The Evidence Lower Bound

François Fleuret

March 2, 2024

Given i.i.d training samples $x_1, \ldots, x_N$ we want to fit a model $p_\theta(x, z)$ to it, maximizing

$$\sum_n \log p_\theta(x_n).$$

If we do not have an analytical form of the marginal $p_\theta(x_n)$ but only the expression of $p_\theta(x_n, z)$, we can get an estimate of the marginal by sampling z with any distribution q

$$\begin{aligned} p_\theta(x_n) &= \int_z p_\theta(x_n, z) dz \\ &= \int_z \frac{p_\theta(x_n, z)}{q(z)} q(z) dz \\ &= \mathbb{E}_{Z \sim q(z)} \left[\frac{p_\theta(x_n, Z)}{q(Z)} \right]. \end{aligned}$$

So if we sample a Z with q and maximize

$$\frac{p_\theta(x_n, Z)}{q(Z)},$$

we do maximize $p_\theta(x_n)$ on average.

But we want to maximize $\sum_n \log p_\theta(x_n)$. If we use the $\log$ of the previous expression, we can decompose its average value as

$$\begin{aligned}
&\mathbb{E}_{Z\sim q(z)}\left[\log \frac{p_\theta(x_n, Z)}{q(Z)}\right] \\
&= \mathbb{E}_{Z\sim q(z)}\left[\log \frac{p_\theta(Z \mid x_n) p_\theta(x_n)}{q(Z)}\right] \\
&= \mathbb{E}_{Z\sim q(z)}\left[\log \frac{p_\theta(Z \mid x_n)}{q(Z)}\right] + \log p_\theta(x_n) \\
&= -\mathbb{D}_{\mathsf{KL}}(q(z) \,\|\, p_\theta(z \mid x_n)) + \log p_\theta(x_n).
\end{aligned}$$

Hence this does not maximize $\log p_\theta(x_n)$ on average, but a *lower bound* of it, since the KL divergence is non-negative. And since this maximization pushes that KL term down, it also aligns $p_\theta(z \mid x_n)$ and $q(z)$, and we may get a worse $p_\theta(x_n)$ to bring $p_\theta(z \mid x_n)$ closer to $q(z)$.

However, all this analysis is still valid if q is a parameterized function $q_\alpha(z \mid x_n)$ of x_n. In that case, if we optimize θ and α to maximize

$$\mathbb{E}_{Z\sim q_\alpha(z \mid x_n)}\left[\log \frac{p_\theta(x_n, Z)}{q_\alpha(Z \mid x_n)}\right],$$

it maximizes $\log p_\theta(x_n)$ and brings $q_\alpha(z \mid x_n)$ close to $p_\theta(z \mid x_n)$.

A point that may be important in practice is

$$
\begin{aligned}
&\mathbb{E}_{Z\sim q_\alpha(z|x_n)}\left[\log\frac{p_\theta(x_n,Z)}{q_\alpha(Z\mid x_n)}\right] \\
&= \mathbb{E}_{Z\sim q_\alpha(z|x_n)}\left[\log\frac{p_\theta(x_n\mid Z)p_\theta(Z)}{q_\alpha(Z\mid x_n)}\right] \\
&= \mathbb{E}_{Z\sim q_\alpha(z|x_n)}\left[\log p_\theta(x_n\mid Z)\right] \\
&\qquad\qquad - \mathbb{D}_{\mathsf{KL}}(q_\alpha(z\mid x_n)\,\|\,p_\theta(z)).
\end{aligned}
$$

This form is useful because for certain p_θ and q_α, for instance if they are Gaussian, the KL term can be computed exactly instead of through sampling, which removes one source of noise in the optimization process.