Variational Inference

Setup: Given some data $\mathcal{D}$, and a model with parameters $\theta \in \mathbb{R}^d$ and a likelihood $p(\mathcal{D}|\theta)$, and a prior $p(\theta)$.

Goal: Approximate

$$\underbrace{p(\theta|\mathcal{D})}_{\text{intractable posterior}} = \frac{p(\mathcal{D}|\theta)\,p(\theta)}{p(\mathcal{D})}, \qquad p(\mathcal{D}) \longrightarrow \int p(\mathcal{D}|\theta)\,p(\theta)\,d\theta$$

Main idea: Approx. $p(\theta|\mathcal{D})$ with a family of distributions that is easy to work with.

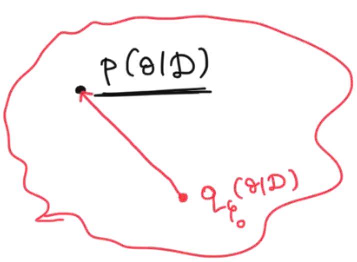
$\theta = (\theta_1, \theta_2, \dots, \theta_d)$

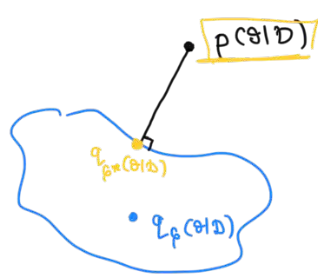
e.g. Mean-field family:

$$p(\theta|\mathcal{D}) \approx q_\varphi(\theta|\mathcal{D}) = \underbrace{\prod_{i=1}^{d} \mathcal{N}(\theta_i \mid \mu_i, \sigma_i^2)}$$

$$\underbrace{\varphi}_{\text{variational parameters}} := \{\mu_1, \sigma_1^2, \mu_2, \sigma_2^2, \dots, \mu_d, \sigma_d^2\}$$

Goal: Find / estimate the variational parameters φ such that $q_\varphi(\theta|\mathcal{D})$ is as close as possible to $p(\theta|\mathcal{D})$.





How to compute the "distance" between $q_\varphi(\theta|\mathcal{D})$ and $p(\theta|\mathcal{D})$?

In practice we use the Kullback-Leibler divergence as a way to compare $q_\varphi(\theta|\mathcal{D})$ to $p(\theta|\mathcal{D})$:

$$\underbrace{\mathbb{KL}[q_\phi(\theta|D) \,\|\, p(\theta|D)]}_{\text{relative entropy}} := \int \log \frac{q_\phi(\theta|D)}{p(\theta|D)}\, q_\phi(\theta|D)\, d\theta$$

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

Why use the KL-divergence?

... It is easy to work with (see below).

... but, it is <u>not</u> a distance, i.e. $\mathbb{KL}[q_\phi \| p] \neq \underbrace{\mathbb{KL}[p \| q_\phi]}_{\text{reverse KL}}$

however, $\mathbb{KL}[q_\phi \| p] = 0 \iff q_\phi = p$.

We want want to find $\boxed{\phi^* = \arg\min_\phi \mathbb{KL}[q_\phi(\theta|D) \| p(\theta|D)]}$

How to estimate the "optimal" variational parameters ϕ^*?

1. Old schoolers had to derive coordinate ascent rules for minimizing the $\mathbb{KL}$. (see chapter 10 in Bishop).

2. New schoolers are using Automatic Differentiation Variational Inferen

<u>ADVI</u>: It is a black-box approach is agnostic to any details about p(θ

Any model for which we can <u>evaluate</u> (and differentiate) the <u>log-likelihood</u> and the <u>log-prior</u> works!

Recall: $\mathbb{KL}[q_\phi(\theta|D) \| p(\theta|D)] = \mathbb{E}_{\theta \sim q_\phi(\theta|D)} [\log q_\phi(\theta|D) - \log p(\theta|D)]$

<u>1st term</u>: $\mathbb{E}_{\theta \sim q_\phi(\theta|D)} [\log q_\phi(\theta|D)] = \underbrace{\int \log q_\phi(\theta|D) \cdot q_\phi(\theta|D)\, d\theta}_{-H[q_\phi(\theta|D)]\text{: negative entropy of } q_\phi(\theta|D)} =$

e.g. mean-field:

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

<u>2^{nd} term</u>: $\mathbb{E}_{\vartheta \sim q_\varphi(\vartheta|\mathcal{D})}\left[\log p(\vartheta|\mathcal{D})\right] =$

constant

$$\overset{\text{Bayes}}{=} \mathbb{E}_{\vartheta \sim q_\varphi(\vartheta|\mathcal{D})}\left[\underbrace{\log p(\mathcal{D}|\vartheta)} + \underbrace{\log p(\vartheta)} - \cancel{\log p(\mathcal{D})}\right]$$

$$\Longrightarrow \mathbb{KL}\left[q_\varphi(\vartheta|\mathcal{D}) \,\|\, p(\vartheta|\mathcal{D})\right] = -\overbrace{H\left[q_\varphi(\vartheta|\mathcal{D})\right]}^{\text{analytically}}$$

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

Now, all terms can be evaluated.

But here's the catch:

$$\left\{ \begin{aligned} \varphi^* &= \arg\min_{\varphi} \mathcal{L}(\varphi) \\ \mathcal{L}(\varphi) &:= -H\left[q_\varphi(\vartheta|\mathcal{D})\right] - \mathbb{E}_{\vartheta \sim q_\varphi(\vartheta|\mathcal{D})}\left[\log p(\mathcal{D}|\vartheta) + \log p(\vartheta)\right] \end{aligned} \right\}$$

Minimization via gradient descent:

$$\varphi_{n+1} = \varphi_n - \eta \underline{\nabla_\varphi \mathcal{L}(\varphi)}$$

<u>Remarks</u>:

1.) All terms in $\mathcal{L}(\varphi)$ can be evaluated. Sometimes this can be done analytically (e.g. linear models with Gaussian likelihood and prior).

2.) We need to compute $\nabla_\varphi \mathcal{L}(\varphi)$:

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

We could sample $\theta_i \sim q_\varphi(\theta|D)$ and use a Monte-Carlo estimator to compute the gradient :

$$\nabla_\varphi \mathbb{E}_{\theta \sim q_\varphi(\theta|D)} \left[\log p(D|\theta) + \log p(\theta) \right] \approx \frac{1}{n} \sum_{i=1}^{n} \left[\underline{\nabla_\varphi} \log p(D|\theta_i) + \nabla_\varphi \log p(\theta_i) \right.$$

$$, \text{where } \theta_i \overset{i.i.d}{\sim} q_{\underset{=}{\varphi}}(\theta_i|D)$$

However, notice that $q_\varphi(\theta_i|D)$ depends on φ, and as φ is changing during optimization, this M.C. estimator will exhibit very high variance. I.e. we will need a very large number of MC samples to get a reasonable approximat of the gradient.

Reparametrization trick (next time) :

Introduce a simple "change of variables" such that the required expectations can be computed with respect to distributions that do not depend on φ.