Notes and Comments on S. Mallat's Lectures at Collège de France (2023)

Multiscale Models and Convolutional Neural Networks

J.E Campagne *

Janv. 2023; rév. 22 septembre 2023

*If you have any comments or suggestions, please send them to `jeaneric DOT campagne AT gmail DOT com`

Table des matières

1. Foreword

***Disclaimer**: What follows are my informal notes in French, translated into rough English, taken on the fly and reformatted with few personal comments ("NDJE" or dedicated sections). It is clear that errors may have crept in, and I apologize in advance for them. You can use the email address provided on the cover page to send me any corrections. I wish you a pleasant read.*

Please note that the Collège de France website has been redesigned. You can find all the course videos, seminars, as well as course notes not only for this year but also for previous years [1].

I would like to thank the entire Collège de France team for producing and editing the videos, without which the preparation of these notes would have been less convenient.

Also, note that S. Mallat [2] provides open access to chapters of his book *"A Wavelet Tour of Signal Processing"*, 3rd edition, as well as other materials on his ENS website.

This year, 2023, marks the sixth year of S. Mallat's Data Science chair, with the theme: **Modeling, Information, and Statistical Physics**.

As part of this year, two books are recommended: **"Elements of Information Theory"** by Thomas Cover and Joy Thomas [3], and **"Information, Physics and Computation"** by Marc Mézard and Andrea Montanari [4].

NDJE. I would like to mention that I have set up a GitHub repository for some digital applications illustrating the course. https://colab.research.google.com/github/jecampagne/cours_mallat_cdf. For 2023, the notebooks can be executed directly on Google Colab. The migration of the 2022 notebooks is also planned.

1. https://www.college-de-france.fr/chaire/stephane-mallat-sciences-des-donnees-chaire-statutaire/events
2. https://www.di.ens.fr/~mallat/CoursCollege.html
3. https://ia801400.us.archive.org/30/items/ElementsOfInformationTheory2ndEd/Wiley_-_2006_-_Elements_of_Information_Theory_2nd_Ed.pdf
4. https://cds.cern.ch/record/1166773/files/9780198570837_TOC.pdf

2. Lecture 18 Jan.

2.1 Introduction

This year's course, as Stéphane Mallat tells us, will revolve around three words: the concept of **High-Dimensional Data** models ($x \in \mathbb{R}^d$), the theory of **Information**, which was already at the heart of the 2022 course, and finally, **Statistical Physics**, which will be our focus this year. The guiding question for our journey is as follows: ***how to think about problems and model data in high dimensions***? In fact, thinking in high dimensions is not intuitive, and this is where Statistical Physics will help us.

Regarding "data," we will be interested in image processing, sound, text, physical measurements, chemistry, and more. Concerning "applications," once we have a data model, we can perform parameter inference, denoise data, solve inverse problems, and consider classification/regression problems. These concepts have already been addressed in previous years. The novelty will be the perspective considered, with underlying concepts of Statistical Physics to construct a model and observe its consequences.

2.2 Statistical Physics: Why?

Let's recall that for about a century, Statistical Physics was the only science dealing with high dimensions before the advent of today's understanding of machine learning. It's worth noting that in 1901, Josiah Willard Gibbs (1839-1903) wrote a book [5] that established a strong connection between Statistical Mechanics and Thermodynamics. He generalized the statistical interpretation of the **entropy** of a system. This entropy is at the heart of Shannon's theory developed in the 1940s and studied in 2022.

By definition, a *mole* of something contains 6×10^{23} entities (cf. Avogadro's number), which sets the dimension [6] of the system to d. We are clearly in the realm of high dimension. In order to understand macroscopic phenomena studied in Physics, several concepts have been developed over time, such as **energy**, which can be conserved, the notions of

5. J. W. Gibbs, *Elementary Principles in Statistical Mechanics developed with especial reference to the Rational Foundation of Thermodynamics*, Yale Univ., published in March 1902.

6. NDJE. In the classical sense, the phase space has a dimension of $6d$ in this case.

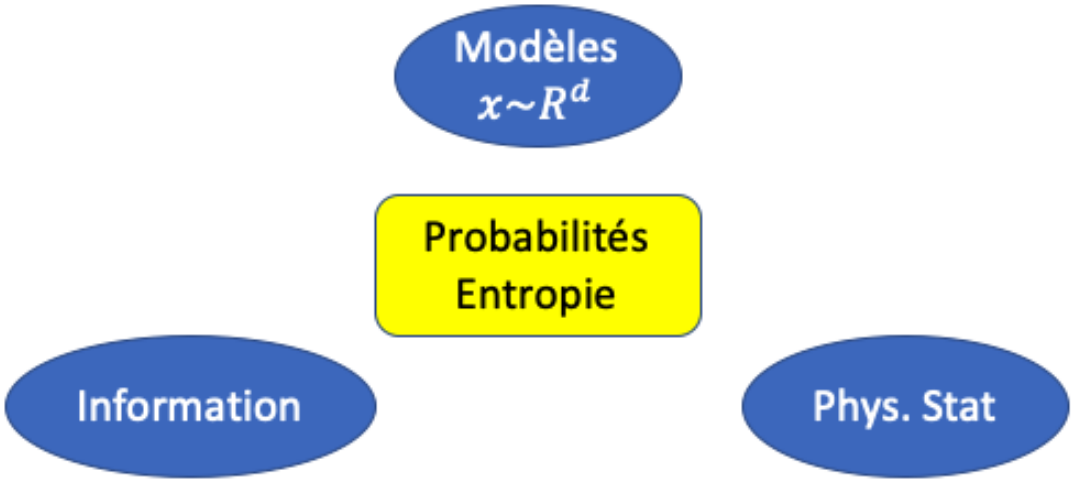


FIGURE 1 – Different concepts covered in the 2023 course.

forces applied to the system, the concept of **equilibrium**, and with Statistical Physics comes the concept of **probability**, and above all, the concept of **entropy**. Furthermore, with the theories underlying the Standard Model of Particle Physics and Cosmology, but also related to Statistical Physics of Phase Transitions, central concepts like **symmetries**, **interactions**, and **scales** emerge.

The idea is to demonstrate that the concepts derived in the field of Statistical Physics will naturally appear in Machine Learning, where *data modeling* also plays a central role. In this context, the **Information Theory** will serve as a mediator, according to Stéphane Mallat's formula. First and foremost, it provides a *mathematical perspective* on the aforementioned concepts, meaning it abstracts these notions from the context in which they originally appeared. Naturally, we encounter **entropy** associated with probabilities, but also **typical sets** with applications in coding, and transmissions (see the 2022 course). Thus, the concepts covered in the 2023 course are schematically represented in Figure 1.

2.3 Statistical Physics Perspective (Brief)

Let's say, in broad strokes, that until the end of the 19th century, fields in Physics such as Solid Mechanics, Continuum Mechanics, Electromagnetism, Optics, Thermodynamics, Chemistry, dealt with macroscopic properties, each with its set of laws and fundamental principles. At the end of the 19th and the beginning of the 20th century, there was a shift in perspective with the advent of Statistical Mechanics (which later became Statistical Physics). It can be said that this began with the assumption that macroscopic properties are, in fact, the result of microscopic-scale collisions between components (e.g., atoms, particles). Even though experimental evidence was lacking at the time, J. Clerk Maxwell (1831-79) developed the theory of velocity distribution in gases, which was generalized in 1896 by Ludwig Boltzmann (1844-1906), who interpreted **entropy**[7] according to the famous formula "S = k log W," which he had engraved on his tombstone. As previously mentioned, J. W. Gibbs established a bridge between Statistical Mechanics and Thermodynamics in 1901.

In doing so, all the aforementioned disciplinary fields were reorganized along a scale axis ranging from the size of elementary particles (1 fermi, 10^{-15}m), the size of atoms (1 angstrom, 10^{-10}m), to astrophysical scales (1 light-year, 10^{16}m), and cosmological scales (Gly), passing through "human" scales. The variables/quantities describing phenomena in Solid Mechanics, Chemistry, and so on are, from a mathematical point of view, quite similar. The problem arises when trying to deduce properties at macroscopic scales from well-established microscopic equations, whether in classical physics (Newton, Maxwell, etc.) or quantum mechanics (Schrödinger, Dirac, etc.).

The main idea of Maxwell, Boltzmann, and Gibbs is that it is probably futile to study the evolution of a particular system with N molecules, especially in the context of Hamiltonian Mechanics. Instead, it is more fruitful to study an ensemble of systems to establish the laws characterizing them, initially at thermodynamic equilibrium. This leads to the introduction of statistical methods in system analysis. The non-obvious point that Gibbs eventually demonstrated is that the quantities and functions developed in Thermodynamics have counterparts in this statistical framework, which, in turn, enriches the description (e.g., the chemical potential introduced by Gibbs). It is remarkable that the results of Quantum Mechanics did not disrupt this theoretical framework, and the statistics

7. The first concept of entropy (1854) was proposed by Rudolf Clausius (1822-88).

of indistinguishable particles of Bose-Einstein (1924-25) and Fermi-Dirac (1926) naturally found their place. Statistical Physics can explain the emergence of phase transitions.

So, Statistical Physics enables us to understand how properties emerge when transitioning from the *microscopic* to the *macroscopic* scale. This evolution is very similar to what we observe in Data Science. Let's also say, in broad terms, that until around 2015, there were distinct fields such as Computer Vision (imaging), Sound (with subfields like speech, music, etc.), Text and Language analysis, and others like Medical Imaging or even the analysis of physical data, etc. There was the underlying theory of Signal Processing, but the data types and algorithms were different, and, therefore, the research communities were largely separated. However, as we now know, deep neural networks have changed all of these disciplines, to the chagrin of various experts as their knowledge has been challenged or sidelined by Machine Learning users. Certainly, this was disheartening for some, but it must be recognized that over the past decade, which can be described as the "pioneers' era," a structure of ***cascade networks of convolution-pooling-non-linearity*** has emerged, as schematically shown in Figure 2, which functions ***universally across all of these fields***. Note that at the input of the neural network, convolutional filters scan *small scales* of the image [8], and as we go deeper into the network, the filters cover increasingly larger areas to ultimately provide global information, e.g., dog/cat categories, the next word in a text, etc. So, we are transitioning from the microscopic to the macroscopic. Similar to Physics, a new axis of organization has emerged, covering all disciplines with the notion of ***multi-scales***, which are very complex due to non-linearities.

Over the past decade, increasingly complex network architectures have emerged with even better successes, and yet there is not much mathematical theory explaining these successes. This is the major difference from Statistical Physics, which, developed over a century, provides a solid theoretical foundation. While it is not surprising that the concepts it has developed can be useful in Machine Learning, it will not be a matter of creating "Applied Statistical Physics." Indeed, ***deep neural networks are capable of solving much more sophisticated problems than those studied in Statistical Physics, but the latter can serve as a conceptual guide, as we will see this year, with the goal of understanding and establishing a theory of neural networks.***

8. We illustrate this with images for illustration purposes, but the schema can be extended to other data types.

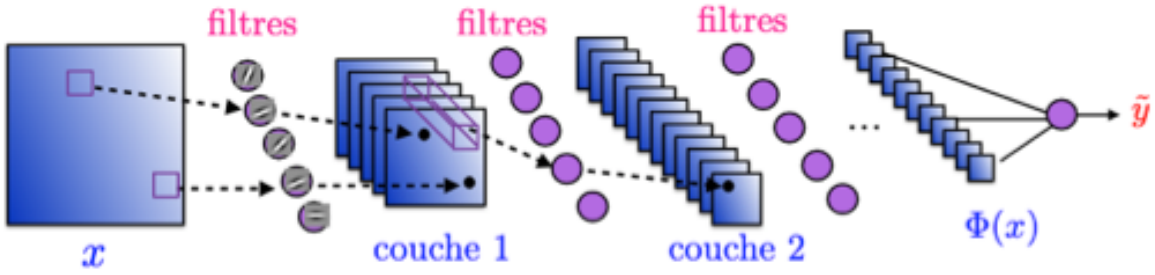


FIGURE 2 – Schematic of a typical deep convolutional neural network.

2.4 Concepts in Statistical Physics

How have the concepts of Statistical Physics emerged over time? Of course, it's impossible to be exhaustive in this historical quest, and in fact, completeness might cloud the message. So, what follows is a perspective that serves as a guide. With that said, we can start the theoretical process with Thermodynamics and its two Principles [9].

The first Principle states that for a system in equilibrium, during a transformation, we can define a function (***internal energy***) with three properties: the state function depends only on the initial and final states of the transformation, it is extensive, and it is conserved for an isolated system. The ***conservation*** of energy is, in fact, a consequence of the ***invariance*** of the laws of Physics with respect to time. It's worth noting that this *relationship between conservation and invariance* is an extremely important and fundamentally critical pair, serving as the basis for creating models in Particle Physics and Gravitation. This relationship is famously encapsulated in Emmy Noether's (1882-1935) theorem.

The second Principle, which has been reinterpreted multiple times, and whose origin is the ***Carnot Principle*** (established in 1824 by Nicolas Léonard Sadi Carnot, 1796-1832), deals with the notion of ***irreversibility*** by stating "that during a transformation, there is a part of the energy that is irreversibly lost as heat." This notion has been connected to that of ***entropy*** (see Sec. 2.3) by formulating the Principle that: "*every transformation*

9. NDJE. especially at the time of their development, the term *Principle* should be understood as a way to *organize* knowledge.

of a thermodynamic system takes place with an increase in the overall entropy, including the entropy of the system and the external surroundings. This is referred to as entropy creation." It is remarkable that there was a gap of about twenty years between the works of R. Clausius (1854) and those of L. Boltzmann (1872) that established the theoretical framework underlying this concept of irreversibility. From a practical standpoint, in the rest of the course, we will denote entropy as H, in accordance with Claude Shannon's notation [10]. Clausius' Principle can be formalized in the following inequality, where δQ is the heat change, T is temperature, and δH is the entropy change:

$$\delta H_{\text{system}} \geq \frac{\delta Q_{\text{exchanged}}}{T} \tag{1}$$

However, this relationship, although established experimentally by Clausius, did not originate from a mathematical theory per se. In a sense, in Thermodynamics, we would be in a situation similar to the current state with neural networks if L. Boltzmann and J. W. Gibbs hadn't intervened.

L. Boltzmann would establish the concept of entropy, as mentioned earlier, by building upon J. C. Maxwell's work on the *kinetics of gases*, now called the Maxwell-Boltzmann theory. It was Maxwell who introduced probabilities into this domain [11], and Boltzmann provided a demonstration of the uniqueness of the Maxwell distribution for gases *without ensemble organization* (*perfect*) [12]. Even if L. Brillouin is sceptical about the validity of the demonstration although he recognises the significance of Boltzmann's H function [13]. So, the distribution of the position r_i of a molecule i in the gas, depending on its velocity v_i, evolves in time $p(r_i, v_i, t)$ according to a differential equation (Boltzmann Equa-

10. NDJE. for the curious, here is the original text of Boltzmann's lectures, translated into French and prefaced by Léon Brillouin: https://gallica.bnf.fr/ark:/12148/bpt6k98134700/f1.item.texteImage; Boltzmann uses the letter H, and it represents the negative of entropy; H in thermodynamics is the notation for enthalpy, and in mechanics, H can also be the notation for the system's Hamiltonian, so be cautious of the context.

11. See Robert E. Robson et al 2017 Eur. J. Phys. 38 065103: *Great moments in kinetic theory: 150 years of Maxwell's (other) equations.* https://iopscience.iop.org/article/10.1088/1361-6404/aa87d4/pdf

12. *molar-ungeordnet*

13. Here is an extract from the afterword of the original text mentioned in footnote Page 12): *"... Apart from the fact that I am concerned about the legitimacy of the division adopted for all space, which is infinite here, I see no reason to propose a law of probability for the speed of an isolated molecule. Let us move on, and adopt Boltzmann's proposal: in this order of ideas, Maxwell's velocity distribution would be the most probable, because it contains the greatest possible number of permutations..."*

tion) that reaches a stationary equilibrium, and the function [14] $H = -\sum_i \int_v p_i \log p_i dv_i$ will depend on the number of possible system configurations (W) and yields the following proportional relationship [15]:

$$H = k_B \log W \tag{2}$$

(NDJE. k_B is a modern notation for Boltzmann's constant). In essence, Boltzmann tells us that the system will evolve to a state where all configurations (*microstates*) are equivalent, which is established by the ***Fundamental Principle of Equiprobability***. These sets of configurations (*micro-canonical*) correspond, as we will see later, to a certain type of data models where probability is completely uniform, and as the system is isolated, its ***energy is fixed*** (a constant). Irreversibility is understood as the evolution of entropy that only increases over time.

The problem with the equiprobability postulate is that in general, the system in question is rarely completely isolated. Therefore, the "universe" consists of the system plus its environment. This is why *Gibbs studies systems in the presence of reservoirs*, which leads him to revisit the concept of entropy. He deepens the relationship between probability and entropy, where $H = -\sum_c p_c \log p_c$ (here with discrete states of configurations), and the concept of ***micro-canonical and macro-canonical ensembles***. We've seen that for a micro-canonical system, the energy is constant, while for a macro-canonical system, the energy fluctuates, and only the average is constant. The reservoir (much larger) serves to fix parameters like temperature. With Gibbs' developments, notions such as the ***Legendre transformation***, ***dual variables*** (temperature-energy, pressure-volume, etc.) emerge. Gibbs' tool is the ***partition function*** Z, from which he retrieves the laws of Thermodynamics.

With the concepts of micro/macro-canonical ensembles in learning, two different types of data models correspond (we will revisit this) as long as the system's dimension d is finite (i.e., a finite number of atoms/molecules/entities), but they become equivalent in the thermodynamic limit, i.e., when $d \to \infty$. So, while numerically the two modeling approaches yield equivalent results, the algorithms differ, as in the example of data generation.

14. NDJE. here is the "modern" sign of entropy.

15. The original expression by Boltzmann considers his function $-H$ from his theorem H. It was Max Planck who wrote the formula $S = k_B \log W$ in a 1923 lecture on the theory of thermal radiation (*Wärmestrahlung*).

2.5 A Brief Historical Detour

NDJE. What follows is, of course, a partial version, and it's impossible to be exhaustive in this course. I've added some personal remarks to the presented timeline by S. Mallat, hoping the reader won't take offense.

It's worth noting that the viewpoints of Boltzmann and Gibbs at the end of the 19th century were based on the ***existence of elementary particles***, without any irrefutable experimental proof to support this hypothesis. Furthermore, other equally famous physicists advocated for continuous matter, such as Marcelin Berthelot (1827-1907) [16]. However, as we know, the evidence would emerge at the turn of the 20th century. We can briefly mention the following sequence of events. Wilhelm Röntgen discovered ***X-rays*** in 1895 (Nobel Prize in Physics in 1901), then Max von Laue (1879-1960) discovered in 1912 that X-rays were diffracted by ***crystals*** (Nobel Prize in Physics in 1914), giving the first glimpse of the "atomic" nature of matter. Finally, in 1909, Hans Geiger (1882-1945), Ernest Marsden (1889-1970), and Ernest Rutherford (1871-1937) (Nobel Prize in Chemistry in 1908) experimentally established ***the planetary model of the atom*** with a central positive charge, refuting J.J. Thomson's (1856-1940) model of a diffuse positive charge. This provided clear confirmation of the corpuscular nature of matter at various scales: the atom and the nucleus.

In doing so, physicist John William Strutt Rayleigh [17] (1842-1919), associated with the mathematician physicist James Jeans (1877-1946), used Statistical Mechanics in 1900 to establish the law, known as the ***Rayleigh-Jeans law***, which expresses the distribution of energy radiated by ***the black body*** as a function of wavelength, valid for long wavelengths. Max Planck [18] (1858-1947) would complement it in the same year using ***the hypothesis of quanta*** (Nobel Prize in Physics in 1918).

On the side of Statistical Physics at the time of Gibbs, there remained the mys-

16. This had unfortunate consequences for the development of French Chemistry; see the article on https://www.universalis.fr/encyclopedie/theorie-atomique/2-la-resistance-de-berthelot/ in the Universalis Encyclopedia.

17. He received the Nobel Prize in Physics in 1904 for studies on gases leading to the discovery of Argon, work conducted jointly with William Ramsay (1852-1916), who received the Nobel Prize in Chemistry in the same year for the discovery of Argon.

18. Initially, Planck favored a continuous matter, rejecting statistical theory, but he realized the evidence and adopted the atomistic viewpoint following experimental results.

tery of describing ***changes of states*** or ***phase transitions*** (solid/liquid/gas transitions, ferro/paramagnetism changes, quantum condensation, etc.) when crossing temperature thresholds within this theoretical framework. The postulates of the theory might not have been right... Note, by the way, that whether it's the partition function Z or Gibbs' free energy, they are both analytical functions of temperature. So why should there be singularities? It took the work of Hendrick Kramers (1894-1952) and Gregory Wannier (1911-83), and finally Lars Onsager (1903-76) and Bruria Kaufman (1918-2010) to show that phase transitions manifest within Gibbs' Statistical Mechanics framework, which was a decisive turning point. The problem Onsager exactly solved in 1944 is the now famous 2D Ising model: this model of interacting spins was introduced by Wilhelm Lenz (1888-1957) in 1920, and his student Ernest Ising (1900-98) had solved it in 1D only and couldn't find a phase transition. Onsager's exact solution clarified its meaning and initiated the study of critical exponents and the development of the ***Renormalization Group Equation*** (RGE) in Statistical Mechanics. It's worth noting that mathematician Hugo Duminil-Copin (1985-) was awarded the Fields Medal in 2022 for his work, among others, on the 3D and 4D Ising model [19], and Giorgio Parisi [20] (1948-) received the Nobel Prize in Physics in 2021 for his work on disordered systems and the study of replica symmetry breaking. So, research in Statistical Physics is still very active.

What's particularly interesting from a historical perspective is that Statistical Physics and Particle Physics have exchanged many concepts. The "Higgs" mechanism of 1964 [21] is an example where Higgs (Particle Physics) uses the spontaneous breaking of the scalar boson of Brout-Englert (Statistical Physics) to give mass to the W and Z bosons, mediators of the weak interaction. The RGE was initiated in Field Theory in Particle Physics in 1954 by Murray Gell-Mann (1929-2019) and Francis E. Low (1921-2007) in the framework of Quantum Electrodynamics (QED), then it was generalized by Curtis Callan and Kurt Symanzik (1923-83) with the establishment of the Callan–Symanzik equations in 1970.

19. See https://www.insmi.cnrs.fr/fr/cnrsinfo/les-travaux-dhugo-duminil-copin

20. He is also known in Particle Physics for establishing, in 1977 with Guido Altarelli, the equations named after them for the evolution of parton densities with energy scale in QCD, subsequently referred to as DGLAP equations, recognizing the earlier work of Dokshitzer, Gribov, and Lipatov.

21. More precisely, the Brout–Englert–Higgs–Guralnik–Hagen–Kibble mechanism, named after Peter Ware Higgs (1929-), Robert Brout (1928-2011), François, Baron Englert (1932-), Gerald Guralnik (1936-2014), Carl Richard Hagen (1937-), and Thomas Walter Bannerman Kibble (1932-2016). Higgs and Englert received the Nobel Prize in Physics in 2013 after the discovery of the Higgs boson at CERN by the teams of the Atlas and CMS experiments.

Developments on phase transitions date back to Leo Philip Kadanoff (1937-) and Kenneth G. Wilson's (1936-2013) Ph.D. thesis, obtained under the supervision of Gell-Mann in 1961. Wilson bridged the gap with developments in Field Theory and developed the theory of critical exponents in relation to phase transitions, which became a prominent theme in the field in the 1970s, like the famous "Les Houches Session XXVIII (1975): Methods in Field Theory" with exceptional contributions.

The key point with the work of Kadanoff-Wilson is that ***probability laws become scale-invariant during phase changes***[22] (e.g., liquid/gas, ferro/para). This point will be of particular interest to us in understanding neural networks.

Now, despite all these very impressive developments, the understanding of a phenomenon like turbulence, which at first glance seems simple, remains in its infancy, despite Kolmogorov's theory. On the other hand, with neural networks, we are tackling much more complex problems, such as face generation. So, it seems that Statistical Physics has remained with much simpler phenomena, but when one takes a step, we understand the "why" and "how," while in Machine Learning, it feels like we're skipping steps.

2.6 The Classical View of Modeling

The subject of data modeling is a distinct field within Probability and Statistics, and as such, many models have been proposed. The first idea that comes to mind when approaching an image or a sound (a signal x in general) and observing that there is clearly ***structure*** is that the signal x is not an arbitrary point in $\mathbb{R}^d$ (d being the number of variables, e.g., the number of pixels). Thus, the number of degrees of freedom m is ***a priori*** much smaller than d (i.e., $m \ll d$), and we will model ***a signal as an element of a surface/manifold*** $S \subset \mathbb{R}^d$ (Fig. 3). These models are of a ***deterministic type***, where the geometry can be illustrated by giving a Riemannian structure to the manifold and studying tangent planes, thus defining a locally parameterized map with variables s, such that the signal x is represented as $x = g(s)$. What needs to be stated clearly is that ***this viewpoint simply does not work in high dimensions***. Firstly, even if $m \ll d$, it still holds that m is large[23], for example, if we compress an image with 10^6 pixels, we achieve a

22. NDJE. the other essential points are the notions of order and symmetry.
23. The "large" dimension can start as soon as $m = 10^3$.

10-fold reduction without significantly degrading the image, but this is only a minimal reduction in dimensionality. However, if m is the dimension of the manifold, based on an argument developed in the 2018-19 courses, in order to estimate the geometry of the manifold, we need to sample it, and in the simple case of regular sampling of $[0,1]^m$ with a spacing of $1/10$ between points, we would need 10^m points. So, with $m = 10^5$, we far exceed the estimated number of atoms in the Universe, which is around 10^{82}, in other words, we are faced with ***the curse of high dimension***. Therefore, we cannot precisely estimate the geometry of the manifold, which means that in this context, ***we cannot use all the tools of differential geometry***.

Another viewpoint is to take a ***probabilistic approach***, where x becomes the realization of a random vector that will have a certain distribution $d\mu(x)$, which we will assume to be smooth with respect to the Lebesgue measure. Thus, it will have a ***probability density*** $p(x)$ such that $p(x)dx = d\mu(x)$, and the challenge is to find $p(x)$ that describes the data. With ***Information Theory***[24], this type of modeling is much richer in high dimensions, and in a certain sense, we can think of the manifold S as a form of ***fuzzy geometry***. *Concentration theorems* provide access to regions where the data have a probability of 1 to be found. By doing so, we have much more flexible mathematical tools at our disposal.

If we briefly return to Statistical Physics and wonder why there was so much resistance to the ideas of Boltzmann and Gibbs, it's necessary to realize that it might have seemed shocking to those immersed in Hamiltonian formalism that ***deterministic laws were the result of stochastic processes described by the Statistical Mechanics formalism*** (even without delving into the quantum realm). Of course, the result of the law of large numbers for particles provides the solution to recover laws interpreted as deterministic. However, such an approach will work on the condition that non-uniform probability laws $p(x)$ will become uniform when the dimensionality d of the problem tends to infinity. ***Thus, by analogy with the thermodynamic limit, we will find phenomena of concentration around manifolds of uniform probabilities***.

24. NDJE. The reader can refer to the 2022 course for more details.

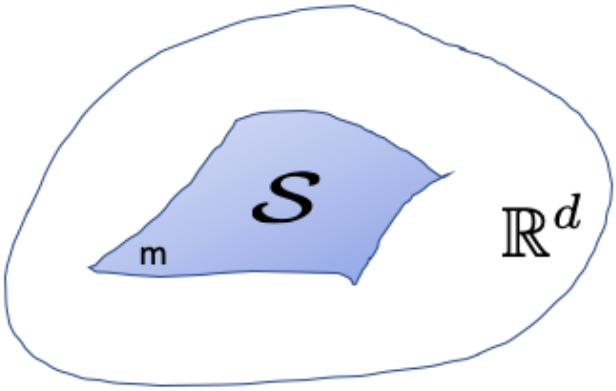


FIGURE 3 – An image, a sound, etc., having structure, can possibly be an element of a surface in $\mathbb{R}^d$, with a much smaller number of degrees of freedom.

2.7 Example of Gaussians

2.7.1 The Case of Variables with Equal Variance

In this example of a Gaussian distribution, we will provide an overview of the issue that we will delve deeper into in the non-Gaussian case. That being said, the Gaussian case is not just a textbook example; it is very important because of the central limit theorem. So, let's assume that x has d variables $x = \{x_i\}_{i \leq d}$, and that the variables x_i are ***independently and identically distributed*** (*iid* hereafter) with a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$. Thus,

$$p(x) = \prod_{i \leq d} p(x_i) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{1}{2}\sum_i \left(\frac{x_i - \mu}{\sigma}\right)^2\right) \tag{3}$$

The probability is maximized when all components x_i are equal to μ. Now, where does a realization of x lie? The result is not intuitive, but x will be found on a quasi-sphere defined by

$$\|x - \mu\|^2 = d\sigma^2 \tag{4}$$

"Quasi" because the thickness ε of the sphere is very small, as we will see. By the way, the argument of the exponential is called the *energy*[25]:

$$U(x) = \frac{1}{2}\sum_i \left(\frac{x_i - \mu}{\sigma}\right)^2 \tag{5}$$

It is always positive and equals 0 for the most probable state. The proportionality constant is denoted as Z^{-1} so that

$$p(x) = Z^{-1}e^{-U(x)} \tag{6}$$

Considering the following sum

$$S_d(x) = \frac{1}{d}\sum_{i=1}^{d} \left(\frac{x_i - \mu}{\sigma}\right)^2 \tag{7}$$

since we are dealing with a sum of independent variables[26],

$$\mathbb{E}_{x\sim p(x)}[S_d] = \mathbb{E}_{z\sim\mathcal{N}(\mu,\sigma^2)}\left[\left(\frac{z-\mu}{\sigma}\right)^2\right] = 1 \tag{8}$$

while

$$\mathbb{V}[S_d] = \frac{1}{d^2}(d \times 2) = \frac{2}{d} \tag{9}$$

so the variance tends to 0 for large dimensions. ***We indeed have a concentration phenomenon, which means that the probabilistic model is very close to a deterministic model because the surface is quasi-zero in thickness, but not negligible, and this is precisely an example of a microcanonical ensemble in Statistical Physics***.

2.7.2 The General Gaussian Case with Covariance

In this model, the probability[27] $p(x)$ is defined according to the following expression:

$$p(x) = \mathcal{N}(\mu, C) = Z^{-1}\exp\left(-\frac{1}{2}(x-\mu)^T C^{-1}(x-\mu)\right), \qquad Z = (2\pi)^{d/2}|\det(C)|^{1/2} \tag{10}$$

25. NDJE. From a notation perspective, $E(x)$ could have been confused with the expected value that we will use extensively, but on the other hand, U is often used in physics for internal energy.

26. $((z-\mu)/\sigma)^2$ with $z \sim \mathcal{N}(\mu, \sigma^2)$ follows a χ^2 distribution with 1 degree of freedom, with a mean of 1 and variance of 2.

27. The term "density" is usually omitted...

where C is the covariance matrix of the variables $\{x_i\}_{i\leq d}$. C is diagonalizable with positive eigenvalues $\text{diag}(\sigma_1^2, \ldots, \sigma_d^2)$, and it introduces "normalized" variables $\{z_i\}_{i\leq d}$ that are the principal axes of a d-dimensional ellipsoid. Thus, the transition from x to z allows us to rewrite $p(x)$ as follows:

$$p(z) = Z^{-1} \exp\left(-\frac{1}{2}\sum_i \left(\frac{z_i}{\sigma_i}\right)^2\right), \qquad Z = (2\pi)^{d/2} \prod_i \sigma_i \tag{11}$$

It is not very difficult to generalize the previous case (cf. the d variables $(z_i/\sigma_i)^2$ with $z_i \sim \mathcal{N}(0, \sigma_i^2)$ follow a χ^2 distribution with 1 degree of freedom), and we realize that the data will concentrate on ***ellipsoidal shells*** centered on μ and whose extensions along the principal axes are given by the values of σ_i and whose thickness will decrease as $1/d$.

Now, this covariance model allows ***selecting the most relevant components*** by taking only those with a value of σ_i above a certain threshold ε (cf. principal component analysis, PCA). Thus, we can perform ***dimensionality reduction***, and similarly, we will find that ***the probability of being on these ellipsoidal shells will be uniform***.

The previous observations of concentration and uniformity of probability on "typical" sets are at the heart of ***Information Theory***, and the reason why ***entropy*** plays such a fundamental role in understanding these ***high-dimensional phenomena***.

2.8 The Perspective of Information Theory in Non-Gaussian Cases

If we generalize to a non-Gaussian case, while preserving the fundamental assumption of *iid*, we have a convergence in probability (law of large numbers) such that

$$-\frac{1}{d}\log p(x_1, x_2, \ldots, x_d) = -\frac{1}{d}\sum_i \log p(x_i) \xrightarrow[d\to\infty]{} \mathbb{E}_{x\sim p(x)}[-\log p(x)] = \bar{\mathbb{H}} \tag{12}$$

where we denote[28] $\bar{\mathbb{H}}$ as the entropy density associated with the distribution $p(x)$ (here, for one of the d components, meaning $d \times \bar{\mathbb{H}}$ is the total entropy associated with the system of d *iid* components, with d playing a role similar to the number N of particles in

28. NDJE. I'm trying to maintain consistent notation with the 2022 course, but there may be variations.

a thermodynamic system; in other words, entropy is an extensive variable). Thus,

$$\forall \varepsilon > 0 \quad \lim_{n\to\infty} \mathbb{P}\left(\left|-\frac{1}{d}\log p(x_1, x_2, \dots, x_d) - \bar{\mathbb{H}}\right| \leq \varepsilon\right) = 1 \tag{13}$$

We come to the idea of the surface S defined by an implicit equation $f(x) = 0$, where we would take

$$f(x) = \left|-\frac{1}{d}\log p(x_1, x_2, \dots, x_d) - \bar{\mathbb{H}}\right| \tag{14}$$

It's not strictly exact that $f(x)$ is zero because there is a thickness ε, but we rediscover the idea of the concentration of $-\log p$ toward entropy, which will lead to a concentration of data on a surface, and we will see that the probability is uniform on it. This is the perspective developed [29] in 1948 by Claude Shannon (1916-2001), which gives rise to the notion of ***typical sets*** that we will encounter this year.

Thus, the following concepts emerge:

- the energy $E(x)$, which is very similar (up to a constant $\log Z$) to $-\log p(x)$;
- the entropy $\mathbb{H}$, which is nothing but the expectation $\mathbb{E}[-\log p(x)]$, which will allow us to understand the complexity of the system (similar to the number of configurations in the style of Boltzmann).

In conclusion, these tools from Statistical Physics and Information Theory allow us to think about high dimensionality. However, ***the major challenge is to construct*** $p(x)$, in other words, to design models that apply to data that are mostly non-Gaussian. In the case of Physics (ideal gas/Ising model), we have models of fairly local interactions between particles/spins, which is absolutely not the case in ***Data Science where we lack a model***. Indeed, what could be the underlying model for images of dogs/cats? That is *the* question, one might say.

2.9 Stationary Ergodic Non-Gaussian Models

In the following, we will focus on ***stationary and ergodic models*** [30]. Ergodicity states that the average value of a quantity calculated statistically (mathematically accessible) is equal to the average of a very large number of measurements taken over time (accessible

29. 2022 course
30. Hypothesis formulated by Boltzmann in 1871

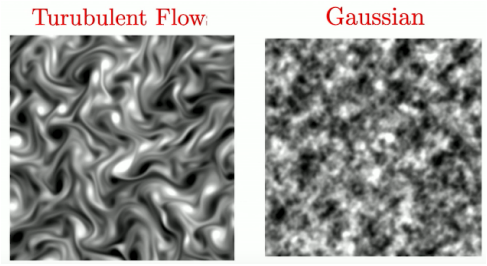


FIGURE 4 – Example of a simulation of 2D turbulent flow (left) and the realization (right) of a Gaussian model that simplifies the problem by considering only the second-order moments of the left distribution, i.e., the covariance matrix.

through experimentation). In the ***Gaussian case***, as we will see, these two assumptions imply that phenomena occurring on two ***distinct scales are independent***. However, in the ***non-Gaussian case***, we observe the emergence of structure (texture, vortex) because phenomena occurring on two ***distinct scales are not independent***. It's this ***notion of dependence*** that we need to understand in order to construct non-Gaussian models.

2.9.1 Turbulence and the Ising Model

To understand non-Gaussianity, S. Mallat provides some images that I'll reproduce here. For example, in Figure 4, we can distinguish the vortices in turbulent fluid on the left. The Gaussian model (on the right), which considers only second-order moments (i.e., the covariance matrix), not only fails to reproduce this structure but is also much more disordered. This disorder reflects higher entropy. So, the challenge in Statistical Physics was to find non-Gaussian models to account for this structure. However, it must be acknowledged that this has not been successful until very recently.

A numerical model that illustrates the properties of the ferro/paramagnetic phase transition is the Ising model. Briefly, the system consists of N *spins* σ with two components $\{-1, +1\}$ placed at the nodes of a lattice (e.g., square lattice). If two spins are neighbors, they experience an interaction with a coupling constant J. To simplify, we define a symmetric interaction matrix J_{ij} with non-zero elements (equal to J, taken as positive here) for only nearest neighbor pairs (i, j). Additionally, the system is placed in an external field H that is assumed to be homogeneous. Thus, each spin is influenced not

only by H but also by the collective action of the other spins in the lattice.

The energy of the system is then given by the expression [31]

$$U(\boldsymbol{\sigma}, J) = -\frac{1}{2} \sum_{(i,j)nn} J_{ij}\sigma_i\sigma_j - h\sum_i \sigma_i \qquad\qquad \boldsymbol{\sigma} \in \{-1, +1\}^N \tag{15}$$

We then define the average magnetization of a configuration "c" of spins as

$$S_c(N, h) = \frac{1}{N}\sum_{i=1}^{N} \sigma_i \tag{16}$$

Note that if $h = 0$, there is a *system symmetry*, meaning that the energy is conserved when all spins are reversed. The system at equilibrium adopts a configuration of minimum energy, with all spins either oriented as $+1$ or all oriented as -1 ($S_c(N, h = 0) = \pm 1$). If h is very strong, then the minimum energy configuration is governed by the second term, and the spin configuration is obtained by orienting all spins in the same direction as the local field. In this case, there is a forced magnetization of the system. The external field sets a direction, and we can choose it as the axis for evaluating spins, so $h > 0$ by convention. Symmetry is broken, and $S_c(N, h \gg 1) = +1$.

Now, let's put the system in contact with a reservoir that sets the temperature T. Intuitively, at high temperature, thermal agitation randomizes the orientation of spins even with $h \neq 0$, and there is no persistent magnetization. There is a competition between temperature (*disorder*) and external field (*order*). But what about the quantity:

$$\lim_{h\to 0}\lim_{N\to\infty} S_c(h, N) = M_0 \tag{17}$$

called *spontaneous magnetization*? That is, ***what happens in the thermodynamic limit when the external field is turned off***[32]? The answer depends on the temperature: if it is below a ***critical temperature*** $T < T_c$, then there is a nonzero spontaneous magnetization (Fig. 5). Furthermore, we observe that there are large regions where spins have the same orientation. At absolute zero temperature, we end up with the configuration where all spins are aligned in the same direction.

31. Note that if H is inhomogeneous, the family of Ising models is known as *Hopfield networks* or *Boltzmann machines* in the neural network community.

32. Note that I leave it to you to guess the answer if we reverse the two limits.

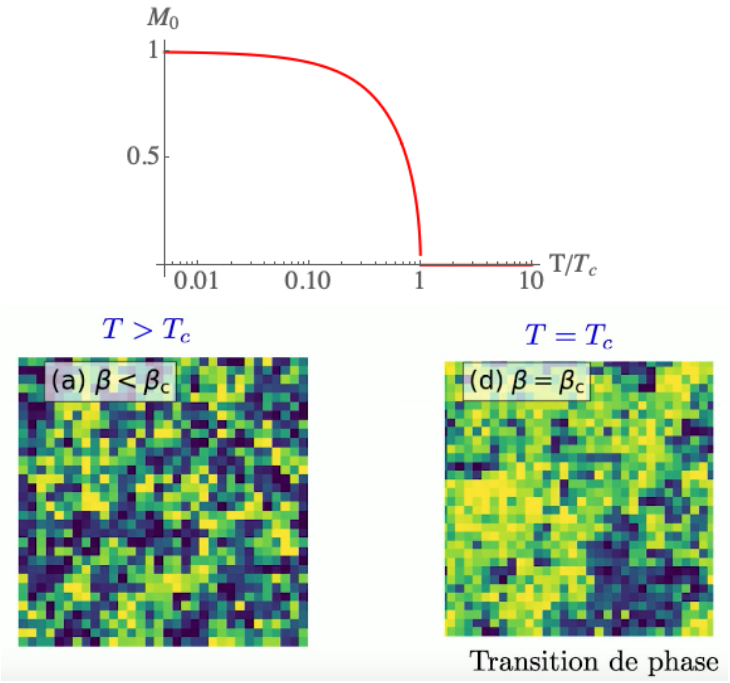


FIGURE 5 – (Top): Spontaneous magnetization M_0 in the case of an Ising model as a function of temperature. Example of a simulated configuration of a spin lattice (Ising here, where spins take continuous values) in contact with a reservoir at a temperature above the critical temperature (left view) or below it (right view). We can appreciate the change in structure that occurs during the phase transition.

This is the entire study of phase transitions that can be studied with this iconic model. In particular, ***how do long-range correlations emerge in the lattice when interactions are only between nearest neighbors*** (local interactions)? Once this problem was understood by Statistical Physics (>1950), attempts were made to apply it to turbulence. Unfortunately, it was a failure again. That is, in trying to build local interactions, we couldn't understand the emergence of the structures in Figure 4.

2.9.2 Sound Frame Generation

NDJE. S. Mallat plays sound frames, so I encourage you to listen to them (again) in the course video around 1:14:00 from the beginning. Also, to understand the concepts

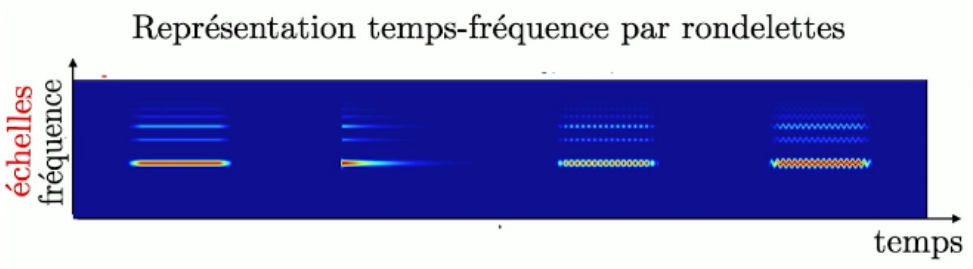


FIGURE 6 – Time-Frequency representation of four music sequences.

of Time-Frequency representations, I recommend referring to the 2020 course on the Fourier Transform (especially windowed) and the Wavelet Transform, whose decomposition algorithm is detailed in the 2021 course.

In Figure 6, we have the Time-Frequency representation of four music sequences: continuous note, "attack," "tremolo," "vibrato." We can distinctly see the differences in these four frames not only in the fundamental frequency but also in the harmonics. Once we understand how such representations are constructed, Figure 7 presents three representations of an "applause": the original sound (left view) exhibits both frequency and temporal structures; in the middle view, generation by *a Gaussian model* that only preserves the original energy at all scales/frequencies (there are *no transient phenomena or correlations between scales* as seen in the original sound); finally, on the right view, generation by a more realistic model that takes into account correlations between scales. While not perfect, this model is much more realistic. We will see how to construct such models and why nonlinearities (e.g., ReLU) are fundamental.

2.9.3 Other Examples

Then, once we understand how to model non-local interactions with correlations between scales, we can, for example, generate images of non-Gaussian fields as illustrated in Figure 8. We can also perform component separation, for example, separating in an image constructed from photons collected by the Planck satellite from the celestial vault, the Gaussian part arising from the cosmic microwave background, and the non-Gaussian part consisting of what is called foregrounds (e.g., emission from intergalactic dust) (Fig. 9). This will be the subject of E. Allys' seminar.

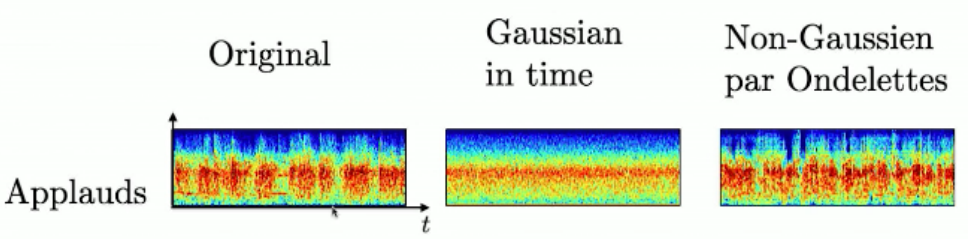


FIGURE 7 – Time-Frequency representations of applause: the original sound on the left; in the middle, generation by a Gaussian model that preserves only the original energy at all scales/frequencies; on the right, generation by a more realistic model that considers correlations between scales.

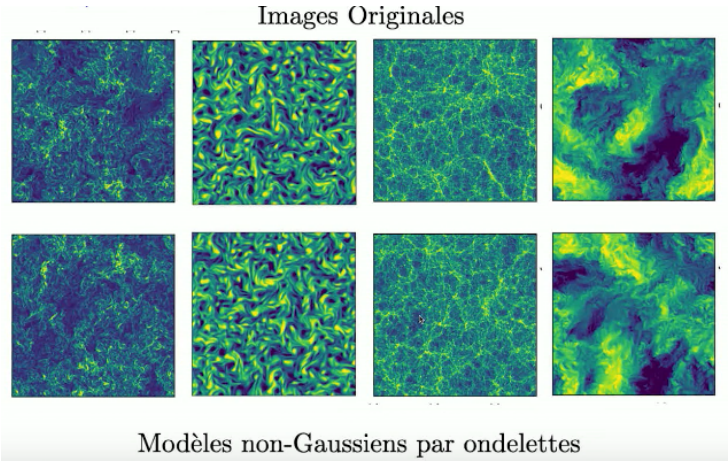


FIGURE 8 – Example of generating non-Gaussian fields using wavelet models: turbulence, cosmic web, etc.

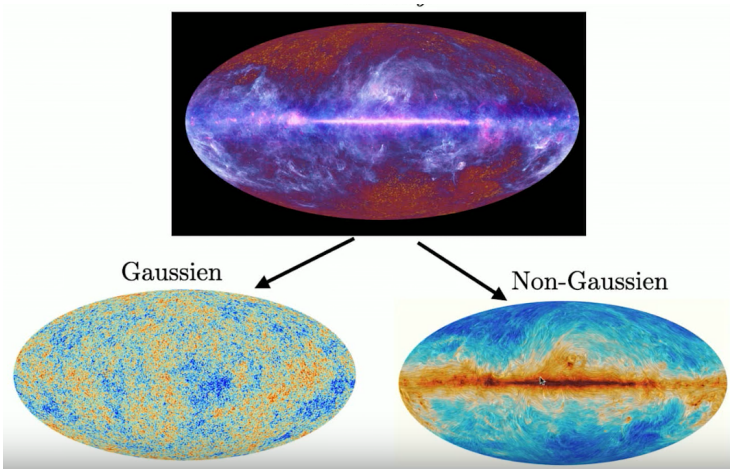


FIGURE 9 – Example of separating Gaussian/non-Gaussian components applied to cosmology.

On the other hand, we won't cover this year, according to S. Mallat, non-stationary and non-ergodic models like generating images of faces or spoken sentences. Neural networks can do this, but the mathematics are more complicated.

2.10 Course Outline [33] for 2023

- We will begin with a ***brief introduction to modeling***, revisiting the curse of dimensionality and introducing Markov fields and maximum entropy models.
- Next, we will review the perspective of ***Statistical Physics***: the *Micro-canonical models*, where we will explore the connection between *energy*, *entropy*, and *temperature*, as well as the concept of *thermodynamic equilibrium*.
- Building on these ideas from Statistical Physics, we will consider the viewpoint of ***Information Theory*** to highlight correspondences. In doing so, we will revisit the concept of entropy in relation to *probabilities* and delve into the notions of *differential entropy* and *conditional/relative entropy* and the "distance" of *Kullback-Leibler*, as well as *entropy density*.

33. NDJE. Keep handy the lecture notes from 2020, 2021, and 2022.

- To understand the notion of equilibrium, we will use ***Markov chains***, which will provide us access to two things: firstly, we will grasp entropy in cases where there is a *form of dependence*, and secondly, how the *evolution towards equilibrium* occurs, including the *Second Law of Thermodynamics*, which we will demonstrate under the Markovian assumption. We will also explore how to sample distributions using ***Monte Carlo algorithms*** with Markov chains (MCMC), which are highly important in practice.
- We will then return to the notion of modeling through ***Maximum Entropy Models***. This will include the concept of *duality* in convex problems (Legendre transformation), *Macro-canonical models*, among others.
- Finally, we will delve into ***non-Gaussian models*** with a focus on ***Harmonic Analysis***. We will explore the *Fourier Transform* (stationary processes that allow us to diagonalize the covariance), the *Wavelet Transform* (multiscale analysis), and all *non-linear interactions*.

3. Lecture 25 Jan.

During this session, we will first and foremost address the question of *how to estimate a model in high dimensions?* We will then explore strategies to overcome *the curse of dimensionality*. Secondly, we will examine the same problem in the context of *Statistical Physics*.

Recall that the challenge is to estimate a probability $p(x)$ where $x \in \mathbb{R}^d$ with $d \gg 1$ (typically $d \sim 10^5 - 10^6$).

To set the framework for the example we are going to study, let's consider $x \in \{-1, +1\}^d$, which can, for example, model the result of a binary questionnaire (Yes/No) for a medical diagnosis, or the values of d spins in an Ising model.

The set $\mathcal{F}$ of probabilities $p(x)$ is such that

$$\mathcal{F} = \left\{ p : \{-1, +1\}^d = \mathcal{E}_d \to [0, 1], \text{such that } \sum_{x_i \in \mathcal{E}_d} p(x_i) = 1 \right\} \tag{18}$$

We can then pose the following problems:

— Firstly, we need to restrict our search to subclasses of $\mathcal{F}$ because without prior information, we will encounter difficulties. Thus, we will need to engage in ***modeling***, which is often parameterized.

— Once we have defined the subclass $\mathcal{C} \subset \mathcal{F}$, we will address the problem of ***estimating*** the model parameters, i.e., the ***learning*** problem.

— We will explore ***inference*** problems, such as calculating *marginal probabilities* (e.g., what is the diagnosis for a particular illness), which requires high-dimensional integration.

— Finally, we may want to ***generate*** new typical data once we know $p(x)$, which is a ***sampling*** problem.

3.1 Frequentist Approach

Let's assume that we have samples, and we can estimate probabilities using histograms. Let $\mathcal{E}_d$ be the set of possibilities for a system with d variables, each of which is binary, i.e., $y \in \{-1, +1\}$, then

$$\mathcal{E}_d = \{y_1, y_2, \ldots, y_K\}, \qquad K = 2^d \tag{19}$$

And we want to estimate $p(x = y_k)$ denoted as θ_k, so we have K parameters. What we know is that

$$\sum_{k=1}^{K} \theta_k = 1 \tag{20}$$

From a geometric point of view, the sought-after K-tuple is located on the "simplex" (Fig. 10). We have at our disposal n data/samples [34] $\{x^i\}_{i \leq n}$.

By hypothesis, the data are *iid* (independently and identically distributed) according to the density $p(x)$ we are looking for; they are examples of a random vector. The estimator $\hat{\theta}_k$ of θ_k is given by [35]

$$\hat{\theta}_k = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{x^i = y_k\}} \tag{21}$$

This is the k-th bin of the histogram, also known as the empirical distribution.

34. NDJE. The superscript i indicates the i-th data point, knowing that it is a vector with d components.
35. $\mathbf{1}_A$ is the indicator function of the set A.

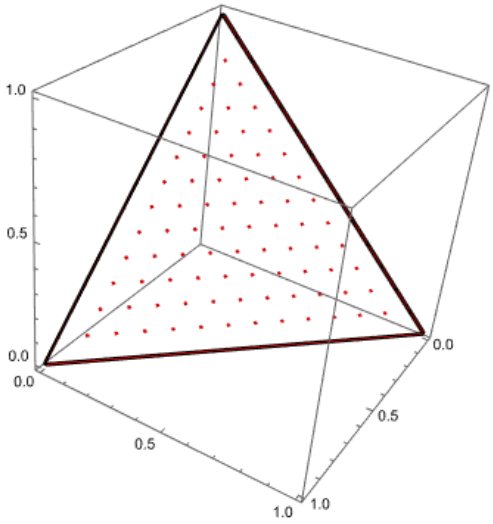


FIGURE 10 – Location of solutions $\sum_k \theta_k = 1$ (here in the case of 3 parameters).

Since the x^i are random variables, we can study the following random variable

$$z_k = \mathbf{1}_{\{x^i = y_k\}} \tag{22}$$

which can take values 0 or 1. In other words, each z_k is a Bernoulli variable. Thus, the expectation of z_k is given by

$$\mathbb{E}[z_k] = \theta_k \tag{23}$$

(indeed, it is the expectation of the "1's"), and its variance is

$$\mathbb{V}[z_k] = \theta_k(1 - \theta_k) \tag{24}$$

Now, regarding the estimator $\hat{\theta}_k$, we can also calculate its mean and variance, yielding:

$$\mathbb{E}[\hat{\theta}_k] = \theta_k \qquad \mathbb{V}[\hat{\theta}_k] = \mathbb{E}[|\hat{\theta}_k - \theta_k|^2] = \frac{\theta_k(1 - \theta_k)}{n} \tag{25}$$

It is unbiased, and the variance decreases with n. However, what interests us is whether we are estimating the probability $p(x)$ correctly. So, if we denote $\Theta = (\theta_k)_{k \leq K}$

(the same for $\hat{\Theta}$), we need to estimate the properties of the normalized vector:

$$\frac{\mathbb{E}[\|\Theta - \hat{\Theta}\|^2]}{\|\Theta\|^2} = \frac{\sum_k \mathbb{E}[|\theta_k - \hat{\theta}_k|^2]}{\|\Theta\|^2} = \frac{\sum_k \theta_k(1-\theta_k)}{n\|\Theta\|^2} = \frac{1 - \|\Theta\|^2}{n\|\Theta\|^2} \tag{26}$$

So, the question now is, if we randomly select a Θ on the simplex (Fig. 10), what is the average error? In high dimensions, the average location of Θ is at the center of the simplex, meaning that for all $k, \theta_k = 1/K$, which gives $\|\Theta\|^2 = K/K^2 = 1/K$. Hence,

$$\frac{\mathbb{E}[\|\Theta - \hat{\Theta}\|^2]}{\|\Theta\|^2} \approx \frac{K}{n} = \frac{\text{number of parameters}}{\text{number of examples}} \tag{27}$$

Now, if we want a good estimate, it means that the number of examples must be large according to the scaling law

$$n \gg 2^d \tag{28}$$

What does this mean when $d = 10^6$! Even with $d = 100$, it is clear that we face the problem of ***the curse of dimensionality***. Conclusion: ***the frequentist approach does not work in high dimensions***.

We need to realize this fact because it is common practice to use this approach by histogramming the data to obtain "experimental probabilities". Therefore, in the case of high dimensionality, we must go against this "intuition". Fortunately, there is another perspective, namely the ***Bayesian approach***[36], which considers probabilities from the perspective of measuring uncertainty or representing partial information.

3.2 Bayesian Approach, Local Markov Model

This is the dominant viewpoint in high dimensions. We will build the model based on the ***known dependencies between variables*** and attempt to represent the *uncertainty about the underlying process* that generated the available data.

That being said, if we have two random variables, the joint probability can be

36. NDJE. see also Section 2.2 of the 2022 course

decomposed as follows, following the idea of Bayes:

$$p(x_1, x_2) = p(x_1)p(x_2|x_1) \tag{29}$$

This generalizes for d variables as follows:

$$\begin{aligned} p(x_1, \ldots, x_d) &= p(x_1)p(x_2|x_1)p(x_3|x_1, x_2)\ldots p(x_d|x_1, \ldots, x_{d-1}) \\ &= p(x_1)\prod_{t=2}^{d} p(x_t|x_1, \ldots, x_{t-1}) \end{aligned} \tag{30}$$

For now, we don't see any gain in making this decomposition. Where we gain is if we can eliminate variables in the set $(x_1, \ldots, x_{t-1})$ when conditioning on x_t. This is the concept of ***conditional independence***, which is fundamental in the process of ***dimensionality reduction***, to break the curse mentioned in the previous section. A notation point:

$$p(x_3|x_1, x_2) = p(x_3|x_2) \Leftrightarrow x_2 \perp x_3|x_1 \tag{31}$$

If we can identify such sources of independence that allow such reductions by removing enough variables, then each conditional probability is a more accessible low-dimensional problem.

Let's examine an "extreme" example of this kind of independence, namely ***Markov chains***. In this case, there is no memory, which translates to

$$p(x_t|x_1, \ldots, x_{t-1}) = p(x_t|x_{t-1}) \tag{32}$$

and thus

$$p(x) = p(x_1)\prod_{t=2}^{d} p(x_t|x_{t-1}) \tag{33}$$

The conditional probabilities (- of transition) are functions of two variables. If, in addition, we are in a stationary regime, these are all identical. Therefore, in this case, the problem is relatively easy, but it is a very specific case.

More generally, ***graphical models*** are models in which conditional dependencies between variables are organized according to a *graph*. This is a form of *topology* given *a priori* on the data. An example is the Ising model, where only interactions between nearest

neighbors are considered. In a somewhat more general sense, however, the (conditional) dependencies will be *local*:

$$p(x_i|x_{j\neq i}) = p(x_i|x_j \text{ for } j \in \mathcal{C}_i) \tag{34}$$

where $\mathcal{C}_i$ are neighborhoods of i. The number of variables that will ultimately intervene in the product of conditional probabilities will directly depend on the number of elements in the neighborhoods. We have the Hammersle-Clifford theorem, which tells us that

$$p(x) = \prod_{i=1}^{d} \psi(x_i, i \in \mathcal{C}_i) \tag{35}$$

This is just another way to write conditional probabilities (Eq. 30) by restricting to the neighborhoods to consider. If $|\mathcal{C}_i| \ll d$, and if there is an invariance in the network, then ***we can overcome the curse of dimensionality***.

However, let's examine the following problem: for example, in an image of the room around you where you might be tempted to take $|\mathcal{C}_i| = 8$ (8 pixels surrounding each pixel), ***how do you account for large-scale structure*** such as the fact that there are very distant pixels that belong to the borders of objects and are therefore correlated? So, this type of modeling works only in certain situations.

On the other hand, in physics/chemistry, we can consider only local interactions at the microscopic level in a first approximation. For example, in a molecule with covalent chemical bonds σ, or ionic bonds in a crystal. But to refine the model, long-range interactions must be taken into account, such as Van der Waals dipole-dipole forces in $1/r^7$, covalent chemical bonds π, or *hydrogen* bonds.

What can help us is *hierarchical modeling* in the intensity of different interactions. In this context, the Markov chain model can serve as the basic model for the most intense local interactions, which can be refined by adding weaker interactions. What we notice in this progression is that the intuition of physics (the notion of *multi-scale*) seems very natural.

3.3 Maximum Entropy Models

If in the previous section ***dimensionality reduction is achieved by considering only local structures*** of the system, we will proceed quite differently by considering ***global measures*** only. This strategy is at the heart of Statistical Physics: we will describe ***the probability distribution in terms of macroscopic quantities***. Formally, we will look at *moments* of the probability distribution:

$$\mu_k = \mathbb{E}_{x\sim p(x)}[\phi_k(x)] = \int_{\mathbb{R}^d} \phi_k(x)p(x)dx \tag{36}$$

Note, by the way, that integration takes place in (very) high dimensions. For example, if $\phi_k(x) = x_k$ (the k-th coordinate of x), $\mu_k = \mathbb{E}[x_k]$. In the *stationary* case, there is invariance $\mu_k = \mu$, and by integrating over all variables, we have a global measure. Another example: $\phi_k(x) = x_i x_{i-k}$ (imagine pixels taken two by two, but all the selected pairs are formed by pixels that are a distance k apart), in the *stationary* case, $\mathbb{E}[\phi_k(x)] = f(k)$, which means that the function f does not depend on i. Of course, you can take much more complex functions. From these moments, how do we obtain $p(x)$?

Of course, in general, it is not enough to have one or two moments to determine a distribution, but we can ask ***what is the smoothest distribution*** that reproduces the given moments? In this context, the question is what is the ***regularity indicator*** that will guide us. The answer is ***entropy*** $\mathbb{H}[p]$, defined as

$$\mathbb{H}[p] = -\int_{\mathbb{R}^d} p(x)\log p(x)dx = \mathbb{E}[-\log p(x)] \tag{37}$$

So, we need to find a distribution $p(x)$ that satisfies constraints both given by the moments $(\mu_k)_{k\leq K}$ and by the quality of maximum regularity/entropy. The parameters of the probability density model p_θ are the $(\mu_k)_{k\leq K}$, so we are looking for

$$p_\theta, \text{ such that } \forall k \mathbb{E}_{p_\theta}[\phi_k] = \mu_k, \text{ and } \max_{p_\theta}(\mathbb{H}[p_\theta]) \tag{38}$$

We will see in more detail in the course that the solution to this convex optimization

problem can be expressed as

$$p_\theta(x) = Z^{-1} \exp\left(- \sum_{k=1}^{K} \theta_k \phi_k(x) \right) \Leftrightarrow - \log p_\theta(x) = \log Z + \sum_{k=1}^{K} \theta_k \phi_k(x) \tag{39}$$

with ***the*** θ_k ***as Lagrange multipliers*** [37] associated with the constraints. Z_θ is the (re)normalization constant [38] so that the integral of p_θ equals 1, which can be written as ($\Theta = (\theta_k)_{k \leq K}$, $\Phi(x) = (\phi_k(x))_{k \leq K}$) [39]

$$Z_\theta = \int_{\mathbb{R}^d} e^{-\Theta^T \Phi(x)} dx \tag{40}$$

From this modeling, we can make some remarks:

— We have two ways to represent the probability density: either on one side the *moments* $(\mu_k)_{k \leq K}$, or on the other side the *parameters* $(\theta_k)_{k \leq K}$;

— We can view the dot product $\Theta^T \Phi(x)$ as an *energy* [40] $U(x)$. Now, by taking the expectation of the right-hand side of Eq. 39, we get

$$\mathbb{H}[p_\theta] = \mathbb{E}_{x \sim p_\theta}[\log Z] + \mathbb{E}_{x \sim p_\theta}[U_\theta(x)] = \log Z + \bar{U}_\theta = -F_\theta + \bar{U}_\theta \tag{41}$$

where we have introduced the ***entropy*** $\mathbb{H}[p_\theta]$, the ***free energy*** F_θ, and the average ***internal energy*** $\bar{U}_\theta$ of the system. This is a relation that we will find in Statistical Physics [41]. This relation can be written as

$$\bar{U}_\theta = \mathbb{H}[p_\theta] + F_\theta \tag{42}$$

In thermodynamics, free energy $F = -\log Z$, one of the tools introduced by Gibbs, depends, for example, on temperature, the number of particles, volume, generalized variables associated with forces, and by differentiation, it allows you to find the average energy, average pressure, average force acting on the system, and so on. Here, we see the statistical view of these concepts.

37. NDJE. See also Sec. 8.3 of the 2018 course.
38. NDJE. In Statistical Physics, this is the Gibbs partition function.
39. NDJE. In machine learning, the vector Φ is often called the *feature vector*.
40. NDJE. From a notation perspective, $E(x)$ could have been confused with expectation, and U is used in thermodynamics for internal energy.
41. NDJE. The equation in thermodynamics involves temperature as a parameter, like $TS = \bar{U} - F$.

3.4 Inference and Applications of Probabilistic Models

3.4.1 Inverse Problems

Once we have a probabilistic model, we can address all classical data analysis problems. The first type of problem concerns data/measurements x obtained (1) from a signal y, transformed for example by a linear operator K assumed for simplicity, and (2) contaminated by noise (η is a random variable). Thus, we have the following relationship between x and y:

$$x = Ky + \eta \tag{43}$$

For example, suppose we take images with a CCD sensor-equipped camera, where each pixel, made up of a silicon photodiode, produces an electrical signal (macroscopic value) proportional to the energy deposited by photons (at the microscopic level) during a clock cycle. The goal then is to recover y from x to obtain, for example, higher-resolution images from local averages. Another example in medical imaging: using X-rays in radiography taken at different angles, we obtain projections of the material-radiation interaction (x), in this case, K is the Radon transform [42], which needs to be inverted to obtain 3D information about the scattering centers in the human body (y). One more example in geophysics: after an explosion in a medium, we collect measurements on the surface (x), and we would like to recover all rock densities in the medium (y) traversed by seismic waves. These are ***inverse problems*** that cover a wide range of disciplines.

Let $\hat{y}$ be the estimator of y, we can say that

$$\hat{y} = \underset{y}{\text{argmax}}\ p(y|x) \tag{44}$$

This is called the ***maximum a posteriori***, also known as MAP, which is the result of a typically Bayesian approach (*a posteriori* because we need to obtain the possibly parameterized probability first). Using Bayes' formula:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} \tag{45}$$

42. Johann Karl August Radon (1887-1956)

where it is customary to call[43] $p(x|y)$ the *likelihood*, $p(y)$ the *prior* on y, and $p(x)$ the *evidence* of the measurements x. Thus, we can write

$$\hat{y} = \underset{y}{\text{argmax}}\ [p(x|y)p(y)] = \underset{y}{\text{argmax}}\ [\log p(x|y) + \log p(y)] \tag{46}$$

So, we need to be able to estimate both $\log p(x|y)$ (*log-likelihood*) and $\log p(y)$.

If we assume that the noise is *white Gaussian* with variance σ^2 ($\eta \sim \mathcal{N}(0, \sigma^2)$), then it is straightforward to conclude that

$$x - Ky \sim \exp\left(-\frac{\|x - Ky\|^2}{2\sigma^2}\right) \Rightarrow \log p(x|y) = -\frac{\|x - Ky\|^2}{2\sigma^2} + \text{const} \tag{47}$$

What about the term $\log p(y)$? It involves constructing a model for the data y. This is where ***modeling the data***, as described in the previous paragraphs, helps in solving the problem (note that optimization is still required to obtain the maximum). With this methodology, we can tackle many physics problems, for example, where the assumption of Gaussian noise in measurements can be a good first approach.

3.4.2 Classification/Regression

We can also address classification/regression problems, where y is the sought-after class given x the signal. The Bayes classifier is again

$$\hat{y} = \underset{y}{\text{argmax}}\ p(y|x) = \underset{y}{\text{argmax}}\ \log p(x|y) = \underset{y}{\text{argmax}}\ [\log p(x|y) + \log p(y)] \tag{48}$$

However, what is "easy" in this case is obtaining $\log p(y)$. Why? In fact, we can use a frequentist approach to *a priori* obtain the probabilities of each class from the available database (e.g., ImageNet). On the other hand, what is difficult is estimating $p(x|y)$, such as what is $p(x)$ (x being the pixels of an image) when we only have images of the "dog" class? Again, this is where ***modeling the probability*** $p(x|y)$ ***helps us solve this problem***, but you need to have the right model!

43. NDJE. Often, we are dealing with measurements x and parameters θ instead of y, and in this context, the *prior* is the information we have *a priori* about these parameters, and $p(\theta|x)$ is the *a posteriori* knowledge about θ once the information contained in the measurements x is taken into account.

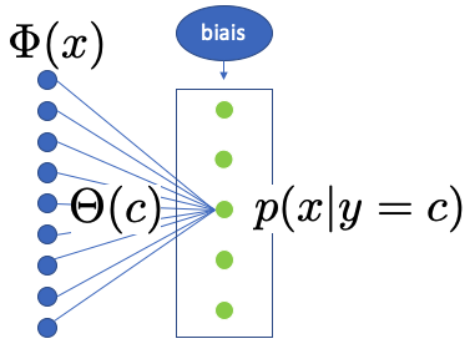


FIGURE 11 – Schematic of linear classification. The representation $\Phi(x)$ is either obtained *a priori* or learned, for example, by a neural network.

What is often used is a ***representation*** $\Phi(x) = \{\phi(x)_k\}_{k\leq K}$ (*feature characteristics*) of the data that is relevant, followed by ***linear classification***[44], also called *logistic regression* (Fig. 11):

$$\log p(x|y) = \sum_{k=1}^{K} \theta_k(y)\phi(x)_k + b = \Theta^T(y)\Phi(x) + b \tag{49}$$

which is identical to the expression in Eq. 39. So, underlying this, if you have moments μ_k such that

$$\mu_k = \mathbb{E}_{x\sim p(x|y)}[\phi_k(x)] \tag{50}$$

then you would have a complete characterization of $p(x|y)$ (the θ_k would be the Lagrange multipliers).

So, using a linear classifier is a way of saying that you have found a data representation that characterizes $p(x|y)$. Typically, $\Phi(x)$ *is either obtained based on a priori arguments* (e.g., symmetries/invariances, kernel methods, etc.) *or through the learning of a neural network*, which may involve convolutional filters, for example.

Thus, the common theme in many problems is the determination of probability distributions in high dimensions. Let's see how this relates to Statistical Physics. This will give us an intuitive foundation for concepts like energy, entropy, etc.

44. NDJE. See Sec. 7.3.3 of the 2019 course, Sec. 5.2 of the 2020 course.

3.5 Statistical Physics: The First Two Principles

We will explore the two principles of Thermodynamics that involve the concepts of energy and entropy.

Regarding ***energy***, it can be considered as a quantity transferred to a system (or internal to it) in the form of heat, work, or interaction. We can consider *kinetic energy* related to motion, various types of *potential energy* (derived from potentials, such as electric, magnetic, gravitational...) which are stored in some way, and also *mass energy* in relativity. The first principle is a conservation principle, which states:

> **First Principle**: *During any transformation of a closed system, the change in its energy is equal to the amount of energy exchanged with the external environment through heat and mechanical work transfer:*
>
> $$\Delta U = \Delta Q + W \tag{51}$$

Consequently, if we consider the ***total energy*** of the system and its surroundings, it remains ***conserved***. Similarly, if ***the system is isolated*** (no exchange), its energy remains constant. For an isolated system, exchanges can occur between different types of energy (e.g., kinetic-potential), which can be expressed as:

$$\Delta E_c + \Delta E_p = \Delta U = 0 \tag{52}$$

The connection to what we discussed in the previous sections is due to Maxwell and Boltzmann in the case of ***studying the kinetics of gases***, a perfect gas of n particles much smaller in size compared to their mean free path. What Boltzmann demonstrates (his theorem H) is that the distribution of velocities over time tends toward a *stationary* distribution given by the Maxwell distribution. What is initially curious (and perhaps shocking) is that ***the laws of collisions considered as elastic***[45] ***are reversible***. However, irrespective of the initial distribution of velocities, we inevitably end up with ***convergence toward the Maxwell distribution***. In other words, ***we have a macroscopic irreversible process produced by underlying microscopic reversible processes***: how is this possible?

45. NDJE. Like billiard ball collisions, as opposed to inelastic collisions like a ball falling into sand.

In fact, the creation of entropy results from the sensitivity of collisions between two particles, which, given their small sizes, can abruptly induce changes in direction (a form of chaos). This is a point that has raised a lot of discussion.

3.5.1 Entropy and Irreversibility

The concept of irreversibility arises from the idea that certain physical processes prevent systems from returning to their initial states. For example, typical cases include the propagation of heat and, more generally, diffusion processes (microscopic scale), as well as convection (macroscopic scale). S. Carnot (1824), followed by R. Clausius (1854) and L. Boltzmann (1872), establish the framework of the second law of thermodynamics:

> **Second Principle**: *Every transformation of a thermodynamic system occurs with an increase in the total entropy, which includes the entropy of the system and the external environment. This is referred to as the creation of entropy.*
>
> $$\Delta\mathbb{H}_{tot.} = \Delta\mathbb{H}_{syst.} + \Delta\mathbb{H}_{ext.} = \mathbb{H}_{created} \geq 0 \tag{53}$$

(NDJE. The system may structure itself during a transformation, and thus $\mathbb{H}_{syst.}$ can be negative, but this is at the expense of the disorder induced in the surroundings. For example, the process of vapor condensation into a liquid releases heat.) The entropy of the system, during a transformation at a constant temperature $T_{ext.}$ set by the external environment, can also be expressed as:

$$\Delta\mathbb{H}_{syst.} = \mathbb{H}_{created} - \Delta\mathbb{H}_{ext.} \geq -\Delta\mathbb{H}_{ext.} \Rightarrow \Delta\mathbb{H}_{syst.} \geq \frac{Q_{ext.\to syst.}}{T_{ext.}} \tag{54}$$

Here, $Q_{ext.\to syst.}$ is the heat exchange transferred to the system from the surroundings. ***Reversible exchanges are characterized by equality, while irreversible exchanges result in a strict inequality*** (Carnot's result). Furthermore, at the ***equilibrium state***, where there is no more creation of entropy ($\mathbb{H}_{created} = 0$), ***entropy is at its maximum***. We won't delve into all the consequences of this principle, whether from an applied perspective (e.g., thermal engines, Carnot cycle, chemical reactions, etc.) or from a philosophical standpoint (e.g., the concept of the arrow of time).

We will see that the ***probability distribution is stationary***, meaning it no longer depends on time, and ***all microstates accessible to the system are equiprobable***. Under these conditions, if we denote Ω as the set of accessible possible configurations for the system, then:

$$P(c \in \Omega) = \frac{1}{|\Omega|}, \qquad \mathbb{H} = k \log |\Omega| \qquad \text{(Boltzmann)} \tag{55}$$

We will revisit these concepts in the upcoming sessions.

4. Lecture 1er Feb.

During this session, we will study probabilistic entropy in both Statistical Physics and Information Theory (See also the 2022 course). First, from Ludwig Boltzmann's perspective, entropy is related to the number of accessible states of a system (at equilibrium). We will then see how pressure, temperature appear as macroscopic quantities defined from entropy variations. Then, Claude Shannon's perspective (1916-2001) will provide us with a much more generic view of entropy, where it becomes a purely mathematical concept, entirely related to probability distributions (independently of the context in which they are employed), and it specifies phenomena of concentration on typical sets. Underlying these typical sets, we find the structure of microcanonical ensembles of Boltzmann. Furthermore, the notion of concentration will provide us with a framework for the intuition that data (e.g., images) aggregates on manifolds. In another session, we will see that if the temporal evolution of distributions is described by a Markov chain, it gives us the second law of thermodynamics, in which entropy inexorably increases (with a condition). Thus, physical properties, elevated to the level of principles in the manner of postulates, somehow emerge from mathematical properties.

4.1 Boltzmann's Perspective: Microcanonical Ensemble

4.1.1 Maximum Entropy, Equiprobability of Microstates

We will consider a system at equilibrium, in which we fix the following state variables as constants: volume V, the number of particles noted as d in this course, and the (internal) energy U. We then consider the concept of a microstate, which consists of all the coordinates and moments that go into the description of the system's Hamiltonian in classical mechanics. In quantum mechanics [46], states have quantized energies, but the philosophy remains unchanged. The set of microstates or configurations is denoted as Ω, and $|\Omega|$ is its cardinality.

NDJE. $|\Omega|$ is the number of configurations that give rise to systems that are certainly different but have energies within the interval $[U, U+\delta U]$. It is also called the thermodynamic weight. To give you an idea, it is given by the expressions

$$|\Omega(U,d,V)| = \frac{1}{\prod_i h^{3N_i} N_i!} \int_{U \leq \mathscr{H} \leq U+\delta U} d\Gamma \qquad (classical) \qquad (56)$$

$$= \sum_{U \leq U_c \leq U+\delta U} 1 \qquad (quantum) \qquad (57)$$

with $d\Gamma$ being the element of the classical phase space of degrees of freedom $\prod_i d^3q_i d^3p_i$ ($\mathscr{H}$ the system's Hamiltonian), and U_c a quantum Hamiltonian eigenvalue.

The second law of Thermodynamics tells us that at equilibrium, the probability of each microstate is a constant with a value of $P = 1/|\Omega|$. This is what we call the ***fundamental principle of equiprobability***. L. Boltzmann then defines entropy as follows (Note that the Boltzmann constant k, sometimes noted as k_B, is equal to $k_\mathrm{B} = 1{,}380\,649 \times 10^{-23}\ \mathrm{J \cdot K^{-1}}$, and it is exact and defines the Kelvin.):

$$\mathbb{H}(U,d,V) = k \log |\Omega(U,d,V)| = -k \log P \qquad (Boltzmann) \qquad (58)$$

How does L. Boltzmann elaborate on this definition? Indeed, the connection with the

46. NDJE. Ludwig Boltzmann had no idea that such a theory would be developed about twenty years after his death.

definition that gives the expression of entropy as follows:

$$\mathbb{H}(U,d,V) = -k\sum_{c\in\Omega} p_c \log p_c \qquad (Gibbs) \tag{59}$$

that is, a formulation more related to probabilities, was made by J. W. Gibbs in 1901. In fact, if we define for any probability distribution (here we consider discrete states to simplify notations), we have

Theorem 1 ***Uniformity of the microcanonical probability distribution, maximum entropy***

$$\sum_{c\in\Omega} p_c = 1, \qquad \mathbb{H}[p] = -\sum_{c\in\Omega} p_c \log p_c \qquad \Longrightarrow \qquad \max_p \mathbb{H}[p] = \log|\Omega| \tag{60}$$

Proof 1.

We use the technique of Lagrange multipliers [47] to which we will return. Thus, we can minimise $-\mathbb{H}[p]$ by studying the saddle point of the function

$$\mathscr{L}(\{p_c\},\lambda) = \sum_{c\in\Omega} p_c \log p_c + \lambda(\sum_{c\in\Omega} p_c - 1) \tag{61}$$

So $\forall i$

$$\frac{\partial\mathscr{L}}{\partial p_i} = \log p_i + 1 + \lambda = 0 \tag{62}$$

so $\log p_i$ is a constant, so is p_i, and using the constraint on the sum of p_i, we conclude that

$$\forall c\in\Omega, \quad p_c = \frac{1}{|\Omega|} \Longrightarrow \mathbb{H}[p] = -\log|\Omega| \tag{63}$$

■

Therefore, the idea of a uniform distribution in the configuration space corresponds well to the principle of maximum entropy.

47. See also Sec. 8.3 of the 2018 Course.

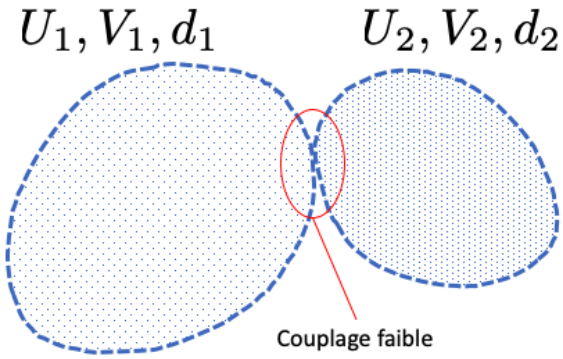


FIGURE 12 – Diagram of two weakly interacting subsystems.

4.1.2 Thermodynamic Equilibrium

If the probability distribution is independent of time, the system is in equilibrium, which in itself can be considered a definition of the equilibrium concept. There is another definition, and thus we have the following equivalent definitions:

> **Definition 1** *(**Equilibrium of a System**)*
> *A system is in equilibrium*
>
> *a) if the probability distribution of microstates is independent of time;*
>
> *b) or if we consider subsystems of a sufficiently large size to associate them with state variables, then they are themselves in equilibrium.*

Let's examine definition (b) where we consider two subsystems of an isolated system, identified by the triplets of state values (U_1, V_1, d_1) and (U_2, V_2, d_2) (Fig. 12). We also assume that the two systems are weakly coupled, meaning that the interactions at the exchange surface are short-ranged and negligible compared to the interactions between particles inside the two systems. Moreover, the statistical fluctuations of the state variables due to exchanges are negligible. The isolation hypothesis indicates that the variables (V, d, U)

$$V = V_1 + V_2 \qquad d = d_1 + d_2 \qquad U = U_1 + U_2 \tag{64}$$

are constants. The number of configurations $|\Omega|$ of the system composed of the two subsystems $i = \{1, 2\}$, under the weak coupling assumption involving independence, is given by the relation

$$|\Omega_{1+2}(V, d, U)| = \prod_i |\Omega(V_i, d_i, U_i)| \tag{65}$$

Thus, the entropy of the overall system $\mathbb{H}_{1+2}$ is nothing but the sum of the entropies of the two subsystems, i.e.,

$$\mathbb{H}_{1+2} = \mathbb{H}_1 + \mathbb{H}_2 \qquad (\textit{entropy additivity}) \tag{66}$$

At equilibrium of the overall system, entropy is constant, so $d\mathbb{H}_{1+2} = 0$.

Let's examine the case where $\{V_i, d_i\}_i = cte$, then

$$d\mathbb{H}_{1+2} = \sum_{i=1,2} \left(\frac{\partial \mathbb{H}_i}{\partial U_i}\right)_{V_i, d_i} dU_i \tag{67}$$

Moreover, $dU = 0$ indicates that $dU_1 = -dU_2$, hence

$$\left(\frac{\partial \mathbb{H}_1}{\partial U_1}\right)_{V_1, d_1} = \left(\frac{\partial \mathbb{H}_2}{\partial U_2}\right)_{V_2, d_2} \tag{68}$$

Now, ***the temperature of a system at equilibrium is related to entropy*** as follows

$$\frac{1}{T} = \left(\frac{\partial \mathbb{H}}{\partial U}\right)_x \tag{69}$$

where x denotes the set of variables other than U (here, V and d). Therefore, at equilibrium, we have

$$T_1 = T_2 \qquad (\textit{equilibrium}) \tag{70}$$

thus ***the temperatures of all sub-states are identical***.

Furthermore, at equilibrium of the total system, with entropy $\mathbb{H}_{1+2}$ being maximal, we have

$$\left(\frac{\partial^2 \mathbb{H}}{\partial U^2}\right)_X \leq 0 \Longrightarrow \left(\frac{\partial T}{\partial U}\right)_X \geq 0 \tag{71}$$

thus at equilibrium, the temperature of a subsystem increases with its internal energy. Thus, $T_1(U_1, X_1 = cte)$ and $T_2(U_2, X_2 = cte) = T_2(U - U_1, , X_2 = cte)$ evolve in opposite directions as U_1 changes: if U_1 increases due to a transfer from $2 \to 1$, then T_1 increases and T_2 decreases, tending to decrease the transfer (if U_1 decreases, we can reason about U_2 increasing). Therefore, ***the transfer of energy tends to homogenize the temperatures*** of the two subsystems, which is also a classical result in thermodynamics.

4.1.3 Free Energy: Variable Volume

Let's relax the assumption that the volumes $(V_i)_i$ are constant while keeping the variables x_i other than energies and volumes constant. However, the global system is isolated and therefore maintains its variables (U, d, V) constant. Consequently, the total differentials of the respective entropies of the two subsystems can be written as

$$\forall i, \quad d\mathbb{H}_i = \left(\frac{\partial \mathbb{H}_i}{\partial U_i}\right)_{V_i, x_i} dU_i + \left(\frac{\partial \mathbb{H}_i}{\partial V_i}\right)_{U_i, x_i} dV_i \tag{72}$$

But $dV_1 = -dV_2$ and $dU_1 = -dU_2$, and at equilibrium $d\mathbb{H}_1 + d\mathbb{H}_2 = 0$, so we can deduce that

$$\left(\frac{\partial \mathbb{H}_1}{\partial U_1}\right)_{V_1, x_1} = \left(\frac{\partial \mathbb{H}_2}{\partial U_2}\right)_{V_2, x_2} \qquad \left(\frac{\partial \mathbb{H}_1}{\partial V_1}\right)_{U_1, x_1} = \left(\frac{\partial \mathbb{H}_2}{\partial V_2}\right)_{U_2, x_2} \tag{73}$$

The first relation reaffirms that ***the temperatures of the two subsystems are equal*** ($T_1 = T_2$). Using the definition of pressure P as follows

$$P = T \left(\frac{\partial \mathbb{H}}{\partial V}\right)_{U, x} \tag{74}$$

with x as the other variables, the second relation results in ***equality of pressures*** at equilibrium

$$P_1 = P_2 \qquad (\textit{equilibrium}) \tag{75}$$

Note that pressure is an *extensive* variable like mass, the number of particles, and energy, while temperature is an *intensive* variable like mass density and concentrations. The

equality of pressures implies that at the interface between the two subsystems, there is no work done by any force.

In general, at equilibrium with P and T fixed, the change in entropy is given by

$$d\mathbb{H} = \frac{1}{T}dU + \frac{P}{T}dV \tag{76}$$

Thus, ***the change in entropy has two contributions*** (in our case), one related to the change in internal energy, and the other related to the change in another quantity, here the volume, and the coefficient of variation defines an associated variable, in this case, pressure, normalized by temperature. This second contribution is nothing but the opposite of external work [48] on the system, W_{ext}. This should be related to Equation 41, where in this case, we would identify the ***free energy*** as the contribution of external work associated with volume change. These two components of entropy variations show a general character:

- one component that in thermodynamics is related to internal energy, which at the microscopic level can be represented as the kinetic agitation of particles, and in probability is represented as $U_\theta(x) = \Theta^T\Phi(x)$;
- and a contribution that appears in thermodynamics as external work, and in probability appears through the (re)normalization constant, i.e., $F_\theta = -\log Z_\theta$.

Hence, a kind of analogy:

Thermo.	Stat.
Internal energy (kinetic)	$U_\theta = \Theta^T\Phi(x)$
Free energy (external work)	$F_\theta = -\log Z_\theta$
$1/T$, P/T, ...	Lagrange multipliers

What must be retained is that ***entropy is the central function, and its maximization describes the system in equilibrium***. The question that will occupy us is to understand why this notion emerges, where does it come from? There is something fundamental underlying it: macroscopic fluctuations of state variables can be neglected and will converge to

48. NDJE. the pressure force is directed outward if P is its internal pressure.

constants, and this is done through the ***law of large numbers that gives the phenomenon of concentration***, but at its core, there is the notion of ***independence*** (or at least weak correlation).

4.2 Information and Coding

4.2.1 The Law of Large Numbers

Let A_d be a random variable that depends on d, the number of observations, such that $A_d \xrightarrow[d\to\infty]{} A$. Then the convergence in probability tells us that

$$\left(\forall \varepsilon > 0,\ \lim_{d\to\infty} \mathbb{P}\left[|A_d - A| \leq \varepsilon\right] = 1\right) \Leftrightarrow \left(A_d \xrightarrow[d\to\infty]{prob.} A\right) \tag{77}$$

In a way, this tells us that it is *rare* for A_d to deviate from its limit value A.

Now, here is the theorem in the case where the random variable, denoted $\bar{X}_d$, is the sample mean of d *i.i.d.* random variables:

Theorem 2 ***(Weak Law of Large Numbers)***
Let $(X_i)_{i\leq d}$ be i.i.d. random variables, and $\mathbb{E}[X_i] = \mu < \infty$. If $\bar{X}_d = \frac{1}{d}\sum_{i=1}^{d} X_i$, then we have a convergence in probability

$$\bar{X}_d \xrightarrow[d\to\infty]{prob.} \mu \tag{78}$$

The proof can be found in Section 3.3 of the 2022 course, based on Bienaymé-Chebyshev's inequality in the case where $\mathbb{E}[X_i^2] < \infty$.

So we have a concentration phenomenon when dealing with *i.i.d.* variables. But why does entropy come into play? For that, we need to refer to Claude Shannon's 1948 paper titled *"A Mathematical Theory of Communication"*[49].

49. C. E. Shannon, The Bell System Technical Journal, Vol. 27, pp. 379–423, 623–656, July, October. https://people.math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf.

4.2.2 Independence and Concentration (Discrete Case)

When considering Information Theory, there are two perspectives, as S. Mallat explains: a probabilistic view, which is likely the most effective, but one can also have a deterministic view of coding by adopting Kolmogorov's viewpoint [50]. Kolmogorov defines the quantity of information through the size of the algorithm, executed on a Turing machine, capable of reproducing a sequence [51]. These two notions are equivalent, primarily due to concentration phenomena. However, performing calculations following Kolmogorov's approach can be very challenging, while Shannon's perspective is well-suited for such calculations.

Let $\{y_k\}_{k\leq K}$ be an alphabet with K symbols denoted as $\mathcal{A}$. Let $(X_i)_i$ be *i.i.d.* random variables taking their values in $\mathcal{A}$, then we have:

$$\forall i,\ \mathbb{P}(X_i = y_k) = p(y_k) \tag{79}$$

where entropy is defined as:

$$\mathbb{H} = -\sum_{y_k \in \Omega} p(y_k) \log p(y_k) \tag{80}$$

Since the probabilities $p(y_k) \in [0, 1]$, and entropy is maximal when all probabilities $p(y_k)$ are equal to $1/|\Omega| = 1/K$, we can write:

$$0 \leq \mathbb{H} \leq \log K \tag{81}$$

Now, consider the case where $p(y_k) = 1$ for $k = k_0$ and 0 if $k \neq k_0$. In this case, the entropy is zero ($\mathbb{H} = 0$). On the other hand, if $\forall k$, $p(y_k) = 1/K$, then $\mathbb{H} = \log K$. It's evident that **entropy is related to the concept of uncertainty**: there is no uncertainty in the first case because all random variables X_i take the value y_{k_0}, while uncertainty is maximal in the second case (Fig. 13).

Let's examine the joint probability of events when drawing the X_i. Since these

50. NDJE. Andrey N. Kolmogorov, starting in 1933 following the work of Emile Borel (1871-1956) and Henri Lebesgue (1875-1941), developed the theory of probability and established a connection between measure and the probability of composite events.

51. For example, π has information in the sense of Kolmogorov because it is defined through a power series that can be encoded and executed on a computer.

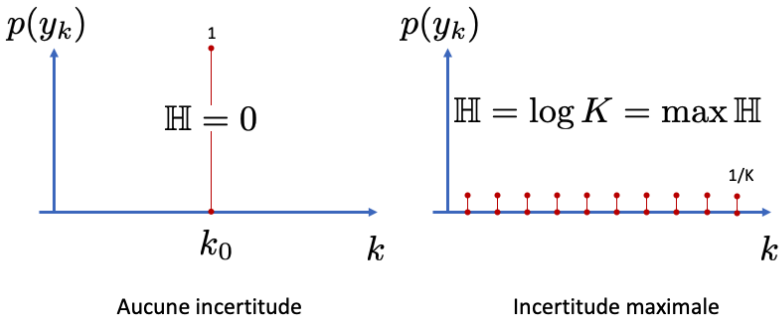


FIGURE 13 – Two extreme scenarios of probability distributions illustrating the measurement of uncertainty by the value of entropy.

variables are *i.i.d.*, for one draw:

$$\mathbb{P}(X_1 = x_1, X_2 = x_2, \ldots, X_d = x_d) = \prod_{i=1}^{d} \mathbb{P}(X_i = x_i) \tag{82}$$

Note that here we obtain one number. However, if we perform multiple draws, $\mathbb{P}(X_1, X_2, \ldots, X_d)$ itself becomes a random variable. Nevertheless, independence gives us the same decomposition:

$$\mathbb{P}(X_1, X_2, \ldots, X_d) = \prod_{i=1}^{d} \mathbb{P}(X_i) \tag{83}$$

Thus, by applying the law of large numbers:

$$-\frac{1}{d} \log \mathbb{P}(X_1, X_2, \ldots, X_d) = \frac{1}{d} \sum_{i=1}^{d} (-\log \mathbb{P}(X_i)) \xrightarrow[d\to\infty]{\text{prob.}} \mathbb{E}[-\log \mathbb{P}(X_i)] = \mathbb{H}[X] \tag{84}$$

(NDJE. we use $\mathbb{H}[X]$ to signify that it's the entropy of any random variable X_i.) Therefore, ***entropy naturally emerges when dealing with independent events***. Consider the following corollary:

Theorem 3 *If the* (X_i) *are i.i.d. with distribution* $\mathbb{P}(X)$*, taking values in* $\mathcal{A} =$

$\{y_k\}_{k\leq K}$ *associated with probabilities* $p(y_k)$*, then*

$$-\frac{1}{d}\log \mathbb{P}(X_1, X_2, \ldots, X_d) \xrightarrow[d\to\infty]{prob.} \mathbb{H}[X] = -\sum_{k=1}^{K} p(y_k)\log p(y_k) \tag{85}$$

We can rewrite the convergence in probability as:

$$\forall \varepsilon > 0, \lim_{d\to\infty} \mathbb{P}\left[\left|-\frac{1}{d}\log \mathbb{P}(X_1, X_2, \ldots, X_d) - \mathbb{H}[X]\right| \leq \varepsilon\right] = 1 \tag{86}$$

This tells us that for a sufficiently large d,

$$\mathbb{P}\left[\left|\underbrace{-\frac{1}{d}\log \mathbb{P}(X_1, X_2, \ldots, X_d)}_{f(\{X_i\}_{i\leq d})} - \mathbb{H}[X]\right| \leq \varepsilon\right] \geq 1 - \varepsilon \tag{87}$$

This means that ***the function with*** d ***variables*** $f(\{X_i\}_{i\leq d})$ ***will concentrate on a surface defined by the equation***

$$f(\{X_i\}_{i\leq d}) \approx \mathbb{H}[X] \tag{88}$$

with a small thickness ε. We can see a parallel with the intuition that images of dogs/cats aggregate on a surface in $\mathbb{R}^d$. To delve further, Claude Shannon developed the concept of a *typical set*.

4.2.3 Typical Set

Based on the above, we can focus on observations that will indeed (for a fixed ε) have an expectation that is within a distance ε from the entropy. Noting $\{X_i\}_{1\leq i\leq d} = \{x\}$, consider the set [52]:

$$T_d^{\varepsilon} = \left\{\{x\} \in \mathcal{A}^d, \left|-\frac{1}{d}\log \mathbb{P}(\{x\}) - \mathbb{H}[X]\right| \leq \varepsilon\right\} \tag{89}$$

52. NDJE. From a notation perspective, there might be confusion between the variable d used previously in the course and n which is generally used in this context and in the 2022 course, for example.

Note that the membership condition for T_d^ε can be expressed as (taking logarithms in base 2):

$$2^{-d(\mathbb{H}[X]+\varepsilon)} \leq \mathbb{P}(\{x\}) \leq 2^{-d(\mathbb{H}[X]-\varepsilon)} \tag{90}$$

This means that up to a non-negligible ε (which governs the convergence of concentration), the probability $\mathbb{P}(\{x\})$ is a constant given by:

$$\mathbb{P}(\{x\}) \approx 2^{-d\mathbb{H}[X]} \tag{91}$$

This indicates that ***within the typical set, probabilities are almost uniform***. Note that due to the additivity of entropy $\mathbb{H}[\{x\}] = d\mathbb{H}[X]$, which helps understand the scaling.

Moreover, we have the following properties. Firstly, for any $\varepsilon > 0$, and d large enough as per Eq. 87:

$$\mathbb{P}[\{x\} \in T_d^\varepsilon] \geq 1 - \varepsilon \tag{92}$$

That is to say, ***almost all realizations will belong to*** T_d^ε***, hence the term "typical set."*** Secondly, the cardinality of the typical set satisfies [53]:

$$(1-\varepsilon)2^{d(\mathbb{H}[X]-\varepsilon)} \leq |T_d^\varepsilon| \leq 2^{d(\mathbb{H}[X]+\varepsilon)} \tag{93}$$

This leads to a relation:

$$\mathbb{P}(\{x\}) \approx \frac{1}{|T_d^\varepsilon|} \tag{94}$$

This relation echoes a similar one in Statistical Physics (Th. 1), establishing ***a close relationship between the microcanonical ensemble*** Ω ***and the typical set*** T_d^ε. The relationships Eqs. 92, 93 form the basis of the famous ***fundamental principle of equiprobability***, defined as a consequence of a probability concentration phenomenon for typical sets (***asymptotic equipartition theorem***). Entropy provides the number of possible

53. NDJE. The proof can be found in the 2022 course Sec. 6.4.

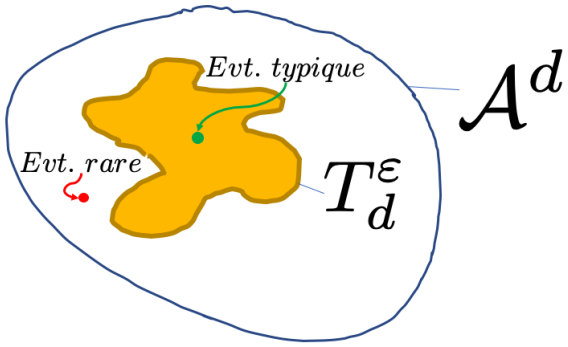


FIGURE 14 – Illustration of a typical set.

states/configurations that have a non-negligible probability, even if it can be very low (NDJE. $|\Omega|$ can be very, very large).

4.2.4 Coding (Typical)

Another concept that we revisit this year concerns *coding*, and how does entropy serve us in this context? Let $X \in \mathcal{A}^d$, it can either be in the typical set T_d^ε with a "high" probability or outside it with a much lower probability (Fig. 14). So, we can think of the following code (called the ***typical code***):

— Either 1 bit that codes whether $X \in T_d^\varepsilon$ or $X \notin T_d^\varepsilon$;

— If $X \in T_d^\varepsilon$, then it can be anywhere, so the code must identify all members of the set. Its length is thus $\ell(X) = \lceil \log_2 |T_d^\varepsilon| \rceil = \lceil d(\mathbb{H}[X] + \varepsilon) \rceil$;

— Or $X \notin T_d^\varepsilon$, and we assume that each of the d components takes values in $\mathcal{A}$ of size K, so we need a code of length $\ell(X) = \left\lceil \log_2 |\mathcal{A}^d| \right\rceil = \lceil d \log_2 K \rceil$.

In a way, ***the idea is to use shorter codes when the probability is high and longer codes when the probability is low***.

What we will optimize is the average number of bits. Hence the following theorem:

Theorem 4 ***(Shannon's Bound)***
The average number of bits per component of $X \in \mathcal{A}^d$ from a typical code is such that:

$$R = \frac{1}{d} \sum_{X \in \mathcal{A}^d} \ell(X)\mathbb{P}(X) \leq \mathbb{H}[X] + C\varepsilon \tag{95}$$

Proof 4. For any X, if we decompose membership in the typical set T_d^ε or not, we have:

$$\begin{aligned} R &= \frac{1}{d} \sum_{X \in T_d^\varepsilon} \ell(X)\mathbb{P}(X) + \frac{1}{d} \sum_{X \notin T_d^\varepsilon} \ell(X)\mathbb{P}(X) \\ &= \frac{1}{d}(\lceil d(\mathbb{H}[X] + \varepsilon) \rceil + 1) \left(\sum_{X \in T_d^\varepsilon} \mathbb{P}(X) \right) + \frac{1}{d}(\lceil d \log_2 K \rceil + 1) \left(\sum_{X \notin T_d^\varepsilon} \mathbb{P}(X) \right) \end{aligned} \tag{96}$$

Now, $\sum_{X \in T_d^\varepsilon} \mathbb{P}(X) \leq 1$ and $\sum_{X \notin T_d^\varepsilon} \mathbb{P}(X) \leq \varepsilon$, and also $\lceil x \rceil = \lfloor x \rfloor + 1 \leq x + 1$, so:

$$R \leq \frac{1}{d}(d(\mathbb{H}[X] + \varepsilon) + 2) + \frac{1}{d}(d \log_2 K + 2)\varepsilon \leq \mathbb{H}[X] + \varepsilon C' + \frac{2}{d} \tag{97}$$

which, for sufficiently large d, provides the result [54]. ■

So, ***the average number of bits per symbol from the typical code is on the order of the entropy of the probability distribution of each symbol***. In fact, we have an upper bound, and the question that arises is whether we can do better? The hope is to find sets that are more efficient than typical sets because, otherwise, Shannon's coding is the best we can do.

So, can we find such more efficient sets, or in other words, ***sets that concentrate even more probability on a smaller number of events*** than what is given by entropy?

Theorem 5 ***(Optimality of Typical Sets)***
Let $X = (x_1, \ldots, x_d)$ where the x_i are iid random variables with distribution $p(x)$

54. NDJE. There was a minor error in the 2022 course's proof, which didn't affect the conclusion, but I'm correcting it this year.

(e.g., the values of x_i belong to $\mathcal{A}$ of size K). Let B_δ^d be the smallest set such that:

$$\mathbb{P}(X \in B_\delta^d) \geq 1 - \delta \tag{98}$$

then, for all $\delta, \delta' > 0$:

$$\mathbb{P}(X \in B_\delta^d) \geq 1 - \delta \Rightarrow \frac{1}{d} \log_2 |B_\delta^d| \geq \mathbb{H}[X] - \delta' \tag{99}$$

The proof given in the 2022 course Sec. 6.6 involves the analysis of the intersection between B_δ^d and T_d^ϵ. It shows that both sets have identical sizes (the intersection probability is nearly 1). ***So, the minimum-sized set that concentrates the observations is indeed the set whose size is given by entropy, namely the typical set***. Therefore, we cannot find a code that surpasses the typical code, and ***the bound given by Theorem 4 is optimal***. At the core of this reasoning, we have ***the law of large numbers***, which gives us relation 84, which is itself the result of the ***independence*** of the X_i.

Now, the typical code is not practical because one needs to know whether an element belongs to the typical set or not. We had seen in the 2022 course ***instantaneous entropy codes*** (Shannon, Huffman) [55]. What will occupy us more this year is ***extending these results to the continuous case***, i.e., when the x_i take values in $\mathbb{R}$ and not in a discrete alphabet (NDJE. pixel intensities of an image, or sound samples are generally digitized, but we can imagine that the quantization step is much smaller than the dynamic range).

5. Lecture 8 Feb.

Recall what we have seen; ***entropy is a mathematical concept*** that has transcended its original domain (as in Statistical Physics) thanks to the work of Cl. Shannon, who made it an intrinsic property of ***probabilities***. Especially in high dimensions, it helps us understand ***concentration phenomena***, which are deeply connected to the concept of ***statistical independence***. However, we need to delve into cases more relevant to real-world data, where ***structure exists*** as soon as we start observing images of objects/scenes,

55. See also some implementations here https://github.com/jecampagne/cours_mallat_cdf/tree/main/cours2022

music/speech waveforms, text, and so on. Thus, we will explore how the notion of entropy from the previous sections can ***generalize in the case of non-independent variables*** and whether there are still concentration phenomena. We will see that this is indeed the case, provided additional assumptions of ***stationarity and ergodicity***. The temporal evolution of entropy will be considered in the Markovian framework.

5.1 Differential Entropy

In the previous session, we discussed entropy in the case of variables taking values in a discrete alphabet. It generalizes to the case of continuous values. Consider the following definition [56]:

Definition 2 *(**Differential Entropy**)*
Let X be a random variable with probability density with respect to Lebesgue measure dx, denoted as $p(x)$ ($x \in \mathbb{R}$ or $\mathbb{R}^n$). The differential entropy is then defined as:

$$\mathbb{H}_d = \mathbb{E}_{x\sim p(x)}[-\log p(x)] = -\int_{\mathbb{R}^n} p(x)\log p(x)\ dx \tag{100}$$

Unlike its "discrete" counterpart, differential entropy is not necessarily positive. A classic example is considering a 1D uniform distribution, $x \sim \mathcal{U}[0,a]$, where the differential entropy is $\mathbb{H}_d = \log a$, which becomes negative when $a < 1$.

Another more interesting example is that of Gaussian distributions:

$$X \sim \mathcal{N}(\mu,\sigma^2) \Rightarrow \mathbb{H}_d = \frac{1}{2} + \frac{1}{2}\log\Big(2\pi\sigma^2\Big) = \frac{1}{2}\log(2\pi e) + \log\sigma \tag{101}$$

So, when $\sigma \ll 1$, $\mathbb{H}_d \ll 0$. Note that the "$\log\sigma$" term reflects a general aspect of the role of a scale factor, as:

$$\forall \alpha > 0, \quad \mathbb{H}_d(\alpha X) = \mathbb{H}_d(X) + \log\alpha \tag{102}$$

56. NDJE. I opt for the change of notation introduced by S. Mallat from d to n for the dimension of space, knowing that d here pertains to *differential*. Also, this way, we keep the notations from the 2022 Course.

This follows from the fact that $\frac{1}{\alpha}p(\frac{x}{\alpha}) = p(x)$. Increasing the variance increases uncertainty about the variable, thus increasing entropy. Note that the measurement interval defines the relativity aspect of entropy's definition.

Now, armed with this new tool, we will revisit the properties seen in the discrete case.

5.2 Asymptotic Equipartition in the Continuous Case, Typical Sets

If we return to the case of independent random variables, we have the following theorem, which can be related to Th. 3 seen in the "discrete" case:

Theorem 6 *If the $(X_i)_{i\leq n}$ are n iid random variables taking their values in $\mathbb{R}$ ($\forall i, X_i \sim p$), then*

$$-\frac{1}{n}\log p(x_1,\dots,x_n) = -\frac{1}{n}\sum_{i=1}^{n}\log p(x_i) \xrightarrow[n\to\infty]{prob.} \mathbb{H}_d(x) = \mathbb{E}_{x\sim p(x)}[-\log p(x)] = -\int_{\mathbb{R}} p(x)\log p(x)dx \quad (103)$$

Therefore,

$$\forall \varepsilon \geq 0, \quad \mathbb{P}\left(\left|-\frac{1}{n}\log p(x_1,\dots,x_n) - \mathbb{H}_d(x)\right| \leq \varepsilon\right) \geq 1-\varepsilon \quad (104)$$

This property naturally leads us to the definition of typical sets, which is identical to the discrete case (Sec. 4.2.3):

$$T_n^\varepsilon = \left\{\{X\} \in \mathbb{R}^n, \left|-\frac{1}{d}\log p(\{X\}) - \mathbb{H}_d(x)\right| \leq \varepsilon\right\} \quad (105)$$

So, we are almost certain that $\{X\}$ is an element of the typical set:

$$\mathbb{P}(\{X\} \in T_n^\varepsilon) \geq 1-\varepsilon \quad (106)$$

and we will rediscover the two properties seen in the discrete case.

Firstly, the expression 105 tells us that

$$2^{-n(\mathbb{H}_d(x)+\varepsilon)} \leq p(\{X\}) \leq 2^{-n(\mathbb{H}_d(x)-\varepsilon)} \tag{107}$$

which provides ***uniformity of probability over the typical set*** up to ε.

Additionally, the size of the typical set, which cannot be defined by counting its elements, has the following property:

Theorem 7 *(**Typical Set Volume**)*
Let the volume of a set Ω be relative to the Lebesgue measure:

$$V(\Omega) := \int_{\Omega} dx$$

For sufficiently large n,

$$(1-\varepsilon)2^{n(\mathbb{H}_d(x)-\varepsilon)} \leq V(T_n^{\varepsilon}) \leq 2^{n(\mathbb{H}_d(x)+\varepsilon)} \tag{108}$$

Proof 7.

First,

$$1 = \mathbb{P}(\{X\} \in \mathbb{R}^n) \geq \mathbb{P}(\{X\} \in T_n^{\varepsilon}) = \int_{T_n^{\varepsilon}} p(x)dx \geq 2^{-n(\mathbb{H}_d(x)+\varepsilon)} \int_{T_n^{\varepsilon}} dx \tag{109}$$

so,

$$V(T_n^{\varepsilon}) \leq 2^{n(\mathbb{H}_d(x)+\varepsilon)} \tag{110}$$

Secondly,

$$1-\varepsilon \leq \mathbb{P}(\{X\} \in T_n^{\varepsilon}) = \int_{T_n^{\varepsilon}} p(x)dx \leq 2^{-n(\mathbb{H}_d)(x)-\varepsilon} \int_{T_n^{\varepsilon}} dx \tag{111}$$

so,

$$V(T_n^{\varepsilon}) \geq (1-\varepsilon)2^{n(\mathbb{H}_d(x)-\varepsilon)} \tag{112}$$

■

It is also found that, approximately, the probability within a typical set is inversely pro-

portional to its size:

$$\mathbb{P}(\{X\} \in T_n^\varepsilon) \approx \frac{1}{V(T_n^\varepsilon)} \tag{113}$$

What we need to remember is that everything is based on the property of ***convergence in probability***, which is itself based on the ***independence*** of the *random variables* (X_i). We are certain that data in high dimensions will end up in T_n^ε with uniform probability governed by the differential entropy. This outlines the support of ***deterministic intuition in a probabilistic approach***. Another result, which we won't prove, is that ***you can't do better in this independence of variables framework***.

As mentioned before, we need to move beyond this framework because most of the time, data interacts. For instance, consider the pixels in an image of an object (vase, table, etc.): the contour of this object necessarily connects several pixels, possibly on a large scale, and to recognize this object among a scene, you must account for these connections/interactions. This is where things become more subtle. Let's keep in mind that entropy is a measure of uncertainty about the sequence of n random variables $(x_1, \ldots, x_n)$. If there are dependencies between these variables, they are not all independent, and there is redundancy, which means less uncertainty, so entropy must decrease, resulting in a smaller size for typical sets. This is a good thing, especially for coding purposes [57]. So, let's delve into the concept of dependency.

5.3 Dependency and Entropy (Joint, Conditional, Relative)

We will use conditional probabilities to define an entropy. The discrete case [58] will provide us with a framework to become familiar while allowing us to handle positive entropies. The continuous framework requires finding the right definition using the Kullback-Leibler divergence.

57. In a side note, S. Mallat tells us that although you can't code a real number with a finite number of bits, if you accept a small quantization error, you can proceed as in the discrete case. These are problems of *Sphere packing*. You can refer to the 2022 Course, Sec. 8. NDJE. In this *Sphere packing* theme, it's worth mentioning that a Fields Medal in 2022 was awarded to Ukrainian-born mathematician Maryna Viazovska: "*[She] is awarded the Fields Medal 2022 for the proof that the E8 lattice provides the densest packing of identical spheres in 8 dimensions, and further contributions to related extremal problems and interpolation problems in Fourier analysis.*"

58. NDJE. See the 2022 course (Sec. 6.2)

Definition 3 *The joint entropy of two random variables X and Y with values in an alphabet $\mathcal{A}$ is defined as:*

$$\mathbb{H}(X,Y) := -\mathbb{E}_{(x,y)\sim p}[\log p(X,Y)] = -\sum_{k,k'} p(X=a_k, Y=a_{k'}) \log p(X=a_k, Y=a_{k'}) \tag{114}$$

and the conditional entropy is defined as:

$$\begin{aligned}\mathbb{H}(Y|X) &:= \sum_k p(X=a_k)\mathbb{H}(Y|X=a_k) \\ &= -\sum_{k,k'} p(X=a_k, Y=a_{k'}) \log p(Y=a_{k'}|X=a_k) \\ &= -\mathbb{E}_{(x,y)\sim p}[\log(Y|X)]\end{aligned} \tag{115}$$

A property connects these two entropies:

Property 1

$$\mathbb{H}(X,Y) = \mathbb{H}(X) + \mathbb{H}(Y|X) \tag{116}$$

If the variables are independent, we then recover the additivity law of entropies. The proof (Course 2022 Sec. 6.2) relies on the conditional probability form of Bayes' theorem. Indeed [59],

$$\begin{aligned}\mathbb{H}(X,Y) - \mathbb{H}(X) &= -\sum_{x,y} p(x,y)\log p(x,y) + \sum_x p(x)\log p(x) \\ &= -\sum_{x,y} p(x,y)\log p(x,y) + \sum_x \left(\sum_y p(x,y)\right)\log p(x) \\ &= -\sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)} = -\sum_{x,y} p(x,y)\log p(y|x) \\ &= \mathbb{H}(Y|X)\end{aligned} \tag{117}$$

Let's now introduce the ***relative entropy***, also known as the ***Kullback-Leibler divergence*** [60]. It is widely used in probability to measure the suitability between two distribu-

59. To simplify notation, $p(x,y) = p(X=x, Y=y)$, the same applies to $p(x)$ and $p(y|x)$.
60. See also the 2019 course (Sec. 7.2.3)

tions, hence its misleading name as the Kullback-Leibler distance (it is not symmetric).

Definition 4 *(**Kullback-Leibler**)*
If the support[a] *of q includes the support of p, then*

$$D_{KL}(p\|q) := \int_{\mathbb{R}^n} p(x) \log \frac{p(x)}{q(x)} dx = \mathbb{E}_{X\sim p}\left[\log \frac{p(X)}{q(X)}\right] \tag{118}$$

The integral can be transformed into a sum if necessary.

a. By convention, we set $0 \log 0 = 0$ and $0 \log(0/0) = 0$.

We can view $D_{KL}(p\|q)$ as a measure of coding inefficiency. The size of the optimal code for $(x_1, \ldots, x_n)$ is on the order of the entropy of the probability associated with x_i. For example, if this is p, the average code size is on the order of $\mathbb{E}_p[-\log p]$. But if we make a mistake[61], thinking that $x \sim q$, then we obtain an average size of $\mathbb{E}_p[-\log q]$ (cf. x is always drawn according to unknown p). Therefore, the difference between the two sizes is nothing other than $D_{KL}(p\|q)$.

The important properties of $D_{KL}(p\|q)$ are as follows:

- $D_{KL}(p\|q) \neq D_{KL}(q\|p)$, which comes from the very definition showing that it is not a distance;
- $D_{KL}(p\|q) \geq 0$, and $D_{KL}(p\|q) = 0 \Leftrightarrow p = q$.

To prove the second property, we can use the following theorem:

Theorem 8 *(**Jensen's Inequality**)*
*Let f be a **convex function** in one dimension (second derivative positive or non-negative). Then, for any random variable X,*

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X]) \tag{119}$$

*and if f is **strictly convex** (second derivative strictly positive), we have equality if*

61. This is a common case because we model data without being certain that the model reflects reality. It is indeed a philosophical debate but not only that.

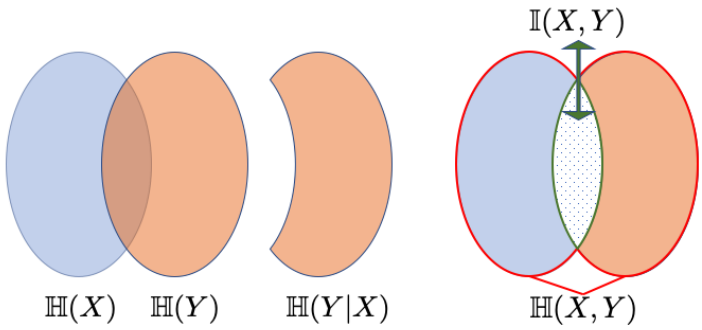


FIGURE 15 – Schematic of entropy $\mathbb{H}(X)$ and $\mathbb{H}(Y)$, conditional entropy $\mathbb{H}(Y|X)$, mutual information, and joint entropy.

and only if the only value taken by X is $\mathbb{E}[X]$.

Thus, since the function log is strictly concave, and therefore $-\log$ is strictly convex, we have:

$$\begin{aligned} D_{KL}(p\|q) &= -\int p(x)\log\frac{q(x)}{p(x)}dx \\ &= \mathbb{E}_p\left[-\log\frac{q(x)}{p(x)}\right] \geq -\log\mathbb{E}_p\left[\frac{q(x)}{p(x)}\right] = -\log(1) = 0 \end{aligned} \tag{120}$$

The strict concavity of the logarithm tells us that the inequality above becomes equality if and only if $p(x)/q(x)$ takes a unique value. Let c be this value, since $\int p(x)dx = \int q(x)dx = 1$ then $c = 1$, and thus we have the second result of the theorem [62]. To understand dependence, let's use the concept of mutual information. But before that, we can prove the following proposition:

Property 2

$$\mathbb{H}(X|Y) \leq \mathbb{H}(X) \tag{121}$$

In other words, conditioning reduces uncertainty.

62. NDJE. We could also have used the fact that for all $x > 0$, $-\log x \geq 1 - x$, with equality being true if and only if $x = 1$.

This can be proven as follows:

$$\begin{aligned}\Delta &= \mathbb{H}(X) - \mathbb{H}(X|Y) = \mathbb{H}(X) + \mathbb{H}(Y) - \mathbb{H}(X,Y) \\ &= \sum_{x,y} p(x,y)\log p(x,y) - \sum_{x} p(x)\log p(x) - \sum_{y} p(y)\log p(y) \\ &= \sum_{x,y} \left[p(x,y)\log p(x,y) - p(x,y)\log p(x) - p(x,y)\log p(y)\right] \\ &= \sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} = D_{KL}(p(x,y)\|p(x)p(y)) \geq 0 \end{aligned} \tag{122}$$

In passing, we can see that equality prevails if the random variables X and Y are independent. From there, it becomes natural to define mutual information (Shannon) as follows:

Definition 5 *(**Mutual Information**)*
Let two random variables X and Y have a joint probability of $p(x,y)$ and marginals $p(x)$, $p(y)$. The mutual information is the following quantity:

$$\mathbb{I}(X,Y) := D(p(x,y)\|p(x)p(y)) = \mathbb{H}(X) + \mathbb{H}(Y) - \mathbb{H}(X,Y) \geq 0 \tag{123}$$

If there is conditioning, entropy is reduced, which is reflected in $\mathbb{I}(X,Y) > 0$, quantifying to some extent the action brought by knowing one variable on another.

A brief summary of these concepts in the form of a diagram is provided in Figure 15. Now, we can tackle the problem of entropy calculation in high dimensions when variables have interdependencies.

5.4 Equipartition with Dependence

5.4.1 Average Entropy, Entropy Rate

When we are in dimension n, the definition of entropy itself is not problematic. Let there be n random variables $\{X\}_n = (X_1, \ldots, X_n)$ (here taking values in an alphabet):

$$\mathbb{H}(\{X\}_n) = -\mathbb{E}_{\{X\}_n \sim p}[\log p(\{X\}_n)] \tag{124}$$

The point of interest is how this behaves as n tends towards infinity. We have the intuition from statistical physics where *entropy is an extensive variable* that grows with the number of particles. Therefore, by studying the average of $\mathbb{H}(\{X\})$, we might expect a constant.

Definition 6 ***entropy rate***

The entropy rate (entropy rate) is called the limit, if it exists, as follows:

$$\mathbb{H}(\chi) := \lim_{n\to\infty} \frac{1}{n}\mathbb{H}(\{X\}_n) \tag{125}$$

It is the average entropy per element/symbol.

In the *independent* case, by the additivity of entropy:

$$\frac{1}{n}\mathbb{H}(\{X\}_n) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{H}(X_i) \tag{126}$$

Does the limit exist? If they have the same law, then the limit exists and gives us asymptotic equipartition. But otherwise, there is no reason for the limit to exist *a priori*. Take the case where each X_i is a Bernoulli variable with probability a_i for 1 (a_i different from one X_i to another).

$$\mathbb{H}[\mathscr{B}(a)] = -a\log(a) - (1-a)\log(1-a) = \begin{cases} 0 & a = 1 \\ 1 & a = 1/2 \end{cases} \tag{127}$$

We can then construct a series of X_i for which the average oscillates in $[0,1]$ without obtaining a limit [63].

That being said, except in the case where all variables X_i are independent, convergence will appear in the context of *stationary processes*. This assumption is not restrictive; on the contrary, if we take an image with a CCD camera, the distribution law of the n random variables consisting of the intensities of the n pixels is roughly the same whether we point the camera here or there (cf. image and translated image). If we pixelate an

63. NDJE. One attempt might be, for example, since we can make $\mathbb{H}[X_i] \in \{0,1\}$, the current averages are elements of $[0,1]$. Then, we can consider the value of the average at step k; then complete the sequence of X_i ($i > k$) with 1's to get as close as we want to 1 (from below), then complete the series with 0's to get as close as we want to 0 (from above), and so on.

entire scene and extract batches of n pixels translated relative to each other, then the assumption tells us that $p(X_1, \ldots, X_n) = p(X_{1+t}, \ldots, X_{n+t})$. *Of course, this assumption is false as soon as we perform operations like centering images* of faces, numbers, galaxies, etc. So be cautious, but the assumption of stationarity related to translation invariance remains well satisfied in practice.

Let's go into detail and start by specifying stationarity.

Definition 7 ***stationarity***

A random process is said to be stationary if and only if the joint probability of any sequence $X_1, X_2, \ldots, X_n$ is translation-invariant:

$$\forall (n,k)\ \forall (x_1,\ldots,x_n), \quad p(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) \\ = p(X_{1+k} = x_1, X_{2+k} = x_2, \ldots, X_{n+k} = x_n) \quad (128)$$

In particular:

$$\forall k, \quad p(X_1 = x_1) = p(X_{1+k} = x_1) \quad (129)$$

This means that the distribution of X_1 is the same as any other variable X_j, ensuring the convergence of the average entropies of the X_i and hence the existence of $\mathbb{H}(\chi)$. Similarly,

$$p(X_1 = x_1 | X_2 = x_2) = \frac{p(X_1 = x_1, X_2 = x_2)}{p(X_2 = x_2)} = \frac{p(X_{1+k} = x_1, X_{2+k} = x_2)}{p(X_{2+k} = x_2)} \\ = p(X_{1+k} = x_1 | X_{2+k} = x_2) \quad (130)$$

This can be generalized to conditional probabilities $p(X_n | X_1, \ldots, X_{n-1})$, for example.

As a small example, where the X_i take values in an alphabet of K symbols with uniform probability $1/K$, the stationarity assumption gives us $\mathbb{H}(\chi) = \log K$. However, in a language like French, letter frequencies are not uniform, and the probability of two consecutive letters is not equal to the product of the probabilities of each. So... it's a different context. However, we will define another average that is more manageable than this average entropy.

5.4.2 Average Conditional Entropy, Another Entropy Rate

If we have a sequence of random variables $X_1, \ldots, X_n$ and we want to code it, one method is to do it iteratively: we start by coding X_1, then we code X_2 knowing X_1, then X_3 knowing X_2, X_1, and so on. Conditional entropy $\mathbb{H}(X_n|X_1, \ldots, X_{n-1})$ involves conditional probability. The question is whether, as n tends to infinity, this conditional entropy converges to a limit. It will then be the additional information that variable X_n provides compared to all the information already collected with the previous $n-1$ variables.

Theorem 9 *In the case of a stationary process, firstly $\mathbb{H}(\chi)$ exists, and secondly:*

$$\mathbb{H}'(\chi) := \lim_{n\to\infty} \mathbb{H}(X_n|X_1, \ldots, X_{n-1}) = \mathbb{H}(\chi) \tag{131}$$

Proof 9. The proof starts by noting that:

$$\mathbb{H}(X_n|X_1, \ldots, X_{n-1}) \leq \mathbb{H}(X_{n-1}|X_1, \ldots, X_{n-2}) \tag{132}$$

Indeed:

$$\begin{aligned} \mathbb{H}(X_n|X_1, \ldots, X_{n-1}) &\leq \mathbb{H}(X_n|X_2, \ldots, X_{n-1}) && \text{(Eq. 121)} \\ &\leq \mathbb{H}(X_{n-1}|X_1, \ldots, X_{n-2}) && \text{(Def. 7)} \end{aligned} \tag{133}$$

Therefore, we have a sequence of positive values (discrete case) that decrease, so there is convergence, and $\mathbb{H}'(\chi)$ exists. Next:

$$\frac{1}{n}\mathbb{H}(X_1, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{H}(X_i|X_1 \ldots, X_{i-1}) \tag{134}$$

which we obtain by iterating Eq. 116. If we take the average, the left-hand side tends to $\mathbb{H}(\chi)$. For the right-hand side, we can use Cesàro's theorem:

Theorem 10 ***(Cesàro's sum)***
Let $(a_i)_{i>0}$ be a sequence of numbers ($\mathbb{R}$ or even $\mathbb{C}$). If the sequence converges to μ,

then the sequence of averages with general term:

$$b_n = \frac{1}{n}\sum_{i=1}^{n} a_i \tag{135}$$

also converges, and its limit is μ.

Proof 10. The proof is as follows. First, the convergence of the sequence (a_i) allows us to say that for all $\varepsilon \geq 0$, there exists N_ε such that for all $i \geq N_\varepsilon$, we have $|a_i - \mu| \leq \varepsilon$. Now consider $n > N_\varepsilon$:

$$\begin{aligned}
|b_n - \mu| &= \left|\frac{1}{n}\sum_{i=1}^{n}(a_i - \mu)\right| = \left|\frac{1}{n}\sum_{i=1}^{N_\varepsilon - 1}(a_i - \mu) + \frac{1}{n}\sum_{N_\varepsilon}^{n}(a_i - \mu)\right| \\
&\leq \left|\frac{1}{n}\sum_{i=1}^{N_\varepsilon}(a_i - \mu)\right| + \left|\frac{1}{n}\sum_{i=N_\varepsilon+1}^{n}(a_i - \mu)\right| \\
&\leq \left|\frac{1}{n}\sum_{i=1}^{N_\varepsilon}(a_i - \mu)\right| + \frac{1}{n}\sum_{i=N_\varepsilon+1}^{n}|a_i - \mu| \\
&\leq \frac{1}{n}\left|\sum_{i=1}^{N_\varepsilon}(a_i - \mu)\right| + \underbrace{\frac{n - N_\varepsilon}{n}\varepsilon}_{\leq \varepsilon}
\end{aligned} \tag{136}$$

The sum in the first term on the right does not depend on n, so there exists a $N'_\varepsilon \geq N_\varepsilon$ such that for all $n \geq N'_\varepsilon$

$$\frac{1}{n}\left|\sum_{i=1}^{N_\varepsilon}(a_i - \mu)\right| \leq \varepsilon$$

Therefore, for all $n \geq N'_\varepsilon$, $|b_n - \mu| \leq 2\varepsilon$, which gives us $\lim_{n\to\infty} b_n = \mu$ (note that we can make the inequality bound ε instead of 2ε). ■

So we can conclude that the right-hand side of Eq. 134 converges to $\mathbb{H}'(\chi)$, and thus we conclude that:

$$\mathbb{H}(\chi) = \mathbb{H}'(\chi)$$

■

We will see how these results give us access to an ***equipartition theorem***; in particular,

we will need to add an ***ergodicity assumption***.

5.5 NDJE. A Brief Guide in the Continuous Setting

Here, I compile some definitions and relationships in the case of random variables taking values in $\mathbb{R}$. In the context of differential entropy, we introduced the following:

$$\mathbb{H}_d(X) = -\int p(x)\log p(x)dx \qquad \text{differential entropy} \tag{137}$$

$$\mathbb{H}_d(X,Y) = -\iint p(x,y)\log p(x,y)dxdy \qquad \text{joint differential entropy} \tag{138}$$

$$\mathbb{H}_d(X|Y) = -\iint p(x,y)\log p(x|y)dxdy \qquad \text{conditional differential entropy} \tag{139}$$

$$\mathbb{I}(X,Y) = D_{KL}(p(x,y)\|p(x)p(y)) \geq 0 \qquad \text{mutual information} \tag{140}$$

with the following relationships:

$$\mathbb{H}_d(X,Y) = \mathbb{H}_d(X|Y) + \mathbb{H}_d(Y) = \mathbb{H}_d(Y|X) + \mathbb{H}_d(X) \tag{141}$$

$$\mathbb{H}_d(X|Y) \leq \mathbb{H}_d(X) \tag{142}$$

$$\mathbb{I}(X,Y) = \mathbb{H}_d(X) + \mathbb{H}_d(Y) - \mathbb{H}_d(X,Y) \tag{143}$$

Most of these relationships have been proven in the case of random variables with values in a discrete alphabet but are also valid in the continuous case.

Similarly, for $X = (X_1, \ldots, X_n)$, we have the following relationships:

$$\mathbb{H}_d(X+c) = \mathbb{H}_d(X) \qquad \forall c \in \mathbb{R} \tag{144}$$

$$\mathbb{H}_d(\alpha X) = \mathbb{H}_d(X) + \log|\alpha| \qquad \forall \alpha \neq 0 \tag{145}$$

$$\mathbb{H}_d(\mathbf{A}X) = \mathbb{H}_d(X) + \log|\det \mathbf{A}| \qquad \forall \mathbf{A} \in GL_n(\mathbb{R}) \tag{146}$$

6. Lecture 15 Feb.

NDJE. I'd like to mention that I've set up a GitHub repository for some numerical applications illustrating the course. You can access it at https:// colab.research.google.

com/github/jecampagne/cours_mallat_cdf. For 2023, the notebooks can be directly executed on Google Colab. The migration of notebooks from 2022 is also planned.

Continuing from the previous section, we will explore more complex systems where variables exhibit interdependencies. It's important to note that this scenario is highly relevant when dealing with structured data. The key concept to keep in mind is that of the ***entropy rate*** (Definition 6), first defined as the limit for $n \to \infty$ of the ***average joint entropy*** of n random variables taking their values in χ (e.g., n pixels of an image or n samples of a sound waveform). As we've seen, for ***stationary processes***, it can also be defined as the ***limit of a sequence of conditional entropies*** through Theorem 9.

We will investigate the implications for coding and data concentration phenomena within a typical set. However, the stationarity assumption will need to be supplemented by ***ergodicity***, leading us to delve into the important ***Shannon–McMillan–Breiman theorem***, which provides insights into the concentration and uniformity of probabilities over typical sets whose characteristics are specified by entropy. In practice, we will study the significant example of ***Markov chains***, which are typical of a physical system where time is discretized, and the state at time t is entirely specified by memoryless variables. We will then explore how the concept of ***irreversibility*** emerges in this context and how entropy increases over time, constituting the ***second law of thermodynamics***.

6.1 Asymptotic Equipartition with Dependence: The Condition

In the independent case (Sec. 4.2.2), the joint probability $\mathbb{P}(X_1, \ldots, X_n)$ is a product, so the logarithm is a sum of n terms, and the average converges to the expected value of the logarithm of the probability, leading to concentration and the natural emergence of entropy in this case (Th. 3 (*discrete case*), 6 (*continuous case*)). Now, let's see how these concepts appear in the case of dependence.

It is still possible to write

$$\mathbb{P}(X_1, \ldots, X_n) = \mathbb{P}(X_1)\mathbb{P}(X_2|X_1) \ldots \mathbb{P}(X_n|X_1, \ldots, X_{n-1}) \tag{147}$$

and therefore

$$-\frac{1}{n}\log\mathbb{P}(X_1,\ldots,X_n) = -\frac{1}{n}\sum_{i=1}^{n}\log\mathbb{P}(X_i|X_1,\ldots,X_{i-1}) \tag{148}$$

It's clear that conditional entropy (Def. 3) [64] will appear in the right-hand side.

Let's imagine for a moment that we are in a situation where

$$-\frac{1}{n}\log\mathbb{P}(X_1,\ldots,X_n) \xrightarrow[n\to\infty]{\text{prob.}} \mathbb{H}(\chi) \tag{149}$$

then using reasoning developed, for example, in the continuous case (Sec. 5.2), we can define the typical set T_n^ε whose volume/size is of the order of $V(T_n^\varepsilon) \approx 2^{n\mathbb{H}(\chi)}$, and the probability that data belongs to this set is quasi-uniform with a value of $p(\{X\}) \approx 1/V(T_n^\varepsilon)$. In this case, we indeed have a data concentration phenomenon within a typical set with quasi-uniform probability. This has immediate consequences, for example, in coding, as the average number of bits required per symbol is on the order of $\mathbb{H}(\chi)$, even in the presence of dependencies.

So, ***the emergence of the concentration phenomenon and its consequences are much more general than in the case of independence of random variables, but it is still necessary to ensure convergence as*** $n \to \infty$**.** One should envision this asymptotic convergence, for example, as if we had a single image but with the number of pixels tending towards infinity.

6.2 Ergodicity, Birkhoff's Theorem

At some point, we'll return to some basic notions to introduce the concept of ergodicity (often referred to as the ergodic hypothesis), first introduced by L. Boltzmann in 1871 for his kinetic theory of gases. We define a probability space by a triplet $(\Omega, \mathscr{B}, \mathbb{P})$, with Ω being the universe, $\mathscr{B}$ the set of measurable sets (sigma-algebra) relative to the probability measure P. A random variable X is then a (measurable) function from Ω to χ a measurable space. The probability associated with X, denoted as $\mathbb{P}_X$, is defined in such a way that the probability that the random variable takes values in $B \subset \chi$ is the

64. or its differential counterpart Sec. 5.5.

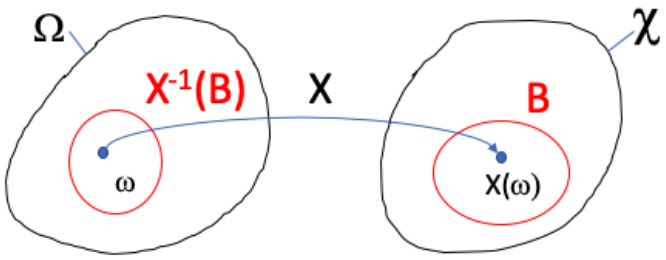


FIGURE 16 – Diagram illustrating the definition of a random variable.

probability of the set $X^{-1}(B)$ (Fig. 16)

$$\mathbb{P}_X(B) = \mathbb{P}(X^{-1}(B)) = \mathbb{P}(\{\omega \in \Omega, x = X(\omega) \in B\}) \tag{150}$$

To this framework, we add the ***stationarity hypothesis*** (Def. 7), meaning that if we take a collection of random variables, they are all of the same distribution, and joint probabilities are translation-invariant. Here, this translates to: for any $A \in \mathscr{B}$, $\mathbb{P}(T(A)) = \mathbb{P}(A)$ (stationarity) (NDJE. one can think of T as a time translation). Now, to this, the ***ergodicity hypothesis*** adds

$$\text{for any } A \in \mathscr{B} \text{ such that } T(A) = A \Leftrightarrow \begin{cases} \text{either } \mathbb{P}(A) & = 0 \\ \text{or } \mathbb{P}(A) & = 1 \end{cases} \quad \text{(ergodicity)} \tag{151}$$

This ergodicity hypothesis tells us that if we apply the transformation T iteratively, we effectively cover the entire set, and we cannot be trapped in a subpart of the set. In fact, if we have a system with 2 subsystems invariant under T, the ergodic hypothesis forces one of the two subsystems to have measure zero; we can almost never be trapped in that one, and we live in the other subsystem. In a sense, in this subsystem with measure 1, we have a *mixing notion*. Note that if we have a transformation that does not preserve the measure, these notions do not apply.

The two properties of stationarity and ergodicity yield the following theorem [65]:

65. George David Birkhoff (1884-1944)

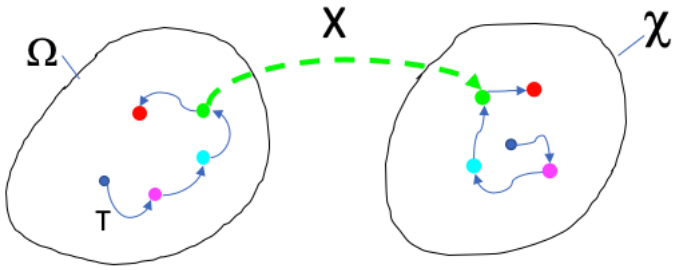


FIGURE 17 – Diagram illustrating the process of ergodic averaging.

Theorem 11 ***(Birkhoff)***

$$\frac{1}{n}\sum_{i=1}^{n} X(T^i(\omega)) \xrightarrow[n\to\infty]{p.s} \mathbb{E}[X] = \int_{\Omega} X(\omega) d\mathbb{P}(\omega) \tag{152}$$

where p.s (almost surely) means with probability 1 [a].

a. $\mathbb{P}(\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\}) = 1$

We understand that when we iterate T, we traverse the space of points in the Ω space, which, by action of X, traverses the χ space (Fig. 17). The ergodicity hypothesis then tells us that the empirical average of the values of X calculated on transformed points converges to the expectation of X with respect to the measure $\mathbb{P}$.

So, if we define the following:

Definition 8 *Let T be an ergodic transformation, we say that the random variables $(X_1, \ldots, X_n)$ such that $X_i = X \circ T^i$ form an ergodic process.*

then Birkhoff's Theorem simply gives us the *p.s* convergence of the empirical average

$$(X_1, \ldots, X_n) \text{ ergodic} \quad \Rightarrow \quad \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{p.s} \mathbb{E}[X] \tag{153}$$

A particular case is the ***dwell time*** in a set $A \subset \Omega$. For this, we note $\mathbf{1}_A$ as the indicator of A, and let's examine what the average of values $\mathbf{1}_A(T^k(\omega))$ yields (Fig. 18).

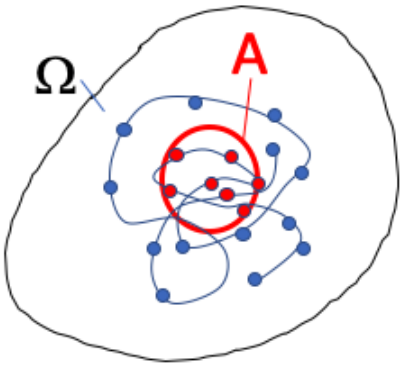


FIGURE 18 – Calculation of the average dwell time in $A \subset \Omega$.

Birkhoff's Theorem then gives us ($\mathbb{P}(\Omega) = 1$ here)

$$\frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_A(T^k(\omega)) \xrightarrow[n\to\infty]{p.s} \mathbb{E}(\mathbf{1}_A) = \int_\Omega \mathbf{1}_A(\omega) d\mathbb{P}(\omega) = \mathbb{P}(A) \qquad \text{(dwell time)} \qquad (154)$$

Therefore, the average number of returns to the set A by applying the ergodic transformation T gives us the probability of being in A, i.e., its measure relative to the probability distribution, in a sense, its size with respect to this measure.

This fundamental property of convergence of sums to expectations is found in physical processes.

6.3 Shannon–McMillan–Breiman Theorem

To return to our initial problem of convergence Eq. 149, suppose that the variables $(X_1, \ldots, X_n)$ form an *ergodic process* (Def. 8). Certainly, the logarithm of the conditional probability $\mathbb{P}(X_i|X_1 \ldots, X_{i-1})$ is a random variable, but we have a boundary problem due to X_i, which prevents us from operating a translation on the indices (cf. stationarity hypothesis). If X_i depended only on $X_{i-\ell} \ldots, X_{i-1}$ $(i > \ell)$, we could attempt a translation, and with the help of ergodicity, we could then have the $p.s$ convergence, hence probability convergence. Here is the theorem that ensures the appearance of concentration phenomena

on typical sets:

Theorem 12 *(**Shannon–McMillan–Breiman**) Let $(X_1, \dots, X_n)$ be a stationary and ergodic process with probability $\mathbb{P}$*

$$-\frac{1}{n}\log \mathbb{P}(X_1, \dots, X_n) \xrightarrow[n\to\infty]{p.s} \mathbb{H}(\chi) \tag{155}$$

Proof 12. We will outline the main steps of the proof. If we return to the decomposition of the joint probability

$$-\frac{1}{n}\log \mathbb{P}(X_1, \dots, X_n) = -\frac{1}{n}\sum_{i=1}^{n} \log \mathbb{P}(X_i|X_1, \dots, X_{i-1}) \tag{156}$$

we need to study the generic term of the sum on the right-hand side. The associated conditional entropy, namely its expectation, that is $(\{X\} = (X_1, \dots, X_n))$

$$\mathbb{H}(X_i|X_{i-1}, \dots, X_1) = \mathbb{E}_{\{X\}\sim\mathbb{P}(\{X\})}[-\log \mathbb{P}(X_i|X_{i-1}, \dots, X_1)] \tag{157}$$

can be bounded by recalling that entropy increases (decreases) if we decrease (increase) the conditioning (Eq. 121 and Sec. 5.5). Thus, for $1 < \ell$

$$\mathbb{E}[-\log \mathbb{P}(X_i|X_{i-\ell}, \dots, X_1)] \leq \mathbb{H}(X_i|X_{i-1}, \dots, X_1) \leq \mathbb{E}[-\log \mathbb{P}(X_i|X_{i-1}, \dots, X_{-\infty})] \tag{158}$$

if we extend the probability law for negative indices up to $-\infty$. The proof[66] is quite technical. However, the philosophy is to play with this bounding to demonstrate the convergence of the lower and upper bounds. This is done using the ergodicity hypothesis and Birkhoff's Theorem. Finally, by showing that the two limits are equal, this ensures the convergence of the conditional entropy taken in between. ■

This theorem is central. For the record, it provides access to all notions of ***concentration, typical set, uniformity of probability***, with entropy at its core, which sets the

66. See Sec. 16.8 of the book by Th. Cover and J. Thomas given as an introduction this year.

characteristics of these notions. The consequences are immediate, such as in coding, for example. However, this theorem may seem abstract, and to better understand its deep significance, we will study it in the context of Markov chains. This tool is very important both conceptually - in understanding stochastic dynamic systems, the limit of Markov chains gives access to stochastic differential equations - and algorithmically for sampling probability distributions in high dimensions (e.g., Monte Carlo Markov Chain). So, we will see an overview of Markov chains that will help us access the Second Law of Thermodynamics with the growth of entropy.

6.4 Markov Chains [67]

6.4.1 Definitions and Properties

The guiding underlying idea is the assumption that the future is accessible by knowing only the present (i.e., no memory). So, here's the following definition:

Definition 9 *(**Markov Chain**)*
A process $(X_1, \dots, X_n)$ *is a Markov chain if*

$$\forall (x_i)_{i \leq n+1} \in \chi^{n+1} \quad \mathbb{P}(X_{n+1} = x_{n+1} | X_n = x_n, \dots, X_1 = x_1) = \mathbb{P}(X_{n+1} = x_{n+1} | X_n = x_n) \quad (159)$$

A practical remark: if we have a finite-sized past (e.g., ℓ) knowing that X_n is the present value, then we can define a process $Y_n = (X_n, \dots, X_{n-\ell})$ which will be Markovian. An immediate consequence of this definition is (note that we later use the notation with (x_i),

67. Andrei Andreyevich Markov (1856-1922)

often referred to as "state" at time/step "i"):

$$\mathbb{P}(X_1 = x_1, \dots, X_n = x_n) = p(x_1, \dots, x_n) = p(x_1) \prod_{i=2}^{n} p(x_i | x_{i-1}, \dots, x_1)$$
$$= \underbrace{p(x_1)}_{\substack{initial \\ distribution}} \prod_{i=2}^{n} \underbrace{p(x_i | x_{i-1})}_{\substack{transition \\ probability}} \tag{160}$$

which introduces the notion of *transition probability* between two neighboring states. The stationarity hypothesis introduces the notion of a *stationary or homogeneous Markov chain*:

Definition 10 *(**Stationary Markov Chain**)*
A Markov chain is stationary if

$$\forall (x, y) \in \chi^2, \forall n, \quad \mathbb{P}(X_{n+1} = y | X_n = x) = \mathbb{P}(X_n = y | X_{n-1} = x) \tag{161}$$

so the transition probabilities do not depend on the time step n, which we can denote as $P_{x,y}$ (be mindful of the order, think "x influences y").

Property 3
In the case of a discrete alphabet (i.e., χ is discrete), then $(P_{x,y})_{(x,y)\in\chi^2}$ is a stochastic transition matrix that has the following properties:

— *Firstly,*

$$\forall x \in \chi, \ \sum_{y \in \chi} P_{x,y} = 1 \qquad (stochastic\ matrix) \tag{162}$$

which comes from the fact that $p(y|x) = p(x, y)/p(x)$ and $\sum_y p(x, y) = p(x)$.

— *Secondly, by noting that*[a]

$$\forall y \in \chi, \ \mathbb{P}(X_{n+1} = y) = \sum_{x \in \chi} \mathbb{P}(X_n = x) P_{x,y} \tag{163}$$

we can use matrix notation by grouping the probabilities at step n as a column

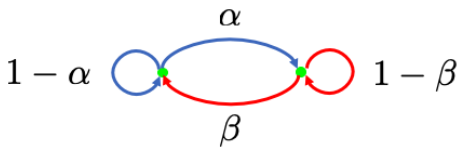


FIGURE 19 – Stationary 2-state Markov process.

vector μ_n*, so*

$$\mu_n := (\mathbb{P}(X_n = x))_{x\in\chi} \Longrightarrow \mu_{n+1} = P^T\ \mu_n \tag{164}$$

a. We can easily see this by expressing that $\mathbb{P}(X_{n+1} = y)$ is a marginal of the joint probability of (X_{n+1}, X_n).

When invoking the ergodicity hypothesis, Birkhoff's Theorem (Th. 11) that ensures convergence invites us to study the iterates of the transformation $P = (P_{x,y})$, which, for example, yields

$$\mu_{n+k} = (P^T)^k\ \mu_n \tag{165}$$

The study of the *eigenvalues of this transition matrix* will inform us about the concepts of *equilibrium* and *invariant measure*.

6.4.2 Some Examples

6.4.2.1 2-State Model

A first simple example (Fig. 19) of a Markovian process is one where the transition matrix is given by

$$P_{x,y} = \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix} \tag{166}$$

6.4.2.2 Random Walk

More complex scenarios can be studied using ***generalized random walks*** that introduce the concept of independent hidden variables.

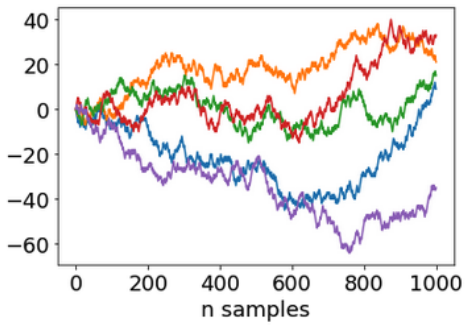


FIGURE 20 – 1D Random Walk: $X_n = X_{n-1} + Z_n$ where Z_n follows a Bernoulli distribution $\{+1, -1\}$ ($p(+1) = 1/2$). All five chains start at the origin $x_1 = 0$. See the notebook *randomwalk.ipynb*.

Theorem 13 *Let $\{Z_i\}_{i\geq 1}$ be iid random variables independent of $(X_j)_{j\geq 1}$, where the state at step n depends on a recurrent function such that*

$$X_{n+1} = f(X_n, Z_{n+1}) \tag{167}$$

This is typical of stochastic differential equations where Z_n is a kind of noise that induces random transitions. Then, $(X_1, \ldots, X_n)$ is a stationary Markov chain.

Proof 13. We need to show the absence of memory, and

$$\mathbb{P}(X_{n+1} = x_{n+1} | X_n = x_n, \ldots, X_1 = x_1) = \mathbb{P}(f(x_n, Z_{n+1}) = x_{n+1} | X_n = x_n, \ldots, X_1 = x_1) \tag{168}$$

Now, $(X_n, \ldots, X_1)$ depends on $(Z_n, \ldots, Z_1, X_1)$, and Z_{n+1} is independent of these variables, so

$$\begin{aligned}\mathbb{P}(X_{n+1} = x_{n+1} | X_n = x_n, \ldots, X_1 = x_1) &= \underbrace{\mathbb{P}(f(X_n = x_n, Z_{n+1}) = x_{n+1})}_{\text{independent of states } 1,\ldots,n-1} \\ &= \mathbb{P}(X_{n+1} = x_{n+1} | X_n = x_n)\end{aligned} \tag{169}$$

So, we indeed have a Markov process. For the process to be stationary, we need that

$\forall (x, y) \in \chi^2$

$$\mathbb{P}(X_{n+1} = y | X_n = x) = \mathbb{P}(X_n = y | X_{n-1} = x) \Leftrightarrow \mathbb{P}(f(x, Z_{n+1}) = y) = \mathbb{P}(f(x, Z_n) = y) \tag{170}$$

Now,

$$\mathbb{P}(f(x, Z_n) = y) = \int \mathbf{1}_{f(x,z)=y} \, d\mathbb{P}_{Z_n}(z) \tag{171}$$

but since the $(Z_i)_i$ are identically distributed, the right-hand side is independent of n. Thus, we have a stationary Markov chain. ■

An example of a random walk on $\mathbb{Z}$ with Z_n being Rademacher variables $\{+1, -1\}$ ($p(+1) = 1/2$) is given in Figure 20:

$$X_{n+1} = X_n + Z_{n+1} \tag{172}$$

You can think of a stochastic automaton as a Turing machine that changes state every time it reads a new symbol (instruction), and the transition depends on a random variable (hidden/unknown).

6.4.2.3 Ehrenfest's Urn

Another model closer to physics is the Ehrenfest's Urn[68] (1907), introduced by Afanasyeva and Paul, which models the diffusion of a perfect gas through a porous wall. It is illustrated in Figure 21. If initially, all balls are in urn A ($n_A(t = 0) = N$), what happens after some time t, knowing that at each discrete time step, a ball is randomly chosen and moved to the other urn: if it was in urn A, it goes to urn B and vice versa. The graph on the right (notebook *urne_Ehrenfest.ipynb*) confirms our everyday experience where we observe a homogenization into equal distributions in both urns. ***There is an irreversible phenomenon at play***. However, this doesn't mean that a ball cannot cross the barrier; on average, there are just as many going from $A \to B$ as from $B \to A$ after a "certain time". ***But this irreversibility behavior when N is small doesn't appear, as illustrated in Figure 22; there is a high likelihood of returning to the initial state! This is***

68. Paul Ehrenfest (1880-1933) made significant contributions, particularly in quantum mechanics and relativity, with his wife Afanasyeva and daughter Tatyana.

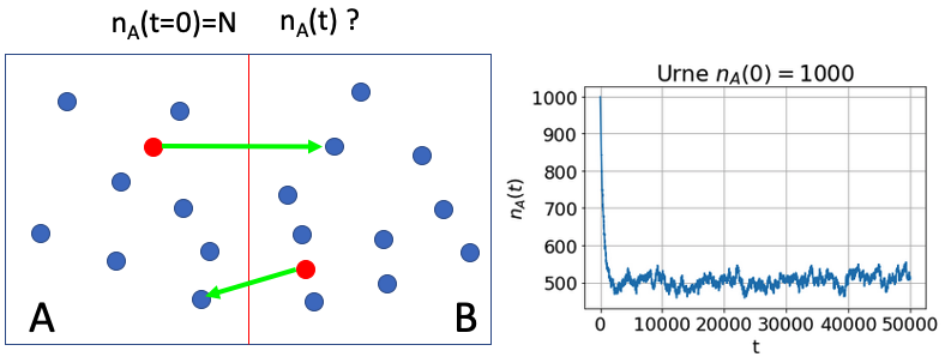


FIGURE 21 – Markov process called "Ehrenfest's urn" modeling a perfect gas diffusing through a porous wall. At each step, a ball is randomly selected and moved from one compartment $A \to B$ or $B \to A$. On the right, the number of balls in urn A after a certain time t. (notebook *urne_Ehrenfest.ipynb*)

extremely unlikely when $N \gg 1$**.** In fact, it can be shown that the average time between two returns to the initial state is $\langle \tau \rangle = 2^N$.

The modeling becomes straightforward if $X_n = n_A(t = n)$, and by defining $X_n = x$, then

$$X_{n+1} = \begin{cases} x - 1 & \text{prob. } \frac{x}{N} \\ x + 1 & \text{prob. } 1 - \frac{x}{N} \end{cases} = X_n + Z_{n+1} \tag{173}$$

where $Z_{n+1} \in \{-1, +1\}$, but it's not a Rademacher variable because

$$\mathbb{P}(Z_{n+1} = -1 | X_n = x) = \frac{x}{N} \tag{174}$$

so there's no independence of the Z_{n+1} distribution concerning the state X_n. Thus, while being a Markov chain, the process is not a random walk as defined in the previous section. It's interesting to see ***under what conditions equilibrium and irreversibility phenomena appear***.

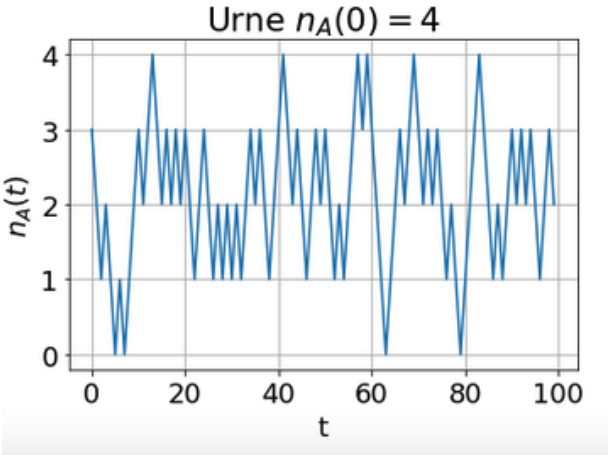


FIGURE 22 – Same experiment as Figure 21 but with $N = 4$. In this case, returning to the initial state is not negligible at all.

6.5 Invariant Law or Stationary Law: Equilibrium

In the context of Markov processes, as seen, for example, in Ehrenfest's urn, where probability distributions evolve over time (n), one might wonder if, after some time ($n \geq n_o$), there is an asymptotic form that makes the system invariant. Consider the following definition:

Definition 11 ***(invariant law)***
An invariant law, denoted as Π*, is a probability measure such that*[a] *(Prop. 3, Eqs. 162, 163) (*P *is the transition matrix* $(P_{x,y})_{(x,y)\in\chi^2}$*)*

$$\sum_{x\in\chi} \Pi(x) = 1 \qquad\qquad \Pi(y) = \sum_{x\in\chi} \Pi(x) P_{x,y} \tag{175}$$

In other words,

$$\Pi = P^T \Pi \tag{176}$$

meaning that Π *is an invariant under the action of the transition matrix; it is an eigenvector associated with the eigenvalue* $\lambda = 1$*.*

a. Note that this is as if we took the limit $n \to \infty$ in Equation 163.

For example, in the 2-state process ($\{x, y\}$) (Sec. 6.4.2.1), for reference

$$P_{x,y} = \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix} \tag{177}$$

whose transpose has 2 eigenvalues $(1, 1-\alpha-\beta)$. So, we do indeed have an invariant distribution ($\lambda = 1$), satisfying not only the unity condition but also

$$-\alpha\Pi(x) + \beta\Pi(y) = 0 \tag{178}$$

hence

$$\Pi(x) = \frac{\beta}{\alpha+\beta} \qquad \Pi(y) = \frac{\alpha}{\alpha+\beta} \tag{179}$$

In general, we have $|\chi|$ equations (the unity condition is redundant as P satisfies it as well) with $|\chi|$ unknowns. So, unless you end up in a degenerate case, there exists a unique invariant distribution.

In the case of the Ehrenfest's urn (Sec. 6.4.2.3), if x is the number of balls in box A, then the equation governing $\Pi(x)$ is given by

$$\Pi(x) = \Pi(x-1)P_{x-1,x} + \Pi(x+1)P_{x+1,x} \tag{180}$$

$$= \Pi(x-1)\left(1 - \frac{x-1}{N}\right) + \Pi(x+1)\frac{x+1}{N} \tag{181}$$

If we fix the condition $\Pi(0) = \Pi(1) \times \frac{1}{N}$, then it can be shown that

$$\Pi(x) = \binom{N}{x}\left(\frac{1}{2}\right)^N \tag{182}$$

This is a binomial distribution with probability $p = 1/2$. This confirms our intuition that at equilibrium, the balls are evenly distributed between the two boxes. See an illustration of this binomial distribution for $N = 10$ and $N = 50$ in Figure 23. When N is large, the distribution converges to a normal distribution $\mathcal{N}(\mu = N/2; \sigma^2 = N/4)$.

The probability measure Π ***is stable for the system; any transformation via*** P

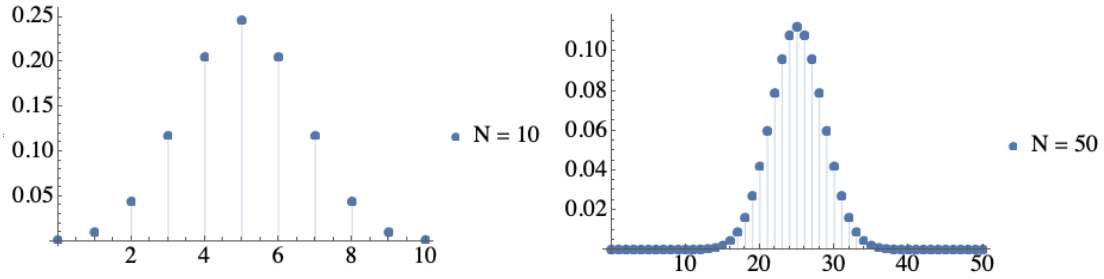


FIGURE 23 – Invariant distribution in the case where Ehrenfest's urn contains $N = 10$ or $N = 50$ balls.

leaves it unchanged. What happens when we start from an unbalanced situation? In the numerical simulations from the previous section (Sec. 6.4.2.3), we noticed that the system can, on average, evolve towards this stable situation, but there are particularly severe fluctuations when N is small. However, when N is large, we observe a kind of *irreversibility*, meaning there is very little (if any in practice) possibility of ending up in a state where all balls are in the same compartment. So, more generally, we are interested in *the system's dynamics to understand when and how it can converge to a stable situation.* The concept of the reversibility of laws while observing an irreversibility phenomenon was quite paradoxical in the early days of thermodynamics. And the concept of entropy (Carnot) emerged in connection with observations of physical systems. Let's see what this looks like from a mathematical perspective.

6.6 Stationary/Invariant Law and Reversibility

The underlying idea is that the existence of the invariant law depends on reversibility. But how do we define the reversibility of a Markov chain? For a Markov chain, we define the transition from X_n to X_{n+1}, so reversing time requires defining how we go from X_{n+1} to X_n.

Definition 12 *(**Inverse Matrix**)*
Consider a stationary process (Def. 10) with a matrix $P_{x,y} = \mathbb{P}(X_{n+1} = y | X_n = x)$ $(\forall (x, y) \in \chi^2)$*. If there exists a stationary law* $\Pi(x) \neq 0$*, meaning all* x *are accessible,*

then we define the following matrix $Q_{x,y}$*:*

$$Q_{x,y} = \frac{\Pi(y)P_{y,x}}{\Pi(x)} \tag{183}$$

Then Q is a **stochastic matrix** (Eq. 162). Indeed, since $\Pi = P^T\Pi$, then $\Pi_x = \sum_y \Pi_y P_{y,x}$, so $\sum_y Q_{x,y} = 1$. Moreover, Q is the transition matrix from X_{n+1} to X_n for the invariant distribution. Using Bayes' theorem, we have

$$\begin{aligned}\mathbb{P}(X_n = y|X_{n+1} = x) &= \frac{\mathbb{P}(X_{n+1} = x|X_n = y)\mathbb{P}(X_n = y)}{\mathbb{P}(X_{n+1} = x)} \\ &= P_{y,x}\frac{\mathbb{P}(X_n = y)}{\mathbb{P}(X_{n+1} = x)}\end{aligned} \tag{184}$$

If we have initialized the process with the invariant distribution Π, then

$$\mathbb{P}(X_n = y|X_{n+1} = x) = P_{y,x}\frac{\Pi(y)}{\Pi(x)} = Q_{x,y} \tag{185}$$

So, if the Markov chain is generated under the conditions where $X_1 \sim \Pi$, and we apply P successively to obtain $X_2, \ldots, X_n$, all the variables X_i follow the distribution law Π, and conversely, we can go back from X_n to $X_{n-1}, \ldots, X_1$ by applying Q. This can be summarized by this small diagram:

$$X_1(\Pi) \underset{Q}{\overset{P}{\rightleftarrows}} X_N(\Pi)$$

Note that the law Π is such that

$$\Pi = P^T\Pi = Q^T\Pi \tag{186}$$

but in any case, it is not necessary to conclude that $P = Q$. Hence the following definition, which specifies the framework of interest for reversible physical laws (except weak interactions, for example [69])

69. NDJE. The laws of the Standard Model conserve CPT with very high precision (C charge conjugation, P parity, and T time reversal, but since the 1964 experiments on neutral Kaons, we know that weak interactions break CP invariance and thus T invariance.

Definition 13 *(**Reversible Markov Chain**)*
We say that a stationary Markov chain with transition matrix P is reversible relative to an invariant probability Π, if, under the conditions of Definition 12, $Q = P$, which translates to [a]

$$\Pi(x)P_{x,y} = \Pi(y)P_{y,x} \qquad (\textit{detailed balance}) \tag{187}$$

(Note that the global balance is given by the equation $\Pi = P^T\Pi$) which is represented by the small diagram:

$$X_1(\Pi) = x \underset{P_{y,x}}{\overset{P_{x,y}}{\rightleftharpoons}} X_2(\Pi) = y$$

a. NDJE. J. C. Maxwell introduced this notion in 1867 in his study of gas kinetics (*principle of sufficient reason*).

The following property is particularly interesting in practice

Property 4 *(**Existence of an Invariant Measure**)*
If a stationary Markov process with transition probability P satisfies a detailed balance, then Π is an invariant measure (i.e., a stationary law).

The proof is simple; we know that $\forall x, y$, $\Pi(x)P_{x,y} = \Pi(y)P_{y,x}$, summing over y on both sides

$$\sum_y \Pi(y)P_{y,x} = \sum_y \Pi(x)P_{x,y} = \Pi(x) \underbrace{\sum_y P_{x,y}}_{=1 \text{ (stochastic matrix)}} = \Pi(x) \tag{188}$$

which is nothing else than $\Pi = P^T\Pi$, so Π is indeed invariant (Prop. 11).

Random walks and Ehrenfest's urn are examples of reversible processes. In the next session, we will see how the ergodicity hypothesis ensures that all states are explored and that the chain does not remain stuck in a recurrent state. Thus, this will ensure convergence to the invariant measure. So, with these properties in hand, we will have access to the entropy rate, the dynamics of entropy, and ultimately the evolution towards maximum entropy.

6.7 NDJE. Stochastic Matrix, Stationary Law, and Reversibility

We have seen that a stochastic matrix P satisfies by definition $\forall x$, $\sum_y P_{x,y} = 1$ (Eq. 162).

- If $P = P^T$, this necessarily implies the existence of a uniform stationary law Π, and the process is reversible. Indeed, if $\Pi(y) = C \neq 0$ for all y, $\Pi(x) = C = \sum_y P_{y,x}\Pi(y)$ gives $\sum_y P_{x,y} = 1$, which is true.
- If we have a stochastic matrix P that additionally satisfies $\sum_y P_{y,x} = 1$ for all x, we call it *bi-stochastic*, and by revisiting the previous proof, we show that the uniform law is stationary. The difference is that the process is not necessarily reversible [70], because in this case, we can have $Q = P^T \neq P$. An example of such a situation is given by a process governed by the following matrices P and Q:

$$P = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \qquad Q = P^T = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \neq P \tag{189}$$

which can be represented by the diagram in Fig. 24.

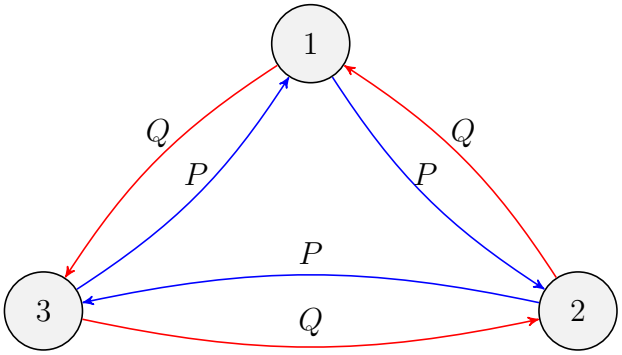


FIGURE 24 – Markov process with *forward* transition matrices P and *backward* transition matrices Q.

70. An example of a 3-state bi-stochastic case where $Q = P^T$ and thus reversible is, for instance:

$$P = \begin{pmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 1/2 & 1/2 \end{pmatrix}$$

.

7. Lecture 22 Feb.

In this section, we will study the dynamics of a system, represented as a Markov chain, as it tends towards equilibrium, where entropy reaches its maximum. Recall that in this context (Definition 9), the system's future is conditioned only on its present state, meaning it does not depend on its past states. Typically, we encounter this behavior when discretizing differential equations where only the first derivative with respect to time is involved. Additionally, we will consider *stationary chains* (Def. 10), where *the transition probability from X_n to X_{n+1} does not depend on "time"*. In this case, we have a potentially infinite matrix [71] denoted as $(P_{x,y})_{(x,y)\in\chi^2}$ (hereafter referred to as P).

As a reminder, $P_{x,y}$ *is a stochastic matrix* that satisfies a normalization relation over y (Eq. 162). This matrix governs the dynamics of the chain generation. We have seen that if we denote $\mu_n = (\mathbb{P}(X_n = x))_{x\in\chi}$, then (Eq. 164)

$$\mu_{n+1} = P^T \mu_n$$

At equilibrium (if it exists), there is an *invariant measure* Π that satisfies the equation [72] (Definition 11)

$$\Pi = P^T \Pi$$

This is an eigenvector of P^T with eigenvalue $\lambda = 1$. This invariant measure tells us that if at the initial time $t = 1$, the system, for example, the set of velocities of gas particles, is such that $X_1 \sim \Pi$, then at time $t = n$, $X_n \sim \Pi$ as well, which is the definition of an equilibrium state.

The concept of *reversibility* led us to consider the matrix Q (Definition 12) defined from P and the invariant measure Π as follows:

$$Q_{x,y} = P_{y,x}\frac{\Pi(y)}{\Pi(x)}$$

It reverses the system's evolution (time reversal). In the scenario where a probability

71. Mnemonic: *"x influences y"*.
72. NDJE. One can also use the transpose $\Pi^T = \Pi^T P$.

distribution satisfies the so-called *detailed balance* relation (Definition 13, Eq. 187)

$$\Pi(x)P_{x,y} = \Pi(y)P_{y,x}$$

then Π is an invariant measure. This can be formulated as *"there are as many states transitioning from x to y as there are states transitioning from y to x"*.

7.1 Random Walk on an Undirected Graph

Let's consider undirected graphs (Fig. 25) where, by definition, transitions $x \to y$ have the same weight as transitions $y \to x$. These graphs can represent transition probabilities of stationary Markov chains.

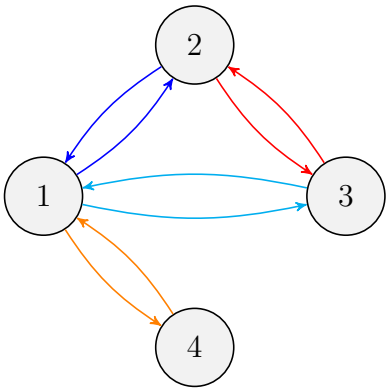
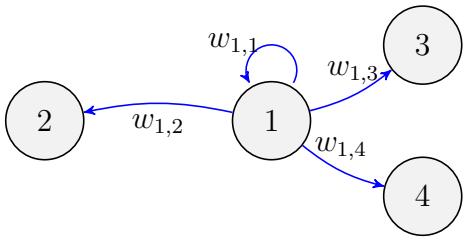


FIGURE 25 – Example of an undirected graph where, for any pair (x, y), the transition $x \to y$ has the same weight as the transition $y \to x$.

The elements of the matrix P are defined as

$$P_{x,y} = \frac{W_{x,y}}{\sum_z W_{x,z}} \tag{190}$$

which means we normalize by the sum of weights of arrows leaving from x, as shown in the diagram below:



What about the invariant measure? Let's follow this reasoning: Consider the sum of weights of all arrows pointing to x:

$$W_x := \sum_y W_{y,x} \tag{191}$$

This represents a kind of measure of the attraction to x. Normalize it by $\sum_x W_x = \sum_{x,y} W_{x,y} := W$ to obtain a probability, and then define

$$\Pi(x) = \frac{W_x}{W} \tag{192}$$

Then, we observe that

$$\Pi(x)P_{x,y} = \frac{\sum_z W_{z,x}}{W} \times \frac{W_{x,y}}{\sum_z W_{x,z}} \tag{193}$$

Now, for any pair (x, z), $W_{z,x} = W_{x,z}$, so we have

$$\Pi(x)P_{x,y} = \frac{W_{x,y}}{W} = \frac{W_{y,x}}{W} = \Pi(y)P_{y,x} \tag{194}$$

Thus, Π satisfies detailed balance, making it an invariant measure.

This particular case of undirected graphs that yield reversible Markov chains is the one commonly encountered in physics and is of primary interest. However, one might wonder if the stationary law (invariant measure) always exists. In the general case, we have Shannon–McMillan–Breiman theorem (Th. 12) that ensures almost sure convergence of $-\frac{1}{n}\log \mathbb{P}(X_1, \ldots, X_n)$ and leads to the phenomenon of concentration on typical sets. But let's see how ergodicity comes into play in the Markovian framework (Birkhoff, Th. 11).

7.2 Ergodicity and Markov Chain

Recall that in the context of an ergodic process governed by a transformation law T, we reach state X_n by applying the transformation T to state X_1 a total of n times, and more generally, $X_{n+k} = T^k X_n$. By Birkhoff's theorem, this assures us that (Eq. 153)

$$(X_1, \dots, X_n) \text{ ergodic} \quad \Rightarrow \quad \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{p.s} \mathbb{E}[X]$$

Now, let's introduce the following definitions that will allow us to ensure the existence of the stationary law.

Definition 14 *(**Irreducible Chain**)*
A Markov chain is irreducible if we can transition from any state x to any other state y, that is

$$\forall (x, y) \in \chi^2, P_{x,y} > 0 \tag{195}$$

Note that there may not necessarily be a direct link between x and y. For example, in the scheme 25, state "4" has no direct connection to states "2" and "3," yet the chain is irreducible. If equation 154 gives us the residence time of a Markov chain in a set A, let's define here the *first return time* to any state.

Definition 15 *(**First Return Time**)*
For a state $X_1 = x$, we define the "time" T_x

$$T_x = \inf\{n; X_n = x\} \tag{196}$$

if we cannot reach x, then $T_x = +\infty$.

NDJE. One can get an idea, for example: we randomly draw a state $x_1 \in \chi$ at time $t = 1$ according to the distribution of X_1; then to determine the new state x_2 at time $t = 2$, we randomly draw according to the probabilities $(P_{x_1,x})_{x\in\chi}$ of the transition matrix P, and so on. We note when we find the initial state again, let's call it $n^{(1)}$. We repeat this exercise K

times, and T_x is the minimum time among the $(n^{(k)})_{k\leq K}$, or rather, the infimum obtained as we consider $K = +\infty$.

If there is a finite number of states, and all are accessible from x_1, then T_x is finite. But in the case of an infinite number of states, the question arises. Note that T_x is itself a *random variable*. Even if we always start from the same initial state, with each new evolution of the Markov chain, the successive intermediate states are randomly drawn according to P before returning to x_1. In other words, there are several possible paths that start from x_1 and return to it. So, consider the following definition.

Definition 16 *(**Positive Recurrent State**)*
A state x is said to be positive recurrent if the expected value of its return time is finite [a]

$$\mathbb{E}_x[T_x] < \infty \tag{197}$$

a. NDJE. keeping in mind $\mathbb{P}(X_1 = x) = 1$, i.e., that we always begin the chain's evolutions from state x.

This means that in the case of an irreducible chain, starting from state x, we can return to it an arbitrarily large number of times with a frequency of $1/T_x$, which measures the attractiveness of state x. However, this attractiveness changes depending on x; some states attract more arrows than others. In the case of the random walk, this is the underlying philosophy behind the definition of W_x. It's tempting to wonder if the stationary law is related to this notion of return time.

Theorem 14 *(**Ergodicity**)*
Let (X_n) be a Markov chain, irreducible and positively recurrent. Then,

1. *considering the average of the number of times $X_k = x$*

$$\frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_{X_k=x} \xrightarrow[n\to\infty]{p.s} \Pi(x) > 0 \tag{198}$$

   where $\Pi(x)$ is an invariant measure;

2. *the previous invariant measure is the unique invariant measure;*
3. *for any function f such that $\sum_x |f(x)|\Pi(x) < \infty$, Birkhoff's theorem tells us*

that

$$\frac{1}{n}\sum_{k=1}^{n} f(X_k) \xrightarrow[n\to\infty]{p.s} \sum_x |f(x)|\Pi(x) = \mathbb{E}_{X\sim\Pi}[f(X)] \tag{199}$$

If we start from x_1 at $t = 1$ by drawing X_1 according to its distribution, we calculate $f(x)$, then we transition the state from $t = 1$ to $t = 2$ and calculate $f(x_2)$, and so on until $f(x_n)$, then we calculate the empirical average of $(f(x_i))_{i\leq n}$. When n tends to infinity, the result is identical to averaging $f(x)$ if we draw x according to the invariant law Π. In other words, temporal averaging is identical to ensemble averaging. In fact, if we add that the chain is aperiodic, then any starting measure converges to the invariant law.

The proof of this theorem is left for another course. It is important because if we have a process described by the evolution of a Markov chain, under the two assumptions of irreducibility and positive recurrence, then there exists an equilibrium state to which the process converges (in distribution). This provides access to the equilibrium entropy.

7.3 Entropy of an Equilibrium Markov Chain

We will revisit concepts of ***entropy rates*** discussed in sections 5.4.1, 5.4.2, and in particular Theorem 9 applicable to stationary processes:

$$\mathbb{H}(\chi) = \mathbb{H}'(\chi) = \lim_{n\to\infty} \mathbb{H}(X_n|X_{n-1},\ldots,X_1) \tag{200}$$

In the case of *Markov chains*, the problem simplifies because conditional entropy reduces to $\mathbb{H}(X_n|X_{n-1})$. Furthermore, in the *stationary case*, this entropy does not depend on n, it is equal, for example, to $\mathbb{H}(X_2|X_1)$. At equilibrium, X_2 and X_1 have the same distribution, namely Π. Therefore, according to the definition of conditional entropy [73] (Definition 3), we have:

$$\mathbb{H}^{(\prime)}(\chi) = \mathbb{H}(X_2|X_1) = \sum_{x\in\chi} \mathbb{P}(X_1 = x)\mathbb{H}(X_2|X_1 = x) \tag{201}$$

73. NDJE. there may be changes in notation along the way.

Now, at equilibrium $\mathbb{P}(X_1 = x) = \Pi(x)$. So, considering that:

$$\begin{aligned}\mathbb{H}(X_2|X_1 = x) &= -\sum_{y\in\chi} \mathbb{P}(X_2 = y|X_1 = x) \log \mathbb{P}(X_2 = y|X_1 = x) \\ &= -\sum_{y} P_{x,y} \log P_{x,y} \end{aligned} \tag{202}$$

It follows:

$$\mathbb{H}^{(\prime)}(\chi) = -\sum_{(x,y)\in\chi} \Pi(x) P_{x,y} \log P_{x,y} \tag{203}$$

Thus, with a stationary equilibrium Markov chain, we can explicitly calculate everything. The next question is how the process evolves when it's not at equilibrium, and where does the Second Law of Thermodynamics come from?

7.4 Markov Chain and the 2nd Law of Thermodynamics

As a preamble, let's recall that if we consider Markov processes, this framework is well-suited for physical processes.

To recap, Boltzmann's perspective (Sec. 4.1) led to considering the entropy of a system with fixed energy (at equilibrium) as the logarithm of the number of accessible microstates, because the probability is uniform. What we will discover is that indeed, ***when the equilibrium probability distribution is uniform, entropy increases towards its maximum***. However, we will also realize that ***if the equilibrium probability distribution is not uniform, then there is no increase in entropy***. So, be cautious.

Consider, at time n, two distributions $\mu_n = (\mathbb{P}(X_n = x))_{x\in\chi}$ and μ'_n. What is the evolution of these distributions when they are governed by the same "transition probability matrix" $(P_{x,y})_{(x,y)\in\chi^2}$? So, at time n, we have (μ_n, μ'_n), and at time $n+1$, we have (μ_{n+1}, μ'_{n+1}). It is interesting to compare the evolution of the Kullback-Leibler divergence between these distributions (Definition 4). We know that the distance is non-negative. What we will show is that the distance decreases as a function of n.

Theorem 15 *(**convergence of distributions**)*
If $\mu_n \xrightarrow{P} \mu_{n+1}$ *and* $\mu'_n \xrightarrow{P} \mu'_{n+1}$ *then*

$$D_{KL}(\mu_n \| \mu'_n) \geq D_{KL}(\mu_{n+1} \| \mu'_{n+1}) \tag{204}$$

Proof 15.
If we consider two random variables X_n and X_{n+1}, we can evaluate their joint probability, denoted here [74] as $p(X_n, X_{n+1})$ in the case of the evolution of μ, and $q(X_n, X_{n+1})$ in the case of μ'. What about their divergence?

$$\begin{aligned}
D_{KL}(p(X_n, X_{n+1}) \| q(X_n, X_{n+1})) &= \sum_{x,y} p(X_n = x, X_{n+1} = y) \log \frac{p(X_n = x, X_{n+1} = y)}{q(X_n = x, X_{n+1} = y)} \\
&= \sum_{x,y} p(X_n = x) p(X_{n+1} = y | X_n = x) \log \frac{p(X_n = x) p(X_{n+1} = y | X_n = x)}{q(X_n = x) q(X_{n+1} = y | X_n = x)} \\
&= \sum_{x,y} p(X_n = x) p(X_{n+1} = y | X_n = x) \log \frac{p(X_n = x)}{q(X_n = x)} \\
&\quad + \sum_{x,y} p(X_n = x) p(X_{n+1} = y | X_n = x) \log \frac{p(X_{n+1} = y | X_n = x)}{q(X_{n+1} = y | X_n = x)} \\
&= \sum_{x} p(X_n = x) \log \frac{p(X_n = x)}{q(X_n = x)} \overbrace{\sum_{y} p(X_{n+1} = y | X_n = x)}^{=1} \\
&\quad + \sum_{x} p(X_n = x) \sum_{y} p(X_{n+1} = y | X_n = x) \log \frac{p(X_{n+1} = y | X_n = x)}{q(X_{n+1} = y | X_n = x)}
\end{aligned}$$

The first term is none other than $D_{KL}(p(X_n) \| q(X_n))$. As for the second term, $p(X_{n+1} = y | X_n = x) = q(X_{n+1} = y | X_n = x) = P_{x,y}$, because it's the same (forward) transition probability matrix governing the evolution of both chains, so the right-hand second term is zero. Thus,

$$D_{KL}(p(X_n, X_{n+1}) \| q(X_n, X_{n+1})) = D_{KL}(p(X_n) \| q(X_n)) \tag{205}$$

74. NDJE. we are trying to simplify notations...

We could equally write

$$\begin{aligned}
&D_{KL}(p(X_n, X_{n+1})\|q(X_n, X_{n+1})) \\
&\quad = \sum_{xy} p(X_{n+1}=x)p(X_n=y|X_{n+1}=x)\log\frac{p(X_{n+1}=x)p(X_n=y|X_{n+1}=x)}{q(X_{n+1}=x)q(X_n=y|X_{n+1}=x)} \\
&\quad = D_{KL}(p(X_{n+1})\|q(X_{n+1})) \\
&\qquad + \sum_x p(X_{n+1}=x)\sum_y p(X_n=y|X_{n+1}=x)\log\frac{p(X_n=y|X_{n+1}=x)}{q(X_n=y|X_{n+1}=x)}
\end{aligned}$$

Now, $p(X_n = y|X_{n+1} = x)$ and $q(X_n = y|X_{n+1} = x)$ are governed by *backward* transition matrices, which might not be the same (we are not at equilibrium), so the right-hand second term is not necessarily zero. However, this term can be written as the expectation of $D_{KL}(p(X_n = y|X_{n+1} = x)\|q(X_n = y|X_{n+1} = x))$, so this term is non-negative. Thus,

$$D_{KL}(p(X_n)\|q(X_n)) \geq D_{KL}(p(X_{n+1})\|q(X_{n+1})) \tag{206}$$

■

So, the evolution of chains with the same forward transition probabilities brings their probability distributions closer together over time.

A special case of the theorem is when ***one of the two distributions is the invariant measure***, e.g., $\forall n, \mu'_n = \Pi$. Then,

$$\forall n, \quad 0 \leq D_{KL}(\mu_{n+1}\|\Pi) \leq D_{KL}(\mu_n\|\Pi) \tag{207}$$

An even more specific case is when ***the invariant measure is the uniform distribution***, meaning $\Pi(x)$ is independent of $x \in \chi$, and it's equal to $1/|\chi|$. Then,

$$D_{KL}(\mu_n\|\Pi) = \sum_x p(X_n = x)\log\frac{p(X_n = x)}{\Pi(x)} = -\mathbb{H}[\mu_n] - \log|\chi| \tag{208}$$

and thus,

Theorem 16 *(**entropy growth**)*
*If $\mu_n \xrightarrow{P} \mu_{n+1}$, and there exists a **uniform invariant measure**, then*

$$\mathbb{H}[\mu_{n+1}] \geq \mathbb{H}[\mu_n] \tag{209}$$

In cases where Π is indeed ***an invariant measure but not uniform***, then the sequence of divergences $D_{KL}(\mu_n \| \Pi)$ will converge (a decreasing positive sequence) but ***the entropy of μ_∞ is not maximal***. So, one may ask: under what conditions is the invariant distribution uniform? To answer this, we need to provide additional characteristics to the Markov process.

Definition 17 *(**bi-stochastic matrix**)*
Let $P = (P_{x,y})_{(x,y)\in\chi^2}$, it is bi-stochastic if and only if

$$\sum_y P_{x,y} = 1 \quad \text{and} \quad \sum_x P_{x,y} = 1 \tag{210}$$

So, we have the following properties:

Property 5

1) If the invariant measure Π is uniform, then P is bi-stochastic.

2) If P is bi-stochastic, then the uniform measure is invariant.

Indeed, for 1), P is stochastic by definition, and moreover:

$$\Pi = P^T\Pi \Rightarrow \Pi(y) = \frac{1}{|\chi|} = \sum_x \Pi(x)P_{x,y} = \frac{1}{|\chi|}\sum_x P_{x,y} \tag{211}$$

So, $\sum_x P_{x,y} = 1$. For 2), we just refer to the previous expression to conclude immediately.

We have a mathematical framework that consolidates the intuition from physics. It tells us that entropy increases as time progresses. If the system is of a Markovian type, then at each step of the time evolution, conditioned on state n, state $n+1$ becomes more uncertain. The distributions approach the invariant distribution that spreads probabilities

over available microstates. The attained entropy is maximum if the invariant distribution is uniform (i.e., maximum spreading). And the tool that served us is the relative entropy.

Now, we will return to modeling high-dimensional data, with the idea that if this data concentrates on surfaces, there is likely an underlying entropy dynamics at play. However, we will return to Statistical Physics with a "canonical" perspective to get an idea. We will end up with richer distributions than the uniform distribution.

7.5 Macrocanonical Ensemble

This concept was introduced by J. W. Gibbs in his 1901 treatise (see note 5). He considers a system $\mathscr{S}$ in contact with a large reservoir $\mathscr{R}$, with no exchange of matter (assumption of very weak interaction). Typically, the reservoir serves to fix the temperature. Thus, there are energy exchanges between $\mathscr{S}$ and $\mathscr{R}$. Additionally, we consider $\mathscr{T} = \mathscr{S} + \mathscr{R}$ to be completely isolated and at equilibrium. Therefore [75],

$$U_{\mathscr{T}} = \text{Const} = U_{\mathscr{R}} + U \tag{212}$$

If we are interested in energy exchanges δU between $\mathscr{S}$ and $\mathscr{R}$, there will be fluctuations that depend on the temperature [76] T. At equilibrium, the total entropy is maximized and decomposes as (see Definition 1),

$$\mathbb{H}[\mathscr{T}] = \mathbb{H}[\mathscr{R}] + \mathbb{H}, \quad d\mathbb{H}[\mathscr{T}] = 0 \quad \text{(equilibrium)} \tag{213}$$

and all else being equal, such as volume V, number of particles (d) (see Sec. 4.1.2), we have [77],

$$d\mathbb{H}[\mathscr{T}] = \left(\frac{\partial \mathbb{H}[\mathscr{R}]}{\partial U_{\mathscr{R}}}\right)_{x_{\mathscr{R}}}(U_{\mathscr{R}})\ dU_{\mathscr{R}} + \left(\frac{\partial \mathbb{H}}{\partial U}\right)_{x}(U)\ dU \tag{214}$$

75. We drop the index $\mathscr{R}$ for the notation of quantities related to the "small" system.

76. Recall that in the microcanonical case, T is related to the derivative of entropy with respect to the system's energy.

77. NDJE. I denote $\left(\frac{\partial f}{\partial x}\right)_z(x_o)$ as the partial derivative of $f(x, z)$ with respect to x, holding z fixed and evaluated at (x_o, z).

Since $dU_{\mathscr{R}} = -dU$, it follows that

$$\left(\frac{\partial \mathbb{H}[\mathscr{R}]}{\partial U_{\mathscr{R}}}\right)(U_{\mathscr{R}}) = \left(\frac{\partial \mathbb{H}}{\partial U}\right)(U) \tag{215}$$

However, there is a difference compared to the equilibrium situation described in Section 4.1.2: we have a very large reservoir and a relatively small system. In particular, $dU_{\mathscr{T}} \approx dU_{\mathscr{R}} \gg U$. Thus,

$$\left(\frac{\partial \mathbb{H}[\mathscr{R}]}{\partial U_{\mathscr{R}}}\right)(U_{\mathscr{R}}) \approx \left(\frac{\partial \mathbb{H}[\mathscr{T}]}{\partial U_{\mathscr{T}}}\right)(U_{\mathscr{T}}) = \frac{1}{T} \tag{216}$$

which defines the *macrocanonical temperature*. How does this condition the states of the (small) system?

We know that the total system is in a microcanonical equilibrium, so all accessible possible states are equiprobable. Among all these states, we count those that fix the energy of the small system to the value U_m, and thus fix the energy of the reservoir to the value $U_{\mathscr{R}} = U_{\mathscr{T}} - U_m$. Let $\Omega_{\mathscr{R}}(U_{\mathscr{T}} - U_m)$ be the set of reservoir states ($|\Omega|$ being the cardinality), $\Omega(U_m)$ be the set of states of the small system, and $\Omega_{\mathscr{T}}(U_{\mathscr{T}})$ be the set of states for the entire system. The probability that the small system has an energy (microcanonical) U_m is given by

$$\mathbb{P}(U_m) = \frac{|\Omega(U_m)||\Omega_{\mathscr{R}}(U_{\mathscr{T}} - U_m)|}{|\Omega_{\mathscr{T}}(U_{\mathscr{T}})|} \tag{217}$$

Now, the micro entropy $\mathbb{H}$ of a system is proportional to $\log|\Omega|$, and we know that it is an extensive quantity proportional to the number of particles present in the system. The reservoir assumption thus makes its statistical weight much greater than that of the small system. Consequently, $\log|\Omega(U_m)|$ is negligible, and $\log|\Omega_{\mathscr{R}}| \approx \log|\Omega_{\mathscr{T}}|$, so [78]

$$\begin{aligned} \log \mathbb{P}(U_m) &\approx \mathbb{H}_{\mathscr{T}}(U_{\mathscr{T}} - U_m) - \mathbb{H}_{\mathscr{T}}(U_{\mathscr{T}}) + \text{Cte} \\ &= -U_m \left(\frac{\partial \mathbb{H}_{\mathscr{T}}}{\partial U_{\mathscr{T}}}\right)(U_{\mathscr{T}}) + \text{Cte} + \dots \end{aligned} \tag{218}$$

78. NDJE. The ... primarily represent a term in U_m/CT with C being the heat capacity of the reservoir, which is very large by nature.

The first term reveals the inverse of the canonical temperature $\beta = 1/T$, so [79]

$$\mathbb{P}(U_m) \approx Z^{-1}\ e^{-\beta U_m} \tag{219}$$

We find the *Maxwell-Boltzmann distribution* in the case of particle velocities in an ideal gas, and Z is the normalization constant that Gibbs calls the *partition function*:

$$Z = \sum_{\substack{\text{states} \\ U=U_m}} e^{-\beta U_m} \tag{220}$$

Therefore, the energy fluctuations (and thus exchanges with the reservoir) of the small system depend on the temperature, that is, the variation of entropy of the reservoir with respect to energy, which does not fluctuate.

This is the perspective of Statistical Physics; let's now explore the mathematical perspective to understand where these exponential laws come from.

7.6 Principle of Maximum Entropy

This is a work undertaken by Edwin Thompson Jaynes (1922-98) in 1957, where he introduced a principle to attempt to reformulate Statistical Mechanics, especially with the aim of addressing out-of-equilibrium problems. Jaynes placed himself in the context of Claude Shannon's Information Theory, where entropy is a concept detached from any physical system. The problem posed is: when we have observations $\phi_k(x)$ where x represents an image, a sound, etc., and ϕ_k can be any function we wish to calculate from x, what can be said about the average/expectation?

$$M_n(\phi_k(x)) = \frac{1}{n}\sum_{i=1}^{n} \phi_k(x_i) \tag{221}$$

We know that if the $(x_i)_i$ are *iid random variables*, then the law of large numbers tells us that M_n converges to the expectation. However, what is the value of $p(x)$, i.e., the (probability) density of x? We would like to construct a model.

79. NDJE. You should interpret T as $k_B T$ if you want to place it in the context of thermodynamics.

Laplace's perspective [80] would be to roughly say: when we know nothing, we can consider the observations as equally probable. This can be understood in a finite discrete case, but what about an infinite continuous case? The uniform distribution over $\mathbb{R}$ poses a problem (a distribution), and the empirical mean does not converge. So, how do we generalize this "intuitive" idea? In fact, the key is to replace *uniformity of probabilities* with *maximizing uncertainty*. In other words, we will try to find the probability $p(x)$ that spreads out the most, or in other words, imposes the least underlying constraints outside of the observations. Hence, the emergence of entropy.

In fact, if we impose that for two independent random variables

$$\mathbb{H}(X, Y) = \mathbb{H}(X) + \mathbb{H}(Y) \tag{222}$$

and that $\mathbb{H}(X) \geq 0$ (in the case of values in an alphabet), Claude Shannon in his 1948 article shows [81] that entropy is, up to a constant, equal to

$$\mathbb{H} = -\sum_i p_i \log p_i \tag{223}$$

We can formulate the problem as follows: what is the probability $p(x)$ such that

1. we have a series of observed empirical means

$$M_n(\phi_k(x)) = \frac{1}{n}\sum_{i=1}^{n} \phi_k(x_i) \tag{224}$$

2. and ***the entropy***, defined as

$$\mathbb{H}[p] = \begin{cases} -\sum_x p(x) \log p(x) & \text{(discrete)} \\ -\int p(x) \log p(x)\ dx & \text{(continuous)} \end{cases} \tag{225}$$

   is ***maximum***.

80. NDJE. "The probability of the existence of an event is simply the ratio of the number of favorable cases to that of all possible cases when we see no reason why one of these cases would occur rather than the other." Extracted from Gérard Jorland https://doi.org/10.4000/ccrh.2772.

81. See the comment in Course 2022 Sec. 6.3

Now, if $x \sim p$, the expectation of $\phi_k(x)$ is simply

$$\mathbb{E}_{x\sim p}[\phi_k(x)] = \int \phi_k(x)p(x)dx \tag{226}$$

Therefore, we will consider the following problem:

Theorem 17 *(**Gibbs**)*
If there exists a probability distribution p such that

$$(\mu_k = \mathbb{E}_{x\sim p}[\phi_k(x)])_{k\leq K} \text{ (constraints)}, \quad p = \underset{p}{\operatorname{argmax}}\ \mathbb{H}[p] \tag{227}$$

then there exist K parameters $\Theta = (\theta_k)_{k\leq K}$ such that

$$p_\Theta(x) = Z_\Theta^{-1} \exp\left\{\sum_k \theta_k\phi_k(x)\right\} = Z_\Theta^{-1}\ e^{\Theta^T\Phi(x)} \tag{228}$$

*with $\Phi(x) = (\phi_k(x))_{k\leq K}$. The constant Z_Θ, **the Gibbs partition function**, ensures the normalization of the probability and has the expression*

$$Z_\Theta = \int e^{\Theta^T\Phi(x)}dx \tag{229}$$

Note that then

$$\mu_k = \mathbb{E}_{x\sim p_\Theta}[\phi_k(x)])_{k\leq K} = \frac{\partial \log Z(\Theta)}{\partial\theta_k} \tag{230}$$

making Z a key piece for the calculation of thermodynamic quantities.

We find an exponential distribution similar to the Maxwell-Boltzmann distribution. Note that in this context

$$U(x) = \Theta^T\Phi(x) \tag{231}$$

and the "temperature" appears as a Lagrange multiplier (up to sign) (e.g., θ_0) that can be factored out, giving rise to the formulation of the argument of the exponential $U(x)/T$ (with the appropriate sign according to the convention).

Proof 17.
We need to maximize a concave function $(-p\log p)$, or equivalently, minimize a convex function, under linear constraints. If a solution exists, it is unique, and it is obtained using

Lagrange multipliers [82]. We then study the function (in the discrete case)

$$\mathcal{L}(\Theta, p) = -\sum_x p(x)\log p(x) + \sum_{k=1}^{K}\theta_k\left(\sum_x p(x)\phi_k(x) - \mu_k\right) + \theta_0\left(\sum_x p(x) - 1\right) \quad (232)$$

The variable(s) are the values of $p(x_i)$, and we then study the derivative (in the continuous case, this would be a functional derivative)

$$\forall x_i, \quad \frac{\partial\mathcal{L}}{\partial p(x_i)} = -1 - \log(p(x_i)) + \sum_k \theta_k\phi_k(x_i) + \theta_0 = 0 \quad (233)$$

From which, for all x, the solution, denoted as p_Θ, is written as

$$p_\Theta(x) = Z_\Theta^{-1}\exp\left\{\sum_k \theta_k\phi_k(x)\right\} \quad (234)$$

But what about the adequacy between p_Θ and the true distribution p? In fact, we have

$$D_{KL}(p\|p_\Theta) = \mathbb{H}[p_\Theta] - \mathbb{H}[p] \geq 0 \quad (235)$$

Indeed,

$$\begin{aligned} D_{KL}(p\|p_\Theta) &= \sum_x p(x)\log\frac{p(x)}{p_\Theta(x)} = -\mathbb{H}[p] - \sum_x p(x)\left(-\log Z + \sum_k \theta_k\phi_k(x)\right) \\ &= -\mathbb{H}[p] + \log Z - \sum_k \theta_k \underbrace{\mathbb{E}_{x\sim p}\left[\phi_k(x)\right]}_{\mu_k} \end{aligned} \quad (236)$$

But we have also imposed that $\mu_k = \mathbb{E}_{x\sim p_\Theta}\left[\phi_k(x)\right]$, so we have

$$D_{KL}(p\|p_\Theta) = -\mathbb{H}[p] + \mathbb{E}_{x\sim p_\Theta}[-\log p_\Theta(x)] = \mathbb{H}[p_\Theta] - \mathbb{H}[p] \quad (237)$$

■

Therefore, having data, we build probabilistic (parameterized) models following Fisher's philosophy (Course 2022 and others). However, p_Θ is not equal to $p(x)$, which we

82. Course 2018 Sec. 8.3

don't know either. So in practice, we define the best model using a guiding principle: 1) it satisfies the observational constraints (the (μ_k)), and 2) the ***principle of maximum entropy***. We have a freedom: what are the measures $(\phi_k(x))_k$ that we take/calculate to constrain the model? For example, if we take means and second-order moments, we typically have a Gaussian model, but we can choose more complex functions, etc. What should we choose? We will choose what reduces uncertainty the most, i.e., $\mathbb{H}[p_\Theta]$, to get as close as possible to $\mathbb{H}[p]$.

So, in the choice of the optimal model, we want to find the measures $(\phi_k(x))_k$ ***that minimize the maximum entropy*** $\mathbb{H}[p_\Theta]$**.** This is a minimax problem. In the case of image processing (see Valentin de Portoli's seminar in 2023), textures' images can be seen as a random process, and we need to establish a stochastic model. What are the measurements we need to consider to establish good models? 20-25 years ago, says Stéphane Mallat, they tried Gaussian models of the form $U(x) = x^T K x$ where $\phi_{k=(i,j)}(x) = x_i x_j$. It's simple, but not sufficient for textures with structure. We can extend to the use of *feature mapping* (Course 2018 Sec. 7.3), but most importantly, what changed the game are neural networks that provide the $\Phi(x)$, i.e., the optimal representations. But what are these $\Phi(x)$? Don't we want to understand what CNNs learn? Don't we have prior information that would be good to put into these features (a theme of Course 2020, among others)? One of the challenges is ***understanding the interactions between different scales of the problem***.

8. Lecture 1st Mar.

We will revisit the model design process guided by the ***principle of maximum entropy*** and by the ***available measurements*** $((\mu_k)_{k\leq K})$ obtained as expectations of data functions $((\mathbb{E}[\phi_k(x)])_k)$. The Gibbs theorem (Th. 17) provides us with access to an approximation of $p(x)$, the probability density, using a ***parametric exponential model*** $p_\theta(x)$ where ***the parameters*** Θ ***are the Lagrange multipliers associated with the measurement constraints***, and hence are the ***dual variables of the measurements***. We will delve into this duality because, as a reminder, these parameters, which may seem like mathematical abstractions, represent tangible quantities in Physics such as temperature, pressure, viscosity, and more.

In the process of entropy maximization, we find $p_\Theta(x)$, which approximates $p(x)$ in

such a way that the divergence between the two distributions is given by

$$D_{KL}(p\|p_\Theta) = \mathbb{H}(p_\Theta) - \mathbb{H}(p) \geq 0 \tag{238}$$

Thus, the error is attributed to an excess of entropy, i.e., too much uncertainty that we lacked to construct the model. Therefore, the problem we face concerns ***the choice of measures*** $\phi_k(x)$ ***that constrain the model*** (i.e., the parameters Θ). Hence, we aim to minimize $D_{KL}(p\|p_\Theta)$ while maximizing entropy. This is achieved by finding ***the most informative measures***.

8.1 Example: The Gaussian Distribution

So, what measurements can we perform? When we have a *random variable*[83] $X = (X_1, \ldots, X_d)^T \in \mathbb{R}^d$, we may want to estimate the mean and also the second-order moments

$$\mu = \mathbb{E}_p(X) = (\mathbb{E}_p(X_i))_{i\leq d} \qquad \Sigma = \mathbb{E}_p(XX^T) = (\mathbb{E}_p(X_iX_j))_{(i,j\leq d)} \tag{239}$$

83. NDJE. To start a bit more simply, let's take a variable $X \in \mathbb{R}$ of which we know the expectation μ and the variance σ^2. The constraints translate to $\mathbb{E}[X] = \mu$ and $\mathbb{E}[X^2] = \sigma^2 + \mu^2$, giving us $\phi_1(x) = x$ and $\phi_2(x) = x^2$. Then we find that

$$p_\Theta(x) = Z^{-1}e^{\theta_1 x + \theta_2 x^2} \qquad Z_\Theta = \int e^{\theta_1 x + \theta_2 x^2} dx$$

Now,

$$\mu = \frac{\partial \log Z}{\partial \theta_1} \quad \text{and} \quad \sigma^2 + \mu^2 = \frac{\partial \log Z}{\partial \theta_2}$$

Furthermore, Z has the expression ($\theta_2 < 0$)

$$Z(\theta_1, \theta_2) = \left(\frac{\pi}{\theta_2}\right)^{1/2} e^{\theta_1^2/(4\theta_2)}$$

Thus, we find that $1/\theta_2 = -2\sigma^2$, $\theta_1 = \mu/\sigma^2$, and $Z = \sqrt{-\pi/\theta_2}e^{-\theta_1^2/(4\theta_2)}$. Finally, the distribution $p_\Theta(x)$ takes the form

$$p_\Theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} = \mathcal{N}(\mu, \sigma^2)$$

The case treated by S. Mallat is a generalization where $X \in \mathbb{R}^d$.

What is the expression for $p_\Theta(x)$? According to theorem 17, then

$$p_\Theta(x_1, \ldots, x_d) = Z_\Theta^{-1} \exp\left\{ \sum_{i=1}^{d} \theta_i x_i + \sum_{i,j=1}^{d} \theta_{ij} x_i x_j \right\} \tag{240}$$

We can "complete the squares," and by expressing the θ in terms of μ, we obtain the expression

$$p_\Theta(x) = Z_\Theta^{-1} \exp\left\{ -\frac{1}{2}(x-\mu)^T C^{-1} (x-\mu) \right\} \tag{241}$$

where the matrix C is the covariance (the expectation of the centered second-order moment)

$$C = \mathbb{E}_p((X-\mu)(X-\mu)^T) = \Sigma - \mu\mu^T \tag{242}$$

and $Z_\Theta = ((2\pi)^d |\det C|)^{1/2}$. It is verified that μ is indeed the mean of p. Regarding C, as a symmetric matrix that is diagonalizable, we can perform rotations to place ourselves in the basis where C is diagonal, allowing us to compute second-order moments through integrals.

This tells us that ***the Gaussian distribution can be interpreted as the maximum entropy distribution constrained by means and second-order moments***. Now, Gaussian distributions are certainly very practical and sufficient in some cases, but ***they cannot capture the structures*** of a turbulent field or textures. So, what are the ϕ_k that we need to add beyond second-order moments?

In theorem 17, there is clearly ***a duality link between observables and Lagrange multipliers***

$$(\mu_k = \mathbb{E}_{x\sim p}[\phi_k(x)])_{k\leq K} \longleftrightarrow (\theta_k)_{k\leq K} \tag{243}$$

Duality tells us that the representation of p_Θ can be written with either set of variables (see, for example, note 83). We will delve into this link because this notion of duality is fundamental, not only in Mathematics in optimization problems, but also in Physics through the Lagrange and Hamiltonian equations. All of this is connected through the Legendre transformation. But before that, let's examine the basic properties of the partition function.

8.2 Partition Function Z_Θ

We have seen in theorem 17 that observables μ_k and the associated Lagrange multiplier θ_k are linked through the partition function as follows:

$$\mu_k = \frac{\partial \log Z(\Theta)}{\partial \theta_k} \tag{244}$$

Let's define the function, called *free energy* in Physics $F(\Theta)$ and in Optimization $A(\Theta)$, *the cumulant function* as follows:

$$A(\Theta) = -F(\Theta) = \log Z(\Theta) \tag{245}$$

Theorem 18 ***(cumulant function)***
In the context of theorem 17, if the function $A(\Theta) = \log Z_\Theta$ has higher-order derivatives on its support Λ, then firstly

$$\nabla_\theta A(\Theta) = \mu = \mathbb{E}(\Phi(x)) \qquad \textit{that is, } \forall k \leq K,\ \mu_k = \frac{\partial A(\Theta)}{\partial \theta_k} = \mathbb{E}(\phi_k(x)) \tag{246}$$

and secondly, if we look at the Hessian, then

$$\nabla^2_\theta A(\Theta) = Cov(\Phi(x)) \geq 0 \qquad \textit{that is, } \forall (k,k') \leq K, \frac{\partial^2 A(\Theta)}{\partial \theta_k \partial \theta_{k'}} = Cov(\phi_k(x)\phi_{k'}(x)) \geq 0 \tag{247}$$

(NDJE. expectations are taken with respect to the distribution $p_\Theta(x)$)

The proof presents no problems; we just need to differentiate under the integral sign, so the conditions for doing so need to be satisfied, but with an exponential, this works fine. This theorem tells us that it is indeed the partition function that characterizes all higher-order moments of p_Θ.

Let's make some observations:

- The Hessian of A is positive (it's a covariance matrix), so ***A is a convex function***.
- Next, consider the following definition:

Definition 18 *(**free family**)*
We call $(\phi)_{k\leq K}$ ***free*** *if it is a* ***linearly independent family****:*

$$\sum_k \beta_k \phi_k = B^T \Phi = \Phi^T B = 0 \Leftrightarrow B = 0 \tag{248}$$

In this case, $C = Cov(\Phi(x)) > 0$, meaning that ***no eigenvector is associated with the eigenvalue zero***. Indeed,

$$\begin{aligned} \forall B \in (\mathbb{R}^K \backslash \{0\}),\ \mathbb{E}[(B^T(\Phi(x) - \mu))^2] &= \mathbb{E}[B^T(\Phi(x) - \mu)(\Phi(x) - \mu)^T B] \\ &= B^T \mathbb{E}[(\Phi(x) - \mu)(\Phi(x) - \mu)^T] B \\ &= B^T C B = f(B) > 0 \end{aligned} \tag{249}$$

Now, $f(B)$ being a strictly positive quadratic form, the eigenvalues of C are *strictly* positive. This is shown by noticing that C being a real symmetric matrix, there exists an orthogonal matrix Q and a diagonal matrix D such that $Q^T C Q = D = diag((\lambda_k)_{k\leq K})$. If $\tilde{B} = QB$, then

$$B^T C B > 0, \forall B \neq 0 \Leftrightarrow \tilde{B}^T D \tilde{B} > 0, \forall \tilde{B} \neq 0 \Leftrightarrow \sum_{k=1}^{K} \lambda_k \tilde{b}_k^2 > 0, \forall \tilde{B} \neq 0 \tag{250}$$

Taking $\tilde{B}$ as the vectors of the canonical basis of $\mathbb{R}^K$, we conclude that all eigenvalues of C are strictly positive.

In this context where $(\phi(x)_k)$ is a ***free family***, then ***the Hessian***, which is equal to the covariance matrix, is ***strictly positive***. ***The problem is then strictly convex***, which is particularly interesting in the use of gradient descent. In particular,

$$\nabla^2 A(\Theta) > 0 \Leftrightarrow (\nabla A(\Theta_1) - \nabla A(\Theta_2)).(\Theta_1 - \Theta_2) > 0 \tag{251}$$

so there cannot be 2 sets of parameters Θ that yield the constraints μ_k. ***There is a unique solution***, meaning that there is only one set of Lagrange multipliers.

8.3 Conjugate Duality: Legendre-Fenchel Transform

Notice that if we have the values of the optimal Lagrange multipliers Θ^*, then we have the probability density p_{Θ^*}. Not only are the $(\mu_k)_k$ accessible, but also any other higher-order moments of the distribution. That being said, how do we compute the vector Θ from the values of $(\mu_k)_k$? This is where the Legendre transform comes into play, and we will take another look at what we are doing, which has a more general character.

Let's place ourselves in a general convex optimization framework. We can reparameterize the function and work in the dual space. To do this, we introduce the Legendre-Fenchel transformation [84].

Definition 19 *(Legendre-Fenchel)*
Let $f : \chi \to \mathbb{R}$ or $\bar{\mathbb{R}}$ (contains infinity) be a convex function. Its conjugate function is defined by the Legendre-Fenchel transformation as follows [a]:

$$L[f](s) = \sup_{x \in \chi} \{s^T x - f(x)\} \tag{252}$$

The solution to the problem is $s = \nabla_x f(x)$, which implicitly provides a relationship between x and s.

a. NDJE. there are several conventions concerning the overall sign of the transformation and the relative sign between the two elements.

(NDJE. While the Legendre-Fenchel transform is initially defined for convex functions and is used as such here, its application to a non-convex function is important to obtain a convex envelope of that function. This property is used in the Landau theory of phase transitions. See section 8.8 for an example of convexifying a Mexican hat.)

84. NDJE. The transformation is named after Adrien-Marie Legendre (1752-1833), who used it in Analytical Mechanics for the transition from the Lagrangian to the Hamiltonian, and in Thermodynamics; and Moritz Werner Fenchel (1905-88), known in particular for his work in convex analysis. I'm using the notation $L[f]$ because $*$ is often used to represent the result of optimization, such as argmax. However, in the literature, you might find f^*.

Property 6 *(**convexity of** $L[f]$)*
The Legendre-Fenchel transform of a convex function is itself a convex function.

Proof 6. We will only show this property in the case where f is convex and twice differentiable (1D) and $f'' > 0$. A solution to the problem is $s = f'(x)$. Note by assumption that f' is strictly monotonic and invertible. So, if we denote $g = (f')^{-1}$, $x = g(s)$, and thus $f'(g(s)) = s$. The function g is also differentiable, $g'(s) = 1/f''(g(s))$. And $L[f](s) = g(s)s - f(g(s))$ is also differentiable, with

$$L[f]'(s) = g(s) + g'(s)(s - \underbrace{f'(g(s))}_{s}) = g(s)$$

and gives

$$L[f]''(s) = g'(s) = 1/f''(g(s)) > 0$$

implying the convexity of the Legendre transformation in this case, but the proof can be generalized [85]. ■

The Legendre transformation $L[f]$ is also called ***the convex conjugate of*** f. The relation $s = \nabla_x f(x)$ allows us to consider parameterizing f not in terms of x but in terms of s.

Theorem 19 *(**synthesis of convex functions**)*
*If f is **convex**, then we can synthesize it using the Legendre transformation applied to $L[f]$:*

$$L[L[f]](x) = L^2[f](x) = \sup_s \{s^T x - L[f](s)\} = f(x) \tag{253}$$

Here again, a solution is provided by $x = \nabla_s L[f](s)$. In this case, $L^2 = Id$ (involution).

(NDJE. An example of a convex function: $f(x) = \frac{1}{4}x^4$. Finding s from x yields $s = f'(x) = x^3$, so here the case is simple, $L[f](s) = \frac{3}{4}s^{4/3}$. Conversely, $x = (L[f])'(s) = s^{1/3}$, so we find that $L[L[f]](x) = x^{4/3} - 3/4x^4 = f(x)$. We recover $f(x)$ by applying the Legendre transformation twice.)

A 1D example is given in Figure 26. There is a geometric interpretation of $L[f](s)$:

85. NDJE. See, for example, https://www.math.univ-toulouse.fr/~weiss/Docs/LectureNotes_ConvexOptimization_PWeiss.pdf.

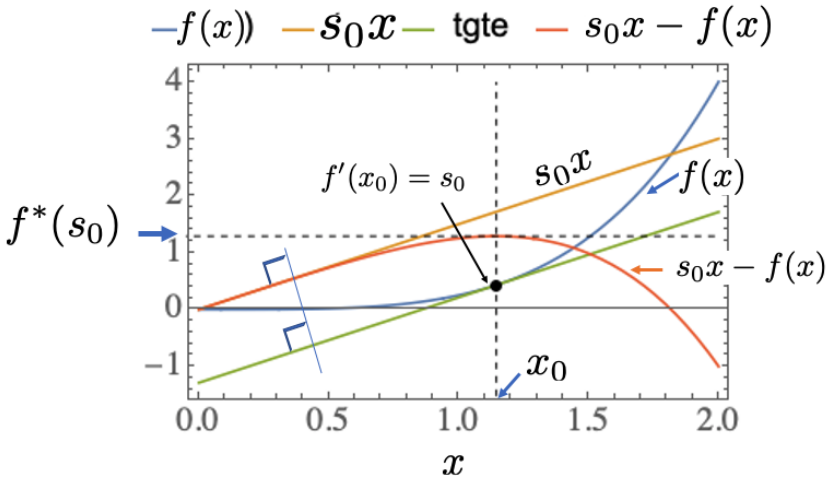


FIGURE 26 – Illustration of the Legendre-Fenchel transformation of the function $f(x)$ to yield the function $L[f](s)$. For a given s_0, the maximum of $s_0x - f(x)$ is reached for x_0 such that $f'(x_0) = s_0$. We have a parametrization of x in terms of s that can be exploited to transform $f(x)$.

the maximum of $sx - f(x)$ is reached at the value of x_0, and the point $(x_0, f(x_0))$ is where the tangent to the graph of f at that point has a slope equal to s_0.

An application in Analytical Mechanics is the transition from the Lagrangian [86] $\mathscr{L}(q, \dot{q}, t)$, which, by the principle of least action, gives the Euler-Lagrange equations, to the Hamiltonian $H(q, p, t)$ defined by the Legendre transformation:

$$p(q, \dot{q}, t) := \frac{\partial \mathscr{L}(q, \dot{q}, t)}{\partial q} \qquad H(q, p, t) = \dot{q}\, p - \mathscr{L}(q, \dot{q}, t) \tag{254}$$

which yields the Hamiltonian equations where the Newtonian force appears as the time derivative of p. Another application is in Statistical Physics for the transition from energy to free energy.

Now, if we inspect theorem 18, we realize that the relation

$$\mu = \nabla_\Theta A(\Theta) \tag{255}$$

86. NDJE. Here, only a single particle is considered, but this generalizes easily. The notation $\dot{q}$ indicates the derivative of $q(t)$ with respect to time t.

plays the same role as $s = \nabla f(x)$. So, we can write

$$L[A](\mu) = \sup_{\Theta}\{\mu^T\Theta - A(\Theta)\} \tag{256}$$

$$L^2[A](\Theta) = \sup_{\mu}\{\mu^T\Theta - L[A](\mu)\} \tag{257}$$

Let's see the connection between the Legendre transform and the optimization of Θ.

8.4 Optimisation of Θ in terms of μ

Our initial goal is to find the parametric model $p_\Theta(x)$ that best approximates $p(x)$ while satisfying the constraints μ (measurements) and following the principle of maximum entropy. If we have n samples $(x_i)_{i\leq n}$ assumed to be randomly drawn from $p(x)$ (i.i.d.), then

$$\mathbb{E}_p[\log p_\Theta] = \int p(x) \log p_\Theta(x)\ dx \approx \frac{1}{n}\sum_{i=1}^{n} \log p_\Theta(x_i) \tag{258}$$

The ***Maximum Likelihood Principle***[87] by R. Fisher tells us that if $p_\Theta(x)$ is truly a good model of $p(x)$ and x_i is a typical example of $p(x)$, then we also expect $p_\Theta(x_i)$ to be large, and thus, the average to be large.

In the case where p_Θ takes the form of a Gibbs distribution, then

$$\log p_\Theta(x) = \Theta^T\Phi(x) - \log Z(\Theta) = \Theta^T\Phi(x) - A(\Theta) \tag{259}$$

So, to obtain the optimal Θ, also known as the Maximum Likelihood Estimator (MLE), we need to perform

$$\max_\Theta \mathbb{E}_p[\log p_\Theta] = \max_\Theta\{\Theta^T\mathbb{E}_p[\Phi(x)] - A(\Theta)\} = \max_\Theta\{\Theta^T\mu - A(\Theta)\} = L[A](\mu) \tag{260}$$

We then find the connection with the Legendre transform of the function $A(\Theta)$ with a correspondence of $s \leftrightarrow \mu$ and $x \leftrightarrow \Theta$ from the general case in the previous section

87. See Course 2022 Sec. 3.5

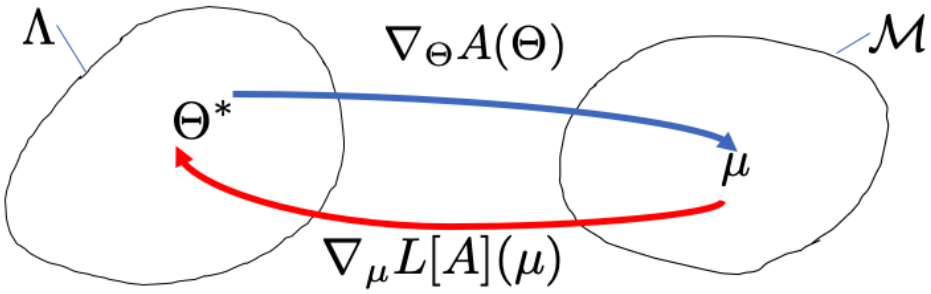


FIGURE 27 – Illustration of the dual relationship between the set of measurements μ and the set of parameters Θ.

(Def. 19). Thus, having the measurements μ, the optimal Θ^* satisfies

$$\nabla_\Theta A(\Theta^*) = \mu \tag{261}$$

However, we know that $L[A]$ is convex (independent of the convexity of A), and therefore,

$$\Theta^* = \nabla_\mu L[A](\mu) \tag{262}$$

We can then summarize the dual relationship between μ and Θ when considering a free family of measurements (Def. 18) in the diagram in Figure 27. In one case, we need to calculate $A(\Theta)$ to obtain its gradient, and in the other case, it is $L[A](\mu)$ that concerns us, and we need to maximize likelihood to obtain the optimal Θ^*.

To obtain Θ^*, we ultimately need to perform

$$\Theta^* = \underset{\Theta}{\text{argmax}} \ (\Theta^T \mu - A(\Theta)) = \underset{\Theta}{\text{argmin}} \ (A(\Theta) - \Theta^T \mu) \tag{263}$$

Typically, in optimization, we can perform a gradient descent [88], so between the parame-

88. See Course 2018 Sec. 10.1, for example

ters obtained at steps t and $t+1$, we have the following relationship:

$$\begin{aligned}\Theta_{t+1} - \Theta_t &= -\nabla_\Theta\{A(\Theta) - \Theta^T\mu\}\Big|_{\Theta=\Theta_t} = -\nabla_\Theta A(\Theta_t) + \mu \\ &= \mu - \mu(\Theta_t)\end{aligned} \tag{264}$$

where, according to the Gibbs parameterization (Th. 17),

$$\mu(\Theta) = \mathbb{E}_{x\sim p_\Theta}[\Phi(x)] \tag{265}$$

So, if we start with an arbitrary value Θ_0, then from the distribution p_{Θ_0}, we can calculate the moments $\mu(\Theta_0)$ using expectations. It is clear that initially these moments will not be equal to the true moments μ which, by hypothesis, are given by

$$\mu = \mathbb{E}_{x\sim p}[\Phi(x)] \tag{266}$$

Therefore, there is an error between μ and $\mu(\Theta_0)$, and the new value Θ_1 must go against the gradient by an amount equal in magnitude to the error on the moments. We iterate the procedure, and as ***we are in a strictly convex case*** (see the observations made after theorem 18), ***the solution exists and is unique***.

So, ***the relationship between the Maximum Entropy Principle and the Maximum Likelihood Principle is a dual relationship that we understand through the Legendre-Fenchel Transform of the logarithm of the partition function*** $Z(\Theta)$, which is called the free energy F in Statistical Physics or the cumulant function A in Optimization.

In conclusion, the mathematical framework described above shows the general nature and removes the concepts of the *partition function* and *free energy* from their framework in Statistical Physics, where Gibbs introduced them in the early 20th century. Also, in the case of optimization, we can view the problem as that of maximizing entropy under constraints (observables, μ), or that of maximum likelihood.

8.5 Problem of Estimating the Quality of Θ_t

If we take a closer look at the equation 264 for gradient descent from Θ_t to Θ_{t+1}, we need to be able to test this new parameter value. To do this, we must be able to calculate

the new estimates of moments. So, in a general sense, we need to be able to calculate $\forall k$

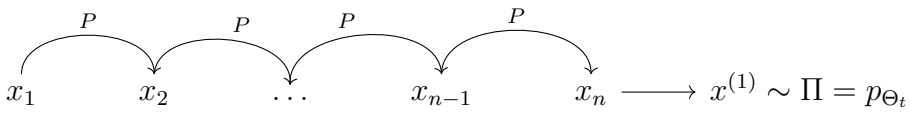
$$\int \phi_k(x)\ p_\Theta(x)\ dx \tag{267}$$

How can we perform this type of integral calculation? One possible method that seems natural is as follows:

1. Generate N samples $x \sim p_\Theta(x)$;
2. Compute the average $1/N \sum_{i=1}^{N} \phi_k(x_i)$. The error scales as $1/\sqrt{N}$.

The challenge lies in step (1) because ***we need to be able to sample from*** $p_\Theta(x)$, which already poses problems in low dimensions if the distribution has multiple modes, but it becomes increasingly difficult in high dimensions, even in cases where there is only one mode.

A commonly used technique is the generation of Markov chains (Sec. 6.4), called ***Monte Carlo Markov Chain*** (MCMC). We construct ***Markov chains whose invariant measure (Def. 11) is precisely*** p_Θ**.** In other words, to produce 1 sample x_i drawn according to p_{Θ_t} and to determine the moments $\mu_k(\Theta_t)$, we need to evolve a Markov chain using a transition matrix such that the invariant distribution is p_{Θ_t}



Then, we repeat the process to obtain $x^{(2)}$ and so on, in order to obtain enough samples to calculate the averages and estimate the quality of $\mu(\Theta_t)$.

In practice, there are several algorithms that generate such chains, such as the ***Metropolis-Hastings*** algorithm *(NDJE. See section 8.9 for an introduction. You can experiment with the 2023 notebooks https://github.com/jecampagne/cours_mallat_cdf/tree/main/cours2023).* We won't go into more detail about MCMC algorithms.

8.6 How to Design $\Phi(x)$?

Summing up the previous sections, ***if we have well-suited*** $(\phi_k)_k$ ***for the problem, we have a formalism that works well***. Especially when we are in the convex case. ***But***

the problem still remains: what are the $(\phi_k)_k$, if we have images of faces, textures, sound patterns of music, speakers, etc., that will allow us to efficiently infer a Gibbs (exponential) probability distribution, which will generate new images, textures, new sound patterns, etc., by sampling. To address this problem, we need to go back to basics and ask the question: ***what properties can we identify and integrate into the form of*** $\phi(x)$? This is the theme of ***a priori information***, which was explored, for example, in the 2020 course.

Now, how do we express this *a priori* information in mathematically exploitable terms? Even if we have precise experimental descriptions, for example, of a turbulent fluid image or the texture of a tree bark, ***the challenge is to concretize it in mathematical terms that define*** $\phi(x)$***: that is the difficult point***.

What has worked so far in the development of Physics (Statistical, Elementary Particles, and Relativity) and in Mathematics as well, is to look at ***the symmetries of the problem***.

8.7 Symmetries of $p(x)$

The question we are tackling is to understand the transformations of x that leave $p(x)$ invariant. Typically, let g be such a transformation, generally an element of a group G, which acts on x to transform it into x' in such a way that

$$g \in G, x' = g.x \xRightarrow{invariance} p(x') = p(x) \tag{268}$$

G is indeed a group because

— $g = Id$ is obviously the identity element of G;

— if $(g_1, g_2) \in G$,

$$p((g_1 \circ g_2).x) = p(g_1.(g_2.x)) = p(g_2.x) = p(x)$$

so $g_1 \circ g_2 \in G$;

— finally, we can impose that g has an inverse such that $g \circ g^{-1} = Id$.

The most classic group G is that of ***translations***. In this case, if we conceive that x depends on a variable u, e.g., $u = (u_1, u_2)$ representing the position of a pixel in an image and $x(u)$

represents the intensity of the signal at that position, then the action of a translation by "vector" g is expressed as

$$x'(u) = x(u - g) = g.x(u) \tag{269}$$

and ***stationarity*** is expressed as

$$p(x) = p(g.x) \Leftrightarrow \forall u,\ p(x(u)) = p(x(u - g)) \tag{270}$$

Stationarity is the manifestation of symmetry with respect to the group of translations. Here's an example where we move from an experimental observation (*"stationarity"*) to a mathematical characterization (*"symmetry with respect to the group of translations"*).

Of course, there can be other types of symmetries: e.g., ***the group of rotations***. But the rotation symmetry might be suitable for certain scenarios, like in Cosmology when observing the sky, but not, for example, in the case of face images where verticality imposes a preferred direction. We can also have ***scale invariance***: for example, in image processing, the distance between the object and the camera does not change the nature of the object, so the collection of all possible images of that object has a scale-invariant probability distribution. But on the other hand, passport photos are not invariant under dilation because preprocessing crops the image to provide an acceptable image for a passport or ID card. All this to say that ***each problem has its own symmetries***. What works in one case/application may not work in another. Complex symmetries are also possible (e.g., local deformations or diffeomorphisms).

But why is it important to know the symmetries of $p(x)$? The main reason is ***the reduction of dimensionality*** they allow us to achieve. This is done by reducing the number K of measurements $(\phi_k(x))_{k \leq K}$, by imposing invariants, which in turn implies a reduction in the number of parameters $(\theta_k)_{k \leq K}$ of the model $p_\Theta(x)$ that approximates $p(x)$.

Note that we have K measurements, and x is of dimension d ($x \in \mathbb{R}^d$). The question we could ask is what is the relationship between K and d? Through reasoning developed in Section 2.6, we expect K to grow exponentially with d: this is the ***curse of dimensionality***. So, K should be huge when $d \sim 10^6$. But if we want to estimate $p(x)$ from a single image, then $K \ll d$. ***This means that a lot of prior information is needed to achieve such dimensionality reduction***. This is where symmetries come to our aid.

Definition 20 *(**Linear Invariant**)*
Let G be a finite group of linear transformations[a] *of x, and let* ***the averaging operator****, denoted by $\mathcal{M}$, be defined as*

$$\mathcal{M}.x = \frac{1}{|G|}\sum_{g\in G} g.x \tag{271}$$

In other words, we take an average over the elements of the orbit of x, which is the set $O_x = \{g.x; g \in G\}$. (NDJE. we can generalize this to a continuous group).

a. e.g., translation, rotation, dilation

$\mathcal{M}$ is an operator invariant under the action of the group. Indeed, for any $g_0 \in G$

$$\mathcal{M}.(g_0.x) = \frac{1}{|G|}\sum_{g\in G}\phi(g.(g_0.x)) = \frac{1}{|G|}\sum_{g\in G}\phi(g.x) = \mathcal{M}.x \tag{272}$$

In the case of an image and G being the group of translations with parameters τ

$$\mathcal{M}.x = \sum_{\tau} x(u-\tau) \tag{273}$$

which is nothing more than the average of the signal contained in the image, and this is indeed an invariant [89].

The consequence is that we can restrict ourselves to $\phi(x)$ that are invariant.

Definition 21 *(**Equivariance**)*
We say that $\phi(x)$ is equivariant under the action of the group G if

$$\forall g \in G,\ \phi(g.x) = g.\phi(x) \tag{274}$$

meaning that ϕ commutes with the elements of G.

89. NDJE: To convince ourselves, let's take a periodic signal $x(t)$ on $[0, 2\pi]$ as an example. The sum over θ is equivalent to an integral $(2\pi)^{-1}\int_0^{2\pi} x(u-\theta)d\theta$, which, through a change of variable $u' = u - \theta$ and the periodicity of the signal, becomes $(2\pi)^{-1}\int_0^{2\pi} x(t)dt$, which is nothing but the average of the signal (

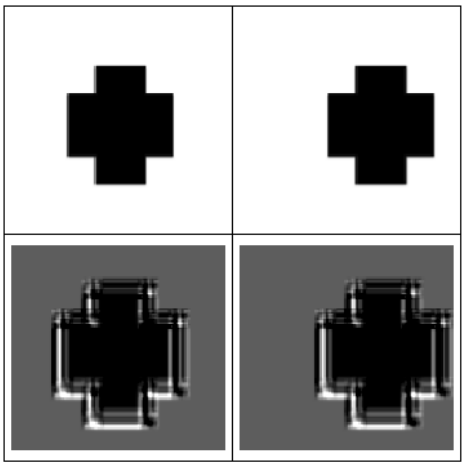


FIGURE 28 – Illustration of the equivariance property of convolution under the application of a translation: (top) from left to right: original image, translated image; (bottom) from left to right: convolution of the original image by a filter, and convolution by the same filter of the translated image. The convolution of the translated image and the translation of the convolution of the original image. If x is the image, f is the convolution, and g is the translation, then $f(g.x) = g.f(x)$.

An illustration of the equivariance property of convolution under the action of the group of translations is provided in Figure 28.

Note that ***equivariance***, also called ***covariance***, is distinct from ***invariance***:

$$\begin{cases} f(g.x) = f(x) & \text{invariance} \\ f(g.x) = g.f(x) & \text{equivariant/covariance} \end{cases} \tag{275}$$

We have an immediate result:

Property 7 *In the case where ϕ is equivariant, the operator*

$$\tilde{\phi}(x) = \mathcal{M}.\phi(x) = \frac{1}{|G|} \sum_{g \in G} g.\phi(x) = \frac{1}{|G|} \sum_{g \in G} \phi(g.x) \tag{276}$$

is ***invariant***.

Indeed, by the same type of change of variables as before,

$$\tilde{\phi}(g'.x) = \frac{1}{|G|}\sum_{g\in G}\phi(g.(g'.x)) = \frac{1}{|G|}\sum_{g\in G}\phi(g.x) = \tilde{\phi}(x) \tag{277}$$

Theorem 20

If Φ is equivariant under G, and if $p(x)$ is invariant under G, then we have the following properties:

1) $\mu = \mathbb{E}_p(\Phi(x)) = \mathbb{E}_p(\mathcal{M}.\Phi(x))$

2) the maximum entropy distribution $p(x)$ constrained by the measurements $\mathbb{E}_p(\Phi(x)) = \mu$ is given by

$$p(x) = Z^{-1}e^{-\Theta^T\mathcal{M}\Phi(x)} \tag{278}$$

Proof 20. For the first property, we can prove it by expanding the expression $\mathbb{E}_p(\mathcal{M}.\phi(x))$ for one of the elements of $\Phi = (\phi_k)_{k\leq K}$, which we denote as ϕ for brevity. It follows:

$$\begin{aligned}
\mathbb{E}_p(\mathcal{M}.\phi(x)) &= \sum_{x\in\chi} p(x)\frac{1}{|G|}\sum_{g\in G}\phi(g.x) && \text{(equivariance of } \phi\text{)}\\
&= \frac{1}{|G|}\sum_{g\in G}\sum_{x\in\chi} p(x)\phi(g.x)\\
&= \frac{1}{|G|}\sum_{g\in G}\sum_{x'\in\chi} p(g^{-1}.x')\phi(x') && (x' = g.x)\\
&= \frac{1}{|G|}\sum_{g\in G}\left(\sum_{x\in\chi} p(x)\phi(x)\right) && \text{(invariance of } p\text{)}\\
&= \sum_{x\in\chi} p(x)\phi(x) && (\sum_{g\in G} a = |G|a)\\
&= \mathbb{E}_p(\phi(x)) && (279)
\end{aligned}$$

The second property follows from the fact that we can seek $p_\Theta(x)$ using the Gibbs theorem (Th. 17) with the constraints $\mu = \mathbb{E}_p(\mathcal{M}.\Phi(x))$. ■

This tells us that ***when we average, we reduce the number of variables by the size of the group***. Ultimately, it is not necessary to take arbitrary $\phi(x)$; ***we only need to keep the averages of*** $\phi(x)$ ***with respect to the symmetry group***. Therefore, we only need to keep the observables "modulo" the symmetries. For example, for second-order moments, we don't consider all possible pairs of points but can group them into batches of pairs with the same relative distance.

Now, as we mentioned earlier, before considering the symmetries of $p(x)$, K is of the order of e^d. If we want to reduce it to $K \sim d$, we need a group whose size is of the order of e^d. This is where we reach the limits of introducing global symmetries to drastically reduce the dimensionality of the problem. In practice, we can assume that $p(x)$ is symmetric under certain groups (translation, rotation, dilation), but ***in the end, we don't have enough knowledge to dramatically reduce the dimensionality of the problem***. Thus, we need to approach the problem more finely. In particular, we need to look at ***the problem's scale dependence***. Indeed, there are not only global symmetries that matter; ***if we observe a phenomenon locally, it tends to propagate across scales***. Expressing this observation in mathematical terms will be the focus of the next session.

8.8 NDJE. Legendre Transformation: Non-Convex Case

This is a small supplement regarding the Legendre transformation. The question is, what happens when the function f is not convex? An example is given with the function $f(x) = 1/4 - x^2 + x^4$, which resembles a Mexican hat (Fig. 29). If we study the function $g(s) = xs - f(x)$, we need to resort to solving a third-degree equation to find the maximum, which cannot be done analytically in its current form. However, we know that $s = f'(x)$ gives the function $s(x)$, which allows us to determine the supremum. We can establish the parametric graph in s given by $\{s = f'(x), g(s(x))\}$ (see Fig. 29 on the right in blue). This graph exhibits a characteristic "swallowtail" shape. The Legendre transform $L[f](s)$ is represented by the dashed red curve. Let's now perform the transform of $L[f](t)$; the key point is that the slope of $L[f](s)$ at the origin is given by $\pm x_0 s$, where $f(\pm x_0) = 0$. However, since $L[f](s)$ is convex, this means that for all $0 < x < x_0$, $xs - L[f](s) \leq 0$, and the supremum with respect to s is reached at $s = 0$, which is 0. Hence, for $0 < x < x_0$, $L[L[f]](x) = 0$, and symmetrically, it is also zero for $-x_0 < x < 0$. Outside the interval $[-x_0, x_0]$, the function $f(x)$ is recovered, and we now have an evidently convex curve

(dashed red curve in Fig. 29 on the left). We have thus convexified the originally non-convex function.

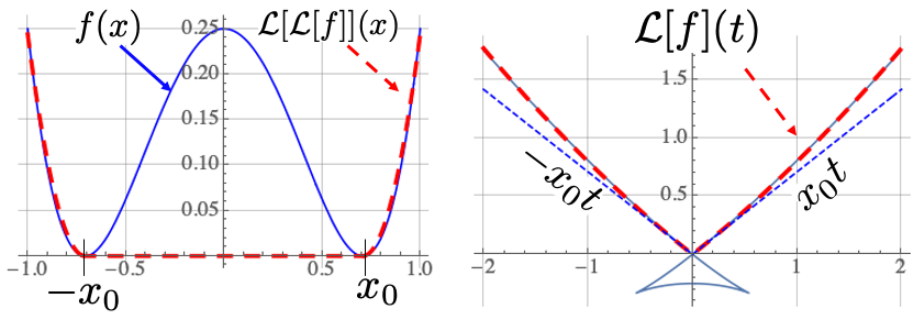


FIGURE 29 – Convexification of the Mexican hat.

8.9 NDJE. Metropolis-Hastings

This is an addition to the course because S. Mallat could not cover it due to time constraints. It is about the Metropolis-Hastings algorithm for generating Markov chains to obtain samples from a target distribution $P(x)$ for which we do not have a direct sampling method, or the other methods (e.g., Importance Sampling) are inefficient.

The method was first developed in 1949 by Nicholas C. Metropolis (1915-99) and Stanisław Ulam (1909-84) and was later elaborated in more detail in 1953 by Metropolis and colleagues and extended in 1970 by Wilfred Hastings (1930-2016). It is called the Metropolis-Hastings (MH) algorithm, even though several authors contributed to it.

So, consider a distribution $P(x)$ from which we want to draw samples. We will construct a Markov chain, for which we will specify the transition matrix $P_{x,x'}$ (transition from x to x') based on a distribution $g(x, x')$ (*proposal*) that allows us to draw a new sample x' given x, and a function $A(x, x')$ that accepts or rejects this new sample to provide a new element of the chain. Otherwise, the chain is completed with the sample x (note that it may happen that the latter is repeated several times in a row). Let's take

the expressions for $P_{x,x'}$ and $A(x,x')$ as follows:

$$P_{x,x'} = g(x,x')A(x,x') \qquad\qquad A(x,x') = \min\left(1, \frac{P(x')g(x',x)}{P(x)g(x,x')}\right) \tag{280}$$

It can be shown that $P(x)$ ***satisfies the detailed balance expression*** (Eq. 187). Indeed, assuming $P(x')g(x',x) > P(x)g(x,x')$, then

$$A(x,x') = 1 \qquad\qquad A(x',x) = \frac{P(x)g(x,x')}{P(x')g(x',x)} \tag{281}$$

so

$$P_{x,x'} = g(x,x') \qquad\qquad P_{x',x} = g(x',x) \times \frac{P(x)g(x,x')}{P(x')g(x',x)} = \frac{P(x)g(x,x')}{P(x')} \tag{282}$$

hence

$$P(x)P_{x,x'} = P(x')P_{x',x} \tag{283}$$

which is the desired expression. The case where $P(x')g(x',x) \leq P(x)g(x,x')$ is treated in the same way and leads to the same conclusion.

Therefore, if $P(x)$ satisfies the detailed balance expression and under the ergodicity conditions that guarantee the uniqueness of the invariant measure, then $P(x)$ ***is the said invariant measure reached asymptotically***.

The algorithm proceeds as follows: a new sample of the chain x_i at step i is generated according to the following steps:

1. A sample $\hat{x}$ is drawn according to $g(x; x_{i-1})$, which can be, for example, a Gaussian distribution with mean x_{i-1};
2. Then, we form the ratio

$$r = \frac{\tilde{P}(\hat{x})}{\tilde{P}(x_{i-1})} \times \frac{g(\hat{x}, x_{i-1})}{g(x_{i-1}, \hat{x})} \qquad\qquad (\text{Metropolis} - \text{Hastings}) \tag{284}$$

   We use the notation $\tilde{P}$ instead of P to indicate that we do not need to know the normalization constant of the target distribution in practice.

3. Finally, we decide on the value of x_i as follows:

$$\text{Let } u \sim \mathcal{U}(0,1), \quad \text{if } r > u \Rightarrow x_i = \hat{x}, \text{ otherwise } x_i = x_{i-1} \tag{285}$$

The initialization of the process is done by randomly generating x_0. If the distribution $g(x, y)$ is invariant under the exchange of variables x and y (e.g., a Gaussian), then the ratio r simplifies to

$$r = \frac{\tilde{P}(\hat{x})}{\tilde{P}(x_{i-1})} \qquad \text{(Metropolis)} \tag{286}$$

Even though this version of the ratio r is particularly dedicated to Metropolis, the terminology Metropolis-Hastings is often used in all cases. The drawback of the method is its random walk nature, especially because the distribution $g(x, y)$ is fixed once and for all. Other methods adapt the shape of this distribution as the process evolves. Finally, a category of "Hamiltonian" methods are Metropolis methods that use gradient information (p is the conjugate variable of x) to reduce the random walk behavior. These methods are particularly effective in high dimensions and cases where the probability density has a complex shape. In cases where the probability density has multiple disjoint modes, a different approach is required.

9. Lecture 8 Mar.

During this session, we will utilize the tools developed in the previous sessions (and lectures) to *model non-Gaussian ergodic stochastic processes* such as sounds, images. Thus, we will revisit the ***maximum entropy models***.

9.1 Maximum Entropy Models

We consider a scenario where, given the data x, we can compute several means of possibly (and almost surely) nonlinear functions $(\mu_k = \mathbb{E}(\phi_k(x)))_{k \leq K}$, and from this, we can infer a probability distribution of maximum entropy constrained by these observations.

This parameterized distribution is expressed as (Th. 17):

$$p_\Theta(x) = Z_\Theta^{-1} \exp\left(-\sum_{k=1}^{K} \theta_k \phi_k(x)\right) \tag{287}$$

with $\Theta = (\theta_k)_k$ being Lagrange multipliers associated with the constraints $(\mu_k)_k$. We have seen that:

$$D_{KL}(p\|p_\Theta) = \mathbb{H}(p_\Theta) - \mathbb{H}(p) \geq 0 \tag{288}$$

Therefore, we need to maximize the entropy of p_Θ while minimizing $D_{KL}(p\|p_\Theta)$. So, ***the choice of*** $\phi_k(x)$ ***or their construction is now important***.

Several issues arise:

— We need to compute expectations (μ_k), which relates to the question of ***ergodicity*** (Sec. 6.2). We use prior information about the studied stochastic process, namely its ***stationarity***;

— In the list of moments μ_k, we will first look at the ***means*** and ***covariances***. One important consequence in the ***stationary assumption*** is that the ***covariance operator is diagonalizable in a Fourier basis***, leading us to the notion of ***spectral power***. However, we will see that the Fourier basis is not suitable in most cases.

— We are in the case of ***a single observation*** (e.g., a single image) as in Cosmology (see the seminar by E. Allys). In general, phenomena are ***multi-scale***. Thus, we need to proceed with ***multi-scale analysis*** (also called multi-resolution), which leads us to the ***Wavelet Transform***. ***NDJE. S. Mallat will provide a brief overview of this transformation; for more details, see Sec. 5.3 of the 2021 course and the included references, as well as the book mentioned in the introduction.***

— The purpose of these concepts is to study ***interactions between scales***. We will see that capturing structures (e.g., textures' filaments in images, impulses in soundtracks, etc.) requires ***understanding how information propagates across scales***, unlike Gaussian processes where different scales are present but independent of each other (they fluctuate without interaction). These interactions are captured with ***non-linearities***, especially in ***the scattering transform*** (Course 2020 Sec. 9.5). We will see some examples of sound and image generation and the limitations of this kind of approach.

9.2 Mean Computation

So, we have a single sample x from a random variable X (e.g., one image, one sound frame), and u represents the underlying variable, such as the location of a pixel in the image ($u = (u_1, u_2)$). This variable u is of low dimension. In the assumption of stationarity, i.e., invariance by translation over the underlying variable, the means and covariances [90] satisfy the following relations:

$$\begin{cases} \mathbb{E}(X(u)) &= \mu(u) = \mu \\ Cov(X(u), X(u')) &= C(u, u') = C(u' - u) \end{cases} \quad \text{(stationary)} \tag{289}$$

Recall that according to ***the stationarity assumption is the manifestation of translation invariance*** of $p(x)$ (Sec. 8.7), the probability density of X used for expectation calculations. In other words, ***expectations do not depend on position, and covariances depend only on relative positions***. In this framework, the estimation of the signal's mean is obtained through the empirical mean:

$$\tilde{\mu} = \frac{1}{N} \sum_{u=1}^{N} X(u) \tag{290}$$

First and foremost, $\tilde{\mu}$ is an unbiased estimator because $\mathbb{E}(\tilde{\mu}) = \mu$ (thanks to the invariance of the expectation of $X(u)$). The question that arises is the *consistency* of this estimator (Course 2022 Sec. 3.4), which means how it behaves as $N \to \infty$ (e.g., the number of pixels tends to infinity).

Property 8

$$\mathbb{E}[(\tilde{\mu} - \mu)^2] = \frac{1}{N} \sum_{\tau=-N+1}^{N-1} \left(1 - \frac{|\tau|}{N}\right) C(\tau) \tag{291}$$

Note that if $\sum_{\tau=0}^{\infty} |C(\tau)| < \infty$, meaning that the correlation between 2 points rapidly decreases with the difference τ between these points, then $\mathbb{E}[(\tilde{\mu} - \mu)^2] = O(1/N)$.

90. $Cov(A, B) = \mathbb{E}((A - \mathbb{E}(A))(B - \mathbb{E}(B))^T)$. Of course, in the case of finite-sized images, sounds, etc., one must consider ad hoc conditions for the boundaries.

Ergodicity (at least for the mean) provides the assumption for this to work.

Proof 8. The proof is done by expressing the value of $\tilde{\mu}$ and performing an appropriate change of variables:

$$\begin{aligned}
\mathbb{E}[(\tilde{\mu}-\mu)^2] &= \mathbb{E}\left[\left(\frac{1}{N}\sum_{u=1}^{N} X(u)-\mu\right)\left(\frac{1}{N}\sum_{u'=1}^{N} X(u')-\mu\right)\right] \\
&= \mathbb{E}\left[\frac{1}{N^2}\sum_{u,u'=1}^{N}(X(u)-\mu)(X(u')-\mu)\right] \\
&= \frac{1}{N^2}\sum_{u,u'=1}^{N} Cov(u-u') \\
&= \frac{1}{N^2}\sum_{\tau=-(N-1)}^{N-1}\underbrace{(N-|\tau|)}_{\#\text{ terms}}C(\tau)
\end{aligned} \tag{292}$$

The number of terms is calculated taking into account that u and u' are in $[\![1\,,\,N]\!]$. Bringing in one of the $1/N$ terms allows us to obtain the result. ■

Now, the formula tells us the following: ***for the empirical mean to converge to the expectation, there must be decorrelation***, meaning that $C(\tau)$ decreases rapidly to make $\sum_\tau |C(\tau)|$ converge. In this case, $\mathbb{E}[(\tilde{\mu}-\mu)^2] = O(1/N)$.

So, ***we need to construct*** $\phi_k(x)$ ***that exhibit decorrelation to ensure the estimation of means with a single sample*** (one image, one sound frame, etc.). However, using only means is not sufficient; we need to consider higher-order moments.

9.3 Second-Order Moments (Covariance), Failure of Fourier

The covariance matrix $C = [C(u,u')]_{u,u'}$ (of size $N \times N$) allows us to calculate correlations between variables obtained by linear combinations of X. In fact, for any A and B:

$$A = \sum_{u=1}^{N} a(u)X(u) = a^T X, \quad B = b^T X \Longrightarrow Cov(A,B) = a^T C b \tag{293}$$

(NDJE. recalling what $Cov(A,B)$ is from footnote 90, we easily obtain the result).

The matrix C is ***symmetric***, ***positive*** because $Cov(A, A) = Var(A) = a^T Ca \geq 0$, so all eigenvectors are associated with positive (or zero) eigenvalues. It is ***diagonalizable***, and working in the eigenvector basis is equivalent to performing Principal Component Analysis (***PCA***, see Course 2021, Sec. 4.3 Th 7). However, we are in a stationarity assumption (Eqs. 289), so C is a symmetric and Toeplitz [91] matrix, which we can write as:

$$C = \begin{pmatrix} C(0) & C(1) & C(2) & \dots & C(N-1) \\ C(1) & C(0) & C(1) & \dots & C(N-2) \\ C(2) & C(1) & C(0) & \dots & C(N-3) \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ C(N-1) & C(N-2) & \dots & C(1) & C(0) \end{pmatrix} \tag{294}$$

Thus, for all b:

$$C(u,.)b = \sum_{u'} C(u,u')b(u') = \sum_{u'} C(u-u')b(u') = (C * b)(u) = \sum_{u'} C(u')b(u-u') \tag{295}$$

The presence of ***convolution*** tells us that ***the diagonalization basis is that of Fourier***. Thus,

$$C * e^{i\omega u} = \sum_{u'} C(u')e^{i\omega(u-u')} = e^{i\omega u} \underbrace{\sum_{u'} C(u')e^{-i\omega u'}}_{\widehat{C}(\omega)} \tag{296}$$

where we recognize ***the (series) Fourier transform of the covariance operator, and the sine waves are the eigenfunctions of this operator***. Since ***the operator is positive***, we have:

$$\widehat{C}(\omega) \geq 0 \qquad \text{(spectral power)} \tag{297}$$

Note that:

$$C(u) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \widehat{C}(\omega) e^{i\omega u} \, d\omega \tag{298}$$

91. Otto Toeplitz (1881-1940).

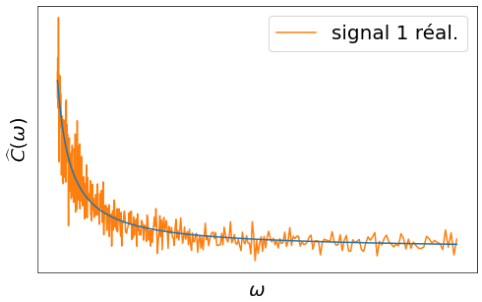


FIGURE 30 – Schematic representation of the evolution of spectral power as a function of ω (blue). For images, $\widehat{C}(\omega) \sim 1/\omega$. When you have only one realization, what you observe is auto-correlation of the signal (orange), and its fluctuations are also governed by $\widehat{C}(\omega)$.

and in particular, for $\tau = 0$:

$$C(0) = \mathbb{E}[|X(u)|^2] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \widehat{C}(\omega)\ d\omega \tag{299}$$

$C(0)$ is the total energy of the signal, and that's why ***the spectral power can be likened to energy density by frequency***.

How do we estimate this covariance if we have only one realization of X? We can project [92] X onto an eigenvector $e_\omega(u) = e^{i\omega u}/\sqrt{N}$:

$$e_\omega^T\ X = \frac{1}{\sqrt{N}} \sum_{u=1}^{N} X(u) e^{-i\omega u} \tag{300}$$

According to property (Eq. 293), since $a = (e_\omega(u))_{1 \leq u \leq N}$ is an eigenvector of C with eigenvalue $\widehat{C}(\omega)$, we deduce that:

$$\mathbb{E}(|e_\omega^T\ X|^2) = \widehat{C}(\omega) \tag{301}$$

A typical evolution of $\widehat{C}(\omega)$ is shown in Figure 30 (blue curve).

However, in the case of *a single realization*, the auto-correlation of the signal fluctuates (orange curve), and the magnitude of these fluctuations is directly related to the

92. NDJE. In the complex case, $a^T b$ should be understood as $(a^*)^T\ b$, where $*$ denotes the complex conjugate.

value of $\widehat{C}(\omega)$. So, there are errors, and the question arises of ***how to obtain a consistent estimator***? To do this, we choose to average the auto-correlation over ranges [93] of ω_k values weighted by well-chosen windows, in order to estimate the covariance only at a few values of ω. Thus, by combining multiple measurements of $\widehat{C}(\omega_k)$, noise can be suppressed. But consistency is only achieved for a few values of ω_k. ***There is an intrinsic problem with the consistency of the covariance estimator using the Fourier transform in practical cases.***

There is a deeper reason why the Fourier transform is not suitable. When analyzing a sound frame with transients, or an image with object contours, the following occurs: when you perform the Fourier transform of a signal like the one shown in Figure 31 (top image, blue curve), you correlate it with sine waves like the orange curve, which means taking a dot product that mixes information from the signal throughout its "temporal" course (note that the phenomenon is similar in 2D for an image). ***In particular, at a given frequency, you indiscriminately mix times when the signal is regular and times when it has a lot of structure***. When you then try to reconstruct the signal by taking only a portion of its spectral components (e.g., ω such that $|\widehat{C}(\omega)| > T$, where T is a threshold), small oscillations appear at all transition moments (red curve, bottom image). In fact, ***the dependence of the signal's Fourier spectrum is dominated by the signal's transients***: from the spectrum, you cannot tell if the signal is regular at a specific part of its temporal course (or part of an image).

9.4 Wavelet Filtering (1D)

Transients are typical of non-Gaussian phenomena, think for example of the localization of vortices in turbulent flow. So, it's crucial for us to be able to handle them as faithfully as possible. We need to have ***a local description of the signal*** and, therefore, be able to ***correlate it with functions that oscillate only over a small "temporal" window*** like the one in green in the top image of Figure 31, which we call ***wavelets***. In order to capture transients over the entire "temporal" range of the signal, we need to be able to ***translate*** these wavelets, and to capture all time scales, we also need to ***dilate/contract*** them as shown in Figure 32. The interest of these wavelets is that ***the correlation coefficient with the signal is nonzero only if the signal has a transient localized in the support of the wavelet***.

93. NDJE. k indicates that discrete values of ω are taken, in practice, when digitizing the signal.

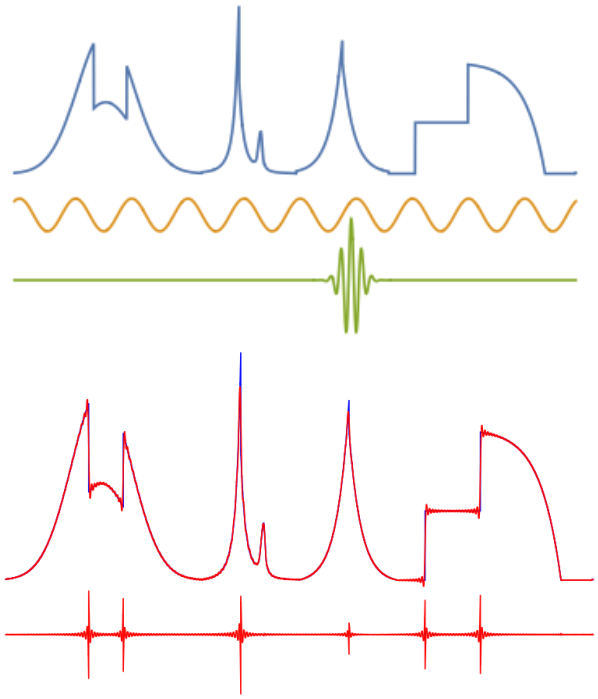


FIGURE 31 – Top image: original signal in blue, typical delocalized sinusoid from Fourier analysis (orange), and typical localized wavelet from Wavelet analysis (green). Bottom image: reconstructed signal in red when taking only part of the Fourier spectrum (the difference from the true signal is shown in the bottom graph). See *gibbs_FFT.ipynb*

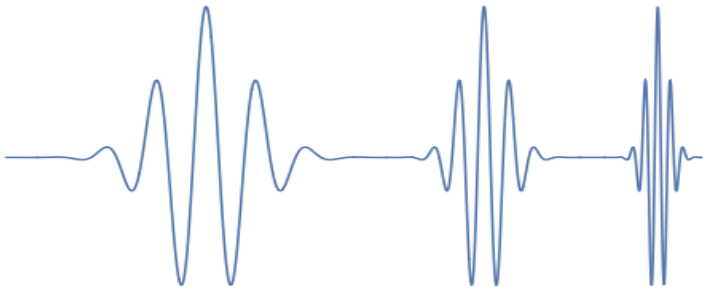


FIGURE 32 – Illustration of translation and scaling operations applied to a wavelet.

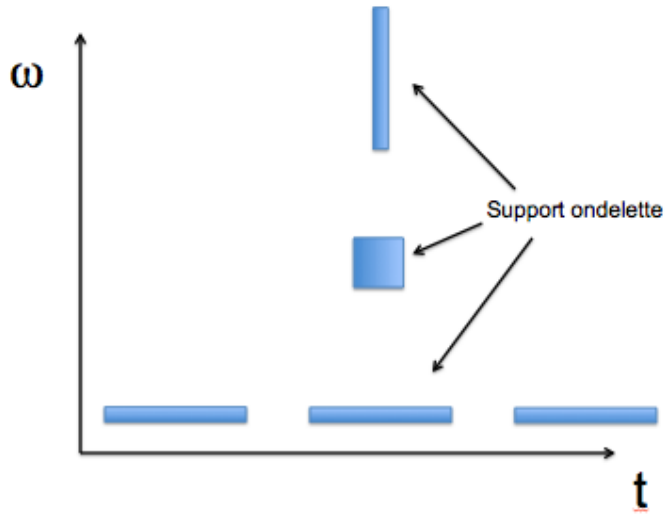


FIGURE 33 – Change in wavelet support size with frequency, and invariance with respect to time translation.

If $\psi_\lambda(u)$ is the "wavelet" function with a scale of λ, then we calculate the wavelet coefficient [94] of $X(u)$ as follows:

$$(X * \psi_\lambda)(u) = \sum_{u'} X(u')\psi_\lambda(u - u') = (W_\lambda x)(u) \tag{302}$$

This coefficient measures the fluctuation of the signal in the vicinity of u ***on a scale of approximately*** λ. How does this solve the problem of the Fourier transform and the fluctuations of $\widehat{C}(\omega)$? The key point is that ***the support of the wavelet is bounded both in time and frequency*** [95], as illustrated in Figure 33. This automatically performs the windowing mentioned in the previous paragraph to obtain consistent estimators of second-order moments.

Let the base wavelet be the following function, known as the Morlet or Gabor wa-

94. NDJE. For a detailed introduction to wavelet transforms, see, for example, the 2020 Course Sec. 7 and the 2021 Course Secs. 7 and 8.

95. NDJE. The same concentration phenomenon also occurs for 2D wavelets.

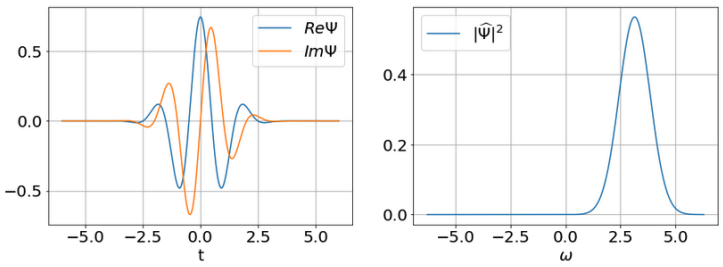


FIGURE 34 – Complex Morlet wavelet Eq. 303 with quasi-zero power spectrum for $\omega < 0$ ($\sigma = \pi$).

velet [96] (See Equation 303):

$$\begin{aligned}\psi(x) &= \frac{e^{-\frac{t^2}{2}}\left(-e^{-\frac{\sigma^2}{2}} + e^{i\sigma t}\right)}{\sqrt[4]{\pi}\sqrt{-e^{-\sigma^2} - 2e^{-\frac{3\sigma^2}{4}} + 1}} \\ \widehat{\psi}(\omega) &= \frac{e^{-\frac{1}{2}(\sigma-\omega)^2} - e^{-\frac{\sigma^2}{2}-\frac{\omega^2}{2}}}{\sqrt[4]{\pi}\sqrt{-e^{-\sigma^2} - 2e^{-\frac{3\sigma^2}{4}} + 1}}\end{aligned} \tag{303}$$

It has the particularity of having a nearly zero Fourier spectrum for $\omega \leq 0$ (admissible wavelet) and concentrated around the frequency ω_0 which satisfies the equation:

$$\omega_0 = \frac{\sigma}{1 - e^{-\sigma\omega_0}}$$

($\sigma \gtrsim 2$, then $\omega_0 \sim \sigma$). Figure 34 illustrates this for $\sigma = \pi$.

Note that one can roughly construct a basic wavelet from a window function (e.g., Gaussian) multiplied by a phase so that the power spectrum is shifted to the desired value in the $\omega > 0$ domain. The fact that $\widehat{\psi}(0) = 0$ means that the wavelet's average is zero:

$$\widehat{\psi}(0) = 0 \Rightarrow \sum_u \psi(u) = 0 \tag{304}$$

96. NDJE. Named after Jean Morlet (1931-2007) and Dennis Gabor (1900-1979). See notebook *morlet_wave_1D_2D.ipynb*

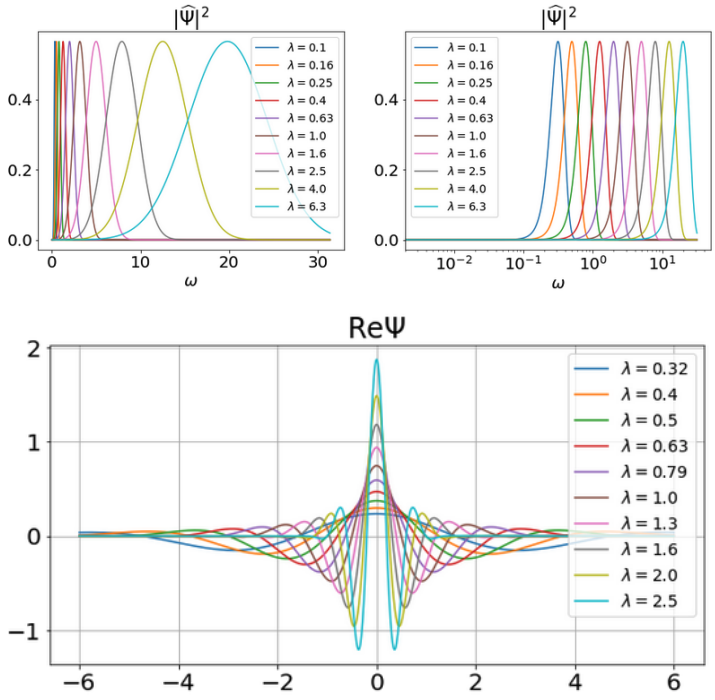


FIGURE 35 – Top image: Fourier spectrum of the Morlet wavelet (Fig. 34) for different scale factors (linear scale on the left, logarithmic on the right). Note the constant width of the spectrum in logarithmic scale. Bottom image: Morlet wavelet in real space for different scale factor values. Note that the characteristic width of the wavelet in real space changes inversely to its width in Fourier space.

Now, if we apply a scale factor λ, it follows that:

$$\psi_\lambda(u) = \lambda\psi(\lambda u) \Longrightarrow \widehat{\psi_\lambda}(\omega) = \widehat{\psi}(\lambda^{-1}\omega) \tag{305}$$

By adjusting the values of λ, we can cover the entire range of ω (Fig. 35 top). Note that the Fourier spectrum of a Morlet wavelet has a constant width in logarithmic scale. Notice a general characteristic that ***the wider the Fourier spectrum, the narrower the characteristic width of the wavelet in real space, and vice versa*** (Fig. 35 bottom).

When we perform a convolution of the signal X with a wavelet ψ_λ in real space, in

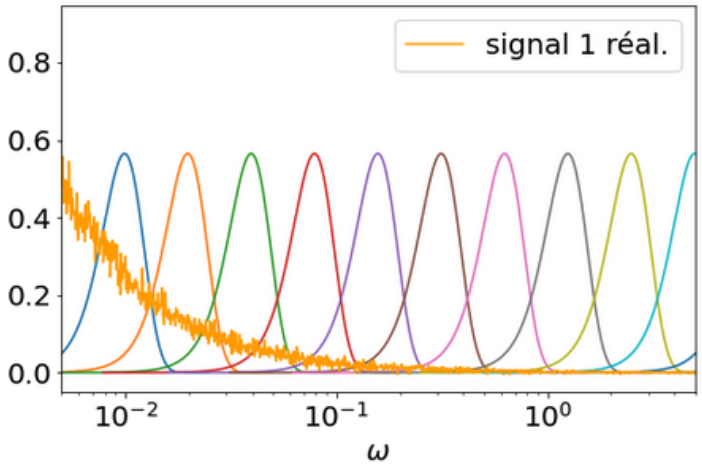


FIGURE 36 – Evolution of $\widehat{C}(\omega)$ (Fig. 30) and spectra of some wavelets obtained with different scale factors.

Fourier space, we perform the product of Fourier transforms as follows:

$$\widehat{W_\lambda x}(\omega) = \widehat{X * \psi_\lambda}(\omega) = \widehat{X}(\omega) \times \widehat{\psi_\lambda}(\omega) \tag{306}$$

So, each wavelet will capture a part of the Fourier spectrum of the signal depending on the value of λ, and at the same time, it will locally capture the signal's behavior in real space.

We will apply this type of filtering in the case of $\widehat{C}(\omega)$ as illustrated in Figure 36. ***Not only does this provide consistent estimators of the spectral power, but in the non-Gaussian case, it also provides useful information about transients***.

NDJE: Before moving on to the 2D case, let's make a small addition that S. Mallat couldn't cover due to time constraints, which you can find, for example, in Secs. 7.5 and 8.1 of the 2020 Course. When looking at the evolution of the Fourier spectra of wavelets (Fig. 35), to cover the entire range of $\omega \in [0, \infty]$, it would require an infinite number of wavelets. Considering high frequencies, we know that the spectral power of the signal decreases, so we can imagine not considering wavelets whose scale factor is beyond a cutoff $\lambda \geq \lambda_{max}$. However, at low frequencies, we cannot do the same because we know that the signal has regions of regularity (think of areas of the same intensity in

any image). Nevertheless, we can complement, for example, the family of band-pass filters $\{\psi_\lambda\}_{\lambda_{min}\leq\lambda\leq\lambda_{max}}$ *with a low-pass filter (e.g., Gaussian)* ϕ*. Then, the Fourier spectrum thus covered allows us to reconstruct the signal from the wavelet coefficients and the equivalent obtained with the* ϕ *filter, that is:*

$$Wx = (\psi_\lambda * X), \phi * X) \tag{307}$$

Finally, following the pattern of characteristic "widths/heights" of the wavelet power in the time-frequency plane (Fig. 33), it can be shown that one can discretize both scales and spatial translations as $(2^j\ n, 2^j)$ *with* $(n, j) \in \mathbb{Z}^2$*, to obtain an orthonormal basis for* $L^2(\mathbb{R})$*. In practice, in audio, it is customary to use some redundancy by taking a scale factor* $2^{j/Q}$ *for wavelets with* $Q \sim 16$*. This is related to human hearing and the concept of an octave (See Course 2020 Sec. 8.1).*

9.5 2D Filtering

NDJE. For the 2D case treated in detail, see Course 2021 Sec. 8.4.

The general principle is as follows: what can be implemented in 1D can be extended to any nD. In Fourier if $X(u = (u_1, u_2))$, then if $\omega = (\omega_1, \omega_2)$ and $\omega.u = \sum_{i=1,2} \omega_i u_i$, synthetically, we have

$$\widehat{X}(\omega) = \sum_u X(u) e^{-i\omega.u} \tag{308}$$

As in the 1D case, we need to design an admissible wavelet whose power spectrum is limited to a region in the Fourier plane. One possible solution is to mimic the philosophy of the 1D Morlet/Gabor wavelet (see the notebook *morlet_wave_1D_2D.ipynb*):

$$\begin{aligned}
\psi(u_1, u_2) &= \left(-e^{-\frac{\xi^2}{2}} + e^{i\xi u_1}\right) e^{-\frac{u_1^2+u_2^2}{2\sigma^2}} \\
\widehat{\psi}(\omega_1, \omega_2) &= \pi\sigma^2 e^{-\frac{1}{2}\sigma^2\omega_2^2} \left(e^{-\frac{1}{2}\sigma^2(\omega_1-\xi)^2} - e^{-\frac{1}{2}(\xi^2+\sigma^2\omega_1^2)}\right)
\end{aligned} \tag{309}$$

The parameter σ controls the characteristic width of the wavelet, while ξ sets the position of the power maximum in Fourier. An illustration is given in Figure 37.

Now, as in 1D, we can change the scale λ (taken dyadic, $\lambda = 2^j$ with $j \in \mathbb{Z}$), but we

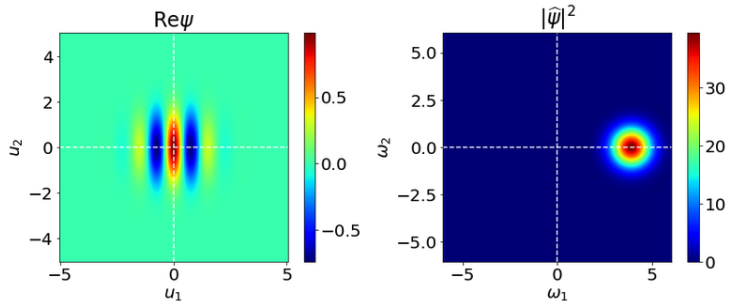


FIGURE 37 – Basic 2D wavelet (Eq. 309) with its Fourier spectrum localized around $\omega_1 \approx \xi$ ($\sigma = 1$, $\xi = 5/4\pi$), with zero power in the $\omega_1 \leq 0$ plane.

can also apply a rotation by an angle θ. It follows [97]

$$\psi_{j,\theta}(u) = 2^{-2j}\psi(2^{-j}r_\theta.u) \Longrightarrow \widehat{\psi_{j,\theta}}(\omega) = \widehat{\psi}(2^j r_{-\theta}.\omega) \tag{310}$$

An illustration of the combined action of scaling and rotation is given in Figure 38.

Then, by combining a collection of wavelets with different parameters j and θ, we can cover the Fourier plane just like in 1D. Figure 39 illustrates this. Of course, we can adapt the base wavelet and the number of rotations to fill any gaps left here just to make the different associated "blobs" visible. However, like in 1D, to cover the low frequencies (the central zone, $\omega \sim (0,0)$), we use a low-pass filter $\phi(\omega)$ like a simple Gaussian.

So, in the end, ***the image (signal) is passed through a series of band-pass filters and a low-pass filter*** ($\theta_k = 2\pi k/K$) just like in 1D:

$$Wx = ((\psi_{j,\theta_k} * X)_{j\in[\![j_{min},j_{max}]\!],k\in[\![0,K-1]\!]}, \phi * X) \tag{311}$$

which allows analyzing the signal at all scales and all orientations. It's also essential to consider translations in real space to have a complete view of transients across the entire image. Note that a particular orientation of a wavelet is sensitive to transients along the perpendicular axis.

97. NDJE. *Firstly*, regarding the 2020 Course Sec. 8.2, I changed the definition of the sense of rotation, which is purely arbitrary. Here, $r_\theta = \left(\begin{smallmatrix}\cos\theta & -\sin\theta \\ \sin\theta & \cos\theta\end{smallmatrix}\right)$. *Secondly*, the normalization factor is indeed λ^2 because we have 2 coordinates.

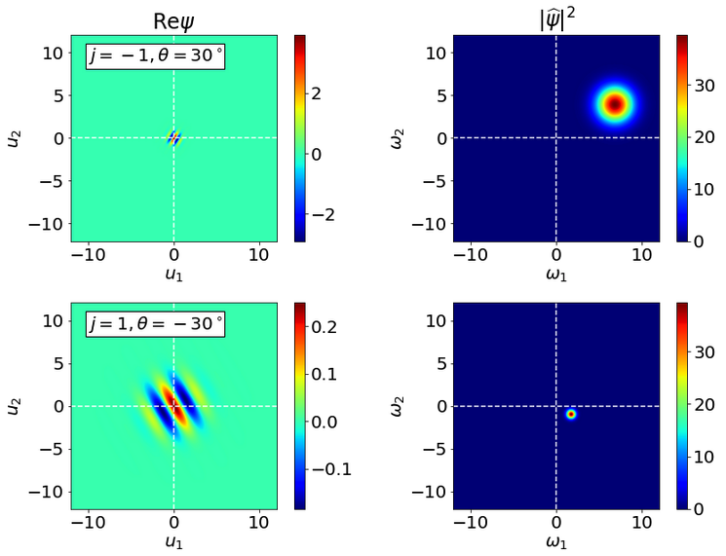


FIGURE 38 – Effect of scaling and rotation on the basic 2D wavelet (Eq. 310): for the top row $j = -1$ and $\theta = 30°$, while for the bottom row $j = 1$ and $\theta = -30°$.

9.6 Examples of Usage

NDJE. S. Mallat projects some examples of wavelet decompositions; I'll try to reproduce them here. Some are from section 2.9.2, which I include here for clarity.

An example of decomposing an audio signal is shown in Figure 40 (also see simple examples in the notebook *wavelet1D.ipynb*). The horizontal axis represents time, while the vertical axis represents frequencies. However, since the size of the band-pass filters is scale-independent in logarithmic scale (e.g., Morlet wavelet), and the position of the maximum is related to the scale, the vertical axis also represents $\log \lambda$. At each point in this plane, the color represents the magnitude of Wx, with blue indicating a value of 0. It's worth noting that the majority of coefficients are zero, and only a few coefficients matter. ***Thus, we have a representation in the time-(log scale) plane of the positions of transients in the signal*** (at what time and at what scale λ). In this case, we can describe the different "attacks" of notes with their fundamental frequencies and harmonics. Thus, we can see the structures of the signal that allow us to discriminate/identify specific instruments, etc.

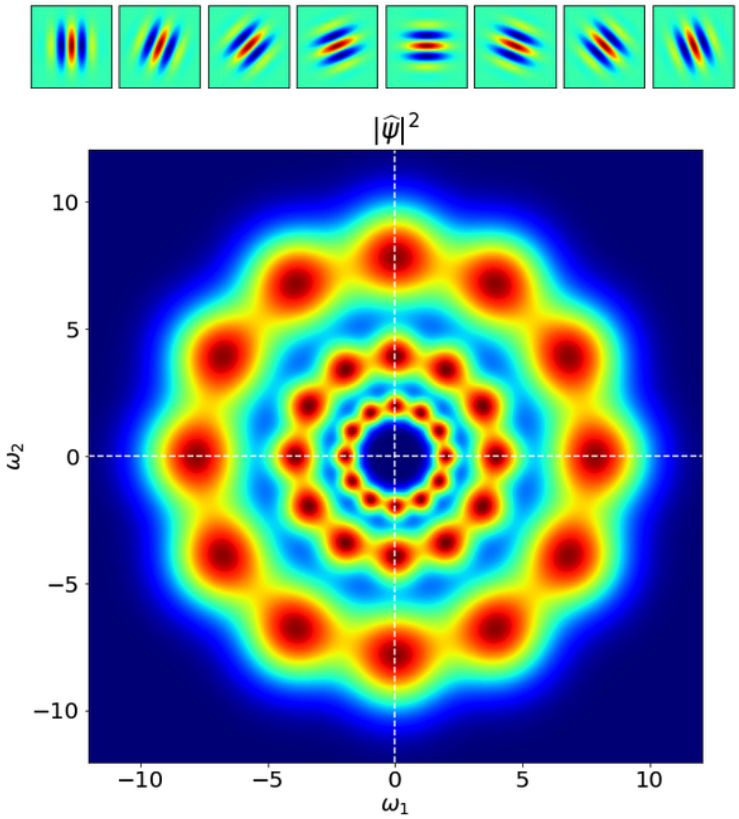


FIGURE 39 – Top: (Real part of) wavelet after rotations of $k\pi/8$ with $k \in [\![0, 7]\!]$. Bottom: Coverage of a portion of the Fourier plane by collecting the spectra of several wavelets with scale factors $j = (-1, 0, 1)$ and undergoing multiple rotations of 30°.

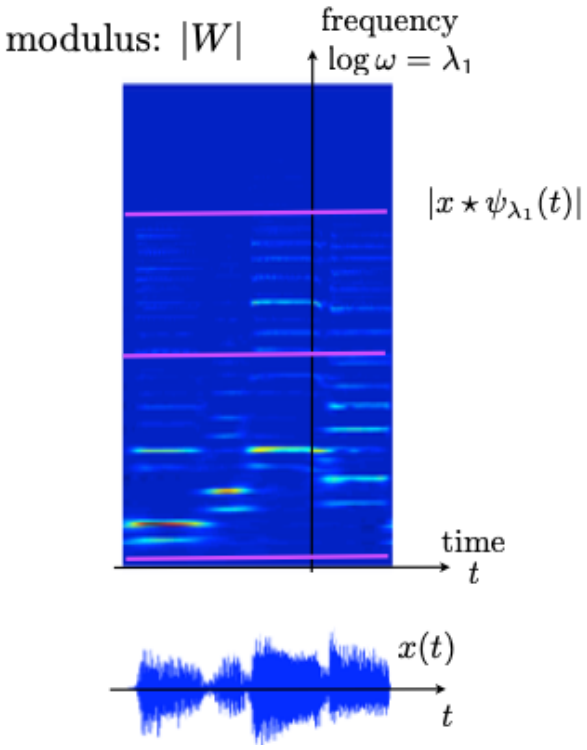


FIGURE 40 – Example of wavelet decomposition of a signal $x(t)$: the color represents the value of $\|Wx\|$ with blue indicating zero values. The sparsity of the frequency decomposition and its temporal evolution can be observed.

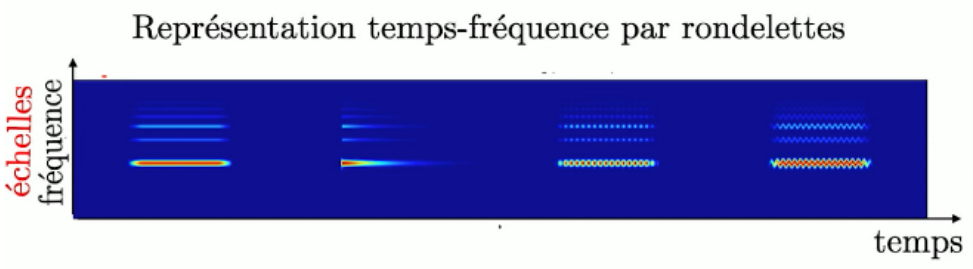


FIGURE 41 – Time-frequency representations of four music sequences (NDJE. the "r" in "rondelettes" is a friendly typo).

The next example is shown in Figure 41, where a note is played by several musical instruments. Wavelet analysis provides a detailed representation of the structures, highlighting the importance of having a network of numerous time-scale filters. ***Thus, starting from a signal*** X ***with*** d ***temporal values, we project it into a higher-dimensional space, which is a general method in Machine Learning***. However, in this case, only a few coefficients contain information that is subsequently used for analysis (e.g., classification...).

The third example concerns Figure 42, which shows three representations of "applause." We can perform a wavelet decomposition of the original sound $X(u)$ and represent it in the time-scale plane as before. However, we can try to reproduce the sound by modeling it as a Gaussian process, capturing the first and second-order moments of the original sound at each discrete instant. Thus, in a time slice, we can model the probability density $p(x)$ of the signal as follows:

$$\tilde{p}(x) \sim \exp\{-\frac{1}{2}(x-\mu)^T C^{-1}(x-\mu)\} \tag{312}$$

and generate a new realization $X'(u)$, which is then analyzed again with wavelets. We notice that the distribution of power across frequencies (scales) of the signal over time is well preserved. However, ***we have lost the signal's structures that are not captured by the mean and covariance***, which we now understand because the Gaussian process cannot account for them. So, we need to find a way to capture the missing structures. The third representation will be explained later in the session.

Similarly, consider a 2D turbulent flow phenomenon (Fig. 43, left image). From $d = 6\ 10^4$ samples of the signal X, we can use $\mu = \mathbb{E}(X)$ and n second-order moments

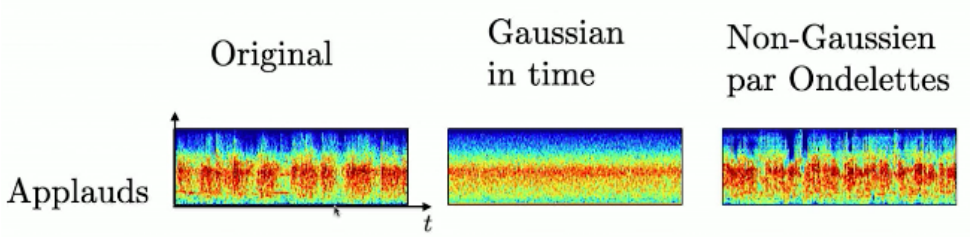


FIGURE 42 – Time-frequency representations of applause: the original sound on the left; in the middle, generation of a Gaussian model that preserves the original energy at all scales/frequencies; on the right, generation by a more realistic model that considers correlations between scales.

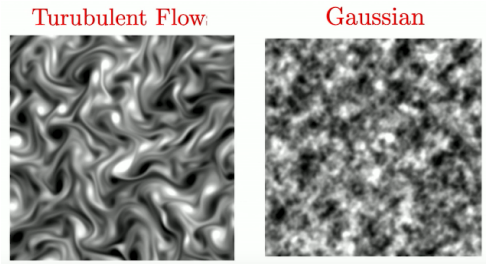


FIGURE 43 – Example of a simulation of 2D turbulent flow (left) and the realization (right) of a Gaussian model that simplifies the problem by considering only the second-order moments of the distribution on the left, i.e., by considering only the covariance matrix.

$C(k) = \mathbb{E}(\phi_k(X))$ defined as $\phi_k(x) = (k.x - \mu)(x - \mu)$ with $k.x$ being a translation of x by $k \in \mathbb{R}^2$. We can then build a Gaussian model of the form

$$\tilde{p}(x) \sim \exp\left(-\sum_k \beta_k \phi_k(x)\right)$$

However, the right image in Figure 43 shows the result of a realization from this Gaussian model. It's evident that while the scales of fluctuations seem preserved, ***the geometry of the structures is not that of the original signal***. Therefore, the problem in physics, but in signal processing in general, is how to capture the signal's structures beyond second-order moments.

When using 2D wavelets with different dyadic scale factors, orientations, and spatial translations, we "explode" a typical image into all these filtering channels and obtain small images that can be represented as shown in Figure 44. It can be observed that at each scale, the orientation of the wavelets makes orthogonal transitions in the original image visible. ***This allows the detection of boundaries between regions of equal intensity, which are the sought-after transient phenomena.***

Notice that this way of operating the various wavelet filters closely resembles ***the structure of convolutional neural networks***.

What we also observe here in 2D, and as we noticed in 1D, is that as soon as there are transients between very regular ranges, the wavelet coefficients are essentially zero, except for a few that are highly localized in the time-scale plane. This is what we call ***sparsity***. Moreover, between scales, we visually notice that the structures are correlated; there is ***a lot of scale-to-scale dependence***. How can we capture these dependencies using statistical quantities?

9.7 Sparsity

NDJE: See also the 2021 course Section 5.1. for the case of an orthonormal basis where thresholding is applied to the decomposition coefficients. Here, S. Mallat uses a different approach within the context of mean estimators.

Suppose we want to use $\mathbb{E}(|X|)$ and $\mathbb{E}(|X|^2)$; we can use estimators (here, u can be

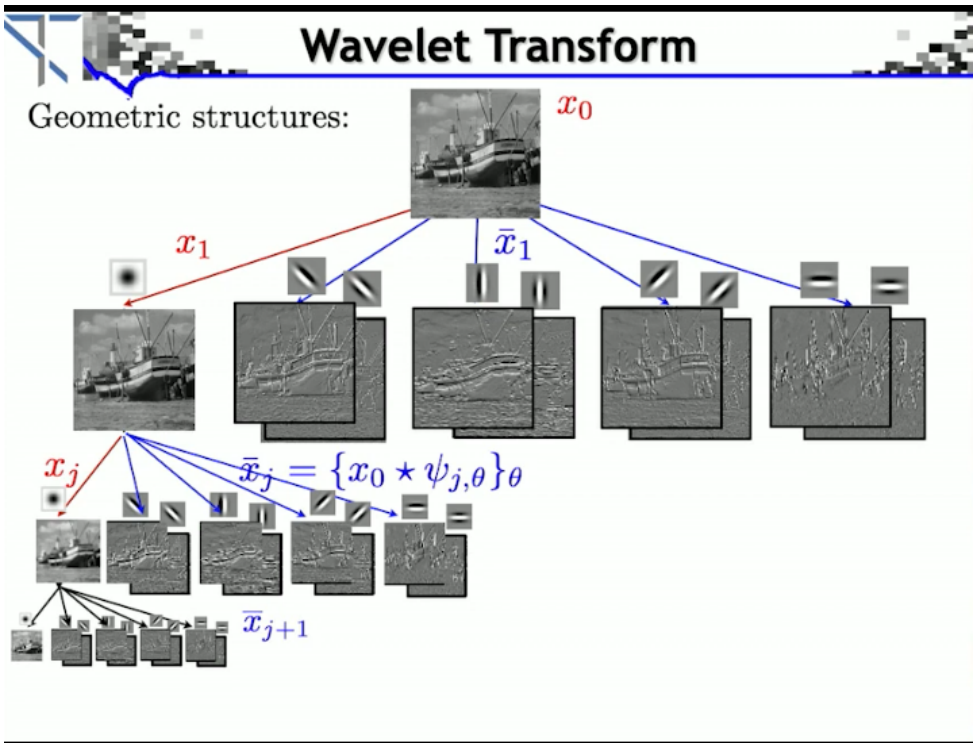


FIGURE 44 – Successive cascade of low-pass (ϕ) and band-pass ψ filter applications at different scales and rotations. At a given scale and rotation, all coefficients obtained through translations are collected to reconstruct image patches. At each change in scale, pooling/averaging by a factor of 2 in both directions is performed, changing the size of the image patches. Dark areas indicate regions where coefficients are zero. The *wavelet2D_sparsity.ipynb* notebook allows you to calculate the first stage of decomposition with a discrete Haar wavelet (Course 2012 Sec. 8.4).

seen as the index of a pixel in 2D):

$$\tilde{\mu}_1 = \frac{1}{N}\sum_{u=1}^{N}|X(u)| \qquad\qquad \tilde{\mu}_2 = \frac{1}{N}\sum_{u=1}^{N}|X(u)|^2 \tag{313}$$

What does the ratio $\tilde{\mu}_1^2/\tilde{\mu}_2$ tell us?

Property 9

$$\frac{1}{N} \leq r = \frac{\tilde{\mu}_1^2}{\tilde{\mu}_2} \leq 1 \tag{314}$$

The case where $r = 1$ corresponds to a constant signal, and $r = 1/N$ when the signal is concentrated on 1 pixel in the case of an image with 1 temporal sample in 1D.

The proof of $r \leq 1$ can be done cleverly using Cauchy-Schwartz:

$$(\sum 1 \times |x(u)|)^2 \leq (\sum 1^2) \times (\sum |x(u)|^2).$$

The first inequality is due to the fact that the square of the sum of absolute values is always greater than the sum of the squares of absolute values.

So, ***the value of r somehow measures the sparsity of the signal*** X: very sparse when r is small, or not sparse at all for $r = 1$. What about a Gaussian signal? For $x \sim \mathcal{N}(0, 1)$ in 1D, then $(\mu_1, \mu_2) = (\sqrt{2/\pi}, 2)$, so $r_g = 2/\pi$. In 2D for $x \sim \mathcal{N}(0, \mathbf{1}_{2\times 2})$ (also for a complex Gaussian), $(\mu_1, \mu_2) = (\sqrt{\pi/2}, 2)$, so $r_g = \pi/4$. What about an image, like the sub-images of the decomposition in Fig. 44? ***If $r < r_g$, then we have sparsity information that is much greater than for a Gaussian process***. ***Note: The notebook wavelet2D_sparsity.ipynb provides an example and values of $r \approx (0.20 - 0.25) < r_g$, illustrating a sparsity of wavelet coefficients much greater than a Gaussian process.***

These notions of sparsity were widely used in the 1980s-2000s, S. Mallat tells us. They are very useful for coding, for example (see the 2022 course Section 9.5), but they provide global information. They are not sufficient to constrain a model for generating a new image of a boat, for example.

As mentioned in the previous section, what needs to be captured are the ***correlations across scales***.

9.8 Interactions between Scales

Understanding these interactions between different scales is truly *the fundamental point* of this entire field of Mathematics and Physics.

By using the various wavelet coefficients, we can study the correlation:

$$C(u, u', \lambda, \lambda') = Cov((X * \psi_\lambda)(u), (X * \psi_{\lambda'})(u')) \tag{315}$$

A property tells us that if the supports of the Fourier transforms of the two wavelets are disjoint, then the correlation coefficient is zero:

$$C(u, u', \lambda, \lambda') = 0 \quad \text{if} \quad \widehat{\psi_\lambda)}(\omega) \times \widehat{\psi_{\lambda'}}(\omega) = 0 \tag{316}$$

This result comes from the filtering of processes. Let X be a stationary random process; we can calculate the convolution with a deterministic filter h. Now,

$$(g_\tau X)(u) = X(u - \tau) \Longrightarrow g_\tau(X * h) = (g_\tau X) * h \tag{317}$$

This is the manifestation of convolution equivariance (Definition 21). And since $g_\tau X = X$ (stationary), $g_\tau(X * h) = X * h$, so $X * h$ is a stationary process. In particular,

$$\mathbb{E}[(X * h)(u)] = \mathbb{E}\left[\sum_v X(u - v)h(v)\right] = \mathbb{E}[X] \sum_v h(v) \tag{318}$$

In the case where h is an admissible wavelet, $\sum_v h(v) = 0$ (Eq. 304), and thus,

$$\mathbb{E}[(X * h)(u)] = 0 \tag{319}$$

Regarding correlations between $X * h$ and $X * g$ taken at 2 different points, then [98]

$$Cov((X * h)(u), (X * g)(u')) = (C * h * \tilde{g})(u - u') = C_{hg}(u - u') \tag{320}$$

with $C(\tau) = \mathbb{E}(X(u)X(u - \tau))$ the two-point correlation function whose Fourier transform is the power spectrum $\hat{C}(\omega)$ (measured, for example, in cosmological analyses of the cosmic microwave background), and $\tilde{g}(u) = g^*(-u)$.

98. Note: By writing the left-hand side and remembering that the means are zero, we obtain a double sum over (v, v'), for example, in which the expectation follows $\mathbb{E}(X(v-u)X^*(v'-u') = C(v-v'+u'-u)$ in the complex weighted case of $h(v)$ and $g^*(v')$. By expressing what $(a * b * c)(x)$ is in a general way, we then reveal the double convolution estimated at $u' - u$. Hence the result when using the definition of $\tilde{g}$.

This tells us that for a stationary process X, if we want to estimate the correlation between two admissible wavelet coefficients, it amounts to filtering the covariance of the process by the corresponding two filters. So, if the supports of $\hat{h}$ and $\hat{g}$ do not overlap, $\widehat{C_{hg}}(\omega) = 0$. Hence the result.

In the concrete example of Figure 44, if we take the coefficients of one of the sub-images denoted $\bar{x}_1$ and its equivalent at $\bar{x}_j$ at a scale 2^j where $j \neq 1$, then the correlation factor is zero. Even though everything seems to suggest otherwise. Why does this happen? ***It's due to destructive interference caused by the phase difference between the two objects***. So, ***if you only want to calculate the correlation between linear measurements, you end up with a failure***, as you obtain no information through the correlation coefficients.

Moreover, it should be noted that if X is a Gaussian process, and $A = X * h$ is a linear combination of Gaussian random variables, it is also a Gaussian random variable. So, if $B = X * g$, we know that A and B are two Gaussian random variables with zero correlation for g and h having disjoint supports. Therefore, A and B are ***two independent random variables***, and all higher-order moments are also zero. ***So, in the Gaussian case, decorrelation means independence between random variables, hence the absence of structure***. This explains the results on the generation of signals (audio or turbulent) by Gaussian models (Figs. 42, 43).

9.9 Scattering Network

To go further and find a way to capture the structure while using correlations, we need to handle phases. ***We can use a non-linearity*** like $|x|$ or alternatively [99] $ReLU(x)$.

Note that if $\psi(u)$ has a Fourier spectrum concentrated around $\omega = \omega_0$, then ψ_λ has its spectrum concentrated around $\omega = \lambda\omega_0$, and the same goes for $X * \psi_\lambda$. So, if we multiply [100] by $e^{-i(\lambda\omega_0)u}$, the spectrum shifts to be concentrated around $\omega = 0$. We can perform this shift for $X * \psi_{\lambda'}$ while leaving $X * \psi_\lambda$ intact if $\lambda' < \lambda$, so that the two spectra partially overlap. However, this is not enough to eliminate destructive interference. But now we can use non-linearity. Considering $|X * \psi_{\lambda'}|$ and $X * \psi_{\lambda'}$, not only do the spectra

99. The ReLU, which cancels the negative part of the signals, indeed works.
100. See, for example, Table 1 Sec. 3.4 in the 2012 lecture.

overlap, but the correlation is non-zero. Thus, we can consider the following covariances, with scales $\lambda = 2^j$ and $\lambda' = 2^{j'}$ such that $j \neq j'$:

$$\begin{cases} Cov((X * \psi_j)(u), |(X * \psi_{j'})(u - \tau)|) & = C^{(1)}(j, j', \tau) \\ Cov(|(X * \psi_j)(u)|, |(X * \psi_{j'})(u - \tau)|) & = C^{(2)}(j, j', \tau) \end{cases} \tag{321}$$

In the Gaussian case, there's no escape from the fact that $C^{(1)} = 0$ and $C^{(2)} = 0$ due to the independence of variables mentioned in the previous section. However, in the non-Gaussian case, *a priori*, we have $C^{(1)} \neq 0$ and $C^{(2)} \neq 0$.

Notice that if we have a signal with N pixels (or discrete samples), we have N possible values of τ, while the number of scales [101] λ (or λ') is of the order of $\log_2 N$. So, ***the number of moments is of the order of*** $N \log_2^2 N$. For example, if $N = 10^6$, this is $4\ 10^8$ coefficients, which is far too many to calculate so many moments. Remember, *first*, we only have one image, and *second*, we want consistent estimators that require weighted averaging over ranges of values. In short, ***we need to find a way to reduce the number of moments***.

We have pointed out that since $X(u)$ is stationary, $|X * \phi_j|(u)$ are stationary processes. In practice, these wavelet coefficients are sparse in the sense that almost all of them are zero except for those corresponding to the transients of $X(u)$. So, by taking their absolute values, we have a new signal that can be analyzed by scale-j_2 wavelets by calculating the coefficients:

$$(|X * \psi_j)| * \psi_{j_2})(u) \tag{322}$$

Then, we only need to consider the following covariances:

$$Cov((|X * \psi_j)| * \psi_{j_2})(u), (|X * \psi_{j'})| * \psi_{j_2})(u)) = C^{(3)}(j, j', j_2) \tag{323}$$

meaning that we don't need to involve the parameter τ, which dramatically reduces the combinatorial complexity. These coefficients are, ***a priori***, non-zero only in the non-Gaussian case.

101. See, for example, Lecture 2018 Sec. 6.6.0.3 for values of j_{max}.

These iterations (cascades) of wavelet convolutions form what S. Mallat calls ***scattering networks*** [102] (also see Lecture 2020 Sec. 9.5). From a network architecture perspective, you can refer [103] to Figure 45:

1. For the top image: Starting from an image x, you begin by applying a cascade similar to the one presented in Figure 44, which yields the green sub-images (low-pass filtering by ϕ) and the blue ones (filtered by ϕ_j), connected by green and black arrows. Then, you apply a non-linearity $\rho(x) = |x|$ in this case, but it could be a ReLU to all coefficients of the sub-images. The next step is to consider each blue sub-image, applying an identical cascade as before, with a new low-pass filter and a set of band-pass filters at a different scale j_2 (red arrows). The notebook *scattering2D.ipynb* allows obtaining coefficients of the form $|x * \psi_\lambda| * \phi$ and $||x * \psi_\lambda| * \psi_{\lambda'}| * \phi$ in 2D. The application of the low-frequency filter allows for translation invariance (Lecture 2020 Sec. 9.4).
2. For the bottom image: At each step, it's necessary to renormalize the coefficients (equivalent to BatchNorm) to condition the problem correctly. Once we have different wavelet coefficients, we can calculate correlations at a given scale (red lines), which in the language of CNN corresponds to ***correlations between channels***. These are the moments that we keep, and their number is roughly $K = \log_2^3 N$, or about 8,000 for $N = 10^6$.

This type of Scattering network is a kind of convolutional network where the weights (i.e., the filters) are known in advance (wavelets). There is no learning here. We impose the wavelets because we have ***prior knowledge***. We want to capture ***transients***, so we need ***spatially localized filters***; furthermore, since the signals are ***stationary***, we need ***consistent moment estimators***, which means we need to average over appropriate frequency ranges. So, ***the filters must also meet this dual requirement***. Inevitably, we are led to consider ***wavelet bases***. Then, to obtain interactions between scales, we need to eliminate phases, hence the need for ***non-linearities*** like the absolute value or ReLU. Thus, the network architecture is ultimately natural.

What does this look like in practice for images like the turbulent flow in Figure 43?

---

102. NDJE. J. Bruna and S. Mallat *"Scattering Convolution Networks"* (2012) https://arxiv.org/pdf/1203.1513.pdf
103. NDJE. I'm taking two images projected by S. Mallat of the same network to explain the process.

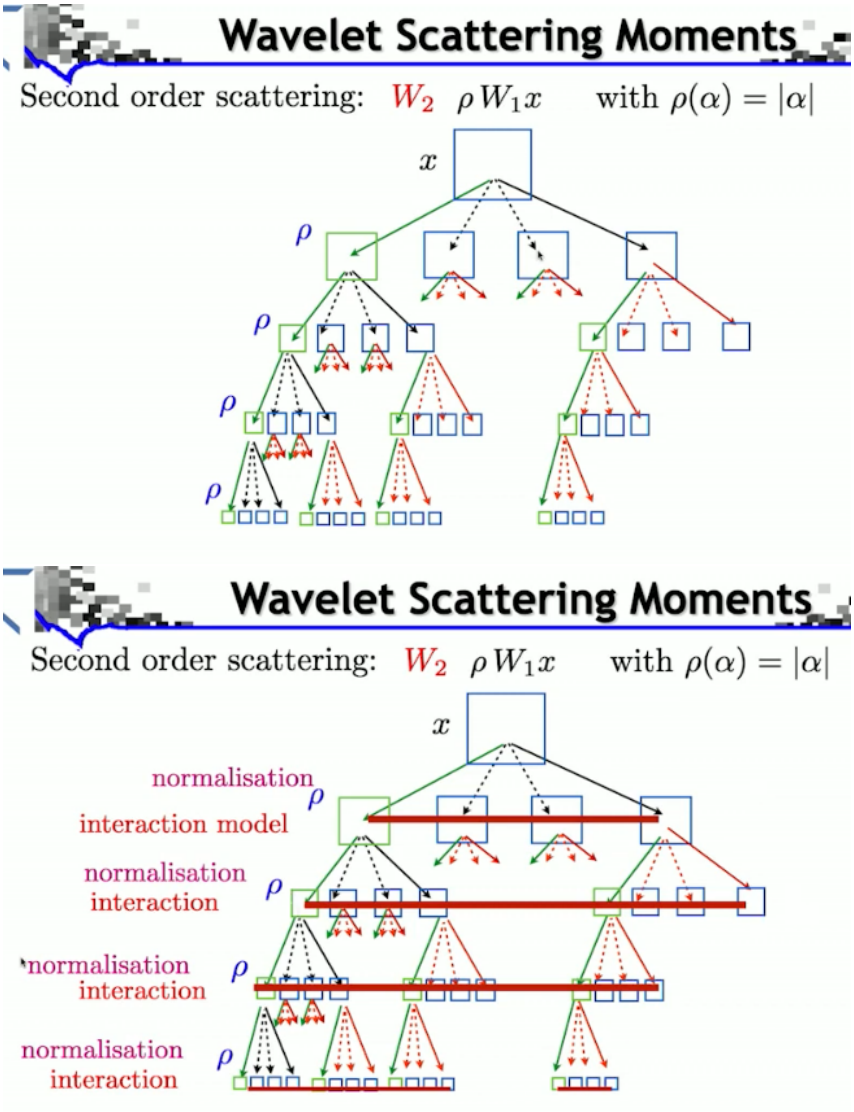


FIGURE 45 – Please refer to the text for descriptions of the images.

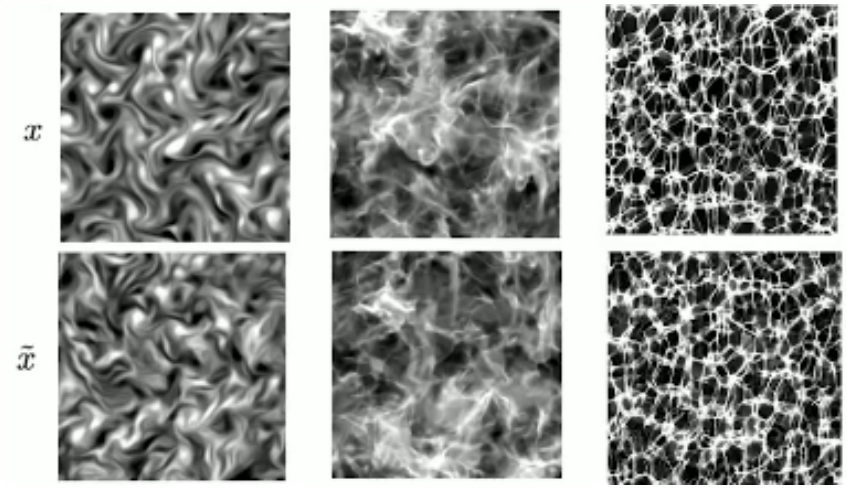


FIGURE 46 – Top row: Original images. Bottom row: New realizations obtained from maximum entropy models based on the coefficients $C^{(3)}(j, j', j_2)$ of a scattering network that calculates correlations between channels. But, limiting it to $\log_2 N$ moments.

With the scattering network, we extract moments [104] $C^{(3)}(j, j', j_2) = \mathbb{E}(\phi_k(X))$, and then we can create a model of $p(x)$ that maximizes entropy, resulting in:

$$\tilde{p}(x) = Z^{-1} \exp\left(- \sum_{k=1}^{K} \beta_k \phi_k(x) \right) \tag{324}$$

We obtain the Lagrange multipliers β_k. Then, we can generate new realizations $\bar{x}$.

The results are shown in Figures 46 and 47. ***We are able to capture the structures unlike the Gaussian model***. And when the original images have high resolution, we can consider more moments. The model is even more faithful and allows generating realizations whose moment statistics are faithful to the originals up to the 4th order.

The description of the Ising model is also non-Gaussian due to local interactions, but also because we impose spin values of ± 1. Here too, the description by scattering networks that provide correlations between scales allows us to capture the multi-scale structuring of the process.

104. NDJE. Here, covariance $Cov(a(u), b(u))$ means $\sum_u a(u)b(u)$ since the means are zero due to the renormalization of coefficients.

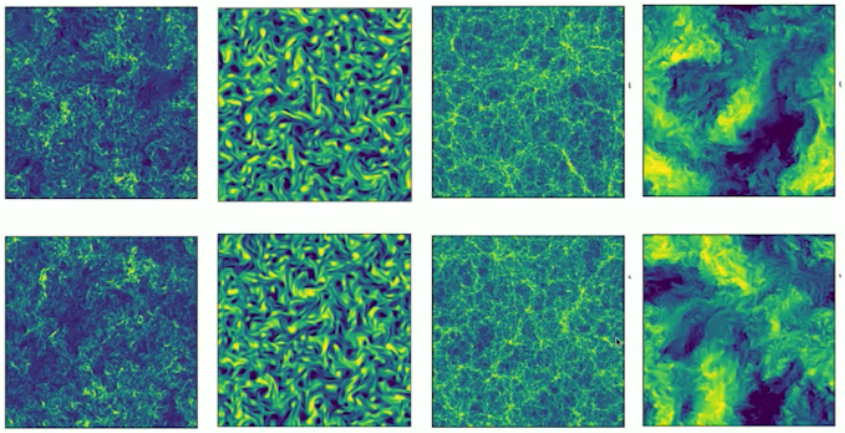


FIGURE 47 – Similar exercise (Fig. 46) of generating new realizations from 2nd-order moments calculated by a scattering network, but considering more coefficients because the original images have higher resolution. In this case, we can reproduce statistics up to the 4th order.

In the case of sounds, we obtain a new synthesis similar to the one on the right side of Figure 42. It works equally well; to the ear, we can recognize the structures of the different original sounds.

10. Epilogue

Is there a limit to describing a phenomenon using 2nd-order coefficients obtained from scattering networks, which, as a reminder, have no learned weights? What about processes like faces? The short answer is: ***everything we've discussed in this course collapses***. The question is why? The ergodicity assumption is false, particularly due to non-stationarity (all images are recentered). But let's assume that one image of a face is a realization of the "face" process; the properties of this process are much more complex than those of the physical processes considered earlier.

Let's summarize the course briefly. ***The central concept is that of entropy***. We can view it as a measure of uncertainty, but we know it's a count of the discrete states the system can take. We were able to define a ***rate of entropy*** for ***stationary processes***,

which was particularly useful in the case of Markov chains (Sec. 7.3). But in the case of a system that becomes increasingly large ($n \to \infty$) for which we don't have these ergodicity/stationarity assumptions, it becomes very complicated to predict what happens. In particular, will entropy fluctuate or converge?

We know that if ***entropy converges***, then we can derive many results. For example, we observe the asymptotic distribution in typical spaces with almost uniform probability, and the number of elements in these sets is given by entropy. So, entropy is the variable at the heart of the system's description. We also saw that the temporal evolution of such a system converges to a maximum entropy (2nd law of thermodynamics) corresponding to the stationary measure (equilibrium) in the Markovian case. In the end, we can summarize by saying that in this case ***we have control over the mathematical properties of the system***. In the Gaussian case, it's even simpler because there is a lack of structure; in the non-Gaussian case over the last 15 years, as S. Mallat tells us, we have been able to master the subject as well (e.g., scattering networks) by understanding scale interactions.

However, in the case of faces, what do we do? As S. Mallat says, we may wonder if there's an underlying process for the "production" of a face image? If not, we cannot apprehend the statistical properties. At some point, we need to calculate expectations, i.e., empirical means in practice, so we need several samples from the same process... But if each image had its own "history," it would not be feasible. However, ***we have algorithms*** (see Valentin de Portoli's 2023 seminar) ***that work spectacularly***: consider Generative Adversarial Networks, Variational Auto-Encoders, or diffusion models... So, ***there is likely an underlying mathematical model***. This is the current area of research.

When we look at the results of these generative models, we see a ***memorization phenomenon***, meaning that parts of the generated image are found in the database. This is not at all what we observe when we proceed with the maximum entropy modeling scheme. When we do so, we favor diversity since we spread the probability over all elements of a typical set. In contrast, in GANs, VAEs, etc., it seems that diversity is not maximized, nor is it simply the trivial case of randomly drawing an element from the database. There is variability without maximizing entropy because we find these structures present in the database. Clearly, a model guided by the principle of maximum entropy would have zero probability of generating even a small portion of the image identical to that in the database. ***Therefore, we must abandon the principle of maximum entropy***. Note that if

we force these generative models to increase entropy, the quality of the images deteriorates, and fine textures disappear (e.g., hair).

According to S. Mallat, we should try to build a bridge between a simple model of face memorization and probability modeling by the maximum entropy principle, which imposes strong ergodicity properties on the process. This is one of the themes discussed by Marc Mézard (*Statistical Physics and Inference: The Challenge of Structured Data*) during this year's seminar.

In conclusion, S. Mallat tells us that about 99% of the activities of researchers and doctoral students are focused on developing architectures to compete boldly to increase performance. However, there is another aspect that is equally fascinating, which is trying to understand what makes a particular architecture successful.

NDJE. (lightly adapted from John Zarka's thesis). "Finding the right balance between theory and practice is always a subtle exercise. Some theoretical frameworks can explain the basic properties of deep architectures well but are limited in practical applications or rely on too restrictive assumptions. On the other hand, recent deep architectures that achieve state-of-the-art results on large-scale problems have reached such a level of technical complexity that it seems difficult to open these black boxes and expose the mathematical foundations that lead to their impressive performance."

So, S. Mallat's message is *"we can try to do less well while trying to understand better."*.