Bayesian, Laplace or De Finetti Statistics?

Christopher Yau

University of Oxford

27 October 2025

Oxford Bayesians

Over the last twenty years, Oxford has been strongly associated with Bayesian Statistics:

- Chris Holmes, Yee Whye Teh, Arnaud Doucet, Tom Rainforth, Geoff Nichols, Christopher Yau, Yarin Gal, Steve Roberts, Mike Osborne, Francois Caron, Mark van der Wilk, Judith Rousseau, ...

... and many graduate students and postdocs.

A motivating example

A billiard ball is rolled onto a table at random, landing somewhere along a line between 0 and 1 (the “true probability” of success).

Then, another ball is rolled repeatedly, and each time we note whether it lands to the left or right of the first ball.

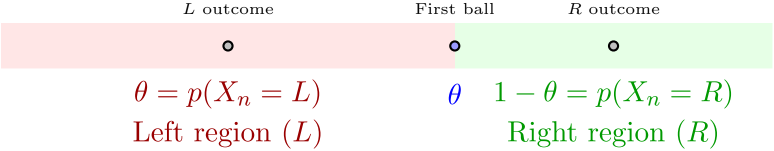
From these outcomes, Bayes asked:

> *Given the number of successes and failures observed, what is the probability that the true chance of success lies within a certain interval?*

A motivating example

Let $\theta \in (0,1)$ denote the position of the first ball.

Let $X_n \in \{L, R\}$ denote the left/right positioning of the n-th following balls.

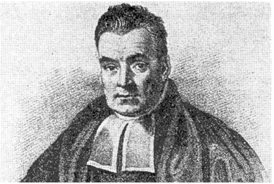


The posterior distribution over θ given $X_1, \ldots, X_N$:

$$\begin{aligned} p(\theta | X_1, \ldots, X_N) &= \frac{p(X_1, \ldots, X_N | \theta) p(\theta)}{p(X_1, \ldots, X_N)}, \\ &\propto \prod_{n=1}^{N} p(X_n | \theta) p(\theta), \\ &\propto \prod_{n=1}^{N} \underbrace{\theta^{\delta(X_n = L)}}_{p(X_n = L)} \underbrace{(1-\theta)^{\delta(X_n = R)}}_{p(X_n = R)} p(\theta), \\ &\propto \theta^{N_L} (1-\theta)^{N_R} p(\theta) \end{aligned}$$

where $N_L = \sum_n \underbrace{\delta(X_n = L)}_{\text{Indicator function}}$ and $N_R = N - N_L$.

1700s: English Reverend Thomas Bayes wanted to know how to infer causes from effects.



Working problem:

> *How could he learn the probability of a future event occurring if he only knew how many times it had occurred or not occurred in the past?*

Bayes' System

Bayes considered problems similar to the motivating example and how you narrow down the area in which the first ball probably sits.

Each new piece of information constrains the area where the first ball probably is.

Bayes' system was:

$$\textit{Initial Belief} + \textit{New Data} \Rightarrow \textit{Improved Belief}.$$

Legacy of Bayes’

There were two enduring criticisms to Bayes’ system:

- Mathematicians (even to this day) were horrified to see something as whimsical as a *guess* play a role in rigorous mathematics,
- Bayes said that if he didn’t know what guess to make, he would just assign all possibilities equal probability to start. This concept of “prior belief” was too much for mathematicians.

Bayes never published his discovery, but his friend Richard Price found it among his notes after Bayes’ death in 1761, re-edited it, and published it (*“An Essay towards solving a Problem in the Doctrine of Chances”*)

Unfortunately, virtually no one seems to have read the paper, and Bayes’ method lay unappreciated until Laplace.

Laplace Statistics?

In 1774, in *"Mémoire sur la probabilité des causes par les événements"*, Pierre Laplace independently identified the same idea but as the greater mathematician was able to systemise it using probability theory (**inductive reasoning**).

Laplace was the first true *Bayesian* and the work was integrated into his monumental work *"Théorie analytique des probabilités"* in 1812.

Strangely, Laplace also identified the **Central Limit Theorem** - a key result - underpinning what has become *frequentist* approaches.

Bayesian Belief Updating

1. Begin with **prior** belief over a parameter θ encapsulated by a probability distribution $p(\theta)$.
2. Acquire data X according to some process/experiment which is controlled by θ. The probability of observing X given θ is known as the **likelihood**, $p(X|\theta)$.
3. Compute the **posterior** probability distribution $p(\theta|X)$ of θ given data X using **Bayes' Theorem**.

Bayesian Belief Updating

The Bayesian belief update is given by:

$$\underbrace{p(\theta|X)}_{\text{Posterior}} = \frac{\overbrace{p(X|\theta)}^{\text{Likelihood}}\,\overbrace{p(\theta)}^{\text{Prior}}}{\underbrace{p(X)}_{\text{Marginal Likelihood / Evidence}}}$$

where θ denotes a parameter and X is data.

The potentially multi-dimensional integral

$$p(X) = \int p(X|\theta)dp(\theta)$$

is the computational barrier to the application of Bayesian belief updating.

The Score Function

Definition

$$s(\theta) = \nabla_\theta \log p(X|\theta)$$

- Measures how sensitive the likelihood is to changes in θ.
- Tells us the *direction* of steepest ascent of the likelihood.

Connection to Bayes' Rule

Taking logs of Bayes' theorem:

$$\log p(\theta|X) = \log p(X|\theta) + \log p(\theta) - \log p(X)$$

Hence:

$$\nabla_\theta \log p(\theta|X) = \underbrace{\nabla_\theta \log p(X|\theta)}_{\text{Score Function}} + \nabla_\theta \log p(\theta) (+\text{no terms involving } p(X))$$

Iterated Bayesian Updating

The update can be iterated:

$$\underbrace{p(\theta|X_{\text{new,old}})}_{\text{Updated Posterior}} = \frac{\overbrace{p(X_{\text{new}}|\theta)}^{\text{Likelihood}} \quad \overbrace{p(\theta|X_{\text{old}})}^{\text{Prior = Previous Posterior}}}{\underbrace{p(X_{\text{new}}|X_{\text{old}})}_{\text{Marginal Likelihood / Evidence}}}$$

where θ denotes a parameter and $X_{\text{old/new}}$ is old and new data respectively.

This alllows *sequential updating* of knowledge.

Bayesianism in decline

Bayesian belief updating appeared on and off over the next couple of centuries but was never the mainstream approach to probability or statistics.

John Stuart Mill denounced probability as:

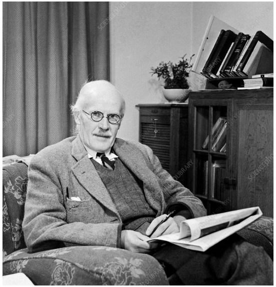
> *"ignorance … coined into science"*

Subjectivity and mathematics did not mix.

Clandestine Bayesianism

Bayesian theory continued to be developed sporadically (Emile Borel, Frank Ramsay, Bruno De Finetti):

- Harold Jeffreys, a geologist, made Bayesian Statistics slightly more fashionable in the 1930s.
- $p(\theta|X)$ vs $p(X|\theta)$.
- Jeffreys was a counterpoint to the great Ronald Fisher - who advocated sampling-based statistical thinking.
- Ironically famous for the *Jeffrey's prior* which is derived using *Fisher information* (variance of the score).



As the more robust personality, Fisher "won".

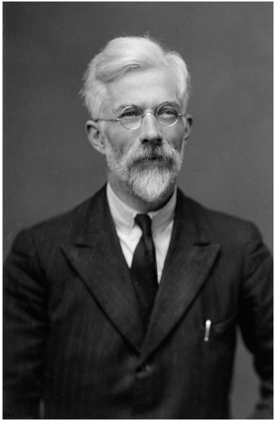
Fisher's Objection

Ronald Fisher and Jerzi Neyman killed Bayesian thinking in the first half of the 1900s.

> *"(Bayes) attempted to find, by observing a sample, the actual probability that the population value lay in any given range."*

> *"Such a problem is indeterminate without knowing the statistical mechanism under which different values of (a parameter) come into existence; it cannot be solved from the data supplied by a sample, or any number of samples, of the population."*

> *"We can know nothing of the probability of hypotheses or hypothetical quantities."*



Fisher was never comfortable with $p(\theta|X)$ and preferred $p(X|\theta)$.

The “importance” of the prior

Fisher’s problematic **prior** is often cited as one of the distinctive and advantageous features of Bayesian Statistics.

It allows you to:

1. Define *preferred* values or ranges for parameters.
2. Regularise and make models more *stable*.
3. Enforce bias or assumptions.

Together with **uncertainty quantification**, these are often reasons given for why it is good to be Bayesian.

Questions for Bayesians

- But how do we know the prior exists?
- How do we know if we keep updating then the true posterior is obtained?
- What is the posterior if we had an infinite number of observations?

A bit of measure theory ...

Definition

A **σ-algebra** $\mathcal{F}$ on a sample space Ω is a collection of subsets of Ω such that:

1. $\Omega \in \mathcal{F}$ (entire sample space is an event),
2. If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$ (closed under complements),
3. If $A_1, A_2, A_3, \dots \in \mathcal{F}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$ (closed under countable unions).

Intuition

$\mathcal{F}$ represents all events that are **knowable or observable**.
Each $A \in \mathcal{F}$ corresponds to a “question” we can ask about the random world.

Example: Coin flips

Let X_t = result of the t-th flip, and $\mathcal{F}_t = \sigma(X_1, \dots, X_t)$.
Then $\mathcal{F}_t$ contains all events determined by the first t outcomes.

Filtrations: Information Over Time

Definition

A **filtration** $(\mathcal{F}_t)_{t \geq 0}$ is an **increasing family of σ-algebras:**

$$\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots$$

Meaning

As time passes, we only gain information:

$$\text{More data} \;\Rightarrow\; \text{more events become measurable.}$$

Intuition

$\mathcal{F}_t$ encodes "everything known up to time t." A filtration represents the process of knowledge accumulation.

Adapted Processes

Definition

A stochastic process (X_t) is **adapted** to a filtration $(\mathcal{F}_t)$ if

$$X_t \text{ is } \mathcal{F}_t\text{-measurable for all } t.$$

Interpretation

- Being adapted means the process respects a causal flow of information.
- X_t depends only on information available up to time t — not on the future.

The Martingale Property

Definition

A process (X_t) is a **martingale** with respect to the filtration $(\mathcal{F}_t)$ if:

$$\mathbb{E}[X_{t+1} \mid \mathcal{F}_t] = X_t,$$

and $\mathbb{E}[|X_t|] < \infty$.

Intuition

- Given all current information, your expected future value is the same as the present value.
- Martingales model a “fair game” — there is no predictable gain over time.

Note: A martingale is defined with respect to a filtration, but the condition at each step uses the σ-algebra $\mathcal{F}_t$ for the corresponding time step in the filtration.

Doob Martingale

Definition

For any integrable random variable X and filtration $(\mathcal{F}_t)$:

$$M_t = \mathbb{E}[X \mid \mathcal{F}_t]$$

defines the **Doob martingale**.

Martingale Property

By the Tower Property (Law of Iterated Conditional Expectations):

$$\mathbb{E}[M_{t+1}|\mathcal{F}_t] = \underbrace{\mathbb{E}[\mathbb{E}[X|\mathcal{F}_{t+1}]|\mathcal{F}_t] = \mathbb{E}[X|\mathcal{F}_t]}_{\text{Tower Property}} = M_t.$$

The next step conditional expectation (given no new information) is the same as the current conditional expectation.

Doob Martingale

Intuition

- Imagine trying to predict a final outcome X (like tomorrow's stock price or total rainfall) as you receive information over time.
- Each $M_t = \mathbb{E}[X|\mathcal{F}_t]$ is your best guess given what you currently know.
- The Doob martingale says that, although your estimates change as new data arrives, they do so fairly — there is no tendency to systematically over- or under-estimate the final outcome.
- In the absence of new information, things stay the same.

Posterior probabilities as a martingale

Setup

Let Θ be a parameter with prior Π_0, and let

$$\mathcal{F}_t = \sigma(X_1, \ldots, X_t)$$

represent the information from the first t data points.

Posterior probability as conditional expectation

For any measurable set $A \subseteq \Theta$:

$$M_t = \mathbb{E}[\mathbf{1}_{\{\Theta \in A\}} \mid \mathcal{F}_t] = \Pi_t(A) = P(\Theta \in A | X_1, \ldots, X_t).$$

Martingale property

$$\mathbb{E}[M_{t+1} \mid \mathcal{F}_t] = M_t.$$

Hence $(\Pi_t(A))_{t \geq 0}$ is a bounded martingale.

Step 1: Martingale convergence

Doob's Martingale Convergence Theorem

If (M_t) is a bounded martingale, then

$$M_t \to M_\infty \quad \text{almost surely as } t \to \infty.$$

Application to posterior probabilities

For every $A \subseteq \Theta$,

$$\Pi_t(A) = \mathbb{E}[\mathbf{1}_{\{\Theta \in A\}} \mid \mathcal{F}_t] \quad \Rightarrow \quad \Pi_t(A) \to \Pi_\infty(A) \text{ a.s.}$$

Thus, each posterior probability converges almost surely.

As t increases, $\mathcal{F}_t$ grows, so your conditional expectation "stabilises" — you have ve learned all there is to learn about X from the filtration.

Step 2: Identifying the limit

Assume data come from the true parameter θ_0

That is, $X_i \sim P_{\theta_0}$ i.i.d.

Likelihood behaviour

For any $\theta \neq \theta_0$,

$$\frac{p_\theta(X_1, \ldots, X_n)}{p_{\theta_0}(X_1, \ldots, X_n)} \to 0 \quad \text{a.s. under } P_{\theta_0}.$$

Posterior collapse

Hence:

$$\Pi_t(A) \to \begin{cases} 1, & \text{if } \theta_0 \in A, \\ 0, & \text{if } \theta_0 \notin A, \end{cases} \qquad P_{\theta_0}\text{-a.s.}$$

The limiting random measure is $\Pi_\infty = \delta_{\theta_0}$.

Doob's Consistency Theorem (1949)

- Each posterior probability $\Pi_t(A)$ is a martingale:

$$\mathbb{E}[\Pi_{t+1}(A) \mid \mathcal{F}_t] = \Pi_t(A).$$

- Bounded martingales converge (Doob, 1953).
- Under the true parameter, posterior mass outside neighborhoods of θ_0 vanishes.
- $\Rightarrow$ Posterior collapses: $\Pi_t \Rightarrow \delta_{\theta_0}$.

Summary

- Bayesian updating = Doob martingale + Martingale convergence theorem $\Rightarrow$ Almost sure posterior consistency.
- Remarkably this is a pure measure-theoretic result with few conditions.
- More detailed exposition: `https://arxiv.org/pdf/1801.03122`

The Prior

The Bayesian belief update is given by:

$$\underbrace{p(\theta|X)}_{\text{Posterior}} = \frac{\overbrace{p(X|\theta)}^{\text{Likelihood}}\,\overbrace{p(\theta)}^{\text{Prior}}}{\underbrace{p(X)}_{\text{Marginal Likelihood / Evidence}}}$$

where θ denotes a parameter and X is data.

How do I now a prior exists?

A thought experiment

Suppose I flip 5 coins:

$$X_1 = H, X_2 = T, X_3 = H, X_4 = H, X_5 = T$$

What is the probability of an H on the 6th flip?

$$P(X_6 = H | X_1 = H, X_2 = T, X_3 = H, X_4 = H, X_5 = T)$$

A thought experiment

If each flip is independent and identically distributed (iid) then:

$$\begin{aligned}
&P(X_6 = H | X_1 = H, X_2 = T, X_3 = H, X_4 = H, X_5 = T) \\
&\quad = \frac{P(X_6 = H, X_1 = H, X_2 = T, X_3 = H, X_4 = H, X_5 = T)}{P(X_1 = H, X_2 = T, X_3 = H, X_4 = H, X_5 = T)}
\end{aligned}$$

A thought experiment

If each flip is independent and identically distributed (iid) then:

$$
\begin{aligned}
&P(X_6 = H | X_1 = H, X_2 = T, X_3 = H, X_4 = H, X_5 = T) = \\
&\qquad = \frac{P(X_6 = H)\cancel{P(X_5 = T)P(X_4 = H)P(X_3 = H)P(X_2 = T)P(X_1 = H)}}{\cancel{P(X_5 = T)P(X_4 = H)P(X_3 = H)P(X_2 = T)P(X_1 = H)}}, \\
&\qquad = P(X_6 = H)
\end{aligned}
$$

The history of coin flips tells me nothing?

A thought experiment

Ah, but there is a parameter - the probability of getting an H which we call θ!

$$\begin{aligned} & P(X_6 = H, \theta | X_1 = H, X_2 = T, X_3 = H, X_4 = H, X_5 = T) = \\ & \qquad = P(X_6 = H | \theta) p(\theta | X_1 = H, X_2 = T, X_3 = H, X_4 = H, X_5 = T) \end{aligned}$$

The history of flips tells us about θ which then helps us to predict X_6.

The coin flips are only iid if I know θ (conditionally iid).

Conditional independence

In general, we can introduce **shared** parameters to make the joint distribution conditionally iid:

$$p(X_1 = x_1, \ldots, X_n = x_n | \theta) = p(X_1 = x_1 | \theta) p(X_2 = x_2 | \theta) \cdots p(X_n = x_n | \theta)$$

The data X is **conditionally** iid **given** parameter θ so historical information can feed into predicting the future:

$$\begin{aligned} p(X_{n+1}, \theta | X_1 = x_1, \ldots, X_n = x_n) = \\ p(X_{n+1} | \theta) p(\theta | X_1 = x_1, X_2 = x_2, X_n = x_n) \end{aligned}$$

Does this construction *always* exist? Must there be a suitable parameter?

Exchangeability

If a sequence of random variables is **exchangeable** then there is no information in their order, i.e.

$$p(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = p(X_{\pi(1)} = x_1, X_{\pi(2)} = x_2, \ldots, X_{\pi(n)} = x_n)$$

where π denotes a permutation operation (e.g. $(1, 2, 3, 4, 5) \to (5, 3, 1, 4, 2)$)

Exchangeability is a *weaker* assumption than independence.

Crucially, exchangeability assumptions apply to many real-world problems.

De Finetti's Representation Theorem

The remarkable finding by De Finetti is that for *any* infinite exchangeable random sequence, it is *always* possible to write:

$$p(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \int \prod_{i=1}^{n} p(X_i|\theta)p(\theta)d\theta$$

as $n \to \infty$.

The implications are profound, there is:

- a parameter θ which makes $X_1, \ldots, X_n$ conditionally independent,
- this parameter must always have a distribution $p(\theta)$ associated with it,
- and there is always a sampling model $p(X|\theta)$,
- exchangeable sequences are mixtures of conditionally iid sequences under the prior measure.

We *should be* Bayesian when exchangeability applies!

We can also use the De Finetti representation theorem in reverse:

$$\int \prod_{i=1}^{n} p(X_i|\theta)p(\theta)d\theta \rightarrow p(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n),$$

If we define a **latent variable model** then the marginal joint distribution of the data must be exchangeable.

This allows us to model complex exchangeable distributions in an *indirect* way.

Caveats

- De Finetti proved for *binary* exchangeable sequences only.
- The Generalised Representation Theorem by Hewitt & Savage appeared later.
- Strictly speaking only applies to *infinite* sequences but Diaconis and Freedman have developed forms/bounds for finite sequences.
- Things become complex for partial exchangeability and more nuanced dependence structures.
- It is an *existence* theorem only.

A more refined understanding of Bayesian Statistics ...

Doob and De Finetti provide strong theoretical justification for Bayesian reasoning:

De Finetti shows that if data are exchangeable, a Bayesian model must exist.

Doob shows that if we update beliefs through Bayes' rule, they will converge consistently.

Exchangeability is not just a theoretical assumption — it is a practical modelling principle.

It gives us a situation when Bayesian inference *optimally* applies, and guides how to build our models.

BREAK: Sentiment Classification (Few-shot Prompt)

Prompt given to an LLM

```
Example 1:  "I loved this movie!" → Positive
Example 2:  "The plot was boring and predictable." → Negative
Example 3:  "The soundtrack was beautiful." → Positive
Now classify:  "The ending was disappointing." → ?
```

Questions:

- What label do you expect the model to output?
- How can it infer that “disappointing” implies a *Negative* label?
- The model’s weights never changed — so where does this learning happen?

Bayesians at War

Throughout its history, Bayesian approaches have always been synonymous with military applications:

- In 1890, Joseph Bertrand helped French/Russia artillery officers to use Bayesian approaches to target enemy locations.
- Kolmogorov suggested the same idea in WW2.
- Turing used it for code cracking.

Modern military information fusion systems are all based on Bayesian frameworks.

Bayesians Return

Following the Second World War:

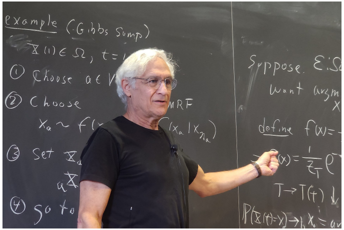
- I.J. Good began to revisit the system of Bayesian Inference.
- Lindley and Savage put mathematical rigour into the approach.
- Lindley began to put Bayesian academics into Professorships.

Computing and Physics

Bayesian approaches had a fundamental problem.

Outside of a few mathematically convenient examples, it was not tractable.

Seminal paper by Geman & Geman (1984) that linked Bayesian computation to Statistical Physics.



Posterior inference

Suppose there are multiple parameters:

$$\theta = \{\theta_1, \theta_2\}.$$

The marginal likelihood:

$$p(X) = \int p(X|\theta)p(\theta)d\theta$$

requires a multi-dimensional integral over the entire parameter space.

Posterior inference

Without the marginal likelihood we cannot directly compute, $p(\theta|X)$.

Geman & Geman overcame this with the Gibbs Sampler:

$$\theta_1 \sim p(\theta_1|\theta_2, X),$$
$$\theta_2 \sim p(\theta_2|\theta_1, X).$$

The Markov chain generated by this process has the posterior distribution as its stationary distribution. As the chain mixes, the samples converge in distribution to draws from this posterior.

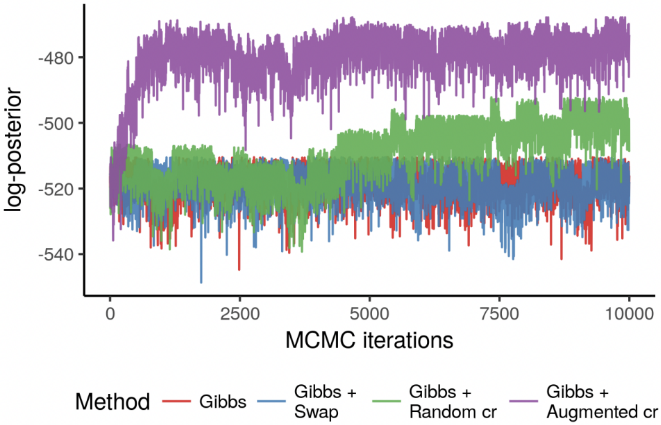
Golden Age

The Gibbs Sampler was one of a family of techniques known now as **Markov Chain Monte Carlo** (MCMC) methods.

Physicists and mathematicians looking at statistical properties of molecular configurations had already worked it out (Metropolis (1953) and Hastings (1970)).

Suddenly through the 1990s and early 2000s, Bayesian computation now enabled widespread use of the technique.

MCMC algorithms only provide *asymptotic* guarantees of convergence.



`http://proceedings.mlr.press/v89/martens19a/martens19a.pdf`

A crisis of BIG DATA

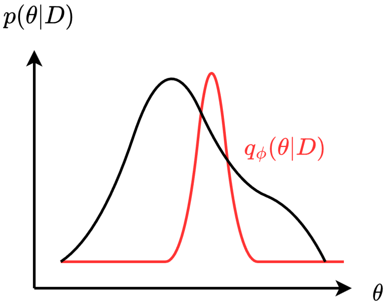
As datasets grew larger and models became higher-dimensional, traditional MCMC-based Bayesian inference began to struggle computationally.

- **Curse of dimensionality** - Sampling efficiency deteriorates exponentially with dimensionality.
- **Conditional sampling** - Sequential conditional updates make MCMC inherently difficult to parallelise.
- **Need to update all model parameters** - each MCMC iteration requires updating all parameters — in some models, the number of parameters grows with the data size.

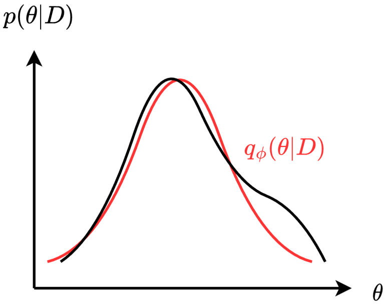
Note - see Chris Nemeth's lecture on Monte Carlo Methods for recent work in this area.

The nearly correct ...

Physics also offered another approach. *What if I approximate the posterior distribution with a simpler form?*



Unoptimised variational approximation



Optimised variational approximation

Variational methods trade *asymptotic exactness* for *optimisation-based approximation.*

The **mean-field approximation** $(O(NP))$:

$$p(\theta_1, \ldots, \theta_P | X) \approx \prod_{p=1}^{P} q_{\phi_p}(\theta_p | X)$$

In mean-field approximation, a set of *variational parameters* are tied to each data point. The use of **amortisation** provides a second layer of approximation $(O(N))$:

$$\phi_p = f_\nu(X_p)$$

where instead of optimising all variational parameters, you learn a function that maps relevant subsets of data to estimates of those parameters.

Note - see Hanwen's lecture this PM

Bayes today

In recent years, *explicit* Bayesian modelling has reduced in popularity or exposure.

But unlike previous centuries, it remains in widespread *implicit* use throughout science and engineering.

Traditional Bayesian analysis is *purist*, pragmatically we need to deal with the following:

- **Prior elicitation** - defining a "good" prior is hard (see George D lecture on Diffusion Models),
- **Uncertainty quantification** - MCMC and VB under-estimate posterior variance,
- **Causality** - Bayesian reasoning is concerned with probabilistic coherence not causal reasoning,
- **Model misspecification** - all the theorems only hold under the *true* model.

Modern Bayesian research has sought to address these issues.

Thinking about the Likelihood

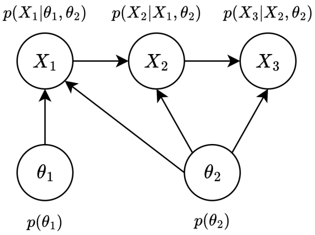
The term originates from *likelihood-based inference* in Statistics but (can) mean much more in Bayesian Statistics:

$$\underbrace{p(\theta|X)}_{\text{Posterior}} = \frac{\overbrace{p(X|\theta)}^{\text{Likelihood}}\ \overbrace{p(\theta)}^{\text{Prior}}}{\underbrace{p(X)}_{\text{Marginal Likelihood / Evidence}}}$$

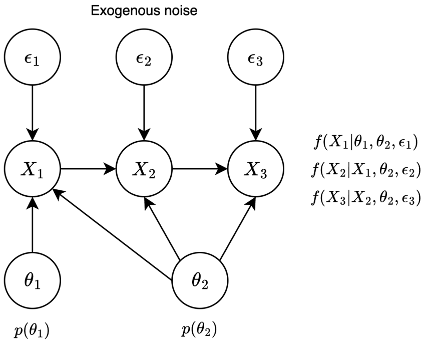
- The likelihood contains a **probabilistic graphical model** (PGM) for your problem.
- The PGM describes the **statistical dependencies** between all quantities in your model.
- Sometimes referred to as harbouring the **data generating process** (DGP) but not actually true ...

A preview of correlation not causation

A **structural causal model** (SCM) allows **interventions** while a **probabilistic graphical model** (PGM) does not.



Probabilistic Graphical Model

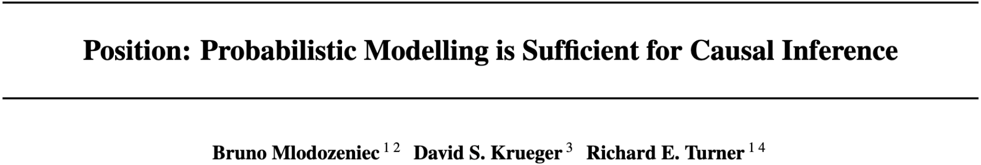


Structural Causal Model

A PGM does not define a data generation model *under interventions*.

Is causality required?

While causal models add structure to describe interventions, Bayesian reasoning doesn't require causality — it operates at the level of probabilistic coherence.



Position: Probabilistic Modelling is Sufficient for Causal Inference

Bruno Mlodozeniec [1 2] **David S. Krueger** [3] **Richard E. Turner** [1 4]

"Always write down the probability of everything."

– Steve Gull (MacKay, 2003, p. 61)

A recent line of thought is to do away with causal inference approaches and to add specific mirror-like "counterfactual" variables to mirror an intervention.[1]

[1] `https://openreview.net/forum?id=V1FP9WDKa7`

Model Misspecification

When discussing model specification, Bayesians refer to **M-closed** and **M-open**.

In the **M-closed world**:

- The true data-generating process lies within our model family $\{p(X|\theta) : \theta \in \Theta\}$.
- Classical Bayes is **well-calibrated and coherent**.
- Posterior concentrates around the true parameter θ^*.

Model Misspecification

In the **M-open world**:

- The true data-generating process is **outside** our model family.
- The likelihood $p(X|\theta)$ is **misspecified**.
- Classical Bayes can be **overconfident or misleading**.
- Traditionally, Bayesian tried to overcome this by using **Bayesian nonparametrics** to define infinite-dimensional function and random probability measures.

In M-open settings, we need a more flexible update rule — **Generalised Bayes** replaces the likelihood with a *loss*.

Bayesian Decision Theory

- Statistics can be viewed as a **decision problem under uncertainty.**
- Let θ denote an unknown state of nature and a an action.
- Decisions are evaluated by a **loss function** $L(a, \theta)$.
- The optimal action minimises **expected loss:**

$$a^* = \arg\min_a \mathbb{E}_{p(\theta|X)}[L(a, \theta)].$$

- Contrast with classical decision theory where the expectation is wrt to sampling distribution of X.

For a Bayesian, rational decisions are those that *minimise* **posterior expected loss**.

From Decision Theory to Generalised Bayes

- Bissiri, Holmes & Walker (2016): updating beliefs **is itself a decision problem**.
- Choose a posterior $q(\theta)$ to minimise:

$$q^*(\theta) = \arg\inf_q \mathbb{E}_q[L(\theta; X)] + \frac{1}{\eta}\mathrm{KL}(q\|p).$$

- Here:
  - $L(\theta; X)$: loss describing the data's informational content (replaces likelihood),
  - $\mathrm{KL}(q\|p)$: penalty for deviating from the prior.
- Solution:

$$q^*(\theta) \propto p(\theta)\exp(-\eta L(\theta; X))$$

— the Generalised Bayes posterior.

Updating = deciding which beliefs to hold, balancing data and prior.

The Generalised Bayesian Update

Also, known as a **Gibbs posterior**:

$$p(\theta|X) \propto p(\theta)\exp(-\eta L(\theta;X))$$

- $p(\theta)$: prior belief
- $L(\theta;X)$: data-dependent loss (not necessarily log-likelihood)
- $\eta > 0$: learning-rate / temperature parameter

Retains the Bayesian **form**, but broadens what “evidence” means.

Note: Standard Bayes is a special case of the Generalised Bayes update.

$$L(\theta;X) = -\log p(X|\theta) \quad \Rightarrow \quad p(\theta|X) \propto p(\theta)p(X|\theta)$$

Why Generalised Bayes Matters in the M-open World

Generalised Bayes offers a way to:

- Retain **coherent belief updating** without a likelihood
- Incorporate **model misspecification and robustness**
- Align inference with **predictive or decision objectives**

Generalised Bayes is Bayesian reasoning for an imperfect world.

Scope of the Generalised Bayes Update

The update:

$$p(\theta|X) \propto p(\theta) \exp(-\eta L(\theta; X))$$

is only valid for **parameters that appear in the loss function** $L(\theta; X)$.

If a component θ_j does not appear in L:

$$p(\theta_j|X) = p(\theta_j)$$

- Coherence requires belief revision only where the loss provides evidence.
- Prevents spurious updating of uninformed parameters.

Generalised Bayes is local in scope: only beliefs connected to the loss are revised.

Applications of Generalised Bayes

- **PAC-Bayes / Learning Theory:** treat L as empirical risk → connects to regularised risk minimisation.
- **Robust Bayes:** replace log-likelihood with β-divergence or Huber loss → down-weight outliers.
- **Online learning:** adapt η dynamically → information-theoretic updates.
- **Approximate inference:** variational or ensemble updates mimic generalised Bayes behaviour.

The Spirit of Bayes Lives On

When our model isn't perfect, *Generalised Bayes keeps us Bayesian in spirit*:

- Still **belief updating**, just under broader notions of evidence
- **Retains coherence** preserves consistency – Doob still applies
- **Bridges with modern approaches in AI**, e.g. gives a framework for what were previously ad-hoc regularisation strategies

For further discussion, see:

"Position: Bayesian Deep Learning is Needed in the Age of Large-Scale AI"
`https://arxiv.org/abs/2402.00809`

BREAK HERE

Latent Variable Model

A simple **latent variable model** has the form:

$$X|Z \sim N(\mu(Z), \sigma^2(Z)), \tag{1}$$
$$Z \sim N(0, 1), \tag{2}$$

and we want to learn $p(Z|X)$ which is the posterior distribution over the unobserved latent variable Z given the observation X.

Frequently used where dimensionality of Z is much lower than X.

Variational Autoencoder

If I choose:

- $\mu(Z)$ is a function described by a neural network,
- $\sigma^2(Z)$ is a function described by a neural network,
- Approximate $p(Z|X)$ via a Normal distribution,
- Apply amortisation to reduce the number of variational parameters.
- Perform optimisation via stochastic gradient descent.

The result is an algorithm called a **Variational Autoencoder** (VAE):

$$[\text{VAE}] = [\text{Latent Variable Model}] + [\text{Inference Choices}]$$

Variational Autoencoder

Encoder - maps data X to (distribution over) latent variables Z

Decoder - maps latent variable Z to (distribution over) data X

Loss function

$$l(\theta, \phi) = -\mathbb{E}_{Z \sim q_\theta(Z|X)}[\log p_\phi(X|Z)] + \mathbb{KL}(q_\theta(Z|X)||p(Z))$$

Black Box Bayesian

Black box building blocks have simplified and normalised the application of Bayesian techniques by unifying models and computation.

$$\underbrace{\underbrace{[\text{Structure}] - [\text{Distributions}]}_{\text{Model}} - [\text{Inference}]}_{\text{Black Box Bayesian}}$$

This convergence underpins the utility of **probabilistic programming languages**, e.g. STAN, Pyro, PyMC.

But black box techniques intrinsically conflate the statistical and computing biases making diagnostic more difficult.

To End: Bayesianism in 2025

Bayesian thinking in 2025 goes beyond the mechanics of applying Bayes' rule and computational inference:

- Foundational motivations for Bayesian reasoning give strong theoretical justification for its optimality
- Existing and new AI approaches should be able to reason and compute in a Bayesian way
- Adds to the battery of expectations we might expect of the highest-performing models.