

AICC II

Arthur Herbette
Prof. Michael Gastpar

Mercredi 5 mars 2025

Table des matières

1	Introduction	5
1.1	About this course	5
1.2	Cours Grading	5
1.2.1	How to be efficient and do well in this course	5
1.3	Initial case : Finite Ω : set of all possible outcomes	6
1.4	Conditional Probability	6
1.5	Conditional probability and Independent Events	7
1.6	Random variable	8
1.6.1	Two random variables	8
1.7	Expected Value	9
1.8	Entropy	10
1.8.1	Information-Theory Inequality	11
1.8.2	Random variables and Entropy	11
1.8.3	Entropy bounds	12
1.9	Source Coding Purpose	12
1.9.1	Setup	12
1.9.2	Codeword length	14
1.9.3	Kraft McMillan	14
1.9.4	Important Consequence of Kraft McMillan	15

Liste des cours

Cours 1 : Discrete Probability — Mardi 18 février 2025	6
Cours 2 : Source and entropy — Mercredi 19 février 2025	10
Cours 2 : suite — Mardi 25 février 2025	12
Cours 4 : Continue — Mercredi 26 février 2025	17
Cours 5 : Conditional Entropy — Mardi 4 mars 2025	17
Cours 6 : Conditional Entropy review — Mardi 4 mars 2025	22

Chapitre 1

Introduction

1.1 About this course

In this course, there will be three main topics that will be studied :

- Communication
- Information and Data science
- Cryptography, Secrecy, Privacy

1.2 Cours Grading

- 90% Final exam during exam period
- 10 % Quizzes (online on Moodle)
 - There will be 6 quizzes. BO5
 - On the quizzes, you can update your answer as many times as you want before the deadline
- Quizzes are highly coorelated with homework.

1.2.1 How to be efficient and do well in this course

Before class :

- Browse through the slides to know what to expect
- review the background material as needed

After class :

- read the notes : they are the reference
- do the review questions

Before the exercice session

- are you up to date with the theory ?
- Solve what you can ahead of time and finish during the exercice session
- write down **your** solution

1.3 Initial case : Finite Ω : set of all possible outcomes

Definition 1 *Sample space Ω is the set of all possible outcomes*

Definition 2 *Event E : a subset of Ω . Since the outcomes are equally likely :*

$$p(E) = \frac{|E|}{|\Omega|}$$

1.4 Conditional Probability

Conditional probability

Definition 3 *The **conditional probability** $p(E|F)$ is the probability that E occurs, given that F has occurred (hence assuming that $|F| \neq 0$) :*

$$p(E|F) = \frac{|E \cap F|}{|F|}$$

Independent Events

Event E and F are called **independent** if $p(E|F) = p(E)$

Personal remark this means that even if we know that F has occurred the probability of E is still the same.

General Case : Finite Ω , arbitrary $p(\omega)$

Having equally likely outcomes is pretty rare in real life, just take two dices and do the sum of the result and you will see that all the possible outcome doesn't have the same probability. In order to express those types of distribution we use the probability mass function :

Definition 4 *Sample space Ω : set of all possible outcomes*
Probability distribution (probability mass function) p :
A function $p : \Omega \rightarrow 1$ such that :

$$\sum_{\omega \in \Omega} p(\omega) = 1$$

If we sum up all the probability it gives us 1.

mass function to a subset Given $E \subset \Omega$ we can define the domain of the probability mass function p is extended to the power set of Ω :

$$p(E) = \sum_{\omega \in E} p(\omega)$$

1.5 Conditional probability and Independent Events

General form The general form for the conditional probability is :

$$p(E|F) = \frac{p(E \cap F)}{p(F)}$$

for F such that $p(F) \neq 0$

Independent events

As before E and F are called independent if $p(E|F) = p(E)$, Equivalently, E and F are independent iff $p(E \cap F) = p(E)p(F)$.

Disjoin event

if E_1 and E_2 are disjoint event then :

$$p(E_1 \cup E_2) = p(E_1) + p(E_2)$$

Law of total probability

For any $F \subseteq \Omega$ and its complement F^c ,

$$p(E) = p(E|F)p(F) + p(E|F^c)p(F^c)$$

which sounds very intuitive because by definition F and F^c are disjoint.

Generally

Theoreme 1 If Ω is the union of disjoint event F_1, F_2, \dots, F_n then :

$$p(E) = p(E|F_1)p(F_1) + p(E|F_2)p(F_2) + \dots + p(E|F_n)p(F_n)$$

Proof

We prove the law of total probability for $\Omega = F \cup F^c$ (the general case follows straightforwardly)

$$\begin{aligned} p(E) &= p(\underbrace{(E \cap F) \cup (E \cap F^c)}_{\text{union of disjoint sets}}) \\ &= p(E \cap F) + p(E \cap F^c) \\ &= \frac{p(E \cap F)}{p(F)}p(F) + \frac{p(E \cap F^c)}{p(F^c)}p(F^c) \\ &= p(E|F)p(F) + p(E|F^c)p(F^c) \end{aligned}$$

Bays' Rule

Theoreme 2

$$p(F|E) = \frac{p(E|F)p(F)}{p(E)}$$

Proof

We use the definition of conditional probability to write $p(E \cap F)$ two ways and solve for $p(F|E)$:

$$p(F|E)p(E) = p(E \cap F) = p(E|F)p(F)$$

1.6 Random variable

Random variable

Definition 5 A Random variable is a function X such as $X : \Omega \rightarrow \mathbb{R}$

Probability distribution

$p_x, p_x(X = x)$ or $p_x(x)$ is the probability that $X = x$, i.e, the probability of the event

$$E = \{\omega \in \Omega : X(\omega) = x\}$$

Hence,

$$p_x(x) = \sum_{w \in E} p(\omega)$$

Example

You rolle a dice.
if the outcome is 6, you receive 10CHF. Otherwise, you pay 1 CHF.

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

$$\text{For each } \omega, p(\omega) = \frac{1}{6}$$

Then define :

$$X(\omega) = \begin{cases} 10, & \omega = 6 \\ -1, & \omega \in \{1, 2, 3, 4, 5\} \end{cases}$$

Hence, we have

$$p_x(X) = \begin{cases} \frac{1}{6}, & x = 10 \\ \frac{5}{6}, & x = -1 \end{cases}$$

1.6.1 Two random variables

Two random variables

Definition 6 Let $X : \Omega \rightarrow \mathbb{R}$ and $Y : \Omega \rightarrow \mathbb{R}$ be two random variables. The probability of the event $E_{x,y} = \{\omega \in \Omega : X(\omega) = x \text{ and } Y(\omega) = y\}$ is :

$$p_{x,y}(x, y) = \sum_{w \in E_{x,y}} p(\omega)$$

- p_x is called **marginal distribution** (of $p_{x,y}(x, y)$ with respect to x)
- p_y can be computed similarly

1.7 Expected Value

Expected value

Definition 7 The expected value $\mathbb{E}[X]$ of a random variable $X : \Omega \rightarrow \mathbb{R}$ is :

$$\begin{aligned}\mathbb{E}[X] &= \sum_{\omega} X(\omega)p(\omega) \\ &= \sum_x xp_x(x)\end{aligned}$$

linearity

Expectation is a linear operation in the following sense :

Let X_1, X_2, \dots, X_n be random variables and $\alpha_1, \alpha_2, \dots, \alpha_n$ be scalars. Then :

$$\mathbb{E}\left[\sum_{i=1}^n X_i \alpha_i\right] = \sum_{i=1}^n \alpha_i \mathbb{E}[X_i]$$

Random variable and independency

Two random variable X and Y are independent if and only if, for all realizations x and y :

$$p(\{X = x\} \cap \{Y = y\}) = p(\{X = x\})p(\{Y = y\})$$

Or, more concisely, iff

$$p_{x,y}(x, y) = p_x(x)p_y(y)$$

Generalization

Theoreme 3 Given n random variables, X_1, \dots, X_n are independent if and only if :

$$p_{x_1, \dots, x_n}(x_1, \dots, x_n) = \prod_{i=1}^n p_{x_i}(x_i)$$

Condition probability

The conditional distribution of Y given X is the function :

$$p_{x,y}(x|y) = \frac{p_{x,y}(x, y)}{p_x(x)}$$

Independent random variables

The following statements are equivalent to the statement that X and Y are two independent random variables :

- $p_{x,y} = p_x p_y$
- $p_{y|x}(y|x) = p_y(y)$
- $p_{y|x}(y|x) = p_y(y)$ is not a function of x
- $p_{x|y}(x|y) = p_x(x)$
- $p_{x|y}(x|y)$ is not a function of y

Summary 1 • Random Variable

- Probability distribution
 - Joint distribution of multiple variables
 - Marginal distribution
 - Conditional distribution
- Independence

Expected value and operation The addition works well with Expectation such that

$$\mathbb{E}[X + Y] = \mathbb{E}[x] + \mathbb{E}[Y]$$

However, the product doesn't work well,

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$$

if and only if X and Y are independent random variables.

1.8 Entropy

Introduction We communicate by revealing the value of sequence of variables that we call **(Symbols), Information**

In modern language, Hartley was saying that the value of a symbol provides information if and only if the symbol is a **random variable**.

How much information is carried by a symbol such as S ?

- Suppose that $S \in \mathcal{A}$ is a symbol that can take $|\mathcal{A}|$ possible values
- The amount of information conveyed by n such symbol should be n times the informations conveyed
- there are $|\mathcal{A}|^n$ possible values for n symbols
- This suggests that $\log |\mathcal{A}|^n = n \log |\mathcal{A}|$ is the appropriate measure for information

However, this approach doesn't work :

Example

Imagine having a town where there are 360 days and 5 rainy days, this leads to have only two possibilities, $|\mathcal{A}| = 2$ which make the quantity of information $\log_2 2 = 1$ bits. Which intuitively sounds kind of false, the forecast doesn't give us that information knowing that it is sunny $\frac{360}{365}$ % of the times, it is kind of expected.

Definition

An article in 1948 from Shannon fixes the problem by defining **Entropy** the **uncertainty** or **entropy** $H(S)$ associated to a discrete random variable S :

Definition 8

$$H_b(S) = - \sum_{S \in \text{supp}(p_s)} p_s(s) \log_b p_s(s)$$

Where $\text{supp}(s) = \{s : p_s(s) > 0\}$.

Few comments

$$H_b(S) = - \sum_{S \in \text{supp}(p_s)} p_s(s) \log_b p_s(s)$$

- The condition $S \in \text{supp}(p_s)$ is needed because $\log_b p_s(s)$ is not define when $p_s(s) = 0$ this convention allows us to use the notation :

$$H_b(S) = - \sum_{s \in \mathcal{A}} p_s(\log_b p_s(s))$$

- The choice of b determines the unit, $b = 2$ is the **bit**

We also can see this as an "average" of $-\log_b p_s(S)$ which is :

$$H(S) = \mathbb{E}[-\log_b p_s(S)]$$

Example

A sequence of 4 decimal digits, s_1, s_2, s_3, s_4 representing the number to open Anne's lock can be senn as the output of a source S_1, S_2, S_3, S_4 with $S_i = \{0, \dots, 9\}$.

If Anne picks all digits at randm and inepedently, the all outcomes are equally likely :

$$p_{S_1, S_2, S_3, S_4}(S_1, S_2, S_3, S_4) = \frac{1}{10^4}$$

If we search the entropy of this we get :

$$H_2(S) = \log_2 |\mathcal{A}| = \log_2 10^4 \approx 13.3 \text{ bits}$$

1.8.1 Information-Theory Inequality

Lemma (IT-Inequality)

lemme 1 For a positive real number r ,

$$\log_b r \leq (r - 1) \log_b(e)$$

with equality if and only if $r = 1$

This proof juste using the deriative

Entropy Bounds

Theoreme 4 The entropy of a discrete random variable $S \in \mathcal{A}$ satisfies :

$$0 \leq H_b(S) \leq \log_b |\mathcal{A}|$$

With equality on the left if and only if $p_s(S) = 1$ and on the right if and only if $p_s(S) = \frac{1}{|\mathcal{A}|}$ for all s .

1.8.2 Random variables and Entropy

n random variable

the formula for entropy can be expanded to any number of random variables. If X and Y are two discrete random variables, with (joint) probability distribution $p_{x,y}$ then :

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

1.4 of text-books

Theoreme 5 Let S_1, \dots, S_n be discrete random variables. Then

$$H(S_1, S_2, \dots, S_n) \leq H(S_1) + H(S_2) + \dots + H(S_n)$$

With equality if and only if S_1, \dots, S_n are independent.

Mardi 25 février 2025 — Cours 2 : suite

**Ex hat party
1950**

- n men, all have the same hat
- they throw hats in a corner
- leaving, they randomly take a hat

Solution

$$\text{Let } R_i = \begin{cases} 1, & \text{if person } i \text{ leaves with their own hat} \\ 0, & \text{otherwise} \end{cases}$$

Entropy

$$H_2(S) = \sum_i p(s) \log \frac{1}{2p(s)} \quad (1.1)$$

$$= \frac{1}{8} \log_2 \frac{8}{2} + \frac{1}{8} \log_2 8 \quad (1.2)$$

$$\approx \frac{1}{8} + \frac{1}{8} \cdot 3 \quad (1.3)$$

*personal re-
mark*

We can see it as an average of "surprise".
Where the average is the randomness. (≈ 0.55)

1.8.3 Entropy bounds

Bound

$$0 \leq H_b(S) \leq \log_b \mathcal{A}$$

1.9 Source Coding Purpose

Source coding is often seen as a way to compress the source.

More generally, the foal of source coding is to efficiently describe how much information there is to a *file*

1.9.1 Setup

Setup

The **encoder** is specified by :

- the input alphabet \mathcal{A} (the same as the source alphabet)
- the output alphabet \mathcal{D} (typically $\mathcal{D} = \{0, 1\}$);
- the codebook \mathcal{C} Which consists of finite sequences over \mathcal{D} ;
- By the one to one encoding map $\Gamma : \mathcal{A}^k \rightarrow \mathcal{C}$ where k is a positive integer.

For now, $k = 1$.

Example	For each code, the encoding map Γ is specified in the following table : A mettre une image. <div> <i>Example</i> Code C or B are uniquely decodable : (A mettre une image 106) </div>
Prefix Free codes	<div> Definition 10 <i>If no codeword is a prefix of another codeword, the code is said to be prefix free.</i> </div> <div> <i>Example</i> The codeword 01 is a prefix of 011. </div> <ul style="list-style-type: none"> • A prefix free code is always uniquely decodable • A uniquely decodable code is not necessarily prefix free <div> <i>A prefix code</i> A prefix free code is also called instantaneous code : <ul style="list-style-type: none"> • Think of phone numbers • Think about streaming : instantaneous codes minimize the decoding delay (for given codeword length) </div>
Code for one random variable	We start by considering codes that encode one single random variable $S \in \mathcal{A}$. To encode a sequence S_1, S_2, \dots of random variables, we encode one random variable at a time.
Complete tree of a code	Slide 113 screen.
Binary tree	<ul style="list-style-type: none"> • There is a root (the beginning) • A vertex (another node) • A leaf is the last vertex • Which is like a (arbre généalogique)
Ternary Tree	The same as a binary tree but with three children.
With/Without prefix	slide 115.
Decoding tree	<ul style="list-style-type: none"> • Obtained from the complete tree by keeping only branches that form a codeword • Useful to visualize the decoding process

Slide 116

1.9.2 Codeword length

- The codeword length is defined the obvious way :
- Example : ct

	\mathcal{A}
	Γ_B
	codeword lengths
	<hr/>
	a
	0
	1
	<hr/>
	b
	10
	2
	<hr/>
	c
	110
	3
	<hr/>
	d
	1110
	4 height

- We would like the average codeword length to be as small as possible.

1.9.3 Kraft McMillan

Part 1. Necessary condition for the code to be uniquely decodable

Theoreme 6 *If a D -ary code is uniquely decodable then its codeword length i_1, \dots, i_M satisfy*

$$D^{-l_1} + \dots + D^{-l_M} \leq 1$$

Kraft's inequality

Example

For code O we have :

$$2^{-2} + 2^{-2} + 2^{-2} + 2^{-2} = 1$$

Recall Kraft McMillan

Theoreme 7

Example A

For code A we have $2^{-1} + 2^{-2} + 2^{-2} + 2^{-2} = 1.25 > 1$.

Kraft-McMillan's inequality is not fulfilled.

There exists no uniquely decodable code with those codeword lengths.

Proof of K-MM Part I

We prove a slightly weaker result, namely that the codeword lengths of prefix free codes satisfy K-MM inequality.

Let $L = \max_i l_i$ be the complete tree's depth.

- There are D^L terminal leaves
- There are D^{L-l_i}
- No two codewords share a terminal leaf (The code is prefix free)
- Hence $D^{L-l_1} + D^{L-l_2} + \dots + D^{L-l_m} \leq D^L$

After dividing both sides by D^L we obtain Kraft's inequality :

$$D^{-l_1} + D^{-l_2} + \dots + D^{-l_M} \leq 1$$

Exercise

What is the **converse** of Kraft McMillan part 1 ?

The **Converse** of Kraft McMillan part 1 is not true (Consider e.g. two codewords : 01 and 0101)

However, the following statement is almost as good :

Theoreme 8 *If the positive integer I_1, \dots, I_M satisfy Kraft's inequality for some positive integer D , then there exists a D -ary **prefix free code** (hence uniquely decodable) that has codewords*

This says that if the inequality is true, then we **can** find D such that there exists a binary prefix which makes it decodable **and** prefix free !

1.9.4 Important Consequence of Kraft McMillan

Part I

Theoreme 9 *If a **D -ary code is uniquely decodable**, then its codeword length I_1, \dots, I_M satisfy Kraft's inequality :*

$$D^{-l_1} + \dots + D^{-l_M} \leq 1$$

Part II

Theoreme 10 *If the positive integer l_1, \dots, l_M satisfy Kraft's inequality for some positive integer D , then there exists a D -ary **prefix free code** that has those codeword lengths.*

The Kraft McMillan theorem implies that any uniquely decodable code can be substituted by a prefix free code of the same codeword lengths.

Prefix free codes

Our focus will be on prefix free codes. Reasons :

- No loss of optimality : codewords can be as short as for any uniquely decodable code ;
- a prefix free codeword is recognized as soon as its last digit is seen :
 - important, e.g. a phone number ;
 - advantageous to limit the decoding delay in, say streaming

Average Code-word length

- The typical use of a code is to encode a sequence of random variables
-

Example

$$\mathcal{A} = \{a, b, c, d\} \quad D = 2$$

Blackboard with table cct $s \in \mathcal{A}$

$$\begin{array}{c} \Gamma(s) \\ l(s) \\ p(s) \\ \hline \end{array}$$

$$\begin{array}{c} a \\ 0 \\ 1 \\ 0.05 \\ \hline b \\ 10 \\ 2 \\ 0.05 \\ c \\ 110 \\ 3 \\ 0.1 \\ d \\ 1111 \\ 4 \\ 0.8 \end{array}$$

$$\mathcal{E}[\text{length}] = 0.05 + 1 + 0.05 \cdot 2$$

Definition 11 Let $l(\Gamma(s))$ be the length of the codeword associated to $s \in \mathcal{A}$. The average codeword length is :

$$L(S, R) = \sum_i p_s(s) l(\Gamma(s))$$

Units

The unit of $L(S, \Gamma)$ are **code symbols**

When $D = 2$, the unit of $L(S, \Gamma)$ are bits.

**Average code-
word length :
Lower Bound**

Theoreme 11 Let $\Gamma : \mathcal{A} \rightarrow \mathcal{C}$ be the encoding map of a D -ary

Proof

We want to prove that :

$$\begin{aligned} H(s) - \sum_s p(s) l(s) &= - \sum_s p(s) \log p(s) - \sum_s p(s) l(s) \\ &= - \sum_s p(s) \log p(s) - \sum_s p(s) \log 2^{l(s)} \\ &= - \sum_s p(s) \log(p(s) \cdot 2^{l(s)}) \leq \dots \end{aligned}$$

Therefore :

$$\begin{aligned}
 &= \sum_s p(s) \log\left(\frac{1}{p(s)} 2^{-l(s)}\right) \\
 &\leq \sum_s p(s) \left(\frac{1}{p(s)} 2^{-l(s)} - 1\right) \cdot C \\
 &= \left(\sum_s 2^{-l(s)} - \sum_s p(s)\right) \cdot C \\
 &\leq 0
 \end{aligned}$$

We know that the left side is less or equal to 1 because of the Kraft Inequality, therefore it is bounded.

Mercredi 26 février 2025 — Cours 4 : Continue

Key observation

The right hand side of :

$$L(S, \Gamma) = \sum_{s \in \mathcal{A}} p(s) l(\Gamma(s))$$

$$H_D(S) = \sum_{s \in \mathcal{A}} p(s) \log_D \frac{1}{p_S(s)}$$

are identical if $l(\Gamma(s))$

- Unfortunately $l(\Gamma(s)) = \log_D \frac{1}{p_S(s)}$ is often not possible (not an integer)
- How about choosing

Theoreme 12 • For every random variable $S \in \mathcal{A}$

Theorem

Theoreme 13 The average codeword length of a D -ary Shannon-Fano code for the random variable S fulfils :

$$H_D(S) \leq L(S, \Gamma_{SF}) < H_D(S) + 1$$

Proof

it suffices to prove the upper bound (we have already proved the lower bound)

First suppose that we could use $l_i = -\log p_i$. The average length would be :

$$L(S, \Gamma) = \sum_i p_i l_i = \sum_i p_i (-\log_D p_i) = H_D(S)$$

Instead we use $l_i = \lceil -\log p_i \rceil < -\log p_i + 1$

Mardi 4 mars 2025 — Cours 5 : Conditional Entropy

Key Idea

Pack multiple symbols into "supersymbols"

- $(S_1, S_2, S_3, \dots, S_n)$

- Now, apply our Main result to such supersymbols

Theoreme 14 *The average codeword-length of a uniquely decodable code Γ for S must satisfy :*

$$H_D(S_1, S_2, \dots, S_n) \leq L((S_1, S_2, \dots, S_n), \Gamma)$$

And there exists a uniquely decodable code Γ_{SF} satisfying :

$$L((S_1, S_2, \dots, S_n), \Gamma_{SF}) < H_D(S_1, S_2, \dots, S_n) + 1$$

Our Next Nug-
get

Understand

Example

Audio recording :

- We can easily anticipate the next image in a video, there

KEy(simple)
Independent

Definition 12 *The source models a seuquence S_1, S_2, \dots, S_n of n coin flips*

So $S_i \in \mathcal{A} = \{H, T\}$ where H stands for heat, T for tails.

$p_{S_i}(H) = p_{S_i}(T) = \frac{1}{2}$ for all $(s_1, S_2, \dots, S_n) \in \mathcal{A}^n$

Not inde-
pendent

Definition 13 *The source models a seuquence S_1, S_2, \dots, S_n of weather conditions.*

So $S_i \in \mathcal{A} = \{S, R\}$, where S stands for sunny and R for rainy

The weather on the first day is uniformly distributed in \mathcal{A} .

For all other days, with probability $q = \frac{6}{7}$ the weather is as for the day before

Conditional
Probability

Recall how to determine the conditional probability :

$$p_{X|Y}(x | y) = \frac{p_{X,Y}(x, y)}{p_Y(y)}$$

It gives the probability of the event $X = x$, given that the event $Y = y$ has occurred.

it is defined for all y for which $p_Y(y) > 0$

Remark

There is good slide with good schema in slide 176-179

Conditional
Expectation of
 X given $Y = y$

$$p_{X|Y}(\cdot | y)$$

is the probability distribution of the alphabet of X , juste like $p_x(\cdot)$

Definition 14 *The conditional expectation of X given $Y = y$ is defined as :*

$$\mathcal{E}[X | Y = y] = \sum_{x \in \mathcal{X}}$$

**Conditional
Entropy of X
given $Y = y$**

$p_{X|Y}(\cdot | y)$ is a probability distribution on the alphabet of X , just like $p_X(\cdot)$. Every probability distribution has an entropy associated to it :

- $p_x(\cdot) \rightarrow H(X)$
- $p_{X|Y}(\cdot | y) \rightarrow H(X | Y = y)$

Definition 15 The conditional entropy of X given $Y = y$ is defined as :

$$H_D(X | Y = y) = - \sum_{x \in \mathcal{X}} p_{X|Y}(x | y) \log_D p_{X|Y}(x | y)$$

Example A faire

**Entropy
Bounds**

Theoreme 15 The conditional entropy of a discrete random variable $X \in \mathcal{X}$ conditioned on $Y = y$ satisfies :

$$0 \leq H_D(X | Y = y) \leq \log_D |\mathcal{X}|$$

With equality on the left iff $p_{X|Y}(x, y) = 1$ for some x , and with equality on the right iff $p_{X|Y}(x | y) = \frac{1}{|\mathcal{X}|}$

The proff is identical to our proof of the basic entropy bounds

Example

Question ?

Do we also have the following entropy bound :

$$H_D(X | Y = y) \stackrel{???}{\leq} H_D(X)?$$

Answer : no.

Example

(Or "counterexample" if better), Juste for ease of calculation, let us set $\delta = 0$ (but this is not necessary for the example to work). Then, we have :

$$H_D(X | Y = 0) = h_D(\varepsilon) \text{ and } H_D(X | Y = 1) = 0$$

where $h_d(\cdot)$ is the binary entropy function (with $\log_D(\cdot)$). But we have :

$$H_D(X) = h_D\left(\frac{1 - \varepsilon}{2}\right)$$

Conditional entropy can either go up or down (if we give the answer the entropy is 0)

**Conditional
Entropy of X
given Y**

The most useful and impactful definition is the *average* conditional entropy of X given $Y = y$, averaged over all values of y under the marginal distribution $p_Y(y)$. Formally, we thus define :

Definition 16 *The conditional entropy X given Y is defined as :*

$$H_D(X | Y) = \sum_{y \in \mathcal{Y}} p_Y(y) \left(- \sum_{x \in \mathcal{X}} p_{X|Y}(x | y) \log_D p_{X|Y}(x | y) \right)$$

Example

For the Bit flipper channel, we have ;

$$H_D(X | Y) = p(Y = 0)H_D(X | Y = 0) + p(Y = 1)H_D(X | Y = 1)$$

We search now :

$$H(X | Y) = p(Y \text{ is Head})H(XY \text{ is head}) + p(Y \text{ is Tail})H(X | Y \text{ is tail}) = \frac{1}{2}.$$

**Conditional
Entropy of X
given Y**

Theoreme 16 *The conditional entropy of discrete random variable $X \in \mathcal{X}$ conditioned on Y satisfies :*

$$0 \leq H_D(X | Y) \leq \log_D |\mathcal{X}|$$

With equality on the left iff for every y there exists and y such that $p_{X|Y}(x | y) = 1$ and with equality on the right iff $p_{X|Y}(x | y) = \frac{1}{|\mathcal{X}|}$ for all x and all y .

This follows directly from our bounds on $H_D(X | Y = y)$

Having $p_{X|Y}$

We know that $p(X | Y) = \frac{1}{|\mathcal{X}|}$ for all y .

$$\begin{aligned} p(x) &= \sum_{y \in \mathcal{Y}} p(y)p(x | y) \\ &= \sum_y p(y) \frac{1}{|\mathcal{X}|} \\ &= \frac{1}{|\mathcal{X}|} \cdot \sum_y \overbrace{p(y)}^{=1} \end{aligned}$$

**Conditioning
Reduces En-
tropy**

The following bound is important and impactful (and also intuitively pleasing!)

Theoreme 17 *For any two discrete random variables X and Y ,*

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

with equality iff X and Y are independent random variables

In words, **On average**, the uncertainty about X can only become smaller if we know Y .

As we have seen, this is not true point-wise : We may have $H_D(X | Y = y) > H_D(X)$ for some values of y .
It works only on average.

Proof

$$\begin{aligned}
 H(X | Y) - H(X) &= \\
 &= \sum_y p(y) \left(- \sum_x p(x | y) \log p(x | y) \right) + \sum_x p(x) \log p(x) \\
 &= \sum_{x,y} p(y) p(x | y) \log \frac{1}{p(x | y)} + \sum_{x,y} p(y | x) p(x) \log p(x) \\
 &= \sum_{x,y} p(x, y) \log \frac{p(x)}{p(x | y)} \\
 &\leq \sum_{x,y} p(x, y) \left(\frac{p(x)}{p(x | y)} - 1 \right) \cdot \log e \\
 &= \sum_{x,y} (p(x)p(y) - p(x, y)) \log(e) \\
 &= \left(\left(\sum_y p(x)p(y) \right) - \left(\sum_x p(x)p(y) \right) \right)
 \end{aligned}$$

Conditional Entropy of $f(x)$ Let X be an arbitrary random variable. Let $f(x)$ be a (deterministic) function of x .

$$H(f(x) | X) = 0$$

Proof

To find this conditional entropy :
Let $Y = f(x)$

$$p(y | y) = \begin{cases} 1, & y = f(x) \\ 0, & y \neq f(x) \end{cases}$$

the probability that y is $f(x)$ is only true if $f(x) = y$.
This implies that the entropy is equal to 0 :

$$H(y | x) = 0$$

Conditioning reduced Entropy

A generalization of the previous bound is also interest to us :

Theoreme 18 For any three discrete random variables X, Y and Z ,

$$H_D(X | Y, Z) \leq H_D(X | Z)$$

With equality iff X and Y are conditionally independent random variables given Z (that is, if and only if $p(x, y | z) = p(x | z)p(y | z)$ for all x, y, z ,

You can see it as make the Z fall which makes it $p(x, y) = p(x)p(y)$

Proof

It is only mathematics :

$$\begin{aligned}
H_D(X | Y, Z) - H_D(X | Z) &= \mathbb{E} \left[\log_D \frac{1}{p_{X|Y,Z}(X | Y, Z)} \right] + \mathbb{E}[\log_D p_{X|Z}(X | Z)] \\
&= \mathbb{E} \left[\log_D \frac{p_{X|Z}(X | Z)}{p_{X|Y}(X | Y, Z)} \right] \\
&= \mathbb{E} \left[\log_D \frac{p_{X|Z}(X | Z) p_{Y|Z}(Y | Z) p_Z(Z)}{niquesamere} \right]
\end{aligned}$$

Mardi 4 mars 2025 — Cours 6 : Conditional Entropy review
Main definitions

We have here two mains definitions :

The entropy for for a "case" of a random variable :

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

And, the conditional entropy on a random variable :

$$\begin{aligned}
H(X | Y) &= \sum_y p(y) H(X | Y = y) \\
&= - \sum_y \sum_x p(x, y) \log p(x | y)
\end{aligned}$$

The main thing to understand here is that $H(X | Y)$ is the *Generalization* of the first definition. It is all the possible values of Y together. This is why we sum up all possible value of y . The second way to write $H(X | Y)$ is like taking all the possible pairs together and calculating the entropy of each pairs.

Main Result : The main result behind this is :

$$\begin{aligned}
0 &\leq H(X | Y = y) \leq \log |\mathcal{X}| \\
0 &\leq H(X | Y) \leq \log |\mathcal{X}|
\end{aligned}$$

And the inequality :

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

Conditional entropy of $f(x)$ Let X be an arbitrary random variable.Let $f(x)$ be a (deterministic) function of x :

$$H(f(x) | x) = 0$$

For example :

$$\begin{aligned}
X &\in \{0, 1, 2, 3\} \\
f(x) &= X \mod 2
\end{aligned}$$

Which is :

$$f(x) = \begin{cases} 0 & \text{if } x \text{ is even} \\ 1 & \text{if } x \text{ is odd} \end{cases}$$

Then,

$$P(f(x) | X) = \begin{cases} 0, & \text{if } x = 0, 2 \\ 1, & \text{if } x = 1, 3 \end{cases}$$

If we now compute the entropy for $X = 0$ and $X = 1$ etc..., we get :

$$H(f(x) | X = 0) = 0$$

$$H(f(x) | X = 1) = 0$$

⋮

Lisa rolls two dice

Lisa rolls two dice and annouces the sum L written as a two digit number. The alphabet of $L = L_1 L_2 = \{02, 03, 04, 05, 06, 07, 08, 09, 10, 11, 12\}$ Where the alphabet of $L_1 = \{0, 1\}$ and the alphabet of $L_2 = \{0, 1, 2, \dots, 9\}$. We are looking for the probability that $L_2 = 2$ knowing that $L_1 = 1$:

$$p_{L_2|L_1}(2 | 1)$$

What we are doing here is the joint distribution :

$$p_{L_2|L_1}(2 | 1) = \frac{P_{L_1, L_2}(1, 2)}{P_{L_1}(1)} = \frac{\frac{1}{36}}{\frac{1}{6}} = \frac{1}{6}$$

After running over all possible values for (i, j) , we obtain :

$L_2 = j \ L_1 = i$	0	1	$p_{L_2}(j)$
0	0	$\frac{3}{36}$	$\frac{3}{36}$
1	0	$\frac{3}{36}$	$\frac{3}{36}$
2	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{2}{36}$
⋮	⋮	⋮	⋮
$p_{L_i}(i)$	$\frac{5}{6}$	$\frac{1}{6}$	height

now you can do the same with conditional probability and then after computing all those value :

$$H(L_2 | L_1) = \frac{5}{6} \cdot 2.857 + \frac{1}{6} \cdot 1.459 = 2.624 \text{ bits}$$

Now we can observe that :

$$2.624 = H(L_2 | L_1) \leq H(L_2) = 3.22$$

Which says that on average, knowing something takes out some randomness

The chain rule for entropy

Recall that the joint entropy of two random variables X, Y is completely naturally defined as :

$$H_D(X, Y) = - \sum_x \sum_y p_{X,Y}(x, y) \log_D p_{X,Y}(x, y)$$

Or as seen earlier :

$$H_D(X, Y) = H_D(X) + H_D(Y)$$

Using the fact the $p_{X,Y}(x, y) = p_X(x)p_{Y|X}(y | x)$, we can write this as :

$$\begin{aligned} H_D(X, Y) &= - \sum_x p_X(x) \left(\sum_Y p_{Y|X}(y | x) \log_D (p_X(x) p_{Y|X}(y | x)) \right) \\ &= - \sum_x p_X(x) \left(\sum_y p_{Y|X}(y | x) (\log_D p_X(x) + \log_D p_{Y|X}(y | x)) \right) \\ &= - \sum_x p_X(x) \left[\left(\sum_y p_{Y|X}(y | x) \log_D p_X(x) \right) + \left(\sum_y p_{Y|X}(y | x) \log_D p_{Y|X}(y | x) \right) \right] \\ &= H(X) + H(Y | X) \end{aligned}$$

Theoreme 19

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

*Professor
remark*

Firstly :

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

Which is either proved by :

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

The relation proved before in words it :

To find the joint entropy of two random variables, we can first calculate the entropy of one of the two, and then add to it the conditional entropy of the second, given the first.

The chain rule entropy

Theoreme 20 *Let S_1, \dots, S_n be discrete random variables. Then :*

$$H_D(S_1, S_2, \dots, S_n) = H_D(S_1) + H_D(S_2 | S_1) + \dots + H_D(S_n | S_1, \dots, S_{n-1})$$

The above result says that the uncertainty of a collection of random variables (in any order) is the uncertainty of the first, plus the uncertainty of the second when the first is known, plus the uncertainty of the third when the first two are known, etc...

Let us see how :

$$\begin{aligned}
 & H(\underbrace{S_1, S_2, \dots, S_{n-1}}_{=Z}, S_n) \\
 &= H(Z) + H(S_n | Z) \\
 &= H(\underbrace{S_1, S_2, \dots, S_{n-2}}_{=Z'}, S_{n-1}) + H(S_n | S_1 \dots, S_{n-1}) \\
 &= H(Z') + H(S_{n-1} | Z') + H(S_n | S_{1,n-1})
 \end{aligned}$$

Until we get $Z' \dots' = S_1$.

Example

Let X, Y, Z be discrete random variables. We have :

$$\begin{aligned}
 H(X, Y, Z) &= H(X) + H(Y | X) + H(Z | X, Y) \\
 &= H(X) + H(Z | X) + H(Y | X, Z) \\
 &= H(Y) + H(X | Y) + H(Z | X, Y) \\
 &= H(Y) + H(Z | Y) + H(X | Y, Z) \\
 &= H(Z) + H(X | Z) + H(Y | X, Z) \\
 &= H(Z) + H(Y | Z) + H(X | Y, Z)
 \end{aligned}$$

Theoreme 21 Let S_1, \dots, S_n be discrete random variables. Then :

$$H(S_1, S_2, \dots, S_n) \leq H(S_1) + H(S_2) + \dots + H(S_n)$$

With equality iff, S_1, \dots, S_n are independent

Proof

$$\begin{aligned}
 H(S_1, S_2, S_3) &= H(S_1) + H(S_2 | S_1) + H(S_3 | S_1, S_2) \\
 &\leq H(S_1) + H(S_2) + H(S_3)
 \end{aligned}$$

**Another way
around**

Sometimes it is convenient to compute the conditional entropy using the chain rule for entropies. For instance :

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

It can be useful to make it easier to compute $H(X | Y)$ because on the right side, it is only marginal entropies with $p \log p$ which are "easy to compute"

corollaire 1

$$\begin{aligned}
 H(X, Y) &\geq H(X) \\
 H(X, Y) &\geq H(Y)
 \end{aligned}$$

The above inequalities follow from the chain rule for entropies and the fact that entropy (condition or not) is nonnegative.

Example

From lisa rolls two dice :

$$H(L_1, L_2) = 3.2744$$

$$H(L_1) = 0.6500$$

$$H(L_2) = 3.2188$$

We compute :

$$H(L_2 \mid L_1) = H(L_1, L_2) - H(L_1) = 3.2744 - 0.6500 = 2.6254$$

$$H(L_1 \mid L_2) = H(L_1, L_2) - H(L_2) = 3.2744 - 3.2188 = 0.056$$

And verify that indeed :

$$H(L_1 \mid L_2) \leq H(L_1) \leq H(L_1, L_2)$$

$$H(L_2 \mid L_1) \leq H(L_2) \leq H(L_1, L_2)$$