MISCELLANEOUS

Bayesian Fundamentals

ar062 · 28 June 2026

A short primer for the vocabulary the neural-Bayesian papers in the Bayesian Literature Review lean on without redefining. The aim is not to teach Bayesian statistics from scratch — it is to fix notation, name the moves, and flag where the hard parts are, so a reader recognises each term on contact when they hit it in Ma, Fiser, Lengyel, or Echeveste.

The core equation

Given a generative model — a forward story for how unobserved causes θ produce observations x — Bayes' rule inverts the story:

$$p(\theta \mid x) = \frac{p(x \mid \theta)p(\theta)}{p(x)} \propto p(x \mid \theta)p(\theta) \quad (1)$$

- θ — the latent (unobserved) cause to be inferred (a tilt angle, a phoneme, a depth);
- x — the observation (the retinal image, the cochlear input);
- $p(\theta)$ — the **prior**, what was believed about θ before seeing x;
- $p(x \mid \theta)$ — the **likelihood**, how probable this x would be if θ were the cause;
- $p(\theta \mid x)$ — the **posterior**, the updated belief about θ after seeing x;
- $p(x) = \int p(x \mid \theta)p(\theta)\,\mathrm{d}\theta$ — the **evidence** (or marginal likelihood), the normaliser.

The proportionality at the right is doing most of the work in practice: every algorithm in the rest of this article exists to avoid computing $p(x)$ exactly.

Generative model vs recognition

A **generative model** is the brain's (or the modeller's) story about how the world produces the data: causes → observations. **Recognition** (or **inference**) is the reverse — recover $p(\theta \mid x)$ from x. These are different computations and the brain need not implement them the same way; a sensory area can run a fast amortised recognition map even if the underlying generative model is much richer. When neural-Bayesian papers say "the cortex represents the posterior", they mean the recognition output.

Marginalisation, and why it's the hard part

Anything you don't care about you must integrate out:

$$p(\theta_1 \mid x) = \int p(\theta_1, \theta_2 \mid x)\,\mathrm{d}\theta_2 \quad (2)$$

- θ_1 — the quantity of interest (the depth of a surface);
- θ_2 — the nuisance variables (lighting, surface reflectance) you marginalise over.

For high-dimensional θ this integral is intractable. Every approximate-inference algorithm below is, at heart, a different way to dodge it.

Conjugacy and the exponential family

A likelihood–prior pair is **conjugate** if the posterior has the same parametric form as the prior. Conjugate pairs (Gaussian–Gaussian, beta–Bernoulli, Dirichlet–multinomial) give closed-form posterior updates and are the only place Bayes is genuinely cheap. They live inside the **exponential family**:

$$p(x \mid \eta) = h(x) \exp(\eta^\top T(x) - A(\eta)) \quad (3)$$

- η — the **natural parameters**;
- $T(x)$ — the **sufficient statistics** (the only function of x the posterior cares about);
- $A(\eta)$ — the **log-partition** (normaliser), whose derivatives give the moments of T;
- $h(x)$ — a base measure.

This matters for cortex because the Probabilistic Population Code framework (Ma et al. 2006, in ar007) writes posteriors as exponential families whose natural parameters are population firing rates: $\eta \equiv \boldsymbol{r}$. Multiplying two PPCs (combining cues) becomes adding their firing rates — the framework's central, very pretty claim.

Three flavours of approximate inference

When the posterior is not conjugate, you approximate. Three families dominate; the neural-Bayesian papers each pick one:

- **Laplace approximation.** Find the posterior mode $\hat{\theta}$, fit a Gaussian whose covariance is $\left(-\nabla^2 \log p(\theta \mid x)\right)^{-1}$ at the mode. Cheap, local, breaks down when the posterior is multimodal.
- **Variational inference (VI).** Pick a tractable family $q_\varphi(\theta)$ and minimise $\mathrm{KL}\big(q_\varphi(\theta) \parallel p(\theta \mid x)\big)$ in φ. Turns inference into optimisation. Fast, scalable, but biased — the bias is whatever p has that q cannot represent.
- **Markov chain Monte Carlo (MCMC).** Construct a Markov chain whose stationary distribution *is* the posterior, then run it. Unbiased in the limit, slow to converge. Variants: Metropolis–Hastings, Gibbs sampling, Langevin dynamics (gradient + noise), Hamiltonian Monte Carlo (gradient + momentum). The Lengyel-group "sampling cortex" papers in ar007 argue that cortical dynamics are a biological MCMC chain — Aitchison & Lengyel cast E/I circuits as a Hamiltonian sampler, with the inhibitory population playing the momentum role.

KL divergence

The asymmetric "distance" between two distributions:

$$\mathrm{KL}(q \parallel p) = \int q(\theta) \log \frac{q(\theta)}{p(\theta)} \, \mathrm{d}\theta \geq 0, \quad \text{with equality iff } \; q = p. \quad (4)$$

Asymmetric because $\mathrm{KL}(q \parallel p) \neq \mathrm{KL}(p \parallel q)$ in general. VI minimises $\mathrm{KL}(q \parallel p)$ (forward KL) which makes q *mode-seeking* (it concentrates on a single mode rather than averaging across modes); some methods minimise $\mathrm{KL}(p \parallel q)$ (reverse KL) which makes q *mass-covering*. The choice shows up in what the approximate posterior looks like.

Posterior summaries (the things a brain might actually read out)

A posterior is a distribution; downstream computation usually wants a number or two. The standard summaries:

- **MAP estimate.** The mode: $\theta_{\text{MAP}} = \arg\max_\theta p(\theta \mid x)$. A single best guess.
- **Posterior mean.** $\mathbb{E}[\theta \mid x]$. The Bayes-optimal point estimate under squared-error loss.
- **Posterior variance / credible interval.** The width — how *uncertain* the inference is. The whole point of being Bayesian rather than maximum-likelihood; if downstream computation needs to know "should I bet on this?" it needs the spread, not just the mean.

The neural-Bayesian split is partly about how the cortex represents this last one — as a population gain (PPC) or as trial-to-trial variability (sampling).

How this lands in the cortex

Three claims recur in the reading list, in roughly this order of strength:

1. **Cortical responses look Bayesian behaviourally.** Cue integration, prior-weighted perception, multisensory fusion — all match Bayes-optimal predictions in many tasks. The papers in ar007 take this as given.
2. **The cortical *representation* of probability is one of two flavours.** Either parametric (firing rates encode posterior parameters — PPC) or sample-based (instantaneous activity is itself a sample). These are different empirical predictions about neural variability, and the camps disagree.
3. **The cortical *dynamics* implement the inference algorithm.** The sampling camp claims recurrent E/I circuits run Langevin or HMC; the PPC camp claims linear combinations of population activity perform the algebra. Echeveste et al. 2020 is the cleanest version of the dynamical claim.

This article stops short of those claims — see ar007 for the reading list that develops them.