Spot Topics for Inference and Statistical Physics

Simone Maria Giancola[1†]

†Bocconi University, Milan

May 19, 2024

[1] simonegiancola09@gmail.com

List of Symbols

The list collects basic symbols used in the document.

Physics

$\langle \cdot \rangle_{.}$	expectation in the Boltzmann distribution
$\mathfrak{F}$	free energy
$\mathscr{N}$	number of micro-states, degeneracy function
$\mathscr{E}$	energy
$\mathscr{O}$	observable
$\mathscr{S}$	Gibbs entropy, Boltzmann entropy
$\mathscr{U}$	internal energy
k_{B}	Boltzmann constant
T, β	temperature, inverse temperature

Statistics and Mathematics

$(\Omega, \mathcal{F}, \mathbb{P})$	probability space
$\boldsymbol{X}, \boldsymbol{x}$	random vector, vector
$\mathcal{Z}$	partition function
$\boldsymbol{\delta}$	Dirac Delta distribution
$\mathbb{E}_{.}[\cdot]$	expectation in a distribution that is not Boltzmann
$\mathcal{H}(\cdot)$	entropy
$\mathcal{H}(\cdot, \cdot)$	cross entropy
$\mathscr{F}[f](\cdot)$	Fourier Transform of f
$\mathscr{L}[f](\cdot), \mathscr{L}_{\pm}[f](\cdot)$	Laplace Transform of f, bilateral Laplace transform
$\mathfrak{L}[f](\cdot)$	Legendre-Fenchel transform of f
$\mathscr{X}$	space of configurations, phase space
$\mathbf{A}$	matrix and random matrix
Ω	sample space
$\phi_X(\cdot)$	Characteristic Function of X
$K_X(\cdot), \kappa_n$	Cumulant Generating Function of X, n^{th} cumulant
$M_X(\cdot)$	Moment Generating Function of X

Contents

Foreword

While diving deep into the intricacies of Statistical Physics, I often found that I was not prepared for some sub-topics and terminology appearing. This document is an attempt to collect part of the aspects that might *bridge the gap*. The intention is not to be comprehensive, but rather to be as self-contained as possible in the discussion of interesting topics.

Clear emphasis is given to techniques and observations that are needed to treat the intersection of Inference and Statistical Physics, which is eventually what I had time to focus on. Said so, this is not the right document if the reader is interested in the branch of the field that still focuses on the purely Physical aspects. For example, a strong influence was the course (Krzakala and Zdeborová 2021), the book (Mezard and Montanari 2009), or in general the works of those researchers that sit at the boundary between Inference, Information Theory, Physics, and Machine Learning.

The topic is very fascinating, wide, and complex. Given that here we focus only on *giving the underlying ideas*, there will be very little emphasis on modern uses of the tools. The reasons are twofold:

1. I wanted to write down a relaxed treatment of what is considered introductory knowledge in the field;

2. the document itself would have been too large and dispersive.

In some sense, the following Chapters could be taken as a quick crash-course on each of the subtopics for inexperienced readers like me, who cannot make sense of some terms and concepts when they first read them.

For a discussion of one of the many leaves of research where Statistical Physics is useful, I have written a thesis (link), of which these pages are effectively a spin-off. **Even more importantly**, any reference cited in the Chapters is a valuable source of interesting and wonderful works. With some care, I have made the effort of combining Publications, Books, Blogs, and Conference papers, as to provide something for every reading taste.

Note for the reader I am very much happy to receive feedback. This is a draft and I do wish to add a Chapter on Graphical Models with some nice results, and maybe also on the REM. . .

Content Outline Chapter I is a brief introduction to Thermodynamics, which one of the fields where the first relevant questions for modern Statistical Physics originated. In Chapter II we collect more formal statements about objects and techniques often used in practice, with an emphasis on those required for replica computations, which is unfortunately not treated here in detail. In Chapter III we zoom out again, and reinterpret some previously stated results in the context of Information Theory, giving further justification towards the fact that all fields have a very wide overlap, obscured by different terminologies, and different questions.

Chapter I

Thermodynamics, Statistical Mechanics

This Chapter is a Physics-style discussion of the foundations of Thermodynamics/Statistical Mechanics. It serves the purpose of introducing terminology and concepts that are not common in Machine Learning. Nice resources that approach the field independently are (S. J. Blundell and K. M. Blundell 2009, Chaps. 1, 2, 4), the lecture notes (Schwartz 2021), especially Lecture 7 and 8, or the manuscript (Tong 2012), which is more comprehensive. Since the subject is very old and attractive, the list is not exhaustive.
In terms of style, we will be as concise as possible and leverage intuition. The practical purpose is providing a definition of entropy and a quick treatment of ensembles and their importance.
In Section I.1, the building blocks of Thermodynamics are presented, to later arrive at Sec. I.2, where Temperature is introduced. Then, we attempt to give a self-contained description of ensembles, in Section I.3. To conclude, we present a discussion on the notion of mean-field approximation and one type of inequality which is widely used under the rug.

Preliminaries Throughout $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space. Emphasis on n is placed especially when the size of the sample will matter. When we need no distinction and the probability space is clear from context, we will use $\mathbb{P}$. As always, random variables may take values $\boldsymbol{X} : \Omega \to \mathscr{X}$, where $\mathscr{X} \subseteq \mathbb{R}^d$. In particular, we choose the following construction, which appears in (Ellis 2006, Sec. I.1). Despite being more general than what we will need, it provides a good understanding of how the probability space is designed and is of great importance for large deviations. We will introduce the concept in Section III.5.
A system is an idealized experiment where a random process is isolated from its surroundings. In a system, there will be a collection of random variables $\{\boldsymbol{X}_i : i \in \mathscr{I}\}$, where $\mathscr{I}$ is an index, defined on the probability space $(\Omega, \mathcal{F}, \mathcal{P})$ taking values in a common *state space* $\mathscr{X}$. Any result will depend solely on the random variables, so a natural choice is to take the sample space as the product space $\mathscr{X}^{\mathscr{I}}$, and $\{\boldsymbol{X}_i\}_{i \in \mathscr{I}}$ is its *coordinate representation process*. Mathematically, for an $\vec{\boldsymbol{\omega}} = \{\boldsymbol{\omega}_i\}_{i \in \mathscr{I}} \in \mathscr{X}^{\mathscr{I}}$ denoting the sample space realization across the index[1] , we let $\boldsymbol{X}_i(\vec{\boldsymbol{\omega}}) = \boldsymbol{\omega}_i$ which is the i^{th} coordinate (vector) of $\vec{\boldsymbol{\omega}}$. The interpretation is that $\Omega = \mathscr{X}^{\mathscr{I}}$ is the set of all possible *micro-states*, i.e. all possible ways in which the phenomenon of randomness realizes, and $\vec{\boldsymbol{\omega}} \in \Omega$ will be a configuration/micro-state of the system. For a suitable choice of $\mathcal{P}$, the construction is valid. We provide the canonical example below, and a simplification in terms of coin tossing.

Example I.0.1. *Adapted from (Ellis 2006, Ex. I.1.1)*
Choose as index set $\mathscr{I} = \mathbb{Z}$, or for a realistic experiment $\{1, \dots, n\}$, which is a subset of it. Let $\mathscr{X} \subseteq \mathbb{R}^d$. Define $\mathcal{P}$ to be the product measure of $\mathscr{X}^{\mathscr{I}}$ of identical

[1] Notice that we somehow abuse notation and use the vector notation on bolds $\vec{\boldsymbol{\omega}}$. This is to be understood as $\vec{\boldsymbol{\omega}}$ being a vector of vectors. Nevertheless, all the construction will be needed only tangentially throughout the document.

d-dimensional marginals on the space $\mathscr{X}$. For the set of integers, it is an infinite product measure, for the experiment set, it is finite. We are describing a set of iid random variables. As a byproduct, given $\mathscr{I}, \mathscr{X}$ it suffices to specify one-sample marginals $\mathcal{P}$ to obtain $\mathbb{P} = \bigotimes_{i \in \mathscr{I}} \mathcal{P}$ uniquely. To conclude that the second example makes sense in the construction, the conclusions of (Ellis 2006, Chap. II) are needed.

In particular, let $\mathscr{I} = \mathbb{Z}$, $\mathscr{X} = \{0,1\}$. The marginals are on a binary space. If they are all identical, then $\mathbb{P} = \bigotimes_{i \in \mathscr{I}} \mathcal{P}$ for a choice $\mathcal{P} \equiv \mathcal{B}ern(\alpha)$ with $\alpha \in [0,1]$. In other words, we are describing an infinite sequence of coin tosses.

In particular, let $\mathscr{I} = \{1, \ldots, n\}, \mathscr{X} = \mathbb{R}^d$, take $\mathcal{P} \equiv \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$. Then $\mathbb{P}$ describes the law of n iid standard multivariate Gaussians.

I.1 Postulates

We place ourselves in the *metaphorical scenario* of a thermodynamic system. For the beginning, we will take the perspective of energy. Theoretically, Energy will be the result of a configuration of random variables with a well defined function. We will reconcile the two notions when “defining” the Hamiltonian, which is the right formalism for the above construction. If it causes difficulties, energy can be directly taken as a function of $\boldsymbol{X}$.

Before getting into concrete matters, we will consider four foundational principles.

(P1) **Energy** $\mathscr{E}$ is a scalar quantity that describes a thermodynamic system. It is a *thermodynamic property* that is to some extent *measurable* or *observable*.

(P2) An *isolated system* is not allowed to interact with external entities. We can think of it as a perimeter with a strong boundary at *fixed volume* that allows for no exchange of energy. Its energy is the sum of a kinetic term and a potential term. The former concerns particle-level energy contributions, the latter is a term accounting for relative positions and interactions intra-particles. A *closed system* does not exchange particles with the environment but can exchange energy.

(P3) An *ideal gas* is a system of n non-interacting particles.[2] Its internal energy is thus only the sum of the individual kinetic energies.

(P4) *Heat* is the phenomenology of energy in transit. For two systems allowed to exchange energy, we say they are in *thermal contact*. Experimentally, heat flows to reach *equilibrium*, with an equal share rather than a polarized scenario.[3]

It is fundamental to pair these principles with some of the famous laws of thermodynamics. While these can be justified on many grounds (e.g. see (S. J. Blundell and K. M. Blundell 2009) or any book on the subject), we will largely avoid such discussion, since it is not relevant.

Zeroth Law of Thermodynamics

> Two systems in equilibrium with a third are in equilibrium with each other. In other words *transitivity* applies.

We consider non-deterministic systems with n particles, where n is large. Due to randomness, a collection of possible realizations in a space Ω is available. We also assume that physical constraints force the statistician to access the configuration only at a global scale, which might however come from many different realizations. Different realizations are called *micro-states*, while what is observable is a *macro-state*. In other words, given the overwhelmingly numerous amount of particles, we will be interested and allowed to measure only macro-states.

[2] the assumption of Molecular Chaos and coarse Graining in Ehrenfest’s and Maxwell’s work (Ehrenfest et al. 1960; Maxwell 1860), (Schwartz 2021, Lecture 3)

[3] Namely, “hot” flows into “cold”, but notice that we are not defining temperature yet! For now, it can be visualized as energy exchange.

In the first appearance of Thermodynamical theories, it was justified by the scaling of the playground. A Physicist deals with $n \sim 10^{23}$ particles, of the order of Avogadro's number, which is the number of atoms in a mole of matter. With such a number of degrees of freedom, it is hopeless to describe the model with Newtonian mechanics, and one must resort to *probabilistic arguments*.
Despite the seemingly tailored adaptation, we will see that many results adapt by analogy to other theoretical frameworks, and more interestingly than others, to high dimensional data. We give a taste of these large scale phenomena in the example below.

Example I.1.1. *Consider a fair coin, where each realization $x_i \in \{0,1\}$. For n throws, we represent a single outcome $\omega \in \Omega_n$ as $\boldsymbol{x}(\omega) = (x_1(\omega), \dots, x_n(\omega))$, which is a string of digits. While every binary string is equally likely with probability (wp) $\frac{1}{2^n}$, there are $\binom{n}{k}$ ways of having exactly k heads. For n big, the probability of having k heads is not uniform in k, since there is not a single occurrence for which k heads are sampled. Consider $n = 20$, $k = 5$, then:*

$$\binom{n}{k} = 15504 \quad \mathbb{P}\left[\sum_{i=1}^{n} x_i = k\right] = \frac{15504}{2^{20}} \approx 0.14 \gg \frac{1}{2^n} \approx 9.53 \cdot 10^{-7}. \tag{I.1.2}$$

In a thermodynamics setting, such notion is justified by the fact that if we have $n \gg 1$ particles, it is very unlikely that we will be able to measure e.g. the velocity of each of them individually. Physicists thus pursue a treatment of the subject that deals with aggregate phenomena stemming from the small scale into the big scale. On a different perspective, the peculiarity of exploding numbers spurs from the small scale indistinguishability, which mathematically translates in factorial expressions that blow up quickly as n grows.

Remark I.1.3. *Notice that we are finding the number of micro-states by fixing the macro-state of observation. It is easy to understand why this choice is made: since for each micro-state there will be only one macro-state, while each macro-state refers to many micro-states. Therefore, the other direction would be meaningless: the function that maps micro-states to macro-states is trivial, though difficult to measure, while the function that maps macro-states to micro-states multiplicity is not friendly, but easy to derive.*

Definition I.1.4 (Observable). *A function $\mathscr{O} : \mathscr{X} \to \mathbb{R}$, which ideally represents the outcome of an experiment. Given the randomness, it will eventually be a random function. An observable is a quantity, a function, a macroscopic property. All names are equivalent.*

Having outlined the starting scenario, we now develop in parallel the treatment of first-principles and their direct consequences.

I.2 A Statistical Definition of Temperature

Consider two systems in thermal contact. When external contributions are negligible, these are thought of as an *isolated system*, for which energy is conserved. We write:

$$\mathscr{E} = \mathscr{E}_A + \mathscr{E}_B + \mathscr{E}_{AB} \qquad \mathrm{d}\mathscr{E} = 0 \tag{I.2.1}$$

to enforce these constraints, where the pedices denote that individual contributions are summed, and the interaction energy $\mathscr{E}_{AB}$ is negligible.[4] Implicitly, we are leveraging another law.

First Law of Thermodynamics

Energy is conserved and may just appear in different forms. The law is equivalent to the more famous statement “No energy is created or destroyed”.

[4] In practice, this is the energy on the layer of contact. For sufficiently large volumes, the surface is negligible.

It must be stressed that Energy is in units of measure, so a different choice of units of measure leads to equivalent notions. Fixing the sum of the two energies to any value thus implies that to know a macro-state of the system it is sufficient to know the value of $\mathscr{E}_A$. If we denote as $\mathscr{N}$ the number of micro-states of a system, system A can be in any of its $\mathscr{N}_A(\mathscr{E}_A)$ micro-states and system B in any of its $\mathscr{N}_B(\mathscr{E}_B)$.

Example I.2.2. *In a probabilistic sense, one can see that $|\Omega| = \mathscr{N}$, where the two are allowed to depend on $(n, \mathscr{E})$. In the two systems setting, only $\mathscr{N}_A$ varies as $\mathscr{E}_A$ varies. Such fact is justified by the observation that the probability space is fixed, and the macroscopic property is itself random. The way of letting $\mathscr{X}$ and $\mathscr{N}$ vary is crucially different. It is done ab-initio, the other is during the calculations. Basically, a choice $(n, \mathscr{E})$ will lead to a system. While we could interpret the evolution at each n, for our purposes, we will define problems at finite n, but when considering a Thermodynamically isolated system, we will see its behavior at $n \to \infty$.*

As already observed in Example I.1.1, many micro-states may map to a single macro-state energy. Hence, it makes sense to introduce an object that collects configurations with the same behavior up to measurements.

Definition I.2.3 (Degeneracy function)**.** *Recall that $\mathscr{E}$ is a scalar function, assume it takes values on a space $E \subseteq \mathbb{R}$. The degeneracy is then a function $\mathscr{N} : E \subseteq \mathbb{R} \to \mathbb{N}$ such that $\mathscr{N}(\mathscr{E}(\omega))$ counts the number of micro-states that attain such energy level. If energy takes continuous values, we take with some care $\mathscr{N}(\mathscr{E})$ to represent the number of micro-states with energy in $[\mathscr{E}, \mathscr{E} + \delta\mathscr{E})$ for a small variation $\delta\mathscr{E}$ of the energy, which is reasonable as long as we implicitly assume that our measurement tools have accuracy less than $\delta\mathscr{E}$.*

Example I.2.4 (Avogadro Scaling of Degeneracy)**.** *For $n_A = 10^{23}$ particles, with possible individual states $\{0, 1\}$, there are $2^{10^{23}}$ arrangements.*

When two systems are in contact, there are $\mathscr{N}_A(\mathscr{E}_A)\mathscr{N}_B(\mathscr{E}_B)$ possible micro-states by a simple combinatorial argument. We imagine that the two are left in contact for enough time t as to reach an equilibrium, which in our case could be regarded with the expressions $\mathrm{d}\mathscr{E}_A = 0, \mathrm{d}\mathscr{E}_B = 0$ (as time dependent quantities).

Assumption I.2.5 (Fundamentals)**.** *We (reasonably) assume that in thermal equilibrium:*

(A1) ***Postulate of equal a priori probabilities:*** *each micro-state is equally likely;*[5]

(A2) ***Variability****: micro-states are continuously changing in time;*

(A3) ***Ergodicity****: for $t \to \infty$ all micro-states will be visited for an equal amount of time. It could also be implied by a detailed balance condition and irreducibility. An argument with such flavour is found in (Jaynes 1957).*

Remark I.2.6. *What is referred to as ergodic hypothesis mostly regards statement (A3). We do not discuss much about this matter, but it is common to take it as granted. The crucial advantage of ergodicity is that* ***time-averages*** *become* ***sample-space*** *averages, and the Central Limit Theorem (CLT) can be used for a sequence of observations in time t rather than in $\omega \in \Omega$. Without it, any computation would be on different grounds, as we are doomed to observe only one ω for each sequence of observations.*

From these foundations, which can be added to (P1)-($P4$), we are ready to state another law.

Second Law of Thermodynamics

The macro-state of an isolated system in Thermal Equilibrium is the one with highest multiplicity of accessible micro-states associated. By *accessible* we mean in the timescale of the observation

[5] With reasonable work, this can be seen as a consequence of Boltzmann's H Theorem (Schwartz 2021, Lecture 3)

Example I.2.7. *Consider the fair coin example, and ignore punctual estimations, it is sensible that by the CLT the average of $n = 10^6$ throws will be very close to $\frac{1}{2}$. In principle exactly $\frac{1}{2}$ is not true at finite n, but if we allow for some error and let n be very big, the almost sure convergence will give the result.*

Coming back to our two systems, we aim to maximize the product of micro-states with respect to (wrt) the energy values at fixed total energy. Mathematically

$$\max_{\mathscr{E}_A : \mathrm{d}\mathscr{E}=0} \mathscr{N}_A(\mathscr{E}_A)\mathscr{N}_B(\mathscr{E}_B), \tag{I.2.8}$$

which by usual rules of differentiation leads to the system:

$$\begin{cases} \dfrac{\mathrm{d}\mathscr{N}_A(\mathscr{E}_A)\mathscr{N}_B(\mathscr{E}_B)}{\mathrm{d}\mathscr{E}_A} = 0 \\ \mathrm{d}\mathscr{E} = 0 \end{cases} = \begin{cases} \mathscr{N}_A(\mathscr{E}_A)\dfrac{\mathrm{d}\mathscr{N}_B(\mathscr{E}_B)}{\mathrm{d}\mathscr{E}_B}\dfrac{\mathrm{d}\mathscr{E}_B}{\mathrm{d}\mathscr{E}_A} + \mathscr{N}_B(\mathscr{E}_B)\dfrac{\mathrm{d}\mathscr{N}_A(\mathscr{E}_A)}{\mathrm{d}\mathscr{E}_A} = 0 \\ \mathrm{d}\mathscr{E}_A + \mathrm{d}\mathscr{E}_B = 0 \end{cases}. \tag{I.2.9}$$

From the second equation (i.e. the constraint of energy conservation), we derive $\mathrm{d}\mathscr{E}_A = -\,\mathrm{d}\mathscr{E}_B$, which substituted into the first gives the following expression:

$$-\mathscr{N}_A(\mathscr{E}_A)\frac{\mathrm{d}\mathscr{N}_B(\mathscr{E}_B)}{\mathrm{d}\mathscr{E}_B} + \mathscr{N}_B(\mathscr{E}_B)\frac{\mathrm{d}\mathscr{N}_A(\mathscr{E}_A)}{\mathrm{d}\mathscr{E}_A} = 0 \tag{I.2.10}$$

multiplying by $\frac{1}{\mathscr{N}_A(\mathscr{E}_A)\mathscr{N}_B(\mathscr{E}_B)}$ and using the identity $\frac{\mathrm{d}\ln f(x)}{\mathrm{d}x} = \frac{1}{f(x)}\frac{\mathrm{d}f(x)}{\mathrm{d}x}$ one gets

$$\frac{\mathrm{d}\ln \mathscr{N}_A(\mathscr{E}_A)}{\mathrm{d}\mathscr{E}_A} = \frac{\mathrm{d}\ln \mathscr{N}_B(\mathscr{E}_B)}{\mathrm{d}\mathscr{E}_B}. \tag{I.2.11}$$

In simple words, two systems isolated from the rest, in contact and at thermal equilibrium will *most likely* be found with a share of the total energies that obeys Eqn. I.2.11.

Remark I.2.12. *Enforcing that energy is conserved and maximizing with respect to the energy of the first system is implicitly assuming that system B admits an energy $\mathscr{E} - \mathscr{E}_A$. It is not in principle guaranteed, but can be if we get back to the reasoning that the energy levels are so fine grained that our tools are not able to be as accurate as their description.*

At this moment, it is worth considering what is the notion of thermal equilibrium in day to day life. Empirically, temperature is a quantity that we measure. In houses there are no energy scales, but rather thermometers, and if two objects are put into contact for enough time they will eventually *average out* their temperatures. Temperature is a rather arbitrary notion[6] like any other quantity in units of measure, but we can formally give it an origin from the exact equilibrium condition. Thus we let:

$$\frac{1}{k_{\mathrm{B}}T} := \frac{\mathrm{d}\ln \mathscr{N}(\mathscr{E})}{\mathrm{d}\mathscr{E}}, \quad k_{\mathrm{B}} = 1.3807 \cdot 10^{-23} JK^{-1} \tag{I.2.13}$$

which formalizes the object and allows for mathematical treatment. Here k_{B} is a constant which is derived by how other objects are defined in terms of units of measure. Since the result is dependent on the reference system, and we are not interested in experimental measurements that depend on Physical units of measure, we might as well make the simplification of taking $k_{\mathrm{B}} = 1$. Note also that by construction $T \geqslant 0$.

Remark I.2.14. *We avoid most of the times the notation T for the more amenable $\beta := \frac{1}{T}$, which might refer to a "final time". This will allow for a unified treatment with the same symbol all around. Obviously, when $T \to 0$ we have $\beta \to \infty$ and when $T \to \infty$, $\beta \to 0$. There two are in a bijection. In rare cases, we will use T and make it explicit.*

Definition I.2.15 (Boltzmann Entropy)**.** *For ease of notation in an isolated system with equally likely micro-states let*

$$\mathscr{S} := k_{\mathrm{B}} \ln \mathscr{N}(\mathscr{E}), \tag{I.2.16}$$

so that we can express Eqn. I.2.11 as $\beta = \frac{\mathrm{d}\mathscr{S}}{\mathrm{d}\mathscr{E}}$.

[6] what is it in principle? Where does the concept of temperature start at all?

Such construction leads to a mathematical statement of the first law of thermodynamics, which reads:

$$\mathrm{d}\mathscr{E} = T\,\mathrm{d}\mathscr{S} - P\,\mathrm{d}V \tag{I.2.17}$$

where (P, V) are pressure and volume. The statement holds in a variety of cases, most importantly with reversible changes and in absence of phase transitions. We will discuss the latter, and avoid dealing with the former. For simplicity, it can be taken to be true *with some care*. On the other hand, the more general statement that $\mathrm{d}\mathscr{E} = \text{đ}\mathscr{Q} - \text{đ}\mathscr{W}$ where $\mathscr{Q}, \mathscr{W}$ are heat and work and the symbol đ denotes an inexact differential is always true but requires different notions. A starting point for the approach is any classical thermodynamics book such as (S. J. Blundell and K. M. Blundell 2009). The discussion is still interesting once the focus moves to a finer analysis of the second law of Thermodynamics.

Remark I.2.18 (An equivalent statement of the second law)**.** *The second law of thermodynamics says that entropy tends to maximize. Taking two systems, not in contact with entropies $(\mathscr{S}_A, \mathscr{S}_B)$, one then derives that when put in contact:*

$$\mathscr{S}_{(A,B)} \geqslant \mathscr{S}_A + \mathscr{S}_B, \tag{I.2.19}$$

which brings to the statement that $\mathrm{d}\mathscr{S} \geqslant 0$. A direct proof is by arguing that the micro-states of A, B are a subset of the micro-states of (A, B) combined. In "microscopic" words, allowing a system for more freedom increases its number of accessible states. However, this is a subtle matter when doing statistical mechanics. Entropy tends to increase, and only increases effectively when the system size is large. We will always consider cases in which the limiting effect holds.

While Boltzmann entropy is just one possible example that restricts to the uniform distribution over micro-states, it is already of great importance. First of all, it can be shown that it implies the thermodynamic definition of entropy change in terms of heat (Schwartz 2021, Lecture 6), which is yet another statement of the second law. Secondly, it has strong links with other definitions of entropy. To partially deal with such matter it is worth introducing descriptive terms of quantities that are common in Physics but not in other fields.

Definition I.2.20 (Extensive and Intensive quantities)**.** *Consider a system A with size n. An observable $\mathscr{O}(x)$ or a property of it is:*

- *extensive if it is proportional to n, i.e. $\mathscr{O}(x; n) \propto n$. It is also called extrinsic.*
- *Intensive if it is independent of n, i.e. if $\mathscr{O}(x; n) \equiv \mathscr{O}(x)$. It is also called intrinsic.*

The definition is somehow sloppy, it could be that extensivity is verified also when proportionality with system size is of higher order. <u>We do not allow this</u>. An observable $\mathscr{O}(x; n) \propto n^2$ is <u>not extensive</u>. To make the claim clearer, extensivity is found by requiring additivity for sub-systems. Then, we give a nicer definition that accounts for all of these facts at the thermodynamic limit $n \to \infty$, which is in the end what we implicitly study. It will however need the notion of density, which is discussed in Remark I.2.23 below.
A quantity is

- *extensive if $\mathscr{O}(x; n) = n o(x) + o(n)$, i.e. it is made of a size times size-independent term and a vanishing term in system size*
- *intensive if $\mathscr{O}(x; n) = O(1)$, i.e. it is constant wrt system size*

Definition I.2.21 (State Quantities)**.** *In a thermodynamic system, a state quantity is a descriptor of the current "appearance" of the system. More formally, it is independent of the dynamics that brought the system to its current form.*

Example I.2.22 (Ideal Scenario)**.** *In general, one can hope/verify/construct structures such that: Energy is extensive, Entropy is extensive, Temperature is intensive. All of them are state variables. An example of a variable that is* **not** *a state variable is heat, but we will not deal much with it.*

Remark I.2.23 (Extensivity implies density notion)**.** *When a quantity is extensive, its thermodynamic limit $n \to \infty$ diverges. Evaluating it is problematic, but the definitional property is helpful. Let an observable be extensive, then*

$$o(x) := \lim_{n \to \infty} \frac{\mathscr{O}(x; n)}{n} \tag{I.2.24}$$

is a well defined object. It is the per-particle equivalent of the original observable, and it is of constant order.

Definition I.2.25 (Conjugate quantities)**.** *Some extensive quantities are conserved (e.g. energy), others are allowed to vary (e.g. entropy). Notably, if we split a system into subsystems (i.e. consider an isolated system, at fixed volume), changes in energy will always be balanced out, while changes in entropy need not to. When the latter happens, an intensive quantity of reference varies as well. In the current entropy-energy comment, it will be temperature. Another example is volume and pressure (volume being extensive, pressure being intensive). When such a pairing between two quantities is found, these are called conjugate, and are denoted inside the brackets $\{\cdot, \cdot\}$.*
Developing additional tools, we will formally see that the relation $\{\mathscr{S}, \beta\}$ holds.

Remark I.2.26 (Entropy Desiderata)**.** *It can be noticed that Entropy as per Def. I.2.15 is extensive. Indeed when we considered two systems there were $\mathscr{N}_A(\mathscr{E}_A)\mathscr{N}_B(\mathscr{E}_B)$ possible micro-states, with associated entropy $\ln(\mathscr{N}_A(\mathscr{E}_A)\mathscr{N}_B(\mathscr{E}_B)) = \mathscr{S}_A + \mathscr{S}_B$. By being the logarithm of a natural number, it is also positive. By analogy, any candidate notion for a comparison must at least satisfy these principles.*

I.3 Ensembles

We are now drawn to a presentation of how micro-states constructions lead to different global scale phenomena. Recall Ex. I.2.2. The set of possible micro-states, i.e. the sample space Ω, will depend on the choice of some parameters *ab-initio*. In the construction carried out to define Boltzmann's Entropy and the temperature we built it such that total energy, number of particles and volume were fixed. Furthermore, it is not the only choice, and different constructions will lead to different collections of micro-states, also known as *ensembles*. An ensemble is made of systems, that can be thought of as different identical copies (i.e. realizations $\omega \in \Omega$) of a probabilistic system. They were first introduced by (Gibbs 1878). We restrict ourselves to two types:

- in the *micro-canonical* ensemble each system has fixed energy, volume and sample size: this is what allowed us to define entropy;
- in the *canonical* ensemble each system is in contact with a large (imaginary) *reservoir of heat* with which it can exchange energy, with volume, temperature and sample size fixed.

It is crucial to understand that the two are just toy examples and serve for the purpose of understanding what will be the equations that describe the system, especially the probability law. Therefore they must be interpreted as thought experiments.

I.3.1 Probability Distributions

Let $\mathscr{N}(n, \mathscr{E}) \equiv \mathscr{N}(n)$. We ignore the volume since it will always be fixed, and will not serve for any discussion.
Since we treat the subject from a Thermodynamics perspective, there is not much to say here in the micro-canonical case. We already know that there are equal a priori probabilities so:

$$\mathbb{P}[\boldsymbol{x}] = \frac{1}{\mathscr{N}(\mathscr{E})}. \tag{I.3.1}$$

In the most immediate terms, the important object in the micro-canonical ensemble is the entropy, not the probability distribution, which is trivial by assumption. On the other hand, a system obeying such construction presents non-trivial properties arising from first principles, and is fundamental to aid the derivation of the canonical probability law, as we will see in the next discussion. To vindicate the usefulness of micro-canonical ensembles independently of others, we report one early result in the box below.

Maxwell-Boltzmann Distribution

In a micro-canonical ensemble of non-relativistic classical mechanical particles with mass m at equilibrium the distribution of velocities $\vec{\boldsymbol{v}}$ in space is:

$$\mathrm{d}\mathbb{P}\left[\vec{\boldsymbol{v}}\right] = d\pi v^2 \left(\frac{m\beta}{2\pi}\right)^{\frac{3}{2}} e^{-\frac{\beta m v^2}{2}}\, \mathrm{d}\vec{\boldsymbol{v}} \tag{I.3.2}$$

Moving to the canonical ensemble, we remove the constraint on energy, which is now allowed to fluctuate. The construction starts with the observation that fixing part of the energy to $\mathscr{E}_i$ we get a number of possible micro-states of $\mathscr{N}(\mathscr{E}-\mathscr{E}_i)$ with associated probability $\mathbb{P}[i] = \frac{\mathscr{N}(\mathscr{E}-\mathscr{E}_i)}{\mathscr{N}(\mathscr{E})}$. Thus, in such a thermodynamical system one can say that $\mathbb{P}[i] \sim e^{-\beta\mathscr{E}_i}$, where i is a collection of states subjected to some constraint on the energies allowed.

It will be shown that the above concept can be greatly generalized to non-physical particles, if one considers the celebrated *principle of maximum entropy* by Jaynes (Jaynes 1957). However, to stick to the simplest discussion possible, we keep the timeline to **before** Information Theory.

The heat reservoir exchanges energy with the system. The system alone is then a closed system, and the dynamics of energy are established by letting it in thermal contact with the reservoir.[7]

Before going deeper, we stress again that when we consider such perspective, all is a thought experiment. Accepting this, we imagine that the bath is so big relative to the system that, despite exchanging energy, the exchange is negligible on its scale, while greatly noticeable on the scale of the system. In some sense, zooming out, the system + reservoir is interpreted as a micro-canonical ensemble where $\mathrm{d}\mathscr{E} = 0$, whilst still most of $\mathscr{E}$ is in the bath. Clearly, the temperature change of the whole system will be negligible, since most of the contribution is from the reservoir, but we are now allowing the energy to vary through the process in the scale of the system.

In this context, we wish to evaluate the probability distribution of the energy of the system $\mathscr{E}_{sys}(\omega)$, where we explicitly denote its randomness. Again recognizing that the total energy is fixed, it holds $\mathscr{E}_{res}(\omega) = \mathscr{E} - \mathscr{E}_{sys}(\omega)$. For the moment, to avoid using ω explicitly, we use lowercase $\varepsilon, \varepsilon_{sys}, \varepsilon_{res}$ letters, to denote that we are dealing at fixed ω. It is then safe to say that:

$$\begin{align}
\mathbb{P}\left[\omega \in \Omega | \mathscr{E}_{sys} = \varepsilon_{sys}\right] &= \mathbb{P}\left[\omega \in \Omega | \mathscr{E}_{res} = \varepsilon - \varepsilon_{sys}\right] \tag{I.3.3}\\
&\propto \mathscr{N}_{res}(\mathscr{E}_{res} = \varepsilon - \varepsilon_{sys}) \tag{I.3.4}\\
&= \mathscr{N}_{res}(\mathscr{E} - \mathscr{E}_{sys} = \varepsilon - \varepsilon_{sys}) \tag{I.3.5}
\end{align}$$

where we have fixed the values to “collect” all micro-states into their realizations. The purpose now becomes finding a polished expression for the energy of the system, by exclusively working at the scale of the heat bath to exploit thermodynamic properties at $n \to \infty$.

Knowing that the heat bath is by construction larger, $\varepsilon_{res} \gg \varepsilon_{sys}$. Thus, we can use $\varepsilon - \varepsilon_{sys} \approx \varepsilon_{sys}$ to Taylor expand the number of micro-states of the reservoir. The clearest way of doing so is to directly expand its logarithm around ε:

$$\ln \mathscr{N}_{res}(\varepsilon - \varepsilon_{sys}) \approx \ln \mathscr{N}_{res}(\varepsilon) - \left.\frac{\mathrm{d}\ln \mathscr{N}_{res}(\mathscr{E}_{res})}{\mathrm{d}\mathscr{E}_{res}}\right|_{\mathscr{E}_{res}=\varepsilon} \varepsilon_{sys} + h.o.t. \tag{I.3.6}$$

where, we have denoted higher order terms (h.o.t.). Neglecting their contribution for small enough ε_{sys} and recalling that at the reservoir level we are dealing with a

[7] Sometimes the reservoir is referred to as a **heat bath**

micro-canonical ensemble one gets that **at equilibrium** via Eqn. I.2.11

$$\ln \mathscr{N}_{res}(\varepsilon - \varepsilon_{sys}) = \ln \mathscr{N}_{res}(\varepsilon) - \beta\varepsilon_{sys} \iff \mathscr{N}_{res}(\varepsilon - \varepsilon_{sys}) = \mathscr{N}_{res}(\varepsilon) e^{-\beta\varepsilon_{sys}}. \tag{I.3.7}$$

Reconciling with Eqn. I.3.3 and isolating the system terms the following expression is derived:

$$\mathbb{P}\left[\omega \in \Omega | \mathscr{E}_{sys} = \varepsilon_{sys}\right] = \frac{1}{\mathcal{Z}(\beta)} e^{-\beta\varepsilon_{sys}} \quad \mathcal{Z}(\beta) \propto \frac{1}{\mathscr{N}_{res}(\varepsilon)} \tag{I.3.8}$$

where $\mathcal{Z}(\beta)$ is a normalization factor to allow the probability to be well defined (i.e. to have total mass one). Recalling that ε is really fixed and $\mathscr{E}_{sys}$ is instead a random variable, by the arbitrariness of ω and the postulate of equal a priori probabilities, the random variable $\mathscr{E}$ has a well defined distribution. Indeed we have the scaling $\sim e^{-\beta\mathscr{E}}$ and know that the rest must be set to make it a valid distribution. A careful retraction of the steps leads us to the following very important remark.[8]

Remark I.3.9. *From now on, we drop the reservoir notion since we found a formula that is independent of it. The symbol $\mathscr{E}$ will denote the energy of the system.*

Definition I.3.10 (Boltzmann Canonical Distribution)**.** *The Boltzmann canonical distribution, in short Boltzmann distribution is the probability density function (pdf) of the energy in a canonical ensemble at the equilibrium temperature. Concisely:*

$$\mathbb{P}\left[\mathscr{E}; \beta\right] := \frac{1}{\mathcal{Z}(\beta)} e^{-\beta\mathscr{E}} \quad \mathcal{Z}(\beta) = \sum_i e^{-\beta\mathscr{E}_i} \tag{I.3.11}$$

where $\mathcal{Z}(\beta)$ is the normalization factor, also known as partition function.
When averaging wrt the Boltzmann distribution, we denote the expectation as $\langle\cdot\rangle_\beta$ or $\langle\cdot\rangle$ if the temperature is clear from context, constant or redundant.

Remark I.3.12 (A first word about the partition function)**.** *Most of this document will be focused on the complications that arise when computing different partition functions. The most formal justification of such claim is in Section II.3.1. For the moment, we just notice that the sum can be rearranged into different forms. Assuming that the degeneracy function is available, if $\mathscr{E}$ takes discrete values in an alphabet:*

$$\mathcal{Z}(\beta) = \sum_{\Omega} e^{-\beta\mathscr{E}(\omega)} = \sum_{\mathscr{E}_i} \mathscr{N}(\mathscr{E}_i) e^{-\beta\mathscr{E}_i}, \tag{I.3.13}$$

while if $\mathscr{E}$ is so fine grained that it is continuous:

$$\mathcal{Z}(\beta) = \int \mathscr{N}(\mathscr{E}) e^{-\beta\mathscr{E}} \, \mathrm{d}\mathscr{E}. \tag{I.3.14}$$

Remark I.3.15 (Probability cross-post)**.** *In Bayesian probability the partition function would be the normalization factor. Some textbooks also use the same terminology of Physics. The scope of $\mathcal{Z}$ goes well beyond Thermodynamics (as all of its ideas). It is not a coincidence that the problems in Bayesian Probability involve computing the very same object. Incidentally, for the same reason, we are taking this starting detour to introduce the terminology and historically justify some tricks.*

We avoid discussing the structural properties of the last object, which will be dealt with at the right time, and opt to continue with the physical interpretation. It is however crucial to informally state one concept, which will be justified later in Section II.3:

The Partition Function is all you need

By manipulating the partition function a statistician can access all the thermodynamical observables, especially the key ones.

[8] As a side note, it also allows to discard completely the thought experiment we carried.

Example I.3.16. *In the canonical ensemble energy is random. Let $\mathscr{E}$ take values in an alphabet. Notice that this is without loss of generality (wlog). Then:*

$$\langle\mathscr{E}\rangle_{=}\sum_i \mathscr{E}_i \mathbb{P}[\mathscr{E}_i] = \sum_i \mathscr{E}_i \frac{e^{-\beta\mathscr{E}_i}}{\mathcal{Z}(\beta)} = \frac{1}{\mathcal{Z}(\beta)}\sum_i -\frac{\partial e^{-\beta\mathscr{E}_i}}{\partial\beta} = -\frac{1}{\mathcal{Z}(\beta)}\overbrace{\partial_\beta \sum_i e^{-\beta\mathscr{E}_i}}^{=\partial_\beta \mathcal{Z}(\beta)} = -\frac{\partial \ln \mathcal{Z}(\beta)}{\partial\beta}, \tag{I.3.17}$$

where we have used ∂ to ensure that the expression is general.

Remark I.3.18. *As discussed previously, energy levels are just a simplification for expository treatment, we will gloss over the complications of the continuous case, but one can consider the energy levels to be so close to each other that macroscopically they look like a continuum.*

In many problems of interest in inference, a realization of the energy is related to the phenomenology of an assumed-to-exist model, and in particular to some "teacher" variable. To make it explicit, we will often use the object below. Notice that it is not the exact Physics definition of Hamiltonian but it will be sufficient for our purpose.

Definition I.3.19 (Hamiltonian)**.** *Consider a random variable $\boldsymbol{X}$. If such random variable is the sole descriptor of the energy of the system, then its randomness is described by how it induces different values of the energy. We make it explicit with the relation $\mathscr{E}(\omega) = \mathscr{H}(\boldsymbol{X}(\omega))$. In this case, the distribution of the random variable reads:*

$$\mathbb{P}\left[\boldsymbol{X} = \boldsymbol{x}\right] = \mathbb{P}\left[\boldsymbol{x};\beta\right] = \frac{1}{\mathcal{Z}(\beta)} e^{-\beta\mathscr{H}(\boldsymbol{x})} \quad \mathcal{Z}(\beta) = \sum_i e^{-\beta\mathscr{E}_i} = \sum_{\mathscr{X}} e^{-\beta\mathscr{H}(\boldsymbol{x})}. \tag{I.3.20}$$

If the random variable $\boldsymbol{X}$ has a measure ν we allow ourselves to write:

$$\mathcal{Z}(\beta) = \int_{\mathscr{X}} e^{-\beta\mathscr{H}(\boldsymbol{x})}\,\mathrm{d}\nu(\boldsymbol{x}). \tag{I.3.21}$$

In particular, we will consider only the case in which $\boldsymbol{X}$ is a continuous random variable for simplicity, but it can be extended.

Remark I.3.22. *Notice that we are moving from a discrete energy to a continuous Hamiltonian when $\boldsymbol{X}$ is continuous. We are not making energy levels continuous.*

Remark I.3.23. *In some sense, $\boldsymbol{X}$ can be seen as the sufficient statistic for inferring the behavior of the Hamiltonian. There is no need to access the abstract micro-states ($\boldsymbol{\omega}$) discussed in the preliminary section. At the same time, $\boldsymbol{X}$ can be on its own high dimensional, therefore making the task of retrieving it non-trivial. When $\boldsymbol{X} \in \mathbb{R}^d$ and $d \gg 1$ is large, it can be thought of as a micro-state on its own, which carries all the information about the system. At the same time, many configurations can map to the same energy value!*
Notice also that nothing says it will also be ***necessary****. As a matter of fact, in some regimes, it will not be.*

I.3.2 Entropies

We first notice that the Boltzmann Entropy (Def. I.2.15) is the adequate entropy notion for a micro-canonical ensemble at thermal equilibrium, where one can exploit the property that all micro-states are equally likely in the only manifested macro-state. Concerning the canonical ensemble, a natural question would be finding a function that can describe settings with non-uniform probabilities. A generalization achieving such objective was argued by Gibbs. We outline below the quickest argument to derive it.
Assume that instead of completely indistinguishable particles, we group them into $i \in [m]$ groups of size n_i such that $\sum_{i=1}^m n_i = n$. This super-construction can be thought of as a set of macro-states, each with probability $p_i = \frac{n_i}{n}$.
Even though the original entropy is $\mathscr{S}_{tot} = \ln(n)$, we assume it is impossible to

measure *a priori* due to experimental limitations. We instead consider the entropy at the level of some α super-sets, denoted as $\mathscr{S}_A$. A statistician with less knowledge could conclude that it is the entropy of the system. Fortunately, a statistician with more fine grained tools arrives. These instruments reveal a tree structure on the α-level micro-states, so that each α there are n_i sub-micro-states. By the combinatorial argument made previously, the total entropy will read $\mathscr{S} = \mathscr{S}_A + \mathscr{S}_B$ where B is the entropy of one of the n_i sized subsets. By assumption it is impossible to measure it directly, so we resort to its mean across subgroups, which is expressed as $\mathscr{S}_B = \sum_i p_i \mathscr{S}_i = \sum_i p_i \ln(n_i)$. Hence, letting $k_\mathrm{B} = 1$:

$$\mathscr{S} = \ln(n) = \mathscr{S}_A + \sum_i p_i \mathscr{S}_i \iff \mathscr{S}_A = \ln(n) - \sum_i p_i \mathscr{S}_i = \sum_i p_i(\ln(n) - \mathscr{S}_i) = -\sum_i p_i \ln(p_i) \tag{I.3.24}$$

where in the last equality we used the definition $\ln(n) - \ln(n_i) = \ln\left(\frac{n_i}{n}\right) = \ln(p_i)$.

Remark I.3.25. *A closer look at how we deal with the different macro-states shows that the energy varies over them, identifying it with a canonical ensemble.*

Another derivation is as follows. Letting the energy vary, we assume it takes a values on a discrete collection, indexed by i. If we had access to a very large number N of copies of the same ensemble, the number of ensembles with energy $\mathscr{E}_i$ would be almost surely $N\mathbb{P}[\mathscr{E}_i; n]$ by a simple application of the Strong Law of Large Numbers (SLLN). Combinatorially, the number of ways in which one can arrange the values of N copies into energy levels $\mathbb{P}[\mathscr{E}_i; n]N$ is:

$$\mathscr{N} = \frac{N!}{\prod_i (\mathbb{P}[\mathscr{E}_i; n]N)!} \implies \mathscr{S}_N = k_\mathrm{B} \ln \mathscr{N} \approx -k_B N \sum_i \mathbb{P}\left[\mathscr{E}_i; n\right] \ln \mathbb{P}\left[\mathscr{E}_i; n\right], \tag{I.3.26}$$

where we have used Stirling's approximation. By extensivity of the entropy, the entropy of the N copies expression suggests that for a single copy:

$$\mathscr{S} = \frac{1}{N}\mathscr{S}_N = -k_\mathrm{B} \sum_{\mathscr{E}_i} \mathbb{P}\left[\mathscr{E}_i; n\right] \ln \mathbb{P}\left[\mathscr{E}_i; n\right] \tag{I.3.27}$$

Before giving it a formal definition, we trivially check that it is extensive and always positive, according to the comments below Remark I.2.26. Additionally, we notice it can be viewed as an extension of Definition I.2.15, which appears for the equally likely distribution $p_i = \frac{1}{\mathscr{N}}$ and $m = |\Omega| = \mathscr{N}$.

Definition I.3.28 (Gibbs Entropy)**.** *The entropy of a canonical ensemble at equilibrium is:*

$$\mathscr{S} := -k_\mathrm{B} \sum_i p_i \ln(p_i) \tag{I.3.29}$$

where i runs over the realizations of the energy $\mathscr{E}_i$, and p_i is its probability, which follows a Boltzmann distribution.

Remark I.3.30. *Notice that the canonical entropy is a function of temperature, differently from the micro-canonical one, which is a function of energy.*

Remark I.3.31. *While we have restricted the discussion to energies, it takes little time to understand that the concept of canonical entropy extends without loss of generality to any distribution. Potentially on a more subtle note, if energy varies, we find that the entropy depends on Temperature. By analogy, we expect that at the right scale of units of measure, any random variable will have entropy dependent on a parameter (a temperature-like parameter), which itself can be seen in the flipped perspective as the random variable of a (micro-canonical-like) ensemble where such parameter is uniformly random in its domain, and has entropy dependent on the random variable (entropy of the micro-canonical ensemble will depend on the energy-like-random variable). This can be seen as a first taste of the notion of conjugacy introduced in Definition I.2.25.*

Remark I.3.32. *The concept of Gibbs entropy adapts to any object that is governed by randomness. In this case, we have opted to represent the entropy of energy levels, but might as well discuss the entropy of the variables $\boldsymbol{X}$ over which the energy depends on.*

Remark I.3.33. *Unless otherwise stated $\mathscr{S}$ will always be a Gibbs entropy. The choice is made to simplify notation. Also, most of the time we will ignore the k_B term in front.*
Later in Sections III.2 and I.3.4 we will argue more connections that will nurture a deeper understanding of the ensembles.

It worth stating that as of now entropy is just a measure of complexity of microstates. Certainly Gibb's entropy is equivalent to Shannon's, an object that we will see in Section III.1, but the two starting points are different. It was not until Jaynes (Jaynes 1957) that the two were reconciliated.

Keeping our eyes on the objectives of this Chapter, we postpone the discussion to the moment when connecting arguments will be possible. In the timeline of Thermodynamics, we are now ready to define a very important object, that properly highlights the importance of the perspective of Statistical Physics.

I.3.3 Free Energies

In an inference problem, the common example brought forward to explain the fallacy of point-estimation (e.g. Maximum-Likelihood, Maximum a Posteriori) is very intuitive. A simple plot of a bimodal distribution is sufficient. As a matter of fact, any researcher interested in inference would like to find solutions that have *stability properties*, to be understood as valleys of the configuration of the object of study where local variations do not affect the global performance. Such trade-off is exactly matched by the notion of free energy, which is studied extensively in Thermodynamics since its early foundation. In words, it is a representation of the randomness of the problem that "characterizes" it,[9] and underlines how the competition between Energy and Entropy needs to be balanced. The former, in the wide sense, is the observable object of interest. The latter is the un-observable disorder of the configurations induced by the randomness and the structure of the problem. For a random problem, the two have to be considered to strike *typical* arrangements.

Definition I.3.34 (Helmoltz Free Energy)**.** *The free energy is a quantity*

$$\mathfrak{F}(\beta) := \langle \mathscr{E} \rangle - \frac{1}{\beta}\mathscr{S} = \mathscr{U} - \frac{1}{\beta}\mathscr{S}. \tag{I.3.35}$$

We use the fraktur symbol since we will mostly deal with its adimensional version $\mathscr{F}$ and identify $\langle \mathscr{E} \rangle := \mathscr{U}$. Free entropy $\mathscr{F} := -\beta\mathfrak{F}$ is more common in Computer Science derivations of the subject. We will present the Computer Science object when needed, and favour free energy for the moment.

As evident from its definition, the free energy/entropy captures the competition between energy and entropy in a common ground of units of measure.
In the next paragraphs, we will see that it is closely linked to many of the quantities we saw earlier. We will also statistically justify why it makes sense that the free energy is as good as observing the system directly (see Sec. II.3, which will give us for free an additional interpretation of Legendre transform.

Fact I.3.36 (Free energy is extensive)**.** *As per the entropy, we can see it by combining two canonical ensembles A, B. Their free energy will be the log of the product of the partition functions. Mathematically:*

$$\mathfrak{F}_{(A,B)}(\beta) = \ln(\mathscr{Z}_{(A,B)}) = \ln(\mathscr{Z}_A\mathscr{Z}_B) = \ln \mathscr{Z}_A + \ln \mathscr{Z}_B = \mathfrak{F}_A(\beta) + \mathfrak{F}_B(\beta). \tag{I.3.37}$$

If we instead take a more formal route, the assumption that the Energy is extensive is sufficient. From extensive energy we have that $\ln \mathscr{Z}(\beta)$ is extensive as well, being the logarithm of the sum of exponentials of an extensive quantity.

[9] as we will see in Section II.3 and throughout all of Chapter III

Canonical Ensemble Version Expanding the Gibbs entropy we find that:

$$\begin{aligned} \mathscr{S} &= -\sum_i p_i \ln(p_i) && \text{(I.3.38)} \\ &= -\sum_i \frac{1}{\mathcal{Z}(\beta)} e^{-\beta \mathscr{E}_i} \ln\left(\frac{1}{\mathcal{Z}(\beta)} e^{-\beta \mathscr{E}_i}\right) && \text{(I.3.39)} \\ &= -\sum_i \frac{e^{-\beta \mathscr{E}_i}}{\mathcal{Z}(\beta)} \left(-\beta \mathscr{E}_i - \ln(\mathcal{Z}(\beta))\right) && \text{(I.3.40)} \\ &= -\ln(\mathcal{Z}(\beta)) + \beta \langle \mathscr{E} \rangle . && \text{(I.3.41)} \end{aligned}$$

A re-arrangment of the objects in the equality allows us to define a very important object in Statistical Physics.

Proposition I.3.42 (Canonical Helmoltz Free Energy)**.** *In a canonical ensemble the free entropy is the logarithm of the partition function, and the free energy follows similarly. Mathematically:*

$$\mathfrak{F}(\beta) \equiv -\frac{1}{\beta} \ln(\mathcal{Z}(\beta)), \quad \mathscr{F}(\beta) = \ln \mathcal{Z}(\beta). \qquad \text{(I.3.43)}$$

Proof. Rearrange the above chain of equalities. □

Adding Variables Allowing for the energy to vary is sufficient for the topics we will encounter. Different generalizations that let the Volume or the number of micro-states vary lead to different ensembles. For some ideas, see (S. J. Blundell and K. M. Blundell 2009; Schwartz 2021). The key point to inspect is always the connection between the free energies, partition functions, duality transforms and learning problems. These will be partially cleared out in the next Chapters, as different ways to generalize the Thermodynamics models we presented. In simple words, to follow the classical result, we should assume that we also allowed the volume V to vary, with its *conjugate* (in the sense of Def. I.2.25) variable being the pressure, and entropy being dependent on both $(V, \mathscr{E})$. In practice, the state variables (Def. I.2.21) wrt which we could express partial derivative equations all depend on each other but on nothing else. Mathematically, a classical model assumes:

$$\mathscr{E} \equiv \mathscr{E}(n, V, \mathscr{S}), \quad \mathscr{S} \equiv \mathscr{S}(\mathscr{E}, n, V), \quad n \equiv n(\mathscr{E}, \mathscr{S}, V), \quad V \equiv (\mathscr{S}, \mathscr{E}, n). \quad \text{(I.3.44)}$$

From these, quantities like temperature[10] will be derived. The inherent difficulty of considering only the above state quantities is that they depend on values that are not straightforward to measure. For example, energy depends on entropy and vice versa. Contrarily, the free energy (which is the log of a normalization function) is independent of these and depends on the derived quantities such as $\beta \equiv \frac{1}{T}$, which is **easy to measure**. We derive dependence by setting aside additional dependences. To begin recognize from Eqn. I.2.11 that the following holds:

$$\mathrm{d}\mathscr{S} = \left(\frac{\partial \mathscr{S}}{\partial \mathscr{E}}\right) \mathrm{d}\langle \mathscr{E} \rangle = \beta \, \mathrm{d}\langle \mathscr{E} \rangle \iff \mathrm{d}\langle \mathscr{E} \rangle = \frac{1}{\beta} \mathrm{d}\mathscr{S}, \qquad \text{(I.3.45)}$$

where the expression is rather general because the identity for temperature is in partial derivatives, and the entropy could have depended on other terms. Then, taking the differential:

$$\mathrm{d}\mathfrak{F} = \mathrm{d}\langle \mathscr{E} \rangle - \frac{1}{\beta} \mathrm{d}\mathscr{S} - \mathscr{S} \, \mathrm{d}\left(\frac{1}{\beta}\right) = -\mathscr{S} \, \mathrm{d}T, \qquad \text{(I.3.46)}$$

where we used Eqn. I.3.45, and can conclude that free energy/entropy depends on measurable (derived) variables.

Remark I.3.47. *From the Physics perspective, it is evident that $\mathfrak{F} \equiv \mathfrak{F}(\beta)$. However, the process is somewhat more convoluted than a Bayesian approach, where one realizes that the partition function $\mathcal{Z}$ will depend on the conditioning variables and parameters only, by definition.*

[10] If we allowed for more variability, Pressure and chemical potential would be in the list.

Mathematically, this *problem switch* from easy to write & hard to measure to easy to measure & hard to write is a well defined concept, that was largely studied in the theory of Legendre Transforms. We will see some fundamental results in Section III.4.

Local Versions Notice that it is straightforward to re-express the probabilities in the canonical ensemble as:

$$p_i = e^{(\beta\mathfrak{F}(\beta)-\mathscr{E}_i)} \quad \mathscr{Z}(\beta) = e^{-\beta\mathfrak{F}(\beta)}, \tag{I.3.48}$$

where the last term can be thought of as the free energy being an energy if only one micro-state is present.
On a similar note, consider a canonical ensemble at constant temperature and volume. Results in Statistical Mechanics show that the Helmoltz free energy, which is only related to the system,[11] will be minimized. Such fact can be understood in at least two ways.
The first and most direct one is by derivation. It holds that $\mathfrak{F} = \langle\mathscr{E}\rangle - T\mathscr{S}$, so since $\mathrm{d}\mathscr{S} \geqslant 0$ and $\langle\mathscr{E}\rangle$ is fixed the free energy will be such that $\mathrm{d}\mathfrak{F} \leqslant 0$.
The second one is borrowed from (Kittel 2004; Young 2012), and has wider generality at the thermodynamic limit. Taking $k_{\mathrm{B}} = 1$ in the canonical ensemble, we have:

$$\mathfrak{F}(\beta) = -\frac{1}{\beta}\ln\mathscr{Z}(\beta) = -\frac{1}{\beta}\ln\sum_i e^{-\beta\mathscr{E}_i} = -\frac{1}{\beta}\ln\sum_{\mathscr{E}}\mathscr{N}(\mathscr{E})e^{-\beta\mathscr{E}}. \tag{I.3.49}$$

Then, assuming that the Gibbs entropy counts the exponential order of the number of configurations (more details about the assumption in the next Section), it holds that $\mathscr{N}(\mathscr{E}) = \exp\{\mathscr{S}(\mathscr{E})\}$ (the degeneracy function of Def. I.2.3), since the RHS expression effectively counts the number of micro-states at energy $\mathscr{E}$. Then, the free energy expression becomes:

$$\mathfrak{F}(\beta) = -\frac{1}{\beta}\ln\sum_{\mathscr{E}} e^{\mathscr{S}(\mathscr{E})-\beta\mathscr{E}} = -\frac{1}{\beta}\ln\sum_{\mathscr{E}} e^{-\beta\tilde{\mathfrak{F}}(\beta;\mathscr{E})}, \tag{I.3.50}$$

where we put emphasis on the fact that $\tilde{\mathfrak{F}}$ is a *generalized free energy*, for a system not necessarily at equilibrium, constrained to have energy $\mathscr{E}$. Similarly, $\mathscr{S}(\mathscr{E})$ is not necessarily at equilibrium. Defining:

$$\mathscr{E}^\star = \arg\max e^{-\beta\tilde{\mathfrak{F}}(\beta;\mathscr{E})} = \arg\min\tilde{\mathfrak{F}}(\beta;\mathscr{E}), \tag{I.3.51}$$

by extensivity of the generalized free energy and large size $n \to \infty$, it is safe to say that the maximum of the first expression will dominate the sum, with $\frac{\ln(n)}{n}$ error.[12] and the free energy takes form:

$$\beta\mathfrak{F}(\beta) = -\ln\sum_{\mathscr{E}} e^{-\beta\tilde{\mathfrak{F}}(\beta;\mathscr{E})} \overset{n\to\infty}{\approx} \beta\tilde{\mathfrak{F}}(\beta;\mathscr{E}^\star). \tag{I.3.52}$$

In words, the free energy at equilibrium on the LHS is the minimum of the generalized free energies on the RHS. Recalling also that in a canonical ensemble $\mathbb{P}[i]\propto e^{-\beta\mathscr{E}_i}$ for a single state and $\mathbb{P}[\mathscr{E}_i]\propto\mathscr{N}(\mathscr{E}_i)e^{-\beta\mathscr{E}_i}$ for an energy level, the probability of the energy can be written down as:

$$\mathbb{P}[\mathscr{E}]\propto e^{-\beta\tilde{\mathfrak{F}}(\beta;\mathscr{E})}, \tag{I.3.53}$$

so the most probable value of the energy is the one that minimizes the generalized free energy, which in turn is up to vanishing corrections the equilibrium free energy,

[11] <u>not</u> also the heat bath

[12] this is a simple principle, *the maximum term method.* Consider a sum of N exponentials, increasingly ordered as $e^{nv_{max}} \leqslant S_n = \sum_{i\geqslant 1} e^{nv_i} = e^{nv_{max}}(1+\sum_{i\geqslant 1} e^{n(v-v_{max})}) \leqslant Ne^{nv_{max}}$. Applying the ln and using continuity, one can show that the limiting density is squeezed by the two limiting densities which are both v_{max}. Notice that $N = O(\ln n)$ usually since $n \sim 10^{23}$ while the number of energy values is implicitly assumed to be bounded, or at least that to a single macro-state there correspond an exponential number of micro-states.

where the dynamics stop. Any system initialized out of equilibrium is drawn to get to $\mathscr{E}^\star$ and minimize the free energy, at least in a *local* neighborhood sense.
In the next subsection, we give more details about the assumption that the Gibbs Entropy can be regarded as a counter of configurations, in analogy with the entropy for micro-canonical ensembles.

I.3.4 Asymptotic Equivalence of micro-canonical and Canonical Ensemble

Before getting into the details, we need a very quick digression on *heat*. Recall that in (P4) we stated that heat, denoted as $\mathscr{Q}$, is energy in transit. Logically, it will be an extensive variable, and *heat capacity*, defined as $\mathcal{C} := \frac{\mathrm{d}\mathscr{Q}}{\mathrm{d}T}$ will be extensive too.[13] Skipping some details, and simplifying for $k_\mathrm{B} = 1$ as always, we observe that for a canonical ensemble:

$$\mathcal{C} = \frac{\partial \langle \mathscr{E} \rangle}{\partial T} = \beta^2 \left\langle \mathscr{E}^2 \right\rangle - \langle \mathscr{E} \rangle \frac{\partial \ln \mathcal{Z}(\beta)}{\partial T} = \beta^2 (\left\langle \mathscr{E}^2 \right\rangle - \langle \mathscr{E} \rangle^2) \tag{I.3.54}$$

where the only informal passage is giving a justification for the first equality, which can be seen to follow from the first law of Thermodynamics (Eqn. I.2.17) at constant volume. The last term on the RHS is a variance wrt the Boltzmann distribution, so we can say that

$$\mathrm{Var}\left[\mathscr{E}\right]_\beta = \frac{\mathcal{C}}{\beta^2}, \tag{I.3.55}$$

and in the thermodynamic limit $n \to \infty$ using extensivity of the quantities, the fluctuations of the energy will be

$$\frac{\sqrt{\mathrm{Var}\left[\mathscr{E}\right]_\beta}}{\mathscr{U}} = \frac{1}{\beta} \frac{\sqrt{\mathcal{C}}}{\mathscr{U}} = O\left(\frac{\sqrt{n}}{n}\right) \overset{n\to\infty}{\to} 0. \tag{I.3.56}$$

In other words, as $n \to \infty$ the average energy will concentrate on a single mean value. Such result solves the potential ambiguity between using the symbol $\langle \mathscr{E} \rangle_\beta$ and $\mathscr{E}$, concluding that it will be a constant, up to vanishing fluctuations. Being that the energy is constant, at the thermodynamic limit a canonical ensemble “is” a micro-canonical ensemble, and the Gibbs entropy inherits the counting properties of Boltmann’s entropy (Def. I.2.15).

A crucial point we have ignored is how to ensure extensivity. In our applications, it is sufficient to take the energy to be extensive, so we will need a Hamiltonian of order $\mathscr{H}(\boldsymbol{X}) \in O(n)$. In general, neglecting long-range interactions makes the system well defined, with a convex in $(n, \mathscr{S})$ energy $\mathscr{U}$ and concave in T entropy $\mathscr{S}$.

Other topics With some care about evaluations of the integrals, one could also have shown this by a saddle point approximation, which will be overviewed in Section II.6.

Further References

For a more formal treatment, one can use Large Deviation Principles (Sec. III.5), where a nice reference is (Touchette 2015). Relaxations to the principle hold in more generality but are not commented on. A review of modern approaches is found in (Tsallis 2019). A quick and clear note on Equivalence of ensembles from the Physics point of view is (Teitel 2021).

For simplicity, we will take equivalence of ensembles as an assumption to deal with the next chapters
In the coming subsection, we introduce the basic terminology to tackle a problem in Statistical Physics.

[13] A fraction of extensive and intensive quantities is extensive

I.4 Phase transitions

In Physics, Phase transitions are phenomenons of co-existance of equilibrium states. They are the result of competition between microscopic effects. Roughly, one will tend to order the system, while the other will induce disorder. The generality of this concept is well understood upon having presented the words used to describe it, as we do below.

By *phase* we mean a region of the phase space that can be summarized under the same macroscopic properties. An *order parameter* is a macroscopic quantity with the further property that it differs greatly across phases. In other words, it is a characterization of the different regions of the phase space, and knowing it allows to be aware of the system phenomenology. When changes in a set of parameters are the cause of the change in the order parameter, the system has a *phase transition.* If the parameters are more than one, there will be a *phase boundary* rather than a point. Often, everything is visualized in a *phase space*, which is just a plot in the Cartesian plane of the parameters that highlights when the order parameter varies. A word of caution is needed here. The free energy is an analytic function of the parameters (Def. II.2.1), and cannot exhibit non-analytic behavior in a system of size n, as we will see in the next Chapter. Nevertheless, the $n \to \infty$ limit of an analytic function need not be analytic, and can have points where it is not. We will refer to these as *singularities*, and identify a phase transition in their location. In other words, an abrupt change in the order parameter is identified when there is a non-analytic behavior of the free energy.

Coming back to our main discussion, we recall that a canonical ensemble is a collection of energy configurations with a Boltzmann distribution. It is at fixed temperature, and has fluctuating energy levels with asymptotic guarantees. In perfect alignment with the order/disorder competition in Phase transitions, its Helmholtz Free Energy (Def. I.3.34, Prop.I.3.42) is sufficient for a full understanding of its macroscopic properties, and expresses the energy/entropy antagonism of the system. It is then natural to claim that free energy will be a good descriptor of the system's phases, and that it will depend on the order parameter(s), identifying phase transitions.

Classifications, Examples Ehrenfest (Ehrenfest et al. 1960) designed a classification that will be sufficient for our purposes. Being at the thermodynamic limit of an extensive quantity, it is natural to work instead with the free entropy density f to have an intensive quantity.

Definition I.4.1 (Erhrenfest classification of Phase transitions)**.** *Let the order parameter be m, and the parameters of the phase space be ϑ. Consider $f(m,\vartheta) = \lim_{n\to\infty} \frac{1}{n}\mathscr{F}(m,\vartheta)$, which is intensive.*

- *A **first order** phase transition happens when the derivative $\frac{\partial f}{\partial \vartheta} = m$ is discontinuous.*
- *A **second order** phase transition happens when the second derivative $\frac{\partial^2 f}{\partial \vartheta^2} = m$ is discontinuous.*
- *Phase transitions of order k follow analogously.*

Example I.4.2. *One can easily see that $\frac{\partial \mathscr{F}}{\partial T} = \mathscr{S}$ so if the phase space parameter is the temperature then the discontinuity is in the entropy, which serves as order parameter. Boiling temperature of water is a practical example.*
Similarly, second order derivatives of the free energy are often associated to susceptibilities, which are order parameters that (roughly) measure the variance of some observable. A second order phase transition will relate to the discontinuous behavior of these quantities.

The minimization of the free energy depends on the realization of the parameters on which it depends. Often in this context the optimization leads to equations of the form:

$$m = \hbar(m;\vartheta), \tag{I.4.3}$$

for which one wishes to find a fixed point. By the usual fixed point theory, different regions of values of ϑ may make certain fixed points stable, unstable, attractive or repulsive. Moreover, the dynamics of the iterative map $m^{t+1} = \mathscr{h}(m^t)$ give great information about the dynamics of procedures to find these non-analytic points.
However, on the phase space, the free energy, despite being convex, can be non strictly convex, and develop local minima. We term these sub-optimal points *metastable* states: despite being stable fixed points, they are not global minima. Out of the thermodynamic limit, they are states of the system that will eventually be escaped in exponential time in size $t \in O(e^n)$, to reach the global minima. Each of these *hops* is the result of overcoming an *energy barrier*. Additionally, in some cases, variations of ϑ make metastable states disappear, and the boundary in the phase space for which the local minima disappears is termed *spinodal*, with symbol $\vartheta_{\mathsf{spinodal}}$.
We report below a summary of one of the first examples in (Krzakala and Zdeborová 2021), which has many of the phenomenological aspects discussed.

Example I.4.4 (Curie-Weiss Model)**.** *Details are found in (Krzakala and Zdeborová 2021, Chap. 1, App. 1.A). We also provide a similar Example in Subsection I.5.1. Assume there are n particles, each with spin $\sigma_i \in \{\pm 1\}$, placed in a fully connected graph. The Hamiltonian of the system reads:*

$$\mathscr{H}(\boldsymbol{\sigma}; n) = -\frac{1}{2n}\sum_{i=1}^{n} \sigma_i \sigma_j - h\sum_{i=1}^{n} \sigma_i, \tag{I.4.5}$$

where $h \in \mathbb{R}$ is a scalar external magnetic field. In the canonical ensemble:

$$\mathbb{P}[\boldsymbol{\sigma} = \boldsymbol{v}; n, \beta, h] = \frac{1}{\mathscr{Z}(\beta; n, h)} e^{-\beta \mathscr{H}(\boldsymbol{v})}. \tag{I.4.6}$$

A good order parameter turns out to be the empirical mean of the magnetization, which is the Boltzmann mean of the random variable:[14]

$$\overline{\boldsymbol{\sigma}} = \frac{1}{n}\sum_{i=1}^{n} \sigma_i \in \mathbb{R}. \tag{I.4.7}$$

It is crucial to notice that $\mathscr{H}(\cdot)$ is extensive when seen as a function of $\overline{\boldsymbol{\sigma}}$ and that the control parameters for the phase space will be (β, h) since $\frac{\partial \mathscr{f}}{\partial h} = m$, where $m = \langle \overline{\boldsymbol{\sigma}} \rangle_\beta$.
After some calculations, one finds that:

$$\lim_{n\to\infty} \mathscr{f}(\beta; n) = \max_{m\in[-1,1]} \phi(m) \quad \phi(m) := \frac{\beta}{2}m^2 + \beta h m + \mathcal{H}\left(\frac{1-m}{2}, \frac{1+m}{2}\right) \tag{I.4.8}$$

$$\mathcal{H}\left(\frac{1-m}{2}, \frac{1+m}{2}\right) = \frac{1-m}{2}\ln\frac{1-m}{2} + \frac{1+m}{2}\ln\frac{1+m}{2}, \tag{I.4.9}$$

where the function above is defined with a symbol consistent with future notation,[15] *and we have used as order parameter $m = \langle \overline{\boldsymbol{\sigma}} \rangle_\beta$. Setting the derivative of $\phi(\cdot)$ to zero gives a function in m of which we seek a fixed point:*

$$\frac{1}{2}\ln\frac{1+m}{1-m} = \beta(m+h) \iff m = \tanh\beta(h+m). \tag{I.4.10}$$

First of all, we notice that despite $\phi(\cdot)$ being analytic, its extremization could lead to non-analyticities. Also, there is a more general phenomenon of energy-entropy competitition, which can be observed by the normalized expression for the definition of free energy $\mathscr{f} = e - \beta\mathscr{s}$. We split the results into some subcases, which are the phases in the phase space values $(\beta, h) \in \mathbb{R}_+ \times \mathbb{R}$.

[14] note the tricky fact that we take the mean of a vector along the particles, therefore obtaining a scalar

[15] it is the Gibbs entropy (Def. I.3.28) of a system with two states and probabilities $(\frac{1-m}{2}, 1 - \frac{1-m}{2})$

- ***paramagnetic***, $h = 0, \beta \to \infty$
  as $\beta \to 0^+$, $\mathscr{s}$ dominates, and it has a unique minimizer at $m_\star = 0$

- ***ferromagnetic*** $h = 0, \beta \geqslant 1$
  the point $m_\star = 0$ becomes a maxima, and two symmetric local minima appear. The first derivative remains continuous since it is proportional to the entropy, suggesting a second order phase transition, which is indeed verified. While at $\beta \leqslant 1$ magnetizations sampled were null, at $\beta \geqslant 1$ the samples "spontaneously" break symmetry, and are either positive or negative. By "spontaneously", we mean without inducing preference, since the symmetry of the $\mathbb{Z}_2$ group in the spins (i.e. flip all the spins) is still satisfied.

- ***explicit symmetry breaking*** $h \neq 0$
  we lose the group symmetry and favour some alignments by construction. The behavior are distinguished into two cases:
  - *$\beta \to 0^+$, convex free energy, single minimum at full alignment with the external field $h = m$*
  - *$\beta \to \infty$ yet another split in terms of a well defined value:*
    - *$h < h_{spinodal}$, a local minima exists as a metastable state, in which the system can be temporarily trapped depending on the initial conditions*
    - *$h \geqslant h_{spinodal}$, the local minima disappears*

Remark I.4.11. *When dealing with phase transitions, the definition of free energy as a simple Legendre transform is inexact. First of all, we are at the thermodynamic limit, so it is infinite, secondly, there are problems with its derivative. For this reason, we must resort to the more general version of Legendre-Fenchel transform. More comments are provided in Section III.4.*

In the last subsection, we provide a simple interpretation of mean field models through a fundamental inequality. While the definition of mean field models is not well-established in inference and machine learning, it serves the purpose of a bonus idea of easy results derived from first principles in Thermodynamics.

I.5 GBF type Inequality and the Mean-Field Approach

It is often the case that a system with Hamiltonian $\mathscr{H}$ is difficult to solve. We will provide a method that gives an approximation and later show how this is justified in an Information Theoretical sense (see Secs. III.2, and III.3.1). Let the randomness follow a Boltzmann canonical distribution. Assume that we can define a different Hamiltonian that satisfies:

$$\langle \mathscr{H}(\boldsymbol{X}) \rangle_\sim = \left\langle \widetilde{\mathscr{H}}(\boldsymbol{X}) \right\rangle_\sim \implies \Delta\mathscr{H} \equiv \mathscr{H} - \widetilde{\mathscr{H}} \quad s.t. \quad \langle \Delta\mathscr{H}(\boldsymbol{X}) \rangle_\sim = 0, \tag{I.5.1}$$

where we emphasized that the expectations are wrt the new Hamiltonian. In a canonical ensemble, the partition function would read:

$$\begin{aligned}
\mathcal{Z}(\beta) &= \int e^{-\beta\mathscr{H}(\boldsymbol{x})} \,\mathrm{d}\boldsymbol{x} = \tilde{\mathcal{Z}}(\beta) \int \frac{1}{\tilde{\mathcal{Z}}(\beta)} e^{-\beta(\mathscr{H}(\boldsymbol{x}))} \,\mathrm{d}\boldsymbol{x} && \text{(I.5.2)}\\
&= \tilde{\mathcal{Z}}(\beta) \int \frac{1}{\tilde{\mathcal{Z}}(\beta)} e^{-\beta(\mathscr{H}(\boldsymbol{x}) + \widetilde{\mathscr{H}}(\boldsymbol{X})) - \widetilde{\mathscr{H}}(\boldsymbol{X}))} \,\mathrm{d}\boldsymbol{x} && \text{(I.5.3)}\\
&= \tilde{\mathcal{Z}}(\beta) \int \frac{e^{-\beta\widetilde{\mathscr{H}}(\boldsymbol{x})}}{\tilde{\mathcal{Z}}(\beta)} e^{-\beta(\mathscr{H}(\boldsymbol{x}) - \widetilde{\mathscr{H}}(\boldsymbol{x}))} \,\mathrm{d}\boldsymbol{x} && \text{(I.5.4)}\\
&= \tilde{\mathcal{Z}}(\beta) \left\langle e^{-\beta\Delta\mathscr{H}(\boldsymbol{X})} \right\rangle_\sim && \text{(I.5.5)}\\
&\geqslant \tilde{\mathcal{Z}}(\beta) e^{-\beta\langle \Delta\mathscr{H}(\boldsymbol{X}) \rangle_\sim} \qquad \text{Jensen's Ineq.} && \text{(I.5.6)}\\
&= \tilde{\mathcal{Z}}(\beta), && \text{(I.5.7)}
\end{aligned}$$

where in the last passage we applied the centrality of the “error” Hamiltonian component. Moving to the free energy, the statement is reformulated as:

$$\mathfrak{F}(\beta) \leqslant \widetilde{\mathfrak{F}}(\beta), \tag{I.5.8}$$

which is a termed a Gibbs-Bogoliubv-Feynman (GBF) type inequality.

In greater generality, assuming simply that $\mathscr{H} \equiv \widetilde{\mathscr{H}} + \Delta\mathscr{H}$ we inherit all the passages until Eqn. I.5.5. To proceed, it is worth seeing the object from a different perspective, since it is not anymore true that $\Delta\mathscr{H}$ is null in expectation. We observe that:

$$\begin{aligned}
\left\langle e^{-\beta\Delta\mathscr{H}(\boldsymbol{X})}\right\rangle_{\sim} &= \left\langle e^{-\beta(\Delta\mathscr{H}(\boldsymbol{X})+\langle\Delta\mathscr{H}(\boldsymbol{X})\rangle_{\sim}-\langle\Delta\mathscr{H}(\boldsymbol{X})\rangle_{\sim}))}\right\rangle_{\sim} && \text{(I.5.9)}\\
&= e^{-\beta\langle\Delta\mathscr{H}(\boldsymbol{X})\rangle_{\sim}}\,\langle\exp\{-\beta(\Delta\mathscr{H}(\boldsymbol{X}) - \langle\Delta\mathscr{H}(\boldsymbol{X})))\rangle_{\sim}\}\rangle_{\sim} && \text{(I.5.10)}\\
&\geqslant e^{-\beta\langle\Delta\mathscr{H}(\boldsymbol{X})\rangle_{\sim}}\,\exp\{-\beta\left(\langle\Delta\mathscr{H}(\boldsymbol{X})\rangle_{\sim} - \langle\Delta\mathscr{H}(\boldsymbol{X})\rangle_{\sim}\right)\} && \text{(I.5.11)}\\
&= e^{-\beta\langle\Delta\mathscr{H}(\boldsymbol{X})\rangle_{\sim}} && \text{(I.5.12)}\\
&\geqslant 0, && \text{(I.5.13)}
\end{aligned}$$

where the first inequality is by Jensen’s. We have thus retrieved a more general Gibbs-Bogoliubv-Feynman type inequality, which is:

$$\mathscr{H} = \widetilde{\mathscr{H}} + \Delta\mathscr{H} \implies \mathscr{Z}(\beta) \geqslant \widetilde{\mathscr{Z}}(\beta)e^{-\beta\langle\Delta\mathscr{H}(\boldsymbol{X})\rangle_{\sim}}. \tag{I.5.14}$$

For sure, it will be the case that the GBF inequality holds in the free energies, but the bound could be tighter or less depending on $\Delta\mathscr{H}$ and its randomness. This is the basis of the variational approach, which we will briefly review in Section III.3.

The best example of an application is given in the context of interacting particle systems and the mean-field theory approximation. There we consider a Hamiltonian such as that of the Ising model over a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$:

$$\mathscr{H}(\boldsymbol{\sigma}) = J\sum_{(i,j)\in\mathcal{E}}\sigma_i\sigma_j + h\sum_i\sigma_i \quad J, h \in \mathbb{R}, \tag{I.5.15}$$

where the spins $\boldsymbol{\sigma} \in \{\pm1\}^d$ are sampled from a Canonical Boltzmann distribution with energy $\mathscr{H}$, i.e. $\mathbb{P}\left[\boldsymbol{\sigma} = \boldsymbol{v}; \beta\right] \cong e^{-\beta\mathscr{H}(\boldsymbol{v})}$ for $v \in \pm\{\pm1\}^d$.
The first term is very difficult to deal with in general. A mean-field approximation consists in allowing each spin to feel the mean spin $\overline{\boldsymbol{\sigma}} = \frac{1}{n}\sum_i\sigma_i$ from its neighbors, instead of the neighbor itself. Therefore, we would derive the new Hamiltonian:

$$\widetilde{\mathscr{H}}(\boldsymbol{\sigma}) = (h + \widetilde{h})\sum_i\sigma_i, \tag{I.5.16}$$

where in particular for a d-regular graph $\widetilde{h} = dJ\langle\sigma_j\rangle$.

I.5.1 Sketch of the Mean-Field Solution of the Ising model

For a relevant example, consider the discussion in (Krzakala and Zdeborová 2021, Chap. I), especially in Section 1.1, where the reasoning is the same, but with a $-\frac{1}{n}$ in front of the Hamiltonian object, and the coupling term J reabsorbed into h. There, the Curie-Weiss model is treated, which is a complete graph, and the average field is actually a full descriptor of the dynamics, making the solution exact.
We *purposely* present a different formulation in Equation I.5.15 to stimulate understanding of the various ways in which these models are presented. We reported part of the argument in Example I.4.4.
Here we opt to give the quickest derivation of the approximation in a setting in which it proves to be *possibly inexact*. It can be found in (Evans 2009), or with more justification in (Utermohlen 2018). A fairly recent and interesting result in these topics is (Basak and Mukherjee 2017).

Following the approximation done, without assuming a d regular graph, we unroll some steps. Starting from the equation of the Hamiltonian, each spin j can be seen to be subject to a “local” Hamiltonian

$$\hbar(\sigma_j) = J\sigma_j\sum_{i\in\mathsf{Neigh}(j)}\sigma_i + h\sigma_j \quad \mathscr{H}(\boldsymbol{\sigma}) = \sum_{j=1}^{n}\hbar(\sigma_j). \tag{I.5.17}$$

The mean-field approximation amounts to postulating that the difficult local sum over neighbors, which will be site-dependent,[16] is replaced by the average of the spins across the spatial model. Mathematically we do the following replacement:

$$\sigma_i \rightsquigarrow m = \frac{1}{n}\sum_{k=1}^{n} \langle \sigma_k \rangle \qquad \boldsymbol{\sigma} = (\sigma_k)_{k=1}^{n} \tag{I.5.18}$$

therefore vindicating the expression mean-field, since each spin is supposed to be subject to the mean external field instead of the j-dependent the force exerted. A little thought then shows that under such construction, if we further require the number of neighbors to have some regularity (as in d-regular graphs), then the computation will greatly simplify. For the sake of the Example, we take d-regular graphs. Hence, we are led to the explicit approximation:

$$\widetilde{\hbar}(\sigma_j) = Jdm\sigma_j + h\sigma_j = h_{\mathsf{MF}}\sigma_j \quad h_{\mathsf{MF}} := Jdm + h. \tag{I.5.19}$$

As a by product, the Boltzmann distribution of the approximated model is very easy, since it factorizes, i.e. for $\boldsymbol{v} \in \{\pm 1\}^n$:

$$\mathbb{P}\left[\boldsymbol{\sigma} = \boldsymbol{v}; \beta, n, (J, h) \sim\right] = \prod_{j=1}^{n} \mathbb{P}\left[\sigma_j = v_j; \beta, m, (J, h), \sim\right] = \left(\mathbb{P}\left[\sigma = v; \beta, m, (J, h), \sim\right]\right)^n, \tag{I.5.20}$$

where the two steps are feasible since the mean-field choice makes the distribution iid.[17] To make matters clear, we have also expressed after the ";" what are the parameters of the local and global probabilities. The expression inside the parentheses is:

$$\mathbb{P}\left[\sigma = v; \beta, m, (J, h), \sim\right] = \frac{1}{\widetilde{\mathfrak{z}}(\beta)} e^{-\beta \widetilde{\hbar}(v)} \qquad \widetilde{\mathfrak{z}}(\beta) = \sum_{v \in \{\pm 1\}} e^{-\beta \widetilde{\hbar}(v)} \tag{I.5.21}$$

$$= \frac{e^{-\beta \widetilde{\hbar}(v)}}{e^{\beta h_{\mathsf{MF}}} + e^{-\beta h_{\mathsf{MF}}}}, \tag{I.5.22}$$

where we have denoted the <u>local</u> partition function as $\widetilde{\mathfrak{z}}(\beta)$. For a model, we need to specify (J, h, β). Since we do not know m in the equations, it is essentially a free parameter. Therefore, two very important remarks must be made.

1. Logically, one degree of freedom must be enforced, as the probability derived in Equation I.5.22 <u>must</u> be such that Eqn. I.5.18 is satisfied.

2. The efficacy of the approximation will depend on whether the predicted phenomenology of the model, at some triplet $(J, h, \beta) \in \mathbb{R}^2 \times \mathbb{R}_+$ and some m is reasonable/exact wrt to rigorous solutions. A key object in the conclusion will be the inverse temperature, by which the Boltzmann distribution is effectively modified in the way it puts weights on configurations. Since the m is not given by the model, but rather part of its phenomenology, we will judge the accuracy of the approximation by comparing m at given (J, β, h).

For the former, we find that, since we have imposed a "vertical" average over $j \in [n]$ in Equation I.5.18, we obtain that a "horizontal" average over the spin values must be satisfied. Indeed, in our notation for the surrogate model:

$$m = \frac{1}{n}\sum_{j=1}^{n} \langle \sigma_j \rangle_{\sim} = \frac{1}{n} n \langle \sigma_j \rangle_{\sim} = \langle \sigma_j \rangle_{\sim}, \tag{I.5.23}$$

by the decoupling, where $\sigma_j \in \{\pm 1\}$. Using the expression for the probability in Equation I.5.22, with less daunting notation:

$$m = \sum_{v \in \{\pm 1\}} \mathbb{P}\left[\sigma = v; \beta, m, (J, h)\right] v \tag{I.5.24}$$

$$= \frac{e^{\beta h_{\mathsf{MF}}} - e^{-\beta h_{\mathsf{MF}}}}{e^{\beta h_{\mathsf{MF}}} + e^{-\beta h_{\mathsf{MF}}}} \tag{I.5.25}$$

$$= \tanh(\beta h_{\mathsf{MF}}). \tag{I.5.26}$$

[16] for each j, the sum depends on j

[17] In Physics, the procedure is termed "decoupling".

Eventually, m must satisfy a so-called **self-consistent equation**:

$$m = \tanh(\beta h + \beta J d m). \tag{I.5.27}$$

Concerning the latter point, we redirect the reader to the nice plots and comments in the references (Evans 2009; Utermohlen 2018); for Curie-Weiss (Krzakala and Zdeborová 2021, Chap. I); and (Basak and Mukherjee 2017) for more advanced topics.

Remark I.5.28. *The factorization property is an aspect that will return, on a different light, in Chapter III, when we will discuss Variational Inference (Sec. III.3).*

Remark I.5.29. *In classic sources, it is always implicitly assumed that the probability depends on (β, J, h, m). In other models, the same will be done for their respective parameters. The choice is often forced: many objects come into the expressions at the same time, and the equations would become too convoluted. As a consequence, the reader must keep in mind that often trivial dependencies are suppressed in notation. To align with the other works, we will do the same from this paragraph onwards.*

Having discussed the foundations of a statistical definition of matter, we move to the foundations of the tools for analyzing it. In Chapter II, we will present some methods that are summoned frequently by Researchers. To give context, we will also try to be as self-contained as possible, outlining the basis and motivations for them.

Further References

For Statistical Physics and Thermodynamics (Arovas 2019; Cross 2006; Mehran 2023a,b)

Chapter II

Tools and Techniques in Statistical Physics

In this Chapter, we will overview some often ignored aspects of the theory underlying results in Statistical Physics. While we will spend sufficient time on each, by no means the following is a comprehensive review.

In Section II.1, we report the basic Gaussian integrals used throughout computations in literature. From these and other first principles, other results can be derived. In Section II.2, we provide a self-contained introduction to Complex Analysis, for the purpose of explaining the nature and peculiarities of an *analytic continuation*, which is an essential step in the Replica formalism. The idea is to reach the criticalities of the method, to shed light on arguments that involve such continuation when reading about papers that develop the heuristic replica method to retrieve the Thermodynamic properties of a model. Then, in Section II.3, we link the notion of free entropy to a purely statistical object, vindicating the claims mumbled in Chapter I. Section II.4 proceeds the discussion with the treatment of similar objects: integral transforms. While we do not spend much time on them, we argue the main principles behind the techniques, which are just nearly sufficient to grasp their power in conjunction with the Legendre Transform, which is discussed in Chapter III. Continuing, the content of Section II.5 is mainly a presentation of Dirac's delta distribution and two nice results involving it in Physics works. The former often appears in replica computations, while the latter is mainly a way to adjust units of measure in computations. Lastly, in II.6, we present two methods to perform integrals when n is large and influences the computations. As in other Sections, this method is often one of the many steps of Replica Theory.

II.1 Some useful Gaussian Integrals

The purpose of this collection is providing the *less* experienced reader with noteworthy mathematical results.

The starting triplet is directly derivable from the first, or *indirectly*, using the fact that a normal density sums to one, and reworking the coefficients. Many more Gaussian integrals can be derived using similar techniques.

Fact II.1.1. *It holds that:*

1. $\int_{-\infty}^{\infty} e^{-x^2}\,\mathrm{d}x = \sqrt{\pi}$

2. $\int_{-\infty}^{\infty} e^{-ax^2}\,\mathrm{d}x = \sqrt{\frac{\pi}{a}} \quad \forall a > 0$

3. $\int_{-\infty}^{\infty} e^{-ax^2+bx}\,\mathrm{d}x = \sqrt{\frac{\pi}{a}} e^{\frac{b^2}{4a}}, \quad \forall a > 0$

Proof. **(Claim #1)** Let $\mathcal{I} = \int_{-\infty}^{\infty} e^{-x^2}\,\mathrm{d}x = \int_{-\infty}^{\infty} e^{-y^2}\,\mathrm{d}y$ since the summand is a dummy index. Then:

$$
\begin{aligned}
\mathcal{I}^2 &= \left(\int_{-\infty}^{\infty} e^{-x^2}\,\mathrm{d}x\right)\left(\int_{-\infty}^{\infty} e^{-y^2}\,\mathrm{d}y\right) && \text{(II.1.2)}\\
&= \int_{-\infty}^{\infty} e^{-x^2}\left(\int_{-\infty}^{\infty} e^{-y^2}\right)\mathrm{d}y\,\mathrm{d}x && \text{(II.1.3)}\\
&= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} e^{-(x^2+y^2)}\right)\mathrm{d}y\,\mathrm{d}x && \text{bring inside } x \text{ argument as } x\perp y \quad \text{(II.1.4)}\\
&= \int_0^{2\pi}\int_0^{\infty} e^{-r^2} r\ \mathrm{d}r\,\mathrm{d}\vartheta && \text{change to polar coordinates } (\star) \quad \text{(II.1.5)}\\
&= 2\pi\int_{-\infty}^{0}\frac{1}{2}e^{s}\ ds && \text{substituting } s=-r^2,\ ds=-2r\,\mathrm{d}r \quad \text{(II.1.6)}\\
&= \pi\int_{-\infty}^{0} e^{s}\ ds && \text{(II.1.7)}\\
&= \pi(e^0-e^{-\infty}) && \text{(II.1.8)}\\
&= \pi \implies \sqrt{\mathcal{I}^2}=\int_{-\infty}^{\infty} e^{-x^2}\,\mathrm{d}x=\sqrt{\pi}. && \text{(II.1.9)}
\end{aligned}
$$

In $(\star)$ we mean:

$$
\begin{cases}
r=\sqrt{x^2+y^2} & \\
\vartheta=\tan^{-1}(\frac{y}{x}) & \\
|J|=r & \text{Jacobian determinant}
\end{cases}. \tag{II.1.10}
$$

Warning: we have **overlooked** some improper integrals to give a sketch of the proof. A more formal treatment is as follows. Let $y=xs, \mathrm{d}y=x\,\mathrm{d}s$. Since e^{-x^2} is even we have the identity $\int_{-\infty}^{\infty} e^{-x^2}\,\mathrm{d}x=2\int_0^{\infty} e^{-x^2}\,\mathrm{d}x$. In the positive half plane of $x\geqslant 0$ the variables (s,x) have the same sign. By the symmetry observed:

$$
\mathcal{I}^2=4\int_0^{\infty}\int_0^{\infty} e^{-(x^2+y^2)}\,\mathrm{d}y\,\mathrm{d}x=4\int_0^{\infty}\int_0^{\infty} e^{-x^2(1+s^2)}x\,\mathrm{d}s\,\mathrm{d}x=4\int_0^{\infty}\int_0^{\infty} e^{-x^2(1+s^2)}x\,\mathrm{d}x\,\mathrm{d}s \tag{II.1.11}
$$

$$
=4\int_0^{\infty}\left(\frac{e^{-x^2(1+s^2)}}{-2(1+s^2)}\bigg|_{x=0}^{x=\infty}\right)\mathrm{d}s=4\frac{1}{2}\int_0^{\infty}\frac{1}{1+s^2}\,\mathrm{d}s=2\arctan s\bigg|_{s=0}^{s=\infty}=\pi, \tag{II.1.12}
$$

where in the first line we have used Fubini's Theorem.
(Claim 2**)** Rework the integral as follows:

$$
\begin{aligned}
\int_{-\infty}^{\infty} e^{-ax^2}\,\mathrm{d}x &= \int_{-\infty}^{\infty} e^{-u^2}\frac{1}{\sqrt{a}}\ \mathrm{d}u && \text{substitute } u=\sqrt{a}x,\ \mathrm{d}u=\sqrt{a}\,\mathrm{d}x \quad \text{(II.1.13)}\\
&= \frac{1}{\sqrt{a}}\int_{-\infty}^{\infty} e^{-u^2}\ \mathrm{d}u && \text{(II.1.14)}\\
&= \frac{1}{\sqrt{a}}\sqrt{\pi} && \text{Claim \#1} \quad \text{(II.1.15)}
\end{aligned}
$$

(Claim 3**)** We prove the result in Lemma II.1.16 which is equivalent. □

Lemma II.1.16. *Let $X\sim\mathcal{N}(0,1)$. Then another perspective for Fct. II.1.1#3 is:*

$$
\mathbb{E}_X\left[e^{\kappa x}\right]=e^{\frac{1}{2}\kappa^2} \tag{II.1.17}
$$

Proof. Expanding the expectation:

$$\begin{aligned}
\mathbb{E}_X\left[e^{\kappa x}\right] &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} e^{\kappa x}\, \mathrm{d}x && \text{(II.1.18)}\\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}+\kappa x}\, \mathrm{d}x && \text{(II.1.19)}\\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x^2-2\kappa x)}\, \mathrm{d}x && \text{(II.1.20)}\\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}[(x-\kappa)^2-\kappa^2]}\, \mathrm{d}x \qquad \text{completing the square} && \text{(II.1.21)}\\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}[(x-\kappa)^2]} e^{-\frac{1}{2}(-\kappa^2)}\, \mathrm{d}x && \text{(II.1.22)}\\
&= e^{\frac{1}{2}\kappa^2} \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}[(x-\kappa)^2]}}_{x\sim\mathcal{N}(\kappa,1)\ \mathit{density}}\, \mathrm{d}x && \text{(II.1.23)}\\
&= e^{\frac{1}{2}\kappa^2}. && \text{(II.1.24)}
\end{aligned}$$

□

Lemma II.1.25. *The following equality holds:*

$$\int_{-\infty}^{\infty} (x-\mu) e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, \mathrm{d}x = \sigma^2, \quad \forall \mu, \sigma \in \mathbb{R}. \tag{II.1.26}$$

Proof. Proceed by substitution, letting:

$$u = -\frac{(x-\mu)^2}{2\sigma^2} \implies \mathrm{d}u = -\frac{x-\mu}{\sigma^2}\, \mathrm{d}x \implies -\sigma^2\, \mathrm{d}u = (x-\mu)\, \mathrm{d}x.$$

We have:

$$\int_{-\infty}^{\infty} exp\left\{ \underbrace{-\frac{(x-\mu)^2}{2\sigma^2}}_{=u} \right\} \underbrace{(x-\mu)\, \mathrm{d}x}_{=-\sigma^2 \mathrm{d}u} = -\sigma^2 \int_0^{\infty} e^u\, \mathrm{d}u = \sigma^2. \tag{II.1.27}$$

□

II.2 A primer on Complex Analysis

Further References

Some resources are (Arfken and Weber 2013), (Evans 2006, Lect. 2), (Weber and Arfken 2004), (Rudin 1987).

In this Section, we collect some important results to develop a theory for free energies, their phase transitions, and the saddle point method.

Definition II.2.1 (Analytic Function)**.** *A function $f : \mathbb{R} \to \mathbb{R}$ is analytic on an open set $D \subset \mathbb{R}$ if for $x_0 \in D$:*

$$f(x) = \sum_{n=0}^{\infty} c_n (x-x_0)^n \quad (c_n) \subset \mathbb{R} \qquad \forall x \in \mathcal{B}_{x_0} \tag{II.2.2}$$

Namely, the function admits a convergent power series expression in a neighborhood of x_0. Similarly a function is complex analytic, or simply analytic, when the previous sentence replaces $\mathbb{R}$ with $\mathbb{C}$.

Remark II.2.3 (Direct property of analytic functions)**.** *A careful evaluation of Definition II.2.1 shows that analytic functions are infinitely differentiable functions that admit a Taylor expansion at x_0 that is convergent **point-wise** in x to $f(x)$. This gives the trivial counterexample that a non differentiable function is non-analytic.*

An analytic function is differentiable in every neighborhood. While the notion is stronger than simple differentiability at a point, we will see that for complex functions it is equivalent.

Definition II.2.4 (Cauchy-Riemann conditions). *Two functions $u, v : \mathbb{R}^2 \to \mathbb{R}$ satisfy the Cauchy-Riemann conditions when:*

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y} \quad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}. \tag{II.2.5}$$

Definition II.2.6 (Laplace's Equation). *For a function in two variables $f : \mathbb{R}^2 \to \mathbb{R}$, we define Laplace's equation as:*

$$\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} = 0 \tag{II.2.7}$$

Having introduced some starting objects, we choose to express an analytic function $f : \mathbb{C} \to \mathbb{C}$ (Def. II.2.1) as:

$$f(z) = \Re(f(z)) + \mathrm{i}\Im(f(z)) := u(x, y) + \mathrm{i}v(x, y) \quad z = x + \mathrm{i}y \in \mathbb{C}, \tag{II.2.8}$$

and aim to define analogues of real-valued functions with real domain in the complex domain-codomain setting. As an example the derivative naturally extends to complex analysis, while the integral needs some care.

Definition II.2.9 (Derivative of complex function). *Let $f : \mathbb{C} \to \mathbb{C}$. Its derivative at $z_0 \in \mathbb{C}$ is defined as:*

$$f'(z_0) := \lim_{h \to 0, h \in \mathbb{C}} \frac{f(z_0 + h) - f(z_0)}{h}, \tag{II.2.10}$$

which as a byproduct, requires that the limit on the LHS does not depend on the path taken to approach 0. Indeed, $f(z)$ can be interpreted as a function of two real variables (x, y) via Eqn. II.2.8

Fact II.2.11. *A complex function as above is differentiable at a point z_0 if and only if it satisfies the Cauchy-Riemann Conditions at z_0 and (u, v) have continuous partial derivatives.*

Proof. By definition. □

Definition II.2.12 (Complex integral, contour integral). *Let $\gamma : [a, b] \to \mathbb{C}$ be a parametric curve, piecewise differentiable, taking values in the complex space. We define the integral of a complex function $f : \mathbb{R} \to \mathbb{C}$ as:*

$$\int_\gamma f(z)\,\mathrm{d}z = \int_a^b f(\gamma(t))\gamma'(t)\,\mathrm{d}t. \tag{II.2.13}$$

To unravel the RHS, we further define for $f(t) = u(t) + \mathrm{i}v(t)$ the integral:

$$\int_a^b f(t)\,\mathrm{d}t = \int_a^b u(t)\,\mathrm{d}t + \mathrm{i}\int_a^b v(t)\,\mathrm{d}t. \tag{II.2.14}$$

Lastly, a contour integral is a complex integral that takes place on a closed path across the complex plane, we denote it as $\oint$. To have a closed path, we will need an orientation, and it will be ***always anti-clockwise****, meaning that γ revolves anti-clockwise in its evolution.*

Definition II.2.15 (Holomorphic function). *A complex valued function taking complex values is holomorphic* ***on an open set*** *if it is complex differentiable for every point in that set. It is holomorphic* ***at a point*** *if it is differentiable in every point on some neighborhood of it.*

Fact II.2.16. *An analytic function is holomorphic.*

Proof. Functions admitting a power series representation in the sense of Def. II.2.1 are differentiable, see Rem. II.2.3. □

In the discussion below, we present a collection of proved & unproved statements to give sufficient acknowledgement to the roles in deriving some foundational results.

Proposition II.2.17 (Green's Theorem)**.** *Let $\mathcal{C} \subset \mathbb{R}^2$ be a countour, that closed, piece-wise, smooth, and simple. It can be seen as the boundary of the region it enclosed, i.e. $\mathcal{C} = \partial R$ of a region $R \subset \mathbb{R}^2$, where R is bounded by ∂R. Further, let f, g have domain D open, and such that $R \subset D$, with continuous partial derivatives in R. Then:*

$$\oint\limits_{\mathcal{C}} f(x,y)\mathrm{d}x + \oint\limits_{\mathcal{C}} g(x,y)\mathrm{d}y = \iint\limits_{R} \left(\frac{\partial g(x,y)}{\partial x} - \frac{\partial f(x,y)}{\partial y}\right) \mathrm{d}x\mathrm{d}y. \tag{II.2.18}$$

Proof. We refer to (Greenlee 2005; Zenisek 1999). □

Remark II.2.19. *For further intuition, we suggest to consult (Weber and Arfken 2004) and (Arfken and Weber 2013, Chap. 1.11). In particular, we report some comments of the latter, and (Arfken and Weber 2013, Eqns. 1.101a-1.104).*
Seen from the practical perspective, the Proposition claims that the integral over a surface of the vector field (f, g) is equivalent to the divergence of the vector field over the whole enclosed region. The notation often seen starts from Gauss' Theorem:

$$\oiint\limits_{\partial R} \boldsymbol{F} \cdot \mathrm{d}\boldsymbol{v} = \iiint\limits_{R} \nabla \cdot \boldsymbol{F} \mathrm{d}r, \tag{II.2.20}$$

and applies the following equalities for scalar valued functions:

$$\nabla \cdot (f\nabla g) = f\nabla \cdot \nabla g + (\nabla f) \cdot (\nabla v) \tag{II.2.21}$$

$$\nabla \cdot (g\nabla f) = g\nabla \cdot \nabla f + (\nabla g) \cdot (\nabla f). \tag{II.2.22}$$

Indeed, subtracting the two, and integrating over $(x,y) \in R$, one obtains the expression:

$$\iiint\limits_{R} (f\nabla \cdot \nabla g - g\nabla \cdot \nabla f)\mathrm{d} = \oiint\limits_{\partial R} (f\nabla g - g\nabla f) \cdot \mathrm{d}\boldsymbol{v}, \tag{II.2.23}$$

which is the same as the statement above, but with different notation, hopefully highlighting the interpretation.

Proposition II.2.24 (Cauchy-Goursat Theorem)**.** *If f is holomorphic on a connected domain $\mathscr{X}$, and $\mathcal{C}$ is a closed contour inside the domain. Then:*

$$\oint\limits_{\mathcal{C}} f(z)\,\mathrm{d}z = 0. \tag{II.2.25}$$

As a consequence, if γ is a curve in $\mathscr{X}$ that has endpoints $[z_1, z_2]$ then:

$$\int_{\gamma} f'(z)\,\mathrm{d}z = f(z_2) - f(z_1). \tag{II.2.26}$$

Namely, the integral is independent of the path. And one can deform it as pleased.

Proof. We prove the statement here for continuous partial derivatives, reporting the standard computation. For the full result, an option is (Moore 1900).
Let $f(z) = u(x,y) + \mathrm{i}v(x,y)$, in its real an imaginary part decomposition as in Definition II.2.12. Its integration will then be in terms of $\mathrm{d}z = \mathrm{d}x + \mathrm{i}\mathrm{d}y$, with an integral that can eb written as:

$$\oint\limits_{\mathcal{C}} f(z)\mathrm{d}z = \oint\limits_{\mathcal{C}} (u + \mathrm{i}v\mathrm{d}(x + \mathrm{i}y) \tag{II.2.27}$$

$$= \oint\limits_{\mathcal{C}} u\mathrm{d}x - \oint\limits_{\mathcal{C}} v\mathrm{d}y + \mathrm{i}\left(\oint\limits_{\mathcal{C}} v\mathrm{d}x + \oint\limits_{\mathcal{C}} u\mathrm{d}y\right), \tag{II.2.28}$$

where the negative coefficient appears since $\mathrm{i}^2 = -1$ and we stress that $u \equiv u(x,y), v \equiv v(x,y)$. Applying Green's Theorem (Prop. II.2.17) to both parts returns two "volume" integrals over R, the region enclosed by $\mathcal{C} \equiv \partial R$:

$$\oint_{\mathcal{C}} u\mathrm{d}x - \oint_{\mathcal{C}} v\mathrm{d}y = \iint_R \left(-\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}\right)\mathrm{d}x\mathrm{d}y \tag{II.2.29}$$

$$\oint_{\mathcal{C}} v\mathrm{d}x + \oint_{\mathcal{C}} u\mathrm{d}y = \iint_R \left(\frac{\partial u}{\partial x} - -\frac{\partial v}{\partial y}\right)\mathrm{d}x\mathrm{d}y. \tag{II.2.30}$$

Recalling Definitions II.2.4, II.2.6 and II.2.11, the function f must satisfy Cauchy-Riemann's conditions, which in this case are:

$$\frac{\partial u(x,y)}{\partial x} = \frac{\partial v(x,y)}{\partial y}, \quad \frac{\partial u(x,y)}{\partial y} = -\frac{\partial v(x,y)}{\partial x}, \tag{II.2.31}$$

Pluggin the first condition (resp. the second) into the second (first) result of Green's Theorem, we find that the integrands are null, and:

$$\oint_{\mathcal{C}} f(z)\mathrm{d}z = \oiint_R 0\mathrm{d}x\mathrm{d}y + \mathrm{i}\oiint_R 0\mathrm{d}x\mathrm{d}y = 0, \tag{II.2.32}$$

as claimed. □

We provide two instrumental examples for a (potentially) non-zero complex integral along a contour.

Example II.2.33. *For any $n \in \mathbb{N}$, let the contour $\mathcal{C}$ be at fixed distance r from the origin. Then:*

$$\oint_{\mathcal{C}} z^n\mathrm{d}z = \int_0^{2\pi} r^n e^{\mathrm{i}n\theta}\mathrm{i}re^{\mathrm{i}\theta}\mathrm{d}\theta = r^{n+1}\frac{e^{2\pi\mathrm{i}n} - 1}{n+1}, \tag{II.2.34}$$

where the first step applies the substitution $z = re^{\mathrm{i}\theta}$ for $\theta \in [0, 2\pi], \mathrm{d}z = \mathrm{i}re^{\mathrm{i}\theta}\mathrm{d}\theta$. We will see that the only removable singularity of this result is $n = -1$.

Example II.2.35. *Let $a \in \mathbb{R}$ be contained in a contour $\mathcal{C}$ which is a ball with radius r centered at the origin. Then:*

$$\oint_{\mathcal{C}} \frac{1}{z-a}\mathrm{d}z = \oint_{\mathcal{C}} \frac{1}{w}\mathrm{d}w \qquad \textit{substitute } w = z - a \tag{II.2.36}$$

$$= \int_0^{2\pi} \frac{1}{r}e^{-\mathrm{i}\theta}\mathrm{i}re^{\mathrm{i}\theta}\mathrm{d}\theta \qquad = 2\pi\mathrm{i} \tag{II.2.37}$$

$$= 2\pi\mathrm{i}, \tag{II.2.38}$$

which shows that if the n is chosen before integrating the integral of z^1 over a closed curve is non-zero. We will use this result in the next Theorem, that also tells us that the other $n \neq -1$ will return null integrals.

With the above results, we can quickly derive the celebrated Cauchy's integral formula.

Theorem II.2.39 (Cauchy's Integral Formula)**.** *Let $f : A \to \mathbb{C}$ be a holomorphic function on an open subset of $\mathbb{C}$, and $\mathcal{C}$ be the anti-clockwise oriented boundary of a region R. Then, for any $z_0 \in R$:*

$$f(z_0) = \frac{1}{2\pi\mathrm{i}}\oint_{\mathcal{C}} \frac{f(z)}{z - z_0}\mathrm{d}z. \tag{II.2.40}$$

More in general, it is also true that the functions are automatically infinitely differentiable (see Fct II.2.49 below) and:

$$f^{(k)}(z_0) = \frac{n!}{2\pi\mathrm{i}}\oint_{\mathcal{C}} \frac{f(z)}{(z - z_0)^{n+1}}\mathrm{d}z, \tag{II.2.41}$$

where $f^{(k)}$ is the k^{th} derivative of f.

Proof. We prove the first statement. The second can be shown by just applying the definition of limit, and observing that it can be brought inside the integral since the denominator is never zero in the Formula of Cauchy when a displacement $h \to 0$ is added to the input.
By the Cauchy-Goursat Theorem (Thm. II.2.24), any arbitrarily small closed curve around z_0 returns the same integration result. Therefore, we choose a circle with radius ϵ, arbitrarily small. Since f is continuous (by f being holomorphic, this follows by Definition), the value of $f(z)$ will be arbitrarily close to $f(z_0)$. Notice also that by the previous examples, we have:

$$2\pi \mathrm{i} = \oint\limits_{\mathcal{C}} \frac{1}{z - z_0} \mathrm{d}z \implies f(z_0) = \oint\limits_{\mathcal{C}} \frac{2\pi \mathrm{i}}{z - z_0} \mathrm{d}z. \tag{II.2.42}$$

Therefore, for arbitrary $z \in \mathbb{C}$ in such circle of radius ϵ, we can express the difference with the usual polar substitution $z \equiv z(\theta)$, with ϵ fixed:

$$\begin{aligned} 0 &\leqslant \left| \frac{1}{2\pi \mathrm{i}} \oint\limits_{\mathcal{C}} \frac{f(z)}{z - z_0} - f(z_0) \mathrm{d}z \right| && \text{(II.2.43)} \\ &= \left| \frac{1}{2\pi \mathrm{i}} \oint\limits_{\mathcal{C}} \frac{f(z) - f(z_0)}{z - z_0} \mathrm{d}z \right| && \text{(II.2.44)} \\ &= \left| \frac{1}{2\pi \mathrm{i}} \int_0^{2\pi} \frac{f(z(\theta)) - f(z_0)}{\epsilon e^{\mathrm{i}\theta}} \mathrm{i}\epsilon e^{\mathrm{i}\theta} \mathrm{d}\theta \right| && \text{(II.2.45)} \\ &\leqslant \frac{1}{2\pi} \int_0^{2\pi} |f(z(\theta)) - f(z_0)| \mathrm{d}\theta, && \text{(II.2.46)} \end{aligned}$$

where the inequality is the simple triangle inequality for integrals. It suffices then to observe that z_0 is at the center of the circle, and $z(\theta)$ lies on the boundary, therefore $|z(\theta) - z_0| = \epsilon$ is constant, and we can bound by the maximum value attained on the circle, i.e.:

$$\frac{1}{2\pi} \int_0^{2\pi} f(z(\theta)) - f(z_0) \mathrm{d}\theta \leqslant \max_{\theta : |z(\theta) - z_0| = \epsilon} \{ |f(z(\theta)) - f(z_0)| \}. \tag{II.2.47}$$

By continuity of f, as $\epsilon \to 0$ the terms $f(z(\theta))$ and $f(z_0)$ get closer: for each set resolution ε on the functions there exists a resolution $\epsilon > 0$ on (z, z_0) that attains it. Then:

$$\lim_{\epsilon \to 0} \max_{\theta : |z(\theta) - z_0| = \epsilon} \{ |f(z(\theta)) - f(z_0)| \} = 0. \tag{II.2.48}$$

By positivity of the modulus, the difference of the objects in the claim, found just after the first inequality, is squeezed in between null expressions. □

Fact II.2.49. *Let f be holomorphic on a set. Then it is infinitely differentiable on that set and it can be expanded as a power series. By combining the two arguments, it is analytic on that set.*

We are eventually brought to consider in any case the words analytic and holomorphic to be equivalent for complex-valued functions.

Theorem II.2.50. *A complex-valued function f is analytic if and only if it is holomorphic.*

Proposition II.2.51 (Morera's Theorem)**.** *Let $f : \mathbb{C} \to \mathbb{C}$ be continuous in a simply connected region*[1] *such that any closed countour integral within the region is null, i.e. $\oint_{\mathcal{C}} f(z)\, \mathrm{d}z = 0$ for all $\mathcal{C}$ in the region. Then f is analytic (Def. II.2.1) in the region.*

[1] this is a notion in topology, it roughly means a space with no holes. We can take it for granted and assume it is the non-degenerate case.

Proof. We report the proof since it is very quick. It is a rewriting of the one in (Weber and Arfken 2004, Sec. 6.4), and build on an explicit construction of the antiderivative. Let z_2, z_1 be the endpoints of a curve in the region. By Cauchy-Goursat Theorem (Prop. II.2.24), integrals only depend on initial and final values of the contour. Then:

$$F(z_2) - F(z_1) = \int_{z_1}^{z_2} f(z)\,\mathrm{d}z \tag{II.2.52}$$

is well defined. By continuity of f and construction the following equalities hold:

$$\lim_{z_2 \to z_1} \frac{F(z_2) - F(z_1)}{z_2 - z_1} - f(z_1) = \lim_{z_2 \to z_1} \frac{\int_{z_1}^{z_2} f(z) - f(z_1)\,\mathrm{d}z}{z_2 - z_1} = 0. \tag{II.2.53}$$

Then, by definition of complex derivative (Def. II.2.9) we have that rearranging the above result:

$$\lim_{z_2 \to z_1} \frac{F(z_2) - f(z_1)}{z_2 - z_1} = F'(z)\Big|_{z=z_1} = f(z_1). \tag{II.2.54}$$

By arbitrariness of z_2, z_1, the claim is proved for any z in the region. Now we claim that the Cauchy integral formula generalized to higher derivatives (Thm. II.2.39) guarantees that taking F on the outer contour, an analytic function has analytic derivatives. So by $F'(z) = f(z)$ and F being analytic (it is holomorphic), we get that $f(z)$ is analytic since it is the derivative of an analytic function. □

Having derived some basic results in complex analysis, we can proceed with those that we actually need for our narrative on Statistical Physics and Machine Learning/Inference.

Analytic Continuation

An interesting property of functions is the possibility of extending them to wider domains. The natural question that comes into mind is if such extension is just arbitrary or unique. We will briefly discuss the *continuation* scheme for analytic functions since it is fundamental for the theory of Replicas in Statistical Physics.

Definition II.2.55 (Continuation)**.** *Let $f : \mathbb{C} \to \mathbb{C}$ have domain $\mathcal{U} \subset \mathbb{C}$, where $\mathcal{U}$ is open. Let g have domain $\mathcal{V} \supset U$. If $g(z) = f(z)$ for all $z \in \mathcal{U}$, g is a continuation of f.*

Analytic (equivalently, holomorphic) functions require a mild condition to be uniquely continued from one set to a larger one. The sufficient condition relates to the smaller set having an accumulation point in the larger one.

Proposition II.2.56 (Identity Theorem for analytic functions)**.** *Let f, g be analytic on a domain $\mathcal{D}$, open and connected. If $f \equiv g$ on $\mathcal{S} \subset \mathcal{D}$ where $\mathcal{S}$ has an accumulation point in $\mathcal{D}$ then $f \equiv g$ on the whole domain.*

By the above result, functions that are equal on a set that is able to arbitrarily approximate values in another larger set makes the candidates equivalent on the latter. We emphasize that the notion of analytic function (Def. II.2.1) is however stronger than simple differentiability. As a corollary, one obtains that analytic continuations are unique as long as two continuations agree on a well behaved set.

Corollary II.2.57 (Analytic continuation is unique)**.** *Let $f : \mathbb{C} \to \mathbb{C}$ be analytic on a domain $\mathcal{U}$ open. Let two candidate continuations g_1, g_2 be analytic and satisfy Def. II.2.55 for a set $\mathcal{U} \supset \mathcal{V}$. Then necessarily $g_1 \equiv g_2$ in $\mathcal{V}$, and continuations are unique.*

Proof. Let $g \equiv g_1 - g_2$. By the hypothesis, it vanishes on the open set $\mathcal{U}$. The zeros of g are not isolated in $\mathcal{V}$ since the subset $\mathcal{U}$ is open. By Prop. II.2.56 $g \equiv 0$ on $\mathcal{V}$ and g_1, g_2 are identical. □

In the case we will see, the condition is not satisfied since we will want to perform a continuation from $\mathcal{U} = \mathbb{N}$ to $\mathbb{R}$ which is not open in $\mathcal{V} = \mathbb{R}$. Nevertheless, a field of study has explored such question and led to some sufficiency statements. We report two of them below with a counterexample. While the literature is out of the scope of this document, these three instances should satisfy the reader wishing to understand if uniqueness of analytic continuations holds from the natural numbers to the real line.

Proposition II.2.58 (Carlson's Theorem). *Let $f : \mathbb{C} \to \mathbb{C}$ satisfy the following:*

1. *f is holomorphic and of sub-exponential type, i.e. $|f(z)| \leqslant Ce^{\tau z}$ for some C, τ and all $z \in \mathbb{C}$*
2. *there exists $k < \pi$ such that $|f(\mathrm{i}x)| \leqslant Ce^{k|x|}$ for some C and all $x \in \mathbb{R}$*
3. *$f(n) \equiv 0$ for all $n \in \mathbb{N}$.*

Then $f \equiv 0$ on $\mathbb{C}$.

Corollary II.2.59 (Relaxations to Carlson's conditions). *The same conclusion of Prop. II.2.58 holds when #1 is replaced with f being analytic in $\Re(z) > 0$ and continuous in $\Re(z) \geqslant 0$ with the same exponential bound. One can also relax #3 by requiring that f is null on a set $A \subset \mathbb{N}$ such that:*

$$\limsup_{n\to\infty} = \frac{|A \cap \{0, 1, \ldots, n-1\}|}{n} = 1, \tag{II.2.60}$$

which was formulated by (Rubel 1956).
If one substitutes these in the statement, the result is sharp. As a consequence, if the updated requirements are not satisfied, the function is <u>not</u> identically zero.

To verify the power of the statement, we briefly outline how to use it. Consider a function f on the natural numbers and two candidate continuations g_1, g_2 which agree on the natural numbers. If #1, #2 are further satisfied by $g \equiv g_1 - g_2$, then the continuation is unique since $g \equiv 0$ on $\mathbb{C}$. Replacing the statements with the relaxed conditions, we also conclude that if the requirements are *not* satisfied, then the continuation is *not* unique.

A similar condition is derived by another combination of statements. To apply them, we will anticipate the Fourier Transform (Def. II.4.4).

Theorem II.2.61 (Paley-Wiener Theorem (Rudin 1987), Thm. 19.3). *Let f satisfy the following:*

- *f is holomorphic on $\mathbb{C}$ and square integrable, i.e.*

$$\int_{-\infty}^{\infty} (f(x))^2 \,\mathrm{d}x < \infty, \tag{II.2.62}$$

- *$f(z) \leqslant ce^{C|z|}$ for some c, C and all $z \in \mathbb{C}$.*

Then f is the inverse Fourier transform (Fct. II.4.9) of some $h \in L^2([-C, C])$, i.e. in our notation

$$f(z) = \int_{-C}^{C} h(x)e^{\mathrm{i}xz} \,\mathrm{d}x \quad \forall z \in \mathbb{C}. \tag{II.2.63}$$

Corollary II.2.64 (Uniqueness of analytic continuation from Naturals II). *Let f be defined on the natural numbers. Consider two continuations g_1, g_2 to the set $\mathbb{R}$. If g_1, g_2 satisfy the hypothesis above, then $g_1 \equiv g_2$ on $\mathbb{R}$.*

Proof. Let $g \equiv g_1 - g_2$. By construction g satisfies the requirements of Thm. II.2.61. Then there exists some $h \in L^2([-C, C])$ such that $h = g^{\mathsf{Fou}}$. Being continuations, they agree on the domain of f, so:

$$f(n) = \int_{-C}^{C} h(x)e^{\mathrm{i}nx} \,\mathrm{d}x = 0 \quad \forall n \in \mathbb{N}. \tag{II.2.65}$$

As a consequence, $h \equiv 0$ by the arbitrariness of n and $g \equiv 0$. □

A straightforward counterexample is the evidence that there is more than one extension of $n!$ to the real numbers. The classic is the Gamma function, but one can also derive the Hadamard Function, and other exotic objects. For historical references, see (Luschny 2010).

The problem in Statistical Physics In Replica Theory, the continuation of a certain Representation of the free energy $\ln \mathcal{Z}$ will be needed. There, the free energy $\ln \mathcal{Z} = \lim_{n\to\infty} \ln \mathcal{Z}_n$ is in general hard to compute. To overcome its computation, the replica trick is implemented

$$\ln \mathcal{Z}_n = \lim_{r\to 0} \frac{\mathcal{Z}_n^r - 1}{r}, \tag{II.2.66}$$

where powers of partition functions are instead easier to compute as products of integrals. As evident from the expression, while powers over the natural numbers of an integral are well defined, the r^{th} power for $r \to 0$ requires to peform an analytic continuation of the function from the naturals to the real line.
However, one will need to apply the limit $n \to \infty$ first to obtain a reliable free energy (since it is an object that exists at the thermodynamic limit), inducing all the potential singularities (phase transitions) discussed in Section I.4. In practice, it amounts to exchanging $\lim_{n\to\infty} \lim_{r\to 0} \mathcal{Z}_n^r$, which would be the mathematically right but physically meaningless object with $\lim_{r\to 0} \lim_{n\to\infty} \mathcal{Z}_n^r = \lim_{r\to 0} \mathcal{Z}^r$, which is representative of the right object, but potentially incorrect from the perspective of Mathematics.

To finally justify why the partition function is such an important object, we take a full Section, discussing very important objects in the Theory of Statistics that provide a different interpretation, with apparently no link with the energy-entropy competition argument.

II.3 Cumulants

In this Section, we give a formal justification of the claim that the partition function/free energy/free entropy is a complete descriptor of the system. The justification is in statistical terms and applies to problems well out of Statistical Physics.

As a first step, we present three objects with very peculiar properties. Despite being very similar , and equivalent in their power in many cases, they appear in different procedures, and thus may provide different perspectives. These are the Moment Generating Function (MGF), the Characteristic Function (CF) and the Cumulant Generating Function (CGF).

Definition II.3.1 (Moment Generating Function)**.** *For a random variable X we define:*

$$M_X(t) := \mathbb{E}\left[e^{tX}\right] \quad t \in \mathbb{R}, \tag{II.3.2}$$

provided that in a neighborhood of t the expectation on the RHS is computable.

Definition II.3.3 (Characteristic function)**.** *Given a random variable X, define:*

$$\phi_X(t) := \mathbb{E}\left[e^{iXt}\right] \quad t \in \mathbb{R}. \tag{II.3.4}$$

A series expansion wrt t of the exponential inside the MGF around zero gives the following nice interpretation:

$$M_X(t) = 1 + t\mathbb{E}\left[X\right] + \frac{t^2}{2}\mathbb{E}\left[X^2\right] + \ldots = 1 + \sum_{n\geqslant 1} \mathbb{E}\left[X^n\right]\frac{t^n}{n!}, \tag{II.3.5}$$

which suggests without the need of a proof that it is possible to determine the moments of the random variable by differentiating:

$$\mathbb{E}\left[X^n\right] = \frac{d^n}{dt^n} M_X(t)\Big|_{t=0}. \tag{II.3.6}$$

Independency or random variables also ensures that the MGF of the sum is the product of the MGFs, thanks to the elementary property of exponentials

$$M_{X+Y}(t) = \mathbb{E}_{X,Y}\left[e^{t(X+Y)}\right] = \mathbb{E}_X\left[e^{tX}\right]\mathbb{E}_Y\left[e^{tY}\right] = M_X(t)M_Y(t). \tag{II.3.7}$$

The characteristic function enjoys similar properties, but is also guaranteed to always exist since $e^{\mathrm{i}tX}$ is always integrable for any distribution X might have. We will see it is as a (weaker) consequence of Fct. II.4.8, since the measure of a random variable is always integrable and $|e^{\mathrm{i}tx}| \leqslant 1$.
Surpsingly enough, these properties extend to the *logarithmic equivalent* in a more comfortable way.

Definition II.3.8 (Cumulant generating function, and cumulants)**.** *For a random variable X the CGF is the logarithm of the MGF.*

$$K_X(t) := \ln\mathbb{E}\left[e^{tX}\right]. \tag{II.3.9}$$

Cumulants are the Taylor series expansion of the CGF around $t = 0$

$$K_X(t) = \sum_{n\geqslant 1}\kappa_n\frac{t^n}{n!} \quad \kappa_n = \frac{d^n}{dt^n}K_X(t)\bigg|_{t=0}. \tag{II.3.10}$$

Remark II.3.11. *The previous identity relating the MGF to moments does not work. Indeed:*

$$K_X(t) = \ln\mathbb{E}\left[e^{tX}\right] = \ln\left(1 + \sum_{n\geqslant 1}\mathbb{E}\left[X^n\right]\frac{t^n}{n!}\right). \tag{II.3.12}$$

The immediate reason is that since the logarithm argument is a sum, ***not a product****, nothing much can be done. In other terms,* ***cumulants are not the logarithm of moments****.*

Nevertheless, calculations show that the cumulants are moments of the random variable up to $n = 2$. For $n > 2$, cumulants are **functions of the moments up to the** n^{th}. We report below a summary of properties of the three with appropriate references to the proofs. After this list, we will establish the existance conditions.

Proposition II.3.13 (Properties of the MGF, CF, and the CGF)**.** *The following results hold:*

1. *The MGF and the CF uniquely determine the distribution of a random variable.*
2. *the MGF and the CF of the sum of independent variables factorize*
3. *the MGF and the CF are linear operators*
4. *the CGF is translation invariant and homogeneous wrt constant factors $c \in \mathbb{R}$, namely for a random variable X we have:*

$$K_{X+c}(t) = K_X(t) \quad K_{c^p X}(t) = c^p K_X(t), \tag{II.3.14}$$

5. *the CGF exists whenever the MGF exists and uniquely determines distributions*
6. *the m^{th} moment of a random variable admits an expression in terms of the MGF and CF derivatives of order m*
7. *the m^{th} moment of a random variable admits an expression in terms of the CGF derivatives of order m for $m \leqslant 4$ and $\leqslant m$ for above the third moment*

Proof. Claims #2, 3, 4, 6 are trivial. Claim #7 involves tedious calculations but is trivial as well. Claim #5 is implied by Claim #1 and noting that the CGF is just the logarithm of the MGF by definition.
Claim #1 is implicitly proved in Fcts. II.4.9, II.4.12, and the discussion just below their statements. □

We motivate the usefulness of the three generating functions by inspecting Prop. II.3.13.

- Firstly, the expressions are in bijection with the distribution they refer to. This allows to use them without ambiguities.
- Secondly, they provide closed form expressions for the moments of the original random variable. Moments describe the shapes that a function has: the mean is the center, the variance is the spread, the skewness is the third, the kurtosis is the fourth, and so on. The more, the better, and having access to all of them is equivalent to being able to *draw* the distribution in a heuristic sense. More formally, having access to more moments means that the shape of the distribution is easier to explore with dynamics.
- Lastly, they obey some nice properties that go beyond. Independent variables factorize in MGFs and CFs when summed. While the evaluation of the distribution of a sum might be daunting, multiplying functions is not.

In the main model of Equilibrium Statistical Mechanics, they provide a clean justification of some of the claims we made in Chapter I, as we will see in the next Subsection.

II.3.1 The Generating functions of the canonical ensemble

In a canonical ensemble the random variable is the energy. Assuming for simplicity that it can take only discrete values the form is as in Chapter I

$$p_i := p(\mathscr{E}_i) = \frac{1}{\mathcal{Z}(\beta)} e^{-\beta\mathscr{E}_i} \qquad \sum_{i=1} p_i = 1 \tag{II.3.15}$$

The MGF is easily found to be:

$$M_{\mathscr{E}}(t) = \left\langle e^{t\mathscr{E}} \right\rangle_\beta = \sum_i p(\mathscr{E}_i) e^{t\mathscr{E}_i} = \frac{1}{\mathcal{Z}(\beta)} \sum_i e^{-(\beta-t)\mathscr{E}_i} = \frac{\mathcal{Z}(\beta-t)}{\mathcal{Z}(\beta)}. \tag{II.3.16}$$

By the discussion above, the moments admit an expression as derivatives of the MGF. A trivial calculation makes the claim explicit.

$$\langle \mathscr{E}^m \rangle = \frac{d^m}{dx^m} M_X(t)\bigg|_{t=0} = \frac{1}{\mathcal{Z}(\beta)} \frac{\partial^m}{\partial t^m} \mathcal{Z}(\beta-t)\bigg|_{t=0} = (-1)^m \frac{\frac{d^m}{d(\beta-t)^m}\mathcal{Z}(\beta-t)}{\mathcal{Z}(\beta)}\Bigg|_{t=0} = (-1)^m \frac{\frac{d^m}{d\beta^m}\mathcal{Z}(\beta)}{\mathcal{Z}(\beta)}, \tag{II.3.17}$$

where it is useful to notice that the derivative is now wrt β. The identity allows us to write nicely the first moment of the energy:

$$\langle \mathscr{E} \rangle = -\frac{1}{\mathcal{Z}(\beta)} \frac{\partial \mathcal{Z}(\beta)}{\partial \beta} = \frac{1}{\mathcal{Z}(\beta)} \sum_i \mathscr{E}_i e^{-\beta\mathscr{E}_i} = \sum_i \mathscr{E}_i p_i. \tag{II.3.18}$$

Remark II.3.19. *We will see later (Subsec. III.2.2) that the identity in Eqn. II.3.17 shows that up to normalization energy and temperature are conjugated (Def. I.2.25).*

From Eqn. II.3.16 we can directly express the CGF as:

$$K_{\mathscr{E}}(t) = \ln M_{\mathscr{E}(t)} = \ln \mathcal{Z}(\beta - t) - \ln \mathcal{Z}(\beta), \tag{II.3.20}$$

and report below the first three cumulants for completeness.

$$\kappa_1 = \frac{d}{dt} K_{\mathscr{E}}(t)\bigg|_{t=0} = -\frac{1}{\mathcal{Z}(\beta)} \frac{d}{d(\beta-t)} \mathcal{Z}(\beta-t) = -\frac{1}{\mathcal{Z}(\beta)} \frac{d}{d\beta} \mathcal{Z}(\beta) = \langle \mathscr{E} \rangle \quad \text{by the result of Eqn. II.3.18} \tag{II.3.21}$$

$$\kappa_2 = \frac{\frac{d^2}{d\beta^2}\mathcal{Z}(\beta)}{\mathcal{Z}(\beta)} - \left(\frac{\frac{d}{d\beta}\mathcal{Z}(\beta)}{\mathcal{Z}(\beta)} \right)^2 = \left\langle \mathscr{E}^2 \right\rangle - \langle \mathscr{E} \rangle^2 \tag{II.3.22}$$

$$\kappa_3 = \ldots = \left\langle (\mathscr{E} - \langle \mathscr{E} \rangle)^3 \right\rangle. \tag{II.3.23}$$

A simple argument allows us to reconcile the partition function with cumulants. The derivative wrt t can be expressed as a derivative wrt β upon a change of sign in the definition of CGF.
Assigning to the CGF $K_{\mathscr{E}}(\beta) = \ln \mathcal{Z}(\beta)$ we recover the same cumulants of before via a slightly modified definition:

$$\kappa_n = (-1)^{n-1} \frac{d^n}{d\beta^n} K_{\mathscr{E}}(\beta). \tag{II.3.24}$$

From now onwards, we will use such perspective when dealing with cumulants.

Remark II.3.25. *Notice that in the new assignment of the CGF and the cumulants the CGF is the Free Entropy $\mathscr{F}$, defined just below the free energy (Def. I.3.34).*

Here the jargon becomes daunting, but it is easy to conclude that the free energy/entropy[2] happens to be the CGF of the canonical ensemble. This explains briefly why working with the partition function/free energy/free entropy is beneficial. In simple terms, we are able to recover any average over the randomness we might desire, and Thermodynamic quantities (i.e. derivatives of the partition function) are just functions of the moments of the random variable of the system.

Remark II.3.26. *Having assumed no thermodynamic setting, the result extends to any partition function object. Such conclusion is very important in the general sense since most of Bayesian statistics seeks to answer questions around $\mathcal{Z}_n$ even without taking the limit $n \to \infty$.*

II.4 Integral Transforms

Further References

A non-exhaustive collection of useful resources is (Figueroa O'Farril 1998; Weber and Arfken 2004), (Evans 2006, Lect. 12).

To begin, we define an object which will be of great interest.

Definition II.4.1 (Laplace Transform)**.** *For a function f defined on the positive real axis, its Laplace transform is:*

$$\mathcal{L}[f](s) := \int_0^\infty f(t) e^{-st}\, \mathrm{d}t \quad s \in \mathbb{C}. \tag{II.4.2}$$

In a similar way we define the bilateral Laplace transform for functions over the real line taking values on the complex plane as:

$$\mathcal{L}_\pm[f](s) := \int_{-\infty}^\infty f(t) e^{-st}\, \mathrm{d}t \quad s \in \mathbb{C}. \tag{II.4.3}$$

Up to constants, the bilateral Laplace transform is strongly related to the Fourier Transform, which we reported below to choose one of the standard formulas for it.

Definition II.4.4 (Fourier Transform)**.** *Given a function $f : \mathbb{R} \to \mathbb{R}$, define:*

$$\mathcal{F}[f](w) := f^{\mathsf{Fou}}(w) = \frac{1}{2\pi} \int_{-\infty}^\infty f(x) e^{-\mathrm{i}wx}\, \mathrm{d}x, \tag{II.4.5}$$

whenever the integral is computable.

Remark II.4.6. *The connection between bilateral Laplace transform and Fourier transform is seen as:*

$$2\pi \mathcal{F}[f](\mathrm{i}w) = \mathcal{L}_\pm[f](w). \tag{II.4.7}$$

Fact II.4.8 (Sufficient conditions for Laplace and Fourier transform)**.** *Let $f : \mathbb{R} \to \mathbb{R}$ be a function on the reals. Then:*

[2] discarding a β factor if needed

1. *if* $\|f\|_2^2 = \int_{-\infty}^{\infty}(f(x))^2\,\mathrm{d}x < \infty$ *then the Fourier transform exists*

2. *if* $|f(x)| \leqslant Ce^{\alpha x}$ *for all* x *and* $\Re(s) > \alpha$ *the Laplace transform exists at* $\mathscr{L}[f](s)$

Fact II.4.9 (Inverse Laplace and Fourier)**.** *One can show that the inverse transformation for Laplace integrals is given by the formula:*

$$\mathscr{L}^{-1}[f^{\mathsf{Lap}}(s)] = \frac{1}{2\pi\mathrm{i}} \lim_{\tau\to\infty} \int_{\gamma-\mathrm{i}\tau}^{\gamma+\mathrm{i}\tau} f^{\mathsf{Lap}}(s)e^{st}\,\mathrm{d}s \qquad \tau \in \mathbb{R} \tag{II.4.10}$$

where f^{Lap} *is the Laplace transform and* γ *is big enough as to have it defined everywhere*
Similarly, if f *is square integrable and continuous, one has:*

$$f(x) = \int_{-\infty}^{\infty} f^{\mathsf{Fou}}(w)e^{\mathrm{i}wx}\,\mathrm{d}w \tag{II.4.11}$$

Sometimes, the function might only be square integrable, and we resort to the more general result below.

Fact II.4.12 (Fourier Inversion Theorem)**.** *Let* f *be in the* $L^2(\mathbb{R})$ *space. Then:*

$$\int_{-\infty}^{\infty} f^{\mathsf{Fou}}(w)e^{\mathrm{i}wx}\,\mathrm{d}w = \begin{cases} f(x) & f \in C^1 \\ \frac{1}{2}\lim_{y\uparrow x} f(y) + \frac{1}{2}\lim_{y\downarrow x} f(y) & \textit{else} \end{cases} \tag{II.4.13}$$

An immediate observation is that the Laplace and Fourier Transform generalize the MGF and CF in the following sense:

- the CF of a distribution $p(x)$ is the complex conjugate of the Fourier Transform of $p(x)$
- the MGF is the bilateral Laplace transform of $p(x)$ with a $-t$ in the argument.

Additionally, Eqn. I.3.14 can be now reinterpreted with the statement that the canonical partition function is the Laplace transform of the microcanonical entropy. We gloss over the discussion on the conditions of existance, but stress on one very important fact: the Laplace transform of L^2 functions is analytic. This is a consequence of its continuity and it being zero on closed countours, which allows to apply Morera's Theorem (Prop. II.2.51).

II.5 Dirac Delta Distribution, Hubbard-Stratonovich Transformation

Further References

To begin, it is possible to consult (Krzakala and Zdeborová 2021; Salasnich n.d.), (Evans 2006, Lect. 6).

We now turn to present two objects that are widely used in Physics.

Definition II.5.1 (Dirac Delta Distribution)**.** *The dirac delta distribution is such that:*

$$\boldsymbol{\delta}(x) \simeq \begin{cases} +\infty & x = 0 \\ 0 & x \neq 0 \end{cases} \quad \textit{s.t.} \quad \int_{-\infty}^{\infty} \boldsymbol{\delta}(x)\,\mathrm{d}x = 1. \tag{II.5.2}$$

One can see that is is not really a function (i.e. the first equation would make the integral zero), but rather a generalization in the sense of limits of valid functions where:

$$\int_{-\infty}^{\infty} \boldsymbol{\delta}(x)\,\mathrm{d}x = \lim_{\epsilon\to 0^+} \int_{-\infty}^{\infty} \boldsymbol{\delta}_\epsilon(x)\,\mathrm{d}x, \quad \lim_{\epsilon\to 0^+} \boldsymbol{\delta}_\epsilon(x) = \begin{cases} \infty & x = 0 \\ 0 & x \neq 0. \end{cases} \tag{II.5.3}$$

These comments are largely indicatory and serve for an understanding. For rigorous definitions, one can resort to Measure Theory or the Theory of Distributions, allowing for some less applicable constructions but more rigor (see (Strichartz 2003)). For what will serve here, we can take it a a heuristic. Also $\boldsymbol{\delta}$ *extends in the most natural way to its multidimensional equivalent.*

Example II.5.4. *An infinite number of functions satisfies Definition II.5.1. One interesting case is the Gaussian measure with vanishing variance*

$$\boldsymbol{\delta}_\epsilon(x) = \frac{1}{\epsilon\sqrt{\pi}} e^{-\frac{x^2}{\epsilon^2}}, \tag{II.5.5}$$

or in the more friendly form $\frac{1}{\sqrt{2\pi\epsilon}} e^{-\frac{1}{2\epsilon}x^2}$. *Alternatively, the simple constant vanishing function:*

$$\boldsymbol{\delta}_\epsilon(x) = \begin{cases} \frac{1}{\epsilon} & |x| \leqslant \frac{\epsilon}{2} \\ 0 & else \end{cases} \tag{II.5.6}$$

is valid.

Lemma II.5.7 (Dirac Delta Property)**.** *for* $f : \mathbb{R} \to \mathbb{R}$ *continuous at* x*:*

$$\int f(m)\boldsymbol{\delta}(m-x)\,\mathrm{d}m = f(x) = \int f(m)\boldsymbol{\delta}(nm-x)\,\mathrm{d}m = \frac{1}{n}f(x). \tag{II.5.8}$$

Proof. We provide a sketch. Depending on the level of formality it can be made more rigorous.
The Dirac delta is concentrated at 0 in its Definition. For the case $\boldsymbol{\delta}(m-x)$ we use the second example above for simplicity.

$$\lim_{\epsilon\to 0^+} \int_{-\infty} \boldsymbol{\delta}_\epsilon(m-x)f(m)\,\mathrm{d}m = \lim_{\epsilon\to 0^+} \frac{1}{\epsilon}\int_{-\frac{\epsilon}{2}+x}^{\frac{\epsilon}{2}+x} f(m)\,\mathrm{d}m = f(x), \tag{II.5.9}$$

by continuity of f. The second equality follows analogously. We recognize that n is fixed and thus $\mathrm{d}nm = n\mathrm{d}m$ and that $\delta(nm-x) = \delta(m-\frac{x}{n})$ as they both concentrate at $x = nm$. Clearly:

$$\int f(m)\boldsymbol{\delta}(nm-x)\,\mathrm{d}m = \frac{1}{n}\int f(m)\boldsymbol{\delta}(nm-x)\,\mathrm{d}nm = \frac{1}{n}f(x). \tag{II.5.10}$$

□

Remark II.5.11. *The second result of Lemma II.5.7 is the most used version by Physicists. Notice that if we take the logarithm we find that:*

$$\frac{\log\frac{f(x)}{n}}{n} = \frac{\log f(x)}{n} - \frac{\log n}{n} \overset{n\to\infty}{\to} \frac{\log f(x)}{n}. \tag{II.5.12}$$

Therefore, the n *at the denominator can be neglected.*

We also derive the relation between Fourier Transform (Def. II.4.4) and Delta distribution by direct computation. If the distribution is centered at x_0, then

$$\mathscr{F}[\boldsymbol{\delta}(x-x_0)] = \frac{1}{2\pi}\int_{-\infty}^{\infty} \boldsymbol{\delta}(x-x_0)e^{\mathrm{i}wx}dx = \frac{1}{2\pi}e^{\mathrm{i}wx_0}, \tag{II.5.13}$$

which by Fourier Inversion (Fct. II.4.12) leads us to the integral representation for the Dirac Delta distribution centered at x_0:

$$\boldsymbol{\delta}(x-x_0) = \mathscr{F}^{-1}[f(x)] = \int_{-\infty}^{\infty} f^{\mathsf{Fou}}(w)e^{-\mathrm{i}wx}\,\mathrm{d}w \implies \boldsymbol{\delta}(x-x_0) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{\mathrm{i}w(x-x_0)}\,\mathrm{d}w. \tag{II.5.14}$$

A better interpretation in the Statistical Physics sense is provided by deriving the identity with the Hubbard-Stratonovich transformation (Stratonovich 1957), which is based on the exact identity:[3]

$$e^{-\frac{a}{2}x^2} = \sqrt{\frac{1}{2\pi a}}\int_{-\infty}^{\infty} e^{-\frac{y^2}{2a}-\mathrm{i}xy}\,\mathrm{d}y, \quad a > 0. \tag{II.5.15}$$

[3] to prove it, one just completes the squares in the exponential and finds a Gaussian integral

In general it is used to change from a particle theory to its respective field theory. In other words, it is applied to decouple the sites x_i, x_j into a system of particles x interacting with a field y.[4] The cardinal example for Statistical Physics is found in the steps of the replica method, where the powers of $\mathcal{Z}^r$ *entangle* configurations across replicas, and such a transformation will represent them as dis-entangled.

To reconnect with Fourier, for a Dirac delta distribution the Hubbard-Stratonovich transformation recovers the integral representation via the Gaussian approximation:

$$\begin{aligned}\boldsymbol{\delta}(x-x_0) &= \int_{-\infty}^{\infty} \boldsymbol{\delta}(x-x_0)\,\mathrm{d}x = \lim_{\epsilon\to 0^+} \frac{1}{\sqrt{2\pi\epsilon}} \int_{-\infty}^{\infty} e^{-\frac{1}{2\epsilon}(x-x_0)^2}\,\mathrm{d}x && \text{(II.5.16)}\\ &= \frac{1}{2\pi} \lim_{\epsilon\to 0^+} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left\{-w^2\frac{\epsilon}{2} + \mathrm{i}w(x-x_0)\right\}\mathrm{d}w\,\mathrm{d}x && \text{(II.5.17)}\\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{\mathrm{i}w(x-x_0)}\,\mathrm{d}w, && \text{(II.5.18)}\end{aligned}$$

where in the last equality we used dominated convergence.

Remark II.5.19. *We are effectively relating the Hubbard-Stratonovich transformation and the Fourier transform for the delta function.*

Remark II.5.20. *Neither the delta distribution nor its Fourier transform are square integrable!*

A suspicious Dirac; intuition on its meaning

In Physics, the dirac delta is often taken as a way to enforce a hard-constraint on an integral. For example, in some cases we might want to restrict the integration of a function $f : \mathbb{R}^d \to \mathbb{R}$ to vectors of some norm, say $\sqrt{d}$. There are at least two options to express this. On one side, a spherical integral $\int_{\mathbb{S}^{d-1}(\sqrt{d})}$ is natural. On the other, we might as well perform the integral over $\mathbb{R}^d$, and enforce the selection of $\boldsymbol{x}$ such that $\|\boldsymbol{x}\|_2^2 = d$, which would be written as:

$$\int_{\mathbb{R}^d} f(\boldsymbol{x})\boldsymbol{\delta}(\|\boldsymbol{x}\|_2^2 - d^2)\mathrm{d}\boldsymbol{x}. \tag{II.5.21}$$

For the latter, Physicists develop tools to approximate the integral especially when the dimensions are large.

On a related note, some papers present a peculiar expression involving the Dirac delta. Without much context or motivation, we present below an example computation. Let us keep the matters simple, and imagine that we are working with real-valued functions on the reals, since the reasoning extends naturally to $\mathbb{R}^d$ integration. Fix a resolution Δ. A collection of n discrete variables $\{x_1, \ldots, x_n\}$, where for each $i \in [n]$ we have $x_i = i\Delta$, approximates in the Riemann sense a continuous line $x \in [0, L]$, with $L = n\Delta$. As usual, we might write the integral as the limit of the Riemann sum, i.e.:[5]

$$\lim_{n\to\infty} \sum_{i=1}^{n} f(x_i) = \int_0^L \frac{f(x)}{\Delta}\mathrm{d}x. \tag{II.5.22}$$

Similarly, in the same limit, we can "identify" for two indexes $i, j \in [n]$ the limit of the Kronecker delta **normalized by the discretization** as Dirac's Delta. We explain this as follows. While Kroenecker's Delta $\boldsymbol{\delta}_{ij}$ is a-dimensional in $\{0, 1\}$, Dirac's delta possesses the dimension of the inputs over which it is evaluated. Intuitively, if x_i, x_j are in units of length, the dirac $\boldsymbol{\delta}(x_i - x_j)$ will be in units of length. A little thought then reveals that the correct assignment is:

$$\lim_{n\to\infty} \frac{1}{\Delta}\boldsymbol{\delta}_{ij} = \lim_{n\to\infty} \frac{n}{L}\boldsymbol{\delta}_{ij} = \boldsymbol{\delta}(x_i - x_j), \tag{II.5.23}$$

[4] Understanding formally what a field is can be avoided. It suffices to notice that the entangling on the LHS is lost on the RHS, and the y is integrated out. The RHS integral will turn out to be easier.

[5] notice we are being sloppy here and leaving Δ on the RHS despite taking $n \to \infty$

and in particular, choosing $i = j$, we obtain:

$$\lim_{n\to\infty} \frac{1}{\Delta} = \boldsymbol{\delta}(0). \tag{II.5.24}$$

Therefore, we can say that:

$$f(x_j) = \lim_{n\to\infty} \sum_{i=1}^{n} \boldsymbol{\delta}_{i,j} f(x_i) = \int_0^L \boldsymbol{\delta}(x - x_j) f(x) \mathrm{d}x, \tag{II.5.25}$$

and regarding Equation II.5.22, the most formal result would be:

$$\int_0^L \boldsymbol{\delta}(0) f(x) \mathrm{d}x. \tag{II.5.26}$$

Remark II.5.27. *On a mathematical basis, we are just performing Riemann integration. On the Physics side, the $\boldsymbol{\delta}(0)$ term serves the purpose of ensuring the same dimensional meaning on both sides of the equation.*

Remark II.5.28. *In the inverted direction, when some Dirac distributions appear in the form $\boldsymbol{\delta}(x_{-} - x_j)$, they can be non-rigorously transformed into $\frac{1}{\Delta}\boldsymbol{\delta}_{i,j}$ to simplify the equations into sums, and then brought back to their original limit taking $\Delta \to 0$. Obviously, this requires reverting an integral and applying the limit $\Delta \to 0$ after some operations, and must be taken with care in the formal sense.*

II.6 Asymptotic Integrals

Remark II.6.1. *Notice that throughout the section and the document we do not perform analysis at the boundaries of integration and think of the integrals as evaluations in the interior of the extremes. The discussion could be generalized, but at the cost of increased space. For related ideas, see (Akbari, Bury, and Phillips 2015; Wong 2001)*

Further References

The topic is very wide but the following are three useful resources (Akbari, Bury, and Phillips 2015; Miller 2006; Wong 2001) and (Evans 2006).

We will just provide the basics of the two methods. The latter is especially quite advanced and involves some knowledge in complex analysis. For the sake of understanding when it is used in the replica computations, we will briefly present it, and always assume that it holds in its *simplest form*. A satisfactory discussion is carried out in (Miller 2006; Wong 2001).

Assume we wish to perform the integral of a function that has exponentially decaying terms in some parameter. Explicitly, in our case will be n. We present a result that is at the basis of two widely used tools.

Lemma II.6.2 (Watson's Lemma (Watson 1918))**.** *A more recent reference is (Miller 2006)[Chap. 2, Sec. 2.2].*
Assume the following conditions hold:

i) $0 < R \leqslant \infty$ *fixed*

ii) $f(x) = x^{\lambda} g(x)$

iii) $g(x)$ *is* C^{∞} *at* $x = 0$ *and* $g(0) \neq 0$

iv) $\lambda > -1$

v) either $|f(x)| \leqslant K e^{bx}$ *for all* $x > 0$ *and* $K, b \perp x$ *or* $\int_0^R |f(x)| \, \mathrm{d}x < \infty$

Then:

1. *an integrability condition holds for the exponentiated function, i.e.:*

$$\left|\int_0^R e^{-nx} f(x)\,\mathrm{d}x\right| < \infty, \tag{II.6.3}$$

2. *and most importantly, the integral of the exponentiated function admits an asymptotic expression in terms of a series expansion of its content:*

$$\int_0^R e^{-nx} f(x)\,\mathrm{d}x \sim \sum_{t=0}^{\infty} \frac{g^{(t)}(0)\Gamma(\lambda+t+1)}{t!n^{\lambda+t+1}} \qquad (n > 0, n \to \infty). \tag{II.6.4}$$

Proof. See (Wong 2001, Chap. I, Sec. 5), (Miller 2006, Chap. 2, 2.2). □

Remark II.6.5. *Notice that we are giving an asymptotic approximation of the Laplace integral!*
The statement could be expanded to Laplace types of integrals, where the x in the exponent is replaced by a function of x.

Consider the following type of integral:

$$\mathcal{I}_n = \int_a^b \phi(x)(f(x))^n\,\mathrm{d}x. \tag{II.6.6}$$

Especially with transforms such as the Fourier or Laplace, some steps require these types of computations. We provide an example of (Wong 2001) below as a motivation, but others are found in Thermodynamics and Statistical Physics.

Example II.6.7. *Let $\{X_i\}_{i=1}^n$ be iid with pdf $f(x)$. Their sum has density $f_n := \circledast_{i=1}^n f$, where $\circledast$ denotes the convolution operator.*[6] *By independence, characteristic functions (Def. II.3.3) $\phi(\zeta)$ multiply with each other and the characteristic function of the sum is $\phi_n(\zeta) = (\phi(\zeta))^n$. Applying the inverse Fourier Transform theorem*[7]*, one finds a representation of the sum pdf as:*

$$f_n(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-\mathrm{i}\zeta x}(\phi(\zeta))^n\,\mathrm{d}\zeta, \tag{II.6.8}$$

which is of the form of Eqn. II.6.6.

Heuristically, for sharply peaked functions inside an integral, with peaking-asymptotics-exogenous parameter n, we should expect that the integral value concentrates around the dominating value as $n \to \infty$. Laplace's result, which is of this type, roughly says that for a function f which over the domain of integration is unimodal concave with maximum at $x_\star \in (a,b)$ one has that:

$$\mathcal{I}_n(x) \overset{n\to\infty}{\sim} \phi(x_\star)(f(x_\star))^{n+\frac{1}{2}}\sqrt{-\frac{2\pi}{nf''(x_\star)}}. \tag{II.6.9}$$

Equivalently, expressing $f(x) = e^{h(x)}$ and assuming that h is unimodal at $x_\star$ and concave, the result is:

$$\mathcal{I}_n(x) = \int_a^b \phi(x)e^{nh(x)}\,\mathrm{d}x \overset{n\to\infty}{\sim} \phi(x_\star)e^{nh(x_\star)}\sqrt{-\frac{2\pi}{nh''(x_\star)}}. \tag{II.6.10}$$

Notice that the argument of the square root is positive since $h''(x_\star) < 0$ by concavity.

[6] As a reminder, for two measures μ, ν on the same sample space $\mathscr{X} \subseteq \mathbb{R}^d$ absolutely continuous wrt a reference measure ρ, the convolution operator is defined as:

$$(\mu \circledast \nu)(z) := \int_{\mathscr{X}} f(x)g(z-x)\mathrm{d}\rho(x),$$

where $z \in \mathscr{X}$ and f, g are the respective densities. Equivalently, f, g *convolve* in the reference measure $\mathrm{d}\rho$.

[7] i.e. the function of interest can be recovered from its Fourier transform, the characteristic function

We will report an informal justification and a rigorous Theorem. The former is achieved by a Taylor expansion around the max $x_\star$ of ϕ and $h(x)$, extending the integration to $\mathbb{R}$, and using properties of Gaussian integrals (Sec. II.1). Despite being unjustified, it gives the same result.

$$\int_a^b \phi(x)e^{nh(x)}\,\mathrm{d}x \approx \int_a^b \phi(x_\star)e^{nh(x_\star)+(x-x_\star)^2\frac{h''(x_\star)}{2}}\,\mathrm{d}x \qquad \text{Taylor} \qquad \text{(II.6.11)}$$

$$\approx \phi(x_\star)e^{nh(x_\star)}\int_{-\infty}^{\infty} e^{n(x-x_\star)^2\frac{h''(x_\star)}{2}}\,\mathrm{d}x \quad \text{domain of integration} \qquad \text{(II.6.12)}$$

$$= \phi(x_\star)e^{nh(x_\star)}\sqrt{\frac{-2\pi}{nh''(x_\star)}} \qquad \text{Fct. II.1.1.} \qquad \text{(II.6.13)}$$

Another very nice set of considerations and exmples on the heuristic method can be found in (Shalizi 2024). Below, we report a formal Proposition.

Proposition II.6.14 (Laplace's Method)**.** *We provide a statement sufficient for our uses, a more general one is found in (Wong 2001, Chap II, Thm. 1.1), or (Miller 2006, Chap. 4). Also (Krzakala and Zdeborová 2021, Chap. 1, Thm. 2) provides a brief presentation.*
Let $h(x) - \frac{1}{n}\log\phi(x)$ be twice differentiable on $[a, b]$ and unimodal, with a maximum at $x_\star \in (a, b)$. Then:

$$\lim_{n\to\infty} \frac{\int_a^b \phi(x)e^{nh(x)}\,\mathrm{d}x}{\phi(x_\star)e^{nh(x_\star)}\sqrt{\frac{-2\pi}{nf''(x_\star)}}} = 1, \qquad \text{(II.6.15)}$$

which is the same as the heuristic guess we computed just above.

Proof. The general statement proof is based on a rearrangement that gets to the Laplace transform. Then Watson's Lem. II.6.2 is applied. So a precise series of sums is found, and the approximation can be made as tight as desired by adding terms. □

II.6.1 Steepest descent

We do not dwell much into the theory because it is extensive, but for complex integrals one can consider as references (Rudin 1987; Weber and Arfken 2004)(Arfken and Weber 2013, Chap. 6, 7). In particular, we build on a combination of arguments from (Wong 2001, Chap. III, Sec. 4), (Akbari, Bury, and Phillips 2015), (Evans 2006, Lect. 5), and (Weber and Arfken 2004, Sec. 6, 7) which present many aspects of a method first devised by Debye (Debye 1909).
Consider integrals of the following form:

$$\mathcal{I}_n = \int_{\mathcal{C}} \phi(z)e^{nh(z)}\,\mathrm{d}z \quad n \gg 1, \quad z \in \mathbb{C}, \qquad \text{(II.6.16)}$$

where $\mathcal{C}$ is a contour as the one in Def. II.2.12, and ϕ, h are analytic functions (Def. II.2.1). We can evaluate it via a combination of the Laplace method and the method of stationary phase.
Letting $f = u + \mathrm{i}v$, we know that u is maximized in the Laplace method and v is stationary, so that oscillatory contributions cancel far away from it and not close to it (Evans 2006, Lec. 5), (Wong 2001, Chap. II, Sec. 3)(Cohn 2007; Rozman 2017). We will therefore argue that the integral above is dominated by the argument evaluated at the stationary point of f.

Remark II.6.17. *While this is just a heuristic, the references explain in detail the various levels of understanding of such idea.*

As a first step, we notice that the extremas of $f, \Re(f)$ are *saddle points.*

Fact II.6.18. *The conditions of maximization of u and stationarity of v imply that the extremas of $f, \Re(f)$ are saddle points.*

Proof. Cauchy-Riemann conditions imply that u, v satisfy Laplace's Equation (Def. II.2.6). The statement is easily seen by:

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial}{\partial y}\left(-\frac{\partial v}{\partial x}\right) = \frac{\partial^2 u}{\partial x^2} - \frac{\partial}{\partial x}\frac{\partial v}{\partial y} = \frac{\partial^2 u}{\partial x^2} - \frac{\partial}{\partial x}\frac{\partial u}{\partial x} = 0 \qquad \text{(II.6.19)}$$

Where we have just exchanged derivatives and applied the conditions multiple times. The result extends to higher order $k \geqslant 2$ derivatives. □

Thanks to the above result, we know that if $\partial_x^2 u > 0$ then $\partial_y^2 u < 0$ at a point of interest. The opposite result also holds (negative, positive). If we denote the point as $z_\star$, we may apply Cauchy's Theorem to deform the countour and let it pass across $z_\star$. The fact that the imaginary part is stationary gets us to evaluating the integral over a real countour that passes through $z_\star$ and has constant imaginary portion:

$$\mathcal{I}_n = \int_{C'} \phi(z) e^{nh(z)}\,\mathrm{d}z = e^{\mathrm{i}nv} \int_{C'} \phi(z) e^{nu(z)}\,\mathrm{d}z \quad n \gg 1, \quad z \in \mathbb{C}, \qquad \text{(II.6.20)}$$

where v is independent of z.
We now wish to find the path of steepest ascent of the real part to the stationary point $z_\star$. Some thought leads to the conclusion that it is exactly realized when the imaginary part variation is null. Indeed:

$$\delta h = h(z) - h(z_\star) = \delta u + \mathrm{i}\delta v \implies |\delta u| \leqslant |\delta h|, \qquad \text{(II.6.21)}$$

and δu is at its steepest when $\delta v = 0$, i.e. the imaginary part is kept constant and $z_\star = (x_\star, y_\star)$ attains a constant value when plugged into v. We further say h has a saddle point of the k^{th} order at $z_\star$ when:

$$\left.\frac{\mathrm{d}^m h}{\mathrm{d}z^m}\right|_{z=z_\star} = 0 \quad \forall\, m \in [k] \quad \left.\frac{\mathrm{d}^m h}{\mathrm{d}z^m}\right|_{z=z_\star} \neq 0 \quad \forall\, m > k. \qquad \text{(II.6.22)}$$

Therefore, at a k order saddle, the Taylor expansion of h reads:

$$h(z) \approx h(z_\star) + \left.\frac{\mathrm{d}^{k+1} h}{\mathrm{d}z^{k+1}}\right|_{z=z_\star} \frac{(z - z_\star)^{k+1}}{(k+1)!} + h.o.t, \qquad \text{(II.6.23)}$$

over which we perform the following change of variables

$$z - z_\star = re^{\mathrm{i}\vartheta} \quad h^{(k+1)}(z_\star) = \rho e^{\mathrm{i}\xi} \quad \vartheta, \xi \in [0, 2\pi], \rho, r \in \mathbb{R}. \qquad \text{(II.6.24)}$$

Focusing on the Taylor expansion, the expression becomes:

$$h(z) - h(z_\star) \approx \frac{r^{k+1} e^{\mathrm{i}(k+1)\vartheta}}{(k+1)!} \rho e^{\mathrm{i}(k+1)\xi} = [\cos(\xi + (k+1)\vartheta) + \mathrm{i}\sin(\xi + (k+1)\vartheta)] \frac{r^{k+1}\rho}{(k+1)!}. \qquad \text{(II.6.25)}$$

Thus, imposing stationarity of the complex part:

$$\Im(h(z) - h(z_\star)) = i\sin(\xi + (k+1)\vartheta) = 0 \iff \xi + (k+1)\vartheta = q\pi, q \in \mathbb{N} \iff \vartheta = -\frac{\xi}{k+1} + q\frac{\pi}{k+1}, q \in \mathbb{N}, \qquad \text{(II.6.26)}$$

and from such condition, we can directly derive the steepest ascent direction:

$$\Re(h(z) - h(z_\star)) < 0 \iff \cos(\xi + (k+1)\vartheta) < 0 \iff \vartheta = -\frac{\xi}{k+1} + (2q+1)\frac{\pi}{k+1}, \qquad \text{(II.6.27)}$$

and the steepest descent direction:

$$\Re(h(z) - h(z_\star)) > 0 \iff \cos(\xi + (k+1)\vartheta) > 0 \iff \vartheta = -\frac{\xi}{k+1} + 2q\frac{\pi}{k+1}. \qquad \text{(II.6.28)}$$

Using such technique, we consider the simplest case in which ϕ is of constant order around $z_\star$ and the saddle is of order $k = 1$, so that we can apply Fct. II.6.18 and know that the maximum of the real part at stationary imaginary part is a saddle point. Taylor expanding the functions inside the integral we set:

$$\phi(z) \approx \phi(z_\star) \quad h(z) \approx h(z_\star) + \frac{1}{2}\frac{\mathrm{d}^2 h}{\mathrm{d}z^2}\bigg|_{z=z_\star}(x - z_\star)^2, \tag{II.6.29}$$

and write the heuristic:[8]

$$\mathcal{I}_n \approx \phi(z_\star)e^{nh(z_\star)} \int_a^b e^{\frac{1}{2}h''(z_\star)(z-z_\star)^2}\,\mathrm{d}z. \tag{II.6.30}$$

To evaluate the integral we perform a change of variables into polar coordinates with:

$$z - z_\star = re^{\mathrm{i}\vartheta} \quad h''(z_\star) = |h''(z_\star)|e^{\mathrm{i}\xi} \quad \vartheta, \xi \in [0, 2\pi], r \in \mathbb{R}_+ \tag{II.6.31}$$

where ϑ is the free to choose angle at which the countour passes through $z_\star$. The integral in polar coordinates is now:

$$\mathcal{I}_n \approx \phi(z_\star)e^{nh(z_\star)} \int_0^\infty e^{\frac{1}{2}|h''(z_\star)|e^{\mathrm{i}\xi}r^2e^{2\mathrm{i}\vartheta}}e^{i\vartheta}\,\mathrm{d}r, \tag{II.6.32}$$

where the integration has moved to arbitrary $r \in \mathbb{R}_+$ without breaking the approximation. This can be verified in terms of the result of Watson's Lem. II.6.2, which roughly says that the integral will be dominated by the terms close to $r = 0$.
A clever choice, which turns out to be the one of steepest descent from the saddle is $\xi + 2\vartheta = \pi$, for which $\vartheta = \frac{\pi-\xi}{2}$. This can be verified with the above reasoning or in (Arfken and Weber 2013, Chap. 7.3). The integral now becomes:

$$\mathcal{I}_n \approx \phi(z_\star)e^{nh(z_\star)}e^{\mathrm{i}\vartheta}\int_0^\infty \exp\left\{\frac{1}{2}n|h''(z_\star)|r^2\right\}\mathrm{d}r \overset{n\to\infty}{\sim} \phi(z_\star)e^{nh(z_\star)}e^{\mathrm{i}\vartheta}\sqrt{\frac{2\pi}{n|h''(z_\star)|}} \quad \vartheta = \frac{\pi-\xi}{2}. \tag{II.6.33}$$

The above approximation result can be made rigorous. A more systematic approach is carried out in (Wong 2001, Chap. II, Sec. 4).

Remark II.6.34 (General remarks)**.** *We give four further heuristic comments.*

1. *If there are multiple saddle points, the contributions sum up;*
2. *the direction in which $z_\star$ is approached is important;*
3. *in absence of a saddle point (e.g. if the function h is monotonic), the biggest contribution is at the boundary, as stated at the beginning;*
4. *in our applications, we will blindly believe that the saddle point method may be applied in the simplest case possible, by which we replace an integral in the form of Eqn. II.6.16 with its maximum/minimum argument of the exponent depending on the sign that the exponential has. Checking the reliability of the method would be tedious, and we rather hope for n to be large enough as to branch out all the unfriendly cases.*

Having introduced some fundamental mathematical notions, we now turn to explaining the connections between Information Theory and Statistical Mechanics, through the lens of the work of Jaynes (Jaynes 1957) and Shannon (Shannon 1948). The loose aim is to give a justification as to why the physical laws just discussed are, in the words of Jaynes "merely an example of statistical inference" (Jaynes 1957). Reconnecting with the discussions in Chapter I, we will still be concerned with describing thermodynamic-like properties of a system at equilibrium, but without seeing it. Subject to some truth-revealing obsevations, we will understand that statistical inference and information are treated under a surrogate of the concepts derived earlier, but this time starting solely from a notion of entropy. The discussion about physical laws and equations of motions mentioned will be latent in the analysis, but still present. As an informal result, Statistics, Information Theory and Thermodynamics have a large overlap.

[8] note that we are not taking out the imaginary component $\Im(h) = v$ explicitly, but inside the general form

Chapter III

Mathematical Topics in Statistical Physics

In this Chapter we formalize and extend some of the statements found in the previous discussions. In Setion III.1, we report the original axiomatic construction of Entropy by Shannon (Shannon 1948). To continue and reconnect with Thermodynamics, we overview the Maximum Entropy Principle of Jaynes and the comments therein about the comparison with Boltzmann's formulation (Jaynes 1957). In Section III.3, we move the narrative to a more grounded justification and contextualization of Free Energy, with an eye on Machine Learning and Bayesian methods. In a related way, we summarize properties and ideas behind the Legendre-Fenchel transform, an extension of the Legendre transform, in Section III.4. To conclude, we just briefly overview two very important fields: large deviations and all of Information Theory. For the former, we do relegate to extensive references the ideas. For the latter, we do still bring some value to the document, by providing a broad validation of Kullback-Leibler divergences and entropy, thanks to their ubiquity in Inference.

Further References

On a similar note, two recent articles by the same author provide a three-way concise jusification of entropy (Lairez 2022), and a quick derivation of fundamental Thermodynamical objects from first principles (Lairez 2023), which is useful in the context of Chapter I.

III.1 Formalizing Uncertainty

We will first take an Information Theoretic approach, which was initiated by Shannon in a foundational publication (Shannon 1948). The Thermodynamics construction can be shown to be equivalent on many terms.

Remark III.1.1 (Units of measure)**.** *The Physics perspective on entropy is at first sight impossible to reconcile with Shannon's notion due to k_B, Boltzmann's factor . However, a careful inspection shows that it is merely a by product of the reference system used, which endows physical quantities up to constants to estimate. A setting in which $k_B = 1$ can be used, without loss of generality.*

We take the discrete case for reference. Therefore, a random variable X takes probabilities p_i over a discrete space $\mathscr{X} = \{x_1, \ldots, x_n\}$, where $|\mathscr{X}| = n$. The aim is finding a nice measure of uncertainty $\mathcal{H}$, or of "how much choice is involved in the selection of an event" (Shannon 1948, Sec. 6) for a probability distribution, denoted wlog as $\{p_i\}_{i=1}^n$. Tautologically, the *nice* notion depends on some properties stated *ab-initio*. We present those used originally below.

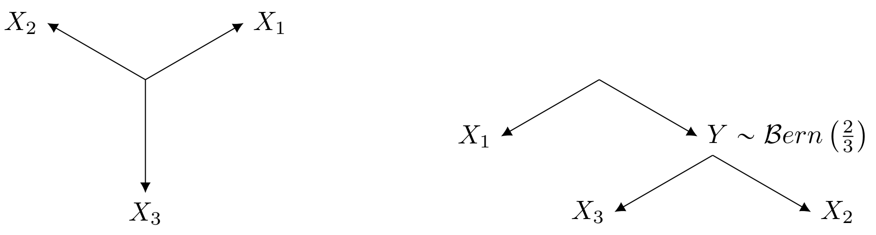

Figure III.1: Two "equally entropic" processes. Source (Shannon 1948, Fig.6) Two sampling systems that are statistically equivalent should exhibit the same entropy.

Shannon's Uncertainty Properties

A good measure of uncertainty satisfies the following properties

(Sh1) **tractability:** continuity in $\{p_i\}_{i=1}^n$ holds

(Sh2) **monotonicity:** if $X \sim \mathcal{U}nif(\mathscr{X})$ then it is monotonically increasing[a] in $|\mathscr{X}|$

(Sh3) **additivity:** for an event decomposed into sub-events, the entropy of the original event is a weighted sum of the entropy of the sub-events

[a] to an increasing number of equally likely outcomes corresponds increased uncertainty

To understand (Sh3), the example of (Shannon 1948, Fig. 6) is instrumental. We report it below.

Example III.1.2. *Given two outcome systems:*

$$X = \begin{cases} X_1 & wp\ \frac{1}{2} \\ X_2 & wp\ \frac{1}{3} \\ X_3 & wp\ \frac{1}{6} \end{cases} \quad X = \begin{cases} X_1 & wp\ \frac{1}{2} \\ X_2 & if\ Y = 1 \\ X_3 & if\ Y = 0 \end{cases} \quad Y \sim \mathcal{B}ern\left(\frac{2}{3}\right), \tag{III.1.3}$$

depicted in Figure III.1.

It is easy to see that an information measure for the former would obey the expression $\mathcal{H}(\frac{1}{2}, \frac{1}{3}, \frac{1}{6})$, while the coarse grained version amounts to having entropy $\mathcal{H}(\frac{1}{2}, \frac{1}{2})$ on the first step (either X_1 or one of X_2, X_3) and $\mathcal{H}(\frac{2}{3}, \frac{1}{3})$ on the second step (when the Y coin is thrown). Given that the probability distribution of X is phenomenologically the same, property (Sh3) enforces that:

$$\mathcal{H}\left(\frac{1}{2}, \frac{1}{3}, \frac{1}{6}\right) = \mathcal{H}\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{1}{2}\mathcal{H}\left(\frac{2}{3}, \frac{1}{3}\right), \tag{III.1.4}$$

where the $\frac{1}{2}$ factor is a weighting for the sub-event since it happens only half of the time.

At first sight one might think that there are plenty $\mathcal{H}$ obeying the uncertainty principles. It turns out that the three are sufficient and necessary for a characterization.

Theorem III.1.5 (Shannon's Uncertainty Measure Characterization, Thm. 2 (Shannon 1948))**.** *There exists one and only one $\mathcal{H}$ satisfying (Sh1)-(Sh3), and it is:*

$$\mathcal{H}(\{p_i\}_{i=1}^n) = -K \sum_{i=1}^n p_i \log_2 p_i \quad K > 0. \tag{III.1.6}$$

Proof. We report the proof for the sake of completeness since it is self-contained and nice to read.
Let $A(n) := \mathcal{H}\left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$. Observe that by (Sh3) we have that $A(t^n) = nA(t)$, since t^n equally likely events decompose into n choices of t equally likely events.

For any pair (n, t) where t is fixed and n is arbitrarily large, there is another pair (s, m) such that $s^m \leqslant t^n \leqslant s^{m+1}$. Taking logarithms and dividing by $n \log s$ one finds:

$$\frac{m}{n} \leqslant \frac{\log t}{\log s} \leqslant \frac{m}{n} + \frac{1}{n}, \tag{III.1.7}$$

or the nicer expression:

$$\left|\frac{m}{n} - \frac{\log t}{\log s}\right| < \epsilon \quad \epsilon \text{ arbitrarily small since } n \text{ arbitrarily large.} \tag{III.1.8}$$

An application of monotonicity (Sh2) and the previous discussion on (Sh3) ensures the chain of inequalities holds after applying $A(\cdot)$

$$mA(s) \leqslant nA(t) \leqslant (m+1)A(s), \tag{III.1.9}$$

which divided by $nA(s)$ gives:

$$\left|\frac{m}{n} - \frac{A(t)}{A(s)}\right| < \epsilon \quad \epsilon \text{ arbitraily small.} \tag{III.1.10}$$

Combining Eqns. III.1.8, III.1.10 we reach the expression:

$$\left|\frac{A(t)}{A(s)} - \frac{\log t}{\log s}\right| < 2\epsilon \quad \epsilon > 0. \tag{III.1.11}$$

Since t, s were arbitrary, it must be that $A(t) = K \log(t), A(s) = K \log(s)$ where $K > 0$ is the constant that cancels out in the division. Note that $K > 0$ is ensured by (Sh2) and the expression is enforced since the bound is for arbitrary $\epsilon > 0$.
In general, for a clusterization of events i with relative frequency n_i we have that an uncertainty/information measure should satisfy for $p_i = \frac{n_i}{\sum n_i}$ the following identity:

$$K \log \sum n_i = \mathcal{H}\left(\{p_i\}_{i\in[n]}\right) + K \sum p_i \log n_i, \tag{III.1.12}$$

where the choice was broken down into n possibilities and n_i is the second (uniform) number of choices.
Eventually:

$$\mathcal{H}(\{p_i\}_{i=1}^n) = -K \sum_i p_i \ln p_i \tag{III.1.13}$$

by the simple observation that $\ln \sum n_i = \sum_j p_j \ln \sum n_i$. Requirement (Sh1) ensures that the definition of p_i is wlog, since by continuity in p_i we can always approximate any p_i' by rationals, and obtain Eqn. III.1.13. □

Definition III.1.14 (Shannon's uncertainty, or entropy)**.** *For a discrete random variable X its entropy is*

$$\mathcal{H}(X) := -\mathbb{E}\left[\ln(\mathbb{P}[X])\right], \tag{III.1.15}$$

where by abuse of notation we write $\mathcal{H}(X)$ without referring to the probabilities. Indeed, the entropy should be in terms of the distribution but we will mostly use it for random variables. Whenever the focus is on a general distribution with probabilities $\{p_i\}_{i\in[n]}$ we write it as $\mathcal{H}(\boldsymbol{p})$ or $\mathcal{H}(\{p_i\}_{i\in[n]})$.
Notice also that the logarithm is in base e. Despite being originally formulated in base 2, a trivial calculation shows that the two definitions are equivalent up to constants.[1]
The terminologies entropy, uncertainty and information measure are used exchangeably. The third and second are opposites (high uncertainty, low information), the first will be shown to be uncertainty in the next Section. By the previous discussion, we are also setting $K = 1$ wlog.

Note that up to a constant factor we have recovered the expression of Def. I.3.28, which has Def. I.2.15 as a special case. Despite being equal, it does not mean that they describe the same *concept* (Jaynes 1957). We will equivalence in a very peculiar way, by establishing a relation in terms of performing least-biased inference when approximating a distribution.

[1] It suffices to use the identity $\log_b a = \frac{\log_d a}{\log_d b}$ for $b = 2, a = \mathbb{P}[X], d = e$ to see that the constant is always $\ln 2$

III.2 The Maximum Entropy Principle

When doing inference, the objective is to minimize the bias of the procedure carried out. In a Physics environment, the concept can be briefly explained as follows. It is often the case that a measurement of an unknown function is available, but the randomness of all the realizations is not. If the random variable takes n finitely many values, the randomness is encapsulated in $\{p_i\}_{i=1}^n$, and the measurement is expressed as a weighted sum:

$$\mathbb{E}\left[f(X)\right] = \sum_{i=1}^{n} p_i f(x_i). \tag{III.2.1}$$

With this information, we would like to estimate the mean of another function $g : \mathscr{X} \to \mathbb{R}$, hoping to minimize the bias and be rightfully representing the truth. Unfortunately, the available information is only in Eqn. III.2.1 and in the fact that probabilities normalize $\sum_i p_i = 1$. The number of constraints is 2, the number of variables is n: we are facing an undetermined problem. In particular, an infinite number of distributions satisfies the constraints.

Historically, there have been many attempts to inject structure as to derive a unique result. The main idea is that one needs to supplement the observable reality with additional hypotheses that make sense. Since inference is on the probabilities of a random variable, the path splits at two main branches, the frequentist-objectivist (Cramér 1999; Feller 2009) and Bayesian-subjectivist (Jeffreys 1998; Keynes 2004) interpretations of probability. Given the unsatisfiable nature of our question, we are forced to implement the latter.

Summarizing, our task is to find a collection $\{\widehat{p}_i\}_{i=1}^n$ that satisfies the constraints and carries no more overflow of information coming from structural assumptions. Maximizing the entropy, to aim for the highest possible amount of uncertainty, is a principled way to solve the search problem. Remarkably, an object that we have already presented appears.

Proposition III.2.2 (Maximum Entropy Principle leads to Boltzmann Distribution (Jaynes 1957))**.** *Consider a random variable X, taking values in a finite discrete space $\mathscr{X} = \{x_1, \dots, x_n\}$, and a function $f : \mathscr{X} \to \mathbb{R}$ of which the expectation $\mathbb{E}\left[f(X)\right]$ is known. Let μ_f be the expectation of f, and $\{p_i\}_{i=1}^n$ represent any assignment of probabilities to the alphabet $\mathscr{X}$. Then, the program:*

$$\left\{ \underset{\{p_i\}_{i=1}^n}{\arg\max}\, \mathcal{H}(X) \quad s.t. \quad \sum_{i=1}^{n} p_i = 1,\ \mathbb{E}\left[f(X)\right] = \mu_f \right\} \tag{III.2.3}$$

is solved by a set of probabilities $\{\widehat{p}_i\}_{i=1}^n$ with canonical Boltzmann distribution (Def. I.3.10) parametrized by (ξ_0, ξ_1). The optimal value of entropy is:

$$\mathcal{H}^\star(X) = \xi_0 + \xi_1 \mu_f. \tag{III.2.4}$$

Proof. We introduce a Lagrange optimization problem for arbitrary distribution $\{p_i\}_{i=1}^n$. Following the usual formalism:

$$\mathcal{L}(p_i; \xi_0, \xi_1) = \mathcal{H}(X) + \xi_0 \left(\sum_i p_i - 1 \right) + \xi_1 \left(\mathbb{E}\left[f(X)\right] - \mu_f \right) \quad \xi_0, \xi_1 \in \mathbb{R}. \tag{III.2.5}$$

Solving for $\nabla \mathcal{L}(p_i; \xi_0, \xi_1) = 0$:

$$\frac{\partial \mathcal{L}}{\partial p_j} = -\sum_i \delta_{ij} \ln p_i - \sum_i p_i \frac{\partial \ln p_i}{\partial p_j} + \xi_0 \sum_i \delta_{ij} + \xi_1 \left(\sum_i \delta_{ij} f(x_i) \right) = 0, \tag{III.2.6}$$

where δ_{ij} are Dirac deltas, along with the two constraints mentioned above. The solution to this problem is:

$$-\ln p_j \underbrace{-1 + \xi_0}_{=\xi_0'} + \xi_1 f(x_j) = 0 \implies \widehat{p}_j = e^{-\xi_0' - \xi_1 f(x_j)} \quad \forall j. \tag{III.2.7}$$

To continue, we use ξ_0 in place of ξ_0' since it is a constant. Working on the derivatives wrt the multipliers:

$$e^{\xi_0} = \sum_i e^{-\xi_1 f(x_i)} = \mathcal{Z}(\xi_1) \iff \xi_0 = \ln \mathcal{Z}(\xi_1), \tag{III.2.8}$$

and

$$\mathbb{E}\left[f(X)\right] = e^{-\xi_0} \sum_i e^{-\xi_1 f(x_i)} f(x_i) = \langle f(X) \rangle_{\xi_1} = -\frac{\partial}{\partial \xi_1} \ln \mathcal{Z}(\xi_1). \tag{III.2.9}$$

The set $\{\widehat{p}_i\}_{i=1}^n$ is the specification of a Boltzmann distribution at temperature ξ_1 and mean energy $\mathscr{U} = \mathbb{E}\left[f(X)\right]$ for the canonical ensemble (Def. I.3.10). If we evaluate the entropy we find:

$$\begin{aligned} \mathcal{H}(X) &= -\sum_{i=1}^n p_i \log p_i = \sum_i e^{-\xi_0 - \xi_1 f(x_i)} \left(\xi_0 + \xi_1 f(x_i)\right) && \text{(III.2.10)} \\ &= \xi_0 \underbrace{\sum_i e^{-\xi_0 - \xi_1 f(x_i)}}_{=1} + \xi_1 \underbrace{\sum_i e^{-\xi_0 - \xi_1 f(x_i)} f(x_i)}_{=\mu_f = \mathbb{E}[f(X)]}. && \text{(III.2.11)} \end{aligned}$$

□

Corollary III.2.12. *The result is readily extended to a set of means* $(\mu_f^{(1)}, \ldots, \mu_f^{(k)})$ *of different functions* $(f^{(1)}, \ldots, f^{(k)})$*, for which one finds that the maximum entropy distribution is:*

$$p_i = \exp\left\{-\xi_0 - \sum_{l=1}^k \xi_l \mu_f^{(l)}(x_i)\right\} \qquad \mathcal{Z}(\xi_1, \ldots, \xi_k) = \sum_i \exp\left\{-\sum_{l=1}^k \xi_l \mu_f^{(l)}(x_i)\right\}; \tag{III.2.13}$$

$$\xi_0 = \ln(\mathcal{Z}(\xi_1, \ldots, \xi_k)) \qquad \mu_f^{(l)} = -\frac{\partial}{\partial \xi_l} \mathcal{Z}(\xi_1, \ldots, \xi_k), \tag{III.2.14}$$

with attained value:

$$\mathcal{H}^\star(X) = \xi_0 + \sum_{l=1}^k \xi_1 \mu_f^{(l)}. \tag{III.2.15}$$

Therefore, the distribution that maximizes uncertainty under the constraint of fixing the expectation of one or more functions is Boltzmann-like, with a well defined structure. However, it should be noted that the second Lagrange multiplier equation is not easy to solve. In the next computation, we show it is implicit.

Fact III.2.16. *One Lagrange multiplier is easy and unique if the other is available, since* $\xi_0 = \ln \mathcal{Z}(\xi_1)$*. Regarding* ξ_1 *it:*

$$\hbar(\xi_1) = \sum_{i=1}^n (f(x_i) - \mu_f) e^{-\xi_1 (f(x_i) - \mu_f)} = 0, \tag{III.2.17}$$

where $\hbar$ *is monotonically decreasing and continuous in* ξ_1 *and thus has only one solution. Unfortunately, the solution has no closed form for any* n*, and is difficult for* $n \gg 1$*. The expression extends to multiple constraints naturally.*

Proof. Starting from the probability expression $p_i = e^{-\xi_0 - \xi_1 f(x_i)}$ is easier. We left and right multiply by $f(x_i) e^{\xi_0}$, then we sum over $i \in [n]$. It gives:

$$\sum_i f(x_i) p_i e^{\xi_0} = \sum_i f(x_i) e^{\xi_0} e^{-\xi_0 - \xi_1 f(x_i)} \iff \mu_f e^{\xi_0} = \sum_i f(x_i) e^{-\xi_1 f(x_i)}. \tag{III.2.18}$$

Using the expression for ξ_0 and noticing that the LHS is also $\sum_i e^{-\xi_1 f(x_i)} \cdot \mu_f$ we can express everything on one side to get:

$$\sum_i (f(x_i) - \mu_f) e^{-\xi_1 f(x_i)} = 0 \iff \sum_i (f(x_i) - \mu_f) e^{-\xi_1 (f(x_i) - \mu_f)} = 0, \tag{III.2.19}$$

where in the last passage we have multiplied and divided by $e^{\xi_1 \mu_f}$. Continuity and strict monotonicity are trivial. □

From Prop. III.2.2, Cor. III.2.12 and Fct. III.2.16 it is understood that a set of contraints by means of expectations on a probability distribution is represented in the *least biased* way by the Boltzmann distribution. Moreover, the parameters are uniquely identified, but not explicitly given. Such an approach covers the missing constraints in the problem presented at the beginning of the subsection. The philosophy is being "maximally noncommittal with regard to missing information" (Jaynes 1957).

Following the approach of Chapter I, and allowing for different energy levels $\mathscr{E}_i$, we can also recover the *dimensional-aware* expressions for Thermodynamic quantities. Considering for simplicity the case in which only one mean is given, we find that $\mu_f = \mathbb{E}\left[f(X)\right] = \langle \mathscr{E} \rangle_{\xi_1} = \mathscr{U}, \beta = \xi_1$ and $\mathfrak{F}(\beta) = -\frac{\xi_0}{\xi_1}$, or more explicitly:

$$\xi_1 = \beta = \frac{1}{k_B T} \qquad \mathscr{U} - \frac{1}{\beta}\mathscr{S} = \mathfrak{F}(\beta) = -k_B T \ln \mathcal{Z}(\beta); \tag{III.2.20}$$

$$\mathscr{S} = -\frac{\partial \mathfrak{F}(T)}{T} = -k_B \sum_i p_i \ln p_i, \tag{III.2.21}$$

where the two entropy concepts differ by a multiplicative constant, K (arbitrary) on one side, k_B on the other. As already argued, the two are just a by product of which measurement units are chosen for experiments, and we could just set them to unity.

An immediate consequence is that if the functions $\left\{f^{(l)}\right\}_{l=1}^{k}$ depend on further parameters of interest $(\rho_1, \rho_2, \ldots)$, then it is possible to recover the *force* associated to any ρ_j via the expression:

$$F_{\rho_j} = \frac{1}{\beta}\frac{\partial}{\partial \rho_j} \ln \mathcal{Z}. \tag{III.2.22}$$

Example III.2.23. *In Physics, one often considers as additional parameters volume, electric/magnetic fields, and associates to them forces such as pressure or electric/magnetic potentials. Under a different light, conjugate quantities introduced in Definition I.2.25 reappear.*

Via simple arguments, we have shown that starting from entropy the thermodynamics arena is established, in a much more straightforward manner. Previously, we had to present various principles and approaches, while such last method just requires a maximization of Shannon's Entropy. The conclusion is that, when considering systems at equilibrium, the laws of physics do not bring any added value if measurements (i.e. means) are available.[2] In the next Subsection, we widen the comparison by narrating the foundational differences of Statistics-based Scientific Methods.

III.2.1 Subjective, Objective Probability, and the concept of macroscopic uniformity

We briefly touch upon one principle that is best explained in (Jaynes 1957, Secs. 3, 4).

In Thermodynamics, there is a correspondence of rules derived from stated laws of nature, and experimental facts. At first, the maximum entropy principle lacks the latter, but we will explain why it is incorrect to say so.

A crucial aspect is that it is an inference structure built upon partial information, and therefore implicitly makes use of the *subjective* interpretation of probability. A more refined argument is that, if we allow to consider Entropy as a measure of information, at each stage the process guarantees that probabilities sharpen towards clear predictions, that should be in line with experiments only when information is sufficient.

Such information, seen as "amount of knowledge", is clearly governed by the number of states n. A little abstraction eventually leads to the observations that:

[2] This specific comment is not alone valid if time is allowed to flow. Allowing particles to move across time requires a specification of rules (equations) of motion.

1. the perspective of maximizing the entropy is a sort of *ergodic hypothesis* (Jaynes 1957);
2. the sharpening of predictions is to be interpreted in terms of the broadest assignment of weights possible that the method ensures;
3. when all micro-states share the same macroscopic properties, sharp predictions arise, and must be compared with experiments.

Conclusion #3 is referred to by Jaynes as *macroscopic uniformity* (Jaynes 1957). A direct consequence of applying it is also that misalingnment of experiments and sharp predictions is evidence for wrongful counting of states. This latter fact is understood as follows. Sharp predictions are one to one with abundance of information, which means that the maximum entropy principle is not flattening the macroscopic distribution due to lack of data. On the other hand, experiments disagree with the theory. Clearly the theory should be wrong, and it is wrong in a specific sense: given the generality of technique, it is implied that the issue is how the entropy was evaluated.

Since enumerating the states of a system is the only step of the process in which laws of Physics are used, such counting process is wrong and the experiments-maximum entropy disagreement is evidence of existence of new laws of physics.[3]

III.2.2 Conjugate quantities

The result of the previous discussions motivates the following comment which is related to the next topic; Legendre transforms. Reconsider the heat bath in Chapter I, for two systems (A, B), where we are given the mean energy of $\mathscr{E}_A$, and the entropy of system A is only an extensive function of it, i.e. $\mathscr{S}_A(\mathscr{E}_A)$. Then, we could define its temperature as $\beta = \frac{\mathrm{d}}{\mathrm{d}\mathscr{E}_A}\,\mathrm{d}\mathscr{S}_A$. The way in which temperature is measured is by means of considering the two systems as a whole, with transitions of the type $i \to j$ for system A and $l \to m$ for system B, that preserve the total energy. As before we would then enforce $\mathrm{d}\mathscr{E} = 0$; or equivalently

$$\mathscr{E}_{i,A} + \mathscr{E}_{l,B} = \mathscr{E}_{j,A} + \mathscr{E}_{m,B}, \tag{III.2.24}$$

where a state of the joint system is denoted as (il) and in this case it transitions to (jm). In practice, given the overwhelming number of states, having access to the matrix of transitions is impossible, but conservation of energy enforces that $p_{(il)}$ is a function of the total energy $\mathscr{E}_{i,A} + \mathscr{E}_{l,B}$. Therefore, the maximum entropy principle for the mean of $\mathscr{E}_A$ and the conservation of the total energy allow to write the partition function and associated condition on the Lagrangian parameter:

$$\mathscr{Z}(\xi_1) = \sum_{(il)} e^{-\xi_1(\mathscr{E}_{i,A}+\mathscr{E}_{l,B})} = \mathscr{Z}_A(\xi_1)\mathscr{Z}_B(\xi_1) \quad \langle \mathscr{E}_A\rangle = -\frac{\partial}{\partial \xi_1}\ln \mathscr{Z}_A(\xi_1). \tag{III.2.25}$$

Thus, the Lagrangian parameter is recovered by invoking the equation of the maximum entropy as $\xi_1 = \frac{\mathrm{d}\mathscr{S}_A}{d\langle \mathscr{E}_A\rangle} = \beta$.

Looking back at how we formulated the canonical ensemble, the last argument shows why it makes sense to assume that the average energy will be fixed, and shows that the two quantities are strongly related, in the sense of being conjugates (Def. I.2.25). In the next Section, we will generalize this last fact with Legendre-Fenchel transforms, that incidentally also allow for the extension of Legendre transforms to a larger class of functions.

Example III.2.26 (Entropy via free entropy)**.** *It is possible to establish the relation between free entropy and entropy. Indeed for a random variable $\mathscr{E}$ with Boltzmann distribution we find*

$$\mathcal{H}(\mathscr{E}) = -\langle \ln \mathbb{P}(\mathscr{E})\rangle = \beta\langle \mathscr{E}\rangle + \ln \mathscr{Z}(\beta) = \beta^2\frac{\partial \mathscr{F}}{\partial \beta} = -\frac{\partial \mathscr{F}}{\partial T}. \tag{III.2.27}$$

[3] A folkloristic example which we do not treat in any formal sense is the appearance of the classical vs quantum mechanics dichotomy.

Remark III.2.28. *As argued before, Shannon Entropy has the same expression of Gibbs's entropy (Def. I.3.28). Provided that a random variable has the Boltzmann distribution, it will return Helmholz free energy-like (Def. I.3.34) quantities.*

We can then eventually say that the free entropy and the Shannon entropy are related by a very nice duality relation via Legendre Transforms (see below at Def. III.4.2, Prop. III.4.27) and we can express the free entropy as the CGF of the model. In mathematical terms, it is understood via the equations:

$$\mathscr{F}(\beta) + \mathcal{H}(\mathscr{E}) = \beta\mathscr{E} \qquad K_{\mathscr{E}}(\beta) = \mathscr{F} = \beta\mathfrak{F}. \tag{III.2.29}$$

Remark III.2.30. *For the interested reader mention three recent papers. The first (Zupanovic and Kuic 2018) connects with Jaynes' work in the modern context of Boltzmann-Shannon Entropies, putting emphasis on their differences at finite size. On a similar note, the authors in (Chakrabarti and De 2000) present different axiomatic derivations of the two versions, in the spirit of (Csiszár 2008). Differently, the interesting result of (Gao 2022) is that Boltzmann's distribution can be derived from more general concepts of ensembles. There are also concerns about the Maximum Entropy principle that are worth exploring (Cardoso Dias and Shimony 1981).*

III.3 Free Energy, Energy, Variational Representations

We report here a collection of arguments that should strengthen further the connections betweeen the various principles. To begin, we remind the reader that the maximum entropy principle has showed us how the canonical ensemble can be recovered from maximizing entropy. In Thermodynamics, it entails the dimensionless relation:

$$\mathscr{f}(\beta) = \inf_{e}\{\beta u + \mathscr{s}\}. \tag{III.3.1}$$

Therefore, for constant entropy systems, the (average) energy will be minimized, while for constant average energy systems, the entropy will be maximized. Similarly, one could derive different relations between the free energy, the internal energy, the entropy and temperature. Rather than dealing with this topic, which is somewhat dependent on thermodynamic arguments, we will comment on a different perspective on free energy, that might be useful for deriving a stronger connection with Bayesian methods. Let us begin with a statement that relates Shannon's Entropy to the "log-sum-exp" function, which is again nothing but a free energy.

Proposition III.3.2. *In this proposition, we are going to give an independent proof that the log-sum-exp is the convex conjugate of the negative entropy. Namely, for any $\gamma > 0$ and $\boldsymbol{p}$ a probability distribution, considering the optimization problem:*

$$\max_{\boldsymbol{q}\in[0,1]^n;\, \sum q_i = 1} \langle \boldsymbol{p}, \boldsymbol{q}\rangle + \gamma\mathcal{H}(\boldsymbol{q}) = \gamma \ln\left[\sum_{i=1}^{n} e^{\frac{p_i}{\gamma}}\right], \tag{III.3.3}$$

where the maximum is attained at:

$$\boldsymbol{q}^\star \quad s.t. \qquad q_i^\star = \frac{e^{\frac{p_i}{\gamma}}}{\sum_{j=1}^{n} e^{\frac{p_j}{\gamma}}}. \tag{III.3.4}$$

We further remark that the optimal distribution is just a Boltzmann canonical distribution and the whole statement is very much linked with the maximum entropy principle discussed earlier.

Proof. For a given index $i \in [n]$, the derivative of the objective $G(\boldsymbol{q};\boldsymbol{p})$ is:

$$\frac{\partial}{\partial q_j} G(\boldsymbol{q};\boldsymbol{p}) = p_j - \gamma(\ln[q_j] + 1), \qquad \frac{\partial^2 G(\boldsymbol{q};\boldsymbol{p})}{\partial q_j\,\partial q_i} = \begin{cases} -\frac{\gamma}{q_i} & i = j \\ 0 & \text{otherwise} \end{cases}. \tag{III.3.5}$$

It is easy to conclude then that the zeros of the gradient will be maximas, since the Hessian is diagonal with negative entries. The solution is easily found to be:

$$\widetilde{q}_i^\star = e^{\frac{p_i}{\gamma}-1} \implies q_i^\star = \frac{e^{\frac{p_i}{\gamma}-1}}{\sum_{j=1}^n e^{\frac{p_j}{\gamma}-1}} = \frac{e^{\frac{p_i}{\gamma}}}{\sum_{j=1}^n e^{\frac{p_j}{\gamma}}} \quad \forall i \in [n], \tag{III.3.6}$$

where in the last step we have reparametrized $\boldsymbol{q}$ to be a probability distribution, without losing anything since we scaled it by a constant (i.e. it is still a stationary point). Such choice makes $\boldsymbol{q}^\star$ lie automatically in the space of discrete distributions of n-sized sample space (the $\mathbb{R}^n$ simplex), making it a maximum of our objective. A tedious calculation allows us to conclude:

$$G(\boldsymbol{q}^\star;\boldsymbol{p}) = \sum_{i=1}^n q_i^\star p_i - \gamma \sum_{i=1}^n \left[\underbrace{\frac{e^{\frac{p_i}{\gamma}}}{\sum_{j=1}^n e^{\frac{p_j}{\gamma}}}}_{\equiv q_i^\star}\right] \ln\left[\frac{e^{\frac{p_i}{\gamma}}}{\sum_{j=1}^n e^{\frac{p_j}{\gamma}}}\right] \tag{III.3.7}$$

$$= \sum_{i=1}^n \frac{e^{\frac{p_i}{\gamma}}}{\sum_{j=1}^n e^{\frac{p_j}{\gamma}}} p_i - \gamma \sum_{i=1}^n \frac{e^{\frac{p_i}{\gamma}}}{\sum_{j=1}^n e^{\frac{p_j}{\gamma}}} \left(\frac{p_i}{\gamma} - \ln\left[\sum_{j=1}^n e^{\frac{p_j}{\gamma}}\right]\right) \tag{III.3.8}$$

$$= \gamma \ln\left[\sum_{j=1}^n e^{\frac{p_j}{\gamma}}\right]. \tag{III.3.9}$$

□

Remark III.3.10. *In the above statement, $\gamma > 0$ plays the role of a regularization term, where the regularization function is Entropy. The result is the Cumulant Generating Function of a Boltzmann canonical distribution where the inverse temperature is played by $\beta = \frac{1}{\gamma}$. We are effectively dealing with a Thermodynamical notion.*

III.3.1 Mean Field Methods and Bayesian Learning

As a first result, we derive an important inequality by Gibbs.

Proposition III.3.11 (Gibbs' Inequality)**.** *Let $\{p_i\}_{i=1}^n, \{q_i\}_{i=1}^n$ be probabilities of two probability distributions. It holds that:*

$$\mathcal{H}(\{p_i\}_{i\in[n]}) \leqslant \mathcal{H}(\{p_i\}_{i\in[n]}, \{q_i\}_{i\in[n]}) \equiv -\sum_{i=1}^n p_i \log q_i. \tag{III.3.12}$$

*Further, the expression are equal if and only if $\{p_i\}_{i=1}^n \equiv \{q_i\}_{i=1}^n$. The RHS of the inequality is often termed **cross entropy**.*

Proof. By Jensen's inequality applied to the log the first result is established:

$$\mathcal{H}(\{p_i\}_{i\in[n]}) + \sum_{i=1}^n p_i \log q_i = \sum_{i=1}^n p_i \log \frac{q_i}{p_i} \leqslant \log \sum_{i=1}^n q_i = \log 1 = 0, \tag{III.3.13}$$

where we have used $\sum_{i=1}^n q_i = 1$.
Given that log is strictly concave, equality holds if and only if the fraction of weights are all equal (maximum of the convex combination), which leads to the added condition $\frac{q_1}{p_1} = \cdots \frac{q_n}{p_n} = k$. The two conditions hold when:

$$\sum_{i=1}^n q_i = \sum_{i=1}^n k p_i = k = 1 \implies \frac{p_i}{q_i} = 1 \quad \forall i \in [n]. \tag{III.3.14}$$

□

The proof technique extends well to countable spaces for the distributions, and needs just some additional details for a measure theoretic statement, which is out of the scope of these notes and will be briefly commented in Section III.5. A proof can be found in the blogpost (Siegel 2019). Additionally, from the result of Gibbs' inequality, it is possible to derive a second proof of the GBF inequality from a more information Theoretic starting point. Before doing so, we take the opportunity for a Definition.

Such composite view on entropy is suggestive of introducing a very important object in Statistics, Machine Learning and Information Theory.

Definition III.3.15 (Kullback-Leibler Divergence)**.** *For distributions (p, q) on a sample space $\mathscr{X}$ the Kullback-Leibler (KL) Divergence is:*

$$\mathsf{d}_{\mathrm{KL}}(p||q) := \mathcal{H}(\boldsymbol{p}, \boldsymbol{q}) - \mathcal{H}(\boldsymbol{p}) = \sum_{\boldsymbol{x} \in \mathscr{X}} p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}, \tag{III.3.16}$$

with the convention that $0 \log 0 = 0$. By Gibbs' inequality, it is always positive. It is also known as relative entropy.

Remark III.3.17. *By the Neyman-Pearson Lemma, the best test for distinguishing probability distributions is the likelihood ratio test. The KL divergence is the expected value of such ratio if the data were distributed according to the density of $\boldsymbol{p}$. Recently, there has been a burst in attempting to connect Statistical Inference, Statistical Physics heuristics and computationally bounded Inference heuristics that arose in Average Case Complexity literature. There, the likelihood ratio plays a pivotal role. For a starting reference, see (Kunisky, Wein, and Bandeira 2019). While the topic is out of the scope of the current document, it might be treated sometime in the future.*

Remark III.3.18. *As we will argue and see across the lines, KL divergence is the most natural generalization of Entropy to continuous distributions that satisfies the constraints to some extent. A more formal discussion can be found in (Hobson 1971).*

A proof of GBF-type inequality is then just a rewriting of properties of the KL. We place it in a statement here to reference it later.

Proposition III.3.19 (GBF inequality)**.** *For two systems with Hamiltonians $\mathscr{H}, \widetilde{\mathscr{H}}$ on the same randomness $\boldsymbol{X}$ it holds:*

$$\mathcal{Z}(\beta) \geqslant \widetilde{\mathcal{Z}}(\beta) \exp\left\{-\beta \left\langle \Delta \mathscr{H}(\boldsymbol{X}) \right\rangle_{\sim}\right\}. \tag{III.3.20}$$

Proof. Let p be the canonical distribution of $\widetilde{\mathscr{H}}$ and q be the distribution of $\mathscr{H}$. The positivity of the KL divergence, which holds by Gibbs' Inequality, is a restatement of the claim. Indeed:

$$\begin{aligned} \mathsf{d}_{\mathrm{KL}}(p||q) &= \sum_{\boldsymbol{x} \in \mathscr{X}} \frac{1}{\widetilde{\mathcal{Z}}(\beta)} e^{-\beta \widetilde{\mathscr{H}}(\boldsymbol{x})} \log \frac{e^{-\beta \widetilde{\mathscr{H}}(\boldsymbol{x})}}{\widetilde{\mathcal{Z}}(\beta)} \frac{\mathcal{Z}(\beta)}{e^{-\beta \mathscr{H}}(\boldsymbol{x})} & \text{(III.3.21)} \\ &= \log \frac{\mathcal{Z}(\beta)}{\widetilde{\mathcal{Z}}(\beta)} + \sum_{\boldsymbol{x} \in \mathscr{X}} \mathbb{P}[\widetilde{\mathscr{H}}(\boldsymbol{x})] \beta(\mathscr{H}(\boldsymbol{x}) - \widetilde{\mathscr{H}}(\boldsymbol{x})). & \text{(III.3.22)} \end{aligned}$$

By positivity, we get:

$$\log \frac{\mathcal{Z}(\beta)}{\widetilde{\mathcal{Z}}(\beta)} \geqslant -\beta \left\langle \mathscr{H} - \widetilde{\mathscr{H}} \right\rangle_{\sim} \iff \mathcal{Z}(\beta) \geqslant \widetilde{\mathcal{Z}}(\beta) \exp\left\{-\beta \left\langle \Delta \mathscr{H}(\boldsymbol{X}) \right\rangle_{\sim}\right\}. \tag{III.3.23}$$

□

Corollary III.3.24. *Gibbs Inequality and the GBF inequality are the same statement. One holds if and only if the other holds. In particular, we can establish that the following equality is true:*

$$\mathcal{H}(\boldsymbol{q}, \boldsymbol{p}) - \mathcal{H}(\boldsymbol{p}) = \log \frac{\mathcal{Z}(\beta)}{\widetilde{\mathcal{Z}}(\beta)} + \left\langle \beta \Delta \mathscr{H}(\boldsymbol{X}) \right\rangle_{\sim}, \tag{III.3.25}$$

where $\boldsymbol{p}$ is the set of probabilities of $\widetilde{\mathscr{H}}$ and $\boldsymbol{q}$ is the set of probabilities of $\mathscr{H}$.

Corollary III.3.26. *As a by-product of the construction, one can realize that for an approximating Hamiltonian $\widetilde{\mathscr{H}}$, another relation holds. Letting $\mathfrak{F}_{\mathsf{app}}$ be the approximating free energy, defined as*

$$\mathfrak{F}_{\mathsf{app}}(\beta) := \widetilde{\mathfrak{F}}(\beta) + \left\langle \mathscr{H}(\boldsymbol{X}) - \widetilde{\mathscr{H}}(\boldsymbol{X}) \right\rangle_{\sim}, \tag{III.3.27}$$

then the GBF inequality introduced in Sec. I.5 holds, i.e.:

$$\mathfrak{F}_{\mathsf{app}}(\beta) \geqslant \mathfrak{F}(\beta), \tag{III.3.28}$$

and in particular, the displacement is:

$$\mathfrak{F}_{\mathsf{app}}(\beta) - \mathfrak{F}(\beta) = \frac{k_{\mathrm{B}}}{\beta} \mathsf{d}_{\mathrm{KL}}(\mu || \widetilde{\mu}) \geqslant 0, \tag{III.3.29}$$

as argued previously, now by positivity of the KL divergence.

Proof. Using the definitions of Chapter I, it is just a matter of computation. □

The GBF procedure, and the use of the KL divergence, are very akin to the concept of variational Bayes inference, briefly outlined below.
When performing Bayesian inference we suppose there exists a joint distribution of some signal $\boldsymbol{x}$ and some vector of observations $\boldsymbol{a}$. Furthermore, we place ourselves in the situation in which we wish to infer unobserved x_i for $i \in [n]$ from a data matrix $\mathbf{A}$, that stores independent $(\boldsymbol{a}_i)_{i=1}^n$, paired to the single signal, implicitly allowing for the existence of a likelihood. Then, Bayes' theorem gives an expression for the posterior:

$$\mathbb{P}\left[\boldsymbol{X}|\mathbf{A}\right] = \frac{\mathbb{P}\left[\mathbf{A}|\boldsymbol{X}\right]\mathbb{P}\left[\boldsymbol{X}\right]}{\mathbb{P}\left[\mathbf{A}\right]} = \frac{\mathbb{P}\left[\mathbf{A}|\boldsymbol{X}\right]\mathbb{P}\left[\boldsymbol{X}\right]}{\sum_{\boldsymbol{x}\in\mathscr{X}} \mathbb{P}\left[\mathbf{A}|\boldsymbol{x}\right]\mathbb{P}\left[\boldsymbol{x}\right]\mathrm{d}\boldsymbol{x}}. \tag{III.3.30}$$

However, the partition function at the denominator is often *hard to compute*, and one resorts to an approximation of the posterior by a more tractable distribution $\mathbb{Q}[\boldsymbol{X}] \approx \mathbb{P}\left[\boldsymbol{X}|\mathbf{A}\right]$. The approximation is in the simplest case evaluated by the KL divergence:

$$\begin{aligned}
\mathsf{d}_{\mathrm{KL}}(\mathbb{Q}||\mathbb{P}) &= \sum_{x\in\mathscr{X}} \mathbb{Q}[\boldsymbol{x}] \left(\frac{\log \mathbb{Q}[\boldsymbol{x}]}{\log \mathbb{P}\left[\boldsymbol{x}|\mathbf{A}\right]} \right) && \text{(III.3.31)} \\
&= \sum_{x\in\mathscr{X}} \mathbb{Q}[\boldsymbol{x}] \left(\frac{\log \mathbb{Q}[\boldsymbol{x}]}{\log \mathbb{P}\left[\boldsymbol{x},\mathbf{A}\right]} + \log \mathbb{P}\left[\mathbf{A}\right] \right) && \text{(III.3.32)} \\
&= \sum_{x\in\mathscr{X}} \mathbb{Q}[\boldsymbol{x}] \left(\log \mathbb{Q}[\boldsymbol{x}] - \log \mathbb{P}\left[\boldsymbol{x},\mathbf{A}\right] \right) + \log \mathbb{P}\left[\mathbf{A}\right] && \text{(III.3.33)} \\
&= \mathbb{E}_{\boldsymbol{X}\sim\mathbb{Q}} \left[\log \mathbb{Q}[\boldsymbol{X}] - \log \mathbb{P}\left[\boldsymbol{X},\mathbf{A}\right] \right] + \log \mathbb{P}\left[\mathbf{A}\right]. && \text{(III.3.34)}
\end{aligned}$$

After a careful inspection, defining the canonical mean energy as:

$$\beta^{-1}\mathscr{U}(\mathbb{Q}) := -\mathbb{E}_{\boldsymbol{X}\sim\mathbb{Q}} \left[\log \mathbb{P}\left[\boldsymbol{X},\mathbf{A}\right] \right], \tag{III.3.35}$$

we have recovered a notion of free entropy-like[4] object $\mathscr{F}(\mathbf{A},\mathbb{Q})$, which is however dependent on the choice of $\mathbb{Q}$. Such term $\mathscr{F}(\mathbf{A};\mathbb{Q})$ is often called *variational Gibbs free entropy* (Krzakala and Zdeborová 2021, Thm. 4) in literature, and we can express it in two ways, either via Bayesian objects, or via Information Theory/Thermodynamics objects. For completeness, we report both explicitly:

$$\begin{aligned}
\beta^{-1}\mathscr{F}(\mathbf{A};\mathbb{Q}) &\equiv \mathbb{E}_{\boldsymbol{X}\sim\mathbb{Q}} \left[\log \mathbb{P}\left[\boldsymbol{X},\mathbf{A}\right] - \log \mathbb{Q}[\boldsymbol{X}] \right] && && \text{(III.3.36)} \\
&= \log \mathbb{P}\left[\mathbf{A}\right] - \mathsf{d}_{\mathrm{KL}}(\mathbb{Q}||\mathbb{P}) && \text{Bayes} && \text{(III.3.37)} \\
&= \mathcal{H}(\mathbb{Q}) - \beta^{-1}\mathscr{U}(\mathbb{Q}) && \text{IT/Thermo.} && \text{(III.3.38)}
\end{aligned}$$

Having recovered a closed form approximation of the true model, we wish it to be as tight as possible. The formulations of the Gibbs Inequality/GBF inequality

[4] in the Machine Learning literature, it would be the Evidence Lower Bound (ELBO)

state that an approximate model is at best a lower bound on the true free entropy, so it holds that for the same temperature:

$$\mathscr{F}(\mathbf{A};\mathbb{Q}) \leqslant \mathscr{F}(\mathbf{A}), \qquad \max_{\mathbb{Q}\in\mathscr{P}(\mathbb{R}^n)} \mathscr{F}(\mathbf{A};\mathbb{Q}) = \mathscr{F}(\mathbf{A}). \tag{III.3.39}$$

Obviously, we are only restating the problem up to now, and the computation is just as hard. Indeed, the principled Bayesian approach is not just to establish this relation, but also to choose an approximation over a space $\mathcal{Q} \subset \mathscr{P}(\mathbb{R}^n)$ which is easy to deal with, to obtain an achievable lower bound. The most trivial way to make the distribution $\mathbb{Q}$ tractable is assuming that it factorizes over $i \in [n]$, so that:

$$\mathbb{Q}[\boldsymbol{x}] = \prod_{i=1}^{n} q(x_i) \quad q \in \mathscr{P}(\mathbb{R}), \tag{III.3.40}$$

which is equivalent to assuming that the approximate Hamiltonian in the canonical ensemble is:

$$\widetilde{\mathscr{H}}(\boldsymbol{x}) = \sum_{i=1}^{n} h_i(x_i), \quad h_i(\cdot) = \log q(\cdot) \quad \forall i \in [n]. \tag{III.3.41}$$

A little thought shows that the analogy underlined in the computations of Mean-Field Approach (Sec. I.5) is very well known in Statistics. The former is an approximation of the Hamiltonian, the latter is an approximation of the distribution leading to an equivalent formulation in a different perspective. Moreover, it is important to stress two facts.

1. On one side, the results we saw apply to general approximations of distributions or Hamiltonian.
2. On the other hand, the words mean-field are self-explanatory, and refer to the precise approximation of assuming that there is a field[5] that acts on particles in place of their interactions. It allows for a nice interpretation, and makes sense in many cases with the appropriate adjustments. One above all, it serves the purpose of being an often-accurate first approximation of a phenomenon.

In addition to this, the equivalence provides a nice interpretation: assuming that the distribution factorizes for each $i \in [n]$ is like assuming that each i, despite being interacting with all the other coordinates, is seen as a particle that is subject to the mean-field of the neighbors, reducing the degrees of freedom from $n-1$ to 1.

Further References

An interesting more Physics-oriented review of the story of mean-field methods is (Kadanoff 2009).

III.4 Legendre Transforms

Remark III.4.1. *For some statements, we will work on the extended real numebrs $\overline{\mathbb{R}}$. Their definition is intuitive and just requires adding some algebraic rules to have nicer sets to work with, and be able to treat the real line as having endpoints.*

Further References

A classic reference is (Zia, Redish, and McKay 2009), but also (Deserno 2012; Touchette 2014; Zanghì 2013). General discussions in (Krzakala and Zdeborová 2021; Mezard and Montanari 2009) are worth consulting.

We take a slightly more mathematical route and aim to present results that specialize to Physics and especially Thermodynamics with nice interpretations.

[5] a field is a proper mathematical object, it can be loosely seen as a constant force at each point of the ambient space

Definition III.4.2 (Legendre-Fenchel Transform)**.** *For a function $f : I \to \mathbb{R}$ with $I \subset \mathbb{R}$ its Legendre-Fenchel (LF) transform is given by*

$$\mathfrak{L}[f](w) = f^{\mathrm{LF}}(w) := \sup_{x \in I} \{wx - f(x)\} \quad f^{\mathrm{LF}} : I^{\mathrm{LF}} \to \mathbb{R}, \tag{III.4.3}$$

where the domain of the transform is $I^{\mathrm{LF}} = \{w \in \mathbb{R} | f^{\mathrm{LF}}(w) < \infty\}$. Compact notation is obvious in the sense that $f^{\mathrm{LF}} = (f)^{\mathrm{LF}}$ is the LF transform of f. The double transform is:

$$\mathfrak{L}[(f^{\mathrm{LF}})](x) = \sup_{w \in I^{\mathrm{LF}}} \{wx - \mathfrak{L}[f](w)\}\,. \tag{III.4.4}$$

We also extend the notion naturally to multivariate real-valued functions $f : I \to \mathbb{R}$ such that $I \subset \mathbb{R}^d$ as

$$\mathfrak{L}[f](\boldsymbol{w}) = \sup_{\boldsymbol{x} \in I} \{\langle \boldsymbol{x}, \boldsymbol{w} \rangle - f(\boldsymbol{x})\} \quad \boldsymbol{w} \in I^{\mathrm{LF}} = \{\boldsymbol{w} \in \mathbb{R}^d | \mathfrak{L}[f](\boldsymbol{w}) < \infty\}, \tag{III.4.5}$$

where $\langle \cdot, \cdot \rangle$ is the inner product.

The definition can be extended further, but due to its easy form, we will stay in the one-dimensional case.

Fact III.4.6. *An equivalent definition of LF transform is*

$$\mathfrak{L}[f](w) = \inf_{x \in \mathbb{R}} \{wx - f(x)\} \tag{III.4.7}$$

Proof. While the statement might be confusing, some changes of variable just give the result. Indeed:

$$-f^{\mathrm{LF}}(w) = -\sup_{x \in I} \{wx - f(x)\} = \inf_{x \in I} \{-wx + f(x)\}, \tag{III.4.8}$$

and for a choice $g(x) = -f(x)$ and $g^{\mathrm{LF}}(w) = -f^{\mathrm{LF}}(-w)$ one gets the identity. □

We say that the Legendre-Fenchel transform is *involutive* when the double LF transform of f is f. This is not guaranteed a priori; an interesting question is exactly which functions present such property.
With such target in mind, we give some operational definitions. Throughout, $f : \mathbb{R} \to \mathbb{R}$ for simplicity and $f < \infty$ by assumption unless otherwise stated.

Definition III.4.9 (Graph and Epigraph)**.** *The graph of a function $f : \mathscr{X} \to \mathbb{R}$ is the collection of its tuples, i.e.*

$$\mathsf{graph}(f) := \{(\boldsymbol{x}, y) \in \mathscr{X} \times \mathbb{R} \mid y = f(\boldsymbol{x})\}; \tag{III.4.10}$$

the epigraph is the set of values above a certain graph, i.e.

$$\mathsf{epigraph}(f) := \{(\boldsymbol{x}, y) \in \mathscr{X} \times y \mid y \geqslant f(\boldsymbol{x})\}. \tag{III.4.11}$$

Fact III.4.12. *A function f is convex if and only if its epigraph is convex.*

Proof. The easiest proof is graphical. We give an equivalent mathematical statement. Assign $E := \mathsf{epigraph}(f)$.
($\Longrightarrow$) Let f be convex, and consider a collection $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \in E$. For arbitrary $\{\lambda_i\}_{i=1}^n$ such that $\sum_{i=1}^n \lambda_i = 1$, inspect the point $(\boldsymbol{x}, y) = (\sum_i \lambda_i \boldsymbol{x}_i, \sum_i \lambda_i y_i)$. It holds:

$$y = \sum_{i=1}^n \lambda_i y_i \geqslant \sum_{i=1}^n \lambda_i f(\boldsymbol{x}_i) \geqslant f\left(\sum_i \lambda_i \boldsymbol{x}_i\right) = f(\boldsymbol{x}), \tag{III.4.13}$$

where in the first passage we used the definition of epigraph and in the second the assumed convexity. The set E is convex since the convex combination $(\boldsymbol{x}, y)$ is in E.
($\Longleftarrow$) Let E be convex. Proceeding in a similar fashion, it is not hard to prove

that for a convex combination of the points in E for which $y_i = f(\boldsymbol{x}_i)$ (i.e. the boundary points), one has the following chain of inequalities:

$$f\left(\sum_{i=1}^{n} \lambda_i \boldsymbol{x}_i\right) = f(\boldsymbol{x}) \leqslant y = \sum_{i=1}^{n} \lambda_i y_i = \sum_{i=1}^{n} f(\boldsymbol{x}_i), \tag{III.4.14}$$

where in the first passage we assign some $\boldsymbol{x}$, in the second we use the convexity of E to say that $f(\boldsymbol{x}) \leqslant y$ for some y and in the third we redirect it to the convexity of E. The fourth passage is by construction and f is convex. □

Definition III.4.15 (Supporting lines)**.** *A function f has a supporting line at $x \in \mathbb{R}$ if:*

$$\exists \alpha \in \mathbb{R} | f(y) \geqslant f(x) + \alpha(y - x) \quad \forall y \in \mathbb{R}. \tag{III.4.16}$$

The line is strictly supporting if the inequality is strict for all $y \neq x$.

We report next a nice implication of the existance of supporting lines, which connects them to convexity.

Fact III.4.17. *If f admits a supporting line for all $x \in [a, b]$ then it is convex in $[a, b]$.*

Proof. Define a region:

$$R := \{(z, v) \in [a, b] \times \mathbb{R},\ \forall x \in [a, b], v \geqslant f(x) + \alpha_x(z - x)\} \tag{III.4.18}$$

Notice that by being an intersection of convex sets R is convex. Additionally, one can see that R is the epigraph of f. Then f is convex by Fct. III.4.12. □

A quick glance at the Figures in (Touchette 2014) provides a geometrical intepretation of supporting lines. In words, for a point x, we have the affine function $g(y; \alpha) = f(x) + \alpha(y - x)$, where x is fixed, for some $\alpha \in R$. Whether the function is supporting or non supporting says many things about the shape of f. The line is supporting if it is always below the function. If it is intersecting the function at one single point x, then it is a strictly supporting line (y is any other point of the function in Def. III.4.15), and the function is strictly convex at x. For another point, we have that the supporting line is not beneath the function, and f is not convex. Therefore, we can infer another statement for free.

Fact III.4.19. *If f has a supporting line at x and the derivative $f'(x)$ exists, then $\alpha = f'(x)$, where α is that of Definition III.4.15.*

III.4.1 A non-exhaustive collection of results for the Legendre-Fenchel transform

The following statements provide a good understanding of what the LF transform does. Their proofs can be found in (Rockafellar 1970). Eventually, we will showcase how to recognize involutive functions.

Proposition III.4.20 (Convexity guarantees)**.** *It holds that:*

1. *$f^{\mathrm{LF}}(w)$ is always convex;*
2. *$(f^{\mathrm{LF}})^{\mathrm{LF}}(x)$ is always convex.*

Hence, both admit a supporting line for any point in their respective domains. As a byproduct:

(bonus) if f is not convex then $(f^{\mathrm{LF}})^{\mathrm{LF}} \neq f$,

which is trivial, and not all functions are involutive.

Proposition III.4.21 (Duality of supporting lines and strict supporting lines)**.** *Let f admit a supporting line at x for a constant $w \in \mathbb{R}$, i.e. $\forall y$ it holds $f(y) \geqslant f(x) + w(y - x)$. Then;*

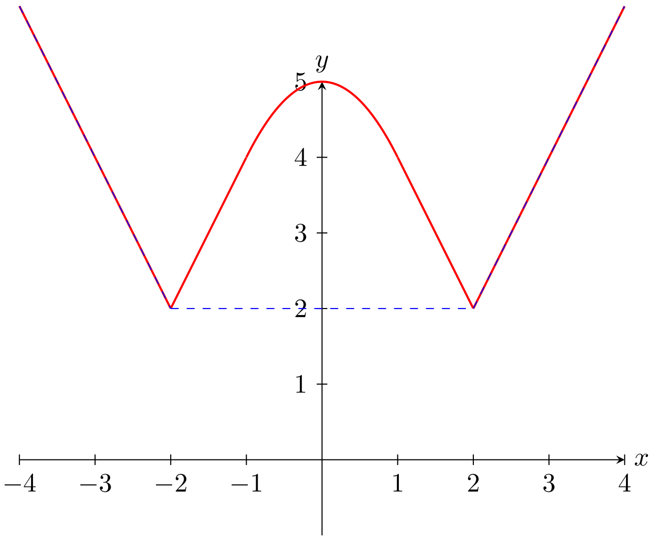

Figure III.2: Convex envelope in dashed blue of a function in red.

1. *f^{LF} admits a supporting line at point w for constant x;*

2. *if the assumption is changed to strict supporting, then the statement is that f^{LF} has derivative w at x, so the supporting line is the tangent as in Fct. III.4.19. Namely, strict support in the original function implies differentiability of the LF transform in a specific "dual" way.*

Proposition III.4.22 (A partial version of the Fenchel-Moreau Theorem)**.** *A characterization of transforms point-wise is:*

$$(f^{\mathrm{LF}})^{\mathrm{LF}}(x) = f(x) \iff f \textit{ admits supporting line at } x. \tag{III.4.23}$$

If f^{LF} is differentiable in a point w, then $f(x) = (f^{\mathrm{LF}})^{\mathrm{LF}}(x)$ where $x = (f^{\mathrm{LF}})'(w)$.

Combining the two results, it is trivial to prove the Corollary below.

Corollary III.4.24. *Let f^{LF} be everywhere differentiable, then $f = (f^{\mathrm{LF}})^{\mathrm{LF}}$ everywhere.*

One of the most important properties of LF transforms is also reported below.

Theorem III.4.25 (Convex envelope)**.** *$(f^{\mathrm{LF}})^{\mathrm{LF}}$ is the largest convex function such that $(f^{\mathrm{LF}})^{\mathrm{LF}}(x) \leqslant f(x)$ for all x.*

For this reason, the double transform is called the *convex envelope* of the function itself. Namely it is convex whenever f is convex and approximates in the convex-best possible way f for each non-convex point (i.e. points with no supporting line). A depiction is Figure III.2. More discussion examples for many cases can be found in (Touchette 2014). Briefly, we find a structure of transforms for a function f that obeys:

$$f \longrightarrow f^{\mathrm{LF}} \longleftrightarrow (f^{\mathrm{LF}})^{\mathrm{LF}}, \tag{III.4.26}$$

where the symbol $\longleftrightarrow$ denotes a bijection. A little thought shows that the statements are already self-explanatory. We report below a step-by-step list:

- the first arrow not coming back is because f is general;
- let $f^{\mathrm{LF}} = g$, by Prop. III.4.20 applied to f, it is convex for any w;
- g^{LF} is convex by the same proposition applied to g;
- by Thm. III.4.25, $(g^{\mathrm{LF}})^{\mathrm{LF}}$ is the convex envelope of a convex function, so it is the function itself. Then f^{LF} and $(f^{\mathrm{LF}})^{\mathrm{LF}}$ are in bijection by invoking the *triple* transform.

Instead, to establish the bijection $f \longleftrightarrow (f^{\mathrm{LF}})^{\mathrm{LF}}$ one necessarily needs f to be convex, for which the two are also equal. This is greatly used in the topics we encountered via the special case Definition of the LF transform, which we report below.

Proposition III.4.27 (Legendre Transform)**.** *Let f be convex and differentiable at each point of its domain, and let f' be an invertible function, then the Legendre-Fenchel transform is simplified to the Legendre Transform:*

$$\mathfrak{L}[f](w) = w\bar{x} - f(\bar{x}) \quad s.t. \quad \bar{x} = f'^{-1}(w), \tag{III.4.28}$$

which for distinction we will call $f^{\mathrm{Leg}}(w)$.

Proof. By convexity, and the definition of LF transform, the superemum is attained, possibly at the boundary. Taking the derivative of $wx - f(x)$ wrt x, and isolating $\bar{x}$, the optimal value, we find the condition claimed on $\bar{x}$, which is well-defined by assumption. □

The special case of the transform is greatly used in Statistical Mechanics and Statistical Physics, and works in conjunction with the saddle point/steepest descent/Laplace method, seen in Section II.6. We see below an example where we provide a different point of view on ensemble equivalence, overviewed in Subsection I.3.4. Recall that micro-canonical entropy is extensive, i.e.

$$\mathscr{S}(\mathscr{E}) = \ln \mathscr{N}(\mathscr{E}; n) = n\mathscr{s}(e) + o\,(n)\,, \tag{III.4.29}$$

where we emphasized that the micro-canonical energy density depends on the per particle energy, again by extensivity. This suggests that the entropy depends on the energy density solely. Considering the partition function of the canonical ensemble, we can equivalently write it as

$$\mathscr{Z}_n(\beta) = \int \mathscr{N}(e; n) e^{-n\beta e}\,\mathrm{d}e, \tag{III.4.30}$$

where the summation is now over the densities, and not the attained energies, but is formally the same as long as we do not give a specification of what kind of integral it is. Then by the method of steepest descent and some structural assumptions (see Sec. II.6), we find that the free energy density admits the expression:

$$\mathscr{f}(\beta) = \lim_{n\to\infty} -\frac{1}{n}\mathscr{Z}_n(\beta) = \inf_e\{\beta e - \mathscr{s}(e)\}, \tag{III.4.31}$$

which by Fct. III.4.6 is the Legendre Transform of the micro-canonical entropy density. In other words, $\mathscr{s}^{\mathrm{Leg}} = \mathscr{f}$.

Remark III.4.32. *Important cases where the thermodynamic ensembles do not give identical results include:*

- *microscopic systems (n small);*
- *large systems at a phase transition (free energy has a non-analytic point);*
- *large systems with long-range interactions (energy becomes non extensive).*

III.4.2 Interpreting the Legendre Transform

The Legendre transform can be seen as a tool to display information from a different point of view. In the interesting paper by (Zia, Redish, and McKay 2009), it is made very clear. We will outline their argument below. Throughout, we refer to the function f and the input x as *dependent* and *independent* variable, like in the economics-based description of linear regression. The terminology choice is not by chance, and underlines even more the power of the contribution of Legendre and Fenchel.

The peculiarity of transformations in general is that of encoding the relationship of two objects *differently.* Since for the moment we restrict to the Legendre

transform, we will implicitly assume that the function is strictly convex and differentiable. In particular, it implies the existence of a well defined Legendre transform (see Prop. III.4.27, for which we assume slightly less).
An immediate implication is that the input and the derivative of the dependent variable $\frac{\mathrm{d}f}{\mathrm{d}x}$ are also unambiguously linked. In fact, we will see that the Legendre transform allows to express the information of $f(x)$ through an encoding, with as input the derivative of f. To begin, express the derivative as $h(x) := \frac{\mathrm{d}f}{\mathrm{d}x}$. By strict convexity, it will be strictly monotonic. Further, by trivial calculus, one can show that $h(x)$ is invertible, and that its inverse is a well-defined function $x(h)$. Then, an equivalent form of f is obtained as $f(x(h))$ for each value h.

The idea, however, is slightly more convoluted, as one must enforce well definedness via affine functions. In what follows, we reroute the reader to (Zia, Redish, and McKay 2009, Fig. 3), and the discussion in text. Alternatively, one can reproduce the drawing, since the most direct interpretation is established geometrically. In a point x, with slope $h(x) = h$, the rectangular triangle formed by:

- the tangent line;
- the projection to the x axis;
- the $y = f(x)$ axis;

has a peculiar property. One can graphically show that $hx = f + g$, where g is a second term in the sum, that depends on how the tangent is oriented wrt the two axes. Since such structure is always verified, the relationship (x, f) is equivalent to the relationship (h, g).
With enough care, we claim a direct translation to the dimensionless[6] information equivalence in Thermodynamics between entropy $\mathscr{S}$, and free entropy $\beta\mathfrak{F} = \mathscr{F}$. Indeed, previous calculations showed that they are related via a Legendre Transform:

$$\mathscr{S}(\mathscr{E}) + \mathscr{F}(\beta) = \beta\mathscr{E}, \tag{III.4.33}$$

where the identity is to be intended with only one of $(\beta, \mathscr{E})$ as independent variable, since $\{\mathscr{S}, \beta\}$ are conjugates in the sense of Def. I.2.25. Then, the differential relation is expressed as $\mathscr{E}(\beta) = \frac{\mathrm{d}\mathscr{F}(\beta)}{\mathrm{d}\beta}$ or $\beta(\mathscr{E}) = \frac{\mathrm{d}\mathscr{S}(\mathscr{E})}{\mathrm{d}\mathscr{E}}$. We can therefore encode the information either via $\{\mathscr{S}, \mathscr{E}\}$ or $\{\mathscr{F}, \beta\}$, as the two are in bijection by the result of Prop. III.4.22 and the reasoning just made.

To corroborate our reasoning, we again go backwards in the computation of the microcanonical-canonical equivalence at $n \to \infty$, discussed originally in Sec. I.3.4, and in the subsection just above. Suppose that the expressions for $\mathscr{N}(\mathscr{E})$ and $\mathcal{Z}(\beta)$, defined in Chapter I, are available. Then, the inverse Laplace transform to retrieve the former from the latter is a complex integral:

$$e^{\mathscr{S}(\mathscr{E})} = \mathscr{N}(\mathscr{E}) = \int_{\mathcal{C}} \mathcal{Z}(\beta) e^{\beta\mathscr{E}}\, \mathrm{d}\beta = \int_{\mathcal{C}} e^{-\mathscr{F}(\beta)+\beta\mathscr{E}}\, \mathrm{d}\beta, \tag{III.4.34}$$

where we have explicitly placed known objects. Assuming further that the free entropy is extensive[7], and that the energy is extensive as well, we can use the steepest descent method. The saddle point condition is:

$$\left.\frac{\mathrm{d}(\mathscr{F}(\beta) - \beta\mathscr{E})}{\mathrm{d}\beta}\right|_{\beta=\beta_\star} = 0 \iff \left.\frac{\mathrm{d}\mathscr{F}(\beta)}{\beta}\right|_{\beta=\beta_\star} = \mathscr{E}, \tag{III.4.35}$$

where $\beta_\star$ is a function of $\mathscr{E}$. Plugging it into the integral after the saddle point result we find:

$$\mathscr{N}(\mathscr{E}) \overset{n\to\infty}{\approx} \exp\{\mathscr{F}(\beta_\star) - \beta_\star\mathscr{E}\} \iff \mathscr{S}(\mathscr{E}) + \mathscr{F}(\beta_\star) = \beta_\star\mathscr{E} \quad \beta_\star \equiv \beta_\star(\mathscr{E}), \tag{III.4.36}$$

where the identification of $\beta_\star$ is in terms of Eqn. III.4.35.

To conclude, we slightly anticipate two very important names in the Theory of Large Deviations, which is treated in part of the next Section.

[6] i.e. with $k_B = 1$

[7] this is true for the canonical ensemble

Proposition III.4.37 (Donkser-Varadhan Duality Variational Formula)**.** *Consider a measurable space $(\mathcal{X}, \mathcal{F})$, endowed with two probability measures $\mu, \nu : \Omega \to \mathbb{R}_+$. Without loss of generality, let ν be absolutely continuous wrt μ, i.e. $\nu \ll \mu$.*[8] *Suppose $h : \mathcal{X} \to \mathbb{R}$ is integrable wir=th respect to $(\mathcal{X}, \mathcal{F}, \mu)$. Then:*

$$\ln \int e^{h(\boldsymbol{x})} \mathrm{d}\mu(\boldsymbol{x}) = \sup_{\nu \ll \mu} \left\{ \int h(\boldsymbol{x}) \mathrm{d}\nu(\boldsymbol{x}) - \mathsf{d}_{\mathrm{KL}} (\nu || \mu) \right\}. \tag{III.4.38}$$

Moreover, the supremum becomes a maximum (i.e. there is an explicit ν attaining it), if and only if:

$$\frac{\mathrm{d}\nu}{\mathrm{d}\mu}(\boldsymbol{x}) \overset{\mathsf{a.s.}}{=} \frac{e^{h(\boldsymbol{x})}}{\int e^{h(\boldsymbol{x})} \mathrm{d}\mu(\boldsymbol{x})} \quad \textit{in} \qquad (\mathcal{X}, \mathcal{F}, \mu). \tag{III.4.39}$$

In particular, we notice that:

- *the first claim is a free energy (dual) "equal to" LF-transform of an energy minus entropy term, where the role of entropy is differential, being that it is the KL divergence;*
- *the RHS of the latter expression is a tilted measure, so a generalization of Boltzmann distributions still maximizes entropy under the constraint of a fixed energy, as pdr the classical arguments.*

Furthermore, integrability of the exponentially weighted function $\mathrm{e}^h(\cdot)$ allow for a short proof.

Proof. We argue by direct plug in, and first principles of the KL divergence, which we will show below. We aim to use the fact that the KL divergence is a valid divergence, and satisfies non-negativity, with nullity if and only if the arguments belogn to the same equivalence class, here identified by $\nu^\star \overset{\mathsf{a.s.}}{=} \nu$ in $(\mathcal{X}, \mathcal{F}, \mu)$. For this purpose, we define the candidate optimal measure:

$$\mathrm{d}\nu^\star(\boldsymbol{x}) \overset{\mathsf{a.s.}}{:=} \frac{e^{h(\boldsymbol{x})}}{\int e^{h(\boldsymbol{x})} \mathrm{d}\mu(\boldsymbol{x})} \mathrm{d}\mu(\boldsymbol{x}). \tag{III.4.40}$$

By construction, one has that $\nu^\star \ll \mu$, and that for any other $\nu \ll \mu$:

$$\mathsf{d}_{\mathrm{KL}}(\nu || \nu^\star) - \mathsf{d}_{\mathrm{KL}}(\nu || \mu) = -\int h(\boldsymbol{x}) \mathrm{d}\nu + \ln \int e^{h(\boldsymbol{x})} \mathrm{d}\mu. \tag{III.4.41}$$

Reordering the above expression, we can isolate the argument of the suprematization on the RHS of the claim, to find:

$$\int h(\boldsymbol{x}) \mathrm{d}\nu - \mathsf{d}_{\mathrm{KL}}(\nu || \mu) = \ln \left[\int e^{h(\boldsymbol{x})} \mathrm{d}\mu \right] - \mathsf{d}_{\mathrm{KL}}(\nu || \nu^\star) \tag{III.4.42}$$

$$\leqslant \ln \left[\int e^{h(\boldsymbol{x})} \mathrm{d}\mu \right], \tag{III.4.43}$$

where the inequality is a simple application of the non-negativity of divergences. Such fact established the first claim, by the RHS being the supremum over any absolutely continuous measure. To obtain the second claim, it suffices to observe that equality holds if and only if $\nu \overset{\mathrm{a.s.}}{=} \nu^\star$ in the space induced by μ, again by the Kullback-Leibler divergence being a divergence. □

Remark III.4.44. *The careful reader will have noticed that in all the steps the only steps used the property that divergences are non-negative. Consequently, the Donsker-Varadhan formula extends to any of these, in the sense of a variational link between:*

- *a free-energy type term;*
- *an entropic term;*
- *an energy term.*

[8] Absolute continuity guarantees that the Radon-Nykodym derivative $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$ is well-defined for each $\boldsymbol{x} \in \mathcal{X}$.

III.5 Spot Topics Deserving More Space

We now turn to the final comments on the path we embarked into in this Chapter. With the aim of justifying partially the results of Chapter I, and using the Theory of Chapter II, we had to introduce many objects and partially formalizing properties of earlier ones. However, there are still gaps between modern research in Statistical Physics-inspired Machine Learning and all of the above. In this Section, we briefly overview two very important fields. Each would require a book on its own, so we will refer to appropriate sources. As far as this Section is concerned, we hope to at least discuss the terminology, which appears often in publications, and the main objectives. The box below will serve as a useful starting point for further details.

III.5.1 More about information Theory and some Statistics

Further References

A comprehensive book in preparation is (Polyanskiy and Wu 2023). There, many of the claims to be presented here can be found. In this document, some proofs are skipped, but hopefully only those that are easily found online. A quick way to understand the connections between the objects we will present, and many that we even avoid nominating, requires just glancing at the map of information distances (Nielsen 2023). The perspective of Information Geometry can be experienced from the eyes of two books (S.-i. Amari 2016; S. Amari et al. 2007), which also contain material to gather some omitted proofs.

The first quick remark is related to what we saw in Section III.2, i.e. that the least biased choice of probability distribution is always Boltzmann-like, without the need to discuss ensembles and their validity. One could indeed argue that the construction of the entropy by the properties (Sh1)-(Sh3) could be *ad-hoc*, but it turns out that the same conclusion can be derived starting from different axioms. A **very good** survey in the matter is (Csiszár 2008).
Having established this, we proceed with other issues.

To expand further the applications of Information Theory, we would like to use it for continuous random variables. Unfortunately, the straightforward expression for entropy

$$-\int_{-\infty}^{\infty} p(\boldsymbol{x}) \ln(p(\boldsymbol{x}))\, \mathrm{d}\boldsymbol{x} \tag{III.5.1}$$

is not appropriate (see (Marsh 2013) for a discussion). Many properties are transferred instead if one considers the KL divergence (Def. III.3.15 here, or "relative entropy" in (Marsh 2013)). We now present a set of objects and properties that generalize the KL divergence greatly and provide some context to it.

Further References

The concurrent definition of Entropy by Kolmogorov is an interesting development of Measure Theory, see (Walters 2000, Chap. 2) for a book exposition and (Kong 2019) for a survey with applications. Another set of interesting references and perspectives can be retrieved from the slides of (Guntuboyina 2012).

Divergences

We first put into perspective the notion of KL divergence from a Statistics standpoint.

Definition III.5.2 (Divergence)**.** *For a differentiable manifold*[9] $\mathscr{M}$ *with* $\mathsf{dim}(\mathscr{M}) = n$ *a divergence is a square differentiable function* $d : \mathscr{M} \times \mathscr{M} \to [0, \infty)$ *such that:*

[9] we avoid discussing Manifolds. It is sufficient to think of it as a space that is locally regular enough to allow for *classic calculus*. For a reference, consider (S.-i. Amari 2016; S. Amari et al. 2007)

1. ***(positivity)*** *$d(p||q) \geqslant 0$ for all $p, q \in \mathscr{M}$;*
2. ***(identification)*** *$d(p||q) = 0$ if and only if $p \equiv q$.*

In Information Geometry a third requirement is present. Informally, the divergence supports the construction of an inner product space on the tangent at a point. For the sake of simplicity, we will avoid it.

Remark III.5.3. *Divergences generalize distances, since they do not require symmetry. For the same reason, they lack a priori a triangular inequality.*

The main idea between divergences is giving different perspectives on how a dissimilarity notion between objects (in such case, probability distributions) can be defined. Being that different works started from different constructions, the result is that there is a plethora of divergences, each coming with its advantages and disadvantages. Introducing the main three groups is instrumental to understand quickly their strengths. In particular, we report some notions to justify the introduction of the KL divergence. As a matter of fact, it was constructed in Definition III.3.15 only as a rearrangement of Gibbs' inequality (Prop. III.3.11), and some notions in Mean-Field interpretations, but it can be derived on different grounds. Our starting objects bear the names of two influential Hungarian Mathematicians.

Definition III.5.4 (Csizár divergence)**.** *Let (μ, ν) on $\mathscr{X}$ be two measures such that μ is absolutely continuous to ν, denoted as $\mu \ll \nu$. Let f be a function $f : [0, \infty) \to (-\infty, \infty]$ such that:*

- *$f(x) < \infty$ for all x;*
- *$f(1) = 0$;*
- *f is continuos at 0, or the convention $f(0) \equiv \lim_{\boldsymbol{x} \to 0_+} f(\boldsymbol{x})$ is enforced.*

We define the Csizár divergence as:

$$\mathsf{d}_f(\mu||\nu) := \int_{\mathscr{X}} f\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right) \mathrm{d}\nu. \tag{III.5.5}$$

In some references, it is termed f-divergence.
Without absolute continuity, for a suitable[10] ρ such that $\mu \ll \rho, \nu \ll \rho$ one defines densities $\mu(\mathrm{d}x) = p(x)\rho(\mathrm{d}x), \nu(\mathrm{d}x) = q(x)\rho(\mathrm{d}x)$. By the observation that f is convex and $f(0) = 1$, the function $\frac{f(x)}{x-1}$ is nondecreasing and the function $\lim_{x \downarrow 0} xf\left(\frac{1}{x}\right)$ takes values in $(-\infty, \infty]$. Accordingly, we redefine the divergence as:

$$\mathsf{d}_f(\mu||\nu) = \int_{p>0} p(x) f\left(\frac{p(x)}{q(x)}\right) \rho(\mathrm{d}x) + f'(\infty)\mu(p=0) \quad f'(\infty) \equiv \lim_{x \downarrow 0} xf\left(\frac{1}{x}\right), \tag{III.5.6}$$

where it is agreed that $\mu(p = 0) = 0$ deletes the second term regardless and we use the standards:

- *$0f\left(\frac{0}{0}\right) = 0$;*
- *$af\left(\frac{a}{0}\right) = af'(\infty)$ for all $a > 0$.*

The power of Csizár divergences is that for an appropriate choice of f, many common divergences are recovered, as the next Definition shows.

Definition III.5.7 (Notable Csizár divergences)**.** *We recover the following well-known objects:*

- *$f(x) = \frac{1}{2}|x - 1|$, total variation distance d_{TV};*
- *$f(x) = x \ln x$, KL divergence;*
- *$f(x) = -\ln x$, reverse-KL divergence;*

[10] it suffices to take $\rho = \nu + \mu$

- $f(x) = (1 - \sqrt{x})^2$, *Hellinger distance;*
- $f(x) = (1 - t)^2$, chi^2 *divergence.*

The fact that we recover many well-known objects in Mathematics is not by chance. As we will see next, the Definition is general enough to include important special cases, without losing their fundamental properties. As an *intermezzo*, we take the occasion to define a very important object in measure theory, and especially the measure-theoretic formulation of probability.

Definition III.5.8 (Transition Kernel)**.** *For a good introduction, refer to (Cinlar 2011, Chap. I, Sec. 6).*
Let $(E, \mathcal{E})$ and $(F, \mathcal{F})$ be measurable spaces. Consider the mapping:

$$T : E \times \mathcal{F} \to \mathbb{R}_+. \tag{III.5.9}$$

We say T is a transition kernel from $(E, \mathcal{E})$ <u>into</u> $(F, \mathcal{F})$ if the following two conditions hold:

- ***$\mathcal{E}$-measurability:*** *the mapping $x \to T(x, B)$ for any fixed $B \in \mathcal{F}$ is $\mathcal{E}$-measurable;*
- ***$\mathcal{F}$-measuring:*** *the mapping $B \to T(x, B)$ for any fixed $x \in \mathcal{E}$ is a measure on $(F, \mathcal{F})$.*

Let us pause a moment to digest such strangely complicated formulation. The name is self-answering: we expect a transition kernel to explicit a change of measure/probability space. To better understand, we provide two progressively practical examples.
It is indeed possible to prove that for a function $\kappa : E \times F \to \mathbb{R}_+$ that is $\mathcal{E} \otimes \mathcal{F}$ measurable the mapping:

$$T(x, B) = \int_B \kappa(x, y) \mathrm{d}\nu(y) \qquad x \in E, B \in \mathcal{F}, \tag{III.5.10}$$

where ν is a finite measure on $(F, \mathcal{F})$ is a transition kernel (see (Cinlar 2011, Chap. I, Sec. 6)). Essentially, the transition kernel is the integral of a weighted function that interpolates between the two measurable spaces, and for each element of the *source* sample space outputs a measure on the *target space*, and for each element of the sigma-algebra of the target space outputs a measurable event in the *source* space. As a matter of fact, it is nothing but a "Hilbert-space" generalization of the Markov Transition matrices, which are taught in basic Probability courses. To see this, let E, F be finite, respectively with m, n sizes. Then, a transition kernel as above is completely specified by the pairs $(x, \{y\})$, where $x \in E$, $\{y\} \subset F$, since we want to have subsets in the second argument.[11] Consequently, $T(x, \{y\}) = T_{x,y}$ can be interpreted as a matrixm with positive entries, weighted such that the columns sum to one (recall that for a fixed x, the kernel has to be a measure).
Having provided basic intuition about transition kernels, we proceed with our discussion of divergences.

Fact III.5.11 (Properties of Csizár divergences)**.** *It holds that:*

1. ***(validity)*** d_f *is a proper divergence in the sense of Def. III.5.2. Hellinger distance and total variation distance are additionally proper distances;*
2. ***(niceness)*** d_f *is linear and jointly convex in the two arguments;*
3. ***(DPI)*** *for a transition kernel T (Def. III.5.8) it holds* $\mathsf{d}_f(T(\mu)||T(\nu)) \leqslant \mathsf{d}_f(\mu||\nu)$*;*
4. ***(bonus)*** *further properties are found across the main references of this Chapter.*

Where the acronym is Data Processing Inequality (DPI).

[11] For a finite sample space, singletons generate the sigma-algebra (Cinlar 2011).

Further References

The direction we take is in some sense non-standard. For other notions of Mutual Information quantities arising from Csizár's construction, see (Sibson 1969) for an example and (**verduAmutualInformation2015**) for a review.

Definition III.5.12 (Csizár Mutual Information). *Let (X, Y) be two discrete random variables with joint measure $(X, Y) \sim \rho$ and individual measures (μ, ν). For a Csizár divergence with associated function f define the Csizár-mutual information of (X, Y) as:*

$$\mathfrak{I}_f(X;Y) := \mathsf{d}_f(\rho||\mu \otimes \nu). \tag{III.5.13}$$

For continuous random variables, the definition is constructive (see Fct. III.5.17). Whenever $f = x\ln x$ and we have the KL divergence, we will just use the symbol $\mathfrak{I}$, without subcript f.

It also turns out that Divergences and Entropy(es) are well related through Mutual Information(s). To understand better, we construct one important object, not to be confused with joint entropy.

Definition III.5.14 (Conditional Shannon Entropy). *Let (X, Y) on spaces $(\mathscr{X}, \mathscr{Y})$ be random variables. If they are discrete:*

$$\mathcal{H}(Y|X) := \sum_{x\in\mathscr{X}} p(x)\mathcal{H}(Y|X = x) = -\sum_{x\in\mathscr{X}}\sum_{y\in\mathscr{Y}} p(x, y)\ln\frac{p(y|x)}{p(x)}. \tag{III.5.15}$$

Conversely, if they have measures absolutely continuous wrt the Lebesgue measure and joint density $p(x, y)$ define:

$$\mathcal{H}(Y|X) := \int_{\mathscr{X}}\int_{\mathscr{Y}} p(x, y)\ln\frac{p(y|x)}{p(x)}\,\mathrm{d}y\,\mathrm{d}x. \tag{III.5.16}$$

We remark that the latter could be negative. To avoid wordiness, it will be termed conditional entropy.

Fact III.5.17 (Properties of conditional entropy). *For the discrete version of conditional entropy with generic discrete random variables (X, Y, Z), the following are true:*

1. $\mathcal{H}((X, Y))$ *is symmetric;*

2. $\mathcal{H}(Y|X) = 0 \iff Y = f(X)$ *where f is deterministic;*

3. $\mathcal{H}(Y|X) = \mathcal{H}(Y) \iff X \perp Y$*;*

4. ***(additive chain rule)*** $\mathcal{H}(Y|X) = \mathcal{H}((X, Y)) - \mathcal{H}(X)$ *and in general* $\mathcal{H}((X_i)_{i\leqslant n}) = \sum_{i=1}^{n}\mathcal{H}(X_i|X_1, \ldots, X_{i=1})$*;*

5. ***(Bayes' rule)*** $\mathcal{H}(Y|X) + \mathcal{H}(X) = \mathcal{H}(X|Y) + \mathcal{H}(Y)$*;*

6. $Y \perp Z|X \implies \mathcal{H}(Y|Z, X) = \mathcal{H}(Y|X)$*;*

7. $\mathcal{H}(Y|X) \leqslant \mathcal{H}(X)$*;*

8. $\mathcal{H}((X, Y)) = \mathcal{H}(X) + \mathcal{H}(Y) - \mathfrak{I}(X;Y)$*;*

9. $\mathfrak{I}(X;Y) \leqslant \mathcal{H}(X)$*;*

10. *while #7 holds, it does not necessarily hold for fixed $y \in \mathscr{Y}$. Namely, the inequality is true once weighting with $(p(y))_{y\in\mathscr{Y}}$.*

For continuous random variables, the facts hold whenever randomness is well behaved, meaning that the objects exist and are finite. When it is possible, we define mutual information for continuous distributions using #8 as:

$$\mathfrak{I}(X;Y) = \mathcal{H}(Y) - \mathcal{H}(Y|X) = \mathcal{H}(X) - \mathcal{H}(X|Y). \tag{III.5.18}$$

Fact III.5.19 (Properties of Csizár Mutual Information)**.** *For random variables (X, Y)*

1. *the Csizár mutual information satisfies a DPI;*
2. *the KL divergence satisfies a chain rule of the form:*

$$\mathsf{d}_{\mathrm{KL}}(\mu(x,y)||\nu(x,y)) = \mathsf{d}_{\mathrm{KL}}(\mu(x)||\nu(x)) + \mathsf{d}_{\mathrm{KL}}(\mu(y \mid x)||\nu(y \mid x)), \quad \text{(III.5.20)}$$

   wher ν, μ are measures on $\mathscr{X} \times \mathscr{Y}$;

3. *the Mutual Information is always positive.*

Proof. Claims #1, #2 follow by the properties of Csizár divergences, since the Mutual Information is a particular divergence. The second property follows by rearrangement. □

As a rough conclusion, the triplet:

- entropy
- divergence
- mutual information,

requires only two elements to define the third unambiguously for non-pathological cases. To give friendlier examples, we present other versions of entropy and divergence next.

Definition III.5.21 (Rényi Entropy)**.** *For $\alpha \in \mathbb{R}_+ \backslash \{0, 1\}$ and a discrete distribution $\boldsymbol{p}$ of a random variable X, Rényi entropy is:*

$$\mathcal{H}_\alpha(X) := \frac{1}{1-\alpha} \ln \left(\sum_{i=1}^{n} p_i^\alpha \right), \quad \text{(III.5.22)}$$

or when possible for absolutely continuous distributions wrt a dominating measure μ:

$$\mathcal{H}_\alpha(X) = \frac{1}{\alpha - 1} \ln \left(\int p^\alpha(x) \mu(\mathrm{d}x) \right). \quad \text{(III.5.23)}$$

In particular, the latter suffers from the same problems of Shannon's continuous entropy.

Definition III.5.24 (Rényi Divergence)**.** *A Rényi divergence of order $\alpha \in \mathbb{R}_+ \backslash \{0, 1\}$ is defined for discrete distributions $(\boldsymbol{p}, \boldsymbol{q})$ lying on the same common space as:*

$$\mathsf{d}_\alpha(\boldsymbol{p}||\boldsymbol{q}) := \frac{1}{1-\alpha} \ln \left(\sum_{i=1}^{n} \frac{p_i^\alpha}{q_i^{\alpha-1}} \right), \quad \text{(III.5.25)}$$

or when possible for absolutely continuous distributions wrt a dominating measure μ:

$$\mathsf{d}_\alpha(p||q) := \frac{1}{\alpha - 1} \int p^\alpha(x) q^{1-\alpha}(x) \mu(\mathrm{d}x). \quad \text{(III.5.26)}$$

In some references, it is termed α-divergence.

For boundary values $\alpha \in \{0, 1, \infty\}$ we extend the definitions applying the limits. For example:

$$\mathcal{H}_{\alpha \to 1}(X) := \lim_{\alpha \to 1} H_\alpha(X), \quad d_{\alpha \to 1}(p||q) := \lim_{\alpha \to 1} \mathsf{d}_\alpha(p||q), \quad \text{(III.5.27)}$$

where the α notation is kept to distinguish it as a Rényi divergence.

Definition III.5.28 (Notable Rényi Entropies and divergences)**.** *Let X be a random variable with probabilities $\boldsymbol{p}$ discrete and finite.*

- *$\mathcal{H}_{\alpha \to 0}(X) = \ln |\mathscr{X}|$, Hartley-entropy, also termed* max-*entropy;*

- $\mathcal{H}_2(X) = -\ln \mathbb{P}[X_1 = X_2]$ *with* $X_1, X_2 \overset{iid}{\sim} \boldsymbol{p}$*, the collision entropy;*
- $\mathcal{H}_{\alpha\to\infty}(X) \asymp -\ln \max_{i\in[n]} p_i$*, the* min*-entropy;*
- $\mathcal{H}_{\alpha\to 1}(X) = \mathcal{H}(X)$*, Shannon's entropy, with associated divergence* $d_{\alpha\to 1}(\boldsymbol{p}||\boldsymbol{q}) = \mathsf{d}_{\mathrm{KL}}(\boldsymbol{p}||\boldsymbol{q})$*, the KL divergence.*

Fact III.5.29 (Properties Rényi entropy and divergence)**.** *The following statements are true.*

1. *For fixed discrete finite random variable X with probabilities $\boldsymbol{p}$, $H_\alpha(X)$ is non-increasing in α.*
2. *for independent random variables $X \perp Y$ with distributions $(\boldsymbol{p}, \boldsymbol{q})$ that are a partition of the phase space:*[12]

$$\mathsf{d}_\alpha(p(X,A)||q(X,A)) = \mathsf{d}_\alpha(p(X)||q(X)) + \mathsf{d}_\alpha(p(Y)||q(Y)). \tag{III.5.30}$$

3. *Rényi divergences are proper divergences.*

Accordingly, we find that the $(\mathcal{H}_\alpha)_\alpha$, where the divergent terms have to be treated carefully, are ordered non-increasingly.

Proof. **(Claim #1)** By simple differentiation:

$$\frac{\mathrm{d}\mathcal{H}_\alpha(X)}{\mathrm{d}\alpha} = -\frac{1}{(1-\alpha)^2}\sum_{i=1}^{n} \alpha p_i^{\alpha-1} \leqslant 0 \quad \forall p. \tag{III.5.31}$$

(Claims #2, #3) The claims can be verified with simple computations. □

For further discussion on Rényi and Csizár divergences, we point the reader to (Polyanskiy and Wu 2023).

Definition III.5.32 (Bregman Divergence)**.** *Let $f : \mathscr{M} \to \mathbb{R}$, where $\mathscr{M}$ is convex, and $f : \mathscr{M} \to \mathbb{R}$ is continuously differentiable and strictly convex. The Bregman divergence of points $\boldsymbol{p}, \boldsymbol{q} \in \mathscr{M}$ is:*

$$\mathsf{d}_{\mathrm{Bre}}(\boldsymbol{p}, \boldsymbol{q}; f) := f(\boldsymbol{p}) - f(\boldsymbol{q}) - \langle \nabla f(\boldsymbol{q}), \boldsymbol{p} - \boldsymbol{q}\rangle . \tag{III.5.33}$$

From the definition, we implicitly require a notion of inner product in the space $\mathscr{M}$.

Remark III.5.34. *We provide two brief comments on the above construction.*

- *The requirement of strict convexity is enforced to make f have a unique minimizer, therefore allowing to avoid discussing which statements hold in more generality.*
- *Intuitively, a Bregman divergence is the difference of a function and its first order approximation.*

Definition III.5.35 (Notable Bregman divergences)**.** *For different choices of $(f, \mathscr{M})$ we recover:*

- *Squared euclidean distance when $f(\boldsymbol{p}) = \|\boldsymbol{p}\|^2, \mathscr{M} = \mathbb{R}^n$;*
- *Squared Mahanobis distance for $f(\boldsymbol{p}) = \frac{1}{2}\boldsymbol{p}^\top \mathbf{C}\boldsymbol{p}$ where $\mathbf{C}$ is positive definite and $\mathscr{M} = \mathbb{R}^n$;*
- *Generalized KL divergence:*

$$\mathsf{d}_{\mathrm{Bre}}(\boldsymbol{p}, \boldsymbol{q}; f) = \sum_{i=1}^{n} p_i \ln\frac{p_i}{q_i} - \sum_{i=1}^{n} p_i - \sum_{i=1}^{n} q_i, \tag{III.5.36}$$

for $f(\boldsymbol{p}) = -\mathcal{H}(\boldsymbol{p})$ and $\mathscr{M} = \mathbb{R}^n$;

[12] i.e. there are well defined functions $p(x), q(x), p(y), q(y)$

- *KL divergence for $f(\boldsymbol{p}) = -\mathcal{H}(\boldsymbol{p})$ and $\mathscr{M} = \boldsymbol{\Delta}_{n-1}$ the simplex.*[13]

Definition III.5.38 (Bregman Projection)**.** *Consider a Bregman divergence for a pair $(f, \mathscr{M})$. Let $\mathcal{U} \subset \mathscr{M}$. For a point $\boldsymbol{q} \in \mathscr{M}$ the Bregman projection onto $\mathcal{U}$ is the closest point to $\mathbf{p}$ in the set. Mathematically:*

$$P^{\mathcal{U}}_{\mathrm{Bre}}(\boldsymbol{p}) = \arg\min_{\boldsymbol{u}\in\mathcal{U}} \mathsf{d}_{\mathrm{Bre}}(\boldsymbol{u}, \boldsymbol{p}; f). \tag{III.5.39}$$

Fact III.5.40 (Properties of Bregman divegence)**.** *The following results can be established for generic $\mathsf{d}_{\mathrm{Bre}}(\cdot,\cdot;f)$ as in Definition III.5.32.*

1. *$\mathsf{d}_{\mathrm{Bre}}(\cdot,\cdot;f)$ is a proper divergence;*

2. ***(identification)** $\mathsf{d}_{\mathrm{Bre}}(\cdot,\cdot;f) = \mathsf{d}_{\mathrm{Bre}}(\cdot,\cdot;g) \iff f - g$ is affine;*[14]

3. ***(convexity)** it is strictly convex in the first argument;*

4. ***(positive linearity)** $\mathsf{d}_{\mathrm{Bre}}(\cdot,\cdot;f + \zeta g) = \mathsf{d}_{\mathrm{Bre}}(\cdot,\cdot;f) + \zeta \mathsf{d}_{\mathrm{Bre}}(\cdot,\cdot;g)$ for all strictly convex differentiable f, g and $\zeta \geqslant 0$;*

5. ***(duality A)** for f^{LF} the Legendre-Fenchel transform (Def. III.4.2) of f we have $\mathsf{d}_{\mathrm{Bre}}(\boldsymbol{p}^{\mathrm{LF}}, \boldsymbol{q}^{\mathrm{LF}}; f^{\mathrm{LF}}) = \mathsf{d}_{\mathrm{Bre}}(\boldsymbol{p}, \boldsymbol{q}; f)$ where $\boldsymbol{p}^{\mathrm{LF}} = \nabla f(\boldsymbol{p}), \boldsymbol{q}^{\mathrm{LF}} = \nabla f(\boldsymbol{q})$ for arbitrary $(\boldsymbol{p}, \boldsymbol{q})$.*

6. ***(duality B)** It holds:*

   $$\mathsf{d}_{\mathrm{Bre}}(\boldsymbol{p}, \boldsymbol{q}; f) = f(\boldsymbol{q}) + f^{\mathrm{LF}}(\boldsymbol{q}^{\mathrm{LF}}) - \left\langle \boldsymbol{p}, \boldsymbol{q}^{\mathrm{LF}} \right\rangle \tag{III.5.41}$$

   for arbitrary $(\boldsymbol{p}, \boldsymbol{q})$.

7. ***(cosine law)** for any $(\boldsymbol{p}, \boldsymbol{q}, \boldsymbol{r})$:*

   $$\mathsf{d}_{\mathrm{Bre}}(\boldsymbol{p}, \boldsymbol{q}; f) = \mathsf{d}_{\mathrm{Bre}}(\boldsymbol{p}, \boldsymbol{r}; f) + \mathsf{d}_{\mathrm{Bre}}(\boldsymbol{r}, \boldsymbol{q}; f) - (\boldsymbol{p} - \boldsymbol{r})^{\top}(\nabla f(\boldsymbol{q}) - \nabla f(\boldsymbol{r})). \tag{III.5.42}$$

8. ***(projections)** If $\mathcal{U}$ is convex, existence of a projection implies uniqueness. In particular, if $\mathcal{U}$ is closed and convex and $\mathscr{M}$ has finite dimension we have existance and uniqueness of the projection;*

9. ***(generalized Pythagora's Theorem)** for $\mathbf{p} \in \mathscr{M}$ and $\mathbf{u} \in \mathcal{U} \subset \mathscr{M}$ it holds:*

   $$\mathsf{d}_{\mathrm{Bre}}(\boldsymbol{p}, \boldsymbol{u}; f) \geqslant \mathsf{d}_{\mathrm{Bre}}(\boldsymbol{u}, P^{\mathcal{U}}_{\mathrm{Bre}}(\boldsymbol{p}); f) + \mathsf{d}_{\mathrm{Bre}}(P^{\mathcal{U}}_{\mathrm{Bre}}(\boldsymbol{p}), \boldsymbol{p}; f). \tag{III.5.43}$$

Further References

A nice generalization that reaches a functional formulation of Bregman divergences is (Frigyik, Srivastava, and Gupta 2006). There, the authors provide also an accessible appendix that reviews necessary material.

In order to introduce our last object, we move to parametric probability distributions. These live on a statistical manifold $\mathscr{M}$ with dimension d (see (S.-i. Amari 2016; S. Amari et al. 2007) for context), and are represented as "points" $p(\boldsymbol{X}; \boldsymbol{\theta})$ where $\boldsymbol{\theta} \in \mathbb{R}^d$.

[13] in a vector space V a k simplex is generated by some $k+1$ points. Precisely, if the points form distinct segments wrt an origin point, i.e. the set of vectors $\{(u_1 - u_0), \ldots, (u_k - u_0)\}$ is an independency, then we define:

$$\mathsf{Simplex}(\{u_i\}_{i=0}^{k}) := \left\{ \sum_{i=0}^{k} \alpha_i u_i \mid \alpha_i \geqslant 0 \; \forall i, \quad \sum_{i=0}^{k} \alpha_i = 1 \right\}. \tag{III.5.37}$$

The condition is also termed "affine independence", and the simplex is also found in literature as *convex hull* of the set.

[14] as a refresher, a map between vector spaces $h : V \to U$ is affine iff it can be represented as a linear transformation plus a translation, i.e. for each $x \in V$ one has $h(x) = Ax + b$ for some A linear map from V to U and $b \in U$.

Definition III.5.44 (Score)**.** *For a parametric model $p(\cdot;\cdot)$ the score is the derivative of the logarithm of the model. It reads:*

$$\nabla_{\boldsymbol{\theta}} \ln p(\boldsymbol{X};\boldsymbol{\theta}). \tag{III.5.45}$$

Assumption III.5.46 (Regularity conditions for parametric estimation)**.** *The following are standard requirements.*

(R1) the gradient wrt $\boldsymbol{\theta}$ of $p(\cdot;\cdot)$ exists almost everywhere;

(R2) differentiation and integration can be exchanged;

(R3) the support $\mathsf{supp}[f(\boldsymbol{X};\boldsymbol{\theta})]$ is independent of $\boldsymbol{\theta}$.

Remark III.5.47. *Requirement (R2) can be verified by an application of Leibniz rule, which is satisfied when the function inside the integral and its derivative are continuous almost everywhere, but can also be assumed. Other sufficient conditions are:*

- *(R3) and bounded support;*
- *or infinite support, (R1) and uniform convergence of the integral for all $\boldsymbol{\theta}$.*

An alternative set of regularity conditions consists in requiring:

- ***identifiability:** for each $\boldsymbol{\theta} \neq \boldsymbol{\theta}'$ the pdfs $f(\boldsymbol{X};\boldsymbol{\theta})$ and $f(\boldsymbol{X};\boldsymbol{\theta}')$ do not coincide;*
- *common support, i.e. (R3) above;*
- ***well-posedness:** the true parameter $\boldsymbol{\theta}^\star$ is in the interior of the paramteric space $\boldsymbol{\Theta} \subseteq \mathbb{R}^d$.*

In what follows, we adapt standard proofs to our discussion.

Fact III.5.48. *Let Ass. III.5.46 hold. If $X \sim p(\cdot;\boldsymbol{\theta})$ for some $\boldsymbol{\theta} \in \mathbb{R}^d$ the expected value of the score function is null at $\boldsymbol{\theta}$.*

Proof. A simple computation gives:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{X}}\left[\nabla_{\boldsymbol{\theta}} \ln p(\boldsymbol{x};\boldsymbol{\theta})\right] &= \int_{\mathscr{X}} \frac{\nabla_{\boldsymbol{\theta}} p(\boldsymbol{x};\boldsymbol{\theta})}{p(\boldsymbol{x};\boldsymbol{\theta})} \mathrm{d}\mu(\boldsymbol{x}) && \text{(III.5.49)}\\
&= \int_{\mathscr{X}} \frac{\nabla_{\boldsymbol{\theta}} p(\boldsymbol{x};\boldsymbol{\theta})}{p(\boldsymbol{x};\boldsymbol{\theta})} p(\boldsymbol{x};\boldsymbol{\theta}) \mathrm{d}\boldsymbol{x} && \text{(III.5.50)}\\
&= \int_{\mathscr{X}} \nabla_{\boldsymbol{\theta}} p(\boldsymbol{x};\boldsymbol{\theta}) \mathrm{d}\boldsymbol{x} && \text{(III.5.51)}\\
&= \nabla_{\boldsymbol{\theta}} \int_{\mathscr{X}} p(\boldsymbol{x};\boldsymbol{\theta}) \mathrm{d}\boldsymbol{x} && \text{(III.5.52)}\\
&= \nabla_{\boldsymbol{\theta}} \cdot 1 && \text{(III.5.53)}\\
&= 0, && \text{(III.5.54)}
\end{aligned}$$

where we were allowed to cancel the densities because we assumed $X \sim p(\cdot;\boldsymbol{\theta})$ and could exchange integral and derivative by the regularity conditions. □

Definition III.5.55 (Fisher information Matrix)**.** *Given a statistical manifold $\mathscr{M}$ with dimension d made of parametrized distributions $p(x;\boldsymbol{\theta})$ where $\theta \in \mathbb{R}^d$ and $\boldsymbol{x} \in \mathbb{R}^d$, define the Fisher Information Matrix (FIM) as the variance of the score when the true parameter is $\boldsymbol{\theta}$. Mathematically:*

$$\mathbf{F}_{\mathsf{IM}}(\boldsymbol{\theta}) = -\mathbb{E}_{\boldsymbol{X}}\left[\left(\nabla_{\boldsymbol{\theta}} \ln p(\boldsymbol{X};\boldsymbol{\theta})\right)\left(\nabla_{\boldsymbol{\theta}} \ln p(\boldsymbol{X};\boldsymbol{\theta})\right)^\top\right] \in \mathbb{R}^{d\times d}, \qquad \boldsymbol{X} \sim p(\cdot;\boldsymbol{\theta}). \tag{III.5.56}$$

where the particular expression is a consequence of the fact that the mean of the score is null.

Fact III.5.57 (Fisher information additional form)**.** *Under the regularity conditions of Ass. III.5.46, the FIM admits a nicer expression:*

$$\mathbf{F}_{\mathsf{IM}}(\boldsymbol{\theta}) = -\mathbb{E}_{\boldsymbol{X}}\left[\nabla^2_{\boldsymbol{\theta}} \ln p(\boldsymbol{X};\boldsymbol{\theta})\right] \in \mathbb{R}^{d\times d}, \tag{III.5.58}$$

Proof. As with the previous proof, it suffices to carefully compute the outer product of gradients in the definition:

$$\begin{aligned}
\mathbf{F}_{\mathsf{IM}}(\boldsymbol{\theta}) &= \int_{\mathscr{X}} (\nabla_{\boldsymbol{\theta}} \ln p(\boldsymbol{x};\boldsymbol{\theta}))\,(\nabla_{\boldsymbol{\theta}} \ln p(\boldsymbol{x};\boldsymbol{\theta}))^{\top}\, p(\boldsymbol{x};\boldsymbol{\theta})\mathrm{d}x && \text{(III.5.59)}\\
&= \int_{\mathscr{X}} \left\{ (\nabla_{\boldsymbol{\theta}} \ln p(\boldsymbol{x};\boldsymbol{\theta}))\,(\nabla_{\boldsymbol{\theta}} \ln p(\boldsymbol{x};\boldsymbol{\theta}))^{\top} - \frac{\nabla^2_{\boldsymbol{\theta}} p(\boldsymbol{x};\boldsymbol{\theta})}{p(\boldsymbol{x};\boldsymbol{\theta})} \right\} p(\boldsymbol{x};\boldsymbol{\theta})\mathrm{d}\boldsymbol{x} && \text{(III.5.60)}\\
&= \int_{\mathscr{X}} \left\{ \left(\frac{\nabla_{\boldsymbol{\theta}} p(\boldsymbol{x};\boldsymbol{\theta})}{p(\boldsymbol{x};\boldsymbol{\theta})}\right)\left(\frac{\nabla_{\boldsymbol{\theta}} p(\boldsymbol{x};\boldsymbol{\theta})}{p(\boldsymbol{x};\boldsymbol{\theta})}\right)^{\top} - \frac{\nabla^2_{\boldsymbol{\theta}} p(\boldsymbol{x};\boldsymbol{\theta})}{p(\boldsymbol{x};\boldsymbol{\theta})} \right\} p(\boldsymbol{x};\boldsymbol{\theta})\mathrm{d}\boldsymbol{x} && \text{(III.5.61)}\\
&= -\int_{\mathscr{X}} \nabla^2_{\boldsymbol{\theta}} \ln p(\boldsymbol{x};\boldsymbol{\theta}) p(\boldsymbol{x};\boldsymbol{\theta})\mathrm{d}\boldsymbol{x} && \text{(III.5.62)}\\
&= -\mathbb{E}_{\boldsymbol{X}}\left[\nabla^2_{\boldsymbol{\theta}} \ln p(\boldsymbol{X};\boldsymbol{\theta})\right] \qquad \boldsymbol{X} \sim p(\cdot;\boldsymbol{\theta}), && \text{(III.5.63)}
\end{aligned}$$

where all the passages are trivial, except the second equality, which holds since for $X \sim p(\cdot;\boldsymbol{\theta})$:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{X}}\left[\frac{\nabla^2_{\boldsymbol{\theta}} p(\boldsymbol{X};\boldsymbol{\theta})}{p(\boldsymbol{X};\boldsymbol{\theta})}\right] &= \int_{\mathscr{X}} \frac{\nabla^2_{\boldsymbol{\theta}} p(\boldsymbol{x};\boldsymbol{\theta})}{p(\boldsymbol{x};\boldsymbol{\theta})} p(\boldsymbol{x};\boldsymbol{\theta})\mathrm{d}\boldsymbol{x} && \text{(III.5.64)}\\
&= \int_{\mathscr{X}} \nabla^2_{\boldsymbol{\theta}} p(\boldsymbol{x};\boldsymbol{\theta})\mathrm{d}\boldsymbol{x} && \text{(III.5.65)}\\
&= \nabla^2_{\boldsymbol{\theta}} \int_{\mathscr{X}} p(\boldsymbol{x};\boldsymbol{\theta})\mathrm{d}\boldsymbol{x} && \text{(III.5.66)}\\
&= \nabla^2_{\boldsymbol{\theta}} \cdot 1 && \text{(III.5.67)}\\
&= 0. && \text{(III.5.68)}
\end{aligned}$$

□

Fact III.5.69 (Properties of the FIM)**.** *Let Assumption III.5.46 hold. Denote the p.s.d. order as* $\preceq, \succeq$*. Then the following statements are verified.*

1. ***(Cramér-Rao Lower bound)*** *For any estimator* $\widehat{\boldsymbol{\theta}} : \mathbb{R}^{n\times d} \to \mathbb{R}^d$ *we have the inequality*

$$\mathrm{CoV}_{\mathbf{X}}\left[\widehat{\boldsymbol{\theta}}\right] \succeq \mathbb{E}_{\mathbf{X}}\left[\widehat{\boldsymbol{\theta}}\right] \mathbf{F}_{\mathsf{IM}}(\boldsymbol{\theta})^{-1} \mathbb{E}_{\mathbf{X}}\left[\widehat{\boldsymbol{\theta}}\right]^{\top}; \tag{III.5.70}$$

2. ***(baby Cramér-Rao)*** *In particular, if the estimator is unbiased, then* $\mathrm{CoV}_{\mathbf{X}}\left[\widehat{\boldsymbol{\theta}}\right] \succeq \mathbf{F}_{\mathsf{IM}}(\boldsymbol{\Theta})^{-1}$ *holds;*

3. ***(chain rule)*** *For random variables* $(\boldsymbol{X}, \boldsymbol{Y})$ *the FIM admits a decomposition similar to mutual information (Def. III.5.12, Fct. III.5.19#1):*

$$\mathbf{F}_{\mathsf{IM}}(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{Y}) = \mathbf{F}_{\mathsf{IM}}(\boldsymbol{\theta}; \boldsymbol{X}) + \mathbf{F}_{\mathsf{IM}}(\boldsymbol{\theta}; \boldsymbol{Y}|\boldsymbol{X}), \tag{III.5.71}$$

where the second term is evaluated as an integral

$$\mathbf{F}_{\mathsf{IM}}(\boldsymbol{\theta}; \boldsymbol{Y}|\boldsymbol{X}) = \mathbb{E}_{X}\left[\mathbf{F}_{\mathsf{IM}}(\boldsymbol{\theta}; \boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x})\right], \tag{III.5.72}$$

which for fixed $\boldsymbol{x}$ *can be computed;*

4. *further properties are found across the main references of this Chapter.*

Fluctuation-Dissipation, FIM and Free Energies

To conclude, we link the Fisher information Matrix to Statistical Mechanics, with the aid of a very insightful technical paper (Crooks 2012). Imagining that the distribution of the model depends smoothly enough on a parameter $\boldsymbol{\theta}$. Further, assume that the properties just listed hold for the FIM. If the distribution of the random variable $\boldsymbol{X}$ is a canonical ensemble, we have:

$$\begin{aligned} p(\boldsymbol{x};\beta,\boldsymbol{\theta}) &= \frac{1}{\mathcal{Z}(\beta,\boldsymbol{\theta})} e^{-\beta\mathscr{E}(\boldsymbol{x};\beta,\boldsymbol{\theta})} && \text{(III.5.73)} \\ &= \exp\left\{-\ln[\mathcal{Z}(\beta;\boldsymbol{\theta})] - \beta\mathscr{E}(\boldsymbol{x};\beta,\boldsymbol{\theta})\right\} && \text{(III.5.74)} \\ &= \exp\left\{\beta\mathfrak{F}(\beta;\boldsymbol{\theta}) - \beta\mathscr{E}(\boldsymbol{x};\beta,\boldsymbol{\theta})\right\} && \text{(III.5.75)} \\ \implies &\begin{cases} \nabla^2_{\boldsymbol{\theta}}\left(\ln p(\boldsymbol{x};\beta,\boldsymbol{\theta})\right) = \nabla^2_{\boldsymbol{\theta}}\left(\beta\mathfrak{F}(\beta;\boldsymbol{\theta}) - \beta\mathscr{E}(\boldsymbol{x};\beta,\boldsymbol{\theta})\right) \\ \nabla_{\boldsymbol{\theta}}\left(\ln p(\boldsymbol{x};\beta,\boldsymbol{\theta})\right) = \nabla_{\boldsymbol{\theta}}\left(\beta\mathfrak{F}(\beta;\boldsymbol{\theta}) - \beta\mathscr{E}(\boldsymbol{x};\beta,\boldsymbol{\theta})\right) \end{cases} . && \text{(III.5.76)} \end{aligned}$$

For simplicity, we choose the former, which after squaring (outer product) and taking Expectations leads to the FIM in the sense of Definition III.5.55. Mathematically:

$$\begin{aligned} \mathbf{F}_{\mathsf{IM}}(\boldsymbol{\theta};\boldsymbol{X}) &= \mathbb{E}_{\boldsymbol{X}}\left[\left(\nabla_{\boldsymbol{\theta}} \ln p(\boldsymbol{X};\boldsymbol{\theta})\right)^2\right] && \text{(III.5.77)} \\ &= \beta^2\mathbb{E}_{\boldsymbol{X}}\left[\left[\nabla_{\boldsymbol{\theta}}\left(\mathfrak{F}(\beta;\boldsymbol{\theta}) - \mathscr{E}(\boldsymbol{X};\beta,\boldsymbol{\theta})\right)\right]\left[\nabla_{\boldsymbol{\theta}}\left(\mathfrak{F}(\beta;\boldsymbol{\theta}) - \mathscr{E}(\boldsymbol{X};\beta,\boldsymbol{\theta})\right)\right]^\top\right]. && \text{(III.5.78)} \\ &= \beta^2\mathbb{E}_{\boldsymbol{X}}\left[\left[\nabla_{\boldsymbol{\theta}}(\mathfrak{F} - \mathscr{E})\right]\left[\nabla_{\boldsymbol{\theta}}(\mathfrak{F} - \mathscr{E})\right]^\top\right], && \text{(III.5.79)} \end{aligned}$$

where in the last step we have compactified notation. It is now customary to summon one of the main properties of Free energy we proposed in Chapter I. Precisely, reminding Equation I.3.17, it is possible to conclude that:

$$\nabla_{\boldsymbol{\theta}}\mathfrak{F} = \mathbb{E}_{\boldsymbol{X}}\left[\nabla_{\boldsymbol{\theta}}\mathscr{E}\right] \implies \mathbf{F}_{\mathsf{IM}}(\boldsymbol{\theta};\boldsymbol{X}) = \beta^2 \mathrm{CoV}_{\boldsymbol{X}}\left[\nabla_{\boldsymbol{\theta}}\mathscr{E}(\boldsymbol{X};\beta,\boldsymbol{\theta})\right] \in \mathbb{R}^{d\times d}. \quad \text{(III.5.80)}$$

In words: the Fisher Information Matrix is the covariance of the gradient of the Energy wrt the parameter of interest for Boltzmann Canonical distributions. Furthermore, there is an apparent relation with the so-called Fluctuation-Dissipation relations (see (Mezard and Montanari 2009, Chap. II)), which we briefly discuss below. To begin, let us translate the computation of (Crooks 2012) in our notation. In attempting to derive a connection between FIM and Free Energy, it turns out that one needs to focus on the second derivative.[15] In mathematical Equations, some work gives as answer.

Fact III.5.81 (Connection Fisher-Free Energy). *Consider the model discussed in the current Subsubsection. Then:*

$$\mathbf{F}_{\mathsf{IM}}(\boldsymbol{\theta};\boldsymbol{X}) = \frac{1}{\beta}\left(\mathbb{E}_{\boldsymbol{X}}\left[\nabla^2_{\boldsymbol{\theta}}\mathscr{E}(\boldsymbol{X};\beta,\boldsymbol{\theta})\right] - \nabla^2_{\boldsymbol{\theta}}\mathfrak{F}(\beta;\boldsymbol{\theta})\right), \quad \text{(III.5.82)}$$

which is a regularized Hessian of the energy at the local point.

Proof. For the sake of simplicity, let us use compact notation. Identify

$$\mathfrak{F} \equiv \mathfrak{F}(\beta;\boldsymbol{\theta}), \qquad \mathscr{E} \equiv \mathscr{E}(\boldsymbol{x};\beta,\boldsymbol{\theta}), \qquad \mathbf{F}_{\mathsf{IM}} \equiv \mathbf{F}_{\mathsf{IM}}(\boldsymbol{\theta};\boldsymbol{x}). \quad \text{(III.5.83)}$$

Let us also precompute a quantity that will be useful in the calculation below. It

[15] Recall that the second derivative of the free energy is associated to a susceptibility term, and to second order phase transitions.

holds that:

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}} \mathscr{Z} &= \int_{\mathscr{X}} e^{-\beta \mathscr{E}} (-\beta \nabla_{\boldsymbol{\theta}} \mathscr{E}) \mathrm{d}\boldsymbol{x} && \text{(III.5.84)} \\
&= -\beta \mathscr{Z} \mathbb{E}_{\boldsymbol{X}} \left[\nabla_{\boldsymbol{\theta}} \mathscr{E} \right]; && \text{(III.5.85)} \\
\nabla_{\boldsymbol{\theta}} \left[\left(\int_{\mathscr{X}} e^{-\beta \mathscr{E}} \mathrm{d}\boldsymbol{x} \right)^{-1} e^{-\beta \mathscr{E}} \right] &= \nabla_{\boldsymbol{\theta}} \left[\frac{1}{\mathscr{Z}} e^{-\beta \mathscr{E}} \right] && \text{(III.5.86)} \\
&= -\frac{1}{\mathscr{Z}^2} (\nabla_{\boldsymbol{\theta}} \mathscr{Z}) e^{-\beta \mathscr{E}} - \beta e^{-\beta \mathscr{E}} \frac{1}{\mathscr{Z}} \nabla_{\boldsymbol{\theta}} \mathscr{E} && \text{(III.5.87)} \\
&= \frac{e^{-\beta \mathscr{E}}}{\mathscr{Z}} \beta \mathbb{E}_{\boldsymbol{X}} \left[\nabla_{\boldsymbol{\theta}} \mathscr{E} \right] - \beta e^{-\beta \mathscr{E}} \frac{1}{\mathscr{Z}} \nabla_{\boldsymbol{\theta}} \mathscr{E}. && \text{(III.5.88)}
\end{aligned}$$

Then:

$$\begin{aligned}
-\beta \nabla^2_{\boldsymbol{\theta}} \mathfrak{F} &= \nabla^2_{\boldsymbol{\theta}} \left[\ln \left\{ \int_{\mathscr{X}} e^{-\beta \mathscr{E}} \mathrm{d}\boldsymbol{x} \right\} \right] && \text{(III.5.89)} \\
&= -\beta \nabla_{\boldsymbol{\theta}} \left[\left(\int_{\mathscr{X}} e^{-\beta \mathscr{E}} \mathrm{d}\tilde{x} \right)^{-1} \int_{\mathscr{X}} e^{-\beta \mathscr{E}} \nabla_{\boldsymbol{\theta}} \mathscr{E} \mathrm{d}\boldsymbol{x} \right] && \text{(III.5.90)} \\
&= -\beta \int_{\mathscr{X}} \nabla_{\boldsymbol{\theta}} \left[\left(\int_{\mathscr{X}} e^{-\beta \mathscr{E}} \mathrm{d}\tilde{x} \right)^{-1} e^{-\beta \mathscr{E}} \nabla_{\boldsymbol{\theta}} \mathscr{E} \right] \mathrm{d}\boldsymbol{x} && \text{(III.5.91)} \\
&= -\beta \int_{\mathscr{X}} \nabla_{\boldsymbol{\theta}} \left[\left(\int_{\mathscr{X}} e^{-\beta \mathscr{E}} \mathrm{d}\tilde{x} \right)^{-1} e^{-\beta \mathscr{E}} \right] \left[\nabla_{\boldsymbol{\theta}} \mathscr{E} \right]^{\top} + \left[\left(\int_{\mathscr{X}} e^{-\beta \mathscr{E}} \mathrm{d}\tilde{x} \right)^{-1} e^{-\beta \mathscr{E}} \right] \left[\nabla^2_{\boldsymbol{\theta}} \mathscr{E} \right] \mathrm{d}\boldsymbol{x} && \text{(III.5.92)} \\
&= -\beta \int_{\mathscr{X}} \nabla_{\boldsymbol{\theta}} \left[\left(\int_{\mathscr{X}} e^{-\beta \mathscr{E}} \mathrm{d}\tilde{x} \right)^{-1} e^{-\beta \mathscr{E}} \right] \left[\nabla_{\boldsymbol{\theta}} \mathscr{E} \right]^{\top} \mathrm{d}\boldsymbol{X} - \beta \int_{\mathscr{X}} \frac{1}{\mathscr{Z}} e^{-\beta \mathscr{E}} \left[\nabla^2_{\boldsymbol{\theta}} \mathscr{E} \right] \mathrm{d}\boldsymbol{x} && \text{(III.5.93)} \\
&= -\beta \int_{\mathscr{X}} \left[\frac{e^{-\beta \mathscr{E}}}{\mathscr{Z}} \beta \mathbb{E}_{\boldsymbol{X}} \left[\nabla_{\boldsymbol{\theta}} \mathscr{E} \right] - \beta e^{-\beta \mathscr{E}} \frac{1}{\mathscr{Z}} \nabla_{\boldsymbol{\theta}} \mathscr{E} \right] \left[\nabla_{\boldsymbol{\theta}} \mathscr{E} \right]^{\top} \mathrm{d}\boldsymbol{x} - \beta \mathbb{E}_{\boldsymbol{X}} \left[\nabla^2_{\boldsymbol{\theta}} \mathscr{E} \right] && \text{(III.5.94)} \\
&= -\beta^2 \left[\mathbb{E}_{\boldsymbol{X}} \left[\nabla_{\boldsymbol{\theta}} \mathscr{E} \right] \right] \left[\mathbb{E}_{\boldsymbol{X}} \left[\nabla_{\boldsymbol{\theta}} \mathscr{E} \right] \right]^{\top} + \beta^2 \mathbb{E}_{\boldsymbol{X}} \left[\left[\nabla_{\boldsymbol{\theta}} \mathscr{E} \right] \left[\nabla_{\boldsymbol{\theta}} \mathscr{E} \right]^{\top} \right] - \beta \mathbb{E}_{\boldsymbol{X}} \left[\nabla^2_{\boldsymbol{\theta}} \mathscr{E} \right] && \text{(III.5.95)} \\
&= \beta^2 \mathrm{CoV}_{\boldsymbol{X}} \left[\nabla_{\boldsymbol{\theta}} \mathscr{E} \right] - \beta \mathbb{E}_{\boldsymbol{X}} \left[\nabla^2_{\boldsymbol{\theta}} \mathscr{E} \right] && \text{(III.5.96)} \\
&= \beta^2 \mathbf{F}_{\mathsf{IM}} - \beta \mathbb{E}_{\boldsymbol{X}} \left[\nabla^2_{\boldsymbol{\theta}} \mathscr{E} \right]. && \text{(III.5.97)}
\end{aligned}$$

Reordering the last equality gives the claim. □

Remark III.5.98. *Curiously, the Fisher Information defines a Riemaniann metric, which is the starting point of Information Geometry (S.-i. Amari 2016; Crooks 2012), and can be furthher used to relate Thermodynamical principles with the Cramer-Rao Inequality (Fct. III.5.69#1), as pointed out in (Crooks 2012) and the references therein.*

Remark III.5.99. *This last proposition is suggestive of an interpretation in terms of Natural Gradient Descent. For more details, see (Martens 2020) and the blog posts (Gibiansky 2014; Kristiadi 2024; Rosse 2013).*

Fact III.5.100 (KL-Fisher Connection)**.** *Let Assumptions III.5.46 hold. Consider two parametrizations $(\boldsymbol{\theta}, \boldsymbol{\theta}^\star)$ where $X \sim p(\cdot; \boldsymbol{\theta}^\star)$. Then:*

$$\nabla^2_{\boldsymbol{\theta}} \mathrm{d}_{\mathrm{KL}}(p(X; \boldsymbol{\theta}^\star) || p(X; \boldsymbol{\theta})) \bigg|_{\theta = \theta^\star} = \mathbf{F}_{\mathsf{IM}}(\boldsymbol{\theta}). \qquad \text{(III.5.101)}$$

Therefore, the local curvature of the KL divergence is the FIM.

Corollary III.5.102. *In the above setting, the KL divergence is locally/asymptotically symmetric.*

Proof. The FIM depends only on one of the two parameters, is well behaved, and such that $\boldsymbol{\theta} \approx \boldsymbol{\theta}^\star$. □

Remark III.5.103. *More information about the connection between KL divergence and FIM are found in the blogpost (Rosse 2013), and the references therein.*

A meta discussion to justify the Kullback-Leibler Divergnce

The above constructions are amenable to highlighting how important the KL divergence is, at least in terms of directly retrieving a large collection of results. In the previous arguments, we reported the basic ones. On top of them, more results are added; there are variational characterizations, inequalities between information measures, and the whole line of research that Information Theory is. Obviously, no summary gives enough credit to it.

As a final remark, we will summarize the results derived for the KL divergence, and close some gaps between divergences.

Fact III.5.104 (KL's uniqueness in terms of divergence types)**.** *The KL divergence is such that:*

1. *if the space of probabilities is finite in size, it is the only divergence that is both an Csizár divergence and a Bregman divergence (Jiao et al. 2014);*
2. *in the formalism of Rényi, the Shannon-KL pair is the only one for which:*
   (a) *the entropy satisfies the additive chain rule of Fct. III.5.17#4, i.e.* $\mathcal{H}((X,Y)) = \mathcal{H}(X) + \mathcal{H}(Y|X)$
   (b) *the divergence satisfies the additive chain rule of Fct. III.5.19#2.*

Proof. For claim #1, we reroute the reader to (Jiao et al. 2014). For the other two claims, it suffices to notice that the properties mentioned are tailored to the singular case of $\alpha = 1$ in the sense of Rényi, and do not hold in general. □

We are then ready to draw a long list of conclusions. Kullback-Leibler divergences (Def. III.3.15):

- can be seen simultaneously into three classes of divergences (Def. III.5.2): in the sense of Csizár, (limiting) Rényi and Bregman (Defs. III.5.4, III.5.24, III.5.32), inheriting all their properties (Fcts. III.5.11, III.5.29, III.5.40);
- admit a notion of Csizár Mutual information (Def. III.5.12) with nice properties (Fct. III.5.19);
- admit a formulation of conditional entropy (Def. III.5.14) which is well behaved and has nice properties specific to the KL in terms of mutual information (Fct. III.5.17);
- have as local representation of their curvature in the space of parametric probability manifolds the FIM (Def. III.5.55), according to the result of Fct. III.5.100), which highlights also some thermodynamic-like behaviors of the FIM (e.g. Fct. III.5.81), as well as some Information-Theoretic results (Fct. III.5.69). The last fact allows for some exchange of results between fields, such as the local symmetry of the KL (Cor. III.5.102);
- present uniqueness results when some properties are enforced, and no other member of a divergence family can have them (Fct. III.5.104);
- play a role in some Large Deviations Results (see e.g. in the next Section Thm. III.5.128).

It is then natural to say that the KL divergence is, for a large class of approaches, a proper object to take into consideration.

III.5.2 Large Deviations Theory

Further References

A comprehensive collection of topics is found in (Ellis 1999, 2006; Krzakala and Zdeborová 2021; Mezard and Montanari 2009; Touchette 2009). For a rigorous but more practical introduction, it is also worth checking out the series of blogposts (Yeo 2013). A collection of examples in the span of a dense course in Stochastic Processes is found in (Shalizi 2006).

Many of the results we have presented so far have been placed into rigorous terms in the context of Large Deviations theory. Since the subject is *very* wide and technical, we will gloss over it and just briefly explain its foundations. For a detailed introduction, one can consider (Dembo and Zeitouni 2010; Ellis 2006; Touchette 2009). As far as this Subsection is concerned, we only introduce the main definitions and the main Theorems. To begin, let us refresh some fundamental notions in analysis.

Definition III.5.105 (Limit superior and Limit inferior)**.** *Recall that given a partially ordered set, denoted with the pair $(\mathcal{A}, \leqslant)$, the supremum and the infimum are, respectively, the least upper bound, and the greatest lower bound. For a sequence of sets $(A_n)_{n\in\mathbb{N}}$, we see them as:*

$$\sup_n A_n = \bigcup_n A_n \qquad \inf_n A_n = \bigcap_n A_n, \tag{III.5.106}$$

which are respectively a non-decreasing and a non-increasing sequence. Analogously, we define the limit supremum and limit infimum of a sequence $(x_n)_{n\in\mathbb{N}} \subset \mathbb{F}$ in a partially ordered set as the infimum and supremum of the limit points of the sequence. In mathematical terms, we mean:

$$\liminf_{n\to\infty} x_n := \lim_{n\to\infty} \inf_{m\geqslant n} x_m \equiv \sup_{n\in\mathbb{N}} \inf_{m\geqslant n} x_n \qquad \limsup_{n\to\infty} x_n := \lim_{n\to\infty} \sup_{m\geqslant n} \{x_m\} \equiv \inf_{n\in\mathbb{N}} \sup_{m\geqslant n} x_m. \tag{III.5.107}$$

To adapt the definition to sequences of sets, it suffices to use the last equivalence for both cases, to find:

$$\liminf_{n\to\infty} A_n = \bigcup_{n\in\mathbb{N}} \bigcap_{m\geqslant n} A_m \qquad \limsup_{n\to\infty} A_n = \bigcap_{n\in\mathbb{N}} \bigcup_{m\geqslant n} A_n. \tag{III.5.108}$$

In statistics, where the sets represent events in a sigma-algebra, it is also very informative to talk about sequences that are true eventually (limit infimum) and infinitely often (limit supremum), where the words are in accordance with the mathematical structure once one accepts that A_n signals an event, with an associated probability. We reroute the reader to (Cinlar 2011, Chap. III, Sec. 2) for context.

Definition III.5.109 (Lower semi-continuity)**.** *A function $f : \mathscr{X} \to [-\infty, \infty]$ is lower semi-continuous (l.s.c.) at a point $\boldsymbol{x}_0$ if for every $y < f(\boldsymbol{x}_0)$ there exists a neighborhood $\mathcal{B}(\boldsymbol{x}_0)$ such that $f(x) > y$ for all $\boldsymbol{x} \in \mathcal{B}(x_0)$. A function is l.s.c. if it is l.s.c. for each point in the domain.*

Fact III.5.110 (Equivalent definitions of lower semicontinuity at a point)**.** *The Following Are Equivalent (TFAE):*

1. *f is l.s.c. at x_0*
2. $\liminf_{x\to x_0} f(x) \geqslant f(x_0)$

Fact III.5.111 (Equivalent definitions of lower semicontinuity)**.** *TFAE:*

1. *f is l.s.c.*
2. *all y-sublevel sets where $y \in \mathbb{R}$ are closed in the domain space. Namely for each $y \in \mathbb{R}$ the set $\{\boldsymbol{x} \in \mathscr{X} : f(\boldsymbol{x}) > y\}$ is closed in $\mathscr{X}$*
3. *the epigraph (Def. III.4.9) is closed in $\mathscr{X} \times \mathbb{R}$*

Definition III.5.112 (Rate function)**.** *A function $\mathscr{I} : \mathscr{X} \to [0,\infty]$ which is not identically ∞ and is lower semicontinuous.*

Definition III.5.113 (Large Deviation Principle)**.** *A collection of measures $(\mu_\rho)_{\rho>0}$ satisfies a Large Deviation Principle (LDP) with rate function $\mathscr{I}$ and rate $\frac{1}{\rho}$ if for closed set A and every open set B in $\mathscr{X}$ we have the following bounds:*

$$\limsup_{\rho\downarrow 0} \rho \ln \mu_\rho(A) \leqslant -\inf_{\boldsymbol{x}\in A} \mathscr{I}(\boldsymbol{x}), \quad \liminf_{\rho\downarrow 0} \rho \ln \mu_\rho(B) \leqslant -\inf_{\boldsymbol{x}\in B} \mathscr{I}(\boldsymbol{x}). \qquad \text{(III.5.114)}$$

In the context of our applications, we will use a rather simplistic view of the concept, which states that for a given random variable $(\boldsymbol{X}_n)_{n\in\mathbb{N}}$ dependent on an index n and a set A in its space of events the LDP is established when:

$$\lim_{n\to\infty} -\frac{1}{n} \ln \mathbb{P}[\boldsymbol{X}_n \in A] = \mathscr{I}(A), \qquad \text{(III.5.115)}$$

which means that at the event A the probability law has a leading exponential behavior, i.e. heuristically $\mathbb{P}[\boldsymbol{X}_n \in [\boldsymbol{x}, \boldsymbol{x} + \mathrm{d}\boldsymbol{x}]] \asymp e^{-n\mathscr{I}(\boldsymbol{x})}\,\mathrm{d}\boldsymbol{x}$. In particular, we are implicitly assuming that the lim sup *and the* lim inf *above are equal.*

To conclude, we provide a dry list of statements.

Fundamental Results in Large Deviations Theory

Theorem III.5.116 (Cramér's Theorem)**.** *Let $(X_i)_{i\geqslant 1}$ be a sequence of iid random variables which admit a MGF (i.e. $\mathbb{E}\left[e^{tX_i}\right] < \infty$ for every $t \in \mathbb{R}$). Then, for $S_n = \sum_{i=1}^n X_i$ it holds*

$$\lim_{n\to\infty} \frac{1}{n} \ln \mathbb{P}\left[S_n \geqslant an\right] = -\mathscr{I}(a) \quad \forall a > \mathbb{E}\left[X_1\right], \qquad \text{(III.5.117)}$$

where $\mathscr{I}(w) := \sup\{wt - \ln \mathbb{E}\left[e^{tX_1}\right]\}$. We recognize that $\mathscr{I}(w) = \mathfrak{L}[K_{X_1}(t)](w)$. In other words, the sum random variable satisfies a LDP with rate function being the negation of the Legendre-Fenchel transform of the CGF.

Proof. A proof and a more general statement for Euclidean vectors in $\mathbb{R}^d$ is found in (Dembo and Zeitouni 2010). □

The independence and identical distribution assumption can be dropped, in favour of a more general statement that highlights the connection between rate functions and CGFs. Let $\mathfrak{z}_n := \frac{1}{n} S_n$, where S_n now is a sum in a Euclidean space $\mathscr{X}$. Denote the sequence of laws of $\mathfrak{z}_n$ as $(\mu_n)_{n\geqslant 1}$.

Definition III.5.118 (Exposed pair)**.** *For a function $f : \mathbb{R}^d \to \overline{\mathbb{R}}$, a vector $\boldsymbol{y} \in \mathbb{R}^d$ is an exposed point if for some $\boldsymbol{t} \in \mathbb{R}^d$ it holds that:*

$$\langle \boldsymbol{t}, \boldsymbol{y}\rangle - f(\boldsymbol{y}) > \langle \boldsymbol{t}, \boldsymbol{x}\rangle - f(\boldsymbol{x}) \quad \forall \boldsymbol{x} \in \mathbb{R}^d. \qquad \text{(III.5.119)}$$

We then call $\boldsymbol{t}$ an exposed hyperplane, and the tuple $(\boldsymbol{y}, \boldsymbol{t})$ is an exposed pair.

Remark III.5.120. *Convex functions such as the LF transform have an exposed point where they are strictly convex.*

Definition III.5.121 (Domain, interior, closure, boundary)**.** *A function $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ has domain $\mathcal{D}_f := \{\boldsymbol{x} \in \mathbb{R}^d : f(\boldsymbol{x}) < \infty\}$. Its interior $\mathsf{int}(\mathcal{D}_f)$ is made of all those points $\boldsymbol{x}_0$ for which an open ball centered at $\boldsymbol{x}_0$ is included in $\mathcal{D}_f$. The closure is the set of all points of closure, such that every open ball contains at least one point in $\mathcal{D}_f$. We write it as $\mathsf{cl}(\mathcal{D}_f)$.The boundary is made of points that are between the closure and the interior. Namely*

$$\partial(\mathcal{D}_f) = \mathsf{cl}(\mathcal{D}_f)\backslash\mathsf{int}(\mathcal{D}_f). \qquad \text{(III.5.122)}$$

Definition III.5.123 (Steep Function)**.** *A function $f : \mathbb{R}^d \to \mathbb{R}$ is steep when for $\boldsymbol{x} \in \partial\mathcal{D}_f$ the boundary of the interior domain, it holds:*

$$\lim_{\boldsymbol{y}\to\boldsymbol{x}} |\nabla f(\boldsymbol{y})| = \infty. \qquad \text{(III.5.124)}$$

Remark III.5.125. *Refer in particular to (Ellis 2006; Krzakala and Zdeborová 2021) and (Shalizi 2006, Lecture 34). Each gives a different perspective on the result.*

Theorem III.5.126 (Gärtner-Ellis Theorem)**.** *With the above setting, assume the following:*

- *the limit of the CGF exists in the sense that:*

$$K_{\mathfrak{z}}(\boldsymbol{t}) = \lim_{n\to\infty} \frac{1}{n} K_{\mathfrak{z}_n}(n\boldsymbol{t}) \in \overline{\mathbb{R}} \quad \forall \boldsymbol{t} \in \mathbb{R}^d \tag{III.5.127}$$

- $\boldsymbol{t} = \boldsymbol{0}$ *is in the finite values of the limiting CGF, i.e.* $K_{\mathfrak{z}}(\boldsymbol{0}) < \infty$.

Then for $\mathcal{I}(\cdot) = \mathfrak{L}[K_{\mathfrak{z}}](\cdot)$ *the LF transform of the CMF it holds:*

1. $\limsup_{n\to\infty} \frac{1}{n} \ln \mu_n(A) \leqslant -\inf_{\boldsymbol{w}\in A} \mathcal{I}(\boldsymbol{w})$ *for all closed* $A \subset \mathscr{X}$*;*
2. $\limsup_{n\to\infty} \frac{1}{n} \ln \mu_n(B) \geqslant -\inf_{\boldsymbol{w}\in B\cap E} \mathcal{I}(\boldsymbol{w})$ *for all open* $B \subset \mathscr{X}$*, where* E *is the set of exposed points of* $\mathcal{I}(\boldsymbol{w})$ *with associated exposed hyperplane in the interior domain of* $M_{\mathfrak{z}}$*;*
3. *if* $K_{\mathfrak{z}}(\cdot)$ *is l.s.c., differentiable on the interior domain and steep then the infimum of #2 is only over open sets* B*;*
4. *if #3 holds, the measures* $(\mu_n)_{n\geqslant 1}$ *satisfy a LDP with rate function being the LF transform of the CGF.*

Lastly, we inspect empirical distributions, which provide a second degree of deviation in the construction of (Ellis 2006, Chaps. I, II). This theory is important since it allows to go beyond conclusions on the mean such as those of Cramér and Gärtner-Ellis (Thms. III.5.116, III.5.126).

Theorem III.5.128 (Sanov's Theorem)**.** *Let* $L_n := \frac{1}{n}\sum_{i=1}^n \delta_{\boldsymbol{X}_i}$ *be the empirical measure of* $\boldsymbol{X}_1, \ldots \boldsymbol{X}_n$ *sampled iid from* μ*. Then, it satisfies a LDP on the space of measures* $\mathscr{P}(\mathscr{X})$ *with rate function* $\mathsf{d}_{\mathrm{KL}}(\cdot||\mu)$*.*[16]

Corollary III.5.129 (Baby Sanov's)**.** *Let* $\{X_i\}_{i=1}^n$ *be iid from* μ *with finitely many values. Lert the empirical distribution be* L_n *as before. Then for a set of measures* $\mathcal{M} \subset \mathscr{P}(\mathbb{R})$*:*

$$\nu_n(L_n \in \mathcal{M}) \leqslant (n+1)^{|\mathscr{X}|} 2^{-n\mathsf{d}_{\mathrm{KL}}(L^\star||\mu)}, \quad L^\star = \operatorname*{arg\,min}_{L\in\mathcal{M}} \mathsf{d}_{\mathrm{KL}}(L||\mu), \tag{III.5.130}$$

where ν_n *is a measure on measures.*
Furthermore, if $\mathcal{M}$ *is the closure of its interior, then we have a clean LDP which reads:*

$$\lim_{n\to\infty} \frac{1}{n} \ln \nu^n(L_n \in \mathcal{M}) = -\mathsf{d}_{\mathrm{KL}}(L^\star||\mu). \tag{III.5.131}$$

Remark III.5.132. *The KL divergence with one fixed measure* $\mathsf{d}_{\mathrm{KL}}(\cdot||\mu)$ *can be seen as a functional equivalent of the LF transform for the space of measures. For more comments, see (Touchette 2009, Sec. IVB). From the baby result, we get the more direct interpretation that any law on empirical measures is bounded from above by a term that depends on:*

- *number of samples* n*;*
- *size of the sample space* $\mathscr{X}$*;*
- *being closest in KL divergence measure to the real distribution.*

All together, the result is quite powerful.

[16] We trivially extend the KL divergence to two measures if they are absolutely continuous, otherwise just set it to infinity. In this case, the empirical measure is absolutely continuous wrt the real measure.

We remark that the above collection is only partially justified. We would require an independent treatment to make it sound better, and there are already very useful resources. The main purpose was showing how some objects of the previous theory can be mathematically recovered in a more rigorous manner all under the same principles. Adding other tools, many concepts seen earlier are then derived on formal grounds. In particular, we find these statements in (Ellis 2006, Chaps. III-V), (Touchette 2009, Sec. V)), where also a formal treatment of ensemble equivalence is developed.

Bibliography

Akbari, Kamran, Thomas Bury, and Brendon Phillips (2015). "The Method of Steepest Descent". University of Waterloo.

Amari, Shun-ichi (2016). *Information Geometry and Its Applications.* Vol. 194. Applied Mathematical Sciences. Tokyo: Springer Japan. ISBN: 978-4-431-55977-1 978-4-431-55978-8. DOI: `10.1007/978-4-431-55978-8`. (Visited on 04/19/2024).

Amari, Shun'ichi et al. (2007). *Methods of Information Geometry.* Trans. by Daishi Harada. Nachdruck. Translations of Mathematical Monographs 191. Providence, Rhode Island: American Mathematical Society. ISBN: 978-0-8218-4302-4 978-0-8218-0531-2.

Arfken, George B. and Hans-Jurgen Weber (2013). *Mathematical Methods for Physicists.* Elsevier. ISBN: 978-0-12-384654-9. DOI: `10.1016/C2009-0-30629-7`. (Visited on 08/18/2023).

Arovas, Daniel (2019). "Lecture Notes on Thermodynamics and Statistical Mechanics". (Visited on 08/11/2023).

Basak, Anirban and Sumit Mukherjee (Aug. 2017). "Universality of the Mean-Field for the Potts Model". In: *Probability Theory and Related Fields* 168.3, pp. 557–600. ISSN: 1432-2064. DOI: `10.1007/s00440-016-0718-0`. (Visited on 03/09/2024).

Blundell, Stephen J. and Katherine M. Blundell (Oct. 2009). *Concepts in Thermal Physics.* 2nd ed. Oxford University PressOxford. ISBN: 978-0-19-956209-1 978-0-19-171823-6. DOI: `10.1093/acprof:oso/9780199562091.001.0001`. (Visited on 07/20/2023).

Cardoso Dias, Penha Maria and Abner Shimony (June 1981). "A Critique of Jaynes' Maximum Entropy Principle". In: *Advances in Applied Mathematics* 2.2, pp. 172–211. ISSN: 0196-8858. DOI: `10.1016/0196-8858(81)90003-8`. (Visited on 04/26/2024).

Chakrabarti, C. G. and Kajal De (2000). "Boltzmann-Gibbs Entropy: Axiomatic Characterization and Application". In: *International Journal of Mathematics and Mathematical Sciences* 23.4, pp. 243–251. ISSN: 0161-1712, 1687-0425. DOI: `10.1155/S0161171200000375`. (Visited on 08/15/2023).

Cinlar, Erhan (2011). *Probability and Stochastics.* Vol. 261. Graduate Texts in Mathematics. New York, NY: Springer. ISBN: 978-0-387-87858-4 978-0-387-87859-1. DOI: `10.1007/978-0-387-87859-1`. (Visited on 11/07/2022).

Cohn, Steve (2007). "Integral Asymptotics 3: Stationary Phase - Math 842-843". (Visited on 05/19/2024).

Cramér, Harald (1999). *Mathematical Methods of Statistics.* Princeton Landmarks in Mathematics and Physics. Princeton: Princeton University Press. ISBN: 978-0-691-00547-8.

Crooks, Gavin E (2012). "Fisher Information and Statistical Mechanics". In.

Cross, Michael (2006). *Physics 127a.* http://www.pmaweb.caltech.edu/~mcc/Ph127/index.html. (Visited on 08/11/2023).

Csiszár, Imre (Sept. 2008). "Axiomatic Characterizations of Information Measures". In: *Entropy* 10.3, pp. 261–273. ISSN: 1099-4300. DOI: `10.3390/e10030261`. (Visited on 08/15/2023).

Debye, P. (Dec. 1909). "Näherungsformeln für die Zylinderfunktionen für große Werte des Arguments und unbeschränkt veränderliche Werte des Index". In: *Mathematische Annalen* 67.4, pp. 535–558. ISSN: 1432-1807. DOI: `10.1007/BF01450097`. (Visited on 09/27/2022).

Dembo, Amir and Ofer Zeitouni (2010). *Large Deviations Techniques and Applications*. Vol. 38. Stochastic Modelling and Applied Probability. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN: 978-3-642-03310-0 978-3-642-03311-7. DOI: `10.1007/978-3-642-03311-7`. (Visited on 09/20/2023).

Deserno, Marcus (2012). “Legendre Transforms”. (Visited on 08/15/2023).

Ehrenfest, Paul et al. (July 1960). “*The Conceptual Foundations of the Statistical Approach in Mechanics*”. In: *Physics Today* 13.7, pp. 50–52. ISSN: 0031-9228, 1945-0699. DOI: `10.1063/1.3057042`. (Visited on 08/05/2023).

Ellis, Richard S. (Sept. 1999). “The Theory of Large Deviations: From Boltzmann’s 1877 Calculation to Equilibrium Macrostates in 2D Turbulence”. In: *Physica D: Nonlinear Phenomena* 133.1-4, pp. 106–136. ISSN: 01672789. DOI: `10.1016/S0167-2789(99)00101-3`. (Visited on 04/26/2024).

— (2006). *Entropy, Large Deviations, and Statistical Mechanics*. Classics in Mathematics. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN: 978-3-540-29059-9 978-3-540-29060-5. DOI: `10.1007/3-540-29060-5`. (Visited on 08/06/2023).

Evans, Martin (2006). *Methods of Mathematical Physics, Lectures 2,5,6, 12*. https://www2.ph.ed.ac.uk/~m (Visited on 08/15/2023).

— (2009). “Mean-Field Theory of the Ising Model”. University of Edinburgh, MP4 Statistical Physics. (Visited on 03/09/2024).

Feller, William (2009). *An Introduction to Probability Theory and Its Applications. Vol. 1*. 3. ed., rev. print., [Nachdr.] Vol. 1. Wiley Series in Probability and Mathematical Statistics. S.l.: Wiley. ISBN: 978-0-471-25708-0.

Figueroa O’Farril, Josè (1998). “Integral Transforms, Mathematical Techniques III, Chapter 3”. (Visited on 08/15/2023).

Frigyik, B. A., S. Srivastava, and M. R. Gupta (Nov. 2006). *Functional Bregman Divergence and Bayesian Estimation of Distributions*. arXiv: `cs/0611123`. (Visited on 04/19/2024).

Gao, Xiang (Mar. 2022). “The Mathematics of the Ensemble Theory”. In: *Results in Physics* 34, p. 105230. ISSN: 22113797. DOI: `10.1016/j.rinp.2022.105230`. (Visited on 08/15/2023).

Gibbs, J. W. (Dec. 1878). “On the Equilibrium of Heterogeneous Substances”. In: *American Journal of Science* s3-16.96, pp. 441–458. ISSN: 0002-9599. DOI: `10.2475/ajs.s3-16.96.441`. (Visited on 08/05/2023).

Gibiansky, Andrew (2014). *Gauss Newton Matrix*. (Visited on 02/26/2024).

Greenlee, Wilfred Martin (2005). “On Green’s Theorem and Cauchy’s Theorem”. In: *Real Analysis Exchange* 30.2, p. 703. ISSN: 01471937. DOI: `10.14321/realanalexch.30.2.0703`. JSTOR: `10.14321/realanalexch.30.2.0703`. (Visited on 03/11/2024).

Guntuboyina, Aditya (2012). *Statistics 212a - Information Theory and Statistics*.

Hobson, Arthur (1971). *Concepts in Statistical Mechanics*. New York, NY: Gordon and Breach. ISBN: 978-0-677-03240-5.

Jaynes, E. T. (May 1957). “Information Theory and Statistical Mechanics”. In: *Physical Review* 106.4, pp. 620–630. ISSN: 0031-899X. DOI: `10.1103/PhysRev.106.620`. (Visited on 07/19/2023).

Jeffreys, Harold (1998). *Theory of Probability*. 3rd ed. Oxford Classic Texts in the Physical Sciences. Oxford [Oxfordshire] : New York: Clarendon Press ; Oxford University Press. ISBN: 978-0-19-850368-2.

Jiao, Jiantao et al. (Dec. 2014). “Information Measures: The Curious Case of the Binary Alphabet”. In: *IEEE Transactions on Information Theory* 60.12, pp. 7616–7626. ISSN: 0018-9448, 1557-9654. DOI: `10.1109/TIT.2014.2360184`. (Visited on 05/04/2024).

Kadanoff, Leo P. (Dec. 2009). “More Is the Same; Phase Transitions and Mean Field Theories”. In: *Journal of Statistical Physics* 137.5-6, pp. 777–797. ISSN: 0022-4715, 1572-9613. DOI: `10.1007/s10955-009-9814-1`. arXiv: `0906.0653 [cond-mat, physics:hep-th, physics:physics]`. (Visited on 08/18/2023).

Keynes, John Maynard (2004). *A Treatise on Probability*. Dover Phoenix Editions. Mineola, N.Y: Dover Publications. ISBN: 978-0-486-49580-4.

Kittel, Charles (2004). *Elementary Statistical Physics*. Dover ed. Mineola, N.Y: Dover Publications. ISBN: 978-0-486-43514-5.

Kong, Tianyu (2019). “ERGODIC THEORY, ENTROPY AND APPLICATION TO STATISTICAL MECHANICS”.

Kristiadi, Augustinus (2024). *Natural Gradient Descent.* (Visited on 02/25/2024).

Krzakala, Florent and Lenka Zdeborová (2021). “Statistical Physics Methods in Optimization and Machine Learning”. In: p. 225.

Kunisky, Dmitriy, Alexander S. Wein, and Afonso S. Bandeira (July 2019). *Notes on Computational Hardness of Hypothesis Testing: Predictions Using the Low-Degree Likelihood Ratio.* arXiv: `1907.11636 [cs, math, stat]`. (Visited on 06/27/2023).

Lairez, Didier (Oct. 2022). *What Entropy Really Is : The Contribution of Information Theory.* arXiv: `2204.05747 [physics]`. (Visited on 04/26/2024).

— (Feb. 2023). *A Short Derivation of Boltzmann Distribution and Gibbs Entropy Formula from the Fundamental Postulate.* DOI: `10.48550/arXiv.2211.02455`. arXiv: `2211.02455 [cond-mat, physics:physics]`. (Visited on 08/15/2023).

Luschny, Peter (2010). *Is the Gamma-function Misdefined?* https://www.luschny.de/math/factorial/hadamard/Hada (Visited on 08/16/2023).

Marsh, Charles O. (2013). “Introduction to Continuous Entropy”. In: (visited on 08/19/2023).

Martens, James (2020). “New Insights and Perspectives on the Natural Gradient Method”. In: *Journal of Machine Learning Research* 21.146, pp. 1–76.

Maxwell, J. C. (Jan. 1860). “V. *Illustrations of the Dynamical Theory of Gases.* —Part I. *On the Motions and Collisions of Perfectly Elastic Spheres*”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 19.124, pp. 19–32. ISSN: 1941-5982, 1941-5990. DOI: `10.1080/14786446008642818`. (Visited on 08/05/2023).

Mehran, Kardar (2023a). *Statistical Mechanics I: Statistical Mechanics of Particles | Physics.* https://ocw.mit.edu/courses/8-333-statistical-mechanics-i-statistical-mechanics-of-particles-fall-2013/. (Visited on 08/11/2023).

— (2023b). *Statistical Mechanics II: Statistical Physics of Fields | Physics.* https://ocw.mit.edu/courses/8-334-statistical-mechanics-ii-statistical-physics-of-fields-spring-2014/. (Visited on 08/11/2023).

Mezard, Marc and Andrea Montanari (2009). *Information, Physics, and Computation.* Oxford Graduate Texts. Oxford ; New York: Oxford University Press. ISBN: 978-0-19-857083-7.

Miller, Peter D. (2006). *Applied Asymptotic Analysis.* Graduate Studies in Mathematics v. 75. Providence, RI: American Mathematical Society. ISBN: 978-0-8218-4078-8.

Moore, Eliakim Hastings (1900). “A Simple Proof of the Fundamental Cauchy-Goursat Theorem”. In: *Transactions of the American Mathematical Society* 1.4, pp. 499–506. ISSN: 0002-9947. DOI: `10.2307/1986368`. JSTOR: `1986368`. (Visited on 03/11/2024).

Nielsen, Frank (2023). *Divergences, Dissimilarities, Discrepancies, Discriminations, Displacements, Diversities, Affinities.* https://franknielsen.github.io/Divergence/index.html. (Visited on 09/16/2023).

Polyanskiy, Yury and Yihong Wu (2023). *Information Theory, from Coding to Learning.* (Visited on 04/13/2024).

Rockafellar, Ralph Tyrell (Dec. 1970). *Convex Analysis:* Princeton University Press. ISBN: 978-1-4008-7317-3. DOI: `10.1515/9781400873173`. (Visited on 11/20/2022).

Rosse, Roger (2013). *Fisher Information.* (Visited on 04/19/2024).

Rozman, Michael G. (2017). “Method of Stationary Phase, Mathematical Methods for the Physical Sciences - Physics 2400”. (Visited on 05/19/2024).

Rubel, L. A. (1956). “Necessary and Sufficient Conditions for Carlson’s Theorem on Entire Functions”. In: *Transactions of the American Mathematical Society* 83.2, pp. 417–429. ISSN: 0002-9947, 1088-6850. DOI: `10.1090/S0002-9947-1956-0081944-8`. (Visited on 08/16/2023).

Rudin, Walter (1987). *Real and Complex Analysis.* 3rd ed. New York: McGraw-Hill. ISBN: 978-0-07-054234-1.

Salasnich, Luca (n.d.). “Lecture Notes on Dirac Delta Function, Fourier Transform, Laplace Transform”. In: ().

Schwartz, Matthew D. (2021). “Statistical Mechanics, Course Notes”. (Visited on 07/19/2023).

Shalizi, Cosma (2006). “Stochastic Processes (Advanced Probability II)”. (Visited on 04/26/2024).

— (2024). *Laplace Approximation.*

Shannon, C. E. (July 1948). “A Mathematical Theory of Communication”. In: *Bell System Technical Journal* 27.3, pp. 379–423. ISSN: 00058580. DOI: `10.1002/j.1538-7305.1948.tb01338.x`. (Visited on 08/06/2023).

Sibson, Robin (June 1969). “Information Radius”. In: *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 14.2, pp. 149–160. ISSN: 1432-2064. DOI: `10.1007/BF00537520`. (Visited on 05/04/2024).

Siegel, Paul (May 2019). *Gibbs’ Inequality.* (Visited on 04/26/2024).

Stratonovich, R. L. (July 1957). “On a Method of Calculating Quantum Distribution Functions”. In: *Soviet Physics Doklady* 2, p. 416. (Visited on 08/01/2023).

Strichartz, Robert S (June 2003). *A Guide to Distribution Theory and Fourier Transforms.* WORLD SCIENTIFIC. ISBN: 978-981-238-421-8 978-981-277-543-6. DOI: `10.1142/5314`. (Visited on 08/15/2023).

Teitel, Stephen (2021). “Equivalence of Ensembles, Phys418S21”. Tex. Rochester University, USA. (Visited on 03/09/2024).

Tong, David (2012). “University of Cambridge Part II Mathematical Tripos”.

Touchette, Hugo (July 2009). “The Large Deviation Approach to Statistical Mechanics”. In: *Physics Reports* 478.1-3, pp. 1–69. ISSN: 03701573. DOI: `10.1016/j.physrep.2009.05.002`. arXiv: `0804.0327 [cond-mat]`. (Visited on 08/15/2023).

— (2014). “Legendre Fenchel Transform in a Nutshell”. In.

— (June 2015). “Equivalence and Nonequivalence of Ensembles: Thermodynamic, Macrostate, and Measure Levels”. In: *Journal of Statistical Physics* 159.5, pp. 987–1016. ISSN: 0022-4715, 1572-9613. DOI: `10.1007/s10955-015-1212-2`. arXiv: `1403.6608 [cond-mat]`. (Visited on 08/11/2023).

Tsallis, Constantino (July 2019). “Beyond Boltzmann–Gibbs–Shannon in Physics and Elsewhere”. In: *Entropy* 21.7, p. 696. ISSN: 1099-4300. DOI: `10.3390/e21070696`. (Visited on 08/11/2023).

Utermohlen, Franz (2018). “Mean Field Theory Solution of the Ising Model”.

Walters, Peter (2000). *An Introduction to Ergodic Theory.* Graduate Texts in Mathematics 79. New York Heidelberg Berlin: Springer. ISBN: 978-0-387-95152-2.

Watson, G. N. (1918). “The Harmonic Functions Associated with the Parabolic Cylinder”. In: *Proceedings of the London Mathematical Society* s2-17.1, pp. 116–148. ISSN: 00246115. DOI: `10.1112/plms/s2-17.1.116`. (Visited on 08/18/2023).

Weber, Hans-Jurgen and George B. Arfken (2004). *Essential Mathematical Methods for Physicists.* San Diego, CA: Academic Press. ISBN: 978-0-12-059877-9.

Wong, R. (Jan. 2001). *Asymptotic Approximations of Integrals.* Society for Industrial and Applied Mathematics. ISBN: 978-0-89871-497-5 978-0-89871-926-0. DOI: `10.1137/1.9780898719260`. (Visited on 08/18/2023).

Yeo, Dominic (Aug. 2013). *Eventually Almost Everywhere; Large Deviation Theory.* (Visited on 04/26/2024).

Young, Peter (2012). *Thermal Physics, Physics 112 Lecture Notes, The (Generalized) Free Energy Is a Minimum in Equilibrium.*

Zanghì, Nino (2013). *CONVEX FUNCTIONS AND THERMODYNAMIC POTENTIALS.* (Visited on 07/19/2023).

Zenisek, Alexander (Feb. 1999). “Green’s Theorem from the Viewpoint of Applications”. In: *Applications of Mathematics* 44.1, pp. 55–80. ISSN: 0862-7940, 1572-9109. DOI: `10.1023/A:1022272204023`. (Visited on 03/11/2024).

Zia, R. K. P., Edward F. Redish, and Susan R. McKay (July 2009). “Making Sense of the Legendre Transform”. In: *American Journal of Physics* 77.7, pp. 614–622. ISSN: 0002-9505, 1943-2909. DOI: `10.1119/1.3119512`. arXiv: `0806.1147 [physics]`. (Visited on 07/13/2023).

Zupanovic, Pasko and Domagoj Kuic (Apr. 2018). “Relation between Boltzmann and Gibbs Entropy and Example with Multinomial Distribution”. In: *Journal of Physics Communications* 2.4, p. 045002. ISSN: 2399-6528. DOI: `10.1088/2399-6528/aab7e1`. (Visited on 08/15/2023).