

Deep Learning

A course about theory & practice



Kullback-Leibler Divergence

Marco Piastra

Entropy of a probability distribution

(A discrete probability setting is adopted for convenience)

■ Shannon's Entropy, definition

Consider a (discrete) random variable

$$X \in \mathcal{X}$$

having distribution

$$P(X)$$

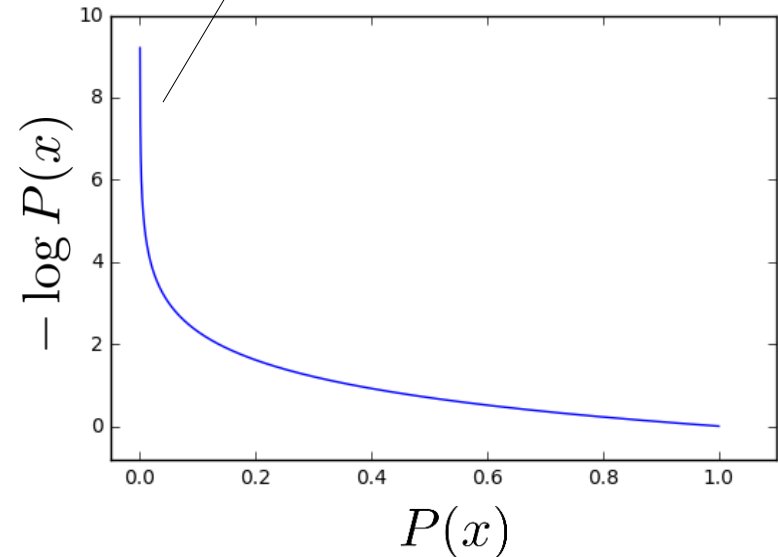
Shannon's Entropy definition:

$$H(X) := - \sum_{x \in \mathcal{X}} P(x) \log P(x)$$

This term is negative or zero

What is the intuitive meaning?

The curve grows exponentially with lower probability values



Intuitively, the entropy is maximal when the probability distribution is 'dispersed' over smaller values and minimal when it is 'one-hot'

Moral: it measures uncertainty

Entropy of a probability distribution

(A discrete probability setting is adopted for convenience)

■ Shannon's Entropy, definition

$$H(X) := - \sum_{x \in \mathcal{X}} P(x) \log P(x)$$

Consider that:

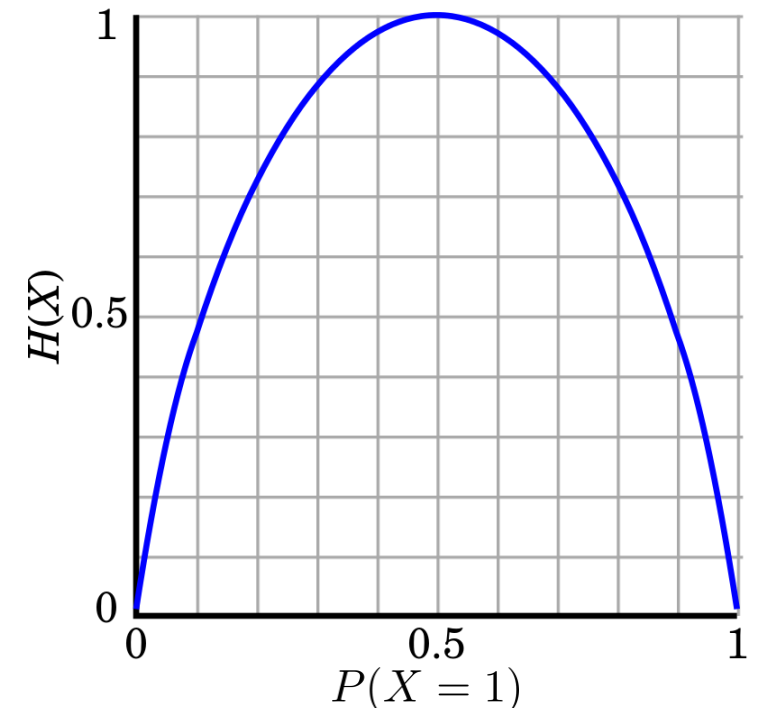
$$P(x) \log P(x) = 0 \text{ when } P(x) = 1$$

$$\lim_{P(x) \rightarrow 0} P(x) \log P(x) = 0$$

■ Head or Tail?

When $X \in \{0, 1\}$ (it is binary) $H(X)$ is maximum for
for $P(X = 0) = P(X = 1) = 0.5$

while it is minimum for $P(X = 0) = 1$ or $P(X = 1) = 1$



*Intuitively, it measures
the level of uncertainty
conveyed by the distribution*

Relative entropy: Kullback–Leibler

(A discrete probability setting is adopted for convenience)

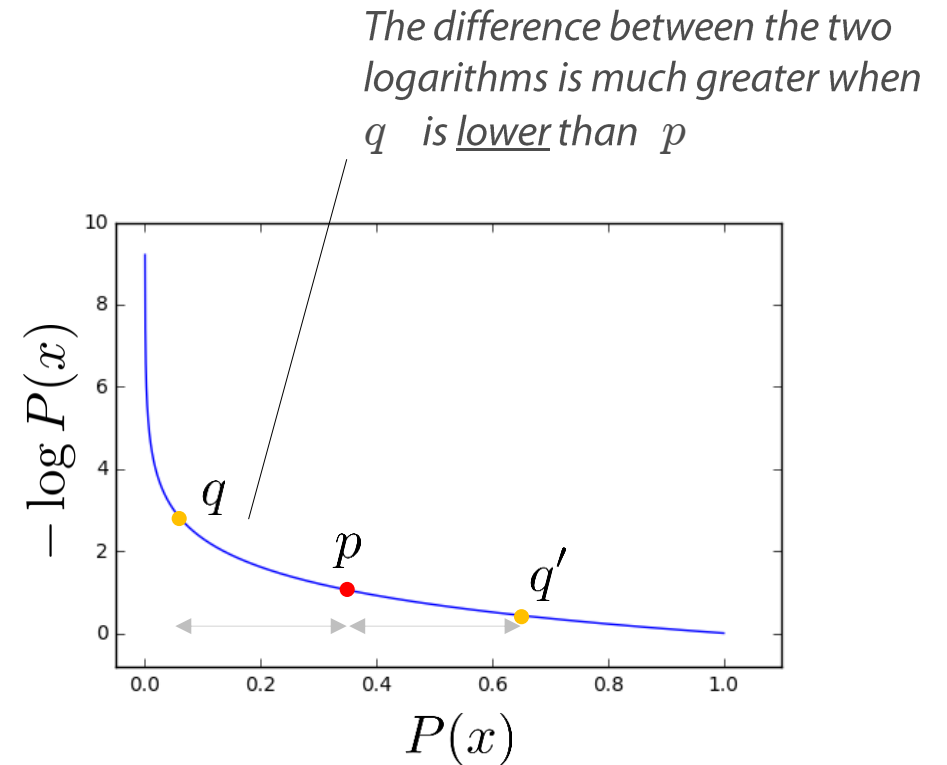
■ Kullback-Leibler divergence, definition

A measure of how one probability distribution $Q(X)$ diverges from a distribution $P(X)$

The Kullback-Leibler divergence is defined as:

$$\begin{aligned} \text{KL}(P \parallel Q) &:= \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)} \\ &= \sum_{x \in \mathcal{X}} P(x) (\log P(x) - \log Q(x)) \end{aligned}$$

Same question:
what is the intuitive meaning?



Given that both distributions sum up to 1 (=normalized) there will be a predominance of positive values

Moral:

$$\text{KL}(P \parallel Q) \geq 0$$

Kullback-Leibler divergence is always non-negative

Relative entropy: Kullback–Leibler

(A discrete probability setting is adopted for convenience)

▪ Kullback-Leibler divergence: probability distribution vs. dataset

Consider a dataset $D := \{x^{(i)}\}_{i=1}^N$

The likelihood of D being *generated* by probability distribution Q is:

$$L(D, Q) := \prod_{i=1}^N Q(x^{(i)}) \quad \text{Likelihood of an I.I.D. dataset = Joint Probability Distribution}$$

$$\text{avg}(\log L(D, Q)) := \frac{1}{N} \sum_{i=1}^N \log Q(x^{(i)})$$

In the limit $N \rightarrow \infty$, it becomes

$$\mathbb{E}[\log L(D, Q)] = \sum_{x \in \mathcal{X}} P(x) \log Q(x)$$

Cross Entropy

The 'true' distribution that generated D

Relative entropy: Kullback–Leibler

(A discrete probability setting is adopted for convenience)

▪ Kullback-Leibler divergence: probability distribution vs. dataset

$$\mathbb{E}[\log L(D, Q)] = \sum_{x \in \mathcal{X}} P(x) \log Q(x)$$

Notice that:

Since this represents the 'true' distribution of the data, this term remains constant during optimization

$$\text{KL}(P \parallel Q) := \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)} = \sum_{x \in \mathcal{X}} P(x) \log P(x) - \sum_{x \in \mathcal{X}} P(x) \log Q(x)$$

(minus) Entropy of P Cross Entropy

Moral: **minimizing** $\text{KL}(P \parallel Q)$ corresponds to **maximizing** $L(D, Q)$

If Q is parametric:

$$\boldsymbol{\vartheta}_{MLE}^* := \operatorname{argmax}_{\boldsymbol{\vartheta}} L(D, Q, \boldsymbol{\vartheta}) = \operatorname{argmin}_{\boldsymbol{\vartheta}} \text{KL}(P \parallel Q, \boldsymbol{\vartheta})$$

Maximum Likelihood Estimator