On Variational Bayesian Methods
DRAFT V3

David Ewing

16th December 2025

Abstract

Probabilistic machine learning often involves models with hidden (latent) variables. The main goal is to uncover the latent structure Z that explains the observed data X. In many practical settings, however, the exact posterior distribution $p(z \mid x)$ cannot be computed in closed form or evaluated efficiently. Variational Inference (VI) addresses this challenge by reframing posterior inference as an optimisation problem. Instead of working with the true posterior, a tractable family of distributions $q_\nu(z)$, is chosen, and parameterised by variational parameters ν. A member of this family that is closest to the true posterior, is chosen, typically measured using Kullback–Leibler divergence. This optimisation is equivalent to maximising the Evidence Lower Bound (ELBO), which provides a lower bound on the log marginal likelihood of the data. In this paper, I outline the core ideas behind VI, explain the role of the ELBO, and show how common choices of variational families—such as fixed-form and mean-field approximations—fit into the broader framework.

1 Probabilistic Modelling and the Inference Problem

Bayesian and frequentist goals (why the choice matters)

Both *Bayesian* and *frequentist* statistics are valid approaches to inference, and each targets different outputs. The choice is driven by the goal of the analysis.

- **Frequentist goal:** estimate unknown parameters (treated as fixed constants) and quantify uncertainty via repeated-sampling properties of the estimator (for example, standard errors and confidence intervals).
- **Bayesian goal:** represent uncertainty about unknown quantities using probability distributions, producing a posterior distribution over latent variables and parameters (for example, posterior means, credible intervals, and posterior predictive distributions).

A simple example illustrates the distinction. Suppose $Y_1, \ldots, Y_n \sim N(\mu, \sigma^2)$ with known σ^2. A frequentist analysis reports an estimate $\hat{\mu} = \bar{Y}$ and a 95% confidence interval $\bar{Y} \pm 1.96\,\sigma/\sqrt{n}$, where the 95% refers to long-run coverage of the interval construction under repeated sampling. A Bayesian analysis places a prior on μ (for example, $\mu \sim N(\mu_0, \tau_0^2)$) and then reports the posterior distribution $p(\mu \mid y)$; a 95% credible interval refers to posterior probability mass given the model, prior, and observed data.

Probabilistic modelling provides a systematic way to connect our assumptions about the world to the data observed. A model is specified in terms of:

- *Observed variables* X, representing the data;
- *Latent variables* or *parameters* Z, representing unknown structure or quantities of interest;

- A joint distribution $p(x, z)$ that encodes our assumptions and prior information.

The central inferential object is the *posterior distribution*

$$p(z \mid x) = \frac{p(x, z)}{p(x)}, \qquad p(x) = \int p(x, z)\, \mathrm{d}z,$$

which tells us how plausible different configurations of Z are, given the observed data X.

In simple models, $p(z \mid x)$ can be derived analytically. However, in many modern applications (hierarchical models, complex latent-variable models, deep generative models, and so on) the marginal likelihood $p(x)$ is intractable: the integral cannot be evaluated exactly and is expensive to approximate directly. This makes exact posterior inference infeasible, and motivates approximate methods such as Markov chain Monte Carlo (MCMC) and Variational Inference.

2 ~~Approximate~~ Bayesian Inference in Practice

Variational inference as Bayesian inference (and how it differs from other Bayesian workflows)

Up to this point, I have described probabilistic modelling in terms of a joint distribution $p(x, z)$ and an inferential target that is naturally expressed as a distribution. In Bayesian work, that target is the posterior distribution $p(z \mid x)$. In realistic models, this posterior is rarely available in closed form, so we turn to *approximate Bayesian inference* methods.

There are several common ways to approximate the posterior:

- **Simulation-based Bayes (MCMC):** construct a Markov chain with stationary distribution $p(z \mid x)$ and use draws to approximate expectations and uncertainty. This is widely treated as a reference standard for accuracy, but can be slow in high dimensions or hierarchical settings.
- **Deterministic local approximations (Laplace):** approximate the posterior by a Gaussian expansion around a mode (often the MAP), giving fast approximate uncertainty when the posterior is close to normal, but potentially misleading summaries when the posterior is skewed or multimodal.
- **Optimisation-based Bayes (Variational Inference):** choose a tractable family $\mathcal{Q}$ and optimise within that family to obtain $q_{\nu^\star}(z)$ as a proxy for $p(z \mid x)$. This is typically much faster than MCMC, but (especially under mean-field factorisations) can underestimate uncertainty.

In other words: VI is not an alternative to Bayesian inference; it is one way of doing Bayesian inference when exact posterior computation is infeasible. The difference is not the inferential target (still a posterior), but the computational route taken to approximate it.

3 Variational Inference as Optimisation

Variational Inference approaches posterior inference through optimisation rather than sampling. The idea is to replace the exact posterior $p(z \mid x)$ with a simpler, tractable distribution $q_\nu(z)$ drawn from a chosen family. Figure 1 provides the standard picture: attention is restricted to a variational family $q(z;\nu)$ and then optimise ν so that $q(z;\nu)$ is close (in KL divergence) to the true posterior $p(z \mid x)$.

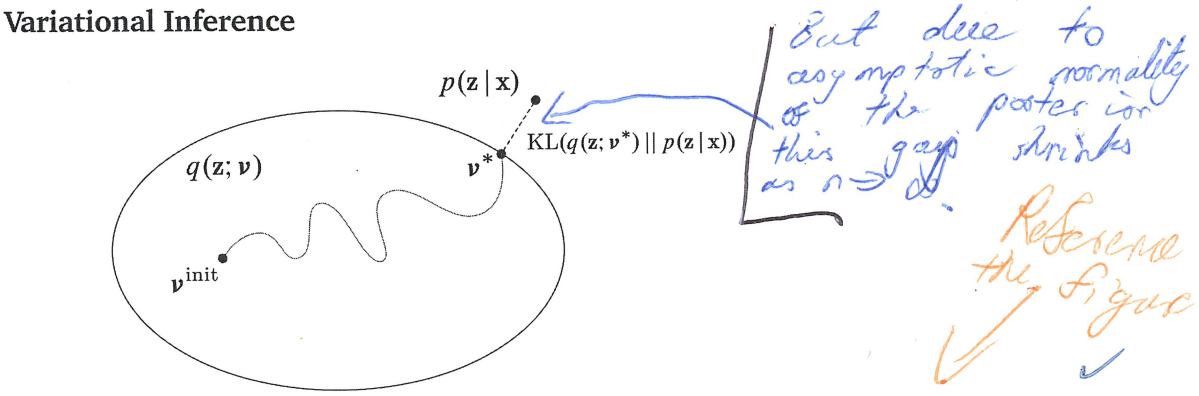

Figure 1: Variational inference as optimisation over variational parameters ν within a variational family $q(z;\nu)$, choosing $\nu^\star$ to minimise $\mathrm{KL}(q(z;\nu) \,\|\, p(z \mid x))$ [1].

The ellipse represents all families of tractable approximations, indexed by the variational parameters ν. Starting from an initial value ν^{init}, an optimisation routine moves through parameter space (the grey path) to reach $\nu^\star$, the best approximation available within the family. The true posterior $p(z \mid x)$ sits outside the ellipse because it is generally too complex to belong to the variational family, and the dashed segment indicates the remaining discrepancy measured by $\mathrm{KL}(q(z;\nu^\star) \,\|\, p(z \mid x))$. In the next subsection we define the family $\mathcal{Q}$ formally and make precise how $\nu^\star$ is determined.

3.1 The Variational Family

We first specify a family of candidate distributions

$$\mathcal{Q} = \{q_\nu(z) : \nu \in \mathcal{V}\},$$

where ν denotes the collection of *variational parameters*. This family might be defined, for example, by:

- A parametric form (e.g. multivariate Gaussian with mean vector and covariance matrix);
- A factorisation assumption (e.g. a product of simpler distributions across components of Z);
- Some combination of structural and parametric choices.

The goal of VI is then to choose

$$\nu^\star = \arg\min_\nu \mathrm{KL}\big(q_\nu(z) \,\|\, p(z \mid x)\big),$$

and to use $q_{\nu^\star}(z)$ as our approximation to the posterior. Once $\nu^\star$ has been found, this approximate posterior can be used for point estimates, interval estimates, prediction, and other downstream tasks.

3.2 KL Divergence and the ELBO

The Kullback–Leibler (KL) divergence between $q_\nu(z)$ and $p(z \mid x)$ is defined as

$$\mathrm{KL}\big(q_\nu(z) \,\|\, p(z \mid x)\big) = \mathbb{E}_{q_\nu}\left[\log \frac{q_\nu(z)}{p(z \mid x)}\right].$$

Directly minimising this quantity is usually not possible, because it depends on the intractable marginal likelihood $p(x)$. Instead, we work with the *Evidence Lower Bound* (ELBO), defined by

$$\mathcal{L}(\nu) = \mathbb{E}_{q_\nu}[\log p(x, z)] - \mathbb{E}_{q_\nu}[\log q_\nu(z)].$$

It can be shown that

$$\log p(x) = \mathcal{L}(\nu) + \mathrm{KL}\big(q_\nu(z) \,\|\, p(z \mid x)\big),$$

so that for fixed data x, maximising $\mathcal{L}(\nu)$ is equivalent to minimising the KL divergence. The ELBO thus serves as a surrogate objective that we can evaluate (or approximate) using only $p(x, z)$ and $q_\nu(z)$.

The ELBO has a useful interpretation as a balance between two terms:

- An *expected log joint* term, $\mathbb{E}_{q_\nu}[\log p(x, z)]$, which encourages $q_\nu(z)$ to place mass on configurations of z that explain the data well.
- An *entropy* term, $-\mathbb{E}_{q_\nu}[\log q_\nu(z)]$, which encourages $q_\nu(z)$ to be diffuse and to avoid collapsing onto a single point.

Optimising the ELBO therefore trades off goodness-of-fit against complexity, in a way that is closely related to (but distinct from) classical regularisation ideas.

3.3 Different Views of the ELBO

Although the ELBO is a single mathematical quantity, it can be written and interpreted in several equivalent ways. Different algorithms tend to emphasise different views. It is helpful to keep three of them in mind.

Projection view (KL divergence)

We can write,

$$\log p(x) = \mathcal{L}(\nu) + \mathrm{KL}\big(q_\nu(z) \,\|\, p(z \mid x)\big).$$

For fixed data x, the term $\log p(x)$ is a constant. Maximising the ELBO is therefore exactly the same as minimising $\mathrm{KL}(q_\nu(z) \,\|\, p(z \mid x))$.

In this view, variational inference is a projection operation: we project the true posterior onto the chosen family of approximations, and the discrepancy is measured by the KL divergence from q_ν to $p(\cdot \mid x)$.

Energy–entropy view

The ELBO can also be written as

$$\mathcal{L}(\nu) = \mathbb{E}_{q_\nu}[\log p(x, z)] - \mathbb{E}_{q_\nu}[\log q_\nu(z)].$$

The first term, $\mathbb{E}_{q_\nu}[\log p(x, z)]$, is sometimes called an "energy" term: it rewards placing probability mass on configurations of z that explain the data and respect the prior. The second term, $-\mathbb{E}_{q_\nu}[\log q_\nu(z)]$, is the entropy of q_ν, which rewards spread or uncertainty in the approximation.

Optimising the ELBO can then be seen as a trade-off:

- improve the fit to the data and prior by increasing the expected log joint;
- avoid collapsing q_ν onto a single point by maintaining entropy.

This energy–entropy view is especially useful when thinking about coordinate-ascent algorithms and mean-field approximations.

Likelihood–regularisation view

A third, equally valid, way to arrange the same terms is

$$\mathcal{L}(\nu) = \mathbb{E}_{q_\nu}[\log p(x \mid z)] - \mathrm{KL}\big(q_\nu(z) \,\|\, p(z)\big).$$

Here the ELBO looks like:

- an expected log-likelihood term, $\mathbb{E}_{q_\nu}[\log p(x \mid z)]$;
- minus a KL penalty that keeps $q_\nu(z)$ close to the prior $p(z)$.

Variational inference resembles a regularised fitting problem: we try to choose q_ν so that it explains the data well on average, but we pay a regularisation cost if q_ν moves too far from the prior.

This perspective connects naturally to fixed-form and black box variational methods, where the ELBO is treated as an objective to optimise numerically [4, 5].

A note on the direction of KL

The divergence that appears is $\mathrm{KL}(q_\nu(z) \,\|\, p(z \mid x))$. This "exclusive" direction penalises q_ν for placing mass in regions where the true posterior density is low.

One practical consequence, which will matter later, is that minimising $\mathrm{KL}(q_\nu \| p)$ tends to produce approximations that concentrate on one main mode of the posterior and avoid spreading mass into low-density regions. This often leads to underestimation of posterior uncertainty, especially in simple mean-field approximations [3, 5].

Other divergences and other directions of KL are possible, but in this introductory note we focus on this standard choice, because it underpins most of the mean-field and fixed-form methods discussed next.

Why these views matter for algorithms

Although the ELBO is the same object in all three views, different formulations support different algorithmic strategies:

- The projection view emphasises VI as "KL minimisation" and is useful for high-level understanding.
- The energy–entropy view fits naturally with coordinate-ascent and mean-field factorisations, where we update one factor at a time to improve the balance between fit and entropy.

- The likelihood–regularisation view fits naturally with fixed-form and black box VI, where we treat the ELBO as a regularised objective and optimise it with gradient-based methods [4, 5].

In latent-variable models, the ELBO is engineered so that maximising it is exactly the same as minimising $\mathrm{KL}(q_\nu(z) \,\|\, p(z \mid x))$*; outside that setting, KL minimisation need not come with an "ELBO" attached.*

In later sections, when we focus on mean-field VI and fixed-form VI, these interpretations will reappear in slightly different guises.

4 Variational Families: Fixed-Form and Mean-Field (Preview)

A key modelling choice in VI is how we define the family $\mathcal{Q}$. Two broad patterns that appear repeatedly in the literature are:

4.1 Fixed-Form Variational Inference

In *fixed-form* variational inference, we choose a specific parametric family (for example, multivariate Gaussian distributions) and constrain $q_\nu(z)$ to lie within that family for all models under consideration. The focus is then on developing general optimisation methods that can handle any model $p(x, z)$ while keeping $q_\nu(z)$ in this shared, tractable form [1, 5].

4.2 Mean-Field Variational Inference

In *mean-field* variational inference, we assume that the latent variables factorise under q_ν:

$$q_\nu(z) = \prod_j q_{\nu_j}(z_j),$$

where $z = (z_1, \ldots, z_J)$ is partitioned into components and each $q_{\nu_j}(z_j)$ is a simpler distribution. This assumption is often unrealistic as a literal description of the true posterior, but it makes the optimisation problem more tractable and can lead to closed-form coordinate ascent updates in many models [1, 5].

Under-dispersion in hierarchical models

A particularly important failure mode of mean-field VI shows up in hierarchical or random-effects models. In these settings, a variance component controls how much the random effects are allowed to vary around a group mean.

When we impose a mean-field factorisation and place strongly correlated parameters (for example, a variance component and its associated random effects) into separate independent blocks, the variational posterior for the variance component is often *under-dispersed*. Its posterior variance is too small, and the mass is pulled towards smaller values compared to a well-calibrated MCMC analysis [3, 5].

In practice, this under-dispersion at the variance level appears as *over-shrinkage*: the random effects are pulled too tightly towards the group mean, so the fitted model looks more confident and more homogeneous than it should. This effect can be subtle in the marginal posteriors for individual random effects, but it becomes very visible when we compare posterior variances from VI and MCMC for the variance component itself.

A detailed comparison of fixed-form, mean-field, and more structured approximations (including their strengths and limitations) will be developed in later sections. Here we simply note that the definition of $\mathcal{Q}$ is central: it encodes both the computational tractability and the expressive power of the variational approximation.

This is not needed.

5 Optimisation

Maximising the ELBO is achieved by numerical optimisation. Common strategies include:

- *Coordinate ascent* algorithms, particularly in conjugate-exponential models, where updates for each factor or parameter have closed-form expressions.
- *Gradient-based* methods, which use derivatives of $\mathcal{L}(\nu)$ with respect to ν and can be combined with modern optimisation techniques (such as variants of stochastic gradient descent) [5, 2].

The ELBO is typically not a convex function of ν, so optimisation procedures may converge to local optima rather than a unique global solution. Nevertheless, in many applications the resulting approximate posteriors are accurate enough to support interpretation and prediction, especially when exact inference is infeasible.

Modern probabilistic programming frameworks and automatic differentiation tools make it increasingly straightforward to implement gradient-based VI for complex models, without having to derive analytic updates by hand. More advanced schemes—such as black box variational inference and reparameterisation-based stochastic gradients—build on the same core ideas presented here [4, 5].

A diagnostic for under-dispersion

One way to quantify under-dispersion is to compare posterior variances from variational Bayes with those from a reference MCMC fit. For each parameter θ, we can form the variance ratio

$$\mathrm{VR}(\theta) = \frac{\mathrm{Var}_{\mathrm{VB}}(\theta)}{\mathrm{Var}_{\mathrm{MCMC}}(\theta)}.$$

Values close to one indicate similar levels of uncertainty, while values substantially below one signal that the variational posterior is under-dispersed [3, 5].

In hierarchical models, variance ratios for variance components are often well below one, and this aligns with the visual impression of over-shrinkage in the random effects. Such variance-ratio summaries provide a compact way to report how far a variational approximation departs from a more accurate MCMC reference.

6) Examples: ✓

To comes.

APPENDIX

A Bayesian vs Frequentist Perspectives

This appendix clarifies key conceptual differences between Bayesian and frequentist approaches to statistical inference, particularly as they relate to variational methods.

A.1 Posterior Distributions

Bayesian perspective: Parameters θ are treated as random variables. The posterior distribution $p(\theta \mid \text{data})$ represents the probability distribution of θ after observing the data. This distribution quantifies uncertainty by specifying how likely different parameter values are given the observed data.

Frequentist perspective: Parameters θ are fixed but unknown constants. There is no probability distribution over parameters. A frequentist estimates θ (for example, $\hat{\theta}$ via maximum likelihood) and may construct confidence intervals, but does not assign probabilities to parameter values. A 95% confidence interval does not mean "θ has 95% probability of lying in this interval"; rather, it means "if we repeated this procedure many times, 95% of such intervals would contain the true θ."

A.2 Prior Distributions

Bayesian perspective: Before observing data, prior beliefs or knowledge about θ are encoded in a prior distribution $p(\theta)$. This prior may be informative (reflecting strong prior knowledge) or vague (reflecting minimal assumptions). Bayes' theorem updates the prior with data via

$$p(\theta \mid \text{data}) \propto p(\text{data} \mid \theta)\, p(\theta).$$

Frequentist perspective: No prior distributions are used. Inference is based solely on the likelihood $p(\text{data} \mid \theta)$ and the data. While frequentist methods may employ regularisation (which has similar mathematical effects to Bayesian priors), this is not framed as encoding beliefs about parameters before observing data.

A.3 Latent Variables and Parameters

Both paradigms use latent (unobserved) variables Z that influence the data. Examples include cluster memberships in mixture models or true abilities in item-response models.

Frequentist approach: Latent variables Z are random, but parameters θ remain fixed. The likelihood is obtained by integrating over Z:

$$L(\theta) = \int p(\text{data}, Z \mid \theta)\, \mathrm{d}Z.$$

One then estimates $\hat{\theta} = \arg\max_\theta L(\theta)$, treating θ as a fixed unknown.

Bayesian approach: Both Z and θ are random. A joint posterior distribution $p(\theta, Z \mid \text{data})$ is obtained, quantifying uncertainty about both latent variables and parameters simultaneously.

A.4 Approximating Posteriors

Bayesian goal: The posterior $p(\theta \mid \text{data})$ is the inferential target, but it is often intractable. Approximation methods include:

- Markov chain Monte Carlo (MCMC), which generates samples from $p(\theta \mid \text{data})$;
- Variational inference, which finds a tractable distribution $q(\theta)$ that approximates $p(\theta \mid \text{data})$ [3, 5].

Frequentist goal: The objective is to estimate a point value $\hat{\theta}$ and quantify sampling uncertainty (how $\hat{\theta}$ would vary across hypothetical repeated samples). There is no notion of "approximating a distribution over θ," because parameters are not viewed as random.

A.5 Random vs Fixed Parameters

Frequentist view: Parameters θ are fixed but unknown constants. Randomness arises only from sampling: different data sets yield different estimates $\hat{\theta}$, but the true θ does not vary. Standard errors describe how $\hat{\theta}$ varies across hypothetical repeated samples.

Bayesian view: Parameters θ are random variables with probability distributions. Uncertainty about θ is represented by a prior distribution before data and a posterior distribution after data. The posterior distribution directly encodes degrees of belief about different parameter values.

A.6 Hierarchical Variance Components

Both frameworks use hierarchical (multilevel) models in which observations are grouped (for example, students within schools). Group-level effects often vary randomly.

Frequentist approach (mixed models):

- Random effects $u_j \sim N(0, \sigma^2)$ for group j are treated as random.
- The variance component σ^2 is a *fixed parameter* to be estimated (for example, via restricted maximum likelihood).
- Inference yields a point estimate $\hat{\sigma}^2$ and possibly a confidence interval.

Bayesian approach:

- Random effects $u_j \sim N(0, \sigma^2)$ for group j.
- The variance component σ^2 also receives a prior distribution $p(\sigma^2)$.
- Inference yields a posterior distribution $p(\sigma^2 \mid \text{data})$, providing full uncertainty quantification.

In the context of mean-field variational Bayes, the approximate posterior $q(\sigma^2)$ for the variance component is often under-dispersed (too narrow) relative to the true posterior obtained via MCMC [3, 5]. This under-dispersion manifests as overconfidence in the estimated variance and can lead to excessive shrinkage of random effects towards their mean. A frequentist analysis would yield a different point estimate but would not frame the issue as "posterior distribution is too narrow," because frequentist inference does not produce posterior distributions in the first place.

A.7 Philosophical Underpinnings

Frequentist: Probability describes long-run frequency in repeated experiments. Parameters are not assigned probabilities; only data (and functions of data) have probability distributions under repeated sampling.

Bayesian: Probability quantifies degree of belief about uncertain quantities, including parameters. Both data and parameters can have probability distributions, and inference updates beliefs via Bayes' theorem.

These foundational differences explain why concepts central to variational inference—such as approximating posterior distributions and quantifying parameter uncertainty through probability distributions—are specific to the Bayesian paradigm and may appear unfamiliar from a frequentist perspective.

References

[1] David Blei, Shakir Mohamed, and Rajesh Ranganath. Variational inference: Foundations and modern methods. NeurIPS 2016 Tutorial (slides), 2016. Slide 20 contains the variational inference optimisation schematic $(\mathrm{q}(z;\nu)ellipse, \nu_i nitto \nu^*, and p(z|x)target)$.

[2] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. Also presented at ICLR 2015.

[3] J. T. Ormerod and M. P. Wand. Explaining variational approximations. *The American Statistician*, 64(2):140–153, 2010.

[4] Rajesh Ranganath, Sean Gerrish, and David M. Blei. Black box variational inference. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pages 814–822, Reykjavik, Iceland, 2014. PMLR.

[5] Minh-Ngoc Tran, Trong-Nghia Nguyen, and Viet-Hung Dao. A practical tutorial on variational bayes. *arXiv preprint arXiv:2103.01327*, 2021.