**Kullback Leibler Divergence -** a distance measure between probability distributions

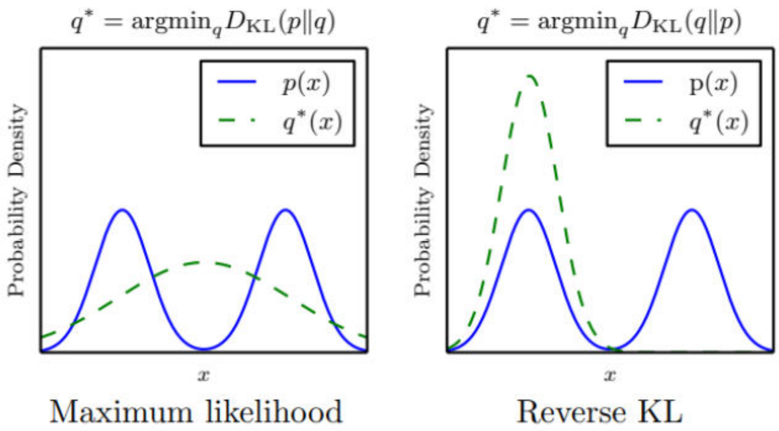
- **Discrete** and **Continuous** probability distributions have different equations
    - We integrate over continuous variables as per usual
- It is a **normalised log ratio** of the probability of occurrences occurring in one distribution vs the other distribution when there are numerous classes
    - It checks how closely the second distribution would be able to generate samples from the first distribution
- With machine learning, we often try to minimise KL divergence between our model and the underlying distribution of our data, i.e. by minimising **cross entropy loss**, which is just minimising KL divergence
- It is **not a distance measure**, as it is not symmetric
- We can use a **smoothing method**, where we add tiny likelihood of occurrences to classes that didn't occur in our samples, so that we don't say that an event for our model  or underlying data is impossible as we might have just been unlucky (prevents issues with log 0 resulting in $D_{KL}$ = 0 or inf)

$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

$$D_{KL}(P||Q) = \int P(x) \log \frac{P(x)}{Q(x)} dx$$

**KL Divergence is not symmetric**

- If **p** is our target distribution, and **q** is the prediction, **$D_{KL}$(p||q)** tries to have **high probability everywhere with data**, while on the other hand, **$D_{KL}$(q||p)** tries to have **low probability everywhere with no data**
- Generations from **$D_{KL}$(q||p)** might have more visually appealing results as there won't be many predictions in places where the target distribution wouldn't have any occurrences
- **$D_{KL}$(p||q)** is the measure of information lost when using **q** to approximate **p**
    - Maximises what proportion of our model **p** is modelled by our estimate **q**
-



Maximum likelihood

Reverse KL

**Reverse KL *(*$D_{KL}$(q||p)*)* -** makes it so you are less likely to predict unlikely occurrences, but you also won't make predictions over the entire range of the target distribution

- **$D_{KL}$(q||p)** is the measure of how much information is kept when using **q** to approximate **p**
- Minimises what proportion of our model **q** is incorrect
- It is said to have a **mode-seeking nature**

**||** operator **-** doesn't mean anything special in $D_{KL}$(p||q), means basically **and**, like a '**,**'

- The reason we use it is to **emphasise that $D_{KL}$** is not a distance, and **the order of p and q matters**

**Zero-forcing -** $D_{KL}$(q||p) is zero-forcing because it makes it so that **whenever p(x) is low, q(x) is low too**

- It means we are modelling the widest component of our distribution, to **minimise false positives**
- Therefore, it often **models the tails of distribution**
- Because it doesn't model the spikes, it **underestimates the variance of p(x)**