Overview

The goal of VI is to estimate the posterior by computable approximation. To this end, there are two approaches:

- MCMC sampling: slow, hard to scale with big data
- Variational inference: Construct a family of distributions over z and find the member of the family that is closest to the posterior. This technique involves optimisation over functions; and can be inexact (high bias, less variance). However, it brings benefits such as:
    - scalable, fast, suitable for big high dimensional data
    - generic and can be used for estimating any unknown function

Inference View

The objective of VI can be formally set as follows,

$$q^*(z) = \underset{q(z)\in\mathcal{Q}}{\arg\min}\, \mathrm{KL}(q(z)||p(z|x)) \tag{1}$$

Note that we are looking a distribution that minimises the KL divergence between it and the true posterior distribution. We can rewrite

$$\begin{align} \mathrm{KL}[q(z||p(z|x)] &= \mathbb{E}_q[\log \frac{q(z)}{p(z|x)}] \tag{2} \\ &= \mathbb{E}_q[\log q(z)] - \mathbb{E}_q[\log p(x,z)] + \log p(x) \tag{3} \end{align}$$

Again the objective function involves the intractable $p(x)$. However, as we formulate the posterior estimation problem as an optimization problem w.r.t q, we can ignore $\log p(x)$ as this term does not depend on q and does not affect the finding of the optimal q^*. In short, minimizing $KL[q(z||p(z|x)]$ is equivalent to maximizing

$$\begin{align} \mathrm{ELBO}(q) &= -\mathbb{E}_q[\log q(z)] + \mathbb{E}_q[\log p(x,z)] \tag{4} \\ &= \mathbb{E}_q[\log p(x|z)] - \mathrm{KL}[q(z)||p(z)] \tag{5} \end{align}$$

Here, all the terms $p(x|z), q(z), p(z)$ are generally computable, thus we have a good change to efficiently solve this optimization problem. The objective function $\mathrm{ELBO}(q)$ is the evidence lower bound because $\mathrm{ELBO}(q) = \log p(x) - KL[q(z||p(z|x)] \leq \log p(x)$ (remember KL is always non-negative).

Notes on the ELBO.

- q is referred to as variational distribution
- Maximising $\mathrm{ELBO}(q)$ also implies maximising the log-likelihood $\log p_\theta(x)$
- Other name: variational lower bound, negative free energy
- Reversing the KL in the original objective (Eq. 1) leads to a different and more challenging optimisation
- We need to solve the optimization in Eq. 1 for each x, a.k.a finding $q^*(z)$ for each x

Fitting View

We can derive the ELBO in a different way by starting from the goal of the fitting problem, which is maximize the log-likelihood. Remember in Lecture 1, directly maximizing the log-likelihood of LVM leads to bias. Hence, it is reasonable to find an alternative objective which is a lower bound of the log-likelihood as follows,

$$\begin{aligned}
\log p(x) &= \log \int_z p(x,z) && (6)\\
&= \log \int_z p(x,z)\frac{q(z)}{q(z)} && (7)\\
&= \log \mathbb{E}_q\left[\frac{p(x,z)}{q(z)}\right] && (8)\\
&\geq \mathbb{E}_q\left[\log\left(\frac{p(x,z)}{q(z)}\right)\right] && (9)\\
&= \mathbb{E}_q[\log p(x,z)] - \mathbb{E}_q[\log q(z)] && (10)
\end{aligned}$$

The RHS is actually the $\mathrm{ELBO}(q)$. Here the inequality is provided by Jensen inequality $\log \mathbb{E}[X] \geq \mathbb{E}[\log X]$ (see geometric illustration in Fig 1).

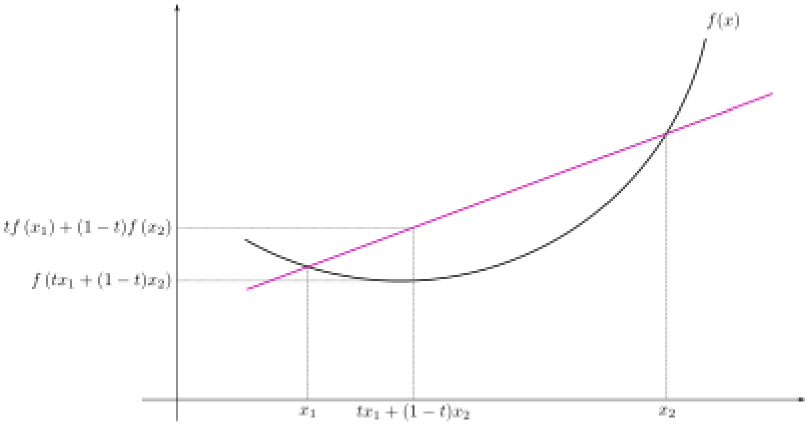

Figure 1: Jensen inequality. Image source: Wikipedia.

Viewing the $\mathrm{ELBO}(q)$ this way, we can see it consists of 2 terms: the negative cross-entropy of the join distribution relative to the variational distribution and the entropy of the variational distribution. Maximising $\mathrm{ELBO}(q)$ is equivalent to minimising the cross-entropy of the join (also maximising the log-likelihood of the join) while maximising the entropy of the variational distribution. Now maximizing the log-likelihood is replaced by

$$\begin{aligned}
\theta^* &= \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{N} \mathbb{E}_q[\log p_\theta(x_i, z)] - \mathbb{E}_q[\log q(z)] && (11)\\
&= \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{N} \mathbb{E}_q[\log p_\theta(x_i, z)] && (12)\\
&= \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{N} \mathbb{E}_q[\log p_\theta(x_i|z)] - \mathrm{KL}[q(z)||p(z)] && (13)\\
&= \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{N} \mathbb{E}_q[\log p_\theta(x_i|z)] && (14)
\end{aligned}$$

Now the objective is no more biased when we use MCMC sampling to estimate the expectation, which means we can safely optimize $\log p_\theta(x_i|z)$ using z sampled from q.

Join optimisation. Two views of the ELBO indicate two optimisation processes on the same objective. If we optimise the ELBO w.r.t q, we are solving the inference problem and when we optimise the ELBO w.r.t θ, we are working on the fitting problem. In practice, we can do both at the same time. Since the process is similar to the classic EM algorithm, people also refer to this joint optimisation as variational EM.

Exercises

1. Prove KL is non-negative using Jensen inequality

2. Prove Eq. (10) is actually the ELBO

3. Why is minimising the cross-entropy equivalent to maximising the log-likelihood