Deep Learning

A course about theory & practice

UNIVERSITÀ DI PAVIA

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

  $$p(X)$$

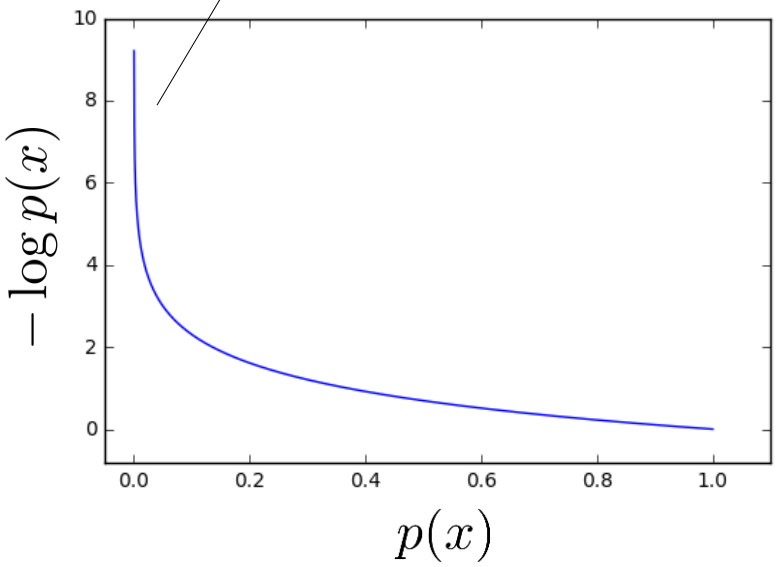
  Shannon's Entropy <u>definition</u>:

  This term is negative or zero

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

  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 <u>uncertainty</u>

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:

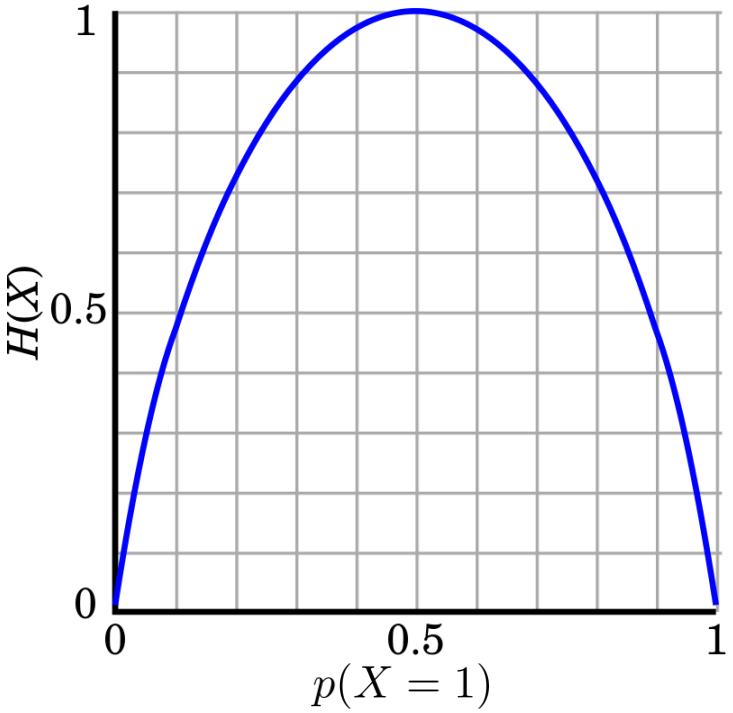
$$p(x) \log p(x) = 0 \quad \text{when} \quad p(x) = 1$$

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

- **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 $p(X = 1) = 1$



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

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 $p(X)$

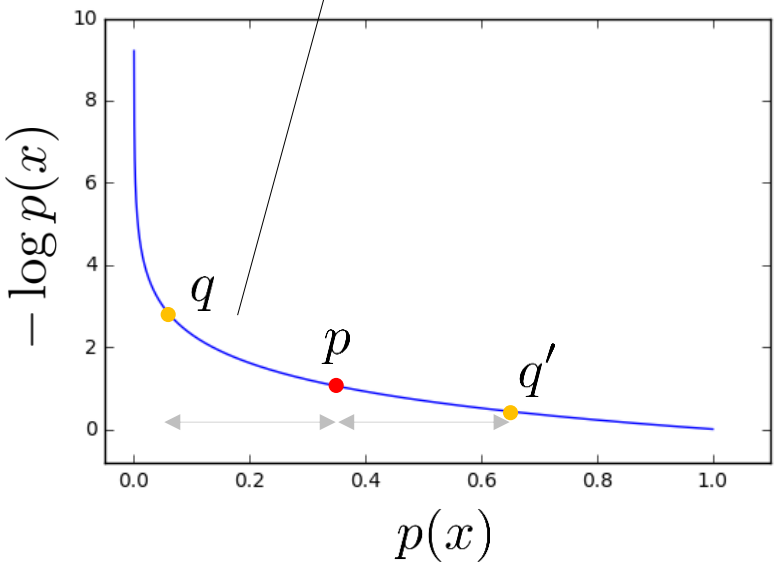
  The Kullback-Leibler divergence is defined as:

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

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

Same question:
what is the intuitive meaning?

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



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

Moral:

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

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:

$$L(D, q) := \prod_{i=1}^{N} q(x^{(i)})$$

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 $N \to \infty$, it becomes

(minus) Cross Entropy

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

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:

$$D_{KL}(p \parallel q) := \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)} = \underbrace{\sum_{x\in\mathcal{X}} p(x)\log p(x)}_{\text{(minus) Entropy of } p} - \underbrace{\sum_{x\in\mathcal{X}} p(x)\log q(x)}_{\text{Cross Entropy}}$$

Moral: minimizing $D_{KL}(p \parallel q)$ *corresponds to maximizing* $L(D,q)$

If q *is parametric:*

$$\vartheta^*_{MLE} := \text{argmax}_\vartheta\ L(D,q,\vartheta) = \text{argmin}_\vartheta\ D_{KL}(p \parallel q,\vartheta)$$

Maximum Likelihood Estimator