# 1 The Theory of Stochastic Convergence

This section serves as an introduction to topics in stochastic convergence. We examine forms of asymptotic behavior of sequences of random elements to develop notions useful for understanding (approximate) properties of statistical inference methods in the following sections. Readers interested in a deeper understanding can refer to a wealth of relevant literature. Indicative references include [1], [3], [4], [5], [7], [13], [14].

Let $X_n$, $X$ be random elements with values in the same codomain, the metric space $S$-equipped with the Borel algebra induced by its metric $d$, $\forall n \in \mathbb{N}$. The $X_n$ form the sequence $(X_n)_{n\in\mathbb{N}}$, while $X$ represents a limit under one of the convergence concepts discussed below, as $n \to \infty$. In the subsequent sections, the involved random elements are assumed to share the same domain, specifically the sample space $\Omega$ of the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. This restriction is not necessary when discussing weak convergence, which refers to the convergence of the distributions induced by the random elements on their common codomain.

In what follows the term *iff* abbreviates the expression 'if and only if'; also the dependence of the random elements involved on (the typically) latent $\omega$, is suppressed when convenient.

# 2 Almost Sure and Convergence in Probability

Conceptually, the simplest (but extremely demanding in the conditions required) notion is pointwise convergence across the entire domain. Thus, we say that $X_n$ converges *surely* (or more appropriately, pointwisely w.r.t.

$\omega$) to $X$, denoted $X_n \to X$, if $X_n(\omega) \to X(\omega)$ for all $\omega \in \Omega$, that is, when for the set for which convergence holds, i.e.,

$$V = \{\omega \in \Omega : \forall\varepsilon > 0, \exists N(\varepsilon, \omega), \forall n \geq N, d(X_n(\omega), X(\omega)) < \varepsilon, \},$$

we have that this coincides with the whole domain of the random elements involved, i.e., $V = \Omega$.

Relaxing the above, we may only require that the set $V$ where convergence occurs has full $\mathbb{P}$-measure. The issue of $V$'s measurability is addressed through the measurability of the involved random elements, the corresponding measurability of the metric $d$ as a continuous function, and the properties of the associated collections of sets at which probabilities can be attributed to. This leads to the concept of almost sure convergence:

**Definition 1.** We say that $X_n$ converges to $X$ with ($\mathbb{P}$-) probability one, or ($\mathbb{P}$-) almost surely, denoted $X_n \overset{\text{a.s.}}{\to} X$, iff $\mathbb{P}(V) = 1$, where $V$ is as defined above.

Schematically, we have $X_n \overset{\text{a.s.}}{\to} X$, if and only if the event that $X_n$ converges pointwise to $X$ is of full probability, i.e.

$$\mathbb{P}\left(\left\{\omega \in \Omega : \lim_{n\to\infty} d(X_n(\omega), X(\omega)) = 0\right\}\right) = 1.$$

The following result equivalently represents this form of convergence as asymptotic uniform convergence (with respect to $n$) of a sequence of probabilities, facilitating the construction of a weaker mode of convergence.

**Theorem 1.** *We have $X_n \overset{a.s.}{\to} X$ if and only if*

$$\lim_{m\to\infty} \mathbb{P}\left[\left\{\omega \in \Omega : \sup_{n\geq m} d\left(X_n\left(\omega\right), X\left(\omega\right)\right) \leq \varepsilon\right\}\right] = 1, \forall \varepsilon > 0.$$

*Proof.* First, note that

$$\lim_{m\to\infty} \mathbb{P}\left[\left\{\omega \in \Omega : \sup_{n\geq m} d\left(X_n\left(\omega\right), X\left(\omega\right)\right) \leq \varepsilon\right\}\right] = 1, \forall \varepsilon > 0$$

is equivalent by definition to

$$\mathbb{P}(A(\varepsilon)) = 1, \forall \varepsilon > 0,$$

where for arbitrary $\varepsilon > 0$:

$$A(\varepsilon) := \cup_{m=1}^{\infty} A_m(\varepsilon),$$

and

$$A_m(\varepsilon) := \cap_{n\geq m} \left\{\omega \in \Omega : \sup_{n} d\left(x_n\left(\omega\right), x\left(\omega\right)\right) \leq \varepsilon\right\},$$

due to the continuity of $\mathbb{P}$. Assume that $\mathbb{P}(V) = 1$ hence almost sure convergence holds. It then suffices to prove $V \subseteq A(\varepsilon), \forall \varepsilon > 0$ due to the monotonicity of $\mathbb{P}$. Indeed, for $\varepsilon > 0$, if $\omega \in V$, then $\omega \in A_{N(\varepsilon,\omega)}(\varepsilon) \Rightarrow \omega \in A(\varepsilon)$, so $V \subseteq A(\varepsilon)$, and thereby $\mathbb{P}(A(\varepsilon)) = 1$ for any $\varepsilon > 0$.

Conversely, assume $\mathbb{P}(A(\varepsilon)) = 1, \forall \varepsilon > 0$. Then for $\varepsilon = \frac{1}{k}, k = 1, 2, \cdots$, define

$$A^{\star} := \cap_{k=1}^{\infty} A\left(\frac{1}{k}\right).$$

We have $\mathbb{P}(A^{\star}) = 1$ due to the De Morgan laws and countable additivity. To show $A^\star \subseteq V$, observe that if $\omega \in A^\star$, then $\omega \in V$ due to the density of the rationals in the set of real numbers. The result follows. □

The theorem informs us that almost sure convergence is a form of uniform convergence over the "tail" of the sequence of random elements.[1] Relaxing the uniformity, which also implies strong requirements, leads to the next (and weaker) form of stochastic convergence, namely convergence in probability:

**Definition 2.** We say that $X_n$ converges to $X$ in ($\mathbb{P}$-) probability (in probability), denoted $X_n \overset{\mathrm{p}}{\rightarrow} X$ (or equivalently $p\lim_{n\rightarrow\infty} X_n = X$), if and only if

$$\forall\varepsilon > 0,\ \lim_{n\rightarrow\infty} \mathbb{P}\left(\{\omega \in \Omega : d\left(X_n\left(\omega\right), X\left(\omega\right)\right) \leq \varepsilon\}\right) = 1.$$

Dually, and due to the law of complementary probability, this is also representable as $\forall\varepsilon > 0,\ \lim_{n\rightarrow\infty} \mathbb{P}\left(\{\omega \in \Omega : d\left(X_n\left(\omega\right), X\left(\omega\right)\right) > \varepsilon\}\right) = 0$. The definition fails, and thus $X_n \overset{\mathrm{p}}{\nrightarrow} X$, iff there exists some positive $\varepsilon, \delta$, such that $\mathbb{P}\left(\{\omega \in \Omega : d\left(X_n\left(\omega\right), X\left(\omega\right)\right) > \varepsilon\}\right) \geq \delta$ for an infinite set of $n$'s. The following lemma is directly derived from the definition and Theorem 1.

**Lemma 1.** $X_n \overset{a.s.}{\rightarrow} X \Rightarrow X_n \overset{p}{\rightarrow} X$.

The converse does not generally hold, as it is possible for there to exist $\varepsilon > 0$ and a non-$\mathbb{P}$-negligible subset of $\Omega$ supporting subsequences of $(d(X_n, X))_{n\in\mathbb{N}}$ that prevent the tail uniformity of Theorem 1, but still converge "individually".

The examples in the remaining paragraph present cases that utilize sequences of random variables. The last of them uses the notion of ergodicity

[1] Another characterization of almost sure convergence, represents it as uniform over subsets of $\Omega$; the related result is termed Egoroff's Theorem.

to formulate what is known as Birkhoff's Law of Large Numbers (Birkhoff's LLN), as well as Example 5.3 later.

**Example 1.** Let $(\Omega, \mathcal{F}, \mathbb{P}) = ([0,1], \mathcal{B}, \lambda)$, where $\lambda$ is the standard uniform probability distribution. Letting $S = \mathbb{N}$ equipped with the usual metric, define:

$$X_n(\omega) = \begin{cases} n, & \omega \in \left[0, \frac{1}{n}\right), \\ 0, & \omega \notin \left[0, \frac{1}{n}\right). \end{cases}$$

Clearly, $V' = \{\omega \in [0,1] : \lim_{n\to\infty} x_n(\omega) \neq 0\} = \{0\}$, and $\lambda(\{0\}) = 0$. Therefore, $X_n \overset{\text{a.s.}}{\to} 0$, and hence $X_n \overset{p}{\to} 0$.

> *Exersice*: If in the previous example the sequence was defined as
>
> $$X_n(\omega) = \begin{cases} n, & \omega \in \left[0, 1 - \frac{1}{n+1}\right), \\ 0, & \omega \notin \left[0, 1 - \frac{1}{n+1}\right), \end{cases}$$
>
> what would happen regarding either modes of convergence?

**Example 2.** Let $S = \mathbb{R}$ with the usual metric, and random variables $X_n \sim$ Ber($^1/_{n^\kappa}$), for some $\kappa \geq 1$. Notice that for arbitrary $\varepsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}(|X_n - 1| > \varepsilon) = \lim_{n\to\infty} \mathbb{P}(X_n = 0) = \lim_{n\to\infty} {}^1/_{n^\kappa} = 0,$$

so $X_n \overset{p}{\to} 1$. Furthermore:

$$\lim_{m\to\infty} \mathbb{P}(\sup_{n\geq m} |X_n - 1| > \varepsilon) = \lim_{m\to\infty} \mathbb{P}(\exists n \geq m, X_n = 0) = \lim_{m\to\infty} \mathbb{P}(\cup_{n\geq m}\{X_n = 0\})$$

$$\leq \lim_{m\to\infty} \sum_{n\geq m} \mathbb{P}(X_n = 0) = \lim_{m\to\infty} \sum_{n\geq m} {}^1/_{n^\kappa}.$$

The inequality in the display above, arises from countable sub-additivity. When $\kappa > 1$, $\lim_{m\to\infty}\sum_{n\geq m} {}^{1}\!/\!{}_{n^\kappa} = 0$, due to the convergence of the super-harmonic series $\sum_{n=0}^{\infty} {}^{1}\!/\!{}_{n^\kappa}$ to $\zeta(\kappa)$ the value of the Riemann zeta function at $\kappa$. Hence, due to complementarity, $\lim_{m\to\infty}\mathbb{P}(\sup_{n\geq m}|X_n-1|\leq\varepsilon)\geq 1$, and thereby the limit equals to one; hence $X_n \overset{\text{a.s.}}{\to} 1$.

When $\kappa \leq 1$ the upper bound in the above display is non informative. Suppose for simplicity that the random variables involved are independent. In this case notice that for any $\varepsilon < 1$:

$$\lim_{m\to\infty}\mathbb{P}(\sup_{n\geq m}|X_n-1|\leq\varepsilon) = \lim_{m\to\infty}\mathbb{P}(\forall n\geq m, X_n = 1) = \lim_{m\to\infty}\mathbb{P}(\cap_{n\geq m}\{X_n = 1\})$$

$$= \lim_{m\to\infty}\prod_{n\geq m}\mathbb{P}(\{X_n = 1\}) = \lim_{m\to\infty}\prod_{n\geq m}(1-\frac{1}{n^\kappa}) = \exp(\lim_{m\to\infty}\sum_{n\geq m}\ln(1-\frac{1}{n^\kappa})).$$

The third equality in the display above arises from independence. Using the approximation $-x\approx\ln(1-x)$ which holds for small $x$, we have that for large enough $m$, $\sum_{n\geq m}\ln(1-\frac{1}{n^\kappa}) = -\sum_{n\geq m}\frac{1}{n^\kappa}$, which as $m\to\infty$ diverges to $-\infty$, due to that the sub-harmonic $\sum_{n=1}^{\infty}\frac{1}{n^\kappa}$ and the harmonic $\sum_{n=1}^{\infty}\frac{1}{n}$ series diverge. Hence, the sequence $(X_n)$ does not converge almost surely to 1, nor does it have an almost sure limit in general (why?). This occurs because the probability of having a subsequence consisting entirely of ones for almost every $\omega$ is zero.

**Example 3.** Let $(X_t)_{t\in\mathbb{Z}}$ be a stationary time series with $\mathbb{E}(|X_0|) < +\infty$, and consider the sequence of arithmetic means $\left(\frac{1}{n}\sum_{t=1}^{n}X_t\right)$. The general version of Birkhoff's Law of Large Numbers implies that

$$\frac{1}{n}\sum_{t=1}^{n} X_t \overset{\text{a.s.}}{\to} \mathbb{E}(X_0/\mathcal{J}_x),$$

where the conditional expectation exists (why?). Hence, if the process is ergodic, the limit is $\mathbb{E}(X_0)$.[2] In this case, under stationarity and ergodicity, the limit is degenerate (why?). For the stationary AR(1) process discussed in the penultimate example of the notes regarding stochastic processes, assuming $\mathbb{E}(|\varepsilon_0|) < +\infty$, it is easily shown that $\mathbb{E}(X_0) = \frac{\mathbb{E}(\varepsilon_0)}{1-\beta}$, so

$$\frac{1}{n}\sum_{t=1}^{n} X_t \overset{\text{a.s.}}{\to} \frac{\mathbb{E}(\varepsilon_0)}{1-\beta}.$$

Similarly, for the GARCH(1,1) process discussed in the last example of the previous chapter, we have almost sure convergence to zero by construction.

> *Exersice*: Provide the details!

## 2.1 Continuous Mapping Theorem

Given that the above pertains to the convergence of sequences of functions, we expect that any continuous transformation of these sequences will transfer the limit. We can prove something even stronger, as we can allow the transformation to not be continuous everywhere:

[2] Such results, where the sequence of arithmetic means almost surely converges to the corresponding expectation, are called Strong Laws of Large Numbers (SLLN). When the convergence is in probability, the law is called Weak (WLLN).

**Theorem 2.** *Let $g : S \to S'$ be Borel measurable. If $\mathcal{C}_g$ is the set of continuity points of $g$ and $\mathbb{P}(X^{-1}(\mathcal{C}_g)) = 1$ for a random element $X$ taking values in $S$, then:*

- *If $X_n \overset{a.s.}{\to} X$, then $g(X_n) \overset{a.s.}{\to} g(X)$.*
- *If $X_n \overset{p}{\to} X$, then $g(X_n) \overset{p}{\to} g(X)$.*

*Proof.* Consider the set $V \cap X^{-1}(\mathcal{C}_g)$. Since $g$ is continuous on $\mathcal{C}_g$, we have that $\forall \omega \in V \cap X^{-1}(\mathcal{C}_g)$, $g(X_n) \to g(X)$. Furthermore,

$$\mathbb{P}(V \cap X^{-1}(\mathcal{C}_g)) = 1 - \mathbb{P}(V' \cap [X^{-1}(\mathcal{C}_g)]') \geq 1 - \mathbb{P}(V') + \mathbb{P}([X^{-1}(\mathcal{C}_g)]') = 1.$$

Thus, $\mathbb{P}(V \cap X^{-1}(\mathcal{C}_g)) = 1$, which proves almost sure convergence.

Now, due to continuity, $\forall \delta > 0, \exists \varepsilon > 0$ such that:

$$X^{-1}(\mathcal{C}_g) \cap \{\omega \in \Omega : d(X_n(\omega), X(\omega)) < \varepsilon\} \subseteq$$

$$\{\omega \in \Omega : d'(g \circ X_n(\omega), g \circ X(\omega)) < \delta\}.$$

In general, if $A, B \in \mathcal{F}$ with $\mathbb{P}(B) = 1$, then $\mathbb{P}(A \cap B) = 1 - \mathbb{P}(A' \cup B') \geq 1 - \mathbb{P}(A') + \mathbb{P}(B') = \mathbb{P}(A)$. Setting $A = \{\omega \in \Omega : d(X_n(\omega), X(\omega)) < \varepsilon\}$ and $B = X^{-1}(\mathcal{C}_g)$, we have:

$$\mathbb{P}(\{\omega \in \Omega : d(X_n(\omega), X(\omega)) < \varepsilon\}) \leq$$

$$\mathbb{P}(X^{-1}(\mathcal{C}_g) \cap \{\omega \in \Omega : d(X_n(\omega), X(\omega)) < \varepsilon\}) \leq$$

$$\mathbb{P}(\{\omega \in \Omega : d'(g \circ X_n(\omega), g \circ X(\omega)) < \delta\}).$$

Taking limits yields:

$$\lim_{n\to\infty} \mathbb{P}(\{\omega \in \Omega : d'(g \circ X_n(\omega), g \circ X(\omega)) < \delta\}) \geq$$

$$\lim_{n\to\infty} \mathbb{P}(\{\omega \in \Omega : d(X_n(\omega), X(\omega)) < \varepsilon\}) = 1,$$

which proves convergence in probability as $\delta$ is arbitrary. □

The continuous mapping result implies many general convergence results. For example, if $S, S'$ have vector space structures with operations continuous under the involved metrics, then, e.g., the respective limits of sums exist and are equal to the sums of the limits. Constructing new random elements via finite Cartesian products of the underlying spaces implies the convergence of vectors formed by these elements, etc.

A simple example involves the Riemann integral:

**Example 4.** If $\Omega = S = \mathcal{C}_{[0,1]}$-i.e. the space of continuous real functions on $[0,1]$, then since the Riemann integration operator $\mathcal{C}_{[0,1]} \ni f \mapsto \int_0^1 f(z)dz \in \mathbb{R}$ is linear and bounded, hence continuous, we have for the following stochastic Riemann integrals:

- If $X_n \overset{\text{a.s.}}{\to} X$, then $\int_0^1 X_n(\omega, z)dz \overset{\text{a.s.}}{\to} \int_0^1 X(\omega, z)dz$.
- If $X_n \overset{\text{p}}{\to} X$, then $\int_0^1 X_n(\omega, z)dz \overset{\text{p}}{\to} \int_0^1 X(\omega, z)dz$.

A more useful (for the course) example relates the Continuous Mapping Theorem, to usual algebraic transformations between finite collections of (sequences of) random variables:

**Example 5.** (Properties of Sums, Products, and Reciprocals of Random Variables Using the Continuous Mapping Theorem] Let $(X_n)_{n\geq 1}$ and $(Y_n)_{n\geq 1}$ be sequences of random variables. Assume that:

$$X_n \overset{\text{a.s.}}{\to} X \quad \text{and} \quad Y_n \overset{\text{a.s.}}{\to} Y \quad \text{(respectively in probability)}.$$

We examine the behavior of sums, products, and reciprocals of these sequences, leveraging the *Continuous Mapping Theorem (CMT)*. Transporting the analysis in $\mathbb{R}^2$, choosing appropriately a metric, and appropriately handling the topological issues and issues of measure, it is not difficult to show that the above convergence premises are equivalent to analogous convergences about random vectors, i.e.

$$(X_n, Y_n) \overset{\text{a.s.}}{\to} (X, Y) \text{(respectively in probability)}.$$

The CMT then allows analysis for the behavior of sums, products, and reciprocals of random variables by treating these operations as continuous functions. The sum of two sequences of random variables $X_n$ and $Y_n$ converges to the sum of their respective limits:

$$X_n + Y_n \overset{\text{a.s.}}{\to} X + Y \quad \text{(respectively in probability)}.$$

This follows because the function $g : \mathbb{R}^2 \to \mathbb{R},\ g(x, y) := x + y$ is continuous.

The product of two sequences of random variables $X_n$ and $Y_n$ converges to the product of their respective limits:

$$X_n Y_n \overset{\text{a.s.}}{\to} XY \quad \text{(respectively in probability)}.$$

This holds because $g : \mathbb{R}^2 \to \mathbb{R},\ g(x, y) = xy$ is a continuous function on $\mathbb{R}^2$.

If $\mathbb{P}(Y \neq 0) = 1$, then the reciprocals $1/Y_n$ converge to $1/Y$:

$$\frac{1}{Y_n} \overset{\text{a.s.}}{\to} \frac{1}{Y} \quad \text{(respectively in probability).}$$

This is valid because $g(y) = 1/y$ is continuous on $\mathbb{R} \setminus \{0\}$, and the condition $Y \neq 0$ ensures the continuity of $g$ at $Y$.

For combinations of sums, products, and reciprocals, the CMT applies iteratively-the composition of continuous functions is continuous. For example:

$$\frac{X_n Y_n}{X_n + Y_n} \overset{\text{a.s.}}{\to} \frac{XY}{X + Y} \quad \text{(respectively in probability),}$$

provided $X + Y \neq 0$, since the mapping $g(x, y) = \frac{xy}{x+y}$ is continuous on its domain.

## 2.2 Convergence in $L^p$ Mode

The third form of convergence can be seen as a generalization of convergence in probability and is based on the $L^p$ norms and the associated metrics. The definition of interest is as follows. Recall that $p$ is a real number greater than or equal to one:

**Definition 3.** We say that $X_n$ converges in $L^p$ metric or in $p$-th mean (in $L^p$, in $p^{th}$ mean) to $X$, and we denote it as $X_n \overset{L^p}{\to} X$, if and only if

$$\lim_{n\to\infty} \mathbb{E}\left([d(X_n, X)]^p\right) = 0.$$

When $p = 2$, this is also called convergence in *quadratic mean.*

This convergence implies that the sequence of random variables $(d(X_n, X))$ is uniformly bounded in the $L^p$ metric, i.e., $\sup_n \mathbb{E}(d(X_n, X)^p) < +\infty$. Furthermore, the behavior of this form of convergence for different values of $p$ is not hard to derive using Jensen's inequality.

**Lemma 2.** *If $1 \leq p^\star < p$ and $X_n \overset{L_p}{\to} X$, then $X_n \overset{L_{p^\star}}{\to} X$.*

*Proof.* Since $1 \leq p^\star < p$, we have

$$\mathbb{E}\left([d(X_n, X)]^p\right) \geq \mathbb{E}\left([d(X_n, X)]^{p^\star}\right)^{\frac{p}{p^\star}}$$

by Jensen's inequality. The result follows by taking limits on both sides of the above inequality. □

Clearly, the above implies that non-convergence for smaller $p$ also implies non-convergence for larger $p$. Thus, the strength of the corresponding convergence increases with $p$. The relative strength of these convergences can be further understood by noting that convergence in $L^p$ implies convergence in probability. Therefore, the latter can also be interpreted as the minimal form of convergence in comparison to $p$-mean convergences.[3]

To relate the above to convergence in probability, we use Markov's inequality, leading to the following result:

**Lemma 3.** *If $X_n \overset{L^p}{\to} X$, then $X_n \overset{p}{\to} X$.*

*Proof.* Using Markov's inequality, for any $\varepsilon > 0$, we have

$$\mathbb{P}\left(d(X_n, X) \geq \varepsilon\right) \leq \frac{\mathbb{E}\left([d(X_n, X)]^p\right)}{\varepsilon^p}.$$

[3] We can symbolically consider convergence in probability as $L^0$ convergence.

Taking limits on both sides of the inequality yields the desired result. □

Similarly, the above is equivalent to stating that non-convergence in probability implies non-convergence in $L^p$ metric. The converse is not generally true. To better understand why the direction cannot be reversed, we need the following definition.

**Definition 4.** A sequence of random variables $(y_n)$ is said to be uniformly $L^p$-integrable if and only if

$$\lim_{K\to\infty}\sup_n \mathbb{E}(|y_n|^p 1_{|y_n|\geq K}) = 0.$$

The definition imposes, in addition to uniform boundedness, restrictions on how "slowly" the tails of the distributions are allowed to decay. Uniform boundedness of the corresponding sequence of moments alone is not sufficient, as the following example illustrates:

**Example 6.** Referring to Example 1, and for $p = 1$, we observe that

$$\sup_n \mathbb{E}(|X_n| 1_{|X_n|\geq K}) = \begin{cases} 1, & K \leq n, \\ 0, & K > n. \end{cases}$$

Thus, while the sequence $(d(X_n, 0))$ is $L^1$-bounded, it is not $L^1$-uniformly integrable. For $p > 1$, it is not even uniformly bounded. Note that in this example, although we have convergence to zero in probability, there is no corresponding $L^p$ convergence for any $p \geq 1$ because $\mathbb{E}(d(X_n, 0)) = 1, \forall n \geq 1$. This observation also informs that almost sure convergence as well as convergence in probability do not necessarily imply $L^p$ convergence-why?

The above example raises the question of whether convergence in probability supplemented with uniform integrability, leads to convergence in $L^p$. The following results confirm this:

**Lemma 4.** *Let* $(y_n)$ *be a sequence of random variables such that:*

- $y_n \xrightarrow{p} 0$, *and,*
- *the sequence is uniformly* $L^p$*-integrable.*

*Then*

$$\lim_{n\to\infty} \mathbb{E}\left(|y_n|^p 1_{|y_n|\geq\varepsilon}\right) = 0,\ \forall\varepsilon > 0.$$

*Proof.* Let $\varepsilon > 0$ and $K > \varepsilon$. Notice that $\mathbb{E}\left(|y_n|^p 1_{|y_n|\geq\varepsilon}\right) = \mathbb{E}\left(|y_n|^p 1_{\varepsilon\leq|y_n|<K}\right) + \mathbb{E}\left(|y_n|^p 1_{|y_n|\geq K}\right)$. Furthermore, $\mathbb{E}\left(|y_n|^p 1_{\varepsilon\leq|y_n|<K}\right) \leq K^p\mathbb{P}(|y_n| \geq \varepsilon)$, due to monotonicity (essentially, $\mathbb{P}(|y_n| > \varepsilon) \geq \mathbb{P}(\varepsilon \leq |y_n| < K)$), and $0 \leq \mathbb{E}\left(|y_n|^p 1_{|y_n|\geq K}\right) \leq \sup_n \mathbb{E}\left(|y_n|^p 1_{|y_n|\geq K}\right)$. Thus, it is obtained that $0 \leq \mathbb{E}\left(|y_n|^p 1_{|y_n|\geq\varepsilon}\right) \leq K^p\mathbb{P}(|y_n| \geq \varepsilon) + \sup_n \mathbb{E}\left(|y_n|^p 1_{|y_n|\geq K}\right)$. Now, for any $\delta > 0$, and due to uniform integrability, there exists some $K_\delta > \varepsilon$ such that $\sup_n \mathbb{E}\left(|y_n|^p 1_{|y_n|\geq K_\delta}\right) \leq \frac{\delta}{2}$; also due to convergence in probability there exists some $n^\star(\delta)$, such that $\mathbb{P}(|y_n| \geq \varepsilon) \leq \frac{\delta}{2K_\delta^p}, \forall n \geq n^\star(\delta)$. Hence, $\mathbb{E}\left(|y_n|^p 1_{|y_n|\geq\varepsilon}\right) \leq K_\delta^p \frac{\delta}{2K_\delta^p} + \frac{\delta}{2} = \delta$. Since $\delta$ is arbitrary this establishes that $\lim_{n\to\infty} \mathbb{E}\left(|y_n|^p 1_{|y_n|\geq\varepsilon}\right) = 0$. The result follows since $\varepsilon$ is arbitrary. □

**Lemma 5.** *If* $X_n \xrightarrow{p} X$ *and* $(d(X_n, X))$ *is* $L^p$*-uniformly integrable, then* $X_n \xrightarrow{L^p} X$.

*Proof.* Starting with $\varepsilon > 0$:

$$\lim_{n\to\infty} \mathbb{E}\left([d(X_n, X)]^p\right) =$$

$$\lim_{n\to\infty} \mathbb{E}\left([d(X_n, X)]^p\, 1_{d(X_n,X)<\varepsilon}\right) + \lim_{n\to\infty} \mathbb{E}\left([d(X_n, X)]^p\, 1_{d(X_n,X)\geq\varepsilon}\right).$$

Clearly, $\lim_{n\to\infty} \mathbb{E}\left([d(X_n,X)]^p\, 1_{d(X_n,X)<\varepsilon}\right) < \varepsilon^p$, and from the previous lemma, $\lim_{n\to\infty} \mathbb{E}\left([d(X_n,X)]^p\, 1_{d(X_n,X)\geq\varepsilon}\right) = 0$. Hence, we conclude that:

$$\lim_{n\to\infty} \mathbb{E}\left([d(X_n,X)]^p\right) < \varepsilon^p.$$

Since $\varepsilon$ is arbitrary, $\lim_{n\to\infty} \mathbb{E}\left([d(X_n,X)]^p\right) = 0$, which is the desired result. □

*Remark* 1. Hence, almost sure convergence complemented with uniform $L^p$-integrability also implies $L^p$ convergence-why?

Finally, note that $L^p$ convergence does not imply almost sure convergence, as shown by the following example:

**Example 7.** Consider a sequence of random variables with $X_1 = 1$, $(X_2, X_3) = (1,0)$ or $(0,1)$ with probability $\frac{1}{2}$. More generally, for $k = 1,2,\ldots$, the sequence section $(X_{\frac{1}{2}k(k-1)+1}, \ldots, X_{\frac{1}{2}k(k+1)})$ equals $(1,0,\ldots,0)$, or $(0,1,\ldots,0)$, or ..., or $(0,0,\ldots,1)$ with probability $\frac{1}{k}$. Thus, for $X_n \in (X_{\frac{1}{2}k(k-1)+1}, \ldots, X_{\frac{1}{2}k(k+1)})$, we have $\mathbb{P}(X_n = 1) = \mathbb{E}(X_n^p) = \frac{1}{k}, \forall p \geq 1$.

It is clear that as $n \to \infty$ (and therefore $k \to \infty$), $X_n \xrightarrow{L^p} 0, \forall p \geq 1$, and consequently, $X_n \xrightarrow{p} 0$. However, we also observe that:

$$\lim_{m\to\infty} \mathbb{P}\left(\sup_{n\geq m} X_n = 0\right) = 0,$$

since beyond the $m$-th component of the sequence, there will almost surely be infinitely many members of the sequence equal to one. Thus, $X_n \nrightarrow^{a.s.} 0$, meaning that convergence in $L^p$ metric does not ensure almost sure convergence.

> *Exersice*: Provide an example in which $L^p$ convergence guarantees almost sure convergence.

## 2.3 Transfer Principle for Lipschitz Transformations

If $(S, d_S)$ and $(S', d_{S'})$ are metric spaces, and $g : S \to S'$ is a function between them, then $g$ is called Lipschitz continuous if and only if there exists a constant $l_g > 0$ such that for every $s_1, s_2 \in S$,

$$d_{S'}(g(s_1), g(s_2)) \leq l_g d_S(s_1, s_2).$$

The constant $l_g$, which is independent of $S$, is not unique but is the smallest upper bound of all such constants and is called the Lipschitz constant of the function. It is proved that every Lipschitz continuous function is also continuous in the usual sense, but the converse is not true. When the spaces involved are Euclidean, it can be shown that Lipschitz continuity is equivalent to being almost everywhere differentiable (in the sense of Lebesgue measure;) with bounded derivatives. For instance, the exponential function is not Lipschitz continuous, though it is continuous. However, it becomes Lipschitz continuous when its domain is restricted to any arbitrary bounded subset of the real numbers. Extending this reasoning, one could develop a localized version of this concept. In any case, this form of continuity is highly useful in convergence analysis, as it can provide insights, for instance, about the rates of convergence when such rates are well-defined.

Regarding the transfer of limits under continuous transformations, we note that general results, such as Theorem 2, are generally not applicable for convergence in the $L^p$ metric. The transformation may not even be a suitably integrable function. Nevertheless, if we strengthen the continuity condition to Lipschitz continuity, the desired transfer of limits can be achieved. Our framework here is similar to that of Paragraph 2.1.

**Theorem 3.** *Let $g : S \to S'$ be Lipschitz continuous with respect to the involved metrics. Then, if $X_n \overset{L^p}{\to} X$, we also have $g(X_n) \overset{L^p}{\to} g(X)$.*

*Proof.* Observe that due to the Lipschitz continuity of $g$ and the monotonicity of the integral,

$$\mathbb{E}((d'(g(X_n), g(X)))^p) \leq l_g^p \mathbb{E}((d(X_n, X))^p),$$

where $l_g$ is the Lipschitz constant. The result follows by taking limits on both sides of the inequality, given the non-negativity of the left-hand side. □

> *Exersice*: Let $S = S' = \mathbb{R}^p$, and let $A$ be a $p \times p$ matrix. If $X_n \overset{L^p}{\to} X$, does it hold that $AX_n \overset{L^p}{\to} AX$?

**Example 8.** Let $X_n = \frac{1}{n}Z + W$, where $Z$ and $W$ are random variables with $\mathbb{E}[Z^2] < \infty$. Then $X_n \xrightarrow{L^2} W$, since $\mathbb{E}(|X_n - W|^2) = \frac{1}{n^2}\mathbb{E}(Z^2) \to 0$.

Let $g(x) = \sin(x)$, which is Lipschitz continuous with Lipschitz constant $l_g = 1$. By the Lipschitz Transfer Principle:

$$\sin(X_n) \xrightarrow{L^2} \sin(W).$$

# 3 Weak Convergence

We now focus on the weakest-compared to the previous-form of stochastic convergence. Our framework now considers a sequence of probability measures $(\mathbb{P}_n)_{n\in\mathbb{N}}$ on $(S, \mathcal{B})$. Compared to the previous sections, $X_n \sim \mathbb{P}_n$, but it is not necessary that $\mathbb{P}_n = \mathbb{P}(X_n^{-1} \in \cdot)$, as we allow the involved random elements to be defined on different probability spaces.

> The set of all bounded functions $S \to \mathbb{R}$ is denoted by $B(S, \mathbb{R})$, while the subset containing functions that are both bounded and continuous, is denoted by $B_c(S, \mathbb{R})$.

> *Exersice*: Show that $B(S, \mathbb{R})$ and $B_c(S, \mathbb{R})$ are non-empty.

We are interested in a concept of convergence of probability distributions, and an intuitively natural way to define such a notion is by requiring the probability assigned by $\mathbb{P}_n$ to any measurable subset of $S$, to converge (as a real number) as $n \to \infty$, to the probability assigned to the same set by $\mathbb{P}$. This is essentially some sort of functional convergence for the probability distributions involved, and it is termed as convergence in total variation. This form of convergence relies on the total variation metric between two distributions. Specifically, for arbitrary distributions $\mathbb{Q}, \mathbb{Q}^\star$ on $S$, the total variation distance is defined as the supremum over the absolute differences in probabilities assigned to elements of $\mathcal{B}$:

$$\text{TV}(\mathbb{Q}, \mathbb{Q}^\star) := \sup_{A\in\mathcal{B}} |\mathbb{Q}(A) - \mathbb{Q}^\star(A)|.$$

It can be shown that the total variation metric is equivalently represented as the uniform distance over bounded measurable functions $S \to \mathbb{R}$ between their integrals with respect to the involved distributions:

$$\text{TV}(\mathbb{Q}, \mathbb{Q}^\star) = \sup_{f \in B(S,\mathbb{R})} \left| \int_S f(s)\, d\mathbb{Q} - \int_S f(s)\, d\mathbb{Q}^\star \right|.$$

Hence $\mathbb{P}_n$ converges in total variation to $\mathbb{P}$ iff $\sup_{A \in \mathcal{B}} |\mathbb{P}_n(A) - \mathbb{P}^\star(A)| \to 0$, or equivalently, $\sup_{f \in B(S,\mathbb{R})} \left| \int_S f(s)\, d\mathbb{P}_n - \int_S f(s)\, d\mathbb{P}^\star \right| \to 0$. Yet, this-intuitively natural mode of convergence is quite strong for our needs; it requires a lot of conditions that can easily fail in circumnstances of interest.

We can obtain weaker forms of convergence by selecting smaller collections of functions to construct the distances between the integrals.

Using $B_c(S, \mathbb{R}) \subset B(S, \mathbb{R})$, we derive a metric that defines weak convergence.

**Definition 5.** Let $(\mathbb{P}_n)_{n \in \mathbb{N}}$ be a sequence of Borel probability measures on $S$, and let $\mathbb{Q}$ be a Borel probability measure. The sequence $(\mathbb{P}_n)_{n \in \mathbb{N}}$ is said to converge weakly to $\mathbb{Q}$, denoted by $\mathbb{P}_n \rightsquigarrow \mathbb{Q}$, if

$$\sup_{f \in BL_1(S,\mathbb{R})} \left| \int_S f(s)\, d\mathbb{P}_n - \int_S f(s)\, d\mathbb{Q} \right| \to 0.$$

Furthermore, if $X_n \sim \mathbb{P}_n$ and $X \sim \mathbb{Q}$, then the random elements $X_n$ are said to converge in distribution to $X$, denoted as $X_n \rightsquigarrow X$.

The above implies that convergence in total variation implies weak convergence, or equivalently, that divergence (weak) implies divergence in total variation. The systematic study of weak convergence is complicated, even in the case $S = \mathbb{R}$, and exceeds the scope of the present notes. Readers interested in the compactness of the converging sequence, characterizations using alternative sets of functions, properties inherited by $S$ on the collection of probability measures when equipped with the metric in the above definition, etc., can refer to the literature (e.g., see [14]).

## 3.1 The Portmanteau Theorem

We have defined the notion of weak convergence, via integrals over a class of functions. But what does this mean for the probabilities of particular subsets of the underlying space? We state-without proof-(a part of) the so-called Portmanteau Theorem (see also [14]), which provides equivalent conditions for weak convergence of measures in metric spaces, thus formalizing the concept.[4] To proceed, recall that if $A \subseteq S$, then $\bar{A}$ denotes the *smallest closed superset* of $A$ (the intersection of all closed subsets in the topology containing $A$, the closure of $A$), and $A^o$ denotes the *largest open subset* of $A$ (the union of all open subsets in the topology contained in $A$, the interior of $A$). Clearly, $\bar{A} \supseteq A^o$, and $\partial A := \bar{A} - A^o$ (the boundary of $A$). Moreover, if $A \in \mathcal{B}_S$, then $\bar{A}, A^o \in \mathcal{B}_S$ because the Borel $\sigma$-algebra includes the topology of $S$ induced by its metric. We then have the following:

**Theorem 4.** *[Portmanteau Theorem] The following are equivalent:*

1. $\mathbb{P}_n \rightsquigarrow \mathbb{P}$.

2. $\int_S f(z) d\mathbb{P}_n \to \int_S f(z) d\mathbb{P}, \forall f \in B_c(S, \mathbb{R})$.

3. $\lim_{n\to\infty} \mathbb{P}_n(A) = \mathbb{P}(A), \forall A \in \mathcal{B}$ *such that* $\mathbb{P}(\bar{A}) = \mathbb{P}(A^o)$.

[4] We present only part of the theorem, omitting, for instance, descriptions involving $\liminf_{n\to\infty} \mathbb{P}_n(A) \geq \mathbb{P}(A)$ for open sets $A$, or $\limsup_{n\to\infty} \mathbb{P}_n(A) \leq \mathbb{P}(A)$ for closed sets $A$.

4. *If $S = \mathbb{R}^n$, whence the corresponding cumulative distribution functions are well-defined, then $\lim_{n\to\infty} F_n(\mathbf{x}) = F(\mathbf{x})$, for all $\mathbf{x} \in \mathbb{R}^n$ at which $F$ is continuous.*

Note that since $\bar{A} = \partial A \cup A^o$ and $\partial A \cap A^o = \emptyset$, the third statement holds for all $A \in \mathcal{B}$ such that $\mathbb{P}(\partial A) = 0$.[5] Furthermore, the fourth statement is simply a specialization of the third for cases where cumulative distribution functions are well-defined.

**Example 9.** Let $S = \mathbb{R}$ and $\mathbb{P}_n = \text{Deg}_{1/n}$. Since for each $n$, $1 = \text{TV}(\mathbb{P}_n, \text{Deg}_0) = |\mathbb{P}_n(\{0\}) - \text{Deg}_0(\{0\})|$, the sequence does not converge in total variation to the degenerate distribution at zero. However, it does converge weakly, as discrete sets containing zero are not continuity sets for the aforementioned distribution. For any continuity set containing $0$, $1/n$ will lie within it for sufficiently large $n$, guaranteeing via the third condition of the Portmanteau Theorem that $\mathbb{P}_n \rightsquigarrow \text{Deg}_0$.

This example, while simple, is strong enough to illustrate that weak convergence may be sufficiently weak to allow limits in cases where stronger forms of convergence do not. The Central Limit Theorems, which will be stated later in this chapter and used subsequently, will demonstrate that weak convergence is also highly useful.

## 3.2 Relations between Modes of Convergence

In this subsection, to examine the relationship between convergence in distribution and other forms of convergence (in probability, $L^p$, and almost

[5] Borel sets for which this holds are called *continuity sets* of $\mathbb{P}$.

sure), we assume $\mathbb{P}_n = \mathbb{P}(X_n \in \cdot)$, i.e., the involved random elements are defined on the same probability space. We state the following:

**Lemma 6.** *If $X_n \xrightarrow{p} X$, then $X_n \rightsquigarrow X$. Thus, both almost sure convergence and convergence in $L^p$ also imply convergence in distribution.*

*Proof.* Let $A \in \mathcal{B}$ be an arbitrary continuity set for $\mathbb{P} := \mathbb{P}(X \in \cdot)$. It can be shown (see, e.g., [6]) that due to convergence in probability, $\lim_{n\to\infty} \mathbb{P}(d(X_n, X) \geq \varepsilon) = 0, \forall \varepsilon > 0$, there exists a sequence of positive numbers $(\epsilon_n)$ such that $\epsilon_n \to 0$ and $\lim_{n\to\infty} \mathbb{P}(d(X_n, X) \geq \epsilon_n) = 0$. Clearly,

$$\{X_n \in A\} = (\{X_n \in A\} \cap \{d(X_n, X) \leq \epsilon_n\}) \cup (\{X_n \in A\} \cap \{d(X_n, X) \geq \epsilon_n\}),$$

and the sets in the union are disjoint. By additivity, and using a slight abuse of notation,

$$\mathbb{P}_n(A) = \mathbb{P}(X_n \in A \cap d(X_n, X) \leq \epsilon_n) + \mathbb{P}(X_n \in A \cap d(X_n, X) \geq \epsilon_n),$$

for all $n$. By monotonicity,

$$\mathbb{P}(X_n \in A \cap d(X_n, X) \geq \epsilon_n) \leq \mathbb{P}(d(X_n, X) \geq \epsilon_n),$$

and the probability on the right-hand side converges to zero due to convergence in probability. By positivity, the same holds for the probability on the left-hand side.

Regarding $\mathbb{P}(X_n \in A \cap d(X_n, X) \leq \epsilon_n)$, it is upper-bounded by $\mathbb{P}(X \in A^{\epsilon_n})$, where $A^{\epsilon} := \{s \in S : d(s, y) \leq \epsilon, \exists y \in A\}$, since if the intersection event occurs, then $X$ must lie within this $\epsilon_n$-"enlargement" of $A$. Consequently, due to the continuity of $\mathbb{P}$ and the fact that $A^{\epsilon_n}$ shrinks under set inclusion as $\epsilon_n$ decreases, we have $\lim_{n\to\infty} \mathbb{P}(X_n \in A \cap d(X_n, X) \leq \epsilon_n) \leq \mathbb{P}(X \in \bar{A}) = \mathbb{P}(X \in A)$, where the last equality follows from $A$ being a continuity set.

Additionally, the intersection of $X \in A$ with $d(X_n, X) \leq \epsilon_n, \forall n$, coincides with $X_n \in A \cap d(X_n, X) \leq \epsilon_n, \forall n$, due to symmetry. Hence, $X \in A \supseteq X_n \in A \cap d(X_n, X) \leq \epsilon_n, \forall n$, and, due to monotonicity and the continuity of $\mathbb{P}$, $\mathbb{P}(X \in A) \geq \lim_{n\to\infty} \mathbb{P}(X_n \in A \cap d(X_n, X) \leq \epsilon_n)$. Combining this with the above, we finally obtain $\mathbb{P}(X \in A) = \lim_{n\to\infty} \mathbb{P}(X_n \in A \cap d(X_n, X) \leq \epsilon_n)$, and the first result follows from condition (3) of the Portmanteau Theorem and the fact that $A$ is arbitrary. The remaining results follow from Lemmas 1 and 3. □

The converse does not hold. Convergence in distribution does not imply convergence in probability and, consequently, the other two forms of convergence, as shown by the following example:

**Example 10.** Let $S = \mathbb{R}$, and $X \sim \text{Unif}_{[-1,1]}$. For $n \geq 1$, let $X_n = -X + \frac{1}{n}$. Then $\mathbb{P}(|X_n - X| \leq \nicefrac{1}{2}) = \mathbb{P}(-\nicefrac{1}{4} + \nicefrac{1}{2n} \leq X \leq \nicefrac{1}{4} - \nicefrac{1}{2n}) \to \nicefrac{1}{2}$—why? Therefore, $X_n \overset{p}{\nrightarrow} X$. Moreover, we have that $F_n(x) = \mathbb{P}(-X \leq x - \nicefrac{1}{n}) = \mathbb{P}(X \geq -x + \nicefrac{1}{n}) =$

$$1 - \mathbb{P}(X \leq -x + \nicefrac{1}{n}) = \begin{cases} 1, & x \geq 1 - \nicefrac{1}{n}, \\ \frac{1+x}{2}, & -1 - \nicefrac{1}{n} \leq x < 1 - \nicefrac{1}{n}, \\ 0, & x \leq -1 - \nicefrac{1}{n} \end{cases} \to \begin{cases} 1, & x \geq 1, \\ \frac{1+x}{2}, & -1 \leq x < 1, \\ 0, & x \leq -1 \end{cases},$$

for every $x \in \mathbb{R}$, and therefore, from the last case of the previous theorem, it follows that $X_n \rightsquigarrow X$.

Thus, convergence in distribution is the weakest among the forms we have discussed. However, there exists a case where convergence in distribution is equivalent to convergence in probability; this occurs when the limit is degenerate:

**Lemma 7.** *If* $X_n \rightsquigarrow s \in S$, *then* $X_n \xrightarrow{p} s$.

*Proof.* Let $\varepsilon > 0$ be arbitrary, and notice that $\mathbb{P}(d(X_n, s) \leq \varepsilon) = \mathbb{P}(X_n \in \bar{B}_s(\varepsilon))$, with $\bar{B}_s(\varepsilon))$ denoting the $d$-closed ball centered at $s$ and of radius $\varepsilon$. Since the distribution of $s$ is degenerate at $s$, $\bar{B}_s(\varepsilon))$ is a continuity set for it, and thereby due to the Portmanteau Theorem $\lim_{n\to\infty} \mathbb{P}(X_n \in \bar{B}_s(\varepsilon)) = \mathbb{P}(s \in \bar{B}_s(\varepsilon)) = 1$, and thereby the result follows from the definition of convergence in probability. □

> *Exersice*: Does $X_n \rightsquigarrow s \in S$ necessarily imply $X_n \xrightarrow{\text{a.s.}} s$ or $X_n \xrightarrow{L^p} s$?

> *Exersice*: Provide an example where Lemma 7 holds and simultaneously $X_n \xrightarrow{L^p} s$.

## 3.3 Continuous Mapping Theorem

Similarly to the corresponding section on earlier forms of stochastic convergence, the question here pertains to what happens to measures when transformed through a continuous (or almost everywhere continuous) measurable function from $(S, d)$ to the metric space $(S', d')$, assuming weak convergence holds for the original measures. We have the following mapping theorem:

**Theorem 5.** *[Continuous Mapping Theorem] Let* $\mathbb{P}_n \rightsquigarrow \mathbb{P}$, *and let* $h : S \to S'$ *be Borel measurable, with* $\mathbb{P}(C_h) = 1$, *where* $C_h$ *consists of the points in* $S$ *where*

*$h$ is continuous. Then, $\mathbb{P}_n(h \in \cdot) \rightsquigarrow \mathbb{P}(h \in \cdot)$. Correspondingly, for the involved random elements $X_n \sim \mathbb{P}_n$ and $X \sim \mathbb{P}$, we have*

$$h\left(X_n\right) \rightsquigarrow h\left(X\right).$$

*Proof.* Initially, note that for all $A \in \mathcal{B}_{S'}$,

$$\mathbb{P} \circ h^{-1}(A) = \mathbb{P}(h^{-1}(A)) = \mathbb{P}(h^{-1}(A) \cap C_h) + \mathbb{P}(h^{-1}(A) \cap C_h^c),$$

since $h^{-1}(A) \cap C_h$ and $h^{-1}(A) \cap C_h^c$ are disjoint by construction. Furthermore, by assumption, $\mathbb{P}(h^{-1}(A) \cap C_h^c) = 0$, as $\mathbb{P}(C_h^c) = 0$, $h^{-1}(A) \cap C_h^c \subseteq C_h^c$, and using the monotonicity of the measure. Thus,

$$\mathbb{P} \circ h^{-1}(A) = \mathbb{P}(h^{-1}(A) \cap C_h). \tag{1}$$

Now, let $B \in \mathcal{B}_{S'}$, such that $\mathbb{P}h^{-1}(\bar{B}) = \mathbb{P}h^{-1}(B^o)$ . Using (1), (2) becomes

$$\mathbb{P}(h^{-1}(\bar{B}) \cap C_h) = \mathbb{P}(h^{-1}(B^o) \cap C_h). \tag{3}$$

From the definition of continuity, we also have

$$\left[h^{-1}(B^o) \cap C_h\right] \subseteq \left[\left(h^{-1}(B)\right)^o \cap C_h\right] \subseteq \left[\overline{h^{-1}(B)} \cap C_h\right] \subseteq \left[h^{-1}(\bar{B}) \cap C_h\right]. \tag{4}$$

Combining (4) with (3) and the monotonicity of the measure, we get

$$\mathbb{P}(\left(h^{-1}(B)\right)^o \cap C_h) = \mathbb{P}(\overline{h^{-1}(B)} \cap C_h). \tag{5}$$

Using (1), (5) implies

$$\mathbb{P}(\big(h^{-1}(B)\big)^o) = \mathbb{P}(\overline{h^{-1}(B)}). \tag{6}$$

From (6) and the weak convergence of the original measures, by the third proposition of the Portmanteau Theorem, we have

$$\mathbb{P}_n\left(h^{-1}(B)\right) \to \mathbb{P}\left(h^{-1}(B)\right). \tag{7}$$

Since $B$ was chosen arbitrarily, (7) implies

$$\mathbb{P}_n \circ h^{-1}(B) \to \mathbb{P} \circ h^{-1}(B), \quad \forall B \in \mathcal{B}_{S'} : \mathbb{P} \circ h^{-1}(\bar{B}) = \mathbb{P} \circ h^{-1}(B^o). \tag{8}$$

Finally, (8) implies the desired convergence by the third proposition of the Portmanteau Theorem. □

Similarly to the analogous section above, the transfer principle here also implies a multitude of general convergence results. For example, if $S, S'$ have vector space structures with operations continuous with respect to the involved metrics, then the corresponding limits of sums exist and coincide with the sums of the limits, etc. Similarly, as previously mentioned, if we construct new random elements through finite Cartesian products of the underlying spaces, the convergence of the constituent elements implies the convergence of the vectors with these components, etc.

> *Exersice*: Formulate and prove the above precisely.

The analogous simple example related to the Riemann integral:

**Example 11.** If $\Omega = S = \mathcal{C}_{[0,1]}$, based on the fact that the Riemann integral operator $\mathcal{C}_{[0,1]} \ni f \mapsto \int_0^1 f(z)dz$ is continuous, we have that if $X_n \rightsquigarrow X$, then

$$\int_0^1 X_n(\omega, z)dz \rightsquigarrow \int_0^1 X(\omega, z)dz.$$

Analogously to the discussion regarding the version of the theorem for almost sure convergence, and convergence in probability, the following example is potentially more useful for what follows:

**Example 12.** Let $(X_n)_{n\geq 1}$ and $(Y_n)_{n\geq 1}$ be sequences of random variables. We examine the behavior of sums, products, and reciprocals of these sequences, leveraging the *Continuous Mapping Theorem (CMT)*. In order to do that it is necessary to begin from a weak convergence premise occuring in $\mathbb{R}^2$. We suppose then that:

$$(X_n, Y_n) \rightsquigarrow (X, Y).$$[6]

The sum of two sequences of random variables $X_n$ and $Y_n$ converges in distribution to the sum of their respective limits:

$$X_n + Y_n \rightsquigarrow X + Y.$$

This follows because the function $g(x, y) = x + y$ is continuous.

[6] Contrary to what occurs regarding almost sure and convergence in probability, this is not equivalent to that $X_n \rightsquigarrow X$ and $Y_n \rightsquigarrow Y$. The former (*joint convergence*) implies the latter (*marginal convergence*) but not vice versa. This is due to the fact that generally, due to dependence, the joint distribution of a random vector, contains more information than the set of the marginal distributions of its constituents random variables. A sufficient condition for equivalence between joint and marginal convergence, is that eventually, $X_n$ be independent of $Y_n$-see the further exercises section in the Addendum.

The product of two sequences of random variables $X_n$ and $Y_n$ converges in distribution to the product of their respective limits:

$$X_n Y_n \rightsquigarrow XY.$$

This holds because $g(x, y) = xy$ is a continuous function on $\mathbb{R}^2$.

If $Y_n \rightsquigarrow Y$ and $Y \neq 0$ almost surely, then the reciprocals $1/Y_n$ converge in distribution to $1/Y$:

$$\frac{1}{Y_n} \rightsquigarrow \frac{1}{Y}.$$

This is valid because $g(y) = 1/y$ is continuous on $\mathbb{R} \setminus \{0\}$, and the condition $Y \neq 0$ ensures the continuity of $g$ at $Y$.

For combinations of sums, products, and reciprocals, the CMT applies iteratively. For example:

$$\frac{X_n Y_n}{X_n + Y_n} \rightsquigarrow \frac{XY}{X + Y},$$

provided $X + Y \neq 0$, since the mapping $g(x, y) = \frac{xy}{x+y}$ is continuous on its domain.

Lemma 7 along with the above example provide directly with a result known as (part of) Slutsky's Lemma:

**Lemma 8.** *(Slutsky) If $X_n \rightsquigarrow X$, and $Y_n \rightsquigarrow s$, then $X_n + Y_n \xrightarrow{p} X + s$ and $X_n Y_n \xrightarrow{p} Xs$.*

*Proof.* The fact that $s$ is degenerate, implies that $X_n$ is asymptotically independent from $Y_n$. Thus marginal convergences imply the joint convergence $(X_n, Y_n) \rightsquigarrow (X, s)$. The results follow from the CMT as in the example above. □

## 3.4 A "General" Central Limit Theorem

Central limit theorems (CLTs) describe cases of convergence in distribution. They concern the asymptotic behavior of partial sums of random variables (or random vectors, or more generally random elements with values in spaces where addition is algebraically feasible), suitably shifted and scaled so that their resulting distributions converge weakly to normal distributions. The latter serve as significant attractors for many such sequences of partial sums. A detailed enumeration and proof of such results are beyond the scope of this book. The following theorem is presented without proof-see [8]-and pertains to stationary and ergodic sequences whose component marginal distributions possess sufficiently high-order moments, and where the dependence between elements diminishes sufficiently fast as they become temporally distant:

**Theorem 6.** *Let $(X_t)_{t\in\mathbb{Z}}$ be a stationary and ergodic sequence satisfying:*[7]

1. *There exists $\varepsilon > 0$ such that $\mathbb{E}(|X_0|^{2+\varepsilon}) < +\infty$, and,*
2. *For every $t > 0$, $\sum_{k=0}^{\infty} |Cov(\mathbb{E}(X_t \mid \sigma(X_n, n \leq 0)), X_k)| < +\infty$.*

*Then, as $n \to \infty$,*

$$Var\left(\frac{1}{\sqrt{n}}\sum_{t=1}^{n}(X_t - \mathbb{E}(X_0))\right) \to v \geq 0. \tag{1}$$

[7] The summation index is now turned to $t$ in order for the time series setting to be stressed.

*If* $v > 0$*, then*

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n}(X_t - \mathbb{E}(X_0)) \rightsquigarrow X \sim N(0, v). \tag{2}$$

*Moreover, if* $\tau \in [0, 1]$*, then*

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{\lfloor n\tau \rfloor}(X_t - \mathbb{E}(X_0)) \rightsquigarrow vW(\tau), \tag{3}$$

*in the space of CADLAG functions on* $[0, 1]$ *equipped with the Skorokhod metric.*

We observe that the properties of stationarity and ergodicity, along with the first condition of the theorem, and the fact that this implies the existence of the first-order moment for the sequence members, already guarantee the validity of Birkhoff's Law of Large Numbers. Consequently, due to the relationships between forms of convergence and the continuous mapping theorem, it holds that $\frac{1}{n}\sum_{t=1}^{n}(X_t - \mathbb{E}(X_0)) \rightsquigarrow 0$ (explain!).

The theorem refines this convergence to zero; it informs us that multiplying $\frac{1}{n}\sum_{t=1}^{n}(X_t - \mathbb{E}(X_0))$ by the rate of convergence $\sqrt{n}$ yields a stochastic limit in distribution—a non-degenerate random variable following a normal distribution. Thus, among other things, it provides information about the rate at which the Birkhoff convergence to the first moment occurs, namely at most at a rate of $\sqrt{n}$. This holds because if we multiplied $\frac{1}{n}\sum_{t=1}^{n}(X_t - \mathbb{E}(X_0))$ by a sequence diverging to infinity faster than $\sqrt{n}$, it would not be difficult to show—using the theorem—that we would obtain a sequence of distributions whose probability mass shifts towards infinity.[8] In such a case, a

[8] This property is known as (non-)uniform tightness, which is related to compactness, briefly mentioned in the introductory paragraph.

weak limit that is a well-defined distribution over the real numbers cannot exist.[9]

The conditions for validity relate to:

1. The existence of moments of sufficiently high order for the random variables forming the partial sum. Any order greater than 2 suffices. Under stronger assumptions for condition (b)—e.g., independence—this condition could be weakened to require moments of order 2. However, the theorem would not hold without at least the second moment since Var$(\frac{1}{\sqrt{n}}\sum_{t=1}^{n}(X_t - \mathbb{E}(X_0)))$ would not be well-defined, and (1) could not hold.

2. The rate at which a specific type of dependence between the constituent random variables decays as they become temporally distant. This is described by the requirement for the convergence of the series of covariances $\sum_{k=0}^{\infty}|$Cov$(\mathbb{E}(X_t|\sigma(X_n, n \leq 0)), X_k)|$. This is ensured when the covariance between the $L^2$-projection of $X_t$ onto $\sigma(X_n, n \leq 0)$ and the $k$-th component converges sufficiently quickly to zero—a property not guaranteed by ergodicity alone.

Detailed conditions like these specialize the theorem to forms of temporal dependence found in time series models used in empirical economics and econometrics. For example, it can be shown to hold for AR(1) and GARCH(1,1) models as presented in the previous chapter.

[9] Under certain conditions, it may be possible to recover some form of limit—specifically, a collection of accumulation points of the sequence, possibly in almost sure convergence—by multiplying with a slower diverging rate. The interested reader could study the concept of the Law of the Iterated Logarithm; see, e.g., [12] for the iid case. Clearly, due to Birkhoff's law and the continuous mapping theorem, multiplication by any convergent or bounded sequence would preserve convergence to 0.

**Example 13** (Classical CLT)**.** When the underlying sequence is iid, then for the second condition it is obtained that $\sum_{k=0}^{\infty} |\text{Cov}(\mathbb{E}(X_t \mid \sigma(X_n, n \leq 0)), X_k)|$ equals $\sum_{k=0}^{\infty} |\text{Cov}(\mathbb{E}(X_0), X_k)| = \sum_{k=0}^{\infty} 0 = 0$, hence the condition holds trivially. Furthermore, Var $\left(\frac{1}{\sqrt{n}} \sum_{t=1}^{n} (X_t - \mathbb{E}(X_0))\right) = \frac{1}{n} \sum_{t=1}^{n} \text{Var}(X_1) = \text{Var}(X_1)$. Hence, if Var$(X_1) > 0$, it is obtained that $\frac{1}{\sqrt{n}} \sum_{t=1}^{n} (X_t - \mathbb{E}(X_0)) \rightsquigarrow X \sim N(0, \text{Var}(X_1))$, which constitutes the classical iid CLT.

**Example 14** (Stationary Ergodic Square Integrable Martingale Difference Sequence CLT)**.** Suppose that $(\mathcal{F}_t)_{t \in \mathbb{Z}}$ is an increasing (i.e. $\mathcal{F}_t \subseteq \mathcal{F}_s$, $t < s$) sequence of information sets ($\sigma$-algebras). The pair $(X_t, \mathcal{F}_t)_{t \in \mathbb{Z}}$ is a martingale difference sequence iff $\mathbb{E}(X_t | \mathcal{F}_t) = 0$, $\forall t$. Suppose then that $(X_t, \mathcal{F}_t)_{t \in \mathbb{Z}}$ is a stationary ergodic martingale difference sequence, and $0 < \mathbb{E}(X_0^2) < +\infty$. Notice that in this case and due to LIE, $\mathbb{E}(X_0) = \mathbb{E}(\mathbb{E}(X_0 \mid \mathcal{F}_0)) = 0$. If for all $t$, $\sigma(X_{t-n}, n > 0)) \subseteq \mathcal{F}_t$, then due to the LIE $\mathbb{E}(X_t \mid \sigma(X_n, n \leq 0)) = \mathbb{E}(\mathbb{E}(X_t \mid \mathcal{F}_t) \mid \sigma(X_n, n \leq 0)) = \mathbb{E}(0 \mid \sigma(X_n, n \leq 0)) = 0$, and thereby $\sum_{k=0}^{\infty} |\text{Cov}(\mathbb{E}(X_t \mid \sigma(X_n, n \leq 0)), X_k)|$ equals $\sum_{k=0}^{\infty} |\text{Cov}(\mathbb{E}(X_0), X_k)| = \sum_{k=0}^{\infty} 0 = 0$, hence the covariance summability condition holds trivially. The previous also directly implies that $Cov(X_t, X_s) = 0$, $\forall t \neq s$. Thus, Var $\left(\frac{1}{\sqrt{n}} \sum_{t=1}^{n} X_t\right) = \frac{1}{n} \sum_{t=1}^{n} \text{Var}(X_1) = \text{Var}(X_1)$. Hence, $\frac{1}{\sqrt{n}} \sum_{t=1}^{n} X_t \rightsquigarrow X \sim N(0, \text{Var}(X_1))$, which constitutes a significant generalization of the previous classical iid CLT (why?).

The first result of the theorem informs us that the variance of the weighted partial sums, which takes the form of an arithmetic mean of the covariances between members of the sequence, converges. Both assumptions of the theorem play a role in this convergence. In certain cases of divergence, it is possible to recover the theorem by incorporating a rate that accounts for the aforementioned divergence. As mentioned above, if the limit of the variance is strictly positive, the weak limit obtained is the normal distribution with zero mean and variance given by the aforementioned series. If the variance limit were zero, the weak limit would instead be $\mathrm{Deg}_0$.

The third result describes something even more complex, encompassing the previous result as a special case. It concerns the asymptotic behavior of the truncated sum at $[n\tau]$, treated as a function of $\tau \in [0, 1]$. The partial sum thus becomes a stochastic process taking values in CADLAG real-valued functions over $[0, 1]$. This space is equipped with the Skorokhod metric-see, for instance, [2]-which modifies and generalizes the uniform metric, incorporating appropriate transformations between the involved functions. These transformations facilitate the study of compactness for the stochastic processes involved. Together with the first part of the theorem, this implies that the weak limit is the distribution of the (non-standardized version) Wiener process. The Wiener process-actually a Gaussian process-serves as the analogue of the normal distribution, acting as an attractor for such processes.

The above results are particularly useful in statistical inference, among other fields. Limit theorems like these can be employed to ascertain rates of convergence and asymptotic distributions—along with their asymptotic properties—of estimators and hypothesis tests in statistical models relevant to Empirical Economics and Econometrics, as discussed in the next part of the book. This demonstrates the utility of weak convergence: it is "weak" enough to enable the derivation of such properties.[10]

[10] We note for the interested reader that asymptotic theories like the one described by

# Addendum

## Asymptotic Tightness

The sequence of random elements $(X_n)_{n\in\mathbb{N}}$ is termed asymptotically tight, if there does not exist some $n$ such that the distribution $\mathbb{P}_n$ of $X_n$ assigns a fixed strictly positive probability to subsets of the underlying space that lie outside every compact subset of it. If $S$ is actually a Euclidean space, this is equivalent to that for any $\epsilon > 0$, there exists $M_\epsilon > 0$, such that $\lim_{n\to\infty} \mathbb{P}(\|X_n\| > M_\varepsilon) < \epsilon$. Asymptotically tight sequences have distributions that do not allow any fixed positive probability mass to be attributed to parts of the underlying space that "escape to infinity" (are not approximable by finitary constructions in the space).

It is not difficult to show that convergence in distribution (and thereby all the modes of convergence examined above-why?) implies asymptotic tightness. There exists a partial converse to this (which forms the interesting part of what is termed Prokhorov's Theorem): a sequence is asymptotically tight, iff every subsequence of it (i.e. an infinite part of it) has a further subsequence that converges in distribution.

the above theorem do not generally hold for other forms of convergence. These typically yield results in the form of laws of large numbers. Nevertheless, under certain conditions, it is possible to represent weak convergence as almost sure convergence: the underlying space is sufficiently augmented, and the involved random elements are extended appropriately to converge almost surely to limits having the distribution resulting from the weak convergence. This process pertains to the so-called Skorokhod representations, which are clearly beyond the scope of this book—see, for instance, [2]. However, it indicates that, in some cases, the convergence is essentially almost sure in suitably "richer" probability spaces.

> *Exersice*: If $(X_n)_{n\in\mathbb{N}}$ is asymptotically tight, and $h$ as in Theorem 5. Show that $(h(X_n))_{n\in\mathbb{N}}$ is asymptotically tight. Hint: combine Theorem 5 with the above characterization of Prokhorov's Theorem.

Asymptotic tightness, is directly related to concepts like the rates of convergence of estimators, the construction of confidence regions, etc, in statistics and econometrics.

> It is often convenient-particularly in proving asymptotic properties-to use concise expressions to denote asymptotic properties, such as convergence in probability or asymptotic tightness of a sequence of random variables. The expression $O_p(1)$ denotes that the sequence is asymptotically tight, while $o_p(1)$ denotes convergence in probability to zero. More generally, if $x_n$, $y_n$, and $z_n$ represent the general terms of related sequences, then $x_n = O_p(z_n)$ means $x_n = y_n z_n$ and $y_n = O_p(1)$. Similarly, $x_n = o_p(z_n)$ means $x_n = y_n z_n$ and $y_n = o_p(1)$. It is clear from the above definitions that $O_p(z_n) = z_n O_p(1)$, $o_p(z_n) = z_n o_p(1)$, and it is easy to prove properties like $o_p(1) + o_p(1) = o_p(1)$, $O_p(1) + O_p(1) = O_p(1)$, $o_p(1) + O_p(1) = O_p(1)$, $o_p(1)o_p(1) = o_p(1)$, $O_p(1)O_p(1) = O_p(1)$, $o_p(1)O_p(1) = o_p(1)$, and so on.

## Markov's Inequality, Tightness, and an iid LLN

Let $\mathbb{P}$ be a probability distribution on $\mathbb{R}$ along with a random variable $X \sim \mathbb{P}$, and $p, \varepsilon > 0$. From the properties of the integral, we have that:

$$\begin{aligned}
\mathbb{E}(|X|^p) = \int_{\mathbb{R}} |z|^p d\mathbb{P} = \int_{-\infty}^{-\varepsilon} |z|^p d\mathbb{P} + \int_{-\varepsilon}^{\varepsilon} |z|^p d\mathbb{P} + \int_{\varepsilon}^{+\infty} |z|^p d\mathbb{P} \\
\geq \int_{-\infty}^{-\varepsilon} \min_{z\in(-\infty,-\varepsilon]} |z|^p d\mathbb{P} + \int_{\varepsilon}^{+\infty} \min_{z\in[\varepsilon,+\infty)} |z|^p d\mathbb{P} \\
= |-\varepsilon|^p \int_{-\infty}^{-\varepsilon} d\mathbb{P} + |\varepsilon|^p \int_{\varepsilon}^{+\infty} d\mathbb{P} \\
= \varepsilon^p \left( \int_{-\infty}^{-\varepsilon} + \int_{\varepsilon}^{+\infty} \right) d\mathbb{P} = \varepsilon^p \left[ \mathbb{P}((-\infty,-\varepsilon]) + \mathbb{P}([\varepsilon,+\infty)) \right] \\
= \varepsilon^p \left[ \mathbb{P}((-\infty,-\varepsilon]) \cup [\varepsilon,+\infty)) \right],
\end{aligned}$$

where the first equality in the above follows from linearity, the first inequality from monotonicity, which also implies that $\int_{-\varepsilon}^{\varepsilon} |z|^p d\mathbb{P} \geq 0$, the second equality from the strict monotonicity of $z^p$ and the fact that its corresponding extrema are independent of $z$, and the last equality follows from the additivity of the distribution. Rearranging the above, we have essentially proven that:

$$\forall \varepsilon > 0,\ p > 0,\ \mathbb{P}((-\infty,-\varepsilon]) \cup [\varepsilon,+\infty)) = \mathbb{P}(|X| > \varepsilon) \leq \frac{\mathbb{E}(|X|^p)}{\varepsilon^p}. \tag{4}$$

The above is called Markov's inequality, and in the special case where $p = 2$, it is called Chebyshev's inequality. We observe that this inequality is trivial when $\varepsilon$ is very small, or when $\mathbb{E}(|X|^p) = +\infty$; in both cases the right-hand side is greater than one (even if it is not a real number). The inequality is informative on how the distribution assigns probabilities only if $\mathbb{E}(|X|^p)$ exists and $\varepsilon$ is large enough that the right-hand side of the inequality is less than one. In this case, the inequality informs us about the decay rate of the probability that the distribution assigns to extreme events. This decay rate is constrained by the absolute moment of highest order that exists and cannot decay faster than $C\varepsilon^{-p}$ as $\varepsilon \to \infty$ (when $p$ is a natural number, this is a polynomial rate in ${}^{1}\!/\!{}_{\varepsilon}$). The above seems to suggest that distributions such as the exponential and the Gaussian, with decay rates faster than the above for all $p$, possess moments of all orders.

Markov's inequality is an example of a class of inequalities that link probabilities to moment-related quantities (see, for example, Hoeffding's inequality and related inequalities, as detailed in [11] and [15]).

Suppose now that $(X_n)_{n\in\mathbb{N}}$ is a sequence of random variables, with a bounded sequence of absolute moments of order $p$, i.e., $\sup_n \mathbb{E}(|X_n|^p) < +\infty$, which is equivalent to that there exists $N > 0 : \mathbb{E}(|X_n|^p) \leq N, \forall n \in \mathbb{N}$. For arbitrary $\epsilon > 0$, consider $(\frac{N}{\epsilon})^{\frac{1}{p}} > 0$. Due to the inequality of Markov, for arbitrary $n$, $\mathbb{P}(|X_n| > (\frac{N}{\epsilon})^{\frac{1}{p}}) \leq \epsilon \frac{\mathbb{E}(|X_n|^p)}{N} \leq \epsilon$. Since the outermost r.h.s. of the previous inequality is independent of $n$, $\lim_{n\to\infty} \mathbb{P}(|X_n| > (\frac{N}{\varepsilon})^{\frac{1}{p}}) \leq \epsilon$, which implies the definition of asymptotic tightness above for $M_\epsilon := (\frac{N}{\epsilon})^{\frac{1}{p}}$. Hence any sequence of random variables with a bounded sequence of absolute moments of some order is asymptotically tight.

Suppose furthermore that $(X_n)_{n\in\mathbb{N}}$ is iid and that $p = 2$. Then, due to independence, for any $n > 0$, $\mathrm{Var}(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_n)$, and the latter, due to homogeneity equals $\frac{\mathrm{Var}(X_1)}{n}$. Remembering that due to the linearity of the integral and homogeneity, $\mathbb{E}(\frac{1}{n}\sum_{i=1}^{n} X_i)) = \mathbb{E}(X_1)$, then for any $n$, and $\varepsilon > 0$, Chebychev's inequality states that $\mathbb{P}(|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}(X_1)| > \varepsilon) \leq \frac{\mathrm{Var}(X_1)}{n\varepsilon^2}$.

Letting $n \to \infty$, the asymptotic version of the inequality forms

$$\forall \varepsilon > 0, \lim_{n \to \infty} \mathbb{P}(|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}(X_1)| > \varepsilon) = 0,$$

which is equivalent to $\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mathbb{E}(X_1)$. Hence, an LLN is obtained: if $(X_n)_{n \in \mathbb{N}}$ is iid with second moments, the sample mean converges in probability to the population mean. A bit more work, which would also utilize Markov's inequality, would imply the validity of the LLN in this context for $p = 1$. The stationary and ergodic LLN described in a previous section vastly generalizes this (in addition to establishing an LLN in the context of the stronger almost sure convergence).

## Jensen's Inequality

The Jensen inequality states that if $h : \mathbb{R} \to \mathbb{R}$ is concave, then $\mathbb{E}(h(X)) \geq h(\mathbb{E}(X))$, for any distribution on the real numbers $\mathbb{P}$ that has a first moment, with $X \sim \mathbb{P}$. The left-hand side of the inequality should be interpreted as $+\infty$ when the integral of $h$ does not exist. The dual version of the inequality pertains to the case where $h$ is convex, in which case $\mathbb{E}(h(X)) \leq h(\mathbb{E}(X))$. In the case where $h$ is also positive, the inequality, for instance, implies the integrability of $h$ when the mean of the distribution exists.

The inequality generalizes the property of linearity of the integral, as it holds as equality in the case where $h$ is linear, making it simultaneously convex and concave (remember that the concept of convexity is dual to concavity, $f$ being convex if and only if $-f$ is concave). The inequality can be proved using subdifferential calculus, which applies to these collections of functions-see, for instance, [9] and [10].

### Hölder's inequality

Hölder's inequality is a fundamental result in mathematical analysis and probability theory, particularly in the study of integrable random variables. It states that for any random variables $X$ and $Y$ and for $p, q > 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$, the following inequality holds:

$$\mathbb{E}[|XY|] \leq \left(\mathbb{E}[|X|^p]\right)^{1/p} \left(\mathbb{E}[|Y|^q]\right)^{1/q}.$$

Hölder's inequality generalizes the Cauchy-Schwarz inequality, which is recovered as the special case when $p = q = 2$. In this case, it simplifies to:

$$\mathbb{E}[|XY|] \leq \sqrt{\mathbb{E}[|X|^2]}\sqrt{\mathbb{E}[|Y|^2]}.$$

These inequalities are widely used in probability, statistics, and functional analysis, providing a powerful tool for bounding expectations and understanding relationships between random variables.

## Wold Device for Random Vectors

Let $(\mathrm{X}_n)_{n \geq 1}$ be a sequence of random vectors in $\mathbb{R}^d$, and let X be a random vector in $\mathbb{R}^d$. Then:

$$\mathbf{X}_n \rightsquigarrow \mathbf{X} \quad \text{if and only if} \quad \mathbf{a}^\top \mathbf{X}_n \rightsquigarrow \mathbf{a}^\top \mathbf{X}, \quad \forall \mathbf{a} \in \mathbb{R}^d.$$

The result follows from the fact that linear tranformations $\mathbf{a}^\top \mathbf{X}_n$ are Lipschitz continuous and bounded functions of $\mathbf{X}_n$, hence (when appropriately scaled) lie inside $BL_1(\mathbb{R}^d, \mathbb{R})$, and convergence in distribution is defined by the behavior of such functions, as noted in the main text.

## Applications

### 1. Multivariate Central Limit Theorem (CLT)

If $(\mathbf{X}_n)$ is an iid sequence of random vectors in $\mathbb{R}^d$ with mean $\mathbf{0}$ and covariance matrix $\Sigma$, then:

$$\sqrt{n}\mathbf{X}_n \rightsquigarrow N(\mathbf{0}, \Sigma).$$

Using the Wold Device, for any $\mathbf{a} \in \mathbb{R}^d$:

$$\mathbf{a}^\top(\sqrt{n}\mathbf{X}_n) \rightsquigarrow N(0, \mathbf{a}^\top \Sigma \mathbf{a}),$$

which confirms the multivariate convergence.

### 2. Linear Transformations

If $\mathbf{X}_n \rightsquigarrow \mathbf{X}$ and $A$ is a fixed matrix, then $A\mathbf{X}_n \rightsquigarrow A\mathbf{X}$. This follows directly from the Wold Device by considering $\mathbf{a}^\top A\mathbf{X}_n$ for all $\mathbf{a} \in \mathbb{R}^k$.

The Wold Device simplifies proving convergence in distribution for random vectors by reducing the problem to verifying the convergence of their real linear transformations.

## Delta Method for Weak Convergence

The delta method is a fundamental tool in asymptotic statistics, used to approximate the distribution of functions of estimators. Suppose we are given the weak convergence result:

$$r_n(M_n - M) \rightsquigarrow Z,$$

where:

- $M_n$ is a sequence of random $p$-vectors,
- $M$ is a non-random $p$-vector,
- $r_n$ is a scaling sequence (typically $r_n = \sqrt{n}$ in many applications),
- $Z$ is a random vector in $\mathbb{R}^p$ that represents the limiting distribution.

Let $F : \mathbb{R}^p \to \mathbb{R}^q$ be a continuously differentiable function. The delta method states that:

$$r_n\big(F(M_n) - F(M)\big) \rightsquigarrow \partial F(M)Z,$$

where $\partial F(M)$ is the $q \times p$ Jacobian matrix of $F$ evaluated at $M$.

### Explanation and Proof Outline

The result follows from an application of the Mean Value Theorem of $F(M_n)$ around $M$, justified by continuous differentiability:

$$F(M_n) = F(M) + \partial F(M_n^\star)(M_n - M),$$

where $M_n^\star$ is a random vector every realization of which lies in the line that connects $M$ with the analogous realization of $M_n$. Since $M_n \xrightarrow{p} M$ (why?), $M_n^\star \xrightarrow{p} M$, and due to the continuity of $\partial F$ and the CMT, $\partial F(M_n^\star) \xrightarrow{p} \partial F(M)$. Thus, collecting terms in the Mean Value expansion and multiplying by $r_n$, we have:

$$r_n\big(F(M_n) - F(M)\big) = \partial F(M_n^\star)\big(r_n(M_n - M)\big),$$

and the result follows by Slutsky's Lemma, the Wold device and the CMT.

## Applications

The delta method is widely used in asymptotic statistics for:

- Transforming estimators, such as using $F(M) = \log(M)$, $F(M) = M^2$, or higher-dimensional transformations.
- Establishing the asymptotic distribution of maximum likelihood estimators (MLEs) under transformations of parameters.
- Deriving the standard errors for non-linear functions of estimators.

## Example

Consider $M_n \sim N(\mu, \sigma^2/n)$, where $M_n$ is a sequence of estimators for $\mu$. If $r_n = \sqrt{n}$, then, trivially:

$$\sqrt{n}(M_n - \mu) \rightsquigarrow N(0, \sigma^2).$$

Let $F(M) = M^2$. Then:

$$\partial F(x) = 2x.$$

By the delta method:

$$\sqrt{n}\big(M_n^2 - \mu^2\big) \rightsquigarrow N(0, 4\mu^2\sigma^2).$$

## Further Exercises

1. Consider a sequence of random variables $(X_n)_{n\geq 1}$ such that $X_n \xrightarrow{p} X$ and $X_n \rightsquigarrow Y$. Prove or disprove: $X = Y$ almost surely.

2. Suppose $(X_n)_{n\geq 1}$ is a sequence of random variables such that $X_n \rightsquigarrow X$. Let $g : \mathbb{R} \to \mathbb{R}$ be a continuous function. Show that $g(X_n) \rightsquigarrow g(X)$ using the Continuous Mapping Theorem.

3. Consider $(X_n)_{n\geq 1}$ where $X_n = \mathbb{I}_{\{U_n > 1/n\}}$, with $U_n \sim$ Uniform$[0, 1]$. Show that $X_n \xrightarrow{p} 1$. Does $X_n \xrightarrow{\text{a.s.}} 1$? Justify your answer.

4. Consider the simple linear regression model:

$$Y_t = \beta_0 + \beta_1 X_t + \epsilon_t, \quad t = 1, \dots, n,$$

   where $(X_t, \epsilon_t)$ are i.i.d., $\mathbb{E}[\epsilon_t] = 0$, and Var$(\epsilon_t) = \sigma^2$. Show that the ordinary least squares (OLS) estimator $\hat{\beta}_1 = \frac{\sum_{t=1}^n X_t Y_t}{\sum_{t=1}^n X_t^2}$ satisfies $\hat{\beta}_1 \xrightarrow{p} \beta_1$.

5. Let $(X_n)_{n\geq 1}$ be a sequence of random variables such that $X_n \rightsquigarrow X$ and $Y_n \rightsquigarrow Y$. Assume $X_n$ and $Y_n$ are independent for all $n$. Show that $(X_n, Y_n) \rightsquigarrow (X, Y)$.

6. Let $(X_n)_{n\in\mathbb{N}}$ be a stationary and ergodic sequence of non-negative random variables such that $\mathbb{E}(X_1) < +\infty$. Derive the asymptotic behaviour

of $(\prod_{i=1}^{n} \exp(X_i))^{1/n}$.

# Epilogue

The modes of convergence of sequences of random elements (e.g., random variables, random vectors, stochastic processes, random functions) are extremely useful for the determination of asymptotic properties of sequences of inferential procedures (e.g. estimators, statistical tests, statistical forecasting procedures, etc.). Those properties are helpful in deciding what procedure to use for a given statistical problem, as well as in designing statistical methodologies with desired properties.

In what follows, we will use notions like almost sure convergence and convergence in probability to ascertain whether an estimator asymptotically locates the parameter value of interest. In such frameworks, tools like the Laws of Large Numbers, and Continuous Mapping Theorems, will be of importance. In many cases the estimators at hand will be defined as optimizers of criteria that somehow reflect the statistical/probabilistic information available to the researcher. Those will be random functions of the parameter of interest; when they appropriately converge in some of the aforementioned modes, to a limiting non-stochastic criterion that is uniquely optimized at the parameter value of interest, their optimizers will accordingly converge to the latter.

We will also use notions like asymptotic tightness, and convergence in distribution, in order to ascertain the rate at which an estimator converges to the parameter value of interest, as well as, the limiting distribution of its deviation from this value, scaled by the rate of convergence. Analogously, suchlike notions will be used in order to study the asymptotic behavior of test statistics under the hypotheses of interest. When the aforementioned criteria are implicated in the construction of the estimators and/or test statistics at hand, then local properties of the criteria will be thus useful. In such derivations Central Limit Theorems and Laws of Large Numbers, as well as tools like the Continuous Mapping Theorem and the Delta Method (see the Wald-type tests paragraph in the notes regarding OE) will be useful.

## References

[1] Herman J Bierens. *Robust methods and asymptotic theory in nonlinear econometrics*. Vol. 192. Springer Science & Business Media, 2012.

[2] Patrick Billingsley. *Probability and measure*. John Wiley & Sons, 2008.

[3] Anirban DasGupta. *Asymptotic theory of statistics and probability*. Vol. 180. Springer, 2008.

[4] James Davidson. *Stochastic limit theory: An introduction for econometricians*. OUP Oxford, 1994.

[5] Brendan McCabe and Andrew Tremayne. *Elements of modern asymptotic theory with statistical applications*. Manchester University Press, 1993.

[6] David Pollard. *A user's guide to measure theoretic probability*. 8. Cambridge University Press, 2002.

[7] Benedikt M Pötscher and Ingmar Prucha. *Dynamic nonlinear econometric models: Asymptotic theory*. Springer Science & Business Media, 1997.

[8] Emmanuel Rio. *Asymptotic theory of weakly dependent random processes*. Vol. 80. Springer, 2017.

[9] R Tyrrell Rockafellar. *Convex analysis*. Vol. 18. Princeton university press, 1970.

[10] R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*. Vol. 317. Springer Science & Business Media, 2009.

[11] Robert J Serfling. "Probability inequalities for the sum in sampling without replacement". In: *The Annals of Statistics* (1974), pp. 39–48.

[12] Volker Strassen. "An invariance principle for the law of the iterated logarithm". In: *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 3.3 (1964), pp. 211–226.

[13] Masanobu Taniguchi and Yoshihide Kakizawa. *Asymptotic theory of statistical inference for time series*. Springer Science & Business Media, 2012.

[14] AW van der Vaart and Jon A Wellner. *Weak convergence and empirical processes with applications to statistics*. Springer Series in Statistics, 1997.

[15] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. Vol. 47. Cambridge university press, 2018.