Variational inference : Re-parametrization trick

<u>Setup</u>: Bayesian inference on a ML model with parameters θ given some data $\mathcal{D}$: $\hookrightarrow \theta \in \mathbb{R}^d$

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)\, p(\theta)}{p(\mathcal{D})}$$

<u>Idea</u>: Approximate $p(\theta|\mathcal{D}) \approx q_\varphi(\theta|\mathcal{D})$ → variational family parametrized by φ

<u>Mean-field family</u>: $q_\varphi(\theta|\mathcal{D}) = \prod_{i=1}^{d} \mathcal{N}(\theta_i \,|\, \mu_i, \sigma_i^2)$

$$\varphi := \{\mu_1, \sigma_1^2, \mu_2, \sigma_2^2, \dots, \mu_d, \sigma_d^2\}$$

<u>Train φ</u>: $\varphi^* = \underset{\varphi}{\arg\min}\ \mathbb{KL}\left[q_\varphi(\theta|\mathcal{D}) \,\|\, p(\theta|\mathcal{D})\right]$

see last lecture $\Longrightarrow$ $\varphi^* = \underset{\varphi}{\arg\min}\ \mathcal{L}(\varphi) \xrightarrow{\text{G.D.}} \varphi^{n+1} = \varphi^n - \eta \nabla_\varphi \mathcal{L}(\varphi)$

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

Remarks on computing $\nabla_\varphi \mathcal{L}(\varphi)$.

Gradient of the second term wrt φ :

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

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

$$= \int \log p(D|\theta)\, \nabla_\phi q_\phi(\theta|D)\, d\theta + \int \log p(\theta)\, \nabla_\phi q_\phi(\theta|D)\, d\theta \quad (1)$$

Recall : $\quad (\log f(x))' = \dfrac{f'(x)}{f(x)}$

$$\left\{ \nabla_\phi \log q_\phi(\theta|D) = \frac{\nabla_\phi q_\phi(\theta|D)}{q_\phi(\theta|D)} \right. \Longrightarrow$$

$$\Rightarrow \nabla_\phi q_\phi(\theta|D) = \nabla_\phi \log q_\phi(\theta|D) \cdot q_\phi(\theta|D) \quad (2)$$

$$(1) \overset{(2)}{=} \int \log p(D|\theta)\, \nabla_\phi \log q_\phi(\theta|D) \cdot q_\phi(\theta|D)\, d\theta \; +$$

$$\int \log p(\theta)\, \nabla_\phi \log q_\phi(\theta|D)\, q_\phi(\theta|D)\, d\theta$$

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

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

$$\approx \frac{1}{n} \sum_{i=1}^{n} \nabla_\phi \log q_\phi(\theta_i|D) \left[\log p(D|\theta_i) + \log p(\theta_i) \right],$$

$$\text{where } \underset{\in \mathbb{R}^d}{\theta_i} \overset{i.i.d.}{\sim} q_\phi(\theta|D) \overset{M.F.}{=} \mathcal{N}(\theta \mid \mu_\phi, \Sigma_\phi)$$

$$\text{where} \quad \mu_\phi = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_d \end{bmatrix}, \quad \Sigma_\phi = \begin{bmatrix} \sigma_1^2 & & & 0 \\ & \sigma_2^2 & & \\ & & \ddots & \\ 0 & & & \sigma_d^2 \end{bmatrix}$$

This Monte Carlo estimator depends on ϕ and in practice exhibits very high variance (i.e. it is very innacurrate unless a very large number of samples n is considered).

Re-parametrization trick:

If we can find a function $\underline{h}: (\varepsilon, \phi) \to \theta$, where ε is a random variable $\varepsilon \sim p(\varepsilon)$, then we can write:

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

i.e. we will try to find a function such that samples from $q_\phi(\theta|D)$ can be written as $\underline{\theta = h_\phi(\varepsilon)}$, $\varepsilon \sim p(\varepsilon)$

__e.g.__ Re-parametrize a Gaussian:

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

In fact, we can generate samples θ using the following re-parametriz

$$\underset{\in \mathbb{R}^d}{\theta} = \underbrace{\mu_\phi + \varepsilon \Sigma_\phi^{\frac{1}{2}}}_{h_\phi(\varepsilon)} \;, \quad \text{where } \underline{\varepsilon} \sim \underline{p(\varepsilon)} = \mathcal{N}(0, I)$$

↪ this does not depend on ϕ!

Now recall the trouble-some gradient term:

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

$$\underbrace{\mathbb{E}_{\varepsilon \sim p(\varepsilon)} \left[\nabla_\phi \log p(D | h_\phi(\varepsilon)) + \nabla_\phi \log p(h_\phi(\varepsilon)) \right]}$$

Now the gradient is __not__ related to the variational parameter. hence it is __not__ related to the distribution with respect to which the expectation is taken.

Summary:

$$\varphi^* = \underset{\varphi}{\operatorname{argmin}} \ \mathbb{KL}[q_\varphi(\theta|\mathcal{D}) \| p(\theta|\mathcal{D})] := \mathcal{L}(\varphi)$$

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

To solve this problem we typically employ SGD:

$$\varphi^{n+1} = \varphi^n - \eta \nabla_\varphi \mathcal{L}(\varphi),$$

For a mean-field approximation, i.e. $q_\varphi(\theta|\mathcal{D}) = \mathcal{N}(\theta|\mu_\varphi, \Sigma_\varphi)$

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

1st term:

$$-H[q_\varphi(\theta|\mathcal{D})] := \mathbb{E}_{\theta\sim q_\varphi(\theta|\mathcal{D})}[\log q_\varphi(\theta|\mathcal{D})] = -\sum_{i=1}^d \log\sigma_i + \text{constant}$$

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

2nd term:

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

$$= \mathbb{E}_{\varepsilon\sim\mathcal{N}(0,I)}\left[\nabla_\varphi \log p(\mathcal{D}|\mu_\varphi + \varepsilon\Sigma_\varphi^{\frac{1}{2}}) + \nabla_\varphi \log p(\mu_\varphi + \varepsilon\Sigma_\varphi^{\frac{1}{2}})\right]$$

$$\approx \frac{1}{n}\sum_{i=1}^n \nabla_\varphi \log p(\mathcal{D}|\mu_\varphi + \varepsilon_i\Sigma_\varphi^{\frac{1}{2}}) + \nabla_\varphi \log p(\mu_\varphi + \varepsilon_i\Sigma_\varphi^{\frac{1}{2}})],$$

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

(*) Typically in practice $n = 1, 10, 20, 50$ samples.