8

Integrating complex functions,
Fourier theory and related topics

The intent of this short chapter is to indicate how the previous theory may be extended in an obvious way to include the integration of complex-valued functions with respect to a measure (or signed measure) $\mu$ on a measurable space $(X, S)$. The primary purpose of this is to discuss Fourier and related transforms which are important in a wide variety of contexts – and in particular the Chapter 12 discussion of characteristic functions of random variables which provide a standard and useful tool in summarizing their probabilistic properties.

Some standard inversion theorems will be proved here to help avoid overload of the Chapter 12 material. However, methods of this chapter also apply to other diverse applications e.g. to Laplace and related transforms used in fields such as physics as well as in probabilistic areas such as stochastic modeling, and may be useful for reference.

Finally it might be emphasized (as noted later) that the integrals considered here involve complex functions as integrands and as for the preceding development, form a “Lebesgue-style” theory. This is in contrast to what is termed “complex variable” methodology, which is a “Riemann-style” theory in which integrals are considered with respect to a complex variable $z$ along some curve in the complex plane. The latter methods – not considered here – can be especially useful in providing means for evaluation of integrals such as characteristic functions which may resist simple real variable techniques.

8.1 Integration of complex functions

Let $(X, S, \mu)$ be a measure space and $f$ a complex-valued function defined on $X$ with real and imaginary parts $u, v$:

$$f(x) = u(x) + iv(x).$$

$f$ is said to be measurable if $u$ and $v$ are measurable functions.
We say $f \in L_1(X, S, \mu)$ if $u$ and $v$ both belong to $L_1(X, S, \mu)$ and write

$$\int f \, d\mu = \int u \, d\mu + i \int v \, d\mu.$$ 

As noted above this is not integration with respect to a complex variable here, i.e. we are not considering contour integrals. The integral involves a complex-valued function, integrated with respect to a (real) measure on $(X, S)$.

Many properties of integrals of real functions hold in the complex case also. Some of the most elementary and obvious ones are given in the following theorem.

**Theorem 8.1.1** Let $(X, S, \mu)$ be a measure space and write $L_1 = L_1(X, S, \mu)$. Let $f$ be a complex measurable function on $X$, $f = u + iv$. Then

(i) $f \in L_1$ if and only if $|f| = (u^2 + v^2)^{1/2} \in L_1$.

(ii) If $f, g \in L_1$, $\alpha, \beta$ complex, then $\alpha f + \beta g \in L_1$ and

$$\int (\alpha f + \beta g) \, d\mu = \alpha \int f \, d\mu + \beta \int g \, d\mu.$$ 

(iii) If $f \in L_1$ then $|\int f \, d\mu| \leq \int |f| \, d\mu$.

**Proof**  
(i) Measurability of $|f|$ follows from that of $u, v$. Also it is easily checked that $|u|, |v| \leq |f| = (u^2 + v^2)^{1/2} \leq |u| + |v|$ from which (i) follows in both directions.

(ii) is easily checked by expressing $f, g, \alpha, \beta$ in terms of their real and imaginary parts and applying the corresponding result for real functions.

(iii) is perhaps slightly more involved to show directly than one might imagine. Write $z = \int f \, d\mu$ and $z = re^{i\theta}$. Then

$$|\int f \, d\mu| = r = e^{-i\theta}z = e^{-i\theta}\int f \, d\mu = \int (e^{-i\theta}f) \, d\mu.$$ 

But since this is real, the imaginary part of the integral must vanish, giving

$$|\int f \, d\mu| = \int \Re[e^{-i\theta}f] \, d\mu \quad (\Re \text{ denoting “real part”})$$

$$\leq \int |e^{-i\theta}f| \, d\mu$$

$$= \int |f| \, d\mu$$

as required. □

Many of the simple results for real functions will be used for complex functions with little if any comment, in view of their obvious nature – e.g. Theorems 4.4.3, 4.4.6, 4.4.8, 4.4.9. Of course some results (e.g. Theorem 4.4.4) simply have no immediate generalization to complex functions.

For the most part the more important and sophisticated theorems also generalize in cases where the generalized statements have meaning. This
is the case for Fubini’s Theorem for $L_1$-functions (Theorem 7.4.2 (iii)), the “Transformation Theorem” (Theorem 4.6.1), Dominated Convergence (Theorem 4.5.5) and the uses of the Radon–Nikodym Theorem such as Theorem 5.6.1 (for complex integrable functions). It may be checked that these results follow from the real counterparts. As an example we prove the dominated convergence theorem in the complex setting.

**Theorem 8.1.2** (Dominated Convergence for complex sequences) Let $\{f_n\}$ be a sequence of complex-valued functions in $L_1(X, S, \mu)$ such that $|f_n| \leq |g|$ a.e. where $g \in L_1$. Let $f$ be a complex measurable function such that $f_n \to f$ a.e. Then $f \in L_1$ and $\int|f_n - f| \, d\mu \to 0$. In particular $\int f_n \, d\mu \to \int f \, d\mu$.

**Proof** Write $f_n = u_n + iv_n, f = u + iv$. Since $f_n \to f$ a.e. it follows that $u_n \to u, v_n \to v$ a.e. Also $|u_n| \leq |g|, |v_n| \leq |g|$. Hence $u, v \in L_1$ by Theorem 4.5.5 (hence $f \in L_1$), and

$$\int|u_n - u| \, d\mu \to 0, \quad \int|v_n - v| \, d\mu \to 0.$$ 

Thus

$$\int|(u_n + iv_n) - (u + iv)| \, d\mu \leq \int(|u_n - u| + |v_n - v|) \, d\mu \to 0$$

or $\int|f_n - f| \, d\mu \to 0$ as required. Finally

$$|\int f_n \, d\mu - \int f \, d\mu| = |\int(f_n - f) \, d\mu| \leq \int|f_n - f| \, d\mu$$

by Theorem 8.1.1 and thus the final statement follows. \qed

We conclude this section with some comments concerning $L_p$-spaces of complex functions, and the Hölder and Minkowski Inequalities.

As for real functions, if $f$ is complex and measurable we define $\|f\|_p = (\int|f|^p \, d\mu)^{1/p}$ for $p > 0$ and say that $f \in L_p$ if $\|f\|_p < \infty$. Clearly such (complex, measurable) $f \in L_p$ if and only if $|f| \in L_p$, i.e. $|f|^p \in L_1$. It is also easily checked that if $f = u + iv$, then $f \in L_p$ if and only if each of $u, v$ are in $L_p$. (For if $f \in L_p, |u|^p \leq |f|^p \in L_1$, whereas if $u, v \in L_p$ then $|u| + |v| \in L_p$ and $|f|^p \leq (|u| + |v|)^p \in L_1$.)

Further if $f, g$ are complex functions in $L_p$, it is readily seen that $f + g \in L_p$ and hence $\alpha f + \beta g \in L_p$ for any complex $\alpha, \beta$. For $|f|, |g|$ are real functions in $L_p$ and hence $|f| + |g| \in L_p$, so that $|f + g| \leq (|f| + |g|) \in L_p$ showing that $|f + g|^p \in L_1$ and hence $f + g \in L_p$.

Hölder’s Inequality generalizes verbatim for complex integrands, since if $f \in L_p, g \in L_q$ for some $p \geq 1, q \geq 1, 1/p + 1/q = 1$, then $|f| \in L_p, |g| \in L_q$ so that $|fg| \in L_1$ by Theorem 6.4.2 and

$$\int|fg| \, d\mu = \int|f||g| \, d\mu \leq (\int|f|^p \, d\mu)^{1/p}(\int|g|^q \, d\mu)^{1/q}.$$
Armed with Hölder’s Inequality, Minkowski’s Inequality follows by the same proof as in the real case.

The complex \(L_p\)-space may be discussed in the same manner as the real \(L_p\)-space (cf. Section 6.4). This is a linear space (over the complex field) and is normed by \(\|f\|_p = (\int |f|^p d\mu)^{1/p} (p \geq 1)\). It is easily checked that if \(f_n \to f\) in \(L_p\) (i.e. \(\|f_n - f\| \to 0\)) and if \(f_n = u_n + iv_n, f = u + iv,\) then \(u_n \to u, v_n \to v\) in \(L_p\), and conversely (e.g. \(|u_n - u|^p \leq |f_n - f|^p\) and hence \(\|u_n - u\| \leq \|f_n - f\|\), whereas also \(\|f_n - f\| \leq \|u_n - u\| + \|v_n - v\|\)). Using these facts, completeness of \(L_p\) follows from the results for the real case. As for the real case \(L_p\) is a complete metric space for \(0 < p < 1\) (Theorem 6.4.7).

8.2 Fourier–Stieltjes, and Fourier Transforms in \(L_1\)

Suppose that \(F\) is a real bounded, nondecreasing function (assumed right-continuous, for convenience) on the real line \(\mathbb{R}\) and defining the measure \(\mu_F\). The Fourier–Stieltjes Transform \(F^*(t)\) of \(F\) is defined as a complex function on \(\mathbb{R}\) by

\[
F^*(t) = \int_{-\infty}^{\infty} e^{itx} dF(x) = \int e^{itx} d\mu_F.
\]

This integral exists since \(|e^{itx}| = 1\) and \(\mu_F(\mathbb{R}) < \infty\).

A function \(F\) on \(\mathbb{R}\) is of bounded variation (b.v.) on \(\mathbb{R}\) (cf. Section 5.7 for finite ranges) if it can be expressed as the difference of two bounded nondecreasing functions, \(F = F_1 - F_2\) (again assume \(F_1, F_2\) to be right-continuous for convenience). If \(F\) is b.v. its Fourier–Stieltjes Transform is defined as

\[
F^*(t) = F_1^*(t) - F_2^*(t).
\]

(Note that this definition is unambiguous since if also \(F = G_1 - G_2\) then \(G_1 + F_2 = G_2 + F_1\), and it is readily checked that \(G_1^* + F_2^* = F_1^* + F_1^*\), giving \(G_1^* - G_2^* = F_1^* - F_2^*\).)

**Theorem 8.2.1** If \(F\) is b.v., its Fourier–Stieltjes Transform \(F^*(t)\) is uniformly continuous on \(\mathbb{R}\).

**Proof** Suppose \(F\) is nondecreasing. For any real \(t, s, t - s = h,\)

\[
|F^*(t) - F^*(s)| = |\int (e^{itx} - e^{isx}) dF(x)|
\]

\[
\leq \int |e^{itx} - e^{isx}| |dF(x)|
\]

\[
= \int |e^{ihx} - 1| |dF(x)|.
\]

As \(h \to 0, |e^{ihx} - 1| \to 0\) and is bounded by \(|e^{ihx}| + 1 = 2\) which is \(dF\)-integrable. Hence by Dominated Convergence (Theorem 8.1.2)
\[ \int |e^{ihx} - 1| dF(x) \to 0 \text{ as } h \to 0 \] (through any sequence and hence generally). Thus given \( \epsilon > 0 \) there exists \( \delta > 0 \) such that \( \int |e^{ihx} - 1| dF(x) < \epsilon \) if \( |h| < \delta \). Then \( |F^s(t) - F^s(s)| < \epsilon \) for all \( t, s \) such that \( |t - s| < \delta \), which proves uniform continuity. If \( F \) is b.v. the result follows by writing \( F = F_1 - F_2 \). \( \square \)

Suppose now that \( f \) is a real Lebesgue measurable function on \( \mathbb{R} \) and \( f \in L_1 = L_1(-\infty, \infty) \) (Lebesgue measure). Then \( f(x)e^{itx} \in L_1 \) for all real \( t \), and we define the \( L_1 \) Fourier Transform \( f^\mathbb{F} \) of \( f \) by

\[
 f^\mathbb{F}(t) = \int_{-\infty}^{\infty} e^{itx} f(x) \, dx.
\]

First note that \( f, g \in L_1 \) then \( (\alpha f + \beta g)^\mathbb{F} = \alpha f^\mathbb{F} + \beta g^\mathbb{F} \) for any real constants \( \alpha, \beta \).

It is also immediate that \( f^\mathbb{F}(t) = F^s(t) \) where \( F(x) = \int_{-\infty}^{x} f(u) \, du \). For if \( f \) is nonnegative, \( F \) is then nondecreasing and

\[
 F^s(t) = \int e^{itx} dF(x) = \int e^{itx} f(x) \, dx
\]

by Theorem 5.6.1. The general case follows by writing \( f = f_+ - f_- \), \( F_1(x) = \int_{-\infty}^{x} f_+(u) \, du \), \( F_2(x) = \int_{-\infty}^{x} f_-(u) \, du \).

If \( f \in L_1 \) it follows from the above fact and Theorem 8.2.1 that \( f^\mathbb{F}(t) \) is uniformly continuous on \( \mathbb{R} \).

It is clear that a general Fourier–Stieltjes Transform \( F^s(t) \) does not have to tend to zero as \( t \to \pm \infty \). For example if \( F(x) \) has a single jump of size \( \alpha \) at \( x = \lambda \), then \( F^s(t) = \alpha e^{i\lambda t} \). However, the Fourier Transform \( f^\mathbb{F}(t) \) of an \( L_1 \)-function \( f \) does tend to zero as \( t \to \pm \infty \) as the important Theorem 8.2.3 shows. This depends on the following useful lemma.

**Lemma 8.2.2** Let \( f \in L_1(-\infty, \infty) \) (Lebesgue measure). Then given \( \epsilon > 0 \) there exists a function \( h \) of the form \( h(x) = \sum_1^n \alpha_j \chi_{I_j}(x) \), where \( I_1, \ldots, I_n \) are (disjoint) bounded intervals, such that \( \int_{-\infty}^{\infty} |h - f| \, dx < \epsilon \).

**Proof** Since \( f \in L_1 \), there exists \( A < \infty \) such that \( \int_{|x|>A} |f(x)| \, dx < \epsilon/3 \), and hence \( \int |g - f| \, dx < \epsilon/3 \) where \( g(x) = f(x) \) for \( |x| < A \), and \( g(x) = 0 \) for \( |x| \geq A \). By the definition of the integral, \( g(x) \) may be approximated by a simple function \( k(x) = \sum_{j=1}^{n} \alpha_j \chi_{B_j}(x) \) where the \( B_j \) are bounded Borel sets and where \( \int |g - k| \, dx < \epsilon/3 \), so that \( \int |f - k| \, dx < 2\epsilon/3 \). Finally for each \( j \) there is a finite union \( I_j \) of bounded intervals such that \( m(B_j \triangle I_j) < \epsilon/(3n \max |\alpha_j|) \) where \( m \) denotes Lebesgue measure (Theorem 2.6.2), so that writing \( h(x) = \sum_1^n \alpha_j \chi_{I_j} \) we have

\[
 \int |k - h| \, dx \leq \sum |\alpha_j| \int |\chi_{I_j} - \chi_{B_j}| \, dx = \sum |\alpha_j|m(I_j \triangle B_j) < \epsilon/3
\]
Integrating complex functions, Fourier theory and related topics

and hence $\int |f - h| \, dx < \epsilon$. The given form of $h$ may now be achieved by a simple change of notation – replacing each $I_j$ by the intervals of which it is composed.

**Theorem 8.2.3 (Riemann–Lebesgue Lemma)** Let $f \in L_1(\mathbb{R})$ (i.e. $f$ is Lebesgue integrable). Then its Fourier Transform $f^\dagger(t) \to 0$ as $t \to \pm\infty$.

**Proof** Let $g$ be any function of the form $c\chi_{[a,b]}$ for finite constants $a, b, c$. Then $g^\dagger(t) = c\int_a^b e^{itx} \, dx = c[e^{ibt} - e^{ita}]/(it)$ which tends to zero as $t \to \pm\infty$.

If $h(x) = \sum_{j=1}^n \alpha_j g_j(x)$ where each $g_j$ is of the above type, then clearly $h^\dagger(t) \to 0$ as $t \to \pm\infty$.

Now given $\epsilon > 0$ there is (by Lemma 8.2.2) a function $h$ of the above type such that $\int |h(x) - f(x)| \, dx < \epsilon$. Hence

$$|f^\dagger(t)| = |\int e^{itx} (f(x) - h(x)) \, dx + h^\dagger(t)|$$

$$\leq \int |f(x) - h(x)| \, dx + |h^\dagger(t)|$$

$$< \epsilon + |h^\dagger(t)|.$$  

Since $h^\dagger(t) \to 0$ it follows that $|f^\dagger(t)|$ can be made arbitrarily small for $t$ sufficiently large (positive or negative) and hence $f^\dagger(t) \to 0$ as $t \to \pm\infty$, as required.

**8.3 Inversion of Fourier–Stieltjes Transforms**

The main result of this section is an inversion formula from which $F$ may be “recovered” from a knowledge of its Fourier–Stieltjes Transform. In fact the formula gives not $F$ itself but $\tilde{F}(x) = \frac{1}{2}[F(x+0) + F(x-0)] = \frac{1}{2}[F(x) + F(x-0)]$, assuming right-continuity. $F$ itself is easily obtained from $\tilde{F}$ since $F = \tilde{F}$ at continuity points, and at discontinuities $F(x) = \tilde{F}(x+0)$.

**Theorem 8.3.1 (Inversion for Fourier–Stieltjes Transforms)** Let $F$ be b.v. with Fourier–Stieltjes Transform $F^\ast$. Then for all real $a, b$ ($a < b$ say) with the above notation,

$$\tilde{F}(b) - \tilde{F}(a) = \lim_{T \to \infty} \frac{1}{2\pi} \int_{-T}^{T} e^{-ibt} - e^{-iat} F^\ast(t) \, dt.$$  

Also, for any real $a$, the jump of $F$ at $a$ is

$$F(a + 0) - F(a - 0) = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} e^{-iat} F^\ast(t) \, dt$$  

(which will be zero if $F$ is continuous at $a$).
Proof  If the result holds for bounded nondecreasing functions, it clearly holds for a b.v. function. Hence we assume that $F$ is nondecreasing and bounded (and right-continuous for convenience). Now

$$
\frac{1}{2\pi} \int_{-T}^{T} e^{-ibt} - e^{-iat} \frac{F^*(t)}{-it} dt = \frac{1}{2\pi} \int_{-T}^{T} e^{-ibt} - e^{-iat} \int_{-\infty}^{\infty} e^{itx} dF(x) dt
$$

$$
= \frac{1}{2\pi} \int_{-\infty}^{\infty} \left( \int_{-T}^{T} e^{it(x-b)} - e^{it(x-a)} \frac{dF(x)}{-it} \right) dF(x)
$$

by an application of Fubini’s Theorem (noting that the integrand may be written as $\int_{x-b}^{x-a} e^{itu} du$ and its modulus therefore does not exceed the constant $(b-a)$ which is integrable with respect to the product of Lebesgue measure on $(-T, T)$ and $F$-measure). Now the inner integral above is

$$
\int_{-T}^{T} \int_{x-b}^{x-a} e^{itu} du \ dt = \int_{x-b}^{x-a} \int_{-T}^{T} e^{itu} \ dt \ du
$$

$$
= 2 \int_{x-b}^{x-a} \sin Tu \ du = 2 \int_{T(x-b)}^{T(x-a)} \sin u \ du
$$

$$
= 2 \left( H[T(x-a)] - H[T(x-b)] \right)
$$

where $H(x) = \int_{0}^{x} \sin u \ du$. As is well known, $H$ is a bounded, odd function which converges to $\frac{\pi}{2}$ as $x \to \infty$. Hence $\lim_{T \to \infty} H[T(x-a)] = -\frac{\pi}{2}, 0$ or $\frac{\pi}{2}$ according as $x < a, x = a,$ or $x > a$. Thus (with the corresponding limit for $H[T(x-b)]$),

$$
\lim_{T \to \infty} \{ H[T(x-a)] - H[T(x-b)] \} = 0 \quad x < a \quad \text{or} \quad x > b
$$

$$
= \frac{\pi}{2} \quad x = a \quad \text{or} \quad x = b
$$

$$
= \pi \quad a < x < b.
$$

Further $\{ H[T(x-a)] - H[T(x-b)] \}$ is dominated in absolute value by a constant (which is $dF$-integrable) and hence, by dominated convergence,

$$
\lim_{T \to \infty} \frac{1}{2\pi} \int_{-T}^{T} e^{-ibt} - e^{-iat} \frac{F^*(t)}{-it} dt
$$

$$
= \frac{2}{2\pi} \left[ \frac{\pi}{2} (F(a) - F(a-0)) + \pi(F(b-0) - F(a)) + \frac{\pi}{2} (F(b) - F(b-0)) \right]
$$

which reduces to $\tilde{F}(b) - \tilde{F}(a)$, as required.

The second expression is obtained similarly. Specifically

$$
\frac{1}{2T} \int_{-T}^{T} e^{-iat} F^*(t) dt = \frac{1}{2T} \int_{-T}^{T} e^{-iat} \int_{-\infty}^{\infty} e^{itx} dF(x) dt
$$

$$
= \frac{1}{2T} \int_{-\infty}^{\infty} \int_{-T}^{T} e^{it(x-a)} dt dF(x) = \int_{-\infty}^{\infty} \frac{\sin T(x-a)}{T(x-a)} dF(x)
$$

where

$$
H(x) = \int_{0}^{x} \sin u \ du.
$$

As is well known, $H$ is a bounded, odd function which converges to $\frac{\pi}{2}$ as $x \to \infty$. Hence $\lim_{T \to \infty} H[T(x-a)] = -\frac{\pi}{2}, 0$ or $\frac{\pi}{2}$ according as $x < a, x = a,$ or $x > a$. Thus (with the corresponding limit for $H[T(x-b)]$),

$$
\lim_{T \to \infty} \{ H[T(x-a)] - H[T(x-b)] \} = 0 \quad x < a \quad \text{or} \quad x > b
$$

$$
= \frac{\pi}{2} \quad x = a \quad \text{or} \quad x = b
$$

$$
= \pi \quad a < x < b.
$$

Further $\{ H[T(x-a)] - H[T(x-b)] \}$ is dominated in absolute value by a constant (which is $dF$-integrable) and hence, by dominated convergence,
Integrating complex functions, Fourier theory and related topics

(Using Fubini) where the value of the integrand at \( x = a \) is unity. The integrand tends to zero as \( T \to \infty \) for all \( x \neq a \) and is bounded by one \((dF\text{-integrable})\). Hence the integral converges as \( T \to \infty \) by dominated convergence, to the value

\[
\mu_F(\{a\}) = F(a) - F(a - 0) = F(a + 0) - F(a - 0)
\]
as required.

\[\square\]

A most interesting case occurs when the (complex) function \( F^*(t) \) is itself in \( L_1(\mathbb{R}) \). First of all it is then immediate that \( F \) must be continuous since dominated convergence gives

\[
\lim_{T \to \infty} \int_{-T}^{T} e^{-iat} F^*(t) \, dt = \int_{-\infty}^{\infty} e^{-iat} F^*(t) \, dt
\]
and hence it follows from the second formula of Theorem 8.3.1 that \( F(a+0) - F(a-0) = 0 \). Similarly, the limit in the first inversion may be written as \( \int_{-\infty}^{\infty} \) instead of \( \lim \int_{-T}^{T} \) (again by dominated convergence) and \( \tilde{F} = F \) (since \( F \) is continuous) giving

\[
F(b) - F(a) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{ibt} - e^{-ibt}}{-it} F^*(t) \, dt.
\]

In fact even more is true and can be shown using the following obvious lemma.

**Lemma 8.3.2** Let \( F = F_1 - F_2 \) be a b.v. function on \( \mathbb{R} \) (\( F_1, F_2 \) bounded nondecreasing) and \( g \) a real function in \( L_1(\mathbb{R}) \) for any finite \( K \), and such that \( F(b) - F(a) = \int_{a}^{b} g(x) \, dx \) for all real \( a < b \). Then \( g \in L_1(\mathbb{R}) \) and \( \mu_F(E) = \int_{E} g(x) \, dx \) for all Borel sets \( E \) (\( \mu_F \) is defined to be \( \mu_{F_1} - \mu_{F_2} \)).

**Proof** Fix \( K \) and define the finite signed measures

\[
\mu(E) = \mu_F(E \cap (-K, K)), \quad \nu(E) = \int_{E \cap (-K, K)} g(x) \, dx.
\]
Clearly \( \mu = \nu \) for all sets of the form \( (a, b] \) and hence for all Borel sets (Lemma 5.2.4). Thus the “total variations” \( |\mu|, |\nu| \) are equal giving

\[
\int_{(-K, K]} g(x) \, dx = |\nu|(-K, K) = |\mu|(-K, K) \leq (\mu_{F_1} + \mu_{F_2})(-K, K) \leq (\mu_{F_1} + \mu_{F_2})(\mathbb{R}) < \infty.
\]
Hence \( g \in L_1(\mathbb{R}) \) by monotone convergence \( (K \to \infty) \). Thus \( \mu_F(E) \) and \( \int_{E} g \, dx \) are two finite signed measures which are equal on sets \( (a, b] \) and thus on \( \mathcal{B} \), as required. \[\square\]
Theorem 8.3.3  Let $F$ be b.v. on $\mathbb{R}$, with Fourier–Stieltjes Transform $F^*$, and assume $F^* \in L_1(-\infty, \infty)$. Then $F$ is absolutely continuous, and specifically

$$F(x) = F(-\infty) + \int_{-\infty}^{x} g(u) \, du$$

where $g(u) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-iat} F^*(t) \, dt$ is real and in $L_1(-\infty, \infty)$.

Proof  The formula just prior to Lemma 8.3.2 gives

$$F(b) - F(a) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{a}^{b} e^{-iat} F^*(t) \, dt \, du$$

by Fubini’s Theorem (since $F^* \in L_1$) and the definition of $g$.

To see that $g$ is real note that the integral of its imaginary part over any finite interval is zero, and it follows that the imaginary part of $g$ has zero integral over any Borel set $E$, and is thus zero a.e. (Theorem 4.4.8). But a function which is continuous and zero a.e. is everywhere zero (as is easily checked) and thus $g$ is real.

The result now follows at once by applying Lemma 8.3.2 to $F$ and $g$. □

We may now obtain an important inversion theorem for $L_1$ Fourier Transforms when the transform is also in $L_1$.

Theorem 8.3.4  Let $f \in L_1(-\infty, \infty)$. Then if its Fourier Transform $f^\dagger(t)$ is in $L_1(-\infty, \infty)$, we have the inversion

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-ixt} f^\dagger(t) \, dt \quad \text{a.e. (Lebesgue measure).}$$

Proof  Write $F(x) = \int_{-\infty}^{x} f(u) \, du$. Then by Theorem 8.3.3, for all $a, b$

$$\int_{a}^{b} f(u) \, du = F(b) - F(a) = \int_{a}^{b} g(u) \, du$$

where $g(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-ixt} f^\dagger(t) \, dt$ is real and in $L_1(-\infty, \infty)$. The finite signed measures $\int_{E} f \, dx$, $\int_{E} g \, dx$ are thus equal for all $E$ of the form $(a, b]$ and hence for all $E \in \mathcal{B}$ (and finally for all Lebesgue measurable sets $E$). Hence $f = g$ a.e. by the corollary to Theorem 4.4.8, as required. □

Note that the expression $f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-ixt} f^\dagger(t) \, dt$ a.e. may be regarded as displaying $f$ as an “inverse Fourier Transform”.

For (apart from the factor $\frac{1}{2\pi}$ and the negative sign in the exponent) this has the form of the Fourier Transform of the (assumed $L_1$) function $f^\dagger$. Of course we have defined Fourier Transforms of real functions since that is our primary interest (and $f^\dagger$ may be complex) but one could also define the transform of a complex
8.4 “Local” inversion for Fourier Transforms

In the last section it was shown that the inversion

\[ f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-ixt} f^\dagger(t) \, dt \quad \text{a.e.} \]

holds when the transform \( f^\dagger(t) \in L_1 \). There are important cases when \( f^\dagger \) does not belong to \( L_1 \) but where an inversion is still possible. For example suppose \( f(x) = 0 \) for \( x < 0 \) and \( f(x) = e^{-x} \) for \( x > 0 \). Then

\[
\begin{align*}
 f^\dagger(t) &= \int_0^\infty e^{-x} e^{ixt} \, dx = \int_0^\infty e^{-x} \cos xt \, dx + i \int_0^\infty e^{-x} \sin xt \, dx \\
 &= \frac{1}{1 + t^2} + \frac{it}{1 + t^2} \\
 &= \frac{1}{1 - it}.
\end{align*}
\]

Clearly \( f^\dagger(t) \notin L_1 \) since \( |f^\dagger(t)| = (1 + t^2)^{-1/2} \).

To obtain an appropriate inversion the following limit is needed.

**Lemma 8.4.1 (Dirichlet Limit)**  If for some \( \delta > 0 \), \( g(x) \) is a bounded nondecreasing function of \( x \) in \((0, \delta)\), then

\[
\frac{2}{\pi} \int_0^\delta \frac{\sin Tx}{x} g(x) \, dx \to g(0+) \quad \text{as } T \to \infty.
\]

*Proof* \( \int_0^\delta \left( \frac{\sin Tx}{x} \right) \, dx = \int_0^{T\delta} \left( \frac{\sin u}{u} \right) \, du \to \frac{\pi}{2} \) as \( T \to \infty \) (cf. proof of Theorem 8.3.1).

Thus it will be sufficient to show that

\[
\int_0^\delta \frac{\sin Tx}{x} (g(x) - g(0+)) \, dx \to 0.
\]

Given \( \epsilon > 0 \) there exists \( \eta > 0 \) such that \( g(\eta) - g(0+) < \epsilon \). Then

\[
\int_0^\eta \frac{\sin Tx}{x} (g(x) - g(0+)) \, dx = [g(\eta) - g(0+)] \int_0^\eta \frac{\sin Tx}{x} \, dx
\]

for some \( \xi \in [0, \eta] \) by the second mean value theorem for integrals. The last expression may be written as

\[
(g(\eta) - g(0+)) \int_{\xi T}^{\eta T} \frac{\sin x}{x} \, dx.
\]
But since \( \int_0^T (\sin u/u) \, du \) is bounded, \( \left| \int_{T_1}^{T_2} (\sin u/u) \, du \right| < A \) for some \( A \) and all \( T_1, T_2 \geq 0 \). Thus for all \( T \)

\[
\left| \int_0^\eta \frac{\sin Tx}{x} (g(x) - g(0)) \, dx \right| \leq \epsilon A.
\]

Now \((g(x) - g(0))/x \in L_1([\eta, \delta])\) (\( g \) being bounded and \( \eta > 0 \)). The Riemann–Lebesgue Lemma (Theorem 8.2.3) applies equally well to a finite range of integration (or the function may be extended to be zero outside such a range). Considering the imaginary part of the integral we see that \( \int_\eta^\delta (g(x) - g(0))(\frac{\sin Tx}{x}) \, dx \to 0 \) as \( T \to \infty \). Hence

\[
\lim_{T \to \infty} \left| \int_0^\eta \frac{\sin Tx}{x} (g(x) - g(0)) \, dx \right| \leq \epsilon A
\]

for any \( \epsilon > 0 \) from which the required result follows. \( \Box \)

Recall from Section 5.7 that a function \( f \) is b.v. \textit{in a finite range} if it can be written as the difference of two bounded nondecreasing functions in that range. The Dirichlet Limit clearly holds for such b.v. functions (in \((0, \delta)\)) also.

The desired inversion may now be obtained.

\textbf{Theorem 8.4.2} (Local Inversion Theorem for \( L_1 \) Transforms) \textit{If} \( f \in L_1 \), \textit{and} \( f \) is b.v. in \((x - \delta, x + \delta)\) \textit{for a fixed given} \( x \) \textit{and for some} \( \delta > 0 \), \textit{then}

\[
\frac{1}{2} \{f(x + 0) + f(x - 0)\} = \lim_{T \to \infty} \frac{1}{2\pi} \int_{-T}^T e^{-itx} f^\delta (t) \, dt.
\]

\textit{Proof}

\[
\frac{1}{2\pi} \int_{-T}^T e^{-itx} f^\delta (t) \, dt = \frac{1}{2\pi} \int_{-T}^T \int_{-\infty}^{\infty} e^{-it(x-y)} f(y) \, dy \, dt
\]

\[= \frac{1}{2\pi} \int_{-\infty}^{\infty} (\int_{-T}^T e^{-it(x-y)} \, dt) f(y) \, dy \quad \text{(Fubini)}
\]

\[= \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\sin T(x-y)}{x-y} f(y) \, dy
\]

\[= \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\sin Tu}{u} f(x+u) \, du.
\]

Now for \( x \) fixed, \( f(x + u)/u \) is in \( L_1(\delta, \infty) \) and \( L_1(-\infty, -\delta) \) for \( \delta > 0 \) so that

\[
\int_{|u|>\delta} (\sin Tu/u) f(x+u) \, du \to 0 \quad \text{as} \quad T \to \infty
\]

by the Riemann–Lebesgue Lemma. Thus we need consider only the range \([-\delta, \delta]\) for the integral. Now \( f(x+u) \) is b.v. in \((0, \delta)\) and by the Dirichlet
Limit $\frac{1}{\pi} \int_{0}^{\delta} (\sin Tu) f(x + u) \, du \rightarrow \frac{1}{2} f(x + 0)$. Similarly $\frac{1}{\pi} \int_{-\delta}^{0} (\sin Tu) f(x + u) \, du \rightarrow \frac{1}{2} f(x - 0)$ and hence

$$\frac{1}{\pi} \int_{-\delta}^{\delta} \frac{\sin Tu}{u} f(x + u) \, du \rightarrow \frac{1}{2} (f(x + 0) + f(x - 0))$$

giving the desired conclusion of the theorem. □

**Corollary** If $f$ is continuous at $x$ the stated inversion formula gives $f(x)$. If also $f^{\dagger} \in L_1$, $f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-ixt} f^{\dagger}(t) \, dt$.

In contrast to the previous inversion formula, that considered here applies to the value of $f$ at a given point $x$ rather than holding a.e. It is often convenient to use complex variable methods (i.e. contour integrals) to evaluate the formula. For example in the case $f^{\dagger}(t) = \frac{1}{1+it}$ one may consider $\frac{1}{2\pi} \int_{C} e^{zit} \, dz$ around upper and lower semicircles to recover $f(x) = 0$ for $x < 0$ and $f(x) = e^{-x}$ for $x > 0$. (The limit as $T \rightarrow \infty$ occurs naturally, making the semicircle larger.) The case $x = 0$ is easily checked directly giving the value $\frac{1}{2} \ (= \ (f(0+) + f(0-))/2)$. 

Foundations of probability

9.1 Probability space and random variables

By a probability space we mean simply a measure space for which the measure of the whole space is unity. It is customary to denote a probability space by \((\Omega, \mathcal{F}, P)\), rather than the \((X, \mathcal{S}, \mu)\) used in previous chapters for general measure spaces. That is, \(P\) is a measure on a \(\sigma\)-field \(\mathcal{F}\) of subsets of a space \(\Omega\), such that \(P(\Omega) = 1\) (and \(P\) is thus called a probability measure).

It will be familiar to the reader that this framework is used to provide a mathematical ("probabilistic") model for physical situations involving randomness i.e. a random experiment \(E\) – which may be very simple, such as the tossing of coins or dice, or quite complex, such as the recording of an entire noise waveform. In this model, each point \(\omega \in \Omega\) represents a possible outcome that \(E\) may have. The measurable sets \(E \in \mathcal{F}\) are termed events. An event \(E\) represents that "physical event" which occurs when the experiment \(E\) is conducted if the actual outcome obtained corresponds to one of the points of \(E\).

It will also be familiar that the complement \(E^c\) of an event \(E\) represents another physical event – which occurs precisely when \(E\) does not occur if \(E\) is conducted. Further, for two events \(E, F\), \(E \cup F\) represents that event which occurs if either or both of \(E, F\) occur, whereas \(E \cap F\) represents occurrence of both these events simultaneously. If \(E \cap F = \emptyset\), the events \(E\) and \(F\) cannot occur together when \(E\) is performed. Similar interpretations hold for other set operations such as \(-, \Delta, \cup^\infty\) and so on.

The probability measure \(P(E)\) (sometimes written also as \(Pr(E)\)) of an event \(E\), is referred to as the "probability that the event \(E\) occurs" when \(E\) is conducted. As is intuitively reasonable, its values lie between zero and one (\(P\) being monotone). If \(E, F\) are events which cannot occur together (i.e. disjoint events – \(E \cap F = \emptyset\)), it is also intuitively plausible that the probability \(P(E \cup F)\) of one or other of \(E, F\) occurring, should be equal to \(P(E) + P(F)\). This is true since the measure \(P\) is additive. (Of course, the
countable additivity of $P$ implies a corresponding statement for a sequence of disjoint events.)

It is worth recalling that these properties are also intuitively desirable from a consideration of the “frequency interpretation” of $P(E)$ as the proportion of times $E$ occurs in very many repetitions of $E$. Thus the requirements which make $P$ a probability measure are consistent with intuitive properties which probability should have.

We turn now to random variables. To conform to the notion of a random variable as a “numerical outcome of a random experiment”, it is intuitively reasonable to consider a function on $\Omega$ (i.e. an assignment of a numerical value to each possible outcome $\omega$). For example for two tosses of a coin we may write $\Omega = (\text{HH, HT, TH, TT})$ and the number of heads $\xi(\omega)$ taking the respective values 2, 1, 1, 0. It will be convenient to allow infinite values on occasions. Precisely, the following definitions will apply.

By an extended (real) random variable we shall mean a measurable function (Section 3.3) $\xi = \xi(\omega)$ defined a.e. on $(\Omega, \mathcal{F}, P)$. If the values of $\xi$ are finite a.e., we shall simply refer to $\xi$ as a random variable (r.v.).

Note that the precise usage of the term random variable is not uniform among different authors. Sometimes it is required that a r.v. be defined and finite for all $\omega$, and sometimes defined for all $\omega$ and finite a.e. The latter definition is inesthetistic since the sum of two such “r.v.’s” need not be defined for all $\omega$, and hence not a r.v. The former can be equally as good as the definition above since a redefinition of an a.e. finite function will lead to one which is everywhere finite, with the “same properties except on a zero measure set” (a fact which will be used from time to time anyway). Which definition is chosen is largely a matter of personal preference since there are compensating advantages and disadvantages of each, and in any case the differences are of no real consequence.

As in previous chapters, $\mathcal{B}$ ($\mathcal{B}^*$) will be used to denote the $\sigma$-field of Borel sets (extended Borel sets – Section 3.1) on the real line $\mathbb{R}$ (extended real line $\mathbb{R}^*$). By a Borel function $f$ on $\mathbb{R}$ ($\mathbb{R}^*$) we mean that $f$ (either real or extended real) is measurable with respect to $\mathcal{B}$ ($\mathcal{B}^*$).

An extended r.v. $\xi$ viewed as a mapping (transformation) from $\Omega$ to $\mathbb{R}^*$, induces the probability measure $P_{\xi^{-1}}$ on $\mathcal{B}^*$ (Section 3.7). As discussed in the next section this is the distribution of $\xi$, using the notation (for $B \in \mathcal{B}^*$),

$$P[\xi \in B] = P(\xi^{-1}B).$$

Similarly other obvious notation (such as $P[\xi \leq a]$ for $P_{\xi^{-1}}(-\infty, a]$) will be clear and used even if not formally defined.

A further convenient notation is the use of the abbreviation “a.s.” (“almost surely”) which is usually preferred over “a.e.” when the measure
involved is a probability measure. This is especially useful when another measure (e.g. Lebesgue) is considered simultaneously with $P$, since then “a.s.” will refer to $P$, and “a.e.” to the other measure. It is also not uncommon to use the phrase “with probability one” instead of “a.s.”. Thus statements (for a Borel set $B$) such as

“$\xi \in B$ a.e. ($P$)”, “$\xi \in B$ a.s.”, “$\xi \in B$ with probability one”, $P\{\xi \in B\} = 1$

are equivalent.

Finally the measures $P, P\xi$–1 may or may not be complete (Section 2.6). Completeness may, of course, be simply achieved where needed or desired by the completion procedure of Theorem 2.6.1.

### 9.2 Distribution function of a random variable

As above a r.v. $\xi$ on $(\Omega, \mathcal{F}, P)$ induces the distribution $P\xi$–1 on $(\mathbb{R}^*, \mathcal{B}^*)$ and also, by restriction, on $(\mathbb{R}, \mathcal{B})$. Further if $A$ denotes the (measurable) set of points $\omega$ where $\xi$ is either not defined or $\xi(\omega) = \pm \infty$ then $P(A) = 0$ and $P\xi$–1($\mathbb{R}$) = $P(\Omega) – P(A) = 1$, so that $P\xi$–1 is a probability measure on $\mathcal{B}$, and, since $P\xi$–1($\mathbb{R}^*$) = 1, also on $\mathcal{B}^*$.

Now $P\xi$–1 as a measure on $(\mathbb{R}, \mathcal{B})$ is a Lebesgue–Stieltjes measure, corresponding to the point function (Theorem 2.8.1) by

$$F(x) = P\xi$–1$((-\infty, x]) = P\{\xi \leq x\},$$

i.e. $P\xi$–1 = $\mu_F$ in the notation of Section 2.8. $F$ is called the distribution function (d.f.) of $\xi$. According to Theorem 2.8.1 $F(x)$ is nondecreasing and continuous to the right. Further it is easily checked, writing $F(–\infty) = \lim_{x \to –\infty} F(x), \; F(\infty) = \lim_{x \to \infty} F(x)$ that $F(–\infty) = 0, \; F(\infty) = 1$. In fact these properties are also sufficient for a function $F$ to be the d.f. of some r.v. $\xi$, as concluded in the following theorem.

**Theorem 9.2.1** (i) For a function $F$ on $\mathbb{R}$ to be the d.f. ($P\{\xi \leq x\}$) of some r.v. $\xi$, it is necessary and sufficient that $F$ be nondecreasing, continuous to the right and that $\lim_{x \to –\infty} F(x) = 0, \lim_{x \to \infty} F(x) = 1$.

(ii) Two r.v.’s $\xi, \eta$ (on the same or different probability spaces) have the same distribution (i.e. $P\xi$–1$B = P\eta$–1$B$ for all $B \in \mathcal{B}^*$) if and only if they have the same d.f. $F$.

**Proof**    The necessity of the conditions in (i) has been shown by the remarks above. Conversely if $F$ is a nondecreasing function with the properties stated in (i), we may define a probability space $(\mathbb{R}, \mathcal{B}, \mu_F)$ where $\mu_F$ is
the measure defined by $F$ (as in Theorem 2.8.1). Since

$$
\mu_F(\mathbb{R}) = \lim_{n \to \infty} \mu_F([-n,n]) = \lim_{n \to \infty} [F(n) - F(-n)] = 1,
$$

it follows that $\mu_F$ is a probability measure. If $\xi$ denotes the “identity r.v.” on $(\mathbb{R}, \mathcal{B}, \mu_F)$ given (for real $\omega$) by $\xi(\omega) = \omega$, its d.f. is

$$
\mu_F\xi^{-1}((\infty, x]) = \mu_F((-\infty, x]) = F(x),
$$

so that $F$ is the d.f. of a r.v. $\xi$ as required.

To prove (ii), note that clearly if $\xi, \eta$ have the same distribution (on either $\mathcal{B}^*$ or $\mathcal{B}$) they have the same d.f. (Take $B = (-\infty, x]$.) Conversely if $\xi, \eta$ have the same d.f., then by the uniqueness part of Theorem 2.8.1, $P\xi^{-1}$ and $P\eta^{-1}$ are equal on $\mathcal{B}$ (being measures on $(\mathbb{R}, \mathcal{B})$ corresponding to the same function $F$), i.e. $P\xi^{-1}(B) = P\eta^{-1}(B)$ for all $B \in \mathcal{B}$. But this also holds if $B$ is replaced by $B \cup \{\infty\}$, $B \cup \{-\infty\}$ or $B \cup \{-\infty\} \cup \{\infty\}$ (since e.g. $P\xi^{-1}(B \cup \{\infty\}) = P\xi^{-1}(B) = P\eta^{-1}(B \cup \{\infty\})$). That is $P\xi^{-1} = P\eta^{-1}$ on $\mathcal{B}^*$ also.  

If two r.v.’s $\xi, \eta$ (on the same or different probability spaces) have the same distribution ($P\xi^{-1}B = P\eta^{-1}B$ for all $B \in \mathcal{B}$, or equivalently for all $B \in \mathcal{B}^*$) we say that they are identically distributed, and write $\xi \overset{d}{=} \eta$. By the theorem it is necessary and sufficient for this that they have the same d.f. It is, incidentally, usually “distributional properties” of a r.v. which are important in probability theory. If $\xi$ is a r.v. on some $(\Omega, \mathcal{F}, P)$, we can always find an identically distributed r.v. on the real line. For if $F$ is the d.f. of $\xi$ a r.v. $\eta$ may be constructed on $(\mathbb{R}, \mathcal{B}, \mu_F)$ as above ($\eta(x) = x$). $\eta$ has the same d.f. $F$ as $\xi$, and hence the same distribution as $\xi$, by Theorem 9.2.1.

As noted, if $F$ is the d.f. of $\xi$, $P\xi^{-1}$ is the Lebesgue–Stieltjes measure $\mu_F$ defined by $F$ as in Section 2.8. However, in addition to being everywhere finite, as required in Section 2.8, a d.f. is bounded (with values between zero and one).

A d.f. $F$ may have discontinuities, but as noted above it is continuous to the right. Also since $F$ is monotone the limit $F(x-0) = \lim_{h \to 0} F(x-h)$ exists for every $x$. The measure of a single point is clearly the jump $\mu_F(x) = F(x) - F(x-0)$. The following useful result follows from Lemma 2.8.2.

**Lemma 9.2.2** Let $F$ be a d.f. (with corresponding probability measure $\mu_F$ on $\mathcal{B}$). Then $\mu_F$ has at most countably many “atoms” (i.e. points $x$ with $\mu_F(x) > 0$). Correspondingly $F$ has at most countably many discontinuity points.
Two extreme kinds of distribution and d.f. are of special interest. The first corresponds to r.v.’s $\xi$ whose distribution $P_{\xi}^{-1}$ on $\mathcal{B}$ is discrete. That is (cf. Section 5.7) there is a countable set $C$ such that $P_{\xi}^{-1}(C^c) = 0$. If $C = \{x_1, x_2, \ldots\}$ and $P_{\xi}^{-1}\{x_i\} = p_i$, we have for any $B \in \mathcal{B}$

$$P_{\xi}^{-1}(B) = P_{\xi}^{-1}(B \cap C) = \sum_{\{x_i \in B\}} P_{\xi}^{-1}\{x_i\} = \sum_{\{x_i \in B\}} p_i$$

and thus for the d.f.

$$F(x) = P_{\xi}^{-1}(-\infty, x] = \sum_{\{x_i \leq x\}} p_i.$$  

$F$ increases by jumps of size $p_i$ at the points $x_i$ and is called a discrete d.f. The r.v. $\xi$ with such a d.f. is also said to be a discrete r.v. Note that such a d.f. may often be visualized as an increasing “step function” with successive stairs of heights $p_i$. This is the case (cf. Section 5.7) if the $x_i$ can be written as a sequence in increasing order of size. However, such size ordering is not always possible – as when the set of $x_i$ consists of all rational numbers.

Two standard examples of discrete r.v.’s are

(i) **Binomial**, where $C = \{0, 1, 2, \ldots, n\}$ and

$$p_r = \binom{n}{r} p^r (1-p)^{n-r}, \ r = 0, 1, \ldots, n \ (0 \leq p \leq 1),$$

(ii) **Poisson**, where $C = \{0, 1, 2, \ldots\}$ and

$$p_r = e^{-m} m^r / r!, \ r = 0, 1, 2, \ldots \ (m > 0).$$

At the “other extreme” the distribution $P_{\xi}^{-1} (= \mu_F)$ of $\xi$ may be absolutely continuous with respect to Lebesgue measure. Then for any $B \in \mathcal{B}$

$$P_{\xi}^{-1}(B) = \int_B f(x) \, dx$$

where the Radon–Nikodym derivative $f$ (of $P_{\xi}^{-1}$ with respect to Lebesgue measure) is nonnegative a.e. and hence may be taken as everywhere nonnegative (by writing e.g. zero instead of negative values). $f$ is in $L_1(-\infty, \infty)$ and its integral is unity. It is called the probability density function (p.d.f.) for $\xi$ and the d.f. is given by

$$F(x) = P_{\xi}^{-1}(-\infty, x] = \int_{-\infty}^x f(u) \, du.$$  

($F$ is thus an absolutely continuous function – cf. Section 5.7.) We then say that $\xi$ has an absolutely continuous distribution or simply that $\xi$ is an absolutely continuous r.v. Common examples are
(i) the *normal distribution* \( N(\mu, \sigma^2) \) where

\[
f(x) = (\sigma \sqrt{2\pi})^{-1} \exp\{- (x - \mu)^2 / 2\sigma^2\} \quad (\mu \text{ real}, \sigma > 0)
\]

(ii) the *gamma distribution* with parameters \( \alpha > 0, \beta > 0 \), where

\[
f(x) = \frac{\alpha \beta}{\Gamma(\beta)} x^{\beta-1} e^{-\alpha x} \quad (x > 0).
\]

The case \( \beta = 1 \) gives the exponential distribution.

There is a third “extreme type” of r.v. which is not typically encountered in classical statistics but has received significant more recent attention in connection with use of fractals in important applied sciences. This is a r.v. \( \xi \) whose distribution is *singular* with respect to Lebesgue measure (Section 5.4) and such that \( P_\xi^{-1}\{x\} = 0 \) for every singleton set \( \{x\} \). That is \( P_\xi^{-1} \) has mass confined to a set \( B \) of Lebesgue measure zero, but unlike a discrete r.v. \( P_\xi^{-1} \) has no atoms in \( B \) (or \( B^c \), of course). The corresponding d.f. is everywhere continuous, but clearly by no means absolutely continuous. Such a d.f. (and the r.v.) will be called singular (though continuous singular would perhaps be a better name).

It is readily seen from Section 5.7 that any d.f. whatsoever may be represented in terms of the three special types considered above, as the following celebrated result shows.

**Theorem 9.2.3** (Lebesgue Decomposition for d.f.’s) *Any d.f. \( F \) may be written as a “convex combination”*

\[
F(x) = \alpha_1 F_1(x) + \alpha_2 F_2(x) + \alpha_3 F_3(x)
\]

where \( F_1, F_2, F_3 \) are d.f.’s, \( F_1 \) being absolutely continuous, \( F_2 \) discrete, \( F_3 \) singular, and where \( \alpha_1, \alpha_2, \alpha_3 \) are nonnegative with \( \alpha_1 + \alpha_2 + \alpha_3 = 1 \). The constants \( \alpha_1, \alpha_2, \alpha_3 \) are unique, and so is the \( F_i \) corresponding to any \( \alpha_i > 0 \) (hence the term \( \alpha_i F_i \) is unique for each \( i \)).

**Proof** By Theorem 5.7.1 (Corollary) we may write \( F(x) = F_1^*(x) + F_2^*(x) + F_3^*(x) \), where \( F_i^* \) are nondecreasing functions defining measures \( \mu_{F_i^*} \) which are respectively absolutely continuous, discrete and singular (for \( i = 1, 2, 3 \)). Further, noting that \( \sum_{i=1}^3 F_i^*(-\infty) = 0 \), we may replace \( F_i^* \) by \( F_i^* - F_i^*(-\infty) \) and hence take \( F_i^*(-\infty) = 0 \) for each \( i \). Write now \( \alpha_i = F_i^*(\infty) \) and \( F_i(x) = F_i^*(x)/\alpha_i \) if \( \alpha_i > 0 \) (and an arbitrary d.f. of “type \( i \)” if \( \alpha_i = 0 \)). Then \( F_i \) is a d.f. and the desired decomposition \( F(x) = \alpha_1 F_1(x) + \alpha_2 F_2(x) + \alpha_3 F_3(x) \) follows. Letting \( x \to \infty \) we see that \( \alpha_1 + \alpha_2 + \alpha_3 = 1 \).

If there is another such decomposition, \( F = \beta_1 G_1 + \beta_2 G_2 + \beta_3 G_3 \) say, then

\[
\mu_{\alpha_1 F_1} + \mu_{\alpha_2 F_2} + \mu_{\alpha_3 F_3} = \mu_{\beta_1 G_1} + \mu_{\beta_2 G_2} + \mu_{\beta_3 G_3}
\]
and hence by Theorem 5.7.1, $\mu_{\alpha_i F_i} = \mu_{\beta_i G_i}$. Hence $\alpha_i F_i$ differs from $\beta_i G_i$ at most by an additive constant which must be zero since $F_i$ and $G_i$ vanish at $-\infty$. Since $F_i(\infty) = G_i(\infty) = 1$ we thus have $\alpha_i = \beta_i$ and hence also $F_i = G_i$ (provided $\alpha_i > 0$).

\[ \square \]

### 9.3 Random elements, vectors and joint distributions

It is natural to extend the concept of a r.v. by considering more general mappings rather than just “measurable functions”. These will be precisely “measurable transformations” as discussed in Chapter 3, but the term “measurable mapping” will be more natural (and thus used) in the present context. Specifically let $\xi$ be a measurable mapping defined a.s. on a probability space $(\Omega, \mathcal{F}, P)$, to a measurable space $(X, \mathcal{S})$ (i.e. $\xi^{-1} E \in \mathcal{F}$ for all $E \in \mathcal{S}$). Then $\xi$ will be called a random element (r.e.) on $(\Omega, \mathcal{F}, P)$ with values in $X$ (or in $(X, \mathcal{S})$). An extended r.v. is thus a r.e. with values in $\mathbb{R}^*$. Another case of importance is when $(X, \mathcal{S}) = (\mathbb{R}^n, \mathcal{B}^n)$ and $\xi(\omega) = (\xi_1(\omega), \ldots, \xi_n(\omega))$. A r.e. of this form and such that each $\xi_i$ is finite a.s. will be called a random vector or vector random variable. Yet more generally a stochastic process may be defined as a r.e. of $(X, \mathcal{S}) = (\mathbb{R}^T, \mathcal{B}^T)$ (cf. Section 7.9) for e.g. an index set $T = \{1, 2, 3, \ldots\}$ or $T = (0, \infty)$. As will be briefly indicated in Chapter 15 this is alternatively described as an infinite (countable or uncountable) family of r.v.’s.

Before pursuing probabilistic properties of random elements it will be convenient to develop some notation and obvious measurability results in the slightly more general framework in which $\xi$ is a mapping defined on a space $\Omega$, not necessarily a probability space, with values in a measurable space $(X, \mathcal{S})$. Apart from notation this is precisely the framework of Section 3.2 replacing $X$ by $\Omega$ and $(Y, T)$ by $(X, \mathcal{S})$, and identifying $\xi$ with the transformation $T$. It will be more natural in the present context to refer to $\xi$ as a mapping rather than a transformation but the results of Section 3.2 apply. For such a mapping $\xi$ the $\sigma$-field $\sigma(\xi)$ generated by $\xi$ is defined on $\Omega$ (cf. Section 3.2, identifying $\xi$ with $T$) by

$$\sigma(\xi) = \sigma(\xi^{-1} S) = \sigma(\xi^{-1} E : E \in \mathcal{S}).$$

As noted in Section 3.3, $\sigma(\xi)$ is the smallest $\sigma$-field $\mathcal{G}$ on $\Omega$ making $\xi \mathcal{G}|S$-measurable. Further if $\xi(\omega)$ is defined for every $\omega$ then the $\sigma$-ring $\xi^{-1}(\mathcal{S})$ contains $\xi^{-1}(X) = \Omega$ and hence is itself the $\sigma$-field $\sigma(\xi)$. Note that $\sigma(\xi)$ depends on the “range” $\sigma$-field $\mathcal{S}$. 

More generally if $C$ is any family of mappings on the same space $\Omega$, but with values in possibly different measurable spaces, we write

$$\sigma(C) = \sigma(\cup_{\xi \in C}\sigma(\xi)).$$

If the family is written as an indexed set $C = \{\xi_\lambda : \lambda \in \Lambda\}$, where $\xi_\lambda$ maps $\Omega$ into $(X_\lambda, S_\lambda)$, we write

$$\sigma(C) = \sigma\{\xi_\lambda : \lambda \in \Lambda\} = \sigma(\cup_{\lambda \in \Lambda}\sigma(\xi_\lambda)).$$

For $\Lambda = \{1, 2, \ldots, n\}$ write $\sigma(C) = \sigma(\xi_1, \xi_2, \ldots, \xi_n)$.

The following lemma, stated for reference, should be proved as an exercise (Ex. 9.7).

**Lemma 9.3.1**

(i) If $C$ is any family of mappings on the space $\Omega$, $\sigma(C)$ is then the unique smallest $\sigma$-field on $\Omega$ with respect to which every $\xi \in C$ is measurable. ($\sigma(C)$ is called the $\sigma$-field generated by $C$.)

(ii) If $C = \{\xi_\lambda : \lambda \in \Lambda\}$, $\xi_\lambda$ taking values in $(X_\lambda, S_\lambda)$, then $\sigma(C) = \sigma(\xi_\lambda^{-1}B_\lambda : B_\lambda \in S_\lambda, \lambda \in \Lambda)$.

(iii) If $C_\lambda$ is a family of mappings on the space $\Omega$ for each $\lambda$ in an index set $\Lambda$ then

$$\sigma(\cup_{\lambda \in \Lambda}C_\lambda) = \sigma(\cup_{\lambda \in \Lambda}\sigma(C_\lambda)).$$

As indicated above, we shall be especially interested in the case where $(X, S) = (\mathbb{R}^n, B^\otimes n)$ leading to random vectors. The following lemma will be applied to show the equivalence of a random vector and its component r.v.'s.

**Lemma 9.3.2** Let $\xi$ be a mapping defined on a space $\Omega$ with values in $(\mathbb{R}^n, B^\otimes n)$ so that $\xi = (\xi_1, \xi_2, \ldots, \xi_n)$ where $\xi_i$ maps $\Omega$ into $(\mathbb{R}^*, B^*)$. Then $\sigma(\xi) = \sigma(\xi_1, \xi_2, \ldots, \xi_n)$. That is the $\sigma$-field generated on $\Omega$ by the mapping $\xi$ into $(\mathbb{R}^n, B^\otimes n)$ is identical to that generated by the family of its components $\xi_i$, each mapping $\Omega$ into $(\mathbb{R}^*, B^*)$.

**Proof** If $B_i \in B^*$ for each $i$, then $\xi^{-1}(B_1 \times B_2 \times \ldots \times B_n) = \cap_{1}^{n}\xi_i^{-1}B_i$. Since the rectangles $B_1 \times B_2 \times \ldots \times B_n$ generate $B^\otimes n$, the corollary to Theorem 3.3.2 gives

$$\sigma(\xi) = \sigma(\cap_{1}^{n}\xi_i^{-1}B_i : B_i \in B^*) = \sigma(\xi_i^{-1}B_i : B_i \in B^*), 1 \leq i \leq n$$

as is easily checked. But this is just $\sigma(\xi_1, \xi_2, \ldots, \xi_n)$ by Lemma 9.3.1 (ii).

We proceed now to consider random vectors – measurable mappings $\xi = (\xi_1, \xi_2, \ldots, \xi_n)$ defined a.s. on a probability space $(\Omega, F, P)$ with values in $(\mathbb{R}^n, B^\otimes n)$ its components $\xi_i$ being finite a.s. (i.e. $\xi \in \mathbb{R}^n$ a.s.).
The following result shows that a random vector \( \xi \) is, equivalently, just a family of \( n \) r.v.’s \( (\xi_1, \ldots, \xi_n) \) (with \( \sigma(\xi) = \sigma(\xi_1, \ldots, \xi_n) \) as shown above).

**Theorem 9.3.3** Let \( \xi \) be a mapping defined a.s. on a probability space \((\Omega, \mathcal{F}, P)\), with values in \( \mathbb{R}^n \). Write \( \xi = (\xi_1, \xi_2, \ldots, \xi_n) \). Then \( \sigma(\xi) = \sigma(\xi_1, \xi_2, \ldots, \xi_n) \). Further, \( \xi \) is a random element in \((\mathbb{R}^n, \mathcal{B}^n)\) (i.e. \( \mathcal{F} | \mathcal{B}^n \)-measurable) if and only if each \( \xi_i \) is an extended r.v. (i.e. \( \mathcal{F} | \mathcal{B}_i \)-measurable). Hence \( \xi \) is a random vector (r.e. of \((\mathbb{R}^n, \mathcal{B}^n)\)) if and only if each \( \xi_i \) is a r.v.

**Proof** That \( \sigma(\xi) = \sigma(\xi_1, \xi_2, \ldots, \xi_n) \) restates Lemma 9.3.2. The mapping \( \xi \) is a r.e. on \((\Omega, \mathcal{F}, P)\) with values in \((\mathbb{R}^n, \mathcal{B}^n)\) iff it is \( \mathcal{F} \)-measurable, i.e. \( \sigma(\xi) \subset \mathcal{F} \). But this is precisely \( \sigma(\xi_1, \xi_2, \ldots, \xi_n) \subset \mathcal{F} \), which holds iff all \( \xi_i \) are extended r.v.’s. The final statement also follows immediately. \( \square \)

The **distribution** of a r.e. \( \xi \) on \((\Omega, \mathcal{F}, P)\) with values in \((X, \mathcal{S})\) is defined to be the probability measure \( P\xi \) on \( \mathcal{S} \) – directly generalizing the distribution of a r.v. Note that a corresponding point function (d.f.) is not defined as before except in special cases where e.g. \( X = \mathbb{R}^n \) (or at least has some “order structure”). The distribution \( P\xi^{-1} \) of a random vector \( \xi = (\xi_1, \ldots, \xi_n) \), is a probability measure on \( \mathcal{B}^n \), and its restriction to \( \mathcal{B}_i \) is a probability measure on \((\mathbb{R}^n, \mathcal{B}_i)\), as in the case \( n = 1 \) considered previously. The corresponding point function (cf. Section 7.8) \( F(x_1, \ldots, x_n) = P[\xi_i \leq x_i, 1 \leq i \leq n] = P\xi^{-1}(\{-\infty, x\}] \) \( (x = (x_1, \ldots, x_n)) \) is the joint distribution function of \( \xi_1, \ldots, \xi_n \). As shown in Theorem 7.8.1, such a function has the following properties:

(i) \( F \) is bounded, nondecreasing and continuous to the right in each \( x_i \).

(ii) For any \( a = (a_1, \ldots, a_n) \), \( b = (b_1, \ldots, b_n) \), \( a_i < b_i \) we have

\[
\sum \text{(*)} (-)^{n-r} F(c_1, c_2, \ldots, c_n) \geq 0
\]

where \( \sum \text{(*)} \) denotes summation over the \( 2^n \) distinct terms with \( c_i = a_i \) or \( b_i \) and \( r \) is the number of \( c_i \) which are \( b_i \)’s.

In addition since \( P\xi^{-1} \) is a probability measure it is easy to check that the following also hold:

(iii) \( 0 \leq F(x_1, \ldots, x_n) \leq 1 \) for all \( x_1, \ldots, x_n \), \( \lim_{x_i \to -\infty} F(x_1, \ldots, x_n) = 0 \) (for any fixed \( i \)), and

\[
\lim_{(x_1, \ldots, x_n) \to (\infty, \ldots, \infty)} F(x_1, \ldots, x_n) = 1.
\]

In fact these conditions are also sufficient for \( F \) to be the joint d.f. of some set of r.v.’s as stated in the following theorem.
Theorem 9.3.4 A function $F$ on $\mathbb{R}^n$ is the joint d.f. of some r.v.’s $\xi_1, \ldots, \xi_n$ if and only if it satisfies Conditions (i)–(iii) above. Then for $a_i \leq b_i$, $1 \leq i \leq n$, $P[a_i < \xi_i \leq b_i, 1 \leq i \leq n]$ is given by the sum in (ii) above.

Sketch of Proof The necessity of the conditions has been noted. The sufficiency follows simply from the fact (Theorem 7.8.1) that $F$ defines a measure $\mu_F$ on $(\mathbb{R}^n, \mathcal{B}^n)$. It is easily checked that $\mu_F$ is a probability measure.

If $\Omega = \mathbb{R}^n$, $\mathcal{F} = \mathcal{B}^n$, $P = \mu_F$ and $\xi_i(x_1, x_2, \ldots, x_n) = x_i$ then $\xi_1, \ldots, \xi_n$ are r.v.’s on $\Omega$ with the joint d.f. $F$. (The details should be worked through as an exercise.)

As in the previous section, it is of particular interest to consider the case when $P_{\xi^{-1}}$ is absolutely continuous with respect to $n$-dimensional Lebesgue measure, i.e. for every $E \in \mathcal{B}^n$,

$$P_{\xi^{-1}}(E) = \int_E f(u_1, \ldots, u_n) \, du_1 \ldots du_n$$

for some Lebesgue integrable $f$ which is thus (Radon–Nikodym Theorem) nonnegative a.e. (hence may be taken everywhere nonnegative) and integrates over $\mathbb{R}^n$ to unity. Equivalently, this holds if and only if

$$F(x_1, \ldots, x_n) = \int_{-\infty}^{x_n} \cdots \int_{-\infty}^{x_1} f(u_1, \ldots, u_n) \, du_1 \ldots du_n$$

for all choices of $x_1, \ldots, x_n$. We say that $f$ is the joint p.d.f. of the r.v.’s $\xi_1, \ldots, \xi_n$ whose d.f. is $F$. As noted above its integral over any set $E \in \mathcal{B}^n$ gives $P_{\xi^{-1}}(E)$ which is the probability $P[\xi \in E]$ that the value of the vector $(\xi_1(\omega), \ldots, \xi_n(\omega))$ lies in the set $E$.

Next note that if the r.v.’s $\xi_1, \ldots, \xi_n$ have joint d.f. $F$, the joint d.f. of any subset, say $\xi_1, \ldots, \xi_k$ of the $\xi$’s may be obtained by letting the remaining $x$’s $(x_{k+1}, \ldots, x_n)$ tend to $+\infty$; e.g. $F(x_1, \ldots, x_{n-1}, \infty) = \lim_{x_n \to \infty} F(x_1, \ldots, x_{n-1}, x_n)$ is the joint d.f. of $\xi_1, \ldots, \xi_{n-1}$. This is easily checked. If $F$ is absolutely continuous, the joint density for $\xi_1, \ldots, \xi_k$ may be obtained by integrating the density $f(x_1, \ldots, x_n)$ (corresponding to $F$) over $x_{k+1}, \ldots, x_n$. Again this is easily checked (Ex. 9.9). Of course, if we “put” $x_2 = x_3 = \cdots = x_n = \infty$ in the joint d.f. (or integrate the joint density over these variables in the absolutely continuous case) we obtain just the d.f. (or p.d.f.) of $\xi_1$. Accordingly the d.f. (or p.d.f.) of $\xi_1$ is called a marginal d.f. (or p.d.f.) obtained from the joint d.f. (or p.d.f.) in this way.

Finally, note that if $\xi_1, \ldots, \xi_n$, $\xi_1^*, \ldots, \xi_n^*$ are r.v.’s such that $\xi_i^* = \xi_i$ a.s. for each $i$, then the joint d.f.’s of the two families $(\xi_1, \ldots, \xi_n)$, $(\xi_1^*, \ldots, \xi_n^*)$ are equal. This is obvious, but should be checked.
9.4 Expectation and moments

Let \((\Omega, \mathcal{F}, P)\) be a probability space. If \(\xi\) is a r.v. or extended r.v. on this space, we write \(E\xi\) to denote \(\int \xi(\omega) dP(\omega)\) whenever this integral is defined, e.g. if \(\xi\) is a.s. nonnegative or \(\xi \in L_1(\Omega, \mathcal{F}, P)\). \(E\) thus simply denotes the operation of integration with respect to \(P\) and \(E\xi\) is termed the mean or expectation of \(\xi\). In the case where \(\xi \in L_1(\Omega, \mathcal{F}, P)\) (and hence in particular \(\xi\) is a.s. finite and thus a r.v.) \(E\xi\) and \(E|\xi|\) are finite (since \(|\xi|\) \(\in L_1\) also). It is then customary to say that the mean of \(\xi\) exists, or that \(\xi\) has a finite mean.

Since \(E\) denotes integration, any theorem of integration theory will be used with this notation without comment.

Suppose now that \(\xi\) is finite a.s. (i.e. is a r.v.) with d.f. \(F\). Let \(g(x) = |x|\), so that \(g(\xi(\omega))\) is defined a.s. and then

\[
E|\xi| = \int_{\Omega} g(\xi(\omega)) dP(\omega) = \int_{\mathbb{R}^*} g(x) dP^{-1}(x)
\]

viewing \(\xi\) as a transformation from \(\Omega\) to \(\mathbb{R}^*\) (Theorem 4.6.1). But this latter integral is just \(\int_{\mathbb{R}} g(x) dP^{-1}(x) = \int |x| dF(x)\) (since \(P^{-1} = \mu_F\) – see Section 4.7) and hence

\[
E|\xi| = \int |x| dF(x) \leq \infty.
\]

\(E|\xi|\) is thus finite if and only if \(\int |x| dF(x) < \infty\), and in this case the same argument but with \(g(x) = x\) gives

\[
E\xi = \int x dF(x).
\]

If also \(\xi\) has an absolutely continuous distribution, with p.d.f. \(f\) then (Theorem 5.6.1)

\[
E\xi = \int x f(x) dx.
\]

On the other hand, if \(\xi\) is discrete with \(P(\xi = x_n) = p_n\), it is easily checked (Ex. 9.12) that \(E|\xi| = \sum p_n |x_n|\) and, when \(E|\xi| < \infty\), that \(E\xi = \sum p_n x_n\).

Suppose now that \(\xi\) is a r.v. on \((\Omega, \mathcal{F}, P)\) and that \(g\) is a real-valued measurable function on \(\mathbb{R}\). Then \(g(\xi(\omega))\) is clearly a r.v. (Theorem 3.4.3) and an argument along the precise lines as that given above at once demonstrates the truth of the following result.

**Theorem 9.4.1** If \(\xi\) is a r.v. and \(g\) is a finite real-valued measurable function on \(\mathbb{R}\), then \(E|g(\xi)| < \infty\) if and only if \(\int |g(x)| dF(x) < \infty\). Then \(Eg(\xi) = \int g(x) dF(x)\).

In particular consider \(g(x) = x^p\) for \(p = 1, 2, 3, \ldots\). We call \(E|\xi|^p\) the \(p\)th absolute moment of \(\xi\) and when it is finite, say that the \(p\)th moment of
ξ exists, given by \( E\xi^p \). This holds equivalently if \( ξ \in L_p(Ω, F, P) \) and the theorem shows that \( E\xi^p = \int x^p \, dF(x) \).

If \( p > 0 \) but \( p \) is not an integer then \( x^p \) is not real-valued for \( x < 0 \) and thus \( E\xi^p(ω) \) is not necessarily defined a.s. However, if \( ξ \) is a nonnegative r.v. (a.s.) \( E\xi^p(ω) \) is defined a.s. and the above remarks hold. In any case one can still consider \( E|ξ|^p \) for all \( p > 0 \) regardless of the signs of the values of \( ξ \).

It will be seen in the next section that if \( ξ \in L_p = L_p(Ω, F, P) \) for some \( p > 1 \) (i.e. \( E|ξ|^p < ∞ \)) then \( ξ \in L_q \) for \( 1 ≤ q ≤ p \). (This fact applies since \( P \) is a finite measure – it does not apply to \( L_p \) classes for general measures.) Thus in this case the mean of \( ξ \) exists in particular, and (since any constant belongs to \( L_p \) on account of the finiteness of \( P \)) if \( p \) is a positive integer, \( ξ – Eξ \in L_p \) or \( E|ξ – Eξ|^p < ∞ \). This quantity is called the \( p \)th absolute central moment of \( ξ \), and \( E(ξ – Eξ)^p \) the \( p \)th central moment,

\[ p = 1, 2, \ldots. \]

If \( p = 2 \), the quantity \( E(ξ – Eξ)^2 \) is the variance of \( ξ \) (denoted by var(\( ξ \)) or \( σ^2_ξ \)). It is readily checked (Ex. 9.13) that a central moment may be expressed in terms of ordinary moments (and conversely) and in particular that \( var(ξ) = Eξ^2 – (Eξ)^2 \).

Joint moments of two or more r.v.’s are also commonly used. For example if \( ξ, η \) have finite second moments (\( ξ, η \in L_2 \)) then as will be seen in Theorems 9.5.2, 9.5.1 they are both in \( L_1 \) and \( (ξ – Eξ)(η – Eη) \in L_1 \). The expectation \( γ = E(ξ – Eξ)(η – Eη) \) is termed the covariance (cov(\( ξ, η \)) of \( ξ \) and \( η \), and \( ρ = γ/(σ_ξσ_η) \) is their correlation, where \( σ^2_ξ = var(ξ) \) and \( σ^2_η = var(η) \). See Ex. 9.20 for some useful interpretations and properties which should be checked.

A most important family of r.v.’s in statistical theory and practice arising from Theorem 9.3.4 is that of multivariate normal r.v.’s \( ξ_1, ξ_2, \ldots, ξ_n \) whose joint distribution is specified by their means, variances and covariances (or correlations). For the nonsingular case they have the joint p.d.f.

\[ f(x_1, x_2, \ldots, x_n) = (2π)^{-n/2} |Λ|^{-1/2} \exp\{-\frac{1}{2} (x – μ)’ Λ^{-1} (x – μ)\} \]

where \( x = (x_1, x_2, \ldots, x_n)’ \), \( μ = (μ_1, μ_2, \ldots, μ_n)’ \), \( (μ_i = Eξ_i) \) and \( Λ \) is the covariance matrix with \((i,j)\)th element \( γ_{ij} = \text{cov}(ξ_i, ξ_j) \), assumed nonsingular (that is, its determinant \(|Λ|\) is not zero). See Exs. 9.21, 9.22 for further details, properties and comments.

9.5 Inequalities for moments and probabilities

There are a number of standard and useful inequalities concerning moments of a r.v., and probabilities of exceeding a given value. A few of
these will be given now, starting with a “translation” of the Hölder and Minkowski Inequalities (Theorems 6.4.2, 6.4.3) into the expectation notation.

**Theorem 9.5.1**  Suppose that $\xi, \eta$ are r.v.’s on $(\Omega, \mathcal{F}, P)$.

(i) (Hölder’s Inequality) If $E|\xi|^p < \infty$, $E|\eta|^q < \infty$ where $1 < p$, $q < \infty$, $1/p + 1/q = 1$, then $E|\xi \eta| < \infty$ and

$$|E\xi \eta| \leq E|\xi| \leq (E|\xi|^p)^{1/p} (E|\eta|^q)^{1/q}$$

with equality in the second inequality only if one of $\xi, \eta$ is zero a.s. or if $|\xi|^p = c|\eta|^q$ a.s. for some constant $c > 0$.

(ii) (Minkowski’s Inequality) If $E|\xi|^p < \infty$, $E|\eta|^p < \infty$ for some $p \geq 1$ then $E|\xi + \eta|^p < \infty$ and

$$(E|\xi + \eta|^p)^{1/p} \leq (E|\xi|^p)^{1/p} + (E|\eta|^p)^{1/p}$$

with equality (if $p > 1$) only if one of $\xi, \eta$ is zero a.s. or if $\xi = c\eta$ a.s. for some constant $c > 0$. For $p = 1$ equality holds if and only if $\xi \eta = 0$ a.s. (see also Ex. 9.19).

(iii) If $0 < p < 1$ and $E|\xi|^p < \infty$, $E|\eta|^p < \infty$, then $E|\xi + \eta|^p < \infty$ and $E|\xi + \eta|^p \leq E|\xi|^p + E|\eta|^p$, with equality iff $\xi \eta = 0$ a.s. (see also Ex. 9.19).

The norm notation – writing $||\xi||_p = (E|\xi|^p)^{1/p}$ – gives the neatest statements of the inequalities as in Section 6.4, in the case $p \geq 1$. For Hölder’s Inequality may be written as $||\xi \eta||_1 \leq ||\xi||_p ||\eta||_q$ and Minkowski’s Inequality as $||\xi + \eta||_p \leq ||\xi||_p + ||\eta||_p$.

The following result, mentioned in the previous section, is an immediate corollary of (i), and restates Theorem 6.4.8 (with $\mu(X) = 1$).

**Theorem 9.5.2**  If $\xi$ is a r.v. on $(\Omega, \mathcal{F}, P)$ and $E|\xi|^p < \infty$ for some $p > 0$, then $E|\xi|^q < \infty$ for $0 < q \leq p$, and $(E|\xi|^q)^{1/q} \leq (E|\xi|^p)^{1/p}$, i.e. $||\xi||_q \leq ||\xi||_p$.

In particular it follows that if $E\xi^2 < \infty$ then $E|\xi| < \infty$ and $(E\xi^2)^2 \leq (E|\xi|^2) \leq E\xi^2$ (which, of course, may be readily shown directly from $E(|\xi|^2 - E|\xi|^2)^2 \geq 0$).

Another very simple class of ("Markov type") inequalities relates probabilities such as $P\{\xi \geq a\}$, $P\{|\xi| \geq a\}$ etc., to moments of $\xi$. The following result gives typical examples of such inequalities.

**Theorem 9.5.3**  Let $g$ be a nonnegative, real-valued function on $\mathbb{R}$, and let $\xi$ be a r.v.
(i) If $g(x)$ is even, and nondecreasing for $0 \leq x < \infty$ then for all $a \geq 0$, with $g(a) \neq 0$,

$$P(|\xi| \geq a) \leq \mathbb{E}[g(\xi)]/g(a).$$

(ii) If $g$ is nondecreasing on $-\infty < x < \infty$ then for all $a$ with $g(a) \neq 0$,

$$P(\xi \geq a) \leq \mathbb{E}[g(\xi)]/g(a).$$

**Proof**  Note first that the monotonicity of $g$ in each case implies its (Borel) measurability (cf. Ex. 3.11). With $g$ as in (i) it is clear that $g(\xi(\omega))$ is defined and finite a.s. and is thus a (nonnegative) r.v. and

$$\mathbb{E}g(\xi) = \int g(\xi(\omega))dP(\omega) \geq \int_{\{\omega:|\xi(\omega)| \geq a\}} g(\xi(\omega))dP(\omega) \geq g(a)P(|\xi| \geq a),$$

since $g(\xi(\omega)) \geq g(a)$ if $|\xi(\omega)| \geq a$. Hence (i) is proved, and the proof of (ii) is similar. □

For an inequality in the opposite direction see Ex. 9.18.

**Corollary**  (i) If $\xi$ is any r.v. and $0 < p < \infty$, $a > 0$, then

$$P(|\xi| \geq a) \leq \mathbb{E}|\xi|^p/a^p.$$  

(ii) If $\xi$ is a r.v. with $\mathbb{E}\xi^2 < \infty$, then for all $a > 0$,

$$P(|\xi - \mathbb{E}\xi| \geq a) \leq \frac{\text{var}(\xi)}{a^2}. $$

The inequality in (i) (which follows by taking $g(x) = |x|^p$) is called “the” Markov Inequality. The case $p = 2$ in (i) is the well known Chebychev Inequality.

The final inequality, which is sometimes very useful, concerns convex functions of a r.v. We recall that a function $g$ defined on the real line is convex if $g(\lambda x + (1 - \lambda)y) \leq \lambda g(x) + (1 - \lambda)g(y)$ for any $x, y, \ 0 \leq \lambda \leq 1$. A convex function is known to be continuous and thus Borel measurable.

**Theorem 9.5.4** (Jensen’s Inequality)  If $\xi$ is a r.v. with $\mathbb{E}|\xi| < \infty$ and $g$ is a convex function on $\mathbb{R}$ such that $\mathbb{E}|g(\xi)| < \infty$, then

$$g(\mathbb{E}\xi) \leq \mathbb{E}g(\xi).$$

**Proof**  Since $g$ is convex it is known that given any $x_0$ there is a real number $h = h(x_0)$ such that $g(x) - g(x_0) \geq (x - x_0)h$ for all $x$. (This may be proved for example by showing that for all $x < x_0 < y$ we have
(\(g(x_0) - g(x))/(x_0 - x) \leq (g(y) - g(x_0))/(y - x_0)\) and taking \(h = \sup_{x < x_0} (g(x_0) - g(x))/(x_0 - x)\). Hence, putting \(x = \xi, x_0 = \mathbb{E}\xi\) we have, a.s.,

\[g(\xi) - g(\mathbb{E}\xi) \geq (\xi - \mathbb{E}\xi)h\ (h = h(\mathbb{E}\xi)).\]

The desired conclusion follows at once by taking expectations of both sides since the expectation of the right hand side is zero. □

### 9.6 Inverse functions and probability transforms

If \(F\) is a strictly increasing continuous function on the real line (or a sub-interval thereof) and \(a = \inf F(x), b = \sup F(x)\), then its inverse function \(F^{-1}\) is immediately defined for \(y \in (a, b)\) by \(F^{-1}(y) = x\), where \(x\) is the unique value such that \(F(x) = y\). Then \(F^{-1}(F(x)) = x\) for all \(x\) in the domain of \(F\) and \(F(F^{-1}(y)) = y\) for all \(y\) in the domain \((a, b)\), of \(F^{-1}\).

If \(F\) is strictly increasing but not everywhere continuous, \(F^{-1}(y)\) is not thus defined in this way for all \(y \in (a, b)\) e.g. if \(x_0\) is a discontinuity point of \(F\) and e.g. \(F(x_0) > F(x_0 - 0)\), there is no \(x\) for which \(F(x) = y\) if \(y \in (F(x_0 - 0), F(x_0))\). On the other hand, if \(F\) is continuous and nondecreasing but not strictly increasing, there is an interval \((x_1, x_2)\), on which \(F\) is constant, i.e. \(F(x) = y\) say for \(x_1 < x < x_2\). Hence there is no unique \(x\) for which \(F(x) = y\).

It is, however, useful to define an inverse function \(F^{-1}\) when \(F\) is nondecreasing (or nonincreasing) but not necessarily strictly monotone or continuous, and this may be done in various equally natural ways to retain some of the useful properties valid for the strictly monotone continuous case. We employ the following (commonly used) form of definition.

Let \(F\) be a nondecreasing function defined on an interval and for \(y \in (\inf F(x), \sup F(x))\) define \(F^{-1}(y)\) by

\[F^{-1}(y) = \inf\{x : F(x) \geq y\}.\]

To see the meaning of this definition it is helpful to visualize its value at points \(y \in (F(x_0 - 0), F(x_0 + 0))\) where \(F\) is discontinuous at \(x_0\) or at points \(y = F(x)\) for \(x\) such that \(F\) is constant in some neighborhood \((x - \epsilon, x + \epsilon)\). It is also helpful to determine the points \(x\) for which \(F^{-1}(F(x)) \neq x, y\) such that \(F(F^{-1}(y)) \neq y\). The following results are examples of many useful properties of this form of the inverse function, the proofs of which may be supplied as exercises by an interested reader.\(^1\)

\(^1\) Or see e.g. [Resnick, Section 0.2] for an excellent detailed treatment.
Lemma 9.6.1  If $F$ is a nondecreasing function on $\mathbb{R}$ with inverse $F^{-1}$ defined as above, then

(i) (a) $F^{-1}$ is nondecreasing and left-continuous ($F^{-1}(y^{-0}) = F^{-1}(y)$)
(b) $F^{-1}(F(x)) \leq x$
(c) If $F$ is strictly increasing from the left at $x$ in the sense that $F(a) < F(x)$ whenever $a < x$, then $F^{-1}(F(x)) = x$.

(ii) If $F$ is right-continuous then
(a) $\{x : F(x) \geq y\}$ is closed for each $y$
(b) $F(F^{-1}(y)) \geq y$
(c) $F^{-1}(y) \leq x$ if and only if $y \leq F(x)$
(d) $x < F^{-1}(y)$ if and only if $F(x) < y$.

(iii) If for a given $y$, $F$ is continuous at $F^{-1}(y)$ then $F(F^{-1}(y)) = y$. Hence if $F$ is everywhere continuous then $F(F^{-1}(y)) = y$ for all $y$.

Results of this type are useful for transformation of r.v.’s to standard distributions (“Probability transformations”). For example, it should be shown as an exercise (Ex. 9.4) that if $\xi$ has a continuous distribution function $F$, then $F(\xi)$ is a uniform r.v. and (Ex. 9.5) that if $\xi$ is a uniform r.v. and $F$ some d.f., then $\eta = F^{-1}(\xi)$ is a r.v. with d.f. $F$. Such results can be useful for simulation and sometimes allow the proof of properties of general r.v.’s to be done just under special assumptions such as uniformity, normality, etc.

We shall be interested later in the topic of “convergence in distribution” involving the convergence of d.f.’s $F_n$ to a d.f. $F$ at continuity points of the latter. The following result (which may be proved as an exercise or reference made to e.g. [Resnick]) involves the more general framework where the $F_n$’s need not be d.f.’s (and convergence at continuity points is then commonly referred to as vague convergence – cf. Section 11.3).

Lemma 9.6.2  If $F_n$, $n \geq 1$, $F$ are nondecreasing and $F_n(x) \to F(x)$ at all continuity points $x$ of $F$, then $F_n^{-1}(y) \to F^{-1}(y)$ at all continuity points $y$ of $F^{-1}$.

Exercises

9.1  Let $p_j \geq 0$, $\sum_{j=1}^{\infty} p_j = 1$, $x_j$ real, $F(x) = \sum_{x_j \leq x} p_j$. Show that $\nu(E) = \sum_{x_j \in E} p_j$ defines a measure on the Borel sets $\mathcal{B}$ and $\nu(E) = \mu_F(E)$ for $E \in \mathcal{B}$. (If $E = \bigcup_{k=1}^{\infty} E_k$ write $\chi_j = \chi_E(x_j)$, $\chi_{jk} = \chi_{E_k}(x_j)$ so that $\nu(E) = \sum \chi_j p_j$, $\nu(E_k) = \sum_j \chi_{jk} p_j$.) Thus for given $p_j \geq 0$, $\sum p_j = 1$, there is a discrete r.v. $\xi$ with $P[\xi = x_j] = p_j$ and $P[\xi \in E] = \sum_{x_j \in E} p_j$. 

9.2 Let $F$ be a d.f. and $F(x) = \int_{-\infty}^{x} f(t) \, dt$ where $f \in L_1(-\infty, \infty)$. (It is not initially assumed that $f \geq 0$.) Define the finite signed measure $\nu(E) = \int_E f \, dx$. Show that $\nu(E) = \mu_f(E)$ on the Borel sets $\mathcal{B}$. (Hint: Use Lemma 5.2.4.) Hence show that $f \geq 0$ a.e.

9.3 Let $\Omega$ be the unit interval, $\mathcal{F}$ its Borel subsets, and $P$ Lebesgue measure on $\mathcal{F}$. Let $\xi(\omega) = \omega$, $\eta(\omega) = 1 - \omega$. Show that $\xi, \eta$ have the same distribution but are not identical. In fact $P(\xi \neq \eta) = 1$.

9.4 Let $\xi$ be a r.v. whose d.f. $F$ is continuous. Let $\eta = F(\xi)$ (i.e. $\eta(\omega) = F(\xi(\omega))$). Show that $\eta$ is uniformly distributed on $(0, 1)$, i.e. that its d.f. $G$ is given by $G(x) = 0$ for $x < 0$, $G(x) = x$ for $0 \leq x \leq 1$ and $G(x) = 1$ for $x > 1$. What if $F$ is not continuous? (For simplicity assume $F$ has just one jump.)

9.5 Let $F$ be any d.f. and define its inverse $F^{-1}$ as in Section 9.6. Show that if $\xi$ is uniformly distributed over $(0, 1)$, then $\eta = F^{-1}(\xi)$ has d.f. $F$.

9.6 If $\xi, \eta$ are discrete r.v.'s, is $\xi + \eta$ discrete? What about $\xi \eta$ and $\xi \eta$? What happens to these combinations if $\xi$ is discrete and $\eta$ continuous?

9.7 Prove Lemma 9.3.1. (Hints: For (i) it may be noted that (a) every $\xi \in C$ is $\sigma(C)$-measurable and (b) if every $\xi \in C$ is $\sigma$-measurable (for some fixed $\sigma$-field $\mathcal{G}$) then $\mathcal{G} \supset \sigma(\xi)$, each $\xi \in C$. Clearly in (ii) the $\sigma$-field on the left contains that on the right. However, each $\xi_1$ is measurable with respect to the $\sigma$-field on the right, which therefore contains the smallest $\sigma$-field yielding measurability of all $\xi_1$, viz. $\sigma(C)$.)

9.8 In Theorem 9.3.3, the $\xi_i$ are all defined on the same subset of $\Omega$ (i.e. where $\xi$ is defined). If we start with mappings $\xi_1, \ldots, \xi_n$ defined (and finite a.s.) on possibly different subsets $D_1, \ldots, D_n$ (with $P(D_i) = 1$) we may define $\xi = (\xi_1, \ldots, \xi_n)$ on $D = \cap_i D_i$. If $\xi_1, \ldots, \xi_n$ are each r.v.'s then $\xi$ is a random vector, as in the theorem. Show that the converse may not be true, that is, if $\xi$ is a random vector, it is not necessarily true that the $\xi_i$ are r.v.'s (it is true if $D_i$ are measurable – e.g. if $P$ is complete).

9.9 Let $F$ be an absolutely continuous d.f. on $\mathbb{R}^n$ (with density $f(x_1, \ldots, x_n)$) for r.v.'s $\xi_1, \ldots, \xi_n$. Show that the r.v.'s $\xi_1, \ldots, \xi_k$ ($k < n$) have an absolutely continuous distribution and find their joint p.d.f.

9.10 The concept of a “continuous singular” d.f. or probability measure in $\mathbb{R}^2$ is more common than in $\mathbb{R}$. For example, let $F$ be any continuous d.f. on $\mathbb{R}$. For any Borel set $B$ in $\mathbb{R}^2$ define $\mu(B) = \mu_f(B^0)$ where $B^0$ is the section of $B$ defined by $y = 0$. Show that $\mu$ has no point atoms but is singular with respect to two-dimensional Lebesgue measure.

9.11 More generally suppose the $C$ is a simple curve in the plane given parametrically as $x = x(s)$, $y = y(s)$, where $x$ and $y$ are (Borel) measurable 1-1 functions of $s$. If $\mu$ is a probability measure on $(\mathbb{R}, \mathcal{B})$ we may define a probability measure on $(\mathbb{R}^2, \mathcal{B}^2)$ by $\nu(E) = \mu T^{-1}(E)$ where $T$ is the measurable transformation $T(s) = (x(s), y(s))$. The measure $\nu$ is singular with respect to Lebesgue measure and has no atoms if $\mu$ has no atoms. If $s$ is distance along the curve, $\nu(E)$ may be regarded as the $\mu$-measure of $E \cap C$. 
Let $F$ be a real-valued r.v. on $(\Omega, \mathcal{F}, P)$ and define $E_n = \{\omega : |\xi(\omega)| \geq n\}$. Show that

$$\sum_{n=1}^{\infty} P(E_n) \leq E|\xi| \leq 1 + \sum_{n=1}^{\infty} P(E_n)$$

and hence that $E|\xi| < \infty$ if and only if $\sum_{n=1}^{\infty} P(E_n) < \infty$. If $\xi$ takes only positive integer values, show that $E|\xi| = \sum_{n=1}^{\infty} P(E_n)$. (Hint: Let $F_n = \{\omega : n \leq |\xi(\omega)| < n + 1\}$ and note that $\sum_{n=1}^{\infty} nP(F_n) = \sum_{n=1}^{\infty} P(E_n)$.)

If $\xi$ is a nonnegative r.v. with d.f. $F$ show that

$$E\xi = \int_0^\infty [1 - F(x)] \, dx.$$  

(Hint: Use Fubini’s Theorem.) If $\xi$ is a real-valued r.v. with d.f. $F$ show that

$$E|\xi| = \int_{-\infty}^{0} F(x) \, dx + \int_{0}^{\infty} [1 - F(x)] \, dx$$

and thus $E|\xi| < \infty$ if and only if $\int_{-\infty}^{0} F(x) \, dx < \infty$ and $\int_{0}^{\infty} [1 - F(x)] \, dx < \infty$, in which case

$$E\xi = \int_{0}^{\infty} [1 - F(x)] \, dx - \int_{-\infty}^{0} F(x) \, dx.$$

Let $F$ be any d.f. Show that, for any $h > 0$, $\int_{-\infty}^{\infty} (F(x + h) - F(x)) \, dx = h$.

Why does this not contradict the obvious statement that $\int_{-\infty}^{\infty} F(x + h) \, dx = \int_{-\infty}^{\infty} F(x) \, dx$?

Let $g$ be a nonnegative bounded function on $\mathbb{R}$, and $\xi$ a r.v. If $g$ is even and nondecreasing on $0 < x < \infty$, show that

$$P(|\xi| \geq a) \geq E(g(\xi) - g(a))/M$$

for any $M < \infty$ such that $g(\xi(\omega)) \leq M$ a.s. (e.g. $M = \sup g(x)$). If $g$ is instead nondecreasing on $(-\infty, \infty)$ show that the same inequality holds with $\xi$ instead of $|\xi|$ on the left.
9.19 Let \( \xi, \eta \) be r.v.’s with \( E|\xi|^p < \infty, \ E|\eta|^p < \infty \). Show that for \( p > 0, \ E|\xi + \eta|^p \leq c_p \{E|\xi|^p + E|\eta|^p\} \) where \( c_p = 1 \) if \( 0 < p \leq 1 \), \( c_p = 2^{p-1} \) if \( p > 1 \). (Hint: \( (1+x)^p \leq c_p(1+x^p) \) for \( x \geq 0 \). Note equality when \( x = 0 \) for \( p \leq 1 \), and \( x = 1 \) for \( p > 1 \) and consider derivatives.)

9.20 Show that the covariance \( \gamma \) of two r.v.’s \( \xi_1, \xi_2 \) satisfies \( |\gamma| \leq \sigma_1 \sigma_2 \) where \( \sigma_i \) is the standard deviation of \( \xi_i, i = 1, 2, \) and hence that the correlation \( \rho \) satisfies \( |\rho| \leq 1 \). The parameters \( \gamma \) and especially \( \rho \) are regarded as simple measures of dependence of \( \xi_1, \xi_2 \). What is the value of \( \rho \) if \( \xi_1 = a \xi_2 \) (a) for some \( a > 0 \), (b) for \( a < 0 \) ?

9.21 Write down the covariance matrix \( \Lambda \) for a pair of r.v.’s \( \xi_1, \xi_2 \) in terms of their means \( \mu_1, \mu_2 \), standard deviations \( \sigma_1, \sigma_2 \) and correlation \( \rho \). Show that \( \Lambda \) is nonsingular if \( |\rho| < 1 \) and then obtain its inverse. Hence write down the joint p.d.f. of \( \xi_1 \) and \( \xi_2 \) in terms of \( \mu_i, \sigma_i, i = 1, 2, \rho \), when \( \xi_1 \) and \( \xi_2 \) are assumed to be jointly normal.

9.22 If \( \xi_1, \xi_2, \ldots, \xi_n \) are jointly normal, means \( \mu_i, 1 \leq i \leq n \), nonsingular covariance matrix \( \Lambda \), show that the members of any subgroup (e.g. \( \xi_1, \xi_2, \ldots, \xi_k, k \leq n \)) are jointly normal, writing down their covariance matrix in terms of \( \Lambda \).
10

Independence

10.1 Independent events and classes

Two events $A, B$ are termed independent if $P(A \cap B) = P(A) \cdot P(B)$. Physically this means (as can be checked by interpreting probabilities as long term frequencies) that the proportion of those times $A$ occurs, for which $B$ also occurs in many repetitions of the experiment $E$, is ultimately the same as the proportion of times $B$ occurs in all. That is, roughly “knowledge of the occurrence or not of $A$ does not affect the probability of $B$” (and conversely). We are, of course, interested primarily in the mathematical definition given, and its consequences.

The definition of independence can be usefully extended to a class of events. We say that $\mathcal{A}$ is a class of independent events (or that the events of a class $\mathcal{A}$ are independent) if for every finite subclass of distinct events $A_1, A_2, \ldots, A_n$ of $\mathcal{A}$, we have $P(\cap_{i=1}^n A_i) = \prod_{i=1}^n P(A_i)$. Note that it is not, in general, sufficient for this that the events of $\mathcal{A}$ be pairwise independent (see Ex. 10.1).

A more general notion concerns a family of independent classes. If $\mathcal{A}_\lambda$ is a class of events for each $\lambda$ in some index set $\Lambda$, $\{\mathcal{A}_\lambda : \lambda \in \Lambda\}$ is said to be a family of independent classes of events (or that the classes $\{\mathcal{A}_\lambda : \lambda \in \Lambda\}$ are independent), if for every choice of one member $A_\lambda$ from each $\mathcal{A}_\lambda$, the events $\{A_\lambda : \lambda \in \Lambda\}$ are independent.

Note that a class $\mathcal{A}$ of independent events may be regarded as a family of independent classes of events, where the classes of the family each consist of just one event of $\mathcal{A}$. This viewpoint is sometimes useful. Note also that while the index set $\Lambda$ may be infinite (of any order) a family $\mathcal{A} = \{\mathcal{A}_\lambda : \lambda \in \Lambda\}$ is independent if and only if every finite subfamily $\{\mathcal{A}_{\lambda_1}, \ldots, \mathcal{A}_{\lambda_n}\}$ is independent (for distinct $\lambda_i$). Thus it usually suffices to consider finite families.

Remark If $\mathcal{A}_1, \ldots, \mathcal{A}_n$ are classes of events such that each $\mathcal{A}_i$ contains a set $C_i$ with $P(C_i) = 1$ (e.g. $C_i = \Omega$) then to show that $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_n$ are
independent classes it is only necessary to show that \( P(\cap_i^n A_i) = \prod_i^n P(A_i) \) for this one \( n \), and all choices of \( A_i \in \mathcal{A}_i \), \( 1 \leq i \leq n \). For this relation then follows at once for subfamilies – e.g.

\[
\prod_{i=1}^{n-1} P(A_i) = \prod_{i=1}^{n-1} P(A_i) P(C_n) = P(\cap_i^{n-1} A_i) P(C_n) = P(\cap_i^{n-1} A_i) - P(\cap_i^{n-1} A_i) P(C_n) = P(\cap_i^{n-1} A_i)
\]

since \( P(C_n) = 0 \).

A family of independent classes may often be enlarged without losing independence. The following is a small result in this direction – its proof is left as an easy exercise (cf. Ex. 10.3).

**Lemma 10.1.1** Let \( \{\mathcal{A}_\lambda : \lambda \in \Lambda\} \) be independent classes of events, and \( \mathcal{A}_\lambda^* = \mathcal{A}_\lambda \cup \mathcal{G}_\lambda \) where, for each \( \lambda \), \( \mathcal{G}_\lambda \) is any class of sets \( E \) such that \( P(E) = 0 \) or 1. Then \( \{\mathcal{A}_\lambda^* : \lambda \in \Lambda\} \) are independent classes.

The next result is somewhat more sophisticated and very useful.

**Theorem 10.1.2** Let \( \{\mathcal{A}_\lambda : \lambda \in \Lambda\} \) be independent classes of events, and such that each \( \mathcal{A}_\lambda \) is closed under finite intersections. Let \( \mathcal{B}_\lambda \) be the \( \sigma \)-field generated by \( \mathcal{A}_\lambda \), \( \mathcal{B}_\lambda = \sigma(\mathcal{A}_\lambda) \). Then \( \{\mathcal{B}_\lambda : \lambda \in \Lambda\} \) are also independent classes.

**Proof** Define \( \mathcal{A}_\lambda^* = \mathcal{A}_\lambda \cup \{\Omega\} \). Then by Lemma 10.1.1 \( \{\mathcal{A}_\lambda^* : \lambda \in \Lambda\} \) are independent classes, and clearly \( \mathcal{B}_\lambda \) is also the \( \sigma \)-field generated by \( \mathcal{A}_\lambda^* \). Thus we assume without loss of generality that \( \Omega \in \mathcal{A}_\lambda \) for each \( \lambda \). In accordance with a remark above, it is sufficient to show that any finite subfamily \( \{\mathcal{B}_{\lambda_1}, \mathcal{B}_{\lambda_2}, \ldots, \mathcal{B}_{\lambda_k}\} \) (with distinct \( \lambda_i \)), are independent classes. If it is shown that \( \{\mathcal{B}_{\lambda_1}, \mathcal{A}_{\lambda_2}, \ldots, \mathcal{A}_{\lambda_k}\} \) are independent classes, the result will then follow inductively.

Let \( \mathcal{G} \) be the class of sets \( E \in \mathcal{F} \) such that \( P(E \cap A_2 \cap \ldots \cap A_n) = P(E)P(A_2) \ldots P(A_n) \) for all \( A_i \in \mathcal{A}_{\lambda_i} \) \( (i = 2, \ldots, n) \). If \( E \in \mathcal{G} \), \( F \in \mathcal{G} \) and \( E \supset F \), \( A_i \in \mathcal{A}_{\lambda_i} \) \( (i = 2, \ldots, n) \),

\[
P((E - F) \cap A_2 \cap \ldots \cap A_n)
= P(E \cap A_2 \cap \ldots \cap A_n) - P(F \cap A_2 \cap \ldots \cap A_n)
= P(E)P(A_2) \ldots P(A_n) - P(F)P(A_2) \ldots P(A_n)
= P(E - F)P(A_2) \ldots P(A_n).
\]
Thus \( E - F \in \mathcal{G} \) and \( \mathcal{G} \) is therefore closed under proper differences. Similarly it is easily checked that \( \mathcal{G} \) is closed under countable disjoint unions so that \( \mathcal{G} \) is a \( \mathcal{D} \)-class. But \( \mathcal{G} \supset \mathcal{A}_1 \), which is closed under intersections and hence by Theorem 1.8.5 (Corollary) \( \mathcal{G} \) contains the \( \sigma \)-ring generated by \( \mathcal{A}_1 \). This \( \sigma \)-ring is the \( \sigma \)-field \( \mathcal{B}_1 \) since \( \Omega \in \mathcal{A}_1 \) and hence \( \mathcal{G} \supset \mathcal{B}_1 \). Hence (using the Remark preceding Lemma 10.1.1) \( \{ \mathcal{B}_1, \mathcal{A}_2, \ldots, \mathcal{A}_n \} \) are independent classes and, as noted, this is sufficient for the result of the theorem.

If a class \( \mathcal{A} \) of independent events is regarded as a family of independent classes in the manner described above (i.e. each class consisting of one member of \( \mathcal{A} \)) we may, according to the theorem, enlarge each (1-member) class \( \{ A \} \) to the \( \sigma \)-field it generates, viz. \( \{ A, A^c, \Omega, \emptyset \} \). Thus these classes constitute, for \( A \in \mathcal{A} \), a family of independent classes. A class of independent events may now be obtained by selecting one event from each \( \{ A, A^c, \Omega, \emptyset \} \). Thus the following corollary to Theorem 10.1.2 holds.

**Corollary** If \( \mathcal{A} \) is a class of independent events, and if some of the events of \( \mathcal{A} \) are replaced by their complements, then the resulting class is again a class of independent events.

This result can, of course, be shown “by hand” from the definition. For example, if \( A, B \) are independent then it follows directly that so are \( A, B^c \) (which should be shown as an exercise).

The final result of this section is a useful extension of Theorem 10.1.2 involving the “grouping” of a family of independent classes. In this, by a *partition* of the set \( \Lambda \) we mean any class of disjoint sets \( \{ \Lambda_\gamma : \gamma \in \Gamma \} \) with \( \cup_{\gamma \in \Gamma} \Lambda_\gamma = \Lambda \). If \( \{ \mathcal{A}_\lambda : \lambda \in \Lambda \} \) are independent classes, clearly the “grouped classes” \( \{ \cup_{\lambda \in \Lambda_\gamma} \mathcal{A}_\lambda : \gamma \in \Gamma \} \) are independent. The following result shows that the same is true for \( \mathcal{B}_\gamma = \sigma(\cup_{\lambda \in \Lambda_\gamma} \mathcal{A}_\lambda) \), \( \gamma \in \Gamma \) provided each \( \mathcal{A}_\lambda \) is closed under finite intersections. This does not follow immediately from Theorem 10.1.2 since \( \cup_{\lambda \in \Lambda_\gamma} \mathcal{A}_\lambda \) need not be closed under intersections, but the classes may be expanded to have this closure property and allow application of the theorem.

**Theorem 10.1.3** Let \( \{ \mathcal{A}_\lambda : \lambda \in \Lambda \} \) be independent classes, each being assumed to be closed under finite intersections. Let \( \{ \Lambda_\gamma : \gamma \in \Gamma \} \) be a partition of \( \Lambda \), and \( \mathcal{B}_\gamma = \sigma(\cup_{\lambda \in \Lambda_\gamma} \mathcal{A}_\lambda) \). Then \( \{ \mathcal{B}_\gamma : \gamma \in \Gamma \} \) are independent classes.

**Proof** For each \( \gamma \in \Gamma \) let \( \mathcal{G}_\gamma \) denote the class of all sets of the form \( A_1 \cap A_2 \cap \ldots \cap A_n \), for \( A_1 \in \mathcal{A}_{\lambda_1} \), where \( \lambda_1, \ldots, \lambda_n \) are any distinct members of \( \Lambda_\gamma \) (\( n = 1, 2, \ldots \)). \( \mathcal{G}_\gamma \) is closed under finite intersections since each
10.2 Independent random elements

We will be primarily concerned with the concept of independence in the context of random variables. However, the definition and results of this section will apply more generally to arbitrary random elements, since this extra generality can be useful.

Specifically, suppose that for each \( \lambda \) in an index set \( \Lambda \), \( \xi_\lambda \) is a random element on a fixed probability space \((\Omega, \mathcal{F}, P)\), with values in a measurable space \((X_\lambda, S_\lambda)\) – which may change with \( \lambda \). (If \( \xi_\lambda \) is a r.v., of course, \( X_\lambda = \mathbb{R}^\ast \), \( S_\lambda = \mathcal{B}^\ast \).) If the classes \( \{\sigma(\xi_\lambda) : \lambda \in \Lambda\} \) are independent, then \( \{\xi_\lambda : \lambda \in \Lambda\} \) is said to be a family of independent r.e.’s or the r.e.’s \( \{\xi_\lambda : \lambda \in \Lambda\} \) are independent.

Since \( \sigma(\xi_\lambda) = \sigma(\xi_\lambda^{-1} S_\lambda) = \sigma(\xi_\lambda^{-1} B : B \in S_\lambda) \) and \( \xi_\lambda^{-1} S_\lambda \) is closed under intersections it follows at once from Theorem 10.1.2 that the following criterion holds – facilitating the verification of independence of r.e.’s.

**Theorem 10.2.1**  The r.e.’s \( \{\xi_\lambda : \lambda \in \Lambda\} \) are independent iff \( \{\xi_\lambda^{-1} S_\lambda : \lambda \in \Lambda\} \) are independent classes, i.e. iff for each \( n = 1, 2, \ldots, \) distinct \( \lambda_i \in \Lambda, B_i \in S_{\lambda_i}, \ 1 \leq i \leq n \)

\[
P\left(\bigcap_1^n \xi_{\lambda_i}^{-1} B_i\right) = \prod_1^n P\left(\xi_{\lambda_i}^{-1} B_i\right).\]

Indeed these conclusions hold if each \( S_\lambda \) is replaced by \( \mathcal{G}_\lambda \) where \( \mathcal{G}_\lambda \) is any class of subsets of \( X_\lambda \), closed under intersections and such that \( S(\mathcal{G}_\lambda) = S_\lambda \) for each \( \lambda \).

**Proof**  The main conclusion follows as noted prior to the statement of the theorem. The final conclusion follows by exactly the same pattern (see Ex. 10.9).

The above definition is readily extended to include independence of families of r.e.’s. Specifically, let \( C_\lambda \) be a family of random elements for each \( \lambda \) in an index set \( \Lambda \). Then if the \( \sigma \)-fields \( \{\sigma(C_\lambda) : \lambda \in \Lambda\} \) are independent classes of events, we shall say that \( \{C_\lambda : \lambda \in \Lambda\} \) are independent families of random elements, or “the classes \( C_\lambda \) of r.e.’s are independent for \( \lambda \in \Lambda \)”. 

\( \mathcal{A}_1 \) is so closed. Further \( \{G_\gamma : \gamma \in \Gamma\} \) are independent classes (which is easily checked from the definition of the sets of \( G_\gamma \)). Hence, by Theorem 10.1.2, the \( \sigma \)-fields \( \{\sigma(G_\gamma) : \gamma \in \Gamma\} \) are independent classes. But clearly \( \bigcup_{\lambda \in \Lambda} A_\lambda \subset G_\gamma \) so that \( B_\gamma \subset \sigma(G_\gamma) \) and hence \( \{B_\gamma : \gamma \in \Gamma\} \) are independent classes, as required. □
Thus we have the notions of independence for random elements, and for families of r.e.’s, parallel to the corresponding notions for events and classes of events. (However, see Ex. 10.10.) Theorem 10.1.3 has the following obvious (and useful) analog for independent random elements.

**Theorem 10.2.2**  Let \( \{C_\lambda : \lambda \in \Lambda\} \) be independent families of random elements on a space \((\Omega, \mathcal{F}, P)\), let \( \{\Lambda_\gamma : \gamma \in \Gamma\} \) be a partition of \( \Lambda \), and write \( \mathcal{H}_\gamma = \cup_{\lambda \in \Lambda_\gamma} C_\lambda \). Then \( \{\mathcal{H}_\gamma : \gamma \in \Gamma\} \) are independent families of random elements.

**Proof**  From Lemma 9.3.1 (iii) we have

\[ \sigma(\mathcal{H}_\gamma) = \sigma(\cup_{\lambda \in \Lambda_\gamma} \sigma(C_\lambda)). \]

But since \( \{\sigma(C_\lambda) : \lambda \in \Lambda\} \) are independent classes (each closed under intersections), it follows from Theorem 10.1.3 that \( \{\sigma(\mathcal{H}_\gamma) : \gamma \in \Gamma\} \) are also independent classes.

The following result gives a useful characterization of independence of r.e.’s in terms of product forms for the distributions of finite subfamilies. This is especially important for the case of r.v.’s considered in the next section.

**Theorem 10.2.3**  Let \( \xi_1, \xi_2, \ldots, \xi_n \) be r.e.’s on \((\Omega, \mathcal{F}, P)\) with values in measurable spaces \((X_i, S_i), 1 \leq i \leq n\). Then \( \xi = (\xi_1, \xi_2, \ldots, \xi_n) \) is a r.e. on \((\Omega, \mathcal{F}, P)\) with values in \((\prod_1^n X_i, \prod_1^n S_i)\), and \( \xi_1, \ldots, \xi_n \) are independent iff

\[
P_{\xi^{-1}} = P_{\xi_1^{-1}} \times P_{\xi_2^{-1}} \times \ldots \times P_{\xi_n^{-1}} \left( = \prod_1^n P_{\xi_i^{-1}} \right)
\]

i.e. the distribution of \( \xi \) is the product (probability) measure having the individual distributions as components.

Thus, for a general index set \( \Lambda \), r.e.’s \( (\xi_\lambda : \lambda \in \Lambda) \) are independent iff the distribution of \( \xi = (\xi_{\lambda_1}, \ldots, \xi_{\lambda_n}) \) factors in the above manner for each \( n \) and choice of distinct \( \lambda_i \).

**Proof**  That \( \xi = (\xi_1, \ldots, \xi_n) \) is a r.e. follows simply (as in Theorem 9.3.3 for the special case of random variables and vectors) and

\[
\xi^{-1}(B_1 \times B_2 \times \ldots \times B_n) = \cap_1^n \xi_i^{-1}(B_i)
\]

for any \( B_i \in S_i, 1 \leq i \leq n \). Thus if \( \xi_i \) are independent, \( P_{\xi^{-1}}(B_1 \times B_2 \times \ldots \times B_n) = \prod_1^n P_{\xi_i^{-1}}B_i \) so that \( P_{\xi^{-1}} \) and the product measure \( \prod_1^n P_{\xi_i^{-1}} \) agree
on measurable rectangles and hence on all sets of $\prod_i S_i$. Conversely if $P_{\xi^{-1}} = \prod_i P_{\xi^{-1}_i}$

$$P \left( \bigcap_{i=1}^n \xi_i^{-1} B_i \right) = P_{\xi^{-1}}(B_1 \times B_2 \times \ldots \times B_n) = \prod_{i=1}^n P_{\xi_i^{-1}}(B_i).$$

As noted the same relation is automatic for subclasses of $(\xi_1, \xi_2, \ldots, \xi_n)$ by writing appropriate $B_i = X_i$, so that independence of $(\xi_1, \ldots, \xi_n)$ follows. □

### 10.3 Independent random variables

The independence properties developed in the last section, of course, apply in particular to random variables, as will be seen in the following results. For simplicity these are mainly stated for finite families since the results for infinite families involve just finite subfamilies.

**Theorem 10.3.1** The following conditions are each necessary and sufficient for independence of r.v.'s $\xi_1, \xi_2, \ldots, \xi_n$ (on a probability space $(\Omega, \mathcal{F}, P)$).

(i) $P(\bigcap_{i=1}^n \xi_i^{-1} B_i) = \prod_{i=1}^n P(\xi_i^{-1} B_i)$ for every choice of extended Borel sets $B_1, \ldots, B_n$.

(ii) (i) holds for all choices of (ordinary) Borel sets $B_1, \ldots, B_n$ (in place of all extended Borel sets).

(iii) The distribution $P_{\xi^{-1}}$ of the random vector $\xi = (\xi_1, \xi_2, \ldots, \xi_n)$ on $(\mathbb{R}^n, \mathcal{B}^n)$ (or $(\mathbb{R}^n, \mathcal{B}'^n)$) is the product of the distributions $P_{\xi_i^{-1}}$ on $(\mathbb{R}, \mathcal{B})$ (or $(\mathbb{R}^*, \mathcal{B}'^*)$), i.e.

$$P_{\xi^{-1}} = P_{\xi_1^{-1}} \times P_{\xi_2^{-1}} \times \ldots \times P_{\xi_n^{-1}}.$$

(iv) The joint d.f. $F_{1,\ldots,n}(x_1, \ldots, x_n)$ of $\xi_1, \ldots, \xi_n$ factors as $\prod_{i=1}^n F_i(x_i)$, where $F_i$ is the d.f. of $\xi_i$.

**Proof** Independence of $(\xi_1, \xi_2, \ldots, \xi_n)$ is readily seen to be equivalent to each of (i)–(iii) using Theorem 10.2.3. (iii) at once implies (iv), and that (iv) implies e.g. (iii) is readily checked. □

The next result is a useful application of Theorem 10.2.2.

**Theorem 10.3.2** Let $(\xi_{i1}, \ldots, \xi_{in_1}, \xi_{i2}, \ldots, \xi_{in_2}, \xi_{i3}, \ldots)$ be independent r.v.'s on a space $(\Omega, \mathcal{F}, P)$. Define random vectors $\xi_1, \xi_2, \ldots, \xi_i$ by $\xi_i = (\xi_{i1}, \xi_{i2}, \ldots, \xi_{in_i})$. Then $(\xi_1, \xi_2, \ldots)$ are independent random vectors. Moreover if $\phi_i$
is a finite-valued measurable function on \((\mathbb{R}^{*n}, \mathcal{B}^{*n})\) for \(i = 1, 2, \ldots\), and \(\eta_i = \phi_i(\xi_i)\), then \((\eta_1, \eta_2, \ldots)\) are independent r.v.’s.

**Proof**  By Theorem 10.2.2 \(\{(\xi_{1i}, \xi_{2i}, \ldots, \xi_{ni}) : i = 1, 2, \ldots\}\) are independent families of r.v.’s so that \(\{\sigma(\xi_{1i}, \ldots, \xi_{ni}) : i = 1, 2, \ldots\}\) are independent classes of events. But, by Lemma 9.3.2, \(\sigma(\xi_i) = \sigma(\xi_{1i}, \ldots, \xi_{ni})\) so that \((\xi_{1i}, \xi_{2i}, \ldots)\) are independent random vectors, as required.

Further, a typical generating set of \(\sigma(\eta_i)\) is \(\eta_i^{-1}B\) for \(B \in \mathcal{B}\). But \(\eta_i^{-1}B = \xi_i^{-1}(\phi_i^{-1}B) \in \sigma(\xi_i)\) so that \(\sigma(\eta_i) \subset \sigma(\xi_i)\). Since \(\{\sigma(\xi_i) : i = 1, 2, \ldots\}\) are independent classes, so are the classes \(\{\sigma(\eta_i), i = 1, 2, \ldots\}\), i.e. \((\eta_1, \eta_2, \ldots)\) are independent r.v.’s, completing the proof. □

**Corollary**  The theorem remains true if the \(\phi_i\) are defined only on (measurable) subsets \(D_i \subset \mathbb{R}^{*n_i}\) such that \(\xi_i \in D_i\) a.s. (so that \(\eta_i\) may be defined at fewer \(\omega\)-points than \(\xi_i\) – though still a.s.). In particular, the theorem holds if \(D_i = \mathbb{R}^{n_i}\) i.e. if the \(\phi_i\) are defined for finite values of their arguments only – the case of practical importance.

**Proof**  Define \(\phi_i^* = \phi_i\) on (the measurable set) \(D_i\) and zero on \(\mathbb{R}^{*n_i} - D_i\). Then if \(\eta_i^* = \phi_i^*\xi_i\) we have \(\eta_i^* = \eta_i\) a.s. Since \((\eta_1^*, \eta_2^*, \ldots)\) are independent by the theorem, so are \((\eta_1, \eta_2, \ldots)\) (Ex. 10.11). □

The next result concerns the existence of a sequence of independent r.v.’s with given d.f.’s.

**Theorem 10.3.3**  Let \(F_i\) be a d.f. for each \(i = 1, 2, \ldots\). Then there is a probability space \((\Omega, \mathcal{F}, P)\) and a sequence \((\xi_1, \xi_2, \ldots)\) of independent r.v.’s such that \(\xi_i\) has d.f. \(F_i\).

**Proof**  Write \(\mu_i\) for the Lebesgue–Stieltjes (probability) measure on \((\mathbb{R}, \mathcal{B})\) corresponding to \(F_i\). Then by Theorem 7.10.4, there exists a probability measure \(P\) on \((\mathbb{R}^{*\infty}, \mathcal{B}^{*\infty})\) such that for any \(n\), Borel sets \(B_1, B_2, \ldots, B_n\),

\[
P(B_1 \times B_2 \times \cdots \times B_n \times \mathbb{R} \times \mathbb{R} \times \cdots) = \prod_{i=1}^{n} \mu_i(B_i).
\]

Write \((\Omega, \mathcal{F}, P)\) for the probability space \((\mathbb{R}^{*\infty}, \mathcal{B}^{*\infty}, P)\) and define \(\xi_1, \xi_2, \ldots\) on this space by \(\xi_i\omega = x_i\) when \(\omega = (x_1, x_2, x_3, \ldots)\). Each \(\xi_i\) is clearly a r.v. and for Borel sets \(B_1, B_2, \ldots, B_n\)

\[
P(\cap_{i=1}^{n} \xi_i^{-1}(B_i)) = P(B_1 \times B_2 \times \cdots \times B_n \times \mathbb{R} \times \mathbb{R} \times \cdots) = \prod_{i=1}^{n} \mu_i(B_i).
\]

In particular, \(B_1 = B_2 = \cdots = B_{n-1} = \mathbb{R}\) gives \(P(\xi_n^{-1}B_n) = \mu_n(B_n)\) for each \(n\) so that (writing \(i\) for \(n\)) \(P(\cap_{i=1}^{n} \xi_i^{-1}B_i) = \prod_{i=1}^{n} P(\xi_i^{-1}B_i)\) and hence the \(\xi_i\) are
10.3 Independent random variables

Let $\xi_1, \xi_2$ be independent r.v.'s with d.f.'s $F_1, F_2$ and let $h$ be a finite measurable function on $(\mathbb{R}^2, \mathcal{B}^2)$. Then $h(\xi_1, \xi_2)$ is a r.v. and

$$E h(\xi_1, \xi_2) = \int_{\Omega} \int_{\Omega} h(\xi_1(\omega_1), \xi_2(\omega_2)) \, dP(\omega_1) \, dP(\omega_2) = \int_{\mathbb{R}} \int_{\mathbb{R}} h(x_1, x_2) \, dF_1(x_1) \, dF_2(x_2),$$

whenever $h$ is nonnegative, or $E|h(\xi_1, \xi_2)| < \infty$.

**Proof** It is clear that $h(\xi_1, \xi_2)$ is a r.v. Writing $\xi = (\xi_1, \xi_2)$ we have

$$E h(\xi_1, \xi_2) = \int_{\Omega} h(\xi(\omega)) \, dP(\omega) = \int_{\mathbb{R}^2} h(x_1, x_2) \, dP_{\xi_1}^{-1}(x_1, x_2) = \int_{\mathbb{R}^2} h(x_1, x_2) \, d(P_{\xi_1}^{-1} \times P_{\xi_2}^{-1})$$

by Theorem 4.6.1 and Theorem 10.3.1 (iii). Fubini’s Theorem (the appropriate version according as $h$ is nonnegative, or $h(\xi_1, \xi_2) \in L_1$) now gives the repeated integral

$$E h(\xi_1, \xi_2) = \int_{\mathbb{R}} \int_{\mathbb{R}} h(x_1, x_2) \, dP_{\xi_1}^{-1}(x_1) \, dP_{\xi_2}^{-1}(x_2)$$

which may be written either as $\int_{\mathbb{R}} \int_{\mathbb{R}} h(x_1, x_2) \, dF_1(x_1) \, dF_2(x_2)$ or, by Theorem 4.6.1 applied in turn to each of $\xi_1, \xi_2$, as $\int_{\Omega} \int_{\Omega} h(\xi_1(\omega_1), \xi_2(\omega_2)) \, dP(\omega_1) \, dP(\omega_2)$. Hence the result follows.

**Theorem 10.3.5** Let $\xi_1, \ldots, \xi_n$ be independent r.v.'s with $E|\xi_i| < \infty$ for each $i$. Then $E|\xi_1 \xi_2 \ldots \xi_n| < \infty$ and $E(\xi_1 \xi_2 \ldots \xi_n) = \prod_1^n E\xi_i$.

**Proof** Since by Theorem 10.3.2, $\xi_1$ and $(\xi_2 \xi_3 \ldots \xi_n)$ are independent the result will follow inductively from that for $n = 2$. The $n = 2$ result follows at once from Theorem 10.3.4 first with $h(x_1, x_2) = |x_1 x_2|$ to give

$$E|\xi_1 \xi_2| = \int_{\Omega} \int_{\Omega} |\xi_1(\omega_1)||\xi_2(\omega_2)| \, dP(\omega_1) \, dP(\omega_2) = E|\xi_1|E|\xi_2| < \infty,$$

and then with $h(x_1, x_2) = x_1 x_2$ to give $E(\xi_1 \xi_2) = E\xi_1 E\xi_2$. 

Note that a more general result of this kind, where the $\xi_i$ need not be independent, will be indicated in Chapter 15 for Stochastic Process Theory.
Corollary If $\xi_1, \ldots, \xi_n$ are independent r.v.’s with $\mathbb{E} \xi_i^2 < \infty$ for each $i$, then the variance of $(\xi_1 + \xi_2 + \cdots + \xi_n)$ is given by

$$\text{var}(\xi_1 + \xi_2 + \cdots + \xi_n) = \text{var}(\xi_1) + \text{var}(\xi_2) + \cdots + \text{var}(\xi_n).$$

The simple proof is left as an exercise.

10.4 Addition of independent random variables

We next obtain the distribution and d.f. of the sum of independent r.v.’s.

Theorem 10.4.1 Let $\xi_1, \xi_2$ be independent r.v.’s with distributions $P_{\xi_1}^{-1} = \pi_1$, $P_{\xi_2}^{-1} = \pi_2$. Then

(i) The distribution $\pi$ of $\xi_1 + \xi_2$ is given for Borel sets $B$ (writing $B - y = \{x - y : x \in B\}$) by

$$\pi(B) = \int_{-\infty}^{\infty} \pi_1(B - y) d\pi_2(y) = \int_{-\infty}^{\infty} \pi_2(B - y) d\pi_1(y) = \pi_1 \ast \pi_2(B),$$

where $\pi_1 \ast \pi_2$ is called the convolution of the measures $\pi_1, \pi_2$ (cf. Section 7.6).

(ii) In particular the d.f. $F$ of $\xi_1 + \xi_2$ is given in terms of the d.f.’s $F_1, F_2$ of $\xi_1, \xi_2$ by

$$F(x) = \int_{-\infty}^{\infty} F_1(x - y) dF_2(y) = \int_{-\infty}^{\infty} F_2(x - y) dF_1(y) = F_1 \ast F_2(x)$$

where $F_1 \ast F_2$ is the (Stieltjes) convolution of $F_1$ and $F_2$.

(iii) If $F_1$ is absolutely continuous with density $f_1$, $F$ is then absolutely continuous with density $f(x) = \int f_1(x - y) dF_2(y)$.

(iv) If also $F_2$ is absolutely continuous (with density $f_2$) then

$$f(x) = \int_{-\infty}^{\infty} f_1(x - y)f_2(y) dy = \int_{-\infty}^{\infty} f_2(x - y)f_1(y) dy = f_1 \ast f_2(x),$$

i.e. the convolution of $f_1$ and $f_2$ (cf. Section 7.6).

Proof If $\phi(x_1, x_2) = x_1 + x_2$ (measurable) and $\xi = (\xi_1, \xi_2)$, we have

$$\pi(B) = P[\xi_1 + \xi_2 \in B] = P[\phi \xi \in B] = P[\xi \in \phi^{-1} B] = \mathbb{E} \chi_{\phi^{-1} B}(\xi) = \int_{\mathbb{R}} \int_{\mathbb{R}} \chi_{\phi^{-1} B}(x_1, x_2) d\pi_1(x_1) d\pi_2(x_2).$$
by Theorem 10.3.4. The integrand is one if \( x_1 + x_2 \in B \), i.e. if \( x_1 \in B - x_2 \), and zero otherwise, so that the inner integral is \( \pi_1(B - x_2) \), measurable by Fubini’s Theorem giving the first result for \( \pi(B) \). The second follows similarly. Thus (i) holds.

The expressions for \( F(x) \) in (ii) follow at once by writing \( B = (-\infty, x] \), where e.g. \( \pi_1(B - y) = F_1(x - y) \) etc.

If \( F_1 \) is absolutely continuous with density \( f_1 \) we have

\[
F(x) = \int_{x}^{\infty} F_1(x - y) \, dF_2(y) = \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{x} f_1(t) \, dt \right\} \, dF_2(y)
\]

by the transformation \( t = u - y \) for fixed \( y \) in the inner integral. Thus

\[
F(x) = \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{x} f_1(u - y) \, dF_2(y) \right\} \, du
\]

by Fubini’s Theorem for nonnegative functions. That is \( F(x) = \int_{-\infty}^{x} f(u) \, du \) where \( f(u) = \int_{-\infty}^{\infty} f_1(u - y) \, dF_2(y) \). It is easily seen that the (nonnegative) function \( f \) is in \( L_1(-\infty, \infty) \) (Lebesgue measure) and thus provides a density for \( F \). Hence (iii) follows, and (iv) is immediate from (iii). □

10.5 Borel–Cantelli Lemma and zero-one law

We recall that if \( A_n \) is any sequence of subsets of the space \( \Omega \), then \( A = \lim A_n = \cap_{n=1}^{\infty} \cup_{m=n}^{\infty} A_m \) is the set of all \( \omega \in \Omega \) which belong to \( A_n \) for infinitely many values of \( n \).

If \( A_n \) are measurable sets (i.e. events), so is \( A \). In intuitive terms, \( A \) occurs if infinitely many of the \( A_n \) occur (simultaneously) when the underlying experiment is performed. The following result gives a simple but very useful condition under which \( P(A) = 0 \), i.e. with probability one only a finite number of \( A_n \) occur.

**Theorem 10.5.1** (Borel–Cantelli Lemma) Let \( \{A_n\} \) be a sequence of events of the probability space \( (\Omega, F, P) \), and \( A = \lim A_n \). If \( \sum_{n=1}^{\infty} P(A_n) < \infty \), then \( P(A) = 0 \).

**Proof** \( P(A) = P(\cap_{n=1}^{\infty} \cup_{m=n}^{\infty} A_m) \leq P(\cup_{m=1}^{\infty} A_m) \) for any \( n = 1, 2, \ldots \) Hence \( P(A) \leq \sum_{m=1}^{\infty} P(A_m) \) for all \( n \), and this tends to zero as \( n \to \infty \) since \( \sum P(A_n) \) converges. Thus \( P(A) = 0 \). □

The converse result is not true in general (Ex. 10.12). However, it is true if the events \( A_n \) form an independent sequence. Indeed, rather more is then true as the following result shows.
Theorem 10.5.2 (Borel–Cantelli Lemma for Independent Events) Let \( \{A_n\} \) be an independent sequence of events on \((\Omega, \mathcal{F}, P)\), and \( A = \lim A_n \). Then \( P(A) \) is zero or one, according as \( \sum_1^\infty P(A_n) < \infty \) or \( \sum_1^\infty P(A_n) = \infty \).

Proof Since \( P(A) = 0 \) when \( \sum P(A_n) < \infty \) it will be sufficient to show that \( P(A) = 1 \) when \( \sum P(A_n) = \infty \). Suppose, then, that \( \sum P(A_n) = \infty \).

Now

\[
P((\bigcup_{m=n}^k A_m)^c) = P(\bigcap_{m=n}^k A_m^c) = \prod_{m=n}^k P(A_m^c),
\]

since the events \( A_n^c, A_{n+1}^c, \ldots, A_k^c \) are independent by Theorem 10.1.2 (Corollary). Thus

\[
P((\bigcup_{m=n}^k A_m)^c) = \prod_{m=n}^k (1 - P(A_m)) \leq \prod_{m=n}^k e^{-P(A_m)}
\]

(by using \( 1 - x \leq e^{-x} \) for all \( 0 \leq x \leq 1 \)). The latter term is \( e^{-\sum_{m=n}^k P(A_m)} \) which tends to zero as \( k \to \infty \) since \( \sum P(A_m) = \infty \). Thus \( \lim_{k \to \infty} P(\bigcup_{m=n}^k A_m) = 1 \), giving \( P(A) = 1 \), as required. \( \Box \)

Note (though not shown here) that this result is in fact true if the \( A_n \) are only assumed to be pairwise independent. (See, for example, Theorem 4.3.2 of [Chung].)

The above theorem states in particular that a certain event \( A \) must have probability zero or one. Results of such a kind are therefore often referred to as “zero-one laws”. A particularly well known result of this type is the “Kolmogorov Zero-One Law”, which is shown next. Theorem 10.5.2 is an example of a zero-one law, together with necessary and sufficient conditions for the two alternatives.

First we require some general terminology. If \( \mathcal{F}_n \) is a sequence of sub-\( \sigma \)-fields of \( \mathcal{F} \), then the \( \sigma \)-fields \( \mathcal{G}_n = \sigma(\bigcup_{k=n+1}^\infty \mathcal{F}_k) \) form a decreasing sequence \( (\mathcal{G}_n \supset \mathcal{G}_{n+1}) \) whose intersection \( \bigcap_1^\infty \mathcal{G}_n = \mathcal{F}_\infty \) (clearly a \( \sigma \)-field) is called the tail \( \sigma \)-field of the sequence \( \mathcal{F}_n \). Sets of \( \mathcal{F}_\infty \) are called tail events and \( \mathcal{F}_\infty \)-measurable functions are called tail functions (or tail r.v.’s if defined and finite a.s.).

Theorem 10.5.3 (Kolmogorov Zero-One Law) Let \((\Omega, \mathcal{F}, P)\) be a probability space. If \( \mathcal{F}_n \) is a sequence of independent sub-\( \sigma \)-fields of \( \mathcal{F} \), then each tail event has probability zero or one, and each tail r.v. is constant a.s.
Proof Write \( H_n = \sigma(\cup_1^n F_i) \) and, as above, \( G_n = \sigma(\cup_{k=n+1}^\infty F_k) \). Then since each \( F_i \) is closed under intersections, it follows simply from Theorem 10.1.3 that \( H_n \) and \( G_n \) are independent classes. Since \( G_n \supseteq F_\infty \), it follows that \( H_n \) and \( F_\infty \) are independent. Now \( \cup_1^\infty H_n \) is a field (note that \( H_n \) is nondecreasing), and hence closed under intersections, so that by Theorem 10.1.2, \( F_\infty \) and \( \sigma(\cup_1^\infty H_n) \) are independent. But clearly \( \sigma(\cup_1^\infty H_n) \supseteq \sigma(\cup_1^\infty F_\infty) = G_0 \supseteq F_\infty \), so that \( F_\infty \) and \( \cup_1^\infty H_n \) are independent. Now \( \cup_1^\infty H_n \) is a field, and hence closed under intersections, so that by Theorem 10.1.2, \( F_\infty \) and \( \sigma(\cup_1^\infty H_n) \) are independent. But clearly \( \sigma(\cup_1^\infty H_n) \supseteq \sigma(\cup_{k=n+1}^\infty F_k) \), so that \( \cup_{k=n+1}^\infty F_k \) is independent. Finally suppose that \( \xi \) is a tail r.v. with d.f. \( F \). For any \( x \), \( \{ \omega : \xi(\omega) \leq x \} \) is a tail event and hence has probability zero or one, i.e. \( F(x) = 0 \) or 1. Since \( F \) is not identically either zero or one it must have a unit jump at a finite point \( a (= \inf(x : F(x) = 1)) \) so that \( P(\xi = a) = 1 \). □

**Corollary 1** Let \( \{ \xi_n : n = 1, 2, \ldots \} \) be a sequence of independent r.v.'s and define the tail \( \sigma \)-field \( F_\infty = \cap_{n=0}^\infty \sigma(\xi_{n+1}, \xi_{n+2}, \ldots) \). Then each tail event has probability zero or one, and each tail r.v. is constant a.s.

*Proof* Identify \( F_n \) with \( \sigma(\xi_n) \) and hence \( G_n = \sigma(\cup_{k=n+1}^\infty \sigma(\xi_k)) = \sigma(\xi_{n+1}, \xi_{n+2}, \ldots) \). □

**Corollary 2** If \( \{ C_n : n = 1, 2, \ldots \} \) is a sequence of independent classes of r.v.'s, the conclusion of the theorem holds, with tail \( \sigma \)-field \( F_\infty = \cap_{n=0}^\infty \sigma(\cup_{k=n+1}^\infty C_k) \).

Corollary 2, which follows by identifying \( F_n \) with \( \sigma(C_n) \), and hence \( G_n \) with \( \sigma(\cup_{k=n+1}^\infty \sigma(C_k)) = \sigma(\cup_{k=n+1}^\infty C_k) \) includes a zero-one law for an independent sequence of stochastic processes.

**Exercises**

10.1 Let \( \Omega \) consist of the integers \( \{1, 2, \ldots, 9\} \) with probabilities \( 1/9 \) each. Show that the events \( \{1, 2, 3\}, \{1, 4, 5\}, \{2, 4, 6\} \) are pairwise independent, but not independent as a class.

10.2 Construct an example of three events \( A, B, C \) which are not independent but which satisfy \( P(A \cap B \cap C) = P(A)P(B)P(C) \).

10.3 Let \( \{ A_\lambda : \lambda \in \Lambda \} \) be a family of independent classes of events. Show that arbitrary events of probability zero or one may be added to any or all \( A_\lambda \) while still preserving independence. Show that if \( B_\lambda \) is formed from \( A_\lambda \)
by including (i) all proper differences of two sets of \( \mathcal{A}_i \), (ii) all countable disjoint unions of sets of \( \mathcal{A}_i \), or (iii) all limits of monotone sequences of sets of \( \mathcal{A}_i \) then \( \{ \mathcal{B}_i : \lambda \in \Omega \} \) is a family of independent classes. (Hint: Consider a finite index set \( \Lambda, \Omega \in \mathcal{A}_i \) and show that independence is preserved when just one \( \mathcal{A}_i \) is replaced by \( \mathcal{B}_i \).

10.4 If \( E_1, E_2, \ldots, E_n \) are independent, show that

\[
\sum_{j=1}^{n} P(E_j) - \sum_{j \neq k} P(E_j)P(E_k) \leq P(\cup_{j=1}^{n} E_j) \leq \sum_{j=1}^{n} P(E_j).
\]

If the events \( E^{(n)}_1, \ldots, E^{(n)}_n \) change with \( n \) so that \( \sum_{j=1}^{n} P(E^{(n)}_j) \to 0 \), show that \( P(\cup_{j=1}^{n} E^{(n)}_j) \sim \sum_{j=1}^{n} P(E^{(n)}_j) \) as \( n \to \infty \).

10.5 Let \( \xi, \eta \) be independent r.v.’s with \( \mathcal{E}|\xi| < \infty \). Show that, for any Borel set \( B \),

\[
\int_{\mathcal{E}} \xi \, dP = \mathcal{E} \xi \, P(\eta \in B).
\]

10.6 Let \( \xi, \eta \) be random variables on the probability space \( (\Omega, \mathcal{F}, P) \), let \( E \in \mathcal{F} \), and let \( f \) be a Borel measurable function on the plane. If \( \xi \) is independent of \( \eta \) and \( E \) (i.e. if the classes of events \( \sigma(\xi) \) and \( \sigma(\eta, E) \) are independent) show that

\[
\int_{E} \int_{\Omega} f(\xi(\omega_1), \eta(\omega_2)) \, dP(\omega_1) \, dP(\omega_2) = \int_{E} f(\xi(\omega), \eta(\omega)) \, dP(\omega)
\]

whenever \( f \) is nonnegative or \( \mathcal{E}|f(\xi, \eta)| < \infty \). (Hint: Prove this first for an indicator function \( f \).) If the random variable \( \zeta \) defined on the probability space \( (\Omega', \mathcal{F}', P') \) has the same distribution as \( \xi \), show that

\[
\int_{E} \int_{\Omega'} f(\zeta(\omega'), \eta(\omega)) \, dP'(\omega') \, dP(\omega) = \int_{E} f(\xi(\omega), \eta(\omega)) \, dP(\omega).
\]

10.7 For \( n = 1, 2, \ldots \) let \( R_n(x) \) be the Rademacher functions \( R_n(x) = +1 \) or \(-1\) according as the integer \( k \) for which \( \frac{k-1}{2n} < x \leq \frac{k}{2n} \) \((0 \leq x \leq 1) \) is odd or even. Let \( (\Omega, \mathcal{F}, P) \) be the “unit interval probability space” (consisting of the unit interval, Lebesgue measurable sets and Lebesgue measure). Prove that \( \{R_n, n = 1, 2, \ldots\} \) are independent r.v.’s with the same d.f. Show that any two of \( R_1, R_2, R_1R_2 \) are independent, but the three together are not.

10.8 A r.v. \( \eta \) is called symmetric if \( \eta \) and \(-\eta \) have the same distribution. Let \( \xi \) be a r.v. Let \( \xi_1 \) and \( \xi_2 \) be two independent r.v.’s each having the same distribution as \( \xi \) and let \( \xi^* = \xi_1 - \xi_2 \).

(a) Show that \( \xi^* \) is symmetric (it is called the symmetrization of \( \xi \)) and that

\[
\mu^*(B) = \int_{-\infty}^{\infty} \mu(x - B) \, d\mu(x) = \int_{-\infty}^{\infty} \mu(x + B) \, d\mu(x)
\]

for all Borel sets \( B \), where \( \mu, \mu^* \) are the distributions of \( \xi, \xi^* \) respectively, and \( x - B = \{x - y : y \in B\} \), \( x + B = \{x + y : y \in B\} \).
(b) Show that for all \( t \geq 0 \), real \( a \)

\[
P(|\xi^*| \geq t) \leq 2P(|\xi - a| \geq t/2).
\]

10.9 Criterion for independence of r.e.’s analogous to Theorem 10.1.2:
Let \( \xi_\lambda \) be a random element on \((\Omega, \mathcal{F}, P)\) with values in \((X_\lambda, S_\lambda)\) say, for each \( \lambda \) in an index set \( \Lambda \). For each \( \lambda \), let \( \mathcal{E}_\lambda \) be a class of subsets of \( X_\lambda \) which is closed under finite intersections and whose generated \( \sigma \)-ring \( S(\mathcal{E}_\lambda) = S_\lambda \), and write \( \mathcal{G}_\lambda = \xi_\lambda^{-1}E_\lambda(= \{ \xi_\lambda^{-1}E : E \in \mathcal{E}_\lambda \}) \).
Then \( \{ \xi_\lambda : \lambda \in \Lambda \} \) is a class of independent random elements if and only if \( \{ \mathcal{G}_\lambda : \lambda \in \Lambda \} \) is a family of independent classes of events.

10.10 A weaker concept of independence of a family of classes of random elements would be the following. Let \( \{ C_\lambda : \lambda \in \Lambda \} \) be a family of classes of random elements and suppose that if for every choice of one member \( \xi_\lambda \) from each \( C_\lambda \), \( \{ \xi_\lambda : \lambda \in \Lambda \} \) is a class of independent random elements. Such a definition would be more strictly analogous to the procedure used for classes of sets. Show that it is, in fact, a weaker requirement than the definition in the text. (E.g. take two classes \( C_1 = \{ \xi \}, C_2 = \{ \eta, \zeta \} \) where any two of \( \xi, \eta, \zeta \) are independent but the three together are not (cf. Ex. 10.7). Show that \( \{ C_1, C_2 \} \) satisfies the weaker definition, but is not independent, however, in the sense of the text.)

10.11 For each \( \lambda \) in an index set \( \Lambda \), let \( \xi_\lambda, \xi^*_\lambda \) be random elements on \((\Omega, \mathcal{F}, P)\), with values in \((X_\lambda, S_\lambda)\) and such that \( \xi_\lambda = \xi^*_\lambda \) a.s. Show that if \( \{ \xi_\lambda : \lambda \in \Lambda \} \) is a class of independent random elements, then so is \( \{ \xi^*_\lambda : \lambda \in \Lambda \} \) (e.g. show \((\bigcap_{i=1}^n \xi^*_\lambda E_i)\Delta (\bigcap_{i=1}^n \xi_\lambda E_i) \subset \bigcup_{i=1}^n \{ \omega : \xi_\lambda(\omega) \neq \xi^*_\lambda(\omega) \})
\)

10.12 A bag contains one black ball and \( m \) white balls. A ball is drawn at random. If it is black it is returned to the bag. If it is white, it and an additional white ball are returned to the bag. Let \( A_n \) denote the event that the black ball is not drawn in the first \( n \) trials. Discuss the (converse to) the Borel–Cantelli Lemma with reference to the events \( A_n \).

10.13 Let \((\Omega, \mathcal{F}, P)\) be the “unit interval probability space” of Ex. 10.7. Define r.v.’s \( \xi_n \) by

\[
\xi_n(\omega) = \chi_{[0, \frac{1}{2} + \frac{1}{n}]}(\omega) + 2\chi_{[\frac{1}{2} + \frac{1}{n}, 1]}(\omega).
\]

Find the tail \( \sigma \)-field of \( \{ \xi_n \} \) and comment on the zero-one law.

10.14 Let \( \xi \) be a r.v. which is independent of itself. Show that \( \xi \) is a constant, with probability one.

10.15 Let \( \{ \xi_n \}_{n=1}^{\infty} \) be a sequence of independent random variables on the probability space \((\Omega, \mathcal{F}, P)\). Prove that the probability of pointwise convergence of

(i) the sequence \( \{ \xi_n(\omega) \}_{n=1}^{\infty} \)

(ii) the series \( \sum_{n=1}^{\infty} \xi_n(\omega) \)
is equal to zero or one, and that whenever (i) converges its limit is equal to a constant a.s. (Hint: Show that the set $C$ of all points $\omega \in \Omega$ for which the sequence $\{\xi_n(\omega)\}_{n=1}^{\infty}$ converges is given by

$$C = \bigcap_{k=1}^{\infty} \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} \bigcap_{m=N}^{\infty} \{\omega \in \Omega : |\xi_n(\omega) - \xi_m(\omega)| \leq 1/k\}.$$ \)

10.16 Prove that a sequence of independent identically distributed random variables converges pointwise with zero probability, except when all random variables are equal to a constant a.s. (Hint: Use the result and the hint of the previous problem.)
11

Convergence and related topics

11.1 Modes of probabilistic convergence

Various modes of convergence of measurable functions to a limit function were considered in Chapter 6, and will be restated here with the special terminology customarily used in the probabilistic context. In this section the modes of convergence all concern a sequence \( \{ \xi_n \} \) of r.v.’s on the same probability space \((\Omega, \mathcal{F}, P)\) such that the values \( \xi_n(\omega) \) “become close” (in some “local” or “global” sense) to a “limiting r.v.” \( \xi(\omega) \) as \( n \to \infty \). In the next section we shall consider the weaker form of convergence where the \( \xi_n \)’s can be defined on different spaces, and where one is interested in only the limiting form of the distribution of the \( \xi_n \) (i.e. \( P_{\xi_n}^{-1}B \) for Borel sets \( B \)). This “convergence in distribution” has wide use in statistical theory and application.

The later sections of the chapter will be concerned with various important relationships between the forms of convergence, convergence of series of independent r.v.’s, and related topics. Note that in certain calculations concerning convergence (especially in Section 11.5) it will be implicitly assumed that the r.v.’s involved are defined for all \( \omega \). No comment will be made in these cases, since it is a trivial matter to obtain these results for r.v.’s \( \xi_n \) not defined everywhere by considering \( \xi_n^* \) defined for all \( \omega \), and equal to \( \xi_n \) a.s.

In this section, then, we shall consider a sequence \( \{ \xi_n \} \) of r.v.’s on the same fixed probability space \((\Omega, \mathcal{F}, P)\). The following definitions will apply:

**Almost sure convergence**

Almost sure convergence of a sequence of r.v.’s \( \xi_n \) to a r.v. \( \xi \) (\( \xi_n \to \xi \) a.s.) is, of course, just a.e. convergence of \( \xi_n \) to \( \xi \) with respect to the probability measure \( P \). This is also termed convergence with probability 1. Similarly to say that \( \{ \xi_n \} \) is Cauchy a.s. means that it is Cauchy a.e. (\( P \)), as defined in Chapter 6.
A useful necessary and sufficient condition for a.s. convergence is provided by Theorem 6.2.4 which is restated in the present context:

**Theorem 11.1.1** \( \xi_n \to \xi \) a.s. if and only if for every \( \varepsilon > 0 \), writing \( E_n(\varepsilon) = \{ \omega : |\xi_n(\omega) - \xi(\omega)| \geq \varepsilon \} \)

\[
\lim_{n \to \infty} P\left( \bigcup_{m=n}^{\infty} E_m(\varepsilon) \right) \left( = P\left( \lim_{n \to \infty} E_n(\varepsilon) \right) \right) = 0.
\]

That is, \( \xi_n \to \xi \) a.s. if (except on a zero probability set) the events \( E_n(\varepsilon) \) occur only finitely often for each \( \varepsilon > 0 \), or, equivalently, the probability that \( |\xi_m - \xi| \geq \varepsilon \) for some \( m \geq n \), tends to zero as \( n \to \infty \).

The following very simple but sometimes useful sufficient condition for a.s. convergence is immediate from the above criterion.

**Theorem 11.1.2** Suppose that, for each \( \varepsilon > 0 \),

\[
\sum_{n=1}^{\infty} P\{|\xi_n - \xi| \geq \varepsilon \} < \infty.
\]

Then \( \xi_n \to \xi \) a.s. as \( n \to \infty \).

**Proof** This is an immediate and obvious application of the Borel–Cantelli Lemma (Theorem 10.5.1).

A corresponding condition for \( \{\xi_n\} \) to be a Cauchy sequence a.s. (and hence convergent a.s. to some \( \xi \)) will now be obtained.

**Theorem 11.1.3** Let \( \{\epsilon_n\} \) be positive constants, \( n = 1, 2, \ldots \) with \( \sum_{n=1}^{\infty} \epsilon_n < \infty \) and suppose that

\[
\sum_{n=1}^{\infty} P\{|\xi_{n+1} - \xi_n| > \epsilon_n \} < \infty.
\]

Then \( \{\xi_n\} \) is a Cauchy sequence a.s. (and hence convergent to some r.v. \( \xi \) a.s.).

**Proof** By the Borel–Cantelli Lemma (Theorem 10.5.1) the probability is zero that \( |\xi_{n+1} - \xi_n| > \epsilon_n \) for infinitely many \( n \). That is for each \( \omega \) except on a set of \( P \)-measure zero, there is a finite \( N = N(\omega) \) such that \( |\xi_{n+1}(\omega) - \xi_n(\omega)| \leq \epsilon_n \) when \( n \geq N(\omega) \). Given \( \epsilon > 0 \) we may (by increasing \( N \) if necessary) require that \( \sum_{n=1}^{\infty} \epsilon_n < \epsilon \) (\( N \) now depends on \( \epsilon \) and \( \omega \), of course). Thus if \( n > m \geq N \),

\[
|\xi_n - \xi_m| \leq \sum_{k=m}^{n-1} |\xi_{k+1} - \xi_k| \leq \sum_{k=N}^{\infty} |\xi_{k+1} - \xi_k| \leq \sum_{k=N}^{\infty} \epsilon_k < \epsilon
\]

and hence \( \{\xi_n(\omega)\} \) is a Cauchy sequence, as required. \( \square \)
Convergence in probability

This is just convergence in measure, with the previous terminology. That is, $\xi_n$ tends to $\xi$ in probability ($\xi_n \overset{P}{\to} \xi$) if for each $\epsilon > 0$,

$$P(\omega : |\xi_n(\omega) - \xi(\omega)| \geq \epsilon) \to 0 \text{ as } n \to \infty$$

i.e. $P(E_n(\epsilon)) \to 0$ as $n \to \infty$, with the notation of Theorem 11.1.1, or in probabilistic language $P(|\xi_n - \xi| \geq \epsilon) \to 0$ for each $\epsilon > 0$. That is, for each (large) $n$ there is high probability that $\xi_n$ will be close to $\xi$ – but not necessarily high probability that $\xi_m$ will be close to $\xi$ simultaneously for all $m \geq n$. Thus convergence in probability is a weaker requirement than almost sure convergence. This is made specific by the corollary to Theorem 6.2.2 (or implied by Theorem 11.1.1) which shows that if $\xi_n \overset{a.s.}{\to} \xi$, then $\xi_n \overset{P}{\to} \xi$.

It also follows (from the corollary to Theorem 6.2.3) that if $\xi_n$ converges to $\xi$ in probability, then a subsequence $\xi_{n_k}$, say, of $\xi_n$ converges to $\xi$ a.s. We state these two results as a theorem:

**Theorem 11.1.4**

(i) If $\xi_n \overset{a.s.}{\to} \xi$, then $\xi_n \overset{P}{\to} \xi$.

(ii) If $\xi_n \overset{P}{\to} \xi$, then there exists a subsequence $\xi_{n_k}$ converging to $\xi$ a.s. ($\{n_k\}$ is the same for all $\omega$).

The following result will be useful for later applications.

**Theorem 11.1.5**

(i) $\xi_n \overset{P}{\to} \xi$ if and only if each subsequence of $\{\xi_n\}$ contains a further subsequence which converges to $\xi$ a.s.

(ii) If $\xi_n \overset{P}{\to} \xi$, and $f$ is a continuous function on $\mathbb{R}$, then $f(\xi_n) \overset{P}{\to} f(\xi)$.

(iii) (ii) holds if $f$ is continuous except for $x \in D$ where $P\xi^{-1}D = 0$.

**Proof** (i) If $\xi_n \overset{P}{\to} \xi$ in probability, any subsequence also converges to $\xi$ in probability, and, by Theorem 11.1.4 (ii), contains a further subsequence converging to $\xi$ a.s.

Conversely suppose that each subsequence of $\{\xi_n\}$ contains a further subsequence converging a.s. to $\xi$. If $\xi_n$ does not converge to $\xi$ in probability, there is some $\epsilon > 0$ with $P(|\xi_n - \xi| \geq \epsilon) \not\to 0$, and hence also some $\delta > 0$ such that $P(|\xi_n - \xi| \geq \epsilon) > \delta$ infinitely often. That is for some subsequence $\{\xi_{n_k}\}$, $P(|\xi_{n_k} - \xi| \geq \epsilon) > \delta$, $k = 1, 2, \ldots$. But this means that no subsequence of $\{\xi_{n_k}\}$ can converge to $\xi$ in probability (and thus certainly not a.s.), so a contradiction results. Hence we must have $\xi_n \overset{P}{\to} \xi$ in probability as asserted.
(ii) Suppose $\xi_n \overset{P}{\to} \xi$ and write $\eta_n = f(\xi_n)$. Any subsequence \{\xi_{n_k}\} of \{\xi_n\} has, by (i), a further subsequence \{\xi_{m_\ell}\}_{\ell=1}^\infty, converging to $\xi$ a.s. Hence, by continuity, $f(\xi_{m_\ell}) \to f(\xi)$ a.s. That is the subsequence \{\eta_{n_\ell}\} of \{\eta_n\} has a further subsequence converging to $\eta$ a.s. and hence, again by (i), $\eta_n \to \eta$ in probability, so that (ii) holds.

For (iii) essentially the same proof applies – noting that $f(\xi_{m_\ell})$ still converges to $f(\xi)$ a.s. since any further points $\omega$ where convergence does not occur, are contained in the zero probability set $\xi^{-1} D$. □

**Convergence in pth order mean**

Again, $L_p$ convergence of measurable functions, ($p > 0$), includes $L_p$ convergence for r.v.'s $\xi_n$. Specifically, if $\xi_n$, $\xi$ have finite $p$th moments (i.e. $\xi_n$, $\xi \in L_p(\Omega, \mathcal{F}, P)$) we say that $\xi_n \to \xi$ in $p$th order mean if $\xi_n \to \xi$ in $L_p$, i.e. if

$$E|\xi_n - \xi|^p = \int |\xi_n - \xi|^p dP \to 0 \text{ as } n \to \infty.$$ 

The reader should review the properties of $L_p$-spaces given in Section 6.4, including the inequalities restated in probabilistic terminology in Section 9.5. Especially recall that $L_p$ is a linear space for all $p > 0$ (if $\xi$, $\eta \in L_p$ then $a\xi + b\eta \in L_p$ for any real $a, b$), and that $L_p$ is complete. Many of the useful results apply whether $0 < p < 1$ or $p \geq 1$ and in particular we shall find the following lemma (which restates part of Theorem 6.4.6 (ii)) to be useful.

**Theorem 11.1.6** Let \{\xi_n\} ($n = 1, 2, \ldots$), $\xi$ be r.v.'s in $L_p$ for some $p > 0$ and $\xi_n \to \xi$ in $L_p$. Then

(i) $\xi_n \overset{p}{\to} \xi$

(ii) $E|\xi_n|^p \to E|\xi|^p$.

By (i) if $\xi_n \to \xi$ in $L_p$ ($p > 0$) then $\xi_n \overset{p}{\to} \xi$. This implies also, of course, that a subsequence $\xi_{n_k} \to \xi$ a.s. (Theorem 11.1.4 (ii)). However, the sequence $\xi_n$ itself does not necessarily converge a.s. Conversely, nor does a.s. convergence of $\xi_n$ necessarily imply convergence in any $L_p$.

There is, however, a converse result when the $\xi_n$ are dominated by an $L_p$ r.v. In particular the case $p = 1$ may be regarded as a form of the dominated convergence theorem applicable to finite measure (e.g. probability) spaces,
with a.s. convergence replaced by convergence in probability. (We shall also see a more general converse later – Theorem 11.4.2.)

**Theorem 11.1.7** Let \( \{\xi_n\} \), \( \xi \) be r.v.’s such that \( \xi_n \overset{p}{\to} \xi \). Suppose \( \eta \in L_p \) for some \( p > 0 \), and \( |\xi_n| \leq \eta \) a.s., \( n = 1, 2, \ldots \). Then \( \xi_n \to \xi \) in \( L_p \).

**Proof** Note first that clearly \( \xi_n \in L_p \). Further, since \( \xi_n \overset{p}{\to} \xi \), a subsequence \( \xi_{n_k} \to \xi \) a.s. so that \( |\xi_n| \leq |\eta| \) a.s. Since \( \eta \in L_p \) it follows that \( \xi \in L_p \). Now \( |\xi_n - \xi|^p \) and hence
\[
E|\xi_n - \xi|^p = \int_{|\xi_n - \xi| < \epsilon} |\xi_n - \xi|^p \, dP + \int_{|\xi_n - \xi| \geq \epsilon} |\xi_n - \xi|^p \, dP \leq \epsilon^p + 2 \int_{|\xi_n - \xi| \geq \epsilon} \eta^p \, dP.
\]
The last term tends to zero by Theorem 4.5.3 since \( P\{|\xi_n - \xi| \geq \epsilon\} \to 0 \) so that \( \lim_{n \to \infty} E|\xi_n - \xi|^p \leq \epsilon^p \). Since \( \epsilon \) is arbitrary, \( \lim_{n \to \infty} E|\xi_n - \xi|^p = 0 \) as required. \( \square \)

### 11.2 Convergence in distribution

As noted in the previous section, it is of interest to consider another form of convergence – involving just the distributions of a sequence of r.v.’s, and not their values at each \( \omega \). That is, given a sequence \( \{\xi_n\} \) of r.v.’s we inquire whether the distributions \( P\{\xi_n \in B\} \) converge to that of a r.v. \( \xi \), i.e. \( P\{\xi \in B\} \), for sets \( B \in \mathcal{B} \).

In fact, it is a little too stringent to require this for all \( B \in \mathcal{B} \). For suppose that \( \xi_n \) has d.f. \( F_n(x) \) which is zero for \( x \leq -1/n \), one for \( x \geq 1/n \) and is linear in \((-1/n, 1/n)\). Clearly one would want to say that the limiting distribution of \( \xi_n \) is the probability measure \( \pi \) with unit mass at zero, i.e. the distribution of the r.v. \( \xi = 0 \). But, taking \( B \) to be the “singleton set” \( \{0\} \), we have \( P\{\xi_n = 0\} = 0 \), which does not converge to \( P\{\xi = 0\} = 1 \).

It is easy to see (at least once one is told!) what should be done to give an appropriate definition. In the above example, the d.f.’s \( F_n(x) \) of \( \xi_n \) converge to a limiting d.f. \( F(x) \) (zero for \( x < 0 \), one for \( x \geq 0 \)) at all points \( x \) other than the discontinuity point \( x = 0 \) of \( F \) at which \( F_n(0) = \frac{1}{2} \). Equivalently, as we shall see, \( P_{\xi_n}^{-1}\{[a, b]\} \to \mu_F\{[a, b]\} \) for all \( a, b \) with \( \mu_F\{a\} = \mu_F\{b\} = 0 \). This is conveniently used as the basis for a definition of convergence in distribution. It will also then be true – though we shall neither need nor show this – that \( P_{\xi_n}^{-1}(B) \to \mu_F(B) \) for all Borel sets \( B \) whose (topological) boundary has \( \mu \)-measure zero. The definition below will be stated in what appears to be a slightly more general form, concerning a sequence \( \{\pi_n\} \) of probability measures on \( \mathbb{R} \) from those on \( \Omega \).
Of course, each $\pi_n$ may be regarded as the distribution of some r.v. (Section 9.2). We shall speak of weak convergence of the sequence $\pi_n$ since it is this terminology which is used in the most abstract and general setting for the subject described in a variety of treatises, beginning with the classic volume [Billingsley].

Suppose, then, that $\{\pi_n\}$ is a sequence of probability measures on $(\mathbb{R}, \mathcal{B})$. Then we say that $\pi_n$ converges weakly to a probability measure $\pi$ on $\mathcal{B}$ ($\pi_n \xrightarrow{w} \pi$) if $\pi_n[(a, b)] \to \pi((a, b))$ for all $a, b$ such that $\pi((a]) = \pi((b]) = 0$, (i.e. each “$\pi$-continuity interval” $(a, b]$). It is readily seen (Ex. 11.10) that open intervals $(a, b)$ or closed intervals $[a, b]$ may replace the semiclosed interval $(a, b]$ in the definition.

Correspondingly if $F_n$ is a d.f. for $n = 1, 2, \ldots$, and $F$ is a d.f. we write $F_n \xrightarrow{w} F$ if $F_n(x) \to F(x)$ for each $x$ at which $F$ is continuous.

It is obvious that if $F_n$ is the d.f. corresponding to $\pi_n$, and $F$ to $\pi$ ($\pi_n = \mu_{F_n}$, $\pi = \mu_F$), then $F_n \xrightarrow{w} F$ implies $\pi_n \xrightarrow{w} \pi$. The converse is also quite easy to prove directly (Ex. 11.9) but will follow in the course of the proof of Theorem 11.2.1 below.

If $\{\xi_n\}$ is a sequence of r.v.’s with d.f.’s $\{F_n\}$, and $\xi$ is a r.v. with d.f. $F$, we say that $\xi_n$ converges in distribution to $\xi$ ($\xi_n \xrightarrow{d} \xi$), if $F_n \xrightarrow{w} F$ (i.e. $P_{\xi_n^{-1}} \xrightarrow{w} P_{\xi^{-1}}$). Note that the $\xi_n$ do not need to be defined on the same probability space for convergence in distribution. Further, even if they are all defined on the same $(\Omega, \mathcal{F}, P)$, the fact that $\xi \xrightarrow{d} \xi$ does not require that the values $\xi_n(\omega)$ approach those of $\xi(\omega)$ in any sense, as $n \to \infty$. This is in contrast to the other forms of convergence already considered and which (as we shall see) imply convergence in distribution. For example, if $\{\xi_n\}$ is any sequence of r.v.’s with the same d.f. $F$, then $\xi_n$ converges in distribution to any r.v. $\xi$ with the d.f. $F$. This emphasizes that convergence in distribution is concerned only with limits of probabilities $P(\xi_n \in B)$ as $n$ becomes large. Relationships with other forms of convergence will be addressed in the next section.

The following result is a central criterion for weak convergence, indeed leading to its definition in more abstract settings, in which the result is sometimes termed the “Portmanteau Theorem” (e.g. [Billingsley]).

**Theorem 11.2.1** Let $\{\pi_n : n = 1, 2, \ldots\}$, $\pi$, be probability measures on $(\mathbb{R}, \mathcal{B})$, with corresponding d.f.’s $\{F_n : n = 1, 2, \ldots\}$, $F$. Then the following are equivalent

1 Strictly we should write $P_n$ since the $\xi_n$ may be defined on different spaces $(\Omega_n, \mathcal{F}_n, P_n)$ but it is conventional to omit the $n$ and unlikely to cause confusion.
11.2 Convergence in distribution

(i) \( F_n \xrightarrow{w} F \)

(ii) \( \pi_n \xrightarrow{w} \pi \)

(iii) \( \int_{-\infty}^{\infty} g \, d\pi_n \to \int_{-\infty}^{\infty} g \, d\pi \) for every real, bounded continuous function \( g \) on \( \mathbb{R} \).

Further, weak limits are unique (e.g. if \( F_n \xrightarrow{w} F \) and \( F_n \xrightarrow{w} G \) then \( F = G \)).

Proof. The uniqueness statement is immediate since, for example, if \( F_n \xrightarrow{w} F \) and \( F_n \xrightarrow{w} G \) then \( F = G \) at all continuity points of both \( F \), \( G \), and hence for all points \( x \) except in a countable set. From this it is seen at once that \( F(x + 0) = G(x + 0) \) for all \( x \), and hence \( F = G \).

It is immediate that (i') implies (i). On the other hand if (i) holds, for given \( x \) choose \( y > x \) such that \( F \) is continuous at \( y \). Then \( \limsup F_n(x) \leq \lim F_n(y) = F(y) \) from which it follows that \( \limsup F_n(x) \leq F(x) \) by letting \( y \downarrow x \). That \( \liminf F_n(x) \geq F(x - 0) \) follows similarly. Hence (i) and (i') are equivalent.

To prove the equivalence of (i), (ii), (iii), note first, as already pointed out above, that (i) clearly implies (ii).

Suppose now that (ii) holds. To show (iii) let \( g \) be a fixed, real, bounded, continuous function on \( \mathbb{R} \), and \( M = \sup_{x \in \mathbb{R}} |g(x)| < \infty \). We shall show that \( \limsup \int g \, d\pi_n \leq \int g \, d\pi \). Then replacing \( g \) by \( -g \) it will follow that \( \liminf \int g \, d\pi_n = - \limsup \int -g \, d\pi_n \geq - \int -g \, d\pi = \int g \, d\pi \), to yield the desired result \( \lim \int g \, d\pi_n = \int g \, d\pi \). It will be slightly more convenient to assume that \( 0 \leq g(x) \leq 1 \) for all \( x \) (which may be done by considering \( (g + M)/2M \) instead of \( g \)).

Let \( D \) be the set of atoms of \( \pi \) (i.e. discontinuities of \( F \)). By Lemma 9.2.2, \( D \) is at most countable and thus every interval contains points of its complement \( D^c \). Let \( \epsilon > 0 \). Since \( \pi(\mathbb{R}) = 1 \) there are thus points \( a, b \) in \( D^c \) such that \( \pi((a, b]) > 1 - \epsilon/2 \). Hence also, since \( \pi_n \xrightarrow{w} \pi \), we must have \( \pi_n((a, b]) > 1 - \epsilon/2 \) for all \( n \geq 1 \). Thus for \( n \geq 1 \),

\[
\int_{-\infty}^{\infty} g \, d\pi_n = \int_{(a,b]} g \, d\pi_n + \int_{(a,b]} g \, d\pi_n \leq \int_{(a,b]} g \, d\pi_n + \epsilon/2
\]

since \( g \leq 1 \) and \( \pi_n((a, b]) \leq \epsilon/2 \) when \( n \geq 1 \). Hence

\[
\limsup_{n \to \infty} \int g \, d\pi_n \leq \limsup_{n \to \infty} \int_{(a,b]} g \, d\pi_n + \epsilon/2.
\]

Now \( g \) is uniformly continuous on the finite interval \( [a, b] \) and hence there exists \( \delta = \delta(\epsilon) \) such that \( |g(x) - g(y)| < \epsilon/4 \) if \( |x - y| < \delta \), \( a \leq x, y \leq b \).
Choose a partition $a = x_0 < x_1 < \ldots < x_m = b$ of $[a, b]$ such that $x_k \notin D$, and $x_k - x_{k-1} < \delta$, $k = 1, \ldots, m$. Then if $x_{k-1} < x \leq x_k$ we have

$$g(x) \leq g(x_k) + \epsilon/4 \leq g(x) + \epsilon/2$$

and hence

$$\int_{[a,b]} g \, d\pi_n \leq \sum_{k=1}^{m} (g(x_k) + \epsilon/4)\pi_n\{(x_{k-1}, x_k]\}.\] Letting $n \to \infty$ (with the partition fixed), $\pi_n\{(x_{k-1}, x_k]\} \to \pi\{(x_{k-1}, x_k]\}$ giving

$$\limsup_{n \to \infty} \int_{[a,b]} g \, d\pi_n \leq \sum_{k=1}^{m} (g(x_k) + \epsilon/4)\pi\{(x_{k-1}, x_k]\} \leq \int_{[a,b]} (g(x) + \epsilon/2) \, d\pi \leq \int_{-\infty}^{\infty} g \, d\pi + \epsilon/2.$$

Thus by gathering facts, we have,

$$\limsup_{n \to \infty} \int_{-\infty}^{\infty} g \, d\pi_n \leq \int_{-\infty}^{\infty} g \, d\pi + \epsilon$$

from which the desired result follows since $\epsilon > 0$ is arbitrary. Thus (ii) implies (iii).

Finally we assume that (iii) holds and show that (i′) follows, i.e. $\limsup_{n} F_n(x) \leq F(x)$, $\liminf_{n} F_n(x) \geq F(x-0)$, for any fixed point $x$.

Let $\epsilon > 0$ and write $g_{\epsilon}(t)$ for the bounded continuous function which is unity for $t \leq x$, decreases linearly to zero at $t = x + \epsilon$, and is zero for $t > x + \epsilon$. Then

$$F_n(x) = \int_{(-\infty,x]} g_{\epsilon}(t) \, d\pi_n(t) \leq \int_{-\infty}^{\infty} g_{\epsilon} \, d\pi_n \to \int_{-\infty}^{\infty} g_{\epsilon} \, d\pi \leq F(x + \epsilon).$$

Hence $\limsup_{n \to \infty} F_n(x) \leq F(x + \epsilon)$ for $\epsilon > 0$, and letting $\epsilon \to 0$ gives $\limsup_{n \to \infty} F_n(x) \leq F(x)$.

It may be similarly shown (by writing $h_{\epsilon}(t) = 1$ for $t \leq x - \epsilon$, zero for $t \geq x$ and linear in $(x - \epsilon, x)$) that $\liminf_{n \to \infty} F_n(x) \geq F(x - \epsilon)$ for all $\epsilon > 0$ and, hence $\liminf_{n \to \infty} F_n(x) \geq F(x - 0)$ as required, so that (iii) implies (i′) and hence (i), completing the proof of the equivalence of (i)–(iii). $\square$

**Corollary 1** If $\pi_n \xrightarrow{w} \pi$ then (iii) also holds for bounded measurable functions $g$ just assumed to be continuous a.e. ($\pi$).

**Proof** It may be assumed (by subtracting its lower bound) that $g$ is non-negative. Then a sequence $\{g_n\}$ of continuous functions may be found (cf.
Ex. 11.11 for a sketch of their construction) such that \(0 \leq g_n(x) \uparrow g(x)\) at each continuity point \(x\) of \(g\). Hence, for fixed \(m\),

\[
\liminf_{n \to \infty} \int g \, d\pi_n \geq \liminf_{n \to \infty} \int g_m \, d\pi_n = \int g_m \, d\pi
\]

by (iii) and hence by monotone convergence, letting \(m \to \infty\),

\[
\liminf_{n \to \infty} \int g \, d\pi_n \geq \int g \, d\pi.
\]

The same argument with \(-g\) shows that

\[
\liminf_{n \to \infty} \int -g \, d\pi_n \geq \int -g \, d\pi
\]

so that \(\limsup_{n \to \infty} \int g \, d\pi_n \leq \int g \, d\pi\) and hence (iii) holds for this \(g\) as required. \(\square\)

The above criteria may be translated as conditions for convergence in distribution of a sequence of r.v.'s, as follows.

**Corollary 2**  If \(\{\xi_n : n = 1, 2, \ldots\}, \xi\) are r.v.'s with d.f.'s \(\{F_n : n = 1, 2, \ldots\}, F\), then the following are equivalent

1. \(\xi_n \xrightarrow{d} \xi\)
2. \(F_n \xrightarrow{w} F\)
3. \(P_{\xi_n^{-1}} \xrightarrow{w} P_{\xi^{-1}}\)
4. \(Eg(\xi_n) \to Eg(\xi)\) for every bounded continuous real function \(g\) on \(\mathbb{R}\).

If (iv) holds for all such \(g\) it also holds if \(g\) is just bounded and continuous a.e. \((P_{\xi^{-1}})\).

**Proof**  These are immediate by identifying \(P_{\xi_n^{-1}}, P_{\xi^{-1}}\) with \(\pi_n, \pi\) of Theorem 11.2.1, and noting that (iv) here becomes the statement of Corollary 1 of the theorem. \(\square\)

The final result of this series is a very useful one which shows that an (a.e.) continuous function of a sequence converging in distribution also converges in distribution.

**Theorem 11.2.2** (Continuous Mapping Theorem)  Let \(\xi_n \xrightarrow{d} \xi\) where \(\xi_n, \xi\) have distributions \(\pi_n, \pi\) and let \(h\) be a measurable function on \(\mathbb{R}\) which is continuous a.e. \((\pi)\). Then \(h(\xi_n) \xrightarrow{d} h(\xi)\).

**Proof**  This follows at once from the final statement in (iv) of Corollary 2 on replacing the bounded continuous \(g\) by its composition \(g \circ h\), which is clearly bounded and continuous a.e. \((\pi)\), giving

\[
Eg(h(\xi_n)) = E(g \circ h)(\xi_n) \to E(g \circ h)(\xi) = Eg(h(\xi)).
\]
Note that this result may be equivalently stated that if \( \pi_n, \pi \) are probability measures on \( B \) such that \( \pi_n \xrightarrow{w} \pi \), then \( \pi_n h^{-1} \xrightarrow{w} \pi h^{-1} \) if \( h \) is continuous a.e. \((\pi)\). More general, useful forms of the mapping theorem are given in [Kallenberg 2, Theorem 3.2.7].

**Remark** The definition of weak convergence \( \pi_n \xrightarrow{w} \pi \) only involved \( \pi_n(a, b) \to \pi(a, b) \) for intervals \((a, b)\) with \( \pi\{a\} = \pi\{b\} = 0 \). It may, however, then be shown that \( \pi_n(B) \to \pi(B) \) for any Borel set \( B \) whose boundary has \( \pi \)-measure zero (so-called “\( \pi \)-continuity sets”). It may also be shown that two useful further necessary and sufficient conditions for weak convergence may be added to those of Theorem 11.2.1, viz.

(iv) \( \limsup_{n \to \infty} \pi_n(F) \leq \pi(F) \) all closed \( F \)

(v) \( \liminf_{n \to \infty} \pi_n(G) \geq \pi(G) \) all open \( G \).

These are readily proved (see e.g. the “Portmanteau Theorem” of [Billingsley]) and, of course, suggest extensions of the theory to more abstract (topological) contexts.

We next obtain a useful and well known result, “Helly’s Selection Theorem”, concerning a sequence of d.f.’s. This theorem states that if \( \{F_n\} \) is any sequence of d.f.’s, a subsequence \( \{F_{n_k}\} \) may be selected such that \( F_{n_k}(x) \) converges to a nondecreasing function \( F(x) \) at all continuity points of the latter. The limit \( F \) need not be a d.f., however, as is easily seen from the example where \( F_n(x) = 0, \ x < -n, \ F_n(x) = 1, \ x > n, \) and \( F_n \) is linear in \([-n, n]\). \( (F_n(x) \to 1/2 \) for all \( x \).) A condition which will be seen to be useful in ensuring that such a limit is, in fact, a d.f., is the following.

A family \( \mathcal{H} \) of probability measures (or corresponding d.f.’s) on \( B \) is called **tight** if given \( \epsilon > 0 \) there exists \( A \) such that \( \pi\{(-A, A]\} > 1 - \epsilon \) for all \( \pi \in \mathcal{H} \) (or \( F(A) - F(-A) > 1 - \epsilon \) for all d.f.’s \( F \) with \( \mu_F \in \mathcal{H} \)). Note that if \( \pi_n \xrightarrow{w} \pi \), it may be readily shown then that the sequence \( \{\pi_n\} \) is tight (Ex. 11.18).

**Theorem 11.2.3** (Helly’s Selection Theorem) Let \( \{F_n : n = 1, 2, \ldots\} \) be a sequence of d.f.’s. Then there is a subsequence \( \{F_{n_k} : k = 1, 2, \ldots\} \) and a nondecreasing, right-continuous function \( F \) with \( 0 \leq F(x) \leq 1 \) for all \( x \in \mathbb{R} \) such that \( F_{n_k}(x) \to F(x) \) as \( k \to \infty \) at all \( x \in \mathbb{R} \) where \( F \) is continuous.

*If in addition the sequence \( \{F_n\} \) is tight, then \( F \) is a d.f. and \( F_{n_k} \xrightarrow{w} F \).*

**Proof** We will choose a subsequence \( F_{n_k} \) whose values converge at all rational numbers. Let \( \{r_i\} \) be an enumeration of the rationals. Since \( \{F_n(r_1) :
11.2 Convergence in distribution

\( n = 1, 2, \ldots \) is bounded, it has at least one limit point, and there is a subsequence \( S_1 \) of \( \{F_n\} \) whose members converge at \( x = r_1 \).

Similarly there is a subsequence \( S_2 \) of \( S_1 \) whose members converge at \( r_2 \) as well as at \( r_1 \). Proceeding in this way we obtain sequences \( S_1, S_2, \ldots \) which are such that \( S_n \) is a subsequence of \( S_{n-1} \) and the members of \( S_n \) converge at \( x = r_1, r_2, \ldots, r_n \).

Let \( S \) be the (infinite) sequence consisting of the first member of \( S_1 \), the second of \( S_2 \), and so on (the “diagonal” sequence). Clearly the members of \( S \) ultimately belong to \( S_n \) and hence converge at \( r_1, r_2, \ldots, r_n \), for any \( n \), i.e. at all \( r_k \).

Write \( S = \{F_{n_k}\} \) and \( G(r) = \lim_{k \to \infty} F_{n_k}(r) \) for each rational \( r \). Clearly \( 0 \leq G(r) \leq 1 \) and \( G(r) \leq G(s) \) if \( r, s \) are rational \( (r < s) \). Now define \( F \) by

\[
F(x) = \inf\{G(r) : r \text{ rational, } r > x\}.
\]

Clearly \( F \) is nondecreasing, \( 0 \leq F(x) \leq 1 \) for all \( x \in \mathbb{R} \) and \( G(x) \leq F(x) \) when \( x \) is rational. To see that \( F \) is right-continuous, fix \( x \in \mathbb{R} \). Then for any \( y \in \mathbb{R} \) and rational \( r \) with \( x < y < r \),

\[
F(x + 0) \leq F(y) \leq G(r)
\]

so that \( F(x + 0) \leq G(r) \) for all rational \( r > x \). Hence

\[
F(x + 0) \leq \inf\{G(r) : r \text{ rational, } r > x\} = F(x),
\]

showing that \( F \) is right-continuous.

Now let \( x \) be a point where \( F \) is continuous. Then given \( \epsilon > 0 \) there exist rational numbers \( r, s, r < x < s \) such that

\[
F(x) - \epsilon < F(r) \leq F(x) \leq G(s) \leq F(s) < F(x) + \epsilon.
\]

Also if \( r' \) is rational, \( r < r' < x \), \( F(r) \leq G(r') \leq F(r') \leq F(x) \) so that

\[
F(x) - \epsilon < G(r') \leq F(x) \leq G(s) \leq F(x) + \epsilon
\]

giving

\[
F(x) - \epsilon < \lim_{k \to \infty} F_{n_k}(r') \leq \lim_{k \to \infty} F_{n_k}(s) < F(x) + \epsilon.
\]

But \( F_{n_k}(r') \leq F_{n_k}(x) \leq F_{n_k}(s) \) and hence

\[
F(x) - \epsilon < \liminf_{k \to \infty} F_{n_k}(x) \leq \limsup_{k \to \infty} F_{n_k}(x) < F(x) + \epsilon
\]

from which it follows by letting \( \epsilon \to 0 \) that \( F_{n_k}(x) \to F(x) \) as required.

The final task is to show that if the sequence \( \{F_n\} \) is tight, then \( F \) is a d.f. Fix \( \epsilon > 0 \) and let \( A \) be such that \( F_n(A) - F_n(-A) > 1 - \epsilon \) for all \( n \). Let
\( \alpha \leq -A, \ \beta \geq A \) be continuity points of \( F \). Then \( F_n(\beta) - F_n(\alpha) > 1 - \epsilon \) for all \( k \), and hence \( F(\beta) - F(\alpha) = \lim(F_n(\beta) - F_n(\alpha)) \geq 1 - \epsilon \). It follows that \( F(\infty) - F(-\infty) \geq 1 - \epsilon \) for all \( \epsilon \) and hence \( F(\infty) - F(-\infty) = 1 \). Thus \( F(\infty) = 1 + F(-\infty) \) gives \( F(-\infty) = 0 \) and \( F(\infty) = 1 \). Thus \( F \) is d.f. and \( F_{n_k} \xrightarrow{w} F \).

□

An important notion closely related to tightness (in fact identical to tightness in this real line context) is that of relative compactness. Specifically a family \( \mathcal{H} \) of probability measures on \( \mathcal{B} \) is called relatively compact if every sequence \( \{\pi_n\} \) of elements of \( \mathcal{H} \) has a weakly convergent subsequence \( \{\pi_{n_k}\} \) (i.e. \( \pi_{n_k} \xrightarrow{w} \pi \) for some probability measure \( \pi \), not necessarily in \( \mathcal{H} \)). If \( \mathcal{H} \) is a sequence this means that every subsequence has a further subsequence which is weakly convergent.

It follows from the previous theorem that a family which is tight is also relatively compact. In fact it is easily seen that the converse is also true (in this real line framework and many other useful topological contexts). This is summarized in the following theorem.

**Theorem 11.2.4 (Prohorov’s Theorem)** A family \( \mathcal{H} \) of probability measures on \( \mathcal{B} \) is relatively compact if and only if it is tight.

**Proof** In view of the preceding paragraph, we need only now prove that if \( \mathcal{H} \) is relatively compact it is also tight. If it is not tight, there is some \( \epsilon > 0 \) such that \( \pi([-a,a]) \leq 1 - \epsilon \) for some \( \pi \in \mathcal{H} \), whatever \( a \) is chosen. This means that for any \( n \), there is a member \( \pi_n \) of \( \mathcal{H} \) with \( \pi_n([-n,n]) \leq 1 - \epsilon \). But since \( \mathcal{H} \) is relatively compact a subsequence \( \pi_{n_k} \xrightarrow{w} \pi \), a probability measure, as \( k \to \infty \).

Let \( a, b \) be any points such that \( \pi([a]) = \pi([b]) = 0 \). Then for sufficiently large \( k \), \( (a,b) \subset (-n_k,n_k] \) and hence \( \pi((a,b]) = \lim_{k \to \infty} \pi_{n_k}([a,b]) \leq \lim \sup_{k} \pi_{n_k}([-n_k,n_k]) \leq 1 - \epsilon \). But this contradicts the fact that we may choose \( a, b \) with \( \pi([a]) = \pi([b]) = 0 \) so that \( \pi((a,b]) > 1 - \epsilon \) (since \( \pi(\mathbb{R}) = 1 \)). Thus \( \mathcal{H} \) is indeed tight. □

It is well known (and easily shown) that if every convergent subsequence of a bounded sequence \( \{a_n\} \) of real numbers, has the same limit \( a \), then \( a_n \to a \) (i.e. the whole sequence converges). The next result demonstrates an analogous property for weak convergence.

**Theorem 11.2.5** Let \( \{F_n\} \) be a tight sequence of d.f.’s such that every weakly convergent subsequence \( \{F_{n_k}\} \) has the same limiting d.f. \( F \). Then \( F_n \xrightarrow{w} F \).
Proof Suppose the result is not true. Then there is a continuity point $x$ of the d.f. $F$ such that $F_n(x) \not\rightarrow F(x)$. By the above result stated for real sequences, there must be a subsequence $\{F_{n_k}\}$ of $\{F_n\}$ such that $F_{n_k}(x) \not\rightarrow \lambda \neq F(x)$. By Theorem 11.2.3, a subsequence $\{F_{m_k}\}$ of $\{F_{n_k}\}$ converges weakly, and by assumption its limit is $F$. Thus $F_{m_k}(x) \rightarrow F(x)$, contradicting the convergence of $F_n(x)$ to $\lambda \neq F(x)$. □

Finally, as indicated earlier, the notion of weak convergence may be generalized to apply to more abstract situations. The most obvious of these replaces $\mathbb{R}$ by $\mathbb{R}^k$ for which the generalization is immediate. Specifically we say that a sequence $\{\pi_n\}$ of probability measures on $\mathcal{B}^k$ converges weakly to a probability measure $\pi$ on $\mathcal{B}^k$ ($\pi_n \wrightarrow \pi$) if $\pi_n(I) \rightarrow \pi(I)$ for every “continuity rectangle” $I$; i.e. any rectangle $I$ whose boundary has $\pi$-measure zero.

In $\mathbb{R}$ the boundary of $I = (a, b]$ is just the two points $\{a, b\}$. In $\mathbb{R}^2$ it is the four edges, and in $\mathbb{R}^k$ it is the $2k$ bounding hyperplanes.

As in $\mathbb{R}$ we say that a sequence $\{F_n\}$ of d.f.’s in $\mathbb{R}^k$ converges weakly to a d.f. $F$, $F_n \wrightarrow F$, if $F_n(x) \rightarrow F(x)$ at all points $x = (x_1, \ldots, x_k)$ at which $F$ is continuous. It may then be shown that $F_n \wrightarrow F$ if and only if the corresponding probability measures converge (i.e. $\pi_n = \mu_{F_n} \wrightarrow \pi = \mu_F$). If $F_n$ is the joint d.f. of r.v.’s $(\xi^{(1)}_n, \ldots, \xi^{(k)}_n) = (\xi_n)$ and $F$ is the joint d.f. of $(\xi^{(1)}, \ldots, \xi^{(k)}) = \xi$, and $F_n \wrightarrow F$ we say that $\xi_n$ converges to $\xi$ in distribution ($\xi_n \xrightarrow{d} \xi$) (i.e. $P_{\xi^{-1}_n} \wrightarrow P_{\xi^{-1}}$).

More abstract (topological) spaces than $\mathbb{R}^k$ do not necessarily have an order structure to support the notions of distribution functions and of rectangles. However, the notion of bounded continuous functions does exist so that (iii) of Theorem 11.2.1 ($\int g \, d\pi_n \rightarrow \int g \, d\pi$ for every bounded continuous function $g$) can be used as the definition of weak convergence of probability measures $\pi_n \wrightarrow \pi$. This is needed for consideration of convergence in distribution of a sequence of random elements (e.g. stochastic processes) to a random element $\xi$ in topological spaces more general than $\mathbb{R}$ ($P_{\xi^{-1}_n} \wrightarrow P_{\xi^{-1}}$) but our primary focus on random variables does not require the generalization here. We refer the interested reader to [Billingsley] for an eminently readable detailed account.

### 11.3 Relationships between forms of convergence

Returning now to the real line context, it is useful to note some relationships between the various forms of convergence.

Convergence a.s. and convergence in $L_p$ both imply convergence in probability. It is also simply shown by the next result that convergence
in probability implies convergence in distribution. (For another proof see Ex. 11.12.)

**Theorem 11.3.1** Let \( \{\xi_n\} \) be a sequence of r.v.’s on the same probability space \((\Omega, \mathcal{F}, P)\) and suppose that \( \xi_n \xrightarrow{p} \xi \) as \( n \to \infty \). Then \( \xi_n \xrightarrow{d} \xi \) as \( n \to \infty \).

**Proof** Let \( g \) be any bounded continuous function on \( \mathbb{R} \). By Theorem 11.1.5 (ii) it follows that \( g(\xi_n) \xrightarrow{p} g(\xi) \). But \( |g(\xi_n)| \) is bounded by a constant and any constant is in \( L_1 \), so that \( g(\xi_n) \to g(\xi) \) in \( L_1 \) by Theorem 11.1.7, and hence, in particular \( \mathcal{E}g(\xi_n) \to \mathcal{E}g(\xi) \). Hence (iv) of Corollary 2 to Theorem 11.2.1 shows that \( \xi_n \xrightarrow{d} \xi \). \( \square \)

Of course, the converse to Theorem 11.3.1 is not true (even though the \( \xi_n \) are defined on the same space). However, if \( \xi_n \) converges in distribution to some constant \( a \), it is easy to show that \( \xi_n \xrightarrow{p} a \) (Ex. 11.13).

Convergence in distribution by no means implies a.s. convergence (even for r.v.’s defined on the same \((\Omega, \mathcal{F}, P)\)). However, the following representation of Skorohod shows that a sequence \( \{\xi_n\} \) convergent in distribution may for some purposes be replaced by an a.s. convergent sequence \( \tilde{\xi}_n \) with the same individual distributions as \( \xi_n \), such that \( \tilde{\xi}_n \) converges a.s. This can enable the use of simpler theory of a.s. convergence in proving results for convergence in distribution.

**Theorem 11.3.2** (Skorohod’s Representation) Let \( \{\xi_n\}, \xi \) be r.v.’s and \( \xi_n \xrightarrow{d} \xi \). Then there exist r.v.’s \( \tilde{\xi}_n, \tilde{\xi} \) on the “unit interval probability space” \(([0, 1], \mathcal{B}([0, 1]), m) \) (where \( m \) is Lebesgue measure) such that

1. \( \tilde{\xi}_n \stackrel{d}{=} \xi_n \) for each \( n \), \( \tilde{\xi} \overset{d}{=} \xi \), and
2. \( \tilde{\xi}_n \to \tilde{\xi} \) a.s.

**Proof** Let \( \xi_n, \xi \) have d.f.’s \( F_n, F \), respectively and let \( U(u) = u \) for \( 0 \leq u \leq 1 \). Then \( U \) is a uniform r.v. on \([0, 1]\) and (cf. Section 9.6 and Ex. 9.5)

\[
\tilde{\xi}_n = F^{-1}_n(U), \quad \tilde{\xi} = F^{-1}(U)
\]

have d.f.’s \( F_n, F \), i.e. \( \tilde{\xi}_n \overset{d}{=} \xi_n \), \( \tilde{\xi} \overset{d}{=} \xi \) so that (i) holds.

Since \( \xi_n \xrightarrow{d} \xi, F_n \xrightarrow{w} F \), and hence by Lemma 9.6.2, \( F_n^{-1} \to F^{-1} \) at continuity points of \( F^{-1} \). Thus

\[
1 \geq m\{u \in [0, 1] : \tilde{\xi}_n(u) \to \tilde{\xi}(u)\} = m\{u \in [0, 1] : F_n^{-1}(u) \to F^{-1}(u)\} = m\{u \in [0, 1] : F^{-1}(U(u)) = F_n^{-1}(u)\} \geq m\{u \in [0, 1] : F^{-1} \text{ is continuous at } u\} = 1,
\]
since the discontinuities of $F^{-1}$ are countable. Hence $\tilde{\xi}_n(u) \to \tilde{\xi}(u)$ for a.e. $u$, giving (ii).

Note that while the r.v.’s $\xi_n$ may be defined on different probability spaces, their “representatives” $\tilde{\xi}_n$ are defined on the same probability space (as they must be if a.s. convergent).

Finally, note that weak convergence, $\pi_n \xrightarrow{w} \pi$, has been defined for probability measures $\pi_n$, $\pi$ but the same definition applies to measures $\mu_n$ and $\mu$ just assumed to be finite on $\mathcal{B}$, i.e. $\mu_n(\mathbb{R}) < \infty$, $\mu(\mathbb{R}) < \infty$. Of course, $\mu_n(\mathbb{R})$ and $\mu(\mathbb{R})$ need not be unity but if $\mu_n \xrightarrow{w} \mu$ it follows in particular that $\mu_n(\mathbb{R}) \to \mu(\mathbb{R})$.

Suppose now that $\mu_n$, $\mu$ are Lebesgue–Stieltjes measures i.e. measures on $\mathcal{B}$ which are finite on bounded sets but possibly having infinite total measure (or equivalently are defined by finite-valued, nondecreasing but not necessarily bounded functions $F$). Then the previous definition of weak convergence could still be used but the important criterion (iii) of Theorem 11.2.1 does not apply sensibly since e.g. the bounded continuous function $g(x) = 1$ may not be integrable. This is the case for Lebesgue measure itself, of course. However, an appropriate extended notion of convergence may be given in this case.

Specifically if $\{\mu_n\}$, $\mu$ are such measures on $\mathcal{B}$ (finite on bounded sets), we say that $\mu_n$ converges vaguely to $\mu$ ($\mu_n \xrightarrow{v} \mu$) if

$$\int f \, d\mu_n \to \int f \, d\mu$$

for every continuous function $f$ with compact support, i.e. such that $f(x) = 0$ if $|x| > a$ for some constant $a$. Clearly $\int f \, d\mu_n$ and $\int f \, d\mu$ are defined and finite for such functions.

The notion of vague convergence applies in particular if $\mu_n$ and $\mu$ are finite measures and is clearly then implied by weak convergence. The following easily proved result (Ex. 11.20) summarizes the relationship between weak and vague convergence in this case when both apply.

**Theorem 11.3.3** Let $\mu_n, \mu$ be finite measures on $\mathcal{B}$ (i.e. $\mu_n(\mathbb{R}) < \infty$, $\mu(\mathbb{R}) < \infty$). Then, as $n \to \infty$, $\mu_n \xrightarrow{w} \mu$ if and only if $\mu_n \xrightarrow{v} \mu$ and $\mu_n(\mathbb{R}) \to \mu(\mathbb{R})$.

As for weak convergence, the notion of vague convergence can be extended to apply in more general topological spaces than the real line. Discussion of these forms of convergence and their relationships may be found in the volumes [Kallenberg] and [Kallenberg 2].
11.4 Uniform integrability

We turn now to the relation between $L_p$ convergence and convergence in probability. $L_p$ convergence implies convergence in probability (Theorem 11.1.6). We have seen that the converse is true provided each term of the sequence is dominated by a fixed $L_p$ r.v. (Theorem 11.1.7). A weaker condition turns out to be necessary and sufficient, and since it is important for other purposes, we investigate this now.

Specifically, a family $\{ξ_λ : λ ∈ Λ\}$ of $(L_1)$ r.v.’s is said to be uniformly integrable if

$$\sup_{λ ∈ Λ} \int_{|ξ_λ(ω)| > a} |ξ_λ(ω)| dP(ω) → 0 \text{ as } a → ∞$$

or equivalently if $\sup_{λ ∈ Λ} \int_{|x| > a} |x| dF_λ(x) → 0 \text{ as } a → ∞$, where $F_λ$ is the d.f. of $ξ_λ$. From this latter form it is evident that (like convergence in distribution (Section 11.2)) uniform integrability does not require the r.v.’s to be defined on the same probability space. Of course, we always have $\int_{|ξ_λ| > a} |ξ_λ| dP → 0 \text{ (} \int_{|x| > a} |x| dF_λ(x) → 0\text{) for each } λ \text{ as } a → ∞$ (dominated convergence). The extra requirement is that these should be uniform in $λ ∈ Λ$. It is clear that identically distributed $(L_1)$ r.v.’s are uniformly integrable since $\int_{|ξ| > a} |ξ| dF(x) → 0$ where $F$ is the common d.f. of the family. It is also immediate that finite families of $(L_1)$ r.v.’s are uniformly integrable, and that an arbitrary family $\{ξ_λ\}$ defined on the same probability space and each dominated (in absolute value) by an integrable r.v. $ξ$, is uniformly integrable. For then $|ξ_λ| χ_{|ξ_λ| ≥ a} ≤ |ξ| χ_{|ξ| ≥ a}$ and hence

$$\int_{|ξ_λ| ≥ a} |ξ_λ| dP ≤ \int_{|ξ| ≥ a} |ξ| dP.$$

The concept of uniform integrability is closely related to what is called “uniform absolute continuity”. If $ξ ∈ L_1$, we know that (the measure) $\int_E |ξ| dP$ is absolutely continuous with respect to $P$. Recall (Theorem 4.5.3) that then, given $ε > 0$ there exists $δ > 0$ such that $\int_E |ξ| dP < ε$ if $P(E) < δ$.

If $\{ξ_λ : λ ∈ Λ\}$ is a family of $(L_1)$ r.v.’s, each indefinite integral $\int_E |ξ_λ| dP$ is absolutely continuous. If for each $ε$, one $δ$ may be found for all $ξ_λ$ (i.e. if $\int_E |ξ_λ| dP < ε$ for all $λ$ when $P(E) < δ$) then the family of indefinite integrals $\{\int_E |ξ_λ| dP : λ ∈ Λ\}$ is called uniformly absolutely continuous.

**Theorem 11.4.1** A family of $L_1$ r.v.’s $\{ξ_λ : λ ∈ Λ\}$ is uniformly integrable if and only if:

(i) the indefinite integrals $\int_E |ξ_λ| dP$ are uniformly absolutely continuous, and
(ii) the expectations $\mathbb{E}|\xi_\lambda|$ are bounded; i.e. $\mathbb{E}|\xi_\lambda| < M$ for some $M < \infty$ and all $\lambda \in \Lambda$.

**Proof** Suppose the family is uniformly integrable. To see that (i) holds, note that for any $E \in \mathcal{F}$, $\lambda \in \Lambda$,

$$
\int_E |\xi_\lambda| \, dP = \int_{E \cap \{|\xi_\lambda| \leq a\}} |\xi_\lambda| \, dP + \int_{E \cap \{|\xi_\lambda| > a\}} |\xi_\lambda| \, dP \leq aP(E) + \int_{\{|\xi_\lambda| > a\}} |\xi_\lambda| \, dP.
$$

Given $\epsilon > 0$ we may choose $a$ so that the last term does not exceed $\epsilon/2$, for all $\lambda \in \Lambda$ by uniform integrability. For $P(E) < \delta = \epsilon/2a$ we thus have $\int_E |\xi_\lambda| \, dP < \epsilon$ for all $\lambda \in \Lambda$, so that (i) follows.

(ii) is even simpler. For we may choose $a$ such that $\int_{\{|\xi_\lambda| > a\}} |\xi_\lambda| \, dP < 1$ for all $\lambda \in \Lambda$ and hence $\mathbb{E}|\xi_\lambda| \leq 1 + \int_{\{|\xi_\lambda| > a\}} |\xi_\lambda| \, dP \leq 1 + a$ which is a suitable upper bound.

Conversely, suppose that (i) and (ii) hold and write

$$
\sup_{\lambda \in \Lambda} \mathbb{E}|\xi_\lambda| = M < \infty.
$$

Then by the Markov Inequality (Theorem 9.5.3 (Corollary)), for all $\lambda \in \Lambda$, and all $a > 0$,

$$
P(|\xi_\lambda| > a) \leq \mathbb{E}|\xi_\lambda|/a \leq M/a.
$$

Given $\epsilon > 0$, choose $\delta = \delta(\epsilon)$ so that $\int_E |\xi_\lambda| \, dP < \epsilon$ for all $\lambda \in \Lambda$ when $P(E) < \delta$. For $a > M/\delta$ we have $P(|\xi_\lambda| > a) < \delta$ and thus $\int_{\{|\xi_\lambda| > a\}} |\xi_\lambda| \, dP < \epsilon$ for all $\lambda \in \Lambda$. But this is just a statement of the required uniform integrability.

The following result shows in detail how $L_p$ convergence and convergence in probability are related, and in particular generalizes the (probabilistic form of) dominated convergence (Theorem 11.1.7), replacing domination by uniform integrability.

**Theorem 11.4.2** If $\xi_n \in L_p$ ($0 < p < \infty$) for all $n = 1, 2, \ldots$, and $\xi_n \xrightarrow{p} \xi$, then the following are equivalent

(i) $\{|\xi_n|^p : n = 1, 2, \ldots\}$ is a uniformly integrable family

(ii) $\xi \in L_p$ and $\xi_n \xrightarrow{L_p} \xi$ in $L_p$ as $n \to \infty$

(iii) $\xi \in L_p$ and $\mathbb{E}|\xi_n|^p \to \mathbb{E}|\xi|^p$ as $n \to \infty$.

**Proof** We show first that (i) implies (ii).

Since $\xi_n \xrightarrow{p} \xi$, a subsequence $\xi_{n_k} \xrightarrow{a.s.} \xi$. Hence, by Fatou’s Lemma, and (ii) of the previous theorem,

$$
\mathbb{E}|\xi|^p \leq \liminf_{k \to \infty} \mathbb{E}|\xi_{n_k}|^p \leq \sup_{n \geq 1} \mathbb{E}|\xi_n|^p < \infty
$$
so that $\xi \in L_p$. Further

$$E|\xi_n - \xi|^p = \int_{|\xi_n - \xi|^p \leq \epsilon} |\xi_n - \xi|^p \, dP + \int_{|\xi_n - \xi|^p > \epsilon} |\xi_n - \xi|^p \, dP$$

$$\leq \epsilon + 2^p \int_{E_n} |\xi_n|^p \, dP + 2^p \int_{E_n} |\xi|^p \, dP$$

where $E_n = \{\omega : |\xi_n - \xi| > \epsilon^{1/p}\}$ (hence $P(E_n) \to 0$) and use has been made of the inequality $|a + b|^p \leq 2^p(|a|^p + |b|^p)$ (cf. proof of Theorem 6.4.1).

Uniform integrability of $|\xi_n|^p$ implies the uniform absolute continuity of $\int_E |\xi|^p \, dP$ (Theorem 11.4.1). Thus $\int_E |\xi|^p \, dP < \epsilon$ when $P(E) < \delta$ (so $\delta(\epsilon)$), for all $n$, and hence there is some $N_1$ (making $P(E_n) < \delta$ for $n \geq N_1$) such that $\int_{E_{n_1}} |\xi_{n_1}|^p \, dP < \epsilon$ when $n \geq N_1$. Correspondingly for $n \geq some N_2$ we have $\int_{E_n} |\xi|^p \, dP < \epsilon$, and hence for $n \geq \max(N_1, N_2)$, $E|\xi_n - \xi|^p < \epsilon + 2^p \epsilon + 2^p \epsilon$, showing that $\xi_n \to \xi$ in $L_p$.

Thus (i) implies (ii). That (ii) implies (iii) follows at once from Theorem 11.1.6.

The proof will be completed by showing that (iii) implies (i). Let $A$ be any fixed nonnegative real number such that $P(|\xi| = A) = 0$, and define the function $h(x) = |x|^p$ for $|x| < A$, $h(x) = 0$ otherwise. Now since $\xi_n \to \xi$ in probability and $h$ is continuous except at $\pm A$ (but $P(\xi = \pm A) = 0$), it follows from Theorem 11.1.5 (iii) that $h(\xi_n) \to h(\xi)$ in probability. Since $h(\xi_n) \leq A^p \in L_1$ it follows from Theorem 11.1.7 that $h(\xi_n) \to h(\xi)$ in $L_1$. Thus $Eh(\xi_n) \to Eh(\xi)$, and hence by (iii),

$$E|\xi_n|^p - Eh(\xi_n) \to E|\xi|^p - Eh(\xi)$$

or

$$\int_{|\xi_n| > A} |\xi_n|^p \, dP \to \int_{|\xi| > A} |\xi|^p \, dP.$$

Now if $\epsilon > 0$ we may choose $A = A(\epsilon)$ such that this limit is less than $\epsilon$ (and $P(|\xi| = A) = 0$), so that there exists $N = N(\epsilon)$ such that

$$\int_{|\xi_n| > A} |\xi_n|^p \, dP < \epsilon$$

for all $n \geq N$. Since as noted above the finite family $\{|\xi_n|^p : n = 1, 2, \ldots, N - 1\}$ is uniformly integrable, we have $\sup_{1 \leq n \leq N - 1} \int_{|\xi_n| > a} |\xi_n|^p \, dP \to 0$ as $a \to \infty$, and hence there exists $A' = A'(\epsilon)$ such that

$$\max_{1 \leq n \leq N - 1} \int_{|\xi_n|^p > A'} |\xi_n|^p \, dP < \epsilon.$$

Now taking $A'' = A''(\epsilon) = \max(A, A')$, we have $\int_{|\xi_n|^p > A''} |\xi_n|^p \, dP < \epsilon$ for all $n$, and hence, finally, $\sup_n \int_{|\xi_n|^p > a} |\xi_n|^p \, dP < \epsilon$ whenever $a > (A''(\epsilon))^p$, demonstrating the desired uniform integrability.
11.5 Series of independent r.v.’s

Note that (iii) states that \( \int g \, d\pi_n \to \int g \, d\pi \) where \( \pi_n, \pi \) are the distributions of \( \xi_n \) and \( \xi \), and \( g \) is the function \( g(x) = |x|^p \). This result would have followed under weak convergence of \( \pi_n \) to \( \pi \) only (i.e. \( \xi_n \xrightarrow{d} \xi \)) if \( g \) were bounded (by Theorem 11.2.1). It is thus the fact the \( |x|^p \) is not bounded that makes the extra conditions necessary.

Finally, also note that while we are used to sufficient (e.g. “domination type”) conditions for (ii) the fact that (i) is actually necessary for (ii) indicates the appropriateness of uniform integrability as the correct condition to consider for sufficiency when \( \xi_n \xrightarrow{p} \xi \).

11.5 Series of independent r.v.’s

It follows (Ex. 10.15) from the zero-one law of Chapter 10 that if \( \{\xi_n\} \) are independent r.v.’s then

\[
P(\omega : \sum_{n=1}^{\infty} \xi_n(\omega) \text{ converges}) = 0 \text{ or } 1.
\]

In this section necessary and sufficient conditions will be obtained for this probability to be unity, i.e. for \( \sum_{n=1}^{\infty} \xi_n \) to converge a.s. First, two inequalities are needed.

**Theorem 11.5.1** (Kolmogorov Inequalities) *Let \( \xi_1, \xi_2, \ldots, \xi_n \) be independent r.v.’s with zero means and (possibly different) finite second moments \( \mathbb{E} \xi_i^2 = \sigma_i^2 \). Write \( S_k = \sum_{j=1}^{k} \xi_j \). Then, for every \( a > 0 \)

(i) \( P(\max_{1 \leq k \leq n} |S_k| \geq a) \leq \sum_{i=1}^{n} \sigma_i^2 / a^2 \).

(ii) If in addition the r.v.’s \( \xi_i \) are bounded, \( |\xi_i| \leq c \) a.s., \( i = 1, 2, \ldots, n \), then

\[
P(\max_{1 \leq k \leq n} |S_k| < a) \leq \frac{(c + a)^2}{\sum_{i=1}^{n} \sigma_i^2}.
\]

**Proof** First we prove (i), so do not assume \( \xi_i \) bounded. Write

\[
E = \{\omega : \max_{1 \leq k \leq n} |S_k(\omega)| \geq a\}
\]

\[
E_1 = \{\omega : |S_1(\omega)| \geq a\}
\]

\[
E_k = \{\omega : |S_k(\omega)| \geq a\} \cap \bigcap_{i=1}^{k-1} \{\omega : |S_i(\omega)| < a\}, \quad k > 1.
\]

It is readily checked that \( \chi_{E_k} \) and \( \chi_{E_k} S_k \) are Borel functions of \( \xi_1, \ldots, \xi_k \). By Theorems 10.3.2 (Corollary) and 10.3.5 it follows that if \( i > k \),

\[
\mathbb{E}(\chi_{E_i} S_k \xi_i) = \mathbb{E}(\chi_{E_i} S_k) \mathbb{E} \xi_i = 0, \quad \mathbb{E}(\chi_{E_i} \xi_i^2) = \mathbb{E} \chi_{E_i} \mathbb{E} \xi_i^2
\]

and for \( j > i > k \)

\[
\mathbb{E}(\chi_{E_i} \xi_i \xi_j) = \mathbb{E} \chi_{E_i} \mathbb{E} \xi_i \mathbb{E} \xi_j = 0.
\]
Hence since
\[ S_n^2 = (S_k + \sum_{k+1}^{n} \xi_i)^2 = S_k^2 + 2S_k \sum_{k+1}^{n} \xi_i + \sum_{k+1}^{n} \xi_i^2 + 2 \sum_{n \geq j > i > k} \xi_i \xi_j \]
it follows that
\[ \mathcal{E}(\chi_{E_k} S_n^2) = \mathcal{E}(\chi_{E_k} S_k^2) + P(E_k) \sum_{k+1}^{n} \sigma_i^2, \] (11.1)
so that
\[ \mathcal{E}(\chi_{E_k} S_n^2) \geq \mathcal{E}(\chi_{E_k} S_k^2) \geq a^2 P(E_k) \]
since \( \chi_{E_k} S_k^2 \geq a^2 \chi_{E_k} \) by definition of \( E_k \). Thus since \( E = \bigcup_1^n E_k \), and the sets \( E_k \) are disjoint, \( \chi_E = \sum_1^n \chi_{E_k} \) and
\[ a^2 P(E) = a^2 \sum_{k=1}^{n} P(E_k) \leq \sum_{k=1}^{n} \mathcal{E}(\chi_{E_k} S_k^2) = \mathcal{E}(S_n^2 \chi_E) \leq \mathcal{E}S_n^2 = \sum_{i=1}^{n} \sigma_i^2 \]
by independence of \( \xi_i \). Thus \( P(E) \leq \sum_{i=1}^{n} \sigma_i^2 / a^2 \), which is the desired result, (i).

To prove (ii) assume now that \( |\xi_i| \leq c \) a.s. for each \( i \), and note that the equality (11.1) still holds, so that
\[ \mathcal{E}(\chi_{E_k} S_n^2) \leq \mathcal{E}(\chi_{E_k} S_k^2) + P(E_k) \sum_{i=1}^{n} \sigma_i^2 \leq (a + c)^2 P(E_k) + P(E_k) \sum_{i=1}^{n} \sigma_i^2 \]
since \( |S_k| \leq |S_{k-1}| + |\xi_k| \leq a + c \) on \( E_k \). Summing over \( k \) from 1 to \( n \) we have
\[ \mathcal{E}(\chi_{E} S_n^2) \leq (a + c)^2 P(E) + P(E) \sum_{i=1}^{n} \sigma_i^2 \]
and thus (noting that \( |S_n| \leq a \) on \( E^c \))
\[ \sum_{i=1}^{n} \sigma_i^2 = \mathcal{E}S_n^2 = \mathcal{E}(\chi_{E} S_n^2) + \mathcal{E}(\chi_{E^c} S_n^2) \]
\[ \leq (a + c)^2 P(E) + P(E) \sum_{i=1}^{n} \sigma_i^2 + a^2 P(E^c) \]
\[ \leq (a + c)^2 + P(E) \sum_{i=1}^{n} \sigma_i^2 . \]
Rearranging gives

\[ P(E^c) \leq (a + c)^2 \sum_{i=1}^{n} \sigma_i^2 \]

or

\[ P(\max_{1 \leq k \leq n} |S_k| < a) \leq (a + c)^2 \sum_{i=1}^{n} \sigma_i^2 \]

which is the desired result. \( \square \)

Note that the inequality (i) is a generalization of the Chebyshev Inequality (which it becomes when \( n = 1 \)). Note also that the same inequality holds for \( P(\max_{1 \leq k \leq n} |S_k| \leq a) \) in (ii) as for \( P(\max_{1 \leq k \leq n} |S_k| < a) \). (For we may replace \( a \) in (ii) by \( a + \epsilon \) and let \( \epsilon \downarrow 0 \).)

The next lemma will be useful in obtaining our main theorems concerning a.s. convergence of series of r.v.’s.

**Lemma 11.5.2** Let \( \{\xi_n\} \) be a sequence of r.v.’s and write \( S_n = \sum_{i=1}^{n} \xi_i \). Then \( \sum_{i=1}^{\infty} \xi_n \) converges a.s. if and only if

\[ \lim_{k \to \infty} \lim_{n \to \infty} \lim_{m \to \infty} P(Emnk) = 0 \]

for each \( \epsilon > 0 \). (Note that the k-limit exists by monotonicity.)

**Proof** Since \( \sum_{i=1}^{\infty} \xi_n \) converges if and only if the sequence \( \{S_n\} \) is Cauchy, it is readily seen that

\[ \{\omega : \sum_{i=1}^{\infty} \xi_n \text{ converges}\} = \bigcap_{m=1}^{\infty} \bigcup_{n=1}^{\infty} \{\omega : |S_i - S_j| \leq 1/m \text{ for all } i, j \geq n\} \]

\[ = \bigcap_{m=1}^{\infty} \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} \{\omega : \max_{n \leq i, j \leq k} |S_i - S_j| \leq 1/m\}. \]

Now if \( E_{mnk}^c \) denotes the set in braces, i.e. \( E_{mnk} = \{\omega : \max_{n \leq i, j \leq k} |S_i - S_j| > 1/m\} \), it is clear that \( E_{mnk} \) is nonincreasing in \( n \) (\( \leq k \)), and nondecreasing in both \( k \) (\( \geq n \)) and \( m \) so that, writing \( D \) for the set where \( \sum_{i=1}^{\infty} \xi_n \) does not converge, we have

\[ P(D) = P(\bigcup_{m=1}^{\infty} \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} E_{mnk}) = \lim_{m \to \infty} \lim_{n \to \infty} \lim_{k \to \infty} P(E_{mnk}). \]

Since \( P(E_{mnk}) \) is nondecreasing in \( m \), \( P(D) = 0 \) if and only if \( \lim_{n \to \infty} \lim_{k \to \infty} P(E_{mnk}) = 0 \) for each \( m \), which clearly holds if and only if

\[ \lim_{k \to \infty} \lim_{n \to \infty} P\left(\max_{n \leq i, j \leq k} |S_i - S_j| > \epsilon\right) \to 0 \text{ as } n \to \infty \]
for each $\epsilon > 0$. But for fixed $n, k$,

$$P(\max_{n \leq i \leq k} |S_i - S_n| > \epsilon) \leq P(\max_{n \leq i \leq j} |S_i - S_j| > \epsilon) \leq P(\max_{n \leq i \leq k} |S_i - S_n| > \epsilon/2)$$

(since $|S_i - S_j| \leq |S_i - S_n| + |S_n - S_j|$), from which it is easily seen that $P(D) = 0$ if and only if $\lim_{k \to \infty} P(\max_{n \leq r \leq k} |S_r - S_n| > \epsilon) \to 0$ as $n \to \infty$ for each $\epsilon > 0$, as required.

The next theorem (which will follow at once from the above results), while not as general as the “Three Series Theorem” to be obtained subsequently nevertheless gives a simple useful condition for a.s. convergence of series of independent r.v.’s when the terms have finite variances.

**Theorem 11.5.3** Let $\{\xi_n\}$ be a sequence of independent r.v.’s with zero means and finite variances $\mathbb{E} \xi_n^2 = \sigma_n^2$. Suppose that $\sum_1^\infty \sigma_n^2 < \infty$. Then $\sum_1^\infty \xi_n$ converges a.s.

**Proof** Writing $S_n = \sum_1^n \xi_i$, and noting that $S_r - S_n$ is (for $r > n$) the sum of $r - n$ r.v.’s $\xi_i$, we have by Theorem 11.5.1

$$P(\max_{n \leq r \leq k} |S_r - S_n| > \epsilon) \leq \sum_{i=n+1}^k \frac{\sigma_i^2}{\epsilon^2}$$

so that

$$\lim_{k \to \infty} P(\max_{n \leq r \leq k} |S_r - S_n| > \epsilon) \leq \sum_{i=n+1}^\infty \frac{\sigma_i^2}{\epsilon^2}$$

which tends to zero as $n \to \infty$ by virtue of the convergence of $\sum_1^\infty \sigma_i^2$. Hence the result follows immediately from Lemma 11.5.2.

The next result is the celebrated “Three Series Theorem”, which gives necessary and sufficient conditions for a.s. convergence of series of independent r.v.’s, without assuming existence of any moments of the terms.

**Theorem 11.5.4** (Kolmogorov’s Three Series Theorem) Let $\{\xi_n : n = 1, 2, \ldots\}$ be independent r.v.’s and let $c$ be a positive constant. Write $E_n = \{\omega : |\xi_n(\omega)| \leq c\}$ and define $\xi_n^c(\omega) = \xi_n(\omega)$ or $c$ according as $\omega \in E_n$ or $\omega \in E_n^c$. Then a necessary and sufficient condition for the convergence (a.s.) of $\sum_1^\infty \xi_n$ is the convergence of all three of the series

- (a) $\sum_1^\infty P(E_n^c)$
- (b) $\sum_1^\infty \mathbb{E} \xi_n^c$
- (c) $\sum_1^\infty \sigma_n^2$

$\sigma_n^2$ being the variance of $\xi_n^c$. 

11.5 Series of independent r.v.’s

Proof To see the sufficiency of the conditions note that (a) may be rewritten as \( \sum P(\xi_n \neq \xi_n') \), and convergence of this series implies (a.s.), by the Borel–Cantelli Lemma, that \( \xi_n(\omega) = \xi_n'(\omega) \) when \( n \) is sufficiently large (how large, depending on \( \omega \)). Hence \( \sum \xi_n \) converges a.s. if and only if \( \sum \xi_n' \) does.

But by Theorem 11.5.3 applied to \( \xi_n' - \mathcal{E}_{\xi_n'} \) (using (c), \( \mathcal{E}(\xi_n' - \mathcal{E}_{\xi_n'})^2 = \sigma_n'^2 \)) we have that \( \sum(\xi_n' - \mathcal{E}_{\xi_n'}) \) converges a.s. Hence by (b) \( \sum \xi_n' \) converges a.s., and, by the discussion above, so does \( \sum \xi_n' \), as required.

Conversely, suppose that \( \sum_1^\infty \xi_n \) converges a.s. Since this implies that \( \xi_n \to 0 \) a.s. we must have \( \xi_n = \xi_n' \) a.s. when \( n \) is sufficiently large, and hence \( \sum P(\xi_n \neq \xi_n') < \infty \) by Theorem 10.5.2. That is, condition (a) holds, and further \( \sum \xi_n' \) converges a.s.

Now let \( \eta_n, \xi_n \) be r.v.’s with the same distributions as \( \xi_n' \) and such that \( \{\eta_n, \xi_n : n = 1, 2, \ldots\} \) are all independent as a family. (Such r.v.’s may be readily constructed using product spaces.) It is easily shown (cf. Ex. 11.30) that \( \sum \eta_n \) and \( \sum \xi_n \) both converge a.s. (since \( \sum \xi_n' \) does) and hence so does \( \sum(\eta_n - \xi_n) \). Writing \( S_k = \sum_1^k(\eta_n - \xi_n) \) we have, in particular, that the series \( \{|S_k| : k = 1, 2, \ldots\} \) is bounded for a.e. \( \omega \), i.e. \( P\{\sup_{k \geq 1} |S_k| < \infty\} = 1 \), and hence \( \lim_{a \to \infty} P\{\sup_{k \geq 1} |S_k| < a\} = 1 \) so that \( P\{\sup_{k \geq 1} |S_k| < a\} > \theta \) for some \( \theta > 0 \), \( a > 0 \). Thus, for any \( n \), \( P\{\max_{1 \leq k \leq n} |S_k| < a\} > \theta \). But Theorem 11.5.1 (ii) applies to the r.v.’s \( \eta_k - \xi_k \) (with variance \( 2\sigma_k^2 \), and writing \( 2c \) for \( c \), to give \( (2c + a)^2/(2 \sum n \sigma_k^2) \) for all \( a \)). That is, for all \( n \)

\[
\sum_1^n \sigma_k^2 < (2c + a)^2/2\theta
\]

which shows that \( \sum_1^\infty \sigma_k^2 \) converges; i.e. (c) holds.

(b) is now easily checked, since the sequence of r.v.’s \( \xi_n' - \mathcal{E}_{\xi_n} \) have zero means, and the sum of their variances \( \sum \sigma_k^2 \) is finite. Hence \( \sum(\xi_n' - \mathcal{E}_{\xi_n'}) \) converges a.s., as does \( \sum \xi_n' \). By choosing some fixed \( \omega \) where convergence (of both) takes place, we see that \( \sum \mathcal{E}_{\xi_n'} \) must converge, concluding the proof of the theorem. \(\square\)

Note that it follows from the theorem that if the series (a), (b), (c) converge for some \( c > 0 \), they converge for all \( c > 0 \). Note also that the proof of the theorem will apply if \( \xi_n'(\omega) \) is defined to be zero (rather than \( c \)) when \( \omega \in \mathcal{E}_n' \). This definition of \( \mathcal{E}_n' \) can be simpler in practice.

Convergence in probability does not usually imply convergence a.s. Our final task in this section is to show, however, that convergence of a series of independent r.v.’s in probability does imply its convergence a.s.
**Theorem 11.5.5** Let \( \{ \xi_n \} \) be a sequence of independent r.v.’s. Then the series \( \sum_1^\infty \xi_n \) converges in probability if and only if it converges a.s.

**Proof** Certainly convergence a.s. implies convergence in probability. By Lemma 11.5.2 (using \( 2\epsilon \) in place of \( \epsilon \)) the result will follow if it is shown that for each \( \epsilon > 0 \)

\[
\lim_{k \to \infty} P\{ \max_{n \leq r \leq k} |S_r - S_n| > 2\epsilon \} \to 0, \text{ as } n \to \infty,
\]

with \( S_n = \sum_1^n \xi_i \). Instead of appealing to Kolmogorov’s Inequality (as in the previous theorem), the convergence in probability may be used to obtain this as follows.

If \( n < r \leq k \) and \( |S_r - S_n| > 2\epsilon, |S_k - S_r| \leq \epsilon \) then

\[
|S_k - S_n| = |(S_r - S_n) - (S_r - S_k)| \geq |S_r - S_n| - |S_r - S_k| > \epsilon
\]

and hence

\[
\bigcup_{r=n+1}^k \{ \omega : \max_{n \leq j < r} |S_j - S_n| \leq 2\epsilon, |S_r - S_n| > 2\epsilon, |S_k - S_r| \leq \epsilon \}
\]

\[
\subset \{ \omega : |S_k - S_n| > \epsilon \}.
\]

The sets of the union are disjoint. Also \( \max_{n < j < r} |S_j - S_n| \) and \( |S_r - S_n| \) depend on \( \xi_{n+1}, \ldots, \xi_r \), whereas \( S_k - S_r \) depends on \( \xi_{r+1}, \ldots, \xi_k \). Hence, using independence of the \( \xi_i \),

\[
\sum_{r=n+1}^k P\{ \max_{n \leq j < r} |S_j - S_n| \leq 2\epsilon, |S_r - S_n| > 2\epsilon \} P\{|S_k - S_r| \leq \epsilon \}
\]

\[
\leq P\{|S_k - S_n| > \epsilon \}.
\]

Since \( \sum_1^\infty \xi_n \) converges in probability, \( \{ S_n \} \) is a Cauchy sequence in probability, and hence, given \( \eta > 0 \), there is an integer \( N \) with \( P\{|S_k - S_n| > \epsilon \} < \eta \) when \( k, n \geq N \). Hence also \( P\{|S_k - S_r| \leq \epsilon \} > 1 - \eta \) if \( k \geq r \geq N \), giving

\[
\sum_{r=n+1}^k P\{ \max_{n \leq j < r} |S_j - S_n| \leq 2\epsilon, |S_r - S_n| > 2\epsilon \} \leq \eta/(1 - \eta)
\]

if \( k > n \geq N \). Rephrasing this, we have

\[
P\{ \max_{n \leq r \leq k} |S_r - S_n| > 2\epsilon \} \leq \eta/(1 - \eta)
\]

and hence

\[
\lim_{k \to \infty} P\{ \max_{n \leq r \leq k} |S_r - S_n| > 2\epsilon \} \leq \eta/(1 - \eta) \text{ for } n \geq N,
\]

concluding the proof. □
It may even be shown that if a series $\sum_1^\infty \xi_n$ of independent r.v.’s converges in distribution it converges in probability and hence a.s. Since we shall use characteristic functions to prove it, the explicit statement and proof of this still stronger result is deferred to the next chapter (Theorem 12.5.2).

11.6 Laws of large numbers

The last section concerned convergence of series of independent r.v.’s $\sum_1^\infty \xi_n$. For convergence it is necessary in particular that the terms tend to zero i.e. $\xi_n \to 0$ a.s. Thus the discussion there certainly does not apply to any (nontrivial) independent sequences for which the terms have the same distributions. It is mainly to such “independent and identically distributed” (i.i.d.) random variables that the present section will apply.

Specifically we shall consider an independent sequence $\{\xi_n\}$ with $S_n = \sum_1^n \xi_i$ and obtain conditions under which the averages $S_n/n$ converge to a constant either in probability or a.s., as $n \to \infty$. For i.i.d. random variables with a finite mean, the constant will turn out to be $\mu = E\xi_i$. Results of this type are usually called laws of large numbers, convergence in probability being called a weak law and convergence with probability one a strong law.

Two versions of the strong law will be given – one applying to independent r.v.’s with finite second moments (but not necessarily having the same distributions), and the other applying to i.i.d. r.v.’s with finite first moments. Since convergence a.s. implies convergence in probability, weak laws will follow trivially as corollaries. However, the weak law for i.i.d. r.v.’s may also be easily obtained directly by use of characteristic functions as will be seen in the next chapter.

**Lemma 11.6.1** If $\{y_n\}$ is a sequence of real numbers such that $\sum_{n=1}^\infty y_n/n$ converges, then $\frac{1}{n} \sum_{i=1}^n y_i \to 0$ as $n \to \infty$.

**Proof** Writing $s_n = \sum_{i=1}^n y_i/i$ ($s_0 = 0$), $t_n = \sum_{i=1}^n y_i$, it is easily checked that $t_n/n = \frac{1}{n} \sum_{i=1}^{n-1} s_i + s_n$. Since $\frac{1}{n} \sum_{i=1}^n s_i$ is well known (or easily shown) to converge to the same limit as $s_n$ it follows that $t_n/n \to 0$, which is the result required. \(\square\)

The first form of the strong law of large numbers requires the independent r.v.’s $\xi_n$ to have finite variances but not necessarily to be identically distributed.
**Theorem 11.6.2** (Strong Law, First Form)  If $\xi_n$ are independent r.v.’s with finite means $\mu_n$ and finite variances $\sigma_n^2$, satisfying $\sum_{n=1}^{\infty} \sigma_n^2/n^2 < \infty$, then

$$\frac{1}{n} \sum_{i=1}^{n} (\xi_i - \mu_i) \to 0 \text{ a.s.}$$

In particular if $\frac{1}{n} \sum_{i=1}^{n} \mu_i \to \mu$ (e.g. if $\mu_n \to \mu$) then $\frac{1}{n} \sum_{i=1}^{n} \xi_i \to \mu$ a.s.

**Proof**  It is sufficient to consider the case where $\mu_n = 0$ for all $n$ since the general case follows by replacing $\xi_i$ by $(\xi_i - \mu_i)$. Assume then that $\mu_n = 0$ for all $n$ and write $\eta_n(\omega) = \xi_n(\omega)/n$. Then $E(\eta_n) = 0$ and

$$\sum_{n=1}^{\infty} \text{var}(\eta_n) = \sum_{n=1}^{\infty} \frac{\sigma_n^2}{n^2} < \infty.$$ 

Thus by Theorem 11.5.3, $\sum_{n=1}^{\infty} \xi_n/n = \sum_{n=1}^{\infty} \eta_n$ converges a.s. and the desired conclusion follows at once from Lemma 11.6.1.

The following result also yields the most common form of the strong law, which applies to i.i.d. r.v.’s (but only assumes the existence of first moments).

**Theorem 11.6.3** (Strong Law, Second Form)  Let $\{\xi_n\}$ be independent and identically distributed r.v.’s with (the same) finite mean $\mu$. Then,

$$\frac{1}{n} \sum_{i=1}^{n} \xi_i \to \mu \text{ a.s. as } n \to \infty.$$ 

**Proof**  Again, if the result holds when $\mu = 0$, replacing $\xi_i$ by $(\xi_i - \mu)$ shows that it holds when $\mu \neq 0$. Hence we assume that $\mu = 0$.

Write $\eta_n(\omega) = \xi_n(\omega)$ if $|\xi_n(\omega)| \leq n$, $\eta_n(\omega) = 0$ otherwise (for $n = 1, 2, \ldots$). First it will be shown that $\frac{1}{n} \sum_{i=1}^{n} (\xi_i - \eta_i) \to 0$ a.s. We have

$$\sum_{n=1}^{\infty} P(\xi_n \neq \eta_n) = \sum_{n=1}^{\infty} P(|\xi_n| > n) = \sum_{n=1}^{\infty} (1 - F(n))$$ 

where $F$ is the (common) d.f. of the $|\xi_n|$. But $1 - F(n) \leq 1 - F(x)$ for $n - 1 < x \leq n$ so that

$$\sum_{n=1}^{\infty} (1 - F(n)) \leq \int_{0}^{\infty} (1 - F(x)) \, dx = E|\xi_1| < \infty$$ 

by e.g. Ex. 9.16, so that $\sum_n P(\xi_n \neq \eta_n) < \infty$. Hence by the Borel–Cantelli Lemma, for a.e. $\omega$, $\xi_n(\omega) = \eta_n(\omega)$ when $n$ is sufficiently large and hence it follows at once that $\frac{1}{n} \sum_{i=1}^{n} (\xi_i - \eta_i) \to 0$ a.s.
The proof will be completed by showing that \( \frac{1}{n} \sum_{i=1}^{n} \eta_i \to 0 \) a.s. Note first that the variance of \( \eta_n \) satisfies

\[
\text{var}(\eta_n) \leq \mathbb{E}\eta_n^2 = \int_{|x| \leq n} x^2 \, dF(x)
\]
since the \( |\xi_i| \) have d.f. \( F \). Hence

\[
\sum_{n=1}^{\infty} n^{-2} \text{var}(\eta_n) \leq \sum_{n=1}^{\infty} n^{-2} \int_{|x| \leq n} x^2 \, dF(x)
\]

\[
= \sum_{n=1}^{\infty} n^{-2} \sum_{k=1}^{n} \int_{|(k-1)\leq x \leq k|} x^2 \, dF(x)
\]

\[
= \sum_{k=1}^{\infty} \int_{|(k-1)\leq x \leq k|} x^2 \, dF(x) \sum_{n=k}^{\infty} n^{-2}
\]

\[
\leq \sum_{k=1}^{\infty} (C/k) \int_{|(k-1)\leq x \leq k|} x^2 \, dF(x)
\]

where \( C \) is a constant such that \( \sum_{n=k}^{\infty} 1/n^2 < C/k \) for all \( k = 1, 2, \ldots \). (It is easily proved that such a \( C \) exists – e.g. by dominating the sum by an integral.) Hence

\[
\sum_{n=1}^{\infty} n^{-2} \text{var}(\eta_n) \leq \sum_{k=1}^{\infty} C \int_{|(k-1)\leq x \leq k|} x \, dF(x) = C\mathbb{E}|\xi_1| < \infty.
\]

It thus follows from Theorem 11.6.2 (since the \( \eta_n \) are clearly independent) that \( n^{-1} \sum_{i=1}^{n} (\eta_i - \mathbb{E}\eta_i) \to 0 \) a.s. But \( \mathbb{E}\eta_n = \mathbb{E}(\xi_n \chi_{|\xi_n| \leq n}) = \mathbb{E}\xi_n - \mathbb{E}(\xi_n \chi_{|\xi_n| > n}) = -\mathbb{E}(\xi_n \chi_{|\xi_n| > n}) \) since \( \mathbb{E}\xi_n = 0 \). Hence \( |\mathbb{E}\eta_n| \leq \mathbb{E}(|\xi_n| \chi_{|\xi_n| > n}) = \int_{n}^{\infty} x \, dF(x) \to 0 \) as \( n \to \infty \) (\( \mathbb{E} |\xi_n| < \infty \)). Thus \( n^{-1} \sum_{i=1}^{n} \mathbb{E}\eta_i \to 0 \) so that by the above \( n^{-1} \sum_{i=1}^{n} \eta_i \to 0 \) a.s., as required to complete the proof. \( \square \)

**Exercises**

11.1 Let \( \{\xi_n\}_{n=1}^{\infty} \) be a sequence of r.v.’s with \( \mathbb{E}\xi_n^2 < \infty \) and let

\[
\mu_n = \mathbb{E}\xi_n, \quad \sigma_n^2 = \text{var}(\xi_n).
\]

If \( \mu_n \to \mu \) and \( \sum_{n=1}^{\infty} \sigma_n^2 < \infty \), show that \( \xi_n \to \mu \) a.s.

11.2 Let \( \{\xi_n\}_{n=1}^{\infty} \) be a sequence of random variables on the probability space \( (\Omega, \mathcal{F}, P) \) and \( \{c_n\}_{n=1}^{\infty} \) a sequence of positive numbers. Define the truncation of \( \xi_n \) at \( c_n \) by \( \eta_n = \xi_n \chi_{A_n} \), where

\[
A_n = \{\omega \in \Omega : |\xi_n(\omega)| > c_n\}.
\]

Prove that if \( \sum_{n=1}^{\infty} P(A_n) < \infty \) and if \( \eta_n \to \xi \) almost surely, then \( \xi_n \to \xi \) almost surely.
11.3 Prove that $\xi_n \to \xi$ in probability if and only if
\[
\lim_{n \to \infty} \mathcal{E} \left( \frac{|\xi_n - \xi|}{1 + |\xi_n - \xi|} \right) = 0.
\]

11.4 The result of Ex. 11.3 may be expressed in terms of a “metric” $d$ on the “space” of r.v.’s, provided we regard two r.v.’s which are equal a.s. as being the same in the space. Define $d(\xi, \eta) = \mathcal{E} \left( \frac{|\xi - \eta|}{1 + |\xi - \eta|} \right)$ ($d$ is well defined for any $\xi, \eta$). Then $d(\xi, \eta) \geq 0$ with equality only if $\xi = \eta$ a.s., and $d(\xi, \eta) = d(\eta, \xi)$ for all $\xi, \eta$. Show that the “triangle inequality” holds, i.e.
\[
d(\xi, \zeta) \leq d(\xi, \eta) + d(\eta, \zeta)
\]
for any $\xi, \eta, \zeta$. (Hint: For any $a, b$ it may be shown that $\frac{|a+b|}{1+|a+b|} \leq \frac{|b|}{1+|b|} + \frac{|a|}{1+|a|}$.)

Ex. 11.3 may then be restated as “$\xi_n \to \xi$ in probability if and only if $d(\xi_n, \xi) \to 0$, i.e. $\xi_n \to \xi$ in this metric space”.

11.5 Show that the statement “If $\mathcal{E}\xi_n \to 0$ then $\xi_n \to 0$ in probability” is false, though the statement “If $\xi_n \geq 0$, and $\mathcal{E}\xi_n \to 0$ then $\xi_n \to 0$ in probability” is true.

11.6 Let $\{\xi_n\}$ be a sequence of r.v.’s. Show that there exist constants $A_n$ such that $\xi_n/A_n \to 0$ a.s.

11.7 If $\xi_n \to \xi$ a.s. show that given $\epsilon > 0$ there exists $M$ such that $P[\sup_{n \geq 1} |\xi_n| \leq M] > 1 - \epsilon$.

11.8 Complement the uniqueness statement in Theorem 11.2.1 by showing explicitly that if $\{\pi_n : n = 1, 2, \ldots\}$, $\pi$, $\pi^*$ are probability measures on $(\mathbb{R}, B)$ such that $\pi_n \rightharpoonup \pi$, $\pi_n \rightharpoonup \pi^*$, then $\pi = \pi^*$ on $B$. (Consider the corresponding d.f.’s.)

11.9 Let $\{F_n\}$ be a sequence of d.f.’s with corresponding probability measures $\{\pi_n\}$. Show directly from the definitions that if $\pi_n \rightharpoonup \pi$ then $F_n \rightharpoonup F$. (Hint: Show that if $a, x$ are continuity points of $F$ then $\lim \inf_{n \to \infty} F_n(x) \geq F(x) - F(a)$, and let $a \to -\infty$.)

11.10 Show that in the definition $\pi_n(a, b] \to \pi(a, b]$ for all finite $a, b$ for weak convergence of probability measures $\pi_n \rightharpoonup \pi$, intervals $(a, b]$ or open intervals $(a, b)$ may be equivalently used. For example show that if $\pi_n \rightharpoonup \pi$ then $\pi_n[b] \to \pi[b]$ for any $b$ such that $\pi\{b\} = 0$, and that this also holds under the alternative assumptions replacing semiclosed intervals by open or by closed intervals.

11.11 Prove the assertion needed in Corollary 1, Theorem 11.2.1 that if $\pi$ is a probability measure on $B$ and $g$ is a nonnegative bounded $B$-measurable function which is continuous a.e. ($\pi$) then a sequence $\{g_n\}$ of continuous functions may be found with $0 \leq g_n(x) \uparrow g(x)$ at each continuity point $x$ of $g$.

This may be shown by defining continuous functions $h_1, h_2, \ldots$ such that $0 \leq h_n(x) \leq g(x)$ and $\sup_n h_n(x) = g(x)$, and writing $g_n(x) = \max_{1 \leq i \leq n} h_i(x)$.
Exercises 251

11.12 Let \( \{\xi_n\}_{n=1}^{\infty} \) be r.v.’s with d.f.’s \( \{F_n\}_{n=1}^{\infty} \), \( F \) respectively. Assume that \( \xi_n \xrightarrow{P} \xi \). Show that given \( \epsilon > 0 \),
\[
F_n(x) \leq F(x + \epsilon) + P(|\xi_n - \xi| \geq \epsilon) \\
F(x - \epsilon) \leq F_n(x) + P(|\xi_n - \xi| \geq \epsilon).
\]
Hence show that \( \xi_n \xrightarrow{d} \xi \) (by this alternative method to that of Theorem 11.3.1).

11.13 Convergence in distribution does not necessarily imply convergence in probability. However, if \( \xi_n \xrightarrow{d} \xi \) and \( \xi(\omega) = a \), constant almost surely then \( \xi_n \xrightarrow{P} \xi \) in probability.

11.14 Let \( \{\xi_n\} \), \( \xi \) be r.v.’s such that \( \xi_n \xrightarrow{d} \xi \).

(i) If each \( \xi_n \) is discrete, can \( \xi \) be absolutely continuous?

(ii) If each \( \xi_n \) is absolutely continuous, can \( \xi \) be discrete?

11.15 Let \( \{\xi_n\}_{n=1}^{\infty} \) and \( \xi \) be random variables on \( (\Omega, \mathcal{F}, \mathbb{P}) \) such that for each \( n \) and \( k = 0, 1, \ldots, n \),
\[
P(\xi_n = k/n) = 1/(n+1),
\]
and \( \xi \) has the uniform distribution on \([0, 1]\). Prove that \( \xi_n \xrightarrow{d} \xi \).

11.16 Let \( \{\xi_n\}_{n=1}^{\infty} \) and \( \xi \) be random variables on \( (\Omega, \mathcal{F}, \mathbb{P}) \) and let \( \xi_n = x_n \) (constant) a.s. for all \( n = 1, 2, \ldots \). Prove that \( \xi_n \xrightarrow{d} \xi \) if and only if the sequence of real numbers \( \{x_n\}_{n=1}^{\infty} \) converges and \( \xi = \lim_n x_n \) a.s.

11.17 Let the random variables \( \{\xi_n\}_{n=1}^{\infty} \) and \( \xi \) have densities \( \{f_n\}_{n=1}^{\infty} \) and \( f \) respectively with respect to Lebesgue measure \( m \). If \( f_n \xrightarrow{a.e.} (m) \) on the real line \( \mathbb{R} \), show that \( \xi_n \xrightarrow{d} \xi \). (Hint: Prove that \( f_n \xrightarrow{f} f \) in \( L_1(\mathbb{R}, \mathcal{B}, m) \) by looking at the positive and negative parts of \( f - f_n \).

11.18 Let \( \{\pi_n\}_{n=1}^{\infty} \), \( \pi \) be probability measures on \( \mathcal{B} \). Show that if \( \pi_n \xrightarrow{w} \pi \) then \( \{\pi_n\}_{n=1}^{\infty} \) is tight.

11.19 Weak convergence of d.f.’s, may also be expressed in terms of a metric. If \( F, G \) are d.f.’s, the “Lévy distance” \( d(F, G) \) is defined by \( d(F, G) = \inf(\epsilon > 0 : G(x - \epsilon) - \epsilon \leq F(x) \leq G(x + \epsilon) + \epsilon \) for all real \( x \), show that \( d \) is a metric, and \( F_n \xrightarrow{d} F \) if and only if \( d(F_n, F) \xrightarrow{0} 0 \).

11.20 Prove Theorem 11.3.3, i.e. that for finite measures \( \mu_n, \mu \) on \( \mathcal{B} \), \( \mu_n \xrightarrow{w} \mu \) if and only if \( \mu_n \xrightarrow{v} \mu \) and \( \mu_n(\mathbb{R}) \xrightarrow{\mu(\mathbb{R})} \) as \( n \xrightarrow{\infty} \).

11.21 Suppose \( \{\xi_u : u \in U\}, \{\eta_v : v \in V\} \) are each uniformly integrable families. Show that the family \( \{\xi_u + \eta_v : u \in U, \ v \in V\} \) is uniformly integrable.

11.22 If the random variables \( \{\xi_n\}_{n=1}^{\infty} \) are identically distributed with finite means, then \( \xi_n \xrightarrow{P} \xi \) in probability if and only if \( \xi_n \xrightarrow{d} \xi \) in \( L_1 \).

11.23 If the random variables \( \{\xi_n\}_{n=1}^{\infty} \) are such that \( \sup_{\theta_n} E(\xi_n^p) < \infty \) for some \( p > 1 \), show that \( \{\xi_n\}_{n=1}^{\infty} \) is uniformly integrable.
As a consequence, show that if the random variables \( \{ \xi_n \}_{n=1}^{\infty} \) have uniformly bounded second moments, then \( \xi_n \to \xi \) in probability if and only if \( \xi_n \to \xi \) in \( L_1 \).

11.24 Let \( \{ \xi_n \} \) be r.v.’s with \( \mathbb{E}[\xi_n] < \infty \) for each \( n \). Show that the family \( \{ \xi_n : n = 1, 2, \ldots \} \) is uniformly integrable if and only if the family \( \{ \xi_n : n \geq N \} \) is uniformly integrable for some integer \( N \). Indeed this holds if given \( \varepsilon > 0 \) there exist \( N = N(\varepsilon), A = A(\varepsilon) \) such that \( \int_{|\xi_n| \geq A} |\xi_n| \, dP < \varepsilon \) for all \( n \geq N(\varepsilon) \), \( A \geq A(\varepsilon) \). Show that a corresponding statement holds for uniform absolute continuity of the families \( \{ \int_\gamma |\xi_n| \, dP : n \geq 1 \} \) and \( \{ \int_\gamma |\xi_n| \, dP : n \geq N \} \).

11.25 Let \( \{ \xi_n \}_{n=1}^{\infty} \) be a sequence of independent random variables such that \( \xi_n = \pm 1 \) each with probability 1/2 and let \( \{ a_n \}_{n=1}^{\infty} \) be a sequence of real numbers.

(i) Find a necessary and sufficient condition for the series \( \sum_{n=1}^{\infty} a_n \xi_n \) to converge a.s.

(ii) If \( a_n = 2^{-n} \) prove that \( \sum_{n=1}^{\infty} a_n \xi_n \) has the uniform distribution over \([-1, 1]\).

11.26 Let \( \{ \xi_n \}_{n=1}^{\infty} \) be a sequence of independent random variables such that for every \( n \), \( \xi_n \) has the uniform distribution on \([-n^{1/3}, n^{1/3}]\). Find the probability of convergence of the series \( \sum_{n=1}^{\infty} \xi_n \) and of the sequence \((1/n) \sum_{k=1}^{n} \xi_k \) as \( n \to \infty \).

11.27 The random series \( \sum_{n=1}^{\infty} \pm 1/n \) is formed where the signs are chosen independently and the probability of a positive sign for the \( n \)th term is \( p_n \). Express the probability of convergence of the series in terms of the sequence \( \{ p_n \}_{n=1}^{\infty} \).

11.28 Let \( \{ \xi_n \}_{n=1}^{\infty} \) be a sequence of independent r.v.’s such that each \( \xi_n \) has the uniform distribution on \([a_n, 2a_n], a_n > 0 \). Show that the series \( \sum_{n=1}^{\infty} \xi_n \) converges a.s. if and only if \( \sum_{n=1}^{\infty} a_n < \infty \). What happens if \( \sum_{n=1}^{\infty} a_n = +\infty ? \)

11.29 Let \( \{ \xi_n \}_{n=1}^{\infty} \) be a sequence of nonnegative random variables such that for each \( n \), \( \xi_n \) has the density \( \lambda_n e^{-\lambda_n x} \) for \( x \geq 0 \), where \( \lambda_n > 0 \).

(i) If \( \sum_{n=1}^{\infty} 1/\lambda_n < \infty \) show that \( \sum_{n=1}^{\infty} \xi_n < \infty \) almost surely.

(ii) If the random variables \( \{ \xi_n \}_{n=1}^{\infty} \) are independent show that

\[
\sum_{n=1}^{\infty} 1/\lambda_n < \infty \text{ if and only if } \sum_{n=1}^{\infty} \xi_n < \infty \text{ a.s.}
\]

and

\[
\sum_{n=1}^{\infty} 1/\lambda_n = \infty \text{ if and only if } \sum_{n=1}^{\infty} \xi_n = \infty \text{ a.s.}
\]

11.30 Let \( \{ \xi_n \}, \{ \xi_n^+ \} \) be two sequences of r.v.’s such that, for each \( n \), the joint distribution of \( (\xi_1, \ldots, \xi_n) \) is the same as that of \( (\xi_1^+, \ldots, \xi_n^+) \). Show that \( P(\sum_{n=1}^{\infty} \xi_n \text{ converges}) = P(\sum_{n=1}^{\infty} \xi_n^+ \text{ converges}) \). (Hint: If \( D, D^+ \) denote respectively the sets where \( \sum \xi_n, \sum \xi_n^+ \) do not converge, use e.g. the expression for \( P(D) \) in
the proof of Lemma 11.5.2, and the corresponding expression for $P(D^n)$ to show that $P(D) = P(D^*)$.

In particular this result applies if $\{\xi_n\}, \{\xi^*_n\}$ are each classes of independent r.v.’s and $\xi_n$ has the same distribution as $\xi^*_n$ for each $n$ – this is the case used in Theorem 11.5.4.)

11.31 For any sequence of random variables $\{\xi_n\}_{n=1}^{\infty}$ prove that

(i) if $\xi_n \to 0$ a.s. then $(1/n) \sum_{k=1}^{n} \xi_k \to 0$ a.s.

(ii) if $\xi_n \to 0$ in $L_p$, $p > 1$, then $(1/n) \sum_{k=1}^{n} \xi_k \to 0$ in $L_p$ and hence also in probability.

11.32 Let $\{\xi_n\}_{n=1}^{\infty}$ be a sequence of independent and identically distributed r.v.’s with

\[ \mathbb{E} \xi_n = \mu \neq 0 \quad \text{and} \quad \mathbb{E} \xi_n^2 = \sigma^2 < \infty. \]

Find the a.s. limit of the sequence

\[ \frac{\xi_1^2 + \cdots + \xi_n^2}{\xi_1 + \cdots + \xi_n}. \]

11.33 Let $\{\xi_n\}_{n=1}^{\infty}$ be a sequence of independent and identically distributed random variables and $S_n = \sum_{i=1}^{n} \xi_i$. If $\mathbb{E}(|\xi_1|) = +\infty$ prove that

\[ \lim_{n \to \infty} \sup \frac{|S_n|}{n} = +\infty \text{ a.s.} \]

It then follows from the strong law of large numbers that $(1/n) \sum_{k=1}^{n} \xi_k$ converges a.s. if and only if $\mathbb{E}(|\xi_1|) < +\infty$.

(Hint: Use Ex. 9.15 to conclude that for every $a > 0$ the events $\{\omega \in \Omega : |\xi_n(\omega)| \geq an\}$ occur infinitely often with probability one.)
12

Characteristic functions and central limit theorems

12.1 Definition and simple properties

This chapter is concerned with one of the most useful tools in probability theory – the characteristic function of a r.v. (not to be confused with the characteristic function (i.e. indicator) of a set). We shall investigate properties of such functions, and some of their many implications especially concerning independent r.v.’s and central limit theory. Chapter 8 should be reviewed for the needed properties of integrals of complex-valued functions and basic Fourier Theory.

If \( \xi \) is a r.v. on a probability space \((\Omega, \mathcal{F}, P)\), \( e^{it\xi(\omega)} \) is a complex \( \mathcal{F} \)-measurable function (Chapter 8) (and therefore will be called a complex r.v.). The integration theory of Section 8.1 applies and \( \mathcal{E}\xi \) will be used for \( \int \xi \; dP \) as for real r.v.’s. Since \( |e^{it\xi}| = 1 \) it follows that \( e^{it\xi} \in L_1(\Omega, \mathcal{F}, P) \). The function \( \phi(t) = \int e^{it\xi(\omega)} \; dP(\omega) \) (= \( \mathcal{E}e^{it\xi} \)) of the real variable \( t \) is termed the characteristic function (c.f.) of the r.v. \( \xi \).

By definition, if \( \xi \) has d.f. \( F \),

\[
\phi(t) = \mathcal{E} \cos t\xi + i\mathcal{E} \sin t\xi
= \int_{-\infty}^{\infty} \cos tx \; dF(x) + i\int_{-\infty}^{\infty} \sin tx \; dF(x)
= \int_{-\infty}^{\infty} e^{itx} \; dF(x).
\]

Thus \( \phi(t) \) is simply the Fourier–Stieltjes Transform \( F^*(t) \) of the d.f. \( F \) of \( \xi \) (cf. Section 8.2). If \( F \) is absolutely continuous, with density \( f \), it is immediate that

\[
\phi(t) = \int_{-\infty}^{\infty} e^{itx} f(x) \; dx,
\]

showing that \( \phi \) is the \( L_1 \) Fourier Transform \( f^\wedge(t) \) of the p.d.f. \( f \). If \( F \) is discrete, with mass \( p_j \) at \( x_j \), \( j = 1, 2, \ldots \), then

\[
\phi(t) = \sum_{j=1}^{\infty} p_j e^{itx_j}.
\]
Some simple properties of a c.f. are summarized in the following theorem.

**Theorem 12.1.1** A c.f. $\phi$ has the following properties

(i) $\phi(0) = 1$,
(ii) $|\phi(t)| \leq 1$, for all $t \in \mathbb{R}$,
(iii) $\phi(-t) = \overline{\phi(t)}$, for all $t \in \mathbb{R}$, where the bar denotes the complex conjugate,
(iv) $\phi$ is uniformly continuous on $\mathbb{R}$ (cf. Theorem 8.2.1).

**Proof**

(i) $\phi(0) = E1 = 1$.

(ii) $|\phi(t)| = |E e^{it\xi}| \leq |E|e^{it\xi}| = E1 = 1$, using Theorem 8.1.1 (iii).

(iii) $\phi(-t) = E e^{-it\xi} = \overline{E e^{it\xi}} = \overline{\phi(t)}$.

(iv) Let $t, s \in \mathbb{R}, \ t - s = h$. Then

$$|\phi(t) - \phi(s)| = \left|E (e^{i(s+h)\xi} - e^{i\xi})\right| = \left|E e^{i\xi} (e^{ih\xi} - 1)\right| \leq E|e^{ih\xi} - 1| (|e^{ih\xi}| = 1).$$

Now for all $\omega$ such that $\xi(\omega)$ is finite, $\lim_{h \to 0} |e^{ih\xi(\omega)} - 1| = 0$ and $|e^{ih\xi(\omega)} - 1| \leq |e^{ih\xi}| + 1 = 2$ (which is $P$-integrable). Thus by dominated convergence, $E|e^{ih\xi} - 1| \to 0$ as $h \to 0$. Finally this means that given $\epsilon > 0$ there exists $\delta > 0$ such that $E|e^{ih\xi} - 1| < \epsilon$ if $|h| < \delta$. Thus $|\phi(t) - \phi(s)| < \epsilon$ for all $t, s$, such that $|t - s| < \delta$ which shows uniform continuity of $\phi(t)$ on $\mathbb{R}$. □

The following result is simple but stated here for completeness.

**Theorem 12.1.2** If a r.v. $\xi$ has c.f. $\phi(t)$, and if $a, b$ are real, then the r.v. $\eta = a\xi + b$ has c.f. $e^{ibt}\phi(at)$. In particular the c.f. of $-\xi$ is $\phi(-t) = \overline{\phi(t)}$.

**Proof**

$$E e^{it(a\xi+b)} = e^{ibt} E e^{ia\xi} = e^{ibt}\phi(at).$$ □

In Theorem 12.1.1 it was shown that $\phi(0) = 1$ and $|\phi(t)| \leq 1$ for all $t$ if $\phi$ is a c.f. We shall see now that if $|\phi(t)| = 1$ for any nonzero $t$ then $\xi$ must be a discrete r.v. of a special kind. We shall say that a r.v. $\xi$ is of lattice type if there are real numbers $a, b (b > 0)$ such that $\xi(\omega)$ belongs to the set \{ $a + nb : n = 0, \pm 1, \pm 2, \ldots$ \} with probability one. The d.f. $F$ of such a r.v. thus has jumps at some or all of these points and is constant between them. The corresponding c.f. is, writing $p_n = P(\xi = a + nb)$,

$$\phi(t) = \sum_{-\infty}^{\infty} p_n e^{i(a+nb)t} = e^{iat} \sum_{-\infty}^{\infty} p_n e^{int}.$$

Hence $|\phi(t)| = |\sum_{-\infty}^{\infty} p_n e^{int}|$ is periodic with period $2\pi/b$. 

**Theorem 12.1.3** Let $\phi(t)$ be the c.f. of a r.v. $\xi$. Then one of the following three cases must hold:

(i) $|\phi(t)| < 1$ for all $t \neq 0$,
(ii) $|\phi(t_0)| = 1$ for some $t_0 > 0$ and $|\phi(t)| < 1$ for $0 < t < t_0$,
(iii) $\phi(t) = e^{iat}$ for all $t$, some real $a$ (and hence $|\phi(t)| = 1$ for all $t$).

In case (ii), $\xi$ is of lattice type, belonging to the set $\{a + n2\pi/t_0 : n = 0, \pm 1, \ldots\}$ a.s., for some real $a$. The absolute value of its c.f. is then periodic with period $t_0$.

In case (iii), $\xi = a$ a.s.

Finally if $\xi$ has an absolutely continuous distribution, then (i) holds. This is also the case if $\xi$ is discrete but not constant or of lattice type.

**Proof** Since $|\phi(t)| \leq 1$ it follows that either (i) holds or that $|\phi(t_0)| = 1$ for some $t_0 \neq 0$. Suppose the latter is the case. Then $\phi(t_0) = e^{i\alpha t_0}$ for some real $a$. Consider the r.v. $\eta = \xi - a$. The c.f. of $\eta$ is $\psi(t) = e^{-iat}\phi(t)$ and $\psi(t_0) = 1$. Hence

$$1 = \mathcal{E}e^{i\alpha t_0} = \int \cos(t_0 \eta(\omega)) dP(\omega)$$

since the imaginary part must vanish (to give the real value 1). Hence

$$\int [1 - \cos(t_0 \eta(\omega))] dP(\omega) = 0.$$  

The integrand is nonnegative and thus must vanish a.s. by Theorem 4.4.7. Hence $\cos(t_0 \eta(\omega)) = 1$ a.s., showing that

$$t_0 \eta(\omega) \in \{2n\pi : n = 0, \pm 1, \ldots\}$$ 

a.s.

and thus

$$\xi(\omega) \in \{a + 2n\pi/t_0 : n = 0, \pm 1, \ldots\}$$ 

a.s.

Hence $\xi$ is a lattice r.v.

Now since we assume that (i) does not hold, either (ii) holds or else every neighborhood of $t = 0$ contains such a $t_0$ with $|\phi(t_0)| = 1$. In this case a sequence $t_k \to 0$ may be found such that $\xi(\omega) \in \{a_k + n2\pi/t_k, \ n = 0, \pm 1 \ldots\}$ a.s. (for some real $a_k$), i.e. for each $k$, $\xi$ belongs to a lattice whose points are $2\pi/t_k$ apart.

At least one of the values $a_1 + 2n\pi/t_1$ has positive probability, and if (ii) does not hold, there cannot be more than one. For if there were two, distance $d$ apart we could choose $k$ so that $2\pi/t_k > d$, and obtain a contradiction since the values of $\xi$ must also lie in a lattice whose points are $2\pi/t_k$. 

apart. Thus if (ii) does not hold we have \( \xi = a \) a.s. where \( a \) is that one value of \( a_1 + 2n\pi/t_1 \) which has nonzero probability, and thus has probability 1. Hence (iii) holds and \( |\phi(t)| = |e^{iat}| = 1 \) for all \( t \); indeed \( \phi(t) = e^{iat} \). Note that if (ii) or (iii) holds, \( \xi \) is discrete. Hence \( |\phi(t)| < 1 \) for all \( t \neq 0 \) if \( \xi \) is absolutely continuous.

One of the most convenient properties of characteristic functions is the simple means of calculating the c.f. of a sum of independent r.v.’s, as contained in the following result.

**Theorem 12.1.4** Let \( \xi_1, \xi_2, \ldots, \xi_n \) be independent r.v.’s with c.f.’s \( \phi_1, \phi_2, \ldots, \phi_n \) respectively. Then the c.f. \( \phi \) of \( \eta = \xi_1 + \xi_2 + \cdots + \xi_n \) is simply the product \( \phi(t) = \phi_1(t)\phi_2(t)\ldots\phi_n(t) \).

**Proof** This follows by the analog of Theorem 10.3.5. For the complex r.v.’s \( e^{it\xi_j}, 1 \leq j \leq n \), are obviously independent, showing that \( E \prod_{1}^{n} e^{it\xi_j} = \prod_{1}^{n} E e^{it\xi_j} \). This may also be shown directly from that result by writing \( e^{it\xi_j} = \cos t\xi_j + i \sin t\xi_j \) and using independence of \( \cos t\xi_j, \sin t\xi_j \) and \( \cos t\xi_k, \sin t\xi_k \) for \( j \neq k \).

We conclude this section with a few examples of c.f.’s.

(i) **Degenerate distribution**

If \( \xi = a \) (constant) a.s. then the c.f. of \( \xi \) is \( \phi(t) = e^{ita} \).

(ii) **Binomial distribution**

\[
P(\xi = r) = \binom{n}{r} p^r (1-p)^{n-r}, \quad r = 0, 1, \ldots, n, \quad 0 < p < 1
\]

\[
\phi(t) = \sum_{r=0}^{n} \binom{n}{r} p^r (1-p)^{n-r} e^{itr} = \sum_{r=0}^{n} \binom{n}{r} (pe^{it})^{r}(1-p)^{n-r} = (1-p + pe^{it})^n = (q + pe^{it})^n,
\]

where \( q = 1-p \).

(iii) **Uniform distribution on \([-a, a]\).**

\( \xi \) has p.d.f. \( \frac{1}{2a}, \quad -a \leq x \leq a, \)

\[
\phi(t) = \frac{1}{2a} \int_{-a}^{a} e^{itx} \, dx = \frac{\sin at}{at} = e^{ita} - e^{-ita}
\]

\( (\phi(0) = 1) \).

(iv) **Normal distribution \( N(\mu, \sigma^2) \)**

\( \xi \) has p.d.f. \( \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ \frac{-(x-\mu)^2}{2\sigma^2} \right\} \)

\[
\phi(t) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{itx} \exp \left\{ \frac{-(x-\mu)^2}{2\sigma^2} \right\} \, dx.
\]
Characteristic functions and central limit theorems

This is perhaps most easily evaluated, first for \( \mu = 0, \sigma = 1 \), as a contour integral, making the substitution \( z = x - it \) to give

\[
(2\pi)^{-1/2} e^{-t^2/2} \int_C e^{-z^2/2} \, dz
\]

where \( C \) is the line \( I(z) = -t \) (\( I \) denoting “imaginary part”). This may be evaluated along the real axis instead of \( C \) (by Cauchy’s Theorem) to give \( e^{-t^2/2} \).

If \( \xi = N(\mu, \sigma^2) \), \( \eta = (\xi - \mu)/\sigma \) is \( N(0, 1) \) and thus has this c.f. \( e^{-t^2/2} \).

By Theorem 12.1.2, \( \xi \) thus has c.f. \( \phi(t) = e^{it\mu - \sigma^2 t^2/2} \).

12.2 Characteristic function and moments

The c.f. of a r.v. \( \xi \) is very useful in determining the moments of \( \xi \) (when they exist), and the d.f. or p.d.f. of \( \xi \). It is especially convenient to use the c.f. for either of these purposes when \( \xi \) is a sum of independent r.v.’s, \( \sum_i \xi_i \) say, for then the c.f. of \( \xi \) is simply obtained as the product of those of the \( \xi_i \)’s. Both uses of the c.f. and related matters are explored here, first considering the relation between existence of moments of \( \xi \) and of derivatives of \( \phi \).

**Theorem 12.2.1** Let \( \xi \) be a r.v. with d.f. \( F \) and c.f. \( \phi \). If \( \mathbb{E}|\xi|^n < \infty \) for some integer \( n \geq 1 \), then \( \phi \) has a (uniformly) continuous derivative of order \( n \) given by

\[
\phi^{(n)}(t) = i^n \mathbb{E}(\xi^n e^{it\xi}) = i^n \int_{-\infty}^{\infty} x^n e^{itx} \, dF(x),
\]

and, in particular, \( \mathbb{E}\xi^n = \phi^{(n)}(0)/n! \).

**Proof** For any \( t \), \( (\phi(t+h) - \phi(t))/h = \int e^{itx}(e^{ihx} - 1)/h \, dF(x) \). Since the function \( (e^{ihx} - 1)/h \to ix \) as \( h \to 0 \) and \( |(e^{ihx} - 1)/h| = |\int_0^x e^{ihy} \, dy| \leq |x| \), dominated convergence shows that \( \lim_{h \to 0} (\phi(t+h) - \phi(t))/h = \int_{-\infty}^{\infty} ix e^{itx} \, dF(x) \), i.e. the derivative \( \phi'(t) \) exists, given by \( \phi'(t) = \int_{-\infty}^{\infty} ix e^{itx} \, dF(x) \).

The proof may be completed by induction using the same arguments. Uniform continuity follows as for \( \phi \) itself.

**Corollary** If for some integer \( n \geq 1 \), \( \mathbb{E}|\xi|^n < \infty \) then, writing \( m_k = \mathbb{E}\xi^k \),

\[
\phi(t) = \sum_{k=0}^{n} \frac{(it)^k}{k!} m_k + o(t^n) = \sum_{k=0}^{n-1} \frac{(it)^k}{k!} m_k + \frac{\theta t^n}{n!} \mathbb{E}|\xi|^n
\]

where \( \theta = \theta_t \) is a complex number with \( |\theta_t| \leq 1 \). (The “\( o(t^n) \)” term above is to be taken as \( t \to 0 \), i.e. \( o(t^n) \) is a function \( \psi(t) \) such that \( \psi(t)/t^n \to 0 \) as \( t \to 0 \).)
Proof  The first relation follows at once from the Taylor series expansion
\[ \phi(t) = \sum_{k=0}^{n} \frac{t^k}{k!} \phi^{(k)}(0) + o(t^n). \]
The second follows from the alternative Taylor expansion
\[ \phi(t) = \sum_{k=0}^{n-1} \frac{t^k}{k!} \phi^{(k)}(0) + \frac{t^n}{n!} \phi^{(n)}(\alpha t) \quad (|\alpha| < 1), \]

defining \( \theta \) by
\[ \theta \mathbb{E}|\xi|^n = \phi^{(n)}(\alpha t) = (i)^n \int_{-\infty}^{\infty} x^n e^{itx} dF(x) \]
from which it follows that
\[ |\theta| \mathbb{E}|\xi|^n \leq \int_{-\infty}^{\infty} |x|^n dF(x) = \mathbb{E}|\xi|^n. \]
Thus \(|\theta| \leq 1\) if \( \mathbb{E}|\xi|^n > 0 \), and in the degenerate case where \( \mathbb{E}|\xi|^n = 0 \), i.e. \( \xi = 0 \) a.s., we may clearly take \( \theta = 0 \). \( \Box \)

The converse to Theorem 12.2.1 holds for derivatives and moments of even order, as shown in the following result (see also Exs. 12.12, 12.13, 12.14).

**Theorem 12.2.2**  Suppose that, for some integer \( n \geq 1 \), the c.f. \( \phi(t) \) of the r.v. \( \xi \) has \( 2n \) finite derivatives at \( t = 0 \). Then \( \mathbb{E}|\xi|^{2n} < \infty \).

**Proof**  Consider first the second derivative (i.e. \( n = 1 \)). Since \( \phi^{''} \) exists at \( t = 0 \) we have
\[ \phi(t) = \phi(0) + t\phi'(0) + \frac{1}{2} t^2 \phi^{''}(0) + o(t^2) \]
\[ \phi(-t) = \phi(0) - t\phi'(0) + \frac{1}{2} t^2 \phi^{''}(0) + o(t^2) \]
and thus by addition of these two equations,
\[ \phi^{''}(0) = \lim_{t \to 0} \frac{\phi(t) - 2\phi(0) + \phi(-t)}{t^2} \]
\[ = \lim_{t \to 0} \int_{-\infty}^{\infty} \frac{e^{itx} - 2 + e^{-itx}}{t^2} \ dF(x) \]
\[ = -2 \lim_{t \to 0} \int_{-\infty}^{\infty} \frac{1 - \cos tx}{t^2} \ dF(x) \]
\( (F \) being the d.f. of \( \xi ) \). But \( (1 - \cos tx)/t^2 \to x^2/2 \) as \( t \to 0 \) and hence by Fatou’s Lemma
\[ -\phi^{''}(0) = 2 \lim_{t \to 0} \int_{-\infty}^{\infty} \frac{1 - \cos tx}{t^2} dF(x) \geq \int_{-\infty}^{\infty} x^2 dF(x). \]
Since \(-\phi''(0)\) is (real and) finite it follows that \(\int x^2 \, dF(x) < \infty\), i.e. \(\mathbb{E} \xi^2 < \infty\).

The case for \(n > 1\) may be obtained inductively from the \(n = 1\) case as follows. Suppose the result is true for \((n - 1)\) and that \(\phi^{(2n)}(0)\) exists. Then \(\mathbb{E} \xi^{2n-2}\) exists by the inductive hypothesis and by Theorem 12.2.1

\[
\phi^{(2n-2)}(0) = (-)^{n-1} \int_{-\infty}^{\infty} x^{2n-2} \, dF(x).
\]

If \(\int_{-\infty}^{\infty} x^{2n-2} \, dF(x) = 0\), \(F\) is the d.f. of the degenerate distribution with all its mass at zero, i.e. \(\xi = 0\) a.s., so that the desired conclusion \(\mathbb{E} \xi^{2n} < \infty\) follows trivially. Otherwise write

\[
G(x) = \int_{-\infty}^{x} u^{2n-2} \, dF(u)/\int_{-\infty}^{\infty} u^{2n-2} \, dF(u).
\]

\(G\) is clearly a d.f. and has c.f. (writing \(\lambda^{-1} = \int_{-\infty}^{\infty} u^{2n-2} \, dF(u)\))

\[
\psi(t) = \int e^{itx} \, dG(x) = \lambda \int_{-\infty}^{\infty} x^{2n-2} e^{itx} \, dF(x) = \lambda (-)^{n-1} \phi^{(2n-2)}(t)
\]

(\(\lambda x^{2n-2}\) being the Radon–Nikodym derivative \(d\mu_\psi/d\mu_\phi\)). Since \(\phi^{(2n)}(0)\) exists so does \(\psi''(0)\) and by the first part of this proof (with \(n = 2\) and \(\psi\) for \(\phi\))

\[
-\psi''(0) \geq \int_{-\infty}^{\infty} x^2 \, dG(x) = \lambda \int_{-\infty}^{\infty} x^{2n} \, dF(x)
\]

(Theorem 5.6.1). Thus \(\int x^{2n} \, dF(x)\) is finite as required. \(\square\)

The corollary to Theorem 12.2.1 provides Taylor expansions of the c.f. \(\phi(t)\) when \(n\) moments exist. The following is an interesting variant of such expansions when an even number of moments exists which sheds light on the nature of the remainder term. It is given here for two moments (which will be useful in the central limit theory to be considered in Section 12.6). The extension to \(2n\) moments is evident.

**Lemma 12.2.3** Let \(\xi\) be a r.v. with zero mean, finite variance \(\sigma^2\), d.f. \(F\), and c.f. \(\phi\). Then \(\phi\) can be written as

\[
\phi(t) = 1 - \frac{1}{2} \sigma^2 \psi(t)
\]

where \(\psi\) is a characteristic function. Specifically \(\psi\) corresponds to the p.d.f.

\[
g(x) = \frac{2}{\sigma^2} \int_{-\infty}^{x} [1 - F(u)] \, du, \quad x \geq 0
\]

\[
= \frac{2}{\sigma^2} \int_{-\infty}^{x} F(u) \, du, \quad x < 0.
\]
Proof  Clearly \( g(x) \geq 0 \). Further, using Fubini’s Theorem

\[
\int_0^\infty g(x) \, dx = \frac{2}{\sigma^2} \int_0^\infty dx \int_y^\infty du \int_{(u,\infty)} dF(y)
\]

\[
= \frac{2}{\sigma^2} \int_{(0,\infty)} dF(y) \int_0^y du \int_0^x dx
\]

\[
= \frac{1}{\sigma^2} \int_{(0,\infty)} y^2 \, dF(y).
\]

Similarly

\[
\int_{-\infty}^0 g(x) \, dx = \frac{1}{\sigma^2} \int_{(-\infty,0]} y^2 \, dF(y)
\]

and hence \( \int_{-\infty}^\infty g(x) \, dx = 1 \). Thus \( g \) is a p.d.f. Now by the same inversion of integration order as above,

\[
\int_0^\infty g(x)e^{itx} \, dx = \frac{2}{it\sigma^2} \int_{(0,\infty)} dF(y) \int_0^y du \int_0^x e^{itx} \, dx
\]

\[
= \frac{2}{it\sigma^2} \int_{(0,\infty)} dF(y) \int_0^y (e^{itx} - 1) \, du
\]

\[
= \frac{2}{(it)^2\sigma^2} \int_{(0,\infty)} (e^{ity} - 1 - ity) \, dF(y).
\]

Similarly

\[
\int_{-\infty}^0 g(x)e^{itx} \, dx = \frac{2}{(it)^2\sigma^2} \int_{(-\infty,0]} (e^{ity} - 1 - ity) \, dF(y)
\]

and hence the c.f. corresponding to \( g \) is

\[
\psi(t) = \int_{-\infty}^\infty e^{itx} g(x) \, dx = \frac{2}{\sigma^2 t^2} (1 - \phi(t))
\]

since \( \int_{-\infty}^\infty y \, dF(y) = E\xi = 0 \). Thus \( \phi(t) = 1 - \frac{1}{2} \sigma^2 t^2 + 1 \sigma^2 t^2 (1 - \psi(t)) \). The final term is \( o(t^2) \) as \( t \to 0 \) since \( \psi(t) \to 1 \) so that the standard representation \( \phi(t) = 1 - \frac{1}{2} \sigma^2 t^2 + o(t^2) \) for a c.f. (with zero mean and finite second moments) also follows from this. However, the present result gives a more specific form for the \( o(t^2) \) term since \( \psi \) is known to be a c.f.

12.3 Inversion and uniqueness

The c.f. completely characterizes the distribution by specifying the d.f. \( F \) precisely. In fact since \( \phi \) is the Fourier–Stieltjes Transform of \( F \), this may
be shown from the inversion formulae of Sections 8.3 and 8.4, which are summarized as follows.

**Theorem 12.3.1** Let \( \phi \) be the c.f. of a r.v. \( \xi \) with d.f. \( F \). Then

(i) If \( \tilde{F}(x) = \frac{1}{2}(F(x) + F(x - 0)) \), for any \( a < b \),

\[
\tilde{F}(b) - \tilde{F}(a) = \lim_{T \to \infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ibt} - e^{-iat}}{-it} \phi(t) \, dt
\]

and for any real \( a \) the jump of \( F \) at \( a \) is

\[
F(a) - F(a - 0) = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} e^{-iat} \phi(t) \, dt.
\]

(ii) If \( \phi \in L_1 \), then \( F \) is absolutely continuous with p.d.f.

\[
f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-ixt} \phi(t) \, dt \quad \text{a.e.}
\]

\( f \) is continuous and thus also is the (continuous) derivative of \( F \) at each \( x \).

(iii) If \( F \) is absolutely continuous with p.d.f. \( f \) which is of bounded variation in a neighborhood of some given point \( x \), then

\[
\frac{1}{2} \{f(x + 0) + f(x - 0)\} = \lim_{T \to \infty} \frac{1}{2\pi} \int_{-T}^{T} e^{-ixt} \phi(t) \, dt.
\]

If \( \phi \in L_1 \) this may again be written as \( \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-ixt} \phi(t) \, dt \).

**Proof** (i) follows from Theorem 8.3.1.

(ii) It follows from Theorem 8.3.3 that \( F(x) = \int_{-\infty}^{x} f(u) \, du \) where \( f \), defined as \( \frac{1}{2\pi} \int e^{-ixt} \phi(t) \, dt \), is real, continuous, and in \( L_1 \). We need to show that \( f \) is nonnegative, whence it will follow that \( f \) is a p.d.f. for \( F \). But if \( f \) were negative for some \( x \) it would, by continuity, be negative in a neighborhood of that \( x \) and hence \( F \) would be decreasing in that interval. Thus \( f(x) \geq 0 \) for all \( x \). Finally since \( f \) is continuous it follows at once that \( F'(x) = \frac{d}{dx} \int_{-\infty}^{x} f(u) \, du = f(x) \) for each \( x \).

(iii) just restates Theorem 8.4.2 and its corollary.

Theorem 12.3.1 shows that there is a one-to-one correspondence between d.f.’s and their c.f.’s and this is now stated separately.

**Theorem 12.3.2** (Uniqueness Theorem) The c.f. of a r.v. uniquely determines its d.f., and hence its distribution, and vice versa, i.e. two d.f.’s \( F_1, F_2 \) are identical if and only if their c.f.’s \( \phi_1, \phi_2 \) are identical.

**Proof** It is clear that \( F_1 \equiv F_2 \) implies \( \phi_1 \equiv \phi_2 \). For the converse assume that \( \phi_1 \equiv \phi_2 \). Then by Theorem 12.3.1 (i), \( \tilde{F}_1(b) - \tilde{F}_1(a) = \tilde{F}_2(b) - \tilde{F}_2(a) \) for
all \( a, b \) and hence, letting \( a \to -\infty \), \( \bar{F}_1(b) = \bar{F}_2(b) \) for all \( b \). But, for any d.f. \( F \), \( \lim_{b \downarrow x} \bar{F}(b) = F(x + 0) = F(x) \) and thus, for all \( x \),

\[
F_1(x) = \lim_{b \downarrow x} \bar{F}_1(b) = \lim_{b \downarrow x} \bar{F}_2(b) = F_2(x)
\]
as required. \( \square \)

### 12.4 Continuity theorem for characteristic functions

In this section we shall relate weak convergence of the previous chapter to pointwise convergence of c.f.’s. It will be useful to first prove the following two results.

**Lemma 12.4.1** If \( \xi \) is a r.v. with d.f. \( F \) and c.f. \( \phi \), there exists a constant \( C > 0 \) such that for all \( a > 0 \)

\[
P(\{ |\xi| \geq a \}) = \int_{|x| \geq a} dF(x) \leq Ca \int_0^{a^{-1}} R[1 - \phi(t)] \, dt
\]

(\( R \) denoting “real part”). \( C \) does not depend on \( \xi \).

**Proof**

\[
a \int_0^{a^{-1}} R(1 - \phi(t)) \, dt = a \int_0^{a^{-1}} \{ \int_{-\infty}^{\infty} (1 - \cos tx) \, dF(x) \} \, dt
\]

\[
= \int_{-\infty}^{\infty} \{ a \int_0^{a^{-1}} (1 - \cos tx) \, dt \} \, dF(x) \quad \text{(Fubini)}
\]

\[
= \int_{-\infty}^{\infty} \left( 1 - \frac{\sin a^{-1}x}{a^{-1}x} \right) \, dF(x) \geq \int_{|x| \geq a} \left( 1 - \frac{\sin a^{-1}x}{a^{-1}x} \right) \, dF(x)
\]

\[
\geq \inf_{|t| \geq 1} \left( 1 - \frac{\sin t}{t} \right) \int_{|x| \geq a} dF(x)
\]

which gives the desired result if \( C^{-1} = \inf_{|t| \geq 1} \left( 1 - \frac{\sin t}{t} \right) \). (Note that \( C^{-1} = 1 - \sin 1 \) so that \( C \) is approximately 6.3.) \( \square \)

The next result uses this one to provide a convenient necessary and sufficient condition for tightness of a sequence of d.f.’s in terms of their c.f.’s.

**Theorem 12.4.2** Let \( \{ F_n \} \) be a sequence of d.f.’s with c.f.’s \( \{ \phi_n \} \). Then \( \{ F_n \} \) is tight if and only if \( \limsup_{n \to \infty} R(1 - \phi_n(t)) \to 0 \) as \( t \to 0 \).

**Proof** If \( \{ F_n \} \) is tight we may, given \( \epsilon > 0 \), choose \( A \) so that \( F_n(-A) < \epsilon/8 \), \( 1 - F_n(A) < \epsilon/8 \) for all \( n \) and hence

\[
R[1 - \phi_n(t)] = \int_{-\infty}^{\infty} (1 - \cos tx) \, dF_n(x) \leq \int_{|x| \leq A} (1 - \cos tx) \, dF_n(x) + \epsilon/2.
\]
Now if \( a > 0 \) and \( aA < \pi \), \( 1 - \cos tx \leq 1 - \cos aA \) for \( |x| \leq A \), \( |t| \leq a \) and thus
\[
\mathcal{R}[1 - \phi_n(t)] \leq (1 - \cos aA) + \epsilon/2
\]
when \( |t| \leq a \). Hence \( \limsup_{n \to \infty} \mathcal{R}[1 - \phi_n(t)] < \epsilon \) for \( |t| \leq a \) if \( a \) is chosen so that \( 1 - \cos aA < \epsilon/2 \), giving the desired conclusion.

Conversely suppose that \( \limsup_{n \to \infty} \mathcal{R}[1 - \phi_n(t)] \to 0 \) as \( t \to 0 \). By Lemma 12.4.1 there exists \( C \) such that for any \( a > 0 \),
\[
\int_{|x| \geq a} dF_n(x) \leq Ca \int_0^{a^{-1}} \mathcal{R}[1 - \phi_n(t)] dt.
\]
Hence by Fatou's Lemma (Theorem 4.5.4) applied to \( 2 - \mathcal{R}[1 - \phi_n(t)] \), or by Ex. 4.17,
\[
\limsup_{n \to \infty} \int_{|x| \geq a} dF_n(x) \leq Ca \int_0^{a^{-1}} \limsup_{n \to \infty} \mathcal{R}[1 - \phi_n(t)] dt.
\]
But given \( \epsilon > 0 \) the integrand on the right tends to zero by assumption and hence may be taken less than \( \epsilon/C \) for \( 0 \leq t \leq a^{-1} \) if \( a = a(\epsilon) \) is chosen to be large, and hence \( \limsup_{n \to \infty} \int_{|x| \geq a} dF_n(x) < \epsilon \). Thus there exists \( N \) such that \( \int_{|x| \geq a} dF_n(x) < \epsilon \) for all \( n \geq N \). Since the finite family \( F_1, F_2, \ldots, F_{N-1} \) is tight, \( \int_{|x| > a'} dF_n(x) < \epsilon \) for some \( a' \), \( n = 1, 2, \ldots, N - 1 \) and hence \( \int_{|x| > a} dF_n(x) < \epsilon \) for all \( n \) if \( A = \max\{a, a'\} \). This exhibits the required tightness of \( \{F_n\} \).

The following is the main result of this section (characterizing weak convergence in terms of c.f.'s).

**Theorem 12.4.3** (Continuity Theorem for c.f.’s) \( \text{Let} \ \{F_n\} \ \text{be a sequence of d.f.'s with c.f.'s} \ \{\phi_n\}. \)

(i) If \( F \) is a d.f. with c.f. \( \phi \) and if \( F_n \overset{w}{\to} F \) then \( \phi_n(t) \to \phi(t) \) for all \( t \in \mathbb{R} \).

(ii) Conversely if \( \phi \) is a complex function such that \( \phi_n(t) \to \phi(t) \) for all \( t \in \mathbb{R} \) and if \( \phi \) is continuous at \( t = 0 \), then \( \phi \) is the c.f. of a d.f. \( F \) and \( F_n \overset{w}{\to} F \).

**Proof**

(i) If \( F_n \overset{w}{\to} F \) then by Theorem 11.2.1,
\[
\int_{-\infty}^{\infty} \cos tx \, dF_n(x) \to \int_{-\infty}^{\infty} \cos tx \, dF(x) \quad \text{and} \quad \int \sin tx \, dF_n(x) \to \int \sin tx \, dF(x)
\]
and hence \( \int_{-\infty}^{\infty} e^{ix} \, dF_n(x) \to \int_{-\infty}^{\infty} e^{ix} \, dF(x) \), or \( \phi_n(t) \to \phi(t) \), as required.
(ii) Since \( \phi_n(t) \to \phi(t) \) for all \( t \), we have \( \phi(0) = \lim \phi_n(0) = 1 \) and
\[
\limsup_{n \to \infty} R[1 - \phi_n(t)] = 1 - R[\phi(t)] \to 0 \text{ as } t \to 0
\]
since \( \phi \) is continuous at \( t = 0 \). Thus by Theorem 12.4.2, \( \{F_n\} \) is tight.

If now \( \{F_{n_k}\} \) is any weakly convergent subsequence of \( \{F_n\} \), \( F_{n_k} \xrightarrow{w} F \) say where \( F \) has c.f. \( \psi \), then, by (i), \( \psi(t) = \lim_{k \to \infty} \phi_{n_k}(t) = \phi(t) \). Hence \( F \) has c.f. \( \phi \). Thus every weakly convergent subsequence has the same weak limit \( F \) (determined by the c.f. \( \phi \)), and the tight sequence \( \{F_n\} \) therefore converges weakly to \( F \) by Theorem 11.2.5, concluding the proof. \( \square \)

**Corollary** If \( \{\xi_n\} \) is a sequence of r.v.’s with d.f.’s \( \{F_n\} \) and c.f.’s \( \{\phi_n\} \), and if \( \xi \) is a r.v. with d.f. \( F \) and c.f. \( \phi \), then \( \xi_n \xrightarrow{d} \xi \) (\( F_n \xrightarrow{w} F \)) if and only if \( \phi_n(t) \to \phi(t) \) for all real \( t \).

This follows at once from the theorem since \( \phi \) is a c.f. and hence continuous at \( t = 0 \).

### 12.5 Some applications

In this section we give some applications of the continuity theorem for characteristic functions, beginning with a useful condition for a sequence of r.v.’s to converge in distribution to zero. By Theorem 12.4.3, Corollary, this is equivalent to the convergence of their c.f.’s to one on the entire real line. As shown next it suffices for this special case that the sequence of c.f.’s converges to one in some neighborhood of zero.

**Theorem 12.5.1** If \( \{\xi_n\} \) is a sequence of r.v.’s with c.f.’s \( \{\phi_n\} \), the following are equivalent

(i) \( \xi_n \to 0 \) in probability,
(ii) \( \xi_n \xrightarrow{d} 0 \),
(iii) \( \phi_n(t) \to 1 \) for all \( t \),
(iv) \( \phi_n(t) \to 1 \) in some neighborhood of \( t = 0 \).

**Proof** The equivalence of (i) and (ii) is already known from Ex. 11.13. If \( \xi_n \xrightarrow{d} 0 \) then by Theorem 12.4.3, \( \phi_n(t) \to 1 \) for all \( t \), so that (ii) implies (iii). Since (iii) implies (iv) trivially the proof will be completed by showing that (iv) implies (ii).

Suppose then that for some \( a > 0 \), \( \phi_n(t) \to 1 \) for all \( t \in [-a, a] \). Then
\[
\limsup_{n} R(1 - \phi_n(t)) = 0 \text{ for } |t| \leq a \text{ and thus Theorem 12.4.2 applies trivially to show that the sequence } \{F_n\} \text{ is tight (where } F_n \text{ is the d.f. of } \xi_n). \text{ Let } \{F_{n_k}\} \text{ be any weakly convergent subsequence of } \{F_n\}, F_{n_k} \xrightarrow{w} F, \text{ say, where } F \text{ has
c.f. $\phi$. Then $\phi_{n_k}(t) \to \phi(t)$ for all $t$ by Theorem 12.4.3 and hence $\phi(t) = 1$ for $|t| \leq a$. Thus by Theorem 12.1.3, $\phi(t) = e^{ibt}$ for all $t$ (some $b$) and since $\phi(t) = 1$ for $|t| < a$ it follows that $b = 0$ and $\phi(t) = 1$ for all $t$ so that $F(x)$ is zero for $x < 0$ and one for $x \geq 0$. This means that any weakly convergent subsequence of the tight sequence $\{F_n\}$ has the weak limit $F$ and hence by Theorem 11.2.5, $F_n \overset{w}{\to} F$. This, restated, is the desired conclusion (ii), $\xi_n \overset{d}{\to} 0$. □

Note that it is not true in general that if a sequence $\{\phi_n\}$ of c.f.’s converges to a c.f. $\phi$ in some neighborhood of $t = 0$ then it converges to $\phi$ for all $t$. It is true, however, as shown in this proof, in the special case where $\phi \equiv 1$. (Cf. Ex. 12.26 also.)

In Theorem 11.5.5 it was shown that convergence of a series of independent r.v.’s in probability implies a.s. convergence. The following result shows that convergence in distribution is even sufficient for a.s. convergence in such a case. It also provides a single necessary and sufficient condition, expressed in terms of c.f.’s, for a.s. convergence of a series of independent r.v.’s and should thus be compared with Kolmogorov’s Three Series Theorem 11.5.4.

**Theorem 12.5.2** Let $\{\xi_n\}$ be a sequence of independent r.v.’s with c.f.’s $\{\phi_n\}$. Then the following are equivalent

(i) The series $\sum_1^\infty \xi_n$ converges a.s.
(ii) $\sum_1^\infty \xi_n$ converges in probability.
(iii) $\sum_1^\infty \xi_n$ converges in distribution.
(iv) The products $\prod_{k=1}^n \phi_k(t)$ converge to a nonzero limit as $n \to \infty$, in some neighborhood of the origin.

**Proof** That (i) and (ii) are equivalent follows from Theorem 11.5.5. Clearly (ii) implies (iii), and (iii) implies (iv). The proof will be completed by showing that (iv) implies (ii).

If (iv) holds, $\prod_{k=1}^n \phi_k(t) \to \phi(t)$, say, where $\phi(t) \neq 0$ for $t \in [-a,a]$, some $a > 0$. Let $\{m_k\}, \{n_k\}$ be sequences tending to infinity as $k \to \infty$, with $n_k > m_k$. Then

$$\prod_{j=m_k}^{n_k} \phi_j(t) = \prod_{j=1}^{n_k} \phi_j(t) / \prod_{j=1}^{m_k-1} \phi_j(t) \to 1 \quad \text{as} \quad k \to \infty \quad \text{for} \quad |t| \leq a.$$ 

By Theorem 12.5.1, $\sum_{j=m_k}^{n_k} \xi_j \to 0$ in probability. Since $\{m_k\}$ and $\{n_k\}$ are arbitrary sequences it is clear that $\sum_1^n \xi_j$ is Cauchy in probability and hence $\sum_1^\infty \xi_j$ is convergent in probability, concluding the proof of the theorem. □
12.5 Some applications

The weak law of large numbers is, of course, an immediate corollary of the strong law (Theorem 11.6.3). However, as noted in Section 11.6, it is useful to also obtain it directly since the use of c.f.'s gives a very easy proof.

**Theorem 12.5.3** Let \{\xi_n\} be a sequence of independent r.v.'s with the same d.f. \(F\) and finite mean \(\mu\). Then

\[
\frac{1}{n} \sum_{i=1}^{n} \xi_i \rightarrow \mu \text{ in probability as } n \rightarrow \infty.
\]

**Proof** If \(\phi\) is the c.f. of each \(\xi_n\), the c.f. of \(S_n = \sum_{1}^{n} \xi_i\) is \((\phi(t))^n\) and that of \(S_n/n\) is \(\psi_n(t) = (\phi(t/n))^n\). But since \(\phi(t) = 1 + i\mu t + o(t)\) (Theorem 12.2.1, Corollary) we have, for any fixed \(t\), \(\phi(t/n) = 1 + i\mu n t + o(1/n)\) as \(n \rightarrow \infty\) and thus

\[
\psi_n(t) = \left(1 + i\mu n t + o\left(\frac{1}{n}\right)\right)^n.
\]

It is well known (and if not should be made so!) that the right hand side converges to \(e^{i\mu t}\) as \(n \rightarrow \infty\). Since \(e^{i\mu t}\) is the c.f. of the constant r.v. \(\mu\) it follows that \(S_n/n \overset{d}{\rightarrow} \mu\) (by Theorem 12.4.3, Corollary) and by Ex. 11.13, \(n^{-1}S_n \rightarrow \mu\) in probability. \(\square\)

The weak law of large numbers just proved shows that the average \(\frac{1}{n} \sum_{1}^{n} \xi_j\) independent and identically distributed (i.i.d.) r.v.'s is likely to lie close to \(\mu = E\xi_1\) as \(n\) becomes large. On the other hand, the simple form of the central limit theorem (CLT) to be given next shows how a limiting distribution may be obtained for \(\frac{1}{n} \sum_{1}^{n} \xi_j\) (suitably normalized). A more general form of the central limit theorem is given in the next section.

**Theorem 12.5.4** (Central Limit Theorem – Elementary Form) Let \{\xi_n\} be a sequence of independent r.v.'s with the same distribution and with finite mean \(\mu\) and variance \(\sigma^2\). Then the sequence of normalized r.v.'s

\[
Z_n = \frac{1}{\sigma \sqrt{n}} \sum_{j=1}^{n} (\xi_j - \mu) = \frac{\sqrt{n}}{\sigma} \left(1 - \frac{1}{n} \sum_{j=1}^{n} (\xi_j - \mu)\right)
\]

converges in distribution to a standard normal r.v. \(Z\) (p.d.f. \((2\pi)^{-1/2} e^{-x^2/2}\)).

**Proof** Write \(Z_n = n^{-1/2} \sum_{1}^{n} \eta_j\) where \(\eta_j = (\xi_j - \mu)/\sigma\) are independent with zero means, unit variances and the same d.f. Let \(\phi(t)\) denote their common c.f. which may (by Theorem 12.2.1, Corollary) be written as

\[
\phi(t) = 1 - t^2/2 + o(t^2).
\]
The characteristic function of $Z_n$ is by Theorems 12.1.2, 12.1.4
\[ \psi_n(t) = [\phi(tn^{-1/2})]^n \]
which may therefore (for fixed $t$, as $n \to \infty$) be written, by the corollary to Theorem 12.2.1
\[ \psi_n(t) = \left[ 1 - \frac{t^2}{2n} + o\left(\frac{1}{n}\right) \right]^n \to e^{-t^2/2} \text{ as } n \to \infty. \]
Since this limit is the c.f. corresponding to the standard normal distribution (Section 12.1), $Z_n \xrightarrow{d} Z$ by Theorem 12.4.3. \square

12.6 Array sums, Lindeberg–Feller Central Limit Theorem

As seen in the elementary form of the CLT (Theorem 12.5.4) the partial sums $\sum_{i=1}^{n} \xi_i$ of i.i.d. r.v.’s with finite second moments have a normal limit when standardized by means and standard deviations i.e.
\[ \frac{1}{\sigma \sqrt{n}} \left( \sum_{i=1}^{n} \xi_i - n\mu \right) \xrightarrow{d} N(0, 1). \]

A more general form of the result allows the $\xi_i$ to have different distributions with finite second moments and gives necessary and sufficient conditions for this normal limit. This is the Lindeberg–Feller result.

It is useful to generalize further by considering a triangular array $\{\xi_{ni} : 1 \leq i \leq k_n, n \geq 1\}$, independent in $i$ for each $n$ rather than just a single sequence (but including that case – with $k_n = n$, $\xi_{ni} = \xi_i$) and consider the limiting distribution of $\sum_{i=1}^{k_n} \xi_{ni}$. This is an extensively studied area, “Central Limit Theory”, where the types of possible limit for such sums are investigated. For the case of pure sums ($\xi_{ni} = \xi_i$) the limits are so-called “stable” r.v.’s (if $\xi, \eta$ are i.i.d. with a stable distribution $\mathcal{G}$, then the linear combination $\alpha \xi + \beta \eta$, $\alpha > 0, \beta > 0$, has the distribution $\mathcal{G}(ax + b)$, some $a > 0, b$).

For array sums the possible limits are (under natural conditions) the more general “infinitely divisible laws” corresponding to r.v.’s which may be split up as the sum of $n$ i.i.d. components for any $n$. Here we look at just the special case of the normal limit for array sums under the so-called Lindeberg conditions using a proof due to W.L. Smith. The following lemma will be useful in proving the main theorem. When unstated the range of $j$ in a sum, or product is from $j = 1$ to $k_n$.

**Lemma 12.6.1** Let $k_n \to \infty$ and let $\{a_{nj} : 1 \leq j \leq k_n, \ n = 1, 2, \ldots\}$ be complex numbers such that
(i) \( \max_j |a_{nj}| \to 0 \) and
(ii) \( \sum_j |a_{nj}| \leq K \) all \( n \), some \( K > 0 \).

Then \( \prod_j (1 - a_{nj}) \exp(\sum_j a_{nj}) \to 1 \) as \( n \to \infty \).

**Proof**  This is perhaps most simply shown by use of the expansion

\[
\log(1 - z) = -z + \psi(z), \quad |\psi(z)| \leq A|z|^2
\]

for complex \( z \), \( |z| < 1 \), valid for the “principal branch” of the logarithm. It may alternatively be shown from the version of this for real \( z \), avoiding the multivalued logarithm but requiring more detailed calculation.

Using the above expansion we have, for sufficiently large \( n \),

\[
|\log\left\{ \prod_j (1 - a_{nj}) \exp(\sum_j a_{nj}) \right\}| = \left| \sum_j (\log(1 - a_{nj}) + a_{nj}) \right|
\]

\[
\leq A \sum_j |a_{nj}|^2
\]

\[
\leq A(\max_j |a_{nj}|) \sum_j |a_{nj}|
\]

which tends to zero by the assumptions and hence the result \( \prod_j (1 - a_{nj}) \times \exp(\sum_j a_{nj}) \to 1 \) as required.

**Theorem 12.6.2** (Array Form of Lindeberg–Feller Central Limit Theorem)

Let \( \{\xi_{nj}, 1 \leq j \leq k_n, \ n = 1, 2, \ldots\} \) be a triangular array of r.v.’s, independent in \( j \) for each \( n \), d.f. \( F_{nj} \), mean zero and finite variance \( \sigma^2_{nj} \) such that \( s_n^2 = \sum_j \sigma^2_{nj} \to 1 \) as \( n \to \infty \). Let \( \xi \) be a standard normal \((N(0, 1))\) r.v. Then

\( \sum_j \xi_{nj} \overset{d}{\to} \xi \) and \( \max_j \sigma^2_{nj} \to 0 \) if and only if the Lindeberg condition (L) holds, viz.,

\[
\sum_j \int_{|x|>\epsilon} x^2 \, dF_{nj}(x) \left( = \sum_j \xi_{nj}^2 \chi_{|\xi_{nj}|>\epsilon} \right) \to 0 \text{ as } n \to \infty, \text{ each } \epsilon > 0. \quad \text{(L)}
\]

**Proof**  Note first that (L) implies that \( \max_j \sigma^2_{nj} \to 0 \) since clearly \( \max_j \sigma^2_{nj} \leq \epsilon^2 + \sum_j \mathbb{E}(\xi_{nj}^2 \chi_{|\xi_{nj}|>\epsilon}) \). Hence \( \max_j \sigma^2_{nj} \to 0 \) may be assumed as a basic condition in the proof in both directions.

Now let \( \phi_{nj} \) be the c.f. of \( \xi_{nj} \) and \( \psi_{nj} \) the corresponding c.f. determined as in Lemma 12.2.3, i.e.

\[
\phi_{nj}(t) = 1 - \frac{1}{2} \sigma^2_{nj} t^2 \psi_{nj}(t).
\]
Then the c.f. of $\zeta_n = \sum_j \xi_{nj}$ is

$$\Phi_n(t) = \prod_j \phi_{nj}(t) = \prod_j (1 - \frac{1}{2} \sigma_{nj}^2 \psi_{nj}(t)).$$

It is easily checked that the conditions of Lemma 12.6.1 are satisfied with

$$a_{nj} = \sigma_{nj}^2 \psi_{nj}(t)/2$$

so that

$$\Phi_n(t) \exp\left(\frac{t^2}{2} s_n^2 \Psi_n(t)\right) \to 1$$

where $\Psi_n(t) = s_n^{-2} \sum_j \sigma_{nj}^2 \psi_{nj}(t)$. Since $s_n \to 1$, if $\Psi_n(t) \to 1$ it follows that $\Phi_n(t) \to e^{-r^2/2}$. Conversely if $\Phi_n(t) \to e^{-r^2/2}$ clearly $\exp(\frac{t^2}{2} s_n^2 (\Psi_n(t) - 1)) \to 1$ (since $s_n \to 1$, so that $\Psi_n(t) \to 1$). Hence $\Phi_n(t) \to e^{-r^2/2}$ if and only if $\Psi_n(t) \to 1$. But $\Psi_n(t)$ is a convex combination of the c.f.’s $\psi_{nj}$ ($\sum \sigma_{nj}^2 = s_n^2$) and hence is clearly itself a c.f. for each $n$ (see also next section). Thus $\zeta_n = \sum_j \xi_{nj}$ (with c.f. $\Phi_n$) converges in distribution to a standard normal r.v. if and only if $\Psi_n(t) \to 1$ for each $t$ or equivalently if and only if the d.f. $G_n$ corresponding to $\Psi_n$ converges weakly to $U(x) = 0$ for $x < 0$ and $1$ for $x \geq 0$.

Now it follows from Lemma 12.2.3 that $\Psi_n$ corresponds to the p.d.f. $g_n$ (d.f. $G_n$) where for $x > 0$

$$g_n(x) = \frac{2}{s_n^2} \sum_j \int_x^\infty (1 - F_{nj}(u)) \, du.$$

Using the same inversions of integration as in Lemma 12.2.3 (or integration by parts) it follows readily that for any $\epsilon > 0$

$$\int_\epsilon^\infty g_n(x) \, dx = \frac{1}{s_n^2} \sum_j \int_\epsilon^\infty (u - \epsilon)^2 \, dF_{nj}(u).$$

This and the corresponding result for $x < 0$ (and noting $s_n \to 1$) show that $G_n \wedge U$ if and only if for each $\epsilon > 0$

$$\sum_j \int_{|x| > \epsilon} (|x| - \epsilon)^2 \, dF_{nj}(x) \to 0 \text{ as } n \to \infty. \quad (L')$$

Now $(L')$ has the same form as $(L)$ with integrand $(|x| - \epsilon)^2$ instead of $x^2$ in the same range $(|x| > \epsilon)$. But in this range $0 < |x| - \epsilon < |x|$ so that $(|x| - \epsilon)^2 \leq x^2$ and hence $(L)$ implies $(L')$. Conversely if $(L')$ holds for each $\epsilon > 0$ it holds with $\epsilon/2$ instead of $\epsilon$ and hence (reducing the integration range)

$$\sum_j \int_{|x| > \epsilon} (|x| - \epsilon/2)^2 \, dF_{nj}(x) \to 0.$$
But in the range $|x| > \epsilon$, $1 - \epsilon/(2|x|) > 1/2$ so that
\[
(|x| - \epsilon/2)^2 = x^2 \left(1 - \frac{\epsilon}{2|x|}\right)^2 > x^2/4
\]
so that (L) holds. Thus (L) and (L') are equivalent, completing the proof. □

**Corollary 1** ("Standard" Form of Lindeberg–Feller Theorem) Let $\{\xi_n\}$ be independent r.v.'s with d.f.'s $\{F_n\}$, zero means, and finite variances $\{\sigma_n^2\}$ with $\sigma_1^2 > 0$. Write $s_n^2 = \sum_{j=1}^n \sigma_j^2$. Then $s_n^{-1} \sum_{j=1}^n \xi_j \overset{d}{\rightarrow} \xi$, standard normal, and $\max_{1 \leq j \leq n} \sigma_j^2/s_n^2 \rightarrow 0$ if and only if the Lindeberg condition
\[
s_n^{-2} \sum_{j=1}^n \int_{|x| > \epsilon_n} x^2 dF_j(x) \rightarrow 0 \text{ as } n \rightarrow \infty, \text{ each } \epsilon > 0. \quad (L'')
\]

**Proof** This follows from the theorem by writing $\xi_{nj} = \xi_j/s_n$, $1 \leq j \leq n$, $n = 1, 2, \ldots$. □

The theorem may also be formulated for r.v.'s with nonzero means in the obvious way:

**Corollary 2** If $\{\xi_n\}$ are independent r.v.'s with d.f.'s $\{F_n\}$, means $\{\mu_n\}$, and finite variances $\{\sigma_n^2\}$ with $\sigma_1^2 > 0$, $s_n^2 = \sum_{j=1}^n \sigma_j^2$, $\max_{1 \leq j \leq n} \sigma_j^2/s_n^2 \rightarrow 0$, then a necessary and sufficient condition for $\frac{1}{s_n} \sum_{j=1}^n (\xi_j - \mu_j)$ to converge in distribution to a standard normal r.v. is the Lindeberg condition
\[
\frac{1}{s_n^2} \sum_{j=1}^n \int_{|x-\mu_j| > \epsilon_n} (x-\mu_j)^2 dF_j(x) \rightarrow 0 \text{ as } n \rightarrow \infty \text{ for each } \epsilon > 0. \quad (L'''')
\]
First of all it should be noted that c.f.’s may sometimes be recognized by virtue of being certain combinations of known c.f.’s (see also [Chung]). For example, if \( \phi_j(t), \ j = 1, \ldots, n, \) are c.f.’s we know that \( \prod^n_1 \phi_j(t) \) is a c.f. (Theorem 12.1.4). So is any “convex combination” \( \sum^n_1 \alpha_j \phi_j(t) \) \( (\alpha_j \geq 0, \sum^n_1 \alpha_j = 1) \) which corresponds to the “mixed” d.f. \( \sum^n_1 \alpha_j F_j(x) \) if \( \phi_j \) corresponds to \( F_j \). Indeed, we may have an infinite convex combination – as should be checked. (See also Ex. 12.11.)

Of course, if \( \phi \) is a c.f. so is \( e^{ibt} \phi(at) \) for any real \( a, b \) (Theorem 12.1.2), and \( \phi(-t) \). But \( \phi(t) = \phi(-t) \) and thus \( |\phi(t)|^2 = \phi(t)\phi(-t) \) is a c.f. also.

In all cases mentioned the reader should determine what r.v.’s the indicated c.f.’s correspond to, where possible. For example, if \( \xi, \eta \) are independent with the same d.f. \( F \) (and c.f. \( \phi \)) it should be checked that the c.f. of \( \xi - \eta \) is \( |\phi(t)|^2 \).

Both Bochner’s Theorem and the criterion for recognizing certain c.f.’s will be consequences of the following lemma.

**Lemma 12.7.1** Let \( \phi(t) \) be a continuous complex function on \( \mathbb{R} \) with \( \phi(0) = 1, \) \( |\phi(t)| \leq 1 \) for all \( t \) and such that for all \( T \)

\[
g(\lambda, T) = \frac{1}{2\pi} \int_{-T}^{T} \mu(t/T) \phi(t) e^{-i\lambda t} \, dt
\]

is real and nonnegative for each real \( \lambda \) where \( \mu(t) \) is \( 1 - |t| \) for \( |t| \leq 1 \) and zero for \( |t| > 1 \). Then

(i) for each fixed \( T, g(\lambda, T) \) is a p.d.f. with corresponding c.f. \( \phi(t)\mu(t/T) \).

(ii) \( \phi(t) \) is a c.f.

**Proof** (ii) will follow at once from (i) by Theorem 12.4.3 since \( \phi(t) = \lim_{T \to \infty} \phi(t)\mu(t/T) \) \( (\mu(t/T) \to 1 \) as \( T \to \infty) \) and \( \phi \) is continuous at \( t = 0 \).

To prove (i) we first show that \( g(\lambda, T) \) is integrable, i.e. \( \int_{-\infty}^{\infty} g(\lambda, T) \, d\lambda < \infty \)

since \( g \) is assumed nonnegative. Let \( M > 0 \). Then \( \left( \int_{-\infty}^{\infty} \right)\)

\[
\int g(\lambda, T) \mu(t) \frac{\lambda}{2M} \, d\lambda = \frac{1}{2\pi} \int \mu\left(\frac{\lambda}{2M}\right) \left( \int \mu\left(\frac{t}{T}\right) \phi(t) e^{-i\lambda t} \, dt \right) \, d\lambda.
\]

By the definition of \( \mu(t) \), both ranges of integration are really finite and since the integrand is bounded \( (|\phi(t)| \leq 1) \) the integration order may be
12.7 Recognizing a c.f. – Bochner’s Theorem

\[ \int g(\lambda) \mu(\frac{\lambda}{2M}) d\lambda = \frac{1}{2\pi} \int \mu\left(\frac{t}{T}\right) \left( \int \mu\left(\frac{\lambda}{2M}\right) e^{-i\lambda t} d\lambda \right) dt = \frac{1}{\lambda} \int \mu\left(\frac{t}{T}\right) \left( \int_{-2M}^{2M} (1 - \frac{\lambda}{2M}) e^{-i\lambda t} d\lambda \right) dt = \frac{1}{\pi} \int \mu\left(\frac{t}{T}\right) \left( \int_{0}^{2M} (1 - \frac{\lambda}{2M}) \cos \lambda t d\lambda \right) dt \]

since \( \cos \lambda t \) is even, and \( \sin \lambda t \) is odd. Integration by parts then gives

\[ \int g(\lambda) \mu(\frac{\lambda}{2M}) d\lambda = \frac{M}{\pi} \int \mu\left(\frac{t}{M}\right) \left( \frac{\sin Mt}{M} \right)^2 \frac{dt}{t} \leq \frac{M}{\pi} \left( \frac{\sin Mt}{M} \right)^2 \int \left( \frac{\sin Mt}{M} \right)^2 dt \]

\[ = \frac{1}{\pi} \int \left( \frac{\sin t}{t} \right)^2 dt = 1, \]

as is well known. Now, letting \( M \to \infty \), monotone convergence (\( \mu(\frac{\lambda}{2M}) \uparrow 1 \)) gives \( \int g(\lambda) d\lambda \leq 1 \).

Thus \( g(\lambda, T) \in L_1(-\infty, \infty) \). To see that its integral is in fact equal to one, note that as defined \( g(\lambda, T) \) is a Fourier Transform \( \int \left( \frac{1}{2\pi} \mu\left(\frac{t}{T}\right) \phi(t) \right) e^{-i\lambda t} dt \) of the \( L_1 \)-function \( \frac{1}{2\pi} \mu\left(\frac{t}{T}\right) \phi(t) \) (zero for \( |t| > T \)). Since \( g(\lambda, T) \) is itself in \( L_1 \), inversion (from Theorem 8.3.4 with obvious sign changes) gives

\[ \frac{1}{2\pi} \mu\left(\frac{t}{T}\right) \phi(t) = \frac{1}{2\pi} \int e^{i\lambda t} g(\lambda, T) d\lambda. \]

This holds a.e. and hence for all \( t \), since both sides are continuous. In particular \( t = 0 \) gives

\[ \int g(\lambda, T) d\lambda = \phi(0) = 1 \]

so that \( g(\lambda, T) \) is a p.d.f. with the corresponding c.f. \( \int e^{i\lambda t} g(\lambda, T) d\lambda = \mu(\frac{r}{T}) \phi(t) \), which completes the proof of (i), and thus of the lemma also. \( \square \)

**Corollary** The function \( \psi(t) = 1 - |t|/T \) for \( |t| \leq T \), and zero for \( |t| > T \) is a c.f.

**Proof** Take \( \phi(t) \equiv 1 \) in the lemma and note (cf. proof) that

\[ \frac{1}{2\pi} \int_{-T}^{T} \left( 1 - \frac{|t|}{T} \right) e^{-i\lambda t} dt = \frac{T}{2\pi} \left( \frac{\sin T\lambda/2}{T\lambda/2} \right)^2 \geq 0. \]

We shall now obtain Bochner’s Theorem as a consequence of this lemma. For this it will first be necessary to define and state some simple properties of positive definite functions.
A complex function \( f(t) \) \((t \in \mathbb{R})\) will be called positive definite (or non-negative definite) if for any integer \( n = 1, 2, 3, \ldots \), and real \( t_1, \ldots, t_n \) and complex \( z_1, \ldots, z_n \) we have

\[
\sum_{j, k=1}^{n} f(t_j - t_k)z_j \overline{z_k} \geq 0 \tag{12.1}
\]

(“\( \geq 0 \)” is here used as a shorthand for the statement “is real and \( \geq 0 \”)”). Notice that by a well known result in positive definite quadratic forms, (12.1) implies that the determinant of the matrix \( \{f(t_j - t_k)\}_{j, k=1}^{n} \) is nonnegative. The needed simple properties of a positive definite function are given in the following theorem.

**Theorem 12.7.2**  If \( f(t) \) is a positive definite function, then

(i) \( f(0) \geq 0 \),
(ii) \( f(-t) = \overline{f(t)} \) for all \( t \),
(iii) \(|f(t)| \leq f(0) \) for all \( t \),
(iv) \(|f(t + h) - f(t)|^2 \leq 4f(0)|f(0) - f(h)| \) for all \( t, h \),
(v) \( f(t) \) is continuous for all \( t \) (indeed uniformly continuous) if it is continuous at \( t = 0 \).

**Proof**

(i) That \( f(0) \) is real and nonnegative follows by taking \( n = 1, t_1 = 0, z_1 = 1 \) in (12.1).

(ii) If \( n = 2, t_1 = 0, t_2 = t, z_1 = z_2 = 1 \) we obtain \( 2f(0) + f(t) + f(-t) \geq 0 \) from (12.1), and hence \( f(t) + f(-t) \) is real (= \( \alpha \), say).

If \( n = 2, t_1 = 0, t_2 = t, z_1 = 1, z_2 = i \) we see that \( if(t) - if(-t) \) is real and hence \( f(t) - f(-t) \) is purely imaginary (= \( i\beta \), say).

Thus \( f(t) = \frac{1}{2}(\alpha + i\beta) \) and \( f(-t) = \frac{1}{2}(\alpha - i\beta) \), giving \( f(-t) = \overline{f(t)} \).

(iii) If \( t_1 - t_2 = t \), nonnegativity of the determinant of the matrix \( \{f(t_j - t_k)\}_{j, k=1}^{2} \) gives \( f^2(0) \geq f(t)f(-t) = |f(t)|^2 \) so that \( |f(t)| \leq f(0) \).

(iv) If \( n = 3, t_1 = 0, t_2 = t, t_3 = t + h, \) then

\[
\det\{f(t_j - t_k)\}_{j, k=1}^{3} = \begin{vmatrix}
 f(0) & f(t) & f(-t) \\
 f(t) & f(0) & f(-h) \\
 f(t + h) & f(h) & f(0)
\end{vmatrix} \geq 0
\]

gives

\[
f^3(0) - f(0)|f(t)|^2 - f(0)|f(t + h)|^2 - f(0)|f(h)|^2 + 2\Re[f(t)f(h)\overline{f(t + h)}] \geq 0
\]
and thus, with obvious use of (iii),
\[
|f(0)f(t + h) - f(t)|^2 = f(0)|f(t + h)|^2 + f(0)|f(t)|^2 - 2f(0)|R[f(t)f(t + h)]|
\]
\[
\leq f^2(0) - f(0)|f(h)|^2 + 2R[f(t)f(t + h)]|f(h) - f(0)|
\]
\[
\leq 2f^2(0)|f(0) - f(h)| + 2f^2(0)|f(0) - f(h)|
\]
\[
\leq 4f^2(0)|f(0) - f(h)|
\]
from which the desired inequality follows (even if \(f(0) = 0\), by (iii)).

(v) is clear from (iv).

\[\Box\]

**Theorem 12.7.3** (Bochner’s Theorem) A complex function \(\phi(t) (t \in \mathbb{R})\) is a c.f. if and only if it is continuous, positive definite, and \(\phi(0) = 1\). By Theorem 12.7.2 (v) continuity for all \(t\) may be replaced by continuity at \(t = 0\).

**Proof** If \(\phi\) is a c.f., it is continuous and \(\phi(0) = 1\). If \(t_1, \ldots, t_n\) are real and \(z_1, \ldots, z_n\) complex (writing \(\phi(t) = \int e^{itx} \, dF(x)\)) then
\[
\sum_{j,k=1}^n \phi(t_j - t_k)z_j \overline{z}_k = \int \left( \sum_{j,k=1}^n e^{i(t_j - t_k)\lambda} z_j \overline{z}_k \right) \, dF(x)
\]
\[
= \int |\sum_{j=1}^n z_j e^{i\lambda x}|^2 \, dF(x) \geq 0
\]
and hence \(\phi\) is positive definite.

Conversely suppose that \(\phi\) is continuous and positive definite with \(\phi(0) = 1\). As in Lemma 12.7.1, define \(g(\lambda, T) = \frac{1}{2\pi} \int_{-T}^{T} (1 - |t|)\phi(t)e^{-i\lambda t} \, dt\). It is easy to see that \(g\) may be written as
\[
g(\lambda, T) = \frac{1}{2\pi T} \int_{-T}^{T} \int_{-T}^{T} \phi(t - u)e^{-i\lambda(t - u)} \, dt \, du
\]
(by splitting the square of integration into two parts above and below the diagonal \(t = u\) and putting \(t - u = s\); see figure below). But this latter integral involves a continuous integrand and may be evaluated as the limit of Riemann sums of the form (using the same dissection \(\{t_j\}\) on each axis)
\[
\frac{1}{2\pi T} \sum_{j,k=1}^n \phi(t_j - t_k)z_j \overline{z}_k
\]
with \(z_j = e^{-i\lambda j}(t_j - t_{j-1})\). Since \(\phi\) is positive definite such sums are non-negative and hence so is \(g(\lambda, T)\).

Since \(|\phi(t)| \leq \phi(0)| by Theorem 12.7.2 (iii) and \(\phi(0) = 1\) the conditions for Lemma 12.7.1 are satisfied and \(\phi\) is thus a c.f. \[\Box\]
We turn now to the “practical criterion” referred to above. As will be seen, this criterion provides sufficient conditions for a function to be a c.f. and, while these are useful, they are far indeed from being necessary. Basically the result gives conditions under which a real function \( \phi(t) \) which is convex on \((0, \infty)\) will be a c.f.

**Theorem 12.7.4** Let \( \phi(t) \) be a real, nonnegative, even, continuous function on \( \mathbb{R} \) such that \( \phi(t) \) is nonincreasing and convex on \( t \geq 0 \), and such that \( \phi(0) = 1 \). Then \( \phi \) is a c.f.

**Proof** Consider first a convex polygon \( \phi(t) \) of the type shown in the figure below with vertices at \( 0 < a_1 < a_2 < \ldots < a_n \) (and constant for \( t > a_n \)).

It is easy to see that \( \phi(t) \) may be written as

\[
\phi(t) = \sum_{k=1}^{n} \lambda_k \mu(t/a_k) + \lambda_{n+1}
\]

where \( \mu(t) = 1 - |t| \) for \( |t| \leq 1 \) and \( \mu(t) = 0 \) otherwise. (This expression is clearly linear between \( a_k \) and \( a_{k+1} \), and at \( a_j \) takes the value \( \phi(a_j) = \sum_{k=j+1}^{n} \lambda_k \mu(a_j/a_k) + \lambda_{n+1} \) so that \( \lambda_{n+1}, \lambda_n, \ldots, \lambda_1 \), may be successively calculated from \( \phi(a_n), \phi(a_{n-1}), \ldots, \phi(a_1) \), \( \phi(0) = 1 \).)

The polygon edge between \( a_j \) and \( a_{j+1} \) has the form \( \sum_{k=j+1}^{n} \lambda_k \mu(t/a_k) + \lambda_{n+1} \) and hence (if continued back) intercepts \( t = 0 \) at height \( \sum_{j=1}^{n} \lambda_k \). By convexity these intercepts decrease as \( j \) increases and hence \( \lambda_j = \sum_{j=1}^{n+1} \lambda_k - \sum_{j=1}^{n} \lambda_k > 0 \). Since \( \phi(0) = 1 \) we also have \( \sum_{j=1}^{n+1} \lambda_j = 1 \).
Now \( \mu(t/a_k) \) is a c.f. (Lemma 12.7.1, Corollary) for each \( k \), and so also is the constant function 1. \( \phi(t) \) is thus seen to be a convex combination of c.f.’s and is thus itself a c.f.

If now \( \phi(t) \) is a function satisfying the conditions of the theorem, it may clearly be expressed as a limit of such convex polygons (e.g. inscribed with vertices at \( r/2^n, r = 0, 1, \ldots, 2^n \)). Hence by Theorem 12.4.3, \( \phi \) is a c.f. □

Applications of this theorem are given in the exercises.

### 12.8 Joint characteristic functions

It is also useful to consider the joint c.f. of \( m \) r.v.’s \( \xi_1, \ldots, \xi_m \) defined for real \( t_1, \ldots, t_m \) by

\[
\phi(t_1, \ldots, t_m) = \mathcal{E}e^{it_1\xi_1 + \cdots + t_m\xi_m}.
\]

We shall not investigate such functions in any great detail here, but will indicate a few of their more important properties. First it is easily shown that if \( F \) is the joint d.f. of \( \xi_1, \ldots, \xi_m \), then

\[
\phi(t_1, \ldots, t_m) = \int_{\mathbb{R}^m} e^{it_1x_1 + \cdots + t_mx_m} dF(x_1, \ldots, x_m)
\]

(whence “\( dF \), of course, means \( d\mu_F = dP(\xi_1, \ldots, \xi_m)^{-1} \) in the notation of Section 9.3). Further, the simplest properties of c.f.’s of a single r.v. clearly generalize easily. For example, it is easily seen that \( \phi(0, \ldots, 0) = 1 \), \( |\phi(t_1, \ldots, t_m)| \leq 1 \), and so on.

The following obvious but useful property should also be pointed out: The joint c.f. of \( \xi_1, \ldots, \xi_m \) is uniquely determined by the c.f.’s of all linear combinations \( a_1\xi_1 + \cdots + a_m\xi_m, \ a_1, \ldots, a_m \in \mathbb{R} \). Indeed if \( \phi_{a_1, \ldots, a_m}(t) \) denotes the c.f. of \( a_1\xi_1 + \cdots + a_m\xi_m \), i.e. \( \mathcal{E}\exp\{it(a_1\xi_1 + \cdots + a_m\xi_m)\} \), it is clear that \( \phi(t_1, \ldots, t_m) = \phi_{a_1, \ldots, a_m}(1) \).
Generalizations of the inversion, uniqueness and continuity theorems are, of course, of interest. First a useful form of the inversion theorem may be stated as follows (cf. Theorem 12.3.1).

**Theorem 12.8.1** Let $F$ and $\phi$ be the joint d.f. and c.f. of the r.v.’s $\xi_1, \ldots, \xi_m$. Then if $I = (a,b]$, $a = (a_1, \ldots, a_m)$, $b = (b_1, \ldots, b_m)$ ($a_i \leq b_i$, $1 \leq i \leq m$) is any continuity rectangle (Section 10.2) for $F$,

$$
\mu_F(I) = \lim_{T \to \infty} \frac{1}{(2\pi)^m} \int_{-T}^T \cdots \int_{-T}^T \prod_{j=1}^m \left( \frac{e^{-ib_j t_j} - e^{-ia_j t_j}}{-it_j} \right) \phi(t_1, \ldots, t_m) \, dt_1 \ldots dt_m
$$

$\mu_F(I)$ is defined as in Lemma 7.8.2.

This result is obtained in a similar manner to Theorem 12.3.1 (from the $m$-dimensional form of Theorem 8.3.1), and we do not give a detailed proof.

To obtain the uniqueness theorem, an $m$-dimensional form is needed of the fact that a d.f. $F$ has at most countably many discontinuities (Lem- ma 9.2.2) (or equivalently that the corresponding measure $\mu_F$ has at most countably many points of positive mass, i.e. $x$ such that $\mu_F(\{x\}) > 0$). Consider the case $m = 2$, and for a given $s$ let $L_s$ denote the line $x = s$, $-\infty < y < \infty$. If $\mu$ is a probability measure on the Borel sets of $\mathbb{R}^2$ then by the same argument as for $m = 1$, there are at most countably many values of $s$ for which $\mu(L_s) > 0$. Similarly there are at most countably many values of $t$ such that $\mu(L_t') > 0$ if $L'_t$ denotes the line $y = t$, $-\infty < x < \infty$. It thus follows that given any values $s_0, t_0$, there are values $s, t$ arbitrarily close to $s_0, t_0$ respectively, such that $\mu(L_s) = \mu(L_t') = 0$. (Such $L_s, L'$ will be called lines of zero $\mu$-mass.) Precisely the same considerations hold in $\mathbb{R}^m$ for $m > 2$, with $(m-1)$-dimensional hyperplanes of the form $\{(x_1, \ldots, x_m) : x_i = \text{constant}\}$ taking the place of lines. With these observations we now obtain the uniqueness theorem for $m$-dimensional c.f.’s.

**Theorem 12.8.2** The joint c.f. of $m$ r.v.’s uniquely determines their joint d.f., and hence their distribution, and conversely; i.e. two d.f.’s $F_1, F_2$ in $\mathbb{R}^m$ are identical if and only if their c.f.’s $\phi_1, \phi_2$ are identical.

**Proof** It is clear that $F_1 \equiv F_2$ implies $\phi_1 \equiv \phi_2$. For the converse assume $\phi_1 \equiv \phi_2$ and consider the case $m = 2$. (The case $m > 2$ follows with the obvious changes.) With the above notation let $(a,b)$ be a point in $\mathbb{R}^2$ such that $L_a^b$ have zero $\mu_{\phi_1}$- and $\mu_{\phi_2}$-mass. Choose $a_k, b_k$, both tending to $-\infty$ as $k \to \infty$, and such that $L_{a_k}^{b_k}$ have zero $\mu_{\phi_1}$- and $\mu_{\phi_2}$-mass (which is possible since only countably many lines have positive ($\mu_{\phi_1} + \mu_{\phi_2}$)-mass).
Then writing $I_k = (a_k, a] \times (b_k, b]$,

$$F_1(a, b) = \lim_{k \to \infty} [F_1(a, b) - F_1(a_k, b) - F_1(a, b_k) + F_1(a, b_k)]$$

$$= \lim_{k \to \infty} \mu_{F_1}(I_k)$$

$$= \lim_{k \to \infty} \mu_{F_2}(I_k)$$

by Theorem 12.8.1, since $I_k$ is a continuity rectangle for both $\mu_{F_1}$ and $\mu_{F_2}$, and $F_1, F_2$ have the same c.f. But by the same argument (with $F_2$ for $F_1$),

$$\lim_{k \to \infty} \mu_{F_2}(I_k) = F_2(a, b).$$

Hence $F_1(a, b) = F_2(a, b)$ for any $(a, b)$ such that $L_a$ and $L_b$ have zero $\mu_{F_1}$- and $\mu_{F_2}$-mass.

Finally for any $a, b, c_k \downarrow a, d_k \downarrow b$ may be chosen such that $L_{c_k}$ and $L_{d_k}$ have zero $\mu_{F_1}$- and $\mu_{F_2}$-mass and hence $F_1(c_k, d_k) = F_2(c_k, d_k)$ by the above. By right-continuity of $F_1$ and $F_2$ in each argument $F_1(a, b) = F_2(a, b)$, as required. □

The following characterization of independence of $n$ r.v.’s $\xi_1, \ldots, \xi_m$ may now be obtained as an application. (Compare this theorem with Theorem 12.1.4.)

**Theorem 12.8.3** The r.v.’s $\xi_1, \ldots, \xi_m$ are independent if and only if their joint c.f. $\phi(t_1, \ldots, t_m) = \prod_{i=1}^m \phi_i(t_i)$ where $\phi_i$ is the c.f. of $\xi_i$.

**Proof** If the r.v.’s are independent

$$\phi(t_1, \ldots, t_m) = \mathcal{E} e^{i(t_1 \xi_1 + \cdots + t_m \xi_m)} = \prod_{j=1}^m \phi_j(t_j)$$

by (the complex r.v. form of) Theorem 10.3.5. Conversely if $\xi_1, \ldots, \xi_m$ have joint d.f. $F$ and individual d.f.’s $F_j$, and $\phi(t_1, \ldots, t_m) = \prod_{j=1}^m \phi_j(t_j)$ for all $t_1, \ldots, t_m$, then $F(x_1, \ldots, x_m)$ and $F_1(x_1) \ldots F_m(x_m)$ are both d.f.’s on $\mathbb{R}^m$ with the same c.f. (clearly $\int e^{i(t_1 x_1 + \cdots + t_m x_m)} d[F_1(x_1) \ldots F_m(x_m)] = \prod_{j=1}^m \phi_j(t_j)$). Hence by the uniqueness theorem, $F(x_1, \ldots, x_m) = F(x_1) \ldots F(x_m)$, so that the r.v.’s are independent by Theorem 10.3.1. □

Finally, weak convergence of d.f.’s in $\mathbb{R}^m$ (Section 11.2) may be considered by means of their c.f.’s, giving rise to the following general version of the continuity theorem (Theorem 12.4.3).

**Theorem 12.8.4** Let $\{F_n(x_1, \ldots, x_m)\}$ be a sequence of $m$-dimensional d.f.’s with c.f.’s $\{\phi_n(t_1, \ldots, t_m)\}$.
(i) If $F(x_1, \ldots, x_m)$ is a d.f. with c.f. $\phi(t_1, \ldots, t_m)$ and if $F_n \xrightarrow{w} F$, then $\phi_n(t_1, \ldots, t_m) \to \phi(t_1, \ldots, t_m)$ as $n \to \infty$, for all $t_1, \ldots, t_m \in \mathbb{R}$.

(ii) If $\phi(t_1, \ldots, t_m)$ is a complex function which is continuous at $(0, \ldots, 0)$ and if $\phi_n(t_1, \ldots, t_m) \to \phi(t_1, \ldots, t_m)$ as $n \to \infty$, for all $t_1, \ldots, t_m \in \mathbb{R}$, then $\phi$ is the c.f. of a (m-dimensional) d.f. $F$ and $F_n \xrightarrow{w} F$.

As a corollary to this result we may obtain an elegant simple device due to H. Cramér and H. Wold, which enables convergence in distribution of random vectors to be reduced to convergence of ordinary r.v.’s.

**Theorem 12.8.5** (Cramér–Wold Device) Let $\xi = (\xi_1, \ldots, \xi_m)$, $\xi_n = (\xi_{n1}, \ldots, \xi_{nm})$, $n = 1, 2, \ldots$, be random vectors. Then

$$\xi_n \xrightarrow{d} \xi \text{ as } n \to \infty$$

if and only if

$$a_1\xi_{n1} + \cdots + a_m\xi_{nm} \xrightarrow{d} a_1\xi_1 + \cdots + a_m\xi_m \text{ as } n \to \infty$$

for all $a_1, \ldots, a_m \in \mathbb{R}$.

**Proof** By the continuity theorems 12.4.3 and 12.8.4, $\xi_n \xrightarrow{d} \xi$ is equivalent to

$$Ee^{it_1\xi_{n1} + \cdots + t_m\xi_{nm}} \xrightarrow{d} Ee^{it_1\xi_1 + \cdots + t_m\xi_m}$$

for all $t_1, \ldots, t_m \in \mathbb{R}$, and $a_1\xi_{n1} + \cdots + a_m\xi_{nm} \xrightarrow{d} a_1\xi_1 + \cdots + a_m\xi_m$ is equivalent to

$$Ee^{it(a_1\xi_{n1} + \cdots + a_m\xi_{nm})} \xrightarrow{d} Ee^{it(a_1\xi_1 + \cdots + a_m\xi_m)}$$

for all $t \in \mathbb{R}$. It is then clear that the former implies the latter (by taking $t_j = ta_j$) and conversely (by taking $t = 1$). \qed

This result shows that to prove convergence in distribution of a sequence of random vectors it is sufficient to consider convergence of arbitrary (but fixed) finite linear combinations of the components. This is especially useful for jointly normal r.v.’s since then each linear combination is also normal.

**Exercises**

12.1 Find the c.f.’s for the following r.v.’s

(a) Geometric: $P(\xi = n) = pq^{n-1}$, $n = 1, 2, 3 \ldots$ ($0 < p < 1, q = 1 - p$)

(b) Poisson: $P(\xi = n) = e^{-\lambda} \lambda^n / n!$, $n = 0, 1, 2 \ldots$ ($\lambda > 0$)
(c) Exponential: p.d.f. \( \lambda e^{-\lambda x}, \ x \geq 0 \ (\lambda > 0) \)
(d) Cauchy: p.d.f. \( \frac{1}{\pi(ax^2+b^2)}, \ -\infty < x < \infty \ (\lambda > 0). \)

12.2 Let \( \xi, \eta \) be independent r.v.’s each being uniformly distributed on \((-1, 1). Evaluate the distribution of \( \xi + \eta \) and hence its c.f. Check this with the square of (the absolute value of) the c.f. of \( \xi. \)

12.3 Let \( \xi \) be a standard normal r.v. Find the p.d.f. and c.f. of \( \xi^2. \)

12.4 If \( \xi_1, \ldots, \xi_n \) are independent standard normal r.v.’s, find the c.f. of \( \sum_{i=1}^{n} \xi_i^2. \) Check that this corresponds to the p.d.f. \( 2^{(n/2)\Gamma((n/2))^{-1}} x^{(n/2)-1} e^{-x^2} (x > 0) \) (\( \chi^2 \) with \( n \) degrees of freedom).

12.5 Find two r.v.’s \( \xi, \eta \) which are not independent but have the same p.d.f., and are such that the p.d.f. of \( \xi + \eta \) is the convolution \( f \ast f. \) (Hint: Try \( \xi = \eta \) with an appropriate d.f.)

12.6 According to Section 7.6 if \( f, g \) are in \( L_1(–\infty, \infty) \) then the convolution \( h = f \ast g \in L_1 \) and has \( L_1 \) Fourier Transform \( \hat{h} = \hat{f} \hat{g} \). In the case where \( f \) and \( g \) are nonnegative (e.g. p.d.f.’s) give an alternative proof of this result based on Theorem 10.4.1 and Section 12.1. Give a corresponding result for Fourier–Stieltjes Transforms of the Stieltjes Convolution \( (F_1 \ast F_2)(x) = \int F_1(x–y) \ dF_2(y) \) of two d.f.’s \( F_1, F_2. \)

12.7 If \( \xi \) is a r.v. with c.f. \( \phi \) show that \( E|\xi| = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{R[1–\phi(t)]}{t^2} \ dt. \)

(Hint: \( \int_{-\infty}^{\infty} (\sin \frac{t}{t})^2 \ dt = \pi. \)

12.8 Let \( \phi \) be the c.f. of a r.v. \( \xi. \) Suppose that \( \lim_{t \to 0} (1–\phi(t))/(t^2) = \sigma^2/2 < \infty. \)

Show that \( E\xi = 0 \) and \( E\xi^2 = \sigma^2. \) In particular if \( \phi(t) = 1 + o(t^2) \) show that \( \xi = 0 \) a.s. (Hints: \( R[1–\phi(t)]/(t^2) = \int [(1–\cos tx)/(t^2)] \ dF(x) \to \sigma^2/2. \) Apply Fatou’s Lemma to show \( \int x^2 \ dF(x) < \infty. \) Then use the corollary of Theorem 12.2.1.)

12.9 A r.v. \( \xi \) is called symmetric if \( \xi \) and \( -\xi \) have the same d.f. Show that \( \xi \) is symmetric if and only if its c.f. \( \phi \) is real-valued.

12.10 Show that the real part of a c.f. is a c.f. but that the same is never true of the imaginary part.

12.11 Let \( \xi_1 \) and \( \xi_2 \) be independent r.v.’s with d.f.’s \( F_1 \) and \( F_2 \) and c.f.’s \( \phi_1 \) and \( \phi_2. \)

(i) Show that the c.f. \( \phi \) of \( \xi_1 \xi_2 \) is given by \( \phi(t) = \int_{-\infty}^{\infty} \phi_1(ty) \ dF_2(y) = \int_{-\infty}^{\infty} \phi_2(tx) \ dF_1(x) \) for all \( t \in \mathbb{R}. \)

(ii) If \( F_2(0–) = F_2(0), \) show that the r.v. \( \xi_1/\xi_2 \) is well defined and its c.f. \( \phi \) is given by \( \phi(t) = \int_{-\infty}^{\infty} \phi_1(ty) \ dF_2(y) \) for all \( t \in \mathbb{R}. \)
As a consequence of (i) and (ii), if $\phi$ is a c.f. and $G$ a d.f., then $\int_{-\infty}^{\infty} \phi(ty) \, dG(y)$ is a c.f. and so is $\int_{-\infty}^{\infty} \phi(ty) \, dG(y)$ if $G(0-) = G(0)$.

12.12 If $f(t)$ is a function defined on the real line write $\Delta_h f(t) = f(t + h) - f(t)$, for real $h$, and say that $f$ has a generalized second derivative at $t$ when the following limit exists and is finite

$$\lim_{h,h' \to 0} \frac{\Delta_h \Delta_h f(t)}{h'h}$$

for all sequences $h \to 0$ and $h' \to 0$. Show that if $f$ has two derivatives at $t$ then it has a generalized second derivative at $t$, and that the converse is not true. If $\phi(t)$ is a characteristic function show that the following are equivalent:

(i) $\phi$ has a generalized second derivative at $t = 0$,
(ii) $\phi$ has two finite derivatives at $t = 0$,
(iii) $\phi$ has two derivatives at every real $t$,
(iv) $\int_{-\infty}^{\infty} x^2 \, dF(x) < \infty$, where $F$ is the d.f. of $\phi$.

12.13 If $f(t)$ is a function defined on the real line its first symmetric difference may be defined by

$$\Delta_1^s f(t) = f(t + s) - f(t - s)$$

for real $s$, and its higher order symmetric differences by

$$\Delta_n^{s+1} f(t) = \Delta_1^s \Delta_n^s f(t)$$

for $n = 1, 2, \ldots$. If the limit

$$\lim_{s \to 0} \frac{\Delta_n^s f(t)}{(2s)^n}$$

exists and is finite, we say that $f$ has $n$th symmetric derivative at $t$. Now let $\phi$ be the c.f. of a r.v. $\xi$, and $n$ a positive integer. Show that if

$$\lim_{s \to 0} \inf \left| \frac{\Delta_n^s \phi(0)}{(2s)^{2n}} \right| < \infty$$

then $\mathbb{E} \xi^{2n} < \infty$. (Hint: Show that

$$\Delta_n^s f(t) = \sum_{k=0}^{n} (-1)^k \binom{n}{k} f[t + (n - 2k)s]$$

and

$$\Delta_n^{2n} \phi(t) = \int_{-\infty}^{\infty} e^{i2s} (2i \sin s) \, dF(x).$$

Show also that the following are equivalent

(i) $\phi$ has $(2n)$th symmetric derivative at $t = 0$,
(ii) $\phi$ has $2n$ finite derivatives at $t = 0$,,
(iii) \( \phi \) has \( 2n \) finite derivatives at every real \( t \),

(iv) \( \mathcal{E} \xi^{2n} < \infty \).

12.14 Let \( \xi \) be a r.v. with c.f. \( \phi \) and denote by \( \rho_n \) the \( n \)th symmetric difference of \( \phi \) at 0:

\[
\rho_n(t) = \Delta_n^0 \phi(0)
\]

(see Ex. 12.13). If \( 0 < p < 2n \), show that \( \mathcal{E}|\xi|^p < \infty \) if and only if

\[
\int_0^\epsilon \frac{|\rho_{2n}(t)|}{t^{1+p}} \, dt < \infty
\]

for some \( \epsilon > 0 \), in which case

\[
\mathcal{E}|\xi|^p = \left\{ 2^{2n} \int_0^\infty \frac{(\sin x)^{2n}}{x^{1+p}} \, dx \right\}^{-1} \int_0^\infty \frac{|\rho_{2n}(t)|}{t^{1+p}} \, dt.
\]

(Hint: Show that

\[
\int_0^\epsilon \frac{|\rho_{2n}(t)|}{t^{1+p}} \, dt = 2^{2n} \int_0^\infty |x|^p \left\{ \int_0^{|x|} \frac{(\sin u)^{2n}}{u^{1+p}} \, du \right\} \, dF(x).
\]

12.15 Let \( \phi \) be the c.f. corresponding to the d.f. \( F \). Note that by Theorem 12.3.1 the jump (if any) of \( F \) at \( x \) may be written as

\[
F(x) - F(x - 0) = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^T e^{-ixt} \phi(t) \, dt.
\]

If \( \phi(t_0) = 1 \) for some \( t_0 \neq 0 \) show that the mass of \( F \) is concentrated on the points \( \{2n\pi/n : n = 0, \pm 1, \ldots \} \) and the \( \mu_x \)-measure of the point \( 2n\pi/n \) is

\[
\frac{1}{n} \int_0^{t_0} \phi(t)e^{-2\pi i/n} \, dt. \text{ (Compare Theorem 12.1.3.)}
\]

12.16 Show that \( |\cos t| \) is not a c.f. (e.g. use the result of Ex. 12.15 with \( n = 4 \)). Hence the absolute value of a c.f. is not necessarily a c.f.

12.17 If \( \phi \) is the c.f. corresponding to the d.f. \( F \) (and measure \( \mu_x \)) prove that

\[
\sum_{x \in \mathbb{R}} \left[ \mu_x((x)) \right]^2 = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^T |\phi(t)|^2 \, dt.
\]

(Hint: Mimic proof of the last part of Theorem 8.3.1 or (more simply) apply the second inversion formula of Theorem 12.3.1 (i) for \( a = 0 \) and \( \xi = \xi_1 - \xi_2 \) where \( \xi_1, \xi_2 \) are i.i.d. with c.f. \( \phi \).) What is the implication of this if \( \phi \in L_2(-\infty, \infty) \)?

12.18 If \( \phi \) is the c.f. corresponding to the d.f. \( F \) and \( \phi \in L_2(-\infty, \infty) \), show that \( F \) is absolutely continuous with density a multiple of the Fourier Transform of \( \phi \). (Hint: Use Parseval’s Theorem.) This is an \( L_2 \) analog of Theorem 12.3.1 (ii).

12.19 Show that the conclusion of the continuity theorem for characteristic functions is not necessarily true if \( \phi \) is not continuous at \( t = 0 \) by considering a sequence of random variables \( \{\xi_n\}_{n=1}^\infty \) such that for each \( n \), \( \xi_n \) has the uniform distribution on \([-n, n]\).
12.20 If \( \phi(t) \) is a characteristic function, then so is \( e^{\lambda \phi(t)} \) for each \( \lambda > 0 \). (Hint: Use \( e^{\lambda \phi(t)} = \lim_n \left( 1 + \frac{\lambda}{n} \phi(t) \right)^n \).)

12.21 If the random variable \( \xi_n \) has a binomial distribution with parameters \( (n, p_n) \), \( n = 1, 2, \ldots \), and \( np_n \to \lambda > 0 \) as \( n \to \infty \), prove that \( \xi_n \) converges in distribution to a random variable which has the Poisson distribution with parameter \( \lambda \). Show also that otherwise as \( p_n \to 0, np_n \to \infty \), then \( \xi_n \) (suitably standardized) has a limiting normal distribution.

12.22 If the r.v.’s \( \xi \) and \( \{\xi_n\}_{n=1}^\infty \) are such that for every \( n \), \( \xi_n \) is normal with mean 0 and variance \( \sigma_n^2 \), show that the following are equivalent

(i) \( \xi_n \to \xi \) in probability

(ii) \( \xi_n \to \xi \) in \( L_2 \)

and that in each case \( \xi \) is normal with zero mean.

12.23 Let \( \{\xi_n\}_{n=1}^\infty \) be a sequence of random variables such that for each \( n \), \( \xi_n \) has a Poisson distribution with parameter \( \lambda_n \). If \( \xi_n \overset{d}{\to} \xi \) (after any normalization needed) show that \( \xi \) has either a Poisson or normal distribution.

12.24 Show that \( \frac{\sin(n^{-1/2}t)}{n^{-1/2}} \) is the c.f. of the uniform distribution on \((-1/2, 1/2)\) and prove by using the c.f.’s that for all real \( t \),

\[
\lim_{n \to \infty} \left( \frac{\sin(n^{-1/2}t)}{n^{-1/2}} \right)^n = e^{-t^2/6}.
\]

12.25 Let \( \{\xi_n\}_{n=1}^\infty \) be independent random variables with finite means \( \mu_n \) and variances \( \sigma_n^2 \), and let \( s_n^2 = \sum_{k=1}^{n} \sigma_k^2 \). Prove that the Lindeberg condition is satisfied, and thus the Lindeberg Central Limit Theorem (Corollary 2 of Theorem 12.6.2) is applicable, if the random variables \( \{\xi_n\}_{n=1}^\infty \):

(i) are uniformly bounded, i.e. for some \( 0 < M < \infty \), \( |\xi_n| \leq M \) a.s. for all \( n \), and \( s_n^2 \to \infty \); or

(ii) are identically distributed; or

(iii) satisfy Liapounov’s condition

\[
\frac{1}{s_n^{2+\delta}} \sum_{k=1}^{n} E(|\xi_k - \mu_k|^{2+\delta}) \to 0 \text{ for some } \delta > 0.
\]

12.26 If two c.f.’s \( \phi_1, \phi_2 \) are equal on a neighborhood of zero then whatever derivatives of \( \phi_1 \) exist at zero must be equal to those of \( \phi_2 \) there. Hence existing moments corresponding to each distribution must be the same. Show that, however, it is not necessarily true that \( \phi_1 = \phi_2 \), everywhere, and hence not necessarily true that the d.f.’s are the same. Note that if \( \phi_2 \equiv 1 \) and \( \phi_1 = \phi_2 \) in a neighborhood of zero it is true that \( \phi_1 = \phi_2 \) everywhere.
13

Conditioning

13.1 Motivation

In this chapter \((\Omega, \mathcal{F}, P)\) will, as usual, denote a fixed probability space. If \(A\) and \(B\) are two events and \(P(B) > 0\), the conditional probability \(P(A|B)\) of \(A\) given \(B\) is defined to be

\[
P(A|B) = \frac{P(A \cap B)}{P(B)}
\]

and has a good interpretation; given that event \(B\) occurs, the probability of event \(A\) is proportional to the probability of the part of \(A\) which lies in \(B\). It has also an appealing frequency interpretation – as the proportion of those repetitions of the experiment in which \(B\) occurs, for which \(A\) also occurs.

It is also important to be able to define \(P(A|B)\) in many cases for which \(P(B) = 0\), for example if \(B\) is the event \(\eta = y\) where \(\eta\) is a continuous r.v. and \(y\) is a fixed value. There are various ways of making an appropriate definition depending on the purpose at hand. Here we are interested in integration over \(y\) to provide formulae such as

\[
P(A) = \int P(A|\eta = y)f(y) \, dy
\]

if \(\eta\) has a density \(f\) which will be a particular case of the general definitions to be given. Other situations require different conditioning definitions – e.g. especially if particular fixed values of \(y\) are involved without integration in a condition \(\eta = y\). A particular such case occurs if \(\eta(t)\) is the value of say temperature at time \(t\) and one is interested in defining \(P(A|\eta(t) = 0)\). The definition used for (13.1) will not have the empirical interpretation as the proportion of those time instants \(t\) where \(\eta(t) = 0\) for which \(A\) occurs. In such cases so-called “Palm distributions” can be appropriate.

Here, however, we consider the definitions of conditional probability and expectation for obtaining the probability \(P(A)\) by conditioning on values of a r.v. \(\eta\) and integrating over those values as in (13.1). This will be achieved
in a much more general setting via the Radon–Nikodym Theorem, (13.1) being a quite special case.

To motivate the approach it is illuminating to proceed from the special case where \( \eta \) is a r.v. which can take one of \( n \) possible values \( y_1, y_2, \ldots, y_n \) with \( P(\eta = y_j) = p_j > 0, 1 \leq j \leq n, \sum_{j=1}^n p_j = 1 \). Then for all \( A \in \mathcal{F} \)
\[
P(A|\eta = y_j) = \frac{P(A \cap \eta^{-1}\{y_j\})}{P(\eta^{-1}\{y_j\})}
\]
so that
\[
P(A) = \sum_j P(A|\eta = y_j)p_j
\]
where \( P(A|\eta = y) \) is \( P(A|\eta = y_j) \) at \( y_j \) and (say) zero otherwise.

More generally it is easily shown that for all \( A \in \mathcal{F} \) and \( B \in \mathcal{B} \)
\[
P(A \cap \eta^{-1}B) = \int_B P(A|\eta = y)\,dP\eta^{-1}(y).
\]
This relation holds in the above case where \( P\eta^{-1} \) is confined to the points \( y_1, y_2, \ldots, y_n \) so that the condition “\( \eta = y \)” has positive probability for each such value. However, in other cases where \( P\eta^{-1} \) need not have atoms, the relation may (as will be seen) be used to provide a definition of \( P\{A|\eta = y\} \).

First, however, note that in the case considered (13.2) may be written with \( g(y) = P(A|\eta = y) \) as
\[
P(A \cap \eta^{-1}B) = \int_B g(y)\,dP\eta^{-1}(y) = \int_{\eta^{-1}B} g(\eta(\omega))\,dP(\omega).
\]
Since \( \sigma(\eta) = \sigma(\eta^{-1}(B) : B \in \mathcal{B}) \) it follows that for \( E \in \sigma(\eta) \)
\[
P(A \cap E) = \int_E g(\eta(\omega))\,dP(\omega).
\]
The function \( g(\eta(\omega)) \) depends on the set \( A \in \mathcal{F} \) and writing it explicitly as \( P(\{A|\eta\})(\omega) \) we have
\[
P(A \cap E) = \int_E P(\{A|\eta\})(\omega)\,dP(\omega)
\]
for each \( A \in \mathcal{F}, E \in \sigma(\eta) \). Since \( g \) is trivially Borel measurable, \( P(\{A|\eta\}) \) as defined on \( \Omega \) is a \( \sigma(\eta) \)-measurable function for each fixed \( A \in \mathcal{F} \) and is referred to as the “conditional probability of \( A \) given \( \eta \)”. This is related to but distinguished from the function \( P(A|\eta = y) \) in (13.2), naturally referred to as the “conditional probability of \( A \) given \( \eta = y \)”.

The version \( P(\{A|\eta\})(\omega) \) leads to a yet more general abstraction. The function \( P(\{A|\eta\})(\omega) \) was defined in such a way that it is \( \sigma(\eta) \)-measurable and satisfies (13.3) for each \( E \in \sigma(\eta) \). These requirements involve \( \eta \) only through its generated \( \sigma \)-field \( \sigma(\eta) (\subset \mathcal{F}) \) and it is therefore natural to write alternatively
\[
P(\{A|\eta\})(\omega) = P(\{A|\sigma(\eta)\})(\omega)
\]
13.2 Conditional expectation given a σ-field

Let ξ be a r.v. with E|ξ| < ∞ and G a sub-σ-field of F. The conditional expectation of ξ given G will be defined in a way which extends the definition of conditional probability suggested in the previous section.

Consider the set function ν defined for all E ∈ G by

$$
ν(E) = \int_E ξ \, dP.
$$

Then ν is a finite signed measure on G and ν ⪯ P_G where P_G denotes the restriction of P from F to G. Thus by the Radon–Nikodym Theorem (Theorem 5.5.3) there is a finite-valued G-measurable and P_G-integrable function f on Ω uniquely determined a.s. (P_G) such that for all E ∈ G,

$$
ν(E) = \int_E f \, dP_G = \int_E f \, dP
$$

(for the second equality see Ex. 4.10). We write f = E(ξ|G) and call it the conditional expectation of ξ given the σ-field G. Thus the conditional expectation E(ξ|G) of ξ given G is a G-measurable and P-integrable r.v. which is determined uniquely a.s. by the equality

$$
\int_E ξ \, dP = \int_E E(ξ|G) \, dP
$$

for all E ∈ G.

It is readily seen that this definition extends that suggested in Section 13.1 when G = σ(η) for a r.v. η taking a finite number of values (Ex. 13.1). The equality may also be rephrased in “E-form” as E(χ_E ξ) = E(χ_E E(ξ|G)) for all E ∈ G.

If η is a r.v. the conditional expectation E(ξ|η) of ξ given η is defined by taking G = σ(η), i.e. E(ξ|η) = E(ξ|σ(η)) so that E(ξ|η) is a σ(η)-measurable function f satisfying ∫_E ξ dP = ∫_E f dP for each E ∈ σ(η). It is enough that this equality holds for all E of the form η^{-1}(B) for B ∈ B since the class of such sets is either σ(η) if η is defined for all ω or otherwise generates σ(η).
For a family \( \{ \eta_\lambda : \lambda \in \Lambda \} \) of r.v.’s \( \text{the conditional expectation } E(\xi|\eta_\lambda : \lambda \in \Lambda) \) of \( \xi \) given \( \{ \eta_\lambda : \lambda \in \Lambda \} \) is defined by

\[
E(\xi|\eta_\lambda : \lambda \in \Lambda) = E(\xi|\sigma(\eta_\lambda : \lambda \in \Lambda))
\]

where \( \sigma(\eta_\lambda : \lambda \in \Lambda) \) is the sub-\( \sigma \)-field of \( F \) generated by the union of the \( \sigma \)-fields \( \{ \sigma(\eta_\lambda) : \lambda \in \Lambda \} \) (cf. Section 9.3).

The simplest properties of conditional expectations are stated in the following result.

**Theorem 13.2.1** \( \xi \) and \( \eta \) are r.v.’s with finite expectations and \( a, b \) real numbers.

(i) \( E(E(\xi|\mathcal{G})) = E\xi \).

(ii) \( E(a\xi + b\eta|\mathcal{G}) = aE(\xi|\mathcal{G}) + bE(\eta|\mathcal{G}) \) a.s.

(iii) If \( \xi = \eta \) a.s. then \( E(\xi|\mathcal{G}) = E(\eta|\mathcal{G}) \) a.s.

(iv) If \( \xi \geq 0 \) a.s., then \( E(\xi|\mathcal{G}) \geq 0 \) a.s. Hence if \( \xi \leq \eta \) a.s., then \( E(\xi|\mathcal{G}) \leq E(\eta|\mathcal{G}) \) a.s.

(v) If \( \xi \) is \( \mathcal{G} \)-measurable then \( E(\xi|\mathcal{G}) = \xi \) a.s.

**Proof**

(i) Since \( \Omega \in \mathcal{G} \) we have

\[
E\xi = \int_{\Omega} \xi \, dP = \int_{\Omega} E(\xi|\mathcal{G}) \, dP = E\{E(\xi|\mathcal{G})\}.
\]

(ii) For every \( E \in \mathcal{G} \) we have

\[
\int_E (a\xi + b\eta) \, dP = a\int_E \xi \, dP + b\int_E \eta \, dP = a\int_E E(\xi|\mathcal{G}) \, dP + b\int_E E(\eta|\mathcal{G}) \, dP = \int_E \{aE(\xi|\mathcal{G}) + bE(\eta|\mathcal{G})\} \, dP \]

and since the r.v. within brackets is \( \mathcal{G} \)-measurable the result follows from the definition.

(iii) This is obvious from the definition of conditional expectation.

(iv) If \( \xi \geq 0 \) a.s., \( \nu \) (as defined at the start of this section, \( \nu(E) = \int_E \xi \, dP \)) is a measure (rather than a signed measure) and from the Radon–Nikodym Theorem we have \( E(\xi|\mathcal{G}) \geq 0 \) a.s. The second part follows from the first part and (ii) since by (ii) \( E(\eta|\mathcal{G}) - E(\xi|\mathcal{G}) = E((\eta - \xi)|\mathcal{G}) \geq 0 \) a.s.

(v) This also follows at once from the definition of conditional expectation. \( \square \)

A variety of general results concerning conditional expectations will now be obtained – some involving conditional versions of standard theorems. The first is an important result on successive conditioning.
Theorem 13.2.2  If $\xi$ is a r.v. with $E|\xi| < \infty$ and $\mathcal{G}_1, \mathcal{G}_2$ two $\sigma$-fields with $\mathcal{G}_2 \subset \mathcal{G}_1 \subset \mathcal{F}$ then

$$E\{E(\xi|\mathcal{G}_1)|\mathcal{G}_2\} = E(\xi|\mathcal{G}_2) = E\{E(\xi|\mathcal{G}_2)|\mathcal{G}_1\} \text{ a.s.}$$

Proof  Repeated use of the definition shows that for all $E \in \mathcal{G}_2 \subset \mathcal{G}_1$,

$$\int_E E\{E(\xi|\mathcal{G}_1)|\mathcal{G}_2\} dP = \int_E E(\xi|\mathcal{G}_1) dP = \int_E \xi dP$$

which implies that $E\{E(\xi|\mathcal{G}_1)|\mathcal{G}_2\} = E(\xi|\mathcal{G}_2) \text{ a.s.}$ The right hand equality follows from Theorem 13.2.1 (v). \qed

The fundamental convergence theorems for integrals and expectations (monotone and dominated convergence, Fatou’s Lemma) have conditional versions. We prove the monotone convergence result – the other two then follow from it in the same way as for the corresponding “unconditional” theorems.

Theorem 13.2.3 (Conditional Monotone Convergence Theorem)  Let $\{\xi_n\}$ be an increasing sequence of nonnegative r.v.’s with $\lim \xi_n = \xi$ a.s., where $E\xi < \infty$. Then

$$E(\xi|\mathcal{G}) = \lim_{n \to \infty} E(\xi_n|\mathcal{G}) \text{ a.s.}$$

Proof  By Theorem 13.2.1 (iv) the sequence $\{E(\xi_n|\mathcal{G})\}$ is increasing and nonnegative a.s. The limit $\lim_{n \to \infty} E(\xi_n|\mathcal{G})$ is then $\mathcal{G}$-measurable and two applications of (ordinary) monotone convergence give, for any $E \in \mathcal{G}$,

$$\int_E \lim_{n \to \infty} E(\xi_n|\mathcal{G}) dP = \lim_{n \to \infty} \int_E E(\xi_n|\mathcal{G}) dP = \lim_{n \to \infty} \int_E \xi_n dP$$

showing that $\lim_{n \to \infty} E(\xi_n|\mathcal{G})$ satisfies the conditions required to be a version of $E(\xi|\mathcal{G})$ and hence the desired result follows. \qed

Theorem 13.2.4 (Conditional Fatou Lemma)  Let $\{\xi_n\}$ be a sequence of nonnegative r.v.’s with $E\xi_n < \infty$ and $E\{\liminf_{n \to \infty} \xi_n\} < \infty$. Then

$$E(\liminf_{n \to \infty} \xi_n|\mathcal{G}) \leq \liminf_{n \to \infty} E(\xi_n|\mathcal{G}) \text{ a.s.}$$

This and the next result will not be proved here since – as already noted – they follow from Theorem 13.2.3 in the same way as the ordinary versions of Fatou’s Lemma and dominated convergence follow from monotone convergence.
Theorem 13.2.5 (Conditional Dominated Convergence Theorem) Let \( \{\xi_n\} \) be a sequence of r.v.’s with \( \xi_n \to \xi \) a.s. and \( |\xi_n| \leq \eta \) a.s. for all \( n \) where \( \mathbb{E}|\eta| < \infty \). Then

\[
\mathbb{E}(\xi|\mathcal{G}) = \lim_{n \to \infty} \mathbb{E}(\xi_n|\mathcal{G}) \text{ a.s.}
\]

The following result is frequently useful.

Theorem 13.2.6 Let \( \xi, \eta \) be r.v.’s with \( \mathbb{E}|\eta| < \infty \), \( \mathbb{E}|\xi\eta| < \infty \) and such that \( \eta \) is \( \mathcal{G} \)-measurable (\( \xi \) being \( \mathcal{F} \)-measurable, of course). Then

\[
\mathbb{E}(\xi\eta|\mathcal{G}) = \eta \mathbb{E}(\xi|\mathcal{G}) \text{ a.s.}
\]

Proof If \( \eta = \chi_G \) for some \( G \in \mathcal{G} \) then \( \eta \mathbb{E}(\xi|\mathcal{G}) \) is \( \mathcal{G} \)-measurable and for any \( E \in \mathcal{G} \),

\[
\int_E \eta \mathbb{E}(\xi|\mathcal{G}) \, dP = \int_{E \cap G} \mathbb{E}(\xi|\mathcal{G}) \, dP = \int_E \mathbb{E}(\xi|\mathcal{G}) \, dP = \int_E \eta \, dP
\]

and hence \( \mathbb{E}(\xi\eta|\mathcal{G}) = \eta \mathbb{E}(\xi|\mathcal{G}) \) a.s. It follows from Theorem 13.2.1 (ii) that the result is true for simple \( \mathcal{G} \)-measurable r.v.’s \( \eta \).

Now if \( \eta \) is an arbitrary \( \mathcal{G} \)-measurable r.v. (with \( \eta \in L_1, \xi \eta \in L_1 \)), let \( \{\eta_n\} \) be a sequence of simple \( \mathcal{G} \)-measurable r.v.’s such that for all \( \omega \in \Omega \), \( \lim_n \eta_n(\omega) = \eta(\omega) \) and \( |\eta_n(\omega)| \leq |\eta(\omega)| \) for all \( n \) (Theorem 3.5.2, Corollary). It then follows from the conditional dominated convergence theorem \( (|\xi\eta_n| \leq |\xi\eta| \in L_1) \) that

\[
\mathbb{E}(\xi\eta|\mathcal{G}) = \lim_{n \to \infty} \mathbb{E}(\eta_n\xi|\mathcal{G}) = \lim_{n \to \infty} \eta_n \mathbb{E}(\xi|\mathcal{G}) = \eta \mathbb{E}(\xi|\mathcal{G}) \text{ a.s.} \quad \square
\]

The next result shows that in the presence of independence conditional expectation is the same as expectation.

Theorem 13.2.7 If \( \xi \) is a r.v. with \( \mathbb{E}|\xi| < \infty \) and \( \sigma(\xi) \) and \( \mathcal{G} \) are independent then

\[
\mathbb{E}(\xi|\mathcal{G}) = \mathbb{E}\xi \text{ a.s.}
\]

In particular if \( \xi \) and \( \eta \) are independent r.v.’s and \( \mathbb{E}|\xi| < \infty \), then \( \mathbb{E}(\xi|\eta) = \mathbb{E}\xi \) a.s.

Proof For any \( E \in \mathcal{G} \) the r.v.’s \( \xi \) and \( \chi_E \) are independent and thus

\[
\int_E \xi \, dP = \mathbb{E}(\xi \chi_E) = \mathbb{E}\xi \cdot \mathbb{E}\chi_E = \int_E \mathbb{E}\xi \, dP.
\]

Since the constant \( \mathbb{E}\xi \) is \( \mathcal{G} \)-measurable, it follows that \( \mathbb{E}(\xi|\mathcal{G}) = \mathbb{E}(\xi) \) a.s. \( \square \)

The conditional expectation \( \mathbb{E}(\xi|\eta) \) of \( \xi \) given a r.v. \( \eta \) is \( \sigma(\eta) \)-measurable and hence it immediately follows as shown in the next result that it is a Borel measurable function of \( \eta \).
Theorem 13.2.8  If $\xi$ and $\eta$ are r.v.’s with $E|\xi| < \infty$ then there is a Borel measurable function $h$ on $\mathbb{R}$ such that

$$E(\xi|\eta) = h(\eta) \text{ a.s.}$$

Proof  This follows immediately from Theorem 3.5.3 since $E(\xi|\eta)$ is $\sigma(\eta)$-measurable, i.e. $E(\xi|\eta)(\omega) = h(\eta(\omega))$ for some (Borel) measurable $h$.

Finally in this list we note the occasionally useful property that conditional expectations satisfy Jensen’s Inequality just as expectations do.

Theorem 13.2.9  If $g$ is a convex function on $\mathbb{R}$ and $\xi$ and $g(\xi)$ have finite expectations then

$$g( E(\xi|G)) \leq E(g(\xi)|G) \text{ a.s.}$$

Proof  As stated in the proof of Theorem 9.5.4, $g(x) \geq g(y) + (x - y)h(y)$ for all $x$ and $y$ and some $h(y)$ which is easily seen to be bounded on closed and bounded intervals. Thus whenever $y_n \to x$, $g(y_n) + (x - y_n)h(y_n) \to g(x)$. Hence for every real $x$,

$$g(x) = \sup_{r: \text{rational}} \{g(r) + (x - r)h(r)\}.$$

Putting $x = \xi$ and $y = r$ in the inequality gives

$$g(\xi) \geq g(r) + (\xi - r)h(r) \text{ a.s.}$$

and by taking conditional expectations and using (ii) and (iv) of Theorem 13.2.1

$$E(g(\xi)|G) \geq g(r) + (E(\xi|G) - r)h(r) \text{ a.s.}$$

Since the last inequality holds for all rational $r$, by taking the supremum of the right hand side and combining a countable set of events of zero probability we find

$$E(g(\xi)|G) \geq \sup_{r: \text{rational}} \{g(r) + (E(\xi|G) - r)h(r)\} = g(E(\xi|G)) \text{ a.s.}$$

A different proof is suggested in Ex. 13.7.

13.3 Conditional probability given a $\sigma$-field

If $A$ is an event in $\mathcal{F}$ and $\mathcal{G}$ is a sub-\(\sigma\)-field of $\mathcal{F}$ the conditional probability $P(A|\mathcal{G})$ of $A$ given $\mathcal{G}$ is defined by

$$P(A|\mathcal{G}) = E(\chi_A|\mathcal{G}).$$
Then for $E \in \mathcal{G}$, $P(A \cap E) = \int_E \chi_A \, dP = \int_E \mathcal{E}(\chi_A|\mathcal{G}) \, dP = \int_E P(A|\mathcal{G}) \, dP$ so that $P(A|\mathcal{G})$ is a $\mathcal{G}$-measurable (and $P$-integrable) r.v. which is determined uniquely a.s. by the equality

$$P(A \cap E) = \int_E P(A|\mathcal{G}) \, dP$$

(i.e. $P(A \cap E) = \mathcal{E}(\chi_E P(A|\mathcal{G}))$). In particular (by putting $E = \Omega$)

$$P(A) = \int_\Omega P(A|\mathcal{G}) \, dP \quad \text{(i.e. } \mathcal{E}P(A|\mathcal{G}) = P(A))$$

for all $A \in \mathcal{F}$. If $\eta$ is a r.v. then the conditional probability $P(A|\eta)$ of $A \in \mathcal{F}$ given $\eta$ is defined as $P(A|\eta) = P(A|\sigma(\eta)) = \mathcal{E}(\chi_A|\eta)$. The particular consequence $\mathcal{E}P(A|\eta) = P(A)$ is, of course, natural.

The properties of conditional probability follow immediately from those of conditional expectation. Some of these properties are collected in the following theorems for ease of reference.

**Theorem 13.3.1**  
(i) If $A \in \mathcal{G}$ then

$$P(A|\mathcal{G})(\omega) = \chi_A(\omega) = \begin{cases} 1 & \text{for } \omega \in A \\ 0 & \text{for } \omega \notin A \end{cases} \quad \text{a.s.}$$

(ii) If the event $A$ is independent of the class $\mathcal{G}$ of events then

$$P(A|\mathcal{G})(\omega) = P(A) \quad \text{a.s.}$$

**Theorem 13.3.2**  
(i) If $A \in \mathcal{F}$ then $0 \leq P(A|\mathcal{G}) \leq 1$ a.s.

(ii) $P(\Omega|\mathcal{G}) = 1$ a.s., $P(\emptyset|\mathcal{G}) = 0$ a.s.

(iii) If $\{A_n\}$ is a disjoint sequence of events in $\mathcal{F}$ and $A = \bigcup_{n=1}^\infty A_n$ then

$$P(A|\mathcal{G}) = \sum_{n=1}^\infty P(A_n|\mathcal{G}) \quad \text{a.s.}$$

(iv) If $A, B \in \mathcal{F}$ and $A \subset B$ then

$$P(A|\mathcal{G}) \leq P(B|\mathcal{G}) \quad \text{a.s.}$$

and

$$P(B - A|\mathcal{G}) = P(B|\mathcal{G}) - P(A|\mathcal{G}) \quad \text{a.s.}$$

(v) If $\{A_n\}_{n=1}^\infty$ is a monotone (increasing or decreasing) sequence of events in $\mathcal{F}$ and $A$ is its limit, then

$$P(A|\mathcal{G}) = \lim_{n \to \infty} P(A_n|\mathcal{G}) \quad \text{a.s.}$$
Proof These conclusions follow readily from the properties established for conditional expectations. For example, to show (iii) note that $\chi_A = \sum_1^\infty \chi_{A_n}$ and conditional monotone convergence (Theorem 13.2.3) gives $E(\chi_A | G) = \sum E(\chi_{A_n} | G)$ a.s. which simply restates (iii). □

The above properties look like those of a probability measure, with the exception that they hold a.s., and it is natural to ask whether for fixed $\omega \in \Omega$, $P(A|G)(\omega)$ as a function of $A \in \mathcal{F}$ is a probability measure. Unfortunately the answer is in general negative and this is due to the fact that the exceptional $G$-measurable set of zero probability that appears in each property of Theorem 13.3.2 depends on the events for which the property is expressed. In particular property (i) stated in detail would read:

(i) For every $A \in \mathcal{F}$ there is $N_A \in \mathcal{G}$ depending on $A$ such that $P(N_A) = 0$ and for all $\omega \notin N_A$

$$0 \leq P(A|G)(\omega) \leq 1.$$ 

It is then clear that the statement

$$0 \leq P(A|G) \leq 1 \text{ for all } A \in \mathcal{F} \text{ a.s.}$$

is not necessarily true in general, since to obtain this we would need to combine the zero probability sets $N_A$ to get a single zero probability set $N$. This can be done (as in the example of Section 13.1) if there are only countably many sets $A \in \mathcal{F}$, but not necessarily otherwise. In fact, in general, there may not even exist an event $E \in \mathcal{G}$ with $P(E) > 0$ such that

$$0 \leq P(A|G)(\omega) \leq 1 \text{ for all } A \in \mathcal{F} \text{ and all } \omega \in E.$$ 

Thus in general there is no event $E \in \mathcal{G}$ with $P(E) > 0$ such that for every fixed $\omega \in E$, $P(A|G)(\omega)$ is a probability measure on $\mathcal{F}$.

In the next section we consider the case where the conditional probability does have a version which is a probability measure for all $\omega$ (a “regular conditional probability”) and show that then conditional expectations can be expressed as integrals with respect to this version.

13.4 Regular conditioning

As seen in the previous section, conditional probabilities are not in general probability measures for fixed $\omega$. If a conditional probability has a version which is a probability measure for all $\omega$, then this version is called a regular conditional probability. Specifically let $G$ be a sub-$\sigma$-field of $\mathcal{F}$. A function $P(A, \omega)$ defined for each $A \in \mathcal{F}$ and $\omega \in \Omega$, with values in $[0, 1]$ is called a regular conditional probability on $\mathcal{F}$ given $G$ if
(i) for each fixed $A \in \mathcal{F}$, $P(A, \omega)$ is a $\mathcal{G}$-measurable function of $\omega$, and for each fixed $\omega \in \Omega$, $P(A, \omega)$ is a probability measure on $\mathcal{F}$, and

(ii) for each fixed $A \in \mathcal{F}$, $P(A, \omega) = P(A|\mathcal{G})(\omega)$ a.s.

Regular conditional probabilities do not always exist without any further assumptions on $\Omega$, $\mathcal{F}$ and $\mathcal{G}$. As we have seen, a simple case when they exist is when $\mathcal{G}$ is the $\sigma$-field generated by a discrete r.v. However, if a regular conditional probability does exist we can express conditional expectations as integrals with respect to it, just as ordinary expectations are expressed as integrals with respect to the probability measure. The notation $\int_{\Omega} \xi(\omega')P(d\omega', \omega)$ will be convenient to indicate integration of $\xi$ with respect to the measure $P(\cdot, \omega)$.

**Theorem 13.4.1** If $\xi$ is a r.v. with $E|\xi| < \infty$, and $P(A, \omega)$ is a regular conditional probability on $\mathcal{F}$ given $\mathcal{G}$, then

$$E(\xi|\mathcal{G})(\omega) = \int_{\Omega} \xi(\omega')P(d\omega', \omega) \text{ a.s.}$$

**Proof** If $\xi = \chi_A$ for some $A \in \mathcal{F}$, then $\int_{\Omega} \xi(\omega')P(d\omega', \omega) = P(A, \omega)$ which is $\mathcal{G}$-measurable and equal a.s. to

$$P(A|\mathcal{G})(\omega) = E(\chi_A|\mathcal{G})(\omega) = E(\xi|\mathcal{G})(\omega).$$

Thus $\int_{\Omega} \xi(\omega')P(d\omega', \omega)$ is $\mathcal{G}$-measurable and equal a.s. to $E(\xi|\mathcal{G})(\omega)$ when $\xi$ is a set indicator. It follows by Theorem 13.2.1 (ii) that the same is true for a simple r.v. $\xi$ and, by using the ordinary and the conditional monotone convergence theorem, it is also true for any r.v. $\xi \geq 0$ with $E\xi < \infty$. Using again Theorem 13.2.1 (ii), this is also true for any r.v. $\xi$ with $E|\xi| < \infty$. □

If one is only interested in expressing a conditional expectation $E[g(\xi)|\mathcal{G}]$ for a particular $\xi$ and Borel measurable $g$, as an integral with respect to a conditional probability (as in the previous theorem) then attention may be restricted to conditional probabilities $P(A|\mathcal{G})$ of events $A$ in $\sigma(\xi)$ since $\mathcal{F}$ may be replaced by $\sigma(\xi)$ in defining integrals of $\xi$ over $\Omega$ (Ex. 4.10). We will call this restriction the conditional probability of $\xi$ given $\mathcal{G}$ and it will be seen in Theorem 13.4.5 that a regular version exists under a simple condition on $\xi$. To be precise let $\xi$ be a r.v. and $\mathcal{G}$ a sub-$\sigma$-field of $\mathcal{F}$. A function $P_{\xi|\mathcal{G}}(A, \omega)$ defined for each $A \in \sigma(\xi)$ and $\omega \in \Omega$, with values in $[0, 1]$ is called a regular conditional probability of $\xi$ given $\mathcal{G}$ if

(i) for each fixed $A \in \sigma(\xi)$, $P_{\xi|\mathcal{G}}(A, \omega)$ is a $\mathcal{G}$-measurable function of $\omega$, and for each fixed $\omega \in \Omega$, $P_{\xi|\mathcal{G}}(A, \omega)$ is a probability measure on $\sigma(\xi)$, and

(ii) for each fixed $A \in \sigma(\xi)$, $P_{\xi|\mathcal{G}}(A, \omega) = P(A|\mathcal{G})(\omega)$ a.s.

Theorem 13.4.5 will show that under a very mild condition on $\xi$ (that the range of $\xi$ is a Borel set) $P_{\xi|\mathcal{G}}$ of $\xi$ given $\mathcal{G}$ exists for all $\mathcal{G}$. Also as
already noted if \( G = \sigma(\eta) \) and \( \eta \) is a discrete r.v. then a regular conditional probability \( P_{\xi|G} \) exists. Two further cases where \( P_{\xi|G} \) exists trivially (in view of Theorem 13.3.1) are the following: (i) if \( \sigma(\xi) \) and \( G \) are independent then

\[
P_{\xi|G}(A, \omega) = P(A) \text{ for all } A \in \sigma(\xi) \text{ and } \omega \in \Omega
\]

and (ii) if \( \xi \) is \( G \)-measurable then

\[
P_{\xi|G}(A, \omega) = \chi_A(\omega) \text{ for all } A \in \sigma(\xi) \text{ and } \omega \in \Omega.
\]

As will now be shown, when a regular conditional probability of \( \xi \) given \( G \) exists, then the conditional expectation of every \( \sigma(\xi) \)-measurable r.v. with finite expectation can be expressed as an integral with respect to the regular conditional probability.

**Theorem 13.4.2** If \( \xi \) is a r.v., \( g \) a Borel measurable function on \( \mathbb{R} \) such that \( E|g(\xi)| < \infty \), and \( P_{\xi|G} \) is a regular conditional probability of \( \xi \) given \( G \), then

\[
E\{g(\xi)|G\}(\omega) = \int_{\Omega} g(\xi(\omega'))P_{\xi|G}(d\omega', \omega) \text{ a.s.}
\]

*Proof* The proof extends that of Theorem 13.4.1, with \( \sigma(\xi) \) replacing \( \mathcal{F} \). If \( A \in \sigma(\xi) \) the r.v. \( \eta = \chi_A \) satisfies \( E(\eta|G)(\omega) = \int \eta(\omega')P_{\xi|G}(d\omega', \omega) \) a.s. This remains true if \( \chi_A \) is replaced by a nonnegative simple \( \sigma(\xi) \)-measurable r.v. \( \eta \) and hence by the standard extension (cf. Theorem 13.4.1) for any \( \sigma(\xi) \)-measurable \( \eta \) with \( E|\eta| < \infty \). But \( g(\xi) \) is such a r.v. and hence the result follows. \( \square \)

The distribution of a r.v. \( \xi \) (Chapter 9) is the probability measure \( P_{\xi}^{-1} \) induced from \( P \) on the Borel sets of the real line by \( \xi \) and expectations of functions of \( \xi \) are expressible as integrals with respect to this distribution. Similarly, conditional distributions on the Borel sets of the real line may be induced from regular conditional probabilities and used to obtain conditional expectations. Indeed if the regular conditional probability \( P_{\xi|G}(A, \omega) \) of \( \xi \) given \( G \) exists then a (regular) conditional distribution \( Q_{\xi|G}(B, \omega) \) of \( \xi \) given \( G \) may be defined for any Borel set \( B \) on the real line (i.e. \( B \in \mathcal{B} \)) and \( \omega \in \Omega \) by

\[
Q_{\xi|G}(B, \omega) = P_{\xi|G}(\xi^{-1}B, \omega) \text{ for all } B \in \mathcal{B}, \omega \in \Omega.
\]

Clearly \( Q_{\xi|G} \) has properties similar to \( P_{\xi|G} \) and the only problem is that this “definition” of \( Q_{\xi|G} \) requires the existence of \( P_{\xi|G} \) (which is not always guaranteed). However, this problem is easily eliminated by defining \( Q_{\xi|G} \) in terms of properties it inherits from \( P_{\xi|G} \) but without reference to the latter. More specifically let \( \xi \) be a r.v. and \( G \) a sub-\( \sigma \)-field of \( \mathcal{F} \). A function
$Q_{\xi|G}(B, \omega)$ defined for each $B \in \mathcal{B}$ and $\omega \in \Omega$, with values in $[0, 1]$ is called a regular conditional distribution of $\xi$ given $G$ if

(i) for each fixed $B \in \mathcal{B}$, $Q_{\xi|G}(B, \omega)$ is a $\mathcal{G}$-measurable function of $\omega$, and for each fixed $\omega \in \Omega$, $Q_{\xi|G}(B, \omega)$ is a probability measure on the Borel sets $\mathcal{B}$, and

(ii) for each fixed $B \in \mathcal{B}$, $Q_{\xi|G}(B, \omega) = P(\xi^{-1}(B)|G)(\omega)$ a.s.

It is clear that if a regular conditional probability $P_{\xi|G}$ of $\xi$ given $G$ exists then $Q_{\xi|G}$ as defined above from it, is a regular conditional distribution of $\xi$ given $G$.

We shall see that, in contrast to regular conditional probability, a regular conditional distribution of $\xi$ given $G$ always exists (Theorem 13.4.3) and that the conditional expectation of every $\sigma(\xi)$-measurable r.v. with finite expectation may be expressed as an integral over $\mathbb{R}$ with respect to the regular conditional distribution (Theorem 13.4.4).

As for the regular conditional probability of $\xi$ given $G$ the following intuitively appealing results hold:

(i) if $\sigma(\xi)$ and $G$ are independent, then

$$Q_{\xi|G}(B, \omega) = P_{\xi}(B)$$

i.e. for each fixed $\omega \in \Omega$ the conditional distribution of $\xi$ given $G$ is just the distribution of $\xi$.

(ii) If $\xi$ is $\mathcal{G}$-measurable, then

$$Q_{\xi|G}(B, \omega) = \chi_{\xi^{-1}(B)}(\omega) = \chi_{\xi}(\omega)$$

i.e. for each fixed $\omega \in \Omega$ the conditional distribution of $\xi$ given $G$ is a probability measure concentrated at the point $\xi(\omega)$.

**Theorem 13.4.3** If $\xi$ is a r.v. and $G$ a sub-$\sigma$-field of $\mathcal{F}$, then there exists a regular conditional distribution of $\xi$ given $G$.

**Proof** Write $A_x = \xi^{-1}(-\infty, x]$ for any real $x$. By Theorem 13.3.2 it is clear that for any fixed $x, y$ with $x \geq y$, $P(A_x|\mathcal{G})(\omega) \geq P(A_y|\mathcal{G})(\omega)$ a.s., for any fixed $x$, $P(A_{x+1/n}|\mathcal{G})(\omega) \rightarrow P(A_x|\mathcal{G})(\omega)$ a.s. as $n \rightarrow \infty$, and for any fixed sequence $\{x_n\}$ with $x_n \rightarrow \infty$ ($-\infty$), $P(A_{x_n}|\mathcal{G})(\omega) \rightarrow 1$ ($0$) a.s. By combining a countable number of zero measure sets in $\mathcal{G}$ we obtain a $\mathcal{G}$-measurable set $N$ with $P(N) = 0$ such that for each $\omega \notin N$

(a) $P(A_x|\mathcal{G})(\omega)$ is a nondecreasing function of rational $x$

(b) $\lim_{n \rightarrow \infty} P(A_{x+1/n}|\mathcal{G})(\omega) = P(A_x|\mathcal{G})(\omega)$ for all rational $x$

(c) $\lim_{x \rightarrow \infty} P(A_x|\mathcal{G})(\omega) = 1$, $\lim_{x \rightarrow -\infty} P(A_x|\mathcal{G})(\omega) = 0$ for rational $x \rightarrow \pm \infty$. 

Define functions $F(x, \omega)$ as follows:

for $\omega \notin \mathbb{N}$: $F(x, \omega) = P(A_x|\mathcal{G})(\omega)$ if $x$ is rational

$$= \lim_{r \downarrow x} [F(r, \omega) : r \text{ rational, } r \downarrow x]$$

if $x$ is irrational

for $\omega \in \mathbb{N}$: $F(x, \omega) = 0$ or $1$ according as $x < 0$ or $x \geq 0$.

Then it is easily checked that $F(x, \omega)$ is a distribution function for each fixed $\omega \in \Omega$ and hence defines a probability measure $Q(B, \omega)$ on the class $\mathcal{B}$ of Borel sets, satisfying $Q((-\infty, x], \omega) = F(x, \omega)$ for each real $x$.

It will follow that $Q(B, \omega)$ is the desired regular conditional distribution of $\xi$ given $\mathcal{G}$ if we show that for each $B \in \mathcal{B}$,

(i) $Q(B, \omega)$ is a $\mathcal{G}$-measurable function of $\omega$

(ii) $Q(B, \omega) = P(\xi^{-1}B|\mathcal{G})(\omega)$ a.s.

Let $\mathcal{D}$ be the class of all Borel sets $B$ for which (i) and (ii) hold. If $x$ is rational and $B = (-\infty, x]$, then $Q(B, \omega) = F(x, \omega)$ which is equal to the $\mathcal{G}$-measurable function $P(A_x|\mathcal{G})(\omega)$ if $\omega \notin \mathbb{N}$ and a constant (0 or 1) if $\omega \in \mathbb{N}$. Further $N \in \mathcal{G}$ and $P(N) = 0$. Since $A_x = \xi^{-1}B$, (i) and (ii) both follow when $B = (-\infty, x]$, for rational $x$. Thus $(-\infty, x] \in \mathcal{D}$ when $x$ is rational.

It is easily checked that $\mathcal{D}$ is a $\mathcal{D}$-class. If $B_i$ are disjoint sets of $\mathcal{D}$, with $B = \bigcup B_i$ we have $Q(B, \omega) = \sum Q(B_i, \omega)$ which is $\mathcal{G}$-measurable since each term is, so that (i) holds. Also, $\sum Q(B_i, \omega) = \sum P(\xi^{-1}B_i|\mathcal{G})(\omega) = P(\bigcup \xi^{-1}B_i|\mathcal{G})(\omega)$ a.s. by Theorem 13.3.2, and this is $P(\xi^{-1}B|\mathcal{G})$, so that $\mathcal{D}$ is closed under countable disjoint unions. Similarly it is closed under proper differences.

Thus $\mathcal{D}$ is a $\mathcal{D}$-class containing the class of all sets of the form $(-\infty, x]$ for rational $x$. But this latter class is closed under intersections, and its generated $\sigma$-ring is $\mathcal{B}$ (cf. Ex. 1.21). Hence $\mathcal{D} \supset \mathcal{B}$, as desired. \hfill \Box

The following result shows in particular that the conditional expectation of a function $g$ of a r.v. $\xi$ may be obtained by integrating $g$ with respect to a regular conditional distribution of $\xi$ (cf. Theorem 13.4.2).

**Theorem 13.4.4** Let $\xi$ be a r.v. and $Q_{\xi|\mathcal{G}}$ a regular conditional distribution of $\xi$ given $\mathcal{G}$. Let $\eta$ be a $\mathcal{G}$-measurable r.v. and $g$ a Borel measurable function on the plane such that $\mathcal{E}|g(\xi, \eta)| < \infty$. Then

$$\mathcal{E}|g(\xi, \eta)|\mathcal{G})(\omega) = \int_{-\infty}^{\infty} g(x, \eta(\omega)) Q_{\xi|\mathcal{G}}(dx, \omega) \text{ a.s.}$$

In particular, if $E$ is a Borel measurable set of the plane and $E^y$ its y-section \{x ∈ \mathbb{R} : (x, y) ∈ E\}, then

$$P((\xi, \eta) ∈ E|\mathcal{G})(\omega) = Q_{\xi|\mathcal{G}}(E^y(\omega), \omega) \text{ a.s.}$$
Proof. We will first show that for every $E \in \mathcal{B}^2$, $Q_{\xi \in \mathcal{G}}(E^{\eta(\omega)}, \omega)$ is $\mathcal{G}$-measurable and $P((\xi, \eta) \in E|\mathcal{G})(\omega) = Q_{\xi \in \mathcal{G}}(E^{\eta(\omega)}, \omega)$ a.s. Let $E = A \times B$ where $A, B \in \mathcal{B}$. Then $Q_{\xi \in \mathcal{G}}(E^{\eta(\omega)}, \omega) = Q_{\xi \in \mathcal{G}}(A, \omega)$ or $Q_{\xi \in \mathcal{G}}(\emptyset, \omega)$ according as $\eta(\omega) \in B$ or $B^c$, so that clearly $Q_{\xi \in \mathcal{G}}(E^{\eta(\omega)}, \omega)$ is $\mathcal{G}$-measurable. Further since $Q_{\xi \in \mathcal{G}}(A, \omega) = P(\xi^{-1}A|\mathcal{G})$ a.s. and $P(\xi^{-1}\emptyset|\mathcal{G}) = 0$ a.s., it follows that

$$Q_{\xi \in \mathcal{G}}(E^{\eta(\omega)}, \omega) = \chi_{\eta^{-1}B}(\omega)P(\xi^{-1}A|\mathcal{G})(\omega) \text{ a.s.}$$

$$= \chi_{\eta^{-1}B}(\omega)E[\chi_{\xi^{-1}A}|\mathcal{G}](\omega) \text{ a.s.}$$

$$= E[\chi_{\xi^{-1}A}\chi_{\eta^{-1}B}|\mathcal{G}](\omega) \text{ a.s.}$$

$$= P((\xi, \eta) \in E|\mathcal{G})(\omega) \text{ a.s.}$$

(since $\chi_{\eta^{-1}B}$ is $\sigma(\eta)$-measurable). Hence $Q_{\xi \in \mathcal{G}}(E^{\eta(\omega)}, \omega)$ is (a version of) $P((\xi, \eta) \in E|\mathcal{G})$ when $E = A \times B$, $A, B \in \mathcal{B}$.

Now denote by $\mathcal{D}$ the class of subsets $E$ of $\mathbb{R}^2$ such that $Q_{\xi \in \mathcal{G}}(E^{\eta(\omega)}, \omega)$ is $\mathcal{G}$-measurable and $P((\xi, \eta) \in E|\mathcal{G})(\omega) = Q_{\xi \in \mathcal{G}}(E^{\eta(\omega)}, \omega)$ a.s. (the exceptional set depending in general on each set $E$). Then by writing $P((\xi, \eta) \in E|\mathcal{G}) = E[\chi_{\{\xi \in \sigma(\eta)\}}|\mathcal{G}]$ and using the properties of conditional expectation and the regular conditional distribution it is seen immediately that $\mathcal{D}$ is a $\mathcal{D}$-class (i.e. closed under countable disjoint unions and proper differences). Since $\mathcal{D}$ contains the Borel measurable rectangles of $\mathbb{R}^2$, it will contain the $\sigma$-field they generate, the Borel sets $\mathcal{B}^2$ of $\mathbb{R}^2$. Hence the second equality of the theorem is proved.

The first equality is then obtained by the usual extension. If $g = \chi_E$, the indicator of a set $E \in \mathcal{B}^2$, then by the above the equality holds. Hence it also holds for a $\mathcal{B}$-measurable simple function $g$. By using the ordinary and the conditional monotone convergence theorem (and Theorem 3.5.2) we see that it is true for all nonnegative $\mathcal{B}^2$-measurable functions $g$ and hence also for all $g$ as in the theorem. \qed

Since a regular conditional distribution $Q_{\xi \in \mathcal{G}}$ of $\xi$ given $\mathcal{G}$ always exists, one may attempt to obtain a regular conditional probability $P_{\xi \in \mathcal{G}}$ of $\xi$ given $\mathcal{G}$ by

$$P_{\xi \in \mathcal{G}}(A, \omega) = Q_{\xi \in \mathcal{G}}(B, \omega) \text{ when } A \in \sigma(\xi), B \in \mathcal{B}, A = \xi^{-1}B$$

(as was pointed out earlier in this section, if $P_{\xi \in \mathcal{G}}$ exists this relationship defines a regular conditional distribution $Q_{\xi \in \mathcal{G}}$). However, given $A \in \sigma(\xi)$ there may be several Borel sets $B$ such that $A = \xi^{-1}B$ for which the values $Q_{\xi \in \mathcal{G}}(B, \omega)$ are not all equal (for fixed $\omega$) and then $P_{\xi \in \mathcal{G}}$ is not defined in the above way. Under a rather mild condition on $\xi$ it is shown in the following theorem that this difficulty is eliminated and a regular conditional probability can then be defined from a regular conditional distribution.
Theorem 13.4.5  Let $\xi$ be a r.v. (for convenience defined for all $\omega$) and $\mathcal{G}$ a sub-$\sigma$-field of $\mathcal{F}$. If the range $E = \{\xi(\omega) : \omega \in \Omega\}$ of $\xi$ is a Borel set then there exists a regular conditional probability of $\xi$ given $\mathcal{G}$.

Proof  Let $Q_{\xi|\mathcal{G}}$ be a regular conditional distribution of $\xi$ given $\mathcal{G}$, which always exists by Theorem 13.4.3. Then since $E \in \mathcal{B}$ and $\xi^{-1}(E) = \Omega$, 

$$Q_{\xi|\mathcal{G}}(E, \omega) = P(\xi^{-1}(E)|\mathcal{G})(\omega) = P(\Omega|\mathcal{G})(\omega) = 1 \text{ a.s.}$$

and thus there is a set $N \in \mathcal{G}$, with $P(N) = 0$, such that for all $\omega \notin N$, $Q_{\xi|\mathcal{G}}(E, \omega) = 1$.

Now fix $A \in \sigma(\xi)$ with $A = \xi^{-1}(B_1) = \xi^{-1}(B_2)$ where $B_1, B_2 \in \mathcal{B}$. Then $B_1 - B_2$ and $B_2 - B_1$ are Borel subsets of $E^c$ and thus for all $\omega \notin N$ (since $Q_{\xi|\mathcal{G}}$ is a measure for every $\omega$)

$$Q_{\xi|\mathcal{G}}(B_1 - B_2, \omega) = 0 = Q_{\xi|\mathcal{G}}(B_2 - B_1, \omega)$$

so that

$$Q_{\xi|\mathcal{G}}(B_1, \omega) = Q_{\xi|\mathcal{G}}(B_1 \cap B_2, \omega) = Q_{\xi|\mathcal{G}}(B_2, \omega).$$

Hence the following definition is unambiguous.

$$P_{\xi|\mathcal{G}}(A, \omega) = \begin{cases} Q_{\xi|\mathcal{G}}(B, \omega) & \text{for } \omega \notin N \\ p(A) & \text{for } \omega \in N \end{cases} \text{ and all } A \in \sigma(\xi)$$

where $B \in \mathcal{B}$ is such that $A = \xi^{-1}(B)$ and $p$ is an arbitrary but fixed probability measure on $\sigma(\xi)$. Since $Q_{\xi|\mathcal{G}}$ is a regular conditional distribution of $\xi$ given $\mathcal{G}$ and since $P(N) = 0$, it is clear that $P_{\xi|\mathcal{G}}$ is a regular conditional probability of $\xi$ given $\mathcal{G}$. $\square$

Finally, if $\eta$ is a r.v. then the following notions

regular conditional probability on $\mathcal{F}$ given $\eta$

regular conditional probability of $\xi$ given $\eta$

regular conditional distribution of $\xi$ given $\eta$

are defined (as usual) as the corresponding quantities introduced in this section with $\mathcal{G} = \sigma(\eta)$, the notation used here for the last two being $P_{\xi|\eta}$ and $Q_{\xi|\eta}$. A regular conditional distribution $Q_{\xi|\eta}$ of $\xi$ given $\eta$ always exists (Theorem 13.4.3) and the conditional expectation given $\eta$ of every $\sigma(\xi, \eta)$-measurable r.v. with finite expectation is expressed as an integral with respect to $Q_{\xi|\eta}$, as follows from Theorem 13.4.4. Thus, if $g$ is a Borel measurable function on the plane such that $E|g(\xi, \eta)| < \infty$, then

$$E[g(\xi, \eta)|\eta](\omega) = \int_{-\infty}^{\infty} g(x, \eta(\omega)) Q_{\xi|\eta}(dx, \omega) \text{ a.s.}$$
In particular, if $E$ is a Borel measurable set of the plane and $E^y$ its $y$-section \( \{ x \in \mathbb{R} : (x, y) \in E \} \), then

\[
P((\xi, \eta) \in E|\eta)(\omega) = Q_{\xi|\eta}(E^\eta(\omega), \omega) \text{ a.s.}
\]

### 13.5 Conditioning on the value of a r.v.

As promised in Section 13.1 we will now define conditional expectation (and hence then also conditional probability) given the event that a r.v. $\eta$ takes the value $y$, which may have probability zero even for all $y$. The conditional expectation given $\eta = y$ will be defined first giving the conditional probability as a particular case. Specifically if $\xi, \eta$ are r.v.’s, with $E|\xi < \infty$, it is known by Theorem 13.2.8 that the conditional expectation of $\xi$ given $\eta$ is a Borel measurable function of $\eta$, i.e. $E(\xi|\eta)(\omega) = h(\eta(\omega))$ for some Borel function $h$. The conditional expectation of $\xi$ given the value $y$ of $\eta$ may then be simply defined by

\[
E(\xi|\eta = y) = h(y)
\]

that is $E(\xi|\eta = y)$ may be regarded as a version of the conditional expectation induced on $\mathbb{R}$ by the transformation $\eta(\omega)$ (and thus Borel, rather than $\sigma(\eta)$-measurable).

If $B \in \mathcal{B}$ it follows at once that

\[
\int_B E(\xi|\eta = y) \, dP_\eta^{-1}(y) = \int_B h(y) \, dP_\eta^{-1}(y) = \int_{\eta^{-1}B} h(\eta(\omega)) \, dP(\omega) = \int_{\eta^{-1}B} E(\xi|\eta)(\omega) \, dP(\omega) = \int_{\eta^{-1}B} E(\xi \mid \eta = y) \, dP.
\]

Since in particular $\int_B h(y) \, dP_\eta^{-1}(y) = \int_{\eta^{-1}B} E(\xi \mid \eta = y) \, dP$, any two choices of $h(y)$ have the same integral $\int_B h \, dP_\eta^{-1}$ for every $B$ and hence must be equal a.s. ($P_\eta^{-1}$) so that $E(\xi|\eta = y)$ is uniquely defined (a.s.).

This is, of course, totally analogous to the defining property for $E(\xi|\eta)$ and may be similarly used as an independent definition of $E(\xi|\eta = y)$ as indicated in the following result.

**Theorem 13.5.1** For a r.v. $\xi$ with $E|\xi < \infty$ and a r.v. $\eta$, the conditional expectation of $\xi$ given $\eta = y$ may be equivalently defined (uniquely a.s. ($P_\eta^{-1}$)) as a $\mathcal{B}$-measurable function $E(\xi|\eta = y)$ satisfying

\[
\int_{\eta^{-1}B} E(\xi \mid \eta = y) \, dP_\eta^{-1}(y) = \int_B E(\xi \mid \eta = y) \, dP_\eta^{-1}(y)
\]

for each $B \in \mathcal{B}$. In particular it follows by taking $B = \mathbb{R}$ that $E\xi = \int E(\xi|\eta = y) \, dP_\eta^{-1}(y) = \int E(\xi|\eta = y) \, dF_\eta(y)$ where $F_\eta$ is the d.f. of $\eta$. 


Proof That \( \mathbb{E}(\xi \mid \eta = y) \) exists satisfying the defining equation and is a.s.
unique follow as above, or may be shown directly from use of the Radon–
Nikodym Theorem similarly to the definition of \( \mathbb{E}(\xi \mid \mathcal{G}) \) in Section 13.2. □

The conditional probability \( P(A \mid \eta = y) \) of \( A \in \mathcal{F} \) given \( \eta = y \) is now
defined as

\[
P(A \mid \eta = y) = \mathbb{E}(\chi_A \mid \eta = y) \quad \text{a.s. (} P\eta^{-1} \text{)}
\]

Thus \( P(A \mid \eta = y) \) is a Borel measurable (and \( P\eta^{-1} \)-integrable) function on \( \mathbb{R} \) which is determined uniquely a.s. (\( P\eta^{-1} \)) by the equality

\[
P(A \cap \eta^{-1} B) = \int_B P(A \mid \eta = y) \, dP\eta^{-1}(y) \text{ for all } B \in \mathcal{B}.
\]

In particular, for \( B = \mathbb{R} \)

\[
P(A) = \int_{-\infty}^{\infty} P(A \mid \eta = y) \, dP\eta^{-1}(y).
\]

Since \( P(A \mid \eta = y) = f(y) \) where \( P(A \mid \eta)(\omega) = f(\eta(\omega)) \), the properties of \( P(A \mid \eta = y) \) are easily deduced from those of \( P(A \mid \eta) \). In particular all properties of Theorem 13.3.2 are valid, with “given \( \mathcal{G} \)” replaced by “given \( \eta = y \)” and “a.s.” replaced by “a.s. (\( P\eta^{-1} \))”.

In a similar way the following notions can be defined for r.v.’s \( \xi, \eta \):

- regular conditional probability of \( \mathcal{F} \) given \( \eta = y \)
- regular conditional probability of \( \xi \) given \( \eta = y \)
- regular conditional distribution of \( \xi \) given \( \eta = y \)

with properties similar to the properties of the corresponding notions “given \( \eta \)” or “given \( \mathcal{G} \)” as developed in Section 13.4. These definitions and properties will not all be listed here, in order to avoid overburdening the text, but as an example consider the third notion (which always exists), defined as follows. A function \( \hat{Q}_{\xi \mid \eta}(B, y) \) defined on \( \mathcal{B} \times \mathbb{R} \) to \([0, 1]\) is called a regular conditional distribution of \( \xi \) given \( \eta = y \) if

(i) for each fixed \( B \in \mathcal{B} \), \( \hat{Q}_{\xi \mid \eta}(B, y) \) is a Borel measurable function of \( y \), and for each fixed \( y \in \mathbb{R} \), \( \hat{Q}_{\xi \mid \eta}(B, y) \) is a probability measure on the Borel sets \( \mathcal{B} \), and

(ii) for each fixed \( B \in \mathcal{B} \), \( \hat{Q}_{\xi \mid \eta}(B, y) = P(\xi^{-1} B \mid \eta = y) \) a.s. (\( P\eta^{-1} \)).

As for a regular conditional distribution of \( \xi \) given \( \eta \) there are the following extreme cases:

(i) if \( \xi \) and \( \eta \) are independent then \( \hat{Q}_{\xi \mid \eta}(B, y) = P\xi^{-1}(B) \) for all \( B \in \mathcal{B} \) and \( y \in \mathbb{R} \), i.e. for every fixed \( y \in \mathbb{R} \), the conditional distribution of \( \xi \) given \( \eta = y \) is equal to the distribution of \( \xi \); and
(ii) if $\xi$ is $\sigma(\eta)$-measurable then $\hat{Q}_{\xi|\eta}(B,y) = \chi_B(f(y))$ for all $B \in \mathcal{B}$ and $y \in \mathbb{R}$; where $f$ is defined by $\xi = f(\eta)$, i.e. for each fixed $y \in \mathbb{R}$, the conditional distribution of $\xi$ given $\eta = y$ is a probability measure concentrated at the point $f(y)$.

The main properties of a regular conditional distribution of $\xi$ given $\eta = y$ are collected in the following result.

**Theorem 13.5.2** Let $\xi$ and $\eta$ be r.v.’s. Then

(i) There exists a regular conditional distribution of $\xi$ given $\eta = y$.

(ii) If $Q_{\xi|\eta}$ and $\hat{Q}_{\xi|\eta}$ are regular conditional distributions of $\xi$ given $\eta$ and given $\eta = y$ respectively, then

$$Q_{\xi|\eta}(B, \omega) = \hat{Q}_{\xi|\eta}(B, \eta(\omega))$$

for all $B \in \mathcal{B}$ and $\omega \notin N$.

(iii) If $g$ is a Borel measurable function on the plane such that $E|g(\xi, \eta)| < \infty$, then

$$E\{g(\xi, \eta)|\eta = y\} = \int_{-\infty}^{\infty} g(x, y) \hat{Q}_{\xi|\eta}(dx, y) \text{ a.s. (}P_\eta^{-1}\text{).}$$

In particular, if $E$ is a Borel measurable set of the plane and $E^y$ its $y$-section $\{x \in \mathbb{R} : (x, y) \in E\}$, then

$$P\{(\xi, \eta) \in E|\eta = y\} = \hat{Q}_{\xi|\eta}(E^y, y) \text{ a.s. (}P_\eta^{-1}\text{).}$$

**Proof** The construction of a regular conditional distribution of $\xi$ given $\eta = y$ follows that of Theorem 13.4.3 in detail, with the obvious adjustments: “given $\mathcal{G}$” is replaced by “given $\eta = y$”, the exceptional $\mathcal{G}$-measurable sets with $P$-measure zero become Borel sets with $P_\eta^{-1}$-measure zero, and instead of defining $F(x, \omega)$ from $\mathbb{R} \times \Omega$ to $[0, 1]$, it is defined from $\mathbb{R} \times \mathbb{R}$ to $[0, 1]$. All the needed properties for conditional probabilities given $\eta = y$ are valid since as already noted Theorem 13.3.2 holds with “$\mathcal{G}$” replaced by “$\eta = y$”.

Now let $Q_{\xi|\eta}$ and $\hat{Q}_{\xi|\eta}$ be a regular conditional distribution of $\xi$ given $\eta$ and $\eta = y$ respectively. Then for each fixed $B \in \mathcal{B}$, $Q_{\xi|\eta}(B, \omega) = P(\xi^{-1}B|\eta)(\omega)$ a.s., $\hat{Q}_{\xi|\eta}(B, y) = P(\xi^{-1}B|\eta = y)$ a.s. ($P_\eta^{-1}$) and it follows from the conditional probability version of Theorem 13.5.1 that

$$Q_{\xi|\eta}(B, \omega) = \hat{Q}_{\xi|\eta}(B, \eta(\omega)) \text{ a.s.}$$

From now on we write $Q$ and $\hat{Q}$ for $Q_{\xi|\eta}$ and $\hat{Q}_{\xi|\eta}$. Let $\{B_n\}$ be a sequence of Borel sets which generates the $\sigma$-field of Borel sets $\mathcal{B}$ (cf. Ex. 1.21).
Then by combining a countable number of \( \sigma(\eta) \)-measurable sets of zero probability we obtain a set \( N \in \sigma(\eta) \) with \( P(N) = 0 \) such that
\[
Q(B_n, \omega) = \hat{Q}(B_n, \eta(\omega)) \quad \text{for all } n \text{ and all } \omega \notin N.
\]

Denote by \( C \) the class of all subsets \( B \) of the real line such that \( Q(B, \omega) = \hat{Q}(B, \eta(\omega)) \) for all \( \omega \notin N \). Since for each \( \omega \in \Omega \), \( Q(B, \omega) \) and \( \hat{Q}(B, \eta(\omega)) \) are probability measures on \( \mathcal{B} \), it follows simply that \( C \) is a \( \sigma \)-field and since it contains \( \{B_n\} \) it will contain its generated \( \sigma \)-field \( \mathcal{B} \). Thus \( Q(B, \omega) = \hat{Q}(B, \eta(\omega)) \) for all \( B \in \mathcal{B} \) and \( \omega \notin N \), i.e. (ii) holds.

(iii) follows immediately from Theorem 13.4.4 (see also the last paragraph of Section 13.4), the relationship between \( Q_{\xi|\eta} \) and \( \hat{Q}_{\xi|\eta} \), and Theorem 13.5.1 in the following form:

If \( E\{g(\xi, \eta)|\eta\}(\omega) = f(\eta(\omega)) \) a.s. then \( E\{g(\xi, \eta)|\eta = y\} = f(y) \) a.s. \( (P\eta^{-1}) \). \( \square \)

### 13.6 Regular conditional densities

For two r.v.'s \( \xi \) and \( \eta \) we have (in Sections 13.4 and 13.5) defined the regular conditional distribution \( Q_{\xi|\eta}(B, \omega) \) of \( \xi \) given \( \eta \) and the regular conditional distribution \( \hat{Q}_{\xi|\eta}(B, y) \) of \( \xi \) given \( \eta = y \), and have shown that both always exist. For each fixed \( \omega \) and \( y \), \( Q_{\xi|\eta}(\cdot, \omega) \) and \( \hat{Q}_{\xi|\eta}(\cdot, y) \) are probability measures on the Borel sets \( \mathcal{B} \), and if they are absolutely continuous with respect to Lebesgue measure it is natural to call their Radon–Nikodym derivatives conditional densities of \( \xi \) given \( \eta \), and given \( \eta = y \) respectively. As is clear from the previous sections regular versions of conditional densities will be of primary interest. To be precise, a function \( f_{\xi|\eta}(x, \omega) \) defined on \( \mathbb{R} \times \Omega \) to \( [0, \infty] \) is called a regular conditional density of \( \xi \) given \( \eta \) if it is \( \mathcal{B} \times \sigma(\eta) \)-measurable, for every fixed \( \omega \), \( f_{\xi|\eta}(x, \omega) \) is a probability density function in \( x \), and for all \( B \in \mathcal{B} \) and \( \omega \in \Omega \),
\[
Q_{\xi|\eta}(B, \omega) = \int_B f_{\xi|\eta}(x, \omega) \, dx.
\]

Similarly a function \( \hat{f}_{\xi|\eta}(x, y) \) defined on \( \mathbb{R}^2 \) to \( [0, \infty] \) is called a regular conditional density of \( \xi \) given \( \eta = y \) if it is \( \mathcal{B} \times \mathcal{B} \)-measurable, for every fixed \( y \), \( \hat{f}_{\xi|\eta}(x, y) \) is a probability density function in \( x \), and for all \( B \in \mathcal{B} \) and \( y \in \mathbb{R} \),
\[
\hat{Q}_{\xi|\eta}(B, y) = \int_B \hat{f}_{\xi|\eta}(x, y) \, dx.
\]

It is easy to see that \( f_{\xi|\eta} \) exists if and only if \( \hat{f}_{\xi|\eta} \) exists and that in this case they are related by
\[
f_{\xi|\eta}(x, \omega) = \hat{f}_{\xi|\eta}(x, \eta(\omega)) \quad \text{a.e.}\]
Conditioning

(with respect to the product of Lebesgue measure and $P$) (cf. Theorem 13.5.2). It is also clear (in view of Theorems 13.4.2 and 13.5.2) that conditional expectations can be expressed in terms of regular conditional densities, whenever the latter exist; for instance if $g$ is a Borel measurable function on the plane such that $E|g(\xi, \eta)| < \infty$ then we have the following:

$$E\{g(\xi, \eta) | \eta = y\} = \int_{-\infty}^{\infty} g(x, y) \hat{f}_{\xi|\eta}(x, y) \, dx \quad \text{a.s. (} P\eta^{-1})$$

$$E\{g(\xi, \eta) | \omega\} = \int_{-\infty}^{\infty} g(x, \eta(\omega)) \hat{f}_{\xi|\eta}(x, \omega) \, dx \quad \text{a.s.}$$

The following result shows that a regular conditional density exists if the r.v.’s $\xi$ and $\eta$ have a joint probability density function. If $f(x, y)$ is a joint p.d.f. of $\xi$ and $\eta$ (assumed defined and nonnegative everywhere) then the functions $f_{\xi}(x)$ and $f_{\eta}(y)$ defined for all $x$ and $y$ by

$$f_{\xi}(x) = \int_{-\infty}^{\infty} f(x, y) \, dy, \quad f_{\eta}(y) = \int_{-\infty}^{\infty} f(x, y) \, dx$$

are p.d.f.’s of $\xi, \eta$ respectively (Section 9.3).

**Theorem 13.6.1** Let $\xi$ and $\eta$ be r.v.’s with joint p.d.f. $f(x, y)$ and $f_{\eta}(y)$ defined as above. Then the function $\hat{f}(x, y)$ defined by

$$\hat{f}(x, y) = \begin{cases} f(x, y)/f_{\eta}(y) & \text{if } f_{\eta}(y) > 0 \\ h(x) & \text{if } f_{\eta}(y) = 0 \end{cases}$$

where $h(x)$ is an arbitrary but fixed p.d.f., is a regular conditional density of $\xi$ given $\eta = y$. Hence a regular conditional density of $\xi$ given $\eta$ is given by $f_{\xi|\eta}(x, \omega) = \hat{f}(x, \eta(\omega))$.

**Proof** Since $f$ is $B \times B$-measurable, it follows by Fubini’s Theorem that $f_{\eta}$ is $B$-measurable and hence $\hat{f}$ is $B \times B$-measurable.

From the definition of $\hat{f}$ it is clear that it is nonnegative and that for every fixed $y$, $\int_{-\infty}^{\infty} \hat{f}(x, y) \, dx = 1$. Hence for fixed $y$, $\hat{f}(x, y)$ is a p.d.f. in $x$.

Now define $\hat{Q}(B, y)$ for all $B \in B$ and $y \in \mathbb{R}$ by

$$\hat{Q}(B, y) = \int_{B} \hat{f}(x, y) \, dx.$$
\[ \hat{Q}(B, y) = P(\xi^{-1}B|\eta = y) \text{ a.s. (} P\eta^{-1} \text{).} \] Now for every fixed \( B \in \mathcal{B} \) and every \( E \in \mathcal{B} \) we have

\[
\int_E \hat{Q}(B, y) \ dP\eta^{-1}(y) = \int_{E \cap \{f_\eta(y) > 0\}} \int_B \hat{f}(x, y) \ dx \ dP\eta^{-1}(y) \\
= \int_{E \cap \{f_\eta(y) > 0\}} \int_B \hat{f}(x, y) f_\eta(y) \ dx \ dy \\
= P \left\{ \xi^{-1}B \cap \eta^{-1}(E \cap \{f_\eta(y) > 0\}) \right\} = P(\xi^{-1}B \cap \eta^{-1}E)
\]

since \( P\eta^{-1}\{f_\eta(y) = 0\} = 0 \). It follows that \( \hat{Q}(B, y) = P(\xi^{-1}B|\eta = y) \text{ a.s. (} P\eta^{-1} \text{)} \) and thus \( \hat{f}(x, y) \) is a regular conditional density of \( \xi \) given \( \eta = y \). \( \square \)

### 13.7 Summary

This is a summary of the main concepts defined in this chapter and their mutual relationships.

**I.** 1. \( \mathcal{E}(\xi|\mathcal{G}) \): conditional expectation of \( \xi \) given \( \mathcal{G} \)

2. \( P(A|\mathcal{G}) \): conditional probability of \( A \in \mathcal{F} \) given \( \mathcal{G} \)

Relationship: \( P(A|\mathcal{G}) = \mathcal{E}(\chi_A|\mathcal{G}). \)

**II.** 1. \( P_{\xi|\mathcal{G}}(A, \omega) \): regular conditional probability of \( \xi \) given \( \mathcal{G} \) (\( A \in \sigma(\xi) \)) (exists if \( \xi(\Omega) \in \mathcal{B} \))

2. \( Q_{\xi|\mathcal{G}}(B, \omega) \): regular conditional distribution of \( \xi \) given \( \mathcal{G} \) (\( B \in \mathcal{B} \)) (always exists)

Relationship, when they both exist:
For a.e. \( \omega \in \Omega \)

\[ Q_{\xi|\mathcal{G}}(B, \omega) = P_{\xi|\mathcal{G}}(\xi^{-1}B, \omega) \text{ for all } B \in \mathcal{B}. \]

If \( \mathcal{G} = \sigma(\eta) \) all concepts in I and II retain their name with “given \( \eta \)” replacing “given \( \mathcal{G} \)”.

**III.** 1. \( \mathcal{E}(\xi|\eta = y) \): conditional expectation of \( \xi \) given \( \eta = y \).

2. \( P(A|\eta = y) \): conditional probability of \( A \in \mathcal{F} \) given \( \eta = y \).

Relationship to I:

\[ \mathcal{E}(\xi|\eta = y) = f(y) \text{ a.e. (} P\eta^{-1} \text{) if and only if } \mathcal{E}(\xi|\eta) = f(\eta) \text{ a.s.} \]

\[ P(A|\eta = y) = f(y) \text{ a.e. (} P\eta^{-1} \text{) if and only if } P(A|\eta) = f(\eta) \text{ a.s.} \]

3. \( \hat{Q}_{\xi|\eta}(B, y) \): regular conditional distribution of \( \xi \) given \( \eta = y \) (\( B \in \mathcal{B} \)) (always exists)
Relationship to II:

\[ Q_{\xi|\eta}(B, \omega) = \hat{Q}_{\xi|\eta}(B, \eta(\omega)) \] for all \( B \in \mathcal{B}, \ \omega \notin N \in \sigma(\eta) \) with \( P(N) = 0 \).

**Exercises**

13.1 Let \( \xi \) be a r.v. with \( E|\xi| < \infty \) and \( \mathcal{G} \) a purely atomic sub-\( \sigma \)-field of \( \mathcal{F} \), i.e. \( \mathcal{G} \) is generated by the disjoint events \( \{E_0, E_1, E_2, \ldots \} \) with \( P(E_0) = 0, \ P(E_n) > 0 \) for \( n = 1, 2, \ldots \) and \( \Omega = \bigcup_{n \geq 0} E_n \). Using the definition of \( E(\xi|\mathcal{G}) \) given in Section 13.2 show that

\[
E(\xi|\mathcal{G}) = \sum_{n \geq 1} X_{E_n} \frac{1}{P(E_n)} \int_{E_n} \xi \, dP \text{ a.s.}
\]

(Hint: Show first that every set \( E \) in \( \mathcal{G} \) is the union of a subsequence of \( \{E_n, n \geq 0\} \).)

13.2 If the r.v.'s \( \xi \) and \( \eta \) are such that \( E|\xi| < \infty \) and \( \eta \) is bounded then show that

\[
E[E(\xi|\mathcal{G})\eta] = E[E(\xi|\mathcal{G})\eta] = E[E(\xi|\mathcal{G})E(\eta|\mathcal{G})].
\]

13.3 Let \( \xi, \eta, \zeta \) be r.v.'s with \( E|\xi| < \infty \) and \( \eta \) independent of the pair \( \xi, \zeta \). Show that

\[
E(\xi|\eta, \zeta) = E(\xi|\zeta) \text{ a.s.}
\]

Show also that if \( \xi \) is a Borel measurable function of \( \eta \) and \( \zeta \) (\( \xi = f(\eta, \zeta) \)) then it is a Borel measurable function of \( \zeta \) only (\( \xi = g(\zeta) \)).

13.4 State and prove the conditional form of the H"older and Minkowski Inequalities.

13.5 If \( \xi \in L_p(\Omega, \mathcal{F}, P), \ p \geq 1 \), show that \( E(\xi|\mathcal{G}) \in L_p(\Omega, \mathcal{F}, P) \) and

\[
||E(\xi|\mathcal{G})||_p = E^{1/p}[||E(\xi|\mathcal{G})||^p] \leq E^{1/p}(|\xi|^p) = ||\xi||_p.
\]

(Hint: Use the Conditional Jensen’s Inequality (Theorem 13.2.9).)

13.6 Two r.v.'s \( \xi \) and \( \eta \) in \( L_2(\Omega, \mathcal{F}, P) \) are called orthogonal if \( E(\xi|\eta) = 0 \). Let \( \xi \in L_2(\Omega, \mathcal{F}, P) \); then \( E(\xi|\mathcal{G}) \in L_2(\Omega, \mathcal{F}, P) \) by Ex. 13.5. Show that \( E(\xi|\mathcal{G}) \) is the unique r.v. \( \eta \in L_2(\Omega, \mathcal{G}, P_G) \) which minimizes \( E(\xi - \eta)^2 \) and that the minimum value is

\[
E(\xi^2) = E(\xi^2|\mathcal{G})\].

\( E(\xi|\mathcal{G}) \) is called the (in general, nonlinear) mean square estimate of \( \xi \) based on \( \mathcal{G} \). (Hint: Show that \( \xi - E(\xi|\mathcal{G}) \) is orthogonal to all r.v.'s in \( L_2(\Omega, \mathcal{G}, P_G) \), so that \( E(\xi|\mathcal{G}) \) is the projection of \( \xi \) onto \( L_2(\Omega, \mathcal{G}, P_G) \), and that for every \( \eta \in L_2(\Omega, \mathcal{G}, P_G) \), \( E(\xi - \eta)^2 = E(\xi - E(\xi|\mathcal{G}))^2 + E(\eta - E(\xi|\mathcal{G}))^2 \).

In particular, if \( \eta \) is a r.v., then \( E(\xi|\eta) \) is the unique r.v. \( \zeta \in L_2(\Omega, \sigma(\eta), P_{\sigma(\eta)}) \) which minimizes \( E(\xi - \zeta)^2 \), or equivalently \( h(\eta) = E(\xi|\eta) \) is the unique function \( g \in L_2(\mathbb{R}, \mathcal{B}, P\eta^{-1}) \) which minimizes \( E(|\xi - g(\eta)|^2) \). \( E(\xi|\eta) \) is called
the (in general, nonlinear) mean square estimate or least square regression of \( \xi \) based on \( \eta \). It follows from Ex. 13.12 that if \( \xi \) and \( \eta \) have a joint normal distribution then \( \mathbb{E}(\xi|\eta) = a + b\eta \) a.s. and thus the least squares regression of \( \xi \) based on \( \eta \) is linear.

13.7 Prove the conditional form of Jensen’s Inequality (Theorem 13.2.9) by using regular conditional distributions and the ordinary form of Jensen’s Inequality (Theorem 9.5.4).

13.8 Let \( \xi \) and \( \eta \) be independent r.v.’s. Show that for every Borel set \( B \),

\[
P(\xi + \eta \in B|\eta)(\omega) = P_{\xi|\eta}^{-1}(B - \eta(\omega)) \quad \text{a.s.}
\]

where \( B - y = \{x : x + y \in B\} \). What is then \( P(\xi + \eta \in B|\eta = y) \) equal to? Show also that

\[
Q_{\xi + \eta}(B, \omega) = P_{\xi|\eta}^{-1}\{B - \eta(\omega)\}
\]

is a regular conditional distribution of \( \xi + \eta \) given \( \eta \).

13.9 Let \( \mathcal{G} \) be a sub-\( \sigma \)-field of \( \mathcal{F} \). We say that a family of classes of events \( \{A_\lambda, \lambda \in \Lambda\} \) is conditionally independent given \( \mathcal{G} \) if

\[
P\left(\bigcap_{k=1}^n A_{\lambda_k}|\mathcal{G}\right) = \prod_{k=1}^n P(A_{\lambda_k}|\mathcal{G}) \quad \text{a.s.}
\]

for any \( n \), any \( \lambda_1, \ldots, \lambda_n \in \Lambda \) and any \( A_{\lambda_k} \in \mathcal{A}_{\lambda_k}, \ k = 1, \ldots, n \). Generalize the Kolmogorov Zero-One Law to conditional independence: if \( \{\xi_n\}_{n=1}^\infty \) is a sequence of conditionally independent r.v.’s given \( \mathcal{G} \) and \( A \) is a tail event, show that

\[
P(A|\mathcal{G}) = 0 \text{ or } 1 \quad \text{a.s.,}
\]

and if \( \xi \) is a tail r.v., show that \( \xi = \eta \) a.s. for some \( \mathcal{G} \)-measurable r.v. \( \eta \).

13.10 Let \( \xi \) and \( \eta \) be r.v.’s with \( \mathbb{E}|\xi| < \infty \). If \( y \in \mathbb{R} \) is such that \( P(\eta = y) > 0 \) then show that \( \mathbb{E}(\xi|\eta = y) \) as defined in Section 13.5 is given by

\[
\mathbb{E}(\xi|\eta = y) = \frac{1}{P(\eta = y)} \int_{\{\eta = y\}} \xi \, dP.
\]

(Hint: Let \( D \) be the at most countable set of points \( y \in \mathbb{R} \) such that \( P(\eta = y) > 0 \). Define \( f : \mathbb{R} \to \mathbb{R} \) by \( f(y) = \frac{1}{P(\eta = y)} \int_{\{\eta = y\}} \xi \, dP \) if \( y \in D \) and \( f(y) = \mathbb{E}(\xi|\eta = y) \) if \( y \notin D \), and show that for all Borel sets \( B \), \( \int_B f \, dP \eta^{-1} = \int_{\eta^{-1}B} \xi \, dP \).

13.11 Let \( \xi \) be a r.v. and \( \eta \) a discrete r.v. with values \( y_1, y_2, \ldots \). Find expressions for the regular conditional probability of \( \xi \) given \( \eta \) and for the regular conditional distribution of \( \xi \) given \( \eta \) and given \( \eta = y \). Simplify further these expressions when \( \xi \) is discrete with values \( x_1, x_2, \ldots \).
13.12 Let the r.v.’s $\xi_1$ and $\xi_2$ have a joint normal distribution with $E(\xi_i) = \mu_i$, \( \text{var}(\xi_i) = \sigma_i^2 > 0, i = 1, 2, \) and $E[(\xi_1 - \mu_1)(\xi_2 - \mu_2)] = \rho \sigma_1 \sigma_2$, $|\rho| < 1$, i.e. $\xi_1$ and $\xi_2$ have the joint p.d.f.

\[
\frac{1}{2\pi\sigma_1\sigma_2 \sqrt{1-\rho^2}} \times \exp \left\{ -\frac{1}{2(1-\rho^2)} \left[ \frac{(x_1 - \mu_1)^2}{\sigma_1^2} - \frac{2\rho(x_1 - \mu_1)(x_2 - \mu_2)}{\sigma_1 \sigma_2} + \frac{(x_2 - \mu_2)^2}{\sigma_2^2} \right] \right\}
\]

Find the regular conditional density of $\xi_1$ given $\xi_2 = x_2$ and show that $E(\xi_1|\xi_2) = \mu_1 + \rho \frac{\sigma_1}{\sigma_2} (\xi_2 - \mu_2)$ a.s.

(What happens when $|\rho| = 1$?)

13.13 Let the r.v.’s $\xi$ and $\eta$ be such that $\xi$ has a uniform distribution on $[0, 1]$ and the (regular) conditional distribution of $\eta$ given $\xi = x, x \in [0, 1]$, is uniform on $[-x, x]$. Find the regular conditional densities of $\xi$ given $\eta = y$ and of $\eta$ given $\xi = x$, and the conditional expectations $E(\xi + \eta|\xi)$ and $E(\xi + \eta|\eta)$. 
Martingales

14.1 Definition and basic properties

In this chapter we consider the notion of a martingale sequence, which has many of the useful properties of a sequence of partial sums of independent r.v.’s (with zero means) and which forms the basis of a significant segment of basic probability theory.

As usual, \((\Omega, \mathcal{F}, P)\) will denote a fixed probability space. Let \(\{\xi_n\}\) be a sequence of r.v.’s and \(\{\mathcal{F}_n\}\) a sequence of sub-\(\sigma\)-fields of \(\mathcal{F}\). Where nothing else is specified in writing sequences such as \(\{\xi_n\}, \{\mathcal{F}_n\}\) etc. it will be assumed that the range of \(n\) is the set of positive integers \(\{1, 2, \ldots\}\). We say that \(\{\xi_n, \mathcal{F}_n\}\) is a martingale (respectively, a submartingale, a supermartingale) if for every \(n\),

(i) \(\mathcal{F}_n \subset \mathcal{F}_{n+1}\)

(ii) \(\xi_n\) is \(\mathcal{F}_n\)-measurable and integrable

(iii) \(E(\xi_{n+1}|\mathcal{F}_n) = \xi_n\) (resp. \(\geq \xi_n, \leq \xi_n\)) a.s.

This definition trivially contains the notion of \(\{\xi_n, \mathcal{F}_n, 1 \leq n \leq N\}\) being a martingale (respectively, a submartingale, a supermartingale); just take \(\xi_n = \xi_{N}\) and \(\mathcal{F}_n = \mathcal{F}_{N}\) for all \(n > N\). Clearly \(\{\xi_n, \mathcal{F}_n\}\) is a submartingale if and only if \(\{-\xi_n, \mathcal{F}_n\}\) is a supermartingale. Thus the properties of supermartingales can be obtained from those of submartingales and in the sequel only martingales and submartingales will typically be considered.

**Example 1** Let \(\{\xi_n\}\) be a sequence of independent r.v.’s in \(L_1\) with zero means and let

\[
S_n = \xi_1 + \cdots + \xi_n, \quad \mathcal{F}_n = \sigma(\xi_1, \ldots, \xi_n), \quad n = 1, 2, \ldots.
\]
Martingales

Then \( \{ S_n, F_n \} \) is a martingale since for every \( n \), \( S_n \) is clearly \( F_n \)-measurable and integrable, and

\[
\mathcal{E}(S_{n+1}|F_n) = \mathcal{E}(\xi_{n+1} + S_n|F_n) \\
= \mathcal{E}(\xi_{n+1}|F_n) + \mathcal{E}(S_n|F_n) \\
= \mathcal{E}\xi_{n+1} + S_n = S_n \text{ a.s.}
\]

since \( S_n \) is \( F_n \)-measurable, \( \sigma(\xi_{n+1}) \) and \( F_n \) are independent, and \( \mathcal{E}\xi_{n+1} = 0 \).

**Example 2** Let \( \{ \xi_n \} \) be a sequence of independent r.v.’s in \( L_1 \) with finite, nonzero means \( \mathcal{E}\xi_n = \mu_n \), and let

\[
\eta_n = \prod_{k=1}^{n} \frac{\xi_k}{\mu_k}, \quad F_n = \sigma(\xi_1, \ldots, \xi_n), \quad n = 1, 2, \ldots
\]

Then \( \{ \eta_n, F_n \} \) is a martingale since for every \( n \), \( \eta_n \) is clearly \( F_n \)-measurable and integrable, and

\[
\mathcal{E}(\eta_{n+1}|F_n) = \mathcal{E}\left( \frac{\xi_{n+1}}{\mu_{n+1}} \eta_n | F_n \right) = \eta_n \mathcal{E}\left( \frac{\xi_{n+1}}{\mu_{n+1}} | F_n \right) \\
= \eta_n \mathcal{E}\frac{\xi_{n+1}}{\mu_{n+1}} = \eta_n \text{ a.s.}
\]

since \( \eta_n \) is \( F_n \)-measurable, and \( \sigma(\xi_{n+1}) \) and \( F_n \) are independent.

**Example 3** Let \( \xi \) be an integrable r.v. and \( \{ F_n \} \) an increasing sequence of sub-\( \sigma \)-fields of \( F \) (i.e. \( F_n \subset F_{n+1}, \quad n = 1, 2, \ldots \)). Let

\[
\xi_n = \mathcal{E}(\xi|F_n) \text{ for } n = 1, 2, \ldots
\]

Then \( \{ \xi_n, F_n \} \) is a martingale since for each \( n \), \( \xi_n \) is \( F_n \)-measurable and integrable, and

\[
\mathcal{E}(\xi_{n+1}|F_n) = \mathcal{E}\{ \mathcal{E}(\xi|F_{n+1}) | F_n \} \\
= \mathcal{E}(\xi|F_n) = \xi_n \text{ a.s.}
\]

by Theorem 13.2.2 since \( F_n \subset F_{n+1} \). It will be shown in Section 14.3 that a martingale \( \{ \xi_n, F_n \} \) is of this type, i.e. \( \xi_n = \mathcal{E}(\xi|F_n) \) for some \( \xi \in L_1 \), if and only if the sequence \( \{ \xi_n \} \) is uniformly integrable.

The following results contain the simplest properties of martingales.

**Theorem 14.1.1** (i) If \( \{ \xi_n, F_n \} \) and \( \{ \eta_n, F_n \} \) are two martingales (resp. submartingales, supermartingales) then for any real numbers \( a \) and \( b \) (resp. nonnegative numbers \( a \) and \( b \)), \( \{ a\xi_n + b\eta_n, F_n \} \) is a martingale (resp. submartingale, supermartingale).
14.1 Definition and basic properties

(ii) If \( \{\xi_n, \mathcal{F}_n\} \) is a martingale (resp. submartingale, supermartingale) then
the sequence \( \{\mathbb{E}\xi_n\} \) is constant (resp. nondecreasing, nonincreasing).

(iii) Let \( \{\xi_n, \mathcal{F}_n\} \) be a submartingale (resp. supermartingale). Then \( \{\xi_n, \mathcal{F}_n\} \)
is a martingale if and only if the sequence \( \{\mathbb{E}\xi_n\} \) is constant.

Proof  (i) is obvious from the linearity of conditional expectation (Theo-
rem 13.2.1 (ii)).

(ii) If \( \{\xi_n, \mathcal{F}_n\} \) is a martingale we have for every \( n = 1, 2, \ldots \),
\( \mathbb{E}(\xi_{n+1}|\mathcal{F}_n) = \xi_n \) a.s. and thus
\[ \mathbb{E}\xi_{n+1} = \mathbb{E}(\mathbb{E}(\xi_{n+1}|\mathcal{F}_n)) = \mathbb{E}\xi_n. \]
Similarly for a sub- and supermartingale.

(iii) The “only if” part follows from (ii). For the “if” part assume that
\( \{\xi_n, \mathcal{F}_n\} \) is a submartingale and that \( \{\mathbb{E}\xi_n\} \) is constant. Then for all \( n \),
\[ \mathbb{E}(\mathbb{E}(\xi_{n+1}|\mathcal{F}_n) - \xi_n) = \mathbb{E}\xi_{n+1} - \mathbb{E}\xi_n = 0 \]
and since \( \mathbb{E}(\xi_{n+1}|\mathcal{F}_n) - \xi_n \geq 0 \) a.s. (from the definition of a submartingale)
and \( \mathbb{E}(\xi_{n+1}|\mathcal{F}_n) - \xi_n \in L_1 \), it follows (Theorem 4.4.7) that
\[ \mathbb{E}(\xi_{n+1}|\mathcal{F}_n) - \xi_n = 0 \text{ a.s.} \]
Hence \( \{\xi_n, \mathcal{F}_n\} \) is a martingale. \( \square \)

The next theorem shows that any martingale is also a martingale relative
to \( \sigma(\xi_1, \ldots, \xi_n) \), and extends property (iii) of the martingale (submartin-
gale, supermartingale) definitions.

**Theorem 14.1.2**  If \( \{\xi_n, \mathcal{F}_n\} \) is a martingale, then so is \( \{\xi_n, \sigma(\xi_1, \ldots, \xi_n)\} \)
and for all \( n, k = 1, 2, \ldots \)
\[ \mathbb{E}(\xi_{n+k}|\mathcal{F}_n) = \xi_n \text{ a.s.} \]
with corresponding statements for sub- and supermartingales.

Proof  If \( \{\xi_n, \mathcal{F}_n\} \) is a martingale, since for every \( n, \xi_n \) is \( \mathcal{F}_n \)-measurable
and \( \mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots \subset \mathcal{F}_n \), we have
\[ \sigma(\xi_1, \ldots, \xi_n) \subset \mathcal{F}_n. \]

It follows from Theorem 13.2.2, and Theorem 13.2.1 (v) that
\[ \mathbb{E}(\xi_{n+1}|\sigma(\xi_1, \ldots, \xi_n)) = \mathbb{E}(\mathbb{E}(\xi_{n+1}|\mathcal{F}_n)|\sigma(\xi_1, \ldots, \xi_n)) \]
\[ = \mathbb{E}(\xi_n|\sigma(\xi_1, \ldots, \xi_n)) \]
\[ = \xi_n \text{ a.s.} \]
so that \( \{\xi_n, \sigma(\xi_1, \ldots, \xi_n)\} \) is indeed a martingale.
The equality $\mathcal{E}(\xi_{n+k}|\mathcal{F}_n) = \xi_n$ a.s. holds for $k = 1$ and all $n$ by the definition of a martingale. If it holds for some $k$ and all $n$, then it also holds for $k + 1$ and all $n$ since

$$\mathcal{E}(\xi_{n+k+1}|\mathcal{F}_n) = \mathcal{E}([\mathcal{E}(\xi_{n+k+1}|\mathcal{F}_{n+k})|\mathcal{F}_n])$$

$$= \mathcal{E}(\xi_{n+k}|\mathcal{F}_n) = \xi_n$$

by Theorem 13.2.2 ($\mathcal{F}_n \subset \mathcal{F}_{n+k}$), the definition of a martingale, and the inductive hypothesis. The result thus follows for all $n$ and $k$.

The corresponding statements for submartingales and supermartingales follow with the obvious changes. □

The following properties follow immediately from this theorem.

**Corollary**

(i) If $\{\xi_n, \mathcal{F}_n\}$ is a submartingale, so is $\{\xi_{n+}, \mathcal{F}_n\}$ (where $\xi_+ = \xi$ for $\xi \geq 0$ and $\xi_+ = 0$ for $\xi < 0$).

(ii) If $\{\xi_n, \mathcal{F}_n\}$ is a martingale then $\{|\xi_n|, \mathcal{F}_n\}$ is a submartingale, and so is $\{|\xi_n|^p, \mathcal{F}_n\}$, $1 < p < \infty$, provided $\xi_n \in L_p$ for all $n$. 

---

**Theorem 14.1.3** Let $\{\xi_n, \mathcal{F}_n\}$ be a martingale (resp. a submartingale) and $g$ a convex (resp. a convex nondecreasing) function on the real line. If $g(\xi_n)$ is integrable for all $n$, then $\{g(\xi_n), \mathcal{F}_n\}$ is a submartingale.

**Proof** Since $g$ is Borel measurable, $g(\xi_n)$ is $\mathcal{F}_n$-measurable for all $n$. Also, since $g$ is convex and $\xi_n$, $g(\xi_n)$ are integrable, Theorem 13.2.9 gives

$$g(\mathcal{E}(\xi_{n+1}|\mathcal{F}_n)) \leq \mathcal{E}(g(\xi_{n+1})|\mathcal{F}_n)$$

for all $n$. If $\{\xi_n, \mathcal{F}_n\}$ is a martingale then $\mathcal{E}(\xi_{n+1}|\mathcal{F}_n) = \xi_n$ a.s. and thus

$$g(\xi_n) \leq \mathcal{E}(g(\xi_{n+1})|\mathcal{F}_n)$$

which shows that $\{g(\xi_n), \mathcal{F}_n\}$ is a submartingale. If $\{\xi_n, \mathcal{F}_n\}$ is a submartingale then $\mathcal{E}(\xi_{n+1}|\mathcal{F}_n) \geq \xi_n$ a.s. and if $g$ is nondecreasing we have

$$g(\xi_n) \leq g(\mathcal{E}(\xi_{n+1}|\mathcal{F}_n)) \leq \mathcal{E}(g(\xi_{n+1})|\mathcal{F}_n)$$

which again shows that $\{g(\xi_n), \mathcal{F}_n\}$ is a submartingale. □
A connection between martingales and submartingales is given in the following.

**Theorem 14.1.4 (Doob’s Decomposition)** Every submartingale \( \{\xi_n, F_n\} \) can be uniquely decomposed as

\[
\xi_n = \eta_n + \zeta_n \text{ for all } n, \text{ a.s.}
\]

where \( \{\eta_n, F_n\} \) is a martingale and the sequence of r.v.'s \( \{\zeta_n\} \) is such that

- \( \zeta_1 = 0 \) a.s.
- \( \zeta_n \leq \zeta_{n+1} \) for all \( n \) a.s.
- \( \zeta_{n+1} \) is \( F_n \)-measurable for all \( n \).
- \( \{\zeta_n\} \) is called the predictable increasing sequence\(^1\) associated with the submartingale \( \{\xi_n\} \).

**Proof** Define

\[
\eta_1 = \xi_1, \quad \zeta_1 = 0
\]

and for \( n \geq 2 \)

\[
\eta_n = \xi_1 + \sum_{k=2}^{n} (\xi_k - \mathbb{E}(\xi_k|F_{k-1})) \quad \zeta_n = \sum_{k=2}^{n} \{\mathbb{E}(\xi_k|F_{k-1}) - \xi_{k-1}\}
\]

or equivalently

\[
\eta_n = \eta_{n-1} + \xi_n - \mathbb{E}(\xi_n|F_{n-1}), \quad \zeta_n = \zeta_{n-1} + \mathbb{E}(\xi_n|F_{n-1}) - \xi_{n-1}.
\]

Then \( \eta_1 + \zeta_1 = \xi_1 \) and for all \( n \geq 2 \)

\[
\eta_n + \zeta_n = \xi_1 + \sum_{k=2}^{n} \xi_k - \sum_{k=2}^{n} \xi_{k-1} = \xi_n \text{ a.s.}
\]

Now \( \{\eta_n, F_n\} \) is a martingale, since for all \( n \), \( \eta_n \) is clearly \( F_n \)-measurable and integrable and

\[
\mathbb{E}(\eta_{n+1}|F_n) = \mathbb{E}\{\eta_n + \xi_{n+1} - \mathbb{E}(\xi_{n+1}|F_n)|F_n\}
\]

\[
= \eta_n + \mathbb{E}(\xi_{n+1}|F_n) - \mathbb{E}(\xi_{n+1}|F_n)
\]

\[
= \eta_n \text{ a.s.}
\]

Also, \( \zeta_1 = 0 \) by definition, and for all \( n \), \( \zeta_{n+1} \) is clearly \( F_n \)-measurable and integrable, and the submartingale property \( \mathbb{E}(\xi_{n+1}|F_n) \geq \xi_n \) a.s. implies that

\[
\zeta_{n+1} = \zeta_n + \mathbb{E}(\xi_{n+1}|F_n) - \xi_n \geq \zeta_n \text{ a.s.}
\]

Thus \( \{\zeta_n\} \) has the stated properties.

\(^1\) This terminology is most evident when e.g. \( F_n = \sigma(\xi_1, \ldots, \xi_n) \) so that \( \xi_{n+1} \in F_n \) implies that \( \xi_{n+1} \) may be written as a function of \( (\xi_1, \ldots, \xi_n) \) so is “predictable” from these values.
The uniqueness of the decomposition is shown as follows. Let \( \xi_n = \eta'_n + \zeta'_n \) be another decomposition with \( \{\eta'_n\} \) and \( \{\zeta'_n\} \) having the same properties as \( \{\eta_n\} \) and \( \{\zeta_n\} \). Then for all \( n \),

\[
\eta_n - \eta'_n = \zeta_n - \zeta'_n = \theta_n,
\]
say. Since \( \{\eta_n, \mathcal{F}_n\} \) and \( \{\eta'_n, \mathcal{F}_n\} \) are martingales, so is \( \{\theta_n, \mathcal{F}_n\} \) so that

\[
\mathcal{E}(\theta_{n+1}|\mathcal{F}_n) = \theta_n \text{ for all } n \text{ a.s.}
\]

Also, since \( \zeta_{n+1} \) and \( \zeta'_{n+1} \) are \( \mathcal{F}_n \)-measurable, so is \( \theta_{n+1} \) and thus

\[
\mathcal{E}(\theta_{n+1}|\mathcal{F}_n) = \theta_{n+1} \text{ for all } n \text{ a.s.}
\]

It follows that \( \theta_1 = \cdots = \theta_n = \theta_{n+1} = \cdots \text{ a.s.} \) and since \( \theta_1 = 0 \) a.s. we have \( \theta_n = 0 \) for all \( n \) a.s. and thus

\[
\eta'_n = \eta_n \text{ and } \zeta'_n = \zeta_n \text{ for all } n \text{ a.s.} \quad \square
\]

### 14.2 Inequalities

There are a number of basic and useful inequalities for probabilities, moments and “crossings” of submartingales, and the simpler of these are given in this section. The first provides a martingale form of Kolmogorov’s Inequality (Theorem 11.5.1).

**Theorem 14.2.1** If \( \{(\xi_n, \mathcal{F}_n): 1 \leq n \leq N\} \) is a submartingale, then for all real \( a \)

\[
aP\{\max_{1 \leq n \leq N} \xi_n \geq a\} \leq \int_{\{\max_{1 \leq n \leq N} \xi_n \geq a\}} \xi_N \, dP \leq \mathcal{E}|\xi_N|.
\]

**Proof** Define (as in the proof of Theorem 11.5.1)

\[
E = \{\omega: \max_{1 \leq n \leq N} \xi_n(\omega) \geq a\}
\]

\[
E_1 = \{\omega: \xi_1(\omega) \geq a\}
\]

\[
E_n = \{\omega: \xi_n(\omega) \geq a\} \cap \bigcap_{k=1}^{n-1} \{\omega: \xi_k(\omega) < a\}, \quad n = 2, \ldots, N.
\]

Then \( E_n \in \mathcal{F}_n \) for all \( n = 1, \ldots, N \), \( \{E_n\} \) are disjoint and \( E = \bigcup_{n=1}^N E_n \). Thus

\[
\int_{E} \xi_N \, dP = \sum_{n=1}^{N} \int_{E_n} \xi_N \, dP.
\]

Now for each \( n = 1, \ldots, N \),

\[
\int_{E_n} \xi_N \, dP = \int_{E_n} \mathcal{E}(\xi_N|\mathcal{F}_n) \, dP \geq \int_{E_n} \xi_n \, dP \geq aP(E_n)
\]
since \( E_n \in \mathcal{F}_n \), \( \mathcal{E}(\xi_N|\mathcal{F}_n) \geq \xi_n \) by Theorem 14.1.2, and \( \xi_n \geq a \) on \( E_n \). It follows that
\[
\int_{E} \xi_N dP \geq a \sum_{n=1}^{N} P(E_n) = aP(E).
\]
This proves the left half of the inequality of the theorem and the right half is obvious.

That Theorem 14.2.1 contains Kolmogorov’s Inequality (Theorem 11.5.1) follows from Example 1 and the following corollary.

**Corollary**  Let \( \{(\xi_n, \mathcal{F}_n) : 1 \leq n \leq N\} \) be a martingale and \( a > 0 \). Then

(i) \( P\{\max_{1 \leq n \leq N} |\xi_n| \geq a\} \leq \frac{1}{a} \int_{\{\max_{1 \leq n \leq N} |\xi_n| \geq a\}} |\xi_N| dP \leq \mathcal{E}|\xi_N|/a.\)

(ii) If also \( \mathcal{E}\xi_N^2 < \infty \), then

\[
P\{\max_{1 \leq n \leq N} |\xi_n| \geq a\} \leq \mathcal{E}\xi_N^2/a^2.
\]

**Proof**  Since \( \{(\xi_n, \mathcal{F}_n) : 1 \leq n \leq N\} \) is a martingale, \( \{(|\xi_n|, \mathcal{F}_n) : 1 \leq n \leq N\} \) is a submartingale ((ii) of Theorem 14.1.3, Corollary) and (i) follows from the theorem.

For (ii) we will show that \( \mathcal{E}\xi_N^2 < \infty \) implies \( \mathcal{E}\xi_n^2 < \infty \) for all \( n = 1, \ldots, N \). Then by part (ii) of the corollary to Theorem 14.1.3, \( \{(\xi_n^2, \mathcal{F}_n) : 1 \leq n \leq N\} \) is a submartingale and (ii) follows from the theorem.

To show that if \( \{(\xi_n, \mathcal{F}_n) : 1 \leq n \leq N\} \) is a martingale and \( \mathcal{E}\xi_N^2 < \infty \), then \( \mathcal{E}\xi_n^2 < \infty \) for all \( n = 1, \ldots, N \), we define \( g_k \) on the real line for each \( k = 1, 2, \ldots \), by

\[
g_k(x) = \begin{cases} 
2k(|x| - k/2) & \text{for } |x| > k \\
x^2 & \text{for } |x| \leq k 
\end{cases}
\]

Then each \( g_k \) is convex and \( g_k(x) \uparrow x^2 \) for all real \( x \). For each fixed \( k = 1, 2, \ldots \), since for all \( n = 1, \ldots, N \),

\[
\mathcal{E}|g_k(\xi_n)| = \int_{|\xi_n| \leq k} g_k^2 dP + \int_{|\xi_n| > k} 2k(|\xi_n| - k/2) dP \\
\leq k^2 + 2k\mathcal{E}|\xi_n| < \infty,
\]

it follows from Theorem 14.1.3 that \( \{(g_k(\xi_n), \mathcal{F}_n) : 1 \leq n \leq N\} \) is a submartingale and thus, by Theorem 14.1.1 (ii),

\[
0 \leq \mathcal{E}(g_k(\xi_1)) \leq \ldots \leq \mathcal{E}(g_k(\xi_N)) < \infty.
\]
Since $g_k(x) \uparrow x^2$ for each $x$ as $k \to \infty$, the monotone convergence theorem implies that for each $n = 1, \ldots, N$, $\mathcal{E}\{g_k(\xi_n)\} \uparrow \mathcal{E}\xi_n^2$. Hence we have

$$0 \leq \mathcal{E}\xi_1^2 \leq \ldots \leq \mathcal{E}\xi_N^2$$

and the result follows since $\mathcal{E}\xi_N^2 < \infty$.

As a consequence of Theorem 14.2.1, the following inequality holds for nonnegative submartingales.

**Theorem 14.2.2** If $\{(\xi_n, \mathcal{F}_n) : 1 \leq n \leq N\}$ is a submartingale such that $\xi_n \geq 0$ a.s. $n = 1, \ldots, N$, then for all $p > 1$,

$$\mathcal{E}(\max_{1 \leq n \leq N} \xi_n^p) \leq \left( \frac{p}{p-1} \right)^p \mathcal{E}\xi_N^p.$$

**Proof** Define $\zeta = \max_{1 \leq n \leq N} \xi_n$ and $\eta = \xi_N$. Then $\zeta, \eta \geq 0$ a.s. and it follows from Theorem 14.2.1 that for all $x > 0$,

$$G(x) = P\{\zeta > x\} \leq \frac{1}{x} \int_{\{\zeta \geq x\}} \eta dP.$$

Now by applying the monotone convergence theorem and Fubini’s Theorem (i.e. integration by parts) we obtain

$$\mathcal{E}(\zeta^p) = \int_0^\infty x^p d\{1 - G(x)\} = \int_0^\infty x^p d\{-G(x)\}$$

$$= \lim_{A \to \infty} \int_0^A x^p d\{-G(x)\}$$

$$= \lim_{A \to \infty} \{p \int_0^A x^{p-1} G(x) dx - A^{p} G(A)\}$$

$$\leq \lim_{A \to \infty} \{p \int_0^A x^{p-1} G(x) dx = p \int_0^\infty x^{p-1} G(x) dx\}$$

$$\leq p \int_0^\infty x^{p-1} \frac{1}{x} \left( \int_{\{\zeta \geq x\}} \eta dP \right) dx$$

by the inequality for $G$ shown above. Change of integration order thus gives

$$\mathcal{E}(\zeta^p) \leq p \int_\Omega \eta(\omega) \left( \int_0^{\zeta(\omega)} x^{p-2} dx \right) dP(\omega)$$

$$= \frac{p}{p-1} \int_\Omega \eta(\omega) \zeta^{p-1}(\omega) dP(\omega) = \frac{p}{p-1} \mathcal{E}(\eta \zeta^{p-1})$$

$$\leq \frac{p}{p-1} \mathcal{E}(\eta^p) \mathcal{E}(\zeta^{p-1}),$$

by Hölder’s Inequality. It follows that $\mathcal{E}(\zeta^p) \leq \frac{p}{p-1} \mathcal{E}(\eta^p)$ which implies the result.

The following corollary follows immediately from the theorem and (ii) of Theorem 14.1.3, Corollary.
Corollary  If \( \{ (\xi_n, \mathcal{F}_n) : 1 \leq n \leq N \} \) is a martingale and \( p > 1 \), then

\[
\mathcal{E}(\max_{1 \leq n \leq N} |\xi_n|^p) \leq \left( \frac{p}{p - 1} \right)^p \mathcal{E}|\xi_N|^p.
\]

The final result of this section is an inequality for the number of “upcrossings” of a submartingale, which will be pivotal in the next section in deriving the submartingale convergence theorem. This requires the following definitions and notation. Let \( \{ x_1, \ldots, x_N \} \) be a finite sequence of real numbers and let \( a < b \) be real numbers. Let \( \tau_1 \) be the first integer in \( \{ 1, \ldots, N \} \) such that \( x_{\tau_1} \leq a \), \( \tau_2 \) be the first integer in \( \{ 1, \ldots, N \} \) larger than \( \tau_1 \) such that \( x_{\tau_2} \geq b \), \( \tau_3 \) be the first integer in \( \{ 1, \ldots, N \} \) larger than \( \tau_2 \) such that \( x_{\tau_3} \leq a \), \( \tau_4 \) be the first integer in \( \{ 1, \ldots, N \} \) larger than \( \tau_3 \) such that \( x_{\tau_4} \geq b \), and so on, and define \( \tau_i = N + 1 \) if the condition cannot be satisfied. In other words,

\[
\begin{align*}
\tau_1 &= \min\{ j : 1 \leq j \leq N, \ x_j \leq a \}, \\
\tau_2 &= \min\{ j : \tau_1 < j \leq N, \ x_j \geq b \}, \\
\tau_{2k+1} &= \min\{ j : \tau_{2k} < j \leq N, \ x_j \leq a \}, \quad 3 \leq 2k + 1 \leq N \\
\tau_{2k+2} &= \min\{ j : \tau_{2k+1} < j \leq N, \ x_j \geq b \}, \quad 4 \leq 2k + 2 \leq N
\end{align*}
\]

and \( \tau_i = N + 1 \) if the corresponding set is empty. Let \( M \) be the number of \( \tau_i \) that do not exceed \( N \). Then the number of upcrossings \( U_{[a,b]} \) of the interval \( [a, b] \) by the sequence \( \{ x_1, \ldots, x_N \} \) is defined by

\[
U_{[a,b]} = \lfloor M/2 \rfloor = \begin{cases} 
M/2 & \text{if } M \text{ is even} \\ 
(M - 1)/2 & \text{if } M \text{ is odd}
\end{cases}
\]

and is the number of times the sequence (completely) crosses from \( \leq a \) to \( \geq b \).

Theorem 14.2.3 Let \( \{ (\xi_n, \mathcal{F}_n) : 1 \leq n \leq N \} \) be a submartingale, \( a < b \) real numbers, and let \( U_{[a,b]}(\omega) \) be the number of upcrossings of the interval \( [a, b] \) by the sequence \( \{ \xi_1(\omega), \ldots, \xi_N(\omega) \} \). Then

\[
\mathcal{E}U_{[a,b]} \leq \frac{\mathcal{E}(\xi_N - a)_+ - \mathcal{E}(\xi_1 - a)_+}{b - a} \leq \frac{\mathcal{E}\xi_{N+} + a_-}{b - a}.
\]

Proof It should be checked that \( U_{[a,b]}(\omega) \) is a r.v. This may be done by first showing that \( \{ \tau_n(\omega) : 1 \leq n \leq N \} \) are r.v.’s and then using the definition of \( U_{[a,b]} \) in terms of the \( \tau_n \)’s.

Next assume first that \( a = 0 \) and \( \xi_n \geq 0 \) for all \( n = 1, \ldots, N \). Define \( \{ \eta_n(\omega) : 1 \leq n \leq N \} \) by

\[
\eta_n(\omega) = \begin{cases} 
1 & \text{if } \tau_{2k-1}(\omega) \leq n < \tau_{2k}(\omega) \text{ for some } k = 1, \ldots, [N/2] \\
0 & \text{otherwise}
\end{cases}
\]
We now show that each $\eta_n$ is an $F_n$-measurable r.v. Since by definition $\{\eta_1 = 1\} = \{\xi_1 = 0\}$, $\eta_1$ is an $F_1$-measurable r.v. If $\eta_n$ is $F_n$-measurable, $1 \leq n \leq N$, then it is clear from the definition of the $\eta_n$’s that

$$\{\eta_{n+1} = 1\} = \{\eta_n = 1, \ 0 \leq \xi_{n+1} < b\} \cup \{\eta_n = 0, \ \xi_{n+1} = 0\}$$

and thus $\eta_{n+1}$ is $F_{n+1}$-measurable. It follows by finite induction that each $\eta_n$ is $F_n$-measurable. Define

$$\zeta = \xi_1 + \sum_{n=1}^{N-1} \eta_n (\xi_{n+1} - \xi_n).$$

If $M(\omega)$ is the number of $\tau_n(\omega)$’s that do not exceed $N$, so that $U_{[0,b]}(\omega) = [M(\omega)/2]$, then if $M$ is even

$$\zeta = \xi_1 + \sum_{k=1}^{U_{[0,b]}(\omega)} (\xi_{2k} - \xi_{2k-1})$$

and if $M$ is odd

$$\zeta = \xi_1 + \sum_{k=1}^{U_{[0,b]}(\omega)} (\xi_{2k} - \xi_{2k-1}) + (\xi_N - \xi_{\tau_M}).$$

Since $\xi_{2k} - \xi_{2k-1} \geq b$ and $\xi_N - \xi_{\tau_M} = \xi_N - 0 \geq 0$, we have in either case, i.e. for all $\omega \in \Omega$,

$$\zeta \geq \xi_1 + bU_{[0,b]}$$

and thus

$$E[U_{[0,b]}] \leq \frac{E[\zeta] - E[\xi_1]}{b}.$$

Also

$$E[\zeta] = E[\xi_1] + \sum_{n=1}^{N-1} E[\eta_n (\xi_{n+1} - \xi_n)].$$

Since $\eta_n$ is $F_n$-measurable, $0 \leq \eta_n \leq 1$, and $E(\xi_{n+1} - \xi_n | F_n) \geq 0$ by the submartingale property, we have for $n = 1, \ldots, N - 1$,

$$E(\eta_n (\xi_{n+1} - \xi_n)) = E(E(\eta_n (\xi_{n+1} - \xi_n) | F_n))$$

$$= E(E(\eta_n | F_n) E(\xi_{n+1} - \xi_n | F_n))$$

$$\leq E(E(\xi_{n+1} - \xi_n | F_n))$$

$$= E(\xi_{n+1} - \xi_n).$$

It follows that

$$E[\zeta] \leq E[\xi_1] + \sum_{n=1}^{N-1} E(\xi_{n+1} - \xi_n) = E[\xi_N]$$
and hence

\[ \mathcal{E}U_{[0,b]} \leq \frac{\xi_N^+ - \xi_1^-}{b} \].

For the general case note that the number of upcrossings of \([a,b]\) by \(\{\xi_n\}_{n=1}^N\) is equal to the number of upcrossings of \([0,b-a]\) by \(\{\xi_n - a\}_{n=1}^N\) and this is also equal to the number of upcrossings of \([0,b-a]\) by \(\{((\xi_n - a)_+) : 1 \leq n \leq N\}\). Since \(\{(\xi_n, \mathcal{F}_n) : 1 \leq n \leq N\}\) is a submartingale, so is \(\{(\xi_n - a, \mathcal{F}_n) : 1 \leq n \leq N\}\) and also \(\{(\xi_n - a)_+, \mathcal{F}_n) : 1 \leq n \leq N\}\) by (i) of Theorem 14.1.3, Corollary. It follows from the particular case just considered that

\[ \mathcal{E}U_{[a,b]} \leq \frac{\xi_N^+ - a - \xi_1^-}{b - a} \leq \frac{\xi_N^+ + a^-}{b - a} \]

since \((\xi_N - a)_+ \leq \xi_N^+ + a^-\). 

\[ \Box \]

### 14.3 Convergence

In this section it is shown that under mild conditions submartingales and martingales (and also supermartingales) converge almost surely. The convergence theorems which follow are very useful in probability and statistics. We start with a sufficient condition for a.s. convergence of a submartingale.

**Theorem 14.3.1** Let \(\{\xi_n, \mathcal{F}_n\}\) be a submartingale. If

\[ \lim_{n \to \infty} \mathcal{E}\xi_{n+}^+ < \infty \]

then there is an integrable r.v. \(\xi_\infty\) such that \(\xi_n \to \xi_\infty\) a.s.

**Proof** For every pair of real numbers \(a < b\), let \(U_{[a,b]}^{(n)}(\omega)\) be the number of upcrossings of \([a,b]\) by \(\{\xi_i(\omega) : 1 \leq i \leq n\}\). Then \(\{U_{[a,b]}^{(n)}(\omega)\}\) is a nondecreasing sequence of random variables and thus has a limit

\[ U_{[a,b]}(\omega) = \lim_{n \to \infty} U_{[a,b]}^{(n)}(\omega) \text{ a.s.} \]

By monotone convergence and Theorem 14.2.3, we have

\[ \mathcal{E}U_{[a,b]} = \lim_{n \to \infty} \mathcal{E}U_{[a,b]}^{(n)} \]

\[ \leq \lim_{n \to \infty} \frac{\xi_{n+}^+ + a^-}{b - a} < \infty, \]
so that $U_{[a,b]} < \infty$ a.s. It follows that if
$$E_{[a,b]} = \{ \omega \in \Omega : \liminf_n \xi_n(\omega) < a < b < \limsup_n \xi_n(\omega) \}$$
then
$$P(E_{[a,b]}) = 0 \text{ for all } a < b.$$ 
Thus if
$$E = \bigcup_{a,b: \text{rational}} E_{[a,b]} = \{ \omega \in \Omega : \liminf_n \xi_n(\omega) < \limsup_n \xi_n(\omega) \}$$
then $P(E) = 0$. It follows that $\liminf_n \xi_n(\omega) = \limsup_n \xi_n(\omega)$ a.s. and thus the limit $\lim_{n \to \infty} \xi_n$ exists a.s. Denote this limit by $\xi_\infty$. Then, by Fatou’s Lemma,
$$\mathcal{E}|\xi_\infty| \leq \liminf_n \mathcal{E}|\xi_n|$$
and since (by Theorem 14.1.1 (ii)) $\mathcal{E}|\xi_n| \geq \mathcal{E}_1$,
$$\mathcal{E}|\xi_n| = \mathcal{E}(2\xi_{n+} - \xi_n) \leq 2\mathcal{E}_{n+} - \mathcal{E}_1$$
we obtain
$$\mathcal{E}|\xi_\infty| \leq \liminf_n \{2\mathcal{E}_{n+} - \mathcal{E}_1\}$$
$$= 2 \lim_n \mathcal{E}_{n+} - \mathcal{E}(\xi_1) < \infty.$$ 
Thus $\xi_\infty$ is integrable. □

The next theorem gives conditions under which the a.s. converging submartingale of Theorem 14.3.1 converges also in $L_1$. Throughout the following, given a sequence of $\sigma$-fields $\{\mathcal{F}_n\}$, we denote by $\mathcal{F}_\infty$ the $\sigma$-field generated by $\bigcup_{n=1}^\infty \mathcal{F}_n$. Also, by including $(\xi_\infty, \mathcal{F}_\infty)$ in the sequence, we call $\{ (\xi_n, \mathcal{F}_n) : n = 1, 2, \ldots, \infty \}$ a martingale (respectively submartingale, supermartingale) if for all $m, n$ in $\{1, 2, \ldots, \infty \}$ with $m < n$,

(i) $\mathcal{F}_m \subset \mathcal{F}_n$
(ii) $\xi_n$ is $\mathcal{F}_n$-measurable and integrable
(iii) $\mathcal{E}(\xi_n|\mathcal{F}_m) = \xi_m$ a.s. (resp. $\geq \xi_m, \leq \xi_m$).

We have the following result.

**Theorem 14.3.2** If $\{ \xi_n, \mathcal{F}_n \}$ is a submartingale, the following are equivalent

(i) the sequence $\{ \xi_n \}$ is uniformly integrable
(ii) the sequence $\{ \xi_n \}$ converges in $L_1$
(iii) the sequence \( \{\xi_n\} \) converges a.s. to an integrable r.v. \( \xi_\infty \) such that 
\( \{(\xi_n, F_n) : n = 1, 2, \ldots, \infty\} \) is a submartingale and \( \lim_n \mathcal{E}_{\xi_n} = \mathcal{E}_{\xi_\infty} \).

**Proof** (i) \( \Rightarrow \) (ii): Since \( \{\xi_n\} \) is uniformly integrable, Theorem 11.4.1 implies \( \sup_n \mathbb{E}|\xi_n| < \infty \) and thus, by Theorem 14.3.1, there is an integrable r.v. \( \xi_\infty \) such that \( \xi_n \to \xi_\infty \) a.s. Since a.s. convergence implies convergence in probability, it follows from Theorem 11.4.2 that \( \xi_n \to \xi_\infty \) in \( L_1 \).

(ii) \( \Rightarrow \) (iii): If \( \xi_n \to \xi_\infty \) in \( L_1 \) we have by Theorem 11.4.2, 
\( \mathbb{E}|\xi_n| \to \mathbb{E}|\xi_\infty| < \infty \) and thus \( \sup_n \mathbb{E}|\xi_n| < \infty \). It then follows from Theorem 14.3.1 that \( \xi_n \to \xi_\infty \) a.s.

In order to show that \( \{(\xi_n, F_n) : n = 1, 2, \ldots, \infty\} \) is a submartingale it suffices to show that for all \( n = 1, 2, \ldots \)
\[ \mathcal{E}(\xi_\infty|F_n) \geq \xi_n \text{ a.s.} \]

For every fixed \( n \) and \( E \in F_n \), using the definition of conditional expectation and the convergence \( \xi_m \to \xi_\infty \) in \( L_1 \) (which implies \( \mathbb{E}\xi_m \to \mathbb{E}\xi_\infty \))
\[ \int_E \mathcal{E}(\xi_\infty|F_n) \, d\mathbb{P} = \int_E \xi_\infty \, d\mathbb{P} \]
\[ = \lim_{m \to \infty} \int_E \xi_m \, d\mathbb{P} \]
\[ = \lim_{m \to \infty} \int_E \mathcal{E}(\xi_m|F_n) \, d\mathbb{P} \]
\[ \geq \int_E \xi_n \, d\mathbb{P} \]

since \( \mathcal{E}(\xi_m|F_n) \geq \xi_n \) a.s. for \( m > n \). Thus \( \mathcal{E}(\xi_\infty|F_n) \geq \xi_n \) a.s. (see Ex. 4.14) and as already noted above \( \lim_n \mathcal{E}_{\xi_n} = \mathcal{E}_{\xi_\infty} \).

(iii) \( \Rightarrow \) (i): Since \( \{(\xi_n, F_n) : n = 1, 2, \ldots, \infty\} \) is a submartingale, so is 
\( \{(\xi_n^+, F_n) : n = 1, 2, \ldots, \infty\} \). Thus using the submartingale property repeatedly we have
\[ \int_{(\xi_n^+ > a)} \xi_n^+ \, d\mathbb{P} \leq \int_{(\xi_n^+ > a)} \mathcal{E}(\xi_\infty^+|F_n) \, d\mathbb{P} = \int_{(\xi_n^+ > a)} \xi_\infty^+ \, d\mathbb{P} \]
and
\[ \mathbb{P}\{\xi_n^+ > a\} \leq \frac{1}{a} \mathcal{E}_{\xi_n^+} \leq \frac{1}{a} \mathbb{E}(\mathcal{E}(\xi_\infty^+|F_n)) = \frac{1}{a} \mathcal{E}_{\xi_\infty^+} \to 0 \text{ as } a \to \infty \]
which clearly imply that \( \{\xi_n^+\} \) is uniformly integrable.

Since \( \xi_n^+ \to \xi_\infty^+ \) a.s. and thus also in probability, and since the sequence is uniformly integrable, it follows by Theorem 11.4.2 that \( \xi_n^+ \to \xi_\infty^+ \) in \( L_1 \), and hence that \( \mathcal{E}_{\xi_n^+} \to \mathcal{E}_{\xi_\infty^+} \). Since by assumption \( \mathcal{E}_{\xi_n} \to \mathcal{E}_{\xi_\infty} \), it also follows that \( \mathcal{E}_{\xi_n^+} \to \mathcal{E}_{\xi_\infty^+} \) and hence in probability, Theorem 11.4.2 implies that \( \{\xi_n^-\} \) is uniformly integrable.

Since \( \xi_n = \xi_n^+-\xi_n^- \), the uniform integrability of \( \{\xi_n : n = 1, 2, \ldots\} \) follows (see Ex. 11.21). □
For martingales the following more detailed and useful result holds.

**Theorem 14.3.3**  If \( \{\xi_n, \mathcal{F}_n\} \) is a martingale, the following are equivalent

(i) the sequence \( \{\xi_n\} \) is uniformly integrable
(ii) the sequence \( \{\xi_n\} \) converges in \( L_1 \)
(iii) the sequence \( \{\xi_n\} \) converges a.s. to an integrable r.v. \( \xi_\infty \) such that \( \{(\xi_n, \mathcal{F}_n) : n = 1, 2, \ldots, \infty\} \) is a martingale
(iv) there is an integrable r.v. \( \eta \) such that \( \xi_n = \mathbb{E}(\eta|\mathcal{F}_n) \) for all \( n = 1, 2, \ldots \) a.s.

**Proof** That (i) implies (ii) and (ii) implies (iii) follow from Theorem 14.3.2. That (iii) implies (i) is shown as in Theorem 14.3.2 by considering \( |\xi_n| \) instead of \( \xi_n \), and it is shown trivially by taking \( \eta = \xi_\infty \) that (iii) implies (iv).

(iv) \( \Rightarrow \) (i): Put \( \xi_\infty = \eta \). Then \( \mathbb{E}(\xi_\infty|\mathcal{F}_n) = \mathbb{E}(\eta|\mathcal{F}_n) = \xi_n \) and clearly \( \{(\xi_n, \mathcal{F}_n) : n = 1, 2, \ldots, \infty\} \) is a martingale and thus \( \{(\xi_n, \mathcal{F}_n) : n = 1, 2, \ldots, \infty\} \) is a submartingale. We thus have

\[
\int_{|\xi_n| > a} |\xi_n| dP \leq \int_{|\xi_n| > a} \mathbb{E}(\xi_\infty|\mathcal{F}_n) dP = \int_{|\xi_n| > a} |\xi_\infty| dP
\]

and

\[
P(|\xi_n| > a) \leq \frac{1}{a} \mathbb{E}|\xi_n| \leq \frac{1}{a} \mathbb{E}|\xi_\infty| \to 0 \text{ as } a \to \infty,
\]

which clearly imply that \( \{\xi_n\} \) is uniformly integrable.

As a simple consequence of the previous theorem we have the following very useful result.

**Theorem 14.3.4**  Let \( \xi \) be an integrable r.v., \( \{\mathcal{F}_n\} \) a sequence of sub-\( \sigma \)-fields of \( \mathcal{F} \) such that \( \mathcal{F}_n \subset \mathcal{F}_{n+1} \) all \( n \), and \( \mathcal{F}_\infty \) the \( \sigma \)-field generated by \( \bigcup_{n=1}^\infty \mathcal{F}_n \). Then

\[
\lim_{n \to \infty} \mathbb{E}(\xi|\mathcal{F}_n) = \mathbb{E}(\xi|\mathcal{F}_\infty) \text{ a.s. and in } L_1.
\]

**Proof**  Let \( \xi_n = \mathbb{E}(\xi|\mathcal{F}_n) \), \( n = 1, 2, \ldots \). Then \( \{\xi_n, \mathcal{F}_n\} \) is a martingale (by Example 3 in Section 14.1) which satisfies (iv) of Theorem 14.3.3. It follows by (ii) and (iii) of that theorem that there is an integrable r.v. \( \xi_\infty \) such that \( \xi_n \to \xi_\infty \) a.s. and in \( L_1 \).

It suffices now to show that \( \mathbb{E}(\xi|\mathcal{F}_n) \to \mathbb{E}(\xi|\mathcal{F}_\infty) \) a.s. Since by (iii) of Theorem 14.3.3, \( \{(\xi_n, \mathcal{F}_n) : n = 1, 2, \ldots, \infty\} \) is a martingale, we have that for all \( E \in \mathcal{F}_n \),

\[
\int_E \xi_\infty dP = \int_E \mathbb{E}(\xi_\infty|\mathcal{F}_n) dP = \int_E \xi_n dP = \int_E \mathbb{E}(\xi|\mathcal{F}_n) dP = \int_E \xi dP.
\]
Hence \( \int \xi_\infty \, dP = \int \xi \, dP \) for all sets \( E \) in \( \mathcal{F}_n \) and thus in \( \bigcup_{n=1}^{\infty} \mathcal{F}_n \). It is clear that the class of sets for which it holds is a \( \mathcal{D} \)-class, and since it contains \( \bigcup_{n=1}^{\infty} \mathcal{F}_n \) (which is closed under intersections) it contains also \( \mathcal{F}_\infty \). Hence

\[
\int \xi_\infty \, dP = \int \xi \, dP \quad \text{for all } E \in \mathcal{F}_\infty
\]

and since \( \xi_\infty = \lim_n \xi_n \) is \( \mathcal{F}_\infty \)-measurable, it follows that \( \xi_\infty = E(\xi|\mathcal{F}_\infty) \) a.s.

A result similar to Theorem 14.3.4 is also true for decreasing (rather than increasing) sequences of \( \sigma \)-fields and follows easily if we introduce the concept of reverse submartingale and martingale as follows. Let \( \{\xi_n\} \) be a sequence of r.v.’s and \( \{\mathcal{F}_n\} \) a sequence of sub-\( \sigma \)-fields of \( \mathcal{F} \). We say that \( \{\xi_n, \mathcal{F}_n\} \) is a reverse martingale (respectively, submartingale, supermartingale) if for every \( n \),

(i) \( \mathcal{F}_n \supset \mathcal{F}_{n+1} \)

(ii) \( \xi_n \) is \( \mathcal{F}_n \)-measurable and integrable

(iii) \( E(\xi_n|\mathcal{F}_{n+1}) = \xi_{n+1} \) (resp. \( \geq \xi_{n+1}, \leq \xi_{n+1} \)) a.s.

The following convergence result corresponds to Theorem 14.3.1.

**Theorem 14.3.5** Let \( \{\xi_n, \mathcal{F}_n\} \) be a reverse submartingale. Then there is a r.v. \( \xi_\infty \) such that \( \xi_n \to \xi_\infty \) a.s. and if

\[
\lim_{n \to \infty} E\xi_n > -\infty
\]

then \( \xi_\infty \) is integrable.

**Proof** The proof is similar to that of Theorem 14.3.1. For each fixed \( n \), define

\[
\eta_k = \xi_{n-k+1}, \quad G_k = \mathcal{F}_{n-k+1} \quad k = 1, 2, \ldots, n,
\]

i.e. \( \{\eta_1, G_1; \eta_2, G_2; \ldots; \eta_n, G_n\} = \{\xi_n, \mathcal{F}_n; \xi_{n-1}, \mathcal{F}_{n-1}; \ldots; \xi_1, \mathcal{F}_1\} \). Then \( \{(\eta_k, G_k) : 1 \leq k \leq n\} \) is a submartingale since

\[
E(\eta_{k+1}|G_k) = E(\xi_{n-k}|\mathcal{F}_{n-k+1}) = \eta_k \text{ a.s.}
\]

If \( U_{[a,b]}^{(n)}(\omega) \) denotes the number of upcrossings of the interval \( [a, b] \) by the sequence \( \{\xi_n(\omega), \xi_{n-1}(\omega), \ldots, \xi_1(\omega)\} \), then \( U_{[a,b]}^{(n)}(\omega) \) is equal to the number
of upcrossings of the interval \([a, b]\) by the submartingale \(\{\eta_1(\omega), \ldots, \eta_n(\omega)\}\)
and by Theorem 14.2.3 we have

\[
E U^{(n)}_{[a, b]} \leq \frac{E\eta_{n+} + a_ -}{b - a} = \frac{E\xi_{1+} + a_ -}{b - a}.
\]

As in the proof of Theorem 14.3.1 it follows that the sequence \(\{\xi_n\}\) converges a.s., i.e. \(\xi_n \to \xi_{\infty}\) a.s. Again as in the proof of Theorem 14.3.1 we have by Fatou’s Lemma,

\[
E|\xi_{\infty}| \leq \liminf_n E|\xi_n| \text{ and } E|\xi_{\infty}| = 2E\xi_{n+} - E\xi_n.
\]

But now

\[
E\xi_{n+} = E\eta_{1+} \leq E\eta_{n+} = E\xi_{1+}
\]

since \(\{(\eta_{k+}, G_k) : 1 \leq k \leq n\}\) is a submartingale. Also \(\{E\xi_n\}\) is clearly a nonincreasing sequence. Since \(\lim_n E\xi_n > -\infty\) it follows that

\[
E|\xi_{\infty}| \leq 2E\xi_{1+} - \lim_{n \to \infty} E\xi_n < \infty
\]

and thus \(\xi_{\infty}\) is integrable. \(\Box\)

**Corollary** If \(\{\xi_n, F_n\}\) is a reverse martingale, then there is an integrable r.v. \(\xi_{\infty}\) such that \(\xi_n \to \xi_{\infty}\) a.s.

**Proof** If \(\{\xi_n, F_n\}\) is a reverse martingale, clearly the sequence \(\{E\xi_n\}\) is constant and thus \(\lim_n E\xi_n = E\xi_1 > -\infty\). The result then follows from the theorem. \(\Box\)

We now prove the result of Theorem 14.3.4 for decreasing sequences of \(\sigma\)-fields.

**Theorem 14.3.6** Let \(\xi\) be an integrable r.v., \(\{F_n\}\) a sequence of sub-\(\sigma\)-fields of \(F\) such that \(F_n \supset F_{n+1}\) for all \(n\), and \(F_{\infty} = \bigcap_{n=1}^{\infty} F_n\). Then

\[
\lim_{n \to \infty} E(\xi | F_n) = E(\xi | F_{\infty}) \text{ a.s. and in } L_1.
\]

**Proof** Let \(\xi_n = E(\xi | F_n)\). Then \(\{\xi_n, F_n\}\) is a reverse martingale since \(F_n \supset F_{n+1}\), \(\xi_n\) is \(F_n\)-measurable and integrable and by Theorem 13.2.2,

\[
E(\xi_n | F_{n+1}) = E(E(\xi | F_n) | F_{n+1}) = E(\xi | F_{n+1}) = \xi_{n+1} \text{ a.s.}
\]

It follows from the corollary of Theorem 14.3.5 that \(\xi_n \to \xi_{\infty}\) a.s. for some integrable r.v. \(\xi_{\infty}\).
We first show that $\xi_n \to \xi_\infty$ in $L_1$ as well. This follows from Theorem 11.4.2 since the sequence $\{\xi_n\}_{n=1}^\infty$ is uniformly integrable as is seen from
\[
\int_{|\xi_n|>a} |\xi_n| dP \leq \int_{|\xi_n|>a} E(|\xi||F_n) \, dP = \int_{|\xi_n|>a} |\xi| \, dP
\]
and
\[
P(|\xi_n| > a) \leq \frac{1}{a} E|\xi_n| \leq \frac{1}{a} E|\xi| \to 0 \text{ as } a \to \infty
\]
since $|\xi_n| = |E(\xi||F_n)| \leq E(|\xi||F_n)$ a.s. and thus $E|\xi_n| \leq E|\xi|$.

We now show that $\xi_\infty = E(\xi||F_\infty)$ a.s. For every $E \in F_\infty$ we have $E \in F_n$ for all $n$ and since $\xi_n = E(\xi||F_\infty)$ and $\xi_n \to \xi_\infty$ in $L_1$,
\[
\int_E \xi \, dP = \int_E \xi_n \, dP \to \int_E \xi_\infty \, dP \text{ as } n \to \infty.
\]
Hence $\int_E \xi \, dP = \int_E \xi_\infty \, dP$ for all $E \in F_\infty$. Also the relations $\xi_\infty = \lim_n \xi_n$ a.s. and $F_n \supset F_{n+1}$ imply that $\xi_\infty$ is $F_n$-measurable for all $n$ and thus $F_\infty$-measurable. It follows that $\xi_\infty = E(\xi||F_\infty)$ a.s. \(\square\)

### 14.4 Centered sequences

In this section the results of Section 14.3 will be used to study the convergence of series and the law of large numbers for "centered" sequences of r.v.'s, a concept which generalizes that of a sequence of independent and zero mean r.v.'s. We will also give martingale proofs for some of the previous convergence results for sequences of independent r.v.'s.

A sequence of r.v.'s $\{\xi_n\}$ is called centered if for every $n = 1, 2, \ldots$, $\xi_n$ is integrable and
\[
E(\xi_n||F_{n-1}) = 0 \text{ a.s.}
\]
where $F_n = \sigma(\xi_1, \ldots, \xi_n)$ and $F_0 = \{\emptyset, \Omega\}$. For $n = 1$ this condition is just $E\xi_1 = 0$ while for $n > 1$ it implies the weaker condition $E\xi_n = 0$. $F_n$ will be assumed to be $\sigma(\xi_1, \ldots, \xi_n)$ throughout this section unless otherwise stated. The basic properties of centered sequences are collected in the following theorem. Property (i) shows that results obtained for centered sequences are directly applicable to arbitrary sequences of integrable r.v.'s appropriately modified, i.e. centered.

**Theorem 14.4.1** 
(i) If $\{\xi_n\}$ is a sequence of integrable r.v.'s then the sequence $\{\xi_n - E(\xi_n||F_{n-1})\}$ is centered.
(ii) The sequence of partial sums of a centered sequence is a zero mean martingale, and conversely, every zero mean martingale is the sequence of partial sums of a centered sequence.
(iii) A sequence of independent r.v.’s \( \{\xi_n\} \) is centered if and only if for each \( n, \xi_n \in L_1 \) and \( E\xi_n = 0 \).

(iv) If the sequence of r.v.’s \( \{\xi_n\} \) is centered and \( \xi_n \in L_2 \) for all \( n \), then the r.v.’s of the sequence are orthogonal: \( E\xi_n \xi_m = 0 \) for all \( n \neq m \).

Proof

(i) is obvious. For (ii) let \( \{\xi_n\} \) be centered and let \( S_n = \xi_1 + \cdots + \xi_n = S_{n-1} + \xi_n \) for \( n = 1, 2, \ldots, \) where \( S_0 = 0 \). Then each \( S_n \) is integrable and \( F_{n-1} \)-measurable and

\[
E(S_n | F_{n-1}) = E(S_{n-1} | F_{n-1}) + E(\xi_n | F_{n-1}) = S_{n-1} \text{ a.s.}
\]

Note that \( F_n = \sigma(\xi_1, \ldots, \xi_n) = \sigma(S_1, \ldots, S_n) \). It follows that \( \{S_n\} \) is a martingale with zero mean since \( E S_1 = E \xi_1 = 0 \). Conversely, if \( \{S_n\} \) is a zero mean martingale, let \( \xi_n = S_n - S_{n-1} \) for \( n = 1, 2, \ldots, \) where \( S_0 = 0 \). Then each \( \xi_n \) is \( F_{n-1} \)-measurable and

\[
E(\xi_n | F_{n-1}) = E(S_n | F_{n-1}) - S_{n-1} = 0 \text{ a.s.}
\]

Hence \( \{\xi_n\} \) is centered and clearly \( \xi_1 + \cdots + \xi_n = S_n - S_0 = S_n \).

(iii) follows immediately from the fact that for independent integrable r.v.’s \( \{\xi_n\} \) and all \( n = 1, 2, \ldots \) we have from Theorem 10.3.2 that the \( \sigma \)-fields \( F_{n-1} \) and \( \sigma(\xi_n) \) are independent and thus by Theorem 13.2.7,

\[
E(\xi_n | F_{n-1}) = E\xi_n \text{ a.s.}
\]

(iv) Let \( \{\xi_n\} \) be centered, \( \xi_n \in L_2 \) for all \( n \), and \( m < n \). Then since \( \xi_n \) is \( F_m \subset F_{n-1} \)-measurable and \( E(\xi_n | F_{n-1}) = 0 \) a.s. we have

\[
E(\xi_n \xi_m) = E(E(\xi_n \xi_m | F_{n-1})) = E(\xi_m E(\xi_n | F_{n-1})) = E(0) = 0.
\]

We now prove for centered sequences of r.v.’s some of the convergence results shown in Sections 11.5 and 11.6 for sequences of independent r.v.’s. In view of Theorem 14.4.1 (iii), the following result on the convergence of series of centered r.v.’s generalizes the corresponding result for series of independent r.v.’s (Theorem 11.5.3).

**Theorem 14.4.2** If \( \{\xi_n\} \) is a centered sequence of r.v.’s and if \( \sum_{n=1}^{\infty} E\xi_n^2 < \infty \), then the series \( \sum_{n=1}^{\infty} \xi_n \) converges a.s. and in \( L_2 \).

Proof Let \( S_n = \sum_{k=1}^{n} \xi_k \). Then \( S_n \in L_2 \) since by assumption \( E\xi_n^2 < \infty \) for all \( n \). It follows from Theorem 14.4.1 (iv) that for all \( m < n \),

\[
E(S_n - S_m)^2 = E \left( \sum_{k=m+1}^{n} \xi_k \right)^2 = \sum_{k=m+1}^{n} E\xi_k^2 \rightarrow 0 \text{ as } m, n \rightarrow \infty
\]
since $\sum_{k=1}^{\infty} E\xi_k^2 < \infty$. Hence $\{S_n\}_{n=1}^{\infty}$ is a Cauchy sequence in $L_2$ and by Theorem 6.4.7 (i) there is a r.v. $S \in L_2$ such that $S_n \to S$ in $L_2$. Thus the series converges in $L_2$. Now Theorem 9.5.2 shows that convergence in $L_2$ implies convergence in $L_1$ and thus $S_n \to S$ in $L_1$. Since by Theorem 14.4.1 (ii), $\{S_n\}_{n=1}^{\infty}$ is a martingale, condition (ii) of Theorem 14.3.3 is satisfied and thus (by (iii) of that theorem) $S_n \to S$ a.s. and the series converges also a.s. \qed

Note that the result of this theorem follows also directly from Ex. 14.8.

We now prove a strong law of large numbers for centered sequences which generalizes the corresponding result for sequences of independent r.v.’s (Theorem 11.6.2).

**Theorem 14.4.3** If $\{\xi_n\}$ is a centered sequence of r.v.’s and if

$$\sum_{n=1}^{\infty} E\xi_n^2 n^2 < \infty$$

then

$$\frac{1}{n} \sum_{k=1}^{n} \xi_k \to 0 \ a.s.$$  

**Proof** This follows from Theorem 14.4.2 and Lemma 11.6.1 in the same way as Theorem 11.6.2 follows from Theorem 11.5.3 and Lemma 11.6.1. \qed

The special convergence results for sequences of independent r.v.’s, i.e. Theorems 11.5.4, 11.6.3 and 12.5.2, can also be obtained as applications of the martingale convergence theorems. As an illustration we include here martingale proofs of the strong law of large numbers (second form, Theorem 11.6.3) and of Theorem 12.5.2.

**Theorem 14.4.4** (Strong Law, Second Form) Let $\{\xi_n\}$ be independent and identically distributed r.v.’s with (the same) finite mean $\mu$. Then

$$\frac{1}{n} \sum_{k=1}^{n} \xi_k \to \mu \ a.s. \ and \ in \ L_1.$$  

**Proof** Let $S_n = \xi_1 + \cdots + \xi_n$. We first show that for each $1 \leq k \leq n$,

$$E(\xi_k | S_n) = \frac{1}{n} S_n \ a.s.$$  


Every set \( E \in \sigma(S_n) \) is of the form \( E = S_n^{-1}(B) \), \( B \in \mathcal{B} \), and thus
\[
\int_E \xi_k \, dP = \mathcal{E}(\xi_k \chi_{\{S_n \in B\}}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} x_k \chi_B(x_1 + \cdots + x_n) \, dF(x_1) \cdots dF(x_n)
\]
where \( F \) is the common d.f. of the \( \xi_n \)'s. It follows from Fubini’s Theorem that the last expression does not depend on \( k \) and thus
\[
\int_E \xi_k \, dP = \frac{1}{n} \sum_{i=1}^{n} \int_E \xi_i \, dP = \frac{1}{n} \int_E S_n \, dP
\]
which implies \( \mathcal{E}(\xi_k | S_n) = \frac{1}{n} S_n \) a.s.

Now let \( \mathcal{F}_n = \sigma(S_n, S_{n+1}, \ldots) \) (hence \( \mathcal{F}_n \supset \mathcal{F}_{n+1} \)) and let \( \mathcal{F}_\infty = \cap_{n=1}^{\infty} \mathcal{F}_n \). Since \( S_{n+1} - S_n = \xi_{n+1} \) it is clear that \( \mathcal{F}_n = \sigma(S_n, \xi_{n+1}, \xi_{n+2}, \ldots) \). Also since the classes of events \( \sigma(\xi_1, S_n) \) and \( \sigma(\xi_{n+1}, \xi_{n+2}, \ldots) \) are independent, an obvious generalization of Ex. 13.3 gives
\[
\mathcal{E}(\xi_1 | S_n) = \mathcal{E}(\xi_1 | \mathcal{F}_n) \text{ a.s.}
\]
Thus
\[
\frac{1}{n} S_n = \mathcal{E}(\xi_1 | \mathcal{F}_n) \text{ a.s.}
\]
and Theorem 14.3.6 implies that
\[
\frac{1}{n} S_n \to \mathcal{E}(\xi_1 | \mathcal{F}_\infty) \text{ a.s. and in } L_1.
\]

Now \( \lim_{n} \frac{1}{n} S_n = \lim_{n} \frac{1}{n} (S_n - S_k) \) implies that \( \lim_{n} \frac{1}{n} S_n \) is a tail r.v. of the independent sequence \( \{\xi_n\} \) and by Kolmogorov’s Zero-One Law (Theorem 10.5.3) it is constant a.s. Hence \( \mathcal{E}(\xi_1 | \mathcal{F}_\infty) \) is constant a.s. and thus \( \mathcal{E}(\xi_1 | \mathcal{F}_\infty) = \mathcal{E} \xi_1 = \mu \) a.s. It follows that \( \frac{1}{n} S_n \to \mu \) a.s. and in \( L_1 \).

The following result gives a martingale proof of Theorem 12.5.2.

**Theorem 14.4.5** Let \( \{\xi_n\} \) be a sequence of independent random variables \( \{\phi_n\} \). Then the following are equivalent:

(i) the series \( \sum_{n=1}^{\infty} \xi_n \) converges a.s.
(ii) the series \( \sum_{n=1}^{\infty} \xi_n \) converges in distribution
(iii) the products \( \prod_{k=1}^{n} \phi_k(t) \) converge to a nonzero limit in some neighborhood of the origin.

**Proof** Clearly, it suffices to show that (iii) implies (i), i.e. assume that
\[
\lim_{n \to \infty} \prod_{k=1}^{n} \phi_k(t) = \phi(t) \neq 0 \text{ for each } t \in [-a, a] \text{ for some } a > 0.
\]
Let $S_n = \sum_{k=1}^n \xi_k$ and $\mathcal{F}_n = \sigma(\xi_1, \ldots, \xi_n) = \sigma(S_1, \ldots, S_n)$. For each fixed $t \in [-a, a]$ the sequence $\left\{e^{itS_n} / \prod_{k=1}^n \phi_k(t) \right\}$ is integrable $(dP)$, indeed uniformly bounded, and it follows from Fubini’s Theorem that $\left\{e^{itS_n} / \prod_{k=1}^n \phi_k(t), \mathcal{F}_n \right\}$ is a martingale, in the sense that its real and imaginary parts are martingales. Since for each $t$ the sequence is uniformly bounded, Theorem 14.3.1 applied to the real and imaginary parts shows that the sequence $e^{itS_n}$ converges a.s. as $n \to \infty$. Since the denominator converges to a nonzero limit, it follows that $e^{itS_n}$ converges a.s. as $n \to \infty$, for each $t \in [-a, a]$. Some analysis using this fact will lead to the conclusion that $S_n$ converges a.s.

We have that for every $t \in [-a, a]$ there is a set $\Omega_t \in \mathcal{F}$ with $P(\Omega_t) = 0$ such that for every $\omega \notin \Omega_t$, $e^{itS_n(\omega)}$ converges. Now consider $e^{itS_n(\omega)}$ as a function of the two variables $(t, \omega)$, i.e. in the product space $([-a, a] \times \Omega, \mathcal{B}_{[-a,a]} \times \mathcal{F})$, where $\mathcal{B}_{[-a,a]}$ is the $\sigma$-field of Borel subsets of $[-a, a]$ and $m$ denotes Lebesgue measure. Then clearly $e^{itS_n(\omega)}$ is product measurable and hence

$$D = \{(t, \omega) \in [-a, a] \times \Omega : e^{itS_n(\omega)} \text{ does not converge}\} \in \mathcal{B}_{[-a,a]} \times \mathcal{F}.$$  

Note that the $t$-section of $D$ is

$$D_t = \{\omega \in \Omega : (t, \omega) \in D \} = \{\omega \in \Omega : e^{itS_n(\omega)} \text{ does not converge}\} = \Omega_t.$$  

It follows from Fubini’s Theorem that

$$(m \times P)(D) = \int_{-a}^a P(D_t) \, dt = \int_{-a}^a 0 \, dt = 0$$

and hence

$$0 = (m \times P)(D) = \int_{\Omega} m(D^\omega) \, dP(\omega).$$

Hence $m(D^\omega) = 0$ a.s., i.e. there is $\Omega_0 \in \mathcal{F}$ with $P(\Omega_0) = 0$ such that $m(D^\omega) = 0$ for all $\omega \notin \Omega_0$. But

$$D^\omega = \{t \in [-a, a] ; (t, \omega) \in D\} = \{t \in [-a, a] : e^{itS_n(\omega)} \text{ does not converge}\}.$$  

Hence for every $\omega \notin \Omega_0$, $P(\Omega_0) = 0$, there is $D^\omega \in \mathcal{B}_{[-a,a]}$ with $m(D^\omega) = 0$ such that $e^{itS_n(\omega)}$ converges for all $t \in [-a, a] - D^\omega$. The proof will be completed by showing that for all $\omega \notin \Omega_0$, $S_n(\omega)$ converges to a finite limit and since $P(\Omega_0) = 0$, this means that $S_n$ converges a.s.

Fix $\omega \notin \Omega_0$. To show the convergence of $S_n(\omega)$, we argue first that the sequence $\{S_n(\omega)\}$ is bounded. Indeed, by passing to a subsequence if necessary, suppose by contradiction that $S_n(\omega) \to \infty$. Denote the limit of $e^{itS_n(\omega)}$
by \( g(t) \), defined a.e. \((m)\) on \([-a, a]\). Dominated convergence yields that
\[
\frac{e^{ius_n(\omega)} - 1}{is_n(\omega)} = \int_0^ue^{its_n(\omega)} \, dt \to \int_0^ug(t) \, dt
\]
for any \( u \in [-a, a] \). But since \( S_n(\omega) \to \infty \), it follows that
\[
\int_0^ug(t) \, dt = 0 \quad \text{for any } u \in [-a, a],
\]
and hence \( g(\omega) = 0 \) a.e. \((m)\) on \([-a, a]\). This is a contradiction since \( |g(t)| = 1 = \lim_n |e^{its_n(\omega)}| \) a.e. \((m)\) on \([-a, a]\). If \( \{S_n(\omega)\} \) is bounded and there are two convergent subsequences \( S_{n_k}(\omega) \to s_1 \) and \( S_{n_k}(\omega) \to s_2 \), then \( e^{its_1} = e^{its_2} \) a.e. \((m)\) on \([-a, a]\). Since \( e^{its} \) is continuous for \( t \in [-a, a] \), it follows that \( e^{its_1} = e^{its_2} \) for all \( t \in [-a, a] \). Differentiating the two sides of the last equality and setting \( t = 0 \) yields \( s_1 = s_2 \) and hence that \( S_n(\omega) \) converges. \( \square \)

### 14.5 Further applications

In this section we give some further applications of the martingale convergence results of Section 14.3. The first application is related to the Lebesgue decomposition of one measure with respect to another, and thus also to the Radon–Nikodym Theorem; it helps to identify Radon–Nikodym derivatives and is also of interest in probability and especially in statistics.

**Theorem 14.5.1** Let \((\Omega, \mathcal{F}, P)\) be a probability space and \(\{\mathcal{F}_n\}\) a sequence of sub-\(\sigma\)-fields of \(\mathcal{F}\) such that \(\mathcal{F}_n \subset \mathcal{F}_{n+1}\) for all \(n\) with \(\sigma(\cup_{n=1}^{\infty} \mathcal{F}_n) = \mathcal{F}\). Let \(Q\) be a finite measure on \((\Omega, \mathcal{F})\) and consider its Lebesgue–Radon–Nikodym decomposition with respect to \(P\):
\[
Q(E) = \int_E \xi \, dP + Q(E \cap N) \quad \text{for all } E \in \mathcal{F}
\]
where \(0 \leq \xi \in L_1(\Omega, \mathcal{F}, P), \ N \in \mathcal{F}\) and \(P(N) = 0\). Denote by \(P_n, \ Q_n\) the restrictions of \(P, \ Q\) to \(\mathcal{F}_n\). If \(Q_n \ll P_n\) for all \(n = 1, 2, \ldots, \) then

(i) \(\{\frac{dQ_n}{dP_n}, \mathcal{F}_n\}\) is a martingale on \((\Omega, \mathcal{F}, P)\) and
\[
\frac{dQ_n}{dP_n} \to \xi \text{ a.s. } (P).
\]

(ii) \(Q \ll P\) if and only if \(\{\frac{dQ_n}{dP_n}\}\) is uniformly integrable on \((\Omega, \mathcal{F}, P)\) in which case
\[
\frac{dQ_n}{dP_n} \to \frac{dQ}{dP} \text{ a.s. } (P) \text{ and in } L_1(\Omega, \mathcal{F}, P).
\]
14.5 Further applications

Proof. (i) Let $\xi_n = \frac{dQ_n}{dP_n}$. Since $Q$ and thus $Q_n$ are finite, it follows that $\xi_n \in L_1(\Omega, \mathcal{F}, P)$, i.e. $\xi_n$ is $\mathcal{F}_n$-measurable and $P$-integrable. For every $E \in \mathcal{F}_n$ we have

$$\int_E \xi_{n+1} dP = \int_E \xi_{n+1} dP_{n+1} = Q_{n+1}(E) = Q_n(E)$$

$$= \int_E \xi_n dP_n = \int_E \xi_n dP.$$ 

Hence $\mathcal{E}(\xi_{n+1}|\mathcal{F}_n) = \xi_n$ for all $n$ a.s. and thus $\{\xi_n, \mathcal{F}_n\}_{n=1}^{\infty}$ is a martingale on $(\Omega, \mathcal{F}, P)$.

We also have $\xi_n \geq 0$ a.s. and

$$\mathcal{E}\xi_n = \int_\Omega \xi_n dP = Q_n(\Omega) = Q(\Omega) < \infty.$$ 

It follows from Theorem 14.3.1 that there is an integrable random variable $\xi_\infty$ such that

$$\xi_n \to \xi_\infty \text{ a.s. (} P).$$

Since $\xi_n \geq 0$ a.s. we have $\xi_\infty \geq 0$ a.s. We now show that $\xi_\infty = \xi$ a.s. Since $\xi_n \to \xi_\infty$ a.s., Fatou’s Lemma gives

$$\int_E \xi_\infty dP \leq \liminf \int_E \xi_n dP \text{ for all } E \in \mathcal{F}.$$ 

Hence for all $E \in \mathcal{F}_n$,

$$\int_E \xi_\infty dP \leq \liminf \int_E \xi_n dP = Q_n(E) = Q(E)$$ 

and thus $\int_E \xi_\infty dP \leq Q(E)$ for all $E \in \bigcup_{n=1}^{\infty} \mathcal{F}_n$. We conclude that the same is true for all $E \in \mathcal{F}$, either from the uniqueness of the extension of the finite measure $\mu(E) = Q(E) - \int_E \xi_\infty dP$ (Theorem 2.5.3) or from the monotone class theorem (Ex. 1.16). Since $P(N) = 0$ it follows that for every $E \in \mathcal{F}$,

$$\int_E \xi_\infty dP = \int_{E \cap N^c} \xi_\infty dP \leq Q(E \cap N^c) = \int_{E \cap N^c} \xi dP = \int_E \xi dP$$ 

and thus $\xi_\infty \leq \xi$ a.s.

For the inverse inequality we have $\int_E \xi dP \leq Q(E)$ for all $E \in \mathcal{F}$, and hence for all $E \in \mathcal{F}_n$,

$$\int_E \xi dP = \int_E \xi|\mathcal{F}_n dP = \int_E \xi dP \leq Q(E) = Q_n(E) = \int_E \xi_n dP.$$ 

Since both $\mathcal{E}(\xi|\mathcal{F}_n)$ and $\xi_n$ are $\mathcal{F}_n$-measurable, it follows as in the previous paragraph that

$$\mathcal{E}(\xi|\mathcal{F}_n) \leq \xi_n \text{ a.s.}$$

Since this is true for all $n$ and since $\xi_n \to \xi_\infty$ a.s. and by Theorem 14.3.4, $\mathcal{E}(\xi|\mathcal{F}_n) \to \mathcal{E}(\xi|\mathcal{F}) = \xi$ a.s., it follows that $\xi \leq \xi_\infty$ a.s. Thus $\xi_\infty = \xi$ a.s., i.e. (i) holds.
Martingales

(ii) First assume that $Q \ll P$. Then $Q(N) = 0$ and $\xi = \frac{dQ}{dP}$. Hence by (i), $\xi_n \to \xi$ a.s. Also for all $E \in \mathcal{F}_n$ we have

$$\int_E \xi_n dP = Q(E) = Q_n(E) = \int_E \xi_n dP_n = \int_E \xi_n dP$$

and thus $\xi_n = \mathcal{E}(\xi|\mathcal{F}_n)$. Hence condition (iv) of Theorem 14.3.3 is satisfied and from (i) and (ii) of the same theorem we have that $\{\xi_n\}_{n=1}^{\infty}$ is uniformly integrable on $(\Omega, \mathcal{F}, P)$, and $\xi_n \to \xi$ a.s.

Also for all $E \in \mathcal{F}_n$ we have

$$\int_E \xi_n dP_n = \int_E \xi_n dP$$

and thus $\xi_n = \mathbb{E}(\xi|\mathcal{F}_n)$. Hence condition (iv) of Theorem 14.3.3 is satisfied and from (i) and (ii) of the same theorem we have that $\{\xi_n\}_{n=1}^{\infty}$ is uniformly integrable on $(\Omega, \mathcal{F}, P)$, and $\xi_n \to \xi$ a.s.

Conversely, assume that the sequence $\{\xi_n\}_{n=1}^{\infty}$ is uniformly integrable on $(\Omega, \mathcal{F}, P)$. Then by Theorem 14.3.3, since $\{\xi_n, \mathcal{F}_n\}_{n=1}^{\infty}$ is a martingale on $(\Omega, \mathcal{F}, P)$, there is a r.v. $\xi \in L_1(\Omega, \mathcal{F}, P)$ such that $\xi_n = \mathbb{E}(\xi|\mathcal{F}_n)$ a.s.

for all $n$. It follows from Theorem 14.3.4 that

$$\xi_n = \mathbb{E}(\xi|\mathcal{F}_n) \to \mathbb{E}(\xi|\mathcal{F}) = \xi \text{ a.s. and in } L_1(\Omega, \mathcal{F}, P).$$

It now suffices to show that $Q \ll P$ and $\xi = \frac{dQ}{dP}$ a.s. Indeed for all $E \in \mathcal{F}_n$ we have

$$Q(E) = Q_n(E) = \int_E \xi_n dP = \int_E \mathbb{E}(\xi|\mathcal{F}_n) dP = \int_E \xi dP.$$ 

Hence $Q(E) = \int_E \xi dP$ for all $E \in \bigcup_{n=1}^{\infty} \mathcal{F}_n$ and since the class of sets for which it is true is clearly a $\sigma$-field, it follows that it is true for all $E \in \mathcal{F}$. Thus $Q \ll P$ and $\xi = \frac{dQ}{dP}$ a.s.  

Application of the theorem to the positive and negative parts in the Jordan decomposition of a finite signed measure gives the following result.

**Corollary 1** The theorem remains true if $Q$ is a finite signed measure.

We now show how Theorem 14.5.1 can be used in finding expressions for Radon–Nikodym derivatives.

**Corollary 2** Let $(\Omega, \mathcal{F}, P)$ be a probability space and $Q$ a finite signed measure on $\mathcal{F}$ such that $Q \ll P$. For every $n$ let $\{E_k^{(n)} : k \geq 1\}$ be a measurable partition of $\Omega$ (i.e. $\Omega = \bigcup_{k=1}^{\infty} E_k^{(n)}$ where the $E_k^{(n)}$ are disjoint sets in $\mathcal{F}$) and let $\mathcal{F}_n$ be the $\sigma$-field it generates. Assume that the partitions become finer as $n$ increases (i.e. each $E_i^{(n)}$ is the union of sets from $\{E_k^{(n+1)}\}$) so that $\mathcal{F}_n \subset \mathcal{F}_{n+1}$. If the partitions are such that $\mathcal{F} = \sigma(\bigcup_{n=1}^{\infty} \mathcal{F}_n)$, then

$$\frac{dQ}{dP}(\omega) = \lim_{n \to \infty} \frac{Q(E_k^{(n)}(\omega))}{P(E_k^{(n)}(\omega))} \text{ a.s. and in } L_1(\Omega, \mathcal{F}, P)$$

where for every $\omega$ and $n$, $k^{(n)}(\omega)$ is the unique $k$ such that $\omega \in E_k^{(n)}$. 

Proof This is obvious from the simple observation that
\[ \frac{dQ_n}{dP_n}(\omega) = \sum_{k=1}^{\infty} \frac{Q(E_k^{(n)})}{P(E_k^{(n)})} \chi_{E_k^{(n)}}(\omega) \text{ a.s.} \]
where \( \frac{Q(E_k^{(n)})}{P(E_k^{(n)})} \) is taken to be zero whenever \( P(E_k^{(n)}) = 0 \). \( \square \)

Since conditional expectations and conditional probabilities as defined in Chapter 13 are Radon–Nikodym derivatives of finite signed measures with respect to probability measures, Corollary 2 can be used to express them as limits and the resulting expressions are also intuitively appealing. Such a result will be stated for a conditional probability given the value of a r.v.

**Corollary 3** Let \( \eta \) be a r.v. on the probability space \((\Omega, \mathcal{F}, P)\) and \( A \in \mathcal{F} \). For each \( n \), let \( \{I_k^{(n)} : -\infty < k < \infty\} \) be a partition of the real line into intervals. Assume that the partitions become finer as \( n \) increases and that
\[ \delta^{(n)} = \sup_k m(I_k^{(n)}) \to 0 \text{ as } n \to \infty \]
\((m = \text{Lebesgue measure})\). Then
\[ P(A|\eta = y) = \lim_{n \to \infty} \frac{P(A \cap \eta^{-1}I_{k^\eta(y)}^{(n)})}{P(\eta^{-1}I_{k^\eta(y)}^{(n)})} \text{ a.s. } (P_{\eta^{-1}}) \text{ and in } L_1(\mathbb{R}, \mathcal{B}, P_{\eta^{-1}}) \]
where for each \( y \) and \( n \), \( k^\eta(y) \) is the unique \( k \) such that \( y \in I_k^{(n)} \).

**Proof** By Section 13.5, \( P(A|\eta = y) \) is the Radon–Nikodym derivative of the finite measure \( \nu \), defined for each \( B \in \mathcal{B} \) by \( \nu(B) = P(A \cap \eta^{-1}B) \), with respect to \( P_{\eta^{-1}} \). The result follows from Corollary 2 and the simple observation that if \( \mathcal{B}_n = \sigma(\bigcup_{k=\infty}^{-1}I_k^{(n)}) \) then \( \mathcal{B}_n \subset \mathcal{B}_{n+1} \) and \( \sigma(\bigcup_{n=1}^{\infty} \mathcal{B}_n) = \mathcal{B} \). \( \square \)

The second application concerns “likelihood ratios” and is related to the principle of maximum likelihood.

**Theorem 14.5.2** Let \( \{\xi_n\} \) be a sequence of r.v.’s on the probability space \((\Omega, \mathcal{F}, P)\), and \( \mathcal{F}_n = \sigma(\xi_1, \ldots, \xi_n) \). Let \( Q \) be another probability measure on \((\Omega, \mathcal{F})\). Assume that for every \( n, \ (\xi_1, \ldots, \xi_n) \) has p.d.f. \( p_n \) under the probability \( P \) and \( q_n \) under the probability \( Q \), and define
\[ \eta_n(\omega) = \begin{cases} \frac{q_n(\xi_1(\omega), \ldots, \xi_n(\omega))}{p_n(\xi_1(\omega), \ldots, \xi_n(\omega))} & \text{if the denominator } \neq 0 \\ 0 & \text{otherwise.} \end{cases} \]
Then \( \{ \eta_n, \mathcal{F}_n \}_{n=1}^{\infty} \) is a supermartingale on \((\Omega, \mathcal{F}, P)\) and there is a \( P \)-integrable r.v. \( \eta_{\infty} \) such that
\[
\eta_n \to \eta_{\infty} \text{ a.s.}
\]
and
\[
0 \leq E\eta_{\infty} \leq E\eta_{n+1} \leq E\eta_n \leq 1 \text{ for all } n.
\]

Proof Since \( p_n \) and \( q_n \) are Borel measurable functions, \( \eta_n \) is \( \mathcal{F}_n \)-measurable. Also \( \eta_n \geq 0 \). If \( A_n = \{(x_1, \ldots, x_n) \in \mathbb{R}^n : p_n(x_1, \ldots, x_n) > 0\} \) then \( P(\xi_1, \ldots, \xi_n)^{-1}(A_n^c) = 0 \) and thus \( P(\xi_1, \ldots, \xi_n, \xi_{n+1})^{-1}(A_{n+1}^c \times \mathbb{R}) = 0 \). Further
\[
E\eta_n = \int_{\Omega} \eta_n dP = \int_{\mathbb{R}^n} q_n \chi_{A_n} \frac{1}{p_n} dP(\xi_1, \ldots, \xi_n)^{-1}
\]
\[
= \int_{\mathbb{R}^n} \frac{q_n}{p_n} \chi_{A_n} p_n dx_1 \ldots dx_n = \int_{\mathbb{R}^n} q_n \chi_{A_n} dx_1 \ldots dx_n
\]
\[
\leq \int_{\mathbb{R}^n} q_n dx_1 \ldots dx_n = 1
\]
and thus \( 0 \leq E\eta_n \leq 1 \).

Also, for every \( E \in \mathcal{F}_n \) there is a \( B \in \mathcal{B}^n \) such that \( E = (\xi_1, \ldots, \xi_n)^{-1}(B) \) and
\[
\int_E \eta_{n+1} dP = \int_{\Omega} \eta_{n+1} \chi_E dP
\]
\[
= \int_{\mathbb{R}^n} q_{n+1} \chi_{A_{n+1}} \frac{1}{p_{n+1}} dP(\xi_1, \ldots, \xi_{n+1})^{-1}
\]
\[
= \int_{A_{n+1}} \frac{q_{n+1}}{p_{n+1}} \chi_{B} dP(\xi_1, \ldots, \xi_{n+1})^{-1}
\]
\[
= \int_{A_{n+1} - A_n \times \mathbb{R}} \frac{q_{n+1}}{p_{n+1}} \chi_{B} dP(\xi_1, \ldots, \xi_{n+1})^{-1}
\]
since \( P(\xi_1, \ldots, \xi_{n+1})^{-1}(A_{n+1}^c \times \mathbb{R}) = 0 \). Hence, since \( A_{n+1} - A_n \times \mathbb{R} \subset A_n \times \mathbb{R} \)
\[
\int_E \eta_{n+1} dP = \int_{A_{n+1} - A_n \times \mathbb{R}} q_{n+1} \chi_B dx_1 \ldots dx_n
\]
\[
\leq \int_{A_n \times \mathbb{R}} q_{n+1} \chi_B dx_1 \ldots dx_n
\]
\[
= \int_{A_n} \left( \int_{\mathbb{R}} q_{n+1}(x_1, \ldots, x_n, x_{n+1}) dx_{n+1} \right) \chi_B dx_1 \ldots dx_n
\]
\[
= \int_{A_n} q_n \chi_B dx_1 \ldots dx_n
\]
\[
= \int_{A_n} q_n \chi_B dP(\xi_1, \ldots, \xi_n)^{-1}
\]
\[
= \int_{\Omega} \eta_n \chi_E dP = \int_{E} \eta_n dP.
\]
It follows that \( E(\eta_{n+1} | \mathcal{F}_n) \leq \eta_n \) for all \( n \), a.s., and thus \( \{ \eta_n, \mathcal{F}_n \}_{n=1}^{\infty} \) is a supermartingale on \((\Omega, \mathcal{F}, P)\). Hence \( \{-\eta_n, \mathcal{F}_n \}_{n=1}^{\infty} \) is a negative submartingale which, by the submartingale convergence Theorem 14.3.1, converges a.s. to
14.5 Further applications

a $P$-integrable r.v. $\eta_\infty$. Then by Theorem 14.1.1 (ii) and the first result of this proof we have $0 \leq E\eta_{n+1} \leq E\eta_n \leq 1$ for all $n$. Finally by Fatou’s Lemma $E\eta_\infty \leq E\eta_n$ and this completes the proof. □

If for each $n$ the distribution of $(\xi_1, \ldots, \xi_n)$ under $Q$ is absolutely continuous with respect to its distribution under $P$ then the following stronger result holds.

**Corollary 1** Under the assumptions of Theorem 14.5.2, if for all $n$, $Q(\xi_1, \ldots, \xi_n)^{-1} \ll P(\xi_1, \ldots, \xi_n)^{-1}$ (which is the case if $q_n = 0$ whenever $p_n = 0$) and $\mathcal{F} = \sigma(\bigcup_{n=1}^{\infty} \mathcal{F}_n)$, then $\{\eta_n, \mathcal{F}_n\}$ is a martingale. Furthermore $Q \ll P$ if and only if $\{\eta_n\}$ is uniformly integrable in which case

$$\eta_n \to \frac{dQ}{dP} \text{ a.s. and in } L_1(\Omega, \mathcal{F}, P), \text{ as } n \to \infty.$$  

**Proof** For each $n$ let $Q_n, P_n$ be the restrictions of $Q, P$ to $\mathcal{F}_n$. For every $E \in \mathcal{F}_n$ we have $E = (\xi_1, \ldots, \xi_n)^{-1}(B), \ B \in \mathcal{B}^n$ and since by absolute continuity $P(\xi_1, \ldots, \xi_n)^{-1}(A_n^c) = 0$ implies $Q(\xi_1, \ldots, \xi_n)^{-1}(A_n^c) = 0$, we have

$$Q_n(E) = Q(\xi_1, \ldots, \xi_n)^{-1}(B) = Q(\xi_1, \ldots, \xi_n)^{-1}(B \cap A_n)$$

$$= \int_{B \cap A_n} q_n \, dx_1 \ldots \, dx_n$$

$$= \int_{B \cap A_n} \frac{q_n}{p_n} \, dP(\xi_1, \ldots, \xi_n)^{-1}$$

$$= \int_B \frac{q_n}{p_n} \, dP(\xi_1, \ldots, \xi_n)^{-1}$$

$$= \int_B \eta_n \, dP_n.$$  

Hence $\frac{dQ_n}{dP_n} = \eta_n$ and the result follows from Theorem 14.5.1. □

When the r.v.’s $\{\xi_n\}$ are i.i.d. under both $P$ and $Q$ the following result provides a test for the distribution of a r.v. using independent observations.

**Corollary 2** Assume that the conditions of Theorem 14.5.2 are satisfied and that under each probability measure $P$, $Q$ the r.v.’s $\{\xi_n\}$ are independent and identically distributed with (common) p.d.f. $p, q$. Then $\eta_n \to 0 \text{ a.s. and } P \perp Q$, provided the distributions determined by $p$ and $q$ are distinct.

**Proof** In this case we have

$$\eta_n = \prod_{k=1}^{n} \frac{q(\xi_k)}{p(\xi_k)} \text{ a.s. } (P)$$
and thus by Theorem 14.5.2,

$$\eta_{\infty} = \prod_{k=1}^{\infty} \frac{q(\xi_k)}{p(\xi_k)} \text{ a.s. (}\mathbf{P}\text{)}.$$

Now let $\{\xi'_n\}$ be an i.i.d. sequence of r.v.'s independent also of the sequence $\{\xi_n\}$, with the same distribution as the sequence $\{\xi_n\}$ (such r.v.'s can always be constructed using product spaces). Let also

$$\eta'_{\infty} = \prod_{k=1}^{\infty} \frac{q(\xi'_k)}{p(\xi'_k)} \text{ a.s. (}\mathbf{P}\text{)}.$$

Then $\eta_{\infty}$ and $\eta_{\infty} \eta'_{\infty}$ are clearly identically distributed and $\eta_{\infty} \eta'_{\infty}$ are independent and identically distributed so that

$$P[\eta_{\infty} = 0] = P[\eta_{\infty} \eta'_{\infty} = 0] = 1 - P[\eta_{\infty} \eta'_{\infty} > 0]$$

$$= 1 - P[\eta_{\infty} > 0]P[\eta'_{\infty} > 0]$$

$$= 1 - [1 - P[\eta_{\infty} = 0]]^2.$$

It follows that $P[\eta_{\infty} = 0] = 0$ or $1$.

Assume now that $P[\eta_{\infty} = 0] = 0$, so that $\eta_{\infty} > 0$ a.s. (P). Then the r.v.'s $\log(\eta_{\infty} \eta'_{\infty}) = \log \eta_{\infty} + \log \eta'_{\infty}$ are identically distributed and $\log \eta_{\infty}$, $\log \eta'_{\infty}$ are independent and identically distributed and thus if $\phi(t)$ is the c.f. of $\log \eta_{\infty}$ we have $\phi^2(t) = \phi(t)$ for all $t \in \mathbb{R}$. Since $\phi(0) = 1$ and $\phi$ is continuous, it follows that $\phi(t) = 1$ for all $t \in \mathbb{R}$ and thus $\eta_{\infty} = 1$ a.s. (P). It follows that $\prod_{k=1}^{\infty} \frac{q(\xi_k)}{p(\xi_k)} = 1$ a.s. (P) and thus $\eta_1 = \frac{q(\xi_1)}{p(\xi_1)} = 1$ a.s. Then for each $B \in \mathcal{B}$ we have, using the notation and facts from the proof of Corollary 1,

$$Q_{\xi_1^{-1}}(B) = Q_{\xi_1^{-1}}(B) = \int_{\xi_1^{-1}(B)} \eta_1 \ dP_1 = P_{1\xi_1^{-1}}(B) = P_{\xi_1^{-1}}(B)$$

which contradicts the assumption that the distributions of $\xi_1$ under $P$ and $Q$ are distinct. (In fact one can similarly show that $Q(\xi_1, \ldots, \xi_n)^{-1}(B) = P(\xi_1, \ldots, \xi_n)^{-1}(B)$ for all $B \in \mathcal{B}^n$ and all $n$, which implies that $P = Q$.)

Hence, under the assumptions of the theorem, $P[\eta_{\infty} = 0] = 1$ and the proof may be completed by showing that $P \perp Q$. By reversing the role of the probability measures $P$ and $Q$ we have that

$$\prod_{k=1}^{n} \frac{p(\xi_k)}{q(\xi_k)} \to 0 \text{ a.s. (}\mathbf{Q}\text{)}.$$

Let $E_Q$ be the set of $\omega \in \Omega$ such that $\prod_{k=1}^{n} \frac{p(\xi_k(\omega))}{q(\xi_k(\omega))} \to 0$ and $E_P$ the set of $\omega \in \Omega$ such that $\prod_{k=1}^{n} \frac{q(\xi_k(\omega))}{p(\xi_k(\omega))} \to 0$. Then $P(E_P) = 1 = Q(E_Q)$ and
Exercises

14.1 Let \( \{\xi_n, \mathcal{F}_n\} \) be a submartingale. Let the sequence of r.v.’s \( \{\epsilon_n\} \) be such that for all \( n \), \( \epsilon_n \) is \( \mathcal{F}_n \)-measurable and takes only the values 0 and 1. Define the sequence of r.v.’s \( \{\eta_n\} \) by

\[
\eta_1 = \xi_1 \\
\eta_{n+1} = \eta_n + \epsilon_n(\xi_{n+1} - \xi_n), \quad n \geq 1.
\]

Show that \( \{\eta_n, \mathcal{F}_n\} \) is also a submartingale and \( \mathbb{E}\eta_n \leq \mathbb{E}\xi_n \) for all \( n \). If \( \{\xi_n, \mathcal{F}_n\} \) is a martingale show that \( \{\eta_n, \mathcal{F}_n\} \) is also a martingale and \( \mathbb{E}\eta_n = \mathbb{E}\xi_n \) for all \( n \). (Do you see any gambling interpretation of this?)

14.2 Prove that every uniformly integrable submartingale \( \{\xi_n, \mathcal{F}_n\} \) can be uniquely decomposed in

\[
\xi_n = \eta_n + \zeta_n \quad \text{for all } n \text{ a.s.}
\]

where \( \{\eta_n, \mathcal{F}_n\} \) is a uniformly integrable martingale and \( \{\zeta_n, \mathcal{F}_n\} \) is a negative (\( \zeta_n \leq 0 \) for all \( n \) a.s.) submartingale such that \( \lim_n \zeta_n = 0 \) a.s. This is called the Riesz decomposition of a submartingale.

14.3 Let \( \{\mathcal{F}_n\} \) be a sequence of sub-\( \sigma \)-fields of \( \mathcal{F} \) such that \( \mathcal{F}_n \subset \mathcal{F}_{n+1} \) for all \( n \) and \( \mathcal{F}_\infty = \sigma(\bigcup_{n=1}^{\infty} \mathcal{F}_n) \). Show that if \( E \in \mathcal{F}_\infty \) then

\[
\lim_{n \to \infty} P(E|\mathcal{F}_n) = \chi_E \text{ a.s.}
\]

14.4 (Polya’s urn scheme) Suppose an urn contains \( b \) blue and \( r \) red balls. At each drawing a ball is drawn at random, its color is noted and the drawn ball together with \( a > 0 \) balls of the same color are added to the urn. Let \( b_n \) be the number of blue balls and \( r_n \) the number of red balls after the \( n \)th drawing and let \( \xi_n = b_n/(b_n + r_n) \) be the proportion of blue balls. Show that \( \{\xi_n\} \) is a martingale and that \( \xi_n \) converges a.s. and in \( L_1 \).

14.5 The inequalities proved in Theorems 14.2.1 and 14.2.2 for finite submartingales depend only on the fact that the submartingales considered have a “last element”. Specifically show that if \( \{\xi_n, \mathcal{F}_n : n = 1, 2, \ldots, \infty\} \) is a submartingale then for all real \( a \),

\[
aP\{ \sup_{1 \leq n \leq \infty} \xi_n \geq a \} \leq \int_{\{\sup_{1 \leq n \leq \infty} \xi_n \geq a\}} \mathbb{E}\xi_\infty \, dP \leq \mathbb{E}[\xi_\infty],
\]

and if also \( \xi_n \geq 0 \) a.s. for all \( n = 1, 2, \ldots, \infty \), then for all \( 1 < p < \infty \),

\[
\mathbb{E}(\sup_{1 \leq n \leq \infty} \xi_n^p) \leq \left(\frac{p}{p-1}\right)^p \mathbb{E}\xi_\infty^p.
\]
14.6 The following is an example of a martingale converging a.s. but not in $L_1$. Let $\Omega$ be the set of all positive integers, $\mathcal{F}$ the $\sigma$-field of all subsets of $\Omega$, and $P$ defined by

$$P(|n|) = \frac{1}{n} - \frac{1}{n+1} \quad \text{for all } n = 1, 2, \ldots.$$  

Let $[n, \infty)$ denote the set of all integers $\geq n$ and define

$$\mathcal{F}_n = \sigma(\{1\}, \{2\}, \ldots, \{n\}, \{n + 1, \infty\})$$

$$\xi_n = (n + 1)\chi_{(n+1, \infty)}$$

for $n = 1, 2, \ldots$. Show that $\{\xi_n, \mathcal{F}_n\}_{n=1}^\infty$ is a martingale with $\mathcal{E}\xi_n = 1$. Show also that $\xi_n$ converges a.s. (and find its limit) but not in $L_1$.

14.7 If $\{\xi_n, \mathcal{F}_n : n = 1, 2, \ldots, \infty\}$ is a nonnegative submartingale, show that $\{\xi_n, n = 1, 2, \ldots\}$ is uniformly integrable (cf. Theorem 14.3.2).

14.8 Let $\{\xi_n, \mathcal{F}_n\}_{n=1}^\infty$ be a martingale or a nonnegative submartingale. If

$$\lim_{n \to \infty} E(|\xi_n|^p) < \infty$$

for some $1 < p < \infty$, show that $\xi_n$ converges a.s. and in $L_p$. (Hint: Use Theorems 14.3.1 and 14.2.2.)

14.9 Let $(\Omega, \mathcal{F}, P)$ be a probability space and $\{\mathcal{F}_n\}_{n=1}^\infty$ a sequence of sub-$\sigma$-fields of $\mathcal{F}$ such that $\mathcal{F}_n \subset \mathcal{F}_{n+1}$ and $\mathcal{F} = \sigma(\bigcup_{n=1}^\infty \mathcal{F}_n)$. Let $Q$ be a finite measure on $(\Omega, \mathcal{F})$. Denote by $P_n, Q_n$ the restriction of $P, Q$ to $\mathcal{F}_n$ and the corresponding Lebesgue–Radon–Nikodym decomposition by

$$Q_n(E) = \int_E \xi_n \, dP_n + Q_n(E \cap N_n), \quad E \in \mathcal{F}_n$$

$$Q(E) = \int_E \xi \, dP + Q(E \cap N), \quad E \in \mathcal{F}$$

where $0 \leq \xi_n \in L_1(\Omega, \mathcal{F}_n, P_n)$, $0 \leq \xi \in L_1(\Omega, \mathcal{F}, P)$, $N_n \in \mathcal{F}_n$, $N \in \mathcal{F}$ and $P_n(N_n) = 0$, $P(N) = 0$. Show that $\{\xi_n, \mathcal{F}_n\}_{n=1}^\infty$ is a supermartingale and that $\xi_n \to \xi$ a.s. ($P$). (Hint: Imitate the proof of Theorem 14.5.1.)

14.10 Let $f$ be a Lebesgue integrable function defined on $[0, 1]$. For each $n$, let $0 = a^{(n)}_0 < a^{(n)}_1 < \ldots < a^{(n)}_n = 1$ be a partition of $[0, 1]$ with $\delta^{(n)} = \sup_{0 \leq k \leq n-1}(a^{(n)}_{k+1} - a^{(n)}_k) \to 0$, and assume that the partitions become finer as $n$ increases. For each $n$, define $f_n$ on $[0, 1]$ by

$$f_n(x) = \frac{1}{a^{(n)}_{k+1} - a^{(n)}_k} \int_{a^{(n)}_k}^{a^{(n)}_{k+1}} f(y) \, dy$$

for $a^{(n)}_k < x \leq a^{(n)}_{k+1}$ and by continuity at $x = 0$. Then show that

$$\lim_{n \to \infty} f_n(x) = f(x) \quad \text{a.e. (m) and in } L_1 \quad (m = \text{Lebesgue measure}).$$

14.11 Let $(\Omega, \mathcal{F})$ be a measurable space and assume that $\mathcal{F}$ is purely atomic, i.e. $\mathcal{F}$ is generated by the disjoint sets $\{E_n\}_{n=1}^\infty$ with $\Omega = \bigcup_{n=1}^\infty E_n$. Let $\{T, \mathcal{T}\}$ be another measurable space, $\{P_t, t \in T\}$ a family of probability measures
on \((\Omega, \mathcal{F})\) and \(\{Q_t, t \in T\}\) a family of signed measures on \((\Omega, \mathcal{F})\). Assume that for each \(t \in T\), \(Q_t \ll P_t\) and that for each \(E \in \mathcal{F}\), \(P_t(E)\) and \(Q_t(E)\) are measurable functions on \((T, \mathcal{T})\). Show that there is a \(T \times \mathcal{F}\)-measurable function \(\xi(t, \omega)\) such that for each fixed \(t \in T\),

\[
\xi(t, \omega) = \frac{dQ_t}{dP_t}(\omega) \text{ a.s. } (P_t).
\]

(Hint: Apply Theorem 14.5.1 with \(\mathcal{F}_n = \sigma(E_1, \ldots, E_n)\).)
Basic structure of stochastic processes

Our aim in this final chapter is to indicate how basic distributional theory for stochastic processes, alias random functions, may be developed from the considerations of Chapters 7 and 9. This is primarily for reference and for readers with a potential interest in the topic. The theory will be first illustrated by a discussion of the definition of the Wiener process, and conditions for sample function continuity. This will be complemented, and the chapter completed with a sketch of construction and basic properties of point processes and random measures in a purely measure-theoretic framework, consistent with the nontopological flavor of the entire volume.

15.1 Random functions and stochastic processes

In this section we introduce some basic distributional theory for stochastic processes and random functions, using the product space measures of Chapter 7 and the random element concepts of Chapter 9.

By a stochastic process one traditionally means a family of real random variables \( \{\xi_t : t \in T\} \) (\( \xi_t = \xi_t(\omega) \)) on a probability space \((\Omega, \mathcal{F}, P)\), \(T\) being a set indexing the \(\xi_t\). If \(T = \{1, 2, 3, \ldots\}\) or \(\{\ldots, -2, -1, 0, 1, 2, \ldots\}\) the family \(\{\xi_n : n = 1, 2, \ldots\}\) or \(\{\xi_n : n = \ldots, -2, -1, 0, 1, 2, \ldots\}\) is referred to as a stochastic sequence or discrete parameter stochastic process, whereas \(\{\xi_t : t \in T\}\) is termed a continuous parameter stochastic process if \(T\) is an interval (finite or infinite).

We assume throughout this chapter that each r.v. \(\xi_t(\omega)\) is defined (and finite) for all \(\omega\) (not just a.e.). Then for a fixed \(\omega\) the values \(\xi_t(\omega)\) define a function \(\xi((\xi(\omega))(t) = \xi_t(\omega), t \in T)\) in \(\mathbb{R}^T\) and the \(\mathcal{F}|\mathcal{B}\)-measurability of each \(\xi_t(\omega)\) implies \(\mathcal{F}|\mathcal{B}^T\)-measurability of \(\xi\) as will be shown in Lemma 15.1.1. The mapping \(\xi\) is thus a random element (r.e.) of \((\mathbb{R}^T, \mathcal{B}^T)\) and is termed a random function (r.f.). As will be seen in Lemma 15.1.1 the converse also holds – if \(\xi\) is a measurable mapping from \((\Omega, \mathcal{F}, P)\) to \((\mathbb{R}^T, \mathcal{B}^T)\) then the \(\omega\)-functions \(\xi_t(\omega) = (\xi(\omega))(t)\) are \(\mathcal{F}|\mathcal{B}\)-measurable for each \(t\), i.e. \(\xi_t\) are
r.v.’s. Thus the notions of a stochastic process (family of r.v.’s) and a r.f. are entirely equivalent. For a fixed \( \omega \), the function \((\xi_\omega)(t), t \in T\), is termed a sample function (or sample path or realization) of the process.

**Lemma 15.1.1** For each \( t \in T \), let \( \xi_t = \xi_t(\omega) \) be a real function of \( \omega \in \Omega \) and let \( \xi \) be the mapping from \( \Omega \) to \( \mathbb{R}^T \) defined as \( \xi_\omega = \{ \xi_t(\omega) : t \in T \} \). Then \( \xi_t \) is \( \mathcal{F}|\mathcal{B} \)-measurable for each \( t \in T \) iff \( \xi \) is \( \mathcal{F}|\mathcal{B}^T \)-measurable (see Section 7.9 for the definition of \( \mathcal{B}^T \)).

**Proof** For \( u = (t_1, \ldots, t_k) \) the projection \( \pi_u = \pi_{t_1, \ldots, t_k} \) from \( \mathbb{R}^T \) to \( \mathbb{R}^k \) is clearly \( \mathcal{B}^T|\mathcal{B}^k \)-measurable since if \( B \in \mathcal{B}^k \), \( \pi_u^{-1}B \) is a cylinder and hence is in \( \mathcal{B}^T \). Hence if \( \xi \in \mathcal{F}|\mathcal{B}^T \)-measurable, \( \xi_t = \pi_t \xi \) is \( \mathcal{F}|\mathcal{B} \)-measurable for each \( t \).

Conversely if each \( \xi_t \) is \( \mathcal{F}|\mathcal{B} \)-measurable, \( (\xi_{t_1}, \ldots, \xi_{t_k}) \) is clearly \( \mathcal{F}|\mathcal{B}^k \)-measurable, i.e. \( \pi_u \xi \) is \( \mathcal{F}|\mathcal{B}^k \)-measurable for \( u = (t_1, \ldots, t_k) \). Hence if \( B \in \mathcal{B}^k \), \( \xi^{-1} \pi_u^{-1}B = (\pi_u \xi)^{-1}B \in \mathcal{F} \) or \( \xi^{-1}E \in \mathcal{F} \) for each cylinder \( E \). Since these cylinders generate \( \mathcal{B}^T \), it follows that \( \xi \) is \( \mathcal{F}|\mathcal{B}^T \)-measurable as required. \( \square \)

Probabilistic properties of individual \( \xi_t \) or finite groups \((\xi_{t_1}, \ldots, \xi_{t_k})\) are, of course, defined by the respective marginal or joint distributions

\[
P_{\xi_t^{-1}}(B) = P\{\omega : \xi_t(\omega) \in B\}, B \in \mathcal{B},
\]

\[
P((\xi_{t_1}, \ldots, \xi_{t_k})^{-1}(B) = P\{\omega : (\xi_{t_1}(\omega), \ldots, \xi_{t_k}(\omega)) \in B\}, B \in \mathcal{B}^k.
\]

These are respectively read as \( P(\xi_t \in B) \), \( P((\xi_{t_1}, \ldots, \xi_{t_k}) \in B) \) and are as noted Lebesgue–Stieltjes measures on \( \mathcal{B} \) and \( \mathcal{B}^k \) corresponding to the distribution functions

\[
F_t(x) = P(\xi_t \leq x), \quad F_{t_1, \ldots, t_k}(x_1, \ldots, x_k) = P(\xi_{t_i} \leq x_i, 1 \leq i \leq k).
\]

These joint distributions of \( \xi_{t_1}, \ldots, \xi_{t_k} \) for \( t_i \in T, 1 \leq i \leq k, k = 1, 2, \ldots, \) are termed the finite-dimensional distributions (fidi’s) of the process \( \{\xi_t : t \in T\} \).

The fidi’s determine many useful probabilistic properties of the process but are restricted to probabilities of sets of values taken by finite groups of \( \xi_t \)’s. On the other hand, one may be interested in the probability that the entire sample function \( \xi_t, t \in T \), lies in a given set of functions, i.e.

\[
P(\xi \in E) = P(\omega : \xi(\omega) \in E) = P_{\xi^{-1}}(E)
\]

which is defined for \( E \in \mathcal{B}^T \). Further assumptions may be needed for sets \( E \) of interest but not in \( \mathcal{B}^T \), e.g. to determine that the sample functions are continuous a.s. (see Sections 15.3, 15.4).
This probability measure $P_{\xi^{-1}}$ on $\mathcal{B}^T$ is called the distribution of (the r.f.) $\xi$ and it encompasses the fidi’s. Specifically, the fidi’s are special cases of values of $P_{\xi^{-1}}$, for example, if $B \in \mathcal{B}^k$

$$P\{ (\xi_{t_1}, \ldots, \xi_{t_k}) \in B \} = P\{ \pi_{t_1, \ldots, t_k} \xi \in B \} = P_{\xi^{-1}}(\pi_{t_1, \ldots, t_k} B)$$

i.e. the probability that the sample function $\xi_\omega$ lies in the cylinder $\pi_{t_1, \ldots, t_k} B$ of $\mathcal{B}^T$. That is the fidi’s have the form $P_{\xi^{-1}}\pi_{t_1, \ldots, t_k}$ for each $k, t_1, \ldots, t_k \in T$. On the other hand, note also that the fidi’s determine the distribution of a stochastic process, that is, if two stochastic processes have the same fidi’s, then they have the same distribution. This follows from Theorem 2.2.7 and the fact that $\mathcal{B}^T$ is generated by the cylinders $\pi_{t_1, \ldots, t_k}(B)$.

The fidi’s of a stochastic process are thus related to the distribution $P_{\xi^{-1}}$ of $\xi$ on $\mathcal{B}^T$ exactly as the measures $\nu_u$ are related to $\mu$ in Section 7.10. In particular the fidi’s are consistent as there defined, i.e. if $u = (t_1, \ldots, t_k), v = (s_1, \ldots, s_l) \subset u$, $\xi_u = (\xi_{t_1}, \ldots, \xi_{t_k}), \xi_v = (\xi_{s_1}, \ldots, \xi_{s_l})$, then $P_{\xi_{t_1}^{-1}}\pi_{u,v}^{-1} = P_{\xi_v^{-1}}^{-1}$, i.e. $P(\pi_{u,v}^{-1})^{-1} = P_{\xi_v^{-1}}^{-1}$. This may be made more transparent by noting its equivalence to consistency of the d.f.’s in the sense that for each $n = 1, 2, \ldots$ any choice of $t_1, \ldots, t_n$ and $x_1, \ldots, x_n$

(i) $F_{t_1, \ldots, t_n}(x_1, \ldots, x_n)$ is unaltered by the same permutation of both $t_1, \ldots, t_n$ and $x_1, \ldots, x_n$,

(ii) $F_{t_1, \ldots, t_n-1}(x_1, \ldots, x_{n-1}) = F_{t_1, \ldots, t_{n-1}, t_n}(x_1, \ldots, x_{n-1}, \infty) = \lim_{x_n \to \infty} F_{t_1, \ldots, t_{n-1}, t_n}(x_1, \ldots, x_{n-1}, x_n)$. The requirement (i) can of course be achieved (on the real line) by defining $F_{t_1, \ldots, t_n}$ for $t_1 < \cdots < t_n$ and rearranging other time sets to natural order, and hence is not an issue when $T$ is a subset of $\mathbb{R}$.

Kolmogorov’s Theorem (Theorem 7.10.3) may then be put in the following form.

**Theorem 15.1.2** Let $\{\nu_u\}$ be as in Theorem 7.10.3, a family of probability measures defined on $(\mathbb{R}^u, \mathcal{B}^u)$ for finite subsets $u$ of an index set $T$. If the family $\{\nu_u\}$ is consistent in the sense that $\nu_u\pi_{u,v}^{-1} = \nu_v$ for each $u, v$ with $v \subset u$, then there is a stochastic process $\{\xi_t : t \in T\}$ (unique in distribution) having $\{\nu_u\}$ as its fidi’s. That is $P\{ (\xi_{t_1}, \ldots, \xi_{t_k}) \in B \} = \nu_u(B)$ for each choice of $k, u = (t_1, \ldots, t_k), B \in \mathcal{B}^k$.

**Proof** Let $P$ denote the unique probability measure on $(\mathbb{R}^T, \mathcal{B}^T)$ in Theorem 7.10.3, satisfying $P\pi_u^{-1} = \nu_u$ for each finite set $u \subset T$. Define the probability space $(\Omega, \mathcal{F}, P)$ as $(\mathbb{R}^T, \mathcal{B}^T, P)$. The projection r.v.’s $\xi_t(\omega) = \pi_t \omega = \omega(t)$ for $\omega \in \mathbb{R}^T$ give the desired stochastic process $\{\xi_t : t \in T\}$ with the given fidi’s $\nu_u$. \(\Box\)
Corollary 1 below restates the theorem in terms of distribution functions. Corollary 2 considers the special case of an independent family.

**Corollary 1** Let \( \{F_{t_1, \ldots, t_k} : t_1, \ldots, t_k \in T, k = 1, 2, \ldots \} \) be a family of \( k \)-dimensional d.f.’s, assumed consistent in the sense described prior to the statement of the theorem. Then there is a stochastic process \( \{\xi_t : t \in T\} \) having these d.f.’s defining its fidi’s, i.e.

\[
P(\xi_{t_i} \leq x_i, 1 \leq i \leq k) = F_{t_1, \ldots, t_k}(x_1, \ldots, x_k)
\]

for each choice of \( k, t_1, \ldots, t_k \).

**Proof** This follows since the d.f.’s \( F_{t_1, \ldots, t_k} \) clearly determine consistent probability distributions \( \nu_u \) for each \( u = (t_1, \ldots, t_k) \).

**Corollary 2** If \( F_i \) are d.f.’s for \( i = 1, 2, \ldots \), there exists a sequence of independent r.v.’s \( \xi_1, \xi_2, \ldots \) such that \( \xi_i \) has d.f. \( F_i \) for each \( i \).

**Proof** This follows from Corollary 1 by noting consistency of the d.f.’s

\[
F_{t_1, \ldots, t_k}(x_1, \ldots, x_k) = \prod_{i=1}^{k} F_{t_i}(x_i).
\]

### 15.2 Construction of the Wiener process in \( \mathbb{R}^{[0,1]} \)

The Wiener process \( W_t \) on \([0, 1]\) (a.k.a. Brownian motion) provides an illuminating and straightforward example of the use of Kolmogorov’s Theorem to construct a stochastic process.

\( W_t \) is to be defined by the requirement that all its fidi’s be normal with zero means and \( \text{cov}(W_s, W_t) = \min(s, t) \). Thus the fidi for \((W_{t_1}, W_{t_2}, \ldots, W_{t_k})\), \( 0 \leq t_1 < t_2 < \cdots < t_k \leq 1 \), is to be normal, with zero means and covariance matrix (see Section 9.4)

\[
\Lambda_{t_1, \ldots, t_k} = \begin{bmatrix}
t_1 & t_1 & \cdots & t_1 \\
t_1 & t_2 & \cdots & t_2 \\
t_1 & t_2 & t_3 & \cdots & t_3 \\
\vdots & \vdots & \ddots & \vdots \\
t_1 & t_2 & t_3 & \cdots & t_k
\end{bmatrix}.
\]

This matrix is readily seen to be nonnegative definite (e.g. its determinant is \( t_1(t_2 - t_1)(t_3 - t_2) \cdots (t_k - t_{k-1}) \) as may be simply shown by subtracting the \((i - 1)\)th row from the \(i\)th for \( i = k, k - 1, \ldots, 2 \)). Thus \( \Lambda_{t_1, \ldots, t_k} \) is a covariance matrix of a \( k \)-dimensional normal distribution, and the elimination of one or more points \( t_j \) gives a matrix of the same form in the remaining
Basic structure of stochastic processes

$t_j$’s, showing the consistency required for Kolmogorov’s Theorem (or Theorem 15.1.2). Hence, by that theorem, there is a process $\{W_t : t \in [0, 1]\}$ with the desired $\xi$’s.

15.3 Processes on special subspaces of $\mathbb{R}^T$

A stochastic process $\xi$ constructed via Kolmogorov’s Theorem is a random element of $(\mathbb{R}^T, \mathcal{B}^T)$. Hence one may determine the probability $P\{\xi \in E\}$ that the sample function $\xi_t$, $t \in T$, lies in the set $E$ of functions, for any $E \in \mathcal{B}^T$. However, one is sometimes interested in sets $E$ which are not in $\mathcal{B}^T$ (as, for example, when $T = [0, 1]$, $E = C[0, 1]$, the set of continuous functions on $[0, 1]$).

A small but useful extension to the framework occurs when $\xi \in A$ a.s. where $A \subset \mathbb{R}^T$ but $A$ may or may not be in $\mathcal{B}^T$. Note that the statement $\xi \in A$ a.s. means that $A^c \subset A_0$ for some $A_0 \in \mathcal{B}^T$, $P\xi^{-1}(A_0) = 0$. The extension may be simply achieved by assuming that the space $(\Omega, \mathcal{F}, P)$ is complete (or if not, by completing it to be so in the standard manner – see Section 2.6). Then with $A, A_0$ as above $\xi^{-1}A^c \in \mathcal{F}$ since $P$ is complete on $\mathcal{F}$. Hence also $\xi^{-1}A \in \mathcal{F}$, $P\xi^{-1}(A^c) = 0$ and $\xi^{-1}(A \cap E) = \xi^{-1}A \cap \xi^{-1}E \in \mathcal{F}$ for all $E \in \mathcal{B}^T$.

Hence if $\xi_t$, $t \in T$, is redefined as a fixed function in $A$ at points $\omega \in \Omega$ for which $\{\xi_t(\omega) : t \in T\} \notin A$ (or if the space $\Omega$ is reduced to eliminate such points), then $A$ includes all the values of $(\xi_t(\omega) : t \in T)$ and may be regarded as a space with a $\sigma$-field $\mathcal{A} = A \cap \mathcal{B}^T$. $\xi$ is then a random element in $(A, \mathcal{A})$ with distributions satisfying $P\xi^{-1}(F) = P\xi^{-1}(E)$ for $F = E \cap A$, $E \in \mathcal{B}^T$.

An interesting and useful special case occurs when $T$ is an interval and $A$ is the set of real, continuous functions on $T$. For example, take $T$ to be the unit interval $[0, 1]$ (with standard notation $A = C[0, 1]$, the space of continuous functions on $[0, 1]$). If a stochastic process $\{\xi_t : t \in [0, 1]\}$ has a.s. continuous sample functions (i.e. $\xi_t(\omega)$ is continuous on $0 \leq t \leq 1$ a.s.), then the r.f. $\xi$ may be regarded as a random element of $(C, C)$ where $C = C[0, 1] (\subset \mathbb{R}^{[0,1]}$) and $C = C \cap \mathcal{B}^{[0,1]}$. This is a natural and simple viewpoint.

It is, of course, possible to regard $C$ as a space of continuous functions, without reference to $\mathbb{R}^T$, and to view it as a metric space, with metric defined by the norm $(||x|| = \sup\{|x(t)| : 0 \leq t \leq 1\})$. The class of Borel sets of such a topological space is then defined to be the $\sigma$-field generated by the open sets. This may be shown to be also generated by the (finite-dimensional) cylinder sets of $C$, i.e. sets of the form $\pi_t^{-1}B$ where $B \in \mathcal{B}^k$. 

and $\pi_{t_1,\ldots,t_n}$ is the usual projection mapping but restricted to $C$ rather than $\mathbb{R}^T$. It may thus be seen that the Borel sets form precisely the same $\sigma$-field $C \cap \mathcal{B}^T$ in $C$ as defined and used above. This connection provides a vehicle for the consideration of properties which involve topology more intimately – such as the development of weak convergence theory in $C$.

### 15.4 Conditions for continuity of sample functions

In view of the above discussion it is of interest to give conditions on a process which will guarantee a.s. continuity of sample functions. The theorem to be shown, generalizing original results of Kolmogorov (see [Loève] and [Cramér & Leadbetter]) gives sufficient conditions for a process $\xi_t$ on $[0, 1]$ to have an equivalent version $\eta_t$ (i.e. $\xi_t = \eta_t$ a.s. for each $t$) with a.s. continuous sample functions.

**Theorem 15.4.1** Let $\xi_t$ be a process on $[0, 1]$ such that for all $t$, $t + h \in [0, 1]$

$$P(|\xi_{t+h} - \xi_t| \geq g(h)) \leq q(h)$$

where $g$, $q$ are nonnegative functions of $h > 0$, nonincreasing as $h \downarrow 0$ and such that $\sum g(2^{-n}) < \infty$, $\sum 2^n q(2^{-n}) < \infty$. Then there exists a process $\eta_t$ on $[0, 1]$ with a.s. continuous sample functions and such that $\xi_t = \eta_t$ a.s. for each $t$. In particular, of course, $\eta$ has the same fidi’s as $\xi$.

**Proof** Approximate $\xi_t$ by piecewise linear processes $\xi_t^n$ with the values $\xi_t$ at $t = t_{n,r} = r/2^n$, $r = 0, 1, \ldots, 2^n$, and linear between such points. Then clearly for $t_{n,r} \leq t \leq t_{n,r+1}$,

$$|\xi_t^{n+1} - \xi_t^n| \leq |\xi_{t_{n+1,2r+1}} - 1/2(\xi_{t_{n+1,2r}} + \xi_{t_{n+1,2r+2}}) - A + 1/2B$$

where

$$A = |\xi_{t_{n+1,2r+1}} - \xi_{t_{n+1,2r+2}}|, \quad B = |\xi_{t_{n+1,2r+1}} - \xi_{t_{n+1,2r+2}}|$$

and hence

$$P\left(\max_{t_{n,r} \leq t \leq t_{n,r+1}} |\xi_t^{n+1} - \xi_t^n| \geq g(2^{-n+1})\right) \leq P(A \geq g(2^{-n+1})) + P(B \geq g(2^{-n+1}))$$

$$\leq 2q(2^{-n+1})$$

so that

$$P\left(\max_{0 \leq t \leq 1} |\xi_t^{n+1} - \xi_t^n| \geq g(2^{-n+1})\right) \leq 2^{n+1}q(2^{-n+1}).$$

Since $\sum 2^n q(2^{-n}) < \infty$ it follows by the Borel–Cantelli Lemma (Theorem 10.5.1) that a.s., $\max_{0 \leq t \leq 1} |\xi_t^{n+1} - \xi_t^n| < g(2^{-n+1})$ for $n \geq n_0 = n_0(\omega)$. 

Since $\sum g(2^{-n}) < \infty$ it follows that $\{\xi^n_t\}$ is uniformly Cauchy a.s. and thus uniformly convergent a.s. to a continuous $\eta_t$ as $n \to \infty$. Also $\eta_t = \xi_t$ a.s. for $t = t_{n,r}$ since $\xi_t^{n,p} = \xi_t$, $p = 0, 1, \dotsc$.

If $t$ is not equal to any $t_{n,r}$, $t = \lim t_{n,r}$, then

$$P(\{|\xi_{t_{n,r}} - \xi_t| \geq g(t - t_{n,r})\} \leq q(t - t_{n,r}) \leq q(2^{-n})$$

so that $P(\{|\xi_{t_{n,r}} - \xi_t| \geq g(2^{-n})\} \leq q(2^{-n})$ and the Borel–Cantelli Lemma gives $\xi_{t_{n,r}} \to \xi_t$ a.s.

Since $\eta_{t_{n,r}} \to \eta_t$ a.s. and $\xi_{t_{n,r}} = \eta_{t_{n,r}}$ a.s., it follows that $\xi_t = \eta_t$ a.s. for each $t$ as required.

\[\square\]

## 15.5 The Wiener process on $C$ and Wiener measure

The preceding theorem readily applies to the Wiener process yielding the following result.

**Theorem 15.5.1** The Wiener process $\{W_t : t \in [0, 1]\}$ may be taken to have a.s. continuous sample functions.

**Proof** This follows from the above result. For $W_{t+h} - W_t$ is normal, zero mean and variance $|h|$. Take $0 < a < 1/2$. Then

$$P(|W_{t+h} - W_t| \geq |h|^a) = 2\{1 - \Phi(|h|^{a-1/2})\} \leq 2|h|^{1/2-a} \phi(|h|^{a-1/2})$$

(where $\Phi$, $\phi$ are the standard normal d.f. and p.d.f. respectively) since $1 - \Phi(x) \leq \phi(x)/x$ for $x > 0$. If $g(h) = |h|^a$, $q(h) = 2|h|^{1/2-a} \phi(|h|^{a-1/2})$ then

$$\sum g(2^{-n}) = \sum 2^{-na} < \infty, \quad \sum 2^n q(2^{-n}) = 2 \sum 2^n(1+2a)/2 \phi(2^{n(1-2a)/2}) < \infty$$

(the last convergence being easily checked). Hence a.s. continuity of (an equivalent version of) $W_t$ follows from Theorem 15.4.1.

\[\square\]

As seen in Section 15.3, a process with a.s. continuous sample functions may be naturally viewed as a random element of $(C, C)$ where $C = C[0, 1]$, and $C = C \cap B^{[0,1]}$. By Theorem 15.5.1, the Wiener process $W_t$ may be so regarded. The steps in the construction were (a) to use Kolmogorov’s Theorem to define a process, say $W_t^0$ in $(\mathbb{R}^T, B^T)$ having the prescribed (normal) fidi’s, (b) to replace $W_t^0$ by an equivalent version $W_t$ with a.s. continuous sample functions, i.e. $W_t = W_t^0$ a.s. for each $t$ (hence with the same fidi’s), and (c) to consider $W = \{W_t : t \in [0, 1]\}$ as a random element of $(C, C)$ by restricting to $C = C[0, 1]$ (and taking $C = C \cap B^{[0,1]}$, equivalently the Borel $\sigma$-field of the topological space $C$ as noted in Section 15.3).
15.6 Point processes and random measures

As a result of this construction a probability measure $PW^{-1}$ (the distribution of $W$) is obtained on the measurable space $(C, C)$. This probability measure is termed Wiener measure and is customarily also denoted by $W$. This measure has, of course, multivariate normal form for the fidi probabilities induced on the sets $B^u, u = (t_1, \ldots, t_k)$ for each $k$. Of course, the space $(C, C, W)$ can be used to be the $(\Omega, F, P)$ on which the Wiener process is defined as the identity mapping $W\omega = \omega$.

Finally, it may be noted that an alternative approach to Wiener measure and the Wiener process is to define the latter as a distributional limit of simple processes of random walk type (cf. [Billingsley]). This is less direct and does require considerable weak convergence machinery but has the advantage of simultaneously producing the “invariance principle” (functional central limit theorem) of Donsker, which has significant use e.g. in applications to areas such as sequential analysis.

15.6 Point processes and random measures

In the preceding sections we have indicated some basic structural theory for stochastic processes with continuous sample functions and given useful sufficient conditions for continuity. This included the construction and continuity of the celebrated Wiener process – a key component along with its various extensions in stochastic modeling in diverse fields.

At the other end of the spectrum are processes whose sample functions are patently discontinuous, which may be used to model random sequences of points (i.e. point processes) and their extensions to more general random measures. A special position among these is held by the Poisson process which is arguably equally as prominent as the Wiener process for its extensions and applications.

There are a number of ways of providing a framework for point processes on the (e.g. positive) real line, perhaps the most obvious being the description as a family $\{\tau_n : n = 0, 1, 2, \ldots\}$ of r.v.’s $0 \leq \tau_1 \leq \tau_2 \leq \cdots$ (defined on $(\Omega, F, P)$), representing the positions of points. To avoid accumulation points it is assumed that $\tau_n \to \infty$ a.s. In particular the assumption that $\tau_1, \tau_2 - \tau_1, \tau_3 - \tau_2, \ldots$ are independent and identically distributed with d.f. $F(\cdot)$ leads to a renewal process and the particular case $F(x) = 1 - e^{-\lambda x}$, $x > 0$, gives a Poisson process with intensity $\lambda$. Fine detailed accounts of these and related processes abound, of which, for example [Feller] may be regarded as a seminal work. Our purpose here is just to indicate how a general abstract framework may arise naturally by adding randomness to the measure-theoretic structure considered throughout this volume in line
with the random element approach to real-valued processes of the preceding sections.

An alternative viewpoint to that above of regarding a point process as the sequence \( \{ \tau_n : 0 < \tau_1 < \tau_2 < \cdots \} \) of its point occurrence times is to consider the family of (extended) r.v.’s \( \xi(B) \) taking values 0, 1, 2, \ldots, \(+\infty\), consisting of the numbers of \( \tau_i \) in (Borel) sets \( B \). The assumption \( \tau_n \to \infty \) means that \( \xi(B) < \infty \) for bounded Borel sets \( B \). Since \( \xi(B) \) is clearly countably additive, it may be regarded as a (random) counting measure on the Borel sets of \([0, \infty)\). The two alternative viewpoints are connected e.g. by relation \( \{ \xi(0, x] \geq n \} = \{ \tau_n \leq x \} \). A simple Poisson process with intensity \( \lambda \) may then be regarded as a random counting measure \( \xi(B) \) as above with \( P[\xi(B) = r] = e^{-\lambda m(B)}(\lambda m(B))^r/r! \) \((m = \text{Lebesgue measure as always})\) for each Borel \( B \subset [0, \infty) \) and such that \( \xi(B_1), \xi(B_2) \) are independent for disjoint such \( B_1, B_2 \).

It is natural to extend this latter view of a point process (a) to include \( \xi(B) \) which are not necessarily integer-valued (i.e. to define random measures (r.m.’s) which are not necessarily point processes) and (b) to consider such concepts on a space more general than the real line, such as \( \mathbb{R}^k \) or a space \( S \) with a topological structure. A detailed, encyclopedic account of r.m.’s may be found in [Kallenberg] for certain metric (“Polish”) spaces. The topological assumptions involved are most useful for consideration of more intricate properties (such as weak convergence) of point processes and r.m.’s. However, for the basic r.m. framework they are primarily used to define a purely measure-theoretic structure involving classes of sets (semirings, rings, \( \sigma \)-fields) considered without topology in this volume. Hence our preferred approach in this brief introduction is to define a “clean” purely measure-theoretic framework in the spirit of this volume, leaving topological consideration for possible later study and as a setting for development of more complex properties of interest.

Our interest in the possible use of a measure-theoretic framework arose from hearing a splendid lecture series on random measures in the early 1970’s by Olav Kallenberg – leading to his subsequent classic book [Kallenberg]. Similar developments were also of interest to others at that time and since – including papers by D.G. Kendall, B.D. Ripley, J. Mecke and a subsequent book on the Poisson processes by J.F.C. Kingman.

### 15.7 A purely measure-theoretic framework for r.m.’s

Let \( S \) be an abstract space on which a r.m. is to be defined and \( S \) a \( \sigma \)-field of subsets of \( S \), i.e. \((S, S)\) is a measurable space (Chapter 3). Our basic
15.7 A purely measure-theoretic framework for r.m.’s

structural assumption about \( S \) is that there is a countable semiring \( \mathcal{P} \) in \( S \) whose members cover \( S \) (i.e., if \( \mathcal{P} = \{E_1, E_2, \ldots\} \), \( \bigcup_1^n E_i = S \)) and such that \( \mathcal{P} \) generates \( S \) (i.e. \( S(\mathcal{P}) = S \)). Note that since \( S = \bigcup_1^n E_i \in S(\mathcal{P}) = S \), \( \mathcal{P} \) also generates \( S \) as a \( \sigma \)-field (\( \sigma(\mathcal{P}) = S(\mathcal{P}) = S \)). We shall refer to a system \((S, S, \mathcal{P})\) satisfying these assumptions as a basic structure for defining a random measure or point process.

Two rings connected with such a basic structure are of interest:

(i) \( \mathcal{R}(\mathcal{P}) \), the ring generated by \( \mathcal{P} \), i.e. the class of all finite (disjoint) unions of sets of \( \mathcal{P} \).

(ii) \( S_0 = S_0(\mathcal{P}) \), the class of all sets \( E \in S \) such that \( E \subset \bigcup_1^n E_i \) for some \( n \) and sets \( E_1, E_2, \ldots, E_n \) in \( \mathcal{P} \).

\( S_0 \) is clearly a ring and \( \mathcal{P} \subset \mathcal{R}(\mathcal{P}) \subset S_0 \subset S \). The ring \( S_0 \) will be referred to as the class of bounded measurable sets, since they play this role in the real line, where \( \mathcal{P} = \{(a, b] : a, b \text{ rational}, -\infty < a < b < \infty\} \). This is incidentally also the case in popular topological frameworks, e.g. where \( S \) is a second countable locally compact Hausdorff space, \( S \) is the class of Borel sets (generated by the open sets) and \( \mathcal{P} \) is the ring generated by a countable base of bounded sets.

In these examples, the ring \( S_0 \) is precisely the class of all bounded measurable sets. As noted \( S_0 \) will be referred to as the “class of bounded measurable sets” even in the general context.

Let \((S, S, \mathcal{P})\) be a basic structure, and \((\Omega, \mathcal{F}, P)\) a probability space. Let \( \xi = \{\xi_\omega(B) : \omega \in \Omega, B \in S\} \) be such that

(i) For each fixed \( \omega \in \Omega \), \( \xi_\omega(B) \) is a measure on \( S \).

(ii) For each fixed \( B \in \mathcal{P} \), \( \xi_\omega(B) \) is a r.v. on \((\Omega, \mathcal{F}, P)\).

Then \( \xi \) is called a random measure (r.m.) on \( S \) (defined with respect to \((\Omega, \mathcal{F}, P)\)). Further if the r.m. \( \xi \) is such that \( \xi_\omega(B) \) is integer-valued a.s. for each \( B \in \mathcal{P} \) we call \( \xi \) a point process.

If \( \xi \) is a r.m., since \( \xi_\omega(B) \) is finite a.s. for each \( B \in \mathcal{P} \) and \( \mathcal{P} \) is countable, the null sets may be combined to give a single null set \( \Lambda \in \mathcal{F}, P(\Lambda) = 0 \) such that \( \xi_\omega(B) \) is finite for all \( B \in \mathcal{P}, \omega \in \Omega - \Lambda \). Indeed \( \xi_\omega(B) < \infty \) for all \( B \in S_0 \) when \( \omega \in \Omega - \Lambda \) since such \( B \) can be covered by finitely many sets of \( \mathcal{P} \). If desired, \( \Omega \) may be reduced to \( \Omega - \Lambda \) thus assuming that \( \xi_\omega(B) \) is finite for all \( \omega, B \in S_0 \).

If \( \xi \) is a r.m., \( \xi_\omega(B) \) is an extended r.v. for each \( B \in S \), and a r.v. for \( B \in S_0 \). For if \( S = \bigcup_1^n B_i \) where \( B_i \) are disjoint sets of \( \mathcal{P}, B = \bigcup_1^n (B \cap B_i) \) so that \( \xi_\omega(B) = \sum_1^n \xi_\omega(B \cap B_i) \) which is the measurable sum of (nonnegative) measurable terms.
If $\xi$ is a r.m., its expectation or intensity measure $\lambda = \mathcal{E}\xi$ is defined by $\lambda(B) = \mathcal{E}\xi(B)$ for $B \in \mathcal{S}$. Countable additivity is immediate (e.g. from Theorem 4.5.2 (Corollary)). Note that $\lambda$ is not necessarily finite, even on $\mathcal{P}$.

Point processes and r.m.’s have numerous properties which we do not consider in detail here. Some of these provide means of defining new r.m.’s from one or more given r.m.’s. An example is the following direct definition of a r.m. as an integral of an existing r.m., proved by $\mathcal{D}$-class methods:

**Theorem 15.7.1** If $\xi$ is a r.m. and $f$ is a nonnegative $\mathcal{S}$-measurable function then $\xi f = \int_{\mathcal{S}} f(s) \, d\xi(s)$ is $\mathcal{F}$-measurable. Furthermore, if $f$ is bounded on each set of $\mathcal{P}$, $\nu_f(B) = \int_{\mathcal{B}} f(s) \, d\xi(s)$, $B \in \mathcal{S}$, is a r.m.

It follows from the first part of this result that $e^{-\xi f} = e^{-\int f \, d\xi}$ is a nonnegative bounded r.v. for each nonnegative $\mathcal{S}$-measurable function $f$ and hence has a finite mean. $L_\xi(f) = \mathcal{E}e^{-\xi f}$ is termed the Laplace Transform (L.T.) of the r.m. $\xi$, and is a useful tool for many calculations. In particular for $B \in \mathcal{S}$, $L_\xi(\mathcal{1}_B) = \mathcal{E}e^{-\xi(\mathcal{1})}$ is the L.T. of the nonnegative r.v. $\xi(B)$, a useful alternative to the c.f. for nonnegative r.v.’s.

### 15.8 Example: The sample point process

Let $\tau$ be a r.e. in our basic space $(\mathcal{S}, \mathcal{S})$, and consider $\delta_\tau(B) = \chi_B(\tau)$ which may be viewed as unit mass at $s$, even if the singleton set $\{s\}$ is not $\mathcal{S}$-measurable. Then it is readily checked that the composition $\delta_\tau(B)$ defines a point process $\xi^{(1)}$ with unit mass at the single point $\tau \omega$. If the r.e. $\tau$ has distribution $\nu = P\tau^{-1}$ (Section 9.3), $\xi^{(1)}$ has intensity $\mathcal{E}\xi^{(1)}(B) = \mathcal{E}\chi_B(\tau \omega) = \mathcal{E}\chi_{\tau^{-1}}(\omega) = P\tau^{-1}(B) = \nu(B)$. Further straightforward calculations show that $\xi^{(1)}$ has L.T.

$$L_{\xi^{(1)}}(f) = \mathcal{E}e^{\tau^{-1}(\omega)} = \int e^{\tau^{-1}(s)} \, d\nu = \nu(e^{-\tau}).$$

Suppose now that $\tau_1, \tau_2, \ldots, \tau_n$ are independent r.e.’s of $\mathcal{S}$ with common distribution $P\tau_j^{-1} = \nu$. Then $f(\tau_1), f(\tau_2), \ldots, f(\tau_n)$ are i.i.d. (extended) r.v.’s for any nonnegative measurable $f$ and in particular $\chi_B(\tau_1), \chi_B(\tau_2), \ldots$, $\chi_B(\tau_n)$ are i.i.d. with $P[\chi_B(\tau_1) = 1] = \nu(B) = 1 - P[\chi_B(\tau_1) = 0]$. Hence if $\xi^{(n)}$ is the point process $\sum_1^n \delta_{\tau_j}$ and $B \in \mathcal{S}$,

$${\xi^{(n)}}(B) = \sum_1^n \delta_{\tau_j}(B) = \sum_1^n \chi_B(\tau_j),$$

so that $\xi^{(n)}(B)$ is binomial with parameters $(n, \nu(B))$. $\xi^{(n)}$ is thus a point process consisting of $n$ events at points $\{\tau_1, \tau_2, \ldots, \tau_n\}$, its intensity being
\[ E \xi(n) = n \nu, \text{ and its L.T. is readily calculated to be } \]
\[ L_{\xi(n)}(f) = E e^{-\sum_{i=1}^n \delta_{\tau_i}(f)} = E e^{-\sum_{i=1}^n f(\tau_i)} = (E e^{-f(\tau_1)})^n = (v(e^{-f}))^n. \]

\( \xi(n) \) is referred to as the sample point process consisting of \( n \) independent points \( \tau_1, \tau_2, \ldots, \tau_n. \)

### 15.9 Random element representation of a r.m.

As seen in Section 15.1, a real-valued stochastic process (family of r.v.’s) \( \{\xi_t : t \in T\} \) may be equivalently viewed as a random function, i.e. r.e. of \( \mathbb{R}^T \). Similarly one may regard a r.m. \( \{\xi(B) : B \in \mathcal{S}\} \) as a mapping \( \xi \) from \( \Omega \) into the space \( \mathcal{M} \) of all measures \( \mu \) on \( \mathcal{S} \) which are finite on \( \mathcal{P} \), i.e. \( \xi \omega \) is the element of \( \mathcal{M} \) defined by \( (\xi \omega)(B) = \xi_\omega(B), B \in \mathcal{S} \). A natural \( \sigma \)-field for the space \( \mathcal{M} \) is that generated by the functions \( \phi_B(\mu) = \mu(B), B \in \mathcal{S} \), i.e. the smallest \( \sigma \)-field \( \mathcal{M} \) making each \( \phi_B \) \( \mathcal{M} \)-measurable (\( \mathcal{M} = \sigma\{\phi_B^{-1} E : B \in \mathcal{S}, E \in \mathcal{B}\} \) (cf. Lemma 9.3.1)).

It may then be readily checked (cf. Section 9.3) that a r.m. \( \xi \) is a measurable mapping from \( (\Omega, \mathcal{F}, \mathcal{P}) \) to \( (\mathcal{M}, \mathcal{M}) \), i.e. a random element of \( (\mathcal{M}, \mathcal{M}) \).

As defined in Section 9.3 for r.e.’s, the distribution of the r.m. \( \xi \) is the probability measure \( P_{\xi^{-1}} \) on \( \mathcal{M} \). It is then true that any probability measure \( \pi \) on \( \mathcal{M} \) may be taken to be the distribution of a r.m., namely the identity r.m. \( \xi(\mu) = \mu \) on the probability space \( (\mathcal{M}, \mathcal{M}, \pi) \).

### 15.10 Mixtures of random measures

As noted r.m.’s may be obtained by specifying their distributions as any probability measures on \( (\mathcal{M}, \mathcal{M}) \). Suppose now that \( (\Theta, \mathcal{T}, Q) \) is a probability space, and for each \( \theta \in \Theta, \xi^{(\theta)} \) is a r.m. in \( (\mathcal{S}, \mathcal{S}) \) with distribution \( \pi_\theta \), \( \pi_\theta(A) = P\{\xi^{(\theta)} \in A\} \) for each \( A \in \mathcal{M} \). (Note that the \( \xi^{(\theta)} \)'s can be defined on different probability spaces.)

If for each \( A \in \mathcal{M}, \pi_\theta(A) \) is a \( \mathcal{T} \)-measurable function of \( \theta \), it follows from Theorem 7.2.1 that
\[ \pi(A) = \int_\Theta \pi_\theta(A) dQ(\theta) \]
is a probability measure on \( \mathcal{M} \), and thus may be taken to be the distribution of a r.m. \( \xi \), which may be called the mixed r.m. formed by mixing \( \xi^{(\theta)} \) with respect to \( Q \). Of course, it is the distribution of \( \xi \) rather than \( \xi \) itself which is uniquely specified.
The following intuitively obvious results are readily shown:

(i) If $\xi$ is the mixture of $\xi^{(\theta)} (P\xi^{-1}(A) = \int P\xi^{(\theta)}(A) dQ(\theta))$ and $B \in S$, the distribution of the (extended) r.v. $\xi(B)$ is (for Borel sets $E$)

\[
P\{\xi(B) \in E\} = P\{\phi_B\xi \in E\} = P\xi^{-1}(\phi_B^{-1}E) = \int P\xi^{(\theta)}(B) \in E\} dQ(\theta).
\]

(ii) The intensity $E\xi$ satisfies (for $B \in S$)

\[
E\xi(B) = \int E\xi^{(\theta)}(B) dQ(\theta).
\]

(iii) The Laplace Transform $L\xi(f)$ is, for nonnegative measurable $f$,

\[
L\xi(f) = \int L\xi^{(\theta)}(f) dQ(\theta).
\]

Example  
Mixing the sample point process.

Write $\xi^{(0)} = 0$ and for $n \geq 1$, $\xi^{(n)} = \sum_1^n \delta_{\tau_j}$ as in Section 15.8, where $\tau_1, \ldots, \tau_n$ are i.i.d. random elements of $(S,S)$ with (common) distribution $P\tau^{-1} = \nu$ say.

Let $\Theta = \{0,1,2,3,\ldots\}$, $T = \text{all subsets of } \Theta$, $Q$ the probability measure with mass $q_n$ at $n = 0,1,\ldots$ ($q_n \geq 0$, $\sum_0^\infty q_n = 1$). Then the mixture $\xi$ has distribution

\[
P\xi^{-1}(A) = \int P\theta(A) dQ(\theta) = \sum_{n=0}^\infty q_n P_n(A)
\]

where $P_n(A) = P\{\xi^{(n)} \in A\}$. For each $B \in S$ the distribution of $\xi(B)$ is given by the probabilities

\[
P\{\xi(B) = r\} = \sum_{n=r}^\infty q_n P\{\xi^{(n)}(B) = r\} = \sum_{n=r}^\infty q_n \left(\begin{array}{c}n \\ r\end{array}\right) \nu(B)^r (1 - \nu(B))^{n-r}
\]

and

\[
E\xi(B) = \sum_{n=0}^\infty q_n \nu(B) = \bar{q}\nu(B)
\]

where $\bar{q}$ is the mean of the distribution $\{q_n\}$. That is $E\xi = \bar{q}\nu$.

The Laplace Transform of $\xi$ is

\[
L\xi(f) = \int L\xi^{(\theta)}(f) dQ(\theta) = \sum_{n=0}^\infty q_n L\xi^{(n)}(f) = \sum_{n=0}^\infty q_n (\nu(e^{-f}))^n = G(\nu(e^{-f}))
\]

where $G$ denotes the probability generating function (p.g.f.) of the distribution $\{q_n\}$. 

15.11 The general Poisson process

We now outline how the general Poisson process may be obtained on our basic space \((S, S)\) from the mixed sample point process considered in the last section.

First define a “finite Poisson process” as simply a mixed sample point process with \(q_n = e^{-a}a^n/n!\) for \(a > 0, n = 0, 1, 2, \ldots\), i.e. Poisson probabilities. For \(B \in S\),

\[
P(\xi(B) = r) = \sum_{n=r}^{\infty} \frac{e^{-a}a^n}{n!} \binom{n}{r} (1 - \nu(B))^{n-r}
\]

which reduces simply to \(e^{-av(B)}(av(B))^r/r!\), \(r = 0, 1, 2, \ldots\), i.e. a Poisson distribution for any \(B \in S\), with mean \(av(B)\). In particular if \(B = S\), \(\xi(S)\) has a Poisson distribution with mean \(a\). This, of course, implies \(\xi(S) < \infty\) a.s. so that the total number of Poisson points in the whole space is finite. This limits the process (ordinarily one thinks of a Poisson process – e.g. on the line – as satisfying \(P(\xi(S) = \infty) = 1\)), which is the reason for referring to this as a “finite Poisson process”. This process has intensity measure \(a\nu(B)\) say, and Laplace Transform \(G(\nu(e^{-s}))\) where \(G(s) = e^{-\alpha(1-s)}\), i.e.

\[
L_\xi(f) = e^{-a(1-\nu(e^{-f}))} = e^{-av(1-e^{-f})} = e^{-\lambda(1-e^{-f})} \quad (\nu(1) = 1).
\]

Any finite (nonzero) measure \(\lambda\) on \(S\) may be taken as the intensity measure of a finite Poisson process (by taking \(a = \lambda(S)\) and \(\nu = \lambda/\lambda(S)\)).

The general Poisson process (for which \(\xi(S)\) can be infinite-valued) can be obtained by summing a sequence of independent finite Poisson processes as we now indicate, following the construction of a sequence of independent r.v.’s as in Corollary 2 of Theorem 15.1.2. Let \(\lambda \in M\) (i.e. a measure on \(S\) which is finite on \(P\)). From the basic assumptions it is readily checked that \(S\) may be written as \(\cup_{i}^{\infty} S_{i}\), where \(S_{i}\) are disjoint sets of \(P\) and we write \(\lambda_{i}(B) = \lambda(B \cap S_{i}), B \in S\). The \(\lambda_{i}(B), i = 1, 2, \ldots,\) are finite measures on \(S\) and may thus be taken as the intensities of independent finite Poisson processes \(\xi_{i}\), whose distributions on \((M, M)\) are \(P_{i}\), say. \((P_{i}\) assigns measure 1 to the set \(\{\mu \in M: \mu(S - S_{i}) = 0\}\).)

Define now \(\xi = \sum_{i}^{\infty} \xi_{j}\). Since, for \(B \in P\), \(\mathcal{E}[\sum_{i}^{\infty} \xi_{j}(B)] = \sum_{i}^{\infty} \lambda_{i}(B) = \sum_{i}^{\infty} \lambda(B \cap S_{j}) = \lambda(B) < \infty\) \((\lambda \in M)\) we see that \(\sum_{i}^{\infty} \xi_{j}(B)\) converges a.s. on \(P\) and hence \(\xi\) is a point process. By the above \(E\xi(B) = \lambda(B)\) so that \(\xi\) has intensity measure \(\lambda\).

\(\xi\) is the promised Poisson process in \(S\) with intensity measure \(\lambda \in M\). Some straightforward calculation using independence and dominated
convergence shows that its L.T. is

\[ L_{\xi}(f) = \lim_{n \to \infty} \prod_{i=1}^{n} L_{\xi_i}(f) = e^{-\sum_{i} \lambda_i (1-e^{-f})} = e^{-\lambda (1-e^{-f})} \]

i.e. the same form as in the finite case.

In summary then the following result holds.

**Theorem 15.11.1** Let \((S, S, \mathcal{P})\) be a basic structure, and let \(\lambda\) be a measure on \(S\) which is finite on \((the\ semiring)\ \mathcal{P}\). Then there exists a Poisson process \(\xi\) on \(S\) with intensity \(\xi = \lambda\), thus having the L.T.

\[ L_{\xi}(f) = e^{-\lambda (1-e^{-f})}. \]

By writing \(f = \sum_{i=1}^{n} t_i \chi_{B_i}\) and using the result for L.T.’s corresponding to Theorem 12.8.3 for c.f.’s (with analogous proof using the uniqueness theorem for L.T.’s, see e.g. [Feller]), it is seen simply that \(\xi(B_i), i = 1, 2, \ldots, n,\) are independent Poisson r.v.’s with means \(\lambda(B_i)\) when \(B_i\) are disjoint sets of \(S\).

### 15.12 Special cases and extensions

As defined the general Poisson process \(\xi\) has intensity \(\xi = \lambda\) where \(\lambda\) is a measure on \(S\) which is finite on \(\mathcal{P}\). The simple familiar stationary Poisson process on the real line is a very special case where \((S, S)\) is \((\mathbb{R}, \mathcal{B})\), \(\mathcal{P}\) can be taken to be the semiclosed intervals \([(a, b] : a, b \text{ rational}, -\infty < a < b < \infty}\) and \(\lambda\) is a multiple of Lebesgue measure, \(\lambda(B) = \lambda m(B)\) for a finite positive constant \(\lambda\), termed the intensity of the simple Poisson process. Nonstationary Poisson processes on the line are simply obtained by taking an intensity measure \(\lambda \ll m\), having a time varying intensity function \(\lambda(t)\), \(\lambda(B) = \int_{B} \lambda(t) dt\). These Poisson processes have no fixed atoms (points \(s\) at which \(P\{\xi(s) > 0\} > 0\)) and no “multiple atoms” (random points \(s\) with \(\xi(s) > 1\)). On the other hand fixed atoms or multiple atoms are possible if a chosen intensity measure has atoms.

Poisson processes’ distributions may be “mixed” to form “mixed Poisson process” or “compound Poisson processes” and intensity measures may themselves be taken to be stochastic to yield “doubly stochastic Poisson processes” (“Cox processes” as they are generally known). These latter are particularly useful for modeling applications involving stochastic occurrence rates.

The very simple definition of a **basic structure** in Section 15.7 suffices admirably for the definition of Poisson processes. However, its extensions such as those above and other random measures typically require at least
a little more structure. One such assumption is that of separation of two points of $S$ by sets of $\mathcal{P}$ – a simple further requirement closely akin to the definition of Hausdorff spaces. Such an assumption typically suffices for the definition and basic framework of many point processes. However, more intricate properties such as a full theory of weak convergence of r.m.’s are usually achieved by the introduction of more topological assumptions about the space $S$. 
References


Index

$L_p$-space, 127
  complex, 180
$\lambda$-system, 19
$\mu^*$-measurable, 29
$\sigma$-algebra
  see $\sigma$-field, 13
$\sigma$-field, 13
  generated by a class of sets, 14
  generated by a random variable, 195
  generated by a transformation, 47
$\sigma$-finite, 22, 86
$\sigma$-ring, 13
  generated by a class of sets, 14
$D$-class, 15
absolute continuity, 94, 105, 110, 193, 199
almost everywhere, 57
almost surely (a.s.), 190
atoms, 192
Banach space, 127
binomial distribution, 193, 257
Bochner’s Theorem, 275
Borel measurable function, 59, 190
Borel sets, 16
  extended, 45
  $n$-dimensional, 158
  two-dimensional, 153
Borel–Cantelli Lemma, 217
bounded variation, 110, 180
Brownian motion
  see Wiener process, 343
Cauchy sequence, 118
  almost surely (a.s.), 224
  almost uniformly, 119
  in measure, 121
  in metric space, 125
  uniformly, 118
centered sequences, 325
central limit theorem
  array form of Lindeberg–Feller, 269
  elementary form, 267
  standard form of Lindeberg–Feller, 271
change of variables in integration, 106
characteristic function (c.f.) of a random variable, 254
  inversion and uniqueness, 261
  inversion theorem, 278
  joint, 277
  recognizing, 271
  uniqueness, 262, 278
Chebychev Inequality, 202, 243
classes of sets, 1, 2
completion, 34, 41, 81
conditional distribution, 295
conditional expectation, 287, 288, 300, 305
conditional probability, 285, 291, 301, 305
conditionally independent, 307
consistency of a family of measures (distributions), 167, 342
continuity theorem for characteristic functions, 264, 279
continuous from above (below), 25
continuous mapping theorem, 231
convergence
  almost everywhere (a.e.), 58
  almost sure (a.s.), with probability one, 223
  almost uniform (a.u.), 119
  in distribution, 227, 228
  in measure, 120
  in probability, 225
  in $p$th order mean ($L_p$-spaces), 226
modes, summary, 134
of integrals, 73
pointwise, 118
convergence (cont.)
uniformly, 118
uniformly a.e., 118
vague, 204, 237
weak, 228
convex, 202
convolution, 153, 216
correlation, 200
counting measure, 41, 81
covariance, 200
Cox processes, 354
Cramér–Wold device, 280
cylinder set, 164
De Morgan laws, 6
degenerate distribution, 257
density function, 105
discrete measures, 104, 105
distribution
marginal, 198, 341
of a random element, 197
of a random measure, 351
of a random variable, 190
distribution function (d.f.), 191
absolutely continuous, 193, 199
discrete, 193
joint, 197
dominated convergence, 76, 92, 179
conditional, 290
Doob’s decomposition, 313
Egoroff’s Theorem, 120
equivalent
signed measures, 95
stochastic processes, 345
essentially unique, 96
event, 189
expectation, 199
extension of measures, 27, 31
Fatou’s Lemma, 76
cConditional, 289
field (algebra), 9
finite-dimensional distributions (fidi’s), 341
Fourier Transform, 181, 254
Dirichlet Limit, 186
inverse, 185
inversion, 182
“local” inversion, 186
local inversion theorem, 187
Fourier–Stieltjes Transform, 180, 254
inversion, 182
Fubini’s Theorem, 150, 158
functional central limit theorem
(invariance principle), 347
gamma distribution, 194
generalized second derivative, 282
Hahn decomposition, 88
minimal property, 90
Hausdorff space, 349, 355
Heine–Borel Theorem, 37
Helly’s Selection Theorem, 232
Hölder’s Inequality, 128, 179, 201
increasing sequence of functions, 55
independent events and their classes, 208
independent random elements and their families, 211
independent random variables, 213
addition, 216
existence, 214
indicator (characteristic) functions, 7
integrability, 68
integrable function, 67–69
integral, 68
defined, 69
indefinite, 66
of complex functions, 177
of nonnegative measurable functions, 63
of nonnegative simple functions, 62
with respect to signed measures, 92
integration by parts, 154
inverse functions, 203
inverse image, 46
Jensen’s Inequality, 202
conditional, 291, 306
Jordan decomposition, 89, 152
Kolmogorov Inequalities, 241, 314
Kolmogorov Zero-One Law, 218
Kolmogorov’s Extension Theorem, 167, 169, 342
Kolmogorov’s Three Series Theorem, 244
Laplace Transform (L.T.), 350
laws of large numbers, 247, 248, 327
Lebesgue decomposition, 96, 98, 106, 194
Lebesgue integrals, 78
Lebesgue measurable function, 59
Lebesgue measurable sets, 38
n-dimensional, 158
two-dimensional, 153
Lebesgue measure, 37
  n-dimensional, 158
  two-dimensional, 153
Lebesgue–Stieltjes integrals, 78, 111
Lebesgue–Stieltjes measures, 39, 78, 111, 158, 162
Lévy distance, 251
Liapounov’s condition, 284
likelihood ratios, 333
Lindeberg condition, 269
linear mapping (transformation), 17, 38
linear space, 126
Markov Inequality, 202
martingale (submartingale, supermartingale), 309, 320
  convergence, 319
  predictable increasing sequence, 313
  reverse, 323
  upcrossings, 317
mean square estimate, 306
measurability criterion, 48
measurable functions, 47
  combining, 50
  complex-valued, 178
  extended, 45
measurable space, 44
measurable transformation, 47
measure space, 44
measures, 22
  complete, 34
  complex, 87
  from outer measures, 29
  induced by transformations, 58
  mixtures of, 143
  on \( \mathbb{R}^T \), 167
  regularity, 162
metric space, 124
  complete, 126
  separable, 126
Minkowski’s Inequality, 129, 180, 201
  reverse, 130
moments, 199
  absolute, 199, 200
  central, 200
  inequalities, 200
monotone class theorem, 14, 19
monotone convergence theorem, 74
  conditional, 289
nonnegative definite, 274
norm, 126
normal distribution, 194, 257
  multivariate, 200
normed linear space, 126
outer measure, 29
Palm distributions, 285
point process, 349
Poisson distribution, 193
Poisson process, 353
  compound, 354
  doubly stochastic, 354
  stationary and nonstationary, 354
Polya’s urn scheme, 337
Portmanteau Theorem, 228
positive definite, 274
probability density function (p.d.f.), 193
  joint, 198
probability measure (probability), 44, 189
  frequency interpretation, 190
  inequalities, 200
probability space, 189
probability transforms, 204
product measurable space, 155
product measure, 149, 156
product spaces, 141
  \( \sigma \)-field, 142, 165
  \( \sigma \)-ring, 141, 142
  diagonal, 171
  finite-dimensional, 155
  space (\( \mathbb{R}^T, B^T \)), 163
Prohorov’s Theorem, 234
projection map, 164, 165
Rademacher functions, 220
Radon–Nikodym derivative, 102
  chain rule, 103
Radon–Nikodym Theorem, 96, 100, 179
random element (r.e.), 195
random experiment, 189
random function (r.f.), 340
random measure (r.m.), 350
  basic structure, 349
  intensity measure, 350
  mixed, 351
  random element representation, 351
random variables (r.v.’s), 190
  absolutely continuous, 193
  discrete, 193
  extended, 190
  identically distributed, 192
  symmetric, 281
random vector, 195, 196
<table>
<thead>
<tr>
<th>Term</th>
<th>Page Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>real line applications</td>
<td>78, 104, 153</td>
</tr>
<tr>
<td>rectangle</td>
<td>141</td>
</tr>
<tr>
<td>regular conditional density</td>
<td>303</td>
</tr>
<tr>
<td>regular conditional distribution</td>
<td>296, 299,</td>
</tr>
<tr>
<td></td>
<td>301, 302, 305</td>
</tr>
<tr>
<td>regular conditional probability</td>
<td>293, 299,</td>
</tr>
<tr>
<td></td>
<td>301, 305</td>
</tr>
<tr>
<td>relatively compact</td>
<td>234</td>
</tr>
<tr>
<td>repeated (iterated) integral</td>
<td>148, 157</td>
</tr>
<tr>
<td>Riemann integrals</td>
<td>79, 80, 84</td>
</tr>
<tr>
<td>Riemann–Lebesgue Lemma</td>
<td>182</td>
</tr>
<tr>
<td>rings</td>
<td>8, 11</td>
</tr>
<tr>
<td>sample functions (paths)</td>
<td>341</td>
</tr>
<tr>
<td>continuity</td>
<td>345</td>
</tr>
<tr>
<td>sample point process</td>
<td>351</td>
</tr>
<tr>
<td>mixing</td>
<td>352</td>
</tr>
<tr>
<td>Schwarz Inequality</td>
<td>129</td>
</tr>
<tr>
<td>section of a set</td>
<td>142</td>
</tr>
<tr>
<td>semiring</td>
<td>10</td>
</tr>
<tr>
<td>set functions, 21</td>
<td></td>
</tr>
<tr>
<td>additive</td>
<td>22</td>
</tr>
<tr>
<td>countable subadditivity</td>
<td>29</td>
</tr>
<tr>
<td>extensions and restrictions</td>
<td>22</td>
</tr>
<tr>
<td>finitely additive (countably additive)</td>
<td>22</td>
</tr>
<tr>
<td>monotone</td>
<td>23</td>
</tr>
<tr>
<td>subtractive</td>
<td>23</td>
</tr>
<tr>
<td>set mapping</td>
<td>46</td>
</tr>
<tr>
<td>sets, 1</td>
<td></td>
</tr>
<tr>
<td>complement of a set</td>
<td>4</td>
</tr>
<tr>
<td>convergent</td>
<td>7</td>
</tr>
<tr>
<td>difference</td>
<td>4</td>
</tr>
<tr>
<td>disjoint</td>
<td>4</td>
</tr>
<tr>
<td>empty</td>
<td>3</td>
</tr>
<tr>
<td>equalities</td>
<td>5</td>
</tr>
<tr>
<td>intersection</td>
<td>3</td>
</tr>
<tr>
<td>limits</td>
<td>6</td>
</tr>
<tr>
<td>lower limit</td>
<td>6</td>
</tr>
<tr>
<td>monotone increasing (decreasing)</td>
<td>7</td>
</tr>
<tr>
<td>proper difference</td>
<td>4</td>
</tr>
<tr>
<td>symmetric difference</td>
<td>4</td>
</tr>
<tr>
<td>union (sum)</td>
<td>3</td>
</tr>
<tr>
<td>upper limit</td>
<td>6</td>
</tr>
<tr>
<td>signed measure</td>
<td>86, 152</td>
</tr>
<tr>
<td>null, negative, positive</td>
<td>87</td>
</tr>
<tr>
<td>total variation</td>
<td>112</td>
</tr>
<tr>
<td>simple functions</td>
<td>54</td>
</tr>
<tr>
<td>singularity</td>
<td>94, 105, 194</td>
</tr>
<tr>
<td>Skorohod’s Representation</td>
<td>236</td>
</tr>
<tr>
<td>stochastic process</td>
<td>195, 340</td>
</tr>
<tr>
<td>continuous parameter</td>
<td>340</td>
</tr>
<tr>
<td>on special subspaces of $\mathbb{R}^T$</td>
<td>344</td>
</tr>
<tr>
<td>realization</td>
<td>341</td>
</tr>
<tr>
<td>stochastic sequence or discrete parameter</td>
<td>340</td>
</tr>
<tr>
<td>tail $\sigma$-field, event and random variables,</td>
<td>218</td>
</tr>
<tr>
<td>three series theorem</td>
<td>244</td>
</tr>
<tr>
<td>tight family</td>
<td>232</td>
</tr>
<tr>
<td>transformation</td>
<td>45</td>
</tr>
<tr>
<td>transformation theorem</td>
<td>77, 93, 179</td>
</tr>
<tr>
<td>triangular array</td>
<td>268</td>
</tr>
<tr>
<td>Tychonoff’s Theorem</td>
<td>168</td>
</tr>
<tr>
<td>uniform absolute continuity</td>
<td>238</td>
</tr>
<tr>
<td>uniform distribution</td>
<td>257</td>
</tr>
<tr>
<td>uniform integrability</td>
<td>238</td>
</tr>
<tr>
<td>variance</td>
<td>200</td>
</tr>
<tr>
<td>Wiener measure</td>
<td>347</td>
</tr>
<tr>
<td>Wiener process</td>
<td>343, 346</td>
</tr>
</tbody>
</table>