Bayesian inference

Setting: We are given a model with unknown parameters $\theta \in \mathbb{R}^d$, some data $\mathcal{D}$ distributed according to the likelihood of the mod[el]

$p(\mathcal{D}|\theta)$, and a prior $p(\theta)$.

Goal: Infer the posterior:

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

i.) Variational infere[nce]

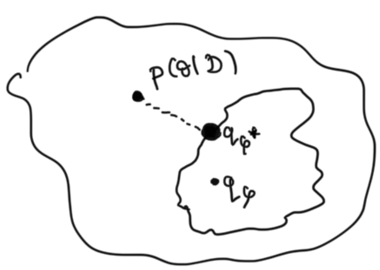
ii.) MCMC

VI: Idea: Approximate the intractable using a family of distributions that is easy to work with.

Most common choice (Mean-field approximation):

$$p(\theta|\mathcal{D}) \approx \overbrace{q_\varphi}^{\text{variational family}}(\theta|\mathcal{D}) = \prod_{i=1}^{d} \mathcal{N}(\theta_i|\mu_i, \sigma$$

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



Goal: Find/estimate $\varphi^* := \overbrace{\{\mu_1, \sigma_1^2, \ldots, \mu_d, \sigma_d^2\}}^{2d}$ such that $q_{\varphi^*}(\theta|\mathcal{D})$ is "as close as possible" to $p(\theta|\mathcal{D})$.

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

<u>Recall</u>: $\mathbb{KL}[q_\varphi(\theta|\mathcal{D}) \,||\, p(\theta|\mathcal{D})] = \int \log \frac{q_\varphi(\theta|\mathcal{D})}{p(\theta|\mathcal{D})} q_\varphi(\theta|\mathcal{D})\, d$

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

Why $\mathbb{KL}$? • It's easy to work with

• $\mathbb{KL}[q_\varphi \,||\, p] = 0 \iff q_\varphi = p$

••• but it's not an actual distance:

$$\mathbb{KL}[q_\varphi||p] \neq \mathbb{KL}[p \,||\, q_\varphi]$$

<u>Remarks</u>:

i.) Typically mean-field VI tends to favor approximatio that capture well the mean of $p(\theta|\mathcal{D})$, but underestimate the variance!

ii) It's hardly ideal, but in cases where other methods don't scale, it still provide useful inference.

How to estimate $\varphi^* = \arg\min_{\varphi} \mathbb{KL}[q_\varphi(\theta|\mathcal{D}) \,||\, p(\theta|\mathcal{D})]$

→ Old-school approaches had to derive coordinate ascent rules for minimizing the $\mathbb{KL}$ (see Ch. 10 Bishop).

→ New-schoolers use Automatic Differentiation Variational Infere

ADVI : It is a "black-box" approach that is agnostic to any details about $p(\theta|D)$: any model for which we can evaluate (and differentiate) its log-likelihood and log-prior distribution works!

Let's see how it works!

$$\mathbb{KL}\left[q_\phi(\theta|D) \,\|\, p(\theta|D)\right] = \mathbb{E}_{\theta\sim q_\phi(\theta|D)}\Big[\underbrace{\log q_\phi(\theta|D)}_{\#1} - \underbrace{\log p(\theta|}_{\#2}$$

1st term : $\mathbb{E}_{\theta\sim q_\phi(\theta|D)}\left[\log q_\phi(\theta|D)\right] = \underbrace{\int \log q_\phi(\theta|D)\, q_\phi(\theta|D)\, d\theta}$

$$= -H\left[q_\phi(\theta|D)\right]$$

Notice that we are free to choose any $q_\phi(\theta|D)$ that suits us, hence we can choose one for which $H[q_\phi]$ is computable :

e.g. MF-VI : $\mathbb{E}_{\theta\sim q_\phi(\theta|D)}\Big[\underbrace{q_\phi(\theta|D)}_{\prod_{i=1}^{d}\mathcal{N}(\theta_i|\mu_i,\sigma_i^2)}\Big] = -\sum_{i=1}^{d}\log\sigma_i + \text{constant}$

Term #2 :

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

$$\mathcal{L}(\phi) := -H\left[q_\phi(\theta|D)\right] - \mathbb{E}\left[\log p(D|\theta) + \log p(\theta)\right]$$

$$\theta \sim q_\varphi(\theta|\mathcal{D})$$

Now we can use gradient-descent to estimate φ^*:

$$\varphi^* = \arg\min_\varphi \mathcal{L}(\varphi)$$

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

All terms in $\mathcal{L}(\varphi)$ can now be evaluated, however, we still need to compute:

$$\boxed{\underbrace{\nabla_\varphi \mathbb{E}_{\theta \sim q_\varphi(\theta|\mathcal{D})} \left[\log p(\mathcal{D}|\theta) + \log p(\theta)\right]}_{\text{exhibits high-variance when approximated via Monte-Carlo sampling}}}$$

Example tutorial:

Given $p(x)$, then try to fit $q_\varphi(x)$ such that

$\mathbb{KL}[q_\varphi(x) || p(x)]$ is minimized.

$$\varphi^* = \arg\min_\varphi \mathbb{E}_{x \sim q_\varphi(x)} \left[\underbrace{\log q_\varphi(x) - \overbrace{\log p(x)}^{\text{Known}}}\right]$$

$$\approx \frac{1}{n} \sum_{i=1}^{n} \overbrace{\left[\log q_\varphi(x_i) - \log p(x_i)\right]},$$

$$x_i \sim q_\varphi(x)$$