Variational inference: Reparametrization tricks

Setup: Bayesian inference for some model with parameters $\theta \in \mathbb{R}^d$ given some data $\mathcal{D}$:

$$\underbrace{p(\theta|\mathcal{D})}_{\text{intractable}} = \frac{p(\mathcal{D}|\theta)\,p(\theta)}{\underbrace{p(\mathcal{D})}}$$

Idea: Approximate $p(\theta|\mathcal{D}) \approx q_\varphi(\theta|\mathcal{D})$

e.g. mean-field: $q_\theta(\theta|\mathcal{D}) = \prod_{i=1}^{d} \mathcal{N}(\theta_i|\mu_i, \sigma_i^2)$

$$\varphi := \{\mu_1, \sigma_1^2, \ldots, \mu_d, \sigma_d^2\}$$

Training: $\varphi^* = \arg\min_\varphi \mathbb{KL}[q_\varphi(\theta|\mathcal{D}) \| p(\theta|\mathcal{D})] := \mathcal{L}(\varphi)$

$$\mathcal{L}(\varphi) := -\underbrace{H[q_\varphi(\theta|\mathcal{D})]}_{\text{computed analytically}} - \underbrace{\mathbb{E}_{\theta\sim q_\varphi(\theta|\mathcal{D})}[\log p(\mathcal{D}|\theta) + \log p(\theta)]}_{\text{computed via sampling}}$$

Optimize via SGD: $\varphi^{n+1} = \varphi^n - \eta \underbrace{\nabla_\varphi \mathcal{L}(\varphi)}$

Compute $\nabla_\varphi \mathcal{L}(\varphi)$:

$$\nabla_\varphi \mathbb{E}_{\theta\sim q_\varphi(\theta|\mathcal{D})}[\log p(\mathcal{D}|\theta) + \log p(\theta)]$$

$$= \nabla_\varphi \int \log p(\mathcal{D}|\theta)\, q_\varphi(\theta|\mathcal{D})\, d\theta + \nabla_\varphi \int \log p(\theta)\, q_\varphi(\theta|\mathcal{D})\, d\theta$$

$\nabla_\varphi q_\varphi(\theta|\mathcal{D})\, d\theta$

$$= \int \log p(\mathcal{D}|\vartheta) \underset{\uparrow}{\boxed{\nabla_\varphi q_\varphi(\vartheta|\mathcal{D})}} d\vartheta + \int \log p(\vartheta) \nabla_\varphi q_\varphi(\vartheta|\mathcal{D}) d\vartheta \quad \textcircled{1}$$

<u>Recall</u>: $\left(\log f(x)\right)' = \dfrac{f'(x)}{f(x)}$

- $$\nabla_\varphi \log q_\varphi(\vartheta|\mathcal{D}) = \frac{\nabla_\varphi q_\varphi(\vartheta|\mathcal{D})}{q_\varphi(\vartheta|\mathcal{D})}$$

$$\Longrightarrow \nabla_\varphi q_\varphi(\vartheta|\mathcal{D}) = \nabla_\varphi \log q_\varphi(\vartheta|\mathcal{D}) \cdot \underline{\underline{q_\varphi(\vartheta|\mathcal{D})}} \quad \textcircled{2}$$

$$\textcircled{1} \overset{\textcircled{2}}{\Longrightarrow} \underbrace{\nabla_\varphi \mathcal{L}(\varphi)} = \int \log p(\mathcal{D}|\vartheta) \nabla_\varphi \log q_\varphi(\vartheta|\mathcal{D}) q_\varphi(\vartheta|\mathcal{D}) d\vartheta$$

$$+ \int \log p(\vartheta) \nabla_\varphi \log q_\varphi(\vartheta|\mathcal{D}) q_\varphi(\vartheta|\mathcal{D}) d\vartheta$$

$$= \mathbb{E}_{\vartheta \sim q_\varphi(\vartheta|\mathcal{D})} \left[\nabla_\varphi \log q_\varphi(\vartheta|\mathcal{D}) \left(\log p(\mathcal{D}|\vartheta) + \log p(\vartheta) \right) \right]$$

$$\approx \frac{1}{S} \sum_{i=1}^{S} \nabla_\varphi \log q_\varphi(\vartheta_i|\mathcal{D}) \left(\log p(\mathcal{D}|\vartheta_i) + \log p(\vartheta_i) \right),$$

$$\vartheta_i \sim q_\varphi(\vartheta|\mathcal{D}) \overset{\text{m.F.}}{=} \mathcal{N}(\vartheta | \mu_\varphi, \Sigma_\varphi)$$

$$\mu_\varphi := \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_d \end{bmatrix}, \quad \Sigma_\varphi = \begin{bmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_d^2 \end{bmatrix}$$

The above Monte Carlo estimator of $\nabla_\varphi \mathcal{L}(\varphi)$ depend on φ and in practice tends to exhibit very high variance (i.e. it is very innaccrurate unless a very large number of considered).

samples is considered , .

Reparametrization trick :

Idea : Introduce a simple "change of variables" such tha
we compute expectations with respect to distributions that do n
~~depend~~ on ϕ.

If we can find a function $h : (\varepsilon, \phi) \longrightarrow \theta$, where
$\varepsilon \sim p(\varepsilon)$, then we can write :

$$\theta_i = h_\phi(\varepsilon) , \quad \varepsilon \sim p(\varepsilon) , \text{ such that } \theta_i \sim q_\phi(\theta|D)$$

e.g. re-parametrize a Gaussian :

$$\theta_i \sim q_\phi(\theta|D) = \mathcal{N}(\theta_i | \mu_\phi, \Sigma_\phi)$$

In fact , we can generate samples θ by sampling $\varepsilon \sim p($

$$\theta = \mu_\phi + \varepsilon \Sigma_\phi^{\frac{1}{2}} , \quad \text{where} \quad \varepsilon \sim p(\varepsilon) = \mathcal{N}(0, I)$$

$$\theta = h_\phi(\varepsilon)$$

Now recall the troublesome gradient term :

$$\nabla_\phi \mathbb{E}_{\theta \sim q_\phi(\theta|D)} \left[\log p(D|\theta) + \log p(\theta) \right] =$$

$$= \mathbb{E}_{\theta \sim q_\phi(\theta|D)} \xcancel{\left[\nabla_\phi \log q_\phi(\theta|D) \left(\log p(D|\theta) + \log p(\theta) \right) \right]}$$

$$\hookrightarrow$$

$$= \nabla_\varphi \, \mathbb{E}_{\varepsilon \sim p(\varepsilon)} \left[\log p(D \mid h_\varphi(\varepsilon)) + \log p(h_\varphi(\varepsilon)) \right]$$

$$\overset{M.F.}{=} \nabla_\varphi \, \mathbb{E}_{\varepsilon \sim \mathcal{N}(0, I)} \left[\log p(D \mid \underbrace{\mu_\varphi + \varepsilon \Sigma_\varphi^{\frac{1}{2}}}) + \log p(\mu_\varphi + \varepsilon \Sigma_\varphi^{1/2}\right.$$

Now the gradient is not related to the variational parau

(i.e. the distribution with respect to which the expectation is taken does not depend on φ.)

Summary: $\varphi^* = \arg\min_\varphi \mathbb{KL}[q_\varphi(\theta|D) \,\|\, p(\theta|D)]$

$$\downarrow$$

$$L(\varphi) := -H[q_\varphi(\theta|D)] - \mathbb{E}_{\theta \sim q_\varphi(\theta|D)} [\log p(D|\theta) + \log p(\theta)]$$

Evidence lower bound (ELBO)

If $\boxed{q_\varphi(\theta|D) = \mathcal{N}(\theta_i | \mu_\varphi, \Sigma_\varphi)}$

1st term: $-H[q_\varphi(\theta|D)] = -\sum_{i=1}^{d} \log \sigma_i + \text{constant}$

$$-\nabla_\varphi H[q_\varphi(\theta|D)] = -\sum_{i=1}^{d} \frac{\partial}{\partial \sigma_i} \log \sigma_i = -\sum_{i=1}^{d} \frac{1}{\sigma_i}$$

2nd term:

$$\nabla_\varphi \, \mathbb{E}_{\theta \sim q_\varphi(\theta|D)} [\log p(D|\theta) + \log p(\theta)] =$$

$$\nabla \, \mathbb{E} \left[\log p(D \mid \mu_\varphi + \varepsilon \Sigma_\varphi^{\frac{1}{2}}) + \log p(\mu_\varphi + \varepsilon \Sigma \right.$$

$$= \nabla_\varphi \mathbb{E}_{\varepsilon\sim\mathcal{N}(0,I)} [$$

$$\approx \frac{1}{S} \sum_{i=1}^{S} \Big[\nabla_\varphi \log p \big(\mathcal{D} \,\big|\, \underbrace{\mu_\varphi + \varepsilon_i \Sigma_\varphi^{\frac{1}{2}}}_{\theta} \big) + \nabla_\varphi \log p (\mu_\varphi + \varepsilon_i \Sigma$$

where $\varepsilon_i \sim \mathcal{N}(0, I)$.