parse-numbers = false per-mode = symbol

AICC II

Arthur Herbette
Prof. Michael Gastpar

25 février 2025

Table des matières

Liste des cours

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

2025-02-18 — **Cours 1 : Discrete Probability**

1.3 Initial case : Finite Ω : set of all possiblie 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 occured (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 occured the probability of E is still the same.

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

Having equally likely outcomes is pretty rare in real life, juste take two dices and do the sum of the result and you will se 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 possiblie outcomes*
Probability distribution (probability mass function) *p :*
A function $p : \Omega \to 1$ such that :

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

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

muss 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 $p(F) \neq 0$

Independet 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 $F_1, F_2, \ldots, F_n$ then :*

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

Proof

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

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

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 $X : \Omega \to \mathbb{R}$*

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

$$\mathbb{E}[X] = \sum_{\omega} X(\omega)p(\omega)$$

$$= \sum_{x} xp_x(x)$$

linearity

Expectation is a linear operation in the folowwing sence :
Let $X_1, X_2, \dots, X_n$ be random variables and $\alpha_1, \alpha_2, \dots, \alpha_n$ be scalars. Then :

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

Random variable and independecy

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 :*

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

Summary 1 *• Random Variable*

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

2025-02-19 — **Cours 2 : Source and entropy**

2025-02-25 — **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

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

$$= \frac{1}{8} \log_2 \frac{8}{2} + \frac{1}{8} \log_2 8 \tag{1.2}$$

$$\approx \frac{1}{8} + \frac{1}{8} \cdot 3 \tag{1.3}$$

> *personal remark* We can see it as an average of ”surprise”.
> Where the average is the randomness. (≈ 0.55)

1.7.1 Entropy bounds

Bound

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

1.8 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*

1.8.1 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 $\Gamma : \mathcal{A}^k \to \mathcal{C}$ 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.

> *Example* Code C or B are uniquely decodable : (A mettre une image 106)

Prefix Free codes

Definition 9 *If no codeword is a prefix of another codeword, the code is said to be prefix free.*

> *Example* The codeword **01** is a prefix of **01**1.

- A prefix free code is always uniquely decodable
- A uniquely decodable code is **not necessarily** prefix free

> *A prefix code* A prefix free code is also called instantaneous code :
>
> - Think of phone numbers
> - Think about streaming : instantaneous codes minimize the decoding delay (for given codeword length)

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

- 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

- Obtained from the complete tree by keeping only branches that form a codeword
- Useful to visualize the decoding process

Slide 116

1.8.2 Codeword length

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

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

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

1.8.3 Kraft McMillan

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

Theoreme 4 *If a D-ary code is uniquely decodable then its codeword length $i_1, \dots, i_M$ satisfy*

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

Kraft's inequality

Example For code O we have :

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

Recall Kraft McMillan

Theoreme 5

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 $D^{L-l_i} + D^{L-l_2} + \dots + D^{L-l_m} \leq D^L$

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

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

Exercice 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 6 *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 !

1.8.4 Important Consequence of Kraft McMillan

Part I

Theoreme 7 *If a* ***D-ary code is uniquely decodable****, then its codeword length $I_1, \dots I_M$ satisfy Kraft's inequality :*

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

Part II

Theoreme 8 *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 Codeword 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 ∈A
$\Gamma(s)$
$l(s)$
$p(s)$

a
0
1
0.05

b
10
2
0.05
c
110
3
0.1
d
1111
4
0.8

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

Definition 10 *Let $l(\Gamma(s))$ be the length of the codeword assiociated to $s \in \mathcal{A}$ The average codeword length is :*

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

Units The unit of $L(S, \Gamma)$ are **code symbols**
When $D = 2$, the unit of $L(S, \Gamma)$ are bits.

Average codeword length : Lower Bound

Theoreme 9 *Let $\Gamma : \mathcal{A} \to \mathcal{C}$ 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 :

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

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