Consistency

Statistics · Module 7 — Asymptotic Theory

Node 7.2

Position in Knowledge Graph

System: Statistics. Structural Domain: Module 7 — Asymptotic Theory. Node 7.2 — Consistency.

Consistency is the minimal asymptotic requirement for an estimator: as the sample size grows, the estimator must converge to the true parameter. This node instantiates the convergence modes of Node 7.1 in the estimation setting, depends on the likelihood framework (Nodes 3.1, 3.2), and enables the rate-of-convergence results in asymptotic normality (Node 7.3) and the general M-estimation framework (Node 7.5).

Formal Definition

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P_\theta$, $\theta \in \Theta \subseteq \mathbb{R}^k$. Let $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ be a sequence of estimators.

Weak consistency. $\hat{\theta}_n$ is weakly consistent for θ if $\hat{\theta}_n \xrightarrow{p} \theta$ under P_θ for all $\theta \in \Theta$.

Strong consistency. $\hat{\theta}_n$ is strongly consistent for θ if $\hat{\theta}_n \xrightarrow{a.s.} \theta$ under P_θ for all $\theta \in \Theta$.

Fisher consistency. A functional $T(F)$ is Fisher consistent for θ if $T(F_\theta) = \theta$ for all $\theta \in \Theta$. Fisher consistency is a population-level property; combined with convergence of the plug-in estimator $T(\hat{F}_n)$ to $T(F)$, it yields statistical consistency.

Intuition

Consistency says that with enough data, the estimator gets the answer right. It does not say how quickly or how precisely — that is the role of asymptotic normality (Node 7.3). An inconsistent estimator is fundamentally broken: no amount of data can rescue it. Consistency is therefore a necessary filter, not a measure of quality.

The conceptual mechanism is always the same: the sample criterion function converges to a population criterion function, and the population criterion has a unique optimizer at the truth. Identifiability ensures the truth is distinguishable; the law of large numbers drives the convergence.

Structural Role

Consistency is the first-order asymptotic property of an estimator. It guarantees that the sampling distribution of $\hat{\theta}_n$ concentrates at the true parameter as $n \to \infty$. It is a necessary but not sufficient condition for good statistical performance: it does not control the rate of convergence, the shape of the sampling distribution, or finite-sample behavior. Rate and distributional shape are characterized in Node 7.3. Consistency of M-estimators is studied in the general framework of Node 7.5.

Mathematical Development

Sufficient condition via moments. If $\mathbb{E}_\theta[\hat{\theta}_n] \to \theta$ and $\mathrm{Var}_\theta(\hat{\theta}_n) \to 0$, then $\hat{\theta}_n \xrightarrow{p} \theta$ by Chebyshev's inequality: $\mathbb{P}(|\hat{\theta}_n - \theta| > \varepsilon) \le \mathrm{Var}(\hat{\theta}_n)/\varepsilon^2 \to 0$. This is the simplest route to consistency but requires both vanishing bias and vanishing variance.

Consistency of the MLE: Wald's proof.

Let $M_n(\theta') = n^{-1}\ell(\theta'; X_1, \ldots, X_n) = n^{-1}\sum_{i=1}^n \log p_{\theta'}(X_i)$ be the normalized log-likelihood. Define the population criterion $M(\theta') = \mathbb{E}_\theta[\log p_{\theta'}(X)]$.

Step 1: Unique maximum. By the information inequality (Gibbs' inequality / KL divergence nonnegativity),

$$M(\theta') = \mathbb{E}_\theta[\log p_{\theta'}(X)] \le \mathbb{E}_\theta[\log p_\theta(X)] = M(\theta),$$

with equality if and only if $p_{\theta'} = p_\theta$ P_θ-a.e. Under identifiability ($p_{\theta'} = p_\theta$ a.e. implies $\theta' = \theta$), the maximum of $M(\theta')$ over Θ is uniquely attained at $\theta' = \theta$.

Step 2: Uniform convergence. Assume $\sup_{\theta' \in \Theta} |M_n(\theta') - M(\theta')| \xrightarrow{p} 0$. This holds under a uniform law of large numbers (ULLN), which requires either compactness of Θ and continuity of $\theta' \mapsto \log p_{\theta'}(x)$, or a Glivenko-Cantelli argument for the class $\{\log p_{\theta'} : \theta' \in \Theta\}$.

Step 3: Argmax continuous mapping theorem. If $M_n \to M$ uniformly and M has a unique maximizer θ, then $\hat{\theta}_n = \arg\max M_n \xrightarrow{p} \theta$.

Proof of step 3. Fix $\varepsilon > 0$. By uniqueness and continuity of M, there exists $\delta > 0$ such that $M(\theta') \le M(\theta) - \delta$ for all $\|\theta' - \theta\| > \varepsilon$. Under uniform convergence, for large n,

$$M_n(\hat{\theta}_n) \ge M_n(\theta) \ge M(\theta) - \delta/3,$$

and $M_n(\hat{\theta}_n) \le M(\hat{\theta}_n) + \delta/3$. If $\|\hat{\theta}_n - \theta\| > \varepsilon$, then $M(\hat{\theta}_n) \le M(\theta) - \delta$, giving $M_n(\hat{\theta}_n) \le M(\theta) - 2\delta/3$, a contradiction. $\square$

Consistency of the method of moments. Let $\mu_k(\theta) = \mathbb{E}_\theta[X^k]$ and $\hat{\mu}_k = n^{-1}\sum X_i^k$. By the WLLN, $\hat{\mu}_k \xrightarrow{p} \mu_k(\theta)$. If the moment map $g : \theta \mapsto (\mu_1(\theta), \ldots, \mu_r(\theta))$ is continuous and injective, the method-of-moments estimator $\hat{\theta}_n = g^{-1}(\hat{\mu}_1, \ldots, \hat{\mu}_r)$ satisfies $\hat{\theta}_n \xrightarrow{p} \theta$ by the continuous mapping theorem.

Uniform law of large numbers (ULLN). A class of functions $\mathcal{F}$ is Glivenko-Cantelli for P if $\sup_{f \in \mathcal{F}} |n^{-1}\sum f(X_i) - \mathbb{E}[f(X)]| \xrightarrow{a.s.} 0$. Sufficient conditions include: (i) $\mathcal{F}$ has finite VC dimension; (ii) $\mathcal{F}$ is uniformly bounded and parametrized by a compact set with equicontinuity. The ULLN is the key technical ingredient linking pointwise LLN convergence to the uniform convergence needed for consistency of extremum estimators.

Rate of convergence and consistency. Consistency alone does not specify a rate. The sample mean converges at rate $n^{-1/2}$ (by the CLT). The MLE for the uniform upper bound θ converges at rate n^{-1} (superefficient). Nonparametric density estimators converge at rate $n^{-2/(4+p)}$ (slower than parametric). These distinctions motivate the refined study of asymptotic normality in Node 7.3.

Consistency under model misspecification. When the true distribution P_0 does not belong to the model $\{P_\theta : \theta \in \Theta\}$, the MLE converges to the pseudo-true parameter

$$\theta^* = \arg\min_{\theta \in \Theta} \mathrm{KL}(P_0 \| P_\theta) = \arg\max_{\theta \in \Theta} \mathbb{E}_0[\log p_\theta(X)].$$

This is the KL projection of P_0 onto the model. The proof of consistency carries through verbatim — the ULLN and argmax CMT arguments do not require correct specification — but the limit θ^* is no longer the "true" parameter. This observation is the foundation for the sandwich variance under misspecification (Node 7.3).

Consistency of Bayesian estimators. The posterior mode (MAP estimator) is consistent under the same conditions as the MLE, since the prior contributes $O(1)$ to the log-posterior while the likelihood contributes $O(n)$. More generally, the posterior distribution itself concentrates at the true parameter (posterior consistency), provided the prior assigns positive mass to every KL neighborhood of the truth (Schwartz's theorem). This connects to the Bernstein-von Mises theorem (Node 6.7).

Well-separated maximum condition. An alternative to compactness for the argmax CMT: if $M(\theta)$ has a well-separated maximum, meaning

$$\sup_{\|\theta' - \theta\| > \varepsilon} M(\theta') < M(\theta) \quad \forall \varepsilon > 0,$$

then uniform convergence $\sup_{\theta'} |M_n(\theta') - M(\theta')| \xrightarrow{p} 0$ implies $\hat{\theta}_n \xrightarrow{p} \theta$. This condition holds when M is strictly concave or when the KL divergence $\mathrm{KL}(P_\theta \| P_{\theta'}) > 0$ for $\theta' \neq \theta$ (identifiability) and M is continuous.

Consistency and efficiency are independent. The sample mean and the sample median are both consistent for the center of a symmetric distribution. However, their rates differ: under normality, the mean converges at rate σ^2/n while the median converges at rate $\pi\sigma^2/(2n)$. Consistency is a qualitative property (yes/no); efficiency is quantitative (how fast).

Consistency of the bootstrap. The bootstrap is consistent if the bootstrap distribution of $\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)$ converges to the same limit as $\sqrt{n}(\hat{\theta}_n - \theta)$. This holds under the conditions for asymptotic normality of $\hat{\theta}_n$, providing nonparametric confidence intervals that are asymptotically correct.

Worked Examples

Example 1. Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathrm{Exp}(\lambda)$. The MLE is $\hat{\lambda}_n = 1/\bar{X}_n$. By the SLLN, $\bar{X}_n \xrightarrow{a.s.} 1/\lambda$. By the continuous mapping theorem with $g(x) = 1/x$, $\hat{\lambda}_n \xrightarrow{a.s.} \lambda$. The MLE is strongly consistent.

Example 2. Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathrm{Uniform}(0, \theta)$. The MLE is $\hat{\theta}_n = X_{(n)} = \max_i X_i$. We have $\mathbb{P}(X_{(n)} \leq t) = (t/\theta)^n$ for $0 \leq t \leq \theta$. Thus $\mathbb{P}(\theta - X_{(n)} > \varepsilon) = (1 - \varepsilon/\theta)^n \to 0$ for any $\varepsilon > 0$, so $\hat{\theta}_n \xrightarrow{p} \theta$. In fact $\sum_n \mathbb{P}(\theta - X_{(n)} > \varepsilon) < \infty$, so by Borel-Cantelli, $X_{(n)} \xrightarrow{a.s.} \theta$. Note that $\hat{\theta}_n$ is biased ($\mathbb{E}[X_{(n)}] = n\theta/(n+1)$) but consistent — consistency does not require unbiasedness.

Example 3 (Inconsistency). Let $\hat{\theta}_n = X_1$ regardless of n. Then $\hat{\theta}_n$ is unbiased but $\mathrm{Var}(\hat{\theta}_n) = \mathrm{Var}(X_1) \not\to 0$. This estimator is not consistent: unbiasedness alone is insufficient.

Example 4 (Method of moments). Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathrm{Gamma}(\alpha, \beta)$ with $\mathbb{E}[X] = \alpha/\beta$ and $\mathrm{Var}(X) = \alpha/\beta^2$. The moment equations are $\hat{\mu}_1 = \hat{\alpha}/\hat{\beta}$ and $\hat{\mu}_2 - \hat{\mu}_1^2 = \hat{\alpha}/\hat{\beta}^2$. Solving: $\hat{\beta} = \hat{\mu}_1/(\hat{\mu}_2 - \hat{\mu}_1^2)$ and $\hat{\alpha} = \hat{\mu}_1^2/(\hat{\mu}_2 - \hat{\mu}_1^2)$. By the WLLN, $\hat{\mu}_k \xrightarrow{p} \mu_k$, and by the continuous mapping theorem, $(\hat{\alpha}, \hat{\beta}) \xrightarrow{p} (\alpha, \beta)$ since the moment map is continuously invertible on $(0, \infty)^2$.

Example 5 (Non-compact parameter space). Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, 1)$ with $\Theta = \mathbb{R}$. The MLE $\hat{\mu} = \bar{X}_n$ is consistent by the WLLN. Here the ULLN holds despite non-compact Θ because the Gaussian log-likelihood has quadratic curvature that prevents the maximizer from escaping. In contrast, for a mixture model with the number of components as a parameter, unbounded Θ can lead to inconsistency of the MLE when the likelihood is unbounded.

Cross-Links

AI. Consistency of empirical risk minimizers (ERM) is the learning-theoretic analogue: under finite VC dimension, the empirical risk minimizer converges to the population risk minimizer. The ULLN over function classes is precisely the Glivenko-Cantelli condition used in both settings.

Economics. Consistency of GMM estimators in econometrics follows the same argmax/argmin continuous mapping theorem structure, with the GMM objective replacing the log-likelihood.

Linear Algebra. The uniform convergence condition can be expressed as operator norm convergence of the empirical measure operator to the population operator in the space of bounded functions on Θ.

Quantum Mechanics. Consistency of quantum state tomography estimators — the reconstructed density matrix converges to the true state — requires analogous identifiability (informationally complete measurements) and ULLN conditions on the measurement statistics.

Structural Compression

Invariant. $\hat{\theta}_n \xrightarrow{p} \theta$: the sampling distribution of $\hat{\theta}_n$ concentrates at θ as $n \to \infty$.

Mechanism. The WLLN drives sample averages to population means; under identifiability, the population criterion is uniquely maximized at the truth; the argmax continuous mapping theorem transfers uniform convergence of the criterion to convergence of the optimizer.

Cross-System Echo. Consistency of the MLE corresponds to the convergence of empirical risk minimizers to the population risk minimizer in statistical learning theory; the structural parallel is: identifiability $\leftrightarrow$ realizability, ULLN $\leftrightarrow$ Glivenko-Cantelli, argmax CMT $\leftrightarrow$ fundamental theorem of statistical learning.

Exercises

1. Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$. Prove that $(\bar{X}_n, S_n^2)$ is strongly consistent for (μ, σ^2).
2. Show that $\hat{\theta}_n = \bar{X}_n + 1/n$ is consistent for μ even though it is biased for every n.
3. Let $\hat{\theta}_n$ be consistent for θ and let g be continuous at θ. Prove that $g(\hat{\theta}_n)$ is consistent for $g(\theta)$.
4. Give an example of a strongly consistent estimator that is not unbiased for any finite n.
5. Verify the ULLN for the family $\{\log p_\lambda : \lambda > 0\}$ where $p_\lambda(x) = \lambda e^{-\lambda x}$ on $x > 0$, by checking equicontinuity on compact subsets of $(0, \infty)$.
6. Suppose Θ is not compact. Construct an example where the MLE is inconsistent because the argmax escapes to infinity, and explain which step of Wald's proof fails.
7. Argue, structurally, why identifiability is necessary for consistency of the MLE, connecting the information inequality to the KL divergence and explaining what fails in the proof when two distinct parameters yield the same distribution.

Statistics · Module 7 — Asymptotic Theory · Node 7.2
Concepts: convergence, maximum-likelihood-estimation, law-of-large-numbers
Prerequisites: 7.1, 3.1, 3.2
Enables: 7.3, 7.5

hbar.university — own.your.cognition