Variational Inference

Xingdong Zuo*

Let $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ be a set of training inputs and let $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$ be a set of training outputs. The objective is to find the parameters $\boldsymbol{\omega}$ of a parameterized function $\mathbf{y} = f_{\boldsymbol{\omega}}(\mathbf{x})$. In Bayesian modelling, we place a prior distribution $p(\boldsymbol{\omega})$ over the parameters, i.e. initial belief of how likely the parameters have generated the observations. We define a likelihood distribution $p(\mathbf{y}|\mathbf{x}, \boldsymbol{\omega})$ to indicate how likely $\mathbf{y}$ is generated by the input $\mathbf{x}$ given the parameters $\boldsymbol{\omega}$. The posterior distribution over the parameters is defined by

$$p(\boldsymbol{\omega}|\mathbf{X}, \mathbf{Y}) = \frac{p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\omega})p(\boldsymbol{\omega})}{p(\mathbf{Y}|\mathbf{X})}$$

Then the predictive distribution of a new test input $\mathbf{x}^*$ can be obtained by integrating model parameters

$$p(\mathbf{y}^*|\mathbf{x}^*, \mathbf{X}, \mathbf{Y}) = \int p(\mathbf{y}^*|\mathbf{x}^*, \boldsymbol{\omega})p(\boldsymbol{\omega}|\mathbf{X}, \mathbf{Y})\, \mathrm{d}\boldsymbol{\omega}$$

However, the denominator of the posterior $p(\mathbf{Y}|\mathbf{X}) = \int p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\omega})p(\boldsymbol{\omega})\, \mathrm{d}\boldsymbol{\omega}$ is intractable to obtain analytically. Because true posterior is intractable, in VI, we use an approximating variational distribution $q_\phi(\boldsymbol{\omega}|\mathbf{X}, \mathbf{Y}) \approx p_\theta(\boldsymbol{\omega}|\mathbf{X}, \mathbf{Y})$. The optimum is obtained by minimizing the KL divergence between two distributions, leading to a fundamental theorem in VI. Intuitively, minimizing KL divergence is equivalent to maximizing the evidence lower bound (ELBO) w.r.t. the variational parameters ϕ.

Theorem 0.1.

$$\begin{aligned} &\min_\phi D_{\mathrm{KL}}\left(q_\phi(\boldsymbol{\omega}|\mathbf{D}) \| p_\theta(\boldsymbol{\omega}|\mathbf{D})\right) \\ \iff &\max_\phi \left\{\mathbb{E}_{q_{\phi(\boldsymbol{\omega}|\mathbf{D})}}\left[\log p_\theta(\mathbf{D}|\boldsymbol{\omega})\right] - D_{\mathrm{KL}}\left(q_\phi(\boldsymbol{\omega}|\mathbf{D}) \| p_\theta(\boldsymbol{\omega})\right)\right\} \end{aligned}$$

First proof

Proof. By the definition of KL divergence, we can obtain

$$\begin{aligned} & D_{\mathrm{KL}}\left(q_\phi(\boldsymbol{\omega}|\mathbf{D}) \| p_\theta(\boldsymbol{\omega}|\mathbf{D})\right) \\ = \quad & \mathbb{E}_{q_\phi(\boldsymbol{\omega}|\mathbf{D})}\left[\log q_\phi(\boldsymbol{\omega}|\mathbf{D})\right] + \mathbb{E}_{q_\phi(\boldsymbol{\omega}|\mathbf{D})}\left[\log p_\theta(\mathbf{D})\right] - \mathbb{E}_{q_\phi(\boldsymbol{\omega}|\mathbf{D})}\left[\log p_\theta(\mathbf{D}, \boldsymbol{\omega})\right]. \end{aligned}$$

Note that $\log p_\theta(\mathbf{D}) = \mathbb{E}_{q_\phi(\boldsymbol{\omega}|\mathbf{D})}\left[\log p_\theta(\mathbf{D})\right]$ and by rearranging the equality, we have

$$\begin{aligned} & \log p_\theta(\mathbf{D}) \\ = \quad & \mathbb{E}_{q_\phi(\boldsymbol{\omega}|\mathbf{D})}\left[\log p_\theta(\mathbf{D}, \boldsymbol{\omega})\right] - \mathbb{E}_{q_\phi(\boldsymbol{\omega}|\mathbf{D})}\left[\log q_\phi(\boldsymbol{\omega}|\mathbf{D})\right] + D_{\mathrm{KL}}\left(q_\phi(\boldsymbol{\omega}|\mathbf{D}) \| p_\theta(\boldsymbol{\omega}|\mathbf{D})\right). \end{aligned}$$

Applying the non-negativity property of the KL divergence i.e. $D_{\mathrm{KL}}\left(q_\phi(\boldsymbol{\omega}|\mathbf{D}) \| p_\theta(\boldsymbol{\omega}|\mathbf{D})\right) \geq 0$ we have the following

$$\begin{aligned} \log p_\theta(\mathbf{D}) &\geq \mathbb{E}_{q_\phi(\boldsymbol{\omega}|\mathbf{D})}\left[\log p_\theta(\mathbf{D}, \boldsymbol{\omega})\right] - \mathbb{E}_{q_\phi(\boldsymbol{\omega}|\mathbf{D})}\left[\log q_\phi(\boldsymbol{\omega}|\mathbf{D})\right] \\ &= \mathbb{E}_{q_\phi(\boldsymbol{\omega}|\mathbf{D})}\left[\log p_\theta(\mathbf{D}|\boldsymbol{\omega})\right] - D_{\mathrm{KL}}\left(q_\phi(\boldsymbol{\omega}|\mathbf{D}) \| p_\theta(\boldsymbol{\omega})\right) \\ &= \mathcal{L} \end{aligned}$$

□

ELBO consists of reconstruction term, which tends to find variational parameters that likely generates the observation, and regularization term, which penalizes the model if it deviates too much from the prior distribution. ELBO is a trade-off between them.

*December 18, 2018

Second proof

We can also prove the theorem by using Jensen's inequality. Let X be a random variable and for a convex function f, the following inequality holds

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]. \tag{1}$$

The inequality is reversed if f is a concave function

$$f(\mathbb{E}[X]) \geq \mathbb{E}[f(X)]. \tag{2}$$

Proof. By applying importance sampling of approximating variational distribution $q_\phi(\boldsymbol{\omega}|\mathbf{D})$ to the marginal likelihood $\log p_\theta(\mathbf{D})$, we have

$$\begin{aligned}
\log p_\theta(\mathbf{D}) &= \log \int p_\theta(\mathbf{D}|\boldsymbol{\omega}) p_\theta(\boldsymbol{\omega}) \frac{q_\phi(\boldsymbol{\omega}|\mathbf{D})}{q_\phi(\boldsymbol{\omega}|\mathbf{D})} \, \mathrm{d}\boldsymbol{\omega} \\
&= \log \mathbb{E}_{q_\phi(\boldsymbol{\omega}|\mathbf{D})} \left[\frac{p_\theta(\mathbf{D}|\boldsymbol{\omega}) p_\theta(\boldsymbol{\omega})}{q_\phi(\boldsymbol{\omega}|\mathbf{D})} \right] \\
&\geq \mathbb{E}_{q_\phi(\boldsymbol{\omega}|\mathbf{D})} \left[\log \frac{p_\theta(\mathbf{D}|\boldsymbol{\omega}) p_\theta(\boldsymbol{\omega})}{q_\phi(\boldsymbol{\omega}|\mathbf{D})} \right] \\
&= \mathbb{E}_{q_\phi(\boldsymbol{\omega}|\mathbf{D})} [\log p_\theta(\mathbf{D}|\boldsymbol{\omega})] - D_{\mathrm{KL}} (q_\phi(\boldsymbol{\omega}|\mathbf{D}) \| p_\theta(\boldsymbol{\omega})) \\
&= \mathcal{L}.
\end{aligned}$$

□

Third proof

We can also have a direct proof by expanding the definition of KL divergence.

Proof.

$$\begin{aligned}
& \min_\phi D_{\mathrm{KL}} (q_\phi(\boldsymbol{\omega}|\mathbf{D}) \| p_\theta(\boldsymbol{\omega}|\mathbf{D})) \\
\iff & \min_\phi \int q_\phi(\boldsymbol{\omega}|\mathbf{D}) \log \frac{q_\phi(\boldsymbol{\omega}|\mathbf{D})}{p_\theta(\mathbf{D}|\boldsymbol{\omega}) p_\theta(\boldsymbol{\omega})} \, \mathrm{d}\boldsymbol{\omega} \\
\iff & \min_\phi \left\{ -\mathbb{E}_{q_\phi(\boldsymbol{\omega}|\mathbf{D})} [\log p_\theta(\mathbf{D}|\boldsymbol{\omega})] + D_{\mathrm{KL}} (q_\phi(\boldsymbol{\omega}|\mathbf{D}) \| p_\theta(\boldsymbol{\omega})) \right\} \\
\iff & \max_\phi \left\{ \mathbb{E}_{q_\phi(\boldsymbol{\omega}|\mathbf{D})} [\log p_\theta(\mathbf{D}|\boldsymbol{\omega})] - D_{\mathrm{KL}} (q_\phi(\boldsymbol{\omega}|\mathbf{D}) \| p_\theta(\boldsymbol{\omega})) \right\}.
\end{aligned}$$

□