Kullback-Leibler divergence

Dripta Mj
Department of Mathematics
Ramakrishna Mission Vivekananda Educational and Research Institute
Belur Math, India

Introduction

- Kullback-Leibler (KL) divergence is a measure of how a probability distribution is different from another one.
- Suppose we have two distributions $p(x)$ and $q(x)$.
- KL divergence between $p(x)$ and $q(x)$ is defined as

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

  - For discrete variables:

$$KL(p||q) = \sum_{x \in \chi} p(x) \log \left(\frac{p(x)}{q(x)} \right)$$

  - For continuous variables:

$$KL(p||q) = \int p(x) \log \left(\frac{p(x)}{q(x)} \right) \mathrm{d}x$$

Properties

- KL divergence is not symmetric, i.e. $KL(p||q) \neq KL(q||p)$.
  - Therefore it is not a distance metric.
- $KL(p||q) \geq 0$.
- $KL(p||q) = 0$ indicates that distributions p and q are identical.
- For KL divergence to be finite, the support of p needs to be contained in the support of q.
  - If $q(x) = 0$ but $p(x) > 0$, then $KL(p||q) = \infty$.

KL divergence: entropy and cross-entropy terms

- Can also express the KL divergence as

$$
\begin{aligned}
KL(p||q) &= \mathbb{E}_{x\sim p}\Big[\log\Big(\frac{p(x)}{q(x)}\Big)\Big] \\
&= \mathbb{E}_{x\sim p}\Big[\log p(x)\Big] - \mathbb{E}_{x\sim p}\Big[\log q(x)\Big] \\
&= -\mathbb{E}_{x\sim p}\Big[-\log p(x)\Big] + \mathbb{E}_{x\sim p}\Big[-\log q(x)\Big] \\
&= -H(p(x)) + H(p(x), q(x))
\end{aligned}
$$

where $H(p(x))$ is the entropy of $p(x)$ and $H(p(x), q(x))$ is the cross-entropy between distributions $p(x)$ and $q(x)$.

Two possibilities

- Suppose we have some true distribution $p(x)$, and we want to estimate using some approximate distribution $q_{\boldsymbol{\theta}}(x)$.
  - Here $\boldsymbol{\theta}$ indicate the parameters of the distribution q.

- There are two possibilities to minimize the divergence:

$$\arg\min_{\boldsymbol{\theta}} KL(p||q_{\boldsymbol{\theta}})$$

  - $KL(p||q_{\boldsymbol{\theta}})$ is known as forward KL divergence

$$\arg\min_{\boldsymbol{\theta}} KL(q_{\boldsymbol{\theta}}||p)$$

  - $KL(q_{\boldsymbol{\theta}}||p)$ is known as reverse KL divergence

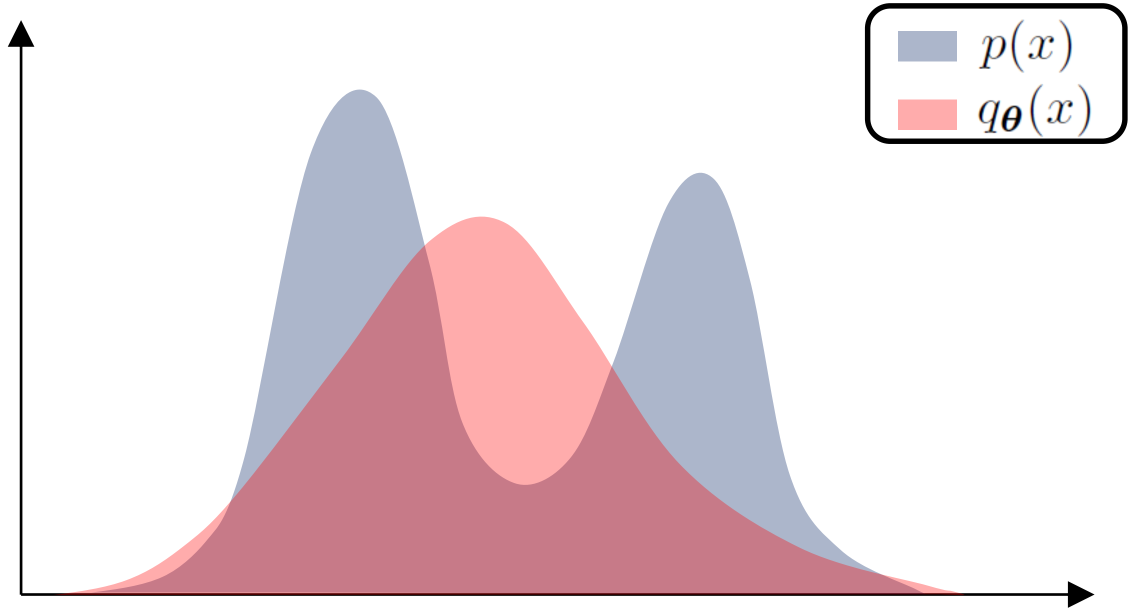
Forward KL divergence

$$\begin{aligned}\arg\min_{\boldsymbol{\theta}} KL(p||q) &= \arg\min_{\boldsymbol{\theta}} -H(p(x)) + \mathbb{E}_{x\sim p}\Big[-\log q(x)\Big] \\ &= \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{x\sim p}\Big[\log q(x)\Big]\end{aligned}$$

- The above objective implies that it will sample points from $p(x)$ and then try to maximize the probability of the sampled points under $q(x)$.
  - Therefore the objective wants to achieve a high probability for $q(x)$ wherever $p(x)$ has a high probability.

- The approximate distribution $q(x)$ tries to cover all modes and regions of high probability in $p(x)$.
  - This is often referred to as the mean-seeking behaviour.

Forward KL divergence: Example



- Consider the case where $p(x)$ is a bimodal distribution.
- Suppose we want to approximate this using a normal distribution $q(x) = \mathcal{N}(\mu, \sigma^2)$.
- The optimal $q(x)$ centers between two modes such that it achieves high coverage for both of them.

Figure for illustration only

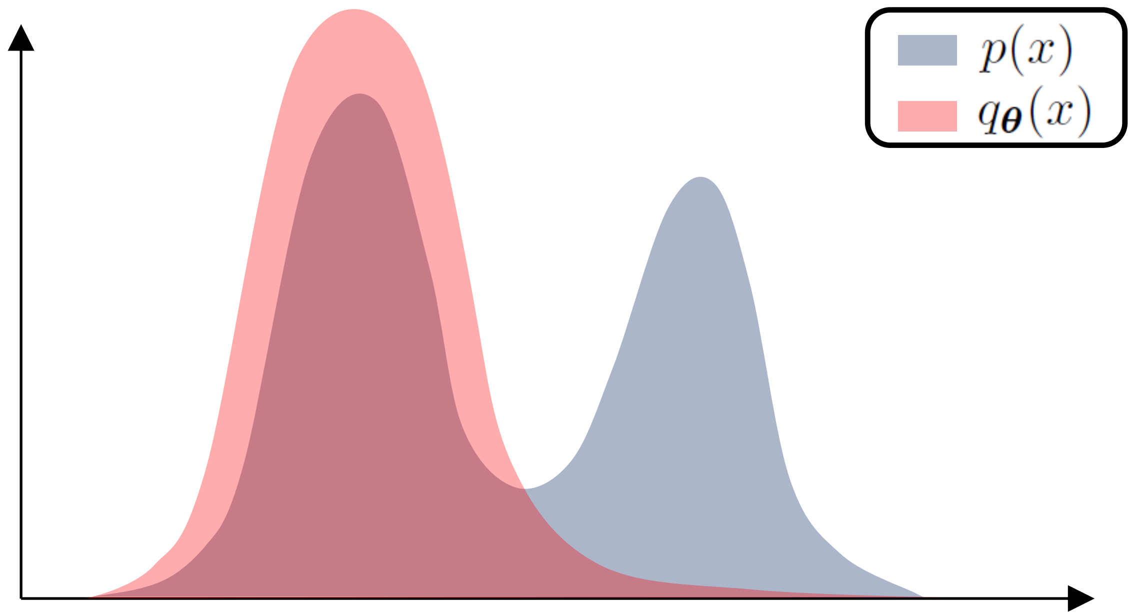
Reverse KL divergence

$$\arg\min_{\boldsymbol{\theta}} KL(q||p) = \arg\min_{\boldsymbol{\theta}} -H(q_{\boldsymbol{\theta}}(x)) + \mathbb{E}_{x\sim q_{\boldsymbol{\theta}}}\Big[-\log p(x)\Big]$$

$$= \arg\max_{\boldsymbol{\theta}} H(q_{\boldsymbol{\theta}}(x)) + \mathbb{E}_{x\sim q_{\boldsymbol{\theta}}}\Big[\log p(x)\Big]$$

- The objective samples points from $q_{\boldsymbol{\theta}}(x))$ and tries to maximize their probability under $p(x)$
- The first term in the objective is the entropy of $q_{\boldsymbol{\theta}}(x))$
  - The entropy term attempts to make $q_{\boldsymbol{\theta}}(x)$ as wide as possible.
- Thus samples from q are required to have a high probability under p, but the entropy term prevents q from collapsing to a narrow mode.

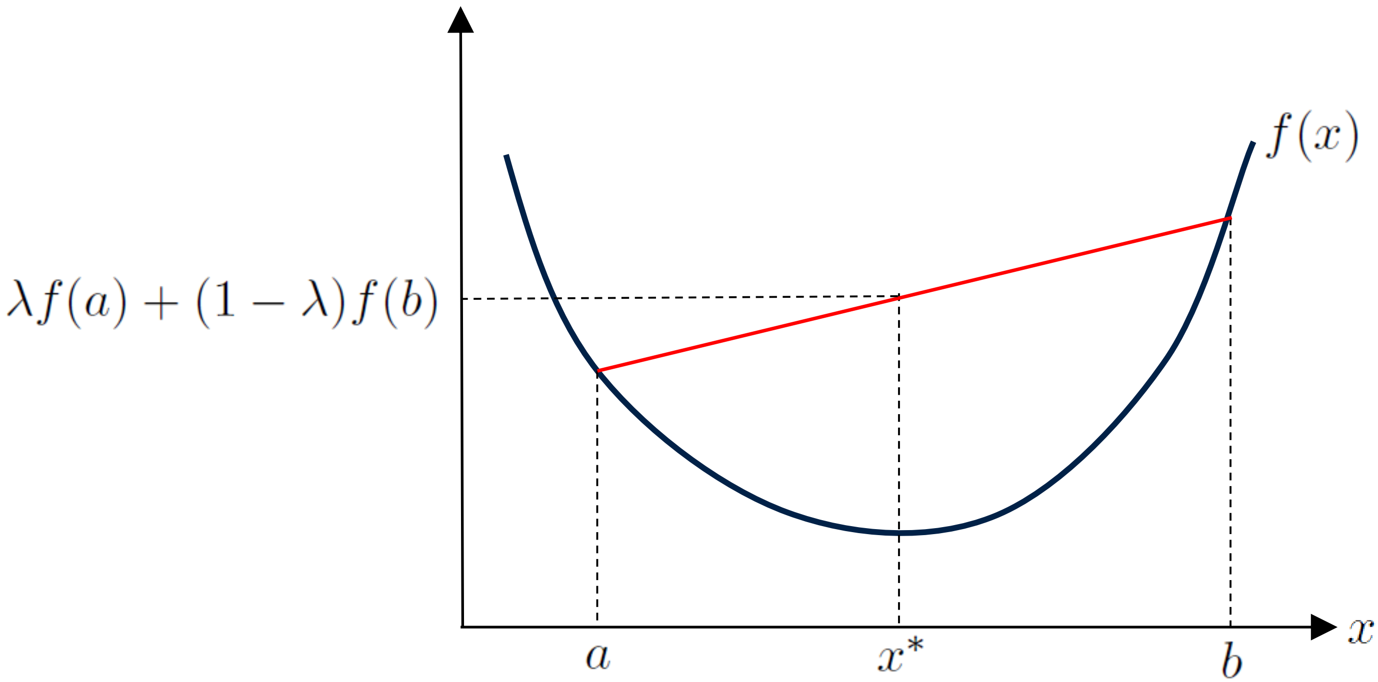
Reverse KL divergence: Example



- This is called "mode-seeking" behaviour

Figure for illustration only

Proof: KL(p||q) ≥ 0

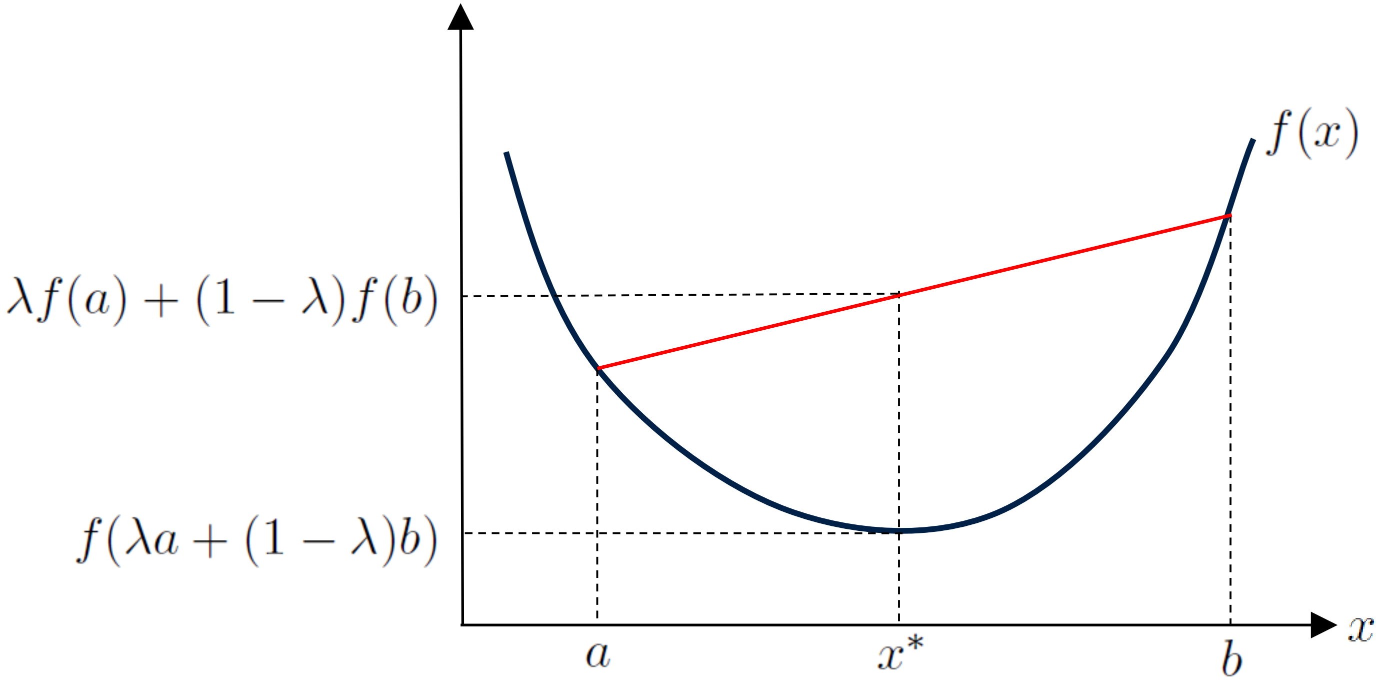


- Consider a convex function $f(x)$
- The value of x at any any point in the interval $a \leq x \leq b$ can be written as

$$x^* = \lambda a + (1-\lambda)b, \text{ where } 0 \leq \lambda \leq 1.$$

- The corresponding point on the chord connecting $f(a)$ with $f(b)$ can be written as $\lambda f(a) + (1-\lambda) f(b)$.

Proof: KL(p||q) ≥ 0



- The value of the function at x^* is $f(\lambda a + (1-\lambda)b)$.
- From convexity we have

$$f\big(\lambda a + (1-\lambda)b\big) \leq \lambda f(a) + (1-\lambda) f(b)$$

Proof: KL(p||q) ≥ 0

- For a convex function, Jensen's inequality states that

$$f\Big(\sum_{m=1}^{M} \lambda_m x_m\Big) \leq \sum_{m=1}^{M} \lambda_m f(x_m)$$

where $\lambda_m \geq 0$ and $\sum_{m=1}^{M} \lambda_m = 1$.

- λ_m can be interpreted as a probability distribution. In that case the above inequality can be rewritten as

$$f\Big(\mathbb{E}[x]\Big) \leq \mathbb{E}[f(x)]$$

- For continuous variables, the inequality is expressed as

$$f\Big(\int x p(x) \mathrm{d}x\Big) \leq \int f(x) p(x) \mathrm{d}x$$

Proof: KL(p||q) ≥ 0

- The KL divergence can then be expressed as

$$
\begin{aligned}
KL(p||q) &= -\int p(x)\log\left(\frac{q(x)}{p(x)}\right)\mathrm{d}x \\
&\geq -\log\left(\int p(x)\frac{q(x)}{p(x)}\mathrm{d}x\right) \\
&\geq -\log\left(\int q(x)\mathrm{d}x\right) \\
&\geq 0
\end{aligned}
$$

Jensen-Shannon divergence

- KL divergence is asymmetric and as such is not a distance metric.
- Jensen-Shannon divergence is a symmetric version of the KL divergence.
- It is defined as

$$JS(p||q) = \frac{1}{2}\Big(KL(p||m) + KL(q||m)\Big)$$

where

$$m = \frac{1}{2}(p + q)$$