Information Theory — Foundations and Applications

Concepts on Classical Information

Lucas Chibebe Céleri

Institute of Physics
Federal University of Goiás

**2024 — Basque Center for Applied Mathematics
University of Basque Country**

Acknowledgments

CAPES

CNPq
Conselho Nacional de Desenvolvimento
Científico e Tecnológico

FAPEG
Fundação de Amparo à Pesquisa
do Estado de Goiás

eman ta zabal zazu

Universidad
del País Vasco

Euskal Herriko
Unibertsitatea

bcam
basque center for applied mathematics

QPEQUI

UFG
UNIVERSIDADE
FEDERAL DE GOIÁS

Outline of the course

Part I — Foundations

- Classical Information Theory
  1. Introductory concepts
  2. First Shannon coding theorem — Data compression
  3. Second Shannon coding theorem — Channel capacity
- Quantum Information Theory
  1. Introductory concepts
  2. Quantum cryptography
  3. Quantum algorithms

Outline of the course

Part II — Applications

- Classical thermodynamics
- Complexity in critical systems
- Quantum thermodynamics as a gauge theory
- Simulating fermionic systems

Part I

Foundations

Introduction

Some bibliography

Here are some interesting books on the subject of this class

1. T. M. Cover and J. A. Thomas, *Elements of information theory* (J. Wiley & Sons, Inc., New Jersey, 2006).

2. M. M. Wilde, *Quantum information theory* (Cambridge University Press, New York, 2013).

3. M. M. Wilde, *From classical to quantum Shannon theory,* https://arxiv.org/abs/1106.1445 (2019).

Shannon theory

Two big questions[1]

- What is the ultimate data compression rate?
- What is the ultimate transmission rate of communication?

The answers changed everything

Fundamental contributions in statistical physics, computer science, statistical inference, economics, and to probability and statistics. Quantum gravity maybe?

But, before answering such questions, we need to introduce some concepts that will be of fundamental importance, starting with the notion of *information*.

[1] C. E. Shannon. Bell System Technical Journal **27**, 379 (1948)

The information unit

One of the central contributions of Shannon is the notion of a **bit** as a measure of information.

Physical bit

"0" or "1": light switch is off or on, a transistor allows current to flow or not, large number of magnetic spins point in one direction or another and so on.

Shannon's bit

Is a measure of the **surprise** upon learning the outcome of a random binary experiment. The outcome of a coin flip resides in a physical bit, but it is the information associated with the random nature of the physical bit that we would like to measure. It is this notion of a bit that is important in information theory.

The measure of information

How can information be quantified?

The first thing we should observe is that every physical system can be described by means of a random variable

$$X = \{p_X(x),\, x \in \mathcal{X}\}$$

To start, let us consider that $\mathcal{X}$ is a finite set, called the alphabet. The cardinality of this set is denoted by $|\mathcal{X}|$.

Observe that every experiment can be put in the form of a *yes* or *no* questions. In this sense, we use the bit as a natural information unity.

The measure of information

Shannon’s notion of information is based on three postulates:

- The information I contained in one event must depend only on the probability of that event to occur.
- I must be a continuous function.
- I must be additive for independent events.

There is only one mathematical function that respect this three postulates: the logarithm!

The measure of information

In order to prove this statement, let us consider s *independent* occurrences of the event x. The information measure should satisfy

$$\begin{aligned} I\left(p_X(x)^s\right) &= I\left(p_X(x)^{s-1}, p_X(x)\right) \overset{\text{Add}}{=} I\left(p_X(x)^{s-1}\right) + I\left(p_X(x)\right) \\ &= I\left(p_X(x)^{s-2}, p_X(x)\right) + I\left(p_X(x)\right) \overset{\text{Add}}{=} I\left(p_X(x)^{s-2}\right) + 2I\left(p_X(x)\right) \\ &= sI\left(p_X(x)\right) \end{aligned}$$

As a consequence

$$I\left(p_X(x)^{1/t}\right) = \frac{1}{t} \times t \times I\left(p_X(x)^{1/t}\right) = \frac{1}{t} I\left(p_X(x)\right)$$

Therefore, for any rational number $r = s/t$ we must have $I\left(p_X(x)^r\right) = rI\left(p_X(x)\right)$.

The measure of information

Any probability $p_X(x)$ can be written as $p_X(x) = 2^{\log p_X(x)}$. Also, any real number can be arbitrarily well approximated by a rational number and, since I must be continuous, we have

$$I\left(p_X(x)\right) = I\left(2^{\log p_X(x)}\right) = \log\left(p_X(x)\right) I\left(2\right)$$

We can choose $I(2)$ as we please. A convenient choice is $I(2) = -1$!

> The amount of information contained in the occurrence of the event x, whose probability is $p_X(x)$, is
>
> $$I\left(p_X(x)\right) = -\log\left(p_X(x)\right)$$

Shannon entropy

$I\left(p_X(x)\right)$ captures the information associated with a **single** occurrence of the random variable X. However, we are often interested in the information content of the physical system, which is our **information source**. The entropy is then defined as the expected information content of this random variable.

Shannon entropy

$$H\left(X\right) = -\sum_{x \in X}^{|\mathcal{X}|} p_X(x) \log\left(p_X(x)\right)$$

The meaning of Shannon entropy

$H(X)$ can be viewed in two distinct ways:

- It is a measure of the uncertainty we have about X.
- It is a measure of the information gain when we learn X.

Such views are completely equivalent!

Some properties of Shannon entropy

$H(X) \geq 0$. It is the expectation value of a non-negative quantity.

$H(X)$ is invariant under permutations of realizations. This follows because it depends only on the probabilities and not on the actual values of the realizations.

$H(X) = 0$ for a deterministic variable, $p_X(x) = \delta_{x,x_0}$.

$H(X) \leq \log|\mathcal{X}|$. Equality is achieved for the uniform distribution, $p_X(x) = 1/|\mathcal{X}| \; \forall\, x$. The inequality an be proved by Lagrange optimization.

The conditional entropy

The conditional entropy of a random variable given another is the expected value of the entropies of the conditional distributions, averaged over the conditioning random variable

$$H(X|Y) = \sum_{y=1}^{|\mathcal{Y}|} p_Y(y) H(X|Y=y) = -\sum_{x=1}^{|\mathcal{X}|}\sum_{y=1}^{|\mathcal{Y}|} p_{X,Y}(x,y) \log p_{X|Y}(x|y)$$

This leads to the chain rule $H(X,Y) = H(Y) + H(X|Y)$ since

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

Note that, in general $H(X|Y) \neq H(Y|X)$. Also

$$0 \leq H(X|Y) \leq H(X)$$

The joint entropy

Let us consider the joint random variable (X, Y). How much uncertain we are regarding the actual occurrence of both x and y? The answer is the joint entropy

$$H(X, Y) = -\sum_{x=1}^{|\mathcal{X}|}\sum_{y=1}^{|\mathcal{Y}|} p_{X,Y}(x, y) \log p_{X,Y}(x, y)$$

It should be clear that, for independent variables

$$H(X, Y) = H(X) + H(Y)$$

since, in this case, $p_{X,Y} = p_X p_Y$. This subadditivity: $H(X, Y) \leq H(X) + H(Y)$ since $H(X) \geq H(X|Y)$. Also

$$H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$$

The mutual information

How much correlation is shared between two random variables?

$$I(X:Y) = \sum_{x=1}^{|\mathcal{X}|} \sum_{y=1}^{|\mathcal{Y}|} p_{X,Y}(x,y) \log \frac{p_{X,Y}(x,y)}{p_X(x) p_Y(y)}$$

It is a measure of how much the actual distribution differs from the product of its marginals. It is clear that $I(X:Y) \geq 0$ for any distribution.
It is not difficult to show that

$$I(X:Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \geq 0$$

The mutual information is the reduction in the uncertainty of X (Y) due to the knowledge of Y (X).

The mutual information

From the above discussion we obtain

$$I(X:Y) = H(X) + H(Y) - H(X,Y)$$

This is fully compatible with our intuition of correlations.

Moreover, we can easily check that

$$I(X:Y) = I(Y:X)$$

and

$$I(X:X) = H(X)$$

The relative entropy

One of the monst important quantities in information theory is the relative entropy, or Kullback-Leibler divergence, defined as

$$D(p_X||q_X) = \sum_x p_X(x) \log \frac{p_X(x)}{q_X(x)} = -H(X) - \sum_x p_X(x) \log q_X(x)$$

This is a measure of how much you are mistaken in taking q_X and the distribution of the random variable X when the true distribution is p_X.

Before stating the properties of D, we note that

$$I(X:Y) = D(p_{X,Y}||p_X \times p_Y)$$

Properties of relative entropy

Using Jensen's inequality[2] we have

$$\begin{aligned}
-D(p_X||q_X) &= \sum_{x\in\text{supp}(p_X)} p_X(x)\log\frac{q_X(x)}{p_X(x)} \overset{\text{JI}}{\leq} \log \sum_{x\in\text{supp}(p_X)} p_X(x)\frac{q_X(x)}{p_X(x)} \\
&= \log \sum_{x\in\text{supp}(p_X)} q_X(x) \leq \log \sum_{x\in\mathcal{X}} q_X(x) = 0
\end{aligned}$$

Thus, we conclude that

$$D(p_X||q_X) \geq 0$$

This is called *information inequality*. It follows that the equality holds if and only if $p_X(x) = q_X(x)$ for all x. This also implies that $I(X:Y) \geq 0$.

[2] For any convex function f it holds that $\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$

Concavity of Shannon entropy

A very important property of $H(X)$ is that it is concave in $p_X(x)$, thus implying that mixing leads to more uncertainty

$$H(\lambda p_X^1 + (1-\lambda)p_X^2) \geq \lambda H(p_X^1) + (1-\lambda)H(p_X^2)$$

Sketch of the proof. First, using the log sum inequality

$$\left(\sum_i a_i\right) \log\left(\frac{\sum a_i}{\sum b_i}\right) \leq \sum a_i \log\left(\frac{a_i}{b_i}\right)$$

it is possible to prove that D is convex. Them, by applying this to $D(p_X||u_X)$, with u_X being the uniform distribution, the concavity of the entropy can be proved.

Data processing inequality

No clever manipulation of the data can improve the inferences that can be made from the data!

Markov chain

The random variables X, Y and Z are said to form a Markov chain $X \to Y \to Z$ if

$$p_{X,Y,Z}(x,y,z) = p_X(x)p_{Y|X}(y|x)p_{Z|Y}(z|y)$$

This can be used to demonstrating that no processing of Y, deterministic or random, can increase the information that Y contains about X. That is **data processing inequality**

$$I(X:Y) \geq I(X:Z)$$

Data processing inequality — Proof

Let us start by the chain rule, which states that

$$I(X:Y,Z)=I(X:Z)+I(X:Y|Z)=I(X:Y)+I(X:Z|Y)$$

Now, since X and Y are conditionally independent

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

we conclude that $I(X:Z|Y)=0$. This implies that

$$I(X:Y)=I(X:Z)+I(X:Y|Z)$$

Data processing inequality — Proof

We also have that $I(X:Y|Z) \geq 0$, thus leading to the final result

$$I(X:Y) \geq I(X:Z)$$

In particular, if $Z = f(Y)$, then $X \to Y \to f(Y)$ and

$$I(X:Y) \geq I(X:f(Y))$$

Processing of information cannot improve inference power!

Data processing inequality

There is another way to state this result, in terms of the relative entropy

Monotonicity of relative entropy
Let Λ be a classical channel. Them, it follows that

$$D(p||q) \geq D(\Lambda(p)||\Lambda(q))$$

If supp$(p) \nsubseteq$ supp(q), them $D(p||q) = \infty$ and the inequality is trivial. supp$(p) \subseteq$ supp(q), the proof rely on the convexity property of the exponential function and on the positivity of the relative entropy.

Fano's inequality

Let us consider the chain $X \to Y \to \hat{X}$.

- X is the input of a classical channel
- Y is the output
- $\hat{X}$ is the best estimation for X given Y.

Let us define the probability of error as

$$p_e = \text{Prob}\left\{\hat{X} \neq X\right\}$$

If the channel is noiseless, $H(X|Y) = 0$. If noise increases, $H(X|Y)$ also increases. Them

$$H(X|Y) \leq H(X|\hat{X}) \leq h_2(p_e) + p_e \log\left(|\mathcal{X}| - 1\right)$$

with $h_2(p) = -p \log p - (1-p)\log(1-p)$.

Thank you for your attention

lucas@qpequi.com

www.qpequi.com