# Bernstein–von Mises Theorem

Statistics · Module 6 — Bayesian Inference

Node 6.7

## Position in Knowledge Graph

**System:** Statistics

**Structural Domain:** Bayesian Inference

**Node:** 6.7

The Bernstein–von Mises theorem is the fundamental asymptotic bridge between the Bayesian and frequentist frameworks. It establishes that in regular parametric models, the posterior distribution is asymptotically Gaussian, centered at the MLE, with variance equal to the inverse Fisher information — regardless of the prior. This node depends on the posterior (Node 6.1), the asymptotic theory of the MLE (Node 3.3), and the Fisher information (Node 7.2). It is the terminal node of Module 6: it closes the loop by showing that Bayesian and frequentist inference converge in large samples.

---

## Formal Definition

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} p_{\theta_0}$ where $\theta_0 \in \Theta \subseteq \mathbb{R}^k$ is the true parameter value. Let $\hat{\theta}_n$ denote the MLE and $I(\theta_0)$ the Fisher information matrix at $\theta_0$.

### Regularity Conditions

1. **Identifiability:** $\theta \neq \theta' \implies p_\theta \neq p_{\theta'}$.
2. **Smoothness:** The log-likelihood $\ell(\theta) = \log p(x \mid \theta)$ is twice continuously differentiable in a neighborhood of $\theta_0$.
3. **Fisher information:** $I(\theta_0) = -\mathbb{E}_{\theta_0}[\nabla^2 \ell(\theta_0)]$ is finite and positive definite.
4. **Prior positivity:** $\pi$ is continuous and positive at $\theta_0$: $\pi(\theta_0) > 0$.
5. **Consistency:** There exists a consistent sequence of estimators (e.g., the MLE $\hat{\theta}_n \to \theta_0$ in probability).
6. **Local asymptotic normality (LAN):** The log-likelihood ratio admits the expansion $\ell_n(\theta_0 + h/\sqrt{n}) - \ell_n(\theta_0) = h^\top \Delta_n - \frac{1}{2} h^\top I(\theta_0) h + o_P(1)$, where $\Delta_n \xrightarrow{d} \mathcal{N}(0, I(\theta_0))$.

### Theorem Statement

Under the regularity conditions above:

$$\left\| \pi\Big(\sqrt{n}(\theta - \hat{\theta}_n) \in \cdot \;\Big|\; X_1, \ldots, X_n\Big) - \mathcal{N}\big(0,\ I(\theta_0)^{-1}\big) \right\|_{\mathrm{TV}} \xrightarrow{P^n_{\theta_0}} 0.$$

Equivalently, the posterior distribution of $\theta$ given the data satisfies

$$\pi(\theta \mid X_1, \ldots, X_n) \approx \mathcal{N}\left(\hat{\theta}_n,\ \frac{I(\theta_0)^{-1}}{n}\right)$$

in the sense of total variation convergence under the true data-generating distribution.

---

## Intuition

As data accumulate, the likelihood dominates the prior. The log-posterior is approximately

$$\log \pi(\theta \mid x) \approx \ell_n(\theta) + \log \pi(\theta) - \text{const},$$

where $\ell_n(\theta) = \sum_{i=1}^n \log p(x_i \mid \theta)$ grows as $O(n)$ while $\log \pi(\theta)$ remains $O(1)$. A Taylor expansion of $\ell_n$ around the MLE produces a quadratic in $\theta$, which is the kernel of a Gaussian.

The theorem says the prior washes out: any two investigators with different priors (both positive and continuous at $\theta_0$) will have posteriors that converge to the same Gaussian as $n \to \infty$. This is the Bayesian analogue of the frequentist result that the MLE is asymptotically normal.

---

## Structural Role

The Bernstein–von Mises theorem is the asymptotic reconciliation of Bayesian and frequentist inference. It depends on the posterior (Node 6.1), asymptotic MLE theory (Node 3.3), and Fisher information (Node 7.2).

It has consequences for every other node in Module 6: - **Node 6.3:** The posterior mean, MAP, and posterior median all converge to the MLE at rate $O_p(n^{-1/2})$. - **Node 6.4:** Credible sets have asymptotic frequentist coverage. - **Node 6.6:** The BIC approximation to the Bayes factor is derived from the same Laplace expansion.

The theorem fails in nonparametric models, misspecified models, and high-dimensional settings, delineating the boundary of Bayesian-frequentist agreement.

---

## Mathematical Development

### Proof Sketch

The argument proceeds in four steps.

**Step 1: Posterior concentration.** By the consistency of the MLE and the positivity of the prior at $\theta_0$, the posterior mass outside any ball $B(\theta_0, \varepsilon)$ converges to zero in probability:

$$\pi\big(\|\theta - \theta_0\| > \varepsilon \mid X_{1:n}\big) \xrightarrow{P} 0.$$

This follows from the exponential growth of the likelihood ratio at the true value relative to distant parameter values.

**Step 2: Log-posterior expansion.** Expand the log-likelihood around the MLE:

$$\ell_n(\theta) = \ell_n(\hat{\theta}_n) - \frac{1}{2}(\theta - \hat{\theta}_n)^\top \hat{J}_n(\theta - \hat{\theta}_n) + R_n(\theta),$$

where $\hat{J}_n = -\nabla^2 \ell_n(\hat{\theta}_n)$ is the observed information and $R_n(\theta)$ is a remainder. By the law of large numbers, $\hat{J}_n/n \to I(\theta_0)$ in probability.

**Step 3: Prior contribution.** On the concentration set $\|\theta - \hat{\theta}_n\| = O(n^{-1/2})$, the prior satisfies $\log \pi(\theta) = \log \pi(\hat{\theta}_n) + O(n^{-1/2})$. Since $\log \pi$ is $O(1)$ and the quadratic term is $O(n)$, the prior contributes a negligible additive correction.

**Step 4: Gaussian identification.** Combining Steps 2 and 3:

$$\log \pi(\theta \mid x) = -\frac{1}{2}(\theta - \hat{\theta}_n)^\top \hat{J}_n(\theta - \hat{\theta}_n) + o_P(1)$$

on the set where $\sqrt{n}(\theta - \hat{\theta}_n)$ is bounded. This is the log-density of $\mathcal{N}(\hat{\theta}_n, \hat{J}_n^{-1})$. Since $\hat{J}_n/n \to I(\theta_0)$, the posterior is asymptotically $\mathcal{N}(\hat{\theta}_n, I(\theta_0)^{-1}/n)$ in total variation. $\square$

### Consequence: Posterior Mean and MLE Agreement

Since the posterior is approximately $\mathcal{N}(\hat{\theta}_n, I(\theta_0)^{-1}/n)$, its mean is

$$\mathbb{E}[\theta \mid X_{1:n}] = \hat{\theta}_n + O_P(n^{-1}).$$

The posterior mean and the MLE agree to first order. Their difference is $O_P(n^{-1})$, a second-order correction driven by the prior and the curvature of the likelihood.

### Consequence: Frequentist Coverage of Credible Sets

Let $C_n(x)$ be any $(1-\alpha)$-credible set derived from the posterior. By the total variation convergence:

$$\pi(\theta \in C_n \mid x) \to 1 - \alpha \implies \Phi\left(\frac{\sqrt{n}\, I(\theta_0)^{1/2}(\theta_0 - \hat{\theta}_n)}{\text{boundary}}\right) \to 1 - \alpha.$$

Since $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, I(\theta_0)^{-1})$ under $P_{\theta_0}$, the event $\theta_0 \in C_n(X)$ has probability converging to $1 - \alpha$:

$$P_{\theta_0}\big(\theta_0 \in C_n(X)\big) \to 1 - \alpha.$$

Asymptotically, Bayesian credible sets are frequentist confidence sets.

### Consequence: HPD and Wald Interval Agreement

The HPD credible set based on the limiting Gaussian posterior is the ellipsoid

$$\big\{\theta : n(\theta - \hat{\theta}_n)^\top I(\theta_0)(\theta - \hat{\theta}_n) \le \chi^2_{k,1-\alpha}\big\},$$

which coincides with the Wald confidence ellipsoid based on the MLE.

### Failure Modes

**Nonparametric models.** When $\Theta$ is infinite-dimensional (e.g., density estimation), the posterior can concentrate at rate slower than $1/\sqrt{n}$, and the Gaussian approximation fails. The Diaconis-Freedman theorem provides examples where the posterior concentrates on the wrong value.

**Misspecified models.** If the true distribution $P_0 \notin \{P_\theta : \theta \in \Theta\}$, the posterior concentrates around the KL-minimizer $\theta^* = \arg\min_\theta \mathrm{KL}(P_0 \| P_\theta)$, but the variance is no longer $I(\theta^*)^{-1}/n$ — a sandwich-type correction is needed.

**High-dimensional models.** When $k = k_n \to \infty$ with $n$, the prior contribution does not vanish relative to the likelihood, and the Bernstein–von Mises conclusion can fail even for parametric models.

**Boundary parameters.** If $\theta_0$ lies on the boundary of $\Theta$, the posterior may not be asymptotically normal (it may concentrate on a half-space).

---

## Worked Examples

### Example 1: Normal Mean

Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, 1)$ with prior $\mu \sim \mathcal{N}(0, \tau^2)$. The exact posterior is $\mu \mid x \sim \mathcal{N}(\mu_n, \sigma_n^2)$ with

$$\sigma_n^2 = \frac{1}{1/\tau^2 + n} = \frac{\tau^2}{1 + n\tau^2}, \qquad \mu_n = \sigma_n^2 \cdot n\bar{x} = \frac{n\tau^2}{1 + n\tau^2}\bar{x}.$$

As $n \to \infty$: $\sigma_n^2 \to 1/n$ and $\mu_n \to \bar{x} = \hat{\mu}_n$. The Fisher information is $I(\mu) = 1$, so the BvM limit is $\mathcal{N}(\hat{\mu}_n, 1/n)$. The exact posterior converges to this limit, confirming the theorem. The convergence rate of the prior correction is $O(1/n)$: the term $1/\tau^2$ in the posterior precision becomes negligible relative to $n$.

### Example 2: Bernoulli Model and Coverage Verification

Let $X_i \sim \text{Bernoulli}(\theta_0)$ with $\theta_0 = 0.3$. Prior: $\text{Beta}(2, 2)$. For large $n$, the posterior is approximately

$$\theta \mid x \mathrel{\dot\sim} \mathcal{N}\left(\hat{\theta}_n,\ \frac{\hat{\theta}_n(1 - \hat{\theta}_n)}{n}\right),$$

since $I(\theta) = 1/[\theta(1 - \theta)]$.

For $n = 500$: if $s = 155$ successes, then $\hat{\theta}_n = 0.31$. The asymptotic posterior variance is $0.31 \cdot 0.69/500 \approx 0.000428$. The 95% HPD interval is approximately $0.31 \pm 1.96\sqrt{0.000428} = 0.31 \pm 0.041 = [0.269, 0.351]$.

The exact posterior is $\text{Beta}(157, 347)$ with mean $157/504 \approx 0.3115$ and variance $157 \cdot 347/(504^2 \cdot 505) \approx 0.000429$. The Gaussian approximation is excellent: the prior contributes only 4 pseudo-observations out of 504 total.

The frequentist coverage: $\theta_0 = 0.3 \in [0.269, 0.351]$, and repeated sampling confirms coverage near 95%.

---

## Cross-Links

- **Linear Algebra:** The quadratic expansion of the log-likelihood around the MLE is governed by the Hessian matrix $\hat{J}_n$. The Bernstein–von Mises theorem requires this matrix to converge (after scaling by $1/n$) to the positive definite Fisher information matrix, a spectral condition.
- **AI:** The Bernstein–von Mises theorem motivates Laplace approximations for Bayesian deep learning: approximate the posterior by a Gaussian centered at the MAP estimate with covariance given by the inverse Hessian. The theorem's failure in high dimensions explains why this approximation is unreliable for overparameterized networks.
- **Economics:** The theorem justifies the common econometric practice of interpreting Bayesian credible intervals as confidence intervals in large samples. It also explains why prior choice matters less in large macro panels than in small cross-sections.
- **Quantum Mechanics:** Quantum local asymptotic normality (quantum LAN) extends the Bernstein–von Mises theorem to quantum statistical models, showing that the posterior over quantum states converges to a Gaussian shift experiment in the limit of many copies of the state.

---

## Structural Compression

**Invariant:** In regular parametric models, the large-sample posterior is asymptotically $\mathcal{N}(\hat{\theta}_n, I(\theta_0)^{-1}/n)$, independent of the prior.

**Mechanism:** A second-order Taylor expansion of the log-likelihood around the MLE produces an $O(n)$ quadratic that dominates the $O(1)$ prior contribution. The resulting Gaussian shape is determined entirely by the Fisher information geometry. The prior survives only in the $O(n^{-1})$ correction to the posterior mean.

**Cross-System Echo:** The Bernstein–von Mises theorem is the statistical manifestation of the Laplace approximation to integrals dominated by a sharp peak. In physics, the saddle-point approximation to partition functions is the same phenomenon. In AI, the Laplace approximation to the neural network posterior and the correspondence between MAP and MLE in the large-data regime are direct applications of this asymptotic equivalence.

---

## Exercises

1. Verify the Bernstein–von Mises conclusion for the Poisson model: let $X_i \sim \text{Poisson}(\lambda_0)$ with prior $\text{Gamma}(\alpha, \beta)$. Show that the exact posterior is Gamma and that it converges in total variation to $\mathcal{N}(\hat{\lambda}_n, \hat{\lambda}_n/n)$ as $n \to \infty$.
2. Show that the Bernstein–von Mises conclusion implies the posterior variance satisfies $\text{Var}(\theta \mid X) \to I(\theta_0)^{-1}/n$ in probability. Derive the rate of convergence of the difference.
3. Construct an explicit example where the Bernstein–von Mises theorem fails: take a uniform prior on a misspecified model where the true density is not in the parametric family, and show the posterior concentrates at the KL-minimizer but with incorrect variance.
4. For the normal location model with known variance and prior $\mathcal{N}(\mu_0, \tau_0^2)$, compute the exact total variation distance between the posterior and the BvM Gaussian limit as a function of $n$, and verify it is $O(n^{-1})$.
5. Prove that the second-order correction to the posterior mean is $\mathbb{E}[\theta \mid x] - \hat{\theta}_n = O_P(n^{-1})$ by expanding the posterior mean in powers of $1/n$ using the exact posterior for a conjugate exponential family.
6. Explain why the Bernstein–von Mises theorem fails in the model $X_i \sim \text{Uniform}(0, \theta)$: identify which regularity condition is violated, and describe the actual asymptotic behavior of the posterior.
7. Argue, structurally, why the Bernstein–von Mises theorem establishes a hierarchy between Bayesian and frequentist inference: in the regular parametric setting, the Bayesian approach subsumes the frequentist one (credible sets are asymptotically confidence sets, posterior mean agrees with MLE), but the theorem's failure modes in nonparametric, misspecified, and high-dimensional settings show where this subsumption breaks down and the two frameworks genuinely diverge.

---

*Statistics · Module 6 — Bayesian Inference · Node 6.7*
**Concepts:** prior-posterior, fisher-information, central-limit-theorem, maximum-likelihood-estimation
**Prerequisites:** 6.1, 3.3, 7.2

*hbar.university — own.your.cognition*