Introduction to Information Theory

Meryem Benammar

15 octobre 2020

Résumé

In this course, we present the various measures of information, defined by C.E Shannon, which allow to describe random variables and their possible interactions. This course is a comprehensive study of these distinct measures as well as the theorems entailed from what Shannon called a "mathematical theory of communications". The proofs are given by means of guidelines and are not given since left to the discretion of the learner.

Table des matières

1 Basics on probability theory

In this section, we present basics on probability theory, and statistics, which will be crucial to the understanding of the course. This is not meant to be a textbook on probability and statistics, but rather a reminder of some definitions, and an introduction to the different notations used in this textbook. For further readings on probability and statistics, do refer to Billignsley's textbook (link on the webpage of the course).

1.1 Scalar random variables

Let X be a random variable defined over the support set $\mathcal{X}$ (set of all possible realizations of X). Depending on whether the support set is continuous (e.g. $\mathcal{X} = \mathbb{R}$), or discrete (e.g. $\mathcal{X} = [1 : M]$), we will distinguish between discrete and continuous random variables.

> Throughout this text book, random variables are always denoted by capital letters X while their realizations are denotes by regular case letters x.

A random variable is defined uniquely by its probability mass function (pmf) P_X if it is discrete, or by its probability density function (pdf) f_X if it is continuous.

A pmf P_X is defined as a function which associates to each possible realization x, the probability that the random variable X be equal to this realization, i.e.,

$$\begin{array}{ccl} \mathcal{X} & \rightarrow & [0,1] \\ x & \rightarrow & P_X(x) = \mathbb{P}(X = x) \end{array}$$

We have by definition of the pmf that

$$\sum_{x \in \mathcal{X}} P_X(x) = 1 \tag{1}$$

> Note that the syntax $\mathbb{P}(X = x)$ reads as the probability of the event that X equals a particular x, hence we will use $\mathbb{P}$ as the probability of event, while the notation $P_X(x)$ reads as the pmf of X evaluated in x.

In the continuous case, the rigorous definition of a pdf invokes more intricate results than the intuitive definition of the pmf. However, one basic property which we will need for pdf is that

$$\int_{x \in \mathcal{X}} f_X(x) \; dx = 1 \tag{2}$$

Definition 1 (Moments) *To each random variable X with pmf P_X/ pdf f_X are associated*

— *An expected value $\mathbb{E}(X)$ (first order moment[a])*

$$\mathbb{E}(X) \triangleq \sum_{x\in\mathcal{X}} x.P_X(x) \text{ or } \int_{x\in\mathcal{X}} x.f_X(x)\ dx. \tag{3}$$

— *A variance $\mathbb{V}(X)$ (second order moment)*

$$\mathbb{V}(X) \triangleq \mathbb{E}\left(X^2\right) - \mathbb{E}^2(X). \tag{4}$$

a. The notation $\triangleq$ is used when defining a notion for the first time.

Properties 1 (Moments) *The expectation and variance have the following properties*

1. *The expectation is a linear transformation, i.e.,*

$$\mathbb{E}(f(X)) = f(X) \tag{5}$$

for any linear transformation f.

2. *For any constant α, we have that*

$$\mathbb{V}(\alpha.X) = \alpha^2\mathbb{V}(X).$$

Examples 1 *In the following, we list some examples of probability laws which will be of interest throughout this course.*

1. *Bernoulli of parameter p*
2. *Discrete uniform over the interval $[1 : K]$*
3. *Binomial with parameter p*
4. *Gaussian with mean μ and variance σ^2*
5. *Continuous uniform over an interval $[a, b]$*

Exercise 1 : *For each of these laws,*

— *give an experiment in which you can encounter the law (throwing dice, picking balls in an urn, ...)*

— *give the formula of the pmf/pdf*

— *compute the expectation*

— *compute the variance.*

1.2 A pair of random variables

In this section, we introduce notations and basics for bi-dimensional distributions (pairs of random variables), before we extend it later to multi-dimensional distributions (vectors of random variables).

Let (X, Y) be a pair of random variables defined over the product support set $\mathcal{X} \times \mathcal{Y}$.

Definition 2 (Joint, marginal and conditional laws)

The joint pmf $P_{X,Y}$ associated with the pair (X, Y) is given by

$$\begin{array}{rcl} \mathcal{X} \times \mathcal{Y} & \to & [0,1] \\ (x,y) & \to & P_{X,Y}(x,y) = \mathbb{P}(X = x \text{ and } Y = y) \end{array}$$

The marginal pmfs associated with $P_{X,Y}$ are defined by

$$P_X(x) = \sum_{y \in \mathcal{Y}} P_{X,Y}(x,y) \ , \ P_Y(y) = \sum_{x \in \mathcal{X}} P_{X,Y}(x,y) \tag{6}$$

The conditional pmfs associated with $P_{X,Y}$ are defined by

$$P_{X|Y}(x|y) = \frac{P_{X,Y}(x,y)}{P_Y(y)} \ , \ P_{Y|X}(y|x) = \frac{P_{X,Y}(x,y)}{P_X(x)} \tag{7}$$

Properties 2

The joint, conditional and marginal distributions verify the following properties.

1. *Independence : X and Y are independent random variables iif,*

$$\forall (x,y) \qquad P_{X,Y}(x,y) = P_X(x) P_Y(y) \tag{8}$$

2. *Bayes' formula : assume that only $P_{Y|X}$ and P_X are known, then*

$$P_{X|Y}(x|y) = \frac{P_X(x) P_{Y|X}(y|x)}{\sum_{x'} P_X(x') P_{Y|X}(y|x')} \tag{9}$$

Exercise 2 : *Prove Bayes' formula.*

Examples 2 (Classical channel models) *In the following, we give two classical examples of channels encountered in communication systems, namely, the Binary Symmetric Channel (BSC) and the Binary Erasure Channel (BEC).*

The Binary Symmetric Channel BSC(p) is defined by $\mathcal{X} = \mathcal{Y} = \{0, 1\}$ and $P_{X,Y}$,

Y, X	0	1
0	$\frac{1-p}{2}$	$\frac{p}{2}$
1	$\frac{p}{2}$	$\frac{1-p}{2}$

$\Rightarrow$

$Y\|X$	0	1
0		
1		

p is called the crossover probability. This channel model can also be written in a additive form,

$$Y = X \oplus W \tag{10}$$

where X is a Bern(0.5) random variable, W is a Bern(p) random variable, X and W are independent, and $\oplus$ is the binary XOR operation.

***Exercise 3** : Given $P_{X,Y}$ of the BSC(p) channel, compute $P_{Y|X}$ for all possible (x, y). Check that the additive channel model yields the same conditional probability.*

Assume that $\mathcal{X} = \{0, 1\}$ and $\mathcal{Y} = \{0, 1, E\}$ and that $P_{X,Y}$ is given in the table

Y, X	0	1
0	$\frac{1-e}{2}$	0
1	0	$\frac{1-e}{2}$
E	$\frac{e}{2}$	$\frac{e}{2}$

$\Rightarrow$

$Y\|X$	0	1
0		
1		
E		

e is called an erasure probability.

***Exercise 4** : Given $P_{X,Y}$ of the BEC(e), compute $P_{Y|X}$ for all possible (x, y). Hint : compute first the marginal law P_X.*

1.3 Random vectors

Let be $(X_1, \ldots, X_n)$ a vector of n random variables, and defined over the support set $\mathcal{X}_1 \times \ldots \mathcal{X}_n$. Similarly to pairs of random variables, we can define a joint pmf and a collection of marginal and conditional pmfs.

Definition 3 (Joint and marginal pmfs)

The joint pmf of the vector can be defined as $P_{X_1,\ldots,X_n}$

$$\begin{aligned} \mathcal{X}_1 \times \ldots \mathcal{X}_n &\rightarrow [0, 1] \\ (x_1, \ldots, x_n) &\rightarrow P_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) \end{aligned}$$

To the joint pmf $P_{X_1,\ldots,X_n}$ are associated n marginal pdfs

$$P_{X_i}(x_i) = \sum_{(x_1,\ldots,x_n)\setminus x_i} P_{X_1,\ldots,X_n}(x_1, \ldots, x_n)$$

In the following, we state one main property which is satisfied by the joint pmf of a vector, and which will be crucial in this course, namely, the chain rule.

Properties 3 (The chain rule)

The joint pmf can be expanded using the so-called chain rule as follows

$$\begin{aligned} P_{X_1,\ldots,X_n}(x_1, \ldots, x_n) &= \prod_{i=1}^{n} P_{X_i|X_1,\ldots,X_{i-1}}(x_i|x_1, \ldots, x_{i-1}) \\ &= \prod_{i=1}^{n} \frac{P_{X_1,\ldots,X_n}(x_1, \ldots, x_n)}{P_{X_1,\ldots,X_{i-1}}(x_1, \ldots, x_{i-1})} \end{aligned}$$

Example 1 (iid variables) *A set of variables $(X_1, \ldots, X_n)$ are deemed pairwise independent, if their joint pmf verifies*

$$P_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} P_{X_i}(x_i),$$

If, further, the variables are identically distributed, i.e., they follow the same law P_X, then

$$P_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} P_X(x_i).$$

Such variables are said to be independent and identically distributed (iid). In the following, we will often encounter such vectors of iid variables.

1.4 Law of large numbers

In the following, we state one of the main results of probability theory and statistics which will be used extensively in Shanon's information theoretic theorems.

Theorem 1 (Law of Large Numbers (LLN))

Let $(X_1, \ldots, X_n)$ be n iid random variables, with pmf P_X, and let $\mu = \mathbb{E}(X)$ be the expectation of X.

The average mean $\bar{X}_n$ of $(X_1, \ldots, X_n)$, defined by

$$\bar{X}_n \triangleq \frac{1}{n} \sum_{i=1}^{n} X_i \tag{11}$$

converges in probability, as n, grow infinite, to μ, i.e.,

$$\mathbb{P}\left(\lim_{n\to\infty} \bar{X}_n = 1 \right) = 1. \tag{12}$$

1.5 Kullback-Leibler divergence

In the following, when introducing information measures, and more specifically, entropy and mutual information, we will need to resort to a measure of distance between two distribution probabilities. There exist many distance measures between probability distributions, but the one which will be of most use to us, will the Kullback Leibler divergence.

Definition 4 (Kullback-Leibler (KL) divergence)

Let P_X and Q_X be two probability distributions defined on a support set $\mathcal{X}$, such that

$$\sum_{x\in\mathcal{X}} P_X(x) = \sum_{x\in\mathcal{X}} P_X(x) = 1.$$

The KL divergence between P_X and Q_X is defined as

$$D_{KL}(P_X||Q_X) \triangleq \sum_{x\in\mathcal{X}} P_X(x) \log\left(\frac{P_X(x)}{Q_X(x)}\right). \tag{13}$$

KL-divergence is defined only between laws which share the same support set.

Properties 4 *The KL divergence verifies a certain set of properties :*

1. *Asymmetry :* $D_{KL}(P_X||Q_X) \neq D_{KL}(Q_X||P_X)$.
2. *Null element :* $D_{KL}(P_X||P_X) = 0$.
3. *Positivity :* $D_{KL}(P_X||Q_X) \geq 0$, *for all laws* (P_X, Q_X)*0*

Due to the fact that the KL-divergence is asymmetric, it is denoted as a divergence and not as a distance. (The triangular inequality is yet to be proved if we had wanted to name it a distance.

KL-divergence however allows to assess the distance between two distributions, in the sense that it is equal to 0 only when P_X and Q_X are equal, and thus, it allows to test whether two distribution probabilities are close enough to one another.

Exercise 5 *: Prove the properties of the KL divergence. Hint : for the positivity, use the convexity of the log function, or use a Lagrangian to find the minimum of the KL divergence, over all* Q_X *with a* P_X *fixed.*

Conclusions 1 *After this part of the course, you should be able to :*

- *List main discrete and continuous probability distributions*
- *Compute their expectation and variance*
- *Distinguish joint conditional and marginal pmfs/pdfs*
- *Enunciate the chain rule, and its applications*
- *Enunciate the law of large numbers and its applications*
- *Define the KL-divergence and assess it*

2 Entropy

Let X be a random variable with probability support set $\mathcal{X}$ (all values of the support set are possible). Depending on whether the support set $\mathcal{X}$ is discrete or continuous, we can define two types of entropy : discrete entropy $H(X)$ and differential entropy $h(X)$. These two entropies share some common characteristics, but differ in many aspects. We start in the following with the discrete entropy, whilst the differential entropy will be introduced later in the textbook.

2.1 Discrete scalar case : discrete entropy

Let us assume that the random variable X is discrete-valued, and that its support set $\mathcal{X}$ is finite with cardinality $|\mathcal{X}|$. Let the probability mass function (pmf) of X be denoted by $P_X(\dot{})$.

Definition 5 (Discrete entropy) *The entropy of X is defined by :*

$$H(X) \triangleq -\sum_{x\in\mathcal{X}} P_X(x)\log_2\left(P_X(x)\right). \tag{14}$$

where $\log_2$ is the logarithm in base 2.

The entropy of a variable X is a measure related only to the values of the pmf P_X and not to the values of X itself. We thus say that the entropy does not carry any semantic information on X, but measures only the quantity of information contained in X. This information is due to the randomness, or equivalently, the uncertainty, on the random variable X. When defined with base-2 logarithms, the entropy is measures in bits, when defined with natural logarithm (ln), the entropy is measured in nats.

Examples 3 *Here are a few special cases of entropy :*

1. *The entropy of a constant random variables, $X = c$ with probability 1, is given by*

$$H(X) = 0. \tag{15}$$

2. *The entropy of a random variable X uniformly distributed over $\mathcal{X}$, c.à.d, $P_X(x) = \dfrac{1}{|\mathcal{X}|}$ for all $x \in \mathcal{X}$, is given by*

$$H(X) = \log_2(|\mathcal{X}|). \tag{16}$$

3. *Let X be a binary Bernoulli random variable $Bern(p)$ where $p = P_X(1)$. The entropy of X is given by*

$$H(X) = H_2(p) \triangleq -p\log_2(p) - (1-p)\log_2(1-p) = H_2(p). \tag{17}$$

$H_2(p)$ is commonly known as the binary entropy function. It is maximal when $p = 0.5$ (equals 1 bit), and is minimal when $p = 0$ or $p = 1$, (equals 0 bits). It is symmetric around $p = 0.5$.

Exercise 6 *: Compute the entropy of the three previous examples. Plot the binary entropy function $H_2(p)$ as a function of p (you can use Matlab for instance)*

In the following, we list some basic properties of discrete entropy.

Properties 5 *Discrete entropy $H(X)$ satisfies the following properties :*

1. *Minimum : $H(X) \geq 0$ for all discrete distributions P_X,and the minimum is achieved for a degenerate distribution, i.e., X is a constant.*
2. *Maximum : $H(X) \leq \log_2(|\mathcal{X}|)$ for all discrete distributions P_X, and the maximum is achieved for the uniform distribution over $\mathcal{X}$.*
3. *Data Processing Inequality (DPI) : the entropy of a function f of X, is no greater than the entropy of X, i.e.,*
$$H(f(X)) \leq H(X) \tag{18}$$
with equality iif f is a one-to-one function.

Exercise 7 *: Prove the second property (maximal value of entropy). Hint : use a Lagrangian in order to optimize over P_X, or use Jensen's inequality (concavity of the logarithm).*

It is only natural that, being a measure of uncertainty, the entropy be minimal for those variables with little uncertainty , constant one in the extreme, and maximal for those very uncertain variables, the uniform one in the extreme.

2.2 Joint and conditionnal entropy

Now that we have defined the entropy of a random variable, we will introduce the joint entropy of a pair of random variables.

Definition 6 (Joint entropy) *Let X and Y be two random variables with respective finite supports $\mathcal{X}$ and mathcalY assumed jointly distributed following $P_{X,Y}$. The joint entropy of (X, Y) is defined by*

$$H(X, Y) \triangleq - \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} P_{X,Y}(x, y) \log_2 (P_{X,Y}(x, y)) \,. \tag{19}$$

The joint entropy describes, similarly to the scalar entropy, the amount of randomness contained in the random pair of variables (X, Y). It verifies the following properties.

Properties 6 *The joint entropy $H(X, Y)$ satisfies the following :*

1. *Symmetry : $H(X, Y) = H(Y, X)$ for all joint distribution $P_{X,Y}$*
2. *Minimum : $H(X, Y) \geq 0$ for all $P_{X,Y}$, and the minimum is achieved when (X, Y) is a pair of constants.*
3. *Upper bound :*
$$H(X, Y) \leq H(X) + H(Y) \tag{20}$$
with equality iif X et Y are independent, i.e., $P_{X,Y} = P_X P_Y$.
4. *Maximum : $H(X, Y) \leq \log_2(|\mathcal{X}|) + \log_2(|\mathcal{Y}|)$ for all $P_{X,Y}$, and the maximum is achieved for a pair of independent uniform random variables (X, Y), i.e.,*
$$P_{X,Y}(x, y) = P_X(x)P_Y(y) = \frac{1}{|\mathcal{X}|}\frac{1}{|\mathcal{Y}|} \text{ pout tout } (x, y) \in \mathcal{X} \times \mathcal{Y} \tag{21}$$

Exercise 8 : *Prove the property 4 (maximum of the joint entropy). Hint : use property 3 to upper bound the joint entropy.*

The fact that the joint entropy is always smaller that the sum of the individual entropy is due to the fact that, when entangled, two random variables exhibit less randomness than on their own. This intuition can be encountered as well in the thermodynamic entropy.

In the following, we define another measure of information induced between two random variables, namely, conditional entropy.

Definition 7 (Conditional entropy) *Let X and Y be two random variables with finite support sets $\mathcal{X}$ and mathcal$Y$ joint pmf $P_{X,Y}$. The conditional entropy of X knowing Y is defined by :*

$$\begin{aligned} H(X|Y) &\triangleq -\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} P_{X,Y}(x,y)\log_2\left(P_{X|Y}(x|y)\right) &(22)\\ &= \sum_{y\in\mathcal{Y}} P_Y(y)H(X|Y=y) &(23)\end{aligned}$$

where $H(X|Y=y)$ is the entropy of the conditional pmf $P_{X|Y=y}$, and can be written as

$$H(X|Y=y) \triangleq -\sum_{x\in\mathcal{X}} P_{X|Y}(x|y)\log_2\left(P_{X|Y}(x|y)\right). \tag{24}$$

Conditional entropy $H(X|Y)$ measures the amount of uncertainty on X remaining after having observed Y. It consists thus in the first measure of correlation between X and Y.

Properties 7 *Conditional entropy $H(X|Y)$ satisfies the following properties*

1. *Asymmetry : $H(X|Y) \neq H(Y|X)$*
2. *Minimum : $H(X|Y) \geq 0$ for all $P_{X,Y}$, and the minimum is achieved when $P_{X|Y}$ is degenrate, i.e., X is a function of Y.*
3. *Upper bound : conditional entropy is always smaller than the individual entropy*
$$H(X|Y) \leq H(X) \tag{25}$$
with equality iif X and Y are independent : $P_{X,Y} = P_X P_Y$.
4. *Maximum : $H(X|Y) \leq \log_2(|\mathcal{X}|)$ for all $P_{X,Y}$ and the maximum is achieved when X and Y are independent, and X is uniform, i.e.,*
$$P_{X|Y}(x|y) = P_X(x)\frac{1}{|\mathcal{X}|} \text{ pout tout } (x,y)\in\mathcal{X}\times\mathcal{Y} \tag{26}$$

Exercise 9 : *Prove the properties 1, 2, and 4. Prove that if X and Y are independent, then, $H(X|Y) = H(X)$.*

Conditional entropy $H(X|Y)$ is always smaller than the individual entropy, since, having observed a variable Y, possibly correlated to X, the uncertainty about X cannot be grater than when nothing is observed. We say thus that *conditioning decreases entropy, hence, uncertainty.*

Exercise 10 *: Compute the conditional entropy $H(Y|X)$ where $Y = X \oplus W$, where X follows a Bern(1/2), independent from W which follows a Bern(p) and $\oplus$ is the binary XOR operation. (Hint : $P_{X,Y}$ and $P_{Y|X}$ were given previously in example 2)*

Let us now relate the different measures of information introduced previously.

Properties 8 *The individual entropies $H(X)$ and $H(Y)$, the joint entropy $H(X,Y)$ and the conditional entropies $H(X|Y)$ et $H(Y|X)$ can be related as follows*

$$\begin{aligned} H(X,Y) &= H(X) + H(Y|X) && (27)\\ &= H(Y) + H(X|Y) && (28) \end{aligned}$$

Exercise 11 *: Prove the relationships listed herebefore.*

2.3 Vector case : vector entropy

In communication systems, and consequently in Shannon's theory as well, the random variables we are dealing with consist in random processes (vectors of random variables) where the dimension of time or frequency is taken into account. To this end, we will need to define information measures in this vector case. D

Definition 8 (Vector entropy) *Let $X^n = (X_1, ..., X_n)$ be a collection of random variables with support set $\mathcal{X}^n = \mathcal{X}_1 \times ... \times \mathcal{X}_n$ and joint pmf $P_{X^n} = P_{X_1,...,X_n}$. The vector joint entropy is defined as*

$$\begin{aligned} H(X^n) = H(X_1, ..., X_n) &= \sum_{x^n \in \mathcal{X}^n} P_{X^n}(x^n) \log_2(P_{X^n}(x^n)) && (29)\\ &= \sum_{(x_1,...,x_n) \in \mathcal{X}^n} P_{X_1,...,X_n}(x_1, ..., x_n) \log_2(P_{X_1,...,X_n}(x_1, ..., x_n)) && (30) \end{aligned}$$

Similarly to the joint entropy of a pair of random variables, the vector entropy verifies a number of properties, listed hereafter.

Properties 9 *The vector entropy satisfies the following :*

- *Symmetry : $H(X^n) = H(\Pi(X^n))$ for all permutation of indices $\Pi()$ over $[1:n]$*
- *Minimum : $H(X^n) \geq 0$ for all joint pmf P_{X^n}, and the minimum is achieved when $(X_1, ..., X_n)$ is a vector of constants.*

— *Upper bound : the joint entropy is no greater than the sum of individual entropies*

$$H(X^n) \leq \sum_{i=1}^{n} H(X_i) \tag{31}$$

with equality iif all X_i are independent.

In practice, we scarcely compute the vector entropy with direct calculations. Rather, we use the following property.

Properties 10 *The vector entropy $H(X^n)$ writes as*

$$H(X^n) = H(X_1, ..., X_n) = \sum_{i=1}^{n} H(X_i | X_1, ..., X_{i-1}) \tag{32}$$

$$= \sum_{i=1}^{n} H(X_i | X_{i+1}, ..., X_n). \tag{33}$$

Exercise 12 : *Prove this property. Hint : use the chain rule of probabilities*

The two ways of writing the vector entropy, are called causal and anti-causal expansions of the vector entropy. Since the joint entropy is invariant to permutations, the causal and anti-causal expressions are just two specific cases of possible joint entropy expansions. Resorting to other permutations, along with the chain rule, could give many more expansions.

Example 2 (Vector of iid random variables) *Let X^n be a vector of n iid random variables $(X_1, ..., X_n)$, with same support set $\mathcal{X}$. We have that :*

$$H(X^n) = nH(X). \tag{34}$$

We say that $H(X^n)$ admits a single letter expression, and this property is key in Shannon's results, in that it is way much easier to compute a scalar entropy rather $H(X)$ than a vector entropy $H(X^n)$. The quantity $\frac{H(X^n)}{n}$ is often called entropy rate.

2.4 Continuous case : differential entropy

Let us assume in the following that the random variable X is a continuous with support $\mathcal{X}$ (often an interval in $\mathbb{R}$ or a convex are in $\mathbb{C}$). Let $f_X(\dot{})$ be the pdf of X. Let us define the differential entropy of X.

Definition 9 (Differential entropy) *The differential entropy of X is given by*

$$h(X) = \int_{x \in \mathcal{X}} f_X(x) \log_2 (f_X(x)) \; dx \tag{35}$$

assuming that the integral does exist.

The differential entropy is denoted by $h(X)$, contrary to the discrete entropy which is denoted $H(X)$. This is to highlight their intrinsic differences.

Similarly to the discrete entropy, differential entropy computes the amount of randomness and uncertainty pertaining a random variable. Yet, the properties of both these measures differ considerably.

Example 3 *Hereafter, we give a few examples of differential entropy, prior to discussing its properties.*

1. *The differential entropy of a real-valued Gaussian random variable $X_G \sim \mathcal{N}(\mu, \sigma^2)$ is given by*
$$h(X_G) = \frac{1}{2} \log_2(2\pi e \sigma^2) \tag{36}$$
2. *The entropy of a circular complex valued Gaussian variable $X_{CG} \sim \mathcal{CN}(\mu, \sigma^2)$ is given by :*
$$h(X_{CG}) = \log_2(2\pi e \sigma^2) \tag{37}$$

Exercise 13 *: Prove that the entropy of a real-valued Gaussian variable is as stated in property 1.*

Properties 11 *Let X be a continuous random variables with finite variance $\mathbb{V}(X)$. The differential entropy of X, $h(X)$, satisfies the following properties.*

— *Maximum : it is maximum for a Gaussian distribution with the same variance as X , i.e., if X is real-valued,*
$$h(X) \leq \frac{1}{2} \log_2 \left(2\pi e \mathbb{V}(X)\right) \tag{38}$$
and if X is complex valued with covariance matrix K_X
$$h(X) \leq \frac{1}{2} \log_2 \left((2\pi e)^2 |K_X|\right) \tag{39}$$
where$|K_X|$ is the determinant of K_X.

— *Differential entropy is not compulsorily positive, for instance, if the variance $\mathbb{V}(X) \leq \dfrac{1}{2\pi e}$, then $h(X) \leq 0$*

— *Data processing inequality does not always hold. Example, if X Gaussian, with de variance σ^2, then $2X$ has a variance $4\sigma^2$. Hence, $h(2X) \geq h(X)$.*

Differential joint and conditional entropies, under the assumption that integrals are finite, are defined in the exact same manner. The relationships between these different entropies are maintained, as well the definitions in the vector case, and the causal/anti-causal expansions. The following properties are thus satisfied.

Properties 12 *1. LEt (X, Y) be a pair of continuous random variables. Let $h(X, Y)$ be their joint differential entropy, and $h(X|Y)$ et $h(Y|X)$ their conditional differential entropies. We have that*

$$\begin{aligned} h(X,Y) &= h(X) + H(Y|X) = h(Y) + h(Y|X) && (40)\\ h(Y|X) &\leq h(Y) && (41)\\ h(X|Y) &\leq h(X) && (42)\\ h(X,Y) &\leq h(X) + h(Y) && (43) \end{aligned}$$

2. Let $(X_1, ..., X_n)$ be n continuous random variables. The vector differential entropy satisfies

$$\begin{aligned} h(X^n) = h(X_1, ..., X_n) &= \sum_{i=1}^{n} h(X_i|X_1, ..., X_{i-1}) && (44)\\ &= \sum_{i=1}^{n} h(X_i|X_{i+1}, ..., X_n) && (45) \end{aligned}$$

Hereafter, we give an example of differential entropy calculations very common in communication systems.

Example 4 (log-det formula) *Let $X^n = (X_1, ..., X_n)$ be n continuous random variables, with Gaussian distributions, with covariance matrix K_X^n. The differential entropy of X^n is given by :*

$$h(X^n) = \frac{1}{2} \log_2 \left((2\pi e)^n |K_X^n|\right) \tag{46}$$

where $|K_X^n|$ is the determinant of the covariance matrix K_X^n. This formula is widely known under the name log-det formula.

When the variables X_i are independent, the covariance matrix is diagonal, and hence

$$K_X^n = diag(\sigma_1^2, ..., \sigma_n^2) \tag{47}$$

and we recover,

$$h(X^n) = \sum_{i=1}^{n} \frac{1}{2} \log_2 \left((2\pi e)\sigma_i^2\right) = \sum_{i=1}^{n} h(X_i) \tag{48}$$

Conclusions 2 *At the end of this section, you should be able to :*

- *Define and list the properties of discrete entropy*
- *Define joint entropy, and conditional entropy*
- *Link entropy, joint entropy, and conditional entropy*
- *Compute entropy for simple examples*
- *List the difference between discrete and differential entropy*
- *Apply the chain rule to vector entropies*

3 Mutual information

In this section, we define a measure of information which is crucial to measure the quantity of information exchanged between two or more random variables, namely, mutual information.

3.1 Discrete scalar case

Let X and Y two discrete random variable with joint pmf $P_{X,Y}$.

Definition 10 (Mutual information) *Mutual information between X and Y is defined by*

$$I(X;Y) \triangleq \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} P_{X,Y}(x,y)\log_2\left(\frac{P_{X,Y}(x,y)}{P_X(x)P_Y(y)}\right) \tag{49}$$

where P_X and P_Y are the marginal pmf associated with $P_{X,Y}$.

Mutual information can be readily seen to correspond to the Kullback-Leibler (KL) divergence between $P_{X,Y}$ and the product of the marginal laws $P_X P_Y$, i.e.,

$$I(X;Y) = D_{KL}\left(P_{X,Y}||P_X P_Y\right), \tag{50}$$

and as such, it exhibits some characteristics.

Properties 13 *The mutual information between X and Y satisfies the following :*

- *Symmetry : $I(X;Y) = I(Y;X)$ for all joint pmf $P_{X,Y}$*
- *Minimum : $I(X;Y) \geq 0$ with equality iif X and Y are independent, i.e., $P_{X,Y} = P_X P_Y$*
- *Maximum : Mutual information is upper bounded by the individual entropies X et Y, c.a.d, $I(X;Y) \leq \min(H(X), H(y))$, with equality iif $X = f(Y)$ and f is a one-to-one function.*

Exercise 14 : *Prove that mutual information is positive. Hint : proof is similar to the proof of positivity of the KL divergence.*

Mutual information can be related in different ways to entropy measures, as follows :

$$\begin{aligned} I(X;Y) &= H(X) - H(Y|X) && (51)\\ &= H(X) - H(Y|X) && (52)\\ &= H(X) + H(Y) - H(X;Y). && (53)\end{aligned}$$

Exercise 15 : *Prove the different equalities of mutual information.*

Mutual information can be interpreted as the difference between the initial uncertainty over X, and the uncertainty on X which remains after we observe Y. As such, it measures the amount of information inherently shared by X and Y. It is somehow also a measure of independence, since, if two variables are independent, then it is minimal (equal to 0), and if they are fully correlated ($X = f(Y)$), it is maximal.

3.2 Joint and conditional mutual information

Let X, Y and Z be three random variables with joint pmf $P_{X,Y,Z}$. We can define different types of measures of information, and relate them in the following.

Definition 11 (Conditional mutual information) *The mutual information between X and Y conditionally to Z is defined by*

$$\begin{aligned} I(X;Y|Z) &\triangleq \sum_{(x,y,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}} P_{X,Y,Z}(x,y,z)\log_2\left(\frac{P_{X,Y|Z}(x,y|z)}{P_{X|Z}(x|z)P_{Y|Z}(y|z)}\right) && (54)\\ &= \sum_{z\in\mathcal{Z}} P_Z(z)\; I(X;Y|Z=z). && (55)\end{aligned}$$

This mutual information describes the interaction between X and Y knowing that we have observed Z which is correlated to both.

Mutual information can be expressed in terms of conditional entropies as follows :

$$\begin{aligned} I(X;Y|Z) &= H(X|Z) - H(X|(Y,Z)) && (56)\\ &= H(Y|Z) - H(Y|(X,Z)) && (57)\\ &= H(X|Z) + H(Y|Z) - H((X,Y)|Z)) && (58)\end{aligned}$$

Exercise 16 : *Prove the equalities listed here above.*

We can also define another type of information measures, which describes the quantity of information between X and the pair (Y,Z) as follows.

Definition 12 (Joint mutual information) *The mutual information between X and the pair (Y,Z) is defined by :*

$$I(X;(Y,Z)) = \sum_{(x,y,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}} P_{X,Y,Z}(x,y,z)\log_2\left(\frac{P_{X,Y,Z}(x,y,z)}{P_X(x)P_{Y,Z}(y,z)}\right) \quad (59)$$

This definition follows the line of the definition of the mutual information except that it treats the pair of random variables (Y,Z) as one joint variable. A common abuse of notation consists in writing $I(X;YZ)$ or $I(X;Y,Z)$.

Joint conditional and scalar mutual information can be related as follows :

$$\begin{aligned} I(X;(Y,Z)) &= I(X;Z) + I(X;Y|Z) && (60)\\ &= I(X;Y) + I(X;Z|Y). && (61)\end{aligned}$$

Exercise 17 : *Prove the equalities listed here above.*

3.3 Continuous case : continuous mutual information

We have previously seen that differential entropy and discrete entropy differ in a given number of properties, among which positivity, Data Processing Inequality (DPI), and other inequalities. For mutual information, we will see that there much fewer differences between the discrete and continuous case, to the extent that we will scarcely make the distinction later in the course.

Definition 13 (Continuous mutual information) *Let X and Y be two continuous random variable with joint pdf $f_{X,Y}$. Continuous mutual information is defined by*

$$I(X;Y) \triangleq \int_{x,y} f_{X,Y}(x,y) \log_2 \left(\frac{f_{X,Y}(x,y)}{f_X(x) f_Y(y)} \right) \; dx \; dy \tag{62}$$

where f_X and f_Y are the marginal pdf associated with $f_{X,Y}$.

Continuous mutual information measures the quantity of information shared by X and Y.

Properties 14 *Continuous mutual information satisfies the following properties*

— *Symmetry ; $I(X;Y) = I(Y;X)$*

— *Minimum : $I(X;Y) \geq 0$ with equality iif X and Y are independent*

— *Maximum : is not always given by individual entropies since the conditional differential entropy can be negative.*

— *Undefined in degenerate case, i.e., $I(X;g(X))$ is undefined for all deterministic functions g.*

However, all other definitions of mutual informations, joint and conditional, are still valid and their relationships as well. As such, we will not make a distinction of the two types of mutual information (discrete and continuous) and note only $I(X;Y)$.

Example 5 (Shannon formula $\log_2(1 + SNR)$) *Let $X \sim \mathcal{N}(0, P)$ and $W \sim \mathcal{N}(0, \sigma^2)$ be two independent Gaussian random variables. Assume that we observe Y given by*

$$Y = X + W. \tag{63}$$

This model describes the so called Additive White Gaussian Noise (AWGN) channel with inpu signal X, additive noise W, and output signal Y.

The mutual information between X and Y is given by

$$I(X;Y) = \frac{1}{2} \log_2 \left(1 + \frac{P}{\sigma^2} \right). \tag{64}$$

This formula is widely known as Shannon's formula for AWGN channels, or the $\log_2(1 + SNR)$ formula, and describes the maximum spectral efficienty (measured in bits/sec/Hz) which can be transmitted over a channel.

Exercise 18 : *Prove the Shannon formula for the AWGN.*

3.4 Vector mutual information

Similarly to the vector entropy, we will define in this section of the notion of vector mutual information. As stated previously, in communication systems, what is of interest to us is rather the interaction between random processes, and not only scalar realizations of these processes.

Properties 15 (Chain-rule) *Let X be a random variable and let $Y^n = (Y_1, ..., Y_n)$ be a random vector. The mutual information between X and Y^n is given by :*

$$I(X;Y^n) = \sum_{i=1}^{n} I(X;Y_i|Y_1,...,Y_{i-1}) = \sum_{i=1}^{n} I(X;Y_i|Y_{i+1},...,Y_n) \tag{65}$$

where we have used the chain rule of entropy.

The chain-rule is of crucial importance since it allows to compute the quantity $I(X;Y^n)$ with having to perform a large marginalization on the vector (X, Y^n).

Properties 16 (Case of iid processes) *Let $X^n = (X_1, ..., X_n)$ and $Y^n = (Y_1, ..., Y_n)$ be two random vectors such that the pairs (X_i, Y_i) are pairwise independent, i.e.,*

$$P_{X^n,Y^n}(x^n, y^n) = \prod_{i=1}^{n} P_{X_i,Y_i}(x_i, y_i). \tag{66}$$

In this case the mutual information writes as

$$I(X^n;Y^n) = \sum_{i=1}^{n} I(X_i;Y_i). \tag{67}$$

If moreover, the pairs (X_i, Y_i) are i.i.d and all follow the same pmf/pdf $P_{X,Y}$, then

$$I(X^n;Y^n) = nI(X;Y). \tag{68}$$

Whether these observations are iif or not, the fraction $\frac{1}{n}I(X^n;Y^n)$ is called *information rate.*

Hereafter, we write a simple example based on this property.

Example 6 (Memoryless AWGN) *Let $X^n = (X_1, ..., X_n)$ n i.i.d Gaussian variables following $\mathcal{N}(0, P)$, and let $W^n = (W_1, ..., W_n)$ n ri.i.d Gaussian variables following $\mathcal{N}(0, \sigma^2)$, independent of X^n. Assume that we observe the random process $Y^n = X^n + W^n$.*

This model describes the so-called memoryless AWGN channel. The mutual information between the input of the channel X^n and its output Y^n is given by

$$I(X^n;Y^n) = nI(X;Y) = \frac{n}{2}\log_2\left(1 + \frac{P}{\sigma^2}\right) \tag{69}$$

Conclusions 3 *At the end of this section, you should be able to :*

- *Define and list the properties of mutual information*
- *Define joint and conditional mutual information*
- *Link entropy, joint entropy, and conditional entropy, to mutual information*
- *Compute mutual information for simple examples*
- *Apply the chain rule to vector mutual information*