Variational Inference

Variational Inference (VI) is a method for approximating a conditional posterior distribution over latent/hidden variables in a Bayesian setting. This is a useful tool, as the resulting posterior distributions can often become computationally complex or entirely intractable.

General Setup

We assume that $x_{1:n} = \{x_1, x_2, \ldots, x_n\}$ are observations, with hidden variables $z_{1:m} = \{z_1, \ldots, z_m\}$ and additional fixed (*hyper*-)parameters α.

We are interested in inference on the hidden variables $z_{1:m}$, which invokes a posterior conditional distribution of the form

$$p(z_{1:m} \mid x_{1:n}, \alpha) = \frac{p(z_{1:m}, x_{1:n} \mid \alpha)}{p(x_{1:n} \mid \alpha)} = \frac{p(z_{1:m}, x_{1:n} \mid \alpha)}{\int_z p(z_{1:m}, x_{1:n} \mid \alpha) dz} \tag{1.1}$$

The denominator for this posterior distribution is often difficult to compute, if not fully intractable, so we must approximate the distribution $\mathbb{P}(z_{1:m} \mid x_{1:n}, \alpha)$. One approach is to consider a **variational family** of distributions $\mathcal{Q} = \{q_\lambda(z_{1:m}) : \lambda \in \Lambda\}$ over the latent variables $z_{1:m}$, and finding the distribution in the family which is the most suitable (i.e. closest) proxy for the 'true' posterior distribution $p(z_{1:m} \mid x_{1:n}, \alpha)$.

Kullback-Leibler Divergence

To measure the 'closeness' of two probability distributions P and Q defined on the same space, we can use the **Kullback-Leibler** (KL) divergence. This divergence is defined as

$$D_{\mathrm{KL}}(P \| Q) := \int P(x) \log\left(\frac{P(x)}{Q(x)}\right) \mathrm{d}P = \mathbb{E}_P\left[\log\left(\frac{P(x)}{Q(x)}\right)\right] \tag{2.1}$$

Note that this is not a distance metric, as $D_{\mathrm{KL}}(P \| Q) \neq D_{\mathrm{KL}}(Q \| P)$. To get a distribution in our variational family which is close to the true posterior, we aim to have a low KL divergence.

Evidence Lower Bound

We define the **Evidence Lower Bound** (ELBO) as a function of our distribution, which we can use to choose the specific variational distribution $q_\lambda(z_{1:m})$, by finding $\lambda \in \Lambda$ to maximize the ELBO.

For probability distributions P, Q, we have the following:

$$\begin{aligned}
\log(P(x)) &= \log\left(\int P(x,z)\mathrm{d}z\right) && \text{(Marginal distribution)} \\
&= \log\left(\int P(x,z)\frac{Q(z)}{Q(z)}\mathrm{d}z\right) \\
&= \log\left(\int Q(z)\left[\frac{P(x,z)}{Q(z)}\right]\mathrm{d}z\right) \\
&= \log\left(\mathbb{E}_Q\left[\frac{P(x,Z)}{Q(Z)}\right]\right) \\
&\geq \mathbb{E}_Q\left[\log\left(\frac{P(x,Z)}{Q(Z)}\right)\right] && \text{(Jensen's Inequality)}
\end{aligned}$$

We define the ELBO as $\mathbb{E}_Q\left[\log\left(\frac{P(x,Z)}{Q(Z)}\right)\right] = \mathbb{E}_Q\left[\log\left(P(x,Z)\right)\right] - \mathbb{E}_Q\left[\log\left(Q(Z)\right)\right]$. Note that $-D_{\mathrm{KL}}(Q\|P) = \mathbb{E}_Q\left[\log\left(\frac{P(x,Z)}{Q(Z)}\right)\right]$, so the ELBO is the negative KL divergence. Finding a distribution $Q(z) \in \mathcal{Q}$ which maximizes the ELBO yields the tightest possible bound on the marginal probability $\log(P(x))$.

Additionally, for some marginal distribution $p(z \mid x)$ and some "variational" distribution $q(z) \in \mathcal{Q}$ we have the following result:

$$
\begin{aligned}
D_{\mathrm{KL}}(q(z)\|p(z \mid x)) &= \mathbb{E}_Q\left[\log\left(\frac{q(Z)}{p(Z \mid x)}\right)\right] \\
&= \mathbb{E}_Q\left[\log\left(\frac{q(Z)}{p(x,Z)/p(x)}\right)\right] \\
&= \mathbb{E}_Q\left[\log\left(q(Z)\right)\right] - \mathbb{E}_Q\left[\log\left(p(x,Z)\right)\right] + \mathbb{E}_Q\left[\log\left(p(x)\right)\right] \\
&= \log(p(x)) - \mathbb{E}_Q\left[\log\left(\frac{p(x,Z)}{q(Z)}\right)\right] && (\log(p(x)) - \text{ELBO}) \\
&= \log(p(x)) + D_{\mathrm{KL}}(q(z)\|p(x,z)) && (\text{Alternative formulation})
\end{aligned}
$$

Thus, the KL divergence between the "variational" distribution $q(z) \in \mathcal{Q}$ and the marginal distribution $p(z \mid x)$ is the difference between the log-marginal distribution and the ELBO, which is the Jensen gap.

As $\log(p(x))$ is constant, we see that maximizing the ELBO is equivalent to minimizing the KL divergence between the conditional posterior and variational distribution.

EULBO

Motivation

For Bayesian Optimization, a variational inference approach can be helpful as a means for approximation since exact Bayesian Optimization via a Gaussian Process requires $\mathcal{O}(n^3)$ runtime.

One potential issue with the use of VI in this setting is that the 'traditional' variational inference setup requires choosing a distribution $q_\lambda(z) \in \mathcal{Q}$ which maximizes the ELBO. However, this is not ideal for BayesOpt, as the goal for BayesOpt is to simply find the global maximum of some unknown function f^*, not to get a good global approximation of f^*.

For a Gaussian Process with an observed dataset $\mathcal{D}$, we can derive the posterior $p(f \mid \mathcal{D})$ for a function f. With a utility function $u(x, f; \mathcal{D}_t)$ (e.g. expected improvement), we can define the **expected utility** as

$$\alpha(x; \mathcal{D}_t) := \int u(x, f; \mathcal{D}_t) p(f \mid \mathcal{D}_t) \mathrm{d}f \tag{4.1}$$

Through variational inference, we may approximate the posterior $p(f \mid \mathcal{D}_t)$ with $q_{\mathbf{S}}(f)$, where $\mathbf{S} \in \mathbb{R}^{n \times k}$ is an n-by-k action matrix, yielding

$$\alpha(x; \mathcal{D}_t) \approx \int u(x, f; \mathcal{D}_t) q_{\mathbf{S}}(f) \mathrm{d}f \tag{4.2}$$

EULBO Derivation

Based on the definitions above, we have the following:

$$
\begin{aligned}
\log\left(\alpha(x;\mathcal{D}_t)\right) &= \log\left(\int u(x,f;\mathcal{D}_t)p(f\mid\mathcal{D}_t)\mathrm{d}f\right) \\
&= \log\left(\int u(x,f;\mathcal{D}_t)p(f\mid\mathcal{D}_t)\left(\frac{q_{\mathbf{S}}(f)}{q_{\mathbf{S}}(f)}\right)\mathrm{d}f\right) \\
&= \log\left(\int u(x,f;\mathcal{D}_t)\left(\frac{p(f,\mathcal{D}_t)}{p(\mathcal{D}_t)}\right)\left(\frac{q_{\mathbf{S}}(f)}{q_{\mathbf{S}}(f)}\right)\mathrm{d}f\right) \\
&= \log\left(\int q_{\mathbf{S}}(f)\left(\frac{u(x,f;\mathcal{D}_t)p(f,\mathcal{D}_t)}{p(\mathcal{D}_t)q_{\mathbf{S}}(f)}\right)\mathrm{d}f\right) \\
&= \log\left(\mathbb{E}_{q_{\mathbf{S}}}\left[\frac{u(x,f;\mathcal{D}_t)p(f,\mathcal{D}_t)}{p(\mathcal{D}_t)q_{\mathbf{S}}(f)}\right]\right) \\
&\geq \mathbb{E}_{q_{\mathbf{S}}}\left[\log\left(\frac{u(x,f;\mathcal{D}_t)p(f,\mathcal{D}_t)}{p(\mathcal{D}_t)q_{\mathbf{S}}(f)}\right)\right] && \text{(Jensen's Inequality)} \\
&= \mathbb{E}_{q_{\mathbf{S}}}\left[\log\left(\frac{p(f,\mathcal{D}_t)}{q_{\mathbf{S}}(f)}\right)\right] + \mathbb{E}_{q_{\mathbf{S}}}\left[\log\left(u(x,f;\mathcal{D}_t)\right)\right] - \log(Z) && (Z = p(\mathcal{D}_t)\text{ is constant})
\end{aligned}
$$

Thus, we can express the EULBO of $q_{\mathbf{S}}$ in terms of the ELBO $\mathbb{E}_{q_{\mathbf{S}}}\left[\log\left(\frac{p(f,\mathcal{D}_t)}{q_{\mathbf{S}}(f)}\right)\right]$ and expected log-utility, $\mathbb{E}_{q_{\mathbf{S}}}\left[\log\left(u(x,f;\mathcal{D}_t)\right)\right]$. As the EULBO is only used for optimization (i.e. selection of $\mathbf{S}$), we do not care about the $\log(Z)$ normalization constant.

Note that the expected log-utility is a function of x defined on the domain of f. This optimization scheme involves a joint optimization to find $(x_{n+1}, \mathbf{S}_{n+1})$, as opposed to individual optimizations.