Entropic Information Theory

Matteo Falcioni
https://github.com/Grufoony/Physics_Unibo

Contents

Chapter 1

Preliminaries

1.1 Introduction

Information theory answers two fundamental questions in communication theory: What is the ultimate data compression (answer: the entropy H), and what is the ultimate transmission rate of communication (answer: the channel capacity C). For this reason some consider information theory to be a subset of communication theory. It can be argued that it is much more. Indeed, it has fundamental contributions to make in statistical physics (thermodynamics), computer science (Kolmogorov complexity or algorithmic complexity), statistical inference (Occam's Razor: "The simplest explanation is best"), and to probability and statistics (error exponents for optimal hypothesis testing and estimation).
Information theory intersects physics (statistical mechanics), mathematics (probability theory), electrical engineering (communication theory), and computer science (algorithmic complexity).

What about ergodicity? The world *ergodic* was introduced by L. Boltzmann in the context of classical (statistical) mechanics to describe the action of the dynamics T_t, $t \in \mathbb{R}$ over an energy surface Σ_E. Boltzmann had hoped that each orbit $\{T_t,\, t \in \mathbb{R}\}$ would equal the whole energy surface Σ_E : he called this statement the *ergodic hypothesis*. The word ergodic comes from the union of the greek words ergon (work) and odos (path). Boltzmann had in fact assumed this hypothesis in order to deduce the equality of time mean and space mean over the phase space, which is a fundamental algorithm in statistical mechanics.
The ergodic hypothesis, as stated above, is ***false****.* The property that phase flows need to satisfy in order for time and space means of real-valued function is now called *ergodicity*.
Usually the term ergodic theory is used to refer to the study of the actions of groups on measure spaces. The actions on topological spaces and smooth manifolds are often called instead topological dynamics and differentiable dynamics.
In the following, we shall study the actions of the group $\mathbb{Z}$ or $\mathbb{N}$ of integers on a space Ω, i.e. we study a transformation $T : \Omega \longrightarrow \Omega$ (the dynamics of our system) and its iterates T^n, $n \in \mathbb{Z}$.

The main sources for these notes will be Walters *An Introduction to Ergodic Theory* [1], Cover and Thomas *Elements of Information Theory, second edition* [2] and course notes at `https://virtuale.unibo.it/course/view.php?id=35914`.

1.2 Measure spaces

Let Ω be a set. A $\sigma-$algebra of subsets of Ω is a collection $\mathscr{B}$ of subsets of Ω satisfying the following three conditions:

1) $\Omega \in \mathscr{B}$

2) if $B \in \mathscr{B}$ then $\Omega \setminus B \in \mathscr{B}$

3) if $B_n \in \mathscr{B}$ for $n \geq 1$ then $\cup_{n=1}^{\infty} B_n \in \mathscr{B}$.

We then call the pair $(\Omega, \mathscr{B})$ a *measurable space*.
A *finite measure* on $(\Omega, \mathscr{B})$ is a function $\mu : \mathscr{B} \longrightarrow \mathbb{R}^+$ satisfying

$a:\ \mu(\emptyset) = 0$

$b:$ if $B_n \in \mathscr{B}$ are disjoint subsets of Ω then $\mu\big(\cup_{n=1}^{\infty} B_n\big) = \sum_{n=1}^{\infty} \mu(B_n)$

A finite-measure space is a triple $(\Omega, \mathscr{B}, \mu)$ where $(\Omega, \mathscr{B})$ is a measurable space and μ is a finite measure on $(\Omega, \mathscr{B})$. We say that $(\Omega, \mathscr{B}, \mu)$ is a *probability space*, or normalised measure pace, if $\mu(\Omega) = 1$. We then say that μ is a probability measure on $(\Omega, \mathscr{B})$. In the following we will usually consider probability spaces.
Another definition that could be useful is that of an algebra: a collection $\mathscr{A}$ of subsets of Ω is called an algebra if it satisfies the following three conditions: *i)* $\emptyset \in \mathscr{A}$; *ii)* if $A, B \in \mathscr{A}$, then $A \cap B \in \mathscr{A}$; *iii)* if $A \in \mathscr{A}$, then $\Omega \setminus A \in \mathscr{A}$.
Also another mathematical object which is similar to an algebra exists, namely the semi-algebra $\mathscr{C}$. The only difference between $\mathscr{C}$ and $\mathscr{A}$ is that condition *iii)* of the definition of algebras should be replaced with the looser condition $*$) if $A \in \mathscr{C}$, then $\Omega \setminus A = \cup_{i=1}^{n} E_i$ where each $E_1 \in \mathscr{C}$ and $E_n, \dots, E_n$ are pairwise disjoint subsets of Ω. From a semi-algebra we can generate an algebra, as stated by the following theorem:

Theorem 1.2.1. *Let $\mathscr{C}$ be a semi-algebra of subsets of Ω. The algebra $\mathscr{A}(\mathscr{C})$ generated by $\mathscr{C}$ consists precisely of those subsets of Ω that can be written in the form $E = \cup_{i=1}^{n} A_i$ where each $A_i \in \mathscr{C}$ and $A_1, \dots, A_n$ are disjoint subsets of Ω.*

We can now define an object which we'll use widely in the following discussions, which is the so called *canonical cilinder*. First, let's define the *measurable rectangles*. For $i \in \mathbb{Z}$ let $(X_i, \mathscr{B}_i, \mu_i)$ be a probability space. Let $X = \prod_{i=-\infty}^{\infty} X_i$. So a point of X is a bisequence $\{x_i\}_{-\infty}^{\infty}$ with $x_i \in X_i$ for each i. We now define a $\sigma-$algebra $\mathscr{B}$ of subsets of X called the product of the $\sigma-$algebras $\mathscr{B}_i$. Let $n \geq 0$, let $A_j \in \mathscr{B}_i$ for $|j| \leq n$, and consider the set

$$\prod_{i=-\infty}^{-(n+1)} X_i \times \prod_{i=-n}^{n} A_j \times \prod_{i=n+1}^{\infty} X_i = \left\{(x_i)_{-\infty}^{\infty} \in X \middle| x_j \in A_j \text{ for } |j| \leq n\right\}. \tag{1.2.1}$$

Such a set is called a *measurable rectangle* and the collection of all such subsets of X forms a semi-algebra $\mathscr{C}$. It simply represents an infinite bi-sequence for which the symbols from the index $-n$ to n are fixed to be part of the set A_j. The $\sigma-$algebra $\mathscr{B}$ is the one generated by $\mathscr{C}$. The measure μ_i can be extended to X by giving the above rectangle the value $\prod_{j=-n}^{n} \mu_j(A_j)$. The probability space $(X, \mathscr{B}, \mu)$ is called the direct product of the spaces $(X_i, \mathscr{B}_i, \mu_i)$.
A special type of product space will be important for us: the case where each space $(X_i, \mathscr{B}_i, \mu_i)$ is the same space $(\mathcal{A}, \mathscr{C}, \mu)$ and $\mathcal{A}$ is a finite set of symbols indexed by an integer k, $\{0, 1, \dots, k-1\}$, $\mathscr{C} = 2^{\mathcal{A}}$, and μ is given by a probability vector $(p_0, p_1, \dots, p_{k-1})$ where $p_i = \mu(\{i\})$. We call these $\mu(\{i\})$ the *marginals* of μ and indicate them equivalently as $\mu(\{k\}) \equiv \mu_k$. We can then take elementary rectangles where each A_j in the description above is taken to be one point of $\mathcal{A}$. So if $n \geq 0$ and $a_j \in \mathcal{A}$, $|j| \leq n$, such an elementary rectangle has the form $\{(x_i)_{-\infty}^{\infty} | x_j = a_j \text{ for } |j| \leq n\}$. We shall denote this set by $[x_{-n}^{n}]$ and call it a *canonical cylinder* or a *block* with end points $-n$ and n.

Let's give a clearer interpretation to all this mathematical description: our space $\mathcal{A}$ is our *alphabet*, the set which contains all of the symbols we can use: for example this could be, for DNA, the set $\mathcal{A} = \{A, C, T, G\}$, or for language the set $\mathcal{A} = \{\text{letters of the italian language}\}$. Then we take its product infinite times, as stated in eq. (1.2.1), to be able to consider infinite sequences (for example infinite strings of DNA). We sometimes denote this infinite product as $\Omega \equiv \mathcal{A}^{\mathbb{Z}}$, to remind ourselves that it is just an infinite product of the same alphabet. Then, we go on to consider infinite sequences for which some characters are fixed: these are our canonical cylinders or blocks. They are nothing more than (infinite) sequences where a certain string of symbols is given.
To make the discussion simpler, most of the times we'll consider just sequences over bisenqueces, and thus consider $\Omega = \mathcal{A}^{\mathbb{N}}$. Also we'll mostly use blocks where just the first n symbols are fixed, therefore considering $[x_1^n]$. Of course this is just to make things easier, as everything that will be said in this setting can be then extended to the more general case.
There are some very useful results we could now state, one above all the Kolmogorov Consistency Theorem. But first, let's make a digression about some other preliminaries which will be useful in understanding better all of this Theory of Information.

1.3 Topology

Consider now $\Omega = \mathcal{A}^{\mathbb{N}}$. [1] It has an elementary topology, induced by a metric. We can define the distance between two characters $z, z' \in \mathcal{A}$ as

$$d_{\mathcal{A}}(z, z') = \begin{cases} 1 & \text{if } z = z' \\ 0 & \text{if } z \neq z' \end{cases} \tag{1.3.1}$$

Then, given $x, y \in \Omega$ we define the metric

$$\tilde{d}(x, y) = \sum_{n=1}^{\infty} 2^{-n} d_{\mathcal{A}}(x_n, y_n) \tag{1.3.2}$$

This function satisfies all properties of a metric. Note that the more two sequences coincide, the smaller $\tilde{d}$ gets.
Sometimes it's useful to define another metric: let $\lambda = \frac{1}{|\mathcal{A}|}$. Define then, given $x, y \in \Omega$,

$$d_1(x, y) = \lambda^{n(x,y)} \tag{1.3.3}$$

where $n(x, y) = \min_k\{k | x_k \neq y_k\}$ is the number of the first n different digits of the two sequences. Notice that again, the more prefixes[2] coincide, the smaller d_1 gets.
The elementary topology in Ω is generated by defining open sets (or *balls*) in Ω by means of our distance d_1:

$$B(x, r) = \{y \in \Omega | d_1(x, y) \leq r\}$$

These are exactly the y's in our space such that the first k digits coincide up to $\log(r)/\log(\lambda)$, since they satisfy $\lambda^n \leq r$. Thus

$$B(x, r) = \{y \in \Omega | x_k = y_k \text{ for } 1 \leq k \leq \frac{\log(r)}{\log(\lambda)}\} \tag{1.3.4}$$

We can now identify cylinders and open balls: notice in fact that, from its definition, a cylinder

$$[x_1^n] = \left\{ y \middle| \begin{array}{c} y_1 = x_1 \\ \vdots \\ y_n = x_n \end{array} \right\}$$

coincides with the open sets defined in eq. (1.3.4)

$$[x_1^{\frac{\log(r)}{\log(\lambda)}}] = B(x, r) \tag{1.3.5}$$

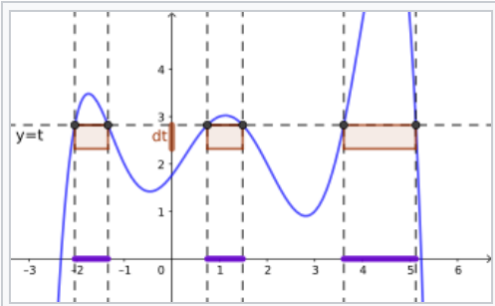
Therefore the canonical cylindera are generators of our topology in Ω.
Also, it can be proven that the two distances $\tilde{d}$ and d_1 induce the same topology, i.e. that open balls defined by $\tilde{d}$ are contained in open balls defined by d_1, and viceversa.

1.4 Lebesgue Integration

[3] The integral of a positive function f between limits a and b can be interpreted as the area under the graph of f. This is straightforward for functions such as polynomials, but what does it mean for more exotic functions? In general, for which class of functions does "area under the curve" make sense? The answer to this question has great theoretical and practical importance.
As part of a general movement toward rigor in mathematics in the nineteenth century, mathematicians attempted to put integral calculus on a firm foundation. The Riemann integral—proposed by Bernhard Riemann (1826–1866)—is a broadly successful attempt to provide such a foundation. Riemann's definition starts with the construction of a sequence of easily calculated areas that converge to the integral of a given function. This definition is successful in the sense that it gives

[1] Note that the cardinality of $\mathcal{A}$ grows exponentially, since $|\mathcal{A}^n| = |\mathcal{A}|^n$.
[2] the first m digits for a certain $m \in \mathbb{N}$

the expected answer for many already-solved problems, and gives useful results for many other problems.
However, Riemann integration does not interact well with taking limits of sequences of functions, making such limiting processes difficult to analyze. This is important, for instance, in the study of Fourier series, Fourier transforms, and other topics. The Lebesgue integral is better able to describe how and when it is possible to take limits under the integral sign (via the monotone convergence theorem and dominated convergence theorem).
While the Riemann integral considers the area under a curve as made out of vertical rectangles, the Lebesgue definition considers horizontal slabs that are not necessarily just rectangles, and so it is more flexible. For this reason, the Lebesgue definition makes it possible to calculate integrals for a broader class of functions. For example, the Dirichlet function, which is 0 where its argument is irrational and 1 otherwise, has a Lebesgue integral, but does not have a Riemann integral. Furthermore, the Lebesgue integral of this function is zero, which agrees with the intuition that when picking a real number uniformly at random from the unit interval, the probability of picking a rational number should be zero. Folland (1984) summarizes the difference between the Riemann and Lebesgue approaches thus: "to compute the Riemann integral of f, one partitions the domain $[a, b]$ into subintervals", while in the Lebesgue integral, "one is in effect partitioning the range of f."
For the Riemann integral, the domain is partitioned into intervals, and bars are constructed to meet the height of the graph. The areas of these bars are added together, and this approximates the integral, in effect by summing areas of the form $f(x)dx$ where $f(x)$ is the height of a rectangle and dx is its width.
For the Lebesgue integral, the range is partitioned into intervals, and so the region under the graph is partitioned into horizontal "slabs" (which may not be connected sets). The area of a small horizontal "slab" under the graph of f, of height dy, is equal to the measure of the slab's width times dy: $\mu(\{x|f(x) > y\})$. The Lebesgue integral may then be defined by adding up the areas of these horizontal slabs. See as an example Fig. 1.1.



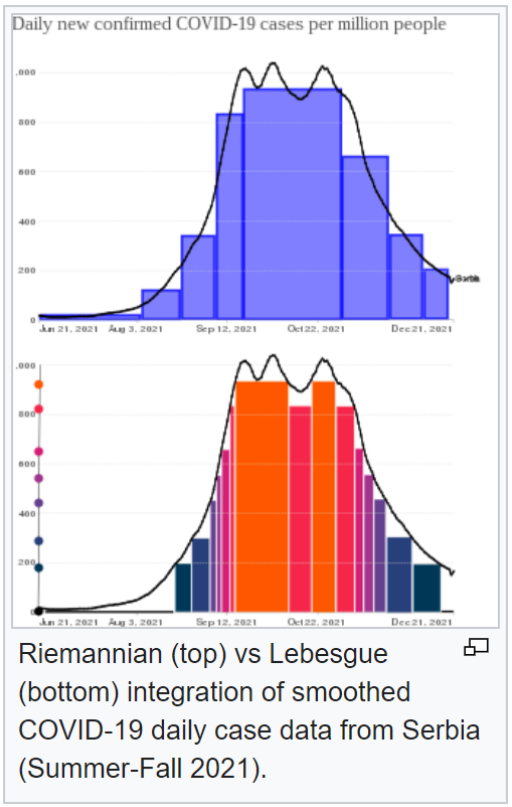
A measurable function is shown, together with the set $\{x : f(x) > t\}$ (on the *x*-axis). The Lebesgue integral is obtained by slicing along the *y*-axis, using the 1-dimensional Lebesgue measure to measure the "width" of the slices.

Figure 1.1

An equivalent way to introduce the Lebesgue integral is to use so-called simple functions, which generalize the step functions of Riemann integration. Consider, for example, determining the cumulative COVID-19 case count from a graph of smoothed new daily cases (Fig. 1.2).

* **The Riemann–Darboux approach**
  Partition the domain (time period) into intervals (eight, in the example at right) and construct bars with heights that meet the graph. The cumulative count is found by summing, over all bars, the product of interval width (time in days) and the bar height (cases per day).

* **The Lebesgue approach**
  Choose a finite number of target values (eight, in the example) in the range of the function. By constructing bars with heights equal to these values, but below the function, they imply a partitioning of the domain into the same number of subsets (subsets, indicated by color in the example, need not be connected). This is a "simple function," as described below. The cumulative count is found by summing, over all subsets of the domain, the product of the measure on that subset (total time in days) and the bar height (cases per day).



Riemannian (top) vs Lebesgue (bottom) integration of smoothed COVID-19 daily case data from Serbia (Summer-Fall 2021).

Figure 1.2

Let's now go back to the mathematics and give a rigorous formulation of the Lebesgue integral. Let $\mathscr{B}(\mathbb{R})$ denote the σ-algebra of Borel subsets of $\mathbb{R}$. This is the σ-algebra generated by all open subsets of $\mathbb{R}$ and is also generated by the collection of all intervals, or by the collection of all intervals of the form (c, ∞).
Let $(X, \mathscr{B}, m)$ a measure space. A function $f : X \to \mathbb{R}$ is *measurable* if $f^{-1}(D) \in \mathscr{B}$ whenever $D \in \mathscr{B}(\mathbb{R})$ or equivalently if $f^{-1}(c, \infty) \in \mathscr{B}$ for all $c \in \mathbb{R}$. A function $f : X \to \mathbb{C}$ is measurable if both its imaginary and real part are measurable. We say $f = g$ *almost everywhere* (a.e.) if $m\big(\{x : f(x) \neq g(x)\}\big) = 0$, i.e. the functions have equal values everywhere except for a set which has null measure.[3] In measure theory, all sets of null measure are "invisible".
Let $(X, \mathscr{B}, m)$ be a probability space. A function $f : X \to \mathbb{R}$ is a *simple function* if it can be written in the form $\sum_{i=1}^{n} a_i \chi_{A_i}$, where $a_i \in \mathbb{R}$, $A_i \in \mathscr{B}$, the sets A_i are disjoint sets of X, and χ_{A_i} denotes the characteristic function of A_i ($\chi(x) = 1$ if $x \in A_i$, 0 otherwise). Simple functions are measurable. We define the integral for simple functions by:

$$\int f \, dm = \sum_{i=1}^{n} a_i m(A_i) \, . \tag{1.4.1}$$

This value is independent of the representation $\sum_{i=1} a_i \chi_{A_i}$.
Suppose $f : X \to \mathbb{R}$ is measurable and $f \geq 0$. Then there exists an increasing sequence of simple

[3] If the only open set of zero measure in a topological space X is the void set, then being equal a.e. for f and g implies being equal.

functions $f_n \to f$. For example, we could take

$$f_n(x) = \begin{cases} \frac{i-1}{2^n}, & \text{if } \frac{i-1}{2^n} \le f(x) \le \frac{i}{2^n}\ i = 1, \dots, n2^n \\ n, & \text{if } f(x) \ge n\,. \end{cases}$$

We define $\int f\,dm = \lim_{n\to\infty} \int f_n\,dm$ and note that this definition is independent of the chosen sequence $\{f_n\}$. We say f is *integrable* if $\int f\,dm < \infty$.
Suppose $f : X \to \mathbb{R}$ is measurable. Then $f = f^+ - f^-$ where $f^+(x) = \max\{f(x), 0\} \ge 0$ and $f^- = \max\{-f(x), 0\} \ge 0$. We say that f is integrable if $\int f^+ dm$, $\int f^- dm < \infty$ and we then define

$$\int f\,dm = \int f^+ dm - \int f^- dm$$

Notice that f is (Lebesgue-)integrable if and only if $|f|$ is integrable. If $f = g$ a.e. then one is integrable if the other is and $\int f\,dm = \int g\,dm$.
The main theorems on integrating sequences as functions are the following:

Theorem 1.4.1 (Dominated Convergence Theorem). *Suppose $f_1 \le f_2 \le f_3 \le \dots$ is an increasing sequence of integrable real-valued functions on $(X, \mathscr{B}, m)$. If $\left\{\int f_n\,dm\right\}$ is a bounded sequence of real numbers, then $\lim_{n\to\infty} f_n$ exists a.e. and is integrable and $\int(\lim_{n\to\infty} f_n)dm = \lim_{n\to\infty}\int f_n$. If instead $\left\{\int f_n\,dm\right\}$ is an unbounded sequence then either $\lim_{n\to\infty} f_n$ is infinite on a set of not zero measure or $\lim_{n\to\infty} f_n$ is not integrable.*

Lemma 1.4.2 (Fatou's Lemma). *Let $\{f_n\}$ be a sequence of measurable real-valued functions on $(X, \mathscr{B}, m)$ which is bounded below by an integrable function. If $\liminf_{n\to\infty}\int f_n\,dm < \infty$ then $\liminf_{n\to\infty} f_n$ is integrable and $\int \liminf_{n\to\infty} f_n\,dm \le \liminf_{n\to\infty}\int f_n\,dm$*

Corollary 1.4.2.1 (Dominated Convergence Theorem). *If $g : X \to \mathbb{R}$ is integrable and $\{f_n\}$ is a sequence of measurable real-valued functions with $|f_n| \le g$ a.e. and $\lim_{n\to\infty} f_n = f$ a.e. then f is integrable and $\lim_{n\to\infty}\int f_n\,dm = \int f\,dm$.*

We denote by $L^1(X)$ the space of all Lebesgue integrable functions. Such space is a Banach space with norm $\|f\|_1 = \int |f|\,dm$. If $f \in L^1(X)$, then $\int_A f\,dm$ denotes $\int f \cdot \chi_A\,dm$.

1.5 Perron-Frobenius Theory

In matrix theory, the Perron–Frobenius theorem, proved by Oskar Perron (1907) and Georg Frobenius (1912), asserts that a real square matrix with positive entries has a unique largest real eigenvalue and that the corresponding eigenvector can be chosen to have strictly positive components, and also asserts a similar statement for certain classes of non-negative matrices.
This theorem has important applications to probability theory (ergodicity of Markov chains); to the theory of dynamical systems (subshifts of finite type); to economics (Okishio's theorem, Hawkins–Simon condition); to demography (Leslie population age distribution model); to social networks (DeGroot learning process); to Internet search engines (PageRank); and even to ranking of football teams.
Let $A = [a_{ij}]$ be a $k \times k$ matrix. We say A is non-negative if $a_{ij} \ge 0$ for all i, j. Such a matrix is called *irreducible* if for any pair i, j there is some $n > 0$ such that $a_{ij}^{(n)} > 0$ where $a_{ij}^{(n)}$ is the (i, j)-th element of A^n. The matrix A is *irreducible* and *aperodic* if there exists $n > 0$ such that $a_{ij}^{(n)} > 0$ for *all* i, j. We shall use the following result:

Theorem 1.5.1 (Perron-Frobenius Theorem). *Let $A = [a_{ij}]$ be a non-negative $k \times k$ matrix. Then:*

(i) There is a non negative eigenvalue λ such that no eigenvalue of A has absolute value greater than λ.

(ii) We have $\min_i(\sum_{j=1}^k a_{ij}) \le \lambda \le \max_i(\sum_{j=1}^k a_{ij})$

(iii) *Corresponding to the eigenvalue λ there is a non-negative left (row) eigenvector $u = (u_1, \ldots, u_k)$ and a non-negative right (column) eigenvector*

$$v = \begin{pmatrix} v_1 \\ \vdots \\ v_k \end{pmatrix}$$

(iv) *If A is irreducible then λ is a simple eigenvalue and the corresponding eigenvectors are strictly positive (i.e. $u_i > 0$, $v_i > 0$ $\forall i$).*

(v) *If A is irreducible then λ is the only eigenvalue of A with a non-negative eigenvector.*

This result, as already said, will become very useful when we'll talk about Markov Chain and Hidden Markov Models.

Chapter 2

Basics of ergodic information theory

2.1 Stochastic processes and properties of the measure

We can now go back to the discussion in §1.2 and state some useful and insightful results.
First of all, recall that we are dealing now with the space of all infinite sequences of our finite alphabet, $\Omega = \mathcal{A}^{\mathbb{N}}$. We introduced earlier a measure μ on this space and defined the marginals of our measure as the functions $\mu\big([x_1^n]\big) = \mu(x_1, \dots, x_n, \dots) \equiv \mu_n$ acting on the so called canonical cylinders. Most of the time, we will be working with the so called *stochastic processes*.

Definition 2.1.1 (Stochastic Process). *A stochastic process is an infinite sequence $X \equiv \{X_n\} = \{X_1, X_2, \dots, X_n, \dots\}$ of random variables X_n with values in $\mathcal{A}$, defined by the $k-th$ order joint distribution on $\mathcal{A}^k$:*

$$\mu_k(a_1^k) \equiv \mathbb{P}\big(X_1^k = a_1^k\big), \quad a_1^k \in \mathcal{A}^k$$

Equivalently, the distribution of a stochastic process can be completely defined by means of the conditional probabilities, by starting from the one-character distribution μ_1 and then computing the successive conditional distributions:

$$\mu(x_n|x_1^{n-1}) = \frac{\mu_n(x_1^n)}{\mu_{n-1}(x_1^{n-1})} \tag{2.1.1}$$

since we could iterate the formula to get

$$\mu(\omega_1, \dots, \omega_n) = \mu(\omega_1)\mu(\omega_2|\omega_1)\mu(\omega_3|\omega_2\omega_1)\dots\mu(\omega_k|\omega_1\dots\omega_{k-1})\dots\mu(\omega_n|\omega_1\dots\omega_{n-1})$$

The process μ is said to be ***stationary*** if

$$\mu(x_1^k) = \mu(x_{n+1}^{n+k})\ \forall a \in \mathcal{A} \tag{2.1.2}$$

We can consider now some dynamics on our probability space $(\Omega, \mathscr{B}, \mu)$, i.e. we consider a transformation $T : \Omega \longrightarrow \Omega'$, where $(\Omega', \mathscr{B}', \mu')$ is another probability space. Most of the times we will actually consider simply consider $\Omega' = \Omega, \mu' = \mu, \mathscr{B}' = \mathscr{B}$.
A transformation $T : \Omega \longrightarrow \Omega'$ is said to be *measurable* if $T^{-1}(\mathscr{B}') \subset \mathscr{B}$ (i.e. $B' \in \mathscr{B}' \Rightarrow T^{-1}B' \in \mathscr{B}$).
A transformation $T : \Omega \to \Omega$ is *measure preserving* if $\forall A \in \mathscr{B}$ we have $\mu\big(T^{-1}(A)\big) = \mu(A)$. We equivalently say then that T is measure preserving or that μ is T*-invariant*, so that $\mu \circ T^{-1} = \mu$. $\mu \circ T^{-1}$ is sometimes referred to as the "pull-back measure" $T^*\mu$. Since Ω is compact, we can consider the set of all measures on Ω,

$$\mathcal{P} = \big\{\mu \big|\, \mu \text{is a finite measure w/ } \mu(\Omega) = 1\big\} \tag{2.1.3}$$

and also we can define

$$\mathcal{P}_I = \big\{\mu \in \mathcal{P} \big|\, \mu \circ T^{-1} = \mu\big\} \tag{2.1.4}$$

as the collection of all invariant measures (depending on the dynamics).
There is a natural dynamics on Ω: consider $x = (x_0, x_1, \dots, x_n, \dots), x_j \in \mathcal{A}$. The ***shift*** is defined as:

$$\begin{aligned} \sigma : \Omega &\longrightarrow \Omega \\ \sigma : (x_0, x_1, \dots, x_n, \dots) &\longrightarrow (x_1, x_2, \dots, x_n, \dots) \end{aligned} \tag{2.1.5}$$

Its name should be self explicative: it shifts the sequences (to the left).
A ***process*** or a ***source*** is a shift-invariant probability measure μ on the topological space $\mathcal{A}^{\mathbb{Z}}$. The process is uniquely defined by the joint distribution, and not really by the exact values of the x_n's. The shift is closely related to the property of stationarity of the measure: actually, both properties are equivalent.

Proposition 2.1.1 (σ-invariancy $\Leftrightarrow$ stationarity). *Shift invariance implies stationarity, and viceversa.*

Proof. By definition, $\sigma^{-1}([x_1^n]) = \{z \in \Omega \big| \sigma(z) \in [x_1^n]\} = \bigsqcup_{a\in\mathcal{A}}[a, x_1, \dots, x_n]$. Then the proof is straightforward, since

$$\mu(\sigma^{-1}([x_1^n])) = \mu([x_1^n]) = \sum_{a\in\mathcal{A}} \mu(a, x_1, \dots, x_n)$$

□

Now we may ask ourselves: what if we take $\mu_{n+1} = \mu\big([x_1 n + 1]\big)$? What is its measure? Notice that

$$[x_1^n] = \bigsqcup_{a\in\mathcal{A}} (x_1, \dots, x_n, a)\,,$$

i.e. the disjoint union of all sequences which have an arbitrary $n+1$ letter. Thus we can now state the *consistency relations* (or *compatibility conditions*):

Proposition 2.1.2 (Consistency relations).
μ is a measure on Ω iff:

(1) $\sum_{a_1,\dots,a_k\in\mathcal{A}} \mu_k(a_1, \dots, a_k) = 1$

(2) $\mu_n(x_1, \dots, x_n, \dots) = \sum_{a\in\mathcal{A}} \mu_{n+1}((x_1, \dots, x_n, \dots, a)$

iff μ is shift invariant,

(3) $\mu_n(x_1, \dots, x_n, \dots) = \sum_{a\in\mathcal{A}} \mu_{n+1}(a, x_1, \dots, x_n, \dots)$

We can now state a very important result:

Theorem 2.1.1 (Kolmogorov representation theorem). *If $\{\mu_k\}$ is a sequence of measure defining a process then there is a **unique** probability measure μ on $\mathcal{A}^{\mathbb{N}}$ such that, for each $k \geq 1$ and for each cylinder x_1^k*

$$\mu\big([x_1^k]\big) = \mu_k(x_1^k)$$

Basically, this is telling us that not only we can reconstruct the measure from the marginals, but that such measure will be the only one we can define on Ω given such marginals.

2.2 Examples of invariant measures

Let's see some basic examples of invariant measures.

2.2.1 i.i.d. (independent identically distributed variables)

Assume that we have a set of measures $p_1, p_2, \ldots, p_n$ over $\mathcal{A}$ such that

$$\begin{gathered} p_k : \mathcal{A} \longrightarrow \mathbb{R}^+ \\ p_k \geq 0 \\ \sum_{a \in \mathcal{A}} p_k(a) = 1 \quad \forall k \end{gathered} \tag{2.2.1}$$

Define then

$$\mu_n(x_1^n) = \mathbb{P}(x_1, \ldots, x_n) = \prod_k p_k(x_k) \tag{2.2.2}$$

choosing x_1 with probability p_1, x_2 with probability p_2 and so on. These marginals satisfy all three conditions; consider the first one for example:

$$\sum_{x_{n+1}} \mu(x_1, \ldots, x_{n+1}) = p_1(x_1) p_2(x_2) \ldots p_n(x_n) \Big[\sum_{x_{n+1}} p_{n+1}(x_{n+1}) \Big] = \mu(x_1, \ldots, x_n)$$

exercise: prove that $\forall p_1, \ldots, p_n$:

i) $\mu = \bigotimes_{j=1}^{\infty} p_j \in \mathcal{P}(\Omega)$

ii) $\mu \in \mathcal{P}_I(\Omega) \Leftrightarrow p_k = p : \mathcal{A} \to \mathbb{R}^+ \, \forall k$

2.2.2 Markov processes

Markov processes are the easiest stochastic processes with memory. Their definition is quite straightforward:

Definition 2.2.1 (Markov of order k).
A process is said to be Markov of order k if

$$\mathbb{P}(x_n = a_n \big| x_1^{n-1}) = \mathbb{P}(x_n = a_n \big| x_{n-k}^{n-1}), \quad n > k \tag{2.2.3}$$

A Markov process of order 1 *is called a Markov chain, i.e. a process is a Markov chain if*

$$\mathbb{P}(x_n = a_n \big| x_1^{n-1}) = \mathbb{P}(x_n = a_n \big| x_{n-1}) \tag{2.2.4}$$

Markov processes have been studied in the past decades to a great extent. We can completely describe a Markov process. To do such thing we have to introduce some objects which simplify our description.
First of all, let $\mathcal{A} = \{a_1, \ldots, a_l\}$ be our alphabet, with $|\mathcal{A}| = l$. In a Markov process, the a_k's are called *states*. Consider then a probability vector

$$\begin{gathered} p = (p_1, \ldots, p_l) \\ p_j \geq 0 \, , \, \sum_{j=1}^{l} p_j = 1 \end{gathered} \tag{2.2.5}$$

Let then $P = [p_{ij}]$ be a *stochastic* $l \times l$ *matrix* , i.e.

$$\begin{gathered} P_{ij} \geq 0 \\ \sum_j P_{ij} = 1 \quad \forall j \end{gathered} \tag{2.2.6}$$

The elements of such matrix represent the probability that from a state i we will transition to a state j in the next iteration of our system, i.e. is a matrix representing transition probabilities. That is why the second condition in eq. (2.2.2) exists: it is telling us that the probability to

transition from a state i to any of the other states j should be 1.
It can be proven that if P is stochastic, then it has an invariant vector $\vec{1}$ of all ones

$$P\vec{1} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

The Perron-Frobenius theorem §1.5 tells us then that all other eigenvalues are contained in a circle of radius 1.
A Markov process is thus defined by a couple $\big(p, P\big)$. p is invariant if it is a left-eigenvalue

$$pP = p \quad (\Leftrightarrow P^T p = p)$$

What measure should we use for a Markov process? A natural definition is, $\forall n$, $\forall x_1, \dots, x_n$, would be

$$\mu_n(x_1, \dots, x_n) \equiv p(x_1) \cdot P_{x_1 x_2} \cdot P_{x_2 x_3} \cdots P_{x_{n-1} x_n} \tag{2.2.7}$$

Proposition 2.2.1. *The measure μ defined by such marginals satisfies all the consistency relations 2.1.2*

Proof.

i) From their definition (2.2.7), $\forall (x_1, \dots, x_n) \in \mathcal{A}^n$ we have $\mu_n \geq 0$; also since p is a probability vector and P a stochastic matrix we will have $\forall n$ that $\sum_{(x_1,\dots,x_n)\in\mathcal{A}^n} \mu_n(x_1, \dots, x_n) = 1$

ii) The second property $\sum_{x_{n+1}} \mu_{n+1}(x_1, \dots, x_{n+1}) = \mu_n(x_1, \dots, x_n)$ also follows from the definition (2.2.7) and stochasticity of P

iii) The only thing left to prove is that $\sum_{x_0} \mu_{n+1}(x_0, x_1, \dots, x_n) = \mu_n(x_1, \dots, x_n)$. This is actually straightforward since

$$\sum_{x_0} p_{x_0} P_{x_0 x_1} P_{x_1 x_2} \cdots P_{x_{n-1} x_n} = p(x_1) P_{x_1 x_2} \cdots P_{x_{n-1} x_n}$$

where we used the fact that p is an eigenvector of P and thus $(pP)_{x_1} = p_{x_1}$

□

By Kolmogorov we then know that this is the unique invariant measure we can define on a Markov process (p, P).
An interesting and useful fact is that a markov of order k can always be reduced to order 1 by using a bigger alphabet (whose cardinality grows exponentially though, so that this becomes difficult to implement in concrete problems). Let's consider for example a Markov of order 2,

$$\mu(x_1, \dots, x_n \big| x_{n+1}) = \mu(x_1, \dots, x_n)\mu(x_{n+1} \big| x_1, \dots, x_n) = \mu(x_1, \dots, x_n \big| x_{n+1} x_n)$$

Now consider

$$\mathcal{A}^{(2)} = \big\{ a_j a_k \equiv b \big|_{a_k \in \mathcal{A}}^{a_j \in \mathcal{A}} \big\}$$

as the set of all bigrams. Of course if μ is a Markov of order 2 on $\mathcal{A}$ it can be seen as a Markov Chain on $\mathcal{A}^{(2)}$.
Consider now a Markov chain over $\mathcal{A}$ with $|\mathcal{A}| = l$; it may happen that some elements of P are zero, for example consider

$$\begin{aligned} p_j > 0 \quad & j = 1, \dots, k \\ p(a_j) = 0 \quad & j = k+1, \dots, l \end{aligned}$$

we can then reduce this to a Markov chain over $\mathcal{A}^k$, a smaller alphabet, just erasing rows and column of P to get a new $\hat{P}$ which will be $k \times k$,

$$(p, P) \text{ on } \mathcal{A}^l \longrightarrow (\hat{p}, \hat{P}) \text{ on } \mathcal{A}^k$$

2.2.3 Hidden Markow Models

A hidden Markov model (HMM) is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobservable ("hidden") states.
Let's consider two alphabets:

$$\begin{aligned} \mathcal{A} &= \{x_1, \dots, x_n\} \quad \text{observable states} \\ \mathcal{L} &= \{y_1, \dots, y_l\} \quad \text{hidden states} \end{aligned} \tag{2.2.8}$$

We assume that there exists a rectangular stochastic matrix R such that

$$\begin{aligned} R &= [R_{yx}]_{y\in\mathcal{L},\, x\in\mathcal{A}} \\ \sum_{x\in\mathcal{A}} R_{yx} &= 1 \quad \forall y \in \mathcal{L} \end{aligned} \tag{2.2.9}$$

The elements of R represent the so called *emission probabilities*, i.e. the probabilities to go from an hidden state $y \in \mathcal{L}$ to an observable state $x \in \mathcal{A}$. Summing on all observable states should amount to 1 since from an hidden state we must go to some observable state sooner or later (see Fig. 2.1).
Surely, we must also assume that there exists also a stochastic matrix for the hidden states with their transition probabilities: this would be the $l \times l$ matrix $P = [P]_{yz}$ $\forall y, z \in \mathcal{L}$. We shall also consider a probability vector p for this Markov system. An HMM is thus defined by a triplet (p, P, R)
Then we define the measure for this system as

$$\mu_n(x_1, \dots, x_n) = \sum_{(y_1,\dots,y_l)\in\mathcal{L}} p_{y_1} R_{y_1x_1} P_{y_1y_2} R_{y_2x_2} \cdots P_{y_{n-1}y_n} R_{y_nx_n} \tag{2.2.10}$$

exercise:

i) prove that μ_n are marginals (they statisfy the first 2 consistency relations)

ii) prove that if $pP = p$ then μ is invariant

There is an equivalent way of defining such measure: we can consider the matrix M_x $\forall x \in \mathcal{A}$ defined as

$$M_x = [m^{(x)}_{yz}]_{y,z\in\mathcal{L}} = \left[R_{yx}P_{yz}\right] \tag{2.2.11}$$

so that $P = \sum_{x\in\mathcal{A}} M_x$. These are again $l \times l$ matrices, and there are $n = |\mathcal{A}|$ of them.
This gives us an equivalent representation of an HMM, since considering a probability vector $p = (p_1, \dots, p_l)$ we can define the marginals as

$$\mu_n(x_1, \dots, x_n) \equiv pM_{x_1}M_{x_2}\cdots M_{x_n} \cdot \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \geq 0 \tag{2.2.12}$$

2.3 Ergodic Theory

We shall now look at some basics results of Ergodic Theory.
Consider a space $\mathcal{M}$ with a measure μ and a certain dynamic T which is measure preserving. Such transformation could be either discrete or continuous (a so called *flow* Φ^t). Given an $x_0 \in \mathcal{M}$ we are interested in studying the orbit under T, i.e. all the points $\mathcal{O}(x) = \{x \in \mathcal{M} \,|\, x_n = T^n(x_0),\, \forall n \in \mathbb{N}\}$.
The first thing that comes natural to study are functions from $\mathcal{M}$ to $\mathbb{R}$, out *observables* $f : \mathcal{M} \longrightarrow \mathbb{R}$.
We are interested in knowing the asymptotic statistical properties of such observables, i.e. we would like to know for each f what happens to the average

$$\lim_{n\to\infty} \frac{f(x_0) + f(x_1) + \cdots + f(x_n)}{n} = \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} f \circ T^k(x) \equiv f^*(x)$$

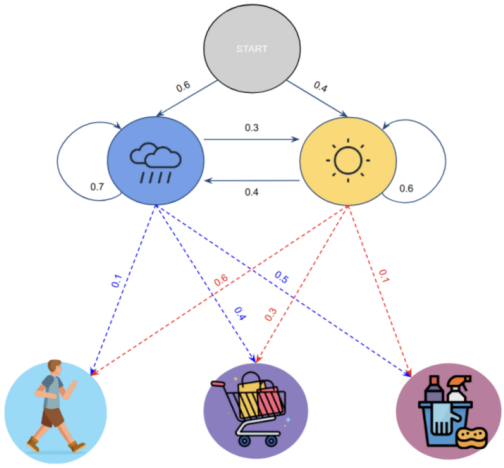


Figure 2.1: Schematic representation of an Hidden Markov Model. My friend lives in the States. Depending on the weather there, he chooses the activity he does every day. I am not able to know the weather where he lives (hidden states), I only know what he is doing that day (observable states).

Remark 2.4. *equivalently for flows we could consider*

$$\frac{1}{T}\int_0^T (f \circ \Phi^t)(x)\, d\mu \tag{2.4.1}$$

and then take the limit for $T \to \infty$

Let's then define the spatial mean of a function f, if it exists, as

$$\bar{f} = \int_{\mathcal{M}} f(x)\, d\mu \tag{2.4.2}$$

A very important classical result is Birkhoff Ergodic Theorem.

Theorem 2.4.1 (Birkhoff Ergodic Theorem). *Let* $(\mathcal{M}, T, \mu)$ *be a dynamical system, and* $f \in L^1(\mathcal{M}, \mu)$ *a summable function with complex values. Then*

a) $f^*(x)$ *exists* μ*-almost everywhere*

b) $f^*(x)$ *is summable and invariant almost everywhere, i.e.*

$$f^*(T^k(x)) = f^*(x) \quad \forall k \text{ a.e.}$$

c)

$$\int_{\mathcal{M}} f^*(x)\, d\mu = \int_{\mathcal{M}} f\, d\mu$$

Let's define a particular class of functions which will be very useful:

Definition 2.4.1 (Characteristic functions). *The characteristic function of an open set* $A \in \mathcal{M}$ *is defined as*

$$\chi_A(x) = \begin{cases} 1 & \text{if} \quad x \in A \\ 0 & \text{if} \quad x \notin A \end{cases} \tag{2.4.3}$$

exercise: prove that $\chi_A \circ T = \chi_{T^{-1}(A)}$, $\forall x \in \mathcal{M}$.

Proposition 2.4.1. *The following are equivalent:*

a) $\forall f \in L^1,\ f^* = c\ \mu\text{-a.e.}$

b) $\forall f \in L^1,\ f^* = \int_{\mathcal{M}} f\, d\mu = \bar{f}$

c) if $f \in L^1$ is invariant $\Rightarrow$ f =constant

d) if $A \subseteq \mathcal{M}$, A measurable and invariant $\Rightarrow$ $\mu(A) = 0, 1$

Proof.

- a) $\Rightarrow$ b):
  if $f \in L^1$ then $\Rightarrow$ f = constant by a). By point b of Birkoff's theorem we have then
  $$\int_{\mathcal{M}} c\, d\mu = \int_{\mathcal{M}} f^*\, d\mu = \int_{\mathcal{M}} f\, d\mu$$
- b) $\Rightarrow$ c):
  take f to be invariant, i.e. $f \circ T = f$: then $f \circ T^k = f\, \forall k \in \mathbb{N} \Rightarrow f^* = f$
- c) $\Rightarrow$ d):
  Apply c) to $f = \chi_A \in L^1$. Since $T^{-1}A = A \bmod \mu \Rightarrow \chi_{T^{-1}A} = \chi_A$ then we have by c) that χ_A =const. By the definition of χ then we must have const= 0 and $A = \emptyset$ or const= 1 and $A = \mathcal{M}$.
- d) $\Rightarrow$ a):
  claim : $f : \mathcal{M} \to \mathbb{R}$ is constant a.e. iff $\nexists l \in \mathbb{R}$ such that $\mu(\{x \in \mathcal{M} | f(x) = l\}) \in (0, 1)$, meaning that we only want μ to have a single jump from 0 to 1 and not multiple jumps.
  Consider now f^*. Suppose that $\nexists l \in \mathbb{R}$ s.t. for $A_l = \{x \in \mathcal{M} | f^*(x) = l\}$ s.t. $\mu(A_l) \in (0, 1)$. Clearly $T^{-1}A_l = A_l$. But then
  $$T^{-1}A_l = \{y \in \mathcal{M} |\, f^*(Ty) \leq l\} = A_l \Rightarrow f^*(Ty) = f^*(y)$$
  and we get a contradiction.

□

Thus ergodicity of a system can be also seen as the **equivalence of time mean and phase space means** of an observable (which is what Boltzmann had hoped was true for any dynamical system).
We can also say that

$$\%\text{ of time spent by } \mathcal{O}(x) \text{ in } A = \frac{\#\{k |\, x_k \in A\}}{n} \xrightarrow{n\to\infty} \mu(A) = \frac{\mu(A)}{\mu(\mathcal{M})}$$

This follows directly from Birkoff's theorem, since if we consider $\chi_A(x)$ then we'll have

$$\chi_A^* = \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \chi_A \circ T^k(x) = \int_{\mathcal{M}} \chi_A\, d\mu$$

Therefore, **in an ergodic system all orbits visit volumes in phase space for a time proportional to the volumes themself**.
Now we could ask ourselves: how does all the discussion about infinite sequences of finite alphabets relate to ergodic theory and its application? The key to deeply connect both arguments will be *partioning*.
Given our measurable dynamical system (X, μ, T), consider X as partitioned in disjoint sets P_a indexed by a finite alphabet $\mathcal{A}$, i.e. consider a finite partition $\mathcal{P}$ of X such that

$$\begin{aligned} \mathcal{P} &= \{P_a\}_{a\in\mathcal{A}} \\ X &= \bigsqcup_{a\in\mathcal{A}} P_a \end{aligned} \tag{2.4.4}$$

The P_a'a only intersect on boundaries (sets of zero measure) as shown in Fig. 2.2 Then associate

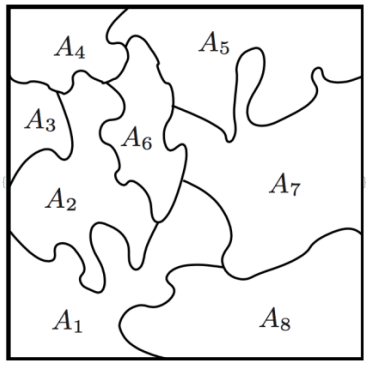


Figure 2.2: Schematic depiction of a partition of a manifold X.

each iteration of the dynamics $x_j = T^J(x_0)$ with a given symbol of $\mathcal{A}$ by the map

$$\begin{aligned} x_1 &\longrightarrow \omega_1 \\ x_2 &\longrightarrow \omega_2 \\ &\vdots \\ x_n &\longrightarrow \omega_n \\ &\vdots \\ \text{where} \quad x_j &\in P_{\omega_j} \end{aligned}$$

This allows us to equivalently consider an orbit $\mathcal{O}(x)$ or a sequence of characters from our alphabet. But what measure should we use for such sequences? A marginal $\mu_k(\omega_1^k)$ will now measure all the points which are in P_{ω_1} at t_1, P_{ω_2} at t_2 and so on (ω_1 and ω_2 possibly being the same symbol). Let's define then

$$\mu_k(\omega_1^k) = \mu\Big(\bigcap_{j=1}^{k} T^{-j+1} P_{\omega_j}\Big) \tag{2.4.5}$$

exercise: prove that such μ_k's are in fact marginals and that the transformation T commutes with the shift σ and that μ is then σ- invariant.

Definition 2.4.2. *The (stationary) process $\{X_n\}_{n\in\mathbb{N}}$ defined by the measure (2.4.5) is called a $(T, \mathcal{P})$-process*

We can now finally move on to the definition of ergodicity. This comes very naturally by the setting we are now working in. First of all, let's see a couple of definitions needed to state ergodicity:

Definition 2.4.3. *A measurable set $B \in X$ is said to be T-invariant if $TB \subseteq B$*

Definition 2.4.4. *The space X is said to be T-decomposable if it can be expressed as the disjoint union $X = X_1 \bigsqcup X_2$ of two measurable invariant[1] sets, each of positive measure. Hence, the space is indecomposable if*

$$T^{-1}B = B \Rightarrow \mu(B) = 0 \vee \mu(B) = 1$$

This is exactly what we mean for ergodicity.

Definition 2.4.5 (Ergodicity). *A measure-preserving tranformation T is said to be **ergodic** if*

$$T^{-1}B = B \Rightarrow \mu(B) = 0 \vee \mu(B) = 1 \tag{2.4.6}$$

Thus **ergodicity of our system means that we cannot subdivide it in smaller parts to study them separately**.
Notice that Hamiltonian systems are not ergodic: we can subdivide them in hypersurfaces where the hamiltonian $H = E$ is constant.

[1] since the sets are invariant, the orbit of a point in X_1 stays in X_2 and the same thing happens for X_2. Therefore we could study the two subsystem separately

Lemma 2.4.2 (Ergodicity equivalents). *The following are equivalent for a μ-measure preserving transformation T:*

1) T is ergodic

2) $T^{-1}B \subseteq B \Rightarrow \mu(B) = 0 \vee \mu(B) = 1$

3) $T^{-1}B \supseteq B \Rightarrow \mu(B) = 0 \vee \mu(B) = 1$

4) $\mu(T^{-1}B \Delta B) = 0 \Rightarrow \mu(B) = 0 \vee \mu(B) = 1$

5) $f \circ T = f$ a.e. $\Rightarrow f$ constant a.e.

The last claim can be proved by recalling that $\chi_A \circ T = \chi_{T^{-1}A}$ (*exercise*).
We will see later on what ergodicity of our system means in the setting we discussed earlier. Basically, it will coincide with an application of ergodicity itself, which is the convergence of empirical probabilities. But for now, let's stick to studying some other basic results of ergodic theory.
Sometimes, to show that a process is ergodic it is even simpler to prove a stronger property, called *mixing*.

Definition 2.4.6 (Mixing). *A measure preserving transformation T is mixing if for any pair of measurable sets C and D we have*

$$\lim_{n\to\infty} \mu\big(T^{-n}C \cap D\big) = \mu(C) \cdot \mu(D) \tag{2.4.7}$$

This definition is telling us that the fact that a system is mixing corresponds to the fact that correlations are asymptotically decaying[2]. Basically then *a mixing system is a system with no memory.*
exercise: prove that mixing $\Rightarrow$ ergodicity (and that the viceversa is not true).
Actually, some sort of viceversa of mixing $\Rightarrow$ ergodicity does exist, but with looser conditions: we can say that "mixing on average" $\Rightarrow$ ergodicity, in the following sense:

Theorem 2.4.3. *The dynamical system (X, μ, T) is ergodic if and only if $\forall A, B \subset X$*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} \mu\big(A \cap T^{-1}B\big) = \mu(A) \cdot \mu(B) \tag{2.4.8}$$

2.4.1 Ergodicity for Markov chains

When is a Markov chain irreducible? When we cannot have a zero probability to go from a state j to a state k.

Definition 2.4.7 (irreducible Markov chains). *The stochastic matrix M is said to be irreducible if for any pair j, k there is a certain sequence $j_0, j_1, \dots, j_n$ with $j_0 = j$ and $j_n = k$ such that each intermediate transition is possible, i.e.*

$$M_{j_m j_{m+1}} > 0\,, \quad m = 0, 1, \dots, n-1 \tag{2.4.9}$$

Namely, given any pair of states j and k, sooner or later strarting from j we'll end up in k.

This actually corresponds (*exercise: chech it*) to the definition of irreducible matrices given in par. §1.5, i.e. $\forall i, j\ \exists n = n(i,j) : (P^n)_{ij} > 0$. If M is irreducible then by Perron-Frobenius (§1.5) there is a **unique** probability vector π (the **equilibrium state**):

$$\pi M = \pi \tag{2.4.10}$$

i.e. the Markov chain with start distribution $\mu_1 = \pi$ and transition matrix M is in fact a stationary process.
Moreover, any initial distribution of characters will *converge to the invariant measure*: given any probability vector μ_1 on $\mathcal{A}$, we have

$$\lim_{n\to\infty} \mu_1 M^n = \pi \tag{2.4.11}$$

We can define the equivalent of the mixing property for a Markov chain:

[2] we could also have "exponential mixing" or "polynomial mixing" based on how the decay evolves with time ($\sim e^{-t}$ or $\sim t^{-\alpha}$)

Definition 2.4.8. *A Markov chain is mixing iff* $\exists\ n$ *s.t.* $\forall j,k$ *we have* M^n_{jk}.

This is a stronger requirement than irreducibility, as we are now asking that such n is not a function of the indexes i,j but instead exists for all couples. This definition of mixing is quite straightforward and equivalent to that given before: let's give it a better look.
Consider a Markov system (p,P) on a space $(\Omega=\mathcal{A}^{\mathbb{N}},\mu,T=\sigma)$. Consider then two cylinders $C=[x_1^m]$ and $D=[y_1^r]$. we know that mixing means for our two sets that

$$\lim_{n\to\infty}\mu(T^{-n}C\cap D)=\mu(C)\,\mu(D)$$

Of course $T^{-n}=\sigma^{-n}$ is just a translation, i.e. $T^{-n}(C)=\left\{z\in\Omega|\, z_{n+j}=y_j,\quad j=1,\dots,r\right\}$. What is then $T^{-n}C\cap D$? Consider an $n>m$ for simplicity. Then

$$T^{-n}C\cap D=\left\{z\in\Omega\middle|\begin{array}{cc} z_1=x_1 & z_{n+1}=y_1\\ \vdots & \vdots\\ z_m=x_m & z_{r+n}=y_r\end{array}\right\}$$

disegnino
The measure of this set in with our Markov measure will then be:

$$\begin{aligned}\mu(T^{-n}C\cap D)&=p_{x_1}P_{x_1x_2}\cdots P_{x_{m-1}x_m}\cdot\left(\sum_{z_1,\dots,z_k}P_{z_1z_2}\cdots P_{z_{t-1}z_t=y_1}\right)P_{y_1y_2}\cdots P_{y_{r-1}y_r}\\&=p_{x_1}P_{x_1x_2}\cdots P_{x_{m-1}x_m}\cdot\left(P^t\right)_{x_my_1}P_{y_1y_2}\cdots P_{y_{r-1}y_r}\end{aligned}\tag{2.4.12}$$

since we must account for all possible paths from x_m to y_1 in t steps. A very useful classical result is the following:

Theorem 2.4.4 (Ergodicity for Markov chains).

(1) A stationary Markov chain is ergodic iff its transition matrix M is irreducible

(2) If the Markov chain is mixing, then it is ergodic

Some more useful results for Markov chains' ergodicity are the following.

Lemma 2.4.5. *(1.18 in Walter) Let P be a stochastic matrix and P^n its power. Take then a "Birkoff average" of P,*

$$Q=\lim_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n-1}P^n\tag{2.4.13}$$

Such matrix is a spectral projector, i.e. it satisfies $Q^2=Q$; moreover, it commutes with P, $QP=PQ$ and $\forall v$ such that $vP=v$ we have that $vQ=Q$.

The existence of Q can be proven by spectral theory. By using Q we can state two theorems which can result useful:

Theorem 2.4.6. *The following are equivalent:*

a) (Ω,μ,T) is ergodic;

b) Q as defined in (2.4.13) is an $l\times l$ matrix of the form

$$Q=\begin{pmatrix}\vec{p}\\ \vec{p}\\ \vdots\\ \vec{p}\end{pmatrix}$$

where p is a probability vector (i.e. $pP=p$, $p_j\geq 0$, $\sum_j p_j=1$)

c) P is irreducible

d) $\vec{1}$ *s.t.* $P\vec{1} = \vec{1}$ *is a simple eigenvector (but not the only eigenvector).*

There is an equivalent of the last theorem for mixing:

Theorem 2.4.7. *The following are equivalent:*

a) (Ω, μ, T) *is mixing;*

b) P *is irreducible and aperiodic*

c) $P^n \xrightarrow{n\to\infty} Q$ *with*

$$Q = \begin{pmatrix} \vec{p} \\ \vec{p} \\ \vdots \\ \vec{p} \end{pmatrix}$$

d) $\vec{1}$ *s.t.* $P\vec{1} = \vec{1}$ *is a simple eigenvector* ***and*** *it is the unique eigenvalue on* S^1 *(the circle).*

2.5 Consequences of ergodicity

Definition 2.5.1 (empirical frequencies). *Given a sequence* $x = x_1^n \in \mathcal{A}^n$ *of length* n *and a* k*-gram* $\omega = \omega_1\omega_2\ldots\omega_k \in \mathcal{A}^k$*, we define the* ***empirical frequency*** *of the word* ω *by*

$$f_x(\omega) \equiv |\{j \in [1, n-k+1] : x_j^{j+k-1} = \omega\}| \tag{2.5.1}$$

Denoting by χ_ω the indicator function of the cylinder $[\omega_1^n]$ we could also write

$$f_x(\omega) = \sum_{j=1}^{n-k+1} \chi_\omega(\sigma^{j-1}x) \tag{2.5.2}$$

Meaning that, since we are looking at how many times a word appears in a sequence, we could do it by sliding a window of size k over our sequence e count a $+1$ for every match.

Definition 2.5.2 (relative frequency). *The relative frequency is instead defined as*

$$\hat{\mu}_n(\omega_1^k|x_1^n) \equiv \frac{1}{n-k+1} f_x(\omega_1^k) \tag{2.5.3}$$

Definition 2.5.3 (typical sequences). *A sequence* $x \in \mathcal{A}^{\mathbb{N}}$ *is said to be frequency* ***typical*** *for the process* μ *if each word* $\omega \in \mathcal{A}^{\mathbb{N}}$ *appears in* x *with the statistics predicted by* μ*, i.e.*

$$\lim_{n\to\infty} \hat{\mu}_n(\omega_1^k|x_1^n) = \mu(\omega_1^k) \tag{2.5.4}$$

The set of all typical sequences is denoted by $\mathcal{T} \subseteq \mathcal{A}^{\mathbb{N}}$*:*

$$\mathcal{T} \equiv \{x \in \mathcal{A}^{\mathbb{N}} : x \text{ is typical}\} \tag{2.5.5}$$

If my system is ergodic and it behaves nicely, then I would expect that if I look at the typical sequence x this should contain ω with the same proportion given by the invariant measure. We do not expect typically to have a million zeros and a single one from a coin toss with a fair coin.
So, **an ergodic system would be one for which we can reconstruct** μ **from the data** (all words appear typically).

Theorem 2.5.1 (Typical-sequence theorem). *If* μ *is ergodic*

$$\mu\big(\mathcal{T}(\mu)\big) = 1 \tag{2.5.6}$$

that is, for all $\omega \in \mathcal{A}^{\mathbb{N}}$

$$\lim_{n\to\infty} \hat{\mu}_n(\omega_1^k|x_1^n) = \mu(\omega_1^k), \quad \text{for almost every } x$$

Note that this coincides with the property of ergodic systems not being decomposable in smaller parts. The last theorem is also true in the other way around:

Theorem 2.5.2 (Typical sequence converse theorem). *Suppose μ is a stationary measure such that for eah $\omega \in \mathcal{A}^{\mathbb{N}}$ the limiting relative frequencies $\hat{\mu}(\omega|x)$ exist and are constant in x, $\mu-$almost surely.*
Then,

$$\mu \text{ is an ergodic measure}$$

Now we want to state consequences of ergodicity in a finite version. To achieve this we must first give two definitions:

Definition 2.5.4 (G and B sets). *Define the good sequences of length n as:*

$$G_n(k,\epsilon) = \{x_1^n : |\hat{\mu}(\omega|x_1^n) - \mu(\omega)| < \epsilon,\, \omega \in \mathcal{A}^k\}\,, \tag{2.5.7}$$

whereas the corresponding bad set is just its complement:

$$B_n(k,\epsilon) = \mathcal{A}^n \setminus G_n(k,\epsilon)$$

We will also have that

$$\frac{|G_n|}{|\mathcal{A}^n} \longrightarrow 0$$

in the sense that objects in G_n have a very big probability to appear, but they are few compared to all of those that can be formed by characters in $\mathcal{A}$. For example in language, there are very few words that appear frequently with respect to all of the sequences of characters that could be formed in language.
Basically, defining a good set is the same as saying that we can reconstruct the measure μ up to ϵ for all ω of such set. We will have "typical sequences" but not in the limit $n \to \infty$.

Theorem 2.5.3 (The good set form of ergodicity).

(1) If μ is ergodic then $x_1^n \in G_n(k\epsilon)$ eventually almost surely

(2) If $\lim_{n\to\infty} \mu_n(G_n(k\epsilon)) = 1$ for every integer $k > 0$ and every $\epsilon > 0$, then μ is ergodic.

Theorem 2.5.4 (The finite form of ergodicity). *A measure μ is ergodic iff for each $\omega \in \mathcal{A}^{\mathbb{N}}$ and $\epsilon > 0$ there is an $N = N(\omega,\epsilon)$ such that if $n > N$ there is a collection $\mathcal{C}_n \subset \mathcal{A}^n$ such that*

(1) $\mu(\mathcal{C}_n) > 1 - \epsilon$

(2) If $x_1^n, y_1^n \in \mathcal{C}_n$ then $|\hat{\mu}(\omega|x_1^n) - \hat{\mu}(\omega|y_1^n)| < \epsilon$

Chapter 3

Entropy

3.1 Entropy and entropy rate

The definition of entropy of a random variable X was originally given by Shannon in 1949: Shannon stated that his main concern while studying such function was how to call it: luckily, while confronting his results wiht mathematician J. Von Neumann, he suggested that Shannon would call his function "Entropy", since, he stated:
"*First: your function of uncertainty is already known in statistichal mechanics with that name; second and most of all, no one actually knows what entropy is, so you'll always have the upper hand in any discussion.*"
Let X be a random variable which takes values in the alphabet $\mathcal{A} = \{a_1, \dots, a_k\}$ with probabilities $\mu_j = \mathbb{P}(X = a_j)$. Then its entropy is defined as

$$H(X) \equiv -\sum_{j=1}^{k} \mu_j \log \mu_j \tag{3.1.1}$$

More generally, one could define Entropy as the expectation value of the information, which is defined as the logarithm of the probability mass function p:

$$H(X) = \mathbb{E}[I(X)] = \mathbb{E}[-\log p(X)]$$

The core idea of information theory is that the "informational value" of a communicated message depends on the degree to which the content of the message is surprising. If a highly likely event occurs, the message carries very little information. On the other hand, if a highly unlikely event occurs, the message is much more informative.
For instance, the knowledge that some particular number will not be the winning number of a lottery provides very little information, because any particular chosen number will almost certainly not win. However, knowledge that a particular number will win a lottery has high informational value because it communicates the outcome of a very low probability event.
Entropy in information theory is directly analogous to the entropy in statistical thermodynamics. The analogy results when the values of the random variable designate energies of microstates, so Gibbs formula for the entropy is formally identical to Shannon's formula.

3.1.1 Basic properties of the Entropy

We now turn to look to some very basic but important properties of Entropy.
Let's be now more precise: consider a finite set $\mathcal{B} = \{b_1, \dots, b_l\}$, so that $|\mathcal{B}| = L$, and let $\mathcal{P}(\mathcal{B}) = \{(p_1, \dots, p_l) : \sum_j p_j = 1,\ p_j \geq 0\}$ be the set of all probability measures on $\mathcal{B}$ such that $p(b_k) = p_k$. Given $p \in \mathcal{P}(\mathcal{B})$ let's again define

$$\mathcal{S} : \mathcal{P}(\mathcal{B}) \longrightarrow \mathbb{R}^+$$

$$\mathcal{S}(p) = -\sum_{b \in \mathcal{B}} p(b) \log p(b) = -\sum_{k=1}^{l} p_k \log p_k \tag{3.1.2}$$

Such function satisfies the following conditions:

Proposition 3.1.1 (Entropy properties).

(1) $\boldsymbol{\mathcal{S}(p) \geq 0}$ ***and*** $\boldsymbol{\mathcal{S}(p) = 0}$ ***iff p is pure****, i.e. p is "concentrated" in one point,* $p = (0, 0, \ldots, 0, 1, 0, \ldots, 0)$*. For example, a source emitting only one symbol with probability* 1*.*

(2) $\boldsymbol{0 \leq \mathcal{S}(p) \leq \log L}$*.*
Specifically, $\mathcal{S}(p) = \log L \Leftrightarrow p = (\frac{1}{L}, \ldots, \frac{1}{L})$*, i.e. when we have a sequence that is "maximally random" (a so called "chaotic sequence").*

(3) $\mathcal{S} : \mathcal{P}(\mathcal{B}) \longrightarrow [0, \log L]$ ***is continous and concave****.*
Concave meaning that if we take $p_1, \ldots p_m \in \mathcal{P}(\mathcal{B})$ *and consider a finite weighted linear combination* $\alpha_1 p_1 + \cdots + \alpha_m p_m$ *where the weights satisfy* $\sum_j \alpha_j = 1$*,* $\alpha_j \geq 0$*, then*

$$\begin{aligned} \mathcal{S}(\alpha_1 p_1 + \cdots + \alpha_m p_m) &\geq \alpha_1 \mathcal{S}(p_1) + \ldots \alpha_m \mathcal{S}(p_m)\,, \\ \mathcal{S}(\alpha_1 p_1 + \cdots + \alpha_m p_m) &= \alpha_1 \mathcal{S}(p_1) + \ldots \alpha_m \mathcal{S}(p_m) \Leftrightarrow p_j = p_k \; \forall k \end{aligned} \tag{3.1.3}$$

(4) $\boldsymbol{\mathcal{S}}$ ***is almost convex****, i.e.*

$$\mathcal{S}(\alpha_1 p_1 + \cdots + \alpha_m p_m) \leq \alpha_1 \mathcal{S}(p_1) + \ldots \alpha_m \mathcal{S}(p_m) + \mathcal{S}(\alpha_1, \ldots, \alpha_n) \tag{3.1.4}$$

where $\mathcal{S}(\alpha_1, \ldots, \alpha_n) = -\sum_j \alpha_j \log \alpha_j$*. Notice that* α *satisfies the properties of a measure, so that the weighted combination between* α *and p can be looked at as a "mixture" of the two, meaning that when two measures "intersect" then their Entropy is no longer just additive.*

Recall that a function $f(x)$ is said to be convex over the interval (a, b) if $\forall x_1, x_2 \in (a, b)$ and $0 \leq \lambda \leq 1$:

$$f(\lambda x_1 + (1 - \lambda) x_2) \leq \lambda f(x_1) + (1 - \lambda) f(x_2)$$

The function f is strictly convex if the equality holds only if $\lambda = 0$ or $\lambda = 1$. Also, recall that (this can be prove by Taylor expansion of f) if the function f has a second derivative that is non-negative (positive) over an interval, the function is convex (strictly convex) over that interval.
Let's now prove the proposition 3.1.1.

Proof.

(2) follows from the Jensen Inequality (stated below)

(1),(3) follow from the fact that $f(x) = -x \log x = 0 \Leftrightarrow x = 0$ or $x = 1$

(4)

$$\begin{aligned} \mathcal{S}(\alpha_1 \mu_1 + \cdots + \alpha_d \mu_d) &= \sum_{k=1}^{d} \sum_{b \in \mathcal{B}} -\alpha_k \mu_k(b) \log \big(\sum_{j=1}^{d} \alpha_j \mu_j(b) \big) \leq \\ &\leq \sum_{k=1}^{d} \sum_{b \in \mathcal{B}} -\alpha_k \mu_k(b) \log(\alpha_k \mu_k(b)) = \\ &= \sum_{k=1}^{d} \sum_{b \in \mathcal{B}} -\alpha_k \mu_k(b) \log \alpha_k - \alpha_k \mu_k(b) \log \alpha_k = \\ &= \sum_{k=1}^{d} \alpha_k \mathcal{S}(\mu_k) + \mathcal{S}(\alpha_1 \mu_1, \ldots, \alpha_d) \end{aligned}$$

where the first inequality is trivial due to the monotonicity of the logarithm.

□

Theorem 3.1.1 (Jensen Inequality). *Let $f : \mathbb{R} \to \mathbb{R}$ be convex. Then if we consider any finite set $\{\alpha_j\}$ such that $\sum_{j=1}^{d} \alpha_j = 1$ we have that*

$$f(\alpha_1 x_1 + \cdots + \alpha_d x_d) \geq \sum_{j=1}^{d} \alpha_j f(x_d)$$

In the context of probability theory, it is generally stated in the following form:
If f is a convex function and X is a random variable, then

$$\mathbb{E}\big[f(X)\big] \geq f\big(\mathbb{E}\big[X\big]\big) \tag{3.1.5}$$

We will prove this in a simple setting, i.e. for atomic measures.

Proof. For a two-mass-point distribution, directly from the definition of convex function we have

$$p_1 f(x_1) + p_2 f(x_2) \geq f(p_1 x_1 + p_2 x_2)$$

So the Jensen inequality is true in this case. Now we use induction. Suppose the theorem being true for $k-1$ atoms. Writing $p_1' = p_i/1 - p_k$ for $i = 1, 2, \ldots, k-1$:

$$\begin{aligned}
\sum_{i=1}^{k} p_i f(x_i) &= p_k f(x_k) + (1 - p_k) \sum_{i=1}^{k-1} p_i' f(x_i) \geq \\
&\geq p_k f(x_k) + (1 - p_k) f\big(\sum_{i=1}^{k-1} p_i' f(x_i)\big) \geq \\
&\geq f\big(p_k x_k + (1 - p_k) \sum_{i=1}^{k-1} p_i' f(x_i)\big) = \\
&= f(\sum_{i=1}^{k} p_i x_i)
\end{aligned}$$

□

3.1.2 n-block Entropy

Let's now generalize to the case where $\mathcal{B} = \mathcal{A}^n$. We define the $n-$block entropy of a stochastic process as

$$H_n(\mu) \equiv - \sum_{\omega_1^n \in \mathcal{A}^n} \mu(\omega) \log \mu(\omega) \tag{3.1.6}$$

which we sometime denote equivalently as $H_n(\omega_1^n) = H(\omega_1, \ldots, \omega_n)$.
Let now $\mathcal{A}$ be separable in the cartesian product of two alphabets, $\mathcal{A}^n = \mathcal{A}^l \times \mathcal{A}^r$, with $l + r = n$.
Let's then denote $b \in \mathcal{A}^n$ as $b = (a^l, a^r)$ with $a^l \in \mathcal{A}^l$ and $a^r \in \mathcal{A}^r$.
Consider now $\mu : \mathcal{A}^n \to [0, 1]$. We know of course that

$$\sum_{b \in \mathcal{A}^n} \mu(b) = \sum_{a^l \in \mathcal{A}^l, a^r \in \mathcal{A}^r} \mu(a^l, a^r) = 1$$

and that we can define the marginals

$$\begin{aligned}
\mu_l(\cdot) &= \sum_{a^r \in \mathcal{A}^r} \mu(\cdot, a^r) \\
\mu_r(\cdot) &= \sum_{a^l \in \mathcal{A}^l} \mu(a^l, \cdot)
\end{aligned}$$

Given now a word $\omega \in \operatorname{supp} \mu_l = \{b \in \mathcal{A} | \mu_l(b) > 0\}$ we can define the **conditional** $\forall \omega' \in \mathcal{A}^r$

$$\mu_{r|l}^{\omega}(\omega') = \frac{\mu(\omega, \omega')}{\mu_l(\omega)} \tag{3.1.7}$$

or equivalently

$$\mu(\omega,\omega') = \mu_l(\omega)\cdot\mu^{\omega}_{r|l}(\omega') \tag{3.1.8}$$

Now in this setup, if we consider again the Entropy $\mathcal{S}:\mathcal{P}(\mathcal{B})\to[0,\log L]$ with $|\mathcal{A}^n|=L$, we can prove that

Proposition 3.1.2 (Conditional Entropy).

(a) $\mathcal{S}(\mu)=\mathcal{S}(\mu_l)+\sum_{\omega\in\mathcal{A}^l}\mu_l(\omega)\mathcal{S}(\mu^{\omega}_{r|l})$

i.e., the Entropy of the whole system is given by the entropy of the subsystem + conditional entropy, weighted on $\mu_l(\omega)$

(b) $\mathcal{S}(\mu)\le\mathcal{S}(\mu_l)+\mathcal{S}(\mu_r)$;
$\mathcal{S}(\mu)=\mathcal{S}(\mu_l)+\mathcal{S}(\mu_r)\Leftrightarrow\mu=\mu_l\otimes\mu_r$

This definition also for $n-$block entropy, namely it is connected to the definition of the Entropy rate:

Definition 3.1.1 (Entropy rate and n-conditional Entropy).

$$h_n(\mu)\equiv H_{n+1}(\mu)-H_n(\mu)=-\sum_{\omega_1^n\in\mathcal{A}^n,a\in\mathcal{A}}\mu(\omega_1^n a)\log\mu(a|\omega_1^n) \tag{3.1.9}$$

Such object can be given the interpretation of information that we get by knowing a new character (adding a character to our sequence). But what exactly is H_{n+1}? How come that conditional Entropy gets in the picture? We can actually directly compute H_{n+1}: consider a decomposition $\mathcal{A}^{n+1}=\mathcal{A}^n\times\mathcal{A}$ so that $(\omega_1\dots\omega_{n+1})\equiv(\omega_1^n,\omega_{n+1})$. Then,

$$\begin{aligned}H_{n+1}&=-\sum_{\omega_1\dots\omega_n,\omega_{n+1}}\log\mu(\omega_1\dots\omega_n\omega_{n+1})=\\&=-\sum_{\omega_1^n\in\mathcal{A}^n,\omega_{n+1}\in\mathcal{A}}\mu(\omega_1^n,\omega_{n+1})\log\big(\mu(\omega_1^n)\mu(\omega_{n+1}|\omega_1^n)\big)=\\&=-\sum_{\omega_1^n\in\mathcal{A}^n,\omega_{n+1}\in\mathcal{A}}\mu(\omega_1^n,\omega_{n+1})\big\{\log\mu(\omega_1^n)+\log\mu(\omega_{n+1}|\omega_1^n)\big\}\end{aligned}$$

Now, since we can first sum on the last character and we know that $\sum_{\omega_{n+1}\in\mathcal{A}}\mu(\omega_1,\dots,\omega_n,\omega_{n+1})=\mu_n(\omega_1,\dots,\omega_n)$, we finally get

$$H_{n+1}=H_n-\sum_{\omega_1^n\in\mathcal{A}^n,\omega_{n+1}\in\mathcal{A}}\mu(\omega_1^n\omega_{n+1})\log\mu(\omega_{n+1}|\omega_1^n)$$

We can also generalize this result to the case of H_{n+m}: consider $\mathcal{A}^{n+m}=\mathcal{A}^n\times\mathcal{A}^m$ and $(\omega_1,\dots,\omega_{n+m})\equiv(\omega^1\in\mathcal{A}^n,\omega_2\in\mathcal{A}^m)$. Then

$$\begin{aligned}H_{n+m}&=-\sum_{\omega^1\in\mathcal{A}^n,\omega_2\in\mathcal{A}^m}\mu(\omega^1,\omega^2)\log\mu(\omega^1,\omega^2)=\\&=-\sum_{\omega^1\in\mathcal{A}^n,\omega_2\in\mathcal{A}^m}\mu(\omega^1,\omega^2)\big(\log\mu(\omega^1)+\log\mu(\omega^2|\omega^1)\big)=\cdots=\\&=H_n(\mu)-\sum_{\omega^1\in\mathcal{A}^n}\mu(\omega^1)H_m(\mu^{\omega^1}(\cdot))\end{aligned}$$

where the first passage comes from the fact that $\mu(\omega^1,\omega^2)=\mu(\omega^1)\mu(\omega^2|\omega^1)$ and $\mu^{\omega^1}(\cdot)$ is the conditional entropy to ω^1 as introduced before.
exercise: finish the demonstration
Also, if we have $\mu:\mathcal{A}^{n+1}\to\mathbb{R}$ and a function $f:\mathcal{A}^{n+1}\to\mathbb{R}$ we can define very naturally the expectation value of f,

$$\mathbb{E}_{\mu_{n+1}}[f]\equiv\sum_{\omega_1\dots\omega_{n+1}\in\mathcal{A}}\mu(\omega_1\dots\omega_{n+1})f(\omega_1\dots\omega_{n+1})$$

In this way we also get

$$h_n(\mu) = \mathbb{E}_{\mu_{n+1}}[\log \mu(a|\omega_1^n)]$$

After the definition of the Entropy rate, we can define the "Entropy per character" $\frac{H_n(\mu)}{n}$ [1].
We then define the Entropy of the source μ to be

Definition 3.1.2.

$$h(\mu) = \lim_{n\to\infty} \frac{H_n(\mu)}{n} = \lim_{n\to\infty} h_n(\mu) = \mathbb{E}_\mu(\log \mu(a|\omega_1^\infty)) \tag{3.1.10}$$

where the last expression denotes the average value of "infinite information", as if we knew all of the history of the system (future and past), i.e. as if we had an infinite data sample.
The existence of the limit (3.1.10) is guaranteed by the subadditivity lemma:

Lemma 3.1.2 (subadditivity lemma). *If $\{x_n\}$ is a sequence of nonnegative numbers, which is subadditice, i.e.*

$$x_{n+m} \leq x_n + x_m$$

then

$$\lim_{n\to\infty} \frac{x_n}{n} = \liminf_{n\to\infty} \frac{x_n}{n}$$

Proof. Left as an exercise □

Notice that since $h_k = H_{k+1} - H_k$ we have that

$$h(\mu) \leq \dots h_k(\mu) \leq h_{k-1}(\mu) \leq \cdots \leq h_1(\mu) \leq H_1(\mu)$$

exercise: prove it

Two classical and easy examples

Let X be a i.i.d. process, with $\mu_k = \mathbb{P}(X_j = a_k)$, $k = 1, \dots, |\mathcal{A}|$.
Then we have that

$$h(X) = \lim_{n\to\infty} \frac{nH(X_j)}{n} = H(X_j) \tag{3.1.11}$$

i.e., the entropy of the process is equal to the entropy of any of its variables.
Let X be a stationary $k-$Markov: the Markov property $\mu(\omega_{n+1}|\omega_1 \dots \omega_n) = \mu(\omega_{n+1}|\omega_{n-k+1} \dots \omega_n)$ ensures that

$$h_n = h_k \quad \forall n \geq k \tag{3.1.12}$$

i.e., the entropy stabilises to a certain value. Since this is an if and only if, **we can also infer if a system is k-Markov if h_n stabilises to a certain value.** We will see the proof of this statement later on (see §3.3).

3.2 Cross and Relative entropy

In this section we focus on properties involving pairs of stochastic sources on the same alphabet with distributions μ and ν: the ***cross entropy*** and the related ***relative entropy*** (or ***Kullback-Leibler divergence***)
Firs, let's define the $n-$conditional cross entropy as

$$h_n(\mu||\nu) = -\sum_{\omega\in\mathcal{A}^n, a\in\mathcal{A}} \mu(\omega a) \log \nu(a|\omega) \tag{3.2.1}$$

and then we define

Definition 3.2.1 (cross entropy).

$$h(\mu||\nu) = \lim_{n\to\infty} h_n(\mu||\nu) \tag{3.2.2}$$

[1] notice that $0 \leq H_n(\mu) \leq \log|\mathcal{A}^n| = \log|\mathcal{A}|^n$ so that $\frac{H_n(\mu)}{n} \leq \log|\mathcal{A}|$

and moreover we define

Definition 3.2.2 (relative entropy (Kullback-Leibler divergence)).

$$d(\mu||\nu) = \lim_{n\to\infty} \mathbb{E}\left[\log \frac{\mu(\omega_n|\omega_1^{n-1})}{\nu(\omega_n|\omega_1^{n-1})}\right] = \tag{3.2.3}$$

$$= \lim_{n\to\infty} \sum_{\omega_1^n \in \mathcal{A}^n} \mu(\omega_1^n) \log \frac{\mu(\omega_n|\omega_1^{n-1})}{\nu(\omega_n|\omega_1^{n-1})} \tag{3.2.4}$$

The Kullback-Leibler divergence is not symmetric, but can easily be symmetrized by defining

$$\Delta(\mu,\nu) = \frac{1}{2}d(\mu||\nu) + \frac{1}{2}d(\nu||\mu) \tag{3.2.5}$$

Now, such function is symmetric and it can be proven (as we will see later on) to be positive, i.e. $\Delta = 0 \Leftrightarrow \mu = \nu$ and $\Delta \geq 0$. Therefore it *looks like a distance*, even though it doesn't satisfy the triangular inequality.
The relative entropy is a measure of the distance between two distributions. In statistics, it arises as an expected logarithm of the likelihood ratio. The relative entropy $d(p||q)$ is a measure of the inefficiency of assuming that the distribution is q when the true distribution is p. For example, if we knew the true distribution p of the random variable, we could construct a code with average description length $H(p)$. If, instead, we used the code for a distribution q, we would need $H(p) + d(p||q)$ bits on the average to describe the random variable.
Entropy and cross entropy can be related to the asymptotic behavior of properly defined *returning times* and *waiting times*, respectively, which are two mathematical objects widely used in applications.

Definition 3.2.3 (Returning time). *I read the first $n-$word and ask myself how much time I have to wait until I see again such sequence:*

$$R(\omega_1^n) = \min k > 1 : \omega_k^{k+n-1} = \omega_1^n \tag{3.2.6}$$

This object is purely combinatoric, it depends only on the sequence and not on the probability of seeing the sequence: but as we'll see, it can give information on the source itself.
If ω is emitted by an ergodic source, such time exists and is finite[2]. Supposing now to have two sources which emit two sequences $z \in \mu$ and $\omega \in \nu$. We can define

Definition 3.2.4 (Waiting time). *I read the first $n-$word in the sequence ω and ask myself how much time I have to wait until I see again such sequence in the other sequence z:*

$$W(\omega_1^n, z) = \min k > 1 : z_k^{k+n-1} = \omega_1^n \tag{3.2.7}$$

Note that $W(\omega_1^n, \omega) = R(\omega_1^n)$.
There are two important results which we are not going to prove about such times we defined above:

Theorem 3.2.1 (Entropy and returning time). *If μ is a stationary and ergodic process, then*

$$\lim_{n\to\infty} \frac{1}{n} \log R(\omega_1^n) = h(\mu) \quad \mu - a.s. \tag{3.2.8}$$

This connects our data with the source. Again, R is a purely topological quantity, combinatorial, unrelated in principle to the source; but if we assume that the sequence ω is typical with respect to the source μ, then R converges to the Entropy. This is a very deep result which we can proved starting from the SMB Theorem (see 4.2.2).
Moreover, we have that

Theorem 3.2.2 (Relative entropy and waiting time). *If μ is a stationary and ergodic process, ν is $k-$Markov and $\mu_n << \nu_n$, then*

$$\lim_{n\to\infty} \log W(\omega_1^n, z) = h(\mu) + d(\mu||\nu) = h(\mu||\nu) \quad \mu \times \nu - a.s. \tag{3.2.9}$$

[2]Poincaré Recurrence Theorem

In relation with the Kullback-Liebler we have, thanks to the Jensen Inequality 3.1.1, an important result:

Theorem 3.2.3 (Information inequality). *For any couple of probability distributions μ and ν for which $d(\mu||\nu)$ is defined,*

$$d(\mu||\nu) \geq 0 \tag{3.2.10}$$

and the equality holds iff $\nu = \mu$.

Proof. Let $A = x : p(x) > 0$ be the support of $p(x)$. Then

$$\begin{aligned} -d(\mu||\nu) = & \\ &= \sum_{a\in A} \mu(a) \log \frac{\mu(a)}{\nu(a)} = \\ &= \sum_{a\in A} \mu(a) \log \frac{\nu(a)}{\mu(a)} \leq \\ &\leq \log \big(\sum_{a\in A} \mu(a) \frac{\nu(a)}{\mu(a)} \big) = \\ &= \log \big(\sum_{a\in A} \nu(a) \big) = \log 1 = 0 \end{aligned}$$

□

3.3 Mutual information

Let's now consider a pair of discrete random variables (x, y) with joint distribution $p(x, y)$ on $\mathcal{A} \times \mathcal{A}$.
We denote by μ and ν the marginals

$$\mu(x) = \sum_{y\in\mathcal{A}} p(x, y)\,, \qquad \nu(y) = \sum_{x\in\mathcal{A}} p(x, y)\,. \tag{3.3.1}$$

The conditional probabilities $p(x|y)$ and $p(y|x)$ are then defined as already seen in par. §3.1.2 by

$$p(x, y) = \mu(x)p(y|x) = \nu(y)p(y|x) \tag{3.3.2}$$

In general of course $p(x, y) \neq \mu(x) \cdot \nu(y) = \mu \otimes \nu$. At this point we define the **joint entropy**

$$h(x, y) = -\sum_{x\in\mathcal{A}} \sum_{y\in\mathcal{A}} p(x, y) \log p(x, y) = -\mathbb{E}_p(\log p(x, y)) \tag{3.3.3}$$

and the **conditional entropy** $h(y|x)$ of the two random variables as

$$h(y|x) = \mathbb{E}_p(\log p(y|x)) = \tag{3.3.4}$$

$$= -\sum_{x\in\mathcal{A}} \sum_{y\in\mathcal{A}} p(x, y) \log p(y|x) = \tag{3.3.5}$$

$$= -\sum_{x\in\mathcal{A}} \mu(x) \sum_{y\in\mathcal{A}} p(y|x) \log p(y|x) \tag{3.3.6}$$

There exists a theorem which tells us that the joint entropy is related to the conditional entropy in a way that resembles the chain rule in analysis:

Theorem 3.3.1 (Chain rule).

$$h(X, Y) = h(X) + h(Y|X) \tag{3.3.7}$$

While of course if X and Y are independent then $h(X, Y) = h(X)+h(Y)$. Otherwise, $h(Y|X) \leq h(Y)$; conditioning y with x makes me loose information about the system.

Corollary 3.3.1.1.

$$h(X,Y|Z) = h(X|Z) + h(Y|X,Z) \tag{3.3.8}$$

and

$$h((X,Y)|Z) \leq h(X|Z) + h(Y|Z) \tag{3.3.9}$$

with equality if and only if X and Y are conditional independent given Z.

Now we can see, thanks to the chain rule corollary, the proof of the statement 3.1.12 which we recall here:

Theorem 3.3.2 (Markov order theorem). *A stationary process μ is Markov of order k if and only if $h_n = h_k$, for all $n \geq k$, i.e. if and only if*

$$H(X_0|X_{-n}^{-1}) = H(X_0|X_0^{-1}) \quad \forall n \geq k\,. \tag{3.3.10}$$

This theorem states a pretty trivial notion about information in Markovian system: the information content of my $n+1$ character knowing all the past ones only depends on the information contained in the last k steps.

Proof.

$$H((X_0, X_{-n}^{-1}|X_{-k}^{-1}) = H(X_{-n}^{-k+1}|X_{-k}^{-1}) + H(X_0|X_{-k}^{-1}, X_{-n}^{-k+1})$$

The second term on the right can then be replaced by $H(X_0|x_{-k}^{-1}) = H(X_0|X_{-n}^{-1})$, for $n \geq k$.
The corollary can then be used to conclude that X_0 and X_{-n}^{-k+1} are conditionally independent given X_{-k}^{-1}.
If this is true for every $n \geq k$ then the process must be Markov of order k. □

Now let's move on to define the so called ***Mutual Information***:

Definition 3.3.1. *The mutual information $I(X:Y)$ is the relative entropy between the joint distribution and the measure given by the product of the marginals (the uncorrelated process):*

$$I(X:Y) = \sum_{x,y\in\mathcal{A}} p(x,y)\log\frac{p(x,y)}{\mu(x)\nu(y)} = \tag{3.3.11}$$

$$= d(p(x,y)||\mu(x)\nu(y)) = \tag{3.3.12}$$

$$= \mathbb{E}_p \log\frac{p(x,y)}{\mu(x)\nu(y)} \tag{3.3.13}$$

We can rewrite the definition of mutual information as (see [2] at pg. 21)

$$I(X:Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \tag{3.3.14}$$

Thus, the mutual information $I(X;Y)$ is the reduction in the uncertainty of X due to the knowledge of Y. By simmetry, X says as much about Y as Y says about X.
By the chain rule (3.3.1) we also have

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

Finally, we note that

$$I(X:X) = H(X) - H(X|X) = H(X) \tag{3.3.15}$$

Thus, the mutual information of a random variable with itself is the entropy of the random variable. This is the reason that entropy is sometimes referred to as self-information.
Collecting these results, we have the following theorem:

Theorem 3.3.3 (Mutual information and entropy).

$$I(X:Y) = H(X) - H(X|Y) \tag{3.3.16}$$

$$I(X:Y) = H(Y) - H(Y|X) \tag{3.3.17}$$

$$I(X:Y) = H(X) + H(Y) - H(X,Y) \tag{3.3.18}$$

$$I(X:Y) = I(Y:X) \tag{3.3.19}$$

$$I(X:X) = H(X) \tag{3.3.20}$$

Proof. Left as an exercise. □

The relationship between mutual Information and Entropy is expressed in the Venn diagram in Fig. 3.1 Then from the theorem we have also that:

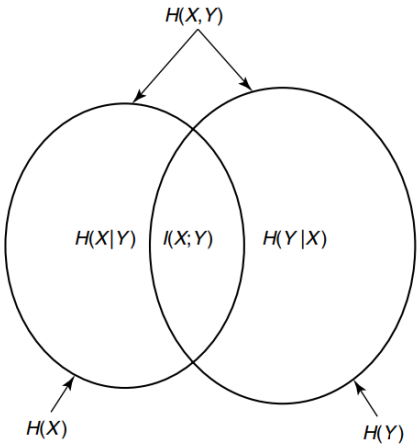


Figure 3.1: Relationship between entropy and mutual information.

Corollary 3.3.3.1 (Non negativity of mutual information). *For any two random variables X and Y,*

$$I(X:Y) \geq 0 \tag{3.3.21}$$

with the equality holding iff X and Y are independent.

Another useful result is

Theorem 3.3.4.

$$h(X) \leq \log |\mathcal{A}|\,, \tag{3.3.22}$$

with the equality holding iff X is uniformly distributed.

Proof. if X is distributed w.r.t. μ and if $p(x) = \frac{1}{|\mathcal{A}|}$ is the uniform distribution,

$$o \leq d(\mu||p) = \sum_{x\in\mathcal{A}} \mu(x) \log \frac{\mu(x)}{p(x)} = \log |\mathcal{A}| - h(X)\,.$$

□

And finally from the positivity condition $0 \leq I(X:Y) = h(X) - h(X|Y)$ we have that

Theorem 3.3.5 (Information can't hurt).

$$h(X|Y) \leq h(X) \tag{3.3.23}$$

with equality holding iff X and Y are independent.

Intuitively, the theorem says that knowing another random variable Y can only reduce the uncertainty in X. Note that this is true only on the average. Specifically, $H(X|Y=y)$ may be greater than or less than or equal to $H(X)$, but on the average $H(X|Y) = \sum_y p(y)H(X|Y=y) \leq H(X)$. For example, in a court case, specific new evidence might increase uncertainty, but on the average evidence decreases uncertainty.
Let's now end this chapter with some last properties of entropy. These come straightforward from the so called *log sum inequality*:

Theorem 3.3.6 (Log sum inequality).

$$\sum_{k=1}^{n} a_k \log \frac{a_k}{b_k} \geq \big(\sum_{k=1}^{n} a_k\big) \log \frac{\sum_{k=1}^{n} a_k}{\sum_{k=1}^{n} b_k} \tag{3.3.24}$$

with equality holding if and only if $\frac{a_k}{b_k}$ *is constant.*

There are two consequences of the log sum inequality:

Theorem 3.3.7.

1) Convexity of relative entropy:
$d(\mu||\nu)$ *is convex in the pair* (μ, ν)*, i.e. for any two pairs* (μ_1, ν_1) *and* (μ_2, ν_2) *and* $\forall 0 \leq \lambda \leq 1$ *we have*

$$d(\lambda\mu_1 + (1-\lambda)\mu_2||\lambda\nu_1 + (1-\lambda)\nu_2) \leq \lambda d(\mu_1||\nu_1) + (1-\lambda)d(\mu_2||\nu_2)$$

2) Concavity of entropy:
$h(\mu)$ *is a concave function of* μ

Proof. Just apply the log sum inequality y to a single term of the left hand side and then sum over $x \in \mathcal{A}$:

$$\lambda\mu_1(x) \log \frac{\lambda\mu_1(x)}{\lambda\nu_1(x)} + (1-\lambda)\mu_2(x) \log \frac{(1-\lambda)\mu_2(x)}{(1-\lambda)\nu_2(x)}$$

Concavity of entropy follows from the convexity of the relative entropy d and from the relation

$$h(\mu) = \log |\mathcal{A}| - d(\mu||p)\,,$$

where $p(x)$ denotes the uniform probability distribution. □

Chapter 4

Ergodic Theorem, AEP Theorem and Entropy Theorem

In this chapter we want to discuss the most important theorems of Information Theory: the *Ergodic theorem*, the *Asymptotic Equipartition Property (AEP) theorem* and the *Shannon-McMillan-Breiman theorem* or *Entropy theorem*.
First we will need to recall some concepts in probability convergence and state some fundamental results.

4.1 Some fundamental results

We say that a sequence of random variables $x_1, x_2, \ldots$, converges to a random variable x:

a) **in probability** if $\mathbb{P}(|x_n - x| > 0) \to 0$ when $n \to \infty$ and for all $\epsilon > 0$

b) **in mean square** if $\mathbb{E}((x_n - x)^2) \to 0$ when $n \to \infty$

c) **almost surely** (also called **with probability 1**) if

$$\mathbb{P}\big(\lim_{n\to\infty} x_n = x\big) = 1 \tag{4.1.1}$$

Of course $c) \Rightarrow b) \Rightarrow a)$.
If $\{P_n\}$ is a sequence of measurable properties (for example: $\frac{\sum_i^n x_i}{n} - \bar{x} <$ smaller or bigger than ϵ) then

1) $P_n(x)$ holds **eventually almost surely**, if for almost every x there is an $N = N(x)$ such that $P_n(x)$ is true for every $n \geq N$

2) $P_n(x)$ holds **infinitely often**, almost surely, if for almost every x there is an increasing sequence $\{n_j\}$ of integers (which may depend on x) such that $P_{n_j}(x)$ is true for $j = 1, 2, \ldots$

Here, $1) \Rightarrow 2)$. We also have three equivalent formulations of almost sure convergence:

Lemma 4.1.1. *The following are equivalent for measurable functions on a probability space:*

1: $f_n \to f$, *almost surely;*

2: $|f_n(x) - f(x)| < \epsilon$, *eventually almost surely, for every* $\epsilon > 0$;

3: Given $\epsilon > 0$, *there is an* N *and a set* G *of measure at least* $1 - \epsilon$ *such that* $|f_n(x) - f(x)| < \epsilon$, $x \in G$, $n \geq N$.

Lemma 4.1.2 (The Markov inequality). *Let* f *be a nonnegative, integrable function on a probability space* (X, Σ, μ).
If

$$\int f \, d\mu \leq \epsilon\delta$$

then

$$f(x) \leq \epsilon$$

except for a set of measure at most δ.

Next, we state a very deep and important result, the Borel-Cantelli principle [4]: such Lemma states that, under certain conditions, an event will have probability of either zero or one. Accordingly, it is the best-known of a class of similar theorems, known as zero-one laws.

Lemma 4.1.3 (The Borel-Cantelli principle). *If* $\{E_n\}$ *is s a sequence of measurable sets (so, a series of events) in a probability space* (X, Σ, μ) *such that*

$$\sum_{n}^{\infty} \mu(E_n) < \infty$$

then

$$x \notin E_n$$

eventually almost surely, or equivalently

$$\mu\big(\limsup_{n\to\infty} E_n\big) = 0$$

Here, "lim sup" denotes limit supremum of the sequence of events, and each event is a set of outcomes. That is, $\limsup E_n$ is the set of outcomes that occur infinitely many times within the infinite sequence of events $\{E_n\}$. Explicitly,

$$\limsup_{n\to\infty} E_n = \cap_{n=1}^{\infty} \cup_{k=n}^{\infty} E_k$$

The set $\limsup E_n$ is sometimes denoted $\{E_n \text{ i.o.}\}$, where "i.o." stands for "infinitely often". The theorem therefore asserts that if the sum of the probabilities of the events E_n is finite, then the set of all outcomes that are "repeated" infinitely many times must occur with probability zero. Note that no assumption of independence is required.
This basically means that such E_n are becoming smaller and smaller with increasing n: thus if for example the sets E_n are "bad" sets in which we do not want our variable x to be, we just need to prove that $\sum_n \mu(E_n) < \infty$ to "neglect them" in the long run.
A related result, sometimes called the **second Borel–Cantelli lemma**, is a partial converse of the first Borel–Cantelli lemma. The lemma states: If the events E_n are independent and the sum of the probabilities of the E_n diverges to infinity, then the probability that infinitely many of them occur is 1. That is:

Lemma 4.1.4 (second Borel–Cantelli lemma). *if* $\sum_{n=1}^{\infty} \mu(E_n) = \infty$ *and the events* $\{E_n\}_{n=1}^{\infty}$ *are* ***pairwise independent****, then*

$$\mu\big(\limsup_{n\to\infty} E_n\big) = 1$$

The infinite monkey theorem, that monkeys endless typing at random will, with probability 1, eventually produce every finite text (such as the works of Shakespeare) [1], amounts to the statement that a (not necessarily fair) coin tossed infinitely often will eventually come up Heads. This is a special case of the second Lemma.
To end this section, we state an important result connecting cardinality and probability:

Lemma 4.1.5 (cardinality bounds). *Let* μ *be a probability measure on the finite set* A*, let* $B \subset A$*, and let* α *be a positive number.*

1) if $a \in B \Rightarrow \mu(a) \geq \alpha$*, then* $|B| \leq 1/\alpha$.

2) For $b \in B$*,* $\mu(b) \geq \alpha/|B|$*, except for a subset of* B *of measure at most* α

Basically, we are saying that given a set in a measurable space, this is **either big in probability or in cardinality**. We cannot have both.

[1] However, the probability that monkeys filling the entire observable universe would type a single complete work, such as Shakespeare's Hamlet, is so tiny that the chance of it occurring during a period of time hundreds of thousands of orders of magnitude longer than the age of the universe is extremely low (but technically not zero). To put it another way, for a one in a trillion chance of success of the monkeys typing Hamlet, there would need to be 10^{360641} observable universes where *each proton* was a monkey typing.

4.2 AEP property and SMB Theorem

In information theory, the analog of the law of large numbers is the asymptotic equipartition property (AEP). It is a direct consequence of the weak law of large numbers.

Theorem 4.2.1 (AEP Theorem).
If $x_1, x_2, x_3, \dots$ are i.i.d. which follow a distribution $p(x)$, then

$$-\frac{1}{n}\log p(x_1, x_2, \dots, x_n) \xrightarrow{\text{in probability}} H(x) \tag{4.2.1}$$

Proof.
Functions of independent random variables are also independent random variables. Thus, since the x_i are i.i.d., so are $\log p(x_i)$. Hence, by the weak law of large numbers,

$$-\frac{1}{n}\log p(x_1, x_2, \dots, x_n) = -\frac{1}{n}\sum_i \log p(x_i) \xrightarrow{\text{in probability}} -\mathbb{E}[\log p(x)] = H(x)$$

which proves the theorem. □

We can also give an alternative (and equivalent) formulation of the AEP Theorem, stated under the assumptions of considering a discrete-time stationary ergodic process μ: the asymptotic equipartition property for such stochastic source is known as the Shannon–McMillan–Breiman theorem (SMB).

Theorem 4.2.2 (SMB Theorem).
*Consider a stationary, ergodic, stochastic source μ with entropy $h = h_\mu > 0$. For simplicity, here we also assume that we are dealing with an i.i.d. process (but the more general case is also true). Then for every $\epsilon > 0$ and all sufficiently large n, one can find a subset $\mathcal{T}_n \subset \mathcal{A}^n$ of **typical sequences** that is **small** in cardinality, but **large** in probability, so that:*

(1) $e^{n(h-\epsilon)} \le |\mathcal{T}_n| \le e^{n(h+\epsilon)}$

(2) $\lim_{n\to\infty} \mu(\mathcal{T}_n) = 1$

(3) For each $\omega \in \mathcal{T}_n$ we have

$$e^{n(h-\epsilon)} \le \mu(\omega) \le e^{n(h+\epsilon)}$$

Let's take a closer look at what this theorem is telling us.
If we consider $|\mathcal{A}| = m$, then $|\mathcal{A}^n| = m^n = e^{n\log m}$. Also, we know from the basic properties of entropy that $0 \le h \le \log n$. If we then look at the ratio $\frac{|\mathcal{T}_n|}{|\mathcal{A}^n|} \sim e^{-n(\log m - h)} \xrightarrow[n\to\infty]{} 0$. This is what we mean by "small" in cardinality: not in a general sense, but small with regard to the dimension of our alphabet.
Also, we are saying that $\mu(\mathcal{T}_n) \to 1$ as $n \to \infty$: i.e., the "density" of $\mathcal{T}_n$ in $\mathcal{A}^n$ is becoming bigger and bigger as we increase n. In the limit of n going to infinity, the measure of our typical set becomes 1: hence, in such limit **all sequences are typical**.
All of this is just to say that under suitable conditions, we will always have in our alphabet a few sequences with are typical with respect to all of those which can be formed: think as always about words in a language versus all the combinations of letters that can be formed.
Also, the last statement is telling us that $\mu(\omega) \sim e^{-nh}$ up to an ϵ, which is equivalent to saying that $\lim_{n\to\infty} \frac{1}{n}\log \mu(\omega_n) = h$ as in the first formulation.

Proof.
Given $\epsilon > 0$, we let $\delta = \delta(\epsilon)$ to be chosen later.
We select finite sequences with good empirical distribution:

$$\mathcal{T}_n \equiv \{\omega \in \mathcal{A}^n : \Big|\frac{f_\omega(a_k)}{n} - \mu_k\Big| \le \delta,\ k = 1, 2, \dots, d\}$$

The second statement follows immediately from the Law of Large Numbers: $\lim_{n\to\infty}\mu(\mathcal{T}_n)=1$. In the i.i.d. case we are discussing here:

$$\begin{aligned}\mu(\omega)&=\prod_{j=1}^{n}\mu(\omega_j)=\prod_{k=1}^{d}\mu_k^{f_\omega(a_k)}=\\&=\exp\big(\sum_{k=1}^{d}f_\omega(a_k)\ln\mu_k\big)=\exp\big(n\sum_{k=1}^{d}\frac{f_\omega(a_k)}{n}\ln\mu_k\big)=\\&=\exp\big(n\sum_{k=1}^{d}\mu_k\ln\mu_k\big)\cdot\exp\big(n\sum_{k=1}^{d}(\frac{f_\omega(a_k)}{n}-\mu_k)\ln\mu_k\big)=\\&=\exp\left(n(-h+\sum_{k=1}^{d}(\frac{f_\omega(a_k)}{n}-\mu_k)\ln\mu_k)\right)\end{aligned}$$

Now we choose δ sufficiently small such that $\big|\sum_{k=1}^{d}(\frac{f_\omega(a_k)}{n}-\mu_k)\ln\mu_k)\big|\le\epsilon$, for $\omega\in\mathcal{A}^n$, and this gives the first statement of the Theorem.
Furthermore,

$$1\ge\mu(\mathcal{T}_n)=\sum_{\omega\in\mathcal{A}^n}\mu(\omega)\ge e^{-n(h+\epsilon)}|\mathcal{T}_n|$$

i.e., $|\mathcal{T}_n|\le e^{n(h+\epsilon)}$.
On the other hand, for sufficiently large n we will have that

$$\frac{1}{2}\le\mu(\mathcal{T}_n)\le e^{-n(h-\frac{\epsilon}{2})}|\mathcal{T}_n|\,,$$

i.e., for sufficiently large n

$$|\mathcal{T}_n|\ge\frac{1}{2}e^{n(h-\frac{\epsilon}{2})}\ge e^{n(h-\epsilon)}$$

and this gives the first statement. □

We now want to prove a a result much stronger than the SMB theorem: the **Entropy Theorem**

4.3 Entropy Theorem and Ergodic Theorem

Theorem 4.3.1 (Entropy Theorem).
Given a stochastic,stationary, ergodic process μ on $\mathcal{A}$,

$$\lim_{n\to\infty}-\frac{1}{n}\log\mu(x_1^n)=h_\mu\qquad\textbf{\textit{almost surely}}\tag{4.3.1}$$

We want to prove this by introducing and using a beautiful technique developed by Ornstein and Weiss, while trying to extend entropy theorem from stochastic processes to random fields. In order to introduce the first fundamental lemmas concerning packings and covering, It is useful to start by proving the Ergodic Theorem.
Recalling that a process μ is ergodic if any shift-invariant subset is either of measure 1 or 0, we state and prove the ergodic theorem in its essential form assuming (without loss of generality) a binary process $\mathcal{A}=\{0,1\}$

Theorem 4.3.2 (Ergodic Theorem).
If μ is a (binary) stationary, ergodic process

$$\lim_{n\to\infty}\frac{1}{n}\sum_{j=1}^{n}x_j\quad\textit{exists }\textbf{\textit{almost surely}}\textit{ on }\mathcal{A}^{\mathbb{N}}$$

In this case, letting $p=\mathbb{E}(x_1)=\mu(x_1=1)$, the theorem asserts that

$$\lim_{n\to\infty}\frac{1}{n}(x_1+x_2+\cdots+x_n)=p\qquad\text{almost surely}$$

Proof. We start by assuming that the Theorem is false: the superior (inferior) limit of the averages is then larger (smaller) than p by a fixed amount on a set of positive measure.
By assumption there exists $\epsilon > 0$ such that if

$$A_\epsilon \equiv \left\{ x \in \mathcal{A}^n : \limsup_{n\to\infty} \frac{x_1 + x_2 + \cdots + x_n}{n} > p + \epsilon \right\}$$

then $\mu(A_\epsilon) > 0$.
It is then easy to see (*exercise - and an important one*) that A_ϵ is *invariant*, hence of measure one:

$$\mu(A_\epsilon) = 1$$

This means that for almost all sequences x, there exists an infinite partition (x-dependent) in disjoint intervals over which the average exceeds $p + \epsilon$.
i.e. given $x = x_1, x_2, \ldots$ and $n \in \mathbb{N}^+$, we denote by $m = m(n) > n$ the first time for which the average over the interval $[n, m]$ exceeds the expected value p by ϵ.
m is also x dependent, and it is also finite for all n and for a set of full measure of x's in $\mathcal{A}^{\mathbb{N}}$.
It follows that for almost all x, we can find a disjoint partition (x dependent):

$$\mathbb{N} = \cup_{k=1}^{\infty} [n_k, m_k], \quad \text{with } m_0 = 0,\ n_k = m_{k-1} + 1$$

and

$$\frac{1}{m_k - n_k + 1} \sum_{j=n_k}^{m_k} x_j > p + \epsilon\,.$$

The important step is now to obtain a control over finite sequences, proving that for sufficiently large N there exists a set $\mathcal{G}_N$, large in probability, on which the average beats $p+\epsilon$. Let's formally state this last passage:

Lemma 4.3.3. *Given $\delta > 0$ there exists a positive integer N and a set $\mathcal{G}_N \subset \mathcal{A}^N$ such that*

1) $\mu(\mathcal{G}_N) > 1 - \delta$

2) For each $\omega_1^n = (\omega_1, \ldots, \omega_n) \in \mathcal{A}^N$ there are disjoint intervals $[n_i, m_i] \subset [1, N]$ such that

a) $\{[n_i, m_i]\}$ is a $\delta-$cover of $[1, N]$, i.e. $\sum_i (m_i - n_i + 1) > (1 - \delta)N$

b) On each element of the coverings, the average exceeds p:

$$\frac{1}{|\omega_{n_i}^{m_i}|} \sum_{j=n_i}^{m_i} \omega_j > p + \epsilon$$

This basic lemma leads to our desired contradiction.
As a matter of fact, if $\omega_1^n \in \mathcal{G}_N$ then

$$\sum_{j=1}^{N} \omega_j \geq \sum_i \sum_{j=n_i}^{m_i} \omega_j \geq (1 - \delta)N(p + \epsilon)$$

hence

$$p = \mathbb{E}\left[\frac{1}{N} \sum_{j=1}^{N} X_j \right] \geq (1 - \delta)(p + \epsilon)\mu(\mathcal{G}_N) \geq (1 - \delta)^2(p + \epsilon)$$

which can not be true for all δ. □

Let's now see for completeness the proof of the Lemma:

proof of Lemma 4.3.3. The first observation is that the lengths $L_n = (m(n) - n + 1)$ of the intervals where the average is larger then $p + \epsilon$ is in fact bounded, except for a set of (arbitrary) small probability.

Namely, there exists an L such that $\mu(L_n > L) < \delta^2/2$.
We now define c_n to be 1 if $L_n > L$ and 0 otherwise, so that, for any N,

$$\mathbb{E}\Big[\frac{1}{N}\sum_{n=1}^{N} c_n\Big] < \frac{\delta^2}{2}$$

The Markov inequality (4.1.2) yields:

$$\mu\Big(\frac{1}{N}\sum_{n=1}^{N} c_n < \frac{\delta^2}{2}\Big) > 1-\delta$$

For $N > L$ we define $\mathcal{G}_N$ to be the event

$$\frac{1}{N-L+1}\sum_{n=1}^{N-L+1} c_n < \frac{\delta}{2}.$$

The Markov inequality gives the first property of the basic Lemma.
For each $(\omega_1, \dots, \omega_N) \in \mathcal{G}_N$ we define the intervals $[n_j, m_j]$ inductively, putting

$$n_1 = \min\{n \geq 1 : L_n \leq L\}, \quad m_1 = m(n_1)$$

and

$$n_{j+1} = \min\{n > m_j : L_n \leq L\}, \quad m_{j+1} = m(n_{j+1}).$$

i.e., we go from a covering to a packing with gaps. The construction stops the first time m_j exceeds $N-L+1$.
The first of the second properties of the Basic Lemma is now true by definition of n_j and m_j.
An integer n in $[1, N]$ fails to be in one of the intervals $[n_j, m_j]$ only if $L_n > L$ or if $n > N-L+1$.
By definition of $\mathcal{G}_N$ there are at most $N\delta/2$ indices n in $[1, N]$ for which $L_n > L$.
Furthermore if we choose $N \geq 2L/\delta$ then $N-L+1, N$ contains at most $N\delta/2$ integers. □

4.4 From coverings to packings

Now we want to discuss four Lemmas : the packing Lemma, the counting Lemma, the doubling Lemma and the strong doubling Lemma; the first two are strictly combinatorial, while the second two are just extensions of the Ergodic Theorem.
The word doubling come from the fact that the applications of these lemmas rely on the success on some other limit theorem, a convergence-in-probability theorem for doubling and an almost-sure theorem for strong doubling.
At the end we will then introduce, discuss and prove the so called strong packing lemma that will be used in the prove of the optimality of the Lempel-Ziv code.

4.4.1 Packing and counting

The packing technique is a method for building "almost" packings of intervals from "almost" covering by subintervals whose left end points already cover most of the interval.
In more mathematical terms:

Definition 4.4.1.
Given an interval of integers $[1, N]$ and a collection $\mathcal{C}$ of distinct subintervals, $\mathcal{C} = \{n_j, m_j\}$ with $1 \leq n_j \leq m_j \leq N$, we say that

*(1) $\mathcal{C}$ is called a **strong $1-\delta$- cover** of $[1, N]$ if the left points cover at least a $1-\delta$ fraction of $[1, N]$:*

$$|\{n_j : [n_j, m_j] \in \mathcal{C}\}| > (1-\delta)N\,.$$

*(2) $\mathcal{C}$ is called a **$1-\delta$- packing** if it is **disjoint** and its union cover at least $(1-\delta)$-fraction of $[1, N]$:*

$$\sum_{[n_j, m_j] \in \mathcal{C}} |(m_j - n_j + 1)| \geq (1-\delta)N\,.$$

*(3) $\mathcal{C}$ is **L− bounded** if $|(m_j - n_j + 1)| \leq L$, for each $[n_j, m_j] \in \mathcal{C}$*

Lemma 4.4.1 (The packing lemma).
if $N \geq L/\delta$, then any $L-$bounded, strong $1-\delta$ cover $\mathcal{C}$ of $[1, N]$ contains a $(1-2\delta)-$packing.

The next counting lemma gives us bounds on the number of sequences that are mostly packed by block drawn from fixed collections whose size are known, which we usually consider to be the collections of typical sequences provided by the AEP theorem, or similar.
First we need few notations: let δ and M be positive numbers: for each $m \geq M$, we have subsets $\mathcal{B}_m \subset \mathcal{A}^m$ and we denote

$$\mathcal{B}_M = \cup_{m \geq M} \mathcal{B}_m$$

Then we define:

Definition 4.4.2. *A sequence $x_1^N \in \mathcal{A}^N$ is said to be $(1-\delta)-$built-up from $\mathcal{B}$ if it can be parsed in variable length blocks as*

$$x_1^N = b_1 b_2 \dots b_k, \quad s.t. \sum_{j: b_j \in \mathcal{B}} |b_j| \geq (1-\delta)N \tag{4.4.1}$$

We denote by G_N the subset of $\mathcal{A}^N$ given by the sequences x_1^N that are $(1-\delta)-$built-up from $\mathcal{B}_M$.

Lemma 4.4.2 (The Counting Lemma).
If $|\mathcal{B}_m| \geq 2^{m(h+\epsilon)}$, $h_b(2/M) \leq \epsilon/2$ and $\delta \log |\mathcal{A}| \leq \epsilon/2$, then

$$|G_N| \leq 2^{N(h+2\epsilon)} .$$

where $h_b(p) = -p \log p - (1-p)\log(1-p)$ denotes the binary entropy function.

Proof.
We first condition on the locations of the blocks that belong to $\mathcal{B}_M$: a **skeleton** $\mathcal{P} = \{[n_j, m_j]\}$ is a disjoint collection of subintervals such that

a) $[n_j, m_j] \subset [1, N]$ and $(m_j - n_j + 1) \geq M$;

b) $\sum_{[n_j, m_j] \in \mathcal{P}} (m_j - n_j + 1) \geq (1-\delta)N$.

We now say that $x_1^N \in G_N$ is compatible with the skeleton $\mathcal{P}$ if the word x_i^j belongs to $\mathcal{B}_M$ whenever the interval $[i, j]$ belongs to $\mathcal{P}$, and we denote by $G_N(\mathcal{P}) \subset G_N$ the set of all the sequences compatible with the skeleton $\mathcal{P}$.
Clearly,

$$G_N \equiv \cup_{\mathcal{P}} G_N(\mathcal{P}) .$$

It is now easy to see that:

$$|G_N(\mathcal{P})| \leq 2^{N(h+\epsilon)} |\mathcal{A}|^{\delta N} .$$

The first term $2^{N(h+\epsilon)}$ come from the fact that any subinterval $[i, j]$ of length m can be filled in at most $2^{m(h+\epsilon)} \geq |\mathcal{B}_M|$ ways, whereas there are at most $|\mathcal{A}|^{\delta N}$ ways to fill the places that do not belong to $\mathcal{P}$.
Each interval in a skeleton has length at least M, so at most $2N/M$ points can be endpoints of its intervals. Thus the number of possible skeletons is upper bounded by

$$\sum_{j \leq 2N/M} \binom{N}{j} \leq 2^{N h_b(2/M)}$$

The set G_N is the union of the sets $G_N(\mathcal{P})$ over all skeletons $\mathcal{P}$, so the cardinality of G_N is upper bounded by the product of the two previous bounds, that is

$$\log |G_N| \leq N(h+\epsilon) + \delta N \log |\mathcal{A}| + N h_b(2/M) .$$

This is bounded by $N(h+2\epsilon)$ if $h_b(2/M) \leq \epsilon/2$ and $\delta \log |\mathcal{A}|$. □

4.4.2 Doubling

We now want to prove that in certain situations, given sequences of blocks $\mathcal{B}_n \subset \mathcal{A}^n$ provided by some convergence in probability limit theorem, eventually almost surely, most indices in x_1^N are in fact starting points of blocks from the $\mathcal{B}_n$'s.

Bibliography

[1] Peter Walters. *An Introduction to Ergodic Theory.* Springer, 1982.

[2] Joy A. Thomas Thomas M. Cover. *Elements of Information Theory, second edition.* Wiley-Interscience, 2006.

[3] Wikipedia. Lebesgue integration. *Wikipedia*, 2023. URL `https://en.wikipedia.org/wiki/Lebesgue_integration`.

[4] Wikipedia. Borel-cantelli lemma. *Wikipedia*, 2023. URL `https://en.wikipedia.org/wiki/Borel-Cantelli_lemma`.