Inference **6 December 2017**

Evidence Lower Bound

Seth Ebner

1.1 Motivation

In inference, we are interested in the posterior distribution $p(y \mid x)$, where y is an unobserved (latent) variable that is related to an observation x (e.g., y generates x, y is a translation of x). According to the rule of conditional probability, we have:

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

The marginal distribution $p(x)$ (called the **evidence**), however, is calculated by summing (or integrating, for continuous distributions) $p(x, y)$ over all values of y, which is intractable in many cases. We'll try to approximate $p(y \mid x)$ with a simpler proposal distribution, $q(y)$.

1.2 Derivation

We want our proposal distribution, $q(y)$, to closely model $p(y \mid x)$, which means minimizing their KL divergence.

$$\begin{aligned}
KL[q(y) \mid\mid p(y \mid x)] &= -\int_y q(y) \log p(y \mid x) + \int_y q(y) \log q(y) \\
&= -\int_y q(y) \log \frac{p(x, y)}{p(x)} + \int_y q(y) \log q(y) \\
&= -\int_y q(y)(\log p(x, y) - \log p(x)) + \int_y q(y) \log q(y) \\
&= -\int_y q(y) \log p(x, y) + \int_y q(y) \log p(x) + \int_y q(y) \log q(y) \\
&= -(\int_y q(y) \log p(x, y) - \int_y q(y) \log q(y)) + \int_y q(y) \log p(x) \\
&= -(\int_y q(y) \log p(x, y) - \int_y q(y) \log q(y)) + \log p(x) \int_y q(y) \\
&= -(E_q[\log p(x, y)] - E_q[\log q(y)]) + \log p(x) \cdot 1 \\
&= -L + \log p(x)
\end{aligned}$$

where

$$\boxed{L = E_q[\log p(x, y)] - E_q[\log q(y)]}$$

1.3 Discussion

We see that $KL[q(y) \mid\mid p(y \mid x)]$ is equal to some term $-L$ dependent on the proposal distribution plus an additive constant, namely $\log p(x)$. Because calculating $p(x)$ is intractable, calculating the KL divergence is intractable. However, note that $\log p(x)$ is independent of the proposal distribution, so optimization with respect to q is not affected by it. **We see that minimizing $KL[q(y) \mid\mid p(y \mid x)]$ is equivalent to minimizing $-L$, which is the same as maximizing L.**

Rearranging:

$$\begin{aligned} L &= \log p(x) - KL[q(y) \mid\mid p(y \mid x)] \\ &\leq \log p(x) \end{aligned}$$

where the inequality arises because $KL[q(\cdot) \mid\mid p(\cdot)] \geq 0$. Equality holds if and only if $q(\cdot)$ perfectly matches $p(\cdot)$ ($KL = 0$).

We see then that L is a lower bound on $p(x)$ (the evidence). We therefore say that L is the **evidence lower bound** (ELBO). L can be calculated without knowing the value of the intractable normalizing constant $p(x)$ and is equal to it when L is maximized (when $q(\cdot)$ matches $p(\cdot)$). That is, we get a tight bound on $p(x)$ by minimizing the KL divergence between $q(y)$ and $p(y \mid x)$, or equivalently, by maximizing L. The family of proposal distributions q is chosen so that L is easily computable.

References

X. Yang, "Understanding the Variational Lower Bound"

B. Moran, "Variational Bayes and the evidence lower bound"

D. Blei "Variational Inference"