Variational Inference and Jensen's Inequality

Jash Vira

PS: Personal notes for intuition building

Variational inference turns posterior inference into optimisation.

x is observed, z is latent, and $p(x, z)$ is the joint model.

The posterior is

$$p(z \mid x) = \frac{p(x, z)}{p(x)}. \tag{1}$$

The difficulty is the evidence (the probability the model assigns to the observed data):

$$p(x) = \int p(x, z)\, dz \tag{2}$$

is often intractable.

So we choose a tractable family $q_\phi(z \mid x)$.

$$q^*_\phi(z \mid x) \approx p(z \mid x) \tag{3}$$

and fit it by

$$\phi^* = \arg\min_\phi \mathrm{KL}\big(q_\phi(z \mid x) \,\|\, p(z \mid x)\big). \tag{4}$$

Why Rewrite the Evidence?

Eq. (4) is the right objective in principle.

Using Eq. (1), it contains the evidence $p(x)$ in the denominator, and Eq. (2) is intractable.

So we rewrite the evidence itself:

$$\begin{aligned} p(x) &= \int p(x, z)\, dz \\ &= \int q_\phi(z \mid x)\, \frac{p(x, z)}{q_\phi(z \mid x)}\, dz \\ &= \mathbb{E}_{q_\phi(z|x)}\left[\frac{p(x, z)}{q_\phi(z \mid x)} \right]. \end{aligned} \tag{5}$$

Taking logs gives

$$\log p(x) = \log \mathbb{E}_{q_\phi(z|x)}\left[\frac{p(x, z)}{q_\phi(z \mid x)} \right]. \tag{6}$$

This is still exact; we have only multiplied and divided by $q_\phi(z \mid x)$.

Now the outer function is log, which is concave, so Jensen applies next.

Apply Jensen

Let

$$Y = \frac{p(x,z)}{q_\phi(z \mid x)}.$$

Jensen's inequality says that for concave f,

$$f(\mathbb{E}[Y]) \geq \mathbb{E}[f(Y)]. \tag{7}$$

Now take $f = \log$. Since log is concave, Eq. (7) gives

$$\log p(x) \geq \mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p(x,z)}{q_\phi(z \mid x)}\right]. \tag{8}$$

So Jensen turns the exact quantity into a lower bound.

A graphical intuition for this Jensen step is in the Appendix.

This Bound Is the ELBO

Using $\log(a/b) = \log a - \log b$ in Eq. (8),

$$\log p(x) \geq \mathbb{E}_{q_\phi(z|x)}[\log p(x,z) - \log q_\phi(z \mid x)] \;=:\; \mathrm{ELBO}(\phi). \tag{9}$$

So the ELBO is a tractable lower bound on $\log p(x)$.

It is the quantity we can optimise when $\log p(x)$ is out of reach.

Raising it means q_ϕ prefers latent states that fit the data better.

ELBO and KL

Start from the KL mismatch:

$$\begin{aligned}
\mathrm{KL}\big(q_\phi(z \mid x) \,\|\, p(z \mid x)\big) &= \mathbb{E}_{q_\phi(z|x)}[\log q_\phi(z \mid x) - \log p(z \mid x)] \\
&= \mathbb{E}_{q_\phi(z|x)}[\log q_\phi(z \mid x) - \log p(x, z) + \log p(x)] \qquad \text{by Eq. (1)} \\
&= \log p(x) - \mathrm{ELBO}(\phi) \qquad \text{by Eq. (9).}
\end{aligned}$$

$$\log p(x) = \mathrm{ELBO}(\phi) + \mathrm{KL}\big(q_\phi(z \mid x) \,\|\, p(z \mid x)\big). \tag{10}$$

So the gap is exactly the KL mismatch.

Since $\log p(x)$ does not depend on ϕ, Eq. (10) implies

$$\max_\phi \ \mathrm{ELBO}(\phi) \quad \Longleftrightarrow \quad \min_\phi \ \mathrm{KL}\big(q_\phi(z \mid x) \,\|\, p(z \mid x)\big). \tag{11}$$

So ELBO maximisation is KL minimisation.

Appendix

Jensen's Inequality

Jensen compares

$$f(\mathbb{E}[X]) \quad \text{and} \quad \mathbb{E}[f(X)].$$

The direction is set by curvature:

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)] \tag{12}$$

if f is convex,

$$f(\mathbb{E}[X]) \geq \mathbb{E}[f(X)] \tag{13}$$

if f is concave.

So curvature tells you which side is larger.

$\boxed{\text{Averaging and nonlinearity do not commute.}}$

The next slide shows this gap geometrically.

How to read the graph

Red construction

Choose $\mathbb{E}[X]$, go up to the curve, then left:

$$\varphi(\mathbb{E}[X]).$$

Black level

Transform first, then average:

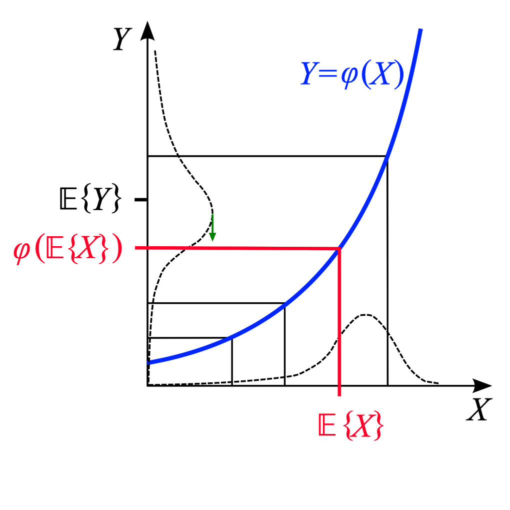
$$\mathbb{E}[Y] = \mathbb{E}[\varphi(X)].$$

Green arrow

This is the vertical gap,

$$\mathbb{E}[\varphi(X)] - \varphi(\mathbb{E}[X]).$$

Here the black level lies above the red one.



This is the Jensen step from Eq. (5) to Eq. (9).

Image: *Jensen graph.svg*, by MHz‘as with later edits by Arkan0ID, via Wikimedia Commons, public domain.