Source: https://www.youtube.com/watch?v=9_eZHt2qJs4&t=16s

KL divergence is measure how one probability distribution is different from second, reference distribution.

X is random variable and small x-s are the states it could take: $X = \{x_1, x_2, x_3, \ldots, x_n\}$

We want to compare different probability distribution $\log p_\theta x_1$ with some other probability distribution $\log q_\phi x_1$. They dont have to be same type of distributions (usually q is simpler to use distribution for modelling)

One way to calculate difference is use subtraction, which because we are using logs is division:

$$\log p_\theta x_1 - \log q_\phi x_1 = \log[\frac{p_\theta x_1}{q_\phi x_1}]$$

This division is called log likelihood ratio. But currently we are calculatiing difference between one sample. We would like to calculate average difference between p and q. In random variables we say what is expected value of the random vairable or what is its central tendency.

For random variables exptected value is weighted average of instances of random variable. Each variable and all states have probability of occurence and average should reflect that. Samples that have higher probability should contribute more to the average.

$$\mathbb{E}_{p_\theta}[X] = \sum_{i=1}^{\infty} x_i p_\theta(x_i)$$

where:

- x_i is state of the random variable
- $p_\theta(x_i)$ is weight of the random variable

Here we showed that calculation should be done for very large number of samples

Here is same formula more in general, instead of random variable we calculate weighted average of a function of random variables

$$\mathbb{E}_{p_\theta}[h(X)] = \sum_{i=1}^{\infty} h(x_i) p_\theta(x_i)$$

So far we have looked at discrete random variables. For continous random variable we would calculate exptected value as such (summation has been replaced by integral):

$$\mathbb{E}_{p_\theta}[h(X)] = \int_{\mathbb{R}} h(x_i) p_\theta(x_i)$$

Our log likelihood ratio is nothig but a function of random variable and since we are interested in average of this function we should be able to use expectation. So we need weights and compute the sums:

$$\sum_{i=1}^{\infty} p_\theta(x_i) \log[\frac{p_\theta x_i}{q_\phi x_i}]$$

To be exact we are calculating exptected value of log likelihood ratio. This is **KL divergence**. We can express it using expectation symbol:

$$\mathbb{E}_{\mathrm{p}}[\log[\frac{p_\theta x_i}{q_\phi x_i}] = \sum_{i=i}^{\infty} p_\theta(x_i) \log[\frac{p_\theta x_i}{q_\phi x_i}]$$

Previous example was for discrete random variable for continuous random variable we replace summation with an integral:

$$\mathbb{E}_{\mathrm{p}}[\log[\frac{p_\theta x_i}{q_\phi x_i}] = \int_{\mathbb{R}} p_\theta(x_i) \log[\frac{p_\theta x_i}{q_\phi x_i}]$$

We have one problem. Integral and summation both go from minus infinity to infinity. We could get help from the **law of large numbers**: "as a sample size grows, its mean gets closer to the average of the whole population" (https://www.investopedia.com/terms/l/lawoflargenumbers.asp). So we can revrite KL divergence as a mean provided we use many samples:

$$\frac{1}{N} \sum_{i=1}^{N} \log[\frac{p_\theta x_i}{q_\phi x_i}]$$

Antoher notation used is:

$$DD_{KL}(p_\theta||q_\phi) = \int_{\mathbb{R}} p_\theta(x) log[\frac{p_\theta(x)}{q_\phi(x)}]dx$$

This on is called **forward KL**

If we would like to use q instead of p for weighting:

$$D_{KL}(q_\phi||p_\theta) = \int_{\mathbb{R}} q_\phi(x) log[\frac{q_\phi(x)}{p_\theta(x)}]dx$$

This on is called **reverse KL**.

These formulations are going to give different values. This is a reason why this is not called metric, but a distance.

Generally we use p for reference distribution and q for approximation.

Which one to use? It depends. Forward KL has mean-seeking behavior as reverse KL has mode seeking behavior.

 image

Here we can see that revrse has picked up mean mode of the distribution. Most of the time when we are doing density estimation and using variatinal inference we use reverse scale. Forward KL is being used a lot in machine learning but you don't see name forward KL a lot. It is being used indirectly. For example when we are using cross-entropy loss in classificatin loss we are using KL indirectly (at least a component of KL divergence).

In [ ]: