Kullback - Leibler divergence

Short talk - 04/10/2021

KL expression

Solomon Kullback

Richard Leibler

- Discrete version:

$$D_{KL}(p \| q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$$

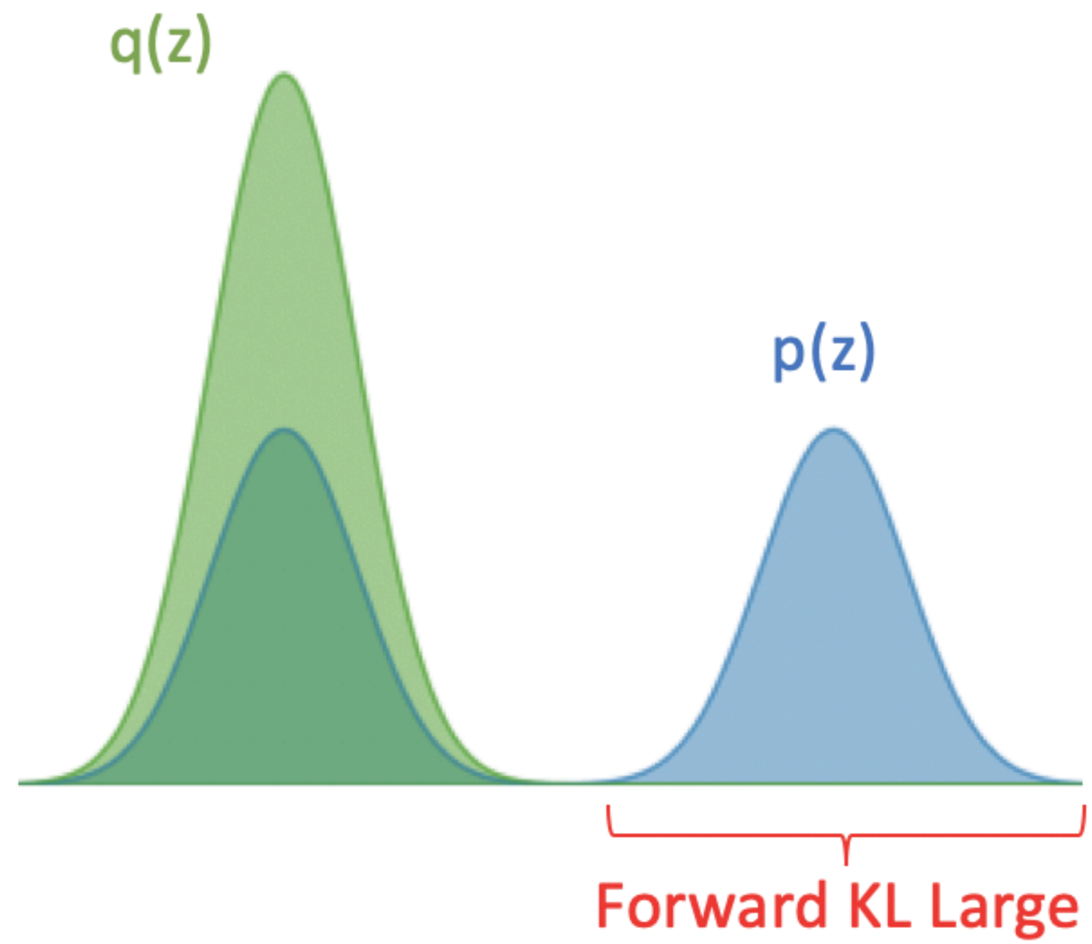
- Continuous version:

$$D_{KL}(p \| q) = \int_{-\infty}^{+\infty} p(x) \log \frac{p(x)}{q(x)} dx$$

If there exists $x \in \mathcal{X}$ **such that** $q(x) = 0$ **and** $p(x) \neq 0$ **then** $D(p \| q) = +\infty$

$$\log \frac{p(x)}{q(x)}$$

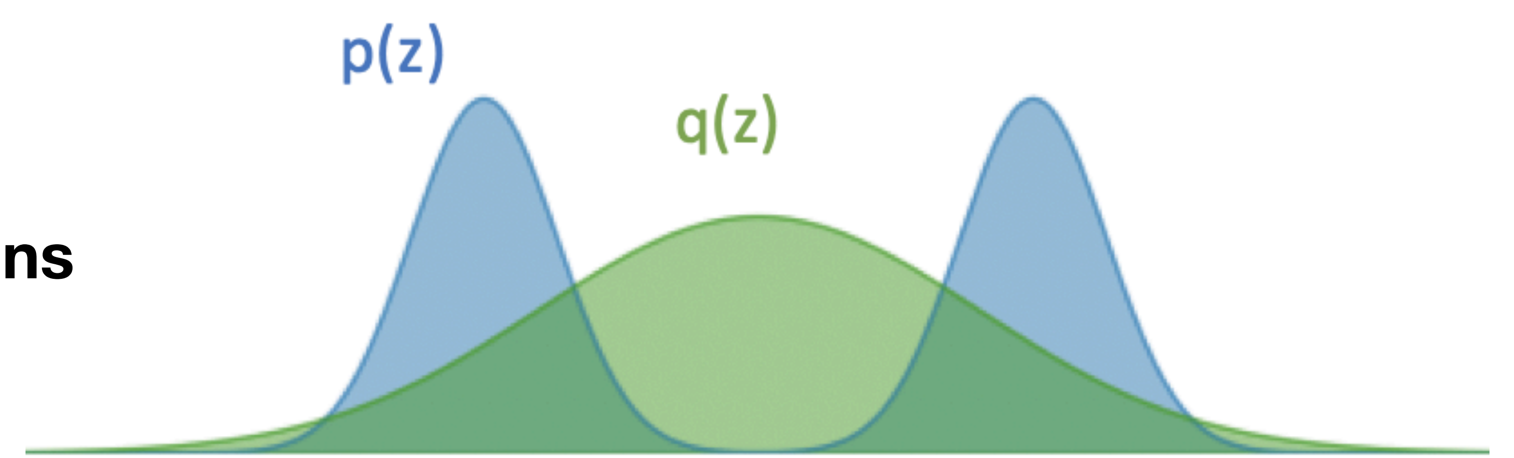
- Positive if $p(x) \geq q(x)$
- Null if $p(x) = q(x)$
- Negative if $q(x) \geq p(x)$



Penalties are weighted by p :

$$D_{KL}(p \| q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$$

Keep in mind that the above functions are probability densities:



$$\sum_{x \in \mathcal{X}} p(x) = 1$$ **and** $$\sum_{x \in \mathcal{X}} q(x) = 1$$

$D_{KL}(p\|q) \geq 0$ **and equality holds if** $p = q$

$$-D_{KL}(p\|q) = -\sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)} = \sum_{x\in\mathcal{X}} p(x)\log\frac{q(x)}{p(x)}$$

$$\leq \log\sum_{x\in\mathcal{X}} p(x)\frac{q(x)}{p(x)} = 0$$ **(Jensen inequality)**

$D_{KL}(p\|q) = 0$ **if equality in Jensen inequality i.e.**

$p = cq \rightarrow p = q$ **since** p **and** q **sum to** 1.

Definition of a distance

Any function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ such that :

- $\forall (a,b) \in \mathcal{X}^2, \quad d(a,b) = 0 \iff a = b$
- $\forall (a,b) \in \mathcal{X}^2, \quad d(a,b) = d(b,a)$
- $\forall (a,b,c) \in \mathcal{X}^3, \quad d(a,c) \leq d(a,b) + d(b,c)$

$D_{KL}(p \| q)$ is not a distance !

Information theory intuition :

$$D_{KL}(p\|q) = \underbrace{\sum_{x\in\mathcal{X}} p(x)\log p(x)}_{H(p)} - \underbrace{\sum_{x\in\mathcal{X}} p(x)\log q(x)}_{H_p(q)}$$

If we knew the true distribution p of the random variable, we could construct a code with average description length $H(p)$.

If, instead, we used the code for a distribution q, we would need $H(p) + D_{KL}(p\|q)$ bits on the average to describe the random variable.

Let $x_1, \ldots, x_N \in \mathcal{X}$ be N i.i.d. observations of a random variable X

$$\hat{p}(x) = \frac{1}{N} \sum_{i=1}^{N} \delta(x - x_i)$$

Let p_θ be a parameterized distribution on $\mathcal{X}$

$$D_{KL}(\hat{p} \| p_\theta) = \sum_{x \in \mathcal{X}} \hat{p}(x) \log \frac{\hat{p}(x)}{p_\theta(x)} = -H(\hat{p}) - \sum_{x \in \mathcal{X}} \hat{p}(x) \log p_\theta(x)$$

$$= -H(\hat{p}) - \sum_{x \in \mathcal{X}} \sum_{i=1}^{N} \delta(x - x_i) \log p_\theta(x) = -H(\hat{p}) - \frac{1}{N} \sum_{i=1}^{N} \log p_\theta(x_i)$$

Maximizing the likelihood $p_\theta(x)$ $\iff$ Minimizing $D_{KL}(\hat{p} \| p_\theta)$

Pinsker's inequality

Total variation distance

$$\delta(p, q) = \sup_{A \in \mathcal{F}} |\, p(A) - q(A)\, |$$

$$\delta(p, q) \leq \sqrt{\frac{1}{2} D_{KL}(p \| q)}$$

In practice

Let x_i **be samples from** $p(x)$:

$$\lim_{N \to +\infty} \frac{1}{N} \sum_{i=0}^{N} \log \frac{p(x_i)}{q(x_i)} = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} = D_{KL}(p \| q)$$