Kullback-Leibler Divergence D(p||q)

What if we encode X with a code optimized for a wrong distribution q?

How many bits would we waste?

$$D(p \| q) = H(p,q) - H(p) = \sum_{x\in\Omega} p(x) \log \frac{p(x)}{q(x)}$$

Relative entropy

Properties:

- D(p||q)≥0
- D(p||q)≠D(q||p)
- D(p||q)=0 iff   p=q

→

KL-divergence is often used to measure the distance between two distributions

Interpretation:

-Fix p, D(p||q) and H(p,q) vary in the same way

-If p is an empirical distribution, minimize D(p||q) or H(p,q) is equivalent to maximizing likelihood