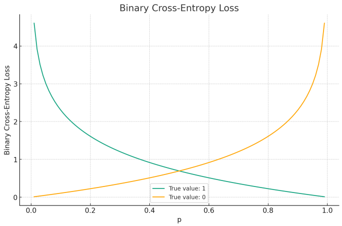
# Introduction to Machine Learning

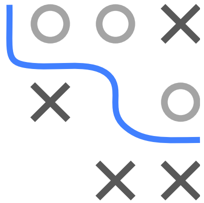
# Information Theory
# Cross-Entropy and KL




## Learning goals

- Know the cross-entropy
- Understand the connection between entropy, cross-entropy, and KL divergence

# CROSS-ENTROPY - DISCRETE CASE



**Cross-entropy** measures the average amount of information required to represent an event from one distribution $p$ using a predictive scheme based on another distribution $q$ (assume they have the same domain $\mathcal{X}$ as in KL).

$$H(p\|q) = \sum_{x \in \mathcal{X}} p(x) \log\left(\frac{1}{q(x)}\right) = -\sum_{x \in \mathcal{X}} p(x) \log\left(q(x)\right) = -\mathbb{E}_{X \sim p}[\log(q(X))]$$

For now, we accept the formula as-is. More on the underlying intuition follows in the content on inf. theory for ML and sourcecoding.

- Entropy = Avg. amount of information if we optimally encode $p$
- Cross-Entropy = Avg. amount of information if we suboptimally encode $p$ with $q$
- $DL_{KL}(p\|q)$: Difference between the two
- $H(p\|q)$ sometimes also denoted as $H_q(p)$ to set it apart from KL

# CROSS-ENTROPY - CONTINUOUS CASE

For continuous density functions $p(x)$ and $q(x)$:

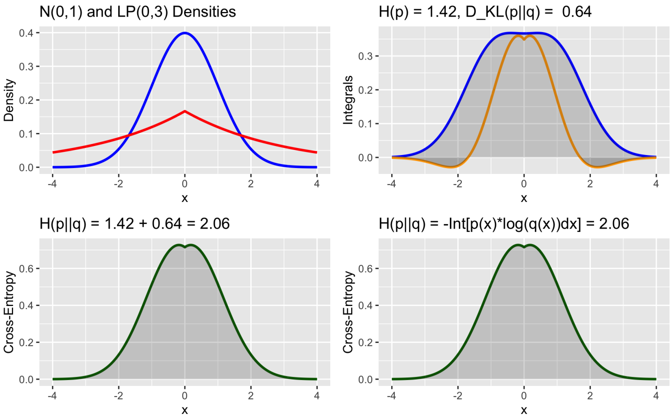
$$H(p\|q) = \int p(x) \log\left(\frac{1}{q(x)}\right) dx = -\int p(x) \log\left(q(x)\right) dx = -\mathbb{E}_{X\sim p}[\log(q(X))]$$

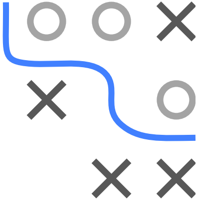
- It is not symmetric.
- As for the discrete case, $H(p\|q) = h(p) + D_{KL}(p\|q)$ holds.
- Can now become negative, as the $h(p)$ can be negative!

# CROSS-ENTROPY EXAMPLE

Let $p(x) = N(0,1)$ and $q(x) = LP(0,3)$. We can visualize

$$H(p\|q) = H(p) + D_{KL}(p\|q)$$

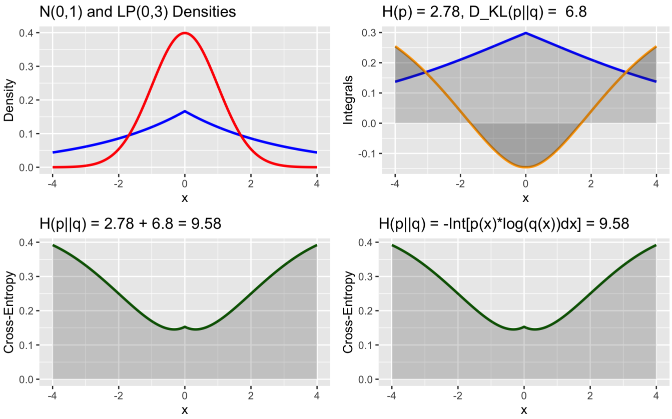


# CROSS-ENTROPY EXAMPLE

Let $p(x) = LP(0, 3)$ and $q(x) = N(0, 1)$. We can visualize

$$H(p\|q) = H(p) + D_{KL}(p\|q)$$



# PROOF: MAXIMUM OF DIFFERENTIAL ENTROPY

**Claim**: For a given variance, the continuous distribution that maximizes differential entropy is the Gaussian.

**Proof**: Let $g(x)$ be a Gaussian with mean $\mu$ and variance $\sigma^2$ and $f(x)$ an arbitrary density function with the same variance. Since differential entropy is translation invariant, we can assume $f(x)$ and $g(x)$ have the same mean.

The KL divergence (which is non-negative) between $f(x)$ and $g(x)$ is:

$$\begin{aligned} 0 \leq D_{KL}(f\|g) &= -h(f) + H(f\|g) \\ &= -h(f) - \int_{-\infty}^{\infty} f(x)\log(g(x))dx \end{aligned} \tag{1}$$