



# Fundamentals of information theory

Enrico Maria Di Mauro  
matr. 062270**1706**



# Acronyms

RV  $\rightarrow$  Random Variable

PMF  $\rightarrow$  Probability Mass Function

## Notation

$X$  and  $Y$  are RVs

$x$  and  $y$  are the values assumed by  $X$  and  $Y$  respectively

$p(\cdot)$  and  $q(\cdot)$  are PMFs of a RV

$u(\cdot)$  is the PMF of a uniform RV

$p(\cdot, \cdot)$  is the joint PMF of a pair of RVs

$p(\cdot | \cdot)$  is the conditional PMF of a pair of RVs

# SUMMARY (introduction)

Entropy

Joint entropy

Conditional entropy

Chain rule

Divergence

Mutual information

Log-Sum inequality

**SUMMARY** (demonstrations)



# Entropy

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

Given a RV  $X$ , entropy represents the information content of  $X$

Information  $\Leftrightarrow$  Uncertainty



# Joint Entropy

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

Given a pair of RVs  $(X, Y)$ , joint entropy represents the information content of  $(X, Y)$



# Conditional Entropy



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

Given two RVs  $X$  and  $Y$ , conditional entropy to a particular value represents the information content of  $X$  given a particular value of  $Y$

$$H(X | Y) = \sum_{y \in \mathbb{Y}} p(y) H(X | Y = y)$$

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

Conditional entropy represents the information content of  $X$  given  $Y$



# Chain rule

$$H(X, Y) = \begin{cases} H(Y) + H(X | Y) \\ H(X) + H(Y | X) \end{cases}$$

The information content of a pair of RVs  $(X, Y)$  corresponds to the information content of  $X$  added to that of  $Y$  given  $X$  or vice versa

Generalizing to  $n$  RVs

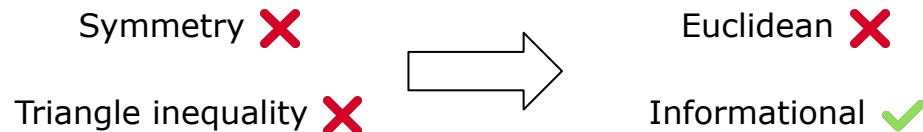
$$H(X_1, \dots, X_n) = \sum_{i=1}^n H(X_i | X_{i-1}, \dots, X_1)$$



# Divergence (Kullback Leibler distance / Relative Entropy)

$$D(p \parallel q) = \sum_{x \in \mathbb{X}} p(x) \log \frac{p(x)}{q(x)}$$

Given two PMF  $p(\cdot)$  and  $q(\cdot)$  defined on the same alphabet, divergence represents a measure of how  $p(\cdot)$  is different from  $q(\cdot)$ . A simple interpretation of the divergence of  $p(\cdot)$  from  $q(\cdot)$  is the expected excess surprise from using  $q(\cdot)$  as a model when the actual distribution is  $p(\cdot)$ . In addition, divergence is interpreted as a distance







# Mutual information



$$I(X; Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} = D(p(x, y) \parallel p(x)p(y))$$

Given two RVs  $X$  and  $Y$ , mutual information represents the information content that  $X$  contains with respect to  $Y$  and vice versa

$$I(X; Y) = \begin{cases} H(X) - H(X | Y) \\ H(Y) - H(Y | X) \\ H(X) + H(Y) - H(X, Y) \end{cases}$$

In addition, mutual information represents the uncertainty reduction of  $X$  when  $Y$  is revealed and vice versa



# Log-Sum inequality

$$\sum_{i=1}^n a_i \log \frac{a_i}{b_i} \geq \left( \sum_{i=1}^n a_i \right) \log \frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i} \quad \text{with "}" } \iff \frac{a_i}{b_i} = \text{const}$$

Log-Sum inequality holds for all  $a_1, \dots, a_n$  and  $b_1, \dots, b_n$  non-negative numbers

## SUMMARY (demonstrations)

1. Non-negative divergence
2. Non-negative mutual information
3. Entropy limits
4. Conditioning reduces entropy
5. Joint entropy with deterministic function
6. Entropy reduction with transformation
7. Entropy reduction with conjunction
8. Entropy increasing linearity
9. Merging reduces entropy
10. Mixing increasing entropy



1.  $D(p \parallel q) \geq 0$  with " $=$ "  $\iff p = q$

For the Log-Sum inequality

$$D(p \parallel q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \geq \left( \sum_x p(x) \right) \log \frac{\sum_x p(x)}{\sum_x q(x)}$$

Since the sum of all PMF values is =1

$$D(p \parallel q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \geq 1 \log \frac{1}{1} = 0$$

$\Downarrow$

$$D(p \parallel q) \geq 0$$

In addition, for the Log-Sum inequality the equality holds  $\Leftrightarrow \frac{p(x)}{q(x)} = \text{const}$

Since this ratio is valid for every  $x$



$$\frac{\sum_x p(x)}{\sum_x q(x)} = \text{const} \implies \text{const} = 1 \implies p = q$$



1.  $D(p \parallel q) \geq 0$  with " = "  $\iff p = q$

Divergence represents a distance, not euclidean but informational, and for this reason it is a non-negative quantity

In addition, this distance is zero when measuring the distance of a PMF from itself





2.  $I(X; Y) \geq 0$  with "="  $\iff$   $X$  and  $Y$  are independent

For the demonstration 1.

$$I(X; Y) = D(p(x, y) \parallel p(x)p(y)) \geq 0$$

In addition, for the demonstration 1. the equality holds  $\iff p(x, y) = p(x)p(y)$



So the equality holds  $\iff X$  and  $Y$  are independent



2.  $I(X; Y) \geq 0$  with "="  $\iff X$  and  $Y$  are independent

Mutual information represents a reduction of uncertainty when the RVs are mutually dependent

If the RVs are independent, there is no uncertainty reduction as they say nothing about each other and there is no mutual information

$$3. \quad 0 \leq H(X) \leq \log n \quad \text{with} \begin{cases} = 0 & \iff X \text{ is degenerate} \\ = \log n & \iff X \text{ is uniform} \end{cases}$$

For definition  $H(X) \geq 0$  and the equality holds  $\Leftrightarrow$  a unique value of  $x$  has the PMF  $=1$  and all others have the PMF  $=0$ , so  $X$  is degenerate

For the demonstration 1.

$$D(p \parallel u) = \sum_{x=1}^n p(x) \log \frac{p(x)}{u(x)} = \sum_{x=1}^n p(x) \log (p(x)n) = \sum_{x=1}^n p(x) \log p(x) + \sum_{x=1}^n p(x) \log n = -H(X) + \log n \geq 0$$



$$\Downarrow$$

$$H(X) \leq \log n$$

In addition, for the demonstration 1. the equality holds  $\Leftrightarrow p(x) = u(x)$



So the equality holds  $\Leftrightarrow X$  is uniform




$$3. \quad 0 \leq H(X) \leq \log n \quad \text{with} \begin{cases} = 0 & \iff X \text{ is degenerate} \\ = \log n & \iff X \text{ is uniform} \end{cases}$$

Entropy is maximum when the RV is uniform since each event has the same probability of happening and therefore the uncertainty is maximum

Entropy is minimum when the RV is degenerate since only one event has the probability of happening =1 and therefore the uncertainty is zero



4.  $H(X | Y) \leq H(X)$   
with " = "  $\iff X$  and  $Y$  are independent



For the demonstration 2

$$I(X; Y) = H(X) - H(X | Y) \geq 0$$

$\Downarrow$



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

In addition, for the demonstration 2 the equality holds  $\iff X$  and  $Y$  are independent



4.  $H(X | Y) \leq H(X)$   
with " = "  $\iff X$  and  $Y$  are independent

The entropy of a RV  $X$  decreases when some information about it is obtained,  
so the entropy of  $X$  decreases when another RV  $Y$  correlated to  $X$  is revealed.  
This does not happen if  $X$  and  $Y$  are independent





5.  $H(Y | X) = 0 \iff Y = g(X)$  with  $g(\cdot)$  deterministic

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

Suppose that  $H(Y | X) = 0 \Rightarrow H(Y | X = x) = 0 \quad \forall x : p(x) > 0$

For the demonstration  $\exists p(y | x)$  is degenerate  $\Rightarrow$  knowing  $x$  it  
is known  $y \Rightarrow y$  is function of  $x \quad \forall x : p(x) > 0$

Suppose that  $Y = g(X) \Rightarrow p(y | x)$  is degenerate  $\forall x : p(x) > 0 \Rightarrow$   
 $H(Y | X = x) = 0 \quad \forall x : p(x) > 0 \Rightarrow H(Y | X) = 0$



5.  $H(Y | X) = 0 \iff Y = g(X)$  with  $g(\cdot)$  deterministic

Given a RV  $X$  and knowing that  $g(X)$  is a deterministic function of it, revealing  $X$  the uncertainty of  $g(X)$  is zeroed



6.  $H(g(X)) \leq H(X)$  with " $=$ "  $\iff g(\cdot)$  is reversible

For the demonstration 5.

$$I(X; g(X)) = \begin{cases} H(X) - H(X | g(X)) \\ H(g(X)) - \underbrace{H(g(X) | X)}_0 \end{cases}$$



$\Downarrow$

$$H(g(X)) = H(X) - H(X | g(X))$$

For definition  $H(X | g(X)) \geq 0 \Rightarrow H(g(X)) \leq H(X)$


In addition, the equality holds  $\Leftrightarrow H(X | g(X)) = 0$

So the equality holds  $\Leftrightarrow g(X)$  is reversible



6.  $H(g(X)) \leq H(X)$  with " $=$ "  $\iff g(\cdot)$  is reversible

Every transformation involves a loss of information  
unless that transformation is reversible



7.  $H(X_1, \dots, X_n) \leq H(X_1) + \dots + H(X_n)$   
with " = "  $\iff X_1, \dots, X_n$  are mutually independent

For the demonstration 2



$$I(X_1; \dots; X_n) = H(X_1) + \dots + H(X_n) - H(X_1, \dots, X_n) \geq 0$$

$\Downarrow$

$$H(X_1, \dots, X_n) \leq H(X_1) + \dots + H(X_n)$$

In addition, the equality holds  $\Leftrightarrow X_1, \dots, X_n$  are mutually independent





7.  $H(X_1, \dots, X_n) \leq H(X_1) + \dots + H(X_n)$   
with "="  $\iff X_1, \dots, X_n$  are mutually independent

Computing the joint entropy of a number of RVs produces a loss of information compared to summing the entropies individually unless they are all mutually independent



8.  $X_1, \dots, X_n \text{ iid} \rightarrow H(X_1, \dots, X_n) = nH(X_1)$

For the demonstration  $\underline{Z}$

$$H(X_1, \dots, X_n) = H(X_1) + \dots + H(X_n)$$

In addition, since all RVs are identically distributed all PMFs are equal  $\Rightarrow$   
all entropies are equal  $H(X_1) = \dots = H(X_n) \Rightarrow H(X_1, \dots, X_n) = nH(X_1)$



8.  $X_1, \dots, X_n \text{ iid} \rightarrow H(X_1, \dots, X_n) = nH(X_1)$

The information grows linearly as the number of data increases, but if data are dependent on each other, the information increases more slowly because every time there is some superfluous information

9.  $H([p_1, \dots, p_{n-1} + p_n]) \leq H([p_1, \dots, p_n])$   
 with " = "  $\iff$  merging is fictitious

Given the entropy of a PMF with and without the merging

$$H([p_1, \dots, p_{n-1} + p_n]) = - \sum_{i=1}^{n-2} [p_i, \dots, p_{n-2}] \log [p_i, \dots, p_{n-2}] - [p_{n-1} + p_n] \log [p_{n-1} + p_n]$$

$$H([p_1, \dots, p_n]) = - \sum_{i=1}^{n-2} [p_i, \dots, p_{n-2}] \log [p_i, \dots, p_{n-2}] - p_{n-1} \log p_{n-1} - p_n \log p_n$$



Deleting the same arguments it is possible to reduce the demonstration to:

$$\begin{aligned} [p_{n-1} + p_n] \log [p_{n-1} + p_n] &\geq p_{n-1} \log p_{n-1} + p_n \log p_n \\ \Downarrow \\ \underline{p_{n-1} \log [p_{n-1} + p_n]} + \underline{p_n \log [p_{n-1} + p_n]} &\geq \underline{p_{n-1} \log p_{n-1}} + \underline{p_n \log p_n} \end{aligned}$$

Since for definition all the PMFs have values between 0 and 1 and that the logarithm is an increasing function

- $p_{n-1} \log [p_{n-1} + p_n] \geq p_{n-1} \log p_{n-1} \Rightarrow$  the inequality is verified
- $p_n \log [p_{n-1} + p_n] \geq p_n \log p_n$

In addition, the equality is verified  $\iff$  merging is done with a null value, so merging is fictitious



9.  $H([p_1, \dots, p_{n-1} + p_n]) \leq H([p_1, \dots, p_n])$   
with " = "  $\iff$  merging is fictitious

Merging reduces information unless it is fictitious, so merging  
reduces information if it is not done with a null value

10.  $H\left(\left[p_1, \dots, \frac{p_i + p_j}{2}, \dots, \frac{p_i + p_j}{2}, \dots, p_n\right]\right) \geq H([p_1, \dots, p_n])$   
 with " = "  $\iff$  mixing is fictitious

Given the entropy of a PMF with and without the mixing

$$H\left(\left[p_1, \dots, \frac{p_i + p_j}{2}, \dots, \frac{p_i + p_j}{2}, \dots, p_n\right]\right) = - \sum_{i=1}^{n-2} [p_i, \dots, p_{n-2}] \log [p_i, \dots, p_{n-2}] - 2 \frac{p_i + p_j}{2} \log \frac{p_i + p_j}{2}$$




$$H([p_1, \dots, p_n]) = - \sum_{i=1}^{n-2} [p_i, \dots, p_{n-2}] \log [p_i, \dots, p_{n-2}] - p_{n-1} \log p_{n-1} - p_n \log p_n$$

Deleting the same arguments it is possible to reduce the demonstration to:

$$[p_i + p_j] \log \frac{p_i + p_j}{2} \leq p_i \log p_i + p_j \log p_j$$

This inequality is verified by the Log-Sum inequality using  $a_1=p_i, a_2=p_j, b_1=b_2=1$

In addition, the equality is verified  $\iff$  mixing is done with same value, so mixing is fictitious



10. 
$$H\left(\left[p_1, \dots, \frac{p_i + p_j}{2}, \dots, \frac{p_i + p_j}{2}, \dots, p_n\right]\right) \geq H([p_1, \dots, p_n])$$
  
with "="  $\iff$  mixing is fictitious

Mixing increases information unless it is fictitious, so mixing increases information if it is not done with the same values



**Thank you** for the  
attention

Enrico Maria Di Mauro  
matr. 062270**1706**