ECE 581: Convergence for Sequences of Random Variables

Henry D. Pfister
Duke University

November 16, 2025

Contents

1 Review: Convergence of Sequences

1.1 Real Numbers

A sequence of real numbers, $a_n \in \mathbb{R}$ for $n \in \mathbb{N}$ (or $a_1, a_2, a_3, \ldots \in \mathbb{R}$), converges to $a \in \mathbb{R}$ if, for every real $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that

$$|a_n - a| < \varepsilon \quad \text{for all } n \geq N.$$

We write $\lim_{n\to\infty} a_n = a$ or $a_n \to a$.

1.2 Vectors in Normed Vector Spaces

Let $(V, \|\cdot\|)$ be a normed vector space. Then, a sequence of vectors $v_n \in V$ (for $n \in \mathbb{N}$) converges to $v \in V$ if the real sequence $\|v_n - v\| \to 0$ or equivalently

$$\lim_{n\to\infty} \|v_n - v\| = 0.$$

Typical examples include:

- $\mathbb{R}^d$ with the Euclidean norm (e.g., think vector calculus),
- normed function spaces such as $L^p([a,b])$ (e.g., think Fourier analysis) with norm

$$\|f\|_p = \left(\int_a^b |f(t)|^p dt\right)^{1/p} \qquad p \in [1,\infty),$$

- the space of real random variables with $\|X\| = \sqrt{\mathbb{E}[X^2]}$ induced by $\langle X, Y\rangle = \mathbb{E}[XY]$.

The last two cases actually coincide in some cases. If we choose $a = 0$, $b = 1$, and $p = 2$, then the set of integrable functions mapping $[0,1]$ to $\mathbb{R}$ can be seen as the set of random variables for the probability space $\Omega = [0,1]$ with a uniform distribution on Ω. In that case, we have

$$\int_0^1 |X(\omega)|^2 \, d\omega = \mathbb{E}[X^2].$$

Another similarity is that, for functions and random variables, how one measures the distance can affect whether or not a sequence converges. Consider functions f_n, f mapping a common domain $\mathcal{X}$ to $\mathbb{R}$. Then, we have

- **Pointwise convergence:** $f_n(x) \to f(x)$ for all $x \in \mathcal{X}$.
- L^p **convergence:** $\int |f_n(x) - f(x)|^p dx \to 0$.
- **Uniform convergence:** $\sup_{x\in[0,1]} |f_n(x) - f(x)| \to 0$.

Example 1.1 (Pointwise but not L^2). Let $f(x) = x^{-1/2}$ for $(0,1]$ with $f(0) = 0$ and define

$$f_n(x) = \begin{cases} 0, & 0 \le x < 1/n, \\ x^{-1/2}, & 1/n \le x \le 1. \end{cases}$$

Then $f_n(x) \to f(x)$ for every $x \in [0,1]$, but

$$\|f_n - f\|_2^2 = \int_{0^+}^{1/n} \frac{1}{x}\, dx = +\infty,$$

so there is no convergence in the L^2 (or mean square) sense.

Example 1.2 (Fourier series). Let $f(x) = \mathbf{1}_{[0,0.5]}(x)$ be one period of a square wave and define

$$f_n(x) = \sum_{k=1}^{n} \frac{4}{(2k-1)\pi} \sin\big(2\pi(2k-1)x\big)$$

to be the first n non-trivial terms of its Fourier series expansion. While Fourier analysis shows that f_n converges to f in $L^2([0,1])$. It does not converge uniformly In fact, the well-known Gibb's Phenomenon is that the overshoot satisfies

$$\sup_{x\in[0,1]} |f_n(x) - f(x)| \to c \approx 0.0895.$$

Remark 1.3. One subtlety here is that we are ignoring exactly how these integrals are defined. Rather than the basic calculus Riemann integral, these spaces use the more general Lebesgue integral. Since the two integrals are equal whenever they both exist (e.g., for continuous functions), this difference is really only important for proofs.

1.3 A Nod in the Direction of Measure Theory

For completeness, we recall the definition of a probability space and introduce the notion of a measurable function.

Definition 1.4 (Probability space)**.** A probability space is defined by a tuple $(\Omega, \mathcal{F}, \mathbb{P})$ where Ω is a sample space, a subset of $E \subseteq \Omega$ is an event, $\mathcal{F}$ is a special collection of events known as a σ-algebra, and the probability function $\mathbb{P} : \mathcal{F} \to [0, 1]$ satisfies:

(A1) **Nonnegativity:** $\mathbb{P}(A) \geq 0$.

(A2) **Normalization:** $\mathbb{P}(\Omega) = 1$.

(A3) **Additivity:** for disjoint A_i, $\mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

Definition 1.5 (Measurable)**.** A function $X : \Omega \to \mathbb{R}$ (i.e., a real random variable) on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is called *measurable*[1] if $X^{-1}\big((a, b)\big) \in \mathcal{F}$ for all $a, b \in \mathbb{R}$.

The set of measurable functions can be understood as the set of well-defined random variables. The space of all real random variables (i.e., measurable functions) forms a vector space over $\mathbb{R}$. The subset with finite second moment also lives in the inner-product (or Hilbert) space of real random variables. In this class, we focus mainly on the Hilbert space formalism and less on its relationship to integration and measure.

2 Some Inequalities Useful for Limit Theorems

2.1 Markov Inequality

Lemma 2.1 (Markov's inequality)**.** *For a nonnegative random variable Y and all $\varepsilon > 0$,*

$$\boxed{\mathbb{P}(Y \geq \varepsilon) \leq \frac{E[Y]}{\varepsilon}}$$

Proof. Since $Y(\omega) \geq \varepsilon \, \mathbf{1}_{\{Y \geq \varepsilon\}}(\omega)$ for all $\varepsilon > 0$ and $\omega \in \Omega$, taking expectations yields

$$E[Y] \geq \varepsilon \, P(Y \geq \varepsilon),$$

which is equivalent to the claimed bound. □

[1] Actually, the stanadard definition uses the term "for all Borel sets" (which we have not defined) but our condition is equivalent for the real numbers.

2.2 Chebyshev Inequality

Lemma 2.2 (Chebyshev's inequality)**.** *Let X have mean μ and variance $\sigma^2 < \infty$. Then for all $\varepsilon > 0$,*

$$\boxed{\mathbb{P}(|X - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}}$$

Proof. Apply Markov's inequality to $Y = (X - \mu)^2$:

$$\mathbb{P}(|X - \mu| \geq \varepsilon) = \mathbb{P}(Y \geq \varepsilon^2) \leq \frac{E[Y]}{\varepsilon^2} = \frac{\sigma^2}{\varepsilon^2}.$$ □

2.3 Chernoff Bound

Lemma 2.3 (Chernoff Bound)**.** *Let X have moment generating function $M_X(s) = \mathbb{E}[e^{sX}]$. Then, for all $s \geq 0$, we have*

$$\mathbb{P}(X \geq t) \leq M_X(s)e^{-st}.$$

A convenient parametrization of this bound is given by

$$\mathbb{P}(X \geq t(s)) \leq M_X(s)e^{-st(s)},$$

where $t(s) = M'_X(s)/M_X(s)$.

Proof. For $s \geq 0$, the events $\{X \geq t\}$ and $\{e^{sX} \geq e^{st}\}$ are equal because the exponential function is strictly increasing. With this Markov's inequality, we get

$$\mathbb{P}(X \geq t) = \mathbb{P}(e^{sX} \geq e^{st})) \leq \frac{\mathbb{E}[e^{sX}]}{e^{st}}.$$

Since this holds for all t and $s \geq 0$, we can minimize bound over $s \geq 0$. Taking the natural log and minimizing over $s \geq 0$ results in the condition $t(s) = M'_X(s)/M_X(s)$. This optimality condition associates a unique t with each s and conveniently gives a closed form bound for $\mathbb{P}(X > t(s))$. □

3 Convergence of Random Variables

Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let $\{X_n\}$ be a sequence of real random variables with limit candidate X. Each convergence type below only makes sense when all random variables are defined on the same probability space, unless stated otherwise.

3.1 Mean–Square Convergence

We say that X_n converges to X in *mean square* (or $X_n \overset{L^2}{\to} X$) if

$$\lim_{n\to\infty} E[(X_n - X)^2] = 0.$$

Because $E[XY]$ defines an inner product on the space of real random variables with finite variance, mean–square convergence is simply convergence in the inner product space $L^2(\Omega, \mathcal{F}, P)$.

Theorem 3.1 (Weak Law of Large Numbers (mean–square form)). *Let $X_1, X_2, \ldots$ be i.i.d. with mean μ and variance $\sigma^2 < \infty$. Then the sample mean*

$$S_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

converges to μ in mean square.

Proof. This follow from

$$\mathbb{E}[(S_n - \mu)^2] = \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)\right)^2\right] = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{Cov}(X_i, X_j) = \frac{\sigma^2}{n} \to 0,$$

where $\mathrm{Cov}(X_i, X_j) = \sigma^2\delta_{i,j}$ due to independence. □

Interpretation. The empirical average of i.i.d. samples approaches the true mean as sample size grows. This convergence is the foundation of statistical estimation.

3.2 Convergence in Probability

We say $X_n \to X$ *in probability* (or $X_n \xrightarrow{p} X$) if, for all $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0.$$

Intuitively, the probability of any arbitrarily small deviation vanishes.

Lemma 3.2 (Mean–square implies in probability). *If $\mathbb{E}[(X_n - X)^2] \to 0$, then $X_n \xrightarrow{p} X$.*

Proof. For all $\varepsilon > 0$, Markov's inequality implies

$$\mathbb{P}(|X_n - X| > \varepsilon) = \mathbb{P}((X_n - X)^2 > \varepsilon^2) \le \frac{\mathbb{E}[(X_n - X)^2]}{\varepsilon^2} \to 0.$$

□

3.3 Almost-Sure Convergence

We say X_n converges to X *almost surely* (or $X_n \xrightarrow{a.s.} X$) if $A = \{\omega \in \Omega \mid X_n(\omega) \to X(\omega)\}$ satisfies

$$\mathbb{P}(A) = 1,$$

where "$X_n(\omega) \to X(\omega)$" means "the sequence of real numbers defined by $X_n(\omega)$ converges to the real number $X(\omega)$ for fixed $\omega \in \Omega$". The idea is that, for each outcome $\omega \in \Omega$, either $\lim_{n\to\infty} X_n(\omega)$ exists and equals $X(\omega)$ or it does not. Almost sure convergence means that the set of outcomes where the $X_n(\omega) \to X(\omega)$ has probability 1. This is stronger than convergence in probability.

Lemma 3.3 (Almost sure implies in probability). *If $X_n \to X$ almost surely, then $X_n \xrightarrow{p} X$.*

Proof. By assumption $X_n \xrightarrow{a.s.} X$, so $\mathbb{P}(A) = 1$ for $A = \{\omega \in \Omega \mid X_n(\omega) \to X(\omega)\}$. Fix any $\varepsilon > 0$. Then, for all $\omega \in A$, there is an $N(\omega) \in \mathbb{N}$ such that $|X_n(\omega) - X(\omega)| \le \varepsilon$ for all $n > N(\omega)$. Let

$$B_n = \{\omega \in \Omega \mid |X_n(\omega) - X(\omega)| > \varepsilon\}.$$

Then, for $\omega \in A$ and $n > N(\omega)$, we see that $\omega \in B_n^c$. Since $0 \leq \mathbf{1}_{B_n^c}(\omega) \leq 1$ and $\lim_{n\to\infty} \mathbf{1}_{B_n^c}(\omega) = \mathbf{1}_A(\omega)$ pointwise, the dominated convergence theorem[2] implies $\lim_{n\to\infty} \mathbb{P}(B_n^c) = \lim_{n\to\infty} \mathbb{E}[\mathbf{1}_{B_n^c}] = \mathbb{E}[\lim_{n\to\infty} \mathbf{1}_{B_n^c}] = \mathbb{P}(A)$. The argument is completed by noting that

$$\lim_{n\to\infty} \mathbb{P}(B_n) = 1 - \lim_{n\to\infty} \mathbb{P}(B_n^c) = 1 - \mathbb{P}(A) = 0. \qquad \square$$

Lemma 3.4 (Summable tail rates imply almost sure convergence)**.** *Let X_n be a sequence of random variables and let X be another random variable. Suppose that for every $\varepsilon > 0$,*

$$\sum_{n=1}^{\infty} \mathbb{P}(|X_n - X| > \varepsilon) < \infty.$$

Then $X_n \to X$ almost surely.

A proof at the right level for these notes can be found here: `https://www.math.ucdavis.edu/~tracy/courses/math135A/UsefullCourseMaterial/lawLargeNo.pdf`

3.4 Convergence in Distribution

We say that X_n converges to X *in distribution* (or $X_n \xrightarrow{d} X$) if

$$\boxed{\lim_{n\to\infty} F_{X_n}(x) = F_X(x) \quad \text{for all } x \text{ where } F_X \text{ is continuous}}$$

We note that convergence in distribution does not compare X_n and X as functions of Ω but instead compares the probability of certain events defined by them. Therefore, X_n and X are not required to be defined on the same probability space. It is one of many ways to compare the distributions of two random variables. See Appendix A for a discussion of other methods.

Example 3.5. Let $X_n \sim \text{Uniform}[0, 1/n]$. Then, $F_{X_n}(x) = 0$ if $x < 0$ and $F_{X_n}(x) = 1$ if $x \geq 1/n$. Hence $F_{X_n}(x) \to F_X(x)$, where $F_X(x) = 0$ for $x < 0$ and $F_X(x) = 1$ for $x \geq 0$. This implies that $X_n \to X$ in distribution where $X = 0$ almost surely.

3.5 Hierarchy of Convergences

a.s. convergence and uniformly bounded ⇒ mean–square convergence ⇒ in probability ⇒ in distribution,

Example 3.6 (Convergence modes). Here are a few sequences of random variables and convergence.

- Mean–square to a constant: Let $Z \sim \mathcal{N}(c, 1)$ and $X_n = Z/n$. Then $\mathbb{E}[X_n^2] = (1 + c^2)/n^2 \to 0$.
- In probability but not mean–square: Let $X_n = n$ with probability $1/n$ and 0 otherwise. Then $X_n \xrightarrow{p} 0$ but $\mathbb{E}[X_n^2] = n \not\to 0$.
- Almost surely but not mean–square: Let $U \sim \text{Uniform}(0, 1)$ and $X_n = \sqrt{n}\,\mathbf{1}_{\{U \leq 1/n\}}$. Then $X_n \to 0$ a.s., but $\mathbb{E}[X_n^2] = 1$ for all n.
- In distribution but not in probability: Let X_n be i.i.d. Uniform$(0, 1)$, and let X be an independent copy. Then, $X_n \xrightarrow{d} X$ but X_n does not converge to X in probability.

[2] The Dominated Convergence Theorem allows one to interchange the limit and expectation when the expected absolute value of all elements in the sequence is upper bounded by a constant.

4 Central Limit Theorem (CLT)

Theorem 4.1 (Central Limit Theorem)**.** *Let $X_1, X_2, \dots$ be i.i.d. random variables with mean μ and variance $\sigma^2 < \infty$. Then, we have*

$$\boxed{Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma} \xrightarrow{d} \mathcal{N}(0,1)}$$

and this implies that

$$\lim_{n\to\infty} \mathbb{P}(Z_n \le z) = \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt.$$

Outline of Proof. Due to normalization, we assume without loss of generality that $\mathbb{E}[X] = 0$ and $\mathrm{Var}(X) = 1$. While the stated result holds in general, we limit our proof to the case where $M_X(s) = E\big[e^{sX}\big]$ is finite for all $s \in [-\delta, \delta]$ for some $\delta > 0$.

Since e^{sx} equals its power series expansion for $s, x \in \mathbb{R}$, we see that $|x|^k \le \frac{k!}{s^k}(e^{sx} + e^{-sx})$ and

$$\mathbb{E}\big[|X|^k\big] \le \frac{k!}{\delta^k}(M_X(\delta) + M_X(-\delta)) < \infty.$$

Thus, all moments exist. Using the fact that $e^{sx} = \sum_{k=0}^{m} \frac{(sx)^k}{k!}$ is increasing in m, the monotone convergence theorem allows one to interchange integration and the infinite sum to see that

$$\mathbb{E}\big[e^{sX}\big] = \sum_{k=0}^{\infty} \frac{s^k}{k!} \mathbb{E}\big[X^k\big].$$

Since this sum converges for $s = \delta$, a standard result for power series expansions implies that it converges for all $|s| < \delta$. Thus, $M_X(s)$ has a continuous 3rd derivative for $|s| < \delta$ and we can apply Taylor's theorem with Lagrange remainder to $g(s) = \log M_X(s)$ about $s = 0$. It follows that there exist constants $C > 0$ and $\delta > 0$ such that for $|s| \le \delta$,

$$\log M_X(s) = \frac{s^2}{2} + R_3(s), \qquad |R_3(s)| \le C|s|^3.$$

Since the moment generating function (mgf) is multiplicative for the sum of independent random variables, we have

$$M_{Z_n}(s) = E\Big[e^{s\frac{1}{\sqrt{n}}\sum_i X_i}\Big] = \left(M_X\left(\frac{s}{\sqrt{n}}\right)\right)^n.$$

Therefore,

$$\log M_{Z_n}(s) = n \log M_X\left(\frac{s}{\sqrt{n}}\right) = \frac{s^2}{2} + n\, R_3\left(\frac{s}{\sqrt{n}}\right),$$

and the bound gives

$$\left| n\, R_3\left(\frac{s}{\sqrt{n}}\right)\right| \le \frac{C|s|^3}{\sqrt{n}} \xrightarrow[n\to\infty]{} 0.$$

Hence $\log M_{Z_n}(s) \to s^2/2$ and $M_{Z_n}(s) \to e^{s^2/2}$ for all s in a neighborhood of 0.

The final step uses Lévy's Continuity Theorem, a key result from probability theory which states that, if $M_{X_n}(s)$ converges pointwise to $M_X(s)$ for all s in a open interval containing 0, then $X_n \xrightarrow{d} X$. Since $M_{Z_n}(s) \to e^{s^2/2}$ for $|s| < \delta$, Z_n converges in distribution to a standard normal (i.e., $Z_n \xrightarrow{d} \mathcal{N}(0,1)$). □

Discussion

The CLT reveals that properly normalized sums of independent variables tend toward Gaussian behavior—regardless of the original distribution (subject to finite variance). This underlies statistical inference, confidence intervals, and error analysis in engineering.

5 Summary

- Deterministic convergence in normed spaces extends to random variables via L^p; for $p = 2$, random variables with finite variance form a Hilbert space with inner product $\langle X, Y\rangle = \mathbb{E}[XY]$.
- Tail bounds: Markov's and Chebyshev's inequalities control probabilities of large deviations and are key tools for proving convergence in probability from L^2 convergence.
- Modes of convergence defined and contrasted: mean–square (L^2), in probability, almost surely, and in distribution, with examples separating these modes.
- Implications: $L^2 \Rightarrow$ in probability; almost surely $\Rightarrow$ in probability; in probability $\Rightarrow$ in distribution.
- Weak Law of Large Numbers (mean–square form): for i.i.d. X_i with variance σ^2, the sample mean S_n satisfies $\mathbb{E}[(S_n - \mu)^2] = \sigma^2/n \to 0$.
- Central Limit Theorem: the normalized sum Z_n converges in distribution to $\mathcal{N}(0, 1)$; proof outline via mgf, Taylor expansion of $\log M_X(s)$, and the continuity theorem.
- Interpretation: empirical averages concentrate around the mean (LLN), and fluctuations about the mean are approximately Gaussian at scale $1/\sqrt{n}$ (CLT).

A Measuring the Distance Between Distributions

A.1 Total Variation Distance

Total variation (TV) distance quantifies how distinguishable two distributions are using the single best yes/no question. For probability distributions P and Q on the same measurable space $(\Omega, \mathcal{F})$, define

$$\|P - Q\|_{\text{TV}} \;=\; \sup_{A\in\mathcal{F}} |P(A) - Q(A)| \;=\; \frac{1}{2} \sup_{\|f\|_\infty \le 1} \big|\mathbb{E}_P[f] - \mathbb{E}_Q[f]\big|.$$

The first equality shows TV is the largest difference in answers to any indicator question $A \mapsto \mathbf{1}_A$; the second shows it is the largest gap in expectations over all bounded measurable tests f with range in $[-1, 1]$.

If P and Q both have well-defined PDFs (i.e., they admit densities p and q with respect to Lebesgue measure), then

$$\|P - Q\|_{\text{TV}} \;=\; \frac{1}{2} \int_{\mathcal{X}} |p(x) - q(x)|\, dx \;=\; 1 - \int_{\mathcal{X}} \min\{p(x), q(x)\}\, dx.$$

Thus, the TV distance is exactly half the L^1 distance between densities.

Optimal binary testing interpretation (Neyman–Pearson/Bayes with equal priors). Consider testing hypotheses $H_0 : P$ versus $H_1 : Q$ with equal priors and 0–1 loss. The minimum achievable error probability satisfies

$$P_e^\star \;=\; \frac{1 - \|P-Q\|_{\mathrm{TV}}}{2},$$

and the optimal test is the likelihood-ratio test. Equivalently, the maximum correct decision probability is $(1 + \|P-Q\|_{\mathrm{TV}})/2$. More generally, for a fixed random variable X defined by a bounded measurable f with $\|f\|_\infty \le 1$,

$$\big|\mathbb{E}_P[X] - \mathbb{E}_Q[X]\big| = \big|\mathbb{E}_P[f] - \mathbb{E}_Q[f]\big| \le 2\|P-Q\|_{\mathrm{TV}},$$

with equality in the supremum. Thus, TV distance is the true operational measure of distinguishability.

A.2 Convergence via Test Functions

A powerful way to express convergence in distribution uses expectations of test functions. By well-known results from probability, convergence in distribution $X_n \xrightarrow{d} X$ occurs if and only if

$$\lim_{n\to\infty} \mathbb{E}\big[f(X_n)\big] \;=\; \mathbb{E}\big[f(X)\big] \quad \text{for all bounded, continuous } f\colon \mathbb{R}\to\mathbb{R}.$$

This "weak" notion depends only on the distribution of the random variables and not on their joint construction.

If one drops the continuity requirement and allows all bounded measurable tests with $\|f\|_\infty \le 1$, then one finds that $X_n \xrightarrow{TV} X$ if and only if

$$\lim_{n\to\infty} \mathbb{E}\big[f(X_n)\big] \;=\; \mathbb{E}\big[f(X)\big] \quad \text{for all bounded, measurable } f\colon \mathbb{R}\to\mathbb{R}.$$

Thus, convergence in total variation implies convergence in distribution but this cannot be reversed.

A.3 Distance via Coupling

A coupling of distributions P and Q on a space $\mathcal{X}$ is any joint distribution π on $\mathcal{X}\times\mathcal{X}$ whose marginals are P and Q. If $(X,Y)\sim\pi$, we also say (X,Y) is a coupling of P and Q.

The total variation distance has the following coupling characterization

$$\|P-Q\|_{\mathrm{TV}} \;=\; \inf_{(X,Y)} \mathbb{P}\{X \ne Y\},$$

where the infimum is over all couplings (X,Y) of P and Q. There exists a maximal (optimal) coupling achieving this infimum, which directly realizes the operational meaning of TV as the minimum mismatch probability when drawing one sample from each distribution.

To describe the optimal coupling for total variation, we assume that P and Q have well-defined PDFs p and q (i.e., they admit densities with respect to Lebesgue measure). Let $r(x) \triangleq \min\{p(x), q(x)\}$ and define the overlap mass

$$m \;=\; \int r(x)\,dx \;=\; 1 - \|P-Q\|_{\mathrm{TV}}.$$

Decompose the residuals $p_{\mathrm{res}}(x) \triangleq p(x) - r(x)$ and $q_{\mathrm{res}}(x) \triangleq q(x) - r(x)$, which are nonnegative and integrate to $\|P-Q\|_{\mathrm{TV}}$. A maximal coupling (X,Y) with marginals P and Q is constructed as follows:

- With probability m, draw Z with density r/m and set $X = Y = Z$.
- With probability $\|P - Q\|_{\mathrm{TV}}$, draw X with density $p_{\mathrm{res}}/\|P - Q\|_{\mathrm{TV}}$ and Y with density $q_{\mathrm{res}}/\|P - Q\|_{\mathrm{TV}}$ independently.

Then, we have $P\{X \neq Y\} = \|P - Q\|_{\mathrm{TV}}$. This is optimal because any coupling satisfies $P\{X \neq Y\} \geq \|P - Q\|_{\mathrm{TV}}$.

Example A.1 (Bernoulli). Let $P = \mathrm{Bern}(p)$ and $Q = \mathrm{Bern}(q)$ on $\{0, 1\}$. Here $\|P - Q\|_{\mathrm{TV}} = |p - q|$ and $m = \min\{p, q\} + \min\{1 - p, 1 - q\}$. A maximal coupling is: with probability $\min\{p, q\}$, set $(X, Y) = (1, 1)$; with probability $\min\{1-p, 1-q\}$, set $(X, Y) = (0, 0)$; with the remaining probability $|p-q|$, put the mass on the unique mismatch: if $p > q$, set $(X, Y) = (1, 0)$ with probability $p - q$; if $q > p$, set $(X, Y) = (0, 1)$ with probability $q - p$. This achieves $P\{X \neq Y\} = |p - q| = \|P - Q\|_{\mathrm{TV}}$.

Wasserstein distances measure discrepancy with respect to a metric topology on $\mathcal{X}$. For a metric d on $\mathcal{X}$ and $p \in [1, \infty)$, the p-Wasserstein distance is

$$W_p(P, Q) \;=\; \inf_{\pi \in \Gamma(P,Q)} \left(\int_{\mathcal{X} \times \mathcal{X}} d(x, y)^p \, d\pi(x, y) \right)^{1/p},$$

where $\Gamma(P, Q)$ is the set of couplings of P and Q. For $p = 1$, the Kantorovich–Rubinstein duality yields

$$W_1(P, Q) \;=\; \sup_{\mathrm{Lip}_d(f) \leq 1} \left| \mathbb{E}_P[f] - \mathbb{E}_Q[f] \right|.$$

Here, $\mathrm{Lip}_d(f) = \sup_{x \neq y} \frac{|f(x) - f(y)|}{d(x,y)}$ is the Lipschitz constant of f relative to the metric d.

Relationships to convergence:

- If $W_1(P_n, P) \to 0$, then $P_n \xrightarrow{d} P$ (convergence in distribution), and first moments converge: $\int \|x\| \, dP_n(x) \to \int \|x\| \, dP(x)$.
- Conversely, if $\int \|x\| \, dP(x) < \infty$ and $P_n \xrightarrow{d} P$ with $\int \|x\| \, dP_n(x) \to \int \|x\| \, dP(x) < \infty$, then $W_1(P_n, P) \to 0$.

Thus W_1 gives a distance based interpretation of convergence in distribution that holds first moments are finite. Here are some examples and counterexamples.

- Simple example where $W_1 \to 0$: on $\{0, 1\}$ with $d(x, y) = |x - y|$, let $P_n = \mathrm{Bern}(p_n)$ and $P = \mathrm{Bern}(p)$. Then $W_1(P_n, P) = |p_n - p| \to 0$, so $P_n \xrightarrow{d} P$.
- Convergence in distribution but not in W_1: let $X_n = n$ with probability $1/n$ and 0 otherwise, and let $X \equiv 0$. Then $X_n \xrightarrow{d} X$, but $W_1(\mathcal{L}(X_n), \delta_0) \geq |\mathbb{E}[X_n] - \mathbb{E}[X]| = 1$, so $W_1 \not\to 0$ (first moments do not converge).
- Weak convergence plus first-moment convergence implies $W_1 \to 0$: if $X_n \xrightarrow{d} X$ and $\mathbb{E}[d(X_n, x_0)] \to \mathbb{E}[d(X, x_0)] < \infty$ for some x_0, then $W_1(\mathcal{L}(X_n), \mathcal{L}(X)) \to 0$. For instance, on $\mathbb{R}$ with $d(x, y) = |x - y|$, $X_n \sim \mathcal{N}(0, 1 + 1/n)$ converge in distribution to $\mathcal{N}(0, 1)$, their first absolute moments converge, and hence $W_1 \to 0$.