

STAT 238 - Bayesian Statistics

Lecture Thirty Seven

Spring 2026, UC Berkeley

Aditya Guntuboyina

29 April 2026

1 Variational Inference

We looked at the basics of variational inference in the last lecture. 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:

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

where the denominator is the evidence $f_y(y) = \int f_{y|\theta}(y)f_\theta(\theta)d\theta$.

The most important quantity in variational inference is the ELBO defined as:

$$\text{ELBO}(q) := \int q(\theta) \log \frac{f_{y,\theta}(y, \theta)}{q(\theta)} d\theta \quad (1)$$

where q is an arbitrary probability density. The key facts about the ELBO are:

$$\text{ELBO}(q) \leq \log f_y(y) \quad \text{for every } q. \quad (2)$$

and

$$\max_q \text{ELBO}(q) = \log f_y(y) \quad \text{with the maximum achieved at } q(\theta) = f_{\theta|y}(\theta). \quad (3)$$

Both these facts are consequences of:

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

Another useful fact about the ELBO (which follows directly from the definition (1) when $f_{y,\theta}(y, \theta)$ is replaced by $f_{y|\theta}(y)f_\theta(\theta)$) is:

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

2 Connection to the EM Algorithm

The relation $\max_q \text{ELBO}(q) = \log f_y(y)$ (with the maximum being achieved at the posterior $q(\theta) = f_{\theta|y}(\theta)$) can be used to understand the EM algorithm.

Consider the same Bayesian setting as in the previous section, but suppose now that there is an additional hyperparameter α , and that both the likelihood $f_{y|\theta,\alpha}(y)$ as well as the prior $f_{\theta|\alpha}(\theta)$ depend on α .

The ELBO now becomes:

$$\text{ELBO}(q, \alpha) = \int q(\theta) \log \frac{f_{y,\theta|\alpha}(y, \theta)}{q(\theta)} d\theta = \int q(\theta) \log \frac{f_{y|\theta,\alpha}(y) f_{\theta|\alpha}(\theta)}{q(\theta)} d\theta$$

and equation (3) now becomes:

$$\max_q \text{ELBO}(q, \alpha) = \log f_{y|\alpha}(y) \quad \text{with the maximum achieved at } q(\theta) = f_{\theta|y,\alpha}(\theta). \quad (4)$$

Suppose now that we want to obtain the MLE $\hat{\alpha}$ of α i.e., we want to solve the optimization:

$$\text{Maximize}_{\alpha} \log f_{y|\alpha}(y).$$

Using (4), we can rewrite the optimization above as:

$$\text{Maximize}_{\alpha} \max_q \text{ELBO}(q, \alpha),$$

which can be viewed as the two parameter maximization problem:

$$\text{Maximize}_{q, \alpha} \text{ELBO}(q, \alpha).$$

It is natural to solve this two-parameter optimization alternatively. Fix α and optimize over q , and then fix q and optimize over α . This leads to the following algorithm.

1. Initialize $\alpha = \alpha^{(0)}$.
2. Repeat the following for $t = 0, 1, 2, \dots$ (until some kind of convergence):
 - a) Take $q^{(t)}$ to be the maximizer of $\text{ELBO}(q, \alpha^{(t)})$ over all q (for fixed $\alpha^{(t)}$)
 - b) Take $\alpha^{(t+1)}$ to be the maximizer of $\text{ELBO}(q^{(t)}, \alpha)$ over all α .

By (4), it is clear that $q^{(t)}$ which maximizes $\text{ELBO}(q, \alpha^{(t)})$ is simply equal to:

$$q^{(t)} = f_{\theta|y,\alpha^{(t)}}.$$

Thus the alternating minimization algorithm above is equivalent to:

1. Initialize $\alpha = \alpha^{(0)}$.
2. Repeat the following for $t = 0, 1, 2, \dots$ (until some kind of convergence):
 - a) Take $\alpha^{(t+1)}$ to be the maximizer of $\text{ELBO}(f_{\theta|y,\alpha^{(t)}}, \alpha)$ over all α .

The quantity $\text{ELBO}(f_{\theta|y,\alpha^{(t)}}, \alpha)$ equals:

$$\begin{aligned} \text{ELBO}(f_{\theta|y,\alpha^{(t)}}, \alpha) &= \int f_{\theta|y,\alpha^{(t)}}(\theta) \log \frac{f_{y,\theta|\alpha}(y, \theta)}{f_{\theta|y,\alpha^{(t)}}(\theta)} d\theta \\ &= \int f_{\theta|y,\alpha^{(t)}}(\theta) \log f_{y,\theta|\alpha}(\theta) d\theta - \int f_{\theta|y,\alpha^{(t)}}(\theta) \log f_{\theta|y,\alpha^{(t)}}(\theta) d\theta. \end{aligned}$$

The second term actually does not depend on α , so maximizing $\text{ELBO}(f_{\theta|y,\alpha^{(t)}}, \alpha)$ over α is equivalent to maximizing the first term above:

$$E(t, \alpha) := \int f_{\theta|y,\alpha^{(t)}}(\theta) \log f_{y,\theta|\alpha}(\theta) d\theta. \quad (5)$$

The alternating maximization algorithm above then becomes:

1. Initialize $\alpha = \alpha^{(0)}$.
2. Repeat the following for $t = 0, 1, 2, \dots$ (until some kind of convergence):
 - a) Calculate $E(t, \alpha)$ given in (5)
 - b) Take $\alpha^{(t+1)}$ to be the maximizer of $E(t, \alpha)$ over all α .

This is the EM algorithm. The calculation of $E(t, \alpha)$ is called the E -step and the maximization of $E(t, \alpha)$ over α is called the M -step. For more on this connection between Variational Inference (VI) and EM, see Neal and Hinton [1].

3 Coordinatewise Variational Inference

The key problem that we need to solve in VI is that of maximizing $\text{ELBO}(q)$. The full maximizer of $\text{ELBO}(q)$ over q is, as already mentioned, $f_{\theta|y}$. But this posterior is intractable in general and the very reason for using variational inference is this intractability. The strategy is to restrict the class of q over which we look for minimizers. This hopefully will lead to a *tractable* minimizer which is close to the actual posterior $f_{\theta|y}$.

Suppose θ can be decomposed into k separate classes of parameters as $\theta = (\theta_1, \dots, \theta_k)$. We look for variational approximations of the form $q(\theta) = q_1(\theta_1) \dots q_k(\theta_k)$. The ELBO maximization problem is:

$$\underset{q_1, \dots, q_k}{\text{maximize}} \text{ELBO}(q_1, \dots, q_k) = \int \dots \int q_1(\theta_1) \dots q_k(\theta_k) \log \frac{f_{y, \theta}(y, \theta)}{q_1(\theta_1) \dots q_k(\theta_k)} d\theta. \quad (6)$$

Alternating minimization leads to the following result.

Proposition 3.1 (CAVI). *Consider the maximization problem (6). Fix $j = 1, \dots, k$ and fix $q_i, i \neq j$. Then the minimizing density q_j^* is given by*

$$q_j^*(\theta_j) \propto \exp \left[\int q_{-j}(\theta_{-j}) \log f_{\theta_j|y, \theta_{-j}}(\theta_j) d\theta_{-j} \right]. \quad (7)$$

where $q_{-j}(\theta_j)$ denotes the product of $q_l(\theta_l)$ over all $l \neq j$ and $d\theta_j$ is the product of $d\theta_l$ over all $l \neq j$.

The CAVI update equation (7) is equivalent to the following:

$$\log q_j^*(\theta_j) = \text{constant} + \mathbb{E}_{\theta_{-j}} \log f_{y, \theta_1, \dots, \theta_k}(y, \theta_1, \dots, \theta_k) \quad (8)$$

where the expectation is taken with respect to $\theta_{-j} \sim q_{-j}$.

Proposition 3.1 is a consequence of the following simple fact.

Lemma 3.2. *1. Suppose g is a nonnegative function that is not necessarily a density function. Then the minimizer of*

$$L(f) := \int f \log \frac{f}{g}$$

over all densities f is given by

$$f^*(x) \propto g(x) \quad (9)$$

which of course means that $f^* = g / (\int g)$.

2. Suppose $f_2(\cdot)$ below is a fixed probability density function, and $g(x_1, x_2)$ is a fixed non-negative function that is not necessarily a density. Then the minimizer of

$$\Gamma(f_1) := \int \int f_1(x_1) f_2(x_2) \log \frac{f_1(x_1) f_2(x_2)}{g(x_1, x_2)} dx_1 dx_2$$

over all densities f_1 is given by

$$f_1^*(x_1) \propto \exp \left(\int f_2(x_2) \log g(x_1, x_2) dx_2 \right). \quad (10)$$

Proof of Lemma 3.2. The first fact is easy because we can write

$$\begin{aligned} L(f) &= \int f \log \frac{f}{g} \\ &= \int f \log \frac{f}{(f g)^{-1} g} - \log \left(\int g \right) \int f \\ &= \int f \log \frac{f}{(f g)^{-1} g} - \log \int g = KL(f \| (\int g)^{-1} g) - \log \int g, \end{aligned}$$

so it is clear that the minimizing f equals $(\int g)^{-1} g$.

For the second fact, write

$$\begin{aligned} \Gamma(f_1) &= \int \int f_1(x_1) f_2(x_2) \log \frac{f_1(x_1) f_2(x_2)}{g(x_1, x_2)} dx_1 dx_2 \\ &= \int \int f_1(x_1) f_2(x_2) \left[\log \frac{f_1(x_1)}{g(x_1, x_2)} + \log f_2(x_2) \right] dx_1 dx_2 \\ &= \int \int f_1(x_1) f_2(x_2) \log \frac{f_1(x_1)}{g(x_1, x_2)} + \int f_2(x_2) \log f_2(x_2) dx_2. \end{aligned}$$

We can ignore the second term as it does not depend on f_1 . So minimizing $\Gamma(f_1)$ is the same as minimizing the first term:

$$\begin{aligned} &\int \int f_1(x_1) f_2(x_2) \log \frac{f_1(x_1)}{g(x_1, x_2)} dx_1 dx_2 \\ &= \int f_1(x_1) \int f_2(x_2) [\log f_1(x_1) - \log g(x_1, x_2)] dx_2 dx_1 \\ &= \int f_1(x_1) \left[\log f_1(x_1) - \int f_2(x_2) \log g(x_1, x_2) dx_2 \right] dx_1 \\ &= \int f_1(x_1) \log \frac{f_1(x_1)}{\exp \left[\int f_2(x_2) \log g(x_1, x_2) dx_2 \right]} dx_1. \end{aligned}$$

We can thus use the first fact with $g(x_1) = \exp \left[\int f_2(x_2) \log g(x_1, x_2) dx_2 \right]$ to claim that the minimizer f_1^* of $\Gamma(f_1)$ satisfies (10). \square

4 Simple CAVI Example

As a simple illustration of how CAVI works, let us consider the problem of estimating μ, σ from observations y_1, \dots, y_n that are i.i.d $N(\mu, \sigma^2)$. As usual, we use the prior

$$\mu, \log \sigma \stackrel{\text{i.i.d}}{\sim} \text{uniform}(-\infty, \infty).$$

In other words, the prior is $f_{\mu, \sigma^2}(\mu, \sigma^2) = \frac{1}{\sigma^2}$ (note that we are treating σ^2 and not σ as the parameter).

The joint density of the data $y = (y_1, \dots, y_n)$ and the parameter $\theta = (\mu, \sigma^2)$ is given by:

$$f_{y, \theta}(y, \theta) = \left(\prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(y_i - \mu)^2}{2\sigma^2}\right) \right) \frac{1}{\sigma^2}$$

so that

$$\begin{aligned} \log f_{y, \theta}(y, \theta) &= -\sum_{i=1}^n \frac{(y_i - \mu)^2}{2\sigma^2} - \left(\frac{n}{2} + 1\right) \log \sigma^2 + \text{constant} \\ &= -\frac{n}{2\sigma^2} (\bar{y} - \mu)^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \bar{y})^2 - \left(\frac{n}{2} + 1\right) \log \sigma^2 + \text{constant}. \end{aligned}$$

CAVI approximates the posterior of $\theta = (\mu, \sigma^2)$ by a product distribution $q_\mu(\mu)q_{\sigma^2}(\sigma^2)$. Below we figure out how to update q_μ given q_{σ^2} and also q_{σ^2} given q_μ .

By (8), given q_{σ^2} , we get

$$\log q_\mu^*(\mu) = -\frac{n}{2} (\bar{y} - \mu)^2 \mathbb{E}_{q_{\sigma^2}} \left(\frac{1}{\sigma^2} \right) + \text{constant},$$

which means that q_μ^* is the density of the normal distribution: with mean μ and variance

$$q_\mu^* = \text{density of } N\left(\bar{y}, \frac{1}{n\mathbb{E}_{q_{\sigma^2}}(1/\sigma^2)}\right). \quad (11)$$

To get $q_{\sigma^2}^*(\sigma^2)$ for fixed q_μ , we again use (8) to get

$$\log q_{\sigma^2}^*(\sigma^2) = -\frac{1}{2\sigma^2} \left(n\mathbb{E}_{q_\mu}(\bar{y} - \mu)^2 + \sum_{i=1}^n (y_i - \bar{y})^2 \right) - \left(\frac{n}{2} + 1\right) \log \sigma^2 + \text{constant}.$$

This has the form of the Inverse Gamma density $IG(a, b)$ given by:

$$-\frac{b}{\sigma^2} - (a + 1) \log \sigma^2.$$

Comparing terms, we get

$$q_{\sigma^2}^* = \text{density of } IG\left(\frac{n}{2}, \frac{n}{2}\mathbb{E}_{q_\mu}(\bar{y} - \mu)^2 + \frac{1}{2}\sum_{i=1}^n (y_i - \bar{y})^2\right). \quad (12)$$

(11) and (12) give the following CAVI algorithm:

1. Initialize with some density $q_{\sigma^2}^{(0)}$ of σ^2 . Suppose $s_0 := \left(n\mathbb{E}_{q_{\sigma^2}^{(0)}}(1/\sigma^2)\right)^{-1}$.
2. Repeat the following for $k = 0, 1, 2, \dots$:
 - a) Take $q_\mu^{(k+1)}$ to be the density of $N(\bar{y}, s_k^2)$.
 - b) Take $q_{\sigma^2}^{(k+1)}$ to be the density of

$$IG\left(\frac{n}{2}, \frac{n}{2}\mathbb{E}_{q_\mu^{(k+1)}}(\bar{y} - \mu)^2 + \frac{1}{2}\sum_{i=1}^n (y_i - \bar{y})^2\right) = IG\left(\frac{n}{2}, \frac{n}{2}s_k^2 + \frac{1}{2}\sum_{i=1}^n (y_i - \bar{y})^2\right)$$

c) Take $s_{k+1}^2 = \left(n \mathbb{E}_{q_{\sigma^2}^{(k+1)}}(1/\sigma^2) \right)^{-1}$ which leads to (using the fact $\mathbb{E}\sigma^{-2} = a/b$ when $\sigma^2 \sim IG(a, b)$)

$$s_{k+1}^2 = \frac{s_k^2}{n} + \frac{1}{n^2} \sum_{i=1}^n (y_i - \bar{y})^2. \quad (13)$$

The limit of the iteration for $\{s_k^2\}$ in (13) can be computed in closed form. Indeed if $s^2 = \lim_{k \rightarrow \infty} s_k^2$, then (13) gives

$$s^2 = \frac{s^2}{n} + \frac{1}{n^2} \sum_{i=1}^n (y_i - \bar{y})^2 \implies s^2 = \frac{1}{n(n-1)} \sum_{i=1}^n (y_i - \bar{y})^2.$$

Thus the VI solution $q_{\mu}^*(\mu)q_{\sigma^2}^*(\sigma^2)$ is given by:

$$q_{\mu}^* = \text{density of } N\left(\bar{y}, \frac{1}{n(n-1)} \sum_{i=1}^n (y_i - \bar{y})^2\right) \quad (14)$$

and

$$\begin{aligned} q_{\sigma^2}^* &= \text{density of } IG\left(\frac{n}{2}, \frac{1}{2(n-1)} \sum_{i=1}^n (y_i - \bar{y})^2 + \frac{1}{2} \sum_{i=1}^n (y_i - \bar{y})^2\right) \\ &= \text{density of } IG\left(\frac{n}{2}, \frac{n}{2(n-1)} \sum_{i=1}^n (y_i - \bar{y})^2\right) \end{aligned} \quad (15)$$

We can compare these VI marginals to those of the exact posterior which we know is given by:

$$\mu \mid \text{data} \sim t_{n-1}\left(\bar{y}, \frac{1}{n(n-1)} \sum_{i=1}^n (y_i - \bar{y})^2\right) \quad \text{and} \quad \sigma^2 \mid \text{data} \sim IG\left(\frac{n-1}{2}, \frac{1}{2} \sum_{i=1}^n (y_i - \bar{y})^2\right).$$

Comparing the above to (14), we see that the VI marginal posterior has a narrower normal density instead of t_{n-1} . Comparing the above to (15), we see that the parameters of the IG distribution are changed slightly (the changes are negligible for large n). Note also that μ and σ^2 are independent in the actual posterior but this VI analysis assumes they are independent.

References

- [1] Neal, R. M. and Hinton, G. E. (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. In M. I. Jordan (ed.), *Learning in Graphical Models*, pages 355–368. Springer Netherlands.