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

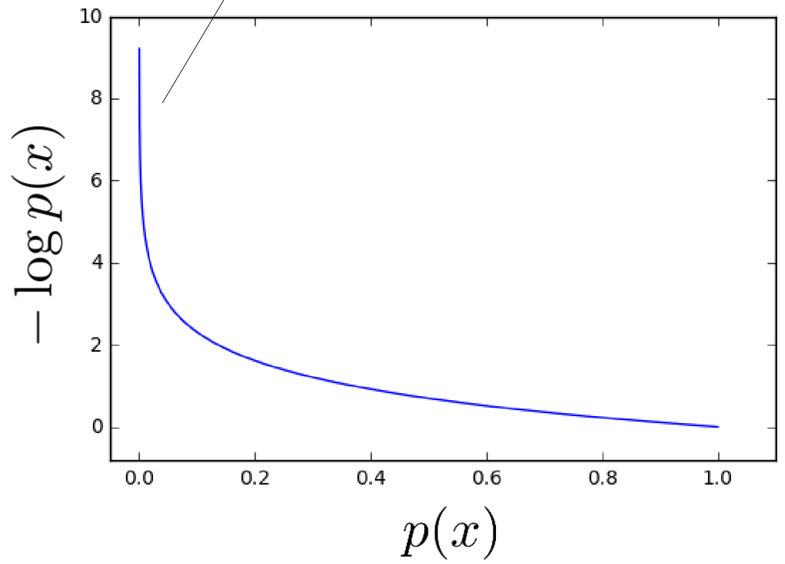
  $$p(X)$$

  Shannon's Entropy is defined as

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

  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**

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

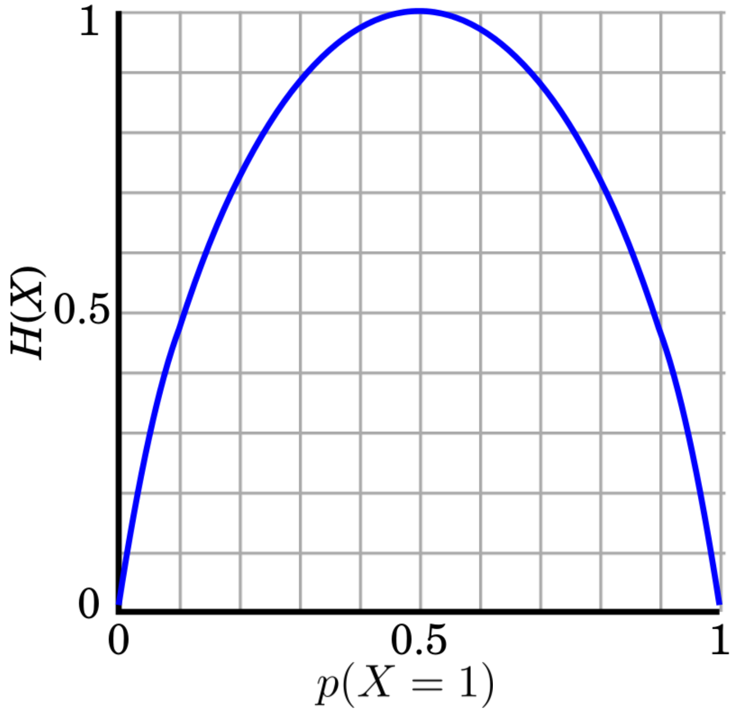
Consider that:

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

$$\lim_{p(x) \to 0} p(x) \log p(x) = 0$$

- **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 when $p(X = 0) = 1$ or $p(X = 1) = 1$



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

[Image from https://en.wikipedia.org/wiki/Entropy_(information_theory)]

Relative entropy: Kullback-Leibler

(A discrete probability setting is adopted for convenience)

- **Kullback-Leibler divergence, definition**

  Consider two (discrete) distributions

  $$p(X),\ q(X)$$

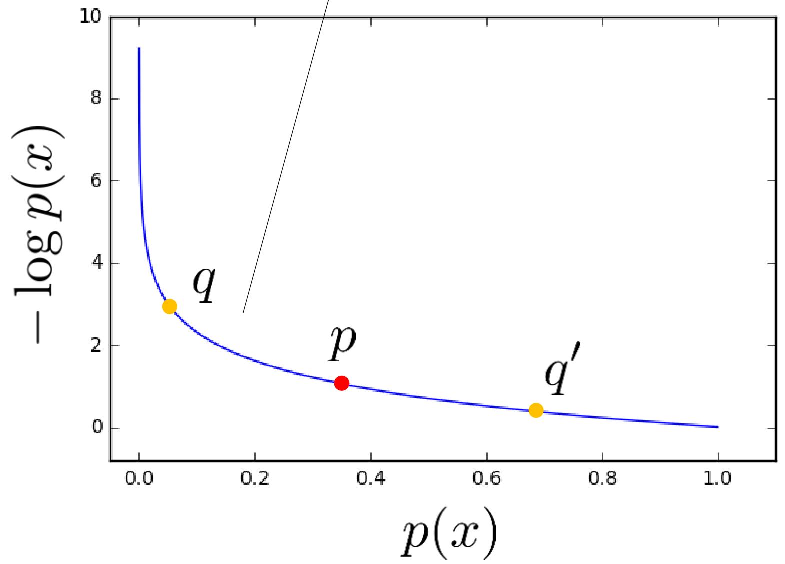
  The Kullback-Leibler divergence is:

  $$D_{KL}(p \parallel q) := \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$$

  Same question:
  what is the intuitive meaning?

  $$= \sum_{x \in \mathcal{X}} p(x)(\log p(x) - \log q(x))$$

The difference between the two logarithms is much greater when q *is lower than* p



Given that both distributions are normalized there will be an excess of positive values

Consider that the above is $-\log p(x)$

Moral:

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

Relative entropy: Kullback-Leibler

(A discrete probability setting is adopted for convenience)

- **Kullback-Leibler divergence, definition**

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

The likelihood of D being *generated* by probability distribution q is

Note that is negative

$$L(D, q) := \prod_{i=1}^{N} q(x^{(i)}) \implies \operatorname{avg}(\log L(D, q)) := \frac{1}{N} \sum_{i=1}^{N} \log q(x^{(i)})$$

In the limit of $N \to \infty$, the latter becomes $\sum_{x \in \mathcal{X}} p(x) \log q(x)$ *Cross Entropy*

The 'true' distribution that generated D

$$D_{KL}(p \parallel q) := \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} = \sum_{x \in \mathcal{X}} p(x) \log p(x) - \sum_{x \in \mathcal{X}} p(x) \log q(x)$$

Entropy of p *Cross Entropy*

Moral: minimizing $D_{KL}(p \parallel q)$ is maximizing $L(D, q)$ (MLE - dataset vs. a model distribution)