Variational Inference

Handout Notes - Generative and Deep Learning (GDL)

Davide Bacciu - University of Pisa

Notation. Random variables are uppercase (e.g., X, Z) and observed values are lowercase (e.g., x, z). A dataset is $\mathcal{D} = \{x_1, \ldots, x_N\}$ with $N = |\mathcal{D}|$. Model parameters are θ. We use $p(\cdot)$ for probability mass functions/densities and $\ell(\theta) = \log p(\mathcal{D} \mid \theta)$ for the log-likelihood. Latent variables are denoted Z (and z for values). The exact posterior is $p(z \mid x, \theta)$. Variational inference introduces a tractable distribution $q(z \mid \phi)$ with variational parameters ϕ.

1 Why variational inference?

Latent-variable models explain observed data using unobserved variables. A generic model specifies a joint distribution

$$p(x, z \mid \theta) = p(x \mid z, \theta)\, p(z \mid \theta).$$

Learning by maximum likelihood aims to maximize

$$\ell(\theta) = \log p(\mathcal{D} \mid \theta) = \sum_{i=1}^{N} \log p(x_i \mid \theta), \qquad p(x \mid \theta) = \int p(x, z \mid \theta)\, dz \quad (\text{or } \sum_{z} \text{ if } Z \text{ is discrete}).$$

The conceptual difficulty is simple: marginalizing out z often produces integrals/sums without closed form. The computational difficulty is sharper: even when $p(x, z \mid \theta)$ is easy to evaluate, the posterior

$$p(z \mid x, \theta) = \frac{p(x, z \mid \theta)}{p(x \mid \theta)}$$

can be intractable because it requires the same marginal likelihood $p(x \mid \theta)$ we cannot compute.

Expectation–Maximization (EM) relies on computing posterior expectations in its E-step. When the posterior is intractable, EM in its exact form cannot be applied. Variational inference offers a systematic workaround: replace *sampling/integration* with *optimization* by introducing a tractable approximation $q(z \mid \phi)$.

2 From calculus to variational calculus (intuition)

Ordinary calculus optimizes over numbers: $\max_x f(x)$. Variational methods optimize over *functions* or *distributions*:

$$\max_{q} \mathcal{F}(q),$$

where $\mathcal{F}$ is a functional. In our setting, the object we optimize over is a distribution $q(z \mid \phi)$, usually restricted to a tractable family (e.g., fully factorized distributions).

3 Two tools: KL divergence and Jensen's inequality

Variational inference is powered by (i) a way to measure how close two distributions are and (ii) a way to lower-bound a difficult quantity.

3.1 Kullback–Leibler divergence

For two distributions $q(z)$ and $p(z)$, the KL divergence is

$$\mathrm{KL}(q\|p) = \mathbb{E}_q\left[\log \frac{q(z)}{p(z)}\right].$$

It is always nonnegative and equals zero if and only if $q(z) = p(z)$ almost everywhere. In latent-variable learning, the key comparison is between the variational approximation and the true posterior:

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

Note that KL is not symmetric, so $\mathrm{KL}(q\|p) \neq \mathrm{KL}(p\|q)$ in general.

3.2 Jensen's inequality in probabilistic form

If f is concave (such as $f = \log$), then

$$f(\mathbb{E}[Y]) \geq \mathbb{E}[f(Y)].$$

This inequality becomes useful when we can express a difficult log-likelihood as the log of an expectation under q.

4 Bounding the log-likelihood: the ELBO

The central idea of variational inference is to derive a tractable lower bound on $\log p(x \mid \theta)$. This lower bound is the *evidence lower bound* (ELBO).

Worked example — Deriving the ELBO with Jensen's inequality

Start from the marginal likelihood for a single observation x:

$$\log p(x \mid \theta) = \log \int p(x, z \mid \theta)\, dz.$$

Introduce any distribution $q(z \mid \phi)$ that is nonzero where $p(x, z \mid \theta)$ is nonzero and use the identity

$$\int p(x, z \mid \theta)\, dz = \int q(z \mid \phi)\, \frac{p(x, z \mid \theta)}{q(z \mid \phi)}\, dz = \mathbb{E}_{q(z|\phi)}\left[\frac{p(x, z \mid \theta)}{q(z \mid \phi)}\right].$$

Therefore,

$$\log p(x \mid \theta) = \log \mathbb{E}_{q(z|\phi)}\left[\frac{p(x, z \mid \theta)}{q(z \mid \phi)}\right] \;\geq\; \mathbb{E}_{q(z|\phi)}\left[\log \frac{p(x, z \mid \theta)}{q(z \mid \phi)}\right],$$

where the inequality is Jensen's (log is concave). Define the ELBO

$$\boxed{\mathcal{L}(x, \theta, \phi) \;=\; \mathbb{E}_{q(z|\phi)}[\log p(x, z \mid \theta)] - \mathbb{E}_{q(z|\phi)}[\log q(z \mid \phi)].}$$

The second term is the entropy of q, i.e., $H(q) = -\mathbb{E}_q[\log q]$, so equivalently

$$\mathcal{L}(x, \theta, \phi) = \mathbb{E}_q[\log p(x, z \mid \theta)] + H(q).$$

For a dataset $\mathcal{D} = \{x_1, \ldots, x_N\}$ with (typically independent) latent variables $z_1, \ldots, z_N$, a common choice is a factorized variational distribution $q(\mathbf{z} \mid \phi) = \prod_{i=1}^{N} q(z_i \mid \phi)$, yielding

$$\log p(\mathcal{D} \mid \theta) = \sum_{i=1}^{N} \log p(x_i \mid \theta) \;\geq\; \sum_{i=1}^{N} \mathcal{L}(x_i, \theta, \phi).$$

5 How tight is the bound? ELBO and KL as an exact decomposition

The ELBO is not only a lower bound; it is part of an equality that explains its meaning.

Worked example — Exact relationship: $\log p(x \mid \theta) = \mathcal{L}(x, \theta, \phi) + \mathrm{KL}(q \| p)$

Start from the KL divergence to the true posterior:

$$\mathrm{KL}\big(q(z \mid \phi) \,\|\, p(z \mid x, \theta)\big) = \mathbb{E}_q[\log q(z \mid \phi) - \log p(z \mid x, \theta)]\,.$$

Use Bayes' rule: $\log p(z \mid x, \theta) = \log p(x, z \mid \theta) - \log p(x \mid \theta)$. Substitute:

$$\mathrm{KL}(q \| p) = \mathbb{E}_q[\log q(z \mid \phi) - \log p(x, z \mid \theta) + \log p(x \mid \theta)].$$

Rearrange terms (note that $\log p(x \mid \theta)$ does not depend on z):

$$\log p(x \mid \theta) = \mathbb{E}_q[\log p(x, z \mid \theta)] - \mathbb{E}_q[\log q(z \mid \phi)] + \mathrm{KL}\big(q(z \mid \phi) \,\|\, p(z \mid x, \theta)\big).$$

Recognizing the ELBO definition yields the identity

$$\boxed{\log p(x \mid \theta) = \mathcal{L}(x, \theta, \phi) + \mathrm{KL}\big(q(z \mid \phi) \,\|\, p(z \mid x, \theta)\big).}$$

Since $\mathrm{KL}(\cdot \| \cdot) \geq 0$, the ELBO is indeed a lower bound. The bound is tight iff $q(z \mid \phi) = p(z \mid x, \theta)$.

This decomposition reveals the operational meaning of variational inference:

> *Maximizing the ELBO is equivalent to minimizing the KL divergence between $q(z \mid \phi)$ and the true posterior, up to the (unknown) constant $\log p(x \mid \theta)$.*

6 Variational inference as a double optimization problem

Variational learning typically solves the coupled optimization

$$\max_{\theta, \phi} \sum_{i=1}^{N} \mathcal{L}(x_i, \theta, \phi).$$

This is a *double optimization*:

- optimize ϕ so that $q(z \mid \phi)$ approximates the posterior well,
- optimize θ so that the model assigns high probability to the data (as measured by the bound).

Two equivalent perspectives are useful in practice:

1. **Bound maximization:** directly maximize $\mathcal{L}(x, \theta, \phi)$ over θ and ϕ.
2. **KL minimization:** for fixed θ, choose ϕ to minimize $\mathrm{KL}(q(z \mid \phi) \| p(z \mid x, \theta))$.

These are the same because of the exact decomposition $\log p(x \mid \theta) = \mathcal{L}(x, \theta, \phi) + \mathrm{KL}(q \| p)$.

7 Variational EM: a generalized form of EM

Classical EM alternates between computing an exact posterior over z (E-step) and maximizing an expected complete-data log-likelihood (M-step). Variational EM keeps the alternating structure but replaces the exact posterior with a tractable approximation.

7.1 The alternating updates

At iteration k:

- **Variational E-step:** update ϕ to improve $q(z \mid \phi)$ under the current model parameters $\theta^{(k)}$:

$$\phi^{(k)} \in \arg\max_{\phi} \sum_{i=1}^{N} \mathcal{L}(x_i, \theta^{(k)}, \phi) \quad \text{(equivalently, minimize KL}(q\|p) \text{ for fixed } \theta^{(k)}).$$

- **M-step:** update θ to improve the bound under the current variational distribution:

$$\theta^{(k+1)} \in \arg\max_{\theta} \sum_{i=1}^{N} \mathcal{L}(x_i, \theta, \phi^{(k)}).$$

Worked example — Recovering exact EM as a special case

If the true posterior $p(z \mid x, \theta)$ is tractable, we may choose the variational family to include it and set

$$q(z \mid \phi^{(k)}) = p(z \mid x, \theta^{(k)}).$$

Then $\mathrm{KL}(q\|p) = 0$ and the ELBO becomes an equality:

$$\mathcal{L}(x, \theta^{(k)}, \phi^{(k)}) = \log p(x \mid \theta^{(k)}).$$

In this case, the variational E-step coincides with the classical EM E-step, and variational EM reduces to exact EM.

8 Pseudocode: variational EM (general template)

Worked example — Variational EM algorithm (template)

Input: dataset $\mathcal{D} = \{x_i\}_{i=1}^N$, model $p(x, z \mid \theta)$, variational family $\{q(z \mid \phi)\}$, tolerance ε, max iterations $K_{\max}$.
Output: parameters θ and variational parameters ϕ.

Initialize: choose $\theta^{(0)}$ and $\phi^{(0)}$. Set $k \leftarrow 0$.

Repeat until convergence or $k = K_{\max} - 1$:

1. **Variational E-step:** update the variational parameters

$$\phi^{(k)} \leftarrow \arg\max_{\phi} \sum_{i=1}^{N} \mathcal{L}(x_i, \theta^{(k)}, \phi),$$

   using (i) closed-form coordinate updates (if available) or (ii) gradient ascent on the ELBO.
2. **M-step:** update the model parameters

$$\theta^{(k+1)} \leftarrow \arg\max_{\theta} \sum_{i=1}^{N} \mathcal{L}(x_i, \theta, \phi^{(k)}),$$

   often resembling maximum likelihood with latent variables where expectations are taken under $q(z \mid \phi^{(k)})$.
3. **Check convergence:** compute the objective

$$\mathcal{J}^{(k+1)} = \sum_{i=1}^{N} \mathcal{L}(x_i, \theta^{(k+1)}, \phi^{(k)}),$$

   and stop if $\left|\mathcal{J}^{(k+1)} - \mathcal{J}^{(k)}\right| \leq \varepsilon$. Otherwise set $k \leftarrow k + 1$.

Return $\theta^{(k+1)}$ and $\phi^{(k)}$.

Practical note. In large-scale settings, it is common to use minibatches and stochastic gradient ascent on the ELBO (Stochastic Variational Inference). When optimizing ϕ with gradients, careful reparameterizations or control variates may be needed to reduce gradient variance, depending on the model and variational family.