Variational Inference

Nipun Batra

November 16, 2023

IIT Gandhinagar

Introduction

Bayesian ML: Recap

- We assume a prior distribution over the parameters of the model given as $P(\theta)$
- We assume a likelihood function $P(D|\theta)$
- We use Bayes' rule to find the posterior distribution of the parameters given the data: $P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$
- Typically, we can not compute the posterior distribution analytically as the denominator is intractable

Bayesian ML: Methods

Laplace Approximation

Approximates the posterior with a Gaussian distribution parameterized by $\Psi = (\mu, \Sigma)$.

$$q_\Psi(\theta) = \mathcal{N}(\mu, \Sigma)$$

where μ is the mode of the posterior and Σ is the negative inverse Hessian of the log joint distribution evaluated at θ_{MAP}.

MCMC (Markov Chain Monte Carlo)

Generates samples from the posterior distribution by constructing a Markov chain.

$$P(\theta|D) \propto P(D|\theta)P(\theta)$$

Variational Inference

Poses posterior inference as an optimization problem. The approximating distribution is parameterized by Ψ.

$$\Psi^* = \arg\min_{\Psi} \text{KL}(q_\Psi(\theta)||P(\theta|D))$$

KL Divergence

- KL divergence is a measure of dissimilarity between two distributions.
- It is defined as: $\mathrm{KL}(q||p) = \int q(\theta) \log \frac{q(\theta)}{p(\theta)} d\theta$
- Or, can be written in terms of expectations as: $\mathrm{KL}(q||p) = \mathbb{E}_{q(\theta)}\left[\log \frac{q(\theta)}{p(\theta)}\right]$

Exercise

Compute the KL divergence between two Gaussian distributions $q(\theta) = \mathcal{N}(\mu_q, \sigma_q^2)$ and $p(\theta) = \mathcal{N}(\mu_p, \sigma_p^2)$.

Exercise

Compute the KL divergence between two Gaussian distributions $q(\theta) = \mathcal{N}(\mu_q, \sigma_q^2)$ and $p(\theta) = \mathcal{N}(\mu_p, \sigma_p^2)$.

- $\mathrm{KL}(q||p) = \mathbb{E}_{q(\theta)}\left[\log \frac{q(\theta)}{p(\theta)}\right]$

Exercise

Compute the KL divergence between two Gaussian distributions $q(\theta) = \mathcal{N}(\mu_q, \sigma_q^2)$ and $p(\theta) = \mathcal{N}(\mu_p, \sigma_p^2)$.

- $\mathrm{KL}(q||p) = \mathbb{E}_{q(\theta)}\left[\log \frac{q(\theta)}{p(\theta)}\right]$
- Expanding $q(\theta) = \mathcal{N}(\mu_q, \sigma_q^2) = \frac{1}{\sqrt{2\pi\sigma_q^2}} \exp\left(-\frac{(\theta-\mu_q)^2}{2\sigma_q^2}\right)$

Exercise

Compute the KL divergence between two Gaussian distributions $q(\theta) = \mathcal{N}(\mu_q, \sigma_q^2)$ and $p(\theta) = \mathcal{N}(\mu_p, \sigma_p^2)$.

- $\mathrm{KL}(q||p) = \mathbb{E}_{q(\theta)}\left[\log \frac{q(\theta)}{p(\theta)}\right]$
- Expanding $q(\theta) = \mathcal{N}(\mu_q, \sigma_q^2) = \frac{1}{\sqrt{2\pi\sigma_q^2}} \exp\left(-\frac{(\theta-\mu_q)^2}{2\sigma_q^2}\right)$
- $\mathrm{KL}(q||p) = \mathbb{E}_{q(\theta)}\left[\log \frac{q(\theta)}{p(\theta)}\right] =$
  $\mathbb{E}_{q(\theta)}\left[\log \frac{\frac{1}{\sqrt{2\pi\sigma_q^2}} \exp\left(-\frac{(\theta-\mu_q)^2}{2\sigma_q^2}\right)}{\frac{1}{\sqrt{2\pi\sigma_p^2}} \exp\left(-\frac{(\theta-\mu_p)^2}{2\sigma_p^2}\right)}\right]$

Exercise

Compute the KL divergence between two Gaussian distributions $q(\theta) = \mathcal{N}(\mu_q, \sigma_q^2)$ and $p(\theta) = \mathcal{N}(\mu_p, \sigma_p^2)$.

- $\mathrm{KL}(q||p) = \mathbb{E}_{q(\theta)}\left[\log \frac{q(\theta)}{p(\theta)}\right]$
- Expanding $q(\theta) = \mathcal{N}(\mu_q, \sigma_q^2) = \frac{1}{\sqrt{2\pi\sigma_q^2}} \exp\left(-\frac{(\theta-\mu_q)^2}{2\sigma_q^2}\right)$
- $\mathrm{KL}(q||p) = \mathbb{E}_{q(\theta)}\left[\log \frac{q(\theta)}{p(\theta)}\right] =$
  $\mathbb{E}_{q(\theta)}\left[\log \frac{\frac{1}{\sqrt{2\pi\sigma_q^2}} \exp\left(-\frac{(\theta-\mu_q)^2}{2\sigma_q^2}\right)}{\frac{1}{\sqrt{2\pi\sigma_p^2}} \exp\left(-\frac{(\theta-\mu_p)^2}{2\sigma_p^2}\right)}\right]$
- $\mathrm{KL}(q||p) = \mathbb{E}_{q(\theta)}\left[\log \frac{\frac{1}{\sqrt{2\pi\sigma_q^2}}}{\frac{1}{\sqrt{2\pi\sigma_p^2}}} + \log \frac{\exp\left(-\frac{(\theta-\mu_q)^2}{2\sigma_q^2}\right)}{\exp\left(-\frac{(\theta-\mu_p)^2}{2\sigma_p^2}\right)}\right]$

Exercise

Compute the KL divergence between two Gaussian distributions $q(\theta) = \mathcal{N}(\mu_q, \sigma_q^2)$ and $p(\theta) = \mathcal{N}(\mu_p, \sigma_p^2)$.

- $\mathrm{KL}(q||p) = \mathbb{E}_{q(\theta)}\left[\log \frac{q(\theta)}{p(\theta)}\right]$
- Expanding $q(\theta) = \mathcal{N}(\mu_q, \sigma_q^2) = \frac{1}{\sqrt{2\pi\sigma_q^2}} \exp\left(-\frac{(\theta-\mu_q)^2}{2\sigma_q^2}\right)$
- $\mathrm{KL}(q||p) = \mathbb{E}_{q(\theta)}\left[\log \frac{q(\theta)}{p(\theta)}\right] =$
  $\mathbb{E}_{q(\theta)}\left[\log \frac{\frac{1}{\sqrt{2\pi\sigma_q^2}} \exp\left(-\frac{(\theta-\mu_q)^2}{2\sigma_q^2}\right)}{\frac{1}{\sqrt{2\pi\sigma_p^2}} \exp\left(-\frac{(\theta-\mu_p)^2}{2\sigma_p^2}\right)}\right]$
- $\mathrm{KL}(q||p) = \mathbb{E}_{q(\theta)}\left[\log \frac{\frac{1}{\sqrt{2\pi\sigma_q^2}}}{\frac{1}{\sqrt{2\pi\sigma_p^2}}} + \log \frac{\exp\left(-\frac{(\theta-\mu_q)^2}{2\sigma_q^2}\right)}{\exp\left(-\frac{(\theta-\mu_p)^2}{2\sigma_p^2}\right)}\right]$

The answer is: $\frac{1}{2}\left(\log \frac{\sigma_2^2}{\sigma_1^2} + \frac{\sigma_1^2+(\mu_1-\mu_2)^2}{\sigma_2^2} - 1\right)$

Exercise

Compute the KL divergence between two Gaussian distributions $q(\theta) = \mathcal{N}(\mu_q, \sigma_q^2)$ and $p(\theta) = \mathcal{N}(\mu_p, \sigma_p^2)$.

Exercise

Compute the KL divergence between two Gaussian distributions $q(\theta) = \mathcal{N}(\mu_q, \sigma_q^2)$ and $p(\theta) = \mathcal{N}(\mu_p, \sigma_p^2)$.

- $\mathrm{KL}(q||p) = \mathbb{E}_{q(\theta)}\left[\log \frac{\frac{1}{\sqrt{2\pi\sigma_q^2}}}{\frac{1}{\sqrt{2\pi\sigma_p^2}}} + \log \frac{\exp\left(-\frac{(\theta-\mu_q)^2}{2\sigma_q^2}\right)}{\exp\left(-\frac{(\theta-\mu_p)^2}{2\sigma_p^2}\right)}\right]$

Exercise

Compute the KL divergence between two Gaussian distributions $q(\theta) = \mathcal{N}(\mu_q, \sigma_q^2)$ and $p(\theta) = \mathcal{N}(\mu_p, \sigma_p^2)$.

- $\mathrm{KL}(q||p) = \mathbb{E}_{q(\theta)}\left[\log \frac{\frac{1}{\sqrt{2\pi\sigma_q^2}}}{\frac{1}{\sqrt{2\pi\sigma_p^2}}} + \log \frac{\exp\left(-\frac{(\theta-\mu_q)^2}{2\sigma_q^2}\right)}{\exp\left(-\frac{(\theta-\mu_p)^2}{2\sigma_p^2}\right)}\right]$
- $\mathrm{KL}(q||p) = \mathbb{E}_{q(\theta)}\left[\log \frac{\frac{1}{\sqrt{2\pi\sigma_q^2}}}{\frac{1}{\sqrt{2\pi\sigma_p^2}}} + \log\exp\left(-\frac{(\theta-\mu_q)^2}{2\sigma_q^2} + \frac{(\theta-\mu_p)^2}{2\sigma_p^2}\right)\right]$

Exercise

Compute the KL divergence between two Gaussian distributions $q(\theta) = \mathcal{N}(\mu_q, \sigma_q^2)$ and $p(\theta) = \mathcal{N}(\mu_p, \sigma_p^2)$.

- $\mathrm{KL}(q||p) = \mathbb{E}_{q(\theta)}\left[\log \frac{\frac{1}{\sqrt{2\pi\sigma_q^2}}}{\frac{1}{\sqrt{2\pi\sigma_p^2}}} + \log \frac{\exp\left(-\frac{(\theta-\mu_q)^2}{2\sigma_q^2}\right)}{\exp\left(-\frac{(\theta-\mu_p)^2}{2\sigma_p^2}\right)}\right]$
- $\mathrm{KL}(q||p) = \mathbb{E}_{q(\theta)}\left[\log \frac{\frac{1}{\sqrt{2\pi\sigma_q^2}}}{\frac{1}{\sqrt{2\pi\sigma_p^2}}} + \log \exp\left(-\frac{(\theta-\mu_q)^2}{2\sigma_q^2} + \frac{(\theta-\mu_p)^2}{2\sigma_p^2}\right)\right]$
- $\mathrm{KL}(q||p) = \mathbb{E}_{q(\theta)}\left[\log \frac{\frac{1}{\sqrt{2\pi\sigma_q^2}}}{\frac{1}{\sqrt{2\pi\sigma_p^2}}} + \left(-\frac{(\theta-\mu_q)^2}{2\sigma_q^2} + \frac{(\theta-\mu_p)^2}{2\sigma_p^2}\right)\right]$

Exercise

Compute the KL divergence between two Gaussian distributions $q(\theta) = \mathcal{N}(\mu_q, \sigma_q^2)$ and $p(\theta) = \mathcal{N}(\mu_p, \sigma_p^2)$.

- $\mathrm{KL}(q||p) = \mathbb{E}_{q(\theta)}\left[\log \frac{\frac{1}{\sqrt{2\pi\sigma_q^2}}}{\frac{1}{\sqrt{2\pi\sigma_p^2}}} + \log \frac{\exp\left(-\frac{(\theta-\mu_q)^2}{2\sigma_q^2}\right)}{\exp\left(-\frac{(\theta-\mu_p)^2}{2\sigma_p^2}\right)}\right]$
- $\mathrm{KL}(q||p) = \mathbb{E}_{q(\theta)}\left[\log \frac{\frac{1}{\sqrt{2\pi\sigma_q^2}}}{\frac{1}{\sqrt{2\pi\sigma_p^2}}} + \log \exp\left(-\frac{(\theta-\mu_q)^2}{2\sigma_q^2} + \frac{(\theta-\mu_p)^2}{2\sigma_p^2}\right)\right]$
- $\mathrm{KL}(q||p) = \mathbb{E}_{q(\theta)}\left[\log \frac{\frac{1}{\sqrt{2\pi\sigma_q^2}}}{\frac{1}{\sqrt{2\pi\sigma_p^2}}} + \left(-\frac{(\theta-\mu_q)^2}{2\sigma_q^2} + \frac{(\theta-\mu_p)^2}{2\sigma_p^2}\right)\right]$
- $\mathrm{KL}(q||p) = \mathbb{E}_{q(\theta)}\left[\log \frac{\frac{1}{\sqrt{2\pi\sigma_q^2}}}{\frac{1}{\sqrt{2\pi\sigma_p^2}}} + \left(-\frac{(\theta^2-2\theta\mu_q+\mu_q^2)}{2\sigma_q^2} + \frac{(\theta^2-2\theta\mu_p+\mu_p^2)}{2\sigma_p^2}\right)\right]$

Exercise

Compute the KL divergence between two Gaussian distributions $q(\theta) = \mathcal{N}(\mu_q, \sigma_q^2)$ and $p(\theta) = \mathcal{N}(\mu_p, \sigma_p^2)$.

Exercise

Compute the KL divergence between two Gaussian distributions $q(\theta) = \mathcal{N}(\mu_q, \sigma_q^2)$ and $p(\theta) = \mathcal{N}(\mu_p, \sigma_p^2)$.

- $\mathrm{KL}(q||p) =$
  $$\mathbb{E}_{q(\theta)}\left[\log \frac{\frac{1}{\sqrt{2\pi\sigma_q^2}}}{\frac{1}{\sqrt{2\pi\sigma_p^2}}} + \left(-\frac{(\theta^2 - 2\theta\mu_q + \mu_q^2)}{2\sigma_q^2} + \frac{(\theta^2 - 2\theta\mu_p + \mu_p^2)}{2\sigma_p^2}\right)\right]$$

Exercise

Compute the KL divergence between two Gaussian distributions $q(\theta) = \mathcal{N}(\mu_q, \sigma_q^2)$ and $p(\theta) = \mathcal{N}(\mu_p, \sigma_p^2)$.

- $\mathrm{KL}(q||p) =$
$$\mathbb{E}_{q(\theta)}\left[\log \frac{\frac{1}{\sqrt{2\pi\sigma_q^2}}}{\frac{1}{\sqrt{2\pi\sigma_p^2}}} + \left(-\frac{(\theta^2 - 2\theta\mu_q + \mu_q^2)}{2\sigma_q^2} + \frac{(\theta^2 - 2\theta\mu_p + \mu_p^2)}{2\sigma_p^2}\right)\right]$$
- $\mathrm{KL}(q||p) =$
$$\mathbb{E}_{q(\theta)}\left[\log \frac{\frac{1}{\sqrt{2\pi\sigma_q^2}}}{\frac{1}{\sqrt{2\pi\sigma_p^2}}} + \left(-\frac{\theta^2}{2\sigma_q^2} + \frac{2\theta\mu_q}{2\sigma_q^2} - \frac{\mu_q^2}{2\sigma_q^2} + \frac{\theta^2}{2\sigma_p^2} - \frac{2\theta\mu_p}{2\sigma_p^2} + \frac{\mu_p^2}{2\sigma_p^2}\right)\right]$$

Exercise

Compute the KL divergence between two Gaussian distributions $q(\theta) = \mathcal{N}(\mu_q, \sigma_q^2)$ and $p(\theta) = \mathcal{N}(\mu_p, \sigma_p^2)$.

- $\mathrm{KL}(q||p) =$
$$\mathbb{E}_{q(\theta)}\left[\log \frac{\frac{1}{\sqrt{2\pi\sigma_q^2}}}{\frac{1}{\sqrt{2\pi\sigma_p^2}}} + \left(-\frac{(\theta^2 - 2\theta\mu_q + \mu_q^2)}{2\sigma_q^2} + \frac{(\theta^2 - 2\theta\mu_p + \mu_p^2)}{2\sigma_p^2}\right)\right]$$
- $\mathrm{KL}(q||p) =$
$$\mathbb{E}_{q(\theta)}\left[\log \frac{\frac{1}{\sqrt{2\pi\sigma_q^2}}}{\frac{1}{\sqrt{2\pi\sigma_p^2}}} + \left(-\frac{\theta^2}{2\sigma_q^2} + \frac{2\theta\mu_q}{2\sigma_q^2} - \frac{\mu_q^2}{2\sigma_q^2} + \frac{\theta^2}{2\sigma_p^2} - \frac{2\theta\mu_p}{2\sigma_p^2} + \frac{\mu_p^2}{2\sigma_p^2}\right)\right]$$
- Now using linearity of expectation, we get:

Exercise

Compute the KL divergence between two Gaussian distributions $q(\theta) = \mathcal{N}(\mu_q, \sigma_q^2)$ and $p(\theta) = \mathcal{N}(\mu_p, \sigma_p^2)$.

- $\mathrm{KL}(q||p) = \mathbb{E}_{q(\theta)}\left[\log \frac{\frac{1}{\sqrt{2\pi\sigma_q^2}}}{\frac{1}{\sqrt{2\pi\sigma_p^2}}} + \left(-\frac{(\theta^2 - 2\theta\mu_q + \mu_q^2)}{2\sigma_q^2} + \frac{(\theta^2 - 2\theta\mu_p + \mu_p^2)}{2\sigma_p^2}\right)\right]$
- $\mathrm{KL}(q||p) = \mathbb{E}_{q(\theta)}\left[\log \frac{\frac{1}{\sqrt{2\pi\sigma_q^2}}}{\frac{1}{\sqrt{2\pi\sigma_p^2}}} + \left(-\frac{\theta^2}{2\sigma_q^2} + \frac{2\theta\mu_q}{2\sigma_q^2} - \frac{\mu_q^2}{2\sigma_q^2} + \frac{\theta^2}{2\sigma_p^2} - \frac{2\theta\mu_p}{2\sigma_p^2} + \frac{\mu_p^2}{2\sigma_p^2}\right)\right]$
- Now using linearity of expectation, we get:
- $\mathrm{KL}(q||p) = \log \frac{\sigma_p}{\sigma_q} + \mathbb{E}_{q(\theta)}(-\frac{\theta^2}{2\sigma_q^2} + \frac{2\theta\mu_q}{2\sigma_q^2} - \frac{\mu_q^2}{2\sigma_q^2} + \frac{\theta^2}{2\sigma_p^2} - \frac{2\theta\mu_p}{2\sigma_p^2} + \frac{\mu_p^2}{2\sigma_p^2})$

Exercise

Aside:

$\theta \sim q(\theta) = \mathcal{N}(\mu_q, \sigma_q^2)$

$\mathbb{E}_{q(\theta)}\left[\theta\right] = \mu_q$

$\mathbb{E}_{q(\theta)}\left[\theta^2\right] = \sigma_q^2 + \mu_q^2$

Exercise

Compute the KL divergence between two Gaussian distributions $q(\theta) = \mathcal{N}(\mu_q, \sigma_q^2)$ and $p(\theta) = \mathcal{N}(\mu_p, \sigma_p^2)$.

Exercise

Compute the KL divergence between two Gaussian distributions $q(\theta) = \mathcal{N}(\mu_q, \sigma_q^2)$ and $p(\theta) = \mathcal{N}(\mu_p, \sigma_p^2)$.

- $\mathrm{KL}(q||p) = \log \frac{\sigma_p}{\sigma_q} + \mathbb{E}_{q(\theta)}\left(-\frac{\theta^2}{2\sigma_q^2} + \frac{2\theta\mu_q}{2\sigma_q^2} - \frac{\mu_q^2}{2\sigma_q^2} + \frac{\theta^2}{2\sigma_p^2} - \frac{2\theta\mu_p}{2\sigma_p^2} + \frac{\mu_p^2}{2\sigma_p^2}\right)$

Exercise

Compute the KL divergence between two Gaussian distributions $q(\theta) = \mathcal{N}(\mu_q, \sigma_q^2)$ and $p(\theta) = \mathcal{N}(\mu_p, \sigma_p^2)$.

- $\mathrm{KL}(q||p) = \log \frac{\sigma_p}{\sigma_q} + \mathbb{E}_{q(\theta)}\left(-\frac{\theta^2}{2\sigma_q^2} + \frac{2\theta\mu_q}{2\sigma_q^2} - \frac{\mu_q^2}{2\sigma_q^2} + \frac{\theta^2}{2\sigma_p^2} - \frac{2\theta\mu_p}{2\sigma_p^2} + \frac{\mu_p^2}{2\sigma_p^2}\right)$
- Using the aside, we expand the expectation:

Exercise

Compute the KL divergence between two Gaussian distributions $q(\theta) = \mathcal{N}(\mu_q, \sigma_q^2)$ and $p(\theta) = \mathcal{N}(\mu_p, \sigma_p^2)$.

- $\text{KL}(q||p) = \log \frac{\sigma_p}{\sigma_q} + \mathbb{E}_{q(\theta)}\left(-\frac{\theta^2}{2\sigma_q^2} + \frac{2\theta\mu_q}{2\sigma_q^2} - \frac{\mu_q^2}{2\sigma_q^2} + \frac{\theta^2}{2\sigma_p^2} - \frac{2\theta\mu_p}{2\sigma_p^2} + \frac{\mu_p^2}{2\sigma_p^2}\right)$
- Using the aside, we expand the expectation:
- $\text{KL}(q||p) =$ Term 1 $+$ Term 2 $+$ Term 3 $+$ Term 4 $+$ Term 5 $+$ Term 6 $+$ Term 7

Exercise

Compute the KL divergence between two Gaussian distributions $q(\theta) = \mathcal{N}(\mu_q, \sigma_q^2)$ and $p(\theta) = \mathcal{N}(\mu_p, \sigma_p^2)$.

- $\mathrm{KL}(q||p) = \log \frac{\sigma_p}{\sigma_q} + \mathbb{E}_{q(\theta)}(-\frac{\theta^2}{2\sigma_q^2} + \frac{2\theta\mu_q}{2\sigma_q^2} - \frac{\mu_q^2}{2\sigma_q^2} + \frac{\theta^2}{2\sigma_p^2} - \frac{2\theta\mu_p}{2\sigma_p^2} + \frac{\mu_p^2}{2\sigma_p^2})$
- Using the aside, we expand the expectation:
- $\mathrm{KL}(q||p) =$ Term 1 $+$ Term 2 $+$ Term 3 $+$ Term 4 $+$ Term 5 $+$ Term 6 $+$ Term 7
- Term 1 $= \log \frac{\sigma_p}{\sigma_q}$

Exercise

Compute the KL divergence between two Gaussian distributions $q(\theta) = \mathcal{N}(\mu_q, \sigma_q^2)$ and $p(\theta) = \mathcal{N}(\mu_p, \sigma_p^2)$.

- $\mathrm{KL}(q||p) = \log \frac{\sigma_p}{\sigma_q} + \mathbb{E}_{q(\theta)}(-\frac{\theta^2}{2\sigma_q^2} + \frac{2\theta\mu_q}{2\sigma_q^2} - \frac{\mu_q^2}{2\sigma_q^2} + \frac{\theta^2}{2\sigma_p^2} - \frac{2\theta\mu_p}{2\sigma_p^2} + \frac{\mu_p^2}{2\sigma_p^2})$
- Using the aside, we expand the expectation:
- $\mathrm{KL}(q||p) =$ Term 1 + Term 2 + Term 3 + Term 4 + Term 5 + Term 6 + Term 7
- Term 1 $= \log \frac{\sigma_p}{\sigma_q}$
- Term 2: $\mathbb{E}_{q(\theta)}\left(-\frac{\theta^2}{2\sigma_q^2}\right) = -\frac{1}{2}\mathbb{E}_{q(\theta)}\left(\frac{\theta^2}{\sigma_q^2}\right) = -\frac{1}{2}\left(\frac{\sigma_q^2+\mu_q^2}{\sigma_q^2}\right)$

Exercise

Compute the KL divergence between two Gaussian distributions $q(\theta) = \mathcal{N}(\mu_q, \sigma_q^2)$ and $p(\theta) = \mathcal{N}(\mu_p, \sigma_p^2)$.

- $\mathrm{KL}(q||p) = \log \frac{\sigma_p}{\sigma_q} + \mathbb{E}_{q(\theta)}(-\frac{\theta^2}{2\sigma_q^2} + \frac{2\theta\mu_q}{2\sigma_q^2} - \frac{\mu_q^2}{2\sigma_q^2} + \frac{\theta^2}{2\sigma_p^2} - \frac{2\theta\mu_p}{2\sigma_p^2} + \frac{\mu_p^2}{2\sigma_p^2})$
- Using the aside, we expand the expectation:
- $\mathrm{KL}(q||p) = \text{Term 1} + \text{Term 2} + \text{Term 3} + \text{Term 4} + \text{Term 5} + \text{Term 6} + \text{Term 7}$
- Term $1 = \log \frac{\sigma_p}{\sigma_q}$
- Term 2: $\mathbb{E}_{q(\theta)}\left(-\frac{\theta^2}{2\sigma_q^2}\right) = -\frac{1}{2}\mathbb{E}_{q(\theta)}\left(\frac{\theta^2}{\sigma_q^2}\right) = -\frac{1}{2}\left(\frac{\sigma_q^2 + \mu_q^2}{\sigma_q^2}\right)$
- Term 3: $\mathbb{E}_{q(\theta)}\left(\frac{2\theta\mu_q}{2\sigma_q^2}\right) = \frac{2\mu_q}{2\sigma_q^2}\mathbb{E}_{q(\theta)}(\theta) = \frac{2\mu_q^2}{2\sigma_q^2}$

Exercise

Compute the KL divergence between two Gaussian distributions $q(\theta) = \mathcal{N}(\mu_q, \sigma_q^2)$ and $p(\theta) = \mathcal{N}(\mu_p, \sigma_p^2)$.

- Term 4: $\mathbb{E}_{q(\theta)}\left(-\frac{\mu_q^2}{2\sigma_q^2}\right) = -\frac{\mu_q^2}{2\sigma_q^2}$

Exercise

Compute the KL divergence between two Gaussian distributions $q(\theta) = \mathcal{N}(\mu_q, \sigma_q^2)$ and $p(\theta) = \mathcal{N}(\mu_p, \sigma_p^2)$.

- Term 4: $\mathbb{E}_{q(\theta)}\left(-\frac{\mu_q^2}{2\sigma_q^2}\right) = -\frac{\mu_q^2}{2\sigma_q^2}$
- Term 5: $\mathbb{E}_{q(\theta)}\left(\frac{\theta^2}{2\sigma_p^2}\right) = \frac{\sigma_q^2+\mu_q^2}{2\sigma_p^2}$

Exercise

Compute the KL divergence between two Gaussian distributions $q(\theta) = \mathcal{N}(\mu_q, \sigma_q^2)$ and $p(\theta) = \mathcal{N}(\mu_p, \sigma_p^2)$.

- Term 4: $\mathbb{E}_{q(\theta)}\left(-\frac{\mu_q^2}{2\sigma_q^2}\right) = -\frac{\mu_q^2}{2\sigma_q^2}$
- Term 5: $\mathbb{E}_{q(\theta)}\left(\frac{\theta^2}{2\sigma_p^2}\right) = \frac{\sigma_q^2+\mu_q^2}{2\sigma_p^2}$
- Term 6: $\mathbb{E}_{q(\theta)}\left(-\frac{2\theta\mu_p}{2\sigma_p^2}\right) = -\frac{2\mu_q\mu_p}{2\sigma_p^2}$

Exercise

Compute the KL divergence between two Gaussian distributions $q(\theta) = \mathcal{N}(\mu_q, \sigma_q^2)$ and $p(\theta) = \mathcal{N}(\mu_p, \sigma_p^2)$.

- Term 4: $\mathbb{E}_{q(\theta)}\left(-\frac{\mu_q^2}{2\sigma_q^2}\right) = -\frac{\mu_q^2}{2\sigma_q^2}$
- Term 5: $\mathbb{E}_{q(\theta)}\left(\frac{\theta^2}{2\sigma_p^2}\right) = \frac{\sigma_q^2+\mu_q^2}{2\sigma_p^2}$
- Term 6: $\mathbb{E}_{q(\theta)}\left(-\frac{2\theta\mu_p}{2\sigma_p^2}\right) = -\frac{2\mu_q\mu_p}{2\sigma_p^2}$
- Term 7: $\mathbb{E}_{q(\theta)}\left(\frac{\mu_p^2}{2\sigma_p^2}\right) = \frac{\mu_p^2}{2\sigma_p^2}$

Exercise

Compute the KL divergence between two Gaussian distributions $q(\theta) = \mathcal{N}(\mu_q, \sigma_q^2)$ and $p(\theta) = \mathcal{N}(\mu_p, \sigma_p^2)$.

- Term 4: $\mathbb{E}_{q(\theta)}\left(-\frac{\mu_q^2}{2\sigma_q^2}\right) = -\frac{\mu_q^2}{2\sigma_q^2}$
- Term 5: $\mathbb{E}_{q(\theta)}\left(\frac{\theta^2}{2\sigma_p^2}\right) = \frac{\sigma_q^2+\mu_q^2}{2\sigma_p^2}$
- Term 6: $\mathbb{E}_{q(\theta)}\left(-\frac{2\theta\mu_p}{2\sigma_p^2}\right) = -\frac{2\mu_q\mu_p}{2\sigma_p^2}$
- Term 7: $\mathbb{E}_{q(\theta)}\left(\frac{\mu_p^2}{2\sigma_p^2}\right) = \frac{\mu_p^2}{2\sigma_p^2}$
- Overall after simplification, we get:
  $\mathrm{KL}(q||p) = \frac{1}{2}[\log\frac{\sigma_p^2}{\sigma_q^2} - 1 + \frac{(\mu_p-\mu_q)^2}{\sigma_p^2} + \frac{\sigma_q^2}{\sigma_p^2}]$

Notebook demo

Optimizing

Notebook demo

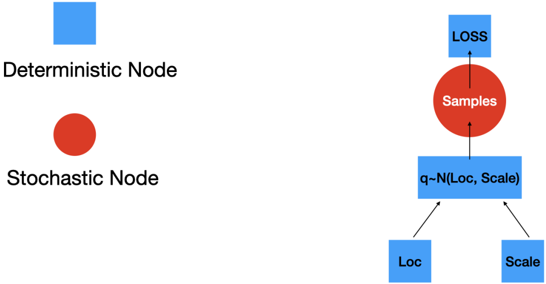
Monte Carlo Sampling

Notebook demo

Repameterization Trick

Original formulation

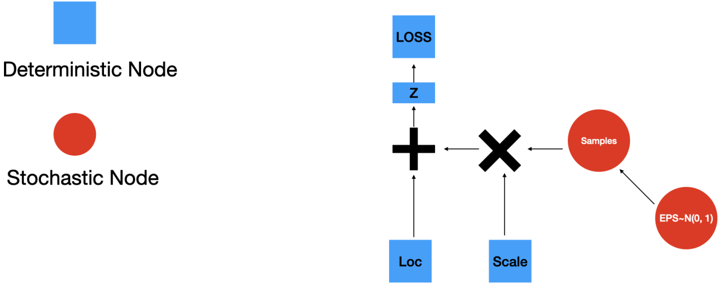
```python
def original_loss(loc, scale):
    q = dist.Normal(loc=loc, scale=scale)
    sample_set = q.sample([n_samples])
    return torch.mean(q.log_prob(sample_set) - p_s.log_prob(sample_set))
```



Repameterization Trick

New formulation (reparameterization trick)

```
def loss(loc, scale):
    q = dist.Normal(loc=loc, scale=scale)
    std_normal = dist.Normal(loc=0.0, scale=1.0)
    sample_set = std_normal.sample([n_samples])
    sample_set = loc + scale * sample_set
    return torch.mean(q.log_prob(sample_set) - p_s.log_prob(sample_set))
```



Notebook demo

Evidence Lower Bound (ELBO)

- Our goal was to find the parameters ψ of the approximating distribution $q_{\Psi}(\theta)$ such that it is as close as possible to the true posterior distribution $P(\theta|D)$.

Evidence Lower Bound (ELBO)

- Our goal was to find the parameters ψ of the approximating distribution $q_{\Psi}(\theta)$ such that it is as close as possible to the true posterior distribution $P(\theta|D)$.
- $\Psi^* = \arg\min_{\Psi} \mathrm{KL}(q_{\Psi}(\theta)||P(\theta|D))$

Evidence Lower Bound (ELBO)

- Our goal was to find the parameters ψ of the approximating distribution $q_\Psi(\theta)$ such that it is as close as possible to the true posterior distribution $P(\theta|D)$.
- $\Psi^* = \arg\min_\Psi \mathrm{KL}(q_\Psi(\theta)||P(\theta|D))$
- But, we do not know the true posterior distribution $P(\theta|D)$.

Evidence Lower Bound (ELBO)

- Our goal was to find the parameters ψ of the approximating distribution $q_\Psi(\theta)$ such that it is as close as possible to the true posterior distribution $P(\theta|D)$.
- $\Psi^* = \arg\min_\Psi \mathrm{KL}(q_\Psi(\theta)||P(\theta|D))$
- But, we do not know the true posterior distribution $P(\theta|D)$.
- Let us focus on the KL divergence term:

Evidence Lower Bound (ELBO)

- Our goal was to find the parameters ψ of the approximating distribution $q_\Psi(\theta)$ such that it is as close as possible to the true posterior distribution $P(\theta|D)$.
- $\Psi^* = \arg\min_\Psi \mathrm{KL}(q_\Psi(\theta)||P(\theta|D))$
- But, we do not know the true posterior distribution $P(\theta|D)$.
- Let us focus on the KL divergence term:
- $\mathrm{KL}(q_\Psi(\theta)||P(\theta|D)) = \mathbb{E}_{q_\Psi(\theta)}\left[\log \frac{q_\Psi(\theta)}{P(\theta|D)}\right]$

Evidence Lower Bound (ELBO)

- Our goal was to find the parameters ψ of the approximating distribution $q_\Psi(\theta)$ such that it is as close as possible to the true posterior distribution $P(\theta|D)$.
- $\Psi^* = \arg\min_\Psi \mathrm{KL}(q_\Psi(\theta)||P(\theta|D))$
- But, we do not know the true posterior distribution $P(\theta|D)$.
- Let us focus on the KL divergence term:
- $\mathrm{KL}(q_\Psi(\theta)||P(\theta|D)) = \mathbb{E}_{q_\Psi(\theta)}\left[\log \frac{q_\Psi(\theta)}{P(\theta|D)}\right]$
- Using Bayes rule, we write: $P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$

ELBO

- $\mathsf{KL}(q_\Psi(\theta)||P(\theta|D)) = \mathbb{E}_{q_\Psi(\theta)}\left[\log \frac{q_\Psi(\theta)}{P(\theta|D)}\right]$
- Using Bayes rule, we write: $P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$

ELBO

- $\mathsf{KL}(q_\Psi(\theta)||P(\theta|D)) = \mathbb{E}_{q_\Psi(\theta)}\left[\log \frac{q_\Psi(\theta)}{P(\theta|D)}\right]$
- Using Bayes rule, we write: $P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$
- Substituting, we get:
  $\mathsf{KL}(q_\Psi(\theta)||P(\theta|D)) = \mathbb{E}_{q_\Psi(\theta)}\left[\log \frac{q_\Psi(\theta)P(D)}{P(D|\theta)P(\theta)}\right]$

ELBO

- $\mathrm{KL}(q_\Psi(\theta)||P(\theta|D)) = \mathbb{E}_{q_\Psi(\theta)}\left[\log \frac{q_\Psi(\theta)}{P(\theta|D)}\right]$
- Using Bayes rule, we write: $P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$
- Substituting, we get:
  $\mathrm{KL}(q_\Psi(\theta)||P(\theta|D)) = \mathbb{E}_{q_\Psi(\theta)}\left[\log \frac{q_\Psi(\theta)P(D)}{P(D|\theta)P(\theta)}\right]$
- Expanding, we get:
  $\mathrm{KL}(q_\Psi(\theta)||P(\theta|D)) = \mathbb{E}_{q_\Psi(\theta)}\left[\log \frac{q_\Psi(\theta)}{P(\theta)P(D|\theta)} + \log P(D)\right]$

ELBO

- $\mathsf{KL}(q_\Psi(\theta)||P(\theta|D)) = \mathbb{E}_{q_\Psi(\theta)}\left[\log \frac{q_\Psi(\theta)}{P(\theta|D)}\right]$
- Using Bayes rule, we write: $P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$
- Substituting, we get:
  $\mathsf{KL}(q_\Psi(\theta)||P(\theta|D)) = \mathbb{E}_{q_\Psi(\theta)}\left[\log \frac{q_\Psi(\theta)P(D)}{P(D|\theta)P(\theta)}\right]$
- Expanding, we get:
  $\mathsf{KL}(q_\Psi(\theta)||P(\theta|D)) = \mathbb{E}_{q_\Psi(\theta)}\left[\log \frac{q_\Psi(\theta)}{P(\theta)P(D|\theta)} + \log P(D)\right]$
- As log-evidence, $\log P(D)$ is independent of Ψ, we get:
  $\mathsf{KL}(q_\Psi(\theta)||P(\theta|D)) = \mathbb{E}_{q_\Psi(\theta)}\left[\log \frac{q_\Psi(\theta)}{P(\theta)P(D|\theta)}\right] + \log P(D)$

ELBO

- $\text{KL}(q_\Psi(\theta)||P(\theta|D)) = \mathbb{E}_{q_\Psi(\theta)}\left[\log \frac{q_\Psi(\theta)}{P(\theta)P(D|\theta)}\right] + \log P(D)$

ELBO

- $\mathrm{KL}(q_\Psi(\theta)||P(\theta|D)) = \mathbb{E}_{q_\Psi(\theta)}\left[\log \frac{q_\Psi(\theta)}{P(\theta)P(D|\theta)}\right] + \log P(D)$
- $\underbrace{\mathrm{KL}(q_\Psi(\theta)||P(\theta|D))}_{\geq 0} = \mathbb{E}_{q_\Psi(\theta)}\left[\log \frac{q_\Psi(\theta)}{P(\theta)P(D|\theta)}\right] + \log P(D)$

ELBO

- $\mathsf{KL}(q_\Psi(\theta)||P(\theta|D)) = \mathbb{E}_{q_\Psi(\theta)}\left[\log \frac{q_\Psi(\theta)}{P(\theta)P(D|\theta)}\right] + \log P(D)$
- $\underbrace{\mathsf{KL}(q_\Psi(\theta)||P(\theta|D))}_{\geq 0} = \mathbb{E}_{q_\Psi(\theta)}\left[\log \frac{q_\Psi(\theta)}{P(\theta)P(D|\theta)}\right] + \log P(D)$
- Let us call the term $\mathbb{E}_{q_\Psi(\theta)}\left[\log \frac{q_\Psi(\theta)}{P(\theta)P(D|\theta)}\right]$ as -**ELBO(q)**

ELBO

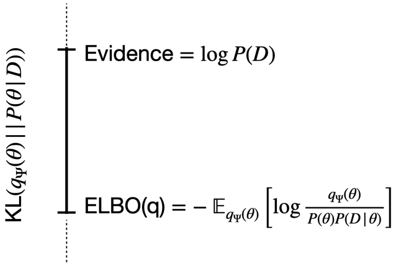
- $\mathrm{KL}(q_\Psi(\theta)||P(\theta|D)) = \mathbb{E}_{q_\Psi(\theta)}\left[\log \frac{q_\Psi(\theta)}{P(\theta)P(D|\theta)}\right] + \log P(D)$
- $\underbrace{\mathrm{KL}(q_\Psi(\theta)||P(\theta|D))}_{\geq 0} = \mathbb{E}_{q_\Psi(\theta)}\left[\log \frac{q_\Psi(\theta)}{P(\theta)P(D|\theta)}\right] + \log P(D)$
- Let us call the term $\mathbb{E}_{q_\Psi(\theta)}\left[\log \frac{q_\Psi(\theta)}{P(\theta)P(D|\theta)}\right]$ as -**ELBO(q)**
- We can see that log-evidence or $\log P(D) \geq$ ELBO(q)

ELBO

- $\mathrm{KL}(q_\Psi(\theta)||P(\theta|D)) = \mathbb{E}_{q_\Psi(\theta)}\left[\log \frac{q_\Psi(\theta)}{P(\theta)P(D|\theta)}\right] + \log P(D)$
- $\underbrace{\mathrm{KL}(q_\Psi(\theta)||P(\theta|D))}_{\geq 0} = \mathbb{E}_{q_\Psi(\theta)}\left[\log \frac{q_\Psi(\theta)}{P(\theta)P(D|\theta)}\right] + \log P(D)$
- Let us call the term $\mathbb{E}_{q_\Psi(\theta)}\left[\log \frac{q_\Psi(\theta)}{P(\theta)P(D|\theta)}\right]$ as -**ELBO(q)**
- We can see that log-evidence or $\log P(D) \geq \mathrm{ELBO(q)}$
- Or, we can see that the **evidence** term is lower bounded by the **ELBO** term

ELBO

- $\text{KL}(q_\Psi(\theta)||P(\theta|D)) = \mathbb{E}_{q_\Psi(\theta)}\left[\log \frac{q_\Psi(\theta)}{P(\theta)P(D|\theta)}\right] + \log P(D)$
- $\underbrace{\text{KL}(q_\Psi(\theta)||P(\theta|D))}_{\geq 0} = \mathbb{E}_{q_\Psi(\theta)}\left[\log \frac{q_\Psi(\theta)}{P(\theta)P(D|\theta)}\right] + \log P(D)$
- Let us call the term $\mathbb{E}_{q_\Psi(\theta)}\left[\log \frac{q_\Psi(\theta)}{P(\theta)P(D|\theta)}\right]$ as -**ELBO(q)**
- We can see that log-evidence or $\log P(D) \geq$ ELBO(q)
- Or, we can see that the **evidence** term is lower bounded by the **ELBO** term



Worked out example: Coin Toss

Worked out example: Linear Regression

Worked out example: Neural Networks