AICC II

Arthur Herbette
Prof. Michael Gastpar

Mardi 4 mars 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

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

Theorem 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

Theorem 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 $X_1, X_2, \ldots, X_n$ be random variables and $\alpha_1, \alpha_2, \ldots, \alpha_n$ 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 indepedecy

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

Theorem 3 *Given n random variables, $X_1, \ldots, X_n$ are independent if and only if :*

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

Condition probability

The conditional distrivution 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 folowig statement are equivalent to the statemant that X and Y are two indepedent 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 be revealing the value of sequence of variables that we call (**Symbols**), **Information**
In modern language, Hartley was saying that the value of a symbole 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 mesure for information

However, this approach doesn't works :

Example Imagine having a town where there are 360 days and 5 rainy days, this leads to have only to possibilities, $|\mathcal{A}| = 2$ which make the quantity of information $\log_2 2 = 1$ bits. Which intiutively 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 excpected.

An article in 1948 from Shannon fixes the problem by defining **Entropy**

Definition

the **uncertainty** or **entropy** $H(S)$ associated to a discrete random variable S :

Definition 8

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

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

Few comments

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

- The condition $S \in supp(p_s)$ 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 $S_i = \{0, \ldots, 9\}$.
If Anne picks all digits at randm and indepedently, 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 :

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

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

Theorem 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 textbooks

Theorem 5 *Let $S_1, \dots, S_n$ be discrete random variables. Then*

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

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

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

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)

1.8.3 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 $\Gamma : \mathcal{A}^k \to \mathcal{C}$ 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 10 *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 $S_1, S_2, \ldots$ 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

1.9.2 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.

1.9.3 Kraft McMillan

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

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

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

Kraft's inequality

Example For code O we have :

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

Recall Kraft McMillan

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

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

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

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 :

Theorem 8 *If the positive integer $I_1, \ldots, I_M$ 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

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

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

Part II

Theorem 10 *If the positive integer $l_1, \ldots, l_M$ 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	a	b	c	d
$\Gamma(s)$	0	10	110	1111
$l(s)$	1	2	3	4
$p(s)$	0.05	0.05	0.1	0.8

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

Definition 11 *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 11 *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.

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

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

Theorem

Theorem 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 $l_i = \lceil -\log p_i \rceil < -\log p_1 + 1$

Mardi 4 mars 2025 — **Cours 5 : Conditional Entropy**

Key Idea

Pack multiple symbols into " *supersymbols*"

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

- Now, apply our Main result to such supersymbols

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

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

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

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

Our Next Nugget

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 $S_1, S_2, \ldots, S_n$ 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 $(s_1, S_2, \ldots, S_n) \in \mathcal{A}^n$

Not independent

Definition 13 *The source models a sequence $S_1, S_2, \ldots, S_n$ 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 :

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

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

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

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 :*

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

Conditional Entropy of X given $Y = y$

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

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

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

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

Example A faire

Entropy Bounds

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

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

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

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

Example

Question ?
Do we also have the following entropy bound :

$$H_D(X \mid >= y) \overbrace{\leq}^{???} H_D(X)?$$

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 :

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

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

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

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 :*

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

Example

For the Bit flipper channel, we have ;

$$H_D(X \mid Y) = p(Y = 0) H_D(X \mid Y = 0) + p(Y = 1) H_D(X \mid Y = 1)$$

We search now :

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

Conditional Entropy of X given Y

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

$$o \leq H_D(X \mid Y) \leq \log_D \mid \mathcal{X} \mid$$

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

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

Having $p_{X|Y}$

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

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

Conditioning Reduces Entropy

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

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

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

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 $H_D(X \mid Y = y) > H_D(X)$ for some values of y.
It works only on average.

Proof

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

Conditional Entropy of $f(x)$

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

$$H(f(x) \mid X) = 0$$

Proof

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

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

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

$$H(y \mid x) = 0$$

Conditioning reduced Entropy

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

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

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

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

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