Introduction to Information Theory

Meryem Benammar

5 novembre 2021

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 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.

1.1 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 1 (Discrete entropy) *The entropy of X is defined by :*

$$H(X) \triangleq -\sum_{x\in\mathcal{X}} P_X(x)\log_2\left(P_X(x)\right). \tag{1}$$

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 1 *Here are a few special cases of entropy :*

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

$$H(X) = 0. \tag{2}$$

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

$$H(X) = \log_2(|\mathcal{X}|). \tag{3}$$

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

$$H(X) = H_2(p) \triangleq -p\log_2(p) - (1-p)\log_2(1-p) = H_2(p). \tag{4}$$

$H_2(p)$ 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 1 *: 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 1 *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.,*

$$H(f(X)) \leq H(X) \tag{5}$$

with equality iif f is a one-to-one function.

Exercise 2 *: 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.

1.2 Entropy and lossless compression

Let us consider a random source (random vector) which produces symbols $x^n = (x_1, \ldots, x_n)$ all drawn iid following a given probability distribution P_X. The symbols could belong to any alphabet $\mathcal{X}$: bits $\mathcal{X} = \{0, 1\}$, DNA nucleotid $\mathcal{X} = \{A, T, G, C\}$, text characters $\mathcal{X} = \{a, b, c, d, ...\}$, pixels $\mathcal{X} = [0 : 255]$. In order for these symbols to be stored in a storage device, they need to be mapped each to a binary word.

If no compression is applied, then, in order to store every possible realization of the alhpabet $\mathcal{X}$, one would need $\log_2(|\mathcal{X}|)$ bits for each symbol. However, by doing this, we do not exploit our prior on these symbols : some symbols are more frequent than others, and hence, it would be clever to use binary words of small length for the more frequency symbols. Inversely, for symbols with very low probability of occurrence, one could afford using binary words which are longer. The resulting average binary word length would then hopefully smaller than the $\log_2(|\mathcal{X}|)$ bits/symbol. This is the basic idea behind lossless compression. It is called lossless since it does not destroy information in the source, and only uses more efficiently the word lengths.

Without having to explicitely derive a lossless compression scheme, Shannon proved that, one can use as few as $H(X)$ bits/symbol to compress losslessly a source, and that if one goes below this number, then there will be losses in the information.

1.3 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 2 (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*

$$H(X,Y) \triangleq -\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} P_{X,Y}(x,y)\log_2\left(P_{X,Y}(x,y)\right). \tag{6}$$

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 2 *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 :*
$$H(X,Y) \leq H(X) + H(Y) \tag{7}$$
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.,*
$$P_{X,Y}(x,y) = P_X(x)P_Y(y) = \frac{1}{|\mathcal{X}|}\frac{1}{|\mathcal{Y}|} \text{ pout tout } (x,y)\in\mathcal{X}\times\mathcal{Y} \tag{8}$$

Exercise 3 *: 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 3 (Conditional entropy) *Let X and Y be two random variables with finite support sets $\mathcal{X}$ and mathcalY joint pmf $P_{X,Y}$. The conditional entropy of X knowing Y is defined by :*

$$\begin{aligned} H(X|Y) &\triangleq -\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} P_{X,Y}(x,y)\log_2\left(P_{X|Y}(x|y)\right) &(9)\\ &= \sum_{y\in\mathcal{Y}} P_Y(y)H(X|Y=y) &(10)\end{aligned}$$

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

$$H(X|Y=y) \triangleq -\sum_{x\in\mathcal{X}} P_{X|Y}(x|y)\log_2\left(P_{X|Y}(x|y)\right). \tag{11}$$

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 3 *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*

$$H(X|Y) \leq H(X) \tag{12}$$

   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.,*

$$P_{X|Y}(x|y) = P_X(x)\frac{1}{|\mathcal{X}|} \text{ pout tout } (x,y) \in \mathcal{X} \times \mathcal{Y} \tag{13}$$

Exercise 4 *: 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 5 *: 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* **??** *)*

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

Properties 4 *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*

$$\begin{aligned} H(X,Y) &= H(X) + H(Y|X) && (14)\\ &= H(Y) + H(X|Y) && (15) \end{aligned}$$

Exercise 6 *: Prove the relationships listed herebefore.*

1.4 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 4 (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*

$$H(X^n) = H(X_1, ..., X_n) = \sum_{x^n \in \mathcal{X}^n} P_{X^n}(x^n) \log_2(P_{X^n}(x^n)) \quad (16)$$

$$= \sum_{(x_1,...,x_n) \in \mathcal{X}^n} P_{X_1,...,X_n}(x_1, ..., x_n) \log_2(P_{X_1,...,X_n}(x_1, ..., x_n)) \quad (17)$$

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

Properties 5 *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*

$$H(X^n) \leq \sum_{i=1}^{n} H(X_i) \quad (18)$$

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

$$H(X^n) = H(X_1, ..., X_n) = \sum_{i=1}^{n} H(X_i | X_1, ..., X_{i-1}) \quad (19)$$

$$= \sum_{i=1}^{n} H(X_i | X_{i+1}, ..., X_n). \quad (20)$$

Exercise 7 : *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 1 (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 :*

$$H(X^n) = nH(X). \tag{21}$$

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.

1.5 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 5 (Differential entropy) *The differential entropy of X is given by*

$$h(X) = \int_{x \in \mathcal{X}} f_X(x) \log_2 \left(f_X(x) \right) \, dx \tag{22}$$

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 2 *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*
$$h(X_G) = \frac{1}{2} \log_2(2\pi e \sigma^2) \tag{23}$$
2. *The entropy of a circular complex valued Gaussian variable $X_{CG} \sim \mathcal{CN}(\mu, \sigma^2)$ is given by :*
$$h(X_{CG}) = \log_2(2\pi e \sigma^2) \tag{24}$$

Exercise 8 *: Prove that the entropy of a real-valued Gaussian variable is as stated in property 1.*

Properties 7 *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,*

$$h(X) \leq \frac{1}{2}\log_2\left(2\pi e\mathbb{V}(X)\right) \tag{25}$$

and if X is complex valued with covariance matrix K_X

$$h(X) \leq \frac{1}{2}\log_2\left((2\pi e)^2|K_X|\right) \tag{26}$$

where$|K_X|$ is the determinant of K_X.

— *Differential entropy is not compulsorily positive, for instance, if the variance $\mathbb{V}(X) \leq \frac{1}{2\pi e}$, 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 8 *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*

$$\begin{aligned} h(X,Y) &= h(X) + H(Y|X) = h(Y) + h(Y|X) && (27)\\ h(Y|X) &\leq h(Y) && (28)\\ h(X|Y) &\leq h(X) && (29)\\ h(X,Y) &\leq h(X) + h(Y) && (30) \end{aligned}$$

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

$$\begin{aligned} h(X^n) = h(X_1, ..., X_n) &= \sum_{i=1}^{n} h(X_i|X_1, ..., X_{i-1}) && (31)\\ &= \sum_{i=1}^{n} h(X_i|X_{i+1}, ..., X_n) && (32) \end{aligned}$$

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

Example 3 (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 :*

$$h(X^n) = \frac{1}{2}\log_2\left((2\pi e)^n|K_X^n|\right) \tag{33}$$

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

$$K_X^n = diag(\sigma_1^2, ..., \sigma_n^2) \tag{34}$$

and we recover,

$$h(X^n) = \sum_{i=1}^{n} \frac{1}{2} \log_2 \left((2\pi e)\sigma_i^2\right) = \sum_{i=1}^{n} h(X_i) \tag{35}$$

Conclusions 1 *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*

2 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.

2.1 Discrete scalar case

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

Definition 6 (Mutual information) *Mutual information between X and Y is defined by*

$$I(X;Y) \triangleq \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} P_{X,Y}(x,y)\log_2\left(\frac{P_{X,Y}(x,y)}{P_X(x)P_Y(y)}\right) \tag{36}$$

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_XP_Y, i.e.,

$$I(X;Y) = D_{KL}\left(P_{X,Y}||P_XP_Y\right), \tag{37}$$

and as such, it exhibits some characteristics.

Properties 9 *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., $P_{X,Y} = P_XP_Y$*
- *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 9 *: 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 :

$$\begin{aligned} I(X;Y) &= H(X) - H(Y|X) &(38)\\ &= H(X) - H(Y|X) &(39)\\ &= H(X) + H(Y) - H(X;Y). &(40)\end{aligned}$$

Exercise 10 *: Prove the different identities 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.

2.2 Mutual information and bit-rate

When transmitting an information (image, sound, text, ...) over a physical medium, this image in affected with possible distortions : non-linearities, delays, Doppler shifts, and most importantly, noise.

Noise is generally due to an underlying random physical process, which we can be modeled as a probability distribution P_N. We will often assume that the noise is additive, and independent from the useful signal, say X. Thus, a random signal with distribution P_X transiting through an additive channel with distribution P_N, will yield an output from this channel

$$Y = X + N. \tag{41}$$

As such, the conditional distribution $P_{Y|X}$ of the channel output Y knowing the input X is given by

$$P_{Y|X}(y|x) = P_{Y-X|X}(y-x|x) = P_{N|X}(y-x|x) = P_N(y-x) \tag{42}$$

where we have used the fact that the random processes N and X are independent.

Having this definition of a channel as a random process, and noticing that the noisier a channel, the less information could transit through it, Shannon sought to characterize the maximum bit-rate (measured in bits/sec/Hz) which could be transmitted through a channel, knowing its probability distribution.

Accordingly, Shannon proved that the maximum bit-rate which could be transmitted through a channel is equal to the mutual information between its input X and its output Y, $I(X;Y)$ and corresponds to the number of bits/sec which can be transmitted per unit frequency.

2.3 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 7 (Conditional mutual information) *The mutual information between X and Y conditionally to Z is defined by*

$$\begin{aligned} I(X;Y|Z) &\triangleq \sum_{(x,y,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}} P_{X,Y,Z}(x,y,z)\log_2\left(\frac{P_{X,Y|Z}(x,y|z)}{P_{X|Z}(x|z)P_{Y|Z}(y|z)}\right) && (43)\\ &= \sum_{z\in\mathcal{Z}} P_Z(z)\ I(X;Y|Z=z). && (44)\end{aligned}$$

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 :

$$\begin{aligned} I(X;Y|Z) &= H(X|Z) - H(X|(Y,Z)) && (45)\\ &= H(Y|Z) - H(Y|(X,Z)) && (46)\\ &= H(X|Z) + H(Y|Z) - H((X,Y)|Z)) && (47)\end{aligned}$$

Exercise 11 : *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 8 (Joint mutual information) *The mutual information between X and the pair (Y, Z) is defined by :*

$$I(X;(Y,Z)) = \sum_{(x,y,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}} P_{X,Y,Z}(x,y,z)\log_2\left(\frac{P_{X,Y,Z}(x,y,z)}{P_X(x)P_{Y,Z}(y,z)}\right) \quad (48)$$

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 :

$$\begin{aligned} I(X;(Y,Z)) &= I(X;Z) + I(X;Y|Z) & (49)\\ &= I(X;Y) + I(X;Z|Y). & (50)\end{aligned}$$

Exercise 12 : *Prove the equalities listed here above.*

2.4 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 9 (Continuous mutual information) *Let X and Y be two continuous random variable with joint pdf $f_{X,Y}$. Continuous mutual information is defined by*

$$I(X;Y) \triangleq \int_{x,y} f_{X,Y}(x,y)\log_2\left(\frac{f_{X,Y}(x,y)}{f_X(x)f_Y(y)}\right)\,dx\,dy \quad (51)$$

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 10 *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 4 (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*

$$Y = X + W. \tag{52}$$

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

$$I(X;Y) = \frac{1}{2}\log_2\left(1+\frac{P}{\sigma^2}\right). \tag{53}$$

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 13 *: Prove the Shannon formula for the AWGN.*

2.5 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 11 (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 :*

$$I(X;Y^n) = \sum_{i=1}^{n} I(X;Y_i|Y_1,...,Y_{i-1}) = \sum_{i=1}^{n} I(X;Y_i|Y_{i+1},...,Y_n) \tag{54}$$

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 12 (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.,*

$$P_{X^n,Y^n}(x^n, y^n) = \prod_{i=1}^{n} P_{X_i,Y_i}(x_i, y_i). \tag{55}$$

In this case the mutual information writes as

$$I(X^n; Y^n) = \sum_{i=1}^{n} I(X_i; Y_i). \tag{56}$$

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

$$I(X^n; Y^n) = nI(X; Y). \tag{57}$$

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 5 (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

$$I(X^n; Y^n) = nI(X; Y) = \frac{n}{2} \log_2 \left(1 + \frac{P}{\sigma^2}\right) \tag{58}$$

Conclusions 2 *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*