

# Cheatsheet InfoTheory

Nicolas Wehrli

June 2024

## 1 Foundations

### Definitions

Information of an outcome  $x$

$$h(x) = -\log(p(x))$$

Cross-Entropy between  $p$  and  $q$

$$H(p; q) = -\sum_x p(x) \log q(x)$$

Shannon Entropy

$$H(p) = H(p; p)$$

### Notation

We identify outcomes  $x$  with integers  $1, \dots, m$  and associate probabilities  $p(x) \geq 0$ .

$H(\frac{1}{m})$  for  $H(p)$  with  $p(x) = \frac{1}{m}$  (uniform)  
 $H(X) = H(p) = \mathbb{E}(-\log(p(X)))$  where  $p$  is the pdf of  $X$

### Jensen's Inequality

Let  $f$  be convex and  $g : [m] \rightarrow \mathbb{R}$  be an arbitrary function that assigns a value to each outcome.

$$f\left(\sum_x p(x)g(x)\right) \leq \sum_x p(x)f(g(x)), \forall p(x) \geq 0, \sum_x p(x) = 1$$

alternatively

$$f(\mathbb{E}(g(X))) \leq \mathbb{E}(f(g(X)))$$

Applying this inequality to relate Cross-Entropy and Entropy, we get the following properties.

By Jensen we have

$$H(p; q) \geq H(p)$$

Defining KL divergence or Relative Entropy as

$$D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

we get

$$H(p; q) = H(p) + D(p||q)$$

Further investigating KL divergence, we find

$$D(p||q) \geq 0 \quad (1)$$

$$D(p||q) = 0 \iff p = q \quad (2)$$

A further consequence of (1) is that the uniform distribution maximizes entropy.

$$H\left(\frac{1}{m}\right) = \max_p H(p)$$

### Definitions - Conditional distributions

Conditional information

$$h(x|y) = -\log p(x|y)$$

Conditional Entropy

$$H(X|Y = y) = -\sum_x p(x|y) \log p(x|y)$$

$$H(X|Y) = \sum_y p(y)H(X|Y = y)$$

Monotonicity of Conditioning

$$H(X|Y) \leq H(X)$$

### Joint Entropy

$$H(X, Y) = -\sum_{x,y} p(x, y) \log p(x, y)$$

### Chain Rule

$$H(X, Y) = H(X|Y) + H(Y)$$

### Subadditivity

$$H(X, Y) \leq H(X) + H(Y)$$

with equality if  $X \perp Y$ .

### Multiple Conditioning

$$H(X|Y, Y') \leq H(X|Y)$$

Generalized to  $X_1, \dots, X_n$  we get

$$H(X_1, \dots, X_n) = \sum_{i=1}^n H(X_i|X_1, \dots, X_{i-1}) \leq \sum_{i=1}^n H(X_i)$$

### Mutual Information

$$I(X; Y) := H(X) - H(X|Y) = H(Y) - H(Y|X)$$

We further have

$$I(X; Y) = D(P(X, Y)||P(X)P(Y))$$

with  $I(X; Y) = 0$  if  $X \perp Y$ .

### Conditional Mutual Information

$$I(X; Y|Z) := H(X|Z) - H(X|Y, Z)$$

### Conditional Independence

If  $X \perp Y|Z$

$$I(X; Y|Z) = 0 \text{ and } I(X; Y) \leq I(X; Z)$$

We can deduce that for any function  $\phi$  on outcomes of  $X$

$$I(\phi(X); Y) \leq I(X; Y)$$

## 2 Compression

### Definition - Code

A code  $C$  is a mapping from outcomes to codewords

$$C : \{1, \dots, m\} \rightarrow \{0, 1\}^*$$

- If there is no codeword that is a prefix of another codeword, the code is a **prefix code**.
- Prefix codes retain injectivity when concatenating codewords.

Sets of codewords fulfilling the prefix property can be uniquely represented by the leaves of a binary tree. Since a leaf node has no children the prefix property is guaranteed.

### Kraft's Inequality

If  $\{c_1, \dots, c_m\}$  are codewords of a prefix code, then

$$\sum_x 2^{-l_x} \leq 1, \text{ where } l_x = |c_x| \quad (3)$$

Conversely, given  $\{l_1, \dots, l_m\} \subset \mathbb{N}$  satisfying (3), there exists a prefix code with those codeword lengths.

- Codes for which Kraft's inequality is strict can be optimized by codeword pruning.
- A prefix is succinct, if Kraft's inequality holds with an equality.
- Succinct codes uniquely define a dyadic probabilistic model

$$q(x) = 2^{-l_x}$$

- Expected codeword length of a prefix code  $C$

$$L(C) = \sum_x p(x)l_x = \sum_x p(x)(-\log q(x)) = H(p; q)$$

- Using  $H(p; q) = H(p) + D(p||q)$  we can deduce that the minimal  $L(C)$  for a binary prefix code  $C$  is

$$L^* = H(p) + \min_{q: \text{dyadic}} D(p||q)$$

- Thus the closer  $q$  is to  $p$ , the more optimal the prefix code is. But since  $p$  doesn't have to be dyadic there can be an inherent suboptimality based on rounding.

### Weak Law of Large Numbers

Let  $Y_1, \dots, Y_n$  be iid. random variables with mean  $\mu$ . Then

$$\bar{Y}_n := \frac{1}{n} \sum_{i=1}^n Y_i \xrightarrow{\mathbb{P}} \mu \iff \lim_{n \rightarrow \infty} \mathbb{P}(|\bar{Y}_n - \mu| < \varepsilon) = 1, \forall \varepsilon > 0$$

### Typicality

Let  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} p$ . The  $\varepsilon$ -typical outcomes are

$$\mathcal{A}_\varepsilon^n = \left\{ x \in \{1, \dots, m\}^n : \left| H(p) + \frac{1}{n} \sum_{i=1}^n \log p(x_i) \right| < \varepsilon \right\}$$