

STAT 238 - Bayesian Statistics

Lecture Thirty Six

Spring 2026, UC Berkeley

Aditya Guntuboyina

27 April 2026

1 Variational Inference

Variational Inference is a method for approximating a posterior distribution by a simpler tractable family of distributions. It uses optimization to select a distribution q that is closest to the true posterior in some metric. In practice, variational inference can be much faster than MCMC, though sometimes less accurate in representing posterior uncertainty.

The basic idea behind variational inference is very simple. Consider the standard Bayesian setup with data given by y and parameters given by θ . The prior is $f_\theta(\theta)$ and the likelihood is $f_{y|\theta}(y)$. The posterior is then:

$$f_{\theta|y}(\theta) = \frac{f_\theta(\theta)f_{y|\theta}(y)}{\int f_\theta(\theta)f_{y|\theta}(y)d\theta}.$$

The denominator is the marginal density $f_y(y)$ of y . The marginal density $f_y(y)$ is also known as the **Evidence**. Using the marginal density for the denominator, we get

$$f_{\theta|y}(\theta) = \frac{f_\theta(\theta)f_{y|\theta}(y)}{f_y(y)}.$$

The posterior is usually intractable because of the integration involved in the computation of the Evidence (denominator). In variational inference, one chooses a simpler family of distributions \mathcal{Q} and then employs optimization techniques to choose a distribution in \mathcal{Q} that is as close as possible to the posterior $f_{\theta|y}(\theta)$. The distance-measure used to measure closeness here is most commonly the following Kullback-Leibler divergence:

$$KL(q||f_{\theta|y}) = \int q(\theta) \log \frac{q(\theta)}{f_{\theta|y}(\theta)} d\theta.$$

Recall that the KL divergence between two densities q and p is given by $\int q \log(q/p)$. It is always nonnegative and it equals zero if and only if $q = p$. It is not symmetric in general (i.e., $KL(q||p) \neq KL(p||q)$).

The optimization that we need to solve therefore is:

$$\operatorname{argmin}_{q \in \mathcal{Q}} KL(q||f_{\theta|y}). \tag{1}$$

The choice of this specific KL divergence is mostly for technical convenience as it makes the resulting optimization tractable. Observe that

$$KL(q\|f_{\theta|y}) = \int q(\theta) \log \frac{q(\theta)}{f_{\theta|y}(\theta)} d\theta = \int q(\theta) \log \frac{q(\theta)}{f_{y,\theta}(y, \theta)} d\theta + \log f_y(y). \quad (2)$$

The second term above (which involves the usually intractable $f_y(y)$) does not depend on q so the optimization (1) is equivalent to:

$$\operatorname{argmin}_{q \in \mathcal{Q}} \int q(\theta) \log \frac{q(\theta)}{f_{y,\theta}(y, \theta)} d\theta.$$

The above objective function is called the Variational Free Energy, or simply, Free Energy $F(q)$:

$$F(q) = \int q(\theta) \log \frac{q(\theta)}{f_{y,\theta}(y, \theta)} d\theta = - \int q(\theta) \log \frac{f_{y,\theta}(y, \theta)}{q(\theta)} d\theta. \quad (3)$$

So the main goal in Variational Inference is to minimize the Free Energy. The negative of the Free Energy is called the Evidence Lower Bound (ELBO):

$$\text{ELBO}(q) = -F(q) = \int q(\theta) \log \frac{f_{y,\theta}(y, \theta)}{q(\theta)} d\theta. \quad (4)$$

So the main goal in Variational Inference can also be said to be the maximization of the ELBO. The name ELBO comes from the fact that $\text{ELBO}(q)$ always satisfies:

$$\text{ELBO}(q) = \int q(\theta) \log \frac{f_{y,\theta}(y, \theta)}{q(\theta)} d\theta \leq \log f_y(y). \quad (5)$$

This is because of the expression (2) which says that the difference between $\log f_y(y)$ and $\text{ELBO}(q)$ equals the Kullback Leibler divergence $KL(q\|f_{\theta|y})$ which is nonnegative. Because $f_y(y)$ is called the evidence and $\text{ELBO}(q)$ gives a lower bound for the logarithm of the evidence, it is given the name Evidence Lower Bound. It is also easy to see that if we maximize $\text{ELBO}(q)$ over *all* probability densities q , then we obtain equality in (5):

$$\sup_q \text{ELBO}(q) = \log f_y(y). \quad (6)$$

The above follows from (2) because

$$\text{ELBO}(q) = \log f_y(y) - KL(q\|f_{\theta|y}).$$

Because $f_{y,\theta}(y, \theta) = f_{\theta}(y) f_{y|\theta}(y)$, the ELBO has the following alternative expression:

$$\begin{aligned} \text{ELBO}(q) &= -F(q) \\ &= \int q(\theta) \log \frac{f_{y,\theta}(y, \theta)}{q(\theta)} d\theta \\ &= \int q(\theta) \log \frac{f_{y|\theta}(y) f_{\theta}(y)}{q(\theta)} d\theta \\ &= \int q(\theta) \log f_{y|\theta}(y) d\theta - \int q(\theta) \log \frac{q(\theta)}{f_{\theta}(y)} d\theta \\ &= \int q(\theta) \log f_{y|\theta}(y) d\theta - KL(q\|f_{\theta}) \end{aligned} \quad (7)$$

In other words, $\text{ELBO}(q)$ represents the average likelihood (with respect to q) minus the KL divergence between q and the prior.

Usually, the main tasks in Variational Inference are (a) to choose a class of distributions \mathcal{Q} , and (b) to maximize $\text{ELBO}(q)$ (or, equivalently, to minimize $F(q)$) over $q \in \mathcal{Q}$. We will see how to do this via some examples.

2 Bayesian Logistic Regression

Consider again the logistic regression setting where we observe data $(x_i, y_i), i = 1, \dots, n$ with $x_i \in \mathbb{R}^d$ and $y_i \in \{0, 1\}$. The likelihood model is:

$$y_i \mid x_i, \beta \sim \text{Bernoulli} \left(\frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}} \right)$$

and we take the improper uniform prior for β . We then have:

$$\begin{aligned} f_{y, \beta}(y, \beta) &= \prod_{i=1}^n \left(\frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}} \right)^{y_i} \left(1 - \frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}} \right)^{1-y_i} \\ &= \prod_{i=1}^n \frac{\exp(y_i x_i^T \beta)}{1 + \exp(x_i^T \beta)} = \exp(\ell(\beta)) \end{aligned}$$

where $\ell(\beta)$ is the log-likelihood is

$$\ell(\beta) = \sum_{i=1}^n (y_i x_i^T \beta - \log(1 + \exp(x_i^T \beta))). \quad (8)$$

Note that, because the prior is one, $f_{y, \beta}(y, \beta)$ equals the likelihood $f_{y|\beta}(y)$. We can thus calculate the ELBO as:

$$\begin{aligned} \text{ELBO}(q) &= \int q(\beta) \log \frac{f_{y, \beta}(y, \beta)}{q(\beta)} d\beta \\ &= \int q(\beta) \ell(\beta) d\beta - \int q(\beta) \log q(\beta) d\beta. \end{aligned}$$

The quantity $-\int q \log q$ is known as the entropy $H(q)$ of q . We thus have

$$\text{ELBO}(q) = \int q(\beta) \ell(\beta) d\beta - H(q).$$

Let us now take \mathcal{Q} to be the class of all normal distributions $N(m, \Sigma)$ as $m \in \mathbb{R}^p$ and Σ varies over the class of all $p \times p$ positive definite matrices Σ . The entropy of the multivariate normal distribution (see https://en.wikipedia.org/wiki/Multivariate_normal_distribution) is:

$$H(N(m, \Sigma)) = \frac{p}{2} (1 + \log(2\pi)) + \frac{1}{2} \log |\Sigma|$$

where $|\Sigma|$ is the determinant of Σ . We thus have:

$$\text{ELBO}(m, \Sigma) = \mathbb{E}_{\beta \sim N(m, \Sigma)} \ell(\beta) + \frac{1}{2} \log |\Sigma| + \frac{p}{2} (1 + \log(2\pi)).$$

This can be maximized over m and Σ to obtain the posterior approximation $N(\hat{m}, \hat{\Sigma})$. Here m can be any vector over \mathbb{R}^p but Σ is constrained to be positive definite. Constrained optimization is usually difficult (for gradient-based methods) so we convert this to unconstrained optimization by taking

$$\Sigma = LL^T$$

where L is a $p \times p$ lower-triangular matrix with non-zero diagonal entries. We can also, without loss of generality, assume that L has strictly positive diagonal entries (given any L ,

we can replace it by LD where D is a diagonal matrix with entries $+1$ if the corresponding entry of L is positive and -1 otherwise). In code, we can represent the diagonal entries of L by $\exp(a_1), \dots, \exp(a_p)$. With this parametrization, we have

$$|\Sigma| = |LL^T| = |L|^2 = \prod_{j=1}^p L_{jj}^2,$$

and thus

$$\text{ELBO}(m, L) = \mathbb{E}_{\beta \sim N(m, LL^T)} \ell(\beta) + \sum_{j=1}^p \log L_{jj} + \frac{p}{2} (1 + \log(2\pi)).$$

In order to maximize the above function, we need to make the term $\mathbb{E}_{\beta \sim N(m, LL^T)} \ell(\beta)$ more explicit. For this, we use the *reparametrization* trick (see https://en.wikipedia.org/wiki/Reparameterization_trick) and write

$$\beta = m + Lz \quad \text{where } z \sim N(0, I_p).$$

This gives

$$\text{ELBO}(m, L) = \mathbb{E}_{z \sim N(0, I_p)} \ell(m + Lz) + \sum_{j=1}^p \log L_{jj} + \frac{p}{2} (1 + \log(2\pi)).$$

The expectation above can be approximated by Monte Carlo. Specifically generate

$$z^{(1)}, \dots, z^{(S)} \stackrel{\text{i.i.d.}}{\sim} N(0, I_p)$$

and use the approximation:

$$\text{ELBO}(m, L) \approx \frac{1}{S} \sum_{s=1}^S \ell(m + Lz^{(s)}) + \sum_{j=1}^p \log L_{jj} + \frac{p}{2} (1 + \log(2\pi)).$$

This can be maximized using standard gradient-based optimization software such as PyTorch.