - \mu}{\sigma}, \quad T_n = Y_1 + \cdots + Y_n \tag{368}$$

Then Y_n has mean zero and variance 1, and

$$T_n = \frac{S_n - n\mu}{\sigma} \tag{369}$$

So it is sufficient to prove that the sequence $T_n/\sqrt{n}$ converges weakly to the standard normal. The strategy of proof uses the moment generating function. Let

$$M_n(t) = \mathbb{E}e^{tT_n/\sqrt{n}}, \quad M(t) = \mathbb{E}e^{tY} \tag{370}$$

Now T_n is a sum of independent random variables, and hence $e^{tT_n/\sqrt{n}}$ is a product:

$$e^{tT_n/\sqrt{n}} = e^{tY_1/\sqrt{n}} \dots e^{tY_n/\sqrt{n}} \tag{371}$$

Because these are independent, the expectation factors:

$$M_n(t) = \mathbb{E}[e^{tY_1/\sqrt{n}}] \dots \mathbb{E}[e^{tY_n/\sqrt{n}}] = M(t/\sqrt{n})^n \tag{372}$$

Now clearly $t/\sqrt{n} \to 0$ as $n \to \infty$, so we can make a good approximation using the Taylor series expansion for e^{tY}:

$$\begin{aligned} M(t/\sqrt{n}) &= \mathbb{E}[1 + \frac{tY}{\sqrt{n}} + \frac{(tY)^2}{2n} + \cdots] \\ &= 1 + \frac{t^2}{2n} + R_n(t) \end{aligned} \tag{373}$$

where $R_n(t)$ denotes the remaining terms in the series, and where we used $\mathbb{E}Y = 0$ and $\mathbb{E}Y^2 = 1$. If we substitute back into (372), this gives

$$\begin{aligned} M_n(t) &= \left(1 + \frac{t^2}{2n} + R_n(t)\right)^n \\ &= \left(1 + \frac{t^2}{2n}\right)^n \left(1 + R_n(t)\ (1 + \frac{t^2}{2n})^{-1}\right)^n \end{aligned} \tag{374}$$

The first factor on the right side of (374) has a simple limit as $n \to \infty$:

$$\lim_{n\to\infty} \left(1 + \frac{t^2}{2n}\right)^n = e^{t^2/2} \tag{375}$$

which of course is the mgf of the standard normal. Therefore the proof of Theorem 52 reduces to the following steps:

Step 1: show that for all t,

$$\lim_{n\to\infty} \left(1 + R_n(t)\ (1 + \frac{t^2}{2n})^{-1}\right)^n = 1 \tag{376}$$

and hence deduce from (374) that $M_n(t) \to e^{t^2/2}$.

Step 2: show that the pointwise convergence of the mgf's $M_n(t)$ to a limit $M_\infty(t)$ implies the pointwise convergence of the cdf's $F_n(x)$ to a cdf $F_\infty(x)$.

Step 3: show that there is only one cdf with mgf $e^{t^2/2}$, and hence conclude that the cdf's $F_n(x)$ converge pointwise to the cdf of the standard normal.

There are serious technical problems to be overcome is completing the three steps described above. Not least is the issue that the mgf of Y may not exist (recall the example of the Cauchy distribution). The way out of this difficulty is peculiar: we use complex values of t. Specifically, the mgf is replaced by the characteristic function (chf)

$$\phi(t) = \mathbb{E}e^{itY} = \int e^{itY}\ dF \tag{377}$$

where $i = \sqrt{-1}$. This clever idea turns out to resolve all the technical difficulties. First, because the magnitude of e^{ity} is one, the integral always exists: there is no integration problem. As a second benefit there is an explicit formula (very similar to the inversion formula for the Fourier transform) which returns $F(y)$ as a function of $\phi(t)$: this settles the problem raised in Step 3. There is still work to be done, but there are no remaining obstacles.

QED

8.1 Applications of CLT

The CLT is useful as it gives an approximation for probabilities when n is large.

Binomial Flip a fair coin and let $X_k = 1$ if Heads, $X_k = 0$ if Tails on k^{th} flip. So S_n is the total number of Heads after n flips. Since $\mathbb{E}X = 1/2$ and $\mathrm{VAR}[X] = 1/4$ the CLT says that

$$\sqrt{\frac{4}{n}}\,(S_n - \frac{n}{2}) \to Z \tag{378}$$

where Z is a standard normal. More precisely stated, for any $a < b$

$$P(a < \sqrt{\frac{4}{n}}\,(S_n - \frac{n}{2}) \le b) \to \frac{1}{\sqrt{2\pi}} \int_a^b e^{-t^2/2}\,dt \tag{379}$$

The CLT is an asymptotic result and gives no information about the rate of convergence. In special cases it is possible to say something about this, as we will see in the next section.

Exercise 102 Let X_i, $i = 1, \ldots, 10$ be independent random variables, each being uniformly distributed over $[0, 1]$. Use the CLT to estimate the probability that $X_1 + \cdots + X_{10}$ exceeds 7.

Exercise 103 If X and Y are Poisson random variables with means λ and μ, then $X + Y$ is Poisson with mean $\lambda + \mu$. Use this fact and the CLT to calculate

$$\lim_{n\to\infty} e^{-n} \sum_{k=0}^{n} \frac{n^k}{k!}$$

[Hint: let X_n be Poisson with mean n, and write the quantity above as the probability of an event involving X_n]

8.2 Rate of convergence in LLN

The LLN says that S_n/n converges to the mean μ, but says little about the rate of convergence. In fact the convergence is exponentially fast, as the following *large deviations* result shows. We need a somewhat stronger assumption to prove the result. This assumption is that the moment generating functions of the variables X_n exist and are finite in some open interval containing zero. Recall that this condition guarantees that all the moments are finite and can be obtained by differentiation of the mgf.

Lemma 53 *Suppose $X_1, X_2, \ldots$ are IID with mean μ. Suppose also that there are $a, b > 0$ such that $M_{X_i}(t) < \infty$ for all $t \in (-a, b)$. Then for all n and all $\epsilon > 0$,*

$$P\left(\frac{S_n}{n} \geq \mu + \epsilon\right) \leq e^{-n\lambda} \tag{380}$$

where

$$e^{-\lambda} = \inf_{0<s<b} (e^{-s(\mu+\epsilon)}\, M_{X_i}(s)) < 1 \tag{381}$$

Proof: first note that for $0 < s < b$, for any random variable Y,

$$\begin{aligned} P(Y \geq 0) &= P(e^{sY} \geq 1) \\ &\leq \frac{\mathbb{E}e^{sY}}{1} \\ &= M_Y(s) \end{aligned} \tag{382}$$

where we used Markov's inequality in the second step. Since this holds for all s we get

$$P(Y \geq 0) \leq \inf_{0<s<b} M_Y(s) \tag{383}$$

Now let $Y_i = X_i - \mu - \epsilon$, then

$$M_{Y_i}(s) = e^{-s(\mu+\epsilon)} M_{X_i}(s) \tag{384}$$

Hence

$$\begin{aligned} P(\frac{S_n}{n} \geq \mu + \epsilon) &= P(Y_1 + \cdots + Y_n \geq 0) \\ &\leq \inf_{0<s<b} M_{Y_1+\cdots+Y_n}(s) \\ &= \inf_{0<s<b} (M_{Y_i}(s))^n \end{aligned} \tag{385}$$

and the result follows.

QED

Exercise 104 Let $X_1, X_2, \ldots$ be IID with distribution $P(X = 1) = P(X = -1) = 1/2$. Find an exponentially decreasing bound for $P(\frac{S_n}{n} \geq 0.1)$.

9 Measure Theory

Here we outline the main ideas which lie behind the construction of probability triples.

9.1 Extension Theorem

Definition 54 *Let Ω be a sample space. A collection of subsets of $\mathcal{S} \subset \Omega$ is a* semialgebra *if it contains $\emptyset$ and Ω, it is closed under finite intersection, and the complement of any set in $\mathcal{S}$ is a finite disjoint union of elements of $\mathcal{S}$.*

The following two examples are the most important and in fact are the motivation for this definition.

Exercise 105 Show that the collection of all intervals $(a, b] \subset \mathbb{R}$ with $-\infty \leq a < b \leq \infty$ is a semialgebra.

Exercise 106 Let Ω be the set of all infinite sequences $(r_1, r_2, \dots)$ where $r_i \in \{1, 2, \dots, N\}$ for each $i = 1, 2, \dots$. For each $n \in \mathbb{N}$ and each $a_1, a_2, \dots, a_n \in \{1, 2, \dots, N\}$ define the *cylinder set*

$$C_{a_1,a_2,\dots,a_n} = \{(r_1, r_2, \dots) \in \Omega \,|\, r_i = a_i \text{ for } 1 \leq i \leq n\} \tag{386}$$

Show that the collection of cylinder sets is a semialgebra.

We will suppose that we have a semialgebra with a proto-measure defined on it, satisfying certain reasonable conditions. The next theorem guarantees that the measure can be extended to a σ-algebra. (We will consider probability measures but the same ideas apply to all finite and σ-finite measures).

Theorem 55 *Let $\mathcal{S}$ be a semialgebra of subsets of Ω. Let $P : \mathcal{S} \to [0, 1]$ satisfy the following conditions:*

a) $P(\emptyset) = 0$, $P(\Omega) = 1$

b) *for any pairwise disjoint finite collection $A_1, \dots, A_n \in \mathcal{S}$, with the union $\bigcup_{i=1}^n A_i \in \mathcal{S}$, we have*

$$P\Big(\bigcup_{i=1}^{n} A_i\Big) \geq \sum_{i=1}^{n} P(A_i) \tag{387}$$

c) *if* $A \subset \bigcup_{n=1}^{\infty} A_n$ *with* $A_1, A_2, \cdots \in \mathcal{S}$ *(countable) then*

$$P(A) \leq \sum_{n=1}^{\infty} P(A_n) \tag{388}$$

The proof is quite long so we break it up into a number of lemmas. First job is to define the outer measure for all sets in Ω:

$$P^*(A) = \inf\{\sum_{i=1}^{\infty} P(A_i) \,|\, A \subset \bigcup_i A_i,\ A_1, A_2, \cdots \in \mathcal{S}\} \tag{389}$$

Next we must verify the properties of P^* which allow the extension to work.

Lemma 56 $P^*(A) = P(A)$ *for all* $A \in \mathcal{S}$.

Proof: by monotonicity of P we have $P(A) \leq \sum_i P(A_i)$ for every covering, hence $P(A) \leq P^*(A)$. On the other hand, take $A_1 = A$ and the rest empty set then $\sum_i P(A_i) = P(A)$ and hence $P^*(A) \leq P(A)$.

QED

Exercise 107 Show that P^* is monotonic, that is if $A \subset B$ then $P^*(A) \leq P^*(B)$.

So this says that P^* agrees with P on the semialgebra $\mathcal{S}$ hence it really is an extension. The next property says that P^* is countably subadditive.

Lemma 57 *For any collection of sets* $B_1, B_2, \cdots \subset \Omega$,

$$P^*\Big(\bigcup_{n=1}^{\infty} B_n\Big) \leq \sum_{n=1}^{\infty} P^*(B_n) \tag{390}$$

Proof: take any $\epsilon > 0$. For each n there is a cover $\{A_{n,k}\}$ such that $B_n \subset \cup_k A_{n,k}$ and

$$P^*(B_n) \geq \sum_k P(A_{n,k}) - \epsilon 2^{-n}$$

Since $\cup_{n,k} A_{n,k}$ is a cover for $\cup_n B_n$ we have

$$P^*\Big(\bigcup_{n=1}^{\infty} B_n\Big) \leq \sum_{n,k} P(A_{n,k}) \leq \sum_n (P^*(B_n) + \epsilon 2^{-n}) \leq \sum_n P^*(B_n) + \epsilon \quad (391)$$

Since this holds for every $\epsilon > 0$, the result follows.

QED

Definition 58 *A set $A \subset \Omega$ is* measurable *if for every $E \subset \Omega$*

$$P^*(E) = P^*(E \cap A) + P^*(E \setminus A) \quad (392)$$

Denote by $\mathcal{M}$ the collection of all measurable sets.

So a measurable set is one that divides every set into two pieces whose measures add up to the original measure. This is the additivity property that we want so we are using P^* to select the sets where it works out.

Lemma 59 *P^* is countably additive on $\mathcal{M}$.*

Proof: Let $A_1, A_2, \cdots \in \mathcal{M}$ be disjoint, we want to show that P^* is additive on their union. First, since $A_1 \in \mathcal{M}$ we have

$$P^*(A_1 \cup A_2) = P^*(A_1) + P^*(A_2) \quad (393)$$

and by extension P^* is finitely additive on $\mathcal{M}$. Furthermore for any m

$$\sum_{n=1}^{m} P^*(A_n) = P^*(\bigcup_{n=1}^{m} A_n) \leq P^*(\bigcup_{n=1}^{\infty} A_n) \quad (394)$$

This holds for all m, hence

$$\sum_{n=1}^{\infty} P^*(A_n) \leq P^*(\bigcup_{n=1}^{\infty} A_n) \quad (395)$$

The reverse inequality holds by countable subadditivity, hence equality holds.

QED

Next we want to show that $\mathcal{M}$ is a σ-algebra. Start with algebra.

Lemma 60 *$\mathcal{M}$ is closed under complement, finite intersections and finite unions, and contains S.*

Proof: Complement is obvious. Also $\Omega \in \mathcal{M}$. Consider the intersection property: let $A, B \in \mathcal{M}$, and $E \subset \Omega$, then it is sufficient to show that

$$P^*(E) \geq P^*(E \cap (A \cap B)) + P^*(E \cap (A \cap B)^c) \tag{396}$$

(the reverse inequality holds by subadditivity). Now

$$P^*(E \cap (A \cap B)^c) \leq P^*(E \cap A^c \cap B^c) + P^*(E \cap A \cap B^c) + P^*(E \cap A^c \cap B) \tag{397}$$

Two applications of the definition of measurability for A and B give the result.

QED

This result implies finite additivity in the following sense: for any disjoint sets $A_1, \ldots, A_n \in \mathcal{M}$, and any $E \subset S$,

$$P^*(E \cap \bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P^*(E \cap A_i) \tag{398}$$

Exercise 108 Prove this.

Lemma 61 *Let $A_1, A_2, \cdots \in \mathcal{M}$ be pairwise disjoint. Then their union is in $\mathcal{M}$.*

Proof: let $B_m = \bigcup_{i=1}^{m} A_i$, then $B_m \in \mathcal{M}$, so for any $E \subset S$

$$\begin{aligned} P^*(E) &= P^*(E \cap B_m) + P^*(E \cap B_m^c) \\ &= \sum_{i=1}^{m} P^*(E \cap A_i) + P^*(E \cap B_m^c) \\ &\geq \sum_{i=1}^{m} P^*(E \cap A_i) + P^*(E \cap B_\infty^c) \end{aligned} \tag{399}$$

where we write B_∞ for the countable union. Since this holds for all m we get

$$\begin{aligned} P^*(E) &\geq \sum_{i=1}^{\infty} P^*(E \cap A_i) + P^*(E \cap B_\infty^c) \\ &\geq P^*(E \cap B_\infty) + P^*(E \cap B_\infty^c) \end{aligned} \tag{400}$$

and this does it.

QED

Now we can easily conclude that $\mathcal{M}$ is a σ-algebra: a general countable union of sets can be written as a countable union of disjoint sets and then the previous lemma says that this is in $\mathcal{M}$.

Exercise 109 Fill in the details of this statement.

The final step is to show that the original semialgebra $\mathcal{S}$ does belong to $\mathcal{M}$.

Lemma 62 *$\mathcal{S} \subset \mathcal{M}$.*

Proof: this needs a bit of work. Let $E \subset \Omega$, then for any $\epsilon > 0$ there are $A_1, A_2, \cdots \in \mathcal{S}$ such that $E \subset \cup_n A_n$ and

$$P^*(E) \geq \sum_n P(A_n) - \epsilon$$

Now let $A \in \mathcal{S}$. Recall that A^c is a disjoint union of elements of $\mathcal{S}$, say $A^c = C_1 \cup \cdots \cup C_k$. Now

$$\begin{aligned} P^*(E \cap A) + P^*(E \cap A^c) &\leq P^*(\cup_n A_n \cap A) + P^*(\cup_n A_n \cap A^c) \\ &= P^*(\cup_n (A_n \cap A)) + P^*(\cup_n \cup_{i=1}^k (A_n \cap C_k)) \\ &\leq \sum_n \left(P(A_n \cap A) + \sum_{i=1}^k P(A_n \cap C_k) \right) \\ &\leq \sum_n P(A_n) \\ &\leq P^*(E) + \epsilon \end{aligned} \tag{401}$$

Since ϵ is arbitrary this does it.

QED

Putting everything together we deduce the result that P^*, when restricted to $\mathcal{M}$, is a measure that extends P. Hence $(\Omega, \mathcal{M}, P^*)$ is a probability triple.

Exercise 110 A measure P on $\mathcal{A}$ is *complete* if the following is true: if $A \in \mathcal{A}$ and $P(A) = 0$, then every subset $B \subset A$ is also in $\mathcal{A}$. Prove that $\mathcal{M}$ as constructed above is complete.

Having this Extension Theorem under our belt provides a way to construct measures, by starting with a semialgebra and a probability function satisfying the hypotheses of the theorem. In many natural cases this can be done.

9.2 The Lebesgue measure

Here the semialgebra consists of the intervals, we can throw in all the open, half-open and closed intervals. The measure P is just the usual length (we won't worry about finiteness here – or we could just look at subsets of $[0, 1]$). So the hard work is to show that the length function on intervals satisfies the conditions of the Extension Theorem. The verification of (387) is straightforward, since there are a finite number of intervals involved. The condition (388) is harder to prove.

Lemma 63 *Let $I \subset \mathbb{R}$ be a closed bounded interval and suppose $I \subset \bigcup_n I_n$ where $I_1, I_2, \ldots$ are open intervals. Then*

$$P(I) \leq \sum_n P(I_n) \tag{402}$$

Proof: by compactness of I there is a finite cover $I_{j_1}, \ldots, I_{j_k}$ (this is the Heine-Borel theorem). So

$$I \subset I_{j_1} \cup \cdots \cup I_{j_k}$$

The result now follows by ordering the intervals and adding up their lengths.

QED

Exercise 111 Show that (388) holds for this semialgebra.

9.3 Independent sequences

Recall the cylinder sets introduced at the start of this section. It was shown in the Exercise that this collection forms a semialgebra. The probability function is defined by

$$P(C_{a_1,a_2,\ldots,a_n}) = P(a_1)\ldots P(a_n) \tag{403}$$

where $\{P(a)\}$ is a probability assignment on $\{1, 2, \ldots, N\}$. The hard part of the Extension Theorem for this example boils down to the following case.

Lemma 64 *Let $C_1, C_2, \ldots$ be a decreasing sequence of cylinder sets such that $\bigcap C_n = \emptyset$. Then*

$$\lim_{n\to\infty} P(C_n) = 0 \tag{404}$$

The proof of this lemma uses compactness of the product space $\{1, 2, \ldots, N\}^{\mathbb{N}}$ when $\{1, 2, \ldots, N\}$ is equipped with the discrete topology. The finite intersection property (f.i.p.) says that a space is compact if and only if every collection of closed sets with an empty intersection has a finite subcollection whose intersection is empty. The cylinder sets are closed and hence the f.i.p. implies the result.

This construction works for any sequence of IID discrete random variables.

9.4 Product measure

Given two probability triples $(S_1, \mathcal{A}_1, P_1)$ and $(S_2, \mathcal{A}_2, P_2)$, let

$$J = \{A \times B \,|\, A \in \mathcal{A}_1, B \in \mathcal{A}_2\} \tag{405}$$

and define a probability function on J by

$$P(A \times B) = P_1(A)\, P_2(B) \tag{406}$$

The elements of J are called measurable rectangles.

Exercise 112 Verify that J is a semialgebra.

The Extension Theorem may now be applied to define a probability measure on a σ-algebra containing J. The resulting probability triple is called the product measure of $(S_1, \mathcal{A}_1, P_1)$ and $(S_2, \mathcal{A}_2, P_2)$.

10 Applications

10.1 Google Page Rank

The worldwide web is a directed graph, where vertices (or nodes) are webpages and whose edges are links between pages. The goal of the Page Rank algorithm is to assign a *rank* to each page. The rank of a webpage is intended to be a measure of interest for the webpage, which is taken as a proxy for the page's importance.

In its simplest version the rank is described by a very simple and intuitive rule. Let x_j be the rank of the j^{th} page, then

$$x_j = \sum_i w_{ij}\, x_i \tag{407}$$

where w_{ij} is the influence of the i^{th} page's rank on the value of the j^{th} page's rank. This influence is assigned by a simple rule: for each vertex i let k_i be the number of *outward* directed links at i. Then

$$w_{ij} = \begin{cases} (k_i)^{-1} & \text{if there is a link } i \mapsto j \\ 0 & \text{otherwise} \end{cases} \tag{408}$$

Each node has a total 'influence' of size 1, and this is spread equally among its outgoing links. So a webpage that links to many others has a diluted effect on the ranks of other webpages. Also, a very popular webpage will have a lot of incoming links and hence will have a higher rank. And note that the rank of a webpage depends on the ranks of the other pages which link into it.

Assuming that the weights $\{w_{ij}\}$ are known the equations (407) can be solved for the ranks $\{x_i\}$. In fact if we define the matrix T with entries $T_{ij} = w_{ij}$, and let x be the row vector with entries x_i then we can write this as a matrix equation:

$$x = x\, T \tag{409}$$

Of course this is exactly the equation satisfied by the stationary distribution of a Markov chain. So as long as T is ergodic (more on this later) we can solve it by iterating the map starting from a seed. Suppose there are N nodes on the graph then the usual choice for a seed is

$$x_0 = \begin{pmatrix} 1/N & 1/N & \cdots & 1/N \end{pmatrix} \tag{410}$$

Then we define

$$x_1 = x_0 T, \quad x_2 = x_1 T = x_0 T^2, \quad \ldots, \quad x_n = x_0 T^n \tag{411}$$

The Perron-Frobenius theorem says that x_n converges to the solution of the equation $x = xT$. So this gives an iterative way to compute x assuming that the weights are known. Note that N is several billion!!

The 'random surfer model' describes a person who randomly clicks on links and so jumps around the web from page to page. We can imagine the surfer's current page as the state of a Markov chain with N states. Then the transition matrix of this chain is exactly T, and hence the surfer's probability distribution will eventually converge to the stationary distribution. This gives another justification for the choice of this model.

There are some problems with this model. One problem is that the matrix T has many zero entries, and checking that it is ergodic is not easy. A more serious problem is the rate of convergence of the iteration $x_{n+1} = x_n T$ as $n \to \infty$. Recall our convergence result from before:

Aside on convergence [Seneta]: another way to express the Perron-Frobenius result is to say that for the matrix P, 1 is the largest eigenvalue (in absolute value) and w is the unique eigenvector (up to scalar multiples). Let λ_2 be the second largest eigenvalue of P so that $1 > |\lambda_2| \geq |\lambda_i|$. Let m_2 be the multiplicity of λ_2. Then the following estimate holds: there is $C < \infty$ such that for all $n \geq 1$

$$\|P^n - ew^T\| \leq C\, n^{m_2 - 1}\, |\lambda_2|^n \tag{412}$$

So the convergence $P^n \to ew^T$ is exponential with rate determined by the first spectral gap.

So to get rapid convergence we want the second eigenvalue of T to be as small as possible. But in general unless something special is happening we expect that $\lambda_2 \sim 1 - 1/N$ where N is the number of states. Since N is huge this is a disaster for the algorithm.

The solution is to modify the weights as follows:

$$w_{ij} = \begin{cases} \alpha(k_i)^{-1} + (1-\alpha)/N & \text{if there is a link } i \mapsto j \\ (1-\alpha)/N & \text{otherwise} \end{cases} \tag{413}$$

Here α is a new parameter which can be chosen freely. Apparently Google uses $\alpha \sim 0.85$.

The meaning in the random surfer model is that the surfer occasionally gets bored and then randomly jumps to any page on the web. So it mixes things up by allowing the surfer to jump between pages which are not linked. Since the probability for this is very small it is reasonable that we end up with a rank not very different from before. But the rate of convergence is hugely increased.

Exercise 113 Define the transition matrix T_α for $0 \leq \alpha \leq 1$ by

$$(T_\alpha)_{ij} = \begin{cases} \alpha(k_i)^{-1} + (1-\alpha)/N & \text{if there is a link } i \mapsto j \\ (1-\alpha)/N & \text{otherwise} \end{cases} \tag{414}$$

So T_1 is the original page rank matrix. Also define the matrix E by

$$(E)_{ij} = (1-\alpha)/N \quad \text{for all } i,j = 1,\ldots,N \tag{415}$$

a) Show that $T_\alpha = \alpha T_1 + (1-\alpha)/N\, E$.
b) Let x, y be probability vectors satisfying $\sum x_i = \sum y_i = 1$. Show that

$$xT - yT = \alpha(xT_1 - yT_1) \tag{416}$$

c) For all probability vectors x, y show that

$$\|xT - yT\| \leq \alpha \, \|x - y\| \tag{417}$$

d) Show that the second largest eigenvalue of T has absolute value at most α.

10.2 Music generation

A student in my class wanted to generate 'Bach-like' music using a Markov chain. He based his analysis on the Back cello suit 1 prelude 1. This piece has 593 notes, of which there are 32 different notes.

Using the relative frequencies of occurrence he computed several different transition matrices and drew samples from each to see how they sound.

First: the original piece.

Second: just using relative frequencies, jumps are completely random.

Third: using relative frequencies for successive two-note sequences.

Fourth: using using relative frequencies for successive three-note sequences.

10.3 Bayesian inference and Maximum entropy principle

Bayesian inference is based on the idea that you have partial knowledge about the outcome of a random experiment. As you acquire more information about the outcome you are able to update your knowledge.

Let X be a discrete random variable whose value we want to estimate. The *prior* distribution P_0 represents our current state of knowledge about X:

$$P(X = x_i) = P_0(x_i) \quad \text{for all } x_i \in \mathrm{Ran}(X) \tag{418}$$

If you know nothing about X then P_0 is the uniform distribution. Conversely if you know for sure that $X = x_i$ then $P_0(x_i) = 1$ and $P_0(x_j) = 0$ for all other values x_j.

Now you measure another random variable Y, and you get the result $Y = y$. Based on this result, how can you update your probability? The answer is provided by Bayes rule:

$$P(X = x_i \,|\, Y = y) = \frac{P(Y = y \,|\, X = x_i)\, P_0(x_i)}{\sum_j P(Y = y \,|\, X = x_j)\, P_0(x_j)} \tag{419}$$

In practice the measurement usually has two ingredients, namely a control part (which you may be able to choose) and the measurement outcome. Let R be the control part and Y the outcome. Then we get

$$P(X = x_i \,|\, Y = y, R) = \frac{P(Y = y \,|\, X = x_i, R)\, P_0(x_i)}{\sum_j P(Y = y \,|\, X = x_j, R)\, P_0(x_j)} \tag{420}$$

10.3.1 Example: locating a segment

Imagine the following task. A 1D array has N adjacent registers. All registers are set to zero except for a connected segment which are all 1's. You have to locate the segment, but you can only make measurements on individual registers. The length and endpoints of the segment are unknown.

Let $X = (a, b)$ be the positions of the two ends of the segment, so $1 \leq a < b \leq N$. A priori we know nothing about X, so the prior distribution is uniform, that is

$$P_0(a, b) = \begin{cases} c & \text{for } a < b \\ 0 & \text{else} \end{cases} \tag{421}$$

where $c^{-1} = N(N-1)/2$. See Fig 1 where $N = 20$.

At the first step we decide to measure the value Y at some register R_1. Depending on the value obtained (either 0 or 1) we update our probability for X using the Bayes rule (420) above. See Fig 2 where $R_1 = 7$, and $Y_1 = 0$.

We continue making measurements. At each step we choose a position R to measure, then based on the outcome Y we update the probability for X. See Figs 3,4 where $R_2 = 15$, $Y_2 = 1$ and $R_3 = 12$, $Y_3 = 0$.

10.3.2 Maximum entropy rule

How do we decide where to make the next measurement? Better yet, how do we program a robot to make this decision? One way is to use the maximum entropy rule.

Let X be a discrete random variable with pmf $\{(x_i, p_i)\}$. The *entropy* of X is

$$H(X) = -\sum_i p_i \log p_i \tag{422}$$

Conventionally use base 2 logarithm, then entropy unit is bits. The entropy satisfies

$$0 \leq H(X) \leq \log N \tag{423}$$

where N is the number of possible values for X. $H(X)$ measures the *amount of randomness* in X: if X is certain then $H(X) = 0$, if X is uniform then $H(X) = \log N$.

Shannon's observation was that entropy is necessary for information transmission. More entropy means more possible variation, and hence greater capacity for storing and transmitting information. Conversely, by measuring a random variable with higher entropy you are able to learn more. This is the principle of maximum entropy: choose the measurement for which the outcome will have the highest entropy.

10.3.3 Back to example

Let's apply this in our example. Fig 5 shows the entropy of the outcomes corresponding to different choices for the initial measurement. It is maximized at 7 and 13. Next, we measure at 7 and get our outcome $Y = 0$,

thus obtaining the updated pmf for X. Now we compute the entropies for outcomes based on this in Fig 6. It is maximized at 13, 15. We measure at 15 and get $Y = 1$. At the next step the entropy is maximized at 10, 11 in Fig 7. We measure at 12 to get $Y = 0$. The entropy is shown in Fig 8. Note that the entropy gets concentrated near the boundary of the segment as we progress, indicating that we will learn most by measuring here.

The preceding example has been extended to 2D location problems, and is being studied as a prototype for computer vision algorithms for autonomous robots.

10.4 Hidden Markov models

The main idea is that there is a Markov chain operating in the background, but we cannot directly observe the states of this chain. Instead we observe a random variable which is correlated with the state, but does not uniquely identify it. We may wish to infer the actual state from the observed state, but we cannot do this with 100% certainty.

10.4.1 The cheating casino

Here is a simple example. A casino uses two kinds of dice. One type is fair, the other is biased with probability 1/2 to come up 6, and equal probabilities for the other values. The casino secretly switches between the two types of dice. Suppose they switch from fair to biased with probability 0.01 before each roll, and switch from biased to fair with probability 0.5. You (the player) do not know which dice are being used, you only see the rolls of the dice.

This is a hidden 2-state Markov model. The two states are the types of dice. The observed rolls occur with different probabilities depending on the current state. You may wish to figure out which dice were used based on the observed sequence of rolls. This is called 'decoding' the state sequence from the observation sequence.

10.4.2 Formal definition

The HMM consists of

an underlying N-state ergodic Markov chain $X_0, X_1, \ldots$ with transition matrix T (ergodicity is not essential but simplifies the analysis)

for each underlying state, a conditional probability distribution for the M possible observations $v_1, \ldots, v_M$:

$$b_j(k) = P(v = k \,|\, X = j), \quad j = 1, \ldots, N, \quad k = 1, \ldots, M \qquad (424)$$

an initial probability distribution for X_0

With this you can in principle compute the probability of any particular sequence of observations.

10.4.3 Applications

Here are two applications. First, in speech recognition: a speech signal is divided into frames (approx 20 ms long). Each frame is assigned to one of a pre-determined set of possible categories. The speech signal is then a long sequence of these categories. The task is to figure out the underlying sequence of words which produced the observed signal. Randomness arises through variations in the sounds uttered and in the time taken to say the words. Second, in DNA sequence analysis: the sequence consists of the 20 amino acids, while the underlying structure is the protein family. Variations arise from insertions and deletions in the protein sequences.

Given a particular HMM, we want to find the likelihood of an observed sequence (for example to compare different parameter values or models). This is done using the forward-backward procedure.

10.4.4 The forward-backward procedure

Let $Y = (y_0, y_1, y_2, \ldots, y_n)$ be the observed sequence, and let $X = (X_0, X_1, \ldots, X_n)$ be an underlying chain sequence. Then

$$\begin{aligned} P(Y) &= \sum_X P(Y \mid X)\, P(X) \\ &= \sum_{j_1,\ldots,j_n} b_{j_1}(y_1) \cdots b_{j_n}(y_n)\, \pi_{j_0}\, p_{j_0,j_1} \cdots p_{j_{n-1},j_n} \end{aligned}$$

where $\pi_j = P(X_0 = j)$ is the initial distribution. The number of operations here scales as N^n, so we need a more efficient method than direct computation.

For all $0 \le m \le n$ define the *forward joint probability*

$$\alpha_m(j) = P(Y_0 = y_0, Y_1 = y_1, \ldots, Y_m = y_m, X_m = j) \tag{425}$$

Note that

$$\begin{aligned} \alpha_0(j) &= P(Y_0 = y_0, X_0 = j) \\ &= P(Y_0 = y_0 \mid X_0 = j) P(X_0 = j) \\ &= b_j(y_0)\, \pi_j \end{aligned}$$

Also we have an induction rule:

$$\alpha_{n+1}(j) = \left[\sum_i \alpha_n(i)\, p_{ij} \right] b_j(y_{n+1}) \tag{426}$$

Exercise 114 Derive (426).

Finally we get our desired probability as

$$P(Y) = \sum_j \alpha_n(j) \tag{427}$$

Note that each step in (426) needs only N^2 operations, so this is much more efficient than before.

In a similar way we define the *backward joint probability*

$$\beta_m(j) = \begin{cases} P(Y_{m+1} = y_{m+1}, \ldots, Y_n = y_n \,|\, X_m = j) & 0 \le m \le n-1 \\ 1 & m = n \end{cases} \tag{428}$$

Exercise 115 Show that for $0 \le m \le n-1$

$$\beta_m(j) = \sum_i p_{ji}\, b_i(y_{m+1})\, \beta_{m+1}(i) \tag{429}$$

10.4.5 Viterbi algorithm

This addresses the following question: given a sequence of observations, what is the 'best' guess for the underlying Markov chain sequence which produced it? The Viterbi algorithm answers this by finding the state sequence which maximizes the joint probability $P(X, Y)$ where Y is the observed sequence. This is equivalent to maximizing the conditional probability $P(X|Y)$.

Define

$$\delta_m(j) = \max_{j_0, \ldots, j_{m-1}} P(X_0 = j_0, X_1 = j_1, \ldots, X_m = j, Y_0 = y_0, \ldots, Y_m = y_m) \tag{430}$$

So $\delta_m(j)$ is the highest probability along a single path of the Markov chain, up to tim em, which accounts for the first m observations and ends in state j. The initialization is

$$\delta_0(j) = \pi_j\, b_j(y_0) \tag{431}$$

We also have an induction step:

$$\delta_{m+1}(j) = \left[\max_i \delta_m(i)\, p_{ij} \right] b_j(y_{m+1}) \tag{432}$$

Exercise 116 Derive (432).

Define

$$\psi_m(j) = \arg\max_i \left[\delta_{m-1}(i)\, p_{ij} \right] \tag{433}$$

The termination step is

$$P^* = \max_i \delta_n(i), \quad y_n^* = \arg\max_i \delta_n(i) \tag{434}$$

The optimal sequence y_m^* is then computed by backtracking:

$$y_m^* = \psi_{m+1}(y_{m+1}^*) \tag{435}$$

References

[1] P. Billingsley, "Probability and Measure", third edition. Wiley (1995). [Wiley series in probability and mathematical statistics]

[2] R. Durrett, "Probability: theory and examples", second edition. Duxbury Press (1996).

[3] W. Feller, "An introduction to probability theory and its applications", volume 1, second edition. Wilet (1957). [Wiley publications in Statistics]

[4] G. R. Grimmett and D. R. Stirzaker, "Probability and Random Processes", second edition. Oxford University Press (1992).

[5] C. M. Grinstead and J. L. Snell, "Introduction to Probability", AMS (2003).

[6] S. I. Resnick, "A Probability Path", Birkhauser (1999).

[7] J. S. Rosenthal, "A first look at rigorous probability theory", second edition. World Scientific (2006).

[8] S. M. Ross, "Introduction to Probability Models", eighth edition. Academic Press (2003)

[9] H. L. Royden, "Real Analysis", third edition. Macmillan (1988).

[10] E. Seneta, "Non-negative matrices and Markov chains", Springer (2006). [Springer series in Statistics]

[11] Y. Suhov and M. Kelbert, "Probability and Statistics by Example", Volume 1, Basic Probability and Statistics Cambridge University Press (2005)