

Lecture 9: Convergence, laws of large numbers, and consistency

Lecturer: Yiqiao Zhong[Report an issue](#)*Last update: October 19, 2026*

These notes support STAT 709 at the University of Wisconsin–Madison. Use them alongside the cited textbooks; corrections and questions are welcome.

High priority *Asymptotic properties of estimators*

Overview. Lecture 8 introduced methods for constructing estimators. This lecture studies conditions under which their estimation errors vanish as the sample size increases. The main tools are convergence in probability, stochastic orders, and laws of large numbers. Laws of large numbers establish convergence of empirical averages to population expectations; continuity and uniform convergence then yield consistency results for moment estimators and minimizers of sample objectives. We distinguish convergence at a fixed argument from uniform convergence over a family of arguments.

We first develop the weak law and use it to justify moment estimators. The optional strong-law and uniform-law sections develop almost-sure and uniform convergence. The final applications concern regression and maximum likelihood, including a uniform-endpoint estimator with a rate faster than $n^{-1/2}$.

Mid priority *Notation and conventions*

We write $\mathbb{E}X$ or $\mathbb{E}[X]$ for expectation and $\mathbb{P}(A)$ for the probability of an event. Uppercase letters usually denote random variables and lowercase letters their possible values; $\mathbb{R}$ denotes the real numbers and $\mathbf{1}_A$ the indicator of A . This follows the probability notation used by Shao and many statistics texts. Machine-learning papers often make different choices, such as bold vectors and a distinction between random variables and their realizations by font. The Deep Learning notation guide is a useful comparison. Check the notation conventions when comparing formulas. Throughout, limits are as $n \rightarrow \infty$, and vectors have a fixed finite dimension unless we explicitly say otherwise.

This begins Part III. The part overview connects these lectures and includes a final self-check.

Recap from earlier lectures

From Lectures 1, 3, 4 and 8: estimator recipes and limits

Reminder only: this block introduces no new exam requirements.

- **Moment and likelihood estimators.** Sample moments are matched to model moments; likelihood is maximized over the parameter. Used in Section 7.
- **Population likelihood.** The KL inequality identifies the truth as a population likelihood maximizer. Used in optional Section 7.3.
- **Markov and Chebyshev.** Moment bounds turn into bounds on error probabilities. Used in Section 4.1.
- **Dominated convergence.** An integrable bound permits the limit to pass through an integral. Used in Section 4.2.
- **Compactness.** A continuous function attains its extrema on a nonempty compact set. Used in Section 7.2.

Learning goals

- Distinguish probability limits, almost-sure limits, and stochastic rates.
- Prove a weak law by variance control and apply the integrable weak law.
- Check the assumptions behind consistency of moments and minimizers.
- Explain why pointwise convergence alone does not control optimization.

Priority guide

High Core concepts and calculations to master; directly examined.

Mid Supporting concepts useful in other lectures; in scope for examinations.

Low Enrichment supporting understanding; will not be examined.

Further reading Optional self-study, references, extensions, and historical notes.

A star (*) marks optional material; priority labels give the scope of each block. Lecture-specific exemptions and the optional status of Part VI override the general exam guide.

1 Exact inference and the asymptotic approach

Mid priority *Motivation for asymptotic analysis*

At a fixed sample size, we study the error of an estimator relative to its population target. Exact calculations of bias or variance are available in some models, but can be difficult in others.

An **asymptotic approach** studies an estimator as the sample size increases. A basic question is whether its error becomes small with high probability.

The sample mean is the starting example. A law of large numbers gives conditions under which this average converges to the population mean, justifying estimation by empirical averages.

Instructor's Note *Asymptotic approximation and model complexity*

Think of increasing the sample size as turning up the resolution of a picture. First we ask whether the estimate settles near the right target; then we can zoom in on the remaining fluctuations. The LLN addresses the first question; the central limit theorem (CLT) often answers the second. Exact calculations, such as the usual t distribution under a Normal model, give another way to study sampling error.

The picture also changes if we add more unknown parameters as we collect more observations. The regression examples in Lecture 7 illustrate this balance between sample size and model complexity. Classical asymptotics often holds dimension fixed; random matrix theory and empirical-process methods help study richer settings.¹ The relative growth of data and dimension matters.

Other methods address uncertainty at the observed sample size: finite-sample corrections and Edgeworth expansions refine an approximation,² the bootstrap creates new samples by drawing with replacement from the observed data, and Bayesian modeling brings domain knowledge into the analysis. Subsampling is a related but different resampling idea. Think of these as different tools for understanding uncertainty. For now, our task is simply to understand why averages stabilize.

¹ An empirical process describes the sampling errors of a family of empirical averages, for example all errors $F_n(x) - F(x)$ as x varies. Its theory includes bounds on these errors over the whole family.

² An Edgeworth expansion adds correction terms to a leading Normal approximation using information such as skewness. Its validity requires additional assumptions; it is not an exact finite-sample formula.

2 Convergence in probability

2.1 Modes of stochastic convergence

High priority *Almost-sure convergence and convergence in probability*

Let $X, X_1, X_2, \dots$ be random vectors in $\mathbb{R}^k$ on a common probability space, with Euclidean norm $\|\cdot\|$.

Definition 1. The sequence converges **almost surely** to X , written $X_n \rightarrow X$ a.s., if

$$\mathbb{P}\{\omega : \|X_n(\omega) - X(\omega)\| \rightarrow 0\} = 1.$$

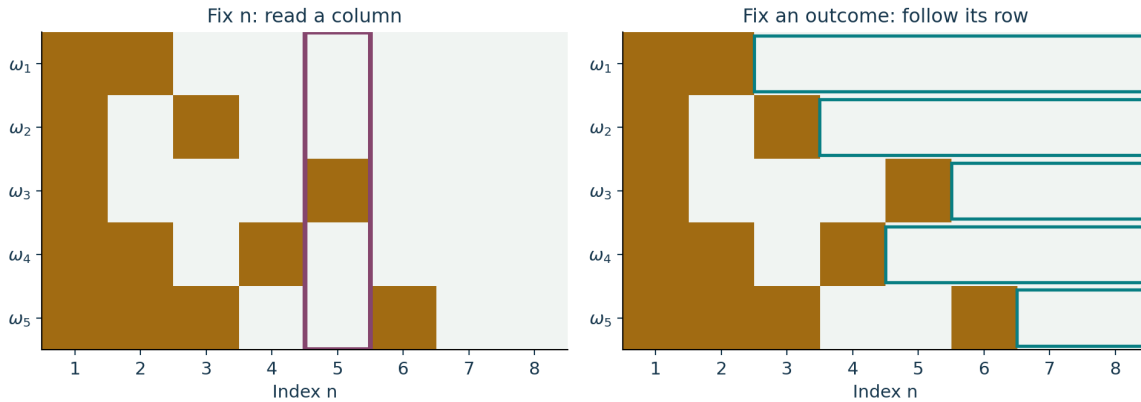
Definition 2. The sequence converges **in probability** to X , written $X_n \xrightarrow{P} X$, if for every $\varepsilon > 0$,

$$\mathbb{P}(\|X_n - X\| > \varepsilon) \rightarrow 0.$$

Convergence to a constant in probability can also be defined when each X_n lives on its own probability space. Almost-sure convergence concerns entire sample paths and therefore requires a common probability space.

Instructor's Note *Marginal probabilities and sample-path behavior*

Intuitively, convergence in probability looks at one photograph for each sample size: what fraction of experiments are still far from the target? Almost-sure convergence watches the whole movie of one experiment as more observations arrive. Along almost every movie, the errors eventually stay inside every fixed tolerance. The random variables have to “work together” across time for this second statement; small error probabilities in separate photographs do not tell the whole story.



Columns compare outcomes at a fixed index; rows follow one outcome through time. Gold cells mark errors above a fixed tolerance. This finite schematic illustrates the quantifiers, not a limit theorem.

High priority *Almost-sure convergence implies convergence in probability*

Fix $\varepsilon > 0$ and put

$$A_{n,\varepsilon} = \{\|X_n - X\| \leq \varepsilon\}, \quad B_{N,\varepsilon} = \bigcap_{n \geq N} A_{n,\varepsilon}.$$

If $X_n \rightarrow X$ a.s., then $B_{N,\varepsilon}$ increases with N and $\mathbb{P}(\bigcup_N B_{N,\varepsilon}) = 1$. Continuity of probability from below and $B_{N,\varepsilon} \subset A_{N,\varepsilon}$ give

$$\mathbb{P}(B_{N,\varepsilon}) \rightarrow 1 \implies \mathbb{P}(A_{N,\varepsilon}) \rightarrow 1. \quad (1)$$

This proves convergence in probability. More explicitly, the pathwise convergence event can be written using the countable tolerances $1/m$ as $\bigcap_{m \geq 1} \bigcup_N B_{N,1/m}$; this also explains its measurability. The converse implication between the two modes is false: see Exercise 7 and its complete counterexample.

High priority *First Borel–Cantelli lemma and almost-sure convergence*

Proposition 1 (First Borel–Cantelli lemma). *If events E_n satisfy $\sum_n \mathbb{P}(E_n) < \infty$, then with probability one only finitely many E_n occur. No independence is needed.*

Indeed,

$$\mathbb{P}\left(\bigcup_{n \geq N} E_n\right) \leq \sum_{n \geq N} \mathbb{P}(E_n) \rightarrow 0.$$

The events on the left decrease to the event that infinitely many E_n occur, so continuity from above proves the claim. In particular, if $\sum_n \mathbb{P}(\|X_n - X\| > \varepsilon) < \infty$ for every $\varepsilon > 0$, apply the proposition at $\varepsilon = 1/m$ and intersect the resulting probability-one events to obtain $X_n \rightarrow X$ a.s. Exercise 8 gives a second proof by counting errors and interchanging a nonnegative sum with expectation.

2.2 Stochastic orders

High priority *Stochastic orders and Gaussian examples*

For deterministic positive a_n , the familiar notation $x_n = o(a_n)$ means $x_n/a_n \rightarrow 0$, while $x_n = O(a_n)$ means that $|x_n|/a_n$ is eventually bounded. Random errors need probabilistic versions.

Definition 3. *For deterministic $a_n > 0$, write $X_n = o_P(a_n)$ if $\|X_n\|/a_n \xrightarrow{P} 0$. Write $X_n = O_P(a_n)$, or say that X_n/a_n is **bounded in probability**, if for every $\eta > 0$ there are finite M and N such that*

$$\sup_{n \geq N} \mathbb{P}(\|X_n\| > Ma_n) < \eta.$$

Thus $o_P(1)$ means convergence in probability to zero; $O_P(1)$ permits persistent random variation. For finite-valued X_n and positive a_n , one may equivalently take the supremum over all $n \geq 1$: enlarge M to control the finitely many omitted indices. The distinction between a small random error and an error that stays on a fixed scale is essential.

Example 1. *For $X_n \sim N(0, \sigma_n^2)$, bounded σ_n gives $O_P(1)$, even if X_n does not converge to zero. If $\sigma_n \rightarrow 0$, then $X_n = o_P(1)$; if $\sigma_n \rightarrow \infty$, it is not $O_P(1)$. No particular dependence between the X_n is needed for these claims.*

Exercise prompt. Use a standard Normal variable to justify the three scale claims in the example. Which one allows a nonvanishing distribution of errors? See Exercise 1 in Section 10 and its complete calculation.

Mid priority *Random scaling and the product rule for stochastic orders*

Random-scale notation can also be useful. If $\mathbb{P}(|Y_n| > 0) = 1$ for every n , define $X_n = o_P(Y_n)$ and $O_P(Y_n)$ by applying the preceding definitions to $\|X_n\|/|Y_n|$. If a denominator can be zero, first specify how that event is handled; an undefined ratio is not a stochastic-order statement.

For deterministic positive scales, $o_P(a_n)$ implies $O_P(a_n)$. Also, for real random variables,

$$U_n = o_P(a_n), \quad V_n = O_P(b_n) \quad \implies \quad U_n V_n = o_P(a_n b_n).$$

There is **no independence assumption**. A useful proof splits according to whether $|V_n|/b_n$ exceeds a large constant; Exercise 6 gives the complete argument.

3 Consistency of estimators

High priority Consistency and rates of convergence

Let $X^{(n)} = (X_1, \dots, X_n)$ be the data, with a distribution determined by P in a model $\mathcal{P}$, and let $\nu(P) \in \mathbb{R}^k$ be the estimand. An estimator $T_n = T_n(X^{(n)})$ is a measurable function of the observations.

Definition 4. The sequence T_n is **consistent** for ν if, for every fixed $P \in \mathcal{P}$, $T_n \xrightarrow{P} \nu(P)$ under P . For deterministic $a_n > 0$ with $a_n \rightarrow \infty$, it is **a_n -consistent** if $a_n\{T_n - \nu(P)\} = O_P(1)$ for every fixed P . It is **strongly consistent** if $T_n \rightarrow \nu(P)$ a.s. under each P , on a common space carrying the data sequence.

Recall the earlier definition of stochastic orders when interpreting the rate. The quantifier “for every fixed P ” describes **pointwise consistency**; it does not assert that the worst error probability over all $P \in \mathcal{P}$ tends to zero. A stochastic rate is an upper bound on the error scale, not a claim that this is the best possible rate. For example, $\sqrt{n}$ -consistency says the error is $O_P(n^{-1/2})$. It neither says that the error is eventually bounded by $n^{-1/2}$ on every sample path nor supplies a limiting distribution.

Exercise prompt. Why does a_n -consistency with $a_n \rightarrow \infty$ imply consistency? Start from the definition of boundedness in probability. See Exercise 2 in Section 10 and its complete proof.

Instructor’s Note Interpretation of consistency and convergence rates

Imagine collecting larger data sets and applying the same estimation procedure each time. Consistency asks whether the estimates concentrate near the population target. A convergence rate describes the scale of the remaining error. Rescaling the errors lets us compare their magnitudes at different sample sizes. This thought experiment helps us understand a procedure even when only one data set is available.

Mid priority Asymptotic and finite-sample guarantees

A finite-sample guarantee might say that, for every $n \geq 3$, an error bound holds with probability at least $1 - 2/n$. Such a statement can imply consistency when its error bound tends to zero, and it also tells us how the guarantee depends on the actual n . Asymptotic statements instead emphasize the eventual behavior, often with simpler formulas.

Instructor’s Note Asymptotic limits and sample-size requirements

Informally, many statistics arguments ask, “What happens as data become abundant?” Many theoretical computer-science arguments ask, “How many samples are enough for this accuracy and failure probability?” It is useful to understand both perspectives. A limit describes the behavior as sample size increases; a finite-sample bound quantifies the sample size needed for a specified accuracy.

Consistency alone does not compare finite-sample risk. The shrinkage and bias–variance examples in Lecture 7 already show why we also care about the magnitude of error at the sample size we have.

4 Laws of large numbers

4.1 Weak law under bounded variances

High priority Weak law for uncorrelated observations

Theorem 1. Suppose $X_1, X_2, \dots$ are pairwise uncorrelated real random variables, $\mathbb{E} X_i = \mu$, and $\text{Var}(X_i) \leq C < \infty$. Then $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i \xrightarrow{P} \mu$.

Proof. Uncorrelatedness eliminates the cross-covariances, so

$$\mathbb{E}(\bar{X}_n - \mu)^2 = \text{Var}(\bar{X}_n) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) \leq \frac{C}{n}. \quad (2)$$

The right side tends to zero: this is L^2 **convergence**. Chebyshev's inequality below gives $\mathbb{P}(|\bar{X}_n - \mu| > \varepsilon) \leq C/(n\varepsilon^2) \rightarrow 0$. $\square$

Theorem 2 (Chebyshev's inequality). *If $\mathbb{E}Y^2 < \infty$ and $\varepsilon > 0$, then*

$$\mathbb{P}(|Y - \mathbb{E}Y| > \varepsilon) \leq \frac{\mathbb{E}[(Y - \mathbb{E}Y)^2]}{\varepsilon^2} = \frac{\text{Var}(Y)}{\varepsilon^2}.$$

It follows immediately from Markov's inequality applied to $(Y - \mathbb{E}Y)^2$. In particular, $\mathbb{E}|Y_n - Y|^2 \rightarrow 0$ implies $Y_n \xrightarrow{P} Y$ by the same argument, without requiring the errors to have mean zero. For i.i.d. variables with variance $\sigma^2 < \infty$, the bound also gives $\sqrt{n}(\bar{X}_n - \mu) = O_P(1)$.

Mid priority *Moment inequalities and integrability conditions*

On a probability space, if $0 < p < q$ and $R = \|X\|$ for a fixed norm, then

$$(\mathbb{E}R^p)^{1/p} \leq (\mathbb{E}R^q)^{1/q}.$$

When $\mathbb{E}R^q < \infty$, Jensen's inequality for the concave function $u \mapsto u^{p/q}$ proves this; if the right side is infinite, the inequality holds in the extended real sense. Thus a finite higher moment implies a finite lower moment. This compares moments of the same nonnegative variable, not different coordinate norms of a vector.

Another useful viewpoint comes from functional analysis: square-integrable random variables belong to $L^2(\Omega)$, with $\|Y\|_{L^2} = (\mathbb{E}|Y|^2)^{1/2}$ (identifying variables equal a.s.). Thus (2) states $\|\bar{X}_n - \mu\|_{L^2} \rightarrow 0$. For bounded R , the moment comparison also has the endpoint $(\mathbb{E}R^p)^{1/p} \leq \text{esssup } R$, the smallest a.s. upper bound.

A finite lower moment does not imply a finite higher one. A Pareto variable with tail $\mathbb{P}(X > t) = t^{-\alpha}$ for $t \geq 1$ and $1 < \alpha < 2$ has finite mean and infinite second moment; Exercise 9 contains the calculation. The variance-based proof does not apply to this example. A weak law remains available under a finite first absolute moment.

Picture: accumulating moment areas. For this Pareto example, which accumulated moment approaches a finite limit as the cutoff increases, and which grows without bound? The tail-integral identity gives, for $R \geq 1$,

$$A_1(R) = \int_0^R \mathbb{P}(X > t) dt = \mathbb{E} \min(X, R), \quad A_2(R) = \int_0^R 2t \mathbb{P}(X > t) dt = \mathbb{E}[\min(X, R)^2].$$

Both integrals contribute area 1 on $[0, 1]$. On $[1, R]$, integrate $t^{-\alpha}$ and $2t^{1-\alpha}$ to obtain

$$A_1(R) = 1 + \frac{1 - R^{1-\alpha}}{\alpha - 1}, \quad A_2(R) = 1 + \frac{2(R^{2-\alpha} - 1)}{2 - \alpha}.$$

These are **capped moments**; capping at R differs from setting observations above R to zero in the truncation argument below. For $\alpha = 3/2$, the following values give a static comparison.

Cutoff R	$A_1(R)$	$A_2(R)$
1	1	1
10	$3 - 2/\sqrt{10} \approx 2.368$	$4\sqrt{10} - 3 \approx 9.649$
100	2.8	37

The first area approaches $\alpha/(\alpha - 1)$; the second grows without bound. A finite picture shows only finite areas. The complete infinite-moment calculation is in Exercise 9 and its solution.

4.2 Truncation and a finite first moment

Mid priority *Weak law with truncated centering*

Theorem 3. Let $X_1, X_2, \dots$ be i.i.d. real random variables. Assume $n\mathbb{P}(|X_1| > n) \rightarrow 0$ and define

$$a_n = \mathbb{E}[X_1 \mathbf{1}_{\{|X_1| \leq n\}}].$$

Then

$$\frac{1}{n} \sum_{j=1}^n X_j - a_n \xrightarrow{P} 0. \quad (3)$$

Low priority *Proof by truncation and variance control*

Proof (optional). Put $Y_{n,j} = X_j \mathbf{1}_{\{|X_j| \leq n\}}$, $S_n = n^{-1} \sum_{j=1}^n X_j$, and $T_n = n^{-1} \sum_{j=1}^n Y_{n,j}$. Within each row n , the $Y_{n,j}$ are i.i.d. and $\mathbb{E} T_n = a_n$. The rows need not be independent. By the union bound,

$$\mathbb{P}(S_n \neq T_n) \leq \mathbb{P}\left(\bigcup_{j=1}^n \{|X_j| > n\}\right) \leq n\mathbb{P}(|X_1| > n) \rightarrow 0. \quad (4)$$

The cutoff includes observations equal to n in absolute value; the discarded event is therefore $|X_j| > n$, also for distributions with atoms. For $Z = |X_1|$, the nonnegative tail-integral identity from Lecture 4 gives

$$\mathbb{E} Y_{n,1}^2 \leq \mathbb{E}[\min(Z, n)^2] = 2 \int_0^n t \mathbb{P}(Z > t) dt.$$

The integer tail condition implies $t\mathbb{P}(Z > t) \rightarrow 0$ for real t : with $m = \lfloor t \rfloor \geq 1$, it is at most $(m+1)\mathbb{P}(Z > m)$. If $g(t) = t\mathbb{P}(Z > t)$, then for any $\delta > 0$ choose R so that $g(t) \leq \delta$ for $t \geq R$. Since $g(t) \leq t$,

$$\frac{1}{n} \int_0^n g(t) dt \leq \frac{1}{n} \int_0^R g(t) dt + \delta \quad (n \geq R).$$

The limsup is at most δ . As δ is arbitrary, the integral average tends to zero. The earlier Chebyshev inequality now yields

$$\mathbb{P}(|T_n - a_n| > \varepsilon) \leq \frac{\text{Var}(Y_{n,1})}{n\varepsilon^2} \leq \frac{2}{n\varepsilon^2} \int_0^n g(t) dt \rightarrow 0.$$

Finally $\mathbb{P}(|S_n - a_n| > \varepsilon)$ is at most the last probability plus $\mathbb{P}(S_n \neq T_n)$, proving the result. $\square$

High priority *Sample mean converges to population mean*

Theorem 4 (Weak LLN for integrable i.i.d. observations). If the X_i are i.i.d. and $\mathbb{E}|X_1| < \infty$, then $\bar{X}_n \xrightarrow{P} \mu = \mathbb{E} X_1$.

Exercise prompt. Use dominated convergence twice to show that $n\mathbb{P}(|X_1| > n) \rightarrow 0$ and $a_n \rightarrow \mu$. How does Theorem 3 then prove the weak LLN? See Exercise 3 in Section 10 and its complete proof.

Instructor's Note *Variance reduction and truncation*

Think of averaging as a noise-canceling operation: centered cross terms cancel in expectation, so the average has much less fluctuation than an individual observation. Rare large observations can dominate the sample mean. Truncation temporarily sets those observations aside. We choose the cutoff so that discarding any observation is unlikely, while the variance of the average of n truncated observations tends to zero. We then recover the original average because it agrees with the truncated average with high probability.

Mid priority *Weak law for random vectors*

If $X_i \in \mathbb{R}^k$ are i.i.d. with $\mathbb{E} \|X_1\| < \infty$, each coordinate is integrable. Coordinatewise weak LLNs and the union bound give

$$\mathbb{P}(\|\bar{X}_n - \mathbb{E} X_1\| > \varepsilon) \leq \sum_{j=1}^k \mathbb{P}(|\bar{X}_{n,j} - \mathbb{E} X_{1,j}| > \varepsilon/\sqrt{k}) \rightarrow 0.$$

Here k is fixed. Increasing k changes both the number of error events and their tolerances, so this proof does not by itself give a high-dimensional law.

5 Strong laws and tail events (optional)

Low priority *Tail events and Kolmogorov's zero-one law*

For real $X_1, X_2, \dots$, put

$$\mathcal{F}'_m = \sigma(X_m, X_{m+1}, \dots), \quad \mathcal{T} = \bigcap_{m \geq 1} \mathcal{F}'_m.$$

The notation $\sigma(X_m, X_{m+1}, \dots)$ means the smallest sigma-algebra making those variables measurable. An event in $\mathcal{T}$ is a **tail event**. For instance,

$$\left\{ \limsup_{n \rightarrow \infty} \frac{X_1 + \dots + X_n}{n} \geq t \right\} \in \mathcal{T}.$$

For each fixed m , deleting $X_1, \dots, X_{m-1}$ changes the average by a finite random sum divided by n , which tends to zero on every path. The limsup event can therefore be expressed using only the variables with indices at least m .

Theorem 5 (Kolmogorov's zero-one law). *If the X_i are independent, every $A \in \mathcal{T}$ has probability zero or one.*

Proof. Such an A is independent of $\sigma(X_1, \dots, X_m)$ for every m . The union of these finite-prefix sigma-algebras is a generating π -system for $\mathcal{F}_\infty = \sigma(X_1, X_2, \dots)$. The events B satisfying $\mathbb{P}(A \cap B) = \mathbb{P}(A) \mathbb{P}(B)$ form a λ -system. The π - λ theorem extends the equality to every $B \in \mathcal{F}_\infty$. Taking $B = A$ gives $\mathbb{P}(A) = \mathbb{P}(A)^2$, hence the conclusion. $\square$

Instructor's Note *Tail-event invariance under finite changes*

A tail event asks a question about the distant future. Changing the opening scene of the movie cannot change its answer. For independent observations, the zero-one law says that such a question has an all-or-nothing answer. This makes a strong law feel plausible: a limiting average is a feature of the long run. We still need to determine which of the two probabilities, zero or one, applies to the event under study.

In particular, the deviation events $\{\limsup_n(\bar{X}_n - \mu) > \varepsilon\}$ and $\{\liminf_n(\bar{X}_n - \mu) < -\varepsilon\}$ must be proved to have probability *zero*. A weak law and the zero-one law alone do not prove this: the counterexample in Exercise 7 has convergence in probability but infinitely many large errors a.s.

Low priority *Strong law with bounded deterministic weights*

Theorem 6. *If X_i are i.i.d., $\mathbb{E}|X_1| < \infty$, and $1 \leq c_i \leq C$ are deterministic, then*

$$\frac{1}{n} \sum_{i=1}^n c_i (X_i - \mathbb{E} X_1) \longrightarrow 0 \quad a.s.$$

In particular, $\bar{X}_n \rightarrow \mathbb{E} X_1$ a.s. when $c_i = 1$.

The proof retains the truncation idea, but must control the whole sequence. For nonnegative observations, first truncate observation i at level i . Chebyshev and Borel–Cantelli then control the centered partial sums on geometric subsequences. Monotonicity of the nonnegative partial sums fills the gaps between subsequence indices. Apply that argument separately to X_i^+ and X_i^- for signed observations. Exercise 10 supplies the complete weighted proof, including the summability estimate and the passage from subsequences to all n .

Instructor's Note *Geometric subsequences in the strong-law proof*

Checking every sample size at once is hard. Instead, put checkpoints at $1, \alpha, \alpha^2, \dots$ with $\alpha > 1$. If the bad-checkpoint probabilities are summable, almost every journey eventually passes all checkpoints successfully. Nonnegative partial sums keep the journey between its neighboring checkpoints. Finally bring the checkpoints closer together by taking α toward one. Truncation, subsequences, and Borel–Cantelli are a useful toolkit to recognize in probability proofs.

Integrability is also necessary here. For i.i.d. variables, an almost-sure finite limit of the sample averages implies $\mathbb{E}|X_1| < \infty$; Exercise 11 proves the necessity.

6 Uniform laws of large numbers (optional)

6.1 Empirical distribution functions

Low priority *Uniform convergence of the empirical CDF*

For i.i.d. observations with CDF F , their **empirical CDF** is $F_n(x) = n^{-1} \sum_{i=1}^n \mathbf{1}_{\{X_i \leq x\}}$. For every fixed x , the strong LLN gives $F_n(x) \rightarrow F(x)$ a.s. The next result strengthens this pointwise statement to uniform convergence over x .

Theorem 7 (Glivenko–Cantelli). *For any CDF F , including CDFs with jumps,*

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \longrightarrow 0 \quad a.s.$$

Proof for continuous F . Fix an integer $m \geq 2$ and choose x_j with $F(x_j) = j/m$ for $j = 1, \dots, m-1$. Put $x_0 = -\infty$, $x_m = \infty$ and give both CDFs the endpoint values 0 and 1. For $x_j \leq x \leq x_{j+1}$, monotonicity bounds both CDFs between their values at the two endpoints, so

$$|F_n(x) - F(x)| \leq \max_{0 \leq j \leq m} |F_n(x_j) - F(x_j)| + \frac{1}{m}.$$

The maximum is over finitely many points and tends to zero a.s. Intersect the probability-one events for integer $m \geq 2$. On that single event, $\limsup_n \sup_x |F_n(x) - F(x)| \leq 1/m$ for every m , and hence the limit is zero. $\square$

A discontinuous CDF need not attain the values j/m , so that grid argument cannot simply be repeated unchanged. Exercise 12 gives a complete extension using quantiles.

Instructor's Note *Interpreting uniform CDF convergence*

Think of accumulating the mass in an empirical histogram from left to right. The empirical CDF records how much of the sample lies below each threshold. As data accumulate, its staircase hugs the population CDF. Uniform convergence asks for the worst vertical gap anywhere on the plot to shrink, not just the gap at a threshold chosen beforehand.

6.2 Uniform convergence over compact parameter spaces

Low priority *Uniform strong law under an integrable envelope*

Theorem 8. Let X_i be i.i.d. and $\Theta \subset \mathbb{R}^p$ be nonempty and compact. Suppose $g(x, \theta)$ is measurable in x for each θ , continuous in θ for each x , and

$$|g(x, \theta)| \leq h(x) \quad \text{for every } \theta \in \Theta, \quad \mathbb{E} h(X_1) < \infty.$$

Then $Q(\theta) = \mathbb{E} g(X_1, \theta)$ is continuous and

$$\sup_{\theta \in \Theta} |Q_n(\theta) - Q(\theta)| \rightarrow 0 \quad \text{a.s.}, \quad Q_n(\theta) = \frac{1}{n} \sum_{i=1}^n g(X_i, \theta).$$

Proof. Dominated convergence proves continuity of Q . For a fixed center θ and radius $r > 0$, set

$$w_{\theta,r}(x) = \sup_{\substack{\eta \in \Theta \\ \|\eta - \theta\| < r}} |g(x, \eta) - g(x, \theta)|.$$

Continuity allows the supremum to be taken over a countable dense subset of the relative open ball, so $w_{\theta,r}$ is measurable. Also $w_{\theta,r} \leq 2h$ and $w_{\theta,r}(x) \rightarrow 0$ as $r \downarrow 0$. Given $\delta > 0$, choose r_θ with $\mathbb{E} w_{\theta,r_\theta}(X_1) < \delta$. Compactness gives finitely many centers $\theta_1, \dots, \theta_J$ whose balls of radii $r_{\theta_j}/2$ cover Θ . Write $w_j = w_{\theta_j, r_{\theta_j}}$ and $P_n w_j = n^{-1} \sum_i w_j(X_i)$. If η lies in the j th covering ball, then

$$|Q_n(\eta) - Q(\eta)| \leq |Q_n(\theta_j) - Q(\theta_j)| + P_n w_j + \mathbb{E} w_j(X_1).$$

The strong law at the finitely many centers and for the finitely many w_j bounds the limsup of the supremum by 2δ a.s. Take a countable sequence $\delta \downarrow 0$ to finish. The supremum in the theorem is itself measurable by continuity and a countable dense subset of Θ . No measurable choice of a maximizing parameter is needed. $\square$

7 Consistency of moment and likelihood estimators

7.1 Moments and continuous transformations

Mid priority *Consistency of sample moments*

The integrable weak LLN gives consistency of the sample mean. More generally, $n^{-1} \sum_i \phi(X_i) \xrightarrow{P} \mathbb{E} \phi(X_1)$ whenever the observations are i.i.d. and $\mathbb{E} |\phi(X_1)| < \infty$. For example, estimating a second moment by an empirical second moment needs $\mathbb{E} X_1^2 < \infty$, but does not need a fourth moment. These mean and variance examples continue Shao's Example 2.33 [3].

Exercise prompt. Assume i.i.d. observations with finite second moment. Prove consistency of both $V_n = n^{-1} \sum_i (X_i - \bar{X}_n)^2$ and the usual sample variance $S_n^2 = (n-1)^{-1} \sum_i (X_i - \bar{X}_n)^2$ for $n \geq 2$. See Exercise 4 in Section 10 and its complete calculation.

Mid priority *Continuous mapping and method-of-moments consistency*

If $Z_n \xrightarrow{P} z$ and f is continuous at the fixed point z , then $f(Z_n) \xrightarrow{P} f(z)$. To see this, choose $\delta > 0$ so that $\|u - z\| < \delta$ implies $\|f(u) - f(z)\| < \varepsilon$. The probability that the transformed variable differs from its limit by more than ε is bounded by $\mathbb{P}(\|Z_n - z\| \geq \delta)$, which tends to zero (bound it by the probability of $\|Z_n - z\| > \delta/2$).

For the method of moments, let $h(\theta)$ be the vector of population moments and $\hat{\mu}_n$ the corresponding empirical moments. Suppose the required absolute moments are finite, the moment map identifies θ_0 , and its inverse h^{-1} is continuous at $\mu = h(\theta_0)$, relative to its domain. If $\mathbb{P}(\hat{\mu}_n \in h(\Theta)) \rightarrow 1$, define $\hat{\theta}_n = h^{-1}(\hat{\mu}_n)$ on that event and assign any fixed parameter value otherwise. Assume the resulting estimator is measurable. Then $\hat{\theta}_n \xrightarrow{P} \theta_0$. Indeed, the probability of estimation error is bounded by the infeasibility probability plus the probability that $\hat{\mu}_n$ is outside the continuity neighborhood of μ .

A global bijection onto all of $\mathbb{R}^k$ is unnecessary. Identification by population moments does not ensure that the empirical moment equations have a solution. Lecture 8's examples illustrate the need to specify the estimator when those equations are infeasible.

Low priority *Strong consistency of moment estimators*

Replacing the weak law by Theorem 6 gives a.s. limits for each integrable sample moment. Continuity then gives strong consistency of the mean and sample variance. The method-of-moments estimator is also strongly consistent if the empirical inverse is defined eventually a.s. and is continuous at the true moment vector. All preceding weak-consistency arguments have complete routes using the high-priority weak law alone.

7.2 Uniform convergence and minimizer consistency

Mid priority *Uniform convergence and separation of the minimum*

Write $Q_n(\theta)$ for a sample objective to minimize and $Q(\theta)$ for its population counterpart. Unlike a fixed moment, the chosen parameter $\hat{\theta}_n$ depends on the data. Pointwise convergence at each fixed θ need not control the data-dependent minimizer $\hat{\theta}_n$.

Theorem 9 (Consistency of approximate minimizers). *Let $\theta_0 \in \Theta \subset \mathbb{R}^p$. Assume $Q(\theta_0)$ is finite and, for each $\varepsilon > 0$, the **separation gap***

$$\Delta_\varepsilon = \inf_{\substack{\theta \in \Theta \\ \|\theta - \theta_0\| \geq \varepsilon}} [Q(\theta) - Q(\theta_0)] > 0$$

whenever the set is nonempty. Suppose the objectives are real-valued, $D_n = \sup_{\theta \in \Theta} |Q_n(\theta) - Q(\theta)|$ is measurable, $D_n \xrightarrow{P} 0$, and a measurable estimate satisfies

$$Q_n(\hat{\theta}_n) \leq Q_n(\theta_0) + r_n, \quad r_n \geq 0, \quad r_n \xrightarrow{P} 0.$$

Then $\hat{\theta}_n \xrightarrow{P} \theta_0$.

Proof. Uniform control allows evaluation at the random $\hat{\theta}_n$:

$$Q(\hat{\theta}_n) - Q(\theta_0) \leq 2D_n + r_n.$$

If $\|\hat{\theta}_n - \theta_0\| \geq \varepsilon$, the left side is at least Δ_ε . Thus the probability of that event is at most $\mathbb{P}(2D_n + r_n \geq \Delta_\varepsilon)$, which tends to zero. If there are no parameters that far away, the event is empty. $\square$

On compact Θ , continuity of Q and a unique minimizer imply separation: the closed set away from θ_0 is compact, so its minimum is attained and is strictly above $Q(\theta_0)$. Strict convexity is one possible route to uniqueness, not a necessary assumption. The optional compact-family uniform law of large numbers (ULLN) gives sufficient conditions for uniform convergence; the theorem above treats it as an explicit assumption and its proof does not require the optional strong law.

Instructor's Note *Pointwise and uniform control of an objective function*

Imagine the sample loss as a landscape that changes with the data. A pointwise limit checks the height at each fixed location. The optimizer can follow a narrow moving valley that those pointwise checks miss. Uniform control keeps the entire sample landscape close to its population shape. If the population objective has a well-separated minimum, its surrounding region also contains the sample minimizer once the uniform error is sufficiently small.

Exercise 13 gives a counterexample in which a localized perturbation shifts the minimizer despite pointwise convergence. The general framework is often called **M-estimation**.³ See [2], Theorem 2.1 and Section 2.3, for a broader treatment of consistency and uniform convergence.

7.3 Regression and likelihood examples

Low priority *Consistency of fixed-design least squares*

Consider $Y_i = z_i^\top \beta_0 + \varepsilon_i$ with deterministic $z_i \in \mathbb{R}^p$, independent centered errors of common variance $\sigma^2 < \infty$, and fixed p . If the errors are Normal, minimizing $\sum_i (Y_i - z_i^\top \beta)^2$ also maximizes the likelihood in β (for known positive noise variance).

Assume $G_n = n^{-1} \sum_i z_i z_i^\top \rightarrow G$ where G is positive definite. Eventually G_n is invertible and

$$\hat{\beta}_n - \beta_0 = G_n^{-1} b_n, \quad b_n = \frac{1}{n} \sum_{i=1}^n z_i \varepsilon_i, \quad \mathbb{E} \|b_n\|^2 = \frac{\sigma^2}{n} \text{tr}(G_n) = O(n^{-1}).$$

The inverse operator norms are eventually bounded because the smallest eigenvalue of G is positive. Markov's inequality therefore gives $\sqrt{n}(\hat{\beta}_n - \beta_0) = O_P(1)$ and consistency. For the finitely many initial sample sizes with a singular design matrix, define the estimator by any measurable rule. This proof accommodates a noncompact parameter space and nonidentical responses; it does not apply an i.i.d. compact-parameter ULLN to fixed-design data.

³ An M -estimator optimizes a data-dependent objective, often an average loss. Maximum likelihood and least squares are examples of this general estimation framework.

Low priority *Likelihood consistency and Kullback–Leibler divergence*

For i.i.d. data with densities p_θ relative to a common measure, MLE minimizes $Q_n(\theta) = -n^{-1} \sum_i \log p_\theta(X_i)$. When the population log losses are finite,

$$Q(\theta) - Q(\theta_0) = \mathbb{E}_{\theta_0} \log \frac{p_{\theta_0}(X_1)}{p_\theta(X_1)} = D_{\text{KL}}(P_{\theta_0} \| P_\theta) \geq 0.$$

The KL inequality from Lecture 8 identifies the population optimum; equality means equality of distributions, so an identifiable model has the unique target θ_0 . For densities, the same Jensen argument uses $R = p_\theta(X_1)/p_{\theta_0}(X_1)$ under P_{θ_0} : $\mathbb{E} R \leq 1$ and $-\mathbb{E} \log R \geq -\log \mathbb{E} R \geq 0$. Equality forces $R = 1$ a.s. and leaves no extra probability mass outside the true support. Consistency still requires separation and control of the sample objective, such as the conditions above. If p_θ vanishes on a set of positive P_{θ_0} probability, the KL divergence is infinite. The real-valued objective theorem cannot then be invoked without additional work. Models with parameter-dependent support often permit a direct consistency proof. General MLE consistency and its regularity conditions will receive more attention in STAT 710. Regular finite-dimensional models often give $\sqrt{n}$ -consistency; nonregular models can have faster or slower rates.

Low priority *Uniform endpoint estimation and its convergence rate*

For X_i i.i.d. $\text{Uniform}[0, \theta]$, $\theta > 0$, the MLE is $M_n = \max_{1 \leq i \leq n} X_i$ (up to null endpoint cases). This is Shao's Example 2.34 [3]. It is consistent and ***n-consistent***. Its rate is determined by the largest observation, which approaches the support boundary.

Exercise prompt. Find $\mathbb{P}(\theta - M_n > \varepsilon)$ and bound $\mathbb{P}(n(\theta - M_n) > t)$. Use the results to prove consistency and *n-consistency*. See Exercise 5 in Section 10 and its complete calculation.

Instructor's Note *Averaging and endpoint estimation*

An average combines information from the full sample, with positive and negative deviations tending to offset each other. An endpoint estimate depends on observations nearest the support boundary. Imagine filling the gap between the largest observation and that boundary. The size of this gap can decrease at a different rate from the fluctuations of an average. Convergence rates reflect both the model and the estimator.

Key ideas

- Almost-sure convergence implies convergence in probability; the converse can fail.
- The first Borel–Cantelli lemma turns summable error probabilities into eventual pathwise control.
- Finite-mean i.i.d. averages obey the weak law; different proof routes use different moment assumptions.
- Consistency combines convergence of the data summaries with continuity or separation of an optimization target.

8 Detailed reference

Notion (linked)	Meaning or useful condition	Priority
Convergence in probability	Each fixed error tolerance is exceeded with probability tending to zero.	High
Almost-sure convergence	Almost every complete sample path converges.	High
Relation between convergence modes	Almost-sure convergence implies convergence in probability; the converse fails.	High
First Borel–Cantelli lemma	Summable event probabilities imply that only finitely many of the events occur a.s.	High
o_P and O_P	Convergence to zero in probability versus boundedness in probability after scaling.	High
Consistency and rates	Convergence to the estimand, evaluated at each fixed model distribution.	High
Variance-based LLN	Uncorrelated observations, common mean, uniformly bounded variances.	High
Weak law by truncation	A vanishing tail condition yields convergence after truncated centering; its proof is optional.	Mid
Integrable weak LLN	I.i.d. observations and a finite first absolute moment.	High
Strong LLN	Integrable i.i.d. averages converge along almost every path.	Low
Kolmogorov’s zero–one law	Independent observations make tail-event probabilities zero or one.	Low
Glivenko–Cantelli	The entire empirical CDF converges uniformly a.s.	Low
Uniform LLN on compact sets	Compact parameters, continuity, and an integrable envelope.	Low
Method-of-moments consistency	Moment LLNs, continuity of the inverse, and empirical feasibility.	Mid
Argmin consistency	Uniform objective control, separation, and vanishing optimization error.	Mid
Least-squares consistency	A positive-definite limiting Gram matrix controls coefficient error.	Low
Uniform endpoint estimation	The maximum estimates the upper endpoint at an n^{-1} scale.	Low

Looking ahead

Lecture 10 goes beyond consistency to describe the scale and distribution of estimation error. The CLT adds a fluctuation approximation to the laws of large numbers studied here. Continue to Lecture 10.

9 Further reading

Further reading *References on limit theorems and estimator consistency*

For weak laws, Borel–Cantelli, strong laws, and related tools, see Durrett [1], Sections 2.2–2.5. The author’s fifth-edition page links an accessible manuscript; Theorems 2.2.12–2.2.14 treat truncation and the integrable weak law, and Theorem 2.4.1 treats the strong law. Shao [3] provides the statistical convergence and consistency framework.

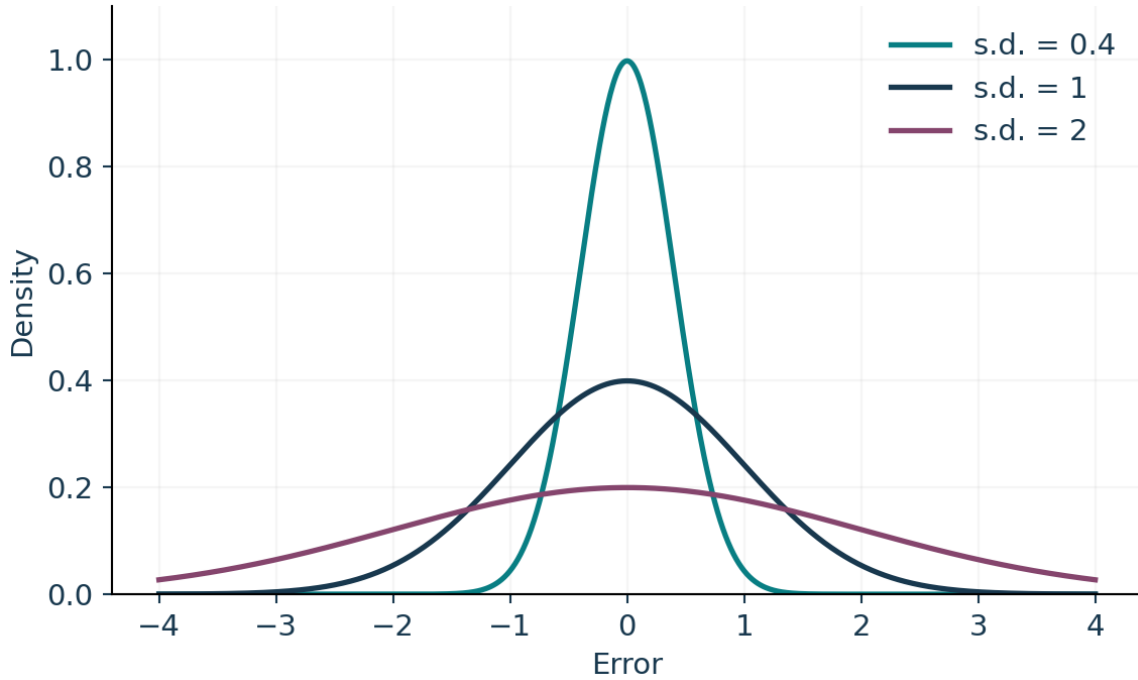
For consistency of extremum estimators, see Newey and McFadden [2], Theorem 2.1 and Lemma 2.4. An accessible copy of their chapter is hosted by UW–Madison. The chapter develops conditions and extensions beyond the elementary versions here. More detailed likelihood asymptotics and inference will be studied later in the course sequence.

10 Exercises and complete solutions

High priority *Stochastic orders of Gaussian sequences*

Exercise 1. Let $X_n \sim N(0, \sigma_n^2)$, allowing $\sigma_n = 0$. Prove the three claims in Example 1.

Solution. For a standard Normal Z , the marginal law of X_n is that of $\sigma_n Z$. If $0 \leq \sigma_n \leq B < \infty$, then for $B > 0$, $\mathbb{P}(|X_n| > M) \leq \mathbb{P}(|Z| > M/B) \rightarrow 0$ as $M \rightarrow \infty$, uniformly in n . If $B = 0$ the assertion is immediate. Hence $X_n = O_P(1)$. In particular $\sigma_n \rightarrow \sigma < \infty$ gives a bounded sequence and $O_P(1)$, though $\sigma_n = 1$ does not give $O_P(1)$.



Centered Normal densities with standard deviations 0.4, 1, 2, on shared axes. Each has total area one; the smaller scale concentrates more probability near zero.

If $\sigma_n \rightarrow 0$, the probability $\mathbb{P}(|X_n| > \varepsilon)$ tends to zero: for nonzero scales it equals $\mathbb{P}(|Z| > \varepsilon/\sigma_n)$, and for zero scales it is zero. If $\sigma_n \rightarrow \infty$, then for every fixed $M > 0$, $\mathbb{P}(|X_n| > M) = \mathbb{P}(|Z| > M/\sigma_n) \rightarrow 1$, since $\mathbb{P}(Z = 0) = 0$. This contradicts the definition of $O_P(1)$, for example with $\eta = 1/2$. All arguments use marginal probabilities only. $\square$

High priority *Consistency from a stochastic rate*

Exercise 2. Prove that $a_n(T_n - \nu) = O_P(1)$ with deterministic $a_n \rightarrow \infty$ implies $T_n \xrightarrow{P} \nu$.

Solution. Fix $\varepsilon, \eta > 0$. Choose M, N so that $\mathbb{P}(a_n \|T_n - \nu\| > M) < \eta$ for $n \geq N$. Eventually $a_n \varepsilon \geq M$, and then

$$\mathbb{P}(\|T_n - \nu\| > \varepsilon) \leq \mathbb{P}(a_n \|T_n - \nu\| > M) < \eta.$$

Since η is arbitrary, the probability tends to zero. In a statistical model this proof is applied under each fixed P , with constants that may depend on P . $\square$

High priority *Weak law under a finite first absolute moment*

Exercise 3. Derive Theorem 4 from the truncation theorem when $\mathbb{E}|X_1| < \infty$.

Solution. The variables $|X_1|\mathbf{1}_{\{|X_1|>n\}}$ tend pointwise to zero and are dominated by the integrable variable $|X_1|$. Hence

$$n\mathbb{P}(|X_1| > n) \leq \mathbb{E}[|X_1|\mathbf{1}_{\{|X_1|>n\}}] \rightarrow 0.$$

Also $X_1\mathbf{1}_{\{|X_1|\leq n\}} \rightarrow X_1$ pointwise, with the same absolute dominator, so $a_n \rightarrow \mu = \mathbb{E}X_1$. Both applications of dominated convergence use one fixed X_1 , not an unproved pathwise limit of the moving observations X_n . For large n , $|a_n - \mu| \leq \varepsilon/2$, so Theorem 3 gives $\mathbb{P}(|\bar{X}_n - \mu| > \varepsilon) \leq \mathbb{P}(|\bar{X}_n - a_n| > \varepsilon/2) \rightarrow 0$. $\square$

Mid priority *Consistency of sample variance*

Exercise 4. For i.i.d. X_i with $\mathbb{E}X_1^2 < \infty$, show that V_n and S_n^2 from the main prompt converge in probability to $\text{Var}(X_1)$.

Solution. Finite second moment implies a finite first absolute moment. Apply the weak LLN to X_i and separately to X_i^2 :

$$\bar{X}_n \xrightarrow{P} \mu, \quad \frac{1}{n} \sum_i X_i^2 \xrightarrow{P} \mathbb{E}X_1^2.$$

The vector of these two quantities converges in probability by the union bound. Continuity of $(u, v) \mapsto v - u^2$ gives

$$V_n = \frac{1}{n} \sum_i X_i^2 - \bar{X}_n^2 \xrightarrow{P} \mathbb{E}X_1^2 - \mu^2 = \text{Var}(X_1).$$

Finally $S_n^2 = \frac{n}{n-1}V_n$ for $n \geq 2$, and $n/(n-1) \rightarrow 1$. Joint convergence with that deterministic factor and continuity of multiplication prove the second claim. No fourth moment was used. $\square$

Low priority *Consistency and rate of the uniform maximum*

Exercise 5. For i.i.d. Uniform $[0, \theta]$ data with $\theta > 0$, prove consistency and n -consistency of $M_n = \max_i X_i$ by computing its error tails.

Solution. For $0 \leq x \leq \theta$, independence gives $\mathbb{P}(M_n \leq x) = (x/\theta)^n$. The distribution is continuous, so strict versus nonstrict inequalities at a fixed x give the same result. For $0 < \varepsilon < \theta$,

$$\mathbb{P}(|M_n - \theta| > \varepsilon) = \left(1 - \frac{\varepsilon}{\theta}\right)^n \rightarrow 0;$$

for $\varepsilon \geq \theta$ the probability is zero. Thus M_n is consistent. For $0 \leq t \leq n\theta$,

$$\mathbb{P}(n|M_n - \theta| > t) = \left(1 - \frac{t}{n\theta}\right)^n \leq e^{-t/\theta},$$

and for $t > n\theta$ the probability is zero. The exponential bound holds uniformly over n . Choosing t large enough makes it less than any given $\eta > 0$, proving $n(M_n - \theta) = O_P(1)$. The sign of $M_n - \theta$ is nonpositive; the absolute error was used only to state its magnitude. $\square$

Mid priority *Product rule for stochastic orders*

Exercise 6. Prove $o_P(1) \subset O_P(1)$ and that $U_n = o_P(1)$, $V_n = O_P(1)$ imply $U_n V_n = o_P(1)$, without independence.

Solution. If $U_n \xrightarrow{P} 0$, then for each $\eta > 0$ we eventually have $\mathbb{P}(|U_n| > 1) < \eta$, which proves $O_P(1)$ using $M = 1$. For the product, fix $\varepsilon, \eta > 0$ and choose $M > 0$ so that $\mathbb{P}(|V_n| > M) < \eta/2$ eventually. The event inclusion

$$\{|U_n V_n| > \varepsilon\} \subset \{|V_n| > M\} \cup \{|U_n| > \varepsilon/M\}$$

gives a probability bound below η for all sufficiently large n . No factorization of probabilities was used. Dividing by deterministic positive scales gives the general product rule in the main text. $\square$

High priority *Convergence in probability without almost-sure convergence*

Exercise 7. Let $B_n, n \geq 2$, be independent Bernoulli variables with $\mathbb{P}(B_n = 1) = 1/n$. Show that $B_n \xrightarrow{P} 0$, but almost surely $B_n = 1$ for infinitely many n .

Solution. For $0 < \varepsilon < 1$, $\mathbb{P}(|B_n| > \varepsilon) = 1/n \rightarrow 0$; for $\varepsilon \geq 1$ the probability is zero. For fixed $N \geq 2$,

$$\mathbb{P}(B_N = \cdots = B_m = 0) = \prod_{n=N}^m (1 - 1/n) = \frac{N-1}{m} \rightarrow 0.$$

Continuity from above implies that the probability of all zeros after N is zero. The union of these events over N still has probability zero. Its complement is the event of infinitely many ones. On that event B_n does not tend to zero, proving the claimed distinction. $\square$

High priority *Borel–Cantelli via the expected number of occurrences*

Exercise 8. Give a second proof of the first Borel–Cantelli lemma using $N = \sum_{n \geq 1} \mathbf{1}_{E_n}$ and nonnegative integration.

Solution. Monotone convergence (equivalently Tonelli for a nonnegative sum) gives $\mathbb{E} N = \sum_n \mathbb{P}(E_n) < \infty$. Markov’s inequality implies $\mathbb{P}(N \geq K) \leq \mathbb{E} N / K$ for each $K > 0$; hence $\mathbb{P}(N = \infty) = 0$. Only finitely many of the events occur a.s. For convergence of errors, apply this result at each tolerance $1/m$ and intersect countably many probability-one events. An uncountable intersection is unnecessary. $\square$

Mid priority *Pareto distribution: finite mean and infinite variance*

Exercise 9. Let $X \geq 1$ have tail $\mathbb{P}(X > t) = t^{-\alpha}$ for $t \geq 1$, with $1 < \alpha < 2$. Calculate its first and second moments and explain which weak LLN applies to i.i.d. copies.

Solution. The tail-integral formulas give

$$\mathbb{E} X = \int_0^1 1 dt + \int_1^\infty t^{-\alpha} dt = 1 + \frac{1}{\alpha - 1} = \frac{\alpha}{\alpha - 1} < \infty,$$

while

$$\mathbb{E} X^2 = 2 \int_0^1 t dt + 2 \int_1^\infty t^{1-\alpha} dt = \infty.$$

The variance bound in Theorem 1 is unavailable. Theorem 4 does apply, so the sample mean converges in probability to $\alpha/(\alpha - 1)$. That theorem supplies consistency, not a $\sqrt{n}$ rate for this example. $\square$

Low priority *Proof of the weighted strong law*

Exercise 10. Prove Theorem 6 by truncation, geometric subsequences, and Borel–Cantelli.

Solution. First suppose $X_i = Z_i \geq 0$, with $\mu = \mathbb{E} Z_1 < \infty$. Put $Y_i = Z_i \mathbf{1}_{\{Z_i \leq i\}}$ and $R_n = \sum_{i=1}^n c_i Y_i$. Since $\sum_{i \geq 1} \mathbb{P}(Z_i > i) \leq \int_0^\infty \mathbb{P}(Z_1 > t) dt = \mu$, the first Borel–Cantelli lemma gives $Y_i = Z_i$ eventually a.s. Their weighted partial sums therefore differ by a finite random quantity for all sufficiently large n .

Summability along geometric subsequences. Tonelli and $\sum_{i \geq k} i^{-2} \leq 2/k$ for $k \geq 1$ imply

$$\sum_{i \geq 1} \frac{\mathbb{E}[Z_1^2 \mathbf{1}_{\{Z_1 \leq i\}}]}{i^2} = \mathbb{E} \left[Z_1^2 \sum_{i \geq \max(1, \lceil Z_1 \rceil)} \frac{1}{i^2} \right] \leq 2 \mathbb{E} Z_1 < \infty.$$

For $0 \leq Z_1 \leq 1$ use $Z_1^2 \leq Z_1$; for $Z_1 > 1$ use $\lceil Z_1 \rceil \geq Z_1$ in the bound. Fix $\alpha > 1$ and $k_m = \lfloor \alpha^m \rfloor$. Ignore the finitely many initial repeated indices. There is a constant $K_\alpha < \infty$ such that $\sum_{m: k_m \geq i} k_m^{-2} \leq K_\alpha/i^2$: for large m , $k_m \geq \alpha^m/2$, and summing the geometric series from the first m with $k_m \geq i$ proves the bound, enlarging the constant for small indices. Independence and $c_i \leq C$ now yield

$$\begin{aligned} \sum_m \mathbb{P}(|R_{k_m} - \mathbb{E} R_{k_m}| > \varepsilon k_m) &\leq \frac{C^2}{\varepsilon^2} \sum_i \mathbb{E} Y_i^2 \sum_{m: k_m \geq i} \frac{1}{k_m^2} \\ &\leq \frac{C^2 K_\alpha}{\varepsilon^2} \sum_i \frac{\mathbb{E} Y_i^2}{i^2} < \infty. \end{aligned} \tag{5}$$

Apply Borel–Cantelli at $\varepsilon = 1, 1/2, 1/3, \dots$ to obtain $(R_{k_m} - \mathbb{E} R_{k_m})/k_m \rightarrow 0$ a.s.

Extension from subsequences to the full sequence. Both R_n and $\mathbb{E} R_n$ are nondecreasing. For $k_m \leq n < k_{m+1}$,

$$\begin{aligned} R_n - \mathbb{E} R_n &\leq R_{k_{m+1}} - \mathbb{E} R_{k_{m+1}} + \mathbb{E} R_{k_{m+1}} - \mathbb{E} R_{k_m}, \\ R_n - \mathbb{E} R_n &\geq R_{k_m} - \mathbb{E} R_{k_m} - (\mathbb{E} R_{k_{m+1}} - \mathbb{E} R_{k_m}). \end{aligned}$$

The expectation increment is at most $C\mu(k_{m+1} - k_m)$. Divide by $n \geq k_m$ and use $k_{m+1}/k_m \rightarrow \alpha$ to conclude $\limsup_n |R_n - \mathbb{E} R_n|/n \leq C\mu(\alpha - 1)$ a.s. Intersect these probability-one events for $\alpha = 1 + 1/j$, $j \geq 1$, and let $j \rightarrow \infty$. Thus $(R_n - \mathbb{E} R_n)/n \rightarrow 0$ a.s.

Truncation bias and recovery of the original sums. Dominated convergence gives $\mathbb{E} Y_i \rightarrow \mu$. Consequently

$$\left| \frac{1}{n} \sum_{i=1}^n c_i (\mathbb{E} Y_i - \mu) \right| \leq \frac{C}{n} \sum_{i=1}^n |\mathbb{E} Y_i - \mu| \rightarrow 0.$$

The last step is the elementary Cesàro fact: the averages of a convergent-to-zero sequence tend to zero, proved by splitting off a fixed finite initial segment. Combining the centered limit, this bias bound, and eventual equality of Y_i, Z_i proves the weighted theorem for nonnegative observations.

For signed X_i , apply the established nonnegative result to the i.i.d. sequences X_i^+ and X_i^- , both integrable. Intersect their two probability-one events and subtract. This step avoids using monotonicity for signed partial sums, which would be invalid. $\square$

Low priority *Necessity of integrability for the strong law*

Exercise 11. Suppose X_i are i.i.d. and $n^{-1} \sum_{i=1}^n X_i$ converges a.s. to a finite limit. Show that $\mathbb{E}|X_1| < \infty$.

Solution. Write $A_n = n^{-1} \sum_{i=1}^n X_i$. Then $X_n/n = A_n - \frac{n-1}{n} A_{n-1} \rightarrow 0$ a.s. If $\sum_n \mathbb{P}(|X_n| > n) = \infty$, independence would give, for each N ,

$$\mathbb{P}(|X_n| \leq n \text{ for all } n \geq N) = \lim_{m \rightarrow \infty} \prod_{n=N}^m (1 - \mathbb{P}(|X_n| > n)) \leq \lim_{m \rightarrow \infty} e^{-\sum_{n=N}^m \mathbb{P}(|X_n| > n)} = 0.$$

Taking the countable union over N contradicts $X_n/n \rightarrow 0$ a.s. This argument used independence for the product and identical distribution to replace X_n by X_1 . Therefore $\sum_n \mathbb{P}(|X_1| > n) < \infty$. Since the tail is decreasing,

$$\mathbb{E}|X_1| = \int_0^\infty \mathbb{P}(|X_1| > t) dt \leq 1 + \sum_{n \geq 1} \mathbb{P}(|X_1| > n) < \infty.$$

□

Low priority *Glivenko–Cantelli for discontinuous distributions*

Exercise 12. Extend the continuous-CDF proof of Glivenko–Cantelli to an arbitrary CDF F using its quantile function.

Solution. For $0 < u < 1$, let $q(u) = \inf\{x : F(x) \geq u\}$. Monotonicity, right-continuity, and the endpoint limits of a CDF give

$$q(u) \leq x \iff u \leq F(x).$$

For completeness, $u \leq F(x)$ places x in the defining set and gives $q(u) \leq x$. In the other direction, the set contains points arbitrarily close above $q(u)$; right-continuity gives $F(q(u)) \geq u$, and monotonicity then gives $F(x) \geq u$. This also proves measurability of q .

Let U_i be i.i.d. Uniform(0, 1) and construct $X'_i = q(U_i)$. Then X'_i are i.i.d. with CDF F . If G_n is the empirical CDF of the U_i , the equivalence gives

$$F'_n(x) = G_n(F(x)), \quad \sup_x |F'_n(x) - F(x)| \leq \sup_{0 \leq u \leq 1} |G_n(u) - u| \rightarrow 0 \quad \text{a.s.}$$

The last result is the already proved continuous-CDF case. The constructed sequence (X'_i) and the original i.i.d. sequence (X_i) have the same joint law. The uniform-convergence event is measurable: by right-continuity, the supremum over $x \in \mathbb{R}$ equals the supremum over rational x . Its probability is therefore the same for the two sequences, and equals one. Atoms cause no difficulty for this argument. □

Mid priority *Failure of argmin consistency under pointwise convergence*

Exercise 13. On $\Theta = [0, 2]$, set $Q(\theta) = \theta^2$ and, for $n \geq 2$, put $a_n = 1 + 1/n$ and

$$Q_n(\theta) = \theta^2 - 4 \max\{1 - n^3|\theta - a_n|, 0\}.$$

Show that Q_n converges pointwise to Q , but any minimizers of Q_n converge to 1 rather than the unique minimizer 0 of Q .

Solution. The perturbation is supported on $[a_n - n^{-3}, a_n + n^{-3}]$. For each fixed $\theta \neq 1$, this interval eventually misses θ because both endpoints tend to 1. It also misses $\theta = 1$ for every $n \geq 2$, since $1/n > 1/n^3$. Thus $Q_n(\theta) = Q(\theta)$ eventually for each fixed θ .

Continuity and compactness give a minimizer of Q_n . At the center $Q_n(a_n) = a_n^2 - 4 < 0$, whereas outside the support $Q_n(\theta) = \theta^2 \geq 0$. Every minimizer must lie within n^{-3} of a_n and hence converges to 1. Nevertheless Q has its unique minimum at 0. Finally $\sup_\theta |Q_n(\theta) - Q(\theta)| = 4$ for every n , displaying exactly the missing uniform control. □

References

- [1] Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 5th edition, 2019.
- [2] Whitney K Newey and Daniel McFadden. Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245, 1994.
- [3] Jun Shao. *Mathematical Statistics*. Springer, 2nd edition, 2003.

What changed

October 19, 2026 (instructor follow-up). Changed Section 7.1’s main results and linked sample-variance exercise to mid priority; retained the optional strong-consistency extension.

October 19, 2026: Corrected notation, wording and cross-references after the course-wide errata review; refreshed the linked HTML and PDF editions.

October 14, 2026. Added issue-reporting links to the webpage and PDF. The mathematics and priority assignments are unchanged in this update.