Probability Theory

Kexing Ying

May 30, 2022

Contents

1 Review of Measure Theory

Modern probability theory is based on measure theory and we will in this section recall some notions from measure theory.

Definition 1.1 (Algebra)**.** Given a set Ω, a set of subsets $\mathcal{A}$ of Ω is an algebra if $\Omega \in \mathcal{A}$ and $\mathcal{A}$ is closed under finite union and complements.

It follows straight away that an algebra is also closed under finite intersections.

Definition 1.2 (Finitely Additive Measure)**.** A function $\mu : \mathcal{A} \to [0, \infty]$ where $\mathcal{A}$ is an algebra, is a finitely additive measure if for any disjoint sets $A, B \in \mathcal{A}$,

$$\mu(A \cup B) = \mu(A) + \mu(B).$$

Definition 1.3 (σ-Algebra)**.** A σ-algebra $\mathcal{F}$ is an algebra that is closed under countable unions.

Similarly, it follows that $\mathcal{F}$ is closed under countable intersections.

Definition 1.4 (Measure)**.** A function $\mu : \mathcal{F} \to [0, \infty]$ where $\mathcal{F}$ is a σ-algebra, is a σ-additive measure (or simply measure) if given a sequence of pairwise disjoint sets $A_1, A_2, ...$ of $\mathcal{F}$, we have

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_n).$$

We call a measure a probability measure if $\mu(\Omega) = 1$.

Definition 1.5 (σ-Finite Measure)**.** A measure μ is said to be σ-finite if there exists a sequence of pairwise disjoint sets $A_1, A_2, ...$ of $\mathcal{F}$, such that $\bigcup_{i=1}^{\infty} A_i = \Omega$ and for all i, $\mu(A_i) < \infty$.

Definition 1.6 (Probability Space)**.** A probability space is the triple $(\Omega, \mathcal{F}, \mathbb{P})$ consisting of a set Ω, a σ-algebra $\mathcal{F}$ on Ω and $\mathbb{P}$ a probability measure on $\mathcal{F}$.

We call elements of $\mathcal{F}$ (i.e. a $\mathcal{F}$-measurable set) an event.

Proposition 1.1 (Continuity of Measures)**.** Let $(A_n)_{n\in\mathbb{N}} \subseteq \mathcal{F}$, then

- (continuity from below) if (A_n) is increasing, then

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} \mathbb{P}(A_n).$$

- (continuity from above) if (A_n) is decreasing, then

$$\mathbb{P}\left(\bigcap_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} \mathbb{P}(A_n).$$

We recall the the finiteness of the measure is vital for continuity from below while continuity from above is also valid for general measures.

Proof. Exercise. □

Proposition 1.2. A finitely additive probability measure on the σ-algebra $\mathcal{F}$ is a probability measure if and only if it is continuous at 0.

Proof. The forward direction follows from above so we will prove the reverse. Suppose μ is finitely additive and for any decreasing $(A_n) \subseteq \mathcal{F}$ with $\bigcap A_n = \emptyset$, we have $\lim_{n\to\infty} \mu(A_n) = 0$. Then, μ is continuous from below, and so for any sequence of disjoint sets (B_n), we have $(C_n) := (\bigcup_{i=1}^{n} B_i)$ is a sequence of increasing sets and thus,

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} B_i\right) = \mathbb{P}\left(\bigcup_{i=1}^{\infty} C_i\right) = \lim_{n\to\infty} \mathbb{P}(C_n) = \lim_{n\to\infty} \mathbb{P}\left(\bigcup_{i=1}^{n} B_i\right) = \lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{P}(B_i)$$

implying μ is σ-additive and so, μ is a measure. □

Proposition 1.3. Given a collection $\{\mathcal{F}_i\}_{i\in I}$ σ-algebras of Ω, $\bigcap_{i\in I} \mathcal{F}_i$ is also a σ-algebra on Ω.

Definition 1.7 (σ-Algebra Generated By Sets)**.** Given a collection of subsets S of Ω, the σ-algebra generated by S is

$$\sigma(S) := \bigcap \{\mathcal{F} \text{ a } \sigma\text{-algebra} \mid S \subseteq \mathcal{F}\}.$$

Definition 1.8 (Borel σ-Algebra)**.** Given a topological space $(X, \mathcal{T})$, the Borel σ-algebra on X is $\mathcal{B}(X) := \sigma(\mathcal{T})$.

Definition 1.9 (Product σ-Algebra)**.** Given measurable spaces $(\Omega_1, \mathcal{F}_1), (\Omega_2, \mathcal{F}_2)$, the product σ-algebra on $\Omega_1 \times \Omega_2$ is

$$\mathcal{F}_1 \otimes \mathcal{F}_2 := \sigma(\mathcal{F}_1 \times \mathcal{F}_2) = \sigma(\{A_1 \times A_2 \mid A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2\}).$$

Definition 1.10 (Cylindrical σ-Algebra)**.** A set $C \subseteq \mathbb{R}^\infty$ is said to be cylindrical if is of the form

$$C = \{x \in \mathbb{R}^\infty \mid (x_1, \cdots, x_n) \in C_n\}$$

where $C_n \in \mathcal{B}(\mathbb{R}^n)$. The set of cylindrical sets $\mathcal{B}(\mathbb{R}^\infty)$ form a σ-algebra on $\mathbb{R}^\infty$ and is called the cylindrical σ-algebra.

Definition 1.11 (Consistent)**.** The sequence of measures $\mathbb{P}_n$ on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ is said to be consistent if for all $n \in \mathbb{N}$, $\mathbb{P}_{n+1}(B_n \times \mathbb{R}) = \mathbb{P}_n(B_n)$ for all $B_n \in \mathcal{B}(\mathbb{R}^n)$.

Theorem 1 (Kolmogorov)**.** Given any consistent sequence of measures $\mathbb{P}_n$, there exists a unique probability measure $\mathbb{P}$ on $(\mathbb{R}^\infty, \mathcal{B}(\mathbb{R}^\infty))$ such that,

$$\mathbb{P}(\{x \in \mathbb{R}^\infty \mid (x_1, \cdots, x_n) \in \mathcal{B}_n\}) = \mathbb{P}_n(B_n)$$

for all $n \geq 1$, $B_n \in \mathcal{B}(\mathbb{R}^n)$.

Proof. Simply define the inner measure on the generating sets as claimed and use the Caratheodory extension (which provides both existence and uniqueness). □

Recall that a nondecreasing function g on $\mathbb{R}$ is continuous up to possibly countably many discontinuities of the first kind. Furthermore, the derivative g' exists λ-a.e. (where λ is the Lebesgue measure on $\mathbb{R}$.

Proposition 1.4. Let $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P})$ be a probability space. Defining $F(x) := \mathbb{P}(-\infty, x]$, we have

- F is nondecreasing;
- $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$;
- F is continuous on the right.

Proof. Clear by the monotonicity, continuity of measures (from above). □

Definition 1.12 (Distribution Function)**.** Any function $F : \mathbb{R} \to [0, 1]$ satisfying the above three properties is said to be a distribution function on $\mathbb{R}$.

It is clear that any probability measure induces a distribution. On the other hand the converse is also true.

Proposition 1.5. Given a distribution function F, there exists a unique probability measure $\mathbb{P}$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that $F(x) = \mathbb{P}(-\infty, x]$ for all $x \in \mathbb{R}$.

Proof. Use Caratheodory extension theorem on the algebra $\{(-\infty, x] \mid x \in \mathbb{R}\}$ mapping $(-\infty, x] \mapsto F(x)$. The uniqueness of the probability measure follows by the uniqueness of the Caratheodory extension. □

Definition 1.13 (Null-set)**.** Given a measure μ, a set $S \subseteq \Omega$ is a null-set if there exists some measurable set $N \subseteq \Omega$ with measure 0 such that $S \subseteq N$.

Definition 1.14 (Complete Measure)**.** A measure μ is complete if every μ-null set is measurable.

If a measure on the σ-algebra Σ is not complete, we may complete the σ-algebra by extending Σ to

$$\overline{\Sigma} := \sigma(\Sigma \cup \{N \mid N \text{ is a null-set}\}).$$

Clearly, the null-sets will have measure 0.

We note that the probability space $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P})$ is not complete as there exists subsets of a Borel null-set which are not Borel. With this in mind, we denote the completion of $\mathcal{B}(\mathbb{R})$ by $\mathcal{M}(\mathbb{R})$ and we say $\mathbb{P}$ is the Lebesgue-Stieltjes measure.

Recall, that in elementary probability theory, we considered three types of distributions, namely discrete, absolutely continuous and singular continuous. Let us now consider them again in a formal measure theoretic setting.

- (Discrete) A random variable $X : \Omega \to \mathbb{R}$ is said to have discrete distribution if there exists some countable (including finite) set $A \subseteq \mathbb{R}$, such that for all $E \in \mathcal{B}(\mathbb{R})$, the push-forward measure satisfies

$$X_*\mathbb{P}(E) = \sum_{x\in A} p(x)\delta_x(E)$$

where $p(x) = X_*\mathbb{P}(\{x\})$ and δ_x is the Dirac measure at x.

We note that a distribution function F corresponds to a discrete random variable if and only if for all $x_0 \in \mathbb{R}$,

$$F(x_0) = \sum_{x \in A \cap \{\leq x_0\}} p(x).$$

It is clear that $\sum_A p(x) = 1$ since $\sum_A p(x) = X_*\mathbb{P}(\mathbb{R}) = 1$.

- (Absolutely continuous) A random variable $X : \Omega \to \mathbb{R}$ is said to be absolutely continuous if $X_*\mathbb{P} = f\lambda$ for some Lebesgue integrable function f and λ denotes Lebesgue measure. Thus, the distribution function corresponding to X satisfies

$$F(x) = \int_{(-\infty, x]} f \mathrm{d}\lambda.$$

In particular, recalling the Radon-Nikodym theorem, we have X is absolutely continuous if and only if $X_*\mathbb{P} \ll \lambda$ (hence the name "absolutely continuous").

Before introducing the last type of distribution, let us consider the following definition.

Definition 1.15 (Concentrated)**.** A measure μ on the measurable space X is said to be concentrated on a measurable set A if $\mu(E) = 0$ for all $E \subseteq X \setminus A$.

- (Singular continuous) A random variable $X : \Omega \to \mathbb{R}$ is singular continuous if its distribution function F is continuous and $X_*\mathbb{P}$ is concentrated on a set A of Lebesgue measure 0 for which $F'(x) = 0$ for all $x \in A$ almost everywhere.

We note that, since $\lambda(A) = 0$, $X_*\mathbb{P} \perp \lambda$ by the set A (hence the name "singular"). Moreover, by continuity, $X_*\mathbb{P}(\{x\}) = 0$ for all $x \in \mathbb{R}$ in contrast to the discrete measure.

Analogous to the Lebesgue decomposition of measures, we may decompose any distribution function in to a discrete, absolutely continuous and singular continuous distribution.

Theorem 2 (Hahn Decomposition for Distributions)**.** Given a distribution function F, there exists $a_1 + a_2 + a_3 = 1$ and $F_{\text{disc}}, F_{\text{ac}}, F_{\text{sc}}$ discrete, absolutely continuous and singular continuous distribution functions respectively, such that

$$F = a_1 F_{\text{disc}} + a_2 F_{\text{ac}} + a_3 F_{\text{sc}}.$$

Proof. Recalling the refinement of the Lebesgue decomposition where we may decompose a measure μ with

$$\mu = \mu_d + \mu_a + \mu_s$$

where μ_d is a discrete measure, $\mu_a \ll \lambda$ and μ_s is singular continuous (i.e. mutually singular with respect to the Lebesgue measure and $\mu_s\{x\}$ for all x). Thus, by simply taking the the decomposition of the measure corresponding to F (i.e. $X_*\mathbb{P}$), we obtain the required decomposition after normalization. □

2 Random Variables

We will continue to let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space.

Definition 2.1 (Random Variable)**.** A function $\xi : \Omega \to \mathbb{R}$ is said to be a random variable if it is $\mathcal{F}$-measurable (i.e. for any $B \in \mathcal{B}(\mathbb{R})$, we have $\xi^{-1}(B) \in \mathcal{F}$).

While we have already introduced the notion of a distribution within the previous section, we will present it here again for organization.

Definition 2.2 (Distribution of a Random Variable)**.** Given a random variable ξ, the distribution of ξ is the push-forward measure of $\mathbb{P}$ along ξ. Furthermore, the distribution function corresponding to ξ is

$$F(x) := X_* \mathbb{P}(-\infty, x].$$

Definition 2.3. Given a random variable ξ, we define $\mathcal{F}_\xi \subseteq \mathcal{F}$ to be the σ-algebra

$$\mathcal{F}_\xi := \{\xi^{-1}(B) \mid B \in \mathcal{B}(\mathbb{R})\}.$$

This is the least σ-algebra for which ξ is measurable.

We will recall some standard results about measurable functions. All proofs are left as exercises and can be found in the second year measure theory notes.

Lemma 2.1. If $\mathcal{B}(\mathbb{R}) = \sigma(\mathcal{D})$ for a collection of sets $\mathcal{D}$, ξ is a random variable if $\xi^{-1}(D) \in \mathcal{F}$ for all $D \in \mathcal{D}$.

Lemma 2.2. Given random variables f, g and $c \in \mathbb{R}$, $f + g, f - g, c \cdot f, |f|, fg$, $\max(f, g)$, and $\min(f, g)$ are all random variables. Furthermore, if $g(x) \neq 0$ for all x, then f/g is also a random variable.

Lemma 2.3. If (f_n) is a sequence of random variables, then

$$\sup_n f_n, \inf_n f_n, \lim_n f_n$$

are random variables if they exist.

Lemma 2.4. If ξ is a random variable and $f : \mathbb{R} \to \mathbb{R}$ is continuous, then $f(\xi)$ is a random variable.

Definition 2.4 (Simple Function)**.** A random variable ξ is simple if there exists a partition of Ω, $D_1, \cdots, D_n$ such that

$$\xi(\omega) = \sum_{i=1}^{n} x_i 1_{D_i}(\omega)$$

for some $x_1, \cdots, x_n$ for all $\omega \in \Omega$.

Lemma 2.5. For any non-negative random variable ξ, there exists a sequence of nondecreasing simple random variables (ξ_n) such that for all $\omega \in \Omega$,

$$\xi_n(\omega) \uparrow \xi(\omega).$$

Definition 2.5 (Random Vector)**.** A function $\xi : \Omega \to \mathbb{R}^n$ is a random vector if it is measurable. Again, we define its distribution to be its push-forward measure.

Lemma 2.6. $\xi : \Omega \to \mathbb{R}^n$ is a random vector if and only if $\xi_i := \mathrm{pr}_i \circ \xi$ is a random variable for all $i = 1, \cdots, n$ (where $\mathrm{pr}_i : \mathbb{R}^n \to \mathbb{R}$ is the i-th projection function).

Definition 2.6 (Independent Random Variables)**.** Two random variables $\xi, \eta : \Omega \to \mathbb{R}$ are said to be independent if

$$(\xi, \eta)_* \mathbb{P} = \xi_* \mathbb{P} \otimes \eta_* \mathbb{P}.$$

Since, to check that two measures are equal, it suffices to check equality on generating sets, ξ, η are independent if for all $A, B \in \mathcal{B}(\mathbb{R})$,

$$\mathbb{P}(\xi \in A, \eta \in B) = \mathbb{P}(\xi \in A)\mathbb{P}(\eta \in B).$$

Let us quickly recall the construction of the Lebesgue integral.

1. Define the Lebesgue integral for simple functions.
2. Define the Lebesgue integral for non-negative functions by taking the limit of the Lebesgue integral of the monotone sequence of simple functions which converge to the said function.
3. Define the Lebesgue integral for general real-valued functions f by taking $\int f = \int f^+ - \int f^-$ if $\int |f| < \infty$.

Definition 2.7 (Expectation)**.** Given a random variable $\xi : \Omega \to \mathbb{R}$, the expectation of ξ is simply

$$\mathbb{E}(\xi) := \int \xi \mathrm{d}\mathbb{P}$$

if it exists. Furthermore, we say ξ is integrable if $\mathbb{E}(|\xi|) < \infty$.

Proposition 2.1. Let ξ, η be integrable random variables and let $c \in \mathbb{R}$, then

- $\mathbb{E}(c) = c$;
- $\mathbb{E}(\xi + \eta) = \mathbb{E}(\xi) + \mathbb{E}(\eta)$;
- $\xi \leq \eta$ a.e. implies $\mathbb{E}(\xi) \leq \mathbb{E}(\eta)$;
- $\xi = \eta$ a.e. implies $\mathbb{E}(\xi) = \mathbb{E}(\eta)$;
- $\xi \geq 0$ a.e. and $\mathbb{E}(\xi) = 0$ implies $\xi = 0$ a.e.

Proof. Follows directly from the properties of the Lebesgue integral. □

Let us recall some convergence theorems for the Lebesgue integral.

Theorem 3 (Dominated Convergence Theorem)**.** Let (ξ_n) be a sequence of random variables such that $\xi_n \to \xi$ almost everywhere. If there exists some integrable η such that $|\xi_n| \leq \eta$ for all n, then, ξ is integrable and

$$\lim_{n \to \infty} \mathbb{E}(\xi_n) = \mathbb{E}(\xi).$$

Theorem 4 (Monotone Convergence Theorem)**.** Let (ξ_n) be a sequence of non-negative increasing random variables. Then,

$$\lim_{n \to \infty} \mathbb{E}(\xi_n) = \mathbb{E} \lim_{n \to \infty} \xi_n.$$

We note that the right hand side limit always exists since for all $\omega \in \Omega$, $\xi_n(\omega)$ is increasing any bounded by ∞.

We remark that the monotone convergence theorem applies if there exists some random variable η such that $\mathbb{E}(\eta) > -\infty$ such that $\eta \leq \xi_n$ for all n by considering $\xi_n - \eta$.

Corollary 4.1. If (η_n) is a sequence of non-negative random variables, then

$$\sum_{i=1}^{\infty} \mathbb{E}(\eta_i) = \mathbb{E}\left(\sum_{i=1}^{\infty} \eta_i\right).$$

Corollary 4.2 (Fatou's lemma)**.** Let ξ_n be a sequence of non-negative random variables. Then,

$$\mathbb{E}(\liminf_n \xi_n) \leq \liminf_n \mathbb{E}\xi_n.$$

Proof. Apply the monotone convergence theorem to $\lambda_n := \inf_{k>n} \xi_k$. □

Again, the non-negative condition can be replaced by the existence of some random variable η such that $\mathbb{E}(\eta) > -\infty$ and $\eta \leq \xi_n$ for all n. On the other hand, if $\mathbb{E}(\eta) < \infty$ and $\xi_n \leq \eta$, the theorem holds with limit supremum instead.

We note that in all above theorems, the statement still holds by replacing Ω with any measurable set by restricting the measure onto that set.

Theorem 5 (Change of Variables)**.** Given a random variable ξ, a measurable function $g : \mathbb{R} \to \mathbb{R}$ and a measurable set A, we have

$$\int_A g \mathrm{d}\xi_* \mathbb{P} = \int_{\xi^{-1}(A)} g \circ \xi \mathrm{d}\mathbb{P},$$

where both integrals either exist or not exist simultaneously.

Proof. Apply usual method where one first prove the statement for indicator functions. Then, it follows that it holds for simple functions by the linearity of the integral. Finally, for any non-negative measurable function, we take a sequence of monotonically increasing simple functions, and apply the monotone convergence theorem. For arbitrary functions, the result follows by taking $f = f^+ - f^-$. □

Corollary 5.1 (Law of the Unconscious Statistician)**.** Given a random variable ξ and a measurable function $g : \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}(g(\xi)) = \int_{\mathbb{R}} g \mathrm{d}\xi_* \mathbb{P}.$$

Corollary 5.2. Let $g : \mathbb{R} \to \mathbb{R}$ be measurable, then, if ξ be a discrete random variable,

$$\mathbb{E}(g(\xi)) = \sum_{x \in A} g(x) p(x).$$

On the other hand, if ξ is absolutely continuous, i.e. there exists some f such that $f\lambda = \xi_* \mathbb{P}$, then

$$\mathbb{E}(g(\xi)) = \int_{\mathbb{R}} g(x) f(x) \lambda(\mathrm{d}x).$$

Proof. In the discrete case, we have

$$\mathbb{E}(g(\xi)) = \int g \mathrm{d}\left(\sum_{x\in A} p(x)\delta_x\right) = \sum_{x\in A} p(x)\int g\mathrm{d}\delta_x.$$

By considering $\int g\mathrm{d}\delta_x = \int_{\{x\}} g\mathrm{d}\delta_x + \int_{\mathbb{R}\setminus\{x\}} g\mathrm{d}\delta_x = \delta_x(\{x\})g(x) + 0 = g(x)$. We have

$$\mathbb{E}(g(\xi)) = \sum_{x\in A} g(x)p(x),$$

as required.

On the other hand, if $f\lambda = \xi_*\mathbb{P}$, we have

$$\mathbb{E}(g(\xi)) = \int_{\mathbb{R}} g\,\mathrm{d}(f\lambda) = \int_{\mathbb{R}} g(x)f(x)\lambda(\mathrm{d}x)$$

as required. □

Theorem 6 (Fubini's Theorem)**.** Let $(E_1, \Sigma_1, \mu_1), (E_2, \Sigma_2, \mu_2)$ be σ-finite measure spaces. Then, for any $\Sigma_1 \otimes \Sigma_2$-measurable functions $g : E_1 \times E_2 \to \mathbb{R}$, $g(\cdot, y_0)$ is Σ_1-measurable for all $y_0 \in E_2$, and $g(x_0, \cdot)$ is Σ_2-measurable for all $x_0 \in E_1$. Furthermore, $\int_{E_1} g\mathrm{d}\mu_1, \int_{E_2} g\mathrm{d}\mu_2$ are Σ_2 and Σ_1-measurable respectively. Finally, if $\int |g|\mathrm{d}\mu_1 \otimes \mu_2 < \infty$, then,

$$\int g\mathrm{d}\mu_1 \otimes \mu_2 = \int \left(\int g(x,y)\mu_2(\mathrm{d}y)\right)\mu_1(\mathrm{d}x) = \int \left(\int g(x,y)\mu_1(\mathrm{d}x)\right)\mu_2(\mathrm{d}y).$$

2.1 Inequalities

Lemma 2.7 (Jensen's Inequality)**.** Let ξ be an integrable random variable and let $g : \mathbb{R} \to \mathbb{R}$ be a measurable, convex function, then,

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

Proof. Recall that the function g is convex if for all $x_0 \in \mathbb{R}$, there exists some λ such that $g(x) \geq g(x_0) + (x - x_0)\lambda$ (graphically, λ is the slope (more accurately, a subderivative) of g at x_0 and so, the inequality is saying that the graph lies above the tangent line).

Setting $x = \xi$ and $x_0 = \mathbb{E}\xi$. Then, the above inequality becomes

$$g(\xi) \geq g(\mathbb{E}\xi) - (\xi - \mathbb{E}\xi)\lambda.$$

Thus, applying the expectation to both sides results in the required inequality by the linearity of the integral. □

Corollary 6.1 (Lyapunov's Inequality)**.** Let ξ be a random variable and let $0 < s < t$ be real numbers, then

$$\mathbb{E}(|\xi|^s)^{1/s} \leq \mathbb{E}(|\xi|^t)^{1/t}.$$

Proof. Use Jensen's inequality with $g(x) = |x|^{t/s}$.

Alternatively, setting $\eta = |\xi|^s$, by Hölder's inequality, we have

$$\|\xi^s\|_1 = \|\eta\|_1 \leq \|\eta\|_{t/s} = \|\xi\|_t^s.$$

Thus, taking both sides to the power of $1/s$, we obtain $\|\xi\|_s = (\|\xi^s\|_1)^{1/s} \leq (\|\xi\|_t^s)^{1/s} = \|\xi\|_t$ as required. □

Proposition 2.2 (Markov Inequality)**.** Let $\xi \geq 0$ be an integrable random variable and let $c > 0$. Then

$$\mathbb{P}(\xi \geq c) \leq \frac{\mathbb{E}(\xi)}{c}.$$

Proof. $\mathbb{E}(\xi) \geq \mathbb{E}(\xi \mathbf{1}_{\xi \geq c}) \geq c\mathbb{E}(\mathbf{1}_{\xi \geq c}) = c\mathbb{P}(\xi \geq c)$. □

Definition 2.8 (Variance)**.** The variance (or dispersion) of a random variable ξ is defined to be

$$V_\xi := \mathbb{E}[(\xi - \mathbb{E}\xi)^2]$$

and we define $\sigma := \sqrt{V_\xi}$ the standard deviation of V_ξ.

By expanding the definition, we fine $V_\xi = \mathbb{E}\xi^2 - (\mathbb{E}\xi)^2$.

Definition 2.9 (Covariance)**.** The covariance of random variables ξ and η is defined to be

$$\mathrm{cov}(\xi, \eta) := \mathbb{E}[(\xi - \mathbb{E})(\eta - \mathbb{E}\eta)].$$

Proposition 2.3. For random variables ξ, η, we have

- $V_{\xi+\eta} = V_\xi + V_\eta + 2\mathrm{cov}(\xi, \eta)$;
- if $\mathrm{cov}(\xi, \eta) = 0$, then $V_{\xi+\eta} = V_\xi + V_\eta$.

Proof. Clear. □

Proposition 2.4 (Chebyshev's Inequality)**.** Let ξ be a integrable random variable. Then, for all $\epsilon > 0$,

$$\mathbb{P}(|\xi - \mathbb{E}\xi| \geq \epsilon) \leq \frac{V_\xi}{\epsilon^2}.$$

Proof. By Markov inequality, for $\xi \geq 0$, we have

$$\mathbb{P}(\xi \geq \epsilon) = \mathbb{P}(\xi^2 \geq \epsilon^2) \leq \frac{\mathbb{E}\xi^2}{\epsilon^2}.$$

Thus, by replacing ξ by $|\xi - \mathbb{E}\xi|$, we have the required inequality. □

Proposition 2.5 (Exponential Chebyshev's Inequality)**.** Let $\xi \geq 0$ be a random variable and let $\epsilon, t > 0$ such that $\xi, e^{t\xi}$ are integrable. Then,

$$\mathbb{P}(\xi \geq \epsilon) \leq e^{-t\epsilon}\mathbb{E}(e^{t\xi}).$$

Proof. We observe, by Markov inequality

$$\mathbb{P}(\xi \geq \epsilon) = \mathbb{P}(e^{t\xi} \geq e^{t\epsilon}) \leq \frac{E(e^{t\xi})}{e^{t\epsilon}}.$$

□

Proposition 2.6 (Tail Probability)**.** Let $\xi \geq 0$ be an integrable rndom variable. Then,

$$\mathbb{E}(\xi) = \int_{[0,\infty)} \mathbb{P}(\xi \geq x)\lambda(\mathrm{d}x).$$

Proof. By change of variable, we have,

$$\mathbb{E}\xi = \int_\Omega \xi \mathrm{d}\mathbb{P} = \int_{[0,\infty)} x(\xi_*\mathbb{P})(\mathrm{d}x) = \int_{[0,\infty)} \int_{[0,x]} \lambda(\mathrm{d}t)(\xi_*\mathbb{P})(\mathrm{d}x).$$

Then, by Fubini's theorem to the function $g : (t, x) \mapsto \mathbf{1}_{[0,x]}(t)$, we have

$$\int_{[0,\infty)} \int_{[0,x]} \lambda(\mathrm{d}t)(\xi_*\mathbb{P})(\mathrm{d}x) = \int_{[0,\infty)^2} g(t,x)\lambda(\mathrm{d}t)(\xi_*\mathbb{P})(\mathrm{d}x) = \int_{[0,\infty)} \mathbb{P}(\xi \geq x)\lambda(\mathrm{d}x)$$

as required. □

Definition 2.10 (Normal Random Variable)**.** A random variable ξ is said to be norm if $\xi_*\mathbb{P} = f\lambda$ where

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

for some $m \in \mathbb{R}, \sigma > 0$. We denote this by $\xi \sim \mathcal{N}(m, \sigma^2)$.

Proposition 2.7. Let $\xi \sim \mathcal{N}(m, \sigma^2)$. Then, $\mathbb{E}\xi = m$ and $V_\xi = \sigma^2$, and so, a normal random variable is fully determined by its mean and variance.

Proof. Exercise. □

Definition 2.11 (Moment)**.** Given a random variable ξ, we define the k-th moment of ξ to be $\mathbb{E}(\xi^k)$.

2.2 Transformation of Random Variables

Let $F_\xi(x)$ be a distribution function of a random variable ξ. Then, if ϕ is a real valued continuous function, we would like to consider the distribution of η where $\eta = \phi(\xi)$. An easy observation is that

$$F_\eta(y) = \mathbb{P}(\eta \leq y) = \mathbb{P}(\eta \in \phi^{-1}(-\infty, y]) = \int_{\phi^{-1}(-\infty,y]} \mathrm{d}(\xi_*\mathbb{P}).$$

As one might expect, elementary methods from first year probability are sufficient for most cases we encounter (consider the case ϕ is linear of quadratic).

Suppose now that ξ is absolutely continuous (and so, has a density by Radon-Nikodym), we would like to find the density of $\eta \coloneqq \phi(\xi)$. Assume first that $\xi(\Omega) \in I$ where I is a finite or infinite open interval and let ϕ be continuously differentiable and strictly increasing on I. Denoting $h(y) = \phi^{-1}(\{y\})$ which is well-defined and differentiable, for all $y \in \phi(I$,

$$F_\eta(y) = \mathbb{P}(\eta \leq y) = \mathbb{P}(\xi \leq \phi^{-1}(\{y\})) = \int_{(-\infty, h(y)]} f_\xi \mathrm{d}\lambda = \int_{(-\infty, y]} f_\xi(h(z))h'(z)\lambda(\mathrm{d}z)$$

where f_ξ is the density of ξ. Hence, the density of η is $f_\xi(h(y))h'(y) = f_\xi(h(y))|h'(y)|$. Similarly, if ϕ is strictly decreasing, η remain to have the density $f_\xi(h(y))|h'(y)|$. With this in mind, we may obtain the density for a large class of transformations by de compositing the density into a strictly increasing and decreasing parts.

In the case that (ξ, η) is a random vector with joint distribution F and let ϕ be a continuous function. Then, $\phi(\xi, \eta)$ has distribution

$$F_{\phi(\xi,\eta)}(z) = \int_{\phi^{-1}(-\infty, z]} \mathrm{d}((\xi, \eta)_* \mathbb{P}) \,.$$

2.3 Independence

We recall that two random variables ξ, η are said to be independent if $(\xi, \eta)_* \mathbb{P} = \xi_* \mathbb{P} \otimes \eta_* \mathbb{P}$. Thus, if F_ξ, F_η are distributions of ξ and η respectively, then, $F_{(\xi,\eta)}(x, y) = F_\xi(x)F_\eta(y)$ for all $x, y \in \mathbb{R}$ where $F_{(\xi,\eta)}$ is a distribution of the random vector (ξ, η).

Proposition 2.8. If ξ, η are independent random variables, then the distribution of $\xi + \eta$ is

$$F_{\xi+\eta}(z) = \int_{\mathbb{R}} F_\eta(z - x)\xi_* \mathbb{P}(\mathrm{d}x) = \int_{\mathbb{R}} F_\xi(z - y)\eta_* \mathbb{P}(\mathrm{d}y).$$

Proof. By taking $\phi : \mathbb{R}^2 \to \mathbb{R} : (x, y) \mapsto x + y$, we have

$$F_{\xi+\eta}(z) = \int_{\phi^{-1}(-\infty, z]} \mathrm{d}((\xi, \eta)_* \mathbb{P}) = \int_{\phi^{-1}(-\infty, z]} \mathrm{d}(\xi_* \mathbb{P} \otimes \eta_* \mathbb{P})$$

by independence. By considering that $(x, y) \in \phi^{-1}(-\infty, z]$ if and only if $x + y \leq< z$, we have by Fubini's theorem,

$$\begin{aligned}
\int_{\phi^{-1}(-\infty, z]} \mathrm{d}(\xi_* \mathbb{P} \otimes \eta_* \mathbb{P}) &= \int_{\mathbb{R}^2} \mathbf{1}_{x+y \leq z} \, \mathrm{d}(\xi_* \mathbb{P} \otimes \eta_* \mathbb{P}) \\
&= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} \mathbf{1}_{x+y \leq z} \eta_* \mathbb{P}(\mathrm{d}y) \right) \xi_* \mathbb{P}(\mathrm{d}x) \\
&= \int_{\mathbb{R}} F_\eta(z - x)\xi_* \mathbb{P}(\mathrm{d}x) = \int_{\mathbb{R}} F_\xi(z - y)\eta_* \mathbb{P}(\mathrm{d}y).
\end{aligned}$$

□

Recalling the definition of the convolution of a function, we may reformulate the above as the following corollaries.

Definition 2.12 (Convolution)**.** Given two real-valued functions $f, g : \Omega \to \mathbb{R}$, we define the convolution of f with g by

$$f * g := t \mapsto \int_{\mathbb{R}} f(x)g(t-x)\mu(\mathrm{d}x).$$

Corollary 6.2. The distribution function of the sum of two independent random variables is the convolution of their distribution functions.

Corollary 6.3. If ξ, η are independent absolutely continuous random variables, then, the density of $\xi + \eta$ is the convolution of their densities.

Proposition 2.9. Let ξ, η be independent integrable random variables. Then, $\xi \cdot \eta$ is integrable and $\mathbb{E}(\xi \cdot \eta) = \mathbb{E}(\xi)\mathbb{E}(\eta)$.

Proof. It is clearly true for indicator functions and so, we may extend to simple function by the linearity of expectation. Hence, by monotone convergence, the statement is true for non-negative random variables and hence true for arbitrary random variables by taking $\xi = \xi^+ - \xi^-$ and $\eta = \eta^+ - \eta^-$. □

Definition 2.13. Random variables ξ, η are said to be uncorrelated if $\mathrm{cov}(\xi, \eta) = 0$.

Proposition 2.10. Independent random variables are uncorrelated.

Proof. Clear since $\mathrm{cov}(\xi, \eta) = \mathbb{E}(\xi \cdot \eta) - \mathbb{E}\xi\mathbb{E}\eta$. □

We note that the converse is not true. Namely, uncorrelated does not imply independence.

2.4 Weak Law of Large Numbers

Consider $(\Omega_n, \mathcal{A}, \mathbb{P}_n)$ as a (finite) probability space such that

$$\Omega_n := \{\omega \mid \omega = (a_1, \cdots, a_n), a_i \in \{0, 1\}\}, \mathcal{A} = \mathcal{P}(\Omega_n)$$

and

$$\mathbb{P}_n(\{\omega\}) = p(\omega) = p^{\sum_{i=1}^n a_i} q^{n - \sum_{i=1}^n a_i}$$

for some $0 < p < 1$ and $q = 1 - p$. Let $\xi_1, \cdots, \xi_n : \Omega_n \to \{0, 1\}$ be random variables such that $\xi_i(\omega) = a_i$. It is easy to check that ξ_i are independent and identically distributed (iid.). Indeed,

$$(\xi_i)_* \mathbb{P}_n(\{1\}) = \mathbb{P}_n(\{\omega \mid a_j(\omega) = 1\}) = p \sum_{k=0}^{n-1} \binom{n-1}{k} p^k q^{n-k} = p$$

and $(\xi_i)_* \mathbb{P}_n(\{0\}) = (\xi_i)_* \mathbb{P}_n(\{1\}^c) = 1 - p = q$.

Now defining $S_n := \sum_{i=1}^n \xi_i$, we observe that

$$\mathbb{E}S_n = \sum_{i=1}^n \mathbb{E}\xi_i = \sum_{i=1}^n p = np.$$

Thus, the expectation of $\frac{1}{n}S_n$ is simply p. We now ask what is $|\frac{1}{n}S_n(\omega)-p|$. Immediately, we observe that $|\frac{1}{n}S_n(\omega)-p|$ cannot tends to 0 point-wise since $S_n(0)=0$ for all n. Nonetheless, we observe that $\mathbb{P}_n(S_n=0)=q^n\to 0$ as $n\to\infty$.

Recalling that the Kolmogorov extension theorem, as $\{\mathbb{P}_n\}$ is consistent, there exists a unique probability measure $\mathbb{P}$ on $(\mathbb{R}^\infty,\mathcal{B}(\mathbb{R}^\infty))$ such that $\mathbb{P}(\xi\in\mathbb{R}^\infty\mid(\xi_1,\cdots,\xi_n)\in B_n)=\mathbb{P}_n((\xi_1,\cdots,\xi_n)\in B_n)$. With this measure in mind, using Chebyshev's inequality, and the fact that they are independent and hence, uncorrelated, we obtain

$$\mathbb{P}\left(\left|\frac{1}{n}S_n-p\right|\geq\epsilon\right)\leq\frac{V(\frac{1}{n}S_n)}{\epsilon^2}=\frac{1}{\epsilon^2}\sum_{j=1}^{n}V\left(\frac{1}{n}\xi_i\right)=\frac{1}{n^2\epsilon^2}\sum_{i=1}^{n}V_{\xi_i}=\frac{npq}{n^2\epsilon^2}\to 0$$

as $n\to\infty$. Thus, $\frac{1}{n}S_n$ converges to p in measure (probability).

This fact is known as Bernoulli's law of large numbers. By noting that we only used the uncorrelated fact, we obtain a more general theorem.

In general (with arbitrary ξ_i), we call the quantity $\frac{1}{n}S_n$ the time average and the (weak) law of large numbers tells us the time average converges in measure to the space average $\mathbb{E}\xi_i$.

Theorem 7 (Weak Law of Large Numbers)**.** let $\xi_1,\xi_2,\cdots$ be integrable random variables. Defining $S_n^{(c)}=\sum_{i=1}^{n}(\xi_i-\mathbb{E}\xi_i)$ (we note that $\mathbb{E}S_n^{(c)}=0$). Then, if $\xi_1,\xi_2\cdots$ are uncorrelated and for all i, $V_{\xi_i}\leq C$ for some $C>0$, for all $\epsilon>0$,

$$\lim_{n\to\infty}\mathbb{P}\left(\left|\frac{1}{n}S_n^{(c)}\right|\geq\epsilon\right)=0$$

i.e. $\frac{1}{n}S_n^{(c)}\to 0$ in measure.

Proof. By Chebyshev's inequality, as ξ_i are uncorrelated,

$$\mathbb{P}\left(\left|\frac{1}{n}S_n^{(c)}\right|\geq\epsilon\right)\leq\frac{V(\frac{S_n^{(c)}}{n})}{\epsilon^2}\leq\frac{c}{n\epsilon^2}\to 0$$

as $n\to\infty$. □

Corollary 7.1. If $\xi_1,\xi_2,\cdots$ integrable iid. random variables such that $V_{\xi_i}<\infty$, then, $\frac{1}{n}\sum_{i=1}^{n}\xi_i$ converges to $\mathbb{E}\xi_i$ in measure.

We would also like to consider the limiting behaviour of the distribution. Let us first recall the big and small-O notation.

Given sequences of functions $(f_n),(g_n)$, we denote $g_n=O_{n\to\infty}(|f_n|)$ if $|g_n/f_n|$ is eventually bounded, i.e. there exists some c,N such that for all $n\geq N$, $|g_n|\leq c|f_n|$. On the other hand, we denote $g_n=o_{n\to\infty}(|f_n|)$ if $\lim_{n\to\infty}|g_n/f_n|=0$.

Returning to our example of the Bernoulli random variables, we have the following result.

Theorem 8 (Local Limit Theorem)**.** For any $0<p<1$,

$$\max_{0\leq k\leq n}\left|\mathbb{P}(S_n=k)-\frac{1}{\sqrt{2\pi np(1-p)}}e^{-\frac{x^2}{2p(1-p)}}\right|=o\left(\frac{1}{\sqrt{n}}\right),$$

where $x = x_{k,n} := \frac{k-np}{\sqrt{n}}$.

Proof. Let $A_n > 0$ such that $A_n = o(n)$ (e.g. $A_n = n^\epsilon$ for some $0 < \epsilon < 1$). Then, let $k \in \mathbb{N}$ such that $|x_{k,n}| \le A_n/\sqrt{n}$ and so,

$$np - A_n \le k \le np + A_n.$$

Then, $k, n-k \to \infty$ as $n \to \infty$ and using Stirling's formula,

$$\begin{aligned}
\mathbb{P}(S_n = k) &= \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \\
&= \frac{\sqrt{2\pi n}}{\sqrt{2\pi k}\sqrt{2\pi(n-k)}} e^{n\log n - k\log k - (n-k)\log(n-k)} p^k (1-p)^{n-k} \left(1 + O\left(\frac{1}{n}\right)\right).
\end{aligned}$$

Then, by noting that $np + x\sqrt{n} = k = np(1 + O(A_n/n))$ and $n(1-p) - x\sqrt{n} = n - k = n(1-p)(1 + O(A_n/n))$, we have

$$\begin{aligned}
\mathbb{P}(S_n = k) &= \sqrt{\frac{n}{np2\pi n(1-p)}} \left(1 + O\left(\frac{A_n}{n}\right)\right) e^{n\log n - (np + x\sqrt{n})(\log(np) + \frac{x}{p\sqrt{n}} - \frac{x^2}{2p^2\sqrt{n}} + O\left(\frac{x}{\sqrt{n}}\right)^3)} \\
&\quad e^{-(n(1-p) - x\sqrt{n})(\log\left(n(1-p) - \frac{x}{(1-p)\sqrt{n}} - \frac{x^2}{2(1-p)^2 n} + O\left(\frac{x}{\sqrt{n}}\right)^3\right))} e^{k\log p + (n-k)\log(1-p)} \\
&= \sqrt{\frac{1}{2\pi np(1-p)}} \left(1 + O\left(\frac{A_n}{n}\right)\right) e^{-(np + x\sqrt{n})(\frac{x}{p\sqrt{n}} - \frac{x^2}{2p^2 n} + O\left(\frac{x}{\sqrt{n}}\right)^3)} \\
&\quad e^{(n(1-p) - x\sqrt{n})(\frac{x}{(1-p)\sqrt{n}} - \frac{x^2}{2(1-p)^2 n} + O\left(\frac{x}{\sqrt{n}}\right)^3)} \\
&= \sqrt{\frac{1}{2\pi np(1-p)}} \left(1 + O\left(\frac{A_n}{n}\right)\right) \\
&\quad e^{-\sqrt{n}x + \frac{x^2}{2p} - \frac{x^2}{p} + \frac{x^3}{2p^2\sqrt{n}}} e^{O\left(\frac{x}{\sqrt{n}}\right)^3 + \sqrt{n}x + \frac{x^2}{2(1-p)} - \frac{x^2}{1-p} - \frac{x^3}{2(1-p^2)\sqrt{n}}} \\
&= \frac{1 + O\left(\frac{A_n}{n}\right)}{\sqrt{2\pi np(1-p)}} e^{-\frac{x^2}{2}(\frac{1}{p} + \frac{1}{1-p}) + O\left(\frac{A_n^3}{\sqrt{n}^3\sqrt{n}}\right)} \\
&= \frac{1}{\sqrt{2\pi p(1-p)n}} e^{-\frac{x^2}{2p(1-p)}} \left(1 + O\left(\frac{A_n^3}{n^2}\right) + O\left(\frac{A_n}{n}\right)\right).
\end{aligned}$$

Hence, choosing $A_n = n^{7/12}$, the result follows for $np - A_n \le k \le np + A_n$. Finally, by monotonicity of $P(S_n = k)$, we have the result follows for all $k \le n$. □

Theorem 9 (de Moivre-Laplace Central Limit Theorem)**.** For any $0 < p < 1$ and $x \in \mathbb{R}$,

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{S_n - np}{\sqrt{np(1-p)}} \le x\right) = \Phi(x),$$

where $\Phi(x)$ is the distribution of the normal random variable with parameters $m = 0, \sigma^2 = 1$.

Proof. Using the local limit theorem, we have

$$\mathbb{P}\left(\frac{S_n - np}{\sqrt{np(1-p)}} \leq x\right) = \mathbb{P}(S_n \leq np + x\sqrt{np(1-p)}) = \sum_{k=0}^{\lfloor np + x\sqrt{np(1-p)}\rfloor} \mathbb{P}(S_n = k)$$
$$= o(1) + \sum_{k=\lfloor np - A_n\rfloor}^{\lfloor np + x\sqrt{np(1-p)}\rfloor} \frac{1}{\sqrt{2\pi}\sqrt{np(1-p)}} e^{-\frac{1}{2}\left(\frac{k-np}{\sqrt{np(1-p)}}\right)^2}$$

which converges to $\int_{(-\infty,x]} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \lambda(\mathrm{d}y)$ as required. □

Finally, as a remark, again using Sterling's formula, we see that

$$\mathbb{P}(S_n = k) = \frac{n(n-1)\cdots(n-k+1)}{k!}\left(\frac{\lambda}{n} + o\left(\frac{1}{n}\right)\right)^k$$
$$= \left(1 - \frac{\lambda}{n} + o\left(\frac{1}{n}\right)\right)^{n-k} \to \frac{1}{k!}\lambda^k e^{-\lambda}$$

as $n \to \infty$ where the limit is the Poisson distribution.

3 Convergence and Characteristic Functions

3.1 Different Modes of Convergence

Let us recall some different notions of convergence in probability theory.

Definition 3.1. A sequence of random variables $(\xi_n)_n$ is said to converge

- almost surely if $\xi_n(\omega) \to \xi(\omega)$ almost everywhere on Ω;
- in probability if for all $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|\xi_n - \xi| \geq \epsilon) = 0;$$

- in L^p for $1 \leq p < \infty$ (also known as convergence in mean of order p) if

$$\lim_{n\to\infty} \mathbb{E}(|\xi_n - \xi|^p) = 0;$$

- in distribution (or weakly) if for any bounded continuous function f,

$$\mathbb{E}(f(\xi_n)) \to \mathbb{E}(f(\xi)).$$

  Namely, this condition requires laws $(\xi_n)_*\mathbb{P}$ converges weakly to $\xi_*\mathbb{P}$. Note that this does not require (ξ_n) to be defined on the same probability space.

Proposition 3.1. We have the following relations between the different notions of convergence.

- Convergence almost surely implies convergence in probability.
- Convergence in probability implies the existence of a convergent almost surely subsequence.
- Convergence in L^p implies convergence in probability.

Proof. Exercise. □

Theorem 10. Let $\xi_n \geq 0$ be a sequence of integrable random variable which converges almost surely to ξ. Then, if $\mathbb{E}\xi_n \to \mathbb{E}\xi$, we have $\xi_n \to \xi$ in L^1.

Proof. We see

$$\mathbb{E}|\xi_n - \xi| = \mathbb{E}(\xi - \xi_n)\mathbf{1}_{\xi \geq \xi_n} + \mathbb{E}(\xi_n - \xi)\mathbf{1}_{\xi_n > \xi} = 2\mathbb{E}(\xi - \xi_n)\mathbf{1}_{\xi \geq \xi_n} + \mathbb{E}(\xi_n - \xi).$$

Thus, the result follows by considering $0 \leq (\xi - \xi_n)\mathbf{1}_{\xi \geq \xi_n} \leq \xi$, and applying dominated convergence. □

Proposition 3.2. Convergence in probability implies convergence in distribution for integrable random variables.

Proof. Let (ξ_n) be a sequence of integrable random variables which converges in probability to ξ. Fix $\epsilon > 0$ and let f be a continuous bounded function such that $|f(x)| \leq C$ for all x. By Markov's inequality, there exists some N such that

$$\mathbb{P}(|\xi| > N) \leq \frac{\epsilon}{4c}.$$

Now, as f is continuous, there exists some $\delta > 0$ such that $|f(x)-f(y)| < \epsilon/2$ for all $|x| \leq N$, $|x-y| \leq \delta$. Then,

$$\begin{aligned}
\mathbb{E}(|f(\xi_n) - f(\xi)|) &= \mathbb{E}(|f(\xi) - f(\xi_n)|)\mathbf{1}_{|\xi_n - \xi| \leq \delta, |\xi| \leq n} \\
&\quad + \mathbb{E}(|f(\xi) - f(\xi_n)|)\mathbf{1}_{|\xi_n - \xi| \leq \delta, |\xi| > n} + \mathbb{E}(|f(\xi) - f(\xi_n)|)\mathbf{1}_{|\xi_n - \xi| > \delta} \\
&\leq \frac{\epsilon}{2} + 2c\frac{\epsilon}{4c} + 2c\mathbb{P}(|\xi_n - \xi| > \delta) \\
&\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} + 2c\mathbb{P}(|\xi_n - \xi| > \delta)
\end{aligned}$$

where the last equality follows as ξ_n converges to ξ in probability. Thus, as the ϵ is arbitrary,

$$\lim_{n\to\infty} \mathbb{E}(|f(\xi_n) - f(\xi)|) = 0.$$

□

3.2 Weak Convergence of Measures

Theorem 11 (Portmanteau Theorem)**.** Let (ξ_n) be a sequence of integrable random variables. Then, the following are equivalent:

1. ξ_n converges in distribution to ξ.
2. $\limsup_n \mathbb{P}(\xi_n \in B) \leq \mathbb{P}(\xi \in B)$ for all closed B.
3. $\liminf_n \mathbb{P}(\xi_n \in A) \geq \mathbb{P}(\xi \in A)$ for all open A.
4. $\lim_n \mathbb{P}(\xi_n \in C) = \mathbb{P}(\xi \in C)$ for all borel C with $\mathbb{P}(\xi \in \partial C) = 0$.
5. The distribution functions F_{ξ_n} converges to F_ξ at all points of continuity of F_ξ.

Proof. $(1 \implies 2)$. Let B be closed and let $f \coloneqq \mathbf{1}_B$ and let $f_\epsilon(x) \coloneqq g(\epsilon^{-1}\mathrm{dist}(x, B))$ where

$$g(t) \coloneqq \begin{cases} 1, & t \leq 0, \\ 1 - t, & 0 \leq t \leq 1, \\ 0, & t > 1. \end{cases}$$

We observe that $f_\epsilon(x)$ is 0 if x is not in a ϵ-neighbourhood of B. In particular, defining $B_\epsilon \coloneqq \{x \mid \mathrm{dist}(x, B)\}$, $f_\epsilon(B_\epsilon^c) = \{0\}$. We note that $B_\epsilon \downarrow B$ as $\epsilon \downarrow 0$ and so $f_\epsilon \to f$. Now, by considering

$$\mathbb{P}(\xi_n \in B) = \int f \,\mathrm{d}(\xi_n)_* \,\mathbb{P} \leq \int f_\epsilon \,\mathrm{d}(\xi_n)_* \,\mathbb{P},$$

we have

$$\limsup_n \mathbb{P}(\xi_n \in B) \leq \limsup_n \int f_\epsilon \,\mathrm{d}(\xi_n)_* \,\mathbb{P} = \int f_\epsilon \mathrm{d}\xi_* \mathbb{P} \leq \mathbb{P}(\xi \in B_\epsilon)$$

where the equality follows by convergence in distribution and the second inequality follows as $f_\epsilon(B_\epsilon^c) \leq 1$. Thus, by the continuity of measure, $\mathbb{P}(\xi \in B_\epsilon) \to \mathbb{P}(\xi \in B)$ as $\epsilon \downarrow 0$ and hence, $\limsup_n \mathbb{P}(\xi_n \in B) \leq \mathbb{P}(\xi \in B)$ as required.

$(2 \iff 3)$. Follows taking the complement.

$(2, 3 \implies 4)$. Taking any borel set C, we recall that $\overline{C} = C \cup \partial C$ and $C^\circ = C \setminus \partial C$. Then, as $\mathbb{P}(\xi \in \partial C) = 0$, we have

$$\limsup_n \mathbb{P}(\xi_n \in C) \leq \limsup_n \mathbb{P}(\xi_n \in \overline{C}) \leq \mathbb{P}(\xi \in \overline{C}) = \mathbb{P}(\xi \in C),$$

and

$$\liminf_n \mathbb{P}(\xi_n \in C) \geq \liminf_n \mathbb{P}(\xi_n \in C^\circ) \geq \mathbb{P}(\xi \in C^\circ) = \mathbb{P}(\xi \in C).$$

Hence, $\lim_n \mathbb{P}(\xi_n \in C) = \mathbb{P}(\xi \in C)$ as required.

$(4 \implies 5)$. Obvious from 5.

$(5 \implies 1)$. This implication is a bit more complicated. The general idea uses the fact that if 5 holds, then there exists a probability space $(\Omega', \mathcal{F}', \mathbb{P}')$ and a sequence of random variables (η_n) on this probability space such that $(\xi_n)_* \mathbb{P}_n = (\eta_n)_* \mathbb{P}'$ (I'm writing $\mathbb{P}_n$ to especially indicate ξ_n are on possibly different measurable spaces) $\xi_* \mathbb{P} = \eta_* \mathbb{P}'$ and $\eta_n \to \eta$ almost surely.

Assuming this theorem (which will will prove below), let f be a continuous bounded function. As f is continuous, $f(\eta_n) \to f(\eta)$ almost surely. Now, as they are bounded, we may apply dominated convergence such that

$$\mathbb{E}(f(\xi_n)) = \mathbb{E}(f(\eta_n)) \to \mathbb{E}(f(\eta)) = \mathbb{E}(f(\xi))$$

where the equalities follows by change of variables. □

Lemma 3.1. Let ξ be a random variable with distribution function F_ξ and let U be a uniformly distributed random variable on the interval $[0, 1]$ where we denote

$$F^{-1}(u) := \sup\{y \mid F_\xi(y) < u\}.$$

Then, $\xi_* \mathbb{P} = (F^{-1}(U))_* \lambda$.

Proof. This follows as $u \leq F_\xi(x) \iff F_\xi^{-1}(u) \leq x$ since F_ξ is nondecreasing and continuous from the right. Indeed, if $u \leq F_\xi(x)$ and suppose for contradiction that $F_\xi^{-1}(u) > x$, then, by definition, $\sup\{y \mid F_\xi(y) < u\} > x$ and so, $\sup\{y \mid F_\xi(y) < u \leq F_\xi(x)\} > x$. However, this is not possible as F_ξ is nondecreasing, thus a contradiction. On the other hand, if $x \leq F_\xi^{-1}(u) = \sup\{y \mid F_\xi(y) < u\}$, as F_ξ is nondecreasing and continuous from the right, $F_\xi(x) \geq F_\xi(\sup\{y \mid F_\xi(y) < u\}) \geq u$. □

Theorem 12 (Single Probability Space Theorem)**.** Let (ξ_n) be a sequence of integrable random variables such that the distribution functions F_{ξ_n} converges to F_ξ at all points of continuity of F_ξ. Then, there exists a probability space $(\Omega', \mathcal{F}', \mathbb{P}')$ and a sequence of random variables (η_n) on this probability space such that $(\xi_n)_* \mathbb{P}_n = (\eta_n)_* \mathbb{P}'$, $\xi_* \mathbb{P} = \eta_* \mathbb{P}'$ and $\eta_n \to \eta$ almost surely.

Proof. Let $(\Omega', \mathcal{F}', \mathbb{P}') := ([0,1], \mathcal{B}([0,1]), \lambda)$. Then, it suffices to prove that $F_{\xi_n}^{-1}(u) \to F_\xi^{-1}(u)$ for all u such that $|F_\xi^{-1}(u)| < \infty$ and the measure of all other points is zero as they form a countable set (nondecreasing function can have at most countably many discontinuities). Indeed, defining $\eta_n := F_{\xi_n}^{-1}(U)$ converges to $\eta := F_\xi^{-1}(U)$ almost surely. Then, by the above lemma, the result follows.

Suppose u is a point such that $|F_\xi^{-1}(u)| < \infty$. Then, for any $x < F_\xi^{-1}(u)$, $F(x) < u$ and if x is a point of continuity of F_ξ, we have $F_{\xi_n}(x) \to F_\xi(x)$ and so, $F_{\xi_n} < u$ for sufficiently large n. Thus, $x \leq F_{\xi_n}^{-1}(u)$ implying $x \leq \liminf_n F_{\xi_n}^{-1}(u)$. Now, by taking a sequence (x_n) of points of continuity of F_ξ such that $x_n \uparrow F^{-1}(u)$, we obtain $F^{-1}(u) \leq \liminf_n F_{\xi_n}^{-1}(u)$.

On the other hand, if $x > F_\xi^{-1}(u)$, we have $F(x) \geq u$. If $F_\xi(x) = u$, then we see that $|F_\xi^{-1}(u)| = \infty$. So, $F_\xi(x) > u$ and from which by the same argument as the less than case, we obtain $F^{-1}(u) \geq \limsup_n F_n^{-1}(u)$ implying convergence as required. □

Definition 3.2 (Relatively Compact)**.** A family of probability measures $\mathcal{P} := \{\mathbb{P}_\alpha \mid \alpha \in A\}$ and the corresponding set of distribution functions F_α is called relatively compact if every sequence of measures of $\mathcal{P}$ has a subsequence which converges weakly (not necessary in $\mathcal{P}$ hence the name "relatively").

We note that the weak convergence of measures is metrizable with the corresponding metric known as the Levy-Prokhorov metric. In this sense, relatively compactness of $\mathcal{P}$ is equivalent to $\overline{\mathcal{P}}$ is compact in this metric space.

Denote $\mathcal{G}$ the set of functions $F : \mathbb{R} \to [0,1]$ which is non-decreasing and continuous from the right (and we call functions of this form generalised distribution functions), we have the following result.

Theorem 13 (Helly's Theorem)**.** The set $\mathcal{G}$ is sequentially compact, i.e. any sequence $(F_n) \subseteq \mathcal{G}$ has a subsequence (F_{n_k}) such that $F_{n_k}(x) \to F(x)$ for all $x \in C_F$.

Proof. Let us enumerate $\mathbb{Q}$ with $(q_i)_{i=1}^\infty$. Then, as $\{F_n(q_1) \mid n\}$ is bounded, it has a convergent subsequence $F_{n_k^{(1)}}(q_1) \to f(q_1)$. Similarly, as $\{F_{n_k^{(1)}}(q_2) \mid n\}$ is bounded, we can find a subsequence $F_{n_k^{(1)}}$, $F_{n_k^{(2)}}$ which converges for both q_1, q_2. Then, iterating this process, for all i, we obtain a subsequence of $F_{n_k^{(i-1)}}$ which converges at $q_1, \cdots, q_i$. Thus, by defining $F_{n_k} := F_{n_k^{(k)}}$, i.e. the diagonal, we have found a subsequence of F_n which converges at all rational numbers. Let us denote the limit of $F_{n_k}(q_i)$ as $f(q_i)$.

Now, defining

$$F(x) := \inf\{f(q) \mid q \in \mathbb{Q}_{>x}\},$$

I claim that $F \in \mathcal{G}$, $F(q) \geq f(q)$ for all $q \in \mathbb{Q}$ and $F(x) \leq f(q)$ for all $x < q$, $q \in \mathbb{Q}$. Indeed, the last inequality is clear by construction while the second inequality follows as f is non-decreasing. Thus, by the definition of inf, F is continuous from the right and so, $F \in \mathcal{G}$.

Finally, it remains to check that $F_{n_k}(x) \to F(x)$ at all points of continuity. Let x be a point of continuity and fix $\epsilon > 0$. As x is a point of continuity, there exists some $y < p < x < q$, such that

$$F(x) - \epsilon < F(y) \leq F(p) \leq F(x) \leq F(q) < F(x) + \epsilon.$$

Then, as $F(y) \leq f(p)$ for all $y < p$, $F(x) - \epsilon < f(p)$. On the other hand, $F(p) \geq f(p)$, so

$$F(x) - \epsilon < f(p) \leq F(p) \leq F(x) \leq f(q) < F(x) + \epsilon.$$

Thus, for sufficiently large k,

$$F(x) - \epsilon < F_{n_k}(p) \leq F_{n_k}(x) \leq F_{n_k}(q) < F(x) + \epsilon,$$

implying $|F_{n_k}(x) - F(x)| < \epsilon$ and hence, $F_{n_k}(x) \to F(x)$ as required. □

Definition 3.3 (Tight)**.** A family of probability measures $\mathcal{P} := \{\mathbb{P}_\alpha \mid \alpha \in A\}$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is called tight if for all $\epsilon > 0$, there exists a compact set $K \subseteq \mathbb{R}$ such that

$$\sup_{\alpha \in A} \mathbb{P}_\alpha(K^c) \leq \epsilon.$$

Theorem 14 (Prokhorov's Theorem)**.** A family of probability measures $\mathcal{P} := \{\mathbb{P}_\alpha \mid \alpha \in A\}$ is tight if and only if it is relatively compact.

Proof. Suppose first $\mathcal{P}$ is relatively compact and suppose for contradiction it is not tight. That is, there exists some $\epsilon > 0$, for all compact $K \subseteq \mathbb{R}$,

$$\sup_{\alpha \in A} \mathbb{P}_\alpha(K^c) > \epsilon.$$

In particular, we may choose $K_n := [-n, n]$ for any n. Thus, for any n, there exists some $\mathbb{P}_{\alpha_n} \in \mathcal{P}$ such that $\mathbb{P}_{\alpha_n}(K_n^c) > \epsilon$. Now, since $\mathcal{P}$ is relatively compact, it has a weakly convergent subsequence $\mathbb{P}_{\alpha_{n_k}} \to \mathbb{P}$ for some probability measure $\mathbb{P}$. As $\mathbb{P}_n$ are probability measures, we have for all $m > k$,

$$\mathbb{P}_{\alpha_m}([-k, k]) \leq \mathbb{P}_{\alpha_m}([-m, m]) < 1 - \epsilon.$$

So, by weak convergence,

$$\mathbb{P}([-k, k]) = \int \mathbf{1}_{[-k,k]} \mathrm{d}\mathbb{P} = \lim_{m \to \infty} \int \mathbf{1}_{[-k,k]} \mathrm{d}\mathbb{P}_{\alpha_{n_m}} < 1 - \epsilon.$$

On the other hand, as $\mathbb{P}$ is a probability measure,

$$1 = \mathbb{P}(\mathbb{R}) = \mathbb{P}\left(\bigcup_{k \in \mathbb{N}} [-k, k]\right) = \lim_{k \to \infty} \mathbb{P}([-k, k]) \leq 1 - \epsilon < 1,$$

which is a contradiction. Hence, $\mathcal{P}$ is tight.

Suppose now $\mathcal{P}$ is tight and let $(\mathbb{P}_n) \subseteq \mathcal{P}$ be a sequence of probability measures with corresponding distribution functions (F_n). By Helly's theorem, there exists a subsequence (F_{n_k}) which converges to some $F \in \mathcal{G}$ at all points of continuity of F. It remains to show $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$.

Fix $\epsilon > 0$. Then, by tightness, there exists some compact K such that $\sup \mathbb{P}_{n_k}(K^c) \leq \epsilon$. Now, since $K \subseteq \mathbb{R}$ it is bounded by some M (choose M such that $\pm M$ is a point of continuity of F) and so, $K \subseteq [-M, M]$ and

$$\sup F_{n_k}(-M) = \sup \mathbb{P}_{n_k}((-\infty, -M]) \leq \sup \mathbb{P}_{n_k}([-M, M]^c) \leq \sup \mathbb{P}_{n_k}(K^c) \leq \epsilon.$$

Thus, as $F(M) \leq \epsilon$, and as it is non-decreasing, $\lim_{x\to-\infty} F(x) = 0$. Similarly, we have

$$\inf F_{n_k}(M) = \inf \mathbb{P}_{n_k}((-\infty, M]) \geq \inf \mathbb{P}_{n_k}([-M, M]) \geq \inf \mathbb{P}_{n_k}(K) \geq 1 - \epsilon.$$

Thus, $\lim_{x\to\infty} F(x) = 1$. Hence, $\mathbb{P}_{n_k}$ converges to the probability measure corresponding to $\mathbb{P}$ weakly by Portmanteau's theorem as required. □

Corollary 14.1. If F_{ξ_n} is a sequence of distribution functions which tends to the distribution function F at all points of continuity of F, then $\{\xi_n\}$ is tight.

To check F is indeed a distribution function, by recalling Helly's theorem, it suffices to check $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to-\infty} F(x) = 1$.

Prokhorov's theorem remains true for in a much general context. Namely, Prokhorov's theorem holds for measures on $\mathbb{R}^n, \mathbb{R}^\infty$ and measures on any complete separable metric space with a Borel σ-algebra.

4 Characteristic Functions

Definition 4.1 (Characteristic Function)**.** The characteristic function of a random variable ξ is

$$\phi_\xi(t) := \mathbb{E}e^{it\xi} = \int e^{itx} \mathrm{d}F_\xi(x),$$

which is the Fourier-Stieltjes transform of the distribution F_ξ.

If $\xi = (\xi_1, \cdots, \xi_n)$ is a random vector, then the characteristic function of ξ is

$$\phi_\xi(t_1, \cdots, t_n) := \mathbb{E}e^{i\sum_{k=1}^n t_k \xi_k}.$$

By the very definition of the characteristic function, we observe the following properties of the characteristic function.

- If ξ is a random variable and a, b are constants, then defining $\eta := a\xi + b$, we have
$$\phi_\eta(t) = e^{itb}\mathbb{E}e^{iat\xi}.$$
- $|\phi(t)| \leq \phi(0) = 1$.
- If ξ is a random variable, then ϕ_ξ is uniformly continuous on $\mathbb{R}$. Indeed, we observe
$$|\phi(t+h) - \phi(t)| = |\mathbb{E}e^{it\xi}(e^{ih\xi} - 1)| \leq \mathbb{E}|e^{ih\xi} - 1|$$
where the last term is independent of t and tends to 0 as $h \to 0$ by dominated convergence.
- If $\xi_1, \cdots, \xi_n$ are independent random variables, and we define $S := \xi_1 + \cdots + \xi_n$, then
$$\phi_S(t) = \prod_{k=1}^n \phi_{\xi_k}(t).$$

Proposition 4.1. If $\xi \sim \mathcal{N}(m, \sigma^2)$, then $\phi_\xi(t) = e^{itm - \frac{t^2\sigma^2}{2}}$.

Proof. Let $\eta = (\xi - m)/\sigma$ so that $\eta \sim \mathcal{N}(0, 1)$. Then, $\xi = \sigma\eta + m$ and hence, by the previous result,

$$\phi_\xi(t) = e^{itm}\mathbb{E}e^{i\sigma\eta}.$$

Thus, it is sufficient to show $\phi_\eta(t) = e^{-t^2/2}$. Indeed, we compute

$$\begin{aligned}\phi_\eta(t) = \mathbb{E}e^{it\eta} &= \frac{1}{\sqrt{2\pi}}\int e^{itx} f_\eta(x)\lambda(\mathrm{d}x) = \frac{1}{\sqrt{2\pi}}\int e^{itx - \frac{x^2}{2}}\lambda(\mathrm{d}x) \\ &= e^{-\frac{t^2}{2}}\frac{1}{\sqrt{2\pi}}\int e^{-\frac{1}{2}(x-it)^2}\lambda(\mathrm{d}x) = e^{-\frac{t^2}{2}}\frac{1}{\sqrt{2\pi}}\int_{-\infty-it}^{\infty+it} e^{-\frac{z^2}{2}}\lambda(\mathrm{d}z).\end{aligned}$$

Then, by contour integration, we find $\int_{-\infty-it}^{\infty+it} e^{-\frac{z^2}{2}}\lambda(\mathrm{d}z) = \sqrt{2\pi}$ and so, $\phi_\eta(t) = e^{-t^2/2}$ as required. □

4.1 Moments and Inversion

Theorem 15. Let ξ be a random variable with characteristic function ϕ_ξ and distribution F_ξ. Then, if

- $\mathbb{E}|\xi|^n < \infty$ for some $n \geq 1$, then $\phi_\xi^{(r)}$ exists for any $0 \leq r \leq n$ and is given by

$$\phi_\xi^{(r)}(t) = \int_{\mathbb{R}} (ix)^r e^{ixt} \mathrm{d}F_\xi(x).$$

  Furthermore, $\mathbb{E}\xi^r = i^{-r}\phi^{(r)}(0)$. By Taylor's theorem,

$$\phi(t) = \sum_{r=0}^{n-1} \frac{(it)^r}{r!}\mathbb{E}\xi^r + \frac{it^n}{n!}\epsilon_n(t)$$

  where $|\epsilon_n(t)| \leq 3\mathbb{E}|\xi|^n$ and $\epsilon_n(t) \to 0$ as $t \to 0$.

- $\mathbb{E}|\xi|^n < \infty$ for all $n \geq 1$ and

$$\limsup_{n\to\infty} \frac{(\mathbb{E}|\xi|^n)^{1/n}}{n} = \frac{1}{eT} < \infty$$

  for some T (note that $(\mathbb{E}|\xi|^n)^{1/n} = \|\xi\|_n$), then

$$\phi(t) = \sum_{n=0}^{\infty} \frac{(it)^n}{n!}\mathbb{E}\xi^n$$

  for all $|t| < T$.

Proof. By noting that $\mathbb{E}|\xi|^n = \|\xi\|_n^n$, and $L_p(\Omega) \subseteq L_q(\Omega)$ for $q \leq p$ and Ω a finite measure space, $\mathbb{E}|\xi|^n < \infty$ implies $\xi \in L_p$ for all $p \leq n$ and thus, $\mathbb{E}|\xi|^p < \infty$ for all $p \leq n$. By noting that

$$\left|\frac{e^{ihx} - 1}{h}\right| \leq x$$

by dominated convergence and l'Hopital's rule,

$$\begin{aligned}\phi'_\xi(t) &= \lim_{h\to 0} \frac{\phi(t+h) - \phi(t)}{h} = \lim_{h\to 0} \mathbb{E}\left(e^{it\xi}\frac{e^{ih\xi} - 1}{h}\right) = i\mathbb{E}\xi e^{it\xi} \\ &= i\int_{\mathbb{R}} xe^{itx}\mathrm{d}F_\xi(x).\end{aligned}$$

With this, we obtain the r-th derivative by induction.

By Taylor's theorem, $e^{it\xi} = \sum_{k=0}^{n-1} \frac{(it\xi)^k}{k!} + \frac{(it\xi)^n}{n!}(\cos t\xi\theta_1(\omega) + i\sin t\xi\theta_2(\omega))$. Thus, as the sum is finite, by the linearity of expectation,

$$\mathbb{E}e^{it\xi} = \sum_{k=0}^{n-1} \frac{(it)^k}{k!}\mathbb{E}\xi^k + \frac{(it\xi)^n}{n!}(\mathbb{E}\xi^n + \epsilon_n(t))$$

where $\epsilon_n(t) = \mathbb{E}(\xi^n(\cos t\xi\theta_1 + i\sin t\xi\theta_2 - 1))$. Thus, $|\epsilon_n(t)| \leq \mathbb{E}|3\xi^n| = 3\mathbb{E}|\xi|^n$ and $\epsilon_n(t) \to 0$ as $t \to 0$ by dominate convergence.

Let $0 < t_0 < T$. Then,

$$\limsup_{n\to\infty} \frac{(\mathbb{E}|\xi|^n)^{1/n}}{n} < \frac{1}{et_0}$$

implying

$$\limsup_{n\to\infty} \left(\frac{\mathbb{E}|\xi|^n}{n^n} t_0^n e^n\right)^{1/n} < 1.$$

We note that this formula is very similar to Stirling's formula and

$$\limsup_{n\to\infty} \left(\frac{\mathbb{E}|\xi|^n t_0^n}{n!}\right)^{1/n} < L < 1$$

and so $\sum_{n=0}^{\infty} \frac{\mathbb{E}|\xi|^n}{n!} t_0^n$ converges by the root test. □

With this theorem in mind, we see that if $\mathbb{E}|\xi|^n < \infty$ exists for all $n \geq 1$, the characteristic function ϕ_ξ is uniquely determined on $[-T, T]$ by the moments of ξ. Furthermore, taking $s = T/2$, by the same proof, we obtain

$$\phi_\xi(t) = \sum_{k=0}^{\infty} i^k \frac{(t-s)^k}{k!} \phi_\xi^{(k)}(s)$$

where $\phi_\xi^{(k)}(s) = \mathbb{E}\xi^k e^{is\xi}$ for $-T/2 < s < T/2$. Thus, ϕ_ξ is uniquely determined by moments on $3/2T$. Continuing this process, we find $\phi_\xi(t)$ is uniquely determined by moments for all t.

A stronger sufficient condition for the unique determination of ϕ_ξ is that

$$\sum_{n=0}^{\infty} \frac{1}{(\mathbb{E}\xi^{2n})^{1/2n}} = \infty.$$

This is called Carleman's test. We will omit the proof. In some sense, if $\mathbb{E}\xi^n$ grows too fast, there may be multiple $\phi(t)$ with these moments.

Theorem 16 (Inversion)**.** Let ξ be a random variable with characteristic function ϕ_ξ and distribution F_ξ. Then, if $a < b$ are points of continuity of F_ξ,

$$F_\xi(b) - F_\xi(a) = \lim_{T\to\infty} \frac{1}{2\pi} \int_{[-T,T]} \frac{e^{-ita} - e^{-itb}}{it} \phi_\xi(t) \lambda(\mathrm{d}t).$$

Furthermore, if $\int |\phi| \mathrm{d}\lambda < \infty$, then F_ξ is absolutely continuous with density f_ξ such that

$$f_\xi(x) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-itx} \phi_\xi(t) \lambda(\mathrm{d}t).$$

By recalling that the characteristic function is the Fourier-Stieltjes transform of F_ξ, this is simply the Fourier-Stieltjes inverse formula.

Proof. If F_ξ is absolutely continuous, the Lebesgue-Stieltjes measure λ_{F_ξ} is absolutely continuous with respect to the Lebesgue measure λ. Thus, by Radon-Nikodym, there exists some f such that $\lambda_{F_\xi} = f\lambda$. Thus,

$$\phi_\xi(t) = \int_{\mathbb{R}} e^{itx} f(x)\lambda(\mathrm{d}x)$$

implying $\int |\phi_\xi|\mathrm{d}\lambda = \int |f|\mathrm{d}\lambda = 1 < \infty$ and so ϕ_ξ is integrable. Then, by the inverse Fourier transform, provided f satisfy some regularity conditions (Dini's condition, c.f. Fourier analysis course notes)

$$f(x) = \frac{1}{2\pi}\int_{\mathbb{R}} e^{-itx}\phi_\xi(t)\lambda(\mathrm{d}t).$$

Hence, integrating this and applying Fubini's theorem,

$$F(b) - F(a) = \int_{(a,b]} f\mathrm{d}\lambda = \frac{1}{2\pi}\int_{(a,b]\times\mathbb{R}} e^{-itx}\phi_\xi(t)\lambda(\mathrm{d}t)\lambda(\mathrm{d}x) = \frac{1}{2\pi}\int_{\mathbb{R}} \frac{e^{-ita} - e^{-itb}}{it}\phi_\xi(t)\lambda(\mathrm{d}t).$$

It remains to show that f satisfy Dini's condition which we shall omit.

Now, let us consider the general case. Define

$$\phi_T := \frac{1}{2\pi}\int_{[-T,T]} \frac{e^{-ita} - e^{-itb}}{it}\phi(t)\lambda(\mathrm{d}t) = \int \psi_T(x)\mathrm{d}F_\xi(\mathrm{d}x)$$

where $\psi_T)x := \frac{1}{2\pi}\int_{[-T,T]} \frac{e^{-ita}-e^{-itb}}{it}e^{itx}\lambda(\mathrm{d}t)$, where the last equality follows by Fubini's theorem. Now, by noting

$$\begin{aligned}\psi_T(x) &= \frac{1}{2\pi}\int_{[-T,T]} \frac{\sin(t(x-a)) - \sin(t(x-b))}{t}\lambda(\mathrm{d}t)\\ &= \frac{1}{2\pi}\left(\int_{-T(x-a)}^{T(x-a)} - \int_{-T(x-b)}^{T(x-b)}\right)\frac{\sin u}{u}\lambda(\mathrm{d}u),\end{aligned}$$

and $\int_{[-R,R]} \frac{\sin u}{u}\lambda(\mathrm{d}u) \to \pi$ as $R \to \infty$, there exists some $c > 0$, such that $|\psi_T(x)| < c$ for all x. Furthermore, by observing that

$$\psi_T(x) \to \psi(x) = \begin{cases} 0, & x \notin [a,b],\\ 1/2, & x \in \{a,b\},\\ 1, & x \in (a,b),\end{cases}$$

by dominated convergence, as $T \to \infty$,

$$\phi_T = \int \psi_T \mathrm{d}F \to \int \psi \mathrm{d}F = F(b-) - F(a) + \frac{1}{2}(F(a) - F(a-) + F(b) - F(b-)).$$

Finally, as a, b are points of continuity, $\phi_T \to F(b) - F(a)$ as claimed.

Now, if $\int |\phi|\mathrm{d}\lambda < \infty$, the function

$$f(x) := \frac{1}{2\pi}\int e^{-itx}\phi(t)\lambda(\mathrm{d}t).$$

exists and by dominated convergence, is continuous, differentiable (and is integrable on $[a, b]$). Then, by Fubini's theorem,

$$\int_{[a,b]} f \mathrm{d}\lambda = \int \frac{1}{2\pi} \phi(t) \frac{e^{-ita} - e^{-itb}}{it} \lambda(\mathrm{d}t) = F(b) - F(a),$$

where $F(x) = \int_{(-\infty,x]} f \mathrm{d}\lambda$. Thus, F is absolutely continuous and by the arguments above, we conclude the proof. □

Corollary 16.1. Probability distributions on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and characteristic functions are in 1-to-1 correspondence.

Proof. Since points of continuity of a distribution function is dense, by the continuity of measure to define $F(b) - F(a)$ uniquely for any a, b from any characteristic functions. □

We will now provide some useful theorems without proof.

Theorem 17 (Bochner)**.** Let ϕ be continuous on $\mathbb{R}$ such that $\phi(0) = 1$. Then, ϕ is a characteristic function if and only if ϕ is positive semi-definite, i.e.

$$\sum_{k,j=1}^{n} \phi(t_j - t_k) \alpha_j \overline{\alpha}_k$$

for all $t_1, \cdots, t_n \in \mathbb{R}$ and $\alpha_1, \cdots, \alpha_n \in \mathbb{C}$.

Theorem 18 (Polya)**.** Let $\phi \geq 0$ be continuous, even, $\phi(0) = 1, \lim_{x\to\infty} \phi(x) = 0$ and is convex on $0 \leq t < \infty$. Then ϕ is a characteristic function.

An example of the above is $e^{-|t|}$.

Theorem 19 (Marcinkievicz)**.** If a characteristic function is of the form $e^{p(t)}$ where $p(t)$ is a polynomial, then p has degree at most 2.

Definition 4.2 (Cumulant)**.** If there exists an expansion such that

$$\log \phi_\xi(t) = \sum_{k=0}^{n} \frac{(it)^k}{k!} S_k + o_{t\to 0}(|t|^n),$$

then the coefficients S_k are called cumulants of ξ.

As an exercise, one can show $\mathbb{E}\xi = S_1$ and $V_\xi = S_2$.

In general, by the Marcinkievicz theorem, if there ξ is a random variable such that $S_k = 0$ for all $k \geq n$ for some n, then, $S_k = 0$ for all $k \geq 3$. One may show that if ξ is normal, then this condition is satisfies and thus, as this determines the characteristic function which in turn determines the distribution, if there exists some n such that $S_k = 0$ for all $k \geq n$, then $\xi \sim \mathcal{N}(S_1, S_2)$.

Theorem 20. Let ϕ be a characteristic function of ξ such that $|\phi(t_0)| = 1$ for some $t_0 \neq 0$. Then, ξ is a pure point random variable.

Proof. As $|\phi(t_0)| = 1$, there exists some $a \in \mathbb{R}$ such that $\phi(t_0) = e^{iat_0}$. Then, by the definition of characteristic functions,

$$\int e^{it_0 x} \mathrm{d}F_\xi(x) = e^{iat_0}.$$

Thus, by comparing the real parts, we obtain

$$1 = \int \cos(t_0(x-a)) \mathrm{d}F_\xi(x)$$

and so $\int 1 - \cos(t_0(x-a)) \mathrm{d}F_\xi(x) = 0$. But $1 - \cos(t_0(x-a)) \geq 0$ implying $\cos(t_0(x-a)) = 1$ almost everywhere with respect to λ_{F_ξ}. Thus, λ_{F_ξ} is concentrated at points of the form $a + \frac{2\pi}{t_0} n$ for $n \in \mathbb{Z}$. □

4.2 Central Limit Theorem

We had previously prove the central limit theorem for Bernoulli random variables. However, as we know, the central limit theorem applies to a much wider class of random variables. We shall in this section prove the general central limit theorem using the characteristic functions.

Theorem 21 (Levy's Continuity Theorem)**.** Let ϕ_n be a sequence of characteristic functions of distribution functions F_n. Then,

- if $F_n \to F$ at all points of continuity (i.e. the corresponding random variables converges weakly) where F is a distribution function, then $\phi_n \to \phi$ point-wise where ϕ is the characteristic function of F, i.e.

$$\phi(t) = \int e^{itx} \mathrm{d}F(x).$$

- if $\lim_{n\to\infty} \phi_n(t) = \phi(t)$ exists and ϕ is continuous at 0, then ϕ is a characteristic function of some distribution F and $F_n \to F$ at all points of continuity.
- if ϕ_n corresponds to F_n and ϕ corresponds to F, then $\phi_n \to \phi$ point-wise if and only if $F_n \to F$ as all points of continuity.

Proof. The first statement is obvious by the very definition of weak convergence. While the third statement is implied by the first two. We omit the second. □

Theorem 22 (Central Limit Theorem for i.i.d. Random Variables)**.** Let (ξ_n) be independent and identically distributed random variables such that $\mathbb{E}\xi_1^2 < \infty$ and nondegenerate (i.e. $V_\xi \neq 0$. Then, defining $S_n = \sum_{i=1}^n \xi_n$,

$$\mathbb{P}\left(\frac{S_n - \mathbb{E}S_n}{\sqrt{V_{S_n}}} \leq x \right) \to \frac{1}{\sqrt{2\pi}} \int_{(-\infty, x]} e^{-y^2/2} \lambda(\mathrm{d}y)$$

as $n \to \infty$ for all $x \in \mathbb{R}$.

Namely, the central limit theorem states that $\frac{S_n - \mathbb{E}S_n}{\sqrt{V_{S_n}}}$ converges in distribution to $\mathcal{N}(0,1)$.

Proof. Suppose $m = \mathbb{E}\xi$ and $\sigma^2 = V_\xi$. Then, setting $\phi(t) = \mathbb{E}e^{it(\xi - m)}$ and $\phi_n(t) = \mathbb{E}e^{it\frac{S_n - \mathbb{E}S_n}{\sqrt{V_{S_n}}}}$. Then, by independence,

$$\phi_n(t) = \left(\phi\left(\frac{t}{\sigma\sqrt{n}}\right)\right)^n.$$

As $\mathbb{E}\xi^2 < \infty$, $\phi(t) = 1 - \frac{\sigma^2 t^2}{2} + o(t^2)$. So, $\phi_n(t) = \left(1 - \frac{t^2}{2n} + o\left(\frac{1}{2n}\right)\right)^n$ which tendsto $e^{-t^2/2}$ as $n \to \infty$. Thus, as this is the characteristic function of the standard normal distribution, by the continuity theorem, we obtain the result. □

Theorem 23 (Central Limit Theorem for Independent Random Variables)**.** Let (ξ_n) be independent with $\mathbb{E}\xi_i^2 < \infty$ and distribution function F_i for all i. Then, denoting $m_i = \mathbb{E}\xi_i$, $\sigma_i^2 = V_{\xi_i}$, $S_n = \sum_{i=1}^n \xi_n$ and $D_n^2 = \sum_{i=1}^n \sigma_i^2$. Then, if the Lindenberg condition holds, namely for all $\epsilon > 0$, $\frac{1}{D_n^2}\sum_{k=1}^n \int_{\{|x-m_k| \geq \epsilon D_n\}} (x - m_k)^2 \mathrm{d}F_k(x)$ which tendsto 0 as $n \to \infty$. Then,

$$\frac{S_n - \mathbb{E}S_n}{\sqrt{V_{S_n}}} \to \mathcal{N}(0,1)$$

in distribution.

Let us note some conditions which implies the Lindenberg condition.

- Lyapunov condition: there exists some $\delta > 0$ such that

  $$\frac{1}{D_n^{2+\delta}} \sum_{k=1}^n (|\xi_k - m_k|^{2+\delta}) \to 0$$

  as $n \to \infty$.

  Proof. Indeed, for all $\epsilon > 0$,

  $$\begin{aligned}\mathbb{E}|\xi_k - m_k|^{2+\delta} &= \int |x - m_k|^{2+\delta} \mathrm{d}F_k(x) \\ &\geq (\epsilon D_n)^\delta \int_{\{|x-m_k| \geq \epsilon D_n\}} |x - m_k|^2 \mathrm{d}F_k(x).\end{aligned}$$

  Thus,

  $$\frac{1}{D_n^2} \sum_{k=1}^n \int_{\{|x-m_k| \geq \epsilon D_n\}} (x - m_k)^2 \mathrm{d}F_k(x) \leq \frac{1}{\epsilon^\delta D^{2+\delta}} \sum_{k=1}^n \mathbb{E}|\xi_k - m_k|^{2+\delta} \to 0$$

  allowing us to conclude. □

- There exists some $K > 0$ such that $|\xi_k| \leq K$ for all k and $D_n \to \infty$ as $n \to \infty$.

  Proof. Exercise. □

We will now consider the rate of convergence.

Theorem 24 (Berry-Esseen Inequality)**.** Let (ξ_k) be i.i.d. random variables with $\mathbb{E}|\xi_1|^3 < \infty$. Then

$$\sup_x \left| \mathbb{P}\left(\frac{S_n - \mathbb{E}S_n}{\sqrt{V_{S_n}}} \leq x \right) - \Phi(x) \right| \leq \frac{c\mathbb{E}|\xi_1 - \mathbb{E}\xi_1|^3}{\sigma^3\sqrt{n}}$$

where Φ is the distribution function of a standard normal distribution and c is a constant such that $\frac{1}{\sqrt{2\pi}} \leq c \leq 1/2$.

Proof. Omitted. □

We remark that the decay $o(1/\sqrt{n})$ is the optimal rate of convergence by considering a sequence of Bernoulli random variables taking values in ± 1.

We note that we required the random variables to have finite second moment in the statement of CLT. This condition is necessary. Indeed, if (ξ_n) is a sequence of i.i.d. Cauchy random variables with density

$$f(x) = \frac{\theta}{\pi(x^2 + \theta^2)}, \theta > 0.$$

It is easy to show that $\mathbb{E}\xi^2 = \infty$ and

$$\phi_\xi(t) = \frac{\theta}{\pi} \int \frac{e^{itx}}{x^2 + \theta^2} \lambda(\mathrm{d}x) = e^{-\theta|t|}.$$

Thus, by independence,

$$\phi_{S_n/n}(t) = \left(e^{-\theta|t|/n}\right)^n = e^{-\theta|t|} = \phi_\xi(t).$$

Hence, S_n/n also has the Cauchy distribution!

5 Results in Probability Theory

We will in this section consider some important results in probability theory. In particular, we will study the Borel-Cantelli lemmas, Kolmogorov's inequality and the strong law of large numbers.

5.1 Borel-Cantelli Lemmas

Theorem 25 (First Borel-Cantelli Lemma)**.** Let (A_n) be a sequence of measurable sets. Then, if $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$, we have

$$\mathbb{P}(\{A_n \text{ i.o.}\}) = \mathbb{P}\left(\limsup_{n\to\infty} A_n\right) = \mathbb{P}\left(\bigcap_{n=0}^{\infty} \bigcup_{k\geq n} A_k\right) = 0.$$

Proof. Define $B_n := \bigcup_{k\geq n} A_k$ so that B_n is decreasing. Then, by continuity,

$$\mathbb{P}\left(\bigcap_{n=0}^{\infty} B_n\right) = \lim_{n\to\infty} \mathbb{P}(B_n) = \lim_{n\to\infty} \mathbb{P}\left(\bigcup_{k\geq n} A_k\right) \leq \lim_{n\to\infty} \sum_{k\geq n} \mathbb{P}(A_k) = 0.$$

Hence, as $\bigcap_{n=0}^{\infty} B_n = \{A_n \text{ i.o.}\}$, we have the required result. □

Theorem 26 (Second Borel-Cantelli Lemma)**.** Let (A_n) be a sequence of **independent** measurable sets. Then, if $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$, we have

$$\mathbb{P}(\{A_n \text{ i.o}\}) = \mathbb{P}\left(\limsup_{n\to\infty} A_n\right) = \mathbb{P}\left(\bigcap_{n=0}^{\infty} \bigcup_{k\geq n} A_k\right) = 1.$$

Proof. Consider $\{A_n \text{ i.o.}\}^c = \bigcup_{n=1}^{\infty} \bigcap_{k\geq n} A_k^c$, so by continuity and independence,

$$1 - \mathbb{P}(\{A_n \text{ i.o.}\}) = \lim_{n\to\infty} \mathbb{P}\left(\bigcap_{k\geq n} A_k^c\right) = \lim_{n\to\infty} \prod_{k\geq n} \mathbb{P}(A_k^c).$$

Now, by noting $\log(1-x) \leq -x$ for all $x \in [0,1)$, we have

$$\log \mathbb{P}\left(\bigcup_{k\geq n} A_k^c\right) = \log \prod_{k\geq n} (1 - \mathbb{P}(A_k)) \leq -\sum_{k\geq n} \mathbb{P}(A_k)$$

where $\sum_{k\geq n} \mathbb{P}(A_k) = \infty$ for all n.

Hence, $\log \mathbb{P}\left(\bigcup_{k\geq n} A_k^c\right) = -\infty$ and thus, $1 - \mathbb{P}(\{A_n \text{ i.o.}\}) = 0$ as required. □

We will provide a new proof for the existence of a subsequence which converges almost everywhere from a sequence which converges in probability.

Corollary 26.1. If $\xi_n \to \xi$ in probability then there exists a subsequence $(\xi_{n_k}) \subseteq (\xi_n)$ which converges almost everywhere to ξ.

Proof. Since $\lim_{n\to\infty} \mathbb{P}(|\xi_n - \xi|) > 1/k) = 0$ for all k, there exists a subsequence (ξ_{n_k}) of (ξ_n) such that

$$\mathbb{P}(|\xi_{n_k} - \xi| > 1/k) \leq 2^{-k}.$$

Hence, as $\sum_{k=1}^{\infty} \mathbb{P}(|\xi_{n_k} - \xi| > 1/k) \leq \sum_{k=1}^{\infty} 2^{-k} = 1 < \infty$, the first Borel-Cantelli lemma implies $\mathbb{P}(\{|\xi_{n_k} - \xi| > 1/k \text{ i.o.}\}) = 0$ which is exactly the set of elements which do not converge. □

Corollary 26.2. If (ξ_n) is a sequence of random variables such that $\xi_1 \geq \xi_2 \geq \cdots \geq 0$ and $\xi_n \to 0$ in probability. Then, ξ_n converges almost everywhere to 0.

Proof. As $\xi_n \to 0$ in probability, it has a subsequence which converges to 0 almost everywhere. Then, as ξ_n is decreasing, a element ω satisfying $\xi_{n_k}(\omega) \to 0$ also satisfies $\xi_n(\omega) \to 0$. Thus, the set of non-convergent elements has measure 0 as required. □

5.2 Strong Law of Large Numbers

Definition 5.1. A sequence of random variables (ξ_n) is said to satisfy the weak/strong law of large numbers (LLN) if ξ_n are integrable and $\frac{S_n - \mathbb{E}S_n}{n}$ converges in probability/almost everywhere to 0 where $S_n := \sum_{i=1}^{n} \xi_i$.

We had previously proved the weak law of large numbers for sequences of i.i.d. random variables with finite variance. We will now consider strong LLN.

Lemma 5.1 (Cantelli's Strong Law of Large Numbers)**.** Let (ξ_n) be a sequence of i.i.d random variables such that $\mathbb{E}\xi_1^4 < \infty$. Then, (ξ_n) satisfies the strong LLN.

Proof. WLOG. we may assume $\mathbb{E}\xi_1 = 0$ and so $\mathbb{E}S_n = 0$. Let $\epsilon > 0$, and define

$$A_n := \{\omega \mid |S_n(\omega)/n| > \epsilon\}.$$

Then, it suffices to show that $\mathbb{P}(\limsup |S_n/n| > \epsilon) = \mathbb{P}(\{A_n \text{ i.o.}\}) = 0$. Indeed, this follows by Borel-Cantelli where

$$\mathbb{P}(A_n) < \frac{1}{\epsilon^4}\mathbb{E}\left(\left|\frac{S_n}{n}\right|^4\right)$$

by Chebyshev's inequality and so, by considering

$$\begin{aligned}\mathbb{E}S_n^4 &= \sum_{i=1}^{n} \mathbb{E}\xi_i^4 + \sum_{i<j} \binom{4}{2} \mathbb{E}\xi_i^2 \mathbb{E}\xi_j^2 + \mathbb{E}\cdots = n\mathbb{E}\xi_1^4 + 3n(n-1)(\mathbb{E}\xi_1^2)^2 \\ &\leq n\mathbb{E}\xi_1^4 + 3n(n-1)\mathbb{E}\xi_1^4 \leq n^2(4\mathbb{E}\xi_1^4)\end{aligned}$$

where $\mathbb{E}\cdots = 0$ as they represent terms with a singular ξ_i, and hence, by independence $\mathbb{E}\cdots = \mathbb{E}\xi_i\mathbb{E}\cdots = 0$. Thus, $\sum_n \mathbb{E}(|S_n/n|^4) \leq \sum_n 1/n^2 < \infty$, implying $\sum_n \mathbb{P}(A_n) < \infty$ which is the condition needed for Borel-Cantelli. □

Theorem 27 (Kolmogorov's Inequality)**.** Let (ξ_n) be a sequence of independent random variables with finite variance (we note that this is implied by finite second (or higher) moment by Lyapunov inequality). Then, for all $n \geq 1$, $\epsilon > 0$,

$$\mathbb{P}\left(\max_{1\leq k\leq n} |S_k - \mathbb{E}S_k| \geq \epsilon\right) \leq \frac{V_{S_n}}{\epsilon^2}.$$

Proof. WLOG. we may assume $\mathbb{E}\xi_i = 0$. Defining $A := \{\max_{1\le k\le n} |S_k| \ge \epsilon\}$, and

$$A_k := \{|S_j| < \epsilon, j = 1, \cdots, k-1, |S_k| \ge \epsilon\}.$$

It is clear that A_k are mutually disjoint, $A = \bigcup_{i=1}^n A_i$ and, $\mathbb{E}S_n^2 \ge \mathbb{E}S_n^2 \mathbf{1}_A = \sum_{i=1}^n \mathbb{E}S_n^2 \mathbf{1}_{A_i}$. Considering

$$\begin{aligned}\mathbb{E}S_n^2 \mathbf{1}_{A_i} &= \mathbb{E}(S_i + \xi_{i+1} + \cdots + \xi_n)^2 \mathbf{1}_{A_i} \\ &= \mathbb{E}S_i^2 \mathbf{1}_{A_i} + 2\mathbb{E}S_i(\xi_{i+1} + \cdots + \xi_n)\mathbf{1}_{A_i} + \mathbb{E}(\xi_{i+1} + \cdots + \xi_n)^2 \mathbf{1}_{A_i}\end{aligned}$$

where by independence,

$$\mathbb{E}S_i(\xi_1 + \cdots + \xi_n)\mathbf{1}_{A_i} = \mathbb{E}S_i \mathbf{1}_{A_i} \mathbb{E}(\xi_1 + \cdots + \xi_n) = 0,$$

and so,

$$\mathbb{E}S_n^2 \mathbf{1}_{A_i} = \mathbb{E}S_i^2 \mathbf{1}_{A_i} + +\mathbb{E}(\xi_{i+1} + \cdots + \xi_n)^2 \mathbf{1}_{A_i} \ge \mathbb{E}S_i^2 \mathbf{1}_{A_i}.$$

Hence, combining the inequalities, we obtain $\mathbb{E}S_n^2 \ge \mathbb{E}S_i^2 \mathbf{1}_A \ge \epsilon^2 \mathbb{P}(A)$ which is exactly the required inequality after dividing both sides by ϵ^2. □

Theorem 28 (Two-series Theorem)**.** Let (ξ_n) be a sequence of independent random variable. If $\sum_{n=1}^\infty \mathbb{E}\xi_n$ and $\sum_{n=1}^\infty V_{\xi_n}$ both converge, then the series $\sum_{n=1}^\infty \xi_n$ converges almost everywhere.

Proof. In the case that $\mathbb{E}\xi_i = 0$, we will show (S_n) is Cauchy almost everywhere, i.e.

$$\lim_{k\to\infty} \sup_{m,n\ge k} |S_m - S_n| = 0$$

almost everywhere. However, defining $T_k := \sup_{m,n\ge k} |S_m - S_n|$, we have a sequence of decreasing, non-negative random variables, and hence, it suffice to show T_n converges to 0 in probability. We note that

$$0 \le \sup_{n,m\ge k} |S_n - S_k| \le \sup_{n,m\ge k} (|S_n - S_k| + |S_k - S_m|) = 2\sup_{n\ge k} |S_n - S_k| =: 2\sigma_k.$$

Hence, it suffice to show $\sigma_k \to 0$ in probability. Fix $\epsilon > 0$, then, by Kolmogorov's inequality,

$$\mathbb{P}(\sigma_k \ge \epsilon) = \lim_{m\to\infty} \mathbb{P}(\max_{k\le n\le m} |S_n - S_k| \ge \epsilon) \le \frac{1}{\epsilon^2} \lim_{m\to\infty} \sum_{n=k+1}^m V_{\xi_n} = \frac{1}{\epsilon^2} \sum_{n=k+1}^m V_{\xi_n}.$$

Now, as $\sum V_{\xi_n}$ converges, the right hand side tends to 0 as $k \to \infty$ and thus, $\sigma_k \to 0$ in probability.

In the case that $\mathbb{E}\xi_j \ne 0$, one may simply consider

$$\sum_{n=1}^\infty \xi_n = \sum_{n=1}^\infty (\xi_n - \mathbb{E}\xi_n) + \sum_{n=1}^\infty \mathbb{E}\xi_n$$

where $\sum_{n=1}^\infty \mathbb{E}\xi_n$ converges by assumption and $\sum_{n=1}^\infty (\xi_n - \mathbb{E}\xi_n)$ converges almost everywhere by the first case. □

Lemma 5.2 (Toeplitz)**.** Let (a_n) be a non-negative real sequence and $b_n := \sum_{k=1}^{n} a_k$ such that $b_n \to \infty$. Furthermore, let $x_n \to x$. Then,

$$\lim_{n\to\infty} \frac{1}{b_n} \sum_{k=1}^{n} a_k x_k = x.$$

Proof. Elementary analysis. □

Lemma 5.3 (Kronecker)**.** Let (a_n) be a non-negative real sequence and $b_n := \sum_{k=1}^{n} a_k$ such that $b_n \to \infty$. Furthermore, let (x_n) be a sequence such that $\sum_{k=1}^{\infty} x_n$ converge. Then,

$$\lim_{n\to\infty} \frac{1}{b_n} \sum_{k=1}^{\infty} b_k x_k = 0.$$

Proof. Follows by summation by parts. □

Theorem 29. Let (ξ_n) be a sequence of independent random variables with finite variance and let (b_n) be a sequence of non-negative real numbers such that $b_n \to \infty$, then, if $\sum_{k=1}^{\infty} \frac{V\xi_k}{b_k^2} < \infty$, we have

$$\frac{S_n - \mathbb{E}S_n}{b_n} \to 0$$

almost everywhere.

We note that if we simply choose $b_n = n$, we obtain a condition for strong LLN.

Proof. We observe that

$$\frac{S_n - \mathbb{E}S_n}{b_n} = \frac{1}{b_n} \sum_{k=1}^{n} b_k \frac{\xi_k - \mathbb{E}\xi_k}{b_k}$$

and $V_{\sum_{k=1}^{n} \frac{\xi_k - \mathbb{E}\xi_k}{b_k}} = \sum_{k=1}^{n} \frac{V\xi_k}{b_k^2}$ which converges by assumption. Then, by the two-series theorem, $\sum_{k=1}^{n} \frac{\xi_k - \mathbb{E}\xi_k}{b_k}$ converges almost surely. Thus, applying Kronecker's lemma at all points of convergence, we obtain the required result. □

Lemma 5.4. Let $\xi \geq 0$ be a non-negative random variable. Then,

$$\sum_{n=1}^{\infty} \mathbb{P}(\xi \geq n) \leq \mathbb{E}\xi \leq 1 + \sum_{n=1}^{\infty} \mathbb{P}(\xi \geq n).$$

Proof. This follows by

$$\begin{aligned}
\sum_{n=1}^{\infty} \mathbb{P}(\xi \geq n) &= \sum_{n=1}^{\infty} \sum_{k\geq n} \mathbb{P}(k \leq \xi \leq k+1) = \sum_{k=1}^{\infty} k\mathbb{P}(k \leq \xi < k+1) \\
&= \sum_{k=1}^{\infty} \mathbb{E}(k\mathbf{1}_{[k,k+1)}(\xi)) \leq \sum_{k=1}^{\infty} \mathbb{E}(\xi\mathbf{1}_{[k,k+1)}(\xi)) = \mathbb{E}\xi \\
&\leq \sum_{k=1}^{\infty} \mathbb{E}((k+1)\mathbf{1}_{[k,k+1)}(\xi)) = 1 + \sum_{n=1}^{\infty} \mathbb{P}(\xi \geq n).
\end{aligned}$$

□

Theorem 30 (Kolmogorov's Strong Law of Large Numbers)**.** Let $(\xi_n)_{n=1}^{\infty}$ be a sequence of i.i.d. *integrable* random variables with finite expectation so $\mathbb{E}\xi_1 = m$, then

$$\frac{1}{n}S_n \to m$$

almost everywhere.

We note that $S_n/n \to m$ almost everywhere is equivalent to $\sum_{k=1}^{n}(\xi_k - \mathbb{E}\xi_k) = o(n)$ as $n \to \infty$ almost everywhere.

Proof. WLOG. we may assume $\mathbb{E}\xi_1 = 0$, and it suffices to show $S_n/n \to 0$ almost everywhere. As $\mathbb{E}|\xi_1| < \infty$, by the above lemma we have $\sum_{n=1}^{\infty}\mathbb{P}(|\xi_n| \geq n) \leq \mathbb{E}|\xi_1| < \infty$. Hence, by Borel-Cantelli, $\mathbb{P}(|\xi_n| \geq n \text{ i.o.}) = 0$ and so $\mathbb{P}(|\xi_n| < n \text{ e.v.}) = 1$.

Now, defining

$$\tilde{\xi}_n = \begin{cases} \xi_n, & |\xi_n| < n, \\ 0, & |\xi_n| \geq n, \end{cases}$$

the above implies $S_n/n \to 0$ almost everywhere if and only if $\sum_{k=1}^{n}\tilde{\xi}_k/n \to 0$ almost everywhere. Noting

$$\mathbb{E}\tilde{\xi}_n = \mathbb{E}\xi_n \mathbf{1}_{|\xi_n|<n} = \mathbb{E}\xi_1 \mathbf{1}_{|\xi_1|<n}$$

tends to $\mathbb{E}\xi_1$ as $n \to \infty$ almost everywhere. Now, by Toeplitz lemma with $x_n := \mathbb{E}\tilde{\xi}_n$, we have $\frac{1}{n}\sum_{k=1}^{n}\mathbb{E}\tilde{\xi}_k \to 0$ almost everywhere as $n \to \infty$. Thus, $S_n/n \to 0$ if and only if $\frac{1}{n}(\sum_k(\tilde{\xi}_k - \mathbb{E}\tilde{\xi}_k)) \to 0$ almost everywhere. Then, applying Kronecker's lemma with $b_n = n, x_n = (\tilde{\xi}_n - \mathbb{E}\tilde{\xi}_n)/n$, we obtain $\frac{1}{n}(\sum_k(\tilde{\xi}_k - \mathbb{E}\tilde{\xi}_k)) \to 0$ if $\sum_{n=1}^{\infty}\frac{\tilde{\xi}_n - \mathbb{E}\tilde{\xi}_n}{n}$ converges. Then, applying the two-series theorem, it suffices to show

$$\sum_{n=1}^{\infty} V\left(\frac{\tilde{\xi}_n - \mathbb{E}\tilde{\xi}_n}{n}\right) = \sum_{n=1}^{\infty}\frac{V(\tilde{\xi}_n - \mathbb{E}\tilde{\xi}_n)}{n^2} < \infty.$$

But this follows by considering

$$\begin{aligned}
\sum_{n=1}^{\infty}\frac{V(\tilde{\xi}_n - \mathbb{E}\tilde{\xi}_n)}{n^2} &\leq \sum_{n=1}^{\infty}\frac{\mathbb{E}\tilde{\xi}_n^2}{n^2} = \sum_{n=1}^{\infty}\frac{1}{n^2}\mathbb{E}(\xi_1^2\mathbf{1}_{|\xi_1|<n}) \\
&= \sum_{n=1}^{\infty}\frac{1}{n^2}\sum_{k=1}^{n}\mathbb{E}(\xi_1^2\mathbf{1}_{k-1\leq|\xi_1|<k}) \\
&= \sum_{k=1}^{\infty}\mathbb{E}(\xi_1^2\mathbf{1}_{k-1\leq|\xi_1|<k})\sum_{n\geq k}\frac{1}{n^2} \\
&\leq 2\sum_{k=1}^{\infty}\mathbb{E}(|\xi_1|\mathbf{1}_{k-1\leq|\xi_1|<k}) = 2\mathbb{E}|\xi_1| < \infty.
\end{aligned}$$

□

5.3 Law of Iterated Logarithm

We have seen for Bernoulli random variables with $\mathbb{E}\xi_n = 0$, that

$$\frac{S_n}{\sqrt{n}\log n} \to 0$$

almost everywhere. Furthermore, by the central limit theorem, $S_n/\sqrt{n}$ converges in distribution to the standard Gaussian implying $S_n/\sqrt{n}$ does not converge to 0.

Definition 5.2. A function $\phi^*(n)$ is called upper for S_n if there exists some n_0 such that $S_n \leq \phi^*(n)$ almost everywhere for all $n \geq n_0$.

Similarly, a function $\phi_*(n)$ is called lower for S_n if $S_n > \phi_*(n)$ for infinitely many n almost everywhere.

For some ϕ, consider the set

$$\begin{aligned}\left\{\limsup_{n\to\infty} \frac{S_n}{\phi(n)} \leq 1\right\} &= \left\{\lim_{n\to\infty} \sup_{m\geq n} \frac{S_m}{\phi(m)} \leq 1\right\} \\ &= \left\{\forall \epsilon > 0, \exists N, \forall n \geq N, \sup_{m\geq n} \frac{S_m}{\phi(m)} \leq 1+\epsilon\right\} \\ &= \{\forall \epsilon > 0, \exists N, \forall m \geq n_1, S_m \leq (1+\epsilon)\phi(m)\}.\end{aligned}$$

Hence, if $\left\{\limsup_{n\to\infty} \frac{S_n}{\phi(n)} \leq 1\right\}$ has probability 1, $(1+\epsilon)\phi$ is an upper of S_n for all ϵ. Similarly, considering the set

$$\begin{aligned}\left\{\limsup_{n\to\infty} \frac{S_n}{\phi(n)} \geq 1\right\} &= \left\{\forall \epsilon > 0, \exists N, \forall n \geq N, \sup_{m\geq n} \frac{S_m}{\phi(m)} \geq 1-\epsilon\right\} \\ &= \{\forall \epsilon > 0, S_m \geq (1-\epsilon)\phi(m) \text{ for infinitely many } m\}.\end{aligned}$$

thus, $\left\{\limsup_{n\to\infty} \frac{S_n}{\phi(n)} \geq 1\right\}$ has probability 1 implies $(1-\epsilon)\phi$ is a lower of S_n.

With this in mind, we have the following theorem.

Theorem 31 (Law of Iterated Logarithm)**.** Let (ξ_n) be i.i.d. random variables with $\mathbb{E}\xi_1 = 0$ and $V_{\xi_1} = \sigma^2 > 0$. Then,

$$\mathbb{P}\left(\limsup_{n\to\infty} \frac{S_n}{\psi(n)} = 1\right) = 1$$

where $\psi(n) = \sqrt{2\sigma^2 n \log\log n}$.

As we have seen from the strong law of large numbers, $\sum_{i=1}^n \xi_i - \mathbb{E}\xi_k = o(n)$ almost everywhere. Now, with the law of iterated logarithm, we also have $\sum_{i=1}^n \xi_i - \mathbb{E}\xi_k = O(\psi(n))$ almost everywhere.

5.4 Kolmogorov's Zero-One Law

Let (ξ_n) be a sequence of random variables on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Denoting $\mathcal{F}_n^{n+k}$ the sub-σ-algebra generated by $\xi_n, \cdots, \xi_{n+k}$. Then, denoting

$$\mathcal{F}_n^{\infty} := \sigma(\xi_n, \cdots) = \sigma\left(\bigcup_{k=1}^{\infty} \mathcal{F}_n^{n+k}\right)$$

we have the following definition.

Definition 5.3 (Tail σ-algebra)**.** The σ-algebra

$$\mathcal{T} := \bigcap_{n=1}^{\infty} \mathcal{F}_n^{\infty}$$

is called the tail σ-algebra. Events of $\mathcal{T}$ are called tail events.

Proposition 5.1. Let $B \in \mathcal{B}(\mathbb{R})$. Then,

$$\{\xi_n \in B \text{ i.o.}\} = \bigcap_{n=1}^{\infty} \bigcup_{k \geq n} \{\xi_n \in B\} \in \mathcal{T}.$$

Proof. Clear. □

Proposition 5.2. The set

$$\left\{ \sum_{k=1}^{\infty} \xi_k \text{ converges} \right\} \in \mathcal{T}.$$

Proof. Clear. □

Lemma 5.5. If $\mathcal{A}, \mathcal{B} \subseteq \mathcal{F}$ are independent (i.e.. $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$ for all $A \in \mathcal{A}, B \in \mathcal{B}$), then $\sigma(\mathcal{A}), \sigma(\mathcal{B})$ are also independent.

Proof. Let $A \in \mathcal{A}$ and define the measures

$$\mathbb{P}_A^1(B) := \mathbb{P}(A \cap B), \mathbb{P}_A^2 := \mathbb{P}(A)\mathbb{P}(B).$$

As $\mathcal{A}, \mathcal{B}$ are independent, they coincide on $\mathcal{B}$. Then, by Caratheodory extension, they also coincide on $\sigma(\mathcal{B})$. Now, taking $B \in \sigma(\mathcal{B})$, applying the same method, we obtain the required result. □

Theorem 32 (Kolmogorov's Zero-One Law)**.** Let (ξ_n) be a sequence of independent random variables. Then, any tail events of (ξ_n) has either probability 0 or 1.

Proof. Note that $\mathcal{F}_1^n$ is independent with $\mathcal{F}_{n+1}^{n+k}$ for all k. Thus, $\mathcal{F}_1^n$ is independent with $\bigcup_{k=1}^{\infty} \mathcal{F}_{n+1}^{n+k}$. Then, by the above lemma, $\mathcal{F}_1^n$ is independent with $\mathcal{F}_{n+1}^{\infty}$ which contains the tail σ-algebra. Hence, $\mathcal{F}_1^n$ is independent with $\mathcal{T}$ and so, $\bigcup_{n=2}^{\infty} \mathcal{F}_1^n$ is independent with $\mathcal{T}$. Hence, as $\mathcal{F}_1^{\infty}$ also contains $\mathcal{T}$, we have $\mathcal{T}$ is independent with itself. Hence, for all $A \in \mathcal{T}$,

$$\mathbb{P}(A) = \mathbb{P}(A \cap A) = \mathbb{P}(A)^2$$

implying $\mathbb{P}(A) = 0$ or 1 as required. □

Corollary 32.1. If (ξ_n) is a sequence of independent random variables, then the event $\{\sum_{n=1}^{\infty} \xi_n \text{ converges}\}$ has either probability 0 or 1. Thus, $\sum_{n=1}^{\infty} \xi_n$ converges almost everywhere or diverges almost everywhere.

Corollary 32.2. Let (A_n) be a sequence of independent sets. Then $\mathbb{P}(A_n \text{ i.o.})$ is either 0 or 1. (Though we already know this from the second Borel-Cantelli.)

6 Conditional Expectation

We recall from second year measure theory the Radon-Nikodym theorem which states that, if μ, ν are σ-finite measures on the measure space $(\Omega, \mathcal{F})$ such that $\mu \ll \nu$, then there exists an essentially unique $\mathcal{F}$-measurable function $\frac{\mathrm{d}\mu}{\mathrm{d}\nu}$ such that for all $A \in \mathcal{F}$,

$$\mu(A) = \int_A \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \mathrm{d}\nu.$$

$\frac{\mathrm{d}\mu}{\mathrm{d}\nu}$ is called the Radon-Nikodym derivative of μ with respect to ν.

Definition 6.1 (Conditional Expectation)**.** Let $\xi \geq 0$ be a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ and let $\mathcal{G} \subseteq \mathcal{F}$ be a sub-σ-algebra. Then, defining $\mathbb{Q}(A) \coloneqq \int_A \xi \mathrm{d}\mathbb{P}$ for all $A \in \mathcal{G}$, it is clear that $\mathbb{Q} \ll \mathbb{P}|_{\mathcal{G}}$ on $(\Omega, \mathcal{G})$. Hence, by the Radon-Nikodym theorem, there exists an essentially unique $\mathcal{G}$-measurable function $\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}|_{\mathcal{G}}}$ such that

$$\mathbb{Q} = \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}|_{\mathcal{G}}} \mathbb{P}|_{\mathcal{G}}$$

We call this function $\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}|_{\mathcal{G}}}$ the conditional expectation of ξ with respect to $\mathcal{G}$ and denote it by $\mathbb{E}(\xi \mid \mathcal{G})$.

Proposition 6.1. By definition, $\mathbb{E}(\xi \mid \mathcal{G})$ is the essentially unique $\mathcal{G}$-measurable function such that for all $A \in \mathcal{G}$,

$$\int_A \mathbb{E}(\xi \mid \mathcal{G}) \mathrm{d}\mathbb{P} = \int_A \xi \mathrm{d}\mathbb{P}.$$

This definition can be easily extended to arbitrary random variables by either considering the signed measure version of Radon-Nikodym, or simply by defining

$$\mathbb{E}(\xi \mid \mathcal{G}) \coloneqq \mathbb{E}(\xi^+ \mid \mathcal{G}) - \mathbb{E}(\xi^- \mid \mathcal{G})$$

if $\min(\mathbb{E}(\xi^+ \mid \mathcal{G}), \mathbb{E}(\xi^- \mid \mathcal{G})) < \infty$ almost everywhere. The two methods of extension coincide.

Definition 6.2 (Conditional Probability)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. The conditional probability of an event $\mathcal{B} \in \mathcal{F}$ with respect to the sub-σ-algebra $\mathcal{G} \subseteq \mathcal{F}$ is the random variable

$$\mathbb{P}(B \mid \mathcal{G}) \coloneqq \mathbb{E}(\mathbf{1}_B \mid \mathcal{G}).$$

It is easy to see that $\int_A \mathbb{P}(B \mid \mathcal{G}) \mathrm{d}\mathbb{P} = \mathbb{P}(A \cap B)$.

Suppose $\mathcal{G} = \sigma(\{D_1, D_2, \cdots\}) \subseteq \mathcal{F}$ be a σ-algebra generated by a partition of Ω. Then, all $\mathcal{G}$-measurable functions have the form $f(\omega) \coloneqq \sum_{j=1}^{\infty} c_i \mathbb{1}_{D_j}(\omega)$. So, given a random variable ξ,

$$\mathbb{E}(\xi \mathbf{1}_{D_i}) = \int_{D_i} \xi \mathrm{d}\mathbb{P} = \int_{D_i} \mathbb{E}(\xi \mid \mathcal{G}) \mathrm{d}\mathbb{P} = \mathbb{E}(\xi \mid D_j) \mathbb{P}(D_j),$$

where $\mathbb{E}(\xi \mid D_j)$ is some constant. Hence,

$$\mathbb{E}(\xi \mid D_i) = \mathbb{E}(\xi \mathbf{1}_{D_i}) / \mathbb{P}(D_i)$$

if $\mathbb{P}(D_i) \neq 0$. Thus, in the case $\xi = \mathbf{1}_B$ for some $B \in \mathcal{F}$,

$$\mathbb{P}(B \mid D_i) = \mathbb{E}(\mathbf{1}_B \mid D_i) = \mathbb{E}(\mathbf{1}_B \mathbf{1}_{D_i})/\mathbb{P}(D_i) = \mathbb{P}(B \cap D_i)/\mathbb{P}(D_i)$$

which coincides with the classical definition of conditional probability.

Theorem 33. Let ξ, η be a random variable on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then, given $\mathcal{G} \subseteq \mathcal{F}$ is a sub-σ-algebra,

- $|\mathbb{E}(\xi \mid \mathcal{G})| \leq \mathbb{E}(|\xi| \mid \mathcal{G})$ almost everywhere;
- $\mathbb{E}(a\xi + b\eta \mid \mathcal{G}) = a\mathbb{E}(\xi \mid \mathcal{G}) + b\mathbb{E}(\eta \mid \mathcal{G})$ for $a, b \in \mathbb{R}$ almost everywhere;
- $\mathbb{E}(\xi \mid \mathcal{G}) \leq \mathbb{E}(\eta \mid \mathcal{G})$ almost everywhere if $\xi \leq \eta$ almost everywhere;
- $\mathbb{E}(\xi \mid \{\emptyset, \Omega\}) = \mathbb{E}\xi$ almost everywhere;
- $\mathbb{E}(\xi \mid \mathcal{F}) = \xi$ almost everywhere;
- if ξ is $\mathcal{G}$-measurable, then $\mathbb{E}(\xi \mid \mathcal{G}) = \xi$ almost everywhere;
- if $\mathcal{G}_1 \subseteq \mathcal{G}_2 \subseteq \mathcal{F}$ are sub-σ-algebras, $\mathbb{E}(\mathbb{E}(\xi \mid \mathcal{G}_2) \mid \mathcal{G}_1) = \mathbb{E}(\xi \mid \mathcal{G}_1)$ and $\mathbb{E}(\mathbb{E}(\xi \mid \mathcal{G}_1) \mid \mathcal{G}_2) = \mathbb{E}(\xi \mid \mathcal{G}_1)$ almost everywhere;
- $\mathbb{E}(\mathbb{E}(\xi \mid \mathcal{G})) = \mathbb{E}(\xi)$;
- if ξ is independent of $\mathcal{G}$ (i.e. $\sigma(\xi)$ and $\mathcal{G}$ are independent), then $\mathbb{E}(\xi \mid \mathcal{G}) = \mathbb{E}\xi$

Proof. Straight forward by the essential uniqueness of the Radon-Nikodym derivative. □

Theorem 34 (Conditional Dominated Convergence)**.** Let $\xi_1, \xi_2, \cdots$ be a sequence of random variables such that $x_n \to \xi$ almost everywhere for some random variable ξ. Then, if $|\xi_n| \leq \eta$ for some random variable η such that $\mathbb{E}\eta < \infty$,

$$\mathbb{E}(\xi_n \mid \mathcal{G}) \to \mathbb{E}(\xi \mid \mathcal{G})$$

almost everywhere for some sub-σ-algebra $\mathcal{G} \subseteq \mathcal{F}$.

Proof. Denote $\zeta_n := \sup_{m \geq n} |\xi_m - \xi|$, then $\zeta_n \to 0$ almost everywhere. then

$$|\mathbb{E}(\xi_n \mid \mathcal{G}) - \mathbb{E}(\xi \mid \mathcal{G})| = |\mathbb{E}(\xi_n - \xi \mid \mathcal{G})| \leq \mathbb{E}(|\xi_n - \xi| \mid \mathcal{G}) \leq \mathbb{E}(\zeta_n \mid \mathcal{G}).$$

Thus, it remains to show $\mathbb{E}(\zeta_n \mid \mathcal{G}) \to 0$ almost everywhere. As ζ_n is monotone decreasing, so is $\mathbb{E}(\zeta_n \mid \mathcal{G})$, then, as $\mathbb{E}(\zeta_n \mid \mathcal{G})$ is bounded below by 0, $\lim_{n\to\infty} \mathbb{E}(\zeta_n \mid \mathcal{G}) =: \phi$ exists almost everywhere. Hence, considering

$$0 \leq \int \phi \mathrm{d}\mathbb{P} \leq \int \mathbb{E}(\zeta \mid \mathcal{G}) \mathrm{d}\mathbb{P} = \int \zeta \mathrm{d}\mathbb{P} \to 0$$

where the limit follows by dominated convergence as $|\zeta_n| \leq 2\eta$, we have $\int \phi \mathrm{d}\mathbb{P} = 0$ and hence, as $\phi \geq 0$ a.e. $\phi = 0$ almost everywhere as required. □

Theorem 35 (Pull-Out Property)**.** Let ξ, η be random variables such that η is $\mathcal{G}$-measurable. Then, if $\mathbb{E}|\xi| < \infty$ and $\mathbb{E}|\xi\eta| < \infty$, we have

$$\mathbb{E}(\xi\eta \mid \mathcal{G}) = \eta\mathbb{E}(\xi \mid \mathcal{G})$$

almost everywhere.

Proof. If $\eta = \mathbf{1}_B$ for some $B \in \mathcal{G}$, then for all $C \in \mathcal{G}$,

$$\int_C \mathbb{E}(\xi\eta \mid \mathcal{G})\mathrm{d}\mathbb{P} = \int_{C\cap B} \xi \mathrm{d}\mathbb{P} = \int_{C\cap B} \mathbb{E}(\xi \mid \mathcal{G})\mathrm{d}\mathbb{P} = \int_C \eta\mathbb{E}(\xi \mid \mathcal{G})\mathrm{d}\mathbb{P}$$

implying $\mathbb{E}(\xi\eta \mid \mathcal{G}) = \eta\mathbb{E}(\xi \mid \mathcal{G})$ as required.

Now, as conditional expectation is linear, if $\eta = \sum_{i=1}^n a_i \mathbf{1}_{B_i}$ is a simple function with respect to $\mathcal{G}$,

$$\mathbb{E}(\xi\eta \mid \mathcal{G}) = \mathbb{E}\left(\xi \sum_i a_i \mathbf{1}_{B_i} \mid \mathcal{G}\right) = \sum_i a_i \mathbb{E}(\xi \mathbf{1}_{B_i} \mid \mathcal{G}) = \sum_i a_i \mathbf{1}_{B_i} \mathbb{E}(\xi \mid \mathcal{G}) ? \eta\mathbb{E}(\xi \mid \mathcal{G})$$

as required.

Finally, for general $\mathcal{G}$-measurable function η, let $\eta_n \uparrow \eta$ be a sequence of monotonically increasing $\mathcal{G}$-simple functions. So, for all $C \in \mathcal{G}$, by monotone convergence and the monotonicity of the conditional expectation

$$\int_C \mathbb{E}(\xi\eta \mid \mathcal{G}) = \lim_{n\to\infty} \int_C \mathbb{E}(\xi\eta_n \mid \mathcal{G}) = \lim_{n\to\infty} \int_C \eta_n \mathbb{E}(\xi \mid \mathcal{G}) = \int_C \eta\mathbb{E}(\xi \mid \mathcal{G})$$

implying $\mathbb{E}(\xi\eta \mid \mathcal{G}) = \eta\mathbb{E}(\xi \mid \mathcal{G})$ almost everywhere as required. □

Corollary 35.1. Under the assumption of the pull-out property,

$$\mathbb{E}(\eta\mathbb{E}(\xi \mid \mathcal{G})) = \mathbb{E}(\xi\eta).$$

Definition 6.3. The conditional expectation of a random variable ξ with respect to the random variable η is the random variable

$$\mathbb{E}(\xi \mid \eta) \coloneqq \mathbb{E}(\xi \mid \sigma(\eta)).$$

Proposition 6.2 (Factorisation Lemma)**.** Let ξ, η are random variables, then ξ is $\sigma(\eta)$-measurable if and only if there exists some Borel-measurable function $f : \mathbb{R} \to \mathbb{R}$ such that $\xi = f \circ \eta$. Furthermore, in this case,

$$\mathbb{E}(\xi \mid \eta) = \xi = f \circ \eta$$

almost everywhere.

Proof. As usual, first show for $\xi = \sum_{i=1}^n a_i \mathbf{1}_{B_i}$ is simple which then, the general case follows by monotone approximation (exercise: fill in the details). □

As a remark, we note that if $\mathbb{E}\xi^2 < \infty$, then

$$\min_f \mathbb{E}(\xi - f(\eta))^2 = \mathbb{E}(\xi - \mathbb{E}(\xi \mid \eta))^2$$

where the minimum is taken over all Borel-measurable functions $f : \mathbb{R} \to \mathbb{R}$ such that $\mathbb{E}f^2(\eta) < \infty$.