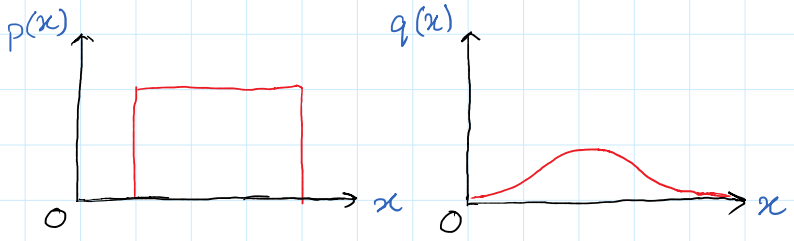
KL Divergence:

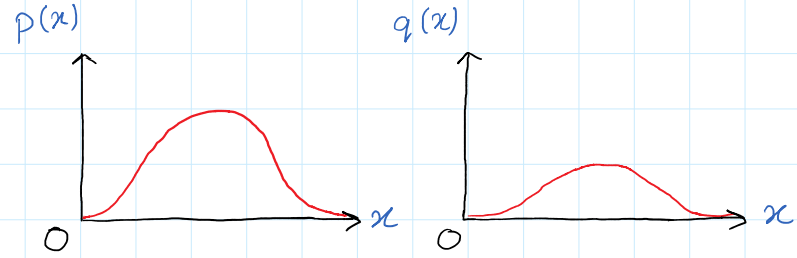
KL Divergence is used to quantify the dissimilarity / difference between probability distributions. It's used in Variational Auto Encoders (VAEs).
Thus, KL Divergence will take a very high value if the two probability distributions are quite different from one another and will take a very small value if they are quite similar to each other.

Eg:



KL Divergence on p(x) and q(x) will have a high value.

Eg:



KL Divergence on p(x) and q(x) will have a lower value.

MATHEMATICAL FORMULATION:

Suppose you have a Discrete Random Variable X. Now, in reality X will be characterized by a single Probability Mass Function, say $p_x(x)$ which is basically $P(X = x)$.

However, practically, you will always have a plethora of 'candidate PMFs' that will be competing for being the best fit PMF to X, thus to compare two different distributions on the same random variable X, we use a divergence

metric called Kullback - Leibler Divergence (KL Divergence).

NOTE: KL Divergence does not represent the distance between two distributions because $D_{KL}(p||q) \neq D_{KL}(q||p)$ and thus one distribution is always being compared with the other for finding the divergence.

Formula -: $$\boxed{D_{KL}(p||q) = E\left[\log\left(\frac{p(x)}{q(x)}\right)\right]} = E\left[\log(\underbrace{p(x)}_{\text{PMF}}) - \log(\underbrace{q(x)}_{\text{PMF}})\right]$$

For Discrete R.V X: $$D_{KL}(p||q) = \sum_{x \in R(X)} p(x) \log\left(\frac{(x)}{(x)}\right)$$

For Cont. R.V. X: $$D_{KL}(p||q) = \int_{-\infty}^{+\infty} p(x) \log\left(\frac{p(x)}{q(x)}\right) dx$$

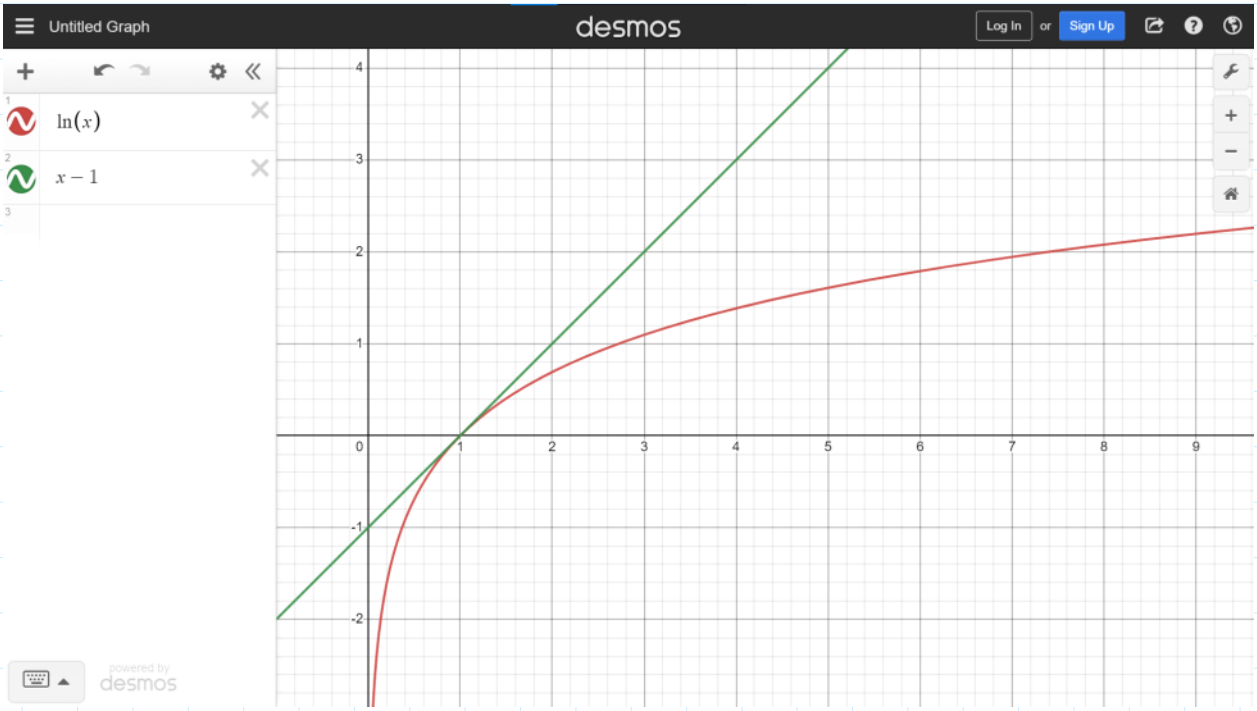
Properties:

1. KL Divergence for two distributions p and q is always non-negative. (V. V. Imp.)

$$D_{KL}(p||q) \geq 0$$

Proof:

We state that, $\ln(c) \leq c - 1 \quad \forall\ c > 0$. — (1)



Now, to show that $D_{KL}(p||q) \geq 0$, we will prove that $-D_{KL}(p||q) \leq 0$.

$$-D_{KL}(p||q) = -\sum_{x \in R(x)} p(x) \ln\left(\frac{p(x)}{q(x)}\right)$$

$$= \sum_{x \in R(x)} (x) \ln\left(\frac{p(x)}{q(x)}\right) \overset{c}{\leq} \sum_{x \in R(x)} p(x)\left(\frac{q(x)}{p(x)} - 1\right)$$

$$= \sum_{x \in R(x)} q(x) - p(x)$$

$$= \sum_{x \in R(x)} q(x) - \sum_{x \in R(x)} p(x)$$

$$= 1 - 1 = \boxed{0}$$

$\therefore D_{KL}(p||q) \geq 0$

2. KL Divergence is not symmetric, i.e.

$$D_{KL}(p||q) \neq D_{KL}(p||q)$$

Types of KL Divergence :

1. Forward KL Divergence : Denoted by $D_{KL}(p||q)$ where p and q are distribution functions on the same R.V.

2. Reverse KL Divergence : Denoted by $D_{KL}(q||p)$ where p and q are distribution functions on the same R.V.

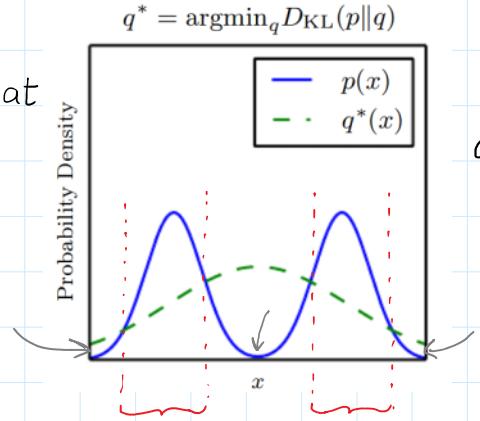
Question : If you want to approximate a distribution $p(x)$ with another distribution $q(x)$, which type of KL Divergence will you minimize ?

Ans : The behavior of $D_{KL}(p||q)$ and $D_{KL}(q||p)$ differ when we try to minimize them.

When $q^*(x) = \underset{q}{\text{argmin}}\ D_{KL}(p||q)$, the effect of minimizing $D_{KL}(p||q)$ is that we try to ensure that $q^*(x)$ is high wherever $p(x)$ is high. This causes $q(x)$ to get streched.

Note that this causes $q^*(x)$ to avoid being equal to 0 at points where $p(x)$ is high. Thus, this is also called zero-avoiding / mean-seeking.

→ : The arrows show that $q^*(x)$ doesn't care much about point where $p(x)$ is close to 0.



Trying to ensure that $q^*(x)$ remains high for high values of $p(x)$.

$$q^* = \underset{q}{\text{argmin}} \left[p(x) \log\left(\frac{p(x)}{q(x)}\right) \right]$$

When $p(x) \approx 0$, $p(x) \log\left[\frac{p(x)}{q(x)}\right]$ is already a small number

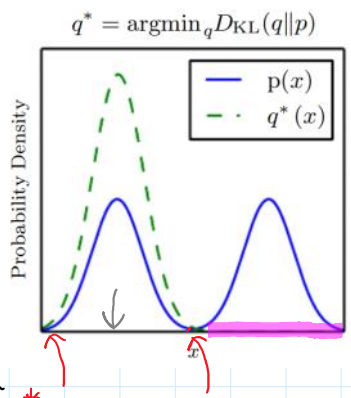
When $p(x) \approx 0$, $p(x) \log\left[\frac{p(x)}{q(x)}\right]$ is already a small number

When $p(x)$ is high, $p(x) \log\left[\frac{p(x)}{q(x)}\right]$ is minimized by making $q(x)$ high.

When $q^*(x) = \underset{q}{\operatorname{argmin}}\ D_{KL}(q\|p)$, the effect of minimizing $D_{KL}(q\|p)$ is that we try to ensure that $q^*(x)$ is low wherever $p(x)$ is low. This causes $q(x)$ to fit to a single mode of $p(x)$.
Note that this causes $q^*(x)$ to avoid having high value at points where $p(x)$ is low. Thus, this is also called zero-avoiding / mean-seeking.

Note: $q^*(x)$ has become a single mode approximation of $p(x)$.

→: Don't really care about what q^* should be here.



$q^*(x)$ is low where $p(x)$ is low.

$$q^* = \underset{q}{\operatorname{argmin}} \left[q(x) \log\left(\frac{q(x)}{p(x)}\right) \right]$$

When $p(x) \approx 0$, $q(x) \log\left[\frac{q(x)}{p(x)}\right]$ can only be minimized by making $q(x) \approx 0$ which would result in $0 \cdot \log\left(\frac{0}{0}\right) = 0$.

For the rest of the part where $q^*(x) = 0$ (highlighted in pink), we are still minimizing the loss here as $0 \cdot \log(0) = 0$.

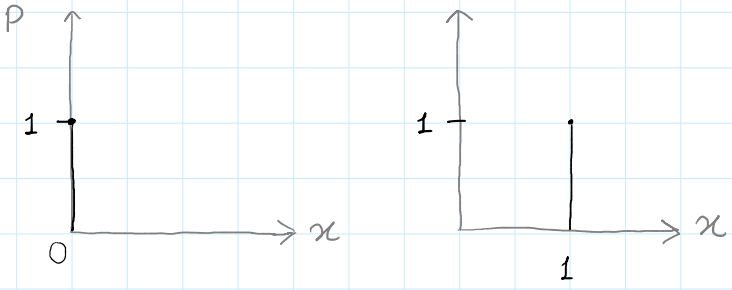
When p has multiple modes (as shown above), $q^*(x)$ will converge to a single mode following the Reverse D_{KL} while maintaining its own distribution shape.

Thus, $D_{KL}(p||q)$ over-estimates the base distribution whereas $D_{KL}(q||p)$ under-estimates the base distribution.

We generally prefer to use Reverse KL Divergence.

PROBLEMS WITH KL DIVERGENCE:

If p and q don't overlap at all and are separated to a large extent (which can happen during the early training stages),



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

$$= p(0) \log\left(\frac{p(0)}{q(0)}\right)$$

$$= 1 \log\left(\frac{1}{0}\right) = \infty$$

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

$$= q(1) \log\left(\frac{q(1)}{p(1)}\right)$$

$$= 1 \log\left(\frac{1}{0}\right) = \infty$$

Thus, both forward and reverse KL Divergences blow up when $p(x)$ and $q(x)$ don't overlap at any x and it's very common to find $p(x)$ and $q(x)$ not overlapping each

other when trying to approximate $p(x)$.

SOLUTION TO THE PROBLEM :

Jensen Shannon Divergence (JSD) is the method used to measure the similarity between two probability distributions.

$$JSD(P||q) = \frac{1}{2}\left[D_{KL}(P||m) + D_{KL}(q||m)\right]$$

where $m = \frac{P+q}{2}$

JSD is used in GANs and thus allows us to avoid the problem mentioned before.

For the example used before,

$$JSD(P||q) = \frac{1}{2}\left[p(0)\log\left(\frac{p(0)}{\frac{p(0)+q(0)}{2}}\right)\right] + \frac{1}{2}\left[q(.5)\log\left(\frac{q(.5)}{\frac{p(.5)+q(.5)}{2}}\right)\right]$$

$$JSD(P||q) = \frac{1}{2}\left[1\cdot\log\left(\frac{1}{0.5}\right)\right] + \frac{1}{2}\left[1\cdot\log\left(\frac{1}{0.5}\right)\right]$$

$$\boxed{JSD(P||q) = \log(2)}$$ which is finite.