Priors and Posteriors

Statistics · Module 6 — Bayesian Inference

Node 6.1

Position in Knowledge Graph

System: Statistics

Structural Domain: Bayesian Inference

Node: 6.1

The prior-to-posterior update is the foundational operation of Bayesian inference. Every subsequent node in this module — conjugate families, point estimation, credible sets, predictive distributions, model comparison, and asymptotic reconciliation — is a derived consequence of the posterior distribution defined here. This node depends on the parametric model framework (Node 2.1), the likelihood function (Node 2.3), and the law of total probability (Node 1.7).

Formal Definition

Let $X = (X_1, \ldots, X_n)$ be a random sample with joint density or mass function $p(x \mid \theta)$ for $\theta \in \Theta \subseteq \mathbb{R}^k$.

Prior Distribution

A **prior distribution** $\pi(\theta)$ is a probability measure on $(\Theta, \mathcal{B}(\Theta))$ specified before observing X. It encodes the state of knowledge about θ prior to data collection. A prior is **proper** if $\int_\Theta \pi(\theta)\, d\theta = 1$, and **improper** if this integral diverges.

Posterior Distribution

Given observed data $X = x$, the **posterior distribution** is

$$\pi(\theta \mid x) = \frac{p(x \mid \theta)\, \pi(\theta)}{m(x)},$$

where the **marginal likelihood** (or evidence) is

$$m(x) = \int_\Theta p(x \mid \theta)\, \pi(\theta)\, d\theta.$$

The posterior exists as a proper distribution whenever $0 < m(x) < \infty$.

Jeffreys Prior

The **Jeffreys prior** is defined as

$$\pi^J(\theta) \propto \sqrt{\det I(\theta)},$$

where $I(\theta)$ is the Fisher information matrix. This prior is invariant under reparametrization: if $\phi = g(\theta)$ is a smooth bijection, then $\pi^J(\phi) \propto \sqrt{\det I_\phi(\phi)}$, where I_ϕ is the Fisher information in the ϕ-parametrization.

Intuition

The posterior is a compromise between what we believed before seeing data (the prior) and what the data tell us (the likelihood). In regions where the likelihood is large, the posterior concentrates mass; in regions where the prior assigns little mass, the posterior is suppressed even if the likelihood is moderate.

The marginal likelihood $m(x)$ serves purely as a normalizing constant for the posterior but carries independent significance as the model's overall predictive score for the observed data (Node 6.6).

Improper priors are useful default specifications — they express maximal pre-data ignorance within a parametric family — but require verification that the resulting posterior is proper.

The Jeffreys prior is the canonical "objective" prior: it assigns prior mass in a way that respects the geometry of the statistical model, treating reparametrizations as coordinate changes on the same underlying manifold.

Structural Role

The prior-to-posterior update is the complete description of Bayesian learning. All Bayesian procedures — point estimates (Node 6.3), credible sets (Node 6.4), posterior predictive distributions (Node 6.5), and Bayes factors (Node 6.6) — are functionals of the posterior.

This node requires: parametric models (Node 2.1), the likelihood principle (Node 2.3), and marginalization via the law of total probability (Node 1.7).

This node enables: every subsequent node in Module 6. The posterior is also the bridge to asymptotic reconciliation with frequentist inference via the Bernstein–von Mises theorem (Node 6.7).

Mathematical Development

Derivation of Bayes' Theorem for Densities

Starting from the joint density of (X, θ):

$$p(x, \theta) = p(x \mid \theta)\, \pi(\theta) = \pi(\theta \mid x)\, m(x).$$

Solving for $\pi(\theta \mid x)$ yields Bayes' theorem. The key requirement is $m(x) > 0$, which holds whenever the data are compatible with at least one parameter value that receives prior mass.

Sequential Updating

For i.i.d. data, the posterior after n observations can be computed sequentially. Let $\pi_0(\theta) = \pi(\theta)$ and define

$$\pi_k(\theta) \propto p(x_k \mid \theta)\, \pi_{k-1}(\theta), \quad k = 1, \ldots, n.$$

Then $\pi_n(\theta) = \pi(\theta \mid x_1, \ldots, x_n)$. The order of updating does not matter:

$$\pi(\theta \mid x_1, \ldots, x_n) \propto \pi(\theta) \prod_{i=1}^{n} p(x_i \mid \theta).$$

Sufficiency and the Posterior

If $T(X)$ is a sufficient statistic for θ, then by the factorization theorem $p(x \mid \theta) = g(T(x), \theta)\, h(x)$, so

$$\pi(\theta \mid x) = \frac{g(T(x), \theta)\, h(x)\, \pi(\theta)}{\int g(T(x), \theta)\, h(x)\, \pi(\theta)\, d\theta} = \frac{g(T(x), \theta)\, \pi(\theta)}{\int g(T(x), \theta)\, \pi(\theta)\, d\theta} = \pi(\theta \mid T(x)).$$

The posterior depends on the data only through the sufficient statistic.

Proper vs Improper Priors

An improper prior $\pi(\theta)$ with $\int_\Theta \pi(\theta)\, d\theta = \infty$ can still yield a proper posterior, provided $m(x) = \int p(x \mid \theta)\, \pi(\theta)\, d\theta < \infty$ for almost all x. For example, the flat prior $\pi(\mu) = 1$ on $\mathbb{R}$ for the normal mean yields proper posterior $\mu \mid x \sim \mathcal{N}(\bar{x}, \sigma^2/n)$.

Derivation of the Jeffreys Prior

Consider a one-parameter model with Fisher information $I(\theta) = -\mathbb{E}_\theta[\ell''(\theta)]$, where $\ell(\theta) = \log p(X \mid \theta)$. Under reparametrization $\phi = g(\theta)$, the Fisher information transforms as

$$I_\phi(\phi) = I(\theta) \left(\frac{d\theta}{d\phi} \right)^2.$$

The prior density transforms as $\pi_\phi(\phi) = \pi(\theta) \left| \frac{d\theta}{d\phi} \right|$. Setting $\pi(\theta) \propto \sqrt{I(\theta)}$ gives

$$\pi_\phi(\phi) \propto \sqrt{I(\theta)} \left| \frac{d\theta}{d\phi} \right| = \sqrt{I(\theta) \left(\frac{d\theta}{d\phi} \right)^2} = \sqrt{I_\phi(\phi)}.$$

Hence the Jeffreys prior is invariant under reparametrization.

Posterior Concentration

As $n \to \infty$, the posterior concentrates around the true parameter value θ_0. Informally, the log-posterior is

$$\log \pi(\theta \mid x) = \sum_{i=1}^{n} \log p(x_i \mid \theta) + \log \pi(\theta) - \log m(x).$$

The likelihood sum grows as $O(n)$ while $\log \pi(\theta)$ is $O(1)$, so the prior is overwhelmed. Under regularity conditions, this is made precise by the Bernstein–von Mises theorem (Node 6.7).

Worked Examples

Example 1: Normal Mean with Flat Prior

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ with σ^2 known. Use the improper flat prior $\pi(\mu) \propto 1$.

The likelihood is

$$p(x \mid \mu) \propto \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right) = \exp\Big(-\frac{n}{2\sigma^2} (\mu - \bar{x})^2 \Big) \cdot C(x),$$

where $C(x)$ does not depend on μ. Thus

$$\pi(\mu \mid x) \propto \exp\Big(-\frac{n}{2\sigma^2}(\mu - \bar{x})^2\Big),$$

which is the kernel of $\mathcal{N}(\bar{x}, \sigma^2/n)$. The posterior is proper despite the improper prior.

Example 2: Jeffreys Prior for the Bernoulli Model

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\theta)$, $\theta \in (0,1)$. The Fisher information is

$$I(\theta) = \frac{1}{\theta(1-\theta)}.$$

The Jeffreys prior is $\pi^J(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$, which is $\text{Beta}(1/2, 1/2)$. With $s = \sum x_i$ successes, the posterior is

$$\pi(\theta \mid x) = \text{Beta}\big(\tfrac{1}{2} + s,\ \tfrac{1}{2} + n - s\big).$$

The posterior mean is $(s + 1/2)/(n+1)$, which shrinks the MLE s/n toward $1/2$ by an amount that vanishes as $n \to \infty$.

Cross-Links

- **Linear Algebra:** The Fisher information matrix $I(\theta)$ is a Riemannian metric on the parameter manifold; the Jeffreys prior is the volume element of this metric, connecting Bayesian inference to differential geometry on statistical manifolds.
- **AI:** Variational inference approximates the posterior $\pi(\theta \mid x)$ by optimizing over a tractable family when the marginal likelihood $m(x)$ is intractable. The prior-to-posterior update is the exact operation that variational and MCMC methods approximate.
- **Economics:** Bayesian agents update beliefs via Bayes' rule; rational expectations equilibria assume all agents compute posteriors from public signals. The prior encodes heterogeneous beliefs across agents.
- **Quantum Mechanics:** State update upon measurement (the Luders rule) is the non-commutative analogue of Bayes' theorem. The prior is a density matrix; the posterior is the post-measurement state.

Structural Compression

Invariant: $\pi(\theta \mid x) \propto p(x \mid \theta)\,\pi(\theta)$. The posterior is proportional to likelihood times prior.

Mechanism: Bayes' theorem normalizes the product $p(x \mid \theta)\,\pi(\theta)$ over Θ to produce a proper probability measure. The likelihood reweights prior mass according to data compatibility. The normalizing constant $m(x)$ ensures coherence.

Cross-System Echo: The posterior update is the universal belief revision rule: it appears as Bayesian filtering in signal processing, as the E-step kernel in EM algorithms, as density matrix conditioning in quantum information, and as belief propagation in graphical models.

Exercises

1. Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Exponential}(\lambda)$. Derive the posterior under the Jeffreys prior $\pi(\lambda) \propto 1/\lambda$. Verify the posterior is proper for $n \geq 1$.

2. Prove that the posterior distribution depends on the data only through a sufficient statistic by applying the factorization theorem.

3. Show that the Jeffreys prior for the normal variance σ^2 (with known mean) is $\pi(\sigma^2) \propto 1/\sigma^2$. Derive the resulting posterior and identify its distributional family.

4. Compute the posterior for θ in the model $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Uniform}(0, \theta)$ under the prior $\pi(\theta) \propto \theta^{-1}$. Show that the MLE and the posterior mode differ.

5. Demonstrate sequential updating: let X_1, X_2, X_3 be i.i.d. Bernoulli(θ) with $\pi(\theta) = \text{Beta}(2, 2)$. Compute the posterior after each observation for the sequence $(1, 0, 1)$ and verify the final posterior matches the batch computation.

6. Prove that if $\pi(\theta)$ is a proper prior and $p(x \mid \theta)$ is a bounded likelihood, then the posterior $\pi(\theta \mid x)$ is proper for all x.

7. Argue, structurally, why the prior is asymptotically irrelevant in regular parametric models: identify the rate at which the likelihood dominates the prior, and explain why this argument fails in nonparametric settings where Θ is infinite-dimensional.

Statistics · Module 6 — Bayesian Inference · Node 6.1
Concepts: prior-posterior, bayesian-updating, likelihood, conditional-probability
Prerequisites: 2.1, 2.3, 1.7
Enables: 6.2, 6.3, 6.4, 6.5, 6.6

hbar.university — own.your.cognition