

Measures of information and uncertainty

Louis Wehenkel

Institut Montefiore, University of Liège, Belgium



ELEN060-2
Information and coding theory
February 2024

- Course organization
- Course objectives
- Introduction to probabilistic reasoning
- Algebra of information measures

- These slides: the slides tend to be self-explanatory; where necessary I have added some notes.

The slides will be available from the WEB:

<https://people.montefiore.uliege.be/lwh/Info/>

- Your personal notes
- Detailed course notes (in french; centrale des cours).
- For further reading, some reference books in english
 - J. Adámek, *Foundations of coding*, Wiley Interscience, 1991.
 - T. M. Cover and J. A. Thomas, *Elements of information theory*, Wiley, 1991.
 - R. Frey, *Graphical models for machine learning and information theory*, MIT Press, 1999.
 - D. Hankerson, G. A. Harris, and P. D. Johnson Jr, *Introduction to information theory and data compression*, CRC Press, 1997.
 - D. J.C. MacKay, *Information theory, inference, and learning algorithms*, Cambridge University Press 2003.
 - D. Welsh, *Codes and cryptography*, Oxford Science Publications, 1998.

Course objectives

1. Introduce information theory
 - Probabilistic (stochastic) systems
 - Reasoning under uncertainty
 - Quantifying information
 - State and discuss coding theorems
2. Give an overview of coding theory and practice
 - Data compression
 - Error-control coding
 - Link with machine learning
3. Illustrate ideas with a large range of practical applications

NB. It is not sure that everything in the slides will be covered during the oral course. You should read the slides and notes (especially those which I will skip) after each course and before the next course.

Notes

The course aims at introducing information theory and the practical aspects of data compression and error-control coding. The theoretical concepts are illustrated using practical examples related to the effective storage and transmission of digital and analog data. Recent developments in the field of channel coding are also discussed (Turbo-codes).

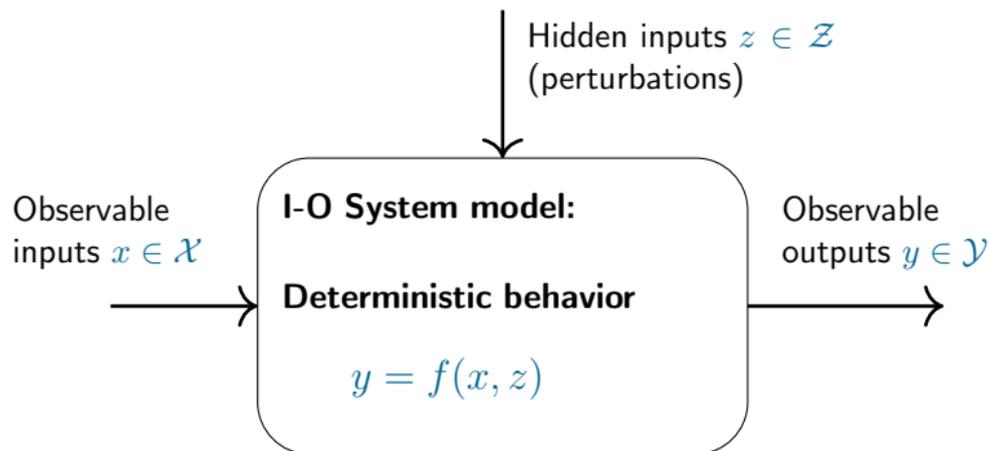
More broadly, the goal of the course is to introduce the basic techniques for reasoning under uncertainty as well as the computational and graphical tools which are broadly used in this area. In particular, Bayesian networks and decision trees will be introduced, as well as elements of automatic learning and data mining.

The theoretical course is complemented by a series of computer laboratories, in which the students can simulate data sources, data transmission channels, and use various software tools for data compression, error-correction, probabilistic reasoning and data mining.

The course is addressed to engineering students (last year), which have some background in computer science, general mathematics and elementary probability theory.

The following two slides aim at clarifying the distinction between deterministic and stochastic systems.

Classical system (deterministic view)



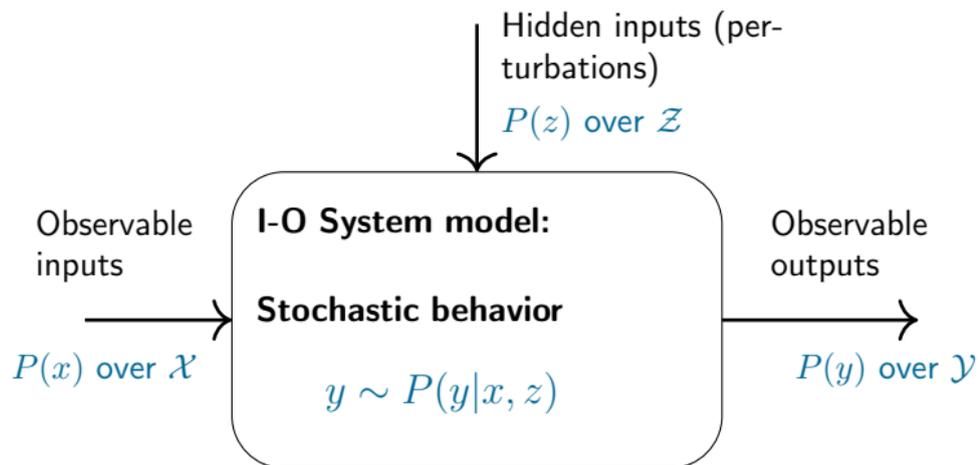
Notes

Classical system theory views a system essentially as a function (or a mapping, in the mathematical sense) between inputs and outputs.

If the system is static, inputs and outputs are scalars (or vectors of scalars). If the system is dynamic, inputs and outputs are temporal signals (continuous or discrete time); a dynamic system is thus viewed as a mapping between input **signals** and output **signals**.

In classical system theory the issue of unobserved inputs and modeling imperfection is handled through stability, sensitivity and robustness theories. In this context uncertainty is essentially modeled by subsets of possible perturbations.

Stochastic system (probabilistic view)



Notes

Here we use probability theory as a tool (a kind of calculus) in order to model and quantify uncertainty. Note that there are other possible choices (e.g. fuzzy set theory, evidence theory...) to model uncertainty, but probability theory is the most mature and most widely accepted approach. Still, there are philosophical arguments and controversies around the interpretation of probability in real-life: e.g. classical (objective) notion of probability vs bayesian (subjective) notion of probability.

Theory of stochastic systems is more complex and more general than deterministic system theory. Nevertheless, the present trend in many fields is to use probability theory and statistics more systematically in order to build and use stochastic system models of reality. This is due to the fact that in many real-life systems, uncertainty plays a major role. Within this context, there are two scientific disciplines which become of growing importance for engineers:

1. **Data mining and machine learning:** how to build models for stochastic systems from observations.
2. **Information theory and probabilistic inference:** how to use stochastic systems in an optimal manner.

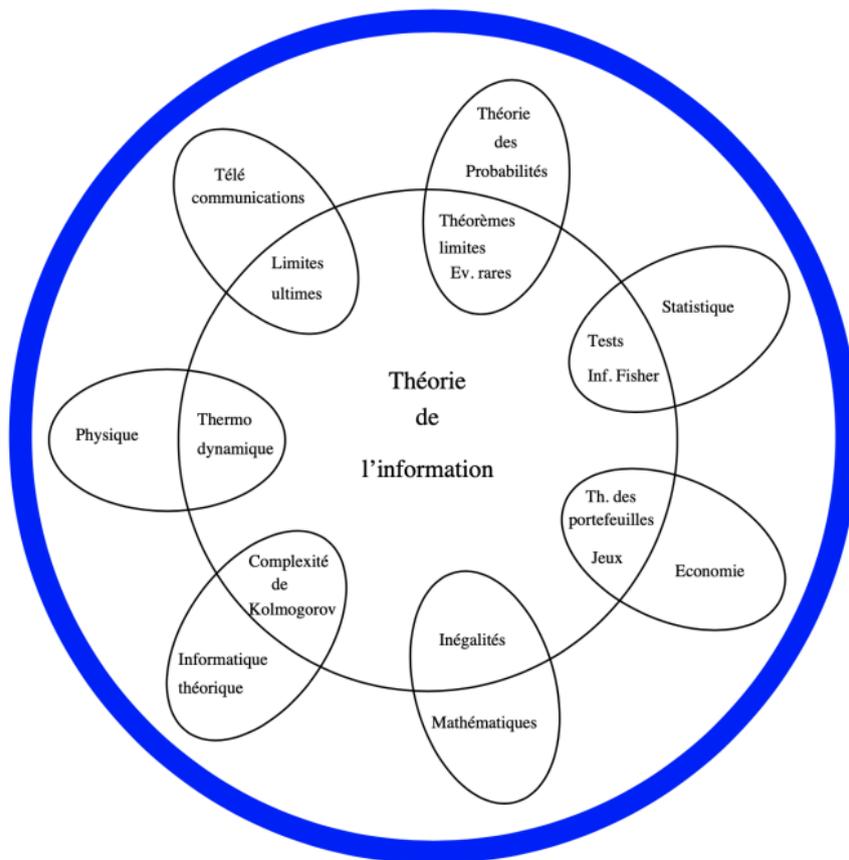
Examples of applications of stochastic system theory:

- Complex systems, where detailed models are intractable (biology, sociology, computer networks...)
- How to take decisions involving an uncertain future (justifications of investments, portfolio management) ?
- How to take decisions under partial/imperfect knowledge (medical diagnosis) ?
- Forecasting the future... (weather, ecosystems, stock exchange...)
- Modeling human behavior (economics, telecommunications, computer networks, road traffic...)
- **Efficient storage and transmission of digital (and analog) data**

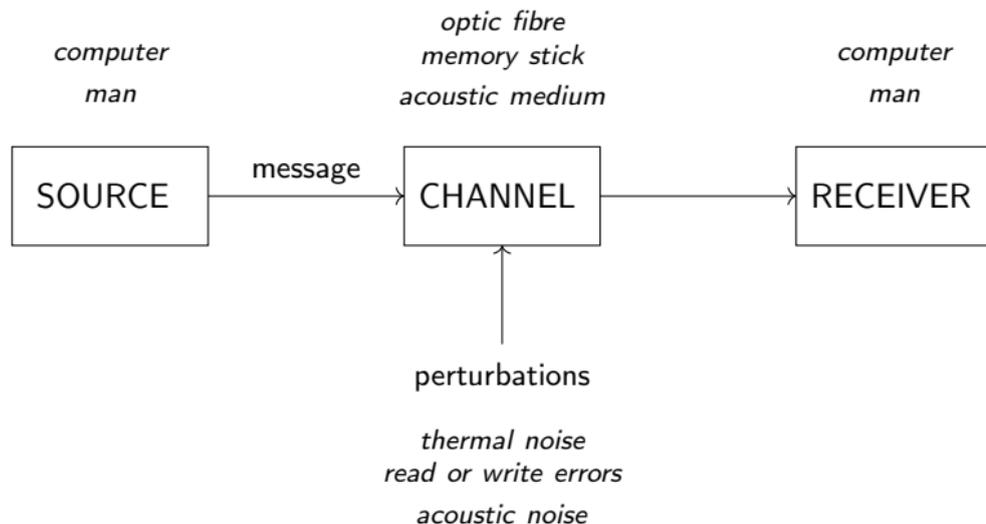
What is it all about ?

- 2 complementary aspects
 - ⇒ Information theory: general theoretical basis
 - ⇒ Coding: compress, fight against noise, encrypt data
- Information theory
 - ⇒ Notions of data source and data transmission channel
 - ⇒ Quantitative measures of information content (or uncertainty)
 - ⇒ Properties: 2 theorems (Shannon theorems) about feasibility limits
 - ⇒ Discrete vs continuous signals
- Applications to coding
 - ⇒ How to reach feasibility limits
 - ⇒ Practical implementation aspects (gzip, Turbo-codes...)

Relations between information theory and other disciplines



Shannon paradigm



Message:

- sequence of symbols, analog signal (sound, image, smell...)
- messages are chosen at **random**
- channel perturbations are **random**

Notes

The foundations of information theory were laid down by Claude E. Shannon, shortly after the end of the second world war in a seminal paper entitled *A mathematical theory of communication (1948)*. In this paper, all the main theoretical ingredients of modern information theory were already present. In particular, as we will see later, Shannon formulated and provided proofs of the two main coding theorems.

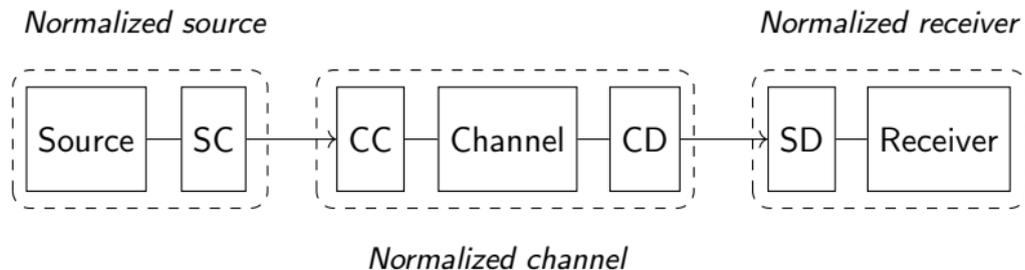
Shannon theory of communication is based on the so-called Shannon paradigm, illustrated on this slide: a data source produces a message which is sent to a receiver through an imperfect communication channel. The possible source messages can generally be modeled by a sequence of symbols, chosen in some way by the source which appears as unpredictable to the receiver. In other words, before the message has been sent, the receiver has some uncertainty about what will be the next message. It is precisely the existence of this uncertainty which makes communication necessary (or useful): after the message has been received, the corresponding uncertainty has been removed. We will see later that information will be measured by this reduction in uncertainty.

Most real life physical channels are imperfect due to the existence of some form of noise. This means that the message sent out will arrive in a corrupted version to the receiver (some received symbols are different from those emitted), and again the corruption is unpredictable for the receiver and for the source of the message.

The two main questions posed by Shannon in his early paper are as follows:

- Suppose the channel is perfect (no corruption), and suppose we have a probabilistic description (model) of the source, what is the maximum rate of communication (source symbols per channel usage), provided that we use an appropriate source code. This problem is presently termed as the *source coding problem* or as the reversible *data compression problem*. We will see that the answer to this question is given by the entropy of the source.
- Suppose now that the channel is noisy, what is then the maximum rate of communication without errors between any source and receiver using this channel? This is the so-called *channel coding problem* or *error-correction coding problem*; we will see that the answer here is the *capacity* of the channel, which is the upper bound of mutual information between input and output messages.

Use of source and channel coding



Source coding (SC/SD): remove redundancy (make message as short as possible)

Channel coding (CC/CD): make data transmission reliable (fight against noise)

Nota bene:

1. Source redundancy may be useful to fight against noise, but is not necessarily adapted to the channel characteristics.
2. Once redundancy has been removed from the source, all sources have the same behavior (completely unpredictable behavior).
3. Channel coding: fight against channel noise without spoiling resources.
4. Coding includes conversion of alphabets.

Notes

The two main results of information theory are thus the characterization of upper bounds in terms of data compression on the one hand, and error-less communication on the other.

A further result of practical importance is that (in most, but not all situations) source and channel coding problems can be decoupled. In other words, data compression algorithms can be designed independently from the type of data communication channel that will be used to transmit (or store) the data. Conversely, channel coding can be carried out irrespectively of the type of data sources that will be used to transmit information over them. This result has led to the partition of coding theory into its two main subparts.

Source coding aims at removing redundancy in the source messages, so as to make them appear shorter and purely random. On the other hand, channel coding aims at introducing redundancy into the message, so as to make it possible to decode the message in spite of the uncertainty introduced by the channel noise.

Because of the decomposition property, these problems are generally solved separately. However, there are examples of situations where the decomposition breaks down (like some multi-user channels) and also situations where from the engineering point of view it is much easier to solve the two problems simultaneously than separately. This latter situation appears when the source redundancy is particularly well adapted to the channel noise (e.g. spoken natural language redundancy is adapted to acoustic noise).

Examples of digital sources are: scanned images, computer files of natural language text, computer programs, binary executable files. . . . From your experience, you already know that compression rates of such different sources may be quite different.

Examples of channels are: AM or FM modulated radio channel, ethernet cable, magnetic storage (tape or hard-disk); computer RAM; CD-ROM. . . . Again, the characteristics of these channels may be quite different and we will see that different coding techniques are also required.

Quantitative notion of information content

Information provided by a message: vague but widely used term

Aspects:

- unpredictable character of a message
- interpretation: truth, value, beauty...

Interpretation depends on the context, on the observer: too complex and probably not appropriate as a measure...

The unpredictable character may be measured \Rightarrow quantitative notion of information

\Rightarrow A message carries more information if it is more unpredictable

\Rightarrow Information quantity: decreasing function of probability of occurrence

Nota bene.

Probability theory (not statistics) provides the main mathematical tool of information theory.

Notations: $(\Omega, \mathcal{E}, P(\cdot)) =$ probability space

Information and coding.

Think of a simple coin flipping experiment (the coin is fair). How much information is gained when you learn (i) the state of a flipped coin; (ii) the states of two flipped coins; (iii) the outcome when a four-sided die is rolled ? How much memory do you need to store these informations on a binary computer ?

Consider now the double coin flipping experiment, where the two coins are thrown together and are indistinguishable once they have been thrown. Both coins are fair. What are the possible outcomes of this experiment ? What are the probabilities of these outcomes ? How much information is gained when you observe any one of these outcomes ? How much is gained in average per experiment (supposing that you repeat it indefinitely) ? Supposing that you have to communicate the result to a friend through a binary channel, how could you code the outcome so that in the average, you will minimize channel use ?

Probability theory reminder: definitions and notations

Probability space: triplet $(\Omega, \mathcal{E}, P_{\Omega}(\cdot))$

Ω : universe of all possible outcomes of a random experiment (sample space)

$\omega, \omega', \omega_i \dots$: elements of Ω (outcomes)

\mathcal{E} : denotes a set of subsets of Ω (called the event space)

$A, B, C \dots$: elements of \mathcal{E} , i.e. events.

The event space models all possible observations.

An event corresponds to a logical statement about ω

Elementary events: singletons $\{\omega_i\}$

$P_{\Omega}(\cdot)$: probability measure (distribution)

$P_{\Omega}(\cdot)$: for each $A \in \mathcal{E}$ provides a number $\in [0, 1]$.

Probability (sample) space: requirements

Event space \mathcal{E} : must satisfy the following properties

- $\Omega \in \mathcal{E}$
- $A \in \mathcal{E} \Rightarrow \neg A \in \mathcal{E}$
- $\forall A_1, A_2, \dots \in \mathcal{E}$ (finite or countable number): $\bigcup_i A_i \in \mathcal{E}$

\Rightarrow one says that \mathcal{E} is a σ -algebra

Measure $P_\Omega(\cdot)$: must satisfy the following properties (Kolmogorov axioms)

- $P_\Omega(A) \in [0, 1], \forall A \in \mathcal{E}$
- $P_\Omega(\Omega) = 1$
- If $A_1, A_2, \dots \in \mathcal{E}$ and $A_i \cap A_j = \emptyset$ for $i \neq j$: $P_\Omega(\bigcup_i A_i) = \sum_i P_\Omega(A_i)$

\Rightarrow one says that $P_\Omega(\cdot)$ is a probability measure

Finite sample spaces

We will generally restrict ourselves to finite sample spaces: i.e. the set Ω has (only) a finite number of elements.

In this context, we will always use the **maximal** σ -algebra: $\mathcal{E} = 2^\Omega$, aka the powerset of Ω which contains all possible subsets of Ω .

Conditional probability $P_{\Omega|B}(\cdot)$ aka $P_\Omega(\cdot|B)$ aka $P(\cdot|B)$

Definition: conditional probability: $P_{\Omega|B}(A) \triangleq \frac{P_\Omega(A \cap B)}{P_\Omega(B)}$ (aka $P(A|B)$)

Careful: $P_{\Omega|B}(A)$ defined only if $P_\Omega(B) > 0$.

Important: If $P_\Omega(B) > 0$, then $P_{\Omega|B}(\cdot)$ is a valid probability measure over (Ω, \mathcal{E}) (it is well defined for any $A \in \mathcal{E}$ and it verifies the Kolomogorov axioms).

Notations: Note that: $P_{\Omega|B}(A|C) = P_{\Omega|C}(A|B) \Rightarrow$ Notation $P(A|B, C) = P(A|B \cap C)$.

Independent events ($A \perp B$)

Definition: $A \perp B$ iff $P(A \cap B) = P(A)P(B)$. (Thus $A \perp B \Leftrightarrow B \perp A$.)

Important: if $P(B) > 0$ then $A \perp B$ iff $P(A|B) = P(A)$. If $P(B) = 0$, $A \perp B$ anyway.

Random variables

From a physical viewpoint, a random variable is an elementary way of perceiving (observing) outcomes.

From a mathematical viewpoint, a random variable is a function defined on Ω (the values of this function may be observed).

Because we restrict ourselves to finite sample spaces, all the random variables are necessarily discrete and finite: they have a finite number of possible values.

Let us denote by $\mathcal{X}(\cdot)$ a function defined on Ω and by $\mathcal{X} = \{X_1, \dots, X_n\}$ its range (i.e. its set of possible values).

We will not distinguish between $\left\{ \begin{array}{l} \text{a value (say } X_i \text{) of } \mathcal{X}(\cdot) \text{ and} \\ \text{the subset } \{\omega \in \Omega \mid \mathcal{X}(\omega) = X_i\} \end{array} \right.$

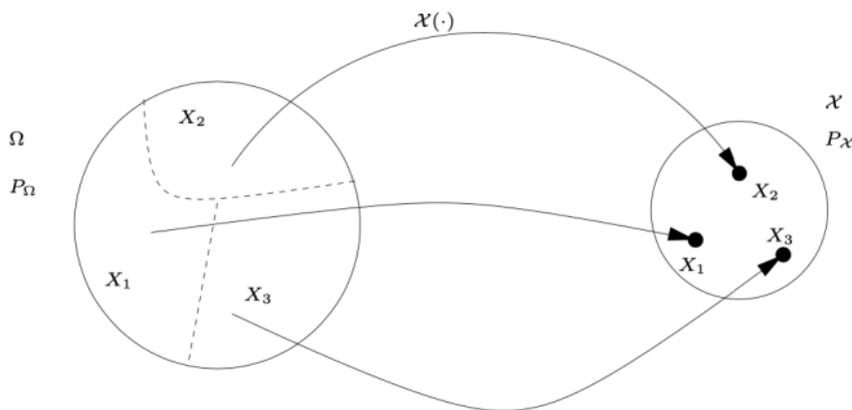
Thus a random variable can also be viewed as a partition $\{X_1, \dots, X_n\}$ of Ω .

Theoretical requirement: $X_i \in \mathcal{E}$ (always true if Ω is finite and \mathcal{E} maximal).

The random variable $\mathcal{X}(\cdot)$ induces a probability measure on $\mathcal{X} = \{X_1, \dots, X_n\}$

$P_{\mathcal{X}}(\mathcal{X} = X_i) \triangleq P_{\Omega}(X_i)$, which we will simply denote by $P(X_i)$.

We will denote by $P(\mathcal{X})$ the measure $P_{\mathcal{X}}(\cdot)$.



A random variable provides *condensed* information about the outcome of a random experiment.

Some more notations ...

\mathcal{X} and \mathcal{Y} two discrete r.v. on $(\Omega, \mathcal{E}, P(\cdot))$.

Notation: $\mathcal{X} = \{X_1, \dots, X_n\}$ et $\mathcal{Y} = \{Y_1, \dots, Y_m\}$. (n and m finite).

X_i (resp. Y_j) value of \mathcal{X} (resp. \mathcal{Y}) \equiv subsets of Ω .

\Rightarrow We identify a r.v. with the partition it induces on Ω .

Contingency table

| | Y_1 | \dots | Y_j | \dots | Y_m | |
|----------|---------|---------|---------------|---------|---------|---------------|
| X_1 | | | \vdots | | | |
| \vdots | | | \vdots | | | |
| X_i | \dots | \dots | $p_{i,j}$ | \dots | \dots | $p_{i,\cdot}$ |
| \vdots | | | \vdots | | | |
| X_n | | | \vdots | | | |
| | | | $p_{\cdot,j}$ | | | |

$$p_{i,\cdot} \equiv P(X_i) \quad \leftarrow *$$

$$\equiv P(\mathcal{X} = X_i)$$

$$p_{\cdot,j} \equiv P(Y_j) \quad \leftarrow *$$

$$\equiv P(\mathcal{Y} = Y_j)$$

$$p_{i,j} \equiv P(X_i \cap Y_j) \equiv P(X_i, Y_j) \quad \leftarrow *$$

$$\equiv P([\mathcal{X} = X_i] \wedge [\mathcal{Y} = Y_j])$$

Complete system of events

Reminder:

- an event denotes a particular subset of Ω
- \mathcal{E} is the σ -algebra of all relevant events for some modeling context

Definition: A_1, \dots, A_n form a *complete system of events* if

- $\forall i \neq j : A_i \cap A_j = \emptyset$ (they are incompatible two by two) and if
- $\bigcup_i^n A_i = \Omega$ (they cover Ω).

Conclusion: a discrete r.v. \equiv complete system of events.

Remark: we suppose that $A_i \neq \emptyset$, but this does not imply that $P(A_i) > 0$!!!

Some authors give a slightly different definition, where the second condition is replaced by: $P(\bigcup_i^n A_i) = 1$.

If then $\bigcup_i^n A_i \neq \Omega$, one may complete such a system by adjoining one more event A_{n+1} (non empty but of zero probability)

Calculus of random variables

On a given Ω we may define an arbitrary number of r.v. In reality, random variables are the only practical way to observe outcomes of a random experiment.

Thus, a random experiment is often *defined* by the properties of a collection of random variables.

Composition of r.v.: $\mathcal{X}(\cdot)$ is a r.v. defined on Ω and $\mathcal{Y}(\cdot)$ is a random variable defined on \mathcal{X} , then $\mathcal{Y}(\mathcal{X}(\cdot))$ is also a r.v. defined on Ω .

Concatenation \mathcal{Z} of $\mathcal{X} = \{X_1, \dots, X_n\}$ and $\mathcal{Y} = \{Y_1, \dots, Y_m\}$ defined on Ω :
 $\mathcal{Z} = \mathcal{X}, \mathcal{Y}$ defined on Ω by $\mathcal{Z}(\omega) = (\mathcal{X}(\omega), \mathcal{Y}(\omega)) \Rightarrow P(\mathcal{Z}) = P(\mathcal{X}, \mathcal{Y})$.

Independence of r.vars. Notation $\mathcal{X} \perp \mathcal{Y}$

- $\mathcal{X} = \{X_1, \dots, X_n\}$ and $\mathcal{Y} = \{Y_1, \dots, Y_m\}$
- $\mathcal{X} \perp \mathcal{Y}$ iff $\forall i \leq n, j \leq m : X_i \perp Y_j$.
- Equivalent to factorisation of probability measure: $P(\mathcal{X}, \mathcal{Y}) = P(\mathcal{X})P(\mathcal{Y})$

In general $P(\mathcal{X}, \mathcal{Y}) = P(\mathcal{X})P(\mathcal{Y}|\mathcal{X}) = P(\mathcal{Y})P(\mathcal{X}|\mathcal{Y})$

Example 1: coin flipping

Experiment: throwing two coins at the same time.

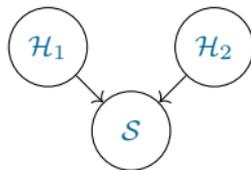
Random variables:

- $\mathcal{H}_1 \in \{T, F\}$ true if first coin falls on heads
- $\mathcal{H}_2 \in \{T, F\}$ true if second coin falls on heads
- $\mathcal{S} \in \{T, F\}$ true if both coins fall on the same face

Thus $\mathcal{S} = (\mathcal{H}_1 \wedge \mathcal{H}_2) \vee (\neg\mathcal{H}_1 \wedge \neg\mathcal{H}_2)$.

Coins are independent: $\mathcal{H}_1 \perp \mathcal{H}_2$, and $P(\mathcal{S}, \mathcal{H}_1, \mathcal{H}_2) = P(\mathcal{S}|\mathcal{H}_1, \mathcal{H}_2)P(\mathcal{H}_1)P(\mathcal{H}_2)$.

Graphically:



A first example of Bayesian network

Suppose the coins are both fair, and compute $P(\mathcal{S})$.

Notes

This is a very simple example (and classical) of a random experiment.

The first structural information we have is that the two coins behave independently (this is a very realistic, but not perfectly true assumption). The second structural information we have is that the third random variable is a (deterministic) function of the other two, in other words its value is a causal consequence of the values of the first two random variables.

Using only the structural information one can depict graphically the relationship between the three variables, as shown on the slide. We will see the precise definition of a Bayesian belief network tomorrow, but this is actually a very simple example of this very rich concept. Note that the absence of any link between the two first variables indicates independence graphically.

To yield a full probabilistic description, we need to specify the following three probability measures: $P(\mathcal{H}_1)$, $P(\mathcal{H}_2)$ and $P(\mathcal{S}|\mathcal{H}_1, \mathcal{H}_2)$, i.e. essentially 2 real numbers (since $P(\mathcal{S}|\mathcal{H}_1, \mathcal{H}_2)$ is already given by the functional description of \mathcal{S}).

If the coins are identical, then we have to specify only one number, e.g. $P(\mathcal{H}_1 = T)$.

If the coins are fair, then we know everything, i.e. $P(\mathcal{H}_1 = T) = 0.5$.

Show that if the coins are fair, we have $\mathcal{H}_1 \perp \mathcal{S} \perp \mathcal{H}_2$. Still, we don't have $\mathcal{H}_1 \perp \mathcal{H}_2|\mathcal{S}$. Explain intuitively.

In general (i.e. for unfair coins) however, we don't have $\mathcal{H}_1 \perp \mathcal{S}$. For example, suppose that both coins are biased towards heads.

You can use the “javabayes” application on the computer to play around with this example. More on this later. . .

Example 2: binary source with independent symbols $\omega_i \in \{0, 1\}$

$P(1)$: probability that the next symbol is 1.

$P(0) = 1 - P(1)$: probability that the next symbol is 0.

Let us denote by $h(\omega)$ the information provided by one symbol ω .

Then: - $h(\omega) = f\left(\frac{1}{P(\omega)}\right)$ where $f(\cdot)$ is increasing and
- $\lim_{x \rightarrow 1} f(x) = 0$ (zero information if event is certain)

On the other hand (symbols are independent):

For two successive symbols ω_1, ω_2 we should have $h(\omega_1, \omega_2) = h(\omega_1) + h(\omega_2)$.

But: $h(\omega_1, \omega_2) = f\left(\frac{1}{P(\omega_1, \omega_2)}\right) = f\left(\frac{1}{P(\omega_1)P(\omega_2)}\right)$
 $\Rightarrow f(xy) = f(x) + f(y) \Rightarrow f(\cdot) \propto \log(\cdot)$

Definition: the *self-information* provided by the observation of an event $A \in \mathcal{E}$ is given by: $h(A) = -\log_2 P(A)$ Shannon

Note: $h(A) \geq 0$. When $P(A) \rightarrow 0 : h(A) \rightarrow +\infty$.

Notes

Comments.

The theory which will be developed is not really dependent on the base used to compute logarithms. Base 2 will be the default, and fits well with binary codes as we will see.

You should convince yourself that the definition of self-information of an event fits with the intuitive requirements of an information measure.

It is possible to show from some explicit hypotheses that there is no other possible choice for the definition of a measure of information (remember the way the notion of thermodynamic entropy is justified).

Nevertheless, some alternative measures of information have been proposed based on relaxing some of the requirements and imposing some others.

To be really convinced that this measure is the right one, it is necessary to wait for the subsequent lectures, so as to see what kind of implications this definition has.

Example: questionnaire, weighting strategies.

Conditional information

Let us consider an event $C = A \cap B$:

$$\begin{aligned}\text{We have: } h(C) &= h(A \cap B) = -\log P(A \cap B) = -\log P(A)P(B|A) \\ &= -\log P(A) - \log P(B|A)\end{aligned}$$

One defines the *conditional* self-information of the event B given that (or supposing that) the event A is true: $h(B|A) = -\log P(B|A)$

Thus, once we know that $\omega \in A$, the information provided by the observation that $\omega \in B$ becomes $-\log P(B|A)$.

Note that: $h(B|A) \geq 0$

One can write: $h(A \cap B) = h(A) + h(B|A) = h(B) + h(A|B)$

In particular: $A \perp B$: $h(A \cap B) = h(A) + h(B)$

Thus: $h(A \cap B) \geq \max\{h(A), h(B)\} \Rightarrow$ **monotonicity of self-information**

Illustration: transmission of information

- $\Omega = \Omega_i \times \Omega_o$: all possible input/output pairs of a channel-source combination.
- A : denotes the observation of a input message; B an output message.
- Linked by transition probability $P(B|A)$ (stochastic channel model).
- $P(A)$: what a receiver can guess about the sent message before it is sent (knowing only the model of the source).
- $P(A|B)$: what a receiver can guess about the sent message after communication has happened and the output message B has been received.
- $P(B|A)$ represents what we can predict about the output, once we know which message will be sent.
- Channel without noise (or deterministic): $P(B|A)$ associates one single possible output to each input.
- For example if inputs and outputs are binary: $P(\omega_i|\omega_o) = \delta_{i,o}$

Mutual information of two events

Definition: $i(A; B) = h(A) - h(A|B)$.

Thus: $i(A; B) = \log \frac{P(A|B)}{P(A)} = \log \frac{P(A \cap B)}{P(A)P(B)} = i(B; A)$

\Rightarrow mutual information is by definition symmetric.

Discussion:

$h(A|B)$ may be larger or smaller than $h(A)$

\Rightarrow mutual information may be positive, zero or negative.

It is equal to zero iff the events are independent.

Particular cases:

If $A \supset B$ then $P(A|B) = 1 \Rightarrow h(A|B) = 0 \Rightarrow i(A; B) = h(A)$.

If $A \supset B$ then $h(B) \geq h(A)$.

But converse is not true...

Exercise

Let us consider the coin throwing experiment.

Suppose that the two coins are identical (not necessarily fair), and say that $p \in [0; 1]$ denotes the probability to get heads for either coin.

Compute¹ the following quantities under the two assumptions $p = \frac{1}{2}$ (fair coins), and $p = 1.0$ (totally biased coins).

$$h(\mathcal{H}_1 = T), h(\mathcal{H}_1 = F), h(\mathcal{H}_2 = T), h(\mathcal{H}_2 = F).$$

$$h([\mathcal{H}_1 = T] \wedge [\mathcal{H}_2 = T])$$

$$h([\mathcal{H}_1 = T] \wedge [\mathcal{H}_1 = F])$$

$$h([\mathcal{S} = T]), h([\mathcal{S} = F]),$$

$$h([\mathcal{H}_1 = T] | [\mathcal{H}_2 = T])$$

$$h([\mathcal{H}_1 = T] | [\mathcal{S} = T])$$

$$h([\mathcal{H}_1 = T] | [\mathcal{S} = T], [\mathcal{H}_2 = T]))$$

¹Use base 2 for the logarithms.

Entropy of a memoryless time-invariant source

Source: at successive times $t \in \{t_0, t_1 \dots\}$ sends symbols $s(t)$ chosen from a *finite* alphabet $S = \{s_1, \dots, s_n\}$.

Assumption: successive symbols are independent and chosen according to the same probability measure (i.e. independent of t) \Rightarrow **Memoryless and time-invariant**

Notation: $p_i = P(s(t) = s_i)$

Definition: **Source entropy**: $H(S) \triangleq E\{h(s)\} = - \sum_{i=1}^n p_i \log p_i$

Entropy: measures average information provided by the symbols sent by the source:
Shannon/symbol

If F denotes the frequency of operation of the source, then $F \cdot H(S)$ measures average information per time unit: Shannon/second.

Note that, because of the law of large numbers the per-symbol information provided by any long message produced by the source converges (almost surely) towards $H(S)$.

Notes

Examples.

Compute the entropy (per symbol) of the following sources:

- A source which always emits the same symbol.
- A source which emits zeroes and ones according to two fair coin flipping processes.

How can you simulate a fair coin flipping process with a coin which is not necessarily fair ?

Generalization: entropy of a random variable

Let \mathcal{X} be a discrete r.v.: \mathcal{X} defines the partition $\{X_1, \dots, X_n\}$ of Ω .

Entropy of \mathcal{X} : $H(\mathcal{X}) \triangleq -\sum_{i=1}^n P(X_i) \log P(X_i)$

What if some probs. are zero: $\lim_{x \rightarrow 0} x \log x = 0$: the terms vanish by continuity.

Note: $H(\mathcal{X})$ does only depend on the values $P(X_i)$

Particular case: $n = 2$ (binary source, an event and its negation)

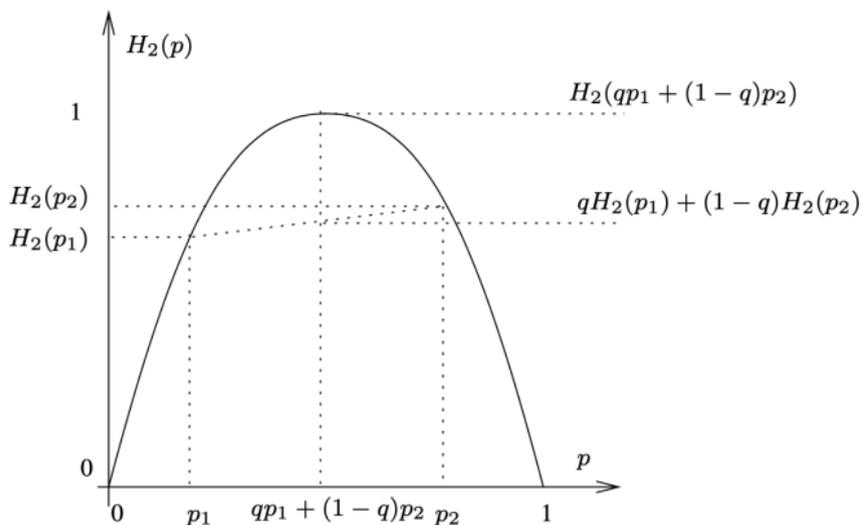
$H(\mathcal{X}) = -p \log p - (1-p) \log(1-p) = H_2(p)$ where p denotes the probability of (any) one of the two values of \mathcal{X} .

Properties of $H_2(p)$:

$$H_2(p) = H_2(1-p)$$

$$H_2(0) = H_2(1) = 0$$

$$H_2(0.5) = 1 \text{ and } H_2(p) \leq 1$$



Another remarkable property: concavity (consequences will appear later).

Means that $\forall p_1 \neq p_2 \in [0, 1], \forall q \in]0, 1[$ we have

$$H_2(qp_1 + (1 - q)p_2) > qH_2(p_1) + (1 - q)H_2(p_2)$$

More definitions

Suppose that $\mathcal{X} = \{X_1, \dots, X_n\}$ and $\mathcal{Y} = \{Y_1, \dots, Y_m\}$ are two (discrete) r.v. defined on a sample space Ω .

Joint entropy of \mathcal{X} and \mathcal{Y} defined by

$$H(\mathcal{X}, \mathcal{Y}) \triangleq - \sum_{i=1}^n \sum_{j=1}^m P(X_i \cap Y_j) \log P(X_i \cap Y_j). \quad (1)$$

Conditional entropy of \mathcal{X} given \mathcal{Y} defined by

$$H(\mathcal{X}|\mathcal{Y}) = - \sum_{i=1}^n \sum_{j=1}^m P(X_i \cap Y_j) \log P(X_i|Y_j). \quad (2)$$

Mutual information defined by

$$I(\mathcal{X}; \mathcal{Y}) = + \sum_{i=1}^n \sum_{j=1}^m P(X_i \cap Y_j) \log \frac{P(X_i \cap Y_j)}{P(X_i)P(Y_j)}. \quad (3)$$

Note that the joint entropy is nothing novel: it is just the entropy of the random variable $\mathcal{Z} = (\mathcal{X}, \mathcal{Y})$. For the time being consider conditional entropy and mutual information as purely mathematical definitions. The fact that these definitions really make sense will become clear from the study of the properties of these measures.

Exercise (computational): Consider our double coin flipping experiment. Suppose the coins are both fair.

Compute $H(\mathcal{H}_1), H(\mathcal{H}_2), H(\mathcal{S})$. Compute $H(\mathcal{H}_1, \mathcal{H}_2), H(\mathcal{H}_2, \mathcal{H}_1), H(\mathcal{H}_1, \mathcal{S}), H(\mathcal{H}_2, \mathcal{S})$ and $H(\mathcal{H}_1, \mathcal{H}_2, \mathcal{S})$

Compute $H(\mathcal{H}_1|\mathcal{H}_2), H(\mathcal{H}_2|\mathcal{H}_1)$ and then $H(\mathcal{S}|\mathcal{H}_1, \mathcal{H}_2)$ and $H(\mathcal{H}_1, \mathcal{H}_2|\mathcal{S})$

Compute $I(\mathcal{H}_1; \mathcal{H}_2), I(\mathcal{H}_2; \mathcal{H}_1)$ and then $I(\mathcal{S}; \mathcal{H}_1)$ and $I(\mathcal{S}; \mathcal{H}_2)$ and $I(\mathcal{S}; \mathcal{H}_1, \mathcal{H}_2)$.

Experiment to work out for tomorrow.

You are given 12 balls, all of which are equal in weight except for one which is either lighter or heavier. You are also given a two-pan balance to use. In each use of the balance you may put any number of the 12 balls on the left pan, and the same number (of the remaining) balls on the right pan. The result may be one of three outcomes: equal weights on both pans; left pan heavier; right pan heavier. Your task is to design a strategy to determine which is the odd ball *and* whether it is lighter or heavier *in as few (expected) uses of the balance as possible*.

While thinking about this problem, you should consider the following questions:

- How can you measure information ? What is the most information you can get from a single weighing ?
- How much information have you gained (in average) when you have identified the odd ball and whether it is lighter or heavier ?
- What is the smallest number of weightings that might conceivably be sufficient to always identify the odd ball and whether it is heavy or light ?
- As you design a strategy you can draw a tree showing for each of the three outcomes of a weighing what weighing to do next. What is the probability of each of the possible outcomes of the first weighing ?

Properties of the function $H_n(\dots)$

Notation: $H_n(p_1, p_2, \dots, p_n) \triangleq - \sum_{i=1}^n p_i \log p_i$

(p_i discrete probability distribution, i.e. $p_i \in [0, 1]$ and $\sum_{i=1}^n p_i = 1$)

Positivity: $H_n(p_1, p_2, \dots, p_n) \geq 0$ (obvious)

Minimal: $H_n(p_1, p_2, \dots, p_n) = 0 \Leftrightarrow p_i = \delta_{i,j}$ for some $j \in \{1, \dots, n\}$ (obvious)

Maximal: $\Leftrightarrow p_i = \frac{1}{n}, \forall i$. (proof follows)

Concavity: (proof follows)

Let (p_1, p_2, \dots, p_n) and (q_1, q_2, \dots, q_n) be two discrete probability distributions and $\lambda \in [0, 1]$ then

$$\begin{aligned} H_n(\lambda p_1 + (1 - \lambda)q_1, \dots, \lambda p_n + (1 - \lambda)q_n) \\ \geq \\ \lambda H_n(p_1, \dots, p_n) + (1 - \lambda)H_n(q_1, \dots, q_n), \end{aligned} \tag{4}$$

Notes

The very important properties of the information and entropy measures which have been introduced before, are all consequences of the mathematical properties of the entropy **function** H_n defined and stated on the slide.

The following slides provide proofs respectively of the maximality and concavity properties which are not trivial.

In addition, let us recall the fact that the function is invariant with respect to any permutation of its arguments.

Questions:

What is the entropy of a uniform distribution ?

Relate entropy function intuitively to uncertainty and thermodynamic entropy.

Nota bene:

Entropy **function**: function defined on a (convex) subset of \mathbb{R}^n .

Entropy **measure**: function defined on the set of random variables defined on a given sample space.

⇒ strictly speaking these are two different notions, and that is the reason to use two different notations (H_n vs H).

Maximum of entropy function (proof)

Gibbs inequality: (Lemma, useful also for the sequel \Rightarrow keep it in mind)

Formulation: let (p_1, p_2, \dots, p_n) and (q_1, q_2, \dots, q_n) be two discrete probability distributions over n values. Then,

$$\sum_{i=1}^n p_i \log \frac{q_i}{p_i} \leq 0, \quad (5)$$

where equality holds if, and only if, $\forall i : p_i = q_i$.

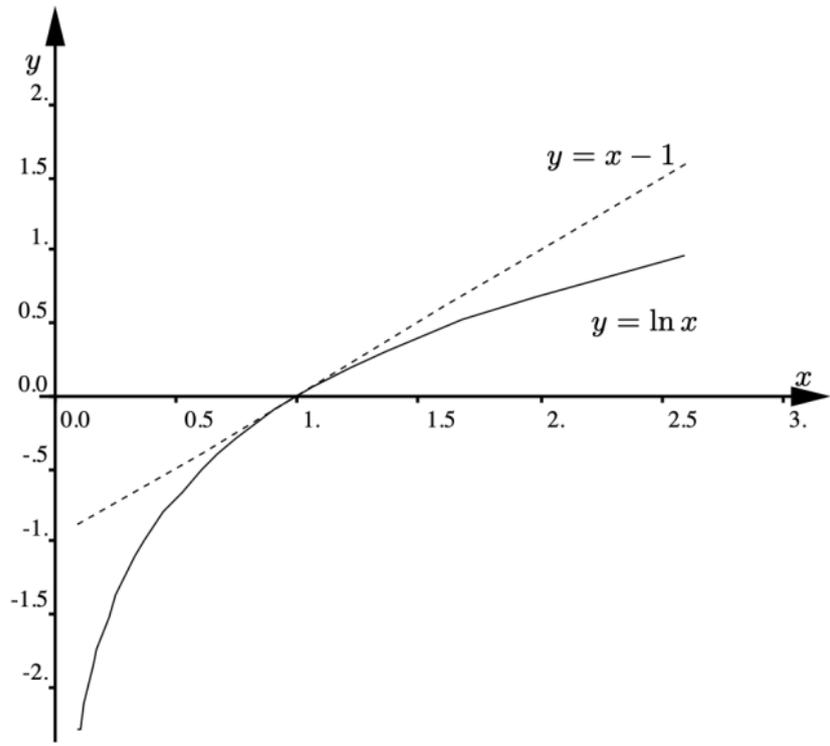
Proof: (we use the fact that: $\ln x \leq x - 1$, with equality $\Leftrightarrow x = 1$, slide below)

Let us prove that $\sum_{i=1}^n p_i \ln \frac{q_i}{p_i} \leq 0$

In $\ln x \leq x - 1$ replace x by $\frac{q_i}{p_i}$, multiply then by p_i sum over index i , which gives (when all p_i are strictly positive)

$$\sum_{i=1}^n p_i \ln \frac{q_i}{p_i} \leq \sum_{i=1}^n p_i \left(\frac{q_i}{p_i} - 1 \right) = \sum_{i=1}^n q_i - \sum_{i=1}^n p_i = 1 - 1 = 0$$

Homework: convince yourself that this remains true even when some of the $p_i = 0$.



Theorem

$$H_n(p_1, p_2, \dots, p_n) \leq \log n, \text{ with equality } \Leftrightarrow \forall i : p_i = \frac{1}{n}.$$

Proof

Let us apply Gibbs inequality with $q_i = \frac{1}{n}$

We find

$$\sum_{i=1}^n p_i \log \frac{1}{np_i} = \sum_{i=1}^n p_i \log \frac{1}{p_i} - \sum_{i=1}^n p_i \log n \leq 0 \Rightarrow$$

$$H_n(p_1, p_2, \dots, p_n) = \sum_{i=1}^n p_i \log \frac{1}{p_i} \leq \sum_{i=1}^n p_i \log n = \log n,$$

where equality holds if, and only if all $p_i = q_i = \frac{1}{n}$ \square

Concavity of entropy function (proof)

Let (p_1, p_2, \dots, p_n) and (q_1, q_2, \dots, q_n) be two probability distributions and $\lambda \in [0, 1]$, then

$$\begin{aligned} H_n(\lambda p_1 + (1 - \lambda)q_1, \dots, \lambda p_n + (1 - \lambda)q_n) \\ \geq \\ \lambda H_n(p_1, \dots, p_n) + (1 - \lambda)H_n(q_1, \dots, q_n), \end{aligned} \tag{6}$$

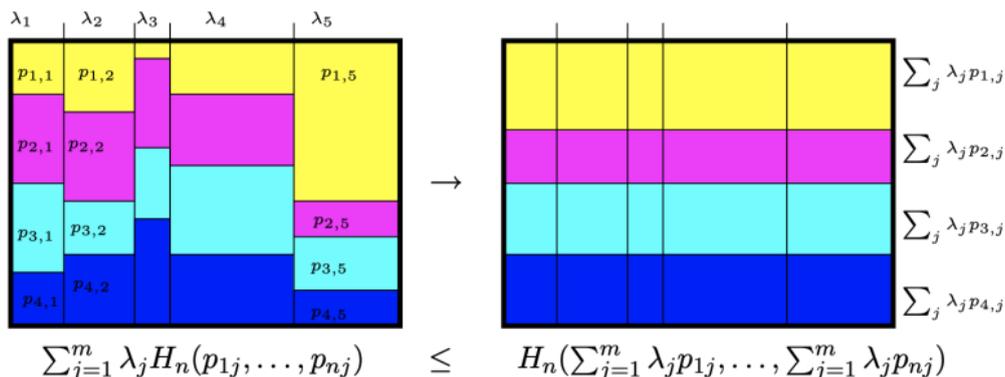
An (apparently) more general (but logically equivalent) formulation:

Mixture of an arbitrary number of probability distributions

$$\begin{aligned} H_n(\sum_{j=1}^m \lambda_j p_{1j}, \dots, \sum_{j=1}^m \lambda_j p_{nj}) \\ \geq \\ \sum_{j=1}^m \lambda_j H_n(p_{1j}, \dots, p_{nj}), \end{aligned} \tag{7}$$

where $\lambda_j \in [0, 1]$, $\sum_{i=j}^m \lambda_j = 1$, et $\forall j : \sum_{i=1}^n p_{ij} = 1$.

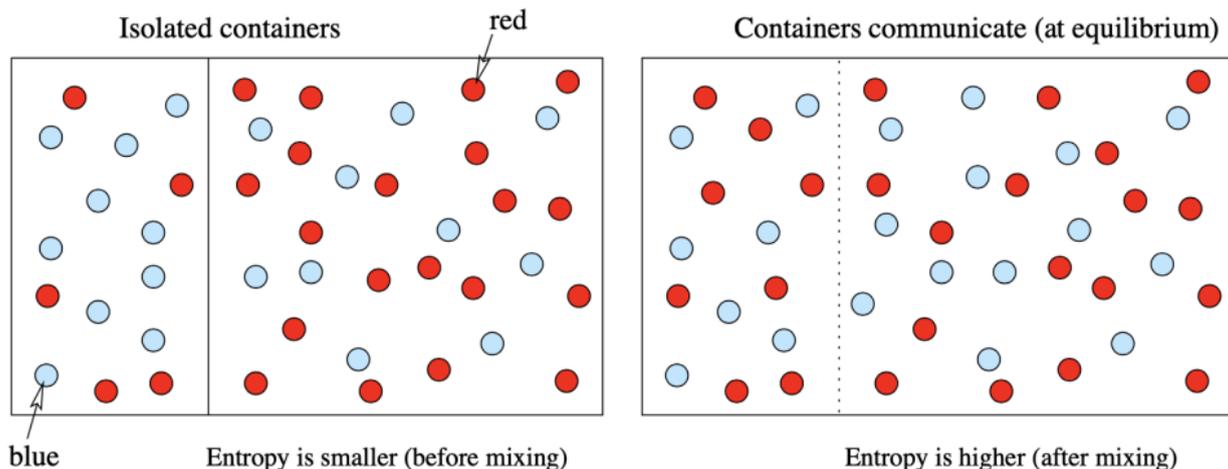
Graphiquely: mixing increases entropy (cf thermodynamics)



Proof: $f(x) = -x \log x$ is concave on $[0, 1]$: $f(\sum_{j=1}^m \lambda_j x_j) \geq \sum_{j=1}^m \lambda_j f(x_j)$.

Thus we have: $H_n(\sum_{j=1}^m \lambda_j p_{1j}, \dots, \sum_{j=1}^m \lambda_j p_{nj}) = \sum_{i=1}^n f(\sum_{j=1}^m \lambda_j p_{ij})$
 $\geq \sum_{i=1}^n [\sum_{j=1}^m \lambda_j f(p_{ij})] = \sum_{j=1}^m \lambda_j \sum_{i=1}^n f(p_{ij})$
 $= \sum_{j=1}^m \lambda_j H_n(p_{1j}, \dots, p_{nj})$

Another interpretation (thermodynamic)



Suppose that you pick a molecule in one of the two cuves and have to guess which kind of molecule you will obtain: in both cases you can take into account in which container you pick the molecule.

⇒ there is less uncertainty in the left cuve than in the right. Compute entropies relevant to this problem.

(Let us open a parenthesis: notions of convexity/concavity)

Convex set : a set which contains all the line segments joining any two of its points.

$C \subset \mathbb{R}^p$ is convex (by def.) if

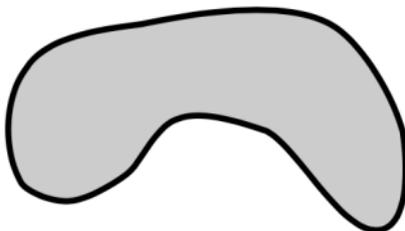
$$x, y \in C, \lambda \in [0, 1] \Rightarrow \lambda x + (1 - \lambda)y \in C.$$

Examples:

Convexe

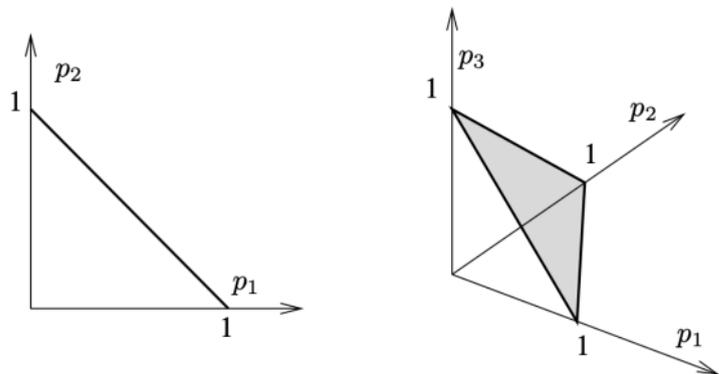


Non convexes



In \mathbb{R} : intervals, semi-intervals, \mathbb{R} .

Examples: sets of probability distributions: $n = 2$ and $n = 3$



More generally:

Linear subspaces, semi-planes are convex

Any intersection of convex sets is also convex (ex. polyhedra)

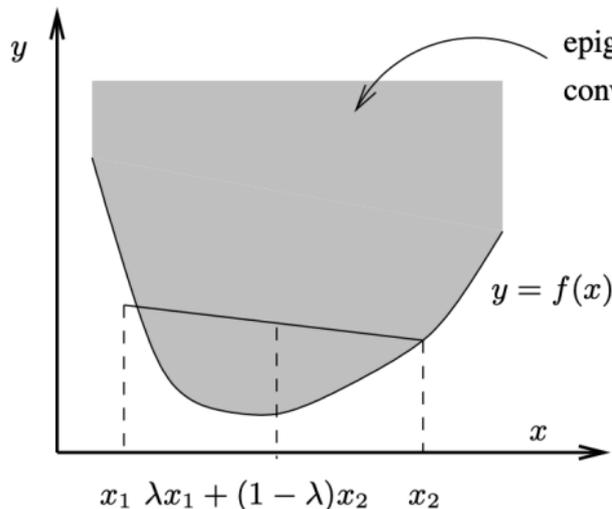
Ellipsoids: $\{\mathbf{x} | (\mathbf{x} - \mathbf{x}_c)^T \mathbf{A}^{-1} (\mathbf{x} - \mathbf{x}_c) \leq 1\}$ (\mathbf{A} positive definite)

Convex functions

$f(\cdot) : \mathbb{R}^p \rightarrow \mathbb{R}$ convex on a convex subset C of \mathbb{R}^p if:

$$\mathbf{x}, \mathbf{y} \in C, \lambda \in [0, 1] \Rightarrow$$

$$f(\lambda \mathbf{x} + (1 - \lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1 - \lambda)f(\mathbf{y}).$$



A sum of convex functions is also convex
If $g(x, y)$ is convex in x then also $\int g(x, y) dy$
 f_i convex, then $\max\{f_i\}$ convex
 $f(x)$ convex $\Rightarrow f(Ax + b)$ convex
Similar properties hold for concave functions...

Strictly convex function

If equality holds only for the trivial cases

$$\lambda \in \{0, 1\} \text{ and/or } x = y$$

(Strictly) concave function

If $-f(\cdot)$ is (strictly) convex.

Important properties

- A convex (or concave) function is continuous inside a convex set
- **Every local minimum of a convex function is also a global minimum**

Criteria:

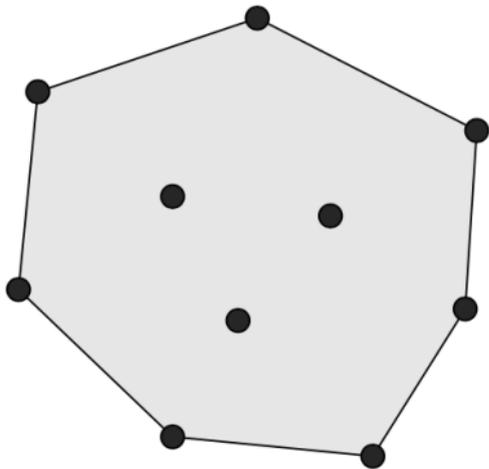
- convex iff convex epigraph
- convex if second derivative (resp. Hessian) positive (resp. positive definite)

In practice: there exist powerful optimization algorithms...

Notion of **convex linear combination**,

$$\left(\sum_j^m \lambda_j \mathbf{x}_j\right), \text{ with } \lambda_j \geq 0 \text{ and } \sum_j^m \lambda_j = 1.$$

Convex hull of some points



Set of points which are convex linear combinations of these points

These are the points which may be written as a kind of weighted average of the starting points (non-negative weights)

Associate weights \Rightarrow center of gravity

A convex hull is a convex set.

In fact, it is the smallest convex set which contains the points \mathbf{x}_j .

(\equiv Intersection of all convex sets containing these points.)

Jensen's inequality:

If $f(\cdot)$ is a convex function defined on $\mathbb{R} \rightarrow \mathbb{R}$ and \mathcal{X} a real random variable

$f(E\{\mathcal{X}\}) \leq E\{f(\mathcal{X})\}$ where, if the function is strictly convex, equality holds iff \mathcal{X} is constant almost surely.

Extension to vectors ...

Concave functions: $\leq \rightarrow \geq \dots$

Particular case: convex linear combinations

The λ_j 's act as a discrete probability measure on $\Omega = \{1, \dots, m\}$.

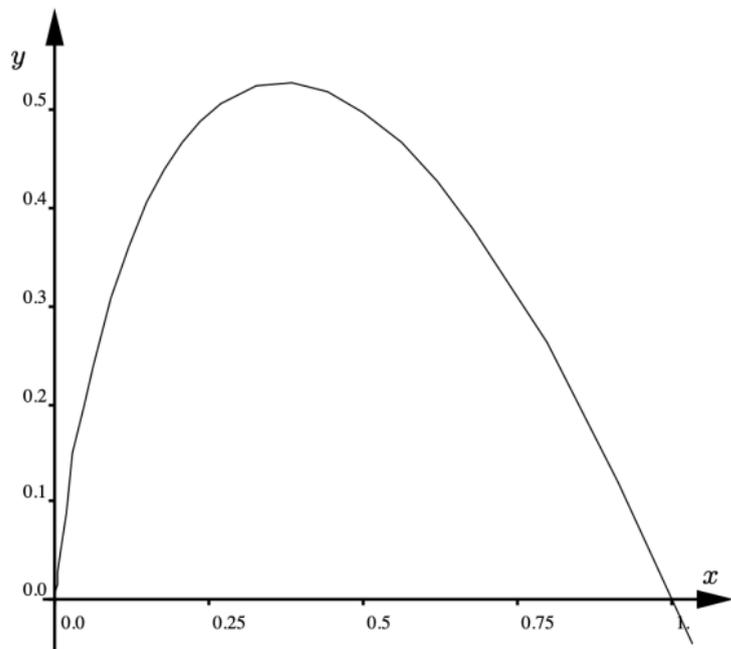
And x_j denotes the value of \mathcal{X} at point $\omega = i$.

Hence: $f(\sum_j^m \lambda_j x_j) \leq \sum_j^m \lambda_j f(x_j)$

Let us close the parenthesis...)

Let us return to the entropies.

Strictly concave functions $f(x) = -x \log x$ and $g(x) = \log x$.



One deduces

$H(\mathcal{X}), H(\mathcal{Y}), H(\mathcal{X}, \mathcal{Y})$ are maximal for uniform distributions.

Also, $H(\mathcal{X}|\mathcal{Y})$ is maximal if, for each j , $P(\mathcal{X}|Y_j)$ is uniform, which is possible only if $P(\mathcal{X})$ is uniform and \mathcal{X} indep. of \mathcal{Y} .

The fact that $f(x) = -x \log x$ is strictly concave is clear from the picture. Clearly $\log x$ is also concave.

All inequalities related to the entropy function may be easily deduced from the concavity of these two functions and Jensen's inequality. For example, let us introduce here a new quantity called relative entropy or Kullback Leibler distance. Let P and Q be two discrete probability distributions defined on a discrete $\Omega = \{\omega_1, \dots, \omega_n\}$. The Kullback Leibler distance (or relative entropy) of P w.r.t. Q is defined by

$$D(P||Q) = \sum_{\omega \in \Omega} P(\omega) \log \frac{P(\omega)}{Q(\omega)} \quad (8)$$

Jensen's inequality allows us to prove that $^2D(P||Q) \geq 0$:

$$-D(P||Q) = - \sum_{\omega \in \Omega} P(\omega) \log \frac{P(\omega)}{Q(\omega)} = \sum_{\omega \in \Omega} P(\omega) \log \frac{Q(\omega)}{P(\omega)} \quad (9)$$

$$\leq \log \left(\sum_{\omega \in \Omega} P(\omega) \frac{Q(\omega)}{P(\omega)} \right) = \log \left(\sum_{\omega \in \Omega} Q(\omega) \right) = \log 1 = 0 \quad (10)$$

where the inequality follows from Jensen's inequality applied to the concave function $\log x$. Because the function is strictly concave, equality holds only if $\frac{Q(\omega)}{P(\omega)}$ is constant over Ω (and hence equal to 1), which justifies the name of *distance* of the relative entropy.

²This is nothing else than Gibbs inequality, which we have already proven without using Jensen's inequality.

- D. MacKay, *Information theory, inference, and learning algorithms*
 - Chapter 2

Frequently asked questions

- Define mathematically the notions of entropy, conditional entropy, joint entropy, and mutual information.