Chapter 1

Basic notions

1.1 A few definitions from measure theory

Given a set $X$, a $\sigma$-algebra on $X$ is a collection $\mathcal{A}$ of subsets of $X$ such that
(1) $\emptyset \in \mathcal{A}$;
(2) if $A \in \mathcal{A}$, then $A^c \in \mathcal{A}$, where $A^c = X \setminus A$;
(3) if $A_1, A_2, \ldots \in \mathcal{A}$, then $\bigcup_{i=1}^{\infty} A_i$ and $\bigcap_{i=1}^{\infty} A_i$ are both in $\mathcal{A}$.

A measure on a pair $(X, \mathcal{A})$ is a function $\mu : \mathcal{A} \to [0, \infty]$ such that
(1) $\mu(\emptyset) = 0$;
(2) $\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i)$ whenever the $A_i$ are in $\mathcal{A}$ and are pairwise disjoint.

A function $f : X \to \mathbb{R}$ is measurable if the set $\{x : f(x) > a\} \in \mathcal{A}$ for all $a \in \mathbb{R}$.

A property holds almost everywhere, written “a.e.,” if the set where it fails has measure 0. For example, $f = g$ a.e. if $\mu(\{x : f(x) \neq g(x)\}) = 0$.

The characteristic function $\chi_A$ of a set in $\mathcal{A}$ is the function that is 1 when $x \in A$ and is zero otherwise. A function of the form $\sum_{i=1}^{n} a_i \chi_{A_i}$ is called a simple function.

If $f$ is simple and of the above form, we define

$$\int f \, d\mu = \sum_{i=1}^{n} a_i \mu(A_i).$$
If $f$ is non-negative and measurable, we define
\[
\int f \, d\mu = \sup \left\{ \int s \, d\mu : 0 \leq s \leq f, s \text{ simple} \right\}.
\]
Provided $\int |f| \, d\mu < \infty$, we define
\[
\int f \, d\mu = \int f^+ \, d\mu - \int f^- \, d\mu,
\]
where $f^+ = \max(f, 0)$, $f^- = \max(-f, 0)$.

### 1.2 Definitions

A probability or probability measure is a measure whose total mass is one. Because the origins of probability are in statistics rather than analysis, some of the terminology is different. For example, instead of denoting a measure space by $(X, \mathcal{A}, \mu)$, probabilists use $(\Omega, \mathcal{F}, \mathbb{P})$. So here $\Omega$ is a set, $\mathcal{F}$ is called a \(\sigma\)-field (which is the same thing as a \(\sigma\)-algebra), and $\mathbb{P}$ is a measure with $\mathbb{P}(\Omega) = 1$. Elements of $\mathcal{F}$ are called events. Elements of $\Omega$ are denoted $\omega$.

Instead of saying a property occurs almost everywhere, we talk about properties occurring almost surely, written a.s.. Real-valued measurable functions from $\Omega$ to $\mathbb{R}$ are called random variables and are usually denoted by $X$ or $Y$ or other capital letters. We often abbreviate ”random variable” by r.v.

We let $A^c = (\omega \in \Omega : \omega \notin A)$ (called the complement of $A$) and $B - A = B \cap A^c$.

Integration (in the sense of Lebesgue) is called expectation or expected value, and we write $\mathbb{E}X$ for $\int X \, d\mathbb{P}$. The notation $\mathbb{E}[X; A]$ is often used for $\int_A X \, d\mathbb{P}$.

The random variable $1_A$ is the function that is one if $\omega \in A$ and zero otherwise. It is called the indicator of $A$ (the name characteristic function in probability refers to the Fourier transform). Events such as $(\omega : X(\omega) > a)$ are almost always abbreviated by $(X > a)$.

Given a random variable $X$, we can define a probability on $\mathbb{R}$ by
\[
\mathbb{P}_X(A) = \mathbb{P}(X \in A), \quad A \subset \mathbb{R}.
\] (1.1)
1.2. DEFINITIONS

The probability \( P_X \) is called the law of \( X \) or the distribution of \( X \). We define \( F_X : \mathbb{R} \to [0, 1] \) by

\[
F_X(x) = P_X((\leq x]) = P(X \leq x).
\] (1.2)

The function \( F_X \) is called the distribution function of \( X \).

As an example, let \( \Omega = \{H, T\} \), \( \mathcal{F} \) all subsets of \( \Omega \) (there are 4 of them), \( \mathbb{P}(H) = \mathbb{P}(T) = \frac{1}{2} \). Let \( X(H) = 1 \) and \( X(T) = 0 \). Then \( F_X = \frac{1}{2} \delta_0 + \frac{1}{2} \delta_1 \), where \( \delta_x \) is point mass at \( x \), that is, \( \delta_x(A) = 1 \) if \( x \in A \) and 0 otherwise. \( F_X(a) = 0 \) if \( a < 0 \), \( \frac{1}{2} \) if \( 0 \leq a < 1 \), and 1 if \( a \geq 1 \).

**Proposition 1.1** The distribution function \( F_X \) of a random variable \( X \) satisfies:

(a) \( F_X \) is nondecreasing;
(b) \( F_X \) is right continuous with left limits;
(c) \( \lim_{x \to \infty} F_X(x) = 1 \) and \( \lim_{x \to -\infty} F_X(x) = 0 \).

**Proof.** We prove the first part of (b) and leave the others to the reader. If \( x_n \downarrow x \), then \( (X \leq x_n) \downarrow (X \leq x) \), and so \( \mathbb{P}(X \leq x_n) \downarrow \mathbb{P}(X \leq x) \) since \( \mathbb{P} \) is a measure. \( \square \)

Note that if \( x_n \uparrow x \), then \( (X \leq x_n) \uparrow (X < x) \), and so \( F_X(x_n) \uparrow \mathbb{P}(X < x) \). Any function \( F : \mathbb{R} \to [0, 1] \) satisfying (a)-(c) of Proposition 1.1 is called a distribution function, whether or not it comes from a random variable.

**Proposition 1.2** Suppose \( F \) is a distribution function. There exists a random variable \( X \) such that \( F = F_X \).

**Proof.** Let \( \Omega = [0, 1] \), \( \mathcal{F} \) the Borel \( \sigma \)-field, and \( \mathbb{P} \) Lebesgue measure. Define \( X(\omega) = \sup\{x : F(x) < \omega\} \). Here the Borel \( \sigma \)-field is the smallest \( \sigma \)-field containing all the open sets.

We check that \( F_X = F \). Suppose \( X(\omega) \leq y \). Then \( F(z) \geq \omega \) if \( z > y \). By the right continuity of \( F \) we have \( F(y) \geq \omega \). Hence \( (X(\omega) \leq y) \subset (\omega \leq F(y)) \).
Suppose \( \omega \leq F(y) \). If \( X(\omega) > y \), then by the definition of \( X \) there exists \( z > y \) such that \( F(z) < \omega \). But then \( \omega \leq F(y) \leq F(z) \), a contradiction. Therefore \( (\omega \leq F(y)) \subset (X(\omega) \leq y) \).

We then have
\[
\mathbb{P}(X(\omega) \leq y) = \mathbb{P}(\omega \leq F(y)) = F(y).
\]

In the above proof, essentially \( X = F^{-1} \). However \( F \) may have jumps or be constant over some intervals, so some care is needed in defining \( X \).

Certain distributions or laws are very common. We list some of them.

(a) **Bernoulli.** A random variable is Bernoulli if \( \mathbb{P}(X = 1) = p, \mathbb{P}(X = 0) = 1 - p \) for some \( p \in [0, 1] \).

(b) **Binomial.** This is defined by \( \mathbb{P}(X = k) = \binom{n}{k} p^k (1 - p)^{n-k} \), where \( n \) is a positive integer, \( 0 \leq k \leq n \), and \( p \in [0, 1] \).

(c) **Geometric.** For \( p \in (0, 1) \) we set \( \mathbb{P}(X = k) = (1 - p)p^k \). Here \( k \) is a nonnegative integer.

(d) **Poisson.** For \( \lambda > 0 \) we set \( \mathbb{P}(X = k) = e^{-\lambda} \lambda^k / k! \). Again \( k \) is a nonnegative integer.

(e) **Uniform.** For some positive integer \( n \), set \( \mathbb{P}(X = k) = 1/n \) for \( 1 \leq k \leq n \).

Suppose \( F \) is absolutely continuous. This is not the definition, but being absolutely continuous is equivalent to the existence of a function \( f \) such that
\[
F(x) = \int_{-\infty}^{x} f(y) \, dy
\]
for all \( x \). We call \( f = F' \) the density of \( F \). Some examples of distributions characterized by densities are the following.

(f) **Uniform on \([a, b]\).** Define \( f(x) = (b - a)^{-1} 1_{[a,b]}(x) \). This means that if \( X \) has a uniform distribution, then
\[
\mathbb{P}(X \in A) = \int_{A} \frac{1}{b - a} 1_{[a,b]}(x) \, dx.
\]
1.3. SOME BASIC FACTS

(g) Exponential. For \( x > 0 \) let \( f(x) = \lambda e^{-\lambda x} \).

(h) Standard normal. Define \( f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \). So

\[
P(X \in A) = \frac{1}{\sqrt{2\pi}} \int_A e^{-x^2/2} \, dx.
\]

(i) \( \mathcal{N}(\mu, \sigma^2) \). We shall see later that a standard normal has mean zero and variance one. If \( Z \) is a standard normal, then a \( \mathcal{N}(\mu, \sigma^2) \) random variable has the same distribution as \( \mu + \sigma Z \). It is an exercise in calculus to check that such a random variable has density

\[
\frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}.
\]

(j) Cauchy. Here

\[
f(x) = \frac{1}{\pi} \frac{1}{1 + x^2}.
\]

1.3 Some basic facts

We can use the law of a random variable to calculate expectations.

**Proposition 1.3** If \( g \) is nonnegative or if \( \mathbb{E} |g(X)| < \infty \), then

\[
\mathbb{E} g(X) = \int g(x) \, \mathbb{P}_X(dx).
\]

**Proof.** If \( g \) is the indicator of an event \( A \), this is just the definition of \( \mathbb{P}_X \). By linearity, the result holds for simple functions. By the monotone convergence theorem, the result holds for nonnegative functions, and by writing \( g = g^+ - g^- \), it holds for \( g \) such that \( \mathbb{E} |g(X)| < \infty \). \( \square \)

If \( F_X \) has a density \( f \), then \( \mathbb{P}_X(dx) = f(x) \, dx \). So, for example, \( \mathbb{E} X = \int x f(x) \, dx \) and \( \mathbb{E} X^2 = \int x^2 f(x) \, dx \). (We need \( \mathbb{E} |X| \) finite to justify this if \( X \) is not necessarily nonnegative.) We define the *mean* of a random variable to be its expectation, and the variance of a random variable is defined by

\[
\text{Var} \, X = \mathbb{E} (X - \mathbb{E} X)^2.
\]
For example, it is routine to see that the mean of a standard normal is zero and its variance is one.

Note

\[ \text{Var } X = \mathbb{E}(X^2 - 2\mathbb{E}X + (\mathbb{E}X)^2) = \mathbb{E}X^2 - (\mathbb{E}X)^2. \]

Another equality that is useful is the following.

**Proposition 1.4** If \( X \geq 0 \) a.s. and \( p > 0 \), then

\[ \mathbb{E}X^p = \int_0^\infty p \lambda^{p-1} \mathbb{P}(X > \lambda) \, d\lambda. \]

The proof will show that this equality is also valid if we replace \( \mathbb{P}(X > \lambda) \) by \( \mathbb{P}(X \geq \lambda) \).

**Proof.** Use Fubini's theorem and write

\[
\int_0^\infty p \lambda^{p-1} \mathbb{P}(X > \lambda) \, d\lambda = \mathbb{E} \int_0^\infty p \lambda^{p-1} 1_{(\lambda,\infty)}(X) \, d\lambda
= \mathbb{E} \int_0^X p \lambda^{p-1} \, d\lambda = \mathbb{E}X^p.
\]

We need two elementary inequalities.

**Proposition 1.5** *Chebyshev’s inequality* If \( X \geq 0 \),

\[ \mathbb{P}(X \geq a) \leq \frac{\mathbb{E}X}{a}. \]

**Proof.** We write

\[ \mathbb{P}(X \geq a) = \mathbb{E}\left[ 1_{[a,\infty)}(X) \right] \leq \mathbb{E}\left[ \frac{X}{a} 1_{[a,\infty)}(X) \right] \leq \mathbb{E}X/a, \]

since \( X/a \) is bigger than 1 when \( X \in [a, \infty) \).
1.3. SOME BASIC FACTS

If we apply this to \( X = (Y - \mathbb{E}Y)^2 \), we obtain

\[
\mathbb{P}(|Y - \mathbb{E}Y| \geq a) = \mathbb{P}((Y - \mathbb{E}Y)^2 \geq a^2) \leq \text{Var} Y/a^2.
\] (1.4)

This special case of Chebyshev's inequality is sometimes itself referred to as Chebyshev's inequality, while Proposition 1.5 is sometimes called the Markov inequality.

The second inequality we need is Jensen's inequality, not to be confused with the Jensen’s formula of complex analysis.

**Proposition 1.6** Suppose \( g \) is convex and and \( X \) and \( g(X) \) are both integrable. Then

\[
g(\mathbb{E}X) \leq \mathbb{E}g(X).
\]

**Proof.** One property of convex functions is that they lie above their tangent lines, and more generally their support lines. So if \( x_0 \in \mathbb{R} \), we have

\[
g(x) \geq g(x_0) + c(x - x_0)
\]

for some constant \( c \). Take \( x = X(\omega) \) and take expectations to obtain

\[
\mathbb{E}g(X) \geq g(x_0) + c(\mathbb{E}X - x_0).
\]

Now set \( x_0 \) equal to \( \mathbb{E}X \). \( \square \)

If \( A_n \) is a sequence of sets, define \((A_n \text{ i.o.})\), read "\( A_n \) infinitely often," by

\[
(A_n \text{ i.o.}) = \cap_{n=1}^{\infty} \cup_{i=n}^{\infty} A_i.
\]

This set consists of those \( \omega \) that are in infinitely many of the \( A_n \).

A simple but very important proposition is the Borel-Cantelli lemma. It has two parts, and we prove the first part here, leaving the second part to the next section.

**Proposition 1.7** (Borel-Cantelli lemma) \( \) If \( \sum_n \mathbb{P}(A_n) < \infty \), then \( \mathbb{P}(A_n \text{ i.o.}) = 0 \).
Proof. We have

$$\mathbb{P}(A_n \text{ i.o.}) = \lim_{n \to \infty} \mathbb{P}(\bigcup_{i=n}^{\infty} A_i).$$

However,

$$\mathbb{P}(\bigcup_{i=n}^{\infty} A_i) \leq \sum_{i=n}^{\infty} \mathbb{P}(A_i),$$

which tends to zero as $n \to \infty$. \qed

1.4 Independence

Let us say two events $A$ and $B$ are independent if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. The events $A_1, \ldots, A_n$ are independent if

$$\mathbb{P}(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_j}) = \mathbb{P}(A_{i_1})\mathbb{P}(A_{i_2}) \cdots \mathbb{P}(A_{i_j})$$

for every subset $\{i_1, \ldots, i_j\}$ of $\{1, 2, \ldots, n\}$.

Proposition 1.8 If $A$ and $B$ are independent, then $A^c$ and $B$ are independent.

Proof. We write

$$\mathbb{P}(A^c \cap B) = \mathbb{P}(B) - \mathbb{P}(A \cap B) = \mathbb{P}(B) - \mathbb{P}(A)\mathbb{P}(B)$$

$$= \mathbb{P}(B)(1 - \mathbb{P}(A)) = \mathbb{P}(B)\mathbb{P}(A^c).$$

\qed

We say two $\sigma$-fields $\mathcal{F}$ and $\mathcal{G}$ are independent if $A$ and $B$ are independent whenever $A \in \mathcal{F}$ and $B \in \mathcal{G}$. The $\sigma$-field generated by a random variable $X$, written $\sigma(X)$, is given by $\{(X \in A) : A \text{ a Borel subset of } \mathbb{R}\}$. Two random variables $X$ and $Y$ are independent if the $\sigma$-field generated by $X$ and the $\sigma$-field generated by $Y$ are independent. We define the independence of $n$ $\sigma$-fields or $n$ random variables in the obvious way.

Proposition 1.8 tells us that $A$ and $B$ are independent if the random variables $1_A$ and $1_B$ are independent, so the definitions above are consistent.
If \( f \) and \( g \) are Borel functions and \( X \) and \( Y \) are independent, then \( f(X) \) and \( g(Y) \) are independent. This follows because the \( \sigma \)-field generated by \( f(X) \) is a sub-\( \sigma \)-field of the one generated by \( X \), and similarly for \( g(Y) \).

Let \( F_{X,Y}(x,y) = P(X \leq x, Y \leq y) \) denote the joint distribution function of \( X \) and \( Y \). (The comma inside the set means ”and.”)

**Proposition 1.9** \( F_{X,Y}(x,y) = F_X(x)F_Y(y) \) if and only if \( X \) and \( Y \) are independent.

**Proof.** If \( X \) and \( Y \) are independent, the \( 1_{(-\infty,x]}(X) \) and \( 1_{(-\infty,y]}(Y) \) are independent by the above comments. Using the above comments and the definition of independence, this shows \( F_{X,Y}(x,y) = F_X(x)F_Y(y) \).

Conversely, if the inequality holds, fix \( y \) and let \( \mathcal{M}_y \) denote the collection of sets \( A \) for which \( P(X \in A, Y \leq y) = P(X \in A)P(Y \leq y) \). \( \mathcal{M}_y \) contains all sets of the form \( (-\infty,x] \). It follows by linearity that \( \mathcal{M}_y \) contains all sets of the form \( (x,z] \), and then by linearity again, by all sets that are the finite union of such half-open, half-closed intervals. Note that the collection of finite unions of such intervals, \( \mathcal{A} \), is an algebra generating the Borel \( \sigma \)-field. It is clear that \( \mathcal{M}_y \) is a monotone class, so by the monotone class lemma, \( \mathcal{M}_y \) contains the Borel \( \sigma \)-field.

For a fixed set \( A \), let \( \mathcal{M}_A \) denote the collection of sets \( B \) for which \( P(X \in A, Y \in B) = P(X \in A)P(Y \in B) \). Again, \( \mathcal{M}_A \) is a monotone class and by the preceding paragraph contains the \( \sigma \)-field generated by the collection of finite unions of intervals of the form \( (x,z] \), hence contains the Borel sets. Therefore \( X \) and \( Y \) are independent. \( \square \)

The following is known as the multiplication theorem.

**Proposition 1.10** If \( X \), \( Y \), and \( XY \) are integrable and \( X \) and \( Y \) are independent, then \( \mathbb{E}[XY] = (\mathbb{E}X)(\mathbb{E}Y) \).

**Proof.** Consider the random variables in \( \sigma(X) \) (the \( \sigma \)-field generated by \( X \)) and \( \sigma(Y) \) for which the multiplication theorem is true. It holds for indicators by the definition of \( X \) and \( Y \) being independent. It holds for simple random variables, that is, linear combinations of indicators, by linearity of both sides.
It holds for nonnegative random variables by monotone convergence. And it holds for integrable random variables by linearity again.

If \( X_1, \ldots, X_n \) are independent, then so are \( X_1 - \mathbb{E} X_1, \ldots, X_n - \mathbb{E} X_n \). Assuming everything is integrable,

\[
\mathbb{E} [(X_1 - \mathbb{E} X_1) + \cdots (X_n - \mathbb{E} X_n)]^2 = \mathbb{E} (X_1 - \mathbb{E} X_1)^2 + \cdots + \mathbb{E} (X_n - \mathbb{E} X_n)^2,
\]

using the multiplication theorem to show that the expectations of the cross product terms are zero. We have thus shown

\[
\text{Var} (X_1 + \cdots + X_n) = \text{Var} X_1 + \cdots + \text{Var} X_n. \tag{1.5}
\]

We finish up this section by proving the second half of the Borel-Cantelli lemma.

**Proposition 1.11** Suppose \( A_n \) is a sequence of independent events. If we have \( \sum_n \mathbb{P}(A_n) = \infty \), then \( \mathbb{P}(A_n \text{ i.o.}) = 1 \).

Note that here the \( A_n \) are independent, while in the first half of the Borel-Cantelli lemma no such assumption was necessary.

**Proof.** Note

\[
\mathbb{P}(\bigcup_{i=n}^N A_i) = 1 - \mathbb{P}(\bigcap_{i=n}^N A_i^c) = 1 - \prod_{i=n}^N \mathbb{P}(A_i^c) = 1 - \prod_{i=n}^N (1 - \mathbb{P}(A_i)).
\]

By the mean value theorem, \( 1 - x \leq e^{-x} \), so we have that the right hand side is greater than or equal to \( 1 - \exp(-\sum_{i=n}^N \mathbb{P}(A_i)) \). As \( N \to \infty \), this tends to 1, so \( \mathbb{P}(\bigcup_{i=n}^\infty A_i) = 1 \). This holds for all \( n \), which proves the result. \( \square \)