4. The error in the ELBO bound

D.J. Wischik

In the Variational Autoencoder, we wish to approximate the true distribution q using a latent-variable generative model $Z \to X$, generated by $Z \sim n$ and $(X|Z=z) \sim p(\cdot|z)$; we'll write $X \sim p$ for the model's marginal distribution. For fitting we use an encoder $(Z|X=x) \sim r_x$.

The ELBO and its error. To bound our model's log likelihood evaluated at datapoint x, we can use importance sampling[1] and Jensen's inequality:

$$\begin{aligned}
\log p(x) &= \log \mathbb{E}_{Z\sim n}\, p(x|Z) \quad \text{law of total probability} \\
&= \log \mathbb{E}_{Z\sim r_x}\Big[p(x|Z)\frac{n(Z)}{r_x(Z)}\Big] \quad \text{importance sampling} \\
&\geq \mathbb{E}_{Z\sim r_x} \log\Big[p(x|Z)\,\frac{n(Z)}{r_x(Z)}\Big] \quad \text{Jensen} \\
&= \underbrace{\mathbb{E}_{Z\sim r_x} \log p(x|Z) - \mathrm{KL}(r_x \,\|\, n)}_{\text{called } \mathrm{ELBO}_x} \quad \text{KL defn.}
\end{aligned}$$

The error in this bound is

$$\log p(x) - \mathrm{ELBO}_x = \log \mathbb{E}_{Z\sim r_x} \frac{n(Z)\,p(x|Z)}{r_x(Z)} - \mathbb{E}_{Z\sim r_x} \log \frac{n(Z)\,p(x|Z)}{r_x(Z)}$$

which may be more clearly written in terms of the perfect sampling distribution, namely the posterior $s_x(z) = n(z)\,p(x|z)/p(x)$, as

$$\begin{aligned}
&= \log \mathbb{E}_{Z\sim r_x} \frac{p(x)\,s_x(Z)}{r_x(Z)} - \mathbb{E}_{Z\sim r_x} \log \frac{p(x)\,s_x(Z)}{r_x(Z)} \\
&= \log p(x) + \log \mathbb{E}_{Z\sim r_x} \frac{s_x(Z)}{r_x(Z)} - \log p(x) + \mathrm{KL}(r_x \,\|\, s_x) \quad \text{by algebra} \\
&= \mathrm{KL}(r_x \,\|\, s_x) \quad \text{by integrating out the density.}
\end{aligned}$$

Putting all this together,

$$\log p(x) = \mathrm{ELBO}_x + \mathrm{KL}(r_x \,\|\, s_x)$$

and averaging over x gives

$$\mathbb{E}_{X\sim q} \log p(X) = \underbrace{\mathbb{E}_{X\sim q}\, \mathrm{ELBO}_X}_{\text{called ELBO}} + \mathbb{E}_{X\sim q}\, \mathrm{KL}(r_X \,\|\, s_X).$$

The bound via KL. Here's a more direct derivation, based on the joint distribution of (Z, X). We have two generative mechanisms for this pair, the first $Z \to X$ which we'll denote (np), and the second $X \to Z$ which we'll denote (qr). We'll use explicit subscript notation to denote marginals from a generative mechanism, and simplify where we can, so that $(qr)_X = q_X$ etc. Using the chain rule of probability on the pair (X, Z),

$$\mathrm{KL}\big((qr)_{X,Z} \| (np)_{X,Z}\big) = \mathrm{KL}(q_X \,\|\, p_X) + \mathbb{E}_{X\sim q_X}\, \mathrm{KL}\big(r_{Z|X} \| (np)_{Z|X}\big) \geq \mathrm{KL}(q_X \,\|\, p_X).$$

The inequality uses $\mathrm{KL} \geq 0$, the grown-up version of our earlier use of Jensen's inequality. Rewriting and expanding the joint KL,

$$\begin{aligned}
\mathrm{KL}(q_X \,\|\, p_X) &\leq \mathrm{KL}\big((qr)_{X,Z} \| (np)_{X,Z}\big) \\
&= \mathbb{E}_{X\sim q_X, Z\sim r_{Z|X}} \log \frac{q_X(X) r_{Z|X}(Z)}{n_Z(Z) p_{X|Z}(X)} \\
&= \underbrace{\mathbb{E}_{X\sim q_X} \log q_X(X)}_{\text{const}} - \mathbb{E}_{X\sim q_X}\, \mathrm{ELBO}_X \ .
\end{aligned}$$

The error in this bound is $\mathbb{E}_{X\sim q_X}\, \mathrm{KL}\big(r_{Z|X} \| (np)_{Z|X}\big)$, or in the earlier notation $\mathbb{E}_{X\sim q}\, \mathrm{KL}(r_X \,\|\, s_X)$.

[1] The importance sampling step requires $r_x(z) > 0$ whenever $n(z)\,p(x|z) > 0$.