Variational Inference: KL Divergence, Jensen, and the ELBO

N-20251010

§1.1 The approximate posterior as an optimization problem

The true posterior in a latent variable model is

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

It is usually intractable because the marginal likelihood

$$p_\theta(x) = \int p_\theta(x, z)\, dz$$

is hard to compute.

Variational inference introduces a tractable family

$$q_\phi(z \mid x)$$

and tries to make it close to the true posterior. A natural measure of closeness is

$$D_{\mathrm{KL}}(q_\phi(z \mid x) \| p_\theta(z \mid x)).$$

This is the reverse direction from what one might first guess: we place expectation under q_ϕ, because q_ϕ is the distribution from which we can sample and whose density we can evaluate.

The obstacle is that this KL divergence contains $p_\theta(z \mid x)$, which itself contains the intractable $p_\theta(x)$. The central algebraic trick is to rewrite it so that $\log p_\theta(x)$ appears as a constant plus a lower bound.

§1.2 The ELBO identity

Start from the KL divergence:

$$D_{\mathrm{KL}}(q_\phi(z \mid x) \| p_\theta(z \mid x)) = \mathbb{E}_{q_\phi(z|x)} \left[\log \frac{q_\phi(z \mid x)}{p_\theta(z \mid x)} \right].$$

Using

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

we get

$$\begin{aligned} D_{\mathrm{KL}}(q_\phi(z \mid x) \| p_\theta(z \mid x)) &= \mathbb{E}_{q_\phi} \left[\log q_\phi(z \mid x) - \log p_\theta(x, z) + \log p_\theta(x) \right] \\ &= \log p_\theta(x) - \mathbb{E}_{q_\phi} \left[\log p_\theta(x, z) - \log q_\phi(z \mid x) \right]. \end{aligned}$$

Define

$$\mathcal{L}(\theta, \phi; x) = \mathbb{E}_{q_\phi(z|x)} \left[\log p_\theta(x, z) - \log q_\phi(z \mid x) \right].$$

Then the identity becomes

$$\log p_\theta(x) = \mathcal{L}(\theta, \phi; x) + D_{\mathrm{KL}}(q_\phi(z \mid x) \| p_\theta(z \mid x)).$$

Since KL divergence is nonnegative,

$$\mathcal{L}(\theta, \phi; x) \leq \log p_\theta(x).$$

This is why $\mathcal{L}$ is called the evidence lower bound, or ELBO.

Remark 1.2.1 — The ELBO is not introduced by magic. It is exactly what remains when the intractable posterior KL is separated from the log marginal likelihood.

§1.3 Jensen's inequality derivation

The same lower bound can be derived directly from Jensen's inequality. Insert any density $q_\phi(z \mid x)$ whose support is compatible with $p_\theta(x, z)$:

$$\begin{aligned}
\log p_\theta(x) &= \log \int p_\theta(x, z)\, dz \\
&= \log \int q_\phi(z \mid x) \frac{p_\theta(x, z)}{q_\phi(z \mid x)}\, dz \\
&= \log \mathbb{E}_{q_\phi(z|x)} \left[\frac{p_\theta(x, z)}{q_\phi(z \mid x)} \right].
\end{aligned}$$

Because log is concave,

$$\begin{aligned}
\log p_\theta(x) &\geq \mathbb{E}_{q_\phi(z|x)} \left[\log \frac{p_\theta(x, z)}{q_\phi(z \mid x)} \right] \\
&= \mathbb{E}_{q_\phi(z|x)} \left[\log p_\theta(x, z) - \log q_\phi(z \mid x) \right] \\
&= \mathcal{L}(\theta, \phi; x).
\end{aligned}$$

The two derivations emphasize different things. The KL derivation shows the gap between the bound and the true log-likelihood. The Jensen derivation shows why a lower bound exists at all.

§1.4 Reconstruction term and regularization term

Using the latent variable factorization

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

the ELBO becomes

$$\begin{aligned}
\mathcal{L}(\theta, \phi; x) &= \mathbb{E}_{q_\phi(z|x)} \left[\log p_\theta(x \mid z) + \log p(z) - \log q_\phi(z \mid x) \right] \\
&= \mathbb{E}_{q_\phi(z|x)} [\log p_\theta(x \mid z)] - D_{\mathrm{KL}}(q_\phi(z \mid x) \| p(z)).
\end{aligned}$$

This is the familiar VAE decomposition:

$$\boxed{\mathcal{L} = \text{expected log-likelihood} - \text{KL to the prior.}}$$

The first term asks whether latent samples from $q_\phi(z \mid x)$ explain x well under the decoder. The second term asks whether the approximate posterior stays close to the prior.

It is common to call the first term the reconstruction term. But mathematically it is an expected log-likelihood. It becomes a squared-error reconstruction loss only under a Gaussian decoder with fixed variance. It becomes a cross-entropy-like loss under a Bernoulli decoder.

§1.5 The ELBO gap

The identity

$$\log p_\theta(x) = \mathcal{L}(\theta, \phi; x) + D_{\mathrm{KL}}(q_\phi(z \mid x) \| p_\theta(z \mid x))$$

shows that there are two goals mixed together.

Increasing $\mathcal{L}$ can improve the model likelihood $\log p_\theta(x)$. But it also tries to reduce the posterior approximation gap

$$D_{\mathrm{KL}}(q_\phi(z \mid x) \| p_\theta(z \mid x)).$$

If q_ϕ were flexible enough to equal the true posterior, then the bound would be tight:

$$\mathcal{L}(\theta, \phi; x) = \log p_\theta(x).$$

In practice, q_ϕ is restricted, often to a diagonal Gaussian, so the bound is generally not exact.

§1.6 Closed-form KL for diagonal Gaussians

The standard VAE uses

$$q_\phi(z \mid x) = \mathcal{N}(z; \mu, \mathrm{diag}(\sigma^2)), \qquad p(z) = \mathcal{N}(0, I).$$

For this pair, the KL divergence has a closed form:

$$D_{\mathrm{KL}}(q \| p) = \frac{1}{2} \sum_{j=1}^{d} \left(\mu_j^2 + \sigma_j^2 - 1 - \log \sigma_j^2 \right).$$

Here is the one-dimensional calculation. Let

$$q(z) = \mathcal{N}(\mu, \sigma^2), \qquad p(z) = \mathcal{N}(0, 1).$$

Then

$$\log q(z) = -\frac{1}{2} \log(2\pi\sigma^2) - \frac{(z-\mu)^2}{2\sigma^2},$$

and

$$\log p(z) = -\frac{1}{2} \log(2\pi) - \frac{z^2}{2}.$$

Therefore

$$\begin{aligned} D_{\mathrm{KL}}(q \| p) &= \mathbb{E}_q[\log q(z) - \log p(z)] \\ &= -\frac{1}{2} \log \sigma^2 - \frac{1}{2} \mathbb{E}_q \left[\frac{(z-\mu)^2}{\sigma^2} \right] + \frac{1}{2} \mathbb{E}_q[z^2] \\ &= -\frac{1}{2} \log \sigma^2 - \frac{1}{2} + \frac{1}{2}(\mu^2 + \sigma^2) \\ &= \frac{1}{2}(\mu^2 + \sigma^2 - 1 - \log \sigma^2). \end{aligned}$$

The multivariate diagonal formula follows by summing over independent coordinates.

This is one reason the diagonal Gaussian encoder is practical: the regularization term does not need Monte Carlo estimation.

§1.7 Negative ELBO as a loss

Training usually minimizes the negative ELBO:

$$-\mathcal{L}(\theta, \phi; x) = -\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x \mid z)] + D_{\mathrm{KL}}(q_\phi(z \mid x) \| p(z)).$$

For a Bernoulli decoder with probabilities $\pi_\theta(z)$,

$$p_\theta(x \mid z) = \prod_i \pi_i(z)^{x_i} (1 - \pi_i(z))^{1 - x_i},$$

so

$$-\log p_\theta(x \mid z) = -\sum_i \left[x_i \log \pi_i(z) + (1 - x_i) \log(1 - \pi_i(z)) \right].$$

For a Gaussian decoder

$$p_\theta(x \mid z) = \mathcal{N}(x; \mu_\theta(z), \sigma^2 I),$$

we have

$$-\log p_\theta(x \mid z) = \frac{1}{2\sigma^2} \|x - \mu_\theta(z)\|^2 + \text{constant}.$$

Thus the common reconstruction losses are not arbitrary engineering choices. They are negative log-likelihoods under different decoder distributions.

§1.8 Empirical lower bounds

The original AEVB experiments plotted the variational lower bound during training. The figure below is useful as a historical reminder: what is being optimized is not raw reconstruction error, but a lower bound $\mathcal{L}$ on marginal likelihood.

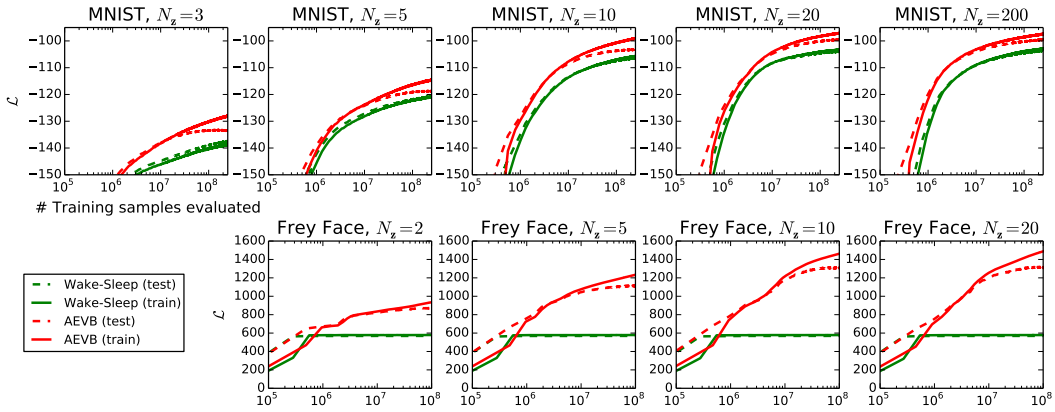

Figure 1.1: Variational lower bound curves in the AEVB experiments.

For the purpose of these notes, the important point is mathematical rather than empirical: the plotted quantity is meaningful because of the identity

$$\log p_\theta(x) = \mathcal{L}(\theta, \phi; x) + D_{\mathrm{KL}}(q_\phi(z \mid x) \| p_\theta(z \mid x)).$$

§1.9 Information-theoretic reading

The ELBO also resembles a rate–distortion tradeoff. The expected log-likelihood asks the latent variable to preserve enough information to explain x. The KL term penalizes deviations of $q_\phi(z \mid x)$ from the prior $p(z)$.

Loosely:

term	role
$-\mathbb{E}_q[\log p_\theta(x \mid z)]$	distortion / reconstruction cost
$D_{\mathrm{KL}}(q_\phi(z \mid x) \| p(z))$	rate / information cost

This interpretation is helpful, but it should not replace the probabilistic derivation. The ELBO is first of all a lower bound on marginal likelihood.

§1.10 Limits of the ELBO

The ELBO is powerful, but it is not magic. Several limitations remain.

First, the variational family may be too small. A diagonal Gaussian $q_\phi(z \mid x)$ cannot represent a complicated multimodal posterior.

Second, the prior may be too simple. The standard normal prior is convenient, but the true aggregated latent structure of data may not be standard Gaussian.

Third, the decoder may be too powerful. If the decoder can model the data without using z, the approximate posterior may collapse toward the prior. This is posterior collapse.

Fourth, maximizing a lower bound is not the same as exactly maximizing $\log p_\theta(x)$. The gap is the posterior KL.

§1.11 What survives later

The VAE objective is not merely a heuristic sum of two losses. It is a variational identity:

$$\log p_\theta(x) = \mathcal{L}(\theta, \phi; x) + D_{\mathrm{KL}}(q_\phi(z \mid x) \| p_\theta(z \mid x)).$$

The ELBO can be written as

$$\mathcal{L}(\theta, \phi; x) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x \mid z)] - D_{\mathrm{KL}}(q_\phi(z \mid x) \| p(z)).$$

The first form explains why it is a lower bound. The second form explains why it looks like reconstruction plus regularization.

The remaining question is computational: how can we differentiate an expectation with respect to parameters that also control the sampling distribution? That is the role of the reparameterization trick.