Information theory

def: Let $x, y \in \{0,1\}^n$. The Hamming distance between x and y is:

$$d_H : \{0,1\}^n \times \{0,1\}^n \to \mathbb{N},$$

$$d_H(x,y) = \sum_{i=1}^{n} |x_i - y_i| = |\{i = 1, \dots, n \text{ s.t. } x_i \neq y_i\}|$$

Entropy is a core concept of Information theory.

Consider two events:
- tonight the moon is full
- tonight the Halley's Comet will be visible in Rome.

The first event occurs every 28 days, whereas the second occurs every 76 years (approx.).

Which of the two events makes us more SURPRISED?

Of course, the second one.

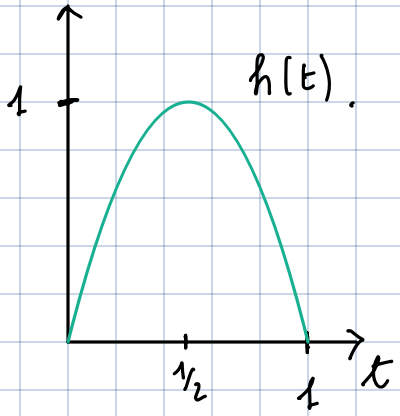
The entropy is a measure of this surprise.

Similarly to the concept of entropy in Physics, it is a measure of how chaotic and unpredictable an event is.

def: let $t \in (0,1)$. The binary entropy of t is

$$h(t) = t \cdot \log_2\left(\frac{1}{t}\right) + (1-t)\log_2\left(\frac{1}{1-t}\right)$$

The graph of the binary entropy is the following:



OBS: $h(t) = h(1-t)$ i.e., the binary entropy is symmetric.

OBS: We can extend the domain of the binary entropy by continuity to $[0,1]$.

def: Let X be a random variable (R.V.) whose values are $\mathcal{X} = \{x_1, \dots, x_n\}$, and $p_i = \mathbb{P}[X = x_i]$ $\forall i = 1, \dots, n$, s.t. $(p_1 \dots p_N)$ is a probability distribution ($\sum_i p_i = 1$, $p_i \geq 0$).

Then, the entropy of X is $H(X) = -\sum_{i=1}^{n} p_i \lg(p_i)$.

Notice that $-\sum_{i=1}^{n} p_i \lg(p_i) = \sum_{i=1}^{n} p_i \lg\left(\frac{1}{p_i}\right)$ (property of logarithms)

* The entropy of a random variable following a probability distribution P is equal to the entropy of P.

def: Similarly, we can define the joint entropy of two (or more) random variables: $x \in \mathcal{X}$, $Y \in \mathcal{Y}$:

$$H(X,Y) := H((X \wedge Y)) = \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \mathbb{P}[X=x, Y=y] \cdot \lg \frac{1}{\mathbb{P}[X=x, Y=y]}$$

We are saying that the entropy of X and Y is defined as the entropy of the couple (X,Y).

def: The INFORMATION CONTENT or SURPRISAL of an event E is a function that grows with the unlikelihood of the event and is defined as:

$$I[E] = -\lg_2 \mathbb{P}[E] = \lg_2 \frac{1}{\mathbb{P}[E]}$$

Hence, entropy measures the expected amount of information conveyed by all possible events of a distribution.

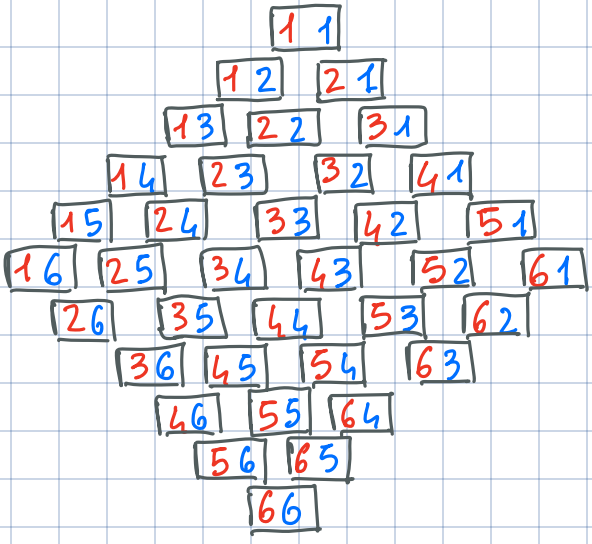
EXAMPLE: we roll a dice with 6 sides. How predictable is the outcome?

→ Totally! Maybe we are not very surprised to see a 6, but actually the dice system is very chaotic.

Two dices, a red and a blue one.

→ How many possible events are there? $6 \times 6 = 36$.

They are the following:



SUM OF THE OUTCOMES	NUMBER of EVENTS PER SUM OUTCOME
2	1
3	2
4	3
5	4
6	5
7	6
8	5
9	4
10	3
11	2
12	1

Let R be the random variable that represents the outcome of rolling the RED DICE. Analogously, we define B for the blue dice.

Notice that each of the 36 events has probability $\frac{1}{36}$, i.e.,

$$\mathbb{P}(R=1, B=6) = \mathbb{P}(R=5, B=3) = \mathbb{P}[R=6, B=6] = \frac{1}{36}$$

Nevertheless, if $S = R+B$, then $\mathbb{P}[S=7] = \frac{6}{36} = \frac{1}{6}$, whereas $\mathbb{P}[S=12] = \frac{1}{36}$. Although the event $(R=4, B=3)$ is just as

likely as the event $(R=6, B=6)$, there are many more events that result in $S=7$, while only one event implies $S=12$.

In fact: $I(S=7) = -\lg \frac{1}{6} \approx 2.6$

$I(S=12) = -\lg \frac{1}{36} \approx 5.2$

} we are much more surprised to observe the event $S=12$ rather than the event $S=7$.

let's compute the entropy of the R.V. $X \sim \text{Dice}_6$

$$H(X) = -6 \cdot \frac{1}{6} \lg_2 \frac{1}{6} = 2.6.$$

Assume now that the dice is not fair and let $\tilde{X}$ be the R.V. describing the outcome of rolling such a dice with distribution:

$$\mathbb{P}[\tilde{X}=1] = \dots = \mathbb{P}[\tilde{X}=5] = \frac{1}{12}, \quad \mathbb{P}[\tilde{X}=6] = \frac{7}{12}$$

Let's compute the entropy of $\tilde{X}$:

$$H(\tilde{X}) = -\frac{5}{12} \lg_2 \frac{1}{12} - \frac{7}{12} \lg_2 \frac{7}{12} \approx 1.3064$$ ↦ it's smaller than the entropy of X.

→ The entropy of a distribution is maximal when the R.V. follows the UNIFORM DISTRIBUTION

Think about it: if a distribution is very skewed, eg. $\mathbb{P}[X=x_1] = \frac{99}{100}$, $\mathbb{P}[X=x_2] = \frac{1}{100}$, we will almost always see $X=x_1$, meaning that the result is easily predictable and does not surprise us.

→ We will see that if a msg is highly predictable, then it is easy to compress.

IF it is unpredictable, it is hard to compress.

Binary Encodings

Let $\mathcal{X}$, $|\mathcal{X}| < \infty$ be an alphabet and let:

$M = \{$words of a language on $\mathcal{X}\}$.

$M^* = \{$sequences of words in $M\}$.

$|M| < \infty$, $|M^*| = \infty$, $M^* = \bigcup_{i=1}^{\infty} M^i$, where $M^i = \{$sequences of i words in $M\}$

We denote a word as $m \in M$, and a sequence of words as $\underline{m} \in M^*$.

def: A VARIABLE LENGTH BINARY ENCODING is an injective function $f: M \to \{0,1\}^*$ that assigns a binary string to each word in M.

def: We can extend this definition by juxstaposing several binary encodings and define the extension by concatenation f^* of the variable length binary encoding f as:

$$f^*: M^* \to \{0,1\}^* \text{ such that } f^*(\underline{m}) = f(m_1) \dots f(m_n)$$

if $\underline{m} = (m_1, \dots, m_n)$.

→ Notice that f^* can be non-injective. For example, let:

$f(m_1) = 0$, $f(m_2) = 1$, $f(m_3) = 01$. Then $f^*(\underline{m} = (m_1, m_2)) = 01 = f^*(\underline{m} = m_3)$.

def: A variable-length binary encoding $f: M \to \{0,1\}^*$ is PREFIX-FREE if $\forall m, m' \in M$, $m \neq m'$, it holds that:

$$f(m) \not\prec f(m'),$$

where $\underline{x} \not\prec \underline{y}$ means "$\underline{x}$ is not a PREFIX of $\underline{y}$".

- We say that $\underline{x} \prec \underline{y}$ ($\underline{x}$ is a prefix of $\underline{y}$) if:

- $\underline{x} = \underline{y}$

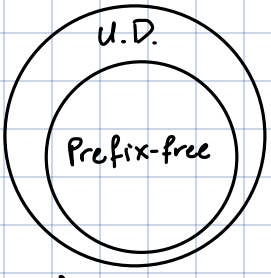
  OR

- $\exists\, \underline{z} \in \{0,1\}^*$ such that $\underline{y} = \underline{x}\,\underline{z}$

→ For example, $\underline{x} = 00$ is a prefix of $\underline{y} = 0001$ ($\underline{z} = 01$) but it is not a prefix of $\underline{y} = 0110$.

def: A binary encoding f is U.D. ("uniquely decodable") if its extension by concatenation f^* is injective.

OBS: PREFIX-FREE $\Rightarrow$ U.D. easy



U.D $\not\Rightarrow$ PREFIX-FREE.

$M = \{m_1, m_2\}$, $f(m_1) = 0$ $f(m_2) = 01$.

f^* is injective, but $f(m_1)$ is a prefix of $f(m_2)$.

↝ being PREFIX-FREE is a stronger condition than uniquely-decodability.

- Why do we like it when f^* is injective?

  let f be the following encoding:

  $$f(m_1) = 00,\ f(m_2) = 01,\ f(m_3) = 000,\ f(m_4) = 1.$$

  notice that f is non U.D, in fact, if $\underline{x} = (m_1, m_2)$ and $\underline{y} = (m_3, m_4)$, it holds that:

  $$f^*(\underline{x}) = 0001$$
  $$f^*(\underline{y}) = 0001$$

  i.e., f^* is not injective.

  Suppose that a receiver receives the message "0001". Then, it cannot possibly know whether the transmitter sent $\underline{x}$ or $\underline{y}$. This causes ambiguity.

→ In contrast, if a binary encoding is uniquely decodable, we can restrict the codomain $\{0,1\}^*$ to just $\mathrm{Im}(f)$ s.t. $f: M \to \mathrm{Im}(f) \subset \{0,1\}^*$ to obtain a surjective, hence bijective function, meaning that $\forall\, \underline{m} \in M$, $\exists!\, \underline{b} \in \mathrm{Im}(f)$ s.t. $f(\underline{m}) = \underline{b}$, hence avoiding ambiguity.

- Why do we like it when a code is prefix-free?

In some channels, such as the telephone line, the communication begins as soon as you dial a valid number.

POLICE TELEPHONE NR: 113
YOUR BF TELEPHONE NR: 1131234.

police telephone number is a prefix of your BF telephone number. If you want to call your friend, as soon as you dial the first 3 digits, you'll get redirected to the police.

<u>def.</u>: Let P be any probability distribution over $\mathcal{M}$, and let $f: \mathcal{M} \to \{0,1\}^*$ be a prefix-free binary encoding.

The AVERAGE LENGTH of f w.r.t. P is

$$\sum_{m \in \mathcal{M}} P(m)\,|f(m)|,$$

where $|f(m)|$ is the length of the string $f(m)$.

($f(m) = t$ if $f(m) \in \{0,1\}^t$).

<u>LEMMA</u> (Kraft's inequality): Let f be a prefix-free binary encoding. Then $\sum_{m \in \mathcal{M}} 2^{-|f(m)|} \leq 1$.

What this result tells us is that for a binary encodining to be prefix-free, it must not use very short encodings because they are the prefix of many other ones.

proof: let L be the maximum length of the binary encodings of the words in M, i.e.:

$$L = \max_{m \in M} |f(m)|.$$

let $Y_L(\underline{x})$ be the set of all the extensions of the string $\underline{x}$ of length L, that is, the set of all the strings of length L s.t. $\underline{x}$ is a prefix:

$$Y_L(\underline{x}) := \{\underline{y} \text{ s.t. } \underline{x} \prec \underline{y}, \underline{y} \in \{0,1\}^L\}.$$

OBS 1: IF $|\underline{x}| > L$, $Y_L(\underline{x}) = \phi$,
IF $|\underline{x}| = L$, $Y_L(\underline{x}) = \{\underline{x}\}$.

OBS 2: IF $\underline{x} \not\prec \underline{z}$, $\underline{z} \not\prec \underline{x}$ $\Rightarrow Y_L(\underline{x}) \cap Y_L(\underline{y}) = \phi$.

proof: by contraddiction, $\exists\, \underline{w} \in Y_L(\underline{x}) \cap Y_L(\underline{y})$.

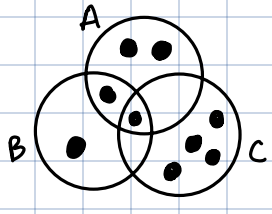
By definition, this means that:
$\underline{x} \prec \underline{w}$ and $\underline{z} \prec \underline{w}$.

but then, either $\underline{x} \prec \underline{z}$ or $\underline{z} \prec \underline{x}$. ↯ □

It holds that $\bigcup_{m \in M} Y_L(f(m)) \subset \{0,1\}^L$, which implies:

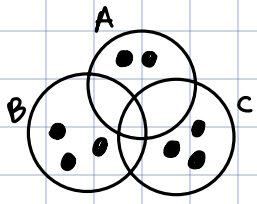
$$\left|\bigcup_{m \in M} Y_L(f(m))\right| < |\{0,1\}^L| = 2^L. \quad (1)$$

Remember that $\left|\bigcup_{m \in M} Y_L(f(m))\right| \leq \sum_{m \in M} |Y_L(f(m))|$,
and equality holds $\Leftrightarrow \bigcap_{m \in M} Y_L(f(m)) = \phi$, that is,
if all the sets $Y_L(f(m))$ are disjoint.



$|A| = 4$, $|B| = 3$, $|C| = 5$.

$|A \cup B \cup C| = 9 < |A| + |B| + |C| = 12$.



$|A|=2,\ |B|=3,\ |C|=3$

$|A \cup B \cup C| = 8 = |A|+|B|+|C|.$

By OBS2, and since f is prefix-free, it follows that:

$$Y_L(f(m)) \cap Y_L(f(m')) = \phi \quad \forall m, m' \in M,\ m \neq m'.$$

Then:

$$\left| \bigcup_{m \in M} Y_L(f(m)) \right| = \sum_{m \in M} |Y_L(f(m))| = \sum_{m \in M} 2^{L-|f(m)|}$$

By equation (1), it follows that:

$$\sum_{m \in M} 2^{L-|f(m)|} \leq 2^L$$

$$2^L \sum_{m \in M} 2^{-|f(m)|} \leq 2^L$$

The thesis follows. □

Why do we need this inequality?

Because it allows us to prove one of the most powerful results in Information Theory.

Entropy is the limit of lossless compression

Theorem: Let $f: M \longrightarrow \{0,1\}^*$ be a binary prefix-free encoding. Let P be a distribution on M. Then, the average length of an encoding f is lower-bounded by the entropy of P:

$$\sum_{m \in M} P(m)\,|f(m)| \geq H(P)$$

proof: the thesis is:

$$\sum_{m \in M} P(m)\,|f(m)| \geq H(P)$$

$$\Leftrightarrow \sum_{m \in M} P(m)|f(m)| - H(P) \geq 0$$

$$\Leftrightarrow \sum_{m \in M} P(m)|f(m)| - \sum_{m \in M} P(m) \lg_2\left(\frac{1}{P(m)}\right) =$$

(by def. of entropy)

$$= \sum_{m \in M} P(m)\left[|f(m)| - \lg_2\left(\frac{1}{P(m)}\right)\right] =$$

$$= \sum_{m \in M} P(m)\left[\lg_2 2^{|f(m)|} - \lg_2\left(\frac{1}{P(m)}\right)\right] =$$

($2^{\lg_2 a} = \lg_2 2^a = a$)

$$= \sum_{m \in M} P(m)\left[\lg_2\left(2^{|f(m)|} P(m)\right)\right] =$$

($\lg_2 a - \lg_2 b = \lg_2 \frac{a}{b}$)

$$= \sum_{m \in M} P(m)\left[\lg_2\left(\frac{P(m)}{2^{-|f(m)|}}\right)\right] =$$

LOG-SUM INEQUALITY:
a_i, b_i, $i = 1, \dots, t$,
$a_i, b_i \geq 0$
$a = \sum_{i=1}^{t} a_i$, $b = \sum_{i=1}^{t} b_i$
Then:
$\sum_{i=1}^{t} a_i \lg \frac{a_i}{b_i} \geq a \lg \frac{a}{b}$

$$\geq \sum_{m \in M} P(m) \cdot \lg_2\left(\frac{\sum_{m \in M} P(m)}{\sum_{m \in M} 2^{-|f(m)|}}\right) \geq \lg 1 = 0$$

(Kraft's inequality)

□

OK, but we asked for the encoding to be prefix-free, which is a strong assumption.

What if we relax this condition and just ask for the encoding to be uniquely decodable? Do we get a better bound?

→ The answer is NO.

ENTROPY IS THE MINIMUM NUMBER OF BITS THAT WE CAN USE TO TRANSMIT A MESSAGE WITHOUT MISSING THE MEANING OF THE MESSAGE, WITHOUT LOOSING SOME INFORMATION

- WE CAN COMPRESS MORE, BUT WE ARE GONNA LOOSE INFORMATION (e.g. JPEG images).

COROLLARY: IF $P \sim \frac{1}{2^{|f(m)|}}$ (uniform distribution) $\Rightarrow \sum_{m \in M} P(m)|f(m)| = H(P)$.

Example: your BF lives in Florence and every day they send you a message to tell you what the weather is like in Florence.

$$\underline{m} \in \{\text{sunny}, \text{foggy}, \text{rainy}, \text{cloudy}\}$$

Assume the 4 weather conditions happen with equal probability:

	S	F	R	C
$P_{\sim}$	1/4	1/4	1/4	1/4

You and your BF can decide for very long prefix-free encodings, eg:

$f(S) = 010010111$
$f(F) = 100010101$
$\vdots$

But if your BF wants to compress information as much as possible, how many bits do they need? Easy, 2.

$$f(S) = 00,\ f(F) = 01,\ f(R) = 10,\ f(C) = 11.$$

What is the entropy of P?

$$H(P) = 4 \cdot \frac{1}{4} \cdot \lg 4 = 2$$ ↦ this means that you BF cannot use less than 2 bits.

What is the average word length?

$4 \cdot \frac{1}{4} \cdot 2 = 2$ ↦ which is equal to $H(P)$ (as expected, since $P_{\sim}$ uniform), this means that we achieved maximum compression

Assume that your BF lives in MILAN.

	S	F	R	C
$P_{\sim}$	1/4	1/2	1/8	1/8

What is a good choice for encoding the messages?

→ Remember: more common words should be shorter!

OPTION 1: $f(S)=01$, $f(F)=1$, $f(R)=000$, $f(C)=001$

→ average message length:

$$= \frac{1}{4}\cdot 2 + \frac{1}{2}\cdot 1 + \frac{1}{8}\cdot 3 + \frac{1}{8}\cdot 3 = \frac{7}{4}.$$

OPTION 2: $f(S)=000$, $f(F)=001$, $f(R)=1$, $f(C)=01$

→ average message length=

$$= \frac{1}{4}\cdot 3 + \frac{1}{2}\cdot 3 + \frac{1}{8}\cdot 1 + \frac{1}{8}\cdot 2 = \frac{21}{8}$$

OPTION 3: $f(S)=00$, $f(F)=01$, $f(R)=10$, $f(C)=11$

average message length:

$$= \frac{1}{4}\cdot 2 + \frac{1}{2}\cdot 2 + \frac{1}{8}\cdot 2 + \frac{1}{8}\cdot 2 = 2$$

OPTION 1 wins! And it is the best encoding, since:

$$H(P) = \frac{1}{4}\lg 4 + \frac{1}{2}\lg 2 + \frac{1}{8}\lg 8 + \frac{1}{8}\lg 8 = \frac{7}{4}$$

Notice that:

1) the entropy of the "Florence" example is larger

than the entropy of the "Milan" example. This is because $P_{Florence}$ is uniform, while P_{Milan} is not.

2) In the "Milan" example, the achieved average length is equal to the entropy, even though P_{Milan} is not uniform.

Nevertheless, in general, if P is not uniform, it could be that the optimal encoding has an average message length that is strictly larger than the entropy:

EXAMPLE:

	S	F	C	R
P ~	5/9	1/9	2/9	1/9

$f(S) = 1$

$f(C) = 01$

$f(F) = 001$

$f(R) = 000$

The average length is:

$$\frac{5}{9} \cdot 1 + \frac{2}{9} \cdot 2 + \frac{2}{9} \cdot 3 = \frac{15}{9} = \frac{5}{3} = 1.\overline{6}$$

Whereas the entropy is:

$$H(P) = \frac{5}{9} \log\left(\frac{9}{5}\right) + \frac{1}{9} \log(9) + \frac{1}{9} \log(9) + \frac{2}{9} \log\left(\frac{9}{2}\right) \approx 1.6577$$

which is strictly smaller.