Supplementary notes for Math 495

Renato Feres

1 Probability Spaces

By definition, a a set A is said to be *countable* if it is either finite or has the form $A = \{a_1, a_2, a_3, \cdots\}$. For example, the set of integers,

$$\mathbb{Z} = \{0, -1, 1, -2, 2, \cdots\},$$

is countable. So is the set of rational numbers,

$$\mathbb{Q} = \left\{ \frac{m}{n} : m, n \in \mathbb{Z}, n \neq 0 \right\}.$$

The latter can be enumerated by listing the pairs $(m, n) \in \mathbb{Z}^2$ in some fashion, for example, as described by the next figure.

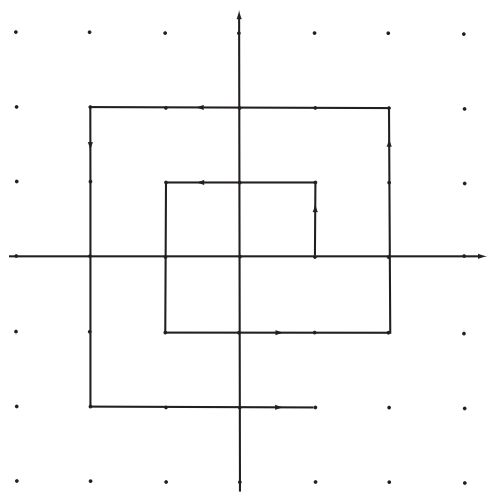

Figure 1: Enumerating pairs of integers

A set B is said to be a *countable union*, or union of countably many sets, if there are sets $B_1, B_2, B_3, \cdots$ such that

$$B = B_1 \cup B_2 \cup B_3 \cup \cdots = \bigcup_{k=1}^{\infty} B_k.$$

Definition 1 (Probability spaces) A *probability space* consists of a triple $(\Omega, \mathcal{F}, P)$, where Ω is a set, called the set of *elementary outcomes* of a given probability experiment; $\mathcal{F}$ is a family of subsets of Ω called the set of *events*; and $P : \mathcal{F} \to [0,1]$ is a function that associates to each event $A \in \mathcal{F}$ its *probability*. $\mathcal{F}$ and P are required to satisfy the following axions:

1. $\Omega \in F$;
2. If $A \in \mathcal{F}$, its complement $\bar{A} = \Omega \setminus A$ is also in $\mathcal{F}$;
3. If $A_1, A_2, \cdots$ are in $\mathcal{F}$, their union $\bigcup_k A_k$ is also in $\mathcal{F}$;
4. $P(\Omega) = 1$;
5. If $A_1, A_2, \cdots$ are disjoint elements of $\mathcal{F}$ (i.e., $A_i \cap A_j = \emptyset$ for all i, j such that $i \neq j$), then
$$P\left(\bigcup_k A_k\right) = \sum_k P(A_k).$$

We refer to $\mathcal{F}$ as a *σ-algebra* (sigma-algebra) of events, and to P as a *probability measure* on $\mathcal{F}$. An event A for which $P(A) = 0$ is called *negligible*. If $P(A) = 1$, we say that A holds *almost surely*.

Example 1 (Flipping a coin once) We can take $\Omega = \{0, 1\}$ and $\mathcal{F}$ the set of all subsets of Ω:
$$\mathcal{F} = \{\emptyset, \{0\}, \{1\}, \Omega\}.$$
The associated probabilities can be set as $P(\{0\}) = p$, $P(\{1\}) = 1 - p$. The probabilities of the other sets are fixed by the axioms: $P(\Omega) = 1$ and $P(\emptyset) = 0$. For a *fair* coin we set $p = 1/2$.

Exercise 1 Let $(\Omega, \mathcal{F}, P)$ be a probability space. Show that:

1. $P(\emptyset) = 0$;
2. Given $A_1, A_2, A_3, \cdots$ in $\mathcal{F}$, then $\bigcap_k A_k$ is also in $\mathcal{F}$;
3. If $A, B \in \mathcal{F}$ satisfy $A \subset B$, then $P(A) \leq P(B)$;
4. If $A_1, A_2, A_3, \cdots$ are elements of $\mathcal{F}$ (not necessarily disjoint), then
$$P\left(\bigcup_k A_k\right) \leq \sum_k P(A_k).$$

A hint: for part 3 of the exercise, write B as a disjoint union: $B = A \cup (B \setminus A)$. For part 4, note that $A_1 \cup A_2 \cup \cdots$ equals the union $A'_1 \cup A'_2 \cup \cdots$, where
$$A'_1 = A_1,\ A'_2 = A_2 \setminus A_1,\ A'_3 = A_3 \setminus (A_1 \cup A_2),\ \ldots.$$

Example 2 (Choosing a number in the unit interval at random) Here $\Omega = [0,1]$. The definition of $\mathcal{F}$ is a little more involved, but essentially amounts to the collection of subsets of $[0,1]$ for which a length can be assigned. More precisely, we define $\mathcal{F}$ as the smallest collection of subsets of $[0,1]$ that contains all intervals $[0,a]$, for $a \in [0,1]$, such that axioms 1, 2, and 3 in the definition of a probability space hold. It can be shown that other types of intervals $((a,b),[a,b),(a,b],[a,b]$, for $a,b \in [0,1])$ also belong to $\mathcal{F}$. (See the exercise below.) The probability measure P is defined on intervals by $P([0,a]) = a$ (the length of $[0,a]$). On other sets in $\mathcal{B}$,the probability is obtained by using the axioms. For example, $P([a,b]) = b - a$. (See the below exercise.) We will reserve the special notation, $\mathcal{B}$, for this family of events and call them *Borel sets*. We will also, occasionally, denote the probability measure by λ and call it the *Lebesgue mesure* on $[0,1]$.

Exercise 2 We consider here the probability space $([0,1],\mathcal{B},\lambda)$ defined above. Show that the following subsets of $[0,1]$ are Borel sets and find their Lebesgue measure:

1. $[a,b]$, $(a,b]$, $[a,b)$, (a,b);
2. $\mathbb{Q} \cap [0,1]$;
3. The set of irrational numbers in $[0,1]$.

A few hints: Intervals $(a,1]$ belong to $\mathcal{B}$ (complement of $[0,a]$) so the intersection $(a,b] = [0,b] \cap (a,1]$ is also in $\mathcal{B}$. The countable family of sets $(c_n,b]$, for an increasing sequence c_n that converges to b, has intersection $\{b\}$, so single point sets are Borel. (What if $b = 0$?) $\mathbb{Q}$ is a countable union of single point sets, so it is also Borel. By taking the probability of the disjoint union $[0,b] = [0,a] \cup (a,b]$, we conclude $\lambda((a,b]) = b - a$. Choose an increasing sequence $c_n \to b$ and write the disjoint union

$$(a,b) = (a,c_1] \cup (c_1,c_2] \cup (c_2,c_3] \cup \cdots .$$

From this, show that

$$\lambda((a,b)) = c_1 - a + \sum_{k=1}^{\infty}(c_{k+1} - c_k) = \lim_{n\to\infty}(c_n - a) = b - a.$$

Using the disjoint union $(a,b] = (a,b) \cup \{b\}$ conclude that $P(\{b\}) = 0$. Therefore $\mathbb{Q} \cap [0,1]$, or any other countable subset of $[0,1]$, has (Lebesgue) probability measure 0. Therefore, such a set is negligible from a probabilistic (measure theoretic) viewpoint. On the other hand, choosing an irrational number at random is an almost sure event, i.e., an event of probability 1.

It can be shown that any set $A \in \mathcal{B}$ has a well-defined probability measure, $\lambda(A)$, obtained by applying the axioms of a probability measure and the definitions in Example 2. It is natural to ask whether there are subsets of $[0,1]$ for which it is impossible to assign a probability without incurring in some logical

contradiction. These non-measurable sets exist, and they are part of the reason for including the family of events $\mathcal{F}$ explicitly into the definition of a probability space. (There are other, more practical, reasons.) These things are explained in courses on Measure Theory. Note that none of this matters for finite probability spaces (that is, when Ω is a finite set).

2 Independence and Random Variables

We fix a probability space $(\Omega, \mathcal{F}, P)$.

Definition 2 (Conditional probability and independence) *Let $A, B \in \mathcal{F}$ be two events such that $P(B) \neq 0$. The* conditional probability *of A given B is defined by*

$$P(A|B) := \frac{P(A \cap B)}{P(B)}.$$

Events $A_1, A_2, \ldots, A_k \in \mathcal{F}$, are said to be independent *if*

$$P(A_1 \cap \cdots \cap A_k) = P(A_1) \cdots P(A_k).$$

Events $A_1, A_2, \ldots$ are independent if $A_1, A_2, \ldots, A_k$ are independent for each k. Note: If $P(B) \neq 0$, A and B are independent if and only if $P(A|B) = P(A)$.

We say that a function $X : \Omega \to \mathbb{R}$ is *measurable* relative to $\mathcal{F}$ (or, simply, *$\mathcal{F}$-measurable*) if for all $a \in \mathbb{R}$, the set $\{\omega \in \Omega : X(\omega) \leq a\}$ belongs to $\mathcal{F}$. In other words, the condition $X \leq a$ defines an event. The probability of this event is usually written

$$P(X \leq a) := P(\{\omega \in \Omega : X(\omega) \leq a\}).$$

The probability of other events, such as $a \leq X \leq b$, are obtained from the values of $P(X \leq c)$, $c \in \mathbb{R}$, by taking intersections, unions, complements, and using the axioms of probability measures. The function

$$F_X(x) := P(X \leq x)$$

is called the *cumulative distribution function* of the random variable X.

Definition 3 (Random variables) *Fix a probability space $(\Omega, \mathcal{F}, P)$. A* random variable *with values in $\mathbb{R}$ is any $\mathcal{F}$-measurable function $X : \Omega \to \mathbb{R}$. That is, any function such that the sets defined by $X \leq a$ belong to $\mathcal{F}$ for all $a \in \mathbb{R}$.*

Having fixed $(\Omega, \mathcal{F}, P)$, any given random variable $X : \Omega \to \mathbb{R}$ gives rise to a probability space $(\mathbb{R}, \mathcal{B}, P_X)$, where $\mathcal{B}$ is the σ-algebra of Borel subsets of $\mathbb{R}$ (which is the smallest σ-algebra of subsets of $\mathbb{R}$ that contain the intervals $(-\infty, a]$), and P_X is defined as follows: for any Borel subset $A \subset \mathbb{R}$,

$$P_X(A) = P(X \in A).$$

P_X is sometimes called the *probability law* of X. Because of this remark, we often do not need to, and will not, make explicit the probability space $(\Omega, \mathcal{F}, P)$; we only use the set of values of X in $\mathbb{R}$ and the probability law P_X (on Borel sets).

Exercise 3 As a simple illustration of the remark on the above paragraph, consider the random variable on $([0,1], \mathcal{B}, \lambda)$ given by $X(x) = 0$ if $x \in [0,p]$ and $X(x) = 1$ if $x \in (p,1]$. The set of values is $\{0,1\}$. Show that P_X is the probability measure on the set $\{0,1\}$ which assigns probability p to 0 and $1-p$ to 1.

We have already defined independence of events. We now define independence of random variables. This will be understood as follows:

Definition 4 (Independent random variables) *Two random variables X and Y are independent if any two events, A and B,* defined in terms of X *and Y, respectively, are independent.*

To make sense of this definition we need to clarify what it means for an event to be defined in terms of a random variable. Whatever the exact definition may be, we certainly want sets of the form $X \in [a,b)$, $X \leq c$, etc., to be events defined in terms of X. This idea can be formalized by means of the following definition.

Definition 5 (Events defined in terms of a random variable) *Let X be a random variable and define $\mathcal{F}_X$ to be the smallest subset of $\mathcal{F}$ that contains the events $X \leq a$, $a \in \mathbb{R}$, and satisfies the axioms for a σ-algebra of events, namely: (i) $\Omega \in \mathcal{F}_X$, (ii) the complement of any event in $\mathcal{F}_X$ is also in $\mathcal{F}_X$, and (iii) the union of countably many events in $\mathcal{F}_X$ is also in $\mathcal{F}_X$. We call $\mathcal{F}_X$ the* σ-algebra of events generated by X. *An event in $\mathcal{F}$ is said to be* defined in terms of X *if it belongs to $\mathcal{F}_X$.*

More generally, we say that random variables $X_1, X_2, \ldots$ are independent if sets $A_1, A_2, \ldots$ defined in terms of the respective X_i are independent.

Definition 6 (I.i.d.) *Random variables $X_1, X_2, X_3, \cdots$ are said to be i.i.d. (independent, identically distributed) if they are independent and, for each $a \in \mathbb{R}$, the probabilities $P(X_i \leq a)$ are the same for all i, that is, they have the same cumulative distribution functions $F_{X_i}(x)$.*

Example 3 (roll of three dice) Set $\Omega = \{1, 2, \ldots, 6\}^3$, $\mathcal{F}$ the collection of all subsets of Ω, and for $A \subset \Omega$ define

$$P(A) := \frac{\#A}{\#\Omega},$$

where $\#A$ denotes the number of elements of A. (Note: $\#\Omega = 216$.) Thus an elementary outcome is a triple $\omega = (i,j,k)$, where i,j,k are integers from 1 to 6. For example, $(1,4,3)$ represents the outcome shown in the next figure.

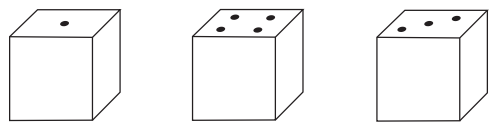

Figure 2: Roll of three dice

Define the random variables X_1, X_2, X_3 by $X_1(i,j,k) = i$, $X_2(i,j,k) = j$, and $X_3(i,j,k) = k$. An event $A \in \mathcal{F}$ is defined in terms of X_2 if and only if it is the union of sets defined by $X_2 = i$. Note that $P(X_l = u) = 1/6$ for any $l = 1,2,3$, and $u = 1,2,\ldots,6$. We have, for example: if A is the event $X_1 = i$ and B is the event $X_2 = j$, then A and B are independent since

$$\begin{aligned} P(A \cap B) &= \frac{\#\{(i,j,1),\ldots,(i,j,6)\}}{6^3} \\ &= \frac{6}{216} \\ &= \frac{1}{6}\frac{1}{6} \\ &= P(A)P(B). \end{aligned}$$

Exercise 4 Check that the random variables X_1, X_2, X_3 are i.i.d.

Continuing with the rolling of three dice experiment, we would like now to find a random variable $X : [0,1] \to \Omega$, defined on the probability space $([0,1], \mathcal{B}, \lambda)$, such that $P_X = P$, as defined in the example. Before seeing how to do it, it is well to justify what use such a random variable would have. Note that, picking a (pseudo-) random number uniformly distributed over the interval $[0,1]$ is the most basic random number most math utility programs, such as Matlab or Mapple, will produce. If what we want is to simulate the dice experiment, then having such a random variable X solves the problem rather easily: we use the appropriate command to get a random $x \in [0,1]$ (in Matlab this is: `x=rand`) and then evaluate $X(x) = (i,j,k)$.

We begin our description of X by a few remarks about representing numbers from $(0,1]$ in base 6. (It will be marginally convenient to exclude 0. There is no harm in doing it since we are only discarding a set of measure zero from $[0,1]$.) Let $K = \{0,1,2,3,4,5\}$ and $I_k = (k/6, (k+1)/6]$. These are the six intervals on the top of Figure 3.

A number $x \in (0,1]$ has expansion in base 6 of the form $0.\omega_1\omega_2\omega_3\ldots$, where $\omega_i \in K$, if we can write

$$x = \frac{\omega_1}{6} + \frac{\omega_2}{6^2} + \frac{\omega_3}{6^3} + \ldots.$$

The representation is in general not unique; for example, $0.1000\cdots = 0.0555\ldots$, since

$$\frac{1}{6} = \frac{0}{6} + \frac{5}{6^2} + \frac{5}{6^3} + \frac{5}{6^4} + \ldots.$$

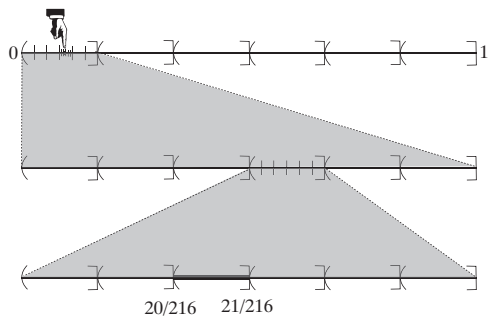

Figure 3: Six-adic interval associated to the dice roll event $(1, 4, 3)$.

To make the choice of ω_i for a given x unique, we can proceed as follows. First, for any $x > 0$ (not necessarily less than 1) define $\eta(x)$ as the non-negative integer (possibly 0) such that

$$\eta(x) < x \leq \eta(x) + 1.$$

Now, define a transformation $T : (0, 1] \to (0, 1]$ by

$$T(x) = 6x - \eta(6x).$$

To obtain a bijective correspondence between $x \in (0, 1]$ and sequences $(\omega_1, \omega_2, \dots)$, define for each $n = 1, 2, \dots$:

$$\omega_n := X_n(x) := k + 1 \Leftrightarrow T^{n-1}(x) \in I_k.$$

In other words, the value of, say, ω_9 is 4, if and only if the 8th iterate of T applied to x, $T^8(x)$ falls into the fifth interval, $I_4 = (4/6, 5/6]$. It is easy to see why this works. For simplicity, assume that $T^n(x)$ never falls on the end point of an interval I_k. (See the below exercise.) Now observe:

$$\begin{aligned} x = T^0(x) &= 0.\omega_1\omega_2\omega_3\dots && \in I_{\omega_1} \\ T^1(x) &= 0.\omega_2\omega_3\omega_4\dots && \in I_{\omega_2} \\ T^2(x) &= 0.\omega_3\omega_4\omega_5\dots && \in I_{\omega_3} \\ &\vdots && \vdots \end{aligned}$$

Exercise 5 Check that:

1. The six-adic representation, $0.\omega_1\omega_2\dots$ ($\omega_i \in \{0, \dots, 5\}$), is ambiguous if and only if for some n, $T^n(x)$ falls at the end point of one of the I_k.

2. The set of $x \in (0, 1]$ having an ambiguous six-adic representation is a countable set. In particular, it is a negligible set from the point of view of probability (with respect to the Lebesgue probability measure).

3. We can lift the ambiguity on the six-adic representation of x by requiring that $0.\omega_1\omega_2\ldots$ contains infinitely many non-zero digits.

We can now define our mathematical model of dice rolling as follows. Choose a point $x \in (0,1]$ at random with uniform probability distribution. (That is, with respect to the Legesgue measure.) Write its six-adic expansion, $x = 0.\omega_1\omega_2\omega_3\ldots$, where ω_n is such that $T^{n-1}(x) \in I_{\omega_n}$. Then the triple

$$X(x) = (\omega_1 + 1, \omega_2 + 1, \omega_3 + 1)$$

can be regarded as the outcome of rolling three dice (or one die three times). The justification for this claim is contained in the next exercise.

Exercise 6 Define the random variables $X_n : (0,1] \to \{1,\ldots,6\}$ on the Borel probability space $((0,1],\mathcal{B},\lambda)$ by

$$X_n(x) := \omega_n + 1 \Leftrightarrow x = 0.\omega_1\omega_2\ldots\omega_n\ldots.$$

Then the sequence $X_1, X_2, \ldots$ is i.i.d., and $P(X_n = k) = 1/6$ for each k.

3 Note on Non-measurable Sets

It was pointed out earlier that one of the reasons for making explicit the σ-algebra of events $\mathcal{F}$ in the definition of probability space is that, for the most important example, $([0,1],\mathcal{B},\lambda)$, there are subsets of Ω for which it is not possible to associate a probability measure. Such subsets cannot be regarded as events of any probability experiment without incurring logical contradictions, and must be excluded from the outset. The purpose of this section is to construct one example of such a non-measurable set.

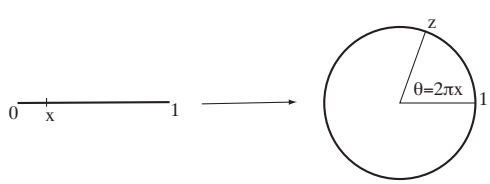

Figure 4: Identifying the interval with the unit circle

Rather than construct the set in $[0,1]$ directly, it is simpler to construct it in the unit circle $S^1 = \{z \in \mathbb{C} : |z| = 1\}$ and note that we can map the interval onto the circle almost bijectively (except for the endpoints 0 and 1, which are sent to $1 \in S^1$) using the map

$$x \in [0,1] \mapsto e^{2\pi i x} \in S^1.$$

Also note that the Lebesgue measure on the interval corresponds under this map to the length of segments of arc on the circle. Therefore, this measure is invariant under rotation. We denote by $R_\theta : S^1 \to S^1$ the rotation by angle θ:

$$R_\theta(z) = e^{i\theta} z.$$

Fix $\alpha = 2\pi\sqrt{2}$. (There is nothing special about $\sqrt{2}$ other than that it is irrational.) Define for each $z \in S^1$ the set

$$I_z := \{R_\alpha^m(z) : m \in \mathbb{Z}\}.$$

In other words, I_z consists of all the points on the circle obtained by rotating z by an integer multiple of the angle α.

If you took Math 310, you should not have much difficulty checking the claims of the following exercise.

Exercise 7 Show that the following claims hold:

1. For any two points $z_1, z_2 \in S^1$, either $I_{z_1} = I_{z_2}$ or $I_{z_1} \cap I_{z_2} = \emptyset$. Note: check that the relation

   $$z_1 \sim z_2 \Leftrightarrow z_2 = R_\alpha^m(z_1), \text{ for some } m \in \mathbb{Z}$$

   is an equivalence relation.

2. By item one, argue that S^1 is an (uncountable) union of sets

   $$S^1 = \bigcup_{z \in A} I_z,$$

   where A is a set that contains a single element from each distinct equivalence class. (Being able to do this requires that you have faith in the axiom of choice. In fact, the set A is defined by choosing from each distinct set I_z a single element.)

3. Let $A_m = R_\alpha^m(A)$ denote the sets obtained by rotating A by the angle $m\alpha$. Using the fact that $\alpha/2\pi$ is irrational, show that

   $$S^1 = \bigcup_{m \in \mathbb{Z}} A_m,$$

   and that the union is disjoint.

4. Now argue that A cannot be a measurable set. In fact, suppose, in order to arrive at a contradiction, that it makes sense to define the arc-length of A. We denote this arc-length by $2\pi\lambda(A)$, where $\lambda(A)$ is the Lebesgue measure the set would have if we view it as a subset of $[0, 1]$ through the identification with S^1 described in Figure 4. Arc-lengths are not changed under rotations, so each set A_m must have the same length as A. But now we have a dilemma: if A has positive length, the length of S^1 would be

infinite, by the axiom of additivity of probability under countable disjoint unions and the claim of Part 3. On the other hand, if A had arc-length 0, S^1 would be forced to have arc-length zero, being a countable union of sets of zero length. Either way we have a contradiction. The only escape is to conclude that A cannot be assigned an arc-length measure consistent with the axioms of probability.

4 Note on Lebesgue Integration

The definition of measurable sets leads in a natural way to a very general and powerful notion of integration, known as the *Lebesgue integral*. We take here a very brief look at this concept. Our main motivation is to have a general notion of expected value of a random variable.

Suppose that $(\Omega, \mathcal{F}, P)$ is a probability space and let $X : \Omega \to \mathbb{R}$ be a random variable. For technical convenience we assume that X is bounded by two numbers: $a \leq X \leq b$, although this is not esssential.

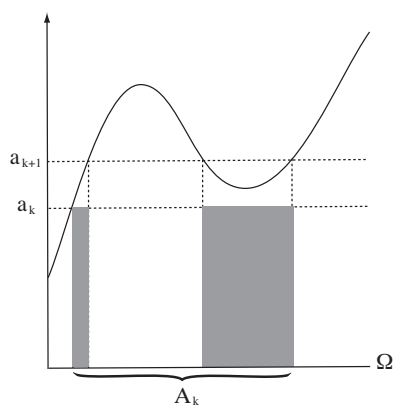

Figure 5: Pre-image of an interval

For a given positive integer n subdivide the interval $[a, b]$ into n equally spaced intervals of length $(b - a)/n$:

$$[a, b] = [a_0, a_1] \cup (a_1, a_2] \cup \cdots \cup (a_{n-1}, a_n],$$

where $a_0 = a$ and $a_n = b$. Let $A_k \subset \Omega$ denote the event $a_k < X \leq a_{k+1}$. Now write

$$I_n(X) = \sum_{k=1}^{n-1} a_k P(A_k).$$

The limit of $I_n(X)$ as $n \to \infty$ is called the *Lebesgue integral* of X and is denoted

$$\int_\Omega X(\omega)dP(\omega) = \lim_{n\to\infty} I_n(X).$$

Sometimes the notation $P(d\omega)$ is used for $dP(\omega)$.

If $f : \mathbb{R} \to \mathbb{R}$ is a Borel measurable (bounded) function and $X : \Omega \to \mathbb{R}$ is a random variable with probability law P_X, it is not too difficult to obtain from the definition that the expected value of $f \circ X$ is given by

$$E[f \circ X] := \int_\Omega f(X(\omega))dP(\omega) = \int_{-\infty}^{\infty} f(x)dP_X(x),$$

where the second integral is the Lebesgue integral on the set of values of X relative to the probability measure P_X.

We note the following remarks. First, what makes this definition much more general than Riemann's definition of integral is that we are freed from the limitation of using simple sets (intervals) to partition the domain of the function X. Here we partition Ω using measurable sets, on each of which X falls into a narrow interval in its image set.

Another feature of the Lebesgue integral is its notational convenience. Notice that Ω could have been continuous or discrete. In the continuous case, the integral may reduce to ordinary Riemann integral, whereas in the discrete case it reduces to a discrete sum. For example, if $\Omega = \{1, 2, \dots, 6\}^3$ as in the rolling of three dice example, and $X(\omega) = (i, j, k)$ is the outcome of a roll, then it can be shown that

$$\int_\Omega f(X(\omega))dP(\omega) = \frac{1}{216} \sum_{(i,j,k)\in\Omega} f(i, j, k).$$

5 Expected Value and Variance

Fix a probability space $(\Omega, \mathcal{F}, P)$. A random variable X is said to be *discrete* if, with probability one, it can take only countably many values. There is, there is a set $\{x_1, x_2, \dots\} \subset \mathbb{R}$ such that

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

X is a *continuous* random variable if there exists a function $f_X : \mathbb{R} \to [0, \infty)$ (not necessarily continuous) such that, for all x,

$$F_X(x) = \int_{-\infty}^{x} f_X(s)ds.$$

Definition 7 (Expectation and variance) *Let X be a random variable. Then its expectation is defined by*

$$\mu := E[X] := \begin{cases} \sum_{i=1}^{\infty} x_i P(X = x_i) & \textit{if } X \textit{ is discrete} \\ \\ \int_{-\infty}^{\infty} x f_X(x) dx & \textit{if } X \textit{ is continuous.} \end{cases}$$

For both cases, the variance of X is defined by

$$Var[X] := E[(X - \mu)^2].$$

Exercise 8 Fix a probability space $(\Omega, \mathcal{F}, P)$.

1. If X and Y are independent random variables, show that

$$E[XY] = E[X]E[Y].$$

2. If $X_1, \dots, X_n$ are independent random variables, show that

$$\mathrm{Var}[X_1 + \cdots + X_n] = \mathrm{Var}[X_1] + \cdots + \mathrm{Var}[X_n].$$

6 Stochastic Processes

Fix a probability space $(\Omega, \mathcal{F}, P)$. (It turns out that for most theoretical or applied purposes no loss of generality is incurred by assuming that this is $([0,1], \mathcal{B}, \lambda)$.)

Let S be an arbitrary set. (We will mostly think of it as a subset of $\mathbb{R}$, although this is not so essential.) By a *stochastic process with state space S* we will understand a family of random variables

$$X_t : \Omega \to S, \ t \in I,$$

where $I \subset \mathbb{R}$ is a set of real numbers which is often interpreted as a time interval or a discrete set of moments in time. Then X_t represents the state at which a given system is at time t.

In this course, we will study stochastic process of the following kinds:

1. *Countable parameter processes with discrete (finite or countably infinite) state space:*

$$I = \{0, 1, 2, \dots\}, \text{ and } S = \{s_1, s_2, \dots, s_k\} \text{ or } S = \{s_1, s_2, \dots\}.$$

   These are studied in Chapters 1 and 2 of the textbook.

2. *Continuous parameter process with discrete state space:*

$$I = [0, \infty) \text{ and } S \text{ is as in the first case.}$$

   These are the subject of chapter 3.

3. *Discrete parameter processes and continuous state space:*

$$I = \{0, 1, 2, \dots\} \text{ and } S = \mathbb{R}^n.$$

These aren't specifically discussed, but they do arise in the text.

4. *Continuous parameter processes and continuous state space:*

$$I = [0, \infty) \text{ and } S = \mathbb{R}^n.$$

Chapters 5 and 6 are mainly about these systems.

Note: If $X_t(\omega)$ is constant in ω, then the process is *deterministic*; there is no randomness involved in specifying the state of X_t.

For the most part we do not refer explicitly to the probability space $(\Omega, \mathcal{F}, P)$, but only to S and the probability measures (*laws*) of the random variables. Recall that if $X : \Omega \to S$ is a random variable, its probability measure is the measure P_X on S defined by

$$P_X(A) = P(\{\omega \in \Omega : X(\omega) \in A\}).$$

(In measure theory we sometimes say that P_X is the *push-forward* of the measure P by the measurable map X.)

Example 4 The dice rolling experiment discussed before is a discrete parameter stochastic process with finite state space, $S = \{1, 2, \dots, 6\}$. The probabilistic structure of this example is simplest we are going to consider since here $X_1, X_2, \dots$ are i.i.d. The probability distribution of the present state, X_n, is independent of the past or future states and unchanged in time. This is an example of a *Bernoulli process*.

7 Discrete Time Markov Chains

Consider a system with the following properties:

1. The system can be at any of a finite or countably infinite number of states. The set of states is $S = \{s_1, s_2, \dots\}$.

2. Starting in some initial state at time $t = 0$, the system changes its state randomly at times $t = 1, 2, \dots$. Representing the state at time k by the random variable X_k, the evolution of the system is described by the chain of random variables
$$X_0 \to X_1 \to X_2 \to \cdots$$

3. At time $t = 0$, the system occupies the state s_k with *initial probability*

$$\pi_0(s_k) = P(X_0 = s_k), \ k = 1, 2, \dots$$

4. Suppose that at any time n the the system has the following history:

$$X_i = x_i, \ i = 0, 1, 2, \ldots, n$$

where $x_n = s_i$ for some i. Then the probability that the system goes into the state $x_{n+1} = s_j \in S$ at the next time step is given by

$$P(X_{n+1} = s_j | X_0 = x_0, X_1 = x_1, \ldots, X_n = s_i).$$

We say that the random process satisfies the *Markov property* if this probability reduces to

$$p_{ij}(n) = P(X_{n+1} = s_j | X_n = s_i).$$

In other words, given the present state of the system, $X_n = s_i$, the future of the system is independent of the past. The numbers $p_{ij}(n)$ are called the *transition probabilities* of the system. If the Markov property holds and the transition probabilities do not depend on n, we say that the process is a *(time) homogeneous Markov chain*. We often omit the term "homogeneous."

8 Examples of Markov Chains

Example 5 (Weather forecast) There are only two possible weather conditions in Sunnyville: s_1 = "fair" and s_2 = "rainy." Empirical observation has shown that the best predictor of Sunnyville's weather tomorrow is today's weather, with the following transition probabilities:

today \ tomorrow	rain	fair
rain	0.75	0.25
fair	0.25	0.75

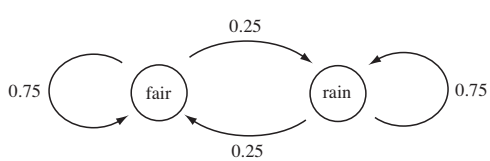

Figure 6: Transition probabilities for weather forecasting Markov chain.

In this case the transition probability matrix is $P = \begin{pmatrix} 0.75 & 0.25 \\ 0.25 & 0.75 \end{pmatrix}$.

Example 6 (Web surfing) Let $S = \{s_1, s_2, \ldots, s_n\}$ denote the set of all web pages on the internet. A page, $s_i \in S$, is linked to a number d_i of other pages. A simple probabilistic model of web surfing can be given by a Markov process, $X_0, X_1, X_2, \ldots$ such that

1. at time $t = 0$, $X_0 =$ "Renato's home page" with probability 1; and
2. if at time $t = k$ the surfer is at the web page s_i, then at time $t = k + 1$ he/she will have jumped to one of the d_i pages linked from s_i with equal probability, $1/d_i$. If a page s_i has no links from it we define $p_{ij} = 0$ for all $j \neq i$, and $p_{ii} = 1$. This is an example of a *random walk on a graph.* (More realistic models of web browsing have been proposed and studied.)

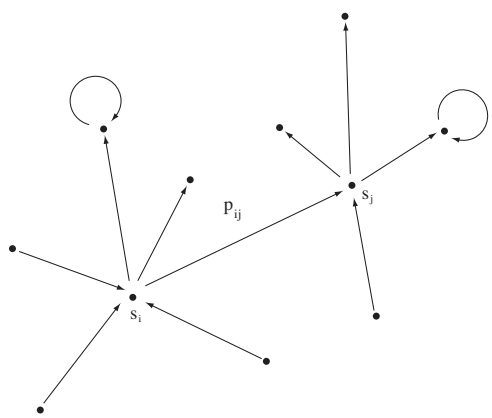

Figure 7: Random walk on the www. The figure shows a small region of the internet.

Example 7 (Shuffling a pile of books) Suppose m books are piled up on your desk. The books are labeled by the integers $1, 2, \ldots, m$. The order of the books from the top of the pile down to the bottom is described by a permutation $s = (i_1, i_2, \ldots, i_m)$, where i_1 is the number of the book on top of the pile, i_2 the second from the top, and so forth. Therefore, the set of possible states the book pile can be in is described by the set of all permutations of m elements. This is a set (a group in the algebraic sense) of $m!$ elements. To each book is associated a probability $p_i, i = 1, \ldots, m$ of being picked at any particular moment. After being picked, the book is returned to the top of the pile.

Let us look at the special case of three books: $m = 3$. We enumerate the states as follows: $s_1 = (1, 2, 3)$, $s_2 = (1, 3, 2)$, $s_3 = (2, 1, 3)$, $s_4 = (2, 3, 1)$, $s_5 = (3, 1, 2)$, $s_6 = (3, 2, 1)$. Having ordered the states of the pile, the transition probability matrix can now be written as follows:

$$P = \begin{pmatrix} p_1 & 0 & p_2 & 0 & p_3 & 0 \\ 0 & p_1 & p_2 & 0 & p_3 & 0 \\ p_1 & 0 & p_2 & 0 & 0 & p_3 \\ p_1 & 0 & 0 & p_2 & 0 & p_3 \\ 0 & p_1 & 0 & p_2 & p_3 & 0 \\ 0 & p_1 & 0 & p_2 & 0 & p_3 \end{pmatrix}.$$

The transition probabilities can also be represented in graphical form as shown in Figure 8.

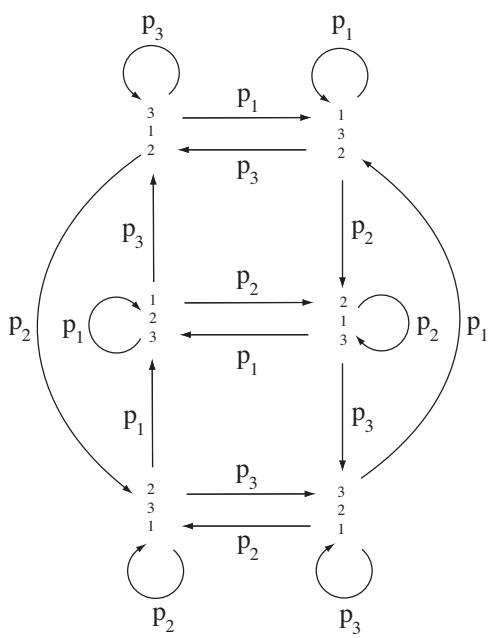

Figure 8: Transition probabilities for the "pile of books" example.

Although there is some risk of notational confusion, I will use P also to denote the transition probability matrix. Observe that any transition probability matrix, $P = (p_{ij})$, satisfies:

1. $p_{ij} \geq 0$ for all i, j in the index set of S; and
2. $p_{i1} + p_{i2} + \cdots = 1$ for each i.

Property 2 expresses the completeness and mutual exclusivity of the set of states:

$$\sum_{s \in S} P(X_{n+1} = s | X_n = s_i) = 1.$$

9 Multi-step Transition Probabilities

The probability structure of the discrete time (homogeneous) Markov chain is completely specified by giving:

1. the initial probability distribution vector (the distribution of X_0):

$$\pi_0 = (\pi_0(1), \mu_0(2), \dots),$$

   where $\pi_0(i) = P(X_0 = s_i)$; and

2. the transition probability matrix P (possibly infinite if there are infinitely many states), whose entries will be denoted $p_{ij} = P(X_{n+1} = s_j | X_n = s_j)$.

Proposition 1 (Probabilities after n steps) *Let $(X_0, X_1, \dots)$ be a homogeneous Markov chain with state space $\{s_1, \dots, s_k\}$, initial distribution vector π_0 and transition matrix P. Then for any n we have that the distribution $\pi_n = (\pi_n(1), \dots, \pi_n(k))$, where $\pi_n(i) = P(X_n = s_i)$, satisfies $\pi_n = \pi_0 P^n$.*

Proof. Consider first the case $n = 1$. We get, for $j = 1, \dots, k$, that

$$\begin{aligned}
\pi_1(j) &= P(X_1 = s_j) \\
&= \sum_{i=1}^{k} P(X_0 = s_i \text{ and } X_1 = s_j) \\
&= \sum_{i=1}^{k} P(X_0 = s_i) P(X_1 = s_j | X_0 = s_i) \\
&= \sum_{i=1}^{k} \pi_0(i) p_{ij} \\
&= (\pi_0 P)(j).
\end{aligned}$$

We now use induction. Fix m and suppose that $\pi_m = \pi_0 P^m$. Then, the probability distribution at time $m + 1$ is given by:

$$\begin{aligned}
\pi_{m+1}(j) &= P(X_{m+1} = s_j) \\
&= \sum_{i=1}^{k} P(X_m = s_i \text{ and } X_{m+1} = s_j) \\
&= \sum_{i=1}^{k} P(X_m = s_i) P(X_{m+1} = s_j | X_m = s_i) \\
&= \sum_{i=1}^{k} \pi_m(i) p_{ij} \\
&= (\pi_m P)(j) \\
&= (\pi_0 P^m P)(j) \\
&= (\pi_0 P^{m+1})(j).
\end{aligned}$$

Therefore, $\pi_{m+1} = \pi_0 P^{m+1}$, proving the claim. $\square$

Let us look at the example of weather forecasting in Sunyville. For the sake of greater generality I will assume that the transition probability matrix has the form

$$P = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}.$$

(In the weather forecast example, $p = q = 0.25$.) The initial probability distribution vector is $\pi_0 = (\pi_0(1), \pi_0(2))$, with $\pi_0(1) + \pi_0(2) = 1$. Then the probability vector of X_n is given by

$$(\pi_n(1), \pi_n(2)) = (\pi_0(1), \pi_0(2)) \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}^n.$$

We can obtain a very explicit description of the $\pi_n(i)$ by expressing the initial probability vector as linear combination of eigenvectors of the transition probability matrix. The eigenvalues of P are the solutions to

$$\det(P - \lambda) = \det \begin{pmatrix} 1-p-\lambda & p \\ q & 1-q-\lambda \end{pmatrix} = \lambda^2 - (1+a)\lambda + a = 0,$$

where $a = 1 - p - q$. This equation is easily solved for λ and the associated eigenvectors. The eigenvalues are: 1 and a; (q, p) is an eigenvector associated to 1 and $(1, -1)$ is eigenvector for a. In other words,

$$(q,p) \begin{pmatrix} 1-p-\lambda & p \\ q & 1-q-\lambda \end{pmatrix} = (q,p)$$
$$(1,-1) \begin{pmatrix} 1-p-\lambda & p \\ q & 1-q-\lambda \end{pmatrix} = a(1,-1).$$

The probability vector π_0 can now be written as a linear combination:

$$\pi_0 = \frac{1}{p+q}(q,p) + \frac{p\pi_0(1) - q\pi_0(2)}{p+q}(1,-1).$$

The probability vector at time n is $\pi_0 P^n$. Therefore,

$$(\pi_n(1), \pi_n(2)) = \frac{1}{p+q}(q,p) + \frac{p\pi_0(1) - q\pi_0(2)}{p+q} a^n (1,-1).$$

Note: $|a| = |1 - p - q| \leq 1$. If the system is not completely deterministic, i.e., not all probabilities are either 0 or 1, then $|a| < 1$. Consequently, $a^n \to 0$ and

$$(\pi_n(1), \pi_n(2)) \to \left(\frac{q}{p+q}, \frac{p}{p+q} \right) \quad \text{as } n \to \infty.$$

For the weather forecast example, $p = q = 0.25$. In this case, the best prediction we can make about the long term weather condition on the basis of our simple model ("the best predictor of tomorrow's weather is today's") is that

$$\pi_n(1) \to 0.5 \text{ as } n \to \infty.$$

10 Computer Simulation of Markov Chains

Suppose that a Markov chain with finite state space is given, with state space $S = \{s_1, \ldots, s_k\}$ and probability data (π_0, P), and we wish to run a computer simulation of the process $X_0, X_1, X_2, \ldots$.

We first need to have available a source of random numbers. This will take the form of a sequence of independent random variables $N_0, N_1, \ldots$ uniformly distributed over the interval $[0, 1]$. (In Matlab, for example, this is obtained by calling the command `rand` each time the next N_k is needed.)

Each run of the program should produce a sequence

$$x_0 \to x_1 \to x_2 \to \ldots$$

in such a way that x_0 is chosen from S according to the probability vector π_0 and each new x_{n+1} is chosen according to the transition probability matrix and the just obtained value of x_n.

To choose the value of X_0 we proceed as follows: Subdivide $[0, 1]$ into k intervals:

$$I_1 = [0, \pi_0(1)],\ I_2 = (\pi_0(1), \pi_0(1) + \pi_0(2)],\ \ldots,\ I_k = (\pi_0(1) + \cdots + \pi_0(k-1), 1].$$

The essential point here is that the length of I_i is $\pi_0(i)$. Now pick a value for N_0 and define:

$$X_0 = s_i \Leftrightarrow N_0 \in I_i.$$

Notice that, by this procedure, we have that $P(X_0 = s_i) = \lambda(I_i) = \pi_0(i)$.

Suppose that we have the values of $X_0, X_1, \ldots, X_{n-1}$ and wish to obtain the value of X_n. Say that $X_{n-1} = s_i$ and let p_{ij} be the transition probability to s_j. Define, for each i, the intervals

$$I_{i1} = [0, p_{i1}],\ I_{i2} = (p_{i1}, p_{i1} + p_{i2}],\ \ldots,\ I_{ik} = (p_{i1} + \cdots + p_{i,k-1}, 1].$$

Set $\mu_{ij} = p_{i1} + \cdots + p_{ij}$. Finally, choose X_n according to the rule:

$$X_n = s_j \Leftrightarrow N_n \in (\mu_{i,j-1}, \mu_{ij}].$$

As an example we implement this method for the Markov chain with initial probability vector

$$\pi_0 = (0.2, 0.5, 0.3)$$

and transition probability matrix

$$P = \begin{pmatrix} 0.8 & 0.1 & 0.1 \\ 0.3 & 0.4 & 0.3 \\ 0.4 & 0.1 & 0.5 \end{pmatrix}.$$

We remark that Matlab expressions such as `(N<=0.7)` or `(X==1)` give 1 if the expression in parenthesis is true, and 0 if not. The script below should be self-explanatory.

```
n=50000 %n+1 is the number of elements in the chain

Y=[]
%this initializes the vector that will contain in
%the end the string of numbers from 1, 2, 3
%which will be interpreted as a run (or trial) of
%length n+1 of the Markov chain.

N=rand;
X=1*(N<=0.2)+2*(N<=0.7 & N>0.2)+3*(0.7<N);
%this is the initial element, X0, of the Markov chain

Y=[Y X];

F=[];
%F will register, for each i=1, 2, ..., n, the frequency
%of occurrence of 1, 2, and 3
%in the first i entries of Y.

for i=1:n
    N=rand;
    a1=1*(N<=0.8)+2*(N<=0.9 & N>0.8)+3*(0.9<N);
    a2=1*(N<=0.3)+2*(N<=0.7 & N>0.3)+3*(0.7<N);
    a3=1*(N<=0.4)+2*(N<=0.5 & N>0.4)+3*(0.5<N);
    X=a1*(X==1)+a2*(X==2)+a3*(X==3);
    Y=[Y X];

    %we have now the string Y of values of the X0, X1, ...
    %for times 1, ..., i
    %and use it to calculate how often each state occurs:
    f1=sum(Y==1)/(i+1);
    f2=sum(Y==2)/(i+1);
    f3=sum(Y==3)/(i+1);
    %f1 is the frequency of occurrence of 1,
    %f2 of 2, and f3 of 3.
    F=[F; f1 f2 f3];
end
```

One run the above script gave the values $f_1 = 0.641$, $f_2 = 0.146$, and $f_3 = 0.213$. You should check as an exercise that the transition probabilities matrix P of this example has a unique (after normalization so as to make the sum of components equal to 1) eigenvector associated to eigenvalue one, which is given by $u = (0.643, 0.143, 0.214)$. (Somehow, I expected a better agreement.) The first 10 elements of a typical run of the chain is

```
[3 3 2 1 1 1 1 1 2 2].
```

The graph of the frequency of $1, 2, 3$ for the partial sums up to $i = 1, \ldots, 50000$ versus the time i is given below.

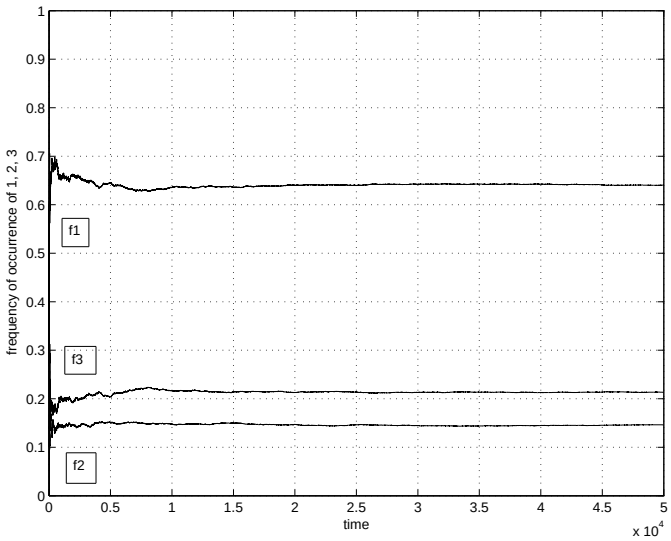

Figure 9: Frequency of occurrence of states $1, 2, 3$.