STAT 870: Applied Probability — Course notes

Richard A. Lockhart

October 10, 2000

Contents

Chapter 1

Introduction

Imagine tossing a coin n times. On trial k write down a 1 if the coin comes up heads and a 0 if it comes up tails. A typical outcome of the experiment will be a sequence, $\omega = (\omega_1, \ldots, \omega_n)$ of zeros and ones. For $n = 3$ for instance there are eight possible outcomes, the elements of the set

$$\Omega = \{(0,0,0),(0,0,1),(0,1,0),(0,1,1),(1,0,0),(1,0,1),(1,1,0),(1,1,1)\}.$$

For the general case the set of all possible outcomes is $\Omega = \{0,1\}^n$ which is just notation for the set of all sequences of n elements chose from $\{0,1\}$; there are 2^n elements in Ω.

Before we do the experiment, we think of it as *random.* I am not going, in this course, to discuss the meaning of the word random, nor the various interpretations of the idea of probability. Instead I want to focus on the mathematics and the modelling. A probability measure on Ω is a function P defined on the set of all subsets of Ω with the following properties:

1. For each subset A of Ω, $P(A)$ is a real number in the set $[0,1]$.
2. If $A_1, \ldots, A_k$ are *pairwise disjoint* (meaning that for $i \neq j$ the intersection $A_i \cap A_j$ which we usually write as $A_i A_j$ is the empty set $\emptyset$) then

$$P(\cup_1^k A_j) = \sum_1^k P(A_j)$$

3. $P(\Omega) = 1$.

Probability modelling is the process of selecting one (or a family) of possible probability measures which makes the best match between the mathematics and the real world phenomenon being described. What constitutes a good match depends on the interpretation of probability being used. In general in this course a long run limiting relative frequency interpretation of $P(A)$ will make the intuition underlying the modelling easiest.

There are actually quite a few possible probability measures on Ω in the examples given. How do we "select" one and how do we explain which one we have selected? Consider for instance the case of $n = 3$. The set Ω has 8 elements and there are $2^8 = 256$ subsets of Ω.

To fully specify a function P on the set of all subsets of Ω, we ought, it seems to specify 256 numbers.

That sort of specification is quite impractical. Instead we *model* by listing some assumptions about the function P. We then try to deduce what P is, or show how to calculate $P(A)$ for any A or at least for lots of interesting A.

Here are three approaches to modelling the coin tossing problem.

The first approach is, essentially, to provide a complete description of P, namely:

$$P(A) = \frac{\text{number of elements in } A}{\text{number of elements in } \Omega}. \tag{1.1}$$

This model has the disadvantage of not telling us how to approach very many other problems.

In fact, this first description is really a deduction from a fairly common model in games of chance, the equally likely outcomes model. In this case our model specifies that if A and B are two singleton sets in Ω that is $A = \{\omega_1\}$ and $B = \{\omega_2\}$ then $P(A) = P(B)$. If there are m elements in Ω, say $\Omega = (\omega_1, \ldots, \omega_m)$ then we have

$$\begin{aligned} P(\Omega) &= P(\cup_1^m \{\omega_j\}) \\ &= \sum_1^m P(\{\omega_j\}) \\ &= mP(\{(\omega_1\}). \end{aligned}$$

Since $P(\Omega) = 1$ we see that $P(\{\omega_i\}) = 1/m$ for each i and then a similar argument proves that this model is the same as the model (1.1).

These two approaches to the fair coin tossing model share a serious defect. They do not naturally make it easy to conceive of experiments in which the set of outcomes is infinite. Consider, for instance the simple game of tossing until you see a head. It is natural to think of Ω as consisting of all sequences of some number k of zeros followed by a one. Or you could just think of Ω as the set $\{0, 1, 2, \ldots\}$ by keeping track of how many tails came before the first head. Either way the natural space Ω is infinite and it is clearly not going to be adequate to assume that each element of Ω has the same probability.

Instead we will describe another method of modelling our coin tossing problem. Consider the model (1.1) for $n = 3$. Consider the event $A = \{\omega : \omega_1 = 1, \omega_2 = 0, \omega_3 = 1\}$. We can write A as the intersection of three events each defined in terms of a different co-ordinate in ω. Put

$$A_1 = \{\omega : \omega_1 = 1\}$$

$$A_2 = \{\omega : \omega_2 = 0\}$$

and

$$A_3 = \{\omega : \omega_3 = 1\}.$$

Then

$$A = A_1 \cap A_2 \cap A_3.$$

Note that $P(A) = 1/8$ and $P(A_i) = 1/2$ for each i so that

$$P(A) = \prod P(A_i).$$

This is a special case of a very general property of our equally likely outcomes model. Consider n tosses. For $i = 1, \ldots, n$ let B_i be some subset of $\{0, 1\}$. (In our little example $B_1 = B_3 = \{1\}$ and $B_2 = \{0\}$.) Define

$$A_i = \{\omega : \omega_i \in B_i\}$$

and

$$A = \cap A_i .$$

It is then possible to prove that

$$P(A) = \prod P(A_i) .$$

Later in the course we will describe this by saying that the random variables X_i defined by $X_i(\omega) = \omega_i$ are independent.

This property forms the basis of the most common modelling tactic for the coin tossing problem. We *assume*

$$P(\{\omega : \omega_i = 1\}) = P(\{\omega : \omega_i = 0\}) = 1/2 \tag{1.2}$$

and that for any set of events of the form given above

$$P(A) = \prod P(A_i) . \tag{1.3}$$

The second assumption is a natural deduction of the long run relative frequency interpretation together with a physical assumption about cause and effect, namely, that the outcome of one toss of the coin is incapable of influencing the outcome of another toss.

The advantage of the second set of assumptions is that it permits generalization in many ways and, in particular, allows infinite spaces Ω. Most of our probability models are like this. We will assume that our probability satisfies identities such as (1.3) and then deduce other properties of P.

Here is the idea. Let us imagine tossing the coin an infinite number of times. Now $\Omega = \{\omega = (\omega_1, \omega_2, \cdots)\}$ is an uncountably infinite set. We will assume that for any n and any event of the form $A = \cap_1^n A_i$ with each $A_i = \{\omega : \omega_i \in B_i\}$ we have

$$P(A) = \prod_1^n P(A_i) . \tag{1.4}$$

This is the assumption that the outcomes of different tosses are independent. This assumption is already strong enough to deduce lots of results. To make it into the fair coin tossing model we add the assumption that

$$P(\{\omega : \omega_i = 1\}) = 1/2 . \tag{1.5}$$

What values of $P(A)$ can be determined from these assumptions? Consider first, any set $A = \{\omega \in \Omega : (\omega_1, \ldots, \omega_n) \in B\}$ where $B \subset \Omega_n = \{0, 1\}^n$. You can prove that the assumptions given guarantee that

$$P(A) = \frac{\text{number of elements in } B}{\text{number of elements in } \Omega_n} . \tag{1.6}$$

In other words, our model specifies that the first n of our infinite sequence of tosses behave like the equally likely outcomes model we started with. What about C_k which is the event that the first head occurs after k consecutive tails? You can write C_k as the intersection

$$C_k = A_1^c \cap A_2^c \cdots \cap A_k^c \cap A_{k+1}$$

where $A_i = \{\omega : \omega_i = 1\}$ and the notation A^c means the complement of A so that $A_i^c = \{\omega : \omega_i = 0\}$. Our assumption guarantees

$$\begin{aligned} P(C_k) &= P(A_1^c \cap A_2^c \cdots \cap A_k^c \cap A_{k+1}) \\ &= P(A_1^c) \cdots P(A_k^c) P(A_{k+1}) \\ &= 2^{-(k+1)} . \end{aligned}$$

It is possible to describe much more complicated events. Here are several which show up in the theorem known as the Strong Law of Large Numbers. See Homework set 1 for problems connected with these events.

$$A_1 \equiv \{\omega : \lim_{n\to\infty} (\omega_1 + \cdots + \omega_n)/n \text{ exists } \}$$

$$A_2 \equiv \{\omega : \lim_{n\to\infty} (\omega_1 + \cdots + \omega_n)/n = 1/2\}$$

$$A_3 \equiv \{\omega : \lim_{n\to\infty} \sum_1^n (2\omega_k - 1)/k \text{ exists } \} .$$

The Strong Law of Large Numbers says, among other things, that for our model $P(A_2) = 1$. In fact, it turns out that $A_3 \subset A_2 \subset A_1$. If $P(A_2) = 1$ then our assumptions guarantee that $P(A_1) = 1$ because $P(A_3)$ must be bigger than $P(A)$ for any subset A of A_1. The proof of the strong law of large numbers actually proves that $P(A_3) = 1$ and so concludes also $P(A_2) = P(A_1) = 1$.

Before we get to that theorem we have some mathematical questions to answer:

1. Do our assumptions (1.4) and (1.5) determine the value of $P(A)$ for any subset A of Ω? [The answer is no.]

2. Do our assumptions determine the value of $P(A_i)$ for $i = 1, 2, 3$? [The answer is yes.]

3. Are our assumptions about the probability P logically consistent? That is, is there any probability measure satisfying (1.4) and (1.5)? [The answer is yes.]

1.1 Probability Definitions

A **Probability Space** is an ordered triple $(\Omega, \mathcal{F}, P)$. The idea is that Ω is the set of possible outcomes of a random experiment, $\mathcal{F}$ is the set of those events, or subsets of Ω whose probability is defined and P is the rule for computing probabilities. Formally:

- Ω is a set.

- $\mathcal{F}$ is a family of subsets of Ω with the property that $\mathcal{F}$ is a σ-field (or σ-algebra):

  1. The empty set $\emptyset$ and Ω are members of $\mathcal{F}$.
  2. $\mathcal{F}$ is closed under complementation. That is, if A is in $\mathcal{F}$ (meaning $P(A)$ is defined) then $A^c = \{\omega \in \Omega : \omega \notin A\}$ is in $\mathcal{F}$ (because we want to be able to say $P(A^c) = 1 - P(A)$).
  3. If $A_1, A_2, \cdots$ are all in $\mathcal{F}$ then so is $A = \cup_{i=1}^{\infty} A_i$. ($A$ is the event that at least one of the A_i happens and we want to be sure that if each of the A_i has a probability then so does this event A.)

- P is a function whose domain is $\mathcal{F}$ and whose range is a subset of $[0, 1]$ which satisfies the axioms for a probability:

  1. $P(\emptyset) = 0$ and $P(\Omega) = 1$.
  2. If $A_1, A_2, \cdots$ are **pairwise disjoint** (or **mutually exclusive**) (meaning for any $j \neq k$ $A_j \cap A_k = \emptyset$) then
  $$P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i) .$$
  This property is called **countable additivity**.

These axioms guarantee that as we compute probabilities by the usual rules, including approximation of an event by a sequence of others we don't get caught in any logical contradictions. Here are some consequences of the axioms for which you should provide a proof.

1. If $A_1, A_2, \cdots, A_n$ are pairwise disjoint then
$$P(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i) .$$
This property is called **finite additivity**.

2. For any event A
$$P(A^c) = 1 - P(A) .$$

3. If $A_1 \subset A_2 \subset \cdots$ are events then
$$P\left(\bigcup_{1}^{\infty} A_i\right) = \lim_{n\to\infty} P(A_n) .$$

4. If $A_1 \subset A_2 \subset \cdots$ then
$$P(\cap_1^{\infty} A_i) = \lim_{n\to\infty} P(A_n) .$$

The most subtle point of the definitions is the σ-field, $\mathcal{F}$. This ingredient is needed to avoid some contradictions which arise if you try to define $P(A)$ for every subset A of Ω when Ω is a set with uncountably many elements.

A vector valued random variable is a function X whose domain is Ω and whose range is in some p dimensional Euclidean space, $\mathbb{R}^p$ with the property that the events whose probabilities we would like to calculate from their definition in terms of X are in $\mathcal{F}$. We will write $X = (X_1, \ldots, X_p)$. We will want to make sense of

$$P(X_1 \le x_1, \ldots, X_p \le x_p)$$

for any constants $(x_1, \ldots, x_p)$. In our formal framework the notation

$$X_1 \le x_1, \ldots, X_p \le x_p$$

is just shorthand for an event, that is a subset of Ω, defined as

$$\{\omega \in \Omega : X_1(\omega) \le x_1, \ldots, X_p(\omega) \le x_p\}$$

Remember that X is a function on Ω so that X_1 is also a function on Ω. In almost all of probability and statistics the dependence of a random variable on a point in the probability space is hidden! You almost always see X not $X(\omega)$.

Now for formal definitions:

The **Borel** σ-field in $\mathbb{R}^p$ is the smallest σ-field in $\mathbb{R}^p$ containing every open ball

$$B_y(r) = \{x \in \mathbb{R}^p : |x - y| < r\}.$$

(To see that there is, in fact, such a "smallest" σ-field you prove the following assertions:

1. The intersection of an arbitrary family of σ-fields is a σ-field. [Homework set 1.]

2. There is at least one σ-field of subsets of $\mathbb{R}^p$ containing every open ball. [Homework set 1.]

Now define the Borel σ-field in $\mathbb{R}^p$ to be

$$\mathcal{B}(\mathbb{R}^p) = \cap \mathcal{F}$$

where the intersection runs over all σ-fields $\mathcal{F}$ which contain every open ball.)

Every common set is a Borel set, that is, in the Borel σ-field. For instance, let O be an open set. Then I will prove that O is Borel. For each x in O there is a point y all of whose co-ordinates are rational numbers and a rational number r such that

$$x \in B_y(r) \subset O$$

Now O is the union of all these $B_y(r)$. (Every $x \in O$ is in one of the $B_y(r)$ and every point in any $B_y(r)$ is in O.) But the union is countable because there are only countably many possible pairs (y, r) with all the co-ordinates rational numbers. [Homework set 1.]

1.2 Random Variables

An $\mathbb{R}^p$ valued **random variable** is a map $X : \Omega \mapsto \mathbb{R}^p$ such that when A is Borel then $\{\omega \in \Omega : X(\omega) \in A\} \in \mathcal{F}$.

Fact: this is equivalent to

$$\{\omega \in \Omega : X_1(\omega) \le x_1, \ldots , X_p(\omega) \le x_p\} \in \mathcal{F}$$

for all $(x_1, \ldots , x_p) \in \mathbb{R}^p$. [Homework set 1.]

Jargon and notation: we write $P(X \in A)$ for $P(\{\omega \in \Omega : X(\omega) \in A)$ and define the **distribution** of X to be the map

$$A \mapsto P(X \in A)$$

which is a probability on the set $\mathbb{R}^p$ with the Borel σ-field rather than the original Ω and $\mathcal{F}$.

The **Cumulative Distribution Function** (or CDF) of X is the function F_X on $\mathbb{R}^p$ defined by

$$F_X(x_1, \ldots , x_p) = P(X_1 \le x_1, \ldots , X_p \le x_p)$$

Properties of F_X (or just F when there's only one CDF under consideration):

1. $0 \le F(x) \le 1$.
2. $x > y \Rightarrow F(x) \ge F(y)$ (F is monotone non-decreasing).
3. $\lim_{x\to-\infty} F(x) = 0$
4. $\lim_{x\to\infty} F(x) = 1$
5. $\lim_{x\searrow y} F(x) = F(y)$ (F is right continuous).
6. $\lim_{x\nearrow y} F(x) \equiv F(y-)$ exists.
7. $F(x) - F(x-) = P(X = x)$.
8. $F_X(t) = F_Y(t)$ for all t implies that X and Y have the same distribution, that is, $P(X \in A) = P(Y \in A)$ for any (Borel) set A.

The distribution of a random variable X is **discrete** (we also call the random variable discrete) if there is a countable set $x_1, x_2, \cdots$ such that

$$P(X \in \{x_1, x_2 \cdots\}) = 1 = \sum_i P(X = x_i)\,.$$

In this case the **discrete density** or **probability mass function** of X is

$$f_X(x) = P(X = x)\,.$$

The distribution of a random variable X is **absolutely continuous** if there is a function f such that

$$P(X \in A) = \int_A f(x)dx$$

for any (Borel) set A. This is a p dimensional integral in general. This condition is equivalent (when $p = 1$) to

$$F(x) = \int_{-\infty}^{x} f(y)\, dy$$

or for general p to

$$F(x_1, \ldots, x_p) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_p} f(y_1, \ldots, y_p) dy_1 \cdots dy_p \, .$$

We call f the **density** of X. For most values of x we then have F is differentiable at x and, for $p = 1$

$$F'(x) = f(x)$$

or in general

$$\frac{\partial^p}{\partial x_1 \cdots \partial x_p} F(x_1, \ldots, x_p) = f(x_1, \ldots, x_p) \, .$$

1.3 Appendix on Measurability

1. A map f from $\mathbb{R}^p$ to $\mathbb{R}^q$ is Borel if for each Borel set $B \subset \mathbb{R}^q$ the inverse image

$$f^{-1}(B) = \{x \in \mathbb{R}^p : f(x) \in B\}$$

   is a Borel set. Students are sometimes confused between inverse functions which exist only when f is bijective and inverse images like $f^{-1}(B)$ which are always defined.

2. If f from $\mathbb{R}^p$ to $\mathbb{R}^q$ and g from $\mathbb{R}^q$ to $\mathbb{R}^r$ are Borel then so is $h = g \circ f$, the composition of g and f from $\mathbb{R}^p$ to $\mathbb{R}^r$.

3. Every continuous function is Borel.

4. If $f : \mathbb{R}^p \mapsto \mathbb{R}^{q_1}$ and $g : \mathbb{R}^p \mapsto \mathbb{R}^{q_2}$ are Borel then $h = (f, g) : \mathbb{R}^p \mapsto \mathbb{R}^{q_1+q_2}$ is Borel.

5. f is Borel if and only if $f^{-1}(O)$ is Borel for each open set O. The analogue for closed sets holds too.

6. If f_n is a sequence of Borel functions then

$$g \equiv \limsup f_n$$

   is Borel. Similarly for $\liminf f_n$ and many other possibilities.

7. If X is an $\mathbb{R}^p$ valued rv and f is a Borel map from $\mathbb{R}^p$ to $\mathbb{R}^q$ then $Y = f(X)$ is an $\mathbb{R}^q$ valued random variable.

As an application suppose that $f_n, n = 1, 2, \cdots$ is a sequence of non-negative real valued Borel functions. Define

$$g(x) = \sum_1^\infty f_n(x)$$

The sum might be infinite for some x values; all the concepts above can be adapted to permit functions to be $+\infty$ on a Borel set and $-\infty$ on some other Borel set and still use the Borel jargon but we will just assume the sum is finite for all x. Let g_n be the nth partial sum:

$$g_n = \sum_1^n f_m \, .$$

If we can prove each g_m is measurable then 6 above shows $g = \lim g_n = \limsup g_n$ is Borel. The function $h : \mathbb{R}^2 \mapsto \mathbb{R}$ defined by

$$h(x, y) = x + y$$

is continuous so Borel. If g_n is Borel for some n then $\phi = (g_n, f_{n+1})$ is Borel by 4 above and then $g_{n+1} = h \circ \phi$ is Borel by 2. Since $g_1 = f_1$ is Borel we conclude by induction that every g_n is Borel.

1.4 Appendix on Lebesgue Measure

There is a probability measure λ defined on the Borel subsets of $[0, 1]$ which agrees with length for intervals: $\lambda([a, b]) = b - q$ for $0 \le a \le b \le 1$. The definition of λ can be extended to arbitrary Borel subsets of $\mathbb{R}$; a p dimensional generalization of volume is also possible. This extension to $\mathbb{R}$ has all the properties of a probability measure except that $\lambda(\mathbb{R}) = \infty$ not 1. Such an object is a **measure**. The measure λ is **translation invariant**:

$$\lambda(c + B) = \lambda(B)$$

for any Borel set B and real number c. The notation $c + B$ means

$$\{y : y = c + x \text{ for some } x \in B\} \, .$$

A Borel set B is called a Lebesgue null set if $\lambda(B) = 0$. Examples include:

1. The rational numbers $\mathbb{Q}$.

2. The Cantor set (all real numbers x in $[0, 1]$ whose expansions in base 3 may be written without the digit 1). The set shows up in lots of math courses as a set whose cardinality is the same as that of $\mathbb{R}$ but whose Lebesgue measure is 0.

If $A \subset B$ and B is a Lebesgue null Borel set then it is natural to define $\lambda(A) = 0$ even if A is not Borel. We call all such A Lebesgue null sets. A set $A \subset \mathbb{R}$ is **Lebesgue measurable** if we can write $A = B \cup N$ with B Borel and N a Lebesgue null set. The family of all Lebesgue measurable sets is a σ-field (which is much larger than the Borel σ-field).

A property of a function $f(x)$ which holds except for a set N of x which is a Lebesgue null set is said to hold **almost everywhere** or *for almost all* x.

Chapter 2

Independence, conditional distributions

Definition: Events A and B are independent if

$$P(AB) = P(A)P(B)\,.$$

(Note the notation: AB is the event that both A and B happen. It is also written $A \cap B$.)
Definition: Events A_i, $i = 1, \ldots, p$ are **independent** if

$$P(A_{i_1} \cdots A_{i_r}) = \prod_{j=1}^{r} P(A_{i_j})$$

for any set of distinct indices $i_1, \ldots, i_r$ between 1 and p.
Example: $p = 3$

$$\begin{aligned} P(A_1 A_2 A_3) &= P(A_1)P(A_2)P(A_3) \\ P(A_1 A_2) &= P(A_1)P(A_2) \\ P(A_1 A_3) &= P(A_1)P(A_3) \\ P(A_2 A_3) &= P(A_2)P(A_3)\,. \end{aligned}$$

You need all these equations to be true for independence!
Example: Toss a coin twice. If A_1 is the event that the first toss is a Head, A_2 is the event that the second toss is a Head and A_3 is the event that the first toss and the second toss are different then $P(A_i) = 1/2$ for each i and for $i \neq j$

$$P(A_i \cap A_j) = \frac{1}{4}$$

but

$$P(A_1 \cap A_2 \cap A_3) = 0 \neq P(A_1)P(A_2)P(A_3)\,.$$

Definition: Random variables X and Y are **independent** if

$$P(X \in A; Y \in B) = P(X \in A)P(Y \in B)$$

for all A and B.
Definition: Random variables $X_1, \ldots, X_p$ are **independent** if

$$P(X_1 \in A_1, \cdots, X_p \in A_p) = \prod P(X_i \in A_i)$$

for any choice of $A_1, \ldots, A_p$.
Definition: σ-fields $\mathcal{F}_1, \ldots, \mathcal{F}_p$ are **independent** if

$$P(A_1 \cap \cdots \cap A_p) = \prod P(A_i)$$

for any choice of events $A_1 \in \mathcal{F}_1, \ldots, A_p \in \mathcal{F}_p$.

Theorem 1

1. *If X and Y are independent then*

$$F_{X,Y}(x,y) = F_X(x)F_Y(y)$$

for all x, y.

2. *If X and Y are independent and have joint density $f_{X,Y}(x,y)$ then X and Y have densities, say f_X and f_Y, and, for almost every pair (x,y)*

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)\,.$$

3. *If X and Y are independent and have marginal densities f_X and f_Y then (X,Y) has joint density $f_{X,Y}(x,y)$ given by*

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)\,.$$

4. *If*

$$F_{X,Y}(x,y) = F_X(x)F_Y(y)$$

for **all** *x, y then X and Y are independent.*

5. *If (X,Y) has density $f(x,y)$ and there are functions $g(x)$ and $h(y)$ such that*

$$f(x,y) = g(x)h(y)$$

for **almost all** *(x,y) then X and Y are independent and they each have a density given by*

$$f_X(x) = g(x) / \int_{-\infty}^{\infty} g(u)du$$

and

$$f_Y(y) = h(y) / \int_{-\infty}^{\infty} h(u)du\,.$$

Proof:

1. Since X and Y are independent so are the events $X \le x$ and $Y \le y$; hence

$$P(X \le x, Y \le y) = P(X \le x)P(Y \le y)$$

2. Suppose $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$. Let $B \subset \mathbb{R}^q$ be Borel. The event $Y \in B$ is the same event as $(X, Y) \in \mathbb{R}^p \times B$. Hence

$$\begin{aligned}
P(Y \in B) &= P((X,Y) \in \mathbb{R}^p \times B) \\
&= \int_{\mathbb{R}^p \times B} f_{X,Y}(x,y) dx\, dy \\
&= \int_B g(y) dy
\end{aligned}$$

where the function g is defined by

$$g(y) = \int_{\mathbb{R}^p} f_{X,Y}(x,y) dx$$

The function g satisfies the definition of density showing that Y has a density. Similarly X has a density. Then for any Borel sets $A \subset \mathbb{R}^p$ and $B \subset \mathbb{R}^q$

$$\begin{aligned}
P(X \in A, Y \in B) &= \int_A \int_B f_{X,Y}(x,y) dy dx \\
P(X \in A) P(Y \in B) &= \int_A f_X(x) dx \int_B f_Y(y) dy \\
&= \int_A \int_B f_X(x) f_Y(y) dy dx
\end{aligned}$$

See the appendix on Lebesgue integration for the theorems of Fubini and Tonelli which give conditions under which the double integral is equal to the iterated integral. These theorems are being used here.

Since $P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$ we see that for any sets A and B

$$\int_A \int_B [f_{X,Y}(x,y) - f_X(x) f_Y(y)] dy dx = 0$$

It follows (see the appendix) that the quantity in [] is 0 (for almost every pair (x, y)).

3. For any A and B we have

$$\begin{aligned}
P(X \in A, Y \in B) &= P(X \in A) P(Y \in B) \\
&= \int_A f_X(x) dx \int_B f_Y(y) dy \\
&= \int_A \int_B f_X(x) f_Y(y) dy dx
\end{aligned}$$

If we **define** $g(x, y) = f_X(x) f_Y(y)$ then we have proved that for $C = A \times B$

$$P((X,Y) \in C) = \int_C g(x,y) dy dx$$

To prove that g is the joint density of (X, Y) we need only prove that this integral formula is valid for an arbitrary Borel set C, not just a rectangle $A \times B$. This is proved via a *monotone class* argument. You prove that the collection of sets C for which the identity holds has closure properties which guarantee that this collection includes the Borel sets.

A **Monotone Class** is a collection $\mathcal{C}$ of subsets of a given set which is closed under the operations of increasing countable unions and decreasing countable intersections. That is it has the properties:

(a) If $A_1 \subset A_2 \subset \cdots$ are in $\mathcal{C}$ then $\cup_1^\infty A_i \in \mathcal{C}$.

(b) If $A_1 \supset A_2 \supset \cdots$ are in $\mathcal{C}$ then $\cap_1^\infty A_i \in \mathcal{C}$.

Lemma 1 *The smallest monotone class containing a field $\mathcal{F}_l$ is a σ-field.*

A field is defined just the same way a σ-field is but with only finite unions and intersections. To prove the Lemma you let $\mathcal{C}$ be the smallest monotone class containing $\mathcal{F}_o$; this is simply the intersection of all monotone classes containing $\mathcal{F}_o$. Notice the similarity to the definition of the Borel σ-field.

Whenever you want to prove a fact about an object defined in this way you will need an indirect proof. Consider first the collection $\mathcal{M}$ of all those sets $A \in \mathcal{C}$ for which $A^c \in \mathcal{C}$. You should prove that $\mathcal{M}$ is a monotone class. Since $\mathcal{F}_o$ is a field and a subset of $\mathcal{C}$ you see that $\mathcal{M}$ contains $\mathcal{F}_o$. Since $\mathcal{C}$ is the smallest monotone class containing $\mathcal{F}_o$ and $\mathcal{M}$ is another monotone class containing $\mathcal{F}_o$ we see t hat $\mathcal{C} \subset \mathcal{M}$. But this just means that every $A \in \mathcal{C}$ has the property $A^c \in \mathcal{C}$. In other words $\mathcal{C}$ is closed under the operation of taking complements, one of the defining properties of a σ-field.

Next fix a set $A \in \mathcal{F}_o$. Let $\mathcal{M}$ be the collection of all $B \in \mathcal{C}$ for which $A \cup B \in \mathcal{C}$. Again you can check that $\mathcal{M}$ is a monotone class and that $\mathcal{M}$ contains $\mathcal{F}_o$. As in the previous part this shows that $\mathcal{C} \subset \mathcal{M}$. In other words for every $A \in \mathcal{F}_o$ and every $B \in \mathcal{C}$ we have $A \cup B \in \mathcal{C}$.

Finally fix a set A in $\mathcal{C}$. Let $\mathcal{M}$ be the collection of all $B \in \mathcal{C}$ for which $A \cup B \in \mathcal{C}$. Again you can check that $\mathcal{M}$ is a monotone class. The previous step showed that $\mathcal{M}$ contains $\mathcal{F}_o$. As in the previous part this shows that $\mathcal{C} \subset \mathcal{M}$. In other words for every $A \in \mathcal{C}$ and every $B \in \mathcal{C}$ we have $A \cup B \in \mathcal{C}$.

These three steps have shown that $\mathcal{C}$ is a field. The fact that $\mathcal{C}$ is a monotone class and a field allows us to rewrite a countable union of sets in $\mathcal{C}$ as an increasing union of sets in $\mathcal{C}$. This proves $\mathcal{C}$ is a σ-field. •

It remains to apply this lemma to the assertion. Call any set C of the form $A \times B$ where A and B are Borel a rectangle. You have proved that

$$P((X,Y) \in C) = \int_C f_X(x) f_Y(y) dx dy \tag{2.1}$$

for each rectangle C. If C is a finite union of rectangles then we may rewrite C as a finite disjoint union of rectangles to see that (2.1) holds for any such C. Similarly the complement of a rectangle may be rewritten as a finite disjoint union of rectangles. It follows that the collection $\mathcal{M}$ of subsets of $\mathbb{R}^{p+q}$ for which (2.1) holds contains the field formed by the collection of all finite disjoin unions of rectangles. Finally $\mathcal{M}$ is a monotone class by the dominated convergence theorem. Here's precisely how the dominated convergence theorem applies.

The function $g(x,y) = f_X(x)f_Y(y)$ is positive and has integral equal to 1 by Tonelli's theorem. If $C_1 \subset C_2 \subset \cdots$ then put

$$g_n(x,y) = g(x,y)1(x \in C_n)$$

Note that $|g_n(x,y)| \le |g(x,y)|$ and that g_n converges for almost all x to

$$g_\infty(x,y) \equiv g(x,y)1(x \in \cup C_n)$$

Thus

$$\begin{aligned}\lim_{n\to\infty} \int g_n(x,y)dxdy &= \lim_{n\to\infty} \int_{C_n} f_X(x)f_Y(y)dxdy \\ &= \int g_\infty(x,y)dxdy \\ &= \int_{\cup C_n} f_X(x)f_Y(y)dxdy\end{aligned}$$

At the same time

$$\lim P((X,Y) \in C_n) = P((X,Y) \in \cup C_n)$$

so that

$$P((X,Y) \in \cup C_n) = \int_{\cup C_n} f_X(x)f_Y(y)dxdy$$

This and a corresponding argument for intersections show that $\mathcal{M}$ is a monotone class containing the field of all finite unions of Borel rectangles. It thus contains the smallest σ field containing this field. But every open rectangle is in this field so in $\mathcal{M}$. Every open set is a countable union of open rectangles and so in $\mathcal{M}$. Thus $\mathcal{M}$ is a σ field containing the open sets and so must contain the Borel σ field.

4. This is proved via another monotone class argument.

5.

$$\begin{aligned}P(X \in A, Y \in B) &= \int_A \int_B g(x)h(y)dydx \\ &= \int_A g(x)dx \int_B h(y)dy\end{aligned}$$

Take $B = \mathbb{R}^1$ to see that

$$P(X \in A) = c_1 \int_A g(x)dx$$

where $c_1 = \int h(y)dy$. From the definition of density we see that $c_1 g$ is the density of X. Since $\iint f_{X,Y}(xy)dxdy = 1$ we see that $\int g(x)dx \int h(y)dy = 1$ so that $c_1 = 1/\int g(x)dx$. A similar argument works for Y.

Theorem 2 *If $X_1, \ldots, X_p$ are independent and $Y_i = g_i(X_i)$ then $Y_1, \ldots, Y_p$ are independent. Moreover, $(X_1, \ldots, X_q)$ and $(X_{q+1}, \ldots, X_p)$ are independent.*

Conditional probability

Definition: $P(A|B) = P(AB)/P(B)$ provided $P(B) \neq 0$.

Definition: For discrete random variables X and Y the conditional probability mass function of Y given X is

$$\begin{aligned} f_{Y|X}(y|x) &= P(Y = y|X = x) \\ &= f_{X,Y}(x,y)/f_X(x) \\ &= f_{X,Y}(x,y)/\sum_t f_{X,Y}(x,t) \end{aligned}$$

For absolutely continuous X the problem is that $P(X = x) = 0$ for all x so how can we define $P(A|X = x)$ or $f_{Y|X}(y|x)$? The solution is to take a limit

$$P(A|X = x) = \lim_{\delta x \to 0} P(A|x \le X \le x + \delta x)$$

If, for instance, X, Y have joint density $f_{X,Y}$ then with $A = \{Y \le y\}$ we have

$$\begin{aligned} P(A|x \le X \le x + \delta x) &= \frac{P(A \cap \{x \le X \le x + \delta x\})}{P(x \le X \le x + \delta x)} \\ &= \frac{\int_{-\infty}^{y} \int_{x}^{x+\delta x} f_{X,Y}(u,v)\,du\,dv}{\int_{x}^{x+\delta x} f_X(u)\,du} \end{aligned}$$

Divide the top and bottom by δx and let δx tend to 0. The denominator converges to $f_X(x)$ while the numerator converges to

$$\int_{-\infty}^{y} f_{X,Y}(x,v)\,dv$$

So we define the conditional CDF of Y given $X = x$ to be

$$P(Y \le y|X = x) = \frac{\int_{-\infty}^{y} f_{X,Y}(x,v)\,dv}{f_X(x)}$$

Differentiate with respect to y to get the definition of the conditional density of Y given $X = x$ namely

$$f_{Y|X}(y|x) = f_{X,Y}(x,y)/f_X(x)$$

or in words "conditional = joint/marginal".

Chapter 3

Expectation

In undergraduate courses we give two definitions of expected value, one for discrete random variables and one for absolutely continuous random variables. In general, there are random variables which are neither absolutely continuous nor discrete. Here's how probabilists define E in general.

Definition: A random variable X is simple if we can write

$$X(\omega) = \sum_1^n a_i 1(\omega \in A_i)$$

for some constants $a_1, \ldots, a_n$ and events A_i.

Definition: For a simple rv X we define

$$\mathrm{E}(X) = \sum a_i P(A_i)$$

Remark: There are many ways to write a simple random variable as a sum $\sum a_i 1(\omega \in A_i)$. You should prove that if

$$P\left[\sum a_i 1(\omega \in A_i) = \sum b_j 1(\omega \in B_j)\right] = 1$$

then

$$\sum a_i P(A_i) = \sum b_j P(B_j)$$

so that our definition is really a definition.

For positive random variables which are not simple we extend our definition by approximation:

Definition: If $X \geq 0$ then

$$\mathrm{E}(X) = \sup\{\mathrm{E}(Y) : 0 \leq Y \leq X, Y \text{ simple}\}$$

Remark: for $X \geq 0$ we are permitting $+\infty$ as a value of E.

Remark: this definition *redefines* E for positive simple X. You must check that if X is simple, say

$$X = \sum a_i 1_{A_i}$$

and $X \ge 0$ then
$$\sup\{E(Y) : 0 \le Y \le X, Y \text{ simple}\} = \sum a_i P(A_i)$$
so that the definition has not been changed.

Definition: We call X **integrable** if
$$E(|X|) < \infty .$$
In this case we define
$$E(X) = E(\max(X, 0)) - E(\max(-X, 0))$$

Remark: Again you must check that you have not changed the definition of $E(X)$ for either of the previous categories of X.

Facts: E is a linear, monotone, positive operator:

1. **Linear**: $E(aX + bY) = aE(X) + bE(Y)$ provided X and Y are integrable.
2. **Positive**: $P(X \ge 0) = 1$ implies $E(X) \ge 0$.
3. **Monotone**: $P(X \ge Y) = 1$ and X, Y integrable implies $E(X) \ge E(Y)$.

Each of these facts is proved first for simple functions then for positive functions then for general integrable functions. You must follow the steps in the definition of E in any proof of general properties of E.

Major technical theorems:

Monotone Convergence: If $0 \le X_1 \le X_2 \le \cdots$ (almost surely, that is, the probability of this event is 1) and $X = \lim X_n$ (which has to exist almost surely) then
$$E(X) = \lim_{n\to\infty} E(X_n)$$

Dominated Convergence: If $|X_n| \le Y_n$ a.s. and there is a random variable X such that $X_n \to X$ almost surely and a random variable Y such that $Y_n \to Y$ almost surely with $E(Y_n) \to E(Y) < \infty$ then
$$E(X_n) \to E(X)$$
This is often used with all Y_n the same random variable Y.

Fatou's Lemma: If $X_n \ge 0$ a.s. then
$$E(\limsup X_n) \le \limsup E(X_n)$$

Theorem: With this definition of E if X has density $f(x)$ (even in $\mathbb{R}^p$ say) and $Y = g(X)$ then
$$E(Y) = \int g(x) f(x) dx .$$
(This could be a multiple integral.) If X has pmf f then
$$E(Y) = \sum_x g(x) f(x) .$$

This works for instance even if X has a density but Y doesn't.

Definition: The r^{th} moment (about the origin) of a real random variable X is $\mu_r' = \mathrm{E}(X^r)$ (provided it exists). We generally use μ for $\mathrm{E}(X)$. The r^{th} central moment is

$$\mu_r = \mathrm{E}[(X-\mu)^r]$$

We call $\sigma^2 = \mu_2$ the variance.

Definition: For an $\mathbb{R}^p$ valued random vector X we define $\mu_X = \mathrm{E}(X)$ to be the vector whose i^{th} entry is $\mathrm{E}(X_i)$ (provided all entries exist).

Definition: The $(p \times p)$ variance covariance matrix of X is

$$Var(X) = \mathrm{E}\left[(X-\mu)(X-\mu)^t\right]$$

which exists provided each component X_i has a finite second moment. More generally if $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ both have all components with finite second moments then

$$\mathrm{Cov}(X,Y) = \mathrm{E}\left[(X-\mu_X)(Y-\mu_Y)^T\right]$$

We have

$$\mathrm{Cov}(AX+b, CY+d) = A\mathrm{Cov}(X,Y)B^T$$

for general (conforming) matrices A, C and vectors b and d.

Moments and probabilities of rare events are closely connected as will be seen in a number of important probability theorems. Here is one version of Markov's inequality (one case is Chebyshev's inequality):

$$\begin{aligned} P(|X-\mu| \ge t) &= \mathrm{E}[1(|X-\mu| \ge t)] \\ &\le \mathrm{E}\left[\frac{|X-\mu|^r}{t^r} 1(|X-\mu| \ge t)\right] \\ &\le \frac{\mathrm{E}[|X-\mu|^r]}{t^r} \end{aligned}$$

The intuition is that if moments are small then large deviations from average are unlikely.

3.1 Moments and independence

Theorem: If $X_1, \ldots, X_p$ are independent and each X_i is integrable then $X = X_1 \cdots X_p$ is integrable and

$$\mathrm{E}(X_1 \cdots X_p) = \mathrm{E}(X_1) \cdots \mathrm{E}(X_p)$$

Proof: Suppose each X_i is simple: $X_i = \sum_j x_{ij} 1(X_i = x_{ij})$ where the x_{ij} are the possible values of X_i. Then

$$\begin{aligned}
\mathrm{E}(X_1 \cdots X_p) &= \sum_{j_1 \ldots j_p} x_{1j_1} \cdots x_{pj_p} \mathrm{E}(1(X_1 = x_{1j_1}) \cdots 1(X_p = x_{pj_p}) \\
&= \sum_{j_1 \ldots j_p} x_{1j_1} \cdots x_{pj_p} P(X_1 = x_{1j_1} \cdots X_p = x_{pj_p}) \\
&= \sum_{j_1 \ldots j_p} x_{1j_1} \cdots x_{pj_p} P(X_1 = x_{1j_1}) \cdots P(X_p = x_{pj_p}) \\
&= \left[\sum_{j_1} x_{1j_1} P(X_1 = x_{1j_1})\right] \cdots \left[\sum_{j_p} x_{pj_p} P(X_p = x_{pj_p})\right] \\
&= \prod \mathrm{E}(X_i)
\end{aligned}$$

For general $X_i \geq 0$ we create a sequence of simple approximations by rounding X_i down to the nearest multiple of 2^{-n} (to a maximum of n). Call this random variable $X_{i,n}$. Then each $X_{i,n}$ is simple and the variables $X_{1,n}, \ldots, X_{p,n}$ are independent. Thus

$$\mathrm{E}(\prod X_{j,n}) = \prod \mathrm{E}(X_{j,n})$$

for each n. If

$$X_n^* = \prod X_{j,n}$$

then

$$0 \leq X_1^* \leq X_2^* \leq \cdots$$

and X_n^* converges to $X^* = \prod X_i$ so that

$$\mathrm{E}(X^*) = \lim \mathrm{E}(X_n^*)$$

by the monotone convergence theorem. Similarly by the monotone convergence theorem

$$\lim \prod \mathrm{E}(X_{j,n}) = \prod \mathrm{E}(X_j)) < \infty$$

This shows both that X^* is integrable and that

$$\mathrm{E}(\prod X_j) = \prod \mathrm{E}(X_j)$$

The general case uses the fact that we can write each X_i as the difference of its positive and negative parts:

$$X_i = \max(X_i, 0) - \max(-X_i, 0)$$

Just expand out the product and use the previous case.

3.2 Appendix on Lebesgue Integration

Corresponding to Lebesgue measure is the Lebesgue integral which is defined in much the same way as E.

We call a Borel measurable function f simple if

$$f(x) = \sum_1^n a_i 1(x \in B_i)$$

for almost all $x \in \mathbb{R}^p$ and some constants a_i and Borel sets A_i with $\lambda(B_i) < \infty$). For such an f we define

$$\int f(x)dx = \sum a_i \lambda(A_i)$$

Again if

$$\sum a_i 1_{A_i} = \sum b_j 1_{B_j}$$

almost everywhere and all A_i and B_j have finite Lebesgue measure you must check that

$$\sum \lambda(A_i) = \sum b_J \lambda(B_j)$$

If $f \geq 0$ almost everywhere and f is Borel define

$$\int f(x)dx = \sup\{\int g(y)dy\}$$

where the sup ranges over all simple functions g such that $0 \leq g(x) \leq f(x)$ for almost all x. Call $f \geq 0$ integrable if $\int f(x)dx < \infty$.

Call a general f integrable if $|f|$ is integrable and define for integrable f

$$\int f(x)dx = \int \max(f(x), 0)dx - \int \max(-f(x), 0)dx$$

Remark: Again you must check that you have not changed the definition of f for either of the previous categories of f.

Facts: $\int$ is a linear, monotone, positive operator:

1. **Linear**: $\int af(x) + bg(x)dx = a\int f(x)dx + b \inf g(x)dx$ provided f and g are integrable.
2. **Positive**: If $f(x) \geq 0$ almost everywhere then $\int f(x)dx \geq 0$.
3. **Monotone**: If $f(x) > g(x)$ almost everywhere and f and g are integrable then $\int f(x)dx \geq \int g(x)dx$.

Each of these facts is proved first for simple functions then for positive functions then for general integrable functions.

Major technical theorems:

Monotone Convergence: If $0 \leq f_1 \leq f_2 \leq \cdots$ almost everywhere and $f = \lim f_n$ (which has to exist almost everywhere) then

$$\int f(x)dx = \lim_{n\to\infty} f_n(x)dx$$

Dominated Convergence: If $|f_n| \leq g_n$ and there is a Borel function f such that $f_n(x) \to f(x)$ for almost all x and a Borel function g such that $g_n(x) \to g(x)$ with $\int g_n(x)dx \to \int g(x)dx < \infty$ then f is integrable and

$$\int f_n(x)dx \to \int f(x)dx$$

Fatou's Lemma: If $f_n \geq 0$ almost everywhere then

$$\int \limsup f_n(x)dx \leq \limsup \int f_n(x)dx$$

Notice the frequent of almost all or almost everywhere in the hypotheses. In our definition of E wherever we require a property of the function $X(\omega)$ we can require it to hold only for a set of ω whose complement has probability 0. In this case we say the property holds **almost surely**. For instance the dominated convergence theorem is usually written:

Dominated Convergence: If $|X_n| \leq Y_n$ almost surely (often abbreviated to a.s.) and there is a random variable X such that $X_n \to X$ a.s. and a random variable Y such that $Y_n \to Y$ almost surely with $\mathrm{E}(Y_n) \to \mathrm{E}(Y) < \infty$ then

$$\mathrm{E}(X_n) \to \mathrm{E}(X)$$

The hypothesis of almost sure convergence can be weakened; I hope to discuss this later in the course.

Multiple Integration

The previous appendix gave a definition of integrals over R^p in terms of Lebesgue measure on R^p. Students will be used to doing integrals one variable at a time. The fact that the multiple integral is equal to the iterated integral is contained in two theorems.

Theorem 3 (Tonelli) *: If $f : \mathbb{R}^{p+q} \mapsto \mathbb{R}$ is Borel and $f \geq 0$ almost everywhere then for almost every $x \in \mathbb{R}^p$ the integral*

$$g(x) \equiv \int f(x,y)dy$$

exists and

$$\int g(x)dx = \int f(x,y)dxdy$$

The right hand side of this formula denotes the $p+q$ dimensional integral defined in the previous appendix.

Theorem 4 (Fubini) *If $f : \mathbb{R}^{p+q} \mapsto \mathbb{R}$ is Borel and integrable then then for almost every $x \in \mathbb{R}^p$ the integral*

$$g(x) \equiv \int f(x,y)dy$$

exists and is finite. Moreover g is integrable and

$$\int g(x)dx = \int f(x,y)dxdy\,.$$

Chapter 4

The Strong Law of Large Numbers

In this chapter I want to prove the strong law of large numbers. Along the way we will examine some standard approaches to probability theorems:

- Many events have sophisticated mathematical definitions (such as the event that a sequence has a limit, or the event that infinitely many times in an infinite sequence of coin tosses there will be more heads than tails). We will see how to convert the formal mathematical definitions into set notation.
- The resulting set theoretic expression may not be convenient. We use theorems of real analysis to re-express the events in more convenient ways.
- In order to prove $P(A) = 0$ it is sometimes convenient to find (guess or whatever) an event B which contains A and has $P(B) = 0$.
- When dealing with an event A defined in terms of infinitely many random variables it is helpful to write A as an increasing union or decreasing intersection of events each defined from only finitely many random variables.
- In the end inequalities play a central role. We will establish Kolmogorov's inequality – an extension of Chebyshev's inequality from one random variable to the maximum of several.

4.1 Events in Set Notation

The event that a sequence of random variables Y_n converges to 0 is

$$A \equiv \{\omega : \lim_{n\to\infty} Y_n(\omega) = 0\}$$

This event is not yet explicitly written in terms of simple events involving only a finite number of Ys at a time. To make the conversion we must recall the basic definition of limit: A sequence y_n converges to 0 if for every $\epsilon > 0$ there exists an N such that for every $n \geq N$ we have $|y_n| \leq \epsilon$.

Now to convert the definition into set theory notation we replace y_n by $Y_n(\omega)$, each *for every* by an intersection and each *there exists* with a union. We get

$$A = \bigcap_{\epsilon>0} \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} \{\omega : |Y_n(\omega)| \leq \epsilon\}$$

Is this an event? As it sits it is not obvious because the intersection over $\epsilon > 0$ is an uncountable intersection. In fact, however, the intersection is countable. Let

$$B_\epsilon \equiv \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} \{\omega : |Y_n(\omega)| \leq \epsilon\}$$

Notice that if $\epsilon' < \epsilon$ then $B_{\epsilon'} \subset B_\epsilon$. This means that

$$\bigcap_{\epsilon>0} B_\epsilon = \bigcap_{m=1}^{\infty} B_{1/m}$$

This shows that A is a countable intersection of countable unions of countable intersections of events and so A is an event.

Here are some other events:

- The sequence S_n has a limit. Formally, a sequence s_n has a limit if there exists s_∞ such that for every $\epsilon > 0$ there exists an N such that for every $n \geq N$ we have $|s_n - s_\infty| \leq \epsilon$. Mechanically this leads to the event

$$\bigcup_s \bigcap_{\epsilon>0} \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} \{\omega : |S_n(\omega) - s| \leq \epsilon\}$$

  Again the union over ϵ can be made countable. The union over s, however, is not easy to make countable. Instead we use a theorem of analysis to describe the existence of a limit in a different way. It is a theorem that a sequence s_n has a limit if and only if the sequence is Cauchy. Here is the definition of Cauchy: for every $\epsilon > 0$ there exists an N such that for every $n \geq N$ we have $|s_n - s_N| \leq \epsilon$. Our event that S_n has a limit is thus

$$\bigcap_{\epsilon>0} \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} \{|S_n - S_N| \leq \epsilon\}$$

  Again the intersection over all $\epsilon > 0$ can be replaced with a countable intersection over $\epsilon = 1/r$ for positive integers r.

- The sequence Y_n is summable. This means that the series of partial sums $S_n = \sum_1^n Y_i$ has a limit so our event is

$$\bigcap_{r=1}^{\infty} \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} \bigcap_{m=n}^{\infty} \{\left|\sum_{n+1}^{m} Y_j\right| \leq 1/r\}$$

- The sequence S_n is positive for infinitely many n. In other words for every N there is an $n \geq N$ such that $S_n > 0$. The event is

$$\bigcap_{N=1}^{\infty} \bigcup_{n=N}^{\infty} \{S_n > 0\}$$

- The limit superior of S_n is 1. It is easiest to write this as the intersection of two events, namely $\limsup S_n \leq 1$ and $\limsup S_n \geq 1$. The former means that for all $\epsilon > 0$ there is an N such that for all $n \geq N$ $S_n \leq 1 + \epsilon$. The latter means that for all $\epsilon > 0$ and all N there is an $n \geq N$ such that $S_n \geq 1 - \epsilon$. So our event is $A^* \cap A_*$ where

$$A^* = \bigcap_{r=1}^{\infty} \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} \{S_n \leq 1 + 1/r\}$$

and

$$A_* = \bigcap_{r=1}^{\infty} \bigcap_{N=1}^{\infty} \bigcup_{n=N}^{\infty} \{S_n \geq 1 + 1/r\}$$

4.2 The Strong Law of Large Numbers with 4 moments

Theorem 5 *Suppose that $X_n; n \geq 1$ is a sequence of independent identically distributed random variables with* $\mathrm{E}(X_n) = 0$. *Then $n^{-1} \sum_1^n X_k \to 0$ almost surely.*

We have seen in the previous section that we are supposed to prove that $P(A) = 1$ where

$$A = \bigcap_{r=1}^{\infty} \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} B_{n,r}$$

where $B_{n,r}$ is the event

$$\{|(X_1 + \cdots X_n)/n| \leq 1/r\}$$

Let

$$C_r = \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} B_{n,r}$$

and notice that

$$C_1 \supset C_2 \supset \cdots$$

so that

$$P(C_1) \geq P(C_2) \geq \cdots \geq \lim P(C_r) = P(A)$$

It is thus necessary and sufficient to prove that $P(C_r) = 1$ for each r.

Remark: This process of recognizing decreasing or increasing sequences of sets plays a substantial role in many probability arguments. With several intersections and unions in a row it can be hard to figure out whether or not the objects really do increase or decrease. However many definitions which begin with "for every $\epsilon > 0$" will be decreasing intersections because the condition in the definition is harder to satisfy for smaller ϵ. Similarly unions in definitions tend to be increasing.

Now each C_r is an increasing union. Let

$$C_{N,r} = \bigcap_{n=N}^{\infty} B_{n,r}$$

and notice that

$$C_{1,r} \subset C_{2,r} \subset \cdots$$

We need to prove that

$$P(C_r) = \lim_N P(C_{N,r}) = 1$$

or equivalently that

$$\lim_N P(C_{N,r}^c) = 0$$

I will now show you one version of the strong law of large numbers which requires more moment conditions than the real thing but has a relatively direct proof. The proof uses a standard strategy worth knowing. It is based on the **Borel-Cantelli** Lemma.

Suppose A_n is a sequence of events. The event

$$B \equiv \{A_n \text{ infinitely often}\} = \{A_n \text{ i.o.}\}$$

is the set of ω belonging to infinitely many of the A_n We have seen above that

$$B = \bigcap_{N=1}^{\infty} \bigcup_{n=N}^{\infty} A_n$$

Lemma 2 (Borel-Cantelli) *If $\sum P(A_n) < \infty$ then $P(A_n \text{ i.o.}) = 0$.*

To prove the lemma we need to prove that

$$\lim_{N\to\infty} P(\bigcup_{n=N}^{\infty} A_n) = 0$$

But (this next inequality is quite widely useful – it applies to an arbitrary sequence of events A_n)

$$P(\bigcup_{n=N}^{\infty} A_n) \le \sum_{n=N}^{\infty} P(A_n)$$

It is a fact from real analysis that a convergent series has tail sums which converge to 0 so

$$\lim_{N\to\infty} \sum_{n=N}^{\infty} P(A_n) = 0$$

which proves the lemma.

Now to apply the lemma notice that the complement of

$$C_r = \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} B_{n,r}$$

is just

$$C_r^c = \bigcap_{N=1}^\infty \bigcup_{n=N}^\infty B_{n,r}^c$$

In words the opposite of the event that there is an N such that $B_{n,r}$ happens for every $n \ge N$ is that for every N there is at least one $n \ge N$ for which $B_{n,r}$ does not happen. That is

$$C_r^c = \{B_{n,r}^c \text{ i.o.}\}$$

Summary: if we prove

$$\sum_{n=1}^\infty P(B_{n,r}^c) < \infty$$

then we will conclude $P(C_r^c) = 0$ so $P(C_r) = 1$.

Eventually in most probability theorems you reach this stage; we need an upper bound to help us compute the sum. I will use one of the versions of Markov's inequality.

$$\begin{aligned} P(B_{n,r}^c) &= P(|X_1 + \cdots + X_n| > n/r) \\ &\le \frac{\mathrm{E}\left[(X_1 + \cdots + X_n)^4\right]}{(n/r)^4} \end{aligned}$$

The fourth moment is (after expanding everything out and collecting all the different kinds of terms

$$\mathrm{E}\left[(X_1 + \cdots + X_n)^4\right] = n\mathrm{E}(X_1^4) + n(n-1)\left[\mathrm{E}(X_1^2)\right]^2$$

This produces the bound (using $n(n-1) \le n^2$)

$$P(B_{n,r}^c) \le \frac{r^4 \mathrm{E}(X_1^4)}{n^3} + \frac{r^4\left[\mathrm{E}(X_1^2)\right]^2}{n^2}$$

Since

$$\sum_n n^{-p} < \infty$$

for any $p > 1$ the infinite sums converge proving the Strong Law of Large Numbers for random variables for which

$$\mathrm{E}(X_1^4) < \infty \quad \text{and} \quad \mathrm{E}(X_1^2) < \infty$$

In fact the inequality $X_1^2 \le 1 + X_1^4$ proves that $\mathrm{E}(X_1^4) < \infty$ implies $\mathrm{E}(X_1^2) < \infty$. There are many inequalities concerning moments which are worth learning to use: Cauchy Schwarz, Minkowski, Holder, Jensen are the names of some of the most famous.

4.3 Proof without 4 finite moments

In probability (and mathematical statistics) considerable effort is devoted to eliminating the need for moment assumptions. The Strong Law of Large Numbers is an example. I am going to weaken the assumption of 4 finite moments to that of 1 finite moment. I will do this in two stages to illustrate the techniques. First I will suppose that each X_i has a finite variance (which is the same as $\mathrm{E}(X_1^2)$ because the random variables have mean 0).

The first standard tactic of probabilists is to use some real variables result to replace the event whose probability you are computing by another whose probability is the same (or almost the same and then take limits). In this case we use Kronecker's Lemma.

Lemma 3 (Kronecker) *Suppose*

$$0 < b_1 < b_2 \cdots$$

and that x_n is a sequence of real numbers. If

$$s_n = \sum_{i=1}^n \frac{x_i}{b_i}$$

is a convergent sequence then

$$\lim_{n\to\infty} \frac{x_1 + \cdots + x_n}{b_n} = 0$$

I won't prove the lemma but observe this. The lemma implies that if

$$A^* = \{\sum_1^n X_k/k \text{ is a convergent sequence}\}$$

then

$$A^* \subset A$$

so that we need only prove that $P(A^*) = 1$.

We have already seen that the event that a sequence is convergent is the same as the event that it is Cauchy so we must compute the probability of the event

$$\bigcap_r \bigcup_{N=1}^\infty \bigcap_{n=N}^\infty \{|S_n - S_N| \le 1/r\}$$

where $S_n = \sum 1^n X_i/i$. The key ingredient in the calculation is a lower bound on

$$P\left(\bigcap_{n=N}^\infty \{|S_n - S_N| \le \epsilon\}\right)$$

This is the same problem as finding an upper bound on the probability of the complement which is to bound

$$P(\bigcup_{n=N}^\infty \{|S_n - S_N| > \epsilon\}) = \lim_{m\to\infty} P(\bigcup_{n=N}^m \{|S_n - S_N| > \epsilon\})$$

Now $S_n - S_N$ is a sum of $n - N$ independent random variables. Our general need is for a bound on

$$P\left(\bigcup_1^n \{|\sum_1^k Z_i| > \epsilon\}\right)$$

where the Z_i are independent mean 0 random variables. We use Kolmogorov's inequality which is a strengthening of Chebyshev's:

Theorem 6 (Kolmogorov) *Suppose that $Z_1, \ldots, Z_n$ are independent random variables with finite variances and mean 0. Put $S_k = \sum_1^k Z_i$. Then*

$$P(\exists k, 1 \le k \le n : |S_k| > \epsilon) \le \frac{\mathrm{Var}(S_n)}{\epsilon^2}$$

The proof uses the idea of a stopping time. Define T to be the least value of $k \in \{1, \ldots, n\}$ for which $|S_k| > \epsilon$ if such a k exists and $T = n$ if no such k exists. Notice that the event $T = k$ is determined by the variables $Z_1, \ldots, Z_k$; that is, there is a function $f_k(Z_1, \ldots, Z_k)$ such that

$$1(T = k) = f_k(Z_1, \ldots, Z_k)$$

Now

$$\begin{aligned} P(\exists k, 1 \le k \le n : |S_k| > \epsilon) &= P(|S_T| > \epsilon) \\ &\le \frac{\mathrm{E}(S_T^2)}{\epsilon^2} \end{aligned}$$

by Chebyshev's inequality.

Next we compare $\mathrm{E}(S_n^2)$ to $\mathrm{E}(S_T^2)$. We have

$$\begin{aligned} \mathrm{E}(S_n^2) &= \mathrm{E}\left[(S_n - S_T + S_T)^2\right] \\ &= \mathrm{E}(S_T^2) + \mathrm{E}\left[(S_n - S_T)^2\right] + 2\mathrm{E}\left[S_T(S_n - S_T)\right] \end{aligned}$$

The first two terms are non-negative so if

$$\mathrm{E}\left[S_T(S_n - S_T)\right] = 0$$

the theorem will be proved.

But

$$\begin{aligned} S_T(S_n - S_T) &= \sum_1^{n-1} S_k(S_n - S_k) 1(T = k) \\ &= \sum_1^{n-1} S_k(S_n - S_k) f_k(Z_1, \ldots, Z_k) \end{aligned}$$

Notice that $S_n - S_k$ and $S_k f_k(Z_1, \ldots, Z_k)$ are independent so

$$\begin{aligned} \mathrm{E}\left[S_T(S_n - S_T)\right] &= \sum_1^{n-1} \mathrm{E}(S_n - S_k) \mathrm{E}\left[S_k 1(T = k)\right] \\ &= 0 \end{aligned}$$

Remark: In fact all I needed to prove Kolmogorov's inequality was to know that $S_n - S_k$ was uncorrelated with any function of $S_1, \ldots, S_k$ (or equivalently uncorrelated with any function of $Z_1, \ldots, Z_k$).

Definition: A sequence $S_1, S_2, \ldots$ is a **martingale** if, for each $k \geq 1$,

$$\mathrm{E}[S_{k+1}|S_1, \ldots, S_k] = S_k$$

(I have yet to define conditional expectation properly but another definition is that

$$\mathrm{E}[(S_{k+1} - S_k) f_k(S_1, \ldots, S_k)] = 0$$

for every bounded Borel function f_k.)

Theorem 7 (Kolmogorov) *Suppose that $S_1, \ldots, S_n$ is a martingale with finite variances and mean 0. Then*

$$P(\exists k, 1 \leq k \leq n : |S_k| > \epsilon) \leq \frac{\mathrm{Var}(S_n)}{\epsilon^2}.$$

Now back to the strong law. Fix an $\epsilon > 0$ and an integer N. The events

$$B_{m,N} = \bigcap_{n=N}^{m} \{|S_n - S_N| \leq \epsilon\}$$

decrease, that is, $B_{1,N} \supset B_{2,N} \supset \cdots$. So

$$P(B_{\infty,N}) \equiv P(\bigcap_{n=N}^{\infty} \{|S_n - S_N| \leq \epsilon\}) = \lim_m P(B_{m,N})$$

Now

$$\begin{aligned} P(B_{m,N}) &= 1 - P(B^c_{m,N}) \\ &\geq 1 - \frac{\mathrm{Var}(S_m - S_N)}{\epsilon^2} \\ &= 1 - \frac{\mathrm{Var}(X_1) \sum_{N+1}^{m} i^{-2}}{\epsilon^2} \\ &\geq 1 - \frac{\mathrm{Var}(X_1)}{\epsilon^2} \sum_{N+1}^{\infty} \frac{1}{i^2} \end{aligned}$$

This shows

$$P(B_{N,\infty}) \geq 1 - \frac{\mathrm{Var}(X_1)}{\epsilon^2} \sum_{N+1}^{\infty} \frac{1}{i^2}$$

Since $\sum_1^\infty i^{-2} < \infty$ the right hand side of this inequality goes to 0 as N goes to ∞. In other words

$$\lim_{N\to\infty} P(\bigcap_{n=N}^{\infty} \{|S_n - S_N| \leq \epsilon\}) = 1$$

This proves

$$P(\bigcup_{m=1}^{\infty} \bigcap_{n=N}^{\infty} \{|S_n - S_N| \leq \epsilon\}) = 1$$

because the union over m is increasing.

Finally we have shown for each r that $P(C_r) = 1$ where

$$C_r = \bigcup_{m=1}^{\infty} \bigcap_{n=N}^{\infty} \{|S_n - S_N| \le 1/r\}$$

Hence

$$P(\bigcap_r C_r) = 1$$

This establishes the strong law of large numbers for variables with a finite variance.

The Strong Law with 1 Moment

Now suppose that we can't assume that $\mathrm{Var}(X_1) < \infty$. In the arguments of the previous section we used this in Kolmogorov's inequality to try to verify that

$$P(\sum_1^n X_k/k \text{ converges}) = 1$$

The widely used probability trick here is called **truncation**. Let

$$X_k^* = X_k 1(|X_k| \le k)$$

For most values of k, X_k and X_k^* are equal. Note that

$$\begin{aligned} P(X_k \ne X_k^*) &= P(|X_k| > k) \\ &= P(|X_1| > k) \end{aligned}$$

Note that

$$\begin{aligned} \sum_{k=1}^{\infty} P(|X| > k) &= \sum_{k=1}^{\infty} \sum_{j=k}^{\infty} P(j < |X| \le j+1) \\ &= \sum_{j=1}^{\infty} \sum_{k=1}^{j} P(j < |X| \le j+1) \\ &= \sum_{j=1}^{\infty} jP(j < |X| \le j+1) \end{aligned}$$

Notice that

$$\begin{aligned} \mathrm{E}\{|X|1(j < |X| \le j+1)\} &\ge j\mathrm{E}\{1(j < |X| \le j+1)\} \\ &= jP(j < |X| \le j+1) \end{aligned}$$

Hence

$$\begin{aligned} \sum_{k=1}^{\infty} P(|X| > k) &\le \sum_{j=1}^{\infty} \mathrm{E}\{|X|1(j < |X| \le j+1)\} \\ &= \mathrm{E}\{|X|1(1 < |X|)\} \\ &\le \mathrm{E}(|X|) \\ &< \infty\,. \end{aligned}$$

Now apply the Borel Cantelli Lemma to prove that

$$P(X_k \neq X_k^* \text{i.o.}) = 0$$

In other words except for a set N of ω with $P(N) = 0$ $X_k(\omega) = X_k^*(\omega)$ except for a finite number of k. If ω is in

$$A^* = \{\sum_1^n X_k^*/n \to 0\}$$

and not in N then ω is in

$$A = \{\sum_1^n X_k/n \to 0\}$$

Hence if $P(A^*) = 1$ then $P(A) = 1$.

We intend to mimic the proof given before but the following differences arise:

- Let

$$\mu_k = \mathrm{E}(X_k^*)$$

  Although $\mathrm{E}(X_k) = 0$ the truncation may make the mean differ from 0. So Kolmogorov's inequality cannot be applied directly to the X_k^* variables. Instead we have to subtract the mean.

- The variable X_k^* is bounded by k so it does have a finite variance. But that variance depends on k so the sum we got above, of $\sum 1/i^2$ will be more complicated.

To deal with the first point we begin with an observation. If Y_n are any random variables and a_n a sequence of constants converging to 0 then the event $\{Y_n \to 0\}$ is almost surely the same as $\{Y_n + a_n \to 0\}$. For clarity the jargon is that two events are almost surely equal if their indicators are almost surely equal. Another way of saying this is

$$P\left(\{Y_n \to 0\}\Delta\{Y_n + a_n \to 0\}\right) = 0$$

where

$$A\Delta B = (AB^c) \cup (A^cB) = (A \cup B) \setminus (A \cap B)$$

is the symmetric difference of two sets.

So we will prove:

Claim 1: $\sum_1^n \mu_k/n$ converges to 0

and

Claim 2:

$$P\left\{\sum_1^n (X_k^* - \mu_k)/n \to 0\right\} = 1$$

For Claim 1 note

$$\begin{aligned}\mu_k &= \mathrm{E}(X_k^*) \\ &= \mathrm{E}(X_k 1(|X_k| \le k)) \\ &= \mathrm{E}\left[X_1 1(|X_1| \le k)\right]\end{aligned}$$

By the dominated convergence theorem

$$\mu_k \to \mathrm{E}(X_1) = 0\,.$$

Finally if $a_n \to 0$ then $(a_1 + \cdots + a_n)/n \to 0$. Hence

$$\sum_1^n \mu_k \to 0$$

which is Claim 1.

Turning to Claim 2 we define

$$A_3^* = \{\sum_1^n (X_k^* - \mu_k)/k \text{ converges}\}$$

According to Kronecker's Lemma Claim 2 (and so the strong law) will follow if we prove $P(A_3^*) = 1$.

Let $\sigma_k^2 = \mathrm{Var}(X_k^*)$ and

$$B_{m,N} = \bigcap_{n=N}^m \{|S_n^* - S_N^*| \le \epsilon\}$$

where $S_n^* = \sum_1^n (X_k^* - \mu_k)/k$. As in the finite variance case:

$$\begin{aligned} P(B_{m,N}) &= 1 - P(B_{m,N}^c) \\ &\ge 1 - \frac{\mathrm{Var}(S_m - S_N)}{\epsilon^2} \\ &= 1 - \frac{\sum_{N+1}^m \sigma_i^2/i^2}{\epsilon^2} \end{aligned}$$

As before we want to let $m \to \infty$ but need to know that

$$\sum_1^\infty \sigma_i^2/i^2 < \infty$$

But

$$\begin{aligned} \sigma_k^2 &\le \mathrm{E}(X_k^2 1(|X_k| \le k)) \\ &= \mathrm{E}(X_1^2 1(|X_1| \le k)) \\ &= \sum_{j=1}^k \mathrm{E}(X_1^2 1(j-1 < |X_1| \le j)) \end{aligned}$$

Sum over k and switch the order of summation to get

$$\begin{aligned}
\sum_1^n \sigma_k^2/k^2 &\le \sum_{j=1}^n \sum_{k=j}^n \mathrm{E}(X_1^2 1(j-1 < |X_1| \le j))/k^2 \\
&\le \sum_{j=1}^n \mathrm{E}(X_1^2 1(j-1 < |X_1| \le j)) \sum_{k=j}^\infty 1/k^2 \\
&\le \sum_{j=1}^n \mathrm{E}(X_1^2 1(j-1 < |X_1| \le j))/j \\
&\le \sum_{j=1}^n \mathrm{E}(|X_1| 1(j-1 < |X_1| \le j)) \\
&= \mathrm{E}(|X_1| 1(|X_1| \le n)) \\
&\le \mathrm{E}(|X_1|)
\end{aligned}$$

This shows

$$P(B_{N,\infty}) \ge 1 - \frac{1}{\epsilon^2} \sum_{N+1}^\infty \sigma_k^2/k^2$$

Since $\sum_1^\infty \sigma_k^2/k^2 < \infty$ the right hand side of this inequality goes to 0 as N goes to ∞. In other words

$$\lim_{N\to\infty} P(\bigcap_{n=N}^\infty \{|S_n^* - S_N^*| \le \epsilon\}) = 1$$

Thus S_n^* is almost surely Cauchy and $P(A_3^*) = 1$.

4.4 Consistency of MLE

Suppose that $X_1, X_2, \ldots$ are independent and identically distributed with density $f(x, \theta_o)$ where

$$\{f(\cdot, \theta); \theta \in \Theta \subset \mathbb{R}\}$$

is a family of densities. In this section we investigate conditions under which the MLE of θ is almost surely consistent. Our focus is on techniques of making the assertion precise in probability theory terms.

Our goal is to find conditions under which we can prove

$$P(\hat\theta_n \to \theta_o) = 1$$

where $\hat\theta_n$ is the mle.

We face the following general technical problems:

- What is the precise definition of $\hat\theta_n$?

- Having settled on some definition is the resulting object a random variable?

We will focus on an example. The Cauchy(θ) density is

$$f(x,\theta) = \frac{1}{\pi\left\{1+(x-\theta)^2\right\}}$$

For a sample $X_1, \ldots, X_n$ the likelihood is

$$\frac{1}{\pi^n \prod_1^n \{1+(X_i-\theta)^2\}}$$

Normally we say that $\hat\theta$ is the value of θ which maximizes this function of θ.

To be precise this is supposed to define $\hat\theta$ as a function of $X_1, \ldots, X_n$. In other words it is supposed that for each $x_1, \ldots, x_n$ there is a unique value $\hat\theta(x_1, \ldots, x_n)$ which maximizes the likelihood. If this were so we would have a definition of a function from $\mathbb{R}^n$ to $\mathbb{R}$.

In order to discuss this problem in general it is convenient to define the log-likelihood:

$$\ell(\theta|x-1, \ldots, x_n) = -n\log(\pi) - \sum_1^n \log(1+(x_i-\theta)^2)$$

It is all very well to say that $\hat\theta_n(x_1, \ldots, x_n)$ maximizes this function of θ for each fixed value of $(x_1, \ldots, x_n)$. But:

1. Is there, for every $(x_1, \ldots, x_n)$ a θ which maximizes ℓ?

2. If so is the θ unique?

3. If so is $\hat\theta_n(x_1, \ldots, xn)$ a Borel function of $x_1, \ldots, x_1$?

Question 1: For the Cauchy density there is always a maximizer. Fix $(x_1, \ldots, x_n)$. As $\theta \to \pm\infty$ it is easy to check that

$$\ell(\theta|x_1, \ldots, x_n) \to -\infty$$

There is then a M such that $|\theta| > M$ implies

$$\begin{aligned}\sup\{\ell(\theta|x_1, \ldots, x_n); |\theta| > M\} & \\ &\le \ell(0|x_1, \ldots, x_n) \\ &\le \sup\{\ell(\theta|x_1, \ldots, x_n); |\theta| \le M\}\end{aligned}$$

Now the function

$$\theta \mapsto \ell(\theta|x_1, \ldots, x_n)$$

is continuous so that it assumes its maximum over $[-M, M]$. This shows the existence of at least one maximizing θ for any set of x values.

Question 2: Consider $n = 2$ and $x_1 = x = -x_2$. Then

$$\ell(\theta|x, -x)$$

is an even function of x. This function has derivative

$$\ell' = \frac{2(\theta - x)}{1 + (x-\theta)^2} + \frac{2(\theta + x)}{1 + (x+\theta)^2}$$

At $\theta = 0$ this evaluates to 0 so that $\theta = 0$ is a critical point of ℓ. The second derivative may be computed and evaluated at 0 to give

$$\ell''(0) = \frac{4(x^2-1)}{(1+x^2)^2}$$

If $|x| < 1$ this is negative so that 0 is a local maximum but if $|x| > 1$ it is a local minimum. In this case, since ℓ is even there must be two maxima n either side of 0. Note that putting the two terms in ℓ' on a common denominator will give a numerator which is a multiple of

$$\theta(\theta^2 - (x^2 - 1))$$

Notice that there are exactly three roots if $x^2 > 1$.

Summary: defining $\hat\theta$ to be the maximizer of ℓ does not actually define a function.

Alternative strategies:

1: You might pick one of the maximizing θ values in an unequivocal way:

$$\hat\theta = \inf\{\theta : \ell(\theta) = \sup \ell\}$$

(The set of such θ is not empty and bounded so there is such a $\hat\theta$ and that $\hat\theta$ is finite. By continuity of ℓ

$$\ell(\hat\theta) = \sup \ell$$

2: You might try defining $\hat\theta$ to be a suitably chosen critical point of ℓ.

3: You might try to prove that

$$P(\text{card}(\{theta : \ell(\theta|X_1, \ldots, X_n) = \sup_\phi \ell(\phi|X_1, \ldots, X_n)\}) = 1) = 1$$

In other words it might be true that the set of θ where ℓ achieves its maximum is almost surely a singleton when the xs are actually a data set.

I am going to follow method 2 since this is the one which works most generally.

For a vector $(x_1, \ldots, x_n)$ we define the order statistics

$$x_{(1)} \le \cdots \le x_{(n)}$$

be the entries in the vector sorted into non-decreasing order. If $n = 2m - 1$ is odd then set

$$g_n(x_1, \ldots, x_n) = x_{(m)}$$

If $n = 2m$ set

$$g_n(x_1, \ldots, x_n) = (x_{(m)} + x_{(m+1)})/2.$$

Now define

$$\tilde\theta_n = g(X_1, \ldots, X_n)$$

Lemma 4 *If $X_1, \ldots, X_n$ are iid from a distribution F with the properties:*

1. *$F(0) = 1/2$.*
2. *For each $\epsilon > 0$*
$$F(-\epsilon) < 1/2 < F(\epsilon)$$

Then $\tilde{\theta}_n$ converges almost surely to 0.

Remark: Part of the theorem is that

$$A \equiv \{\omega : \tilde{\theta}_n \to 0\}$$

is an event. Here is how we work our way back to primitive notions to prove this. Our formal description, in set notation, of this event shows that A is an event if each $\tilde{\theta}_n$ is a random variable. In turn, since the X_i are random variables, we need only check that g_n is Borel. This reduces to the assertion that

$$(x_1, \ldots, x_n) \mapsto x_{(k)}$$

is Borel for each k and n. But

$$\{(x_1, \ldots, x_n) : x_{(k)} < t\} = \{(x_1, \ldots, x_n) : \sum_1^n 1(x_i < t) \geq k\}$$

Since a sum of Borel functions is Borel we are reduced to proving that the map

$$(x_1, \ldots, x_n) \mapsto 1(x_i < t)$$

is Borel for each t and i. This is equivalent to

$$\{(x_1, \ldots, x_n) : x_i < t\}$$

is Borel. Since this last set is actually open it is Borel.

Now to prove the lemma we begin by formalizing an argument we have used several times.

Lemma 5 *Suppose Y_n is a sequence of random variables. Then $Y_n \to 0$ almost surely is equivalent to $P(C_\epsilon) = 1$ for each $\epsilon > 0$ where*

$$C_\epsilon = \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} \{|Y_n| \leq \epsilon\}$$

So fix $\epsilon > 0$. For each x the random variables $Y_1, Y_2, \ldots$ defined by $Y_k = 1(X_k \leq x) - F(x)$ are independent and identically distributed with mean 0. According to the SLLN there is a null set N_x such that for all $\omega \notin N_x$ we have

$$\frac{1}{n} \sum_1^n Y_k \to 0\,.$$

Let $N = N_\epsilon \cup N_{-\epsilon}$. Then N is a null set. If $\omega \notin N$ then

$$\frac{1}{n}\sum_1^n 1(X_k \le \epsilon) \to F(\epsilon) > 1/2$$

and

$$\frac{1}{n}\sum_1^n 1(X_k \le -\epsilon) \to F(-\epsilon) < 1/2$$

For any such ω there is an M such that for all $n \ge M$ the number of X_i exceeding ϵ is less than $n/2$ and the number of X_i less than $-\epsilon$ is less than $n/2$. Thus for such ω, and $n \ge M$

$$-\epsilon \le \tilde\theta_n \le \epsilon$$

In other words the set $C_\epsilon^c \subset N$ so $P(C_\epsilon) = 1$. •

Now back to the Cauchy problem. For a vector $(x_1, \ldots, x_n)$ we define $h_n(x_1, \ldots, x_n)$ to be that root of

$$\ell'(\theta|x_1, \ldots, x_n) = \sum \frac{2(x_i - \theta)}{1 + (x_i - \theta)^2} = 0$$

which is closest to $g_n(x_1, \ldots, x_n)$ (here g_n is the Borel function used in defining the median above). If $g_n(x_1, \ldots, x_n)$ is exactly midway between two roots which are tied for closest we define h_n to be the root smaller than g_n.

It is possible to prove that this defines a Borel function from $\mathbb{R}^n$ to $\mathbb{R}$. Now define

$$\hat\theta_n = h_n(X_1, \ldots, X_n)$$

I claim that if $X_1, X_2, \ldots$ are iid Cauchy(0) then

$$\hat\theta_n \to 0 \tag{4.1}$$

almost surely.

To prove this fix $\epsilon > 0$ we prove that $P(C_\epsilon) = 1$ where

$$C_\epsilon = \bigcup_{N=1}^\infty \bigcap_{n=N}^\infty \{|\hat\theta_n| \le \epsilon\}$$

In the discussion which follows we suppress dependence of ℓ on the data whenever possible. We use superscript $'$ to denote differentiation with respect to θ. We will have use of the notation:

$$\begin{aligned}
L_i(\theta) &= -\log(f(X_i, \theta)) + \log(f(X_i, \theta_o)) \\
U_i(\theta) &= L_i'(\theta) \\
V_i(\theta(&= L_i''(\theta) \\
W_i(\theta) &= L_i'''(\theta)
\end{aligned}$$

Here is how we proceed. We find an event D_ϵ inside C_ϵ with $P(D_\epsilon) = 1$. To define this new event note that if

1. ℓ' has a unique root over $[-3\epsilon, 3\epsilon]$ and

2. that root is actually in $[-\epsilon, \epsilon]$ and

3. $|\tilde{\theta}_n| \le \epsilon$

then the root of ℓ' closest to $\tilde{\theta}_n$ is actually the root in points 1 and 2 and so

$$|\hat{\theta}_n| \le \epsilon$$

Define $D_\epsilon^{(1)}$ to be the event that there is an N such that for all $n \ge N$ and all $|\theta| \le 3\epsilon$ we have

$$\ell''(\theta|X_1, \ldots, X_n) < 0$$

Define $D_\epsilon^{(2)}$ to be the event that there is an N such that for all $n \ge N$

$$\{\ell'(\epsilon|X_1, \ldots, X_n) < 0$$

and

$$\{\ell'(-\epsilon|X_1, \ldots, X_n) > 0$$

Finally define $D_\epsilon^{(3)}$ to be the event that there is an N such that for all $n \ge N$

$$|\tilde{\theta}_n| \le \epsilon$$

We have already shown that $P(D_\epsilon^{(3)}) = 1$. Next I show that $P(D_\epsilon^{(2)}) = 1$. For the Cauchy case

$$U_k(\epsilon) = \frac{2(X_k - \epsilon)}{1 + (X_k - \epsilon)^2}$$

and

$$U_k(-\epsilon) = \frac{2(X_k + \epsilon)}{1 + (X_k + \epsilon)^2}$$

Then

$$\frac{1}{n}\ell''(\epsilon|X_1, \ldots, X_n) = \overline{U_n(\epsilon)}$$

and

$$\frac{1}{n}\ell''(-\epsilon|X_1, \ldots, X_n) = \overline{U_n(-\epsilon)}$$

Thus $D_\epsilon^{(2)}$ is the event that there is an N such that for all $n \ge N$

$$\overline{U_n(\epsilon)} < 0 \quad \text{and} \quad \overline{U_n(-\epsilon)} > 0$$

Since each of $\overline{U_n(-\epsilon)}$ and $\overline{U_n(\epsilon)}$ is an average of iid variates it suffices to show that

$$\mathrm{E}(U_k(\epsilon)) < 0$$
$$\mathrm{E}(U_k(-\epsilon)) > 0\,.$$

In fact

$$\mathrm{E}(U_k(\epsilon)) = \frac{-2\pi\epsilon)}{\epsilon^2+4} < 0$$
$$\mathrm{E}(V_k(-\epsilon)) = \frac{2\pi\epsilon)}{\epsilon^2+4} > 0\,.$$

This argument is not very easy to generalize since it hinged on an exact computation of a moment. A much more general tactic uses Jensen's inequality.

If ℓ is smaller at $-\epsilon$ and at ϵ than it is at 0 then there must be a critical point in $[-\epsilon,\epsilon]$, that is, a root of ℓ'. To deal with this recall the definition

$$L_i(\theta) = \log(1+X_i^2) - \log(1+(X_i-\theta)^2)\,.$$

Define $D_\epsilon^{(4)}$ be the event that there is an N such that for all $n \ge N$

$$\{\sum L_i(\epsilon) < 0, \sum L_i(-\epsilon) < 0\}\,.$$

Define

$$\mu(\epsilon) = \mathrm{E}(L_i(\epsilon))$$

The strong law of large numbers shows that $P(D_\epsilon^{(5)}) = 1$ where

$$D_\epsilon^{(5)} = \{\sum L_i(\epsilon)/n \to \mu(\epsilon), \sum L_i(-\epsilon)/n \to \mu(-\epsilon)\}$$

I claim that for all $\epsilon \ne 0$ $\mu(\epsilon) < 0$. If so then

$$D_\epsilon^{(5)} \subset D_\epsilon^{(4)}$$

and so $P(D_\epsilon^{(4)}) = 1$.

To prove the claim we apply Jensen's inequality:

Proposition 1 (Jensen) *Suppose Y is a random variable and ϕ is a function which is convex on an interval (a,b) with $P(a<Y<b)=1$ Assume $\mathrm{E}(|Y|)<\infty$. Then*

$$\phi(\mathrm{E}(Y)) \le \mathrm{E}(phi(Y))$$

If ϕ is strictly convex then the inequality is strict unless $\mathrm{Var}(Y) = 0$.

Jargon: ϕ is convex if for each x,y and $\lambda \in (0,1)$

$$\phi(\lambda x + (1-\lambda)y) \le \lambda\phi(x) + (1-\lambda)\phi(y)$$

We call ϕ strictly convex if the inequality is strict.

If ϕ is twice differentiable and $\phi'' \ge 0$ then ϕ is convex; a strict inequality shows ϕ is strictly convex. We apply Jensen's inequality with $\phi(x) = -\log(x)$. We apply it to $Y = g(X)/f(X)$ where X has density f and g is some other density. Then

$$\mathrm{E}\{-\log(Y)\} > -\log\{\mathrm{E}(Y)\}$$

But the latter is

$$\log\left\{\int \frac{g(x)}{f(x)} f(x)dx\right\} = \log(1) = 0$$

Technically we ought to be careful. The interval (a, b) in the inequality is $(0, \infty)$. The assumption

$$P(0 < Y < \infty) = 1$$

deserves some discussion. If $f(x) = 0$ for some places where $g(x)$ is not 0 then

$$\mathrm{E}\left\{\frac{g(X)}{f(X)}\right\} = \int g(x)1(f(X) > 0)dx$$

which might be less than 1. This just makes the inequality stronger, however.

The other technical detail is that $g(x)$ might be 0 some places where $f(x)$ is not 0. This might mean $P(Y = 0) > 0$. On the event $Y = 0$ we will agree to take $-\log(Y) = \infty$ and conclude

$$\mathrm{E}\{-log(Y)\} = \infty$$

In any case we find

$$\mathrm{E}\{-log(Y)\} > 0$$

or

$$\mathrm{E}[\log\{g(X)\} - \log\{f(X)\}] < 0\,.$$

Applied to our Cauchy problem we have shown $\mu(\theta) < 0$ for all $\theta \neq 0$. Hence $P(D_\epsilon^{(4)}) = 1$.

Finally we consider $D_\epsilon^{(1)}$. Up to now we have been able to make do with an arbitrary ϵ. In this case, however, the result holds only for small $\epsilon > 0$. First consider

$$\frac{1}{n}\ell''(0|X_1, \ldots, X_n)$$

According to the strong law of large numbers this converges almost surely to

$$\mathrm{E}\left\{\frac{2(1-X^2)}{(1+X^2)^2}\right\} = -\frac{1}{2} < 0$$

Now you can check that

$$\left|\frac{1}{n}\ell'''(\theta|X_1, \ldots, X_n)\right| < 4\,.$$

(In fact each term in ℓ''' may be shown to be bounded by $3/2 + \sqrt{2}$.) As a result

$$\frac{1}{n}\left|\ell''(\theta|X_1, \ldots, X_n) - \ell''(0|X_1, \ldots, X_n)\right| \leq 4|\theta|$$

Pick $\epsilon > 0$ so that $4\epsilon < \pi/2$. If B is the event

$$\frac{1}{n}\ell''(0|X_1, \ldots, X_n) \to -\frac{1}{2}$$

and ω is in B then there is an N such that for $n \geq N$ we have

$$\frac{1}{n}\ell''(\theta|X_1, \ldots, X_n) < 0$$

for all $|\theta| < \epsilon$.

This proves that for all $3\epsilon < \pi/8$

$$P(D^{(1)}(\epsilon)) = 1\,.$$

We have now shown that for $\epsilon < \pi/24$

$$P(D^{(1)}(\epsilon) \cap D^{(2)}_\epsilon \cap D^{(3)}_\epsilon) = 1$$

For ω in this event we have that there is an N such that

$$|\hat{\theta}_n| \leq \epsilon$$

for all $n \geq N$. This establishes the result.

General Case

Now consider a parametric family

$$\{f(x|\theta); a < \theta < b\}$$

Let θ_o be the true value of θ, that is, the value used in making probability calculations. Let A_ϵ be the event that there is an N such that for all $n \geq N$ the log-likelihood has a local maximum on the interval $[\theta_o - \epsilon, \theta_o + \epsilon]$. We have proved quite generally that

$$P(A_\epsilon) = 1 \tag{4.2}$$

If we add the assumption

$$\ell hasacontinuousderivative \tag{\textbf{A}}$$

then letting B_ϵ be the event that there is an N such that for all $n \geq N$ there is a critical point of ℓ in $(\theta_o - \epsilon, \theta_o + \epsilon)$ which is a local maximum of ℓ we have proved

$$P(B_\epsilon) = 1 \tag{4.3}$$

The event $B = \cap_\epsilon B_\epsilon$ then has probability 1. On this event there is a sequence of roots of the likelihood equations which is consistent.

The remaining question is whether or not we can prove, under reasonably general conditions, that there is probably only one root near θ_o.

Now consider the event that ℓ' is monotone on $[\theta_o - \epsilon, \theta_o + \epsilon]$. Our previous proof was based on showing that the next derivative was negative at θ_o and did not change much over a small enough interval.

Behaviour at θ_o is essentially the behaviour of

$$\frac{1}{n}\sum V_i(\theta_o)$$

which converges almost surely to

$$\mathrm{E}(L_1''(\theta_o))$$

I claim this is negative for regular families.

Begin with

$$1 = \int f(x,\theta)dx$$

Differentiating with respect to θ gives

$$\begin{aligned} 0 &= \frac{d}{d\theta}\int f(x,\theta)dx \\ &= \lim_{\epsilon\to 0}\int \frac{f(x,\theta+\epsilon)-f(x,\theta)}{\epsilon}dx \end{aligned}$$

In order to take the limit inside the integral sign we must prove that for any sequence $\epsilon_n \to 0$

$$\lim \int \frac{f(x,\theta+\epsilon_n)-f(x,\theta)}{\epsilon_n}dx$$

$$= \int \frac{\partial}{\partial\theta} f(x,\theta)dx$$

This is normally done by applying the dominated convergence theorem. If f is continuously differentiable with respect to θ then the difference quotient is exactly equal to

$$\frac{\partial}{\partial\theta} f(x,\theta_n^*)$$

The quantity θ_n^* depends on both n and x. One tactic is to try to compute

$$M(x,\epsilon) = \sup\{|\frac{\partial}{\partial\theta} f(x,\theta)|; |theta-\theta_o| \le \epsilon\}$$

and show that

$$\int M(x,\epsilon)dx < \infty$$

which would permit application of dominated convergence.

Assuming that we can apply the dominated convergence theorem we get

$$\begin{aligned} 0 &= \int \frac{\partial}{\partial\theta} f(x,\theta)dx \\ &= \int \frac{\partial \log f(x,\theta)}{\partial\theta} f(x,\theta)dx \\ &= \mathrm{E}(U_k(\theta_o)) \end{aligned}$$

Differentiating again and again passing limits through integrals gives

$$\mathrm{E}(U_k^2(\theta_o)) = -\mathrm{E}(V_k(\theta_o))$$

This shows that

$$\frac{1}{n}\ell''(\theta_o) \to \mathrm{E}(V_k(\theta_o)) < 0$$

almost surely.

Next we consider

$$\frac{1}{n}\ell''(\theta) - \ell''(\theta_o)$$

For a three times continuously differentiable ℓ there is a θ_n^* (which is random but between θ_o and θ) such that this difference is

$$\frac{\theta - \theta_o}{n} \sum W_i(\theta_n^*)$$

Define

$$M_i(\epsilon) = \sup\{|W_i(\theta)| : |\theta - \theta_o| \le \epsilon\}$$

The M_i are iid. If for some $\epsilon > 0$ the M_i are integrable then the strong law of large numbers shows that

$$\limsup \left| \frac{1}{n} \{\ell''(\theta) - \ell''(\theta_o)\} \right| \le \epsilon \mathrm{E}(M_1(\epsilon))$$

almost surely. The right hand side of this inequality can be made arbitrarily small by choosing ϵ small enough. Pick ϵ so small that the bound is strictly smaller than

$$I(\theta_o) \equiv -\mathrm{E}(V_k(\theta_o))$$

we have the result that

$$P(E_\epsilon) = 1$$

where E_ϵ is the event that there is an N such that for all $n \ge N$ we have ℓ' is monotone decreasing on $[\theta_o - \epsilon, \theta_o + \epsilon]$.

Chapter 5

Markov Chains

Definition: A stochastic process is a family $\{X_i; i \in I\}$ of random variables where I is a set, called the index set. Often I is a subset of $\mathbb{R}$ such as $[0, \infty)$, $[0, 1]$ $\mathbb{Z}$ or $\mathbb{N}$. When I is an interval we speak of a continuous time stochastic process; often we use the letter t as a typical element of the index set. When $I \subset \mathbb{Z}$ we call X a discrete time stochastic process.

Generally we expect all the X_n to take values in the same **state space** S. In this chapter we will take S to be a finite or countable set so that each X_n is discrete. Usually we identify S with one of $\mathbb{Z}$, $\mathbb{N}$ or $\{0, \ldots, m\}$ for some finite m.

Definition: A Markov Chain is a stochastic process $X_n; n \in \mathbb{N}$. taking values in a finite or countable set S such that for every n and every event of the form

$$A = \{(X_0, \ldots, X_{n-1}) \in B \subset S^{n+1}\}$$

we have

$$P(X_{n+1} = j | X_n = i, A) = P(X_1 = j | X_0 = i) \tag{5.1}$$

We then adopt the notation that $\mathbf{P}$ is the (possibly infinite) array with elements

$$P_{ij} = P(X_1 = j | X_0 = i)$$

indexed by $i, j \in S$.

Conditional Probability: An Aside

If $P(A) > 0$ we define

$$P(B|A) = \frac{P(AB)}{P(A)}$$

In this chapter all random variables are discrete and so when we condition on values of a finite list of variables we will be conditioning on an event of positive probability. If we try to move to continuous state spaces we will have to improve our understanding of conditional probability and expectation.

WARNING: in (5.1) we require the condition to hold **only** when

$$P(X_n = i, A) > 0$$

End of Aside

Definition: $\mathbf{P}$ is the (one step) **transition matrix** of the Markov Chain.

Evidently the entries in $\mathbf{P}$ are non-negative and

$$\sum_j P_{ij} = 1 \tag{5.2}$$

for all $i \in S$. Any such matrix is called **stochastic**.

We define powers of $\mathbf{P}$ by

$$(\mathbf{P}^n)_{ij} = \sum_k \left(\mathbf{P}^{n-1}\right)_{ik} P_{kj}$$

Notice that even if S is infinite these sums converge absolutely.

5.1 Chapman-Kolmogorov Equations

A central tactic in Markov Chain calculations is conditioning on intermediate steps to compute probabilities of events. We do that now to compute

$$P(X_{l+n} = j | X_l = i)$$

We will condition on X_{l+n-1}:

$$\begin{aligned}
P(X_{l+n} = j | X_l = i) &= \sum_k P(X_{l+n} = j, X_{l+n-1} = k | X_l = i) \\
&= \sum_k P(X_{l+n} = j | X_{l+n-1} = k, X_l = i) P(X_{l+n-1} = k | X_l = i) \\
&= \sum_k P(X_1 = j | X_0 = k) P(X_{l+n-1} = k | X_l = i) \\
&= \sum_k P(X_{l+n-1} = k | X_l = i) \mathbf{P}_{ik}
\end{aligned}$$

Now condition on X_{l+n-2} to get

$$P(X_{l+n} = j | X_l = i) = \sum_{k_1 k_2} \mathbf{P}_{k_1 k_2} \mathbf{P}_{k_2 j} P(X_{l+n-2} = k_1 | X_l = i)$$

Notice that the sum over k_2 computes the K_1, j entry in the matrix $\mathbf{PP}$ which we denote $\mathbf{P}^2$. In other words

$$P(X_{l+n} = j | X_l = i) = \sum_{k_1} (\mathbf{P}^2)_{k_1, j} P(X_{l+n-2} = k_1 | X_l = i)$$

We may now prove by induction on n that

$$P(X_{l+n} = j | X_l = i) = (\mathbf{P}^n)_{ij} \,.$$

Finally we have now proved the Chapman-Kolmogorov equations:

$$P(X_{l+m+n} = j|X_l = i) = \sum_k P(X_{l+m} = k|X_l = i)P(X_{l+m+n} = j|X_{l+m} = k)$$

These are simply a restatement of the identity

$$\mathbf{P}^{n+m} = \mathbf{P}^n\mathbf{P}^m .$$

Remark: It is important to notice that these probabilities depend on m and n but **not** on l. We say the chain has **stationary** transition probabilities. A more general definition of Markov chain than (5.1) is

$$P(X_{n+1} = j|X_n = i, A) = P(X_{n+1} = j|X_n = i) .$$

Notice that the right hand side of this formula is now permitted to depend on n. If we define $\mathbf{P}^{n,m}$ to be the matrix with i, jth entry

$$P(X_m = j|X_n = i)$$

for $m > n$ then we can prove

$$\mathbf{P}^{r,s}\mathbf{P}^{s,t} = \mathbf{P}^{r,t}$$

These equations are also called Chapman-Kolmogorov equations. This chain does not have stationary transitions.

Remark: The calculations above involve sums in which all terms are positive. They therefore apply even if the state space S is countably infinite.

5.1.1 Extensions of the Markov Property

Consider a function $f(x_0, x_1, \dots)$ defined on the set S^∞ of all infinite sequences of points in S. (This family of functions includes, in a natural way, all functions which depend on only finitely many of the co-ordinates.) Let B_n be the event

$$f(X_n, X_{n+1}, \dots) \in C$$

for some suitable C in the range space of f. Then

$$P(B_n|X_n = x, A) = P(B_0|X_0 = x) \tag{5.3}$$

for any event A of the form

$$\{(X_0, \dots, X_{n-1}) \in D\}$$

Also

$$P(AB_n|X_n = x) = P(A|X_n = x)P(B_n|X_n = x) \tag{5.4}$$

The latter is usually expressed as "given the present the past and future are conditionally independent."

To prove the first of these assertions you use monotone class arguments. The general result can be deduced from the special case where

$$B_n = \{(X_{n+1} = x_1, \cdots, X_{n+m} = x_m\}$$

For this B_n the left hand side of (5.3) may be evaluated by repeated conditioning as in Chapman-Kolmogorov to get

$$\mathbf{P}_{x,x_1}\mathbf{P}_{x_1,x_2}\cdots\mathbf{P}_{x_{m-1},x_m}$$

and so may the right. For general events defined only from $X_n, \ldots, X_{n+m}$ you sum over appropriate vectors $x, x_1, \ldots, x_m$. The general case requires a measure theory argument via montone class techniques.

To prove (5.4) write

$$\begin{aligned} P(AB_n|X_n = x) &= P(B_n|X_n = x, A)P(A|X_n = x) \\ &= P(B_n|X_n = x)P(A|X_n = x) \end{aligned}$$

using (5.3).

5.2 Classification of States

If an entry $\mathbf{P}_{ij}$ is 0 it is not possible to go from state i to state j in one step. It may be possible to make the transition in some larger number of steps, however. We say i **leads to** j (or j is accessible from i) if there is an integer $n \ge 0$ such that

$$P(X_n = j|X_0 = i) > 0\,.$$

We use the notation $i \rightsquigarrow j$. By defining $\mathbf{P}^0$ to be the identity matrix $\mathbf{I}$ we arrive at $i \rightsquigarrow j$ if there is an $n \ge 0$ for which $(\mathbf{P}^n)_{ij} > 0$.

We say states i and j **communicate** if $i \rightsquigarrow j$ and $j \rightsquigarrow i$. We write $i \leftrightarrow j$ if i and j communicate. Communication is an equivalence relation, that is, a reflexive, symmetric and transitive relation on the states of S. More precisely:

Reflexive: for all i we have $i \leftrightarrow j$.

Symmetric: if $i \leftrightarrow j$ then $j \leftrightarrow i$.

Transitive: if $i \leftrightarrow j$ and $j \leftrightarrow k$ then $i \leftrightarrow k$.

Proof: Reflexive follows from our inclusion of $n = 0$ in the definition of leads to. Symmetry is obvious. To check transitivity it suffices to check that $i \rightsquigarrow j$ and $j \rightsquigarrow k$ imply that $i \rightsquigarrow k$. But if $(\mathbf{P}^m)_{ij} > 0$ and $(\mathbf{P}^n)_{jk} > 0$ then

$$(\mathbf{P}^{m+n})_{ik} = \sum_l (\mathbf{P}^m)_{il}(\mathbf{P}^n)_{lk} \ge (\mathbf{P}^m)_{ij}(\mathbf{P}^n)_{jk} > 0$$

Any equivalence relation on a set partitions the set into **equivalence classes**; two elements are in the same equivalence class if and only if they are equivalent. So communication partitions the state space into equivalence classes called **communicating classes**.

Here is an example:

$$\mathbf{P} = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

To find communicating classes you start with say state 1 and see where it leads. We see $1 \rightsquigarrow 2$, $1 \rightsquigarrow 3$ and $1 \rightsquigarrow 4$ in row 1. Turning to row 4 we see that $4 \rightsquigarrow 1$. This proves (via transitivity) that 1, 2, 3 and 4 are all in the same communicating class. On the other hand I claim none of these states leads to any of 5, 6, 7 or 8. This is almost obvious because of the zeroes in the upper right hand corner of $\mathbf{P}$. For clarity, however, I will give a precise proof. So suppose $i \in \{1,2,3,4\}$ and $j \in \{5,6,7,8\}$. Then $(\mathbf{P}^n)_{ij}$ is a sum of products of terms of the form $\mathbf{P}_{kl}$. This cannot be positive unless there is a sequence $i_0 = i, i_1, \ldots, i_n = j$ with $\mathbf{P}_{i_{k-1},i_k} > 0$ for $k = 1, \ldots, n$. Consider the first k for which $i_k \in \{5,6,7,8\}$ Then $i_{k-1} \in \{1,2,3,4\}$ and so $\mathbf{P}_{i_{k-1},i_k} = 0$.

This proves $\{1,2,3,4\}$ is a communicating class. Now look at state 5. Clearly $5 \rightsquigarrow 1$, $5 \rightsquigarrow 2$, $5 \rightsquigarrow 3$ and $5 \rightsquigarrow 4$. None of these lead to any of $\{5,6,7,8\}$ so $\{5\}$ must be communicating class. Similarly you may check $\{6\}$ and $\{7,8\}$ are communicating classes.

In the example the states 5 and 6 have a special property. Each time you are in either state you run a risk of going to one of the states 1, 2, 3 or 4. Eventually you will make such a transition and then never return to state 5 or 6. States 5 and 6 are transient. To make this precise we define hitting times:

$$T_k = \min\{n > 0 : X_n = k\}$$

We define

$$f_k = P(T_k < \infty | X_0 = k)$$

A state k is called transient if $f_k < 1$ and recurrent if $f_k = 1$.

Let N_k be the number of times the chain ever is in state k. I will show that if $f_i < 1$ then N_k has a Geometric distribution

$$P(N_k = r | X_0 = k) = f_k^{r-1}(1 - f_k)$$

for $r = 1, 2, \ldots$. On the other hand if $f_i = 1$ then I will show

$$P(N_k = \infty | X_0 = k) = 1$$

To prove this theorem I am going to strengthen the Markov property to what is called the strong Markov Property. A stopping time for the Markov chain is a random variable T taking values in $\{0, 1, \cdots\} \cup \{\infty\}$ such that for each finite k there is a function f_k such that

$$1(T = k) = f_k(X_0, \ldots, X_k)$$

Notice that T_k is a stopping time.

It is now convenient to introduce some standard shorthand notation. By

$$P^x(A)$$

we mean

$$P(A|X_0 = x)$$

and similarly we define

$$\mathrm{E}^x(Y) = \mathrm{E}(Y|X_0 = x)$$

Our goal is to explain and prove the formula

$$\mathrm{E}(f(X_T,\ldots)|X_T,\ldots,X_0) = \mathrm{E}^{X_T}(f(X_0,\ldots))$$

I will do a special case for illustration and then discuss conditional probabilities and expectations.

Consider first

$$P(X_{T+1} = j|X_T = i)$$

I will prove this is just

$$\mathbf{P}_{ij} = P^i(X_1 = j)$$

We evaluate

$$\begin{aligned}
P(X_{T+1} = j|X_T = i) &= \frac{P(X_{T+1} = j, X_T = i)}{P(X_T = i)} \\
&= \frac{\sum_k P(X_{T+1} = j, X_T = i, T = k)}{\sum_k P(X_T = i, T = k)} \\
&= \frac{\sum_k P(X_{k+1} = j, X_k = i, T = k)}{\sum_k P(X_k = i, T = k)} \\
&= \frac{\sum_k P(X_{k+1} = j|X_k = i, T = k)P(X_k = i, T = k)}{\sum_k P(X_k = i, T = k)} \\
&= \frac{\sum_k P(X_1 = j|X_0 = i)P(X_k = i, T = k)}{\sum_k P(X_k = i, T = k)} \\
\mathbf{P}_{i,j}
\end{aligned}$$

Notice the use in deleting $T = k$ from the condition of the fact that $T = k$ is an event defined in terms of $X_0,\ldots,X_k$.

There are some technical problems with this proof, however:

- It might be that $P(T = \infty) > 0$. In that case what are X_T and X_{T+1} on the event $T = \infty$. The answer is that the formula must be revised to condition also on $T < \infty$.

- It is not obvious that the event being conditioned on has positive probability. The answer is that we must seek to prove the formula only for those stopping times where $\{T < \infty\} \cap \{X_T = i\}$ has positive probability.

We will now fix up these technical details.

5.2.1 Conditional distributions and expectations

When X and Y are discrete we have

$$\mathrm{E}(Y|X=x) = \sum_y P(Y=y|X=x)$$

for any x for which $P(X=x)$ is positive. This defines a function of x. When this function is evaluated at X we get a random variable which is a function of X denoted

$$\mathrm{E}(Y|X)$$

Here are some properties of that function.

1. Suppose A is a function defined on the range of X. Then

$$\mathrm{E}(A(X)Y|X=x) = A(x)\mathrm{E}(Y|X=x)$$

and so

$$\mathrm{E}(A(X)Y|X) = A(X)\mathrm{E}(Y|X)$$

The second assertion follows by definition from the first. To prove the first note that if $Z = A(X)Y$ then Z is discrete and

$$P(Z=z) = \sum_{x,y} P(Y=y, X=x)1(z=A(x)y)$$

Also

$$\begin{aligned} P(Z=z|X=x) &= \frac{\sum_y P(Y=y,X=x)1(z=A(x)y)}{P(X=x)} \\ &= \sum_y P(Y=y|X=x)1(z=A(x)y) \end{aligned}$$

Thus

$$\begin{aligned} \mathrm{E}(Z|X=x) &= \sum_z zP(Z=z|X=x) \\ &= \sum_z \sum_y zP(Y=y|X=x)1(z=A(x)y) \\ &= \sum_z \sum_y A(x)yP(Y=y|X=x)1(z=A(x)y) \\ &= A(x)\sum_y yP(Y=y|X=x)\sum_z 1(z=A(x)y) \\ &= A(x)\sum_y yP(Y=y|X=x) \end{aligned}$$

2. Repeated conditioning: if X, Y and Z discrete then

$$\begin{aligned} \mathrm{E}\left\{\mathrm{E}(Z|X,Y)|X\right\} &= \mathrm{E}(Z|X) \\ \mathrm{E}\left\{\mathrm{E}(Y|X)\right\} &= \mathrm{E}(Y) \end{aligned}$$

3. Additivity
$$E(Y+Z|X) = E(Y|X) + E(Z|X)$$

4. Putting the first two items together gives
$$E\left\{E(A(X)Y|X)\right\} E(A(X)Y) \tag{5.5}$$
In fact the general definition of $E(Y|X)$ when X and Y are not assumed to be discrete is that it is a random variable which is a measurable function of X satisfying(5.5). The existence of this function is a measure theory problem.

Now suppose that $f(x_0, x_1, \ldots)$ is a (measurable) function on $S^{\mathbb{N}}$. Put
$$Y_n = f(X_n, X_{n+1}, \ldots).$$
Assume that $E(\,|Y_0|\,|X_0 = x) < \infty$ for all x. Then I claim
$$E(Y_n|X_n, A) = E^{X_n}(Y_0) \tag{5.6}$$
whenever A is any event defined in terms of $X_0, \ldots, X_n$.

Proof: The family of functions f for which the claim is true includes all indicators since then the expectations are just probabilities and (5.6) reduces to (5.3). The family of functions for which the claim is true is a vector space (the point being that if f and g are in the family then so is $af + bg$ for any constants a and b. Thus the family of functions f for which the claim is true includes all simple functions. The family of functions f for which the claim is true is closed under the action of taking monotone increasing limits (of non-negative f_n) by the Monotone Convergence theorem. Hence the claim is true for every non-negative integrable f. Finally the claim follows for every integrable f by linearity.

Aside on "measurable": what sorts of events can be defined in terms of a family $\{Y_i : i\; inI\}$? It seems natural to regard any event of the form $(Y_{i_1}, \ldots, Y_{i_k}) \in C$ as a definition in terms of the family for any finite set $i_1, \ldots, i_k$ and any (Borel) set C in S^k. When the state space is countable, incidentally, the natural notion of Borel is to make each single point $(s_1, \ldots, s_k) \in S^k$ Borel. This makes every subset of S^k Borel.

It is also natural to agree that if you can define each of a sequence of events A_n in terms of the Ys then the definition "there exists an n such that (definition of A_n) ..." defines $\cup A_n$. Similarly if A is definable in terms of the Ys then A^c can be defined from the Ys by just inserting the phrase "It is not true that" in front of the definition of A. In other words the family of events definable in terms of the family $\{Y_i : i\; inI\}$ is a σ-field which includes every event of the form $(Y_{i_1}, \ldots, Y_{i_k}) \in C$. We call the smallest such σ-field, $\mathcal{F}(\{Y_i : i\; inI\})$, the σ-field generated by the family $\{Y_i : i\; inI\}$.

This discussion permits some shorthand. We define
$$\mathcal{F}_n = \mathcal{F}(\{X_0, \ldots, X_n\})$$
and
$$\mathcal{F}_\infty = \mathcal{F}_(\{X_0, X_1, \ldots\})$$

Finally I want to discuss conditioning on $X_0, \dots, X_T$ where T is a stopping time. There are several ways to deal with this problem. One is to simply say that the vector $X_0, \dots, X_T$ takes values in a somewhat exotic space, namely:

$$\left(\bigcup_{n=1}^{\infty} S^n\right) \bigcup S^{\mathbb{N}}$$

We define

$$\mathcal{F}_T = \mathcal{F}(X_0, \dots, X_T)$$

The random vector $X_0, \dots, X_T$ is discrete on the event $T < \infty$ but not discrete on the event $T = \infty$. Our interest, however, is really on the event $T < \infty$.

Suppose X is discrete and $X^* = g(X)$ is a one to one transformation of X. Since $X = x$ is the same event as $X^* = g(x)$ we find

$$\mathrm{E}(Y|X = x) = \mathrm{E}(Y|X^* = g(x))$$

Let $h * (u)$ denote the function $\mathrm{E}(Y|X^* = u)$ and $h(u) = \mathrm{E}(Y|X = u)$. Then

$$h(x) = h^*(g(x))$$

Thus

$$h(X) = h^*(g(X) = h^*(X^*)$$

This just means

$$\mathrm{E}(Y|X) = \mathrm{E}(Y|X^*)$$

This formula deserves some interpretation. First of all it says something obvious. Here's an example. I toss a coin $n = 20$ times. Let Y denote the indicator that the first toss is a heads. Let X be the number of heads and X^* be the number of tails. Then what is being said is

$$\mathrm{E}(Y|X = 17) = \mathrm{E}(Y|X^* = 3)$$

In fact for a general k and n

$$\mathrm{E}(Y|X = k) = \frac{k}{n}$$

so

$$\mathrm{E}(Y|X) = \frac{X}{n}$$

At the same time

$$\mathrm{E}(Y|X^* = j) = \frac{n - j}{n}$$

so

$$\mathrm{E}(Y|X^*) = \frac{n - X^*}{n}$$

But of course $X = n - X^*$ so these are just two ways of describing the same random variable.

Another interpretation is this. The random variable X partitions Ω into a countable set of events of the form $X = x$. For any other random variable X^* which partitions Ω into the same events the values of $\mathrm{E}(Y|X^* = x^*)$ are just the same as the values of $\mathrm{E}(Y|X = x)$ but labelled. When we form $\mathrm{E}(Y|X)$ we are simply taking the value ω, computing the value of $X(\omega)$ to see which member A of the partition we are talking about and then writing down the corresponding $\mathrm{E}(Y|A)$. This should make it apparent that we would get the same answer if we used X^* as long as X and X^* define the same partition of Ω.

For random variables which are not discrete the notion of a partition must be replace by a σ-field. In fact suppose X and X^* are two random variables such that

$$\mathcal{F}(X) = \mathcal{F}(X^*)$$

Then:

- There is a Borel g which is one to one and has a one to one Borel inverse such that $X^* = g(X)$.
- $\mathrm{E}(Y|X) = \mathrm{E}(Y|X^*)$ almost surely.

In other words $\mathrm{E}(Y|X)$ depends *only* on the σ-field generated by X. We write

$$\mathrm{E}(Y|\mathcal{F}(X)) = \mathrm{E}(Y|X)$$

Definition: Suppose $\mathcal{G}$ is a sub-σ-field of $\mathcal{F}$. We say that X is $\mathcal{G}$ measurable if, for every Borel B

$$\{\omega : X(\omega) \in B\} \in \mathcal{G} .$$

Definition: $\mathrm{E}(Y|\mathcal{G})$ is any $\mathcal{G}$ measurable random variable such that for every $\mathcal{G}$ measurable random variable A we have

$$\mathrm{E}(AY) = \mathrm{E}\left\{A\mathrm{E}(Y|\mathcal{G})\right\} .$$

Again the existence of such a random variable is a measure theory problem.

Now we define $\mathcal{F}_T$ to be the collection of all events B such that

$$B \cap \{T \leq n\} \in \mathcal{F}_n$$

for each n. It is easy to prove that $\mathcal{F}_T$ is a σ-field and that T and X_T are $\mathcal{F}_T$ measurable.

Finally I rewrite the Strong Markov Property: if

- $f : S^{\mathbb{N}} \mapsto \mathbb{R}$ is Borel and
- $Y_T = f(X_T, X_{T+1}, \ldots)1(T < \infty)$ and
- $Y_0 = f(X_0, X_1, \ldots)$ and
- $\mathrm{E}^x(|Y_0|) < \infty$ for all x.

then

$$1(T < \infty)\mathrm{E}(f(X_T, X_{T+1}, \ldots)|\mathcal{F}_T) = 1(T < \infty)\mathrm{E}^{X_T}(f(X_0, X_1, \ldots)) \tag{5.7}$$

Proof of the Strong Markov Property.

Since T is $\mathcal{F}_T$ measurable we have

$$1(T<\infty)\mathrm{E}\left\{f(X_T,X_{T+1},\dots)|\mathcal{F}_T\right\} = \mathrm{E}\left\{1(T<\infty)f(X_T,X_{T+1},\dots)|\mathcal{F}_T\right\}$$

Now suppose A is $\mathcal{F}_T$ measurable. Then we need only prove that

$$\mathrm{E}\left[A1(T<\infty)\mathrm{E}^{X_T}\left\{f(X_0,X_1,\dots)\right\}\right] = \mathrm{E}\left\{A1(T<\infty)f(X_T,X_{T+1},\dots)\right\}$$

We evaluate these by breaking up the events into the various possible values of T:

$$\begin{aligned} A1(T<\infty)\mathrm{E}^{X_T}\left\{f(X_0,X_1,\dots)\right\} &= \sum_k A1(T=k)\mathrm{E}^{X_T}\left\{f(X_0,X_1,\dots)\right\} \\ &= \sum_k A1(T=k)\mathrm{E}^{X_k}\left\{f(X_0,X_1,\dots)\right\} \end{aligned}$$

On the other hand

$$\begin{aligned} A1(T<\infty)f(X_T,X_{T+1},\dots) &= \sum_k A1(T=k)f(X_T,X_{T+1},\dots) \\ &= \sum_k A1(T=k)f(X_k,X_{k+1},\dots) \end{aligned}$$

But

$$\begin{aligned} \mathrm{E}\left\{A1(T=k)f(X_k,X_{k+1},\dots)\right\} &= \mathrm{E}\left[\mathrm{E}\left\{A1(T=k)f(X_k,X_{k+1},\dots)|\mathcal{F}_k\right\}\right] \\ &= \mathrm{E}\left[A1(T=k)\mathrm{E}\left\{f(X_k,X_{k+1},\dots)|\mathcal{F}_k\right\}\right] \\ &= \mathrm{E}\left[A1(T=k)\mathrm{E}^{X_k}\left\{f(X_0,X_1,\dots)\right\}\right] \end{aligned}$$

by (5.3). But this is exactly the expectation of the kth term in (5.2.1). This proves the Strong Markov Property.

5.3 Initial Distributions

Up to now I have hidden the nature of the unconditional expected values. The Markov property (5.1) specifies only conditional probabilities. There is no way to deduce the marginal distributions. Instead, for every possible distribution π on S and each possible transition matrix $\mathbf{P}$ there is a a stochastic process $X_0, X_1, \dots$ with

$$P(X_0 = k) = \pi_k$$

and which is a Markov Chain with transition matrix $\mathbf{P}$.

In our proof of the strong Markov property we used unconditional expectations. Part of the point of the calculation is that it works for any initial distribution. Notation: if π is an initial distribution (a probability measure on S) then E^π denotes expected values for chains with initial distribution π; we use $\mathbf{P}^\pi$ for the corresponding probability measure.

Summary of easily verified facts:

- For any sequence of states $i_0, \ldots, i_k$

$$P(X_0 = i_0, \ldots, X_k = i_k) = \pi_{i_0} \mathbf{P}_{i_0 i_1} \cdots \mathbf{P}_{i_{k-1} i_k}$$

- For any event A:

$$\mathbf{P}^{\pi}(A) = \sum_k \pi_k \mathbf{P}^k(A)$$

- For any bounded random variable $Y = f(X_0, \ldots)$

$$\mathrm{E}^{\pi}(Y) = \sum_k \pi_k \mathrm{E}^k(A)$$

5.4 Recurrence and Transience

Now consider a transient state k, that is, a state for which

$$f_k = P^k(T_k < \infty) < 1$$

Note that $T_k = \min\{n > 0 : X_n = k\}$ is a stopping time. Let N_k be the number of visits to state k. That is

$$N_k = \sum_{n=0}^{\infty} 1(X_n = k)$$

Notice that if we define the function

$$f(x_0, x_1, \ldots) = \sum_{n=0}^{\infty} 1(x_n = k)$$

then

$$N_k = f(X_0, X_1, \ldots)$$

Notice, also, that on the event $T_k < \infty$

$$N_k = 1 + f(X_{T_k}, X_{T_k+1}, \ldots)$$

and on the event $T_k = \infty$ we have

$$N_k = 1$$

In short:

$$N_k = 1 + f(X_{T_k}, X_{T_k+1}, \ldots) 1(T_k < \infty)$$

Hence

$$\begin{aligned}
\mathbf{P}^k(N_k = r) &= \mathrm{E}^k\left\{P(N_k = r | calF_T)\right\} \\
&= \mathrm{E}^k\left\{\mathbf{P}(1 + f(X_{T_k}, X_{T_k+1}, \ldots)1(T_k < \infty) = r | calF_T)\right\} \\
&= \mathrm{E}^k\left[1(T_k < \infty) P^{X_{T_k}}\left\{f(X_0, X_1, \ldots) = r - 1\right\}\right] \\
&= \mathrm{E}^k\left\{1(T_k < \infty) P^k(N_k = r - 1)\right\} \\
&= \mathrm{E}^k\left\{1(T_k < \infty)\right\} P^k(N_k = r - 1) \\
&= f_k P^k(N_k = r - 1)
\end{aligned}$$

It is easily verified by induction, then, that

$$\mathbf{P}^k(N_k = r) = f_k^{r-1} P^k(N_k = 1)$$

But $N_k = 1$ if and only if $T_k = \infty$ so

$$\mathbf{P}^k(N_k = r) = f_k^{r-1}(1 - f_k)$$

which shows that N_k has (if the chain starts from state k) a Geometric distribution with mean $1/(1 - f_k)$. Notice that the argument also shows that if $f_k = 1$ then

$$P^k(N_k = 1) = P^k(N_k = 2) = \cdots$$

which can only happen if all these probabilities are 0. Thus

$$P(N_k = \infty) = 1$$

if $f_k = 1$.

Now

$$N_k = \sum_{n=0}^{\infty} 1(X_n = k)$$

Taking expectations gives

$$\mathrm{E}^k(N_k) = \sum_{n=0}^{\infty} (\mathbf{P}^n)_{kk}$$

We therefore learn:
State k is transient if and only if

$$\sum_{n=0}^{\infty} (\mathbf{P}^n)_{kk} < \infty$$

and that this sum is $1/(1 - f_k)$.

Proposition 2 *Recurrence (or transience) is a class property. That is, if i and j are in the same communicating class then i is recurrent (respectively transient) if and only if j is recurrent (respectively transient).*

Proof: Suppose i is recurrent and $i \leftrightarrow j$. There are integers m and n such that

$$(\mathbf{P}^m)_{ji} > 0 \quad \text{and} \quad (\mathbf{P}^n)_{ij} > 0$$

Then

$$\begin{aligned}
\sum_k (\mathbf{P}^k)_{jj} &\geq \sum_{k \geq 0} (\mathbf{P}^{m+k+n})_{jj} \\
&\geq \sum_{k \geq 0} (\mathbf{P}^m)_{ji} (\mathbf{P}^k)_{ii} (\mathbf{P}^n)_{ij} \\
&= (\mathbf{P}^m)_{ji} \left\{ \sum_{k \geq 0} (\mathbf{P}^k)_{ii} \right\} (\mathbf{P}^n)_{ij}
\end{aligned}$$

The middle term is infinite and the two outside terms positive so

$$\sum_k (\mathbf{P}^k)_{jj} = \infty$$

which shows j is recurrent. •

In a finite state space chain there is at least one recurrent state. If all states we transient we would have for each k $P(N_k < \infty) = 1$. This would mean $P(\forall k . N_k < \infty) = 1$. But for any ω there must be at least one k for which $N_k = \infty$ (the total of a finite list of finite numbers is finite).

For an infinite state space chain this argument does not work. The chain X_n satisfying $X_{n+1} = X_n + 1$ on the integers has all states transient.

A more interesting example is this. Toss a coin repeatedly. Let X_n be X_0 plus the number of heads minus the number of tails in the first n tosses. Let p denote the probability of heads on an individual trial.

Notice that $X_n - X_0$ is a sum of n iid random variables Y_i where $P(Y_i = 1) = p$ and $P(Y_i = -1) = 1-p$. The Strong Law of Large Numbers shows X_n/n converges almost surely to $2p-1$. If $p \neq 1/2$ this is not 0. In order for X_n/n to have a positive limit we must have $X_n \to \infty$ almost surely so all states are visited only finitely many times. That is, all states are transient. Similarly for $p < 1/2$ $X_n \to -\infty$ almost surely and all states are transient.

Now look at $p = 1/2$. The law of large numbers argument no long shows anything. I will show that all states are recurrent.

Proof: We show $\sum_n (\mathbf{P}^n)_{00}$ and show the sum is infinite. If n is odd then $(p_n)_{00} = 0$ so we evaluate

$$\sum_m (\mathbf{P}^{2m})_0 0$$

Now

$$(\mathbf{P}^{2m})_0 0 = \binom{2m}{m} 2^{-2m}$$

According to Stirling's approximation

$$\lim_{m\to\infty} \frac{m!}{m^{m+1/2} e^{-m} \sqrt{2\pi}} = 1$$

Hence

$$\lim_{m\to\infty} \sqrt{m} (\mathbf{P}^{2m})_0 0 = \frac{1}{\sqrt{\pi}}$$

Since

$$\sum \frac{1}{\sqrt{m}} = \infty$$

we are done.

5.5 Mean return times

There is a substantial difference between the recurrent states in a finite state space Markov Chain and the recurrent states in our last example. The first step in examining the difference is to compute expected times to return. Again, for $x \in S$ let T_x denote the hitting time for x. Consider a recurrent state x. Since the chain, once having entered the class containing x cannot leave it we will simply assume that there is only one communicating class for the chain. We now derive a set of linear equations between the various expected values of different T_x. Note that each T_x is a certain function f_x applied to $X_1, \ldots$. Setting $\mu_{ij} = \mathrm{E}^i(T_j)$ we find

$$\mu_{ij} = \sum_k \mathrm{E}^i(T_j 1(X_1 = k))$$

Note that if $X_1 = x$ then $T_x = 1$ so

$$\mathrm{E}^i(T_j 1(X_1 = j)) = \mathbf{P}_{ij}$$

For $k \neq j$

$$T_x = 1 + f_x(X_2, X_3, \ldots)$$

and, by conditioning on $X_1 = k$ we find

$$\mathrm{E}^i(T_j 1(X_1 = k)) = \mathbf{P}_{ik}\left\{1 + \mathrm{E}^k(T_j)\right\}$$

This gives

$$\mu_{ij} = 1 + \sum_{k \neq j} \mathbf{P}_{ik}\mu_{kj} \tag{5.8}$$

Technically, I should check that the expectations in (5.8) are finite. All the random variables involved are non-negative, however, and the equation actually makes sense even if some terms are infinite. (To prove this you actually study

$$T_{x,n} = \min(T_x, n)$$

deriving an identity for a fixed n, letting $n \to \infty$ and applying the monotone convergence theorem.)

Here is a simple example:

$$\mathbf{P} = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix}$$

The identity (5.8) becomes

$$\begin{aligned}
\mu_{1,1} &= 1 + \frac{1}{2}\mu_{2,1} + \frac{1}{2}\mu_{3,1} \\
\mu_{1,2} &= 1 + \frac{1}{2}\mu_{3,1} \\
\mu_{1,3} &= 1 + \frac{1}{2}\mu_{2,1} \\
\mu_{2,1} &= 1 + \frac{1}{2}\mu_{3,1} \\
\mu_{2,2} &= 1 + \frac{1}{2}\mu_{1,2} + \frac{1}{2}\mu_{3,2} \\
\mu_{2,3} &= 1 + \frac{1}{2}\mu_{1,3} \\
\mu_{3,1} &= 1 + \frac{1}{2}\mu_{2,1} \\
\mu_{3,2} &= 1 + \frac{1}{2}\mu_{1,2} \\
\mu_{3,3} &= 1 + \frac{1}{2}\mu_{1,3} + \frac{1}{2}\mu_{2,3}
\end{aligned}$$

The seventh and fourth together give

$$\mu_{2,1} = \mu_{3,1}$$

Similar calculations lead to $\mu_{ii} = 3$ and, for $i \neq j$ $\mu_{i,j} = 2$.
Example: The coin tossing Markov Chain with $p = 1/2$ shows that the situation can be different when S is infinite. The equations above become:

$$\begin{aligned}
m_{0,0} &= 1 + \frac{1}{2}m_{1,0} + \frac{1}{2}m_{-1,0} \\
m_{1,0} &= 1 + \frac{1}{2}m_{2,0}
\end{aligned}$$

and many more.

Here are some observations:

- You have to go through 1 to get to 0 from 2 so

$$m_{2,0} = m_{2,1} + m_{1,0}\,.$$

- The chain which counts tails minus heads has exacly the same distribution as X_n by symmetry. So:

$$m_{1,0} = m_{-1,0}$$

- The transition probabilities are **spatially homogeneous**, meaning $\mathbf{P}_{ij}$ is a function of $j - i$ only. Hence

$$m_{2,1} = m_{1,0}$$

Conclusion:

$$\begin{aligned} m_{0,0} &= 1 + m_{1,0} \\ &= 1 + 1 + \frac{1}{2} m_{2,0} \\ &= 2 + m_{1,0} \end{aligned}$$

Notice that there are **no** finite solutions!

Summary of the situation:

- Every state is recurrent.
- All the expected hitting times m_{ij} are infinite.
- All entries $\mathbf{P}^n_{ij}$ converge to 0.

Jargon: The states in this chain are **null recurrent**.

Definition: A recurrent state x is positive recurrent if $\mathrm{E}^x(T_x) < \infty$ otherwise x is null recurrent.

5.6 The ergodic theorem

Consider a finite state space chain. If x is a vector then the ith entry in $\mathbf{P}x$ is

$$\sum_j \mathbf{P}_{ij} x_j$$

This is, since the rows of $\mathbf{P}$ are probability vectors, a weighted average of the entries in x. If the weights are strictly between 0 and 1 and the largest and smallest entries in x are not the same then $\sum_j \mathbf{P}_{ij} x_j$ is strictly between the largest and smallest entries in x. In fact

$$\sum_j \mathbf{P}_{ij} x_j - \min(x_k) \ge \min_j \{p_{ij}\}(\max\{x_k\} - \min\{x_k\})$$

and

$$\max\{x_j\} - \sum_j \mathbf{P}_{ij} x_j \ge \min_j \{p_{ij}\}(\max\{x_k\} - \min\{x_k\})$$

Now consider what happens when we multiply $\mathbf{P}^r$ by $\mathbf{P}^m$. The ijth entry in $\mathbf{P}^{r+m}$ is a weighted average of the jth column of $\mathbf{P}^m$ and so, if all the entries in row i of $\mathbf{P}^r$ are positive and the jth column of $\mathbf{P}^m$ is not constant, the ith entry in the jth column of $\mathbf{P}^{r+m}$ must be strictly between the minimum and maximum entries of the jth column of $\mathbf{P}^m$. In fact, fix a j and let $\overline{x}_m$ be the maximum entry in column j of $\mathbf{P}^m$ and $\underline{x}_m$ the minimum entry. Suppose all entries of $\mathbf{P}^r$ are positive. Let $\delta > 0$ be the smallest entry in $\mathbf{P}^r$. Our argument above shows that