LECTURE 1
INTRODUCTION TO INFORMATION THEORY

Information Theory

Thomas Debris-Alazard

Inria, École Polytechnique

Information Theory: the great Shannon

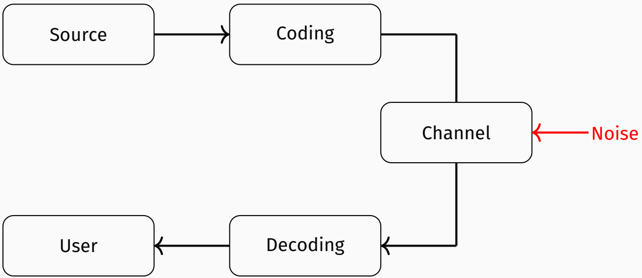
⟶ Without Shannon: no efficient communications, storages!

But implications are much deeper . . .

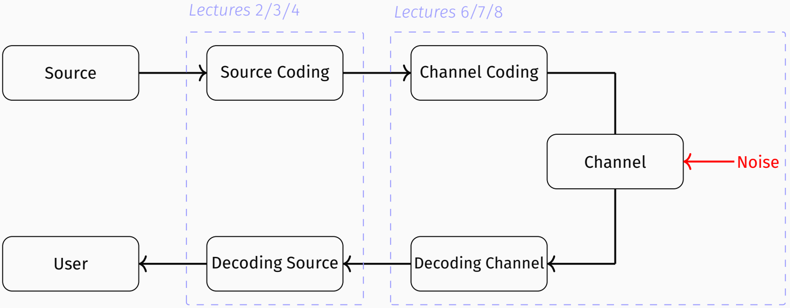
BUT DO WE CARE?

If communications, storages are not efficient, do we only need to improve physical devices?

$\longrightarrow$ Information theory and coding theory offer an alternative (and much more exciting)!



- Source: text, voice, image, video, . . .
- Channel: radio, optical fiber, magnetic support, . . .
- Noise: electromagnetic disturbance , scratches, . . .



- Efficiency: transmit a given quantity of "information" with the minimal amount of resources
- Reliability: provide to users a sufficiently accurate information from the source

- Source coding: remove redundancy/compress as much as possible

An example: compress the language

In French, E is frequent, Z is not

$\longrightarrow$ E is compressed with fewer “symbols” than Z

- Channel coding: add redundancy to recover messages in the presence of noise

An example: spell your name over the phone, send first names!

M like Mike, O like Oscar, R like Romeo, A like Alpha, I like India and N like November

M: **message** ; Mike: **encoding**

Source and Channel coding are “dual”

- Given a source, what is the ultimate data **compression**?

  $\longrightarrow$ Answer: the **entropy** H

- Given a noisy channel, what is the best transmission **rate** of communication?

  $\longrightarrow$ Answer: the **channel capacity** C

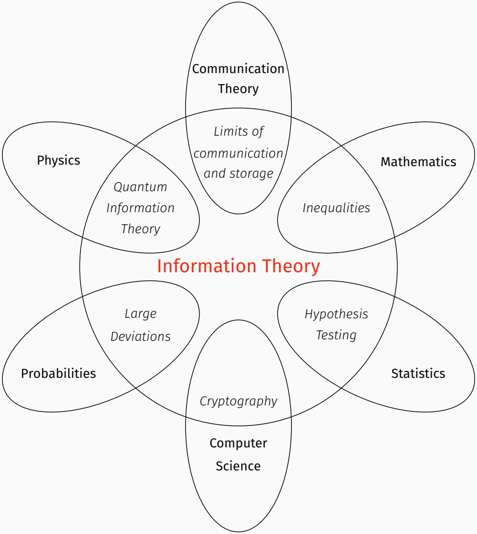
Can we do better?

No!

Can we reach these theoretical limits?

Yes! And we know $\left(\text{surprisingly}\right)$ **efficient solutions/algorithms**!

Information theory is not only about communication and storages. . .



$\longrightarrow$ Basics of information theory and some of its applications

- Theoretical limits for compression and transmission and how to reach them efficiently
- Application to probability and statistics (typical sequences, large deviations)
- Study of linear error correcting codes

References:

- Cover and Thomas, *Elements of Information Theory*,

  $\longrightarrow$ Classical introduction to information theory
- Sendrier's lecture notes: https://www.rocq.inria.fr/secret/Nicolas.Sendrier/thinfo.pdf,

  $\longrightarrow$ Nice for an "algorithmic" point of view
- MacKay, *Information Theory, Inference, and Learning Algorithms*,

  $\longrightarrow$ Nice to get many "intuitions"

EVALUATION OF THIS COURSE

1. An exam (3 hours): 4 pages of personal notes are allowed

   $\longrightarrow$ Three exercises seen during the Exercise Sessions will be at the exam

2. Presentation of a research article or a programming project (30min)

A WARNING

We will be doing a lot of discrete probabilities

$\longrightarrow$ Discrete probabilities need enumeration, no Lebesgue integration

In particular: no hard formalism is involved!

DISCRETE PROBABILITIES

HOW TO MODELIZE A SOURCE? A NOISY CHANNEL?

- A source $\big($language, computer code, ...$\big)$ is modelized according to a discrete random variable

  $\longrightarrow$ See the programming project or ... any generative AI!

- A noisy channel $\big($scratch your parents' CD-ROMs, download a video stored across the world, ...$\big)$ is modelized according to a discrete random variable

  $\longrightarrow$ Very accurate in practice $\big($otherwise no Internet$\big)$

- An alphabet: $\mathcal{X}$ **discrete** $\big($finite in almost all cases in this course$\big)$
- An event: $\mathcal{E} \subseteq \mathcal{X}$
- Random variable: $\mathsf{X} : \Omega \to \mathcal{X}$ $\big($we don't care of $\Omega\big)$
- Probability law / Associated **distribution**: $\big(\mathbb{P}(\mathsf{X} = x)\big)_{x \in \mathcal{X}}$

Abuse of notation:

$$\mathbb{P}(\mathsf{X} = x) = \mathbb{P}_{\mathsf{X}}(x) = p(x)$$

Be careful: given random variables X and Y,

$$p(x) = \mathbb{P}(\mathsf{X} = x) \quad \text{and} \quad p(y) = \mathbb{P}(\mathsf{Y} = y)$$

Remark: the probability law uniquely determines the random variable

Whatever is the event $\mathcal{E}$,

$$\mathbb{P}(\mathsf{X} \in \mathcal{E}) = \sum_{x \in \mathcal{E}} p(x)$$

Notation: $p(x, y)$ denotes

$$\mathbb{P}(X = x \text{ and } Y = y) = \mathbb{P}(X = x,\ Y = y)$$

Random variables X and Y are said to be **independent** if

$$p(x, y) = p(x) \cdot p(y)$$

Important notation: i.i.d.

$X_1, \ldots, X_n$ are said Independent and Identically Distributed (i.d.d.) when they are

1. independent, $\forall \mathcal{I} \subseteq \{1, \ldots, n\}, \forall (x_i)_{i \in \mathcal{I}},\ \mathbb{P}(X_i = x_i,\ i \in \mathcal{I}) = \prod_{i \in \mathcal{I}} \mathbb{P}(X_i = x_i)$
2. identically distributed: $\forall i, j, x,\ \ \mathbb{P}(X_i = x) = \mathbb{P}(X_j = x)$

$$X : \Omega \longrightarrow \mathcal{X}$$

$$\mathbb{E}(X) \stackrel{\text{def}}{=} \sum_{x \in \mathcal{X}} x \, p(x)$$

Transfer formula:

Given $f : \mathcal{X} \longrightarrow \mathbb{C}$,

$$\mathbb{E}\Big(f(X)\Big) = \sum_{x \in \mathcal{X}} f(x) \, p(x)$$

Be careful!

$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$$

is always true (linearity of the expectation)! No independence condition. . .

Exercise: Bernoulli random variables and expectation

Given $X_1, \ldots, X_n$ i.d.d. as Bernoulli random variables of parameter p, *i.e.*, $X_i : \Omega \rightarrow \{0, 1\}$ and $\mathbb{P}(X_i = 1) = p$. Compute,

$$\mathbb{E}\left(\sum_{i=1}^{n} X_i\right)$$

Theorem: weak law of large numbers

$X_1, \ldots, X_n$ be **i.i.d.** with expected value $\mu = \mathbb{E}(X_1) = \cdots = \mathbb{E}(X_n)$. Let,

$$\bar{X}_n \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} X_i$$

Then,

$$\bar{X}_n \xrightarrow[n \to +\infty]{P} \mu = \mathbb{E}(\bar{X}_n), \quad \textit{i.e.}, \forall \varepsilon > 0, \lim_{n \to +\infty} \mathbb{P}\left(\left| \bar{X}_n - \mu \right| < \varepsilon \right) = 1$$

Taking the average of the results obtained from a large number of independent and identical trials tends to become closer to the expected value as more trials are performed

Is expectation enough to "describe" a random variable?

(Spoil: no, but in many cases it is almost enough, it gives us "what we expect")

VARIANCE AND STANDARD DEVIATION

$$X : \Omega \longrightarrow \mathcal{X}$$

Variance: $$\mathbb{V}(X) \overset{\text{def}}{=} \mathbb{E}\Big((X - \mathbb{E}(X))^2 \Big) \underbrace{=}_{\text{linearity of } \mathbb{E}(\cdot)} \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \sum_{x \in \mathcal{X}} x^2\, p(x) - \Big(\sum_{x \in \mathcal{X}} x\, p(x) \Big)^2$$

Standard Deviation: $$\sigma(X) \overset{\text{def}}{=} \sqrt{\mathbb{V}(X)}$$

In practice: expectation good approximation

$X \approx \mathbb{E}(X)$, or more precisely: $X \in [\mathbb{E}(X) - \sigma(X), \mathbb{E}(X) + \sigma(X)]$ with good probability

$\longrightarrow$ **Large deviation theory**: study $\mathbb{P}\big(X \gg \mathbb{E}(X)\big)$

Be careful!

X and Y independent $\Longrightarrow \mathbb{V}(X + Y) = \mathbb{V}(X) + \mathbb{V}(Y)$ $\Big($*the variance is not necessarily additive*$\Big)$

Alphabet $\mathcal{X} \times \mathcal{Y}$ endowed with the probability law $p(x, y)$,

Marginal Law	Conditional Probability
$\mathbb{P}(\mathsf{X} = x) = p(x) = \sum_{y \in \mathcal{Y}} p(x, y)$	$\mathbb{P}(\mathsf{X} = x \mid \mathsf{Y} = y) = p(x\|y) = \frac{p(x,y)}{p(y)}$ (when $p(y) \neq 0$)
$\mathbb{P}(\mathsf{Y} = y) = p(y) = \sum_{x \in \mathcal{X}} p(x, y)$	$\mathbb{P}(\mathsf{Y} = y \mid \mathsf{X} = x) = p(y\|x) = \frac{p(x,y)}{p(x)}$ (when $p(x) \neq 0$)

- Marginal law: the knowledge of $\left(p(x, y)\right)_{(x,y) \in \mathcal{X} \times \mathcal{Y}}$ is enough to know $\left(p(x)\right)_{x \in \mathcal{X}}$
- Conditional probability: what is the probability of x knowing that y_0 happened? Enough to know $\left(p(x, y)\right)_{(x,y) \in \mathcal{X} \times \mathcal{Y}}$.

Law of total probability:

Given disjoint and complete events $\mathcal{B}_1, \ldots, \mathcal{B}_n$, *i.e.*,

1. $\mathcal{B}_i \cap \mathcal{B}_j = \emptyset$ if $i \neq j$
2. $\bigcup_{i=1}^{n} \mathcal{B}_i = \Omega$

Then,

$$\mathbb{P}(X \in \mathcal{E}) = \sum_{i=1}^{n} \mathbb{P}\left(X \in \mathcal{E} \mid \mathcal{B}_i\right) \mathbb{P}(\mathcal{B}_i)$$

One of the most useful fact in probability computations!

Exercise:

A box contains two coins, one is biased to head with probability $1/2 + \varepsilon$, the other one is biased to tail with probability $1/2 + \varepsilon$. You choose a coin uniformly at random and you throw it. What is the probability to get head?

OVERVIEW OF INFORMATION THEORY

Information Theory answers the following two (fundamental) questions:

- Ultimate data compression? Entropy
- Ultimate transmission rate of communication? Channel capacity

$\longrightarrow$ Information Theory is much more!

A common denominator: typical sequences/realisations!

Anecdote:

At the police station, is it easier to answer the following questions: what were you doing three Monday ago? or what were you doing a typical Monday?

$\longrightarrow$ Typical realisations: simple mean to answer hard questions!

TYPICAL OR MOST PROBABLE?

$X_1, \dots, X_n$ be i.i.d. with $\mathbb{P}(X_i = 1) = p < 1/2$

What is the most probable sequence/realisation?

TYPICAL OR MOST PROBABLE?

$X_1, \dots, X_n$ be i.i.d. with $\mathbb{P}(X_i = 1) = p < 1/2$

What is the **most probable sequence/realisation**?

$0 \dots 0$ appears with probability: $(1-p)^n$

$\longrightarrow$ Most probable event!

But do you expect this realisation?

$X_1, \dots, X_n$ be i.i.d. with $\mathbb{P}(X_i = 1) = p < 1/2$

What is the **most probable sequence/realisation**?

$0 \dots 0$ appears with probability: $(1-p)^n$

$\longrightarrow$ Most probable event!

But do you expect this realisation? **No!**

$X_1, \ldots, X_n$ be i.i.d. with $\mathbb{P}(X_i = 1) = p < 1/2$

What is the most probable sequence/realisation?

$0 \ldots 0$ appears with probability: $(1-p)^n$

$\longrightarrow$ Most probable event!

But do you expect this realisation? No!

Hamming weight:

Given $\mathbf{x} = (x_1, \ldots, x_n) \in \{0,1\}^n$, its Hamming weight is defined as

$$|\mathbf{x}| \stackrel{\text{def}}{=} \sharp \{i : \; x_i \neq 0\}$$

Chernoff's bound:

$$\forall \varepsilon > 0, \;\; \mathbb{P}\left(\left|\sum_{i=1}^{n} X_i - np\right| \geq \varepsilon n\right) \leq 2e^{-2\varepsilon^2 n}$$

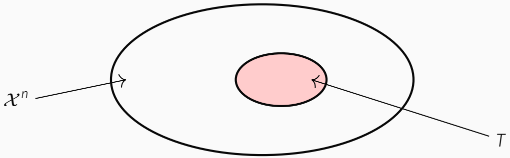
Typical sequence/realisation: $\mathbf{x}$'s such that $|\mathbf{x}| \approx np$

Typical events are an **extremely powerful** tools for proofs!

$\longrightarrow$ and the most important “spirit” of this course. . .

Given a classical source of information $(X_1, \ldots, X_n) \in \mathcal{X}^n$

Your new motto: focus on typical sequences!



$$T \stackrel{\text{def}}{=} \text{typical sequences}$$
$$\mathbb{P}\left((X_1, \ldots, X_n) \in T\right) \approx 1$$

Crucial question:

How many typical sequences are there?

Entropy (informal definition):

$$\text{Entropy}(X_1, \ldots, X_n) \stackrel{\text{def}}{=} \log_2 \sharp T \iff \sharp T = 2^{\text{Entropy}(X_1, \ldots, X_n)}$$

Entropy:

$$H(X_1,\ldots,X_n) \stackrel{\text{def}}{=} -\mathbb{E}\Big(\log_2 \mathbb{P}(X_1,\ldots,X_n)\Big) = -\sum_{x_1,\ldots,x_n\in\mathcal{X}} p(x_1,\ldots,x_n)\cdot\log_2 p(x_1,\ldots,x_n)$$

$\Big(\log_2 \mathbb{P}(X_1,\ldots,X_n)$ random variable outputting $\log_2 p(x_1,\ldots,x_n)$ with probability $p(x_1,\ldots,x_n)\Big)$

Our reasoning to get this formula:

- Non typical sequences $(x_1,\ldots,x_n)$ never appear, *i.e.*,
$$\mathbb{P}(X_1 = x_1,\ldots,X_n = x_n)\approx 0$$
- Typical sequences $(x_1,\ldots,x_n)$ all appear with the "same" probability $\Big($those with smaller probabilities are non-typical$\Big)$ given by their expected value to appear, *i.e.*,
$$\log_2 \mathbb{P}(X_1 = x_1,\ldots,X_n = x_n)\approx \mathbb{E}\Big(\log_2 \mathbb{P}(X_1,\ldots,X_n)\Big) = -H(X_1,\ldots,X_n)$$
$$\textit{i.e.,}\ \ \mathbb{P}(X_1 = x_1,\ldots,X_n = x_n)\approx 2^{-H(X_1,\ldots,X_n)}$$

Conclusion $\Big($informal$\Big)$: T be the set of typical sequences

$$1 = \sum_{x_1,\ldots,x_n} p(x_1,\ldots,x_n) \approx \sum_{(x_1,\ldots,x_n)\in T} p(x_1,\ldots,x_n) \approx \sum_{(x_1,\ldots,x_n)\in T} 2^{-H(X_1,\ldots,X_n)} \approx \sharp T\cdot 2^{-H(X_1,\ldots,X_n)}$$

BINARY ENTROPY

Let's focus on a simple case: $X_1, \ldots, X_n \in \{0,1\}^n$ *be i.i.d with* $p \overset{def}{=} \mathbb{P}(X_i = 1)$

$$H(X_1, \ldots, X_n) = nh(p) \quad \text{where } h(p) \overset{\text{def}}{=} -p \log_2 p - (1-p)\log_2(1-p) \quad \Big(\text{binary entropy}\Big)$$

WHERE (BINARY) ENTROPY IS COMING FROM

Given $(X_1, \cdots, X_n) \in \{0,1\}^n$ be i.i.d with $p \stackrel{\text{def}}{=} \mathbb{P}(X_i = 1)$

Entropy formula is coming from two facts:

(*i*) $\log_2$ maps product into sum

(*ii*) a random variable concentrates around its expectation

$$\begin{aligned}
\log_2 \mathbb{P}\Big((X_1, \ldots, X_n)\Big) &\stackrel{\text{indep}}{=} \log_2 \prod_i \mathbb{P}\Big(X_i\Big) \\
&\stackrel{(i)}{=} \log_2 \mathbb{P}(X_1) + \cdots + \log_2 \mathbb{P}(X_n) \\
&\stackrel{(ii)}{\approx} \mathbb{E}\Big(\log_2 \mathbb{P}(X_1)\Big) + \cdots + \mathbb{E}\Big(\log_2 \mathbb{P}(X_n)\Big) \\
&= (p \log_2 p + (1-p)\log_2(1-p)) + \cdots + (p \log_2 p + (1-p)\log_2(1-p)) \\
&= -nh(p)
\end{aligned}$$

Conclusion (informal):

All non-zero $\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)$ verify

$$\log_2 \mathbb{P}\Big(X_1 = x_1, \cdots, X_n = x_n\Big) \approx -nh(p), \text{ i.e., } \mathbb{P}(X_1 = x_1, \cdots, X_n = x_1) \approx 2^{-nh(p)}$$

$\longrightarrow$ We expect $2^{nh(p)}$ typical sequences $\Big($by using $\sum_{x_1,\ldots,x_n} \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = 1\Big)$!

Two Problematics:

- Source coding: efficient compression of a given source with a maximal compression rate

Realisation: $\mathbf{x} = (x_1, \ldots, x_n) \in \{0,1\}^n$ where $\mathbb{P}(x_i = 1) = p$

Optimal compression size $\approx nh(p)$ bits

- Channel Coding: efficient transmission of a given source through a noisy channel with the minimal amount of redundancy; maximal amount of information bits

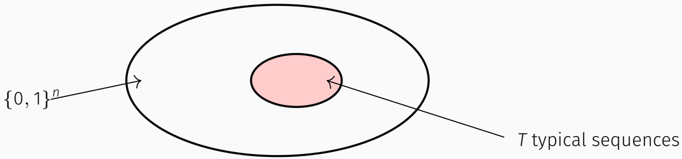
Realisation: $\mathbf{x} = (x_1, \ldots, x_n) \in \{0,1\}^n \rightsquigarrow \mathbf{y} = (y_1, \ldots, y_n) \in \{0,1\}^n$ where $\mathbb{P}(y_i \neq x_i) = p$

Optimal number of bits to transmit $\approx n(1 - h(p))$ bits $\big(nh(p)$ bits of redundancy$\big)$

A common quantity quantifies these limits: entropy (binary entropy in this case)

$$h(p) \stackrel{\text{def}}{=} -p \log_2 p - (1-p)\log_2(1-p)$$

$(X_1, \cdots, X_n) \in \{0,1\}^n$ be i.i.d. with $p \stackrel{\text{def}}{=} \mathbb{P}(X_i = 1)$



$$\mathbb{P}\Big((X_1, \ldots, X_n) \in T \Big) \approx 1$$

Compression algorithm

1. Describe elements of T with bits: it requires $\approx nh(p)$ bits as $\sharp T \approx 2^{nh(p)}$
2. Given a realisation $\mathbf{x}$: if $\mathbf{x} \in T$ describe it with bits, otherwise output fail $\perp$

The compression works with probability ≈ 1 and to decompress we just inverse the bit description of elements in T

Conclusion:

We can compress with $nh(p)$ bits with a success probability ≈ 1

The set of typical sequences T is the smallest set such that $\mathbb{P}\Big((X_1, \dots, X_n) \in T \Big) \approx 1$

$\longrightarrow$ By smaller we mean *exponentially smaller*, *i.e.*, it does not exist S such that $\sharp S = 2^{-cn} \cdot \sharp T$

for some $c > 0$ such that

$$\mathbb{P}\Big((X_1, \dots, X_n) \in S \Big) \approx 1$$

Remark:

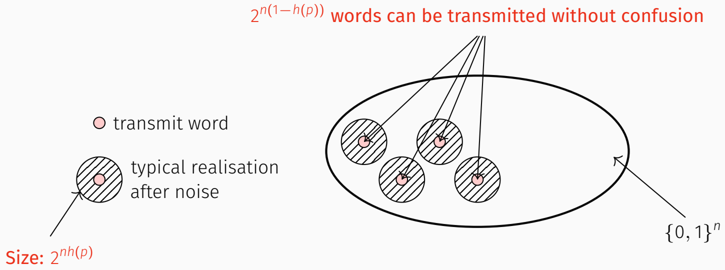
Up to now we did not define rigorously what do we mean by "typical set", wait Lecture 2 and 3

$\longrightarrow$ Conclusion: $\log_2 \sharp T$ is the optimal number of bits to compress!

CHANNEL CODING: GEOMETRIC INTERPRETATION

$(X_1, \cdots, X_n) \in \{0,1\}^n$ be i.i.d. with $p \overset{\text{def}}{=} \mathbb{P}(X_i = 1)$

- Channel Coding: we transmit $\mathbf{c} = (c_1, \ldots, c_n) \in \{0,1\}^n$, the receiver gets $(c_1 + X_1, \ldots, c_n + X_n)$ and wants to recover $\mathbf{c}$



size ball $\times$ words which can be transmitted without confusion $\approx 2^n$

$$\left(2^{nh(p)} \times 2^{n(1-h(p))} = 2^n\right)$$

Typical sequences seem to be useful to prove $\left(\text{sequences } X_i \text{ i.d.d. Bernoulli of parameter } p\right)$

- $nh(p)$ bits for optimal compression
- $n(1 - h(p))$ optimal number of bits which can be transmitted when the noise rate is p

But how to reach these theoretical limits for compression and transmission?

$\longrightarrow$ We will use mathematical objects known as **codes**!

Let's focus on the case of transmission of information

$$\text{To transmit } \mathbf{m} \in \{0,1\}^{k} \xrightarrow{\text{(encoding)}} \mathbf{c} \in \{0,1\}^{n} \xrightarrow[\text{channel}]{\text{noisy}} \mathbf{y} = \mathbf{c} + \mathbf{e}$$

Aim: recover $\mathbf{m}$ from $\mathbf{y}$!

Important Remark:

We mapped k to $n > k$ bits $\left(\text{redundancy}\right)$: $\mathbf{c}$ encoding of $\mathbf{m}$

Your first $\left(\text{error correcting}\right)$ code: 3-repetition code

Encoding 1 bit into 3 bits,

$$\begin{array}{ccc} 0 & \mapsto & 000 \\ 1 & \mapsto & 111 \end{array}$$

$\left\{(000, 111)\right\}$ is called the **three repetition code**!

Exercise:

What does it mean to successfully remove an error with the above encoding? Which error can you successfully remove? Why didn't we introduce the 2-repetition code?

- Encoding: $b \in \{0,1\} \longmapsto bbb \in \{0,1\}^3$
- Noisy Channel: $bbb \longmapsto c_1c_2c_3$ where $\mathbb{P}(c_i \neq b) = p$
- Decoding Strategy: given $c_1c_2c_3 \in \{0,1\}^3$, choose the majority bit

$$001 \longmapsto 0,\ 011 \longmapsto 1,\ 101 \longmapsto 1,\ \text{etc.}\ldots$$

$\longrightarrow$ This strategy is successful if there are < 2 errors

Successful Decoding with probability	Unsuccessful Decoding with probability
$(1-p)^3 + 3p(1-p)^2$	$p^3 + 3(1-p)p^2$

Suppose that $p = 0.01$,

- The decoding procedure fails with probability 3×10^{-4}
- The same decoding procedure with the 5 repetition code fails with probability $\approx 10^{-5}$

Which code will you use for communication?

prob. successfully decoding 5-repetition code $\gg$ prob. successfully decoding 3-repetition code

But. . .

prob. successfully decoding 5-repetition code $\gg$ prob. successfully decoding 3-repetition code

But. . .

Encoding 1 bit necessitates $5 > 3$ bits!

$\longrightarrow$ Higher communication cost with the 5-repetition code. . .

- The 3-repetition code has **rate** $1/3 = 0.33\ldots$
- The 5-repetition code has **rate** $1/5 = 0.2$

Is the rate necessarily go to 0 in order to successfully decoding with probability tending to 1?

No! Second Shannon's theorem

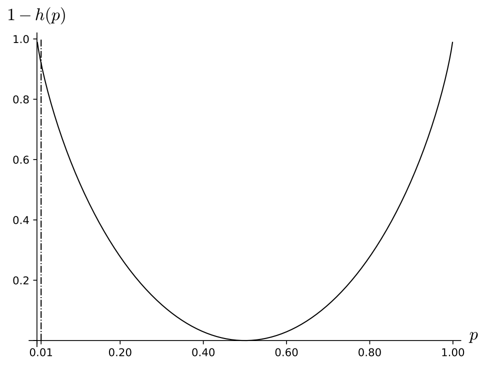
$\longrightarrow \forall$ Rate $\leq$ Channel Capacity

It is possible to decode with probability of success tending to 1!

$p = 0.01$: the 3-repetition code fails to decode with probability 3×10^{-4} with a rate $0.33\ldots$

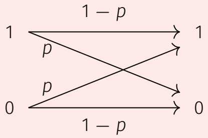
But capacity for 0.01: $C(0.01) = 1 - h(0.01) = 0.919$

We can do **much** better! Even with success probability tending to 1



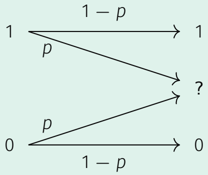
Up to now we considered the following noise model:

Binary Symmetric Channel BSC(p):



$\longrightarrow$ There many other (realistic) channel models! For instance by scratching a CD-ROM you remove bits:

Exercise: Binary Erasure Channel (BEC(p))



Is it "easier" to decode the 3-repetition repetition when BSC or BEC? What do you conclude?

ENTROPY, WHAT ELSE?

Entropy is defined such that number of typical sequences of a random variable X is given by

$$2^{\text{Entropy}(\mathsf{X})}$$

$\longrightarrow$ We need Entropy(X) to describe realisations of X $\left(\text{non-typical sequences "never" appear}\right)$

Informal reasoning:

To enumerate typical sequences:

1. We compute the expected value of $-\log_2 \mathbb{P}(\mathsf{X} = x)$ $\left(\text{over } x\right)$
2. This expected value is defined as the entropy
3. We deduce that the $\mathbb{P}(\mathsf{X} = x)$ are "equal" to $2^{-\text{Entropy}(\mathsf{X})}$ $\left(\text{for typical sequences}\right)$ or 0 $\left(\text{for non-typical sequences}\right)$
4. As probabilities sum to 1, there are $2^{\text{Entropy}(\mathsf{X})}$ typical sequences

Motivated by our discussion on typical sequences, entropy of X *is defined as the average value of*

$$-\log_2 \mathbb{P}(\mathsf{X} = x) = -\log_2 p(x) \quad \left(\textit{over } x \in \mathcal{X}\right)$$

Entropy:

Given $\mathsf{X} : \Omega \to \mathcal{X}$, its entropy is defined as:

$$H(\mathsf{X}) \stackrel{\text{def}}{=} -\sum_{x \in \mathcal{X}} p(x) \cdot \log_2 p(x)$$

with the convention that $0 \cdot \log_2 0 = 0$

Remark:

Given some random variables $\mathsf{X}_1 \in \mathcal{X}_1, \ldots, \mathsf{X}_n \in \mathcal{X}_n$, their $\left(\text{joint}\right)$ entropy is defined as the entropy of $\mathsf{X} \stackrel{\text{def}}{=} (\mathsf{X}_1, \ldots, \mathsf{X}_n)$, *i.e.*

$$H(\mathsf{X}_1, \ldots, \mathsf{X}_n) = -\sum_{x_1 \in \mathcal{X}_1, \ldots, x_n \in \mathcal{X}_n} p(x_1, \ldots, x_n) \log_2 p(x_1, \ldots, x_n)$$

Proposition:

Suppose that $X_1, \ldots, X_n$ are independent, then,

$$H(X_1, \ldots, X_n) = \sum_{i=1}^{n} H(X_i)$$

Proof:

See Exercise Session

Entropy: amount of bits to describe the outcome of a random variable

(think about the example of the compression)

How many bits do we need to describe **X** but when we only know the outcome of **Y**?

$\longrightarrow$ The average value of $-\log_2 \mathbb{P}(\mathsf{X} = x \mid \mathsf{Y} = y) = -\log_2 p(x \mid y)$ $\big($over $x \in \mathcal{X}$ and $y \in \mathcal{Y}\big)$

Conditional entropy:

Given **X** and **Y**, their conditional entropy is defined as,

$$H(\mathsf{X} \mid \mathsf{Y}) \stackrel{\text{def}}{=} -\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x, y) \log_2 p(x \mid y)$$

- $H(\mathrm{X})$: amount of bits to describe possible realisations of X
- $H(\mathrm{X} \mid \mathrm{Y})$: amount of bits to describe realisation of X knowing the realisation of Y

Are Y outcomes help to describe realisation of X?

Mutual information:

Given X and Y, their mutual information is defined as,

$$I(\mathrm{X}, \mathrm{Y}) = H(\mathrm{X}) - H(\mathrm{X} \mid \mathrm{Y})$$

Mutual information is also a measure of dependence between X *and* Y*. If outcomes of* Y *help to describe outcomes of* X*, random variables are dependent whereas in the opposite case they are independent*

Some properties:

- Entropy is maximized when $X : \Omega \to \mathcal{X}$ is uniform,

$$H(X) \leq \log_2 \sharp\mathcal{X} \quad \text{with equality if and only if } X \text{ is uniform}$$

- Mutual information is symmetric,

$$I(X, Y) = I(Y, X)$$

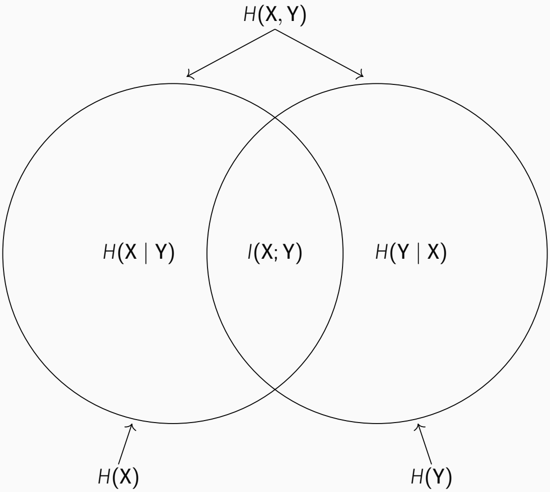
- Mutual information is positive $\Big(\text{how do you interpret this result?}\Big)$

$$I(X; Y) \geq 0 \quad \Big(H(X \mid Y) \leq H(X)\Big)$$

- $H(X, Y) = H(X) + H(Y)$ if X and Y are independent $\Big(\text{how do you interpret this result?}\Big)$

Proof:

See Exercise Session



Usefulness of this picture: for instance (see exercise session for a proof):

$$H(X \mid Y) + H(Y) = H(X, Y) \quad \text{and} \quad H(Y \mid X) + H(X) = H(X, Y)$$

Motivation:

Suppose that we know how X is distributed. But sadly: we are given a random variable $\mathsf{Y} \neq \mathsf{X}$

(you know how to compress outputs of X, not Y)

What do we loose if we would consider that X were given rather than Y?

$\longrightarrow$ Kullback Divergence: measure of the distance between two distributions

(it measures the inefficiency of assuming that X is given when the true random variable is Y)

Kullback-Leibler divergence:

Let $p(x) \stackrel{\text{def}}{=} \mathbb{P}(\mathsf{X} = x)$ and $q(x) \stackrel{\text{def}}{=} \mathbb{P}(\mathsf{Y} = x)$,

$$D_{\mathrm{KL}}(\mathsf{X}||\mathsf{Y}) \stackrel{\text{def}}{=} \sum_{x} p(x) \ \log_2 \frac{p(x)}{q(x)} \in \mathbb{R} \cup \{+\infty\}$$

Be careful: $D_{\mathrm{KL}}(\cdot||\cdot)$ is not symmetric (assuming X given Y $\neq$ assuming Y given X)

Gibb's inequality:

$D_{KL}(X||Y) \geq 0$ with equality if and only if $X = Y$

Gibbs' inequality is probably one of the ***most important inequality*** *in information theory*

Proof:

See Exercise Session

$D_{KL}(\cdot||\cdot)$ *is often useful, not in itself, but because other entropy quantities can be regarded as a special case of* $D_{KL}(\cdot||\cdot)$

EXERCISE SESSION