Chapter 5: Bayesian Inference

Advanced Topics in Statistical Machine Learning

Eugenio Clerico

Hilary 2026

eugenio.clerico@stats.ox.ac.uk

Bayesian Probability is All About Belief

Frequentist Probability

The frequentist interpretation of probability is that it is the **average proportion of the time an event will occur if a trial is repeated infinitely many times**.

Bayesian Probability

The Bayesian interpretation of probability is that it is the **subjective belief that an event will occur in the presence of incomplete information**

In ML this means that rather than looking for a **single** optimal value for the parameters, we seek to quantify our **degree of belief** for each possible value in the parameter space. The parameters themselves become *random quantities*.

Maximum Likelihood Estimator (MLE)

We have a parameterised model $(p_\theta)_{\theta\in\Theta}$, whose probability densities try to explain how the observed data $\hat{\mathcal{D}}$ have been generated. We define the likelihood $\ell(\theta) = p_\theta(\hat{\mathcal{D}})$.
MLE: find the **single best** point estimate $\hat{\theta}_{\mathrm{MLE}}$ that maximises ℓ:

$$\hat{\theta}_{\mathrm{MLE}} = \arg\max_{\theta\in\Theta} \ell(\theta)\,.$$

- We typically maximise the **log-likelihood**, $\log p_\theta(\hat{\mathcal{D}})$, for numerical stability and to turn products into sums.
- Parameters are treated as *fixed but unknown* constants.
- **Limitation:** It ignores prior knowledge and provides no measure of uncertainty. This can be partially addressed via **regularisation** or, more fundamentally, by adopting the **Bayesian perspective**, which replaces the point estimates with a full distribution representing our belief/confidence.

Incorporating Prior Knowledge

In contrast to MLE, the Bayesian approach begins with a **prior distribution** $\pi(\theta)$ on the parameters/models. In this way the model treats both the data and the parameter as random variables, with joint density $\Pi(\theta, \mathcal{D}) = p_\theta(\mathcal{D})\pi(\theta)$.[1] Yet, the randomness of θ is **epistemic** rather than intrinsic, as it only encodes our believes and lack of knowledge.

- π encodes our initial belief about the parameters *before* observing any data.
- π defines which regions of the parameter space are more plausible based on domain expertise or previous experiments.
- From a technical standpoint, a prior can act as a regulariser (e.g., a Gaussian prior centred at zero is equivalent to L2 regularisation).

[1] We implicitly assume that $(\theta, \mathcal{D}) \mapsto p_\theta(\mathcal{D})$ is measurable.

Bayes' Rule and Inference

We combine our **prior belief** with the **information from the data** (likelihood) using Bayes' Rule:

$$\pi(\theta|\mathcal{D}) = \frac{p_\theta(\mathcal{D})\,\pi(\theta)}{p_\pi(\mathcal{D})}\,.$$

- **Posterior** $\pi(\theta|\mathcal{D})$**:** Our updated belief about θ after incorporating data.
- **Evidence** $p_\pi(\mathcal{D})$**:** A normalisation constant (marginal likelihood) calculated as $p_\pi(\mathcal{D}) = \int p_\theta(\mathcal{D})\pi(\theta)\mathrm{d}\theta$.
- Since $p(\mathcal{D})$ does not depend on θ, we often focus on the unnormalised posterior: $\pi(\theta|\mathcal{D}) \propto p_\theta(\mathcal{D})\pi(\theta)$

Finding the posterior is what we call **Bayesian inference**.

Maximum a Posteriori (MAP) Estimation

The MAP estimate is a single parameter that maximises the posterior density for the observed data $\hat{\mathcal{D}}$. It can be seen as a regularised MLE.

Definition

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} \pi(\theta|\hat{\mathcal{D}}) = \arg\max_{\theta} p_{\theta}(\hat{\mathcal{D}})\pi(\theta)$$

- It uses the prior into the optimisation, unlike vanilla MLE.
- It is a **point estimate**: single value instead of full distribution.
- **Drawback:** much less informative than the posterior.

Under mild conditions, in the limit of infinite data, the likelihood dominates the prior, and the MAP estimate converges to the MLE.

Example: A Van Gogh Discovery?

You find a painting signed **“Vincent”** in an old frame. What is the probability that it is a genuine Van Gogh?

Well... It depends heavily on **where** you found it and the **prevalence** of replicas in that context!

Example: Discovery at a Flea Market

- Let $\theta = 1$ be 'Authentic' and $\theta = 0$ be 'Forgery/Replica'.
- In a random flea market, perhaps only 1 in $10,000$ paintings is a lost masterpiece. We set our prior: $\pi(\theta = 1) = 0.0001$.
- Assume a *signature analysis* is 95% accurate: it identifies 95% of originals but also gives a positive for 5% of forgeries: $p_{\theta=1}(\text{pos}) = 0.95$ and $p_{\theta=0}(\text{pos}) = 0.05$.

Applying Bayes' rule:

$$\begin{aligned}\pi(\theta = 1|\text{pos}) &= \frac{p_{\theta=1}(\text{pos})\pi(\theta = 1)}{p_{\theta=1}(\text{pos})\pi(\theta = 1) + p_{\theta=0}(\text{pos})\pi(\theta = 0)} \\ &= \frac{0.95 \times 0.0001}{0.95 \times 0.0001 + 0.05 \times 0.9999} \\ &\approx 0.0019\end{aligned}$$

Even with a positive signature analysis, there is only a 0.2% chance it is real. The low prior dominates the result.

Example: Inside the Van Gogh Museum, Amsterdam

- Now imagine you are looking at a painting inside the **Van Gogh Museum**.
- Because of the curated environment, your prior belief that a painting there is authentic is very high: $\pi(\theta = 1) = 0.99$.

Bayes' rule with the same signature analysis:

$$\begin{aligned}\pi(\theta = 1|\text{pos}) &= \frac{0.95 \times 0.99}{0.95 \times 0.99 + 0.05 \times 0.01} \\ &\approx 0.9995\end{aligned}$$

Now, the 'positive' result confirms what you already strongly suspected.

Take home: **The prior matters!**

Multiple Observations: Using the Posterior as the Prior

- One of the key characteristics of Bayes' rule is that it is **self-similar** under multiple observations
- We can use the posterior after our first observation as the prior when considering the next: (with some abuse of notation...)

$$\begin{aligned}\pi(\theta|\mathcal{D}_1, \mathcal{D}_2) &= \frac{p_\theta(\mathcal{D}_2|\mathcal{D}_1)\pi(\theta|\mathcal{D}_1)}{p_\pi(\mathcal{D}_2|\mathcal{D}_1)} \\ &= \frac{p_\theta(\mathcal{D}_1, \mathcal{D}_2)\pi(\theta)}{p_\pi(\mathcal{D}_1, \mathcal{D}_2)}\end{aligned}$$

- We can thinking of this as continuous updating of beliefs as we receive more information

Making Predictions

- Prediction in Bayesian models is done using the **posterior predictive distribution**
- This is defined by taking the expectation of a predictive model for new data, $p_\theta(\mathcal{D}^*)$, with respect to the posterior:

$$p_\pi(\mathcal{D}^*|\mathcal{D}) = \int \Pi(\mathcal{D}^*, \theta|\mathcal{D}) d\theta = \int p_\theta(\mathcal{D}^*|\mathcal{D}) \pi(\theta|\mathcal{D}) \mathrm{d}\theta$$
$$= \mathbb{E}_{\theta \sim \pi(\cdot|\mathcal{D})}[p_\theta(\mathcal{D}^*)].$$

- Note here that we are making the standard assumption that the data is conditionally independent given θ.
- Prediction is often done dependent on an input point such that we actually calculate $p_\pi(y|x, \mathcal{D}) = \mathbb{E}_{\theta \sim \pi(\cdot|\mathcal{D})}[p_\theta(y|x)]$.
- Note that this can be very expensive: typically requires approximations.

Bayesian Model Selection

Given two models $\mathcal{M}_1 = (p_\theta)_{\theta\in\Theta}$ and $\mathcal{M}_2 = (q_\xi)_{\xi\in\Xi}$, with prior π_1 on Θ and π_2 on Ξ, which is most relevant after observing data $\mathcal{D}$?

- **The Model Evidence:** For each model we compute the predictive prior density at the observed data:
  $E_{\mathcal{M}_1}(\mathcal{D}) = \int_\Theta p_\theta(\mathcal{D})\pi_1(\theta)d\theta \qquad E_{\mathcal{M}_2}(\mathcal{D}) = \int_\Xi q_\xi(\mathcal{D})\pi_2(\theta)d\xi$
- **Bayes Factor:** Models are compared using the ratio of their evidences: $K = \frac{E_{\mathcal{M}_1}(\mathcal{D})}{E_{\mathcal{M}_2}(\mathcal{D})}$
- **Bayesian Occam's Razor:** The evidence naturally penalises models that are "too flexible" (prior mass too spreaded).
- **Hierarchical Prior:** If we define a prior ν over the models themselves, we obtain the posterior probability of model $\mathcal{M}_k$:

$$\nu(\mathcal{M}_k|\mathcal{D}) = \frac{E_{\mathcal{M}_k}(\mathcal{D})\nu(\mathcal{M}_k)}{\sum_j E_{\mathcal{M}_j}(\mathcal{D})\nu(\mathcal{M}_j)}$$

Bayesian Occam's Razor

Occam's Razor states that if two explanations are able to explain a set of observations, the simpler one should be preferred.
Imagine a hypothetical order on datasets where they get more complicated as we move away from the origin. The model with highest evidence is the one that is powerful enough to explain that data but not anything more complicated.

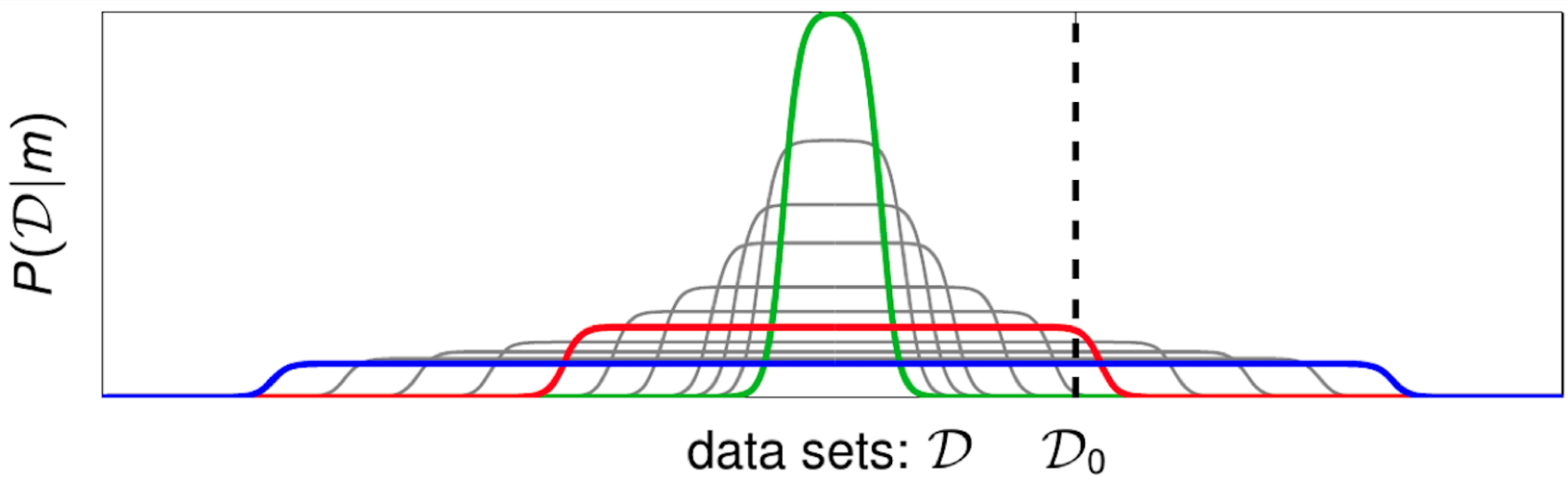

Image credit: Maneesh Sahani

Full Bayesian Inference is Hard

- By Bayes' rule, the posterior is $\pi(\theta|\mathcal{D}) \propto p_\theta(\mathcal{D})\pi(\theta)$. However, calculating the normalisation constant is akin to solving a high-dimensional integral:

$$p_\pi(\mathcal{D}) = \int p_\theta(\mathcal{D})\pi(\theta)\mathrm{d}\theta$$

- Without knowing $p_\pi(\mathcal{D})$, we lack the normalisation constant required to evaluate the density.
- Without $p_\pi(\mathcal{D})$ we have no concept of how much probability mass is missing, or how significant a point is compared to the distribution as a whole.
- **Computational Complexity:** In general, Bayesian inference is an **NP-hard problem**.

Monte Carlo Estimators

- If we can draw N samples from the posterior, we can form **Monte Carlo estimates** for any expectation:

$$\mathbb{E}_{\theta \sim \pi(\cdot|\mathcal{D})}[f(\theta)] \approx \frac{1}{N} \sum_{n=1}^{N} f(\hat{\theta}_n) \text{ where } \hat{\theta}_n \sim \pi(\cdot|\mathcal{D})$$

- This produces an estimator whose **mean squared error** is $O(1/N)$.
- **Issue:** We cannot usually draw exactly from the posterior.
- We instead construct methods to produce **approximate samples** via two main strategies:
  - **Weighted samples:** Equivalent to exact samples in expectation (e.g. Importance Sampling).
  - **Markov chains:** Generating a sequence of samples that converges to the posterior (e.g. MCMC).

Importance Sampling (IS)

- Importance sampling uses a **proposal distribution** ρ for θ to draw samples.
- We apply **corrective importance weights** to account for samples drawn from the *wrong* distribution:

$$\mathbb{E}_{\theta\sim\pi(\cdot|\mathcal{D})}[f(\theta)] = \mathbb{E}_{\theta\sim\rho}\left[\frac{\pi(\theta|\mathcal{D})}{\rho(\theta)}f(\theta)\right]$$

- **The Self-Normalisation Trick:** $p(\mathcal{D})$ is unknown, so we use unnormalised weights $w(\theta) = \Pi(\theta, \mathcal{D})/\rho(\theta)$:
  1. Draw N i.i.d. samples $\hat{\theta}_n \sim q$.
  2. Assign weights $w_n = \Pi(\hat{\theta}_n, \mathcal{D})/\rho(\hat{\theta}_n)$.
  3. Self-normalise: $\bar{w}_n = w_n/(\sum_{m=1}^{N} w_m)$.
  4. Estimate: $\sum_{i=1}^{N} \bar{w}_n f(\hat{\theta}_n)$
- This avoids direct calculation of the marginal likelihood while providing a consistent estimator.

Variational Inference and the ELBO

- **Inference as Optimisation:** Rather than sampling, we introduce a parameterised surrogate distribution ρ_ϕ and minimise its 'distance' to the true posterior.
- **KL Divergence:** We seek to find the optimal variational parameters ϕ^* by minimising the Kullback-Leibler divergence:

$$\phi^* = \arg\min_\phi \mathrm{KL}(\rho_\phi || \pi(\cdot|\mathcal{D}))$$

- **The Intractability Problem:** We cannot calculate the KL directly because it still requires $\pi(\cdot|\mathcal{D})$.
- **Evidence Lower Bound (ELBO):** Rather than minimising the KL divergence, we instead **maximise** the ELBO:

$$\mathcal{L}(\phi) = \mathbb{E}_{\theta\sim\rho_\phi}[\log \Pi(\mathcal{D}, \theta) - \log \rho_\phi(\theta)]$$

- **Efficiency:** Maximising the ELBO only requires evaluating the joint distribution $\Pi(\mathcal{D}, \theta) = p_\theta(\mathcal{D})\pi(\theta)$.

Equivalence: KL and ELBO optimisation

Minimising the KL divergence and maximising ELBO are **equivalent**! We have

$$\begin{aligned}
\mathrm{KL}(\rho_\phi \,\|\, \pi(\cdot \mid \mathcal{D})) &= \mathbb{E}_{\theta\sim\rho_\phi}\left[\log \frac{\rho_\phi(\theta)}{\pi(\theta \mid \mathcal{D})}\right] \\
&= \int \rho_\phi(\theta)\, \log \frac{\rho_\phi(\theta)}{\pi(\theta \mid \mathcal{D})}\, \mathrm{d}\theta \\
&= \int \rho_\phi(\theta)\, \log \frac{\rho_\phi(\theta)}{\frac{\Pi(\theta,\mathcal{D})}{p_\pi(\mathcal{D})}}\, \mathrm{d}\theta \qquad \text{(Bayes' rule)} \\
&= \int \rho_\phi(\theta) \Big(\log \rho_\phi(\theta) - \log \Pi(\theta, \mathcal{D}) + \log p_\pi(\mathcal{D}) \Big)\, \mathrm{d}\theta \\
&= -\mathcal{L}(\phi) + \log p(\mathcal{D}).
\end{aligned}$$

Since $\log p_\pi(\mathcal{D})$ is constant w.r.t. ϕ, minimising the KL is equivalent to maximising the ELBO $\mathcal{L}(\phi)$.

Why Should we Take a Bayesian Approach?

- Bayesian methods allow us to construct models that return principled uncertainty estimates
- Bayesian modelling allows us to combine information from data with that from **prior expertise**
- This means we can exploit existing knowledge, rather than purely relying on black-box processing of data
- Models make clear assumptions and are **explainable**
- Bayesian models are often **interpretable**; they can be easily queried, criticised, and built on by humans

Shortfalls [Non Exhaustive]

- Bayesian inference is typically very difficult and expensive: getting around the proportionality constant in Bayes rule is surprisingly challenging
- All models are approximations of the world
  - Constructing accurate models can be very difficult
  - We will always impart incorrect assumptions on our model, particular in our likelihood function
  - For large datasets, the bias from these can usually be avoided by using a powerful discriminative method
- Bayesian reasoning only incorporates uncertainty that is within our model: it does not account for unknown unknowns
  - This can lead to overconfidence
  - Our probabilities/uncertainties are always inherently subjective
- Can struggle to deal with outliers in the data because likelihood terms are multiplicative

Further Reading

- Various examples in the notes
- Chapter 1 of C Robert. **The Bayesian choice: from decision-theoretic foundations to computational implementation.** 2007. https://www.researchgate.net/publication/41222434_The_Bayesian_Choice_From_Decision_Theoretic_Foundations_to_Computational_Implementation.
- Michael I Jordan. Are you a Bayesian or a frequentist? Video lecture, 2009. http://videolectures.net/mlss09uk_jordan_bfway/
- Chapters 6 and 7 Tom Rainford's course: https://www.cs.ox.ac.uk/files/11549/main.pdf