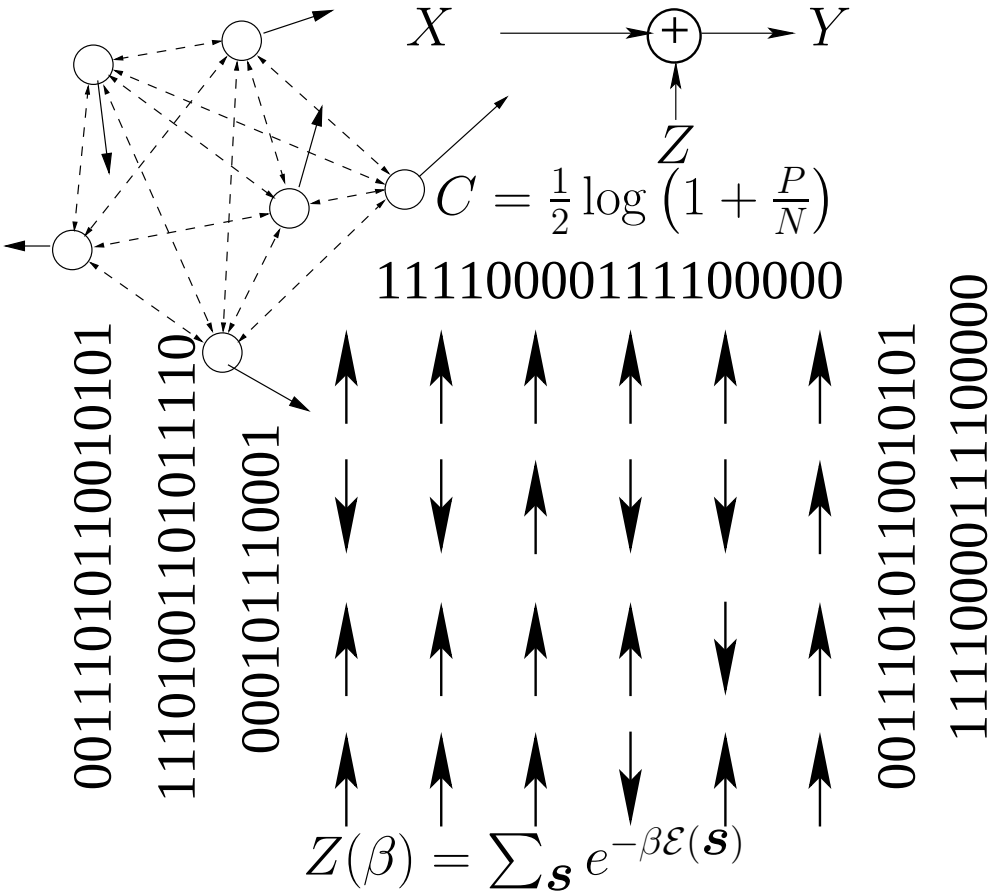
Statistical Physics and Information Theory

Neri Merhav



Statistical Physics and Information Theory

Neri Merhav
Department of Electrical Engineering
Technion - Israel Institute of Technology
Haifa 32000, ISRAEL
merhav@ee.technion.ac.il

Abstract

This is a set of lecture notes for a graduate course, which focuses on the relations between Information Theory and Statistical Physics. The course is aimed at EE graduate students in the area of Communications and Information Theory, as well as to graduate students in Physics who have basic background in Information Theory. Strong emphasis is given to the analogy and parallelism between Information Theory and Statistical Physics, as well as to the insights, the analysis tools and techniques that can be borrowed from Statistical Physics and 'imported' to certain problem areas in Information Theory. This is a research trend that has been very active in the last few decades, and the hope is that by exposing the student to the meeting points between these two disciplines, we will enhance his/her background and perspective to carry out research in the field.

Contents

1 Introduction

This work focuses on some of the relationships and the interplay between information theory and statistical physics – a branch of physics that deals with many–particle systems using probabilistic and statistical methods in the microscopic level.

The relationships between information theory and statistical thermodynamics are by no means new, and many researchers have been exploiting them for many years. Perhaps the first relation, or analogy, that crosses one's mind is that in both fields there is a fundamental notion of *entropy*. Actually, in information theory, the term entropy was coined in the footsteps of the thermodynamic entropy. The thermodynamic entropy was first introduced by Clausius in 1850, and its probabilistic–statistical interpretation was established by Boltzmann in 1872. It is virtually impossible to miss the functional resemblance between the two notions of entropy, and indeed it was recognized by Shannon and von Neumann. The well–known anecdote on this tells that von Neumann advised Shannon to adopt this term because it would provide him with *"... a great edge in debates because nobody really knows what entropy is anyway."*

But the relationships between the two fields go far beyond the fact that both share the notion of entropy. In fact, these relationships have many aspects. We will not cover all of them in this work, but just to taste the flavor of their scope, we will mention just a few.

The maximum entropy (ME) principle. This is perhaps the oldest concept that ties the two fields and it has attracted a great deal of attention, not only of information theorists, but also that of researchers in related fields like signal processing and image processing. The ME principle evolves around a philosophy, or a belief, which, in a nutshell, is the following: If in a certain problem, the observed data comes from an unknown probability distribution, but we do have some knowledge (that stems, e.g., from measurements) of certain moments of the underlying quantity/signal/random–variable, then assume that the unknown underlying probability distribution is the one with *maximum entropy* subject to (s.t.) moment constraints corresponding to this knowledge. For example, if we know the first and the

second moment, then the ME distribution is Gaussian with matching first and second order moments. Indeed, the Gaussian model is perhaps the most common model for physical processes in information theory as well as in signal– and image processing. But why maximum entropy? The answer to this philosophical question is rooted in the *second law of thermodynamics*, which asserts that in an isolated system, the entropy cannot decrease, and hence, when the system reaches thermal equilibrium, its entropy reaches its maximum. Of course, when it comes to problems in information theory and other related fields, this principle becomes quite heuristic, and so, one may question its justification, but nevertheless, this approach has had an enormous impact on research trends throughout the last fifty years, after being proposed by Jaynes in the late fifties of the previous century [44],[45], and further advocated by Shore and Johnson afterwards [109]. In the book by Cover and Thomas [13, Chapter 12], there is a good exposition on this topic. We will not put much emphasis on the ME principle in this work.

Landauer's erasure principle. Another aspect of these relations has to do with a theory whose underlying guiding principle is that information is a physical entity. Specifically, Landauer's erasure principle [62] (see also [6]), which is based on this physical theory of information, asserts that every bit that one erases, increases the entropy of the universe by $k \ln 2$, where k is Boltzmann's constant. The more comprehensive picture behind Landauer's principle, is that "any logically irreversible manipulation of information, such as the erasure of a bit or the merging of two computation paths, must be accompanied by a corresponding entropy increase in non-information bearing degrees of freedom of the information processing apparatus or its environment." (see [6]). This means that each lost information bit leads to the release of an amount $kT \ln 2$ of heat. By contrast, if no information is erased, computation may, in principle, be achieved in a way which is thermodynamically a reversible process, and hence requires no release of heat. This has had a considerable impact on the study of reversible computing. Landauer's principle is commonly accepted as a law of physics. However, there has also been some considerable dispute among physicists on this. This topic is not going to be included either in this work.

Large deviations theory as a bridge between information theory and statistical physics. Both information theory and statistical physics have an intimate relation to large deviations theory, a branch of probability theory which focuses on the assessment of the exponential rates of decay of probabilities of rare events, where one of the most elementary mathematical tools is the Legendre transform, which stands at the basis of the Chernoff bound. This topic will be covered quite thoroughly, mostly in Section 3.2.

Random matrix theory. How do the eigenvalues (or, more generally, the singular values) of random matrices behave when these matrices have very large dimensions or if they result from products of many randomly selected matrices? This is a very active area in probability theory with many applications, both in statistical physics and information theory, especially in modern theories of wireless communication (e.g., MIMO systems). This is again outside the scope of this course, but the interested reader is referred to [117] for a comprehensive introduction on the subject.

Spin glasses and coding theory. As was first observed by Sourlas [111] (see also [112]) and further advocated by many others, it turns out that many problems in channel coding theory (and also to some extent, source coding theory) can be mapped almost verbatim to parallel problems in the field of physics of *spin glasses* – amorphic magnetic materials with a high degree of disorder and very complicated physical behavior, which is customarily treated using statistical–mechanical approaches. It has been many years that researchers have made attempts to 'import' analysis techniques rooted in statistical physics of spin glasses and to apply them to analogous coding problems, with various degrees of success. This is one of main subjects of this course and we will study it extensively, at least from some aspects.

The above list of examples is by no means exhaustive. We could have gone much further and add many more examples of these very fascinating meeting points between information theory and statistical physics, but most of them will not be touched upon in this work. Many modern analyzes concerning multiuser situations, such as MIMO channels, CDMA, etc., and more recently, also in compressed sensing, are based on statistical–mechanical

techniques. But even if we limit ourselves to single–user communication systems, yet another very active problem area under this category is that of codes on graphs, iterative decoding, belief propagation, and density evolution. The main reason for not including it in this work is that it is already very well covered in recent textbooks, such as the one Mézard and Montanari [80] as well as the one by Richardson and Urbanke [100]. Another comprehensive exposition of graphical models, with a fairly strong statistical–mechanical flavor, was written by Wainwright and Jordan [120].

As will be seen, the physics and the information–theoretic subjects are interlaced with each other, rather than being given in two continuous, separate parts. This way, it is hoped that the relations between information theory and statistical physics will be made more apparent. We shall see that, not only these relations between information theory and statistical physics are interesting academically on their own right, but moreover, they also prove useful and beneficial in that they provide us with new insights and mathematical tools to deal with information–theoretic problems. These mathematical tools sometimes prove a lot more efficient than traditional tools used in information theory, and they may give either simpler expressions for performance analysis, or improved bounds, or both.

Having said that, a certain digression is in order. The reader should not expect to see too many real breakthroughs, which are allowed exclusively by statistical–mechanical methods, but could not have been achieved otherwise. Perhaps one exception to this rule is the replica method of statistical mechanics, which will be reviewed in this work, but not in great depth, because of two reasons: first, it is not rigorous (and so, any comparison to rigorous information–theoretic methods would not be fair), and secondly, because it is already very well covered in existing textbooks, such as [80] and [87]. If one cares about rigor, however, then there are no miracles. Everything, at the end of the day, boils down to mathematics. The point then is which culture, or scientific community, has developed the suitable mathematical techniques and what are the new insights that they provide; in many cases, it is the community of statistical physicists.

There are several examples of such techniques and insights, which are emphasized rather

strongly in this work. One example is the use of integrals in the complex plane and the saddle–point method. Among other things, this should be considered as a good substitute to the method of types, with the bonus of lending itself to extensions that include the countable and the continuous alphabet case (rather than just the finite alphabet case). Another example is the analysis technique of error exponents, which stems from the random energy model (see Chapter 6 and onward), along with its insights about phase transitions. Again, in retrospect, these analyzes are just mathematics and therefore could have been carried out without relying on any knowledge in physics. But it is nevertheless the physical point of view that provides the trigger for its use. Moreover, there are situations (see, e.g., Section 7.3), where results from statistical mechanics can be used almost verbatim in order to obtain stronger coding theorems. The point is then that it is not the physics itself that may be useful, it is the way in which physicists use mathematical tools.

One of the main take–home messages, that will hopefully remain with the reader after reading this work, is that whatever the field of statistical mechanics has to offer to us, as information theorists, goes much beyond the replica method. It is believed that this message is timely, because the vast majority of papers at the interface between the two disciplines are about applying the replica method to some information–theoretic problem.

The outline of the remaining part of this work is as follows: In Chapter 2, we give some elementary background in statistical physics and we relate fundamental thermodynamic potentials, like thermodynamical entropy and free energy with fundamental information measures, like the Shannon entropy and the Kullback–Leibler divergence. In Chapter 3, we explore a few aspects of physical interpretations of some fundamental results in information theory, like non–negativity of the Kullback–Leibler divergence, the data processing inequality, and the elementary coding theorems of information theory. In Chapter 4, we review some analysis tools commonly used in statistical physics, like the Laplace integration method, the saddle point method, and the replica method, all accompanied by examples. Chapter 5 is devoted to a (mostly descriptive) exposition of systems with interacting particles and phase transitions, both in physics and information theory. Chapter 6 focuses on one particular

model of a disordered physical system with interacting particles – the random energy model, which is highly relevant to the analysis of random code ensembles. Chapter 7 extends the random energy model in several directions, all relevant to problems in information theory. Finally, Chapter 8 contains a summary and an outlook on the interplay between information theory and statistical mechanics.

As with every paper published in *Foundations and Trends in Communications and Information Theory*, the reader is, of course, assumed to have some solid background in information theory. Concerning the physics part, prior background in statistical mechanics does not harm, but is not necessary. This work is intended to be self–contained as far as the physics background goes.

In a closing note, it is emphasized again that the coverage of topics, in this work, is by no means intended to be fully comprehensive, nor is it aimed at providing the complete plethora of problem areas, methods and results. The choice of topics, the approach, the flavor, and the style are nothing but the mirror image of the author's personal bias, perspective, and research interests in the field. Therefore, this work should actually be viewed mostly as a monograph, and not quite as a review or a tutorial paper. This is also the reason that a considerable part of the topics, covered in this work, are taken from articles in which the author has been involved.

2 Basic Background in Statistical Physics

In this chapter, we begin with some elementary background in statistical physics, and also relate some of the thermodynamic potentials, like entropy and free energy, to information measures, like the entropy and the Kullback–Leibler divergence.

2.1 What is Statistical Physics?

Statistical physics is a branch in physics which deals with systems with a huge number of particles (or any other elementary units). For example, *Avogadro's number*, which is about 6×10^{23}, is the number of molecules in 22.4 liters of ideal gas at standard temperature and pressure. Evidently, when it comes to systems with such an enormously large number of particles, there is no hope to keep track of the physical state (e.g., position and momentum) of each and every individual particle by means of the classical methods in physics, that is, by solving a gigantic system of differential equations pertaining to Newton's laws for all particles. Moreover, even if these differential equations could have been solved somehow (at least approximately), the information that they would give us would be virtually useless. What we normally really want to know about our physical system boils down to a fairly short list of *macroscopic* parameters, such as energy, heat, pressure, temperature, volume, magnetization, and the like. In other words, while we continue to believe in the good old laws of physics that we have known for some time, even the classical ones, we no longer use them in the ordinary way that we are familiar with from elementary physics courses. Instead, we think of the state of the system, at any given moment, as a realization of a certain *probabilistic ensemble*. This is to say that we approach the problem from a probabilistic (or a statistical) point of view. The beauty of statistical physics is that it derives the *macroscopic* theory of thermodynamics (i.e., the relationships between thermodynamical potentials, temperature, pressure, etc.) as *ensemble averages* that stem from this probabilistic *microscopic* theory, in the limit of an infinite number of particles, that is, the *thermodynamic limit*. As we shall see throughout this work, this thermodynamic limit is parallel to the asymptotic regimes that

we are used to in information theory, most notably, the one pertaining to a certain 'block length' that goes to infinity.

2.2 Basic Postulates and the Microcanonical Ensemble

For the sake of concreteness, let us consider the example where our many–particle system is a gas, namely, a system with a very large number N of mobile particles, which are free to move in a given volume. The *microscopic state* (or *microstate*, for short) of the system, at each time instant t, consists, in this example, of the position vector $\vec{r}_i(t)$ and the momentum vector $\vec{p}_i(t)$ of each and every particle, $1 \leq i \leq N$. Since each one of these is a vector of three components, the microstate is then given by a $(6N)$–dimensional vector $\vec{\boldsymbol{x}}(t) = \{(\vec{r}_i(t), \vec{p}_i(t)) : \ i = 1, 2, \ldots, N\}$, whose trajectory along the time axis, in the *phase space*, $\mathbb{R}^{6N}$, is called the *phase trajectory*.

Let us assume that the system is closed, i.e., *isolated* from its environment, in the sense that no energy flows inside or out. Imagine that the phase space $\mathbb{R}^{6N}$ is partitioned into very small hypercubes (or cells) $\Delta\vec{p} \times \Delta\vec{r}$. One of the basic postulates of statistical mechanics is the following: In the very long range, the relative amount of time which $\vec{\boldsymbol{x}}(t)$ spends at each such cell converges to a certain number between 0 and 1, which can be given the meaning of the *probability* of this cell. Thus, there is an underlying assumption of equivalence between temporal averages and ensemble averages, namely, this is the postulate of *ergodicity*. Considerable efforts were dedicated to the proof of the ergodic hypothesis at least in some cases (see e.g., [88], [110] and many references therein). As reasonable and natural as it may seem, the ergodic hypothesis should not be taken for granted. It does not hold for every system but only if no other conservation law holds. For example, the ideal gas in a box (to be discussed soon), is non–ergodic, as every particle retains its momentum (assuming perfectly elastic collisions with the walls).

What are then the probabilities of the above–mentioned phase–space cells? We would like to derive these probabilities from first principles, based on as few as possible basic postulates. Our second postulate is that for an isolated system (i.e., whose energy is fixed) all microscopic

states $\{\vec{\boldsymbol{x}}(t)\}$ are equiprobable. The rationale behind this postulate is twofold:

- In the absence of additional information, there is no apparent reason that certain regions in phase space would have preference relative to any others.
- This postulate is in harmony with a basic result in kinetic theory of gases – *the Liouville theorem*, which we will not touch upon in this work, but in a nutshell, it asserts that the phase trajectories must lie along hyper-surfaces of constant probability density.[1]

Before we proceed, let us slightly broaden the scope of our discussion. In a more general context, associated with our N–particle physical system, is a certain instantaneous microstate, generically denoted by $\boldsymbol{x} = (x_1, x_2, \ldots, x_N)$, where each x_i, $1 \leq i \leq N$, may itself be a vector of several physical quantities associated particle number i, e.g., its position, momentum, angular momentum, magnetic moment, spin, and so on, depending on the type and the nature of the physical system. For each possible value of $\boldsymbol{x}$, there is a certain *Hamiltonian* (i.e., energy function) that assigns to $\boldsymbol{x}$ a certain energy $\mathcal{E}(\boldsymbol{x})$.[2] Now, let us denote by $\Omega(E)$ the *density–of–states* function, i.e., the volume of the shell $\{\boldsymbol{x} : \mathcal{E}(\boldsymbol{x}) = E\}$, or, more precisely, $\Omega(E)\mathrm{d}E = \mathrm{Vol}\{\boldsymbol{x} : \ E \leq \mathcal{E}(\boldsymbol{x}) \leq E + \mathrm{d}E\}$, which will be denoted also as $\mathrm{Vol}\{\boldsymbol{x} : \ \mathcal{E}(\boldsymbol{x}) \approx E\}$, where the dependence on $\mathrm{d}E$ will normally be ignored since $\Omega(E)$ is typically exponential in N and $\mathrm{d}E$ will have virtually no effect on its exponential order as long as it is small. Then, our above postulate concerning the ensemble of an isolated system, which is called the *microcanonincal ensemble*, is that the probability density $P(\boldsymbol{x})$ is given by

$$P(\boldsymbol{x}) = \begin{cases} \frac{1}{\Omega(E)} & \mathcal{E}(\boldsymbol{x}) \approx E \\ 0 & \text{elsewhere} \end{cases} \tag{1}$$

In the discrete case, things are simpler, of course: Here, $\Omega(E)$ is the number of microstates with $\mathcal{E}(\boldsymbol{x}) = E$ (exactly) and $P(\boldsymbol{x})$ is the uniform probability mass function over this set of

[1] This is a result of the energy conservation law along with the fact that probability mass behaves like an incompressible fluid in the sense that whatever mass that flows into a certain region from some direction must be equal to the outgoing flow from some other direction. This is reflected in the so called continuity equation.

[2] For example, in the case of an *ideal gas*, $\mathcal{E}(\boldsymbol{x}) = \sum_{i=1}^{N} \|\vec{p}_i\|^2/(2m)$, where m is the mass of each molecule, namely, it accounts for the contribution of the kinetic energies only. In more complicated situations, there might be additional contributions of potential energy, which depend on the positions.

states. In this case, $\Omega(E)$ is analogous to the size of a *type class* in information theory [16], and $P(\boldsymbol{x})$ is the uniform distribution over this type class.

Back to the continuous case, note that $\Omega(E)$ is, in general, not dimensionless: In the above example of a gas, it has the physical units of $[\text{length} \times \text{momentum}]^{3N}$, but we must eliminate these physical units because very soon we are going to apply non–linear functions on $\Omega(E)$, like the logarithmic function. To this end, we normalize this volume by the volume of an elementary reference volume. In the gas example, this reference volume is taken to be h^{3N}, where h is *Planck's constant* ($h \approx 6.62 \times 10^{-34}$ Joules·sec). Informally, the intuition comes from the fact that h is our best available "resolution" in the plane spanned by each component of $\vec{r}_i$ and the corresponding component of $\vec{p}_i$, owing to the *uncertainty principle* in quantum mechanics, which tells that the product of the standard deviations $\Delta p_a \cdot \Delta r_a$ of each component a ($a = x, y, z$) is lower bounded by $\hbar/2$, where $\hbar = h/(2\pi)$. More formally, this reference volume is obtained in a natural manner from quantum statistical mechanics: by changing the integration variable $\vec{p}$ to $\vec{k}$ by using $\vec{p} = \hbar\vec{k}$, where $\vec{k}$ is the wave vector. This is a well–known relationship pertaining to particle–wave duality. Now, having redefined $\Omega(E)$ in units of this reference volume, which makes it then a dimensionless quantity, the *entropy* is defined as

$$S(E) = k \ln \Omega(E), \tag{2}$$

where k is *Boltzmann's constant* ($k \approx 1.38 \times 10^{-23}$ Joule/degree). We will soon see what is the relationship between $S(E)$ and the information–theoretic entropy, on the one hand, and what is the relationship between $S(E)$ and the classical thermodynamical entropy, due to Clausius. As it will turn out, all three are equivalent to one another.

To get some feeling of this, it should be noted that normally, $\Omega(E)$ behaves as an exponential function of N (at least asymptotically), and so, $S(E)$ is roughly linear in N. For example, if $\mathcal{E}(\boldsymbol{x}) = \sum_{i=1}^{N} \frac{\|\vec{p}_i\|^2}{2m}$, then $\Omega(E)$ is the volume of a shell or surface of a $(3N)$–dimensional sphere with radius $\sqrt{2mE}$, which is proportional to $(2mE)^{3N/2}V^N$, where V is the volume, but we should divide this by $N!$ to account for the fact that the particles are

indistinguishable and we do not count permutations as distinct physical states in this case.[3] More precisely, one obtains:

$$\begin{aligned} S(E) &= k \ln \left[\left(\frac{4\pi m E}{3N} \right)^{3N/2} \cdot \frac{V^N}{N! h^{3N}} \right] + \frac{3}{2} N k \\ &\approx N k \ln \left[\left(\frac{4\pi m E}{3N} \right)^{3/2} \cdot \frac{V}{N h^3} \right] + \frac{5}{2} N k. \end{aligned} \tag{3}$$

Assuming that E and V are both proportional to N, it is readily seen that $S(E)$ is also proportional to N. A physical quantity that has a linear dependence on the size of the system N, is called an *extensive quantity*. Energy, volume and entropy are then extensive quantities. Other quantities, which are not extensive, i.e., independent of the system size, like temperature and pressure, are called *intensive*.

It is interesting to point out that from the function $S(E)$, or actually, the function $S(E, V, N)$, one can obtain the entire information about the relevant macroscopic physical quantities of the system, e.g., temperature, pressure, and so on. Specifically, the *temperature* T of the system is defined according to:

$$\frac{1}{T} = \left[\frac{\partial S(E)}{\partial E} \right]_V \tag{4}$$

where $[\cdot]_V$ means that the derivative is taken in constant volume. One may wonder, at this point, what is the justification for *defining* temperature this way. We will get back to this point a bit later, but for now, let us see that this is indeed true at least for the ideal gas, as by taking the derivative of (3) w.r.t. E, we get

$$\frac{\partial S(E)}{\partial E} = \frac{3Nk}{2E}, \tag{5}$$

but in the case of the ideal gas, one can readily derive (based on the equation of state, $PV = NkT$) a simple expression of the energy E, which depends only on T (see, for example, [103, Sect. 20–4, pp. 353–355]):

$$E = \frac{3}{2} \cdot PV = \frac{3NkT}{2}, \tag{6}$$

[3] Since the particles are mobile and since they have no colors and no identity certificates, there is no distinction between a state where particle no. 15 has position $\vec{r}$ and momentum $\vec{p}$ while particle no. 437 has position $\vec{r}'$ and momentum $\vec{p}'$ and a state where these two particles are swapped.

which when plugged back into (5), gives immediately $1/T$.[4]

Intuitively, in most situations, we expect that $S(E)$ would be an increasing function of E (although this is not strictly always the case), which means $T \geq 0$. But T is also expected to be increasing with E (or equivalently, E is increasing with T, as otherwise, the heat capacity $\mathrm{d}E/\mathrm{d}T < 0$). Thus, $1/T$ should decrease with E, which means that the increase of S in E slows down as E grows. In other words, we expect $S(E)$ to be a concave function of E. In the above example, indeed, $S(E)$ is logarithmic in E and $E = 3NkT/2$, as we have seen.

How can we convince ourselves, in mathematical terms, that under "conceivable conditions", $S(E)$ is a concave function? We know that the Shannon entropy is also a concave functional of the probability distribution. Is this related? The answer may be given by a simple superadditivity argument: As both E and S are extensive quantities, let us define $E = N\epsilon$ and

$$s(\epsilon) = \lim_{N\to\infty} \frac{S(N\epsilon)}{N}, \tag{7}$$

i.e., the per–particle entropy as a function of the per–particle energy, where we assume that the limit exists (see, e.g., [108]). Consider the case where the Hamiltonian is additive, i.e.,

$$\mathcal{E}(\boldsymbol{x}) = \sum_{i=1}^{N} \mathcal{E}(x_i) \tag{8}$$

just like in the above example where $\mathcal{E}(\boldsymbol{x}) = \sum_{i=1}^{N} \frac{\|\vec{p}_i\|^2}{2m}$. Then, the inequality

$$\Omega(N_1\epsilon_1 + N_2\epsilon_2) \geq \Omega(N_1\epsilon_1) \cdot \Omega(N_2\epsilon_2), \tag{9}$$

expresses the simple fact that if our system is partitioned into two parts, one with N_1 particles, and the other with $N_2 = N - N_1$ particles, then every combination of individual microstates with energies $N_1\epsilon_1$ and $N_2\epsilon_2$ corresponds to a combined microstate with a total energy of $N_1\epsilon_1 + N_2\epsilon_2$ (but there are more ways to split this total energy between the two

[4] In fact, the above–mentioned simple derivation leads to the relation $PV = \frac{2}{3}E$. Now, two points of view are possible: The first is to accept the equation of state, $PV = NkT$, as an empirical experimental fact, and then deduce the relation $E = \frac{3}{2}NkT$ from $NkT = PV = 2E/3$. The second point of view is to define temperature as $T = 2E/(3Nk) = 2\epsilon/(3k)$ (i.e., as a measure of the average kinetic energy ϵ of each particle) and then deduce the equation of state from $PV = 2E/3$.

parts). Thus,

$$\begin{aligned}\frac{k\ln\Omega(N_1\epsilon_1+N_2\epsilon_2)}{N_1+N_2} &\geq \frac{k\ln\Omega(N_1\epsilon_1)}{N_1+N_2}+\frac{k\ln\Omega(N_2\epsilon_2)}{N_1+N_2}\\ &= \frac{N_1}{N_1+N_2}\cdot\frac{k\ln\Omega(N_1\epsilon_1)}{N_1}+\\ &\quad\frac{N_2}{N_1+N_2}\cdot\frac{k\ln\Omega(N_2\epsilon_2)}{N_2}.\end{aligned} \tag{10}$$

and so, by taking N_1 and N_2 to ∞, with $N_1/(N_1+N_2)\to\lambda\in(0,1)$, we get:

$$s(\lambda\epsilon_1+(1-\lambda)\epsilon_2)\geq\lambda s(\epsilon_1)+(1-\lambda)s(\epsilon_2), \tag{11}$$

which establishes the concavity of $s(\cdot)$ at least in the case of an additive Hamiltonian, which means that the entropy of mixing two systems of particles is greater than the total entropy before they are mixed (the second law). A similar proof can be generalized to the case where $\mathcal{E}(\boldsymbol{x})$ includes also a limited degree of interactions (short range interactions), e.g., $\mathcal{E}(\boldsymbol{x})=\sum_{i=1}^{N}\mathcal{E}(x_i,x_{i+1})$, but this requires somewhat more caution. In general, however, concavity may no longer hold when there are long range interactions, e.g., where some terms of $\mathcal{E}(\boldsymbol{x})$ depend on a linear subset of particles. Simple examples can be found in [116].

Example – Schottky defects. In a certain crystal, the atoms are located in a lattice, and at any positive temperature there may be defects, where some of the atoms are dislocated (see Fig. 1). Assuming that defects are sparse enough, such that around each dislocated atom all neighbors are in place, the activation energy, ϵ_0, required for dislocation is fixed. Denoting the total number of atoms by N and the number of defected ones by n, the total energy is then $E=n\epsilon_0$, and so,

$$\Omega(E)=\left(\begin{array}{c}N\\n\end{array}\right)=\frac{N!}{n!(N-n)!}, \tag{12}$$

or, equivalently,

$$\begin{aligned}S(E)=k\ln\Omega(E)&=k\ln\left[\frac{N!}{n!(N-n)!}\right]\\ &\approx k[N\ln N-n\ln n-(N-n)\ln(N-n)]\end{aligned} \tag{13}$$

where in the last passage we have used the Sterling approximation. Thus,[5]

$$\frac{1}{T} = \frac{\partial S}{\partial E} = \frac{\mathrm{d}S}{\mathrm{d}n} \cdot \frac{\mathrm{d}n}{\mathrm{d}E} = \frac{1}{\epsilon_0} \cdot k \ln \frac{N-n}{n}, \tag{14}$$

which gives the number of defects as

$$n = \frac{N}{\exp(\epsilon_0/kT) + 1}. \tag{15}$$

At $T = 0$, there are no defects, but their number increases gradually with T, approximately

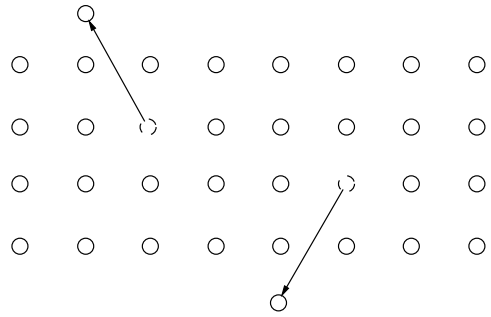


Figure 1: Schottky defects in a crystal lattice.

according to $\exp(-\epsilon_0/kT)$. Note that from a slightly more information–theoretic point of view,

$$\begin{aligned} S(E) &= k \ln \left(\begin{array}{c} N \\ n \end{array} \right) \\ &\approx kNh_2\left(\frac{n}{N}\right) \\ &= kNh_2\left(\frac{E}{N\epsilon_0}\right) \\ &= kNh_2\left(\frac{\epsilon}{\epsilon_0}\right), \end{aligned} \tag{16}$$

where

$$h_2(x) \stackrel{\Delta}{=} -x \ln x - (1-x)\ln(1-x).$$

[5]Here and in the sequel, the reader might wonder about the meaning of taking derivatives of, and with respect to, integer valued variables, like the number of dislocated particles, n. To this end, imagine an approximation where n is interpolated to be a continuous valued variable.

Thus, the thermodynamical entropy is intimately related to the Shannon entropy. We will see shortly that this is no coincidence. Note also that $s(\epsilon) = kh_2(\epsilon/\epsilon_0)$ is indeed concave in this example. $\square$

What happens if we have two independent systems with total energy E, which lie in equilibrium with each other? What is the temperature T and how does the energy split between them? The two systems exchange energy between one another in the form of heat (as no work is done). The number of combined microstates where system no. 1 has energy E_1 and system no. 2 has energy $E_2 = E - E_1$ is $\Omega_1(E_1) \cdot \Omega_2(E - E_1)$. If the combined system is isolated, then the probability of such a combined microstate is proportional to $\Omega_1(E_1) \cdot \Omega_2(E - E_1)$. Keeping in mind that normally, Ω_1 and Ω_2 are exponential in N, then for large N, this product is dominated by the value of E_1 for which it is maximum, or equivalently, the sum of logarithms, $S_1(E_1)+S_2(E-E_1)$, is maximum, i.e., it is a **maximum entropy** situation, which is **the second law of thermodynamics**. This maximum is normally achieved at the value of E_1 for which the derivative vanishes, i.e.,

$$S_1'(E_1) - S_2'(E - E_1) = 0 \tag{17}$$

or

$$S_1'(E_1) - S_2'(E_2) = 0 \tag{18}$$

which means

$$\frac{1}{T_1} \equiv S_1'(E_1) = S_2'(E_2) \equiv \frac{1}{T_2}. \tag{19}$$

Thus, in equilibrium, which is the maximum entropy situation, the energy splits in a way that temperatures are the same.

At this point, we are ready to justify why $S'(E)$ is equal to $1/T$ in general, as was promised earlier. Although it is natural to expect that equality between $S_1'(E_1)$ and $S_2'(E_2)$, in thermal equilibrium, is related equality between T_1 and T_2, this does not automatically mean that the derivative of each entropy is given by one over its temperature. On the face of it, for the purpose of this implication, this derivative could have been equal any one–to–one function of temperature $f(T)$. To see why $f(T) = 1/T$ indeed, imagine that we have a

system with an entropy function $S_0(E)$ and that we let it interact thermally with an ideal gas whose entropy function, which we shall denote now by $S_g(E)$, is given as in eq. (3). Now, at equilibrium $S_0'(E_0) = S_g'(E_g)$, but as we have seen already, $S_g'(E_g) = 1/T_g$, where T_g is the temperature of the ideal gas. But in thermal equilibrium the temperatures equalize, i.e., $T_g = T_0$, where T_0 is the temperature of the system of interest. It then follows eventually that $S_0'(E_0) = 1/T_0$, which now means that in equilibrium, the derivative of entropy of the system of interest is equal to the reciprocal of its temperature *in general*, and not only for the ideal gas! At this point, the fact that our system has interacted and equilibrated with an ideal gas is not important anymore and it does not limit the generality this statement. In simple words, our system does not 'care' what kind system it has interacted with, whether ideal gas or any other. This follows from a fundamental principle in thermodynamics, called the zero–th law, which states that thermal equilibrium has a transitive property: If system A is in equilibrium with system B and system B is in equilibrium with system C, then A is in equilibrium with C.

So we have seen that $\partial S/\partial E = 1/T$, or equivalently, $\mathrm{d}S = \mathrm{d}E/T$. But in the absence of any mechanical work applied to the system (fixed volume), $\mathrm{d}E = \mathrm{d}Q$, where Q is the heat intake. Thus, $\mathrm{d}S = \mathrm{d}Q/T$. But this is exactly the definition of the classical thermodynamical entropy due to Clausius. Thus, at least for the case where no mechanical work is involved, we have demonstrated the equivalence of the two notions of entropy, the statistical notion due to Boltzmann $S = k \ln \Omega$, and the thermodynamical entropy due to Clausius, $S = \int \mathrm{d}Q/T$. The more general proof, that allows mechanical work, has to take into account also partial derivative of S w.r.t. volume, which is related to pressure. We will not delve into this any further, but the line of thought is very similar.

2.3 The Canonical Ensemble

So far we have assumed that our system is isolated, and therefore has a strictly fixed energy E. Let us now relax this assumption and assume that our system is free to exchange energy with its large environment (heat bath) and that the total energy of the heat bath E_0 is by

far larger than the typical energy of the system. The combined system, composed of our original system plus the heat bath, is now an isolated system at temperature T.

Similarly as before, since the combined system is isolated, it is governed by the microcanonical ensemble. The only difference is that now we assume that one of the systems (the heat bath) is very large compared to the other (our test system). This means that if our small system is in microstate $\boldsymbol{x}$ (for whatever definition of the microstate vector) with energy $\mathcal{E}(\boldsymbol{x})$, then the heat bath must have energy $E_0 - \mathcal{E}(\boldsymbol{x})$ to complement the total energy to E_0. The number of ways that the heat bath may have energy $E_0 - \mathcal{E}(\boldsymbol{x})$ is $\Omega_B(E_0 - \mathcal{E}(\boldsymbol{x}))$, where $\Omega_B(\cdot)$ is the density–of–states function pertaining to the heat bath. In other words, the number of microstates of the *combined* system for which the small subsystem is in microstate $\boldsymbol{x}$ is $\Omega_B(E_0 - \mathcal{E}(\boldsymbol{x}))$. Since the combined system is governed by the microcanonical ensemble, the probability of this is proportional to $\Omega_B(E_0 - \mathcal{E}(\boldsymbol{x}))$. More precisely:

$$P(\boldsymbol{x}) = \frac{\Omega_B(E_0 - \mathcal{E}(\boldsymbol{x}))}{\sum_{\boldsymbol{x}'} \Omega_B(E_0 - \mathcal{E}(\boldsymbol{x}'))}. \tag{20}$$

Let us focus on the numerator for now, and normalize the result at the end. Then,

$$\begin{aligned}
P(\boldsymbol{x}) &\propto \Omega_B(E_0 - \mathcal{E}(\boldsymbol{x})) \\
&= \exp\{S_B(E_0 - \mathcal{E}(\boldsymbol{x}))/k\} \\
&\approx \exp\left\{\frac{S_B(E_0)}{k} - \frac{1}{k}\frac{\partial S_B(E)}{\partial E}\bigg|_{E=E_0} \cdot \mathcal{E}(\boldsymbol{x})\right\} \\
&= \exp\left\{\frac{S_B(E_0)}{k} - \frac{1}{kT} \cdot \mathcal{E}(\boldsymbol{x})\right\} \\
&\propto \exp\{-\mathcal{E}(\boldsymbol{x})/(kT)\}.
\end{aligned} \tag{21}$$

It is customary to work with the so called *inverse temperature*:

$$\beta = \frac{1}{kT} \tag{22}$$

and so,

$$P(\boldsymbol{x}) \propto e^{-\beta\mathcal{E}(\boldsymbol{x})}. \tag{23}$$

Thus, all that remains to do is to normalize, and we then obtain the *Boltzmann–Gibbs* (B–G) distribution, or the *canonical ensemble*, which describes the underlying probability law in equilibrium:

$$P(\boldsymbol{x}) = \frac{\exp\{-\beta\mathcal{E}(\boldsymbol{x})\}}{Z(\beta)} \tag{24}$$

where $Z(\beta)$ is the normalization factor:

$$Z(\beta) = \sum_{\boldsymbol{x}} \exp\{-\beta\mathcal{E}(\boldsymbol{x})\} \tag{25}$$

in the discrete case, or

$$Z(\beta) = \int \mathrm{d}\boldsymbol{x} \exp\{-\beta\mathcal{E}(\boldsymbol{x})\} \tag{26}$$

in the continuous case.

This is one of the most fundamental results in statistical mechanics, which was obtained solely from the energy conservation law and the postulate that in an isolated system the distribution is uniform. The function $Z(\beta)$ is called the *partition function*, and as we shall see, its meaning is by far deeper than just being a normalization constant. Interestingly, a great deal of the macroscopic physical quantities, like the internal energy, the free energy, the entropy, the heat capacity, the pressure, etc., can be obtained from the partition function. This is in analogy to the fact that in the microcanonical ensemble, $S(E)$ (or, more generally, $S(E, V, N)$) was pivotal to the derivation of all macroscopic physical quantities of interest.

The B–G distribution tells us then that the system "prefers" to visit its low energy states more than the high energy states, and what counts is only energy differences, not absolute energies: If we add to all states a fixed amount of energy E_0, this will result in an extra factor of $e^{-\beta E_0}$ both in the numerator and in the denominator of the B–G distribution, which will, of course, cancel out. Another obvious observation is that when the Hamiltonian is additive, that is, $\mathcal{E}(\boldsymbol{x}) = \sum_{i=1}^{N} \mathcal{E}(x_i)$, the various particles are statistically independent: Additive Hamiltonians correspond to non–interacting particles. In other words, the $\{x_i\}$'s behave as if they were drawn from a memoryless source. By the law of large numbers $\frac{1}{N}\sum_{i=1}^{N} \mathcal{E}(x_i)$ will tend (almost surely) to $\epsilon = \boldsymbol{E}\{\mathcal{E}(X_i)\}$. Nonetheless, this is different from

the microcanonical ensemble where $\frac{1}{N}\sum_{i=1}^{N}\mathcal{E}(x_i)$ was held strictly at the value of ϵ. The parallelism to well known concepts in information theory is quite clear: The microcanonical ensemble is parallel to the uniform distribution over a type class and the canonical ensemble is parallel to a memoryless system (source or channel).

The two ensembles are asymptotically equivalent as far as expectations go. They continue to be such even in cases of interactions, as long as these are short range. It is instructive to point out that the B–G distribution could have been obtained also in a different manner, owing to the maximum–entropy principle that we mentioned in the Introduction. Specifically, consider the following optimization problem:

$$\begin{aligned} &\max \ H(\boldsymbol{X}) \\ &\text{s.t. } \langle\mathcal{E}(\boldsymbol{X})\rangle = E \end{aligned} \tag{27}$$

where here and throughout the sequel, the operator $\langle\cdot\rangle$ designates expectation. This notation, which is customary in the physics literature, will be used interchangeably with the notation $\boldsymbol{E}\{\cdot\}$, which is more common in other scientific communities. By formalizing the equivalent Lagrange problem, where β now plays the role of a Lagrange multiplier:

$$\max \ \left\{H(\boldsymbol{X}) + \beta\left[E - \sum_{\boldsymbol{x}} P(\boldsymbol{x})\mathcal{E}(\boldsymbol{x})\right]\right\}, \tag{28}$$

or equivalently,

$$\min \ \left\{\sum_{\boldsymbol{x}} P(\boldsymbol{x})\mathcal{E}(\boldsymbol{x}) - \frac{H(\boldsymbol{X})}{\beta}\right\} \tag{29}$$

one readily verifies that the solution to this problem is the B-G distribution where the choice of β **controls** the average energy E. In many physical systems, the Hamiltonian is a quadratic (or "harmonic") function, e.g., $\frac{1}{2}mv^2$, $\frac{1}{2}kx^2$, $\frac{1}{2}CV^2$, $\frac{1}{2}LI^2$, $\frac{1}{2}I\omega^2$, etc., in which case the resulting B–G distribution turns out to be Gaussian. This is at least part of the explanation why the Gaussian distribution is so frequently encountered in Nature. Indeed, the Gaussian density is well known (see, e.g., [13, p. 411, Example 12.2.1]) to maximize the differential entropy subject to a second order moment constraint, which is equivalent to our average energy constraint.

2.4 Properties of the Partition Function and the Free Energy

Let us now examine more closely the partition function and make a few observations about its basic properties. For simplicity, we shall assume that $\boldsymbol{x}$ is discrete. First, let's look at the limits: Obviously, $Z(0)$ is equal to the size of the entire set of microstates, which is also $\sum_E \Omega(E)$, This is the high temperature limit, where all microstates are equiprobable. At the other extreme, we have:

$$\lim_{\beta\to\infty} \frac{\ln Z(\beta)}{\beta} = -\min_{\boldsymbol{x}} \mathcal{E}(\boldsymbol{x}) \stackrel{\Delta}{=} -E_{GS} \tag{30}$$

which describes the situation where the system is frozen to the absolute zero. Only states with minimum energy – the *ground–state energy*, prevail.

Another important property of $Z(\beta)$, or more precisely, of $\ln Z(\beta)$, is that it is a cumulant generating function: By taking derivatives of $\ln Z(\beta)$, we can obtain cumulants of $\mathcal{E}(\boldsymbol{X})$. For the first cumulant, we have

$$\boldsymbol{E}\{\mathcal{E}(\boldsymbol{X})\} \equiv \langle \mathcal{E}(\boldsymbol{X})\rangle = \frac{\sum_{\boldsymbol{x}} \mathcal{E}(\boldsymbol{x}) e^{-\beta\mathcal{E}(\boldsymbol{x})}}{\sum_{\boldsymbol{x}} e^{-\beta\mathcal{E}(\boldsymbol{x})}} = -\frac{\mathrm{d}\ln Z(\beta)}{\mathrm{d}\beta}. \tag{31}$$

Similarly, it is easy to show that

$$\mathrm{Var}\{\mathcal{E}(\boldsymbol{X})\} = \langle \mathcal{E}^2(\boldsymbol{X})\rangle - \langle \mathcal{E}(\boldsymbol{X})\rangle^2 = \frac{\mathrm{d}^2\ln Z(\beta)}{\mathrm{d}\beta^2}. \tag{32}$$

This in turn implies that

$$\frac{\mathrm{d}^2\ln Z(\beta)}{\mathrm{d}\beta^2} \geq 0, \tag{33}$$

which means that $\ln Z(\beta)$ must always be a convex function. Note also that

$$\begin{aligned} \frac{\mathrm{d}^2\ln Z(\beta)}{\mathrm{d}\beta^2} &= -\frac{\mathrm{d}\langle \mathcal{E}(\boldsymbol{X})\rangle}{\mathrm{d}\beta} \\ &= -\frac{\mathrm{d}\langle \mathcal{E}(\boldsymbol{X})\rangle}{\mathrm{d}T}\cdot\frac{\mathrm{d}T}{\mathrm{d}\beta} \\ &= kT^2C(T) \end{aligned} \tag{34}$$

where $C(T) = \mathrm{d}\langle \mathcal{E}(\boldsymbol{X})\rangle/\mathrm{d}T$ is the heat capacity (at constant volume). Thus, the convexity of $\ln Z(\beta)$ is intimately related to the physical fact that the heat capacity of the system is positive.

Next, we look at the function $Z(\beta)$ slightly differently. Instead of summing the terms $\{e^{-\beta\mathcal{E}(\boldsymbol{x})}\}$ over all states individually, we sum them by energy levels, in a collective manner, similarly as in the method of types [16]. This amounts to:

$$\begin{aligned}
Z(\beta) &= \sum_{\boldsymbol{x}} e^{-\beta\mathcal{E}(\boldsymbol{x})} \\
&= \sum_{E} \Omega(E) e^{-\beta E} \\
&\approx \sum_{\epsilon} e^{Ns(\epsilon)/k} \cdot e^{-\beta N\epsilon} \\
&= \sum_{\epsilon} \exp\{-N\beta[\epsilon - Ts(\epsilon)]\} \\
&\dot{=} \max_{\epsilon} \exp\{-N\beta[\epsilon - Ts(\epsilon)]\} \\
&= \exp\{-N\beta \min_{\epsilon}[\epsilon - Ts(\epsilon)]\} \\
&\stackrel{\Delta}{=} \exp\{-N\beta[\epsilon^* - Ts(\epsilon^*)]\} \\
&\stackrel{\Delta}{=} e^{-\beta F},
\end{aligned} \tag{35}$$

where here and throughout the sequel, the notation $\dot{=}$ means asymptotic equivalence in the exponential scale. More precisely, $a_N \dot{=} b_N$ for two positive sequences $\{a_N\}$ and $\{b_N\}$, means that $\lim_{N\to\infty} \frac{1}{N} \ln(a_N/b_N) = 0$.

The quantity $f \stackrel{\Delta}{=} \epsilon - Ts(\epsilon)$ is the (per–particle) *free energy*. Similarly, the entire free energy, F, is defined as

$$F = E - TS = -\frac{\ln Z(\beta)}{\beta}. \tag{36}$$

The physical meaning of the free energy, or more precisely, the difference between two free energies F_1 and F_2, is the minimum amount of work that it takes to transfer the system from equilibrium state 1 to another equilibrium state 2 in an isothermal (fixed temperature) process. This minimum is achieved when the process is *reversible*, i.e., so slow that the system is always almost in equilibrium. Equivalently, $-\Delta F$ is the maximum amount energy in the system, that is *free* and useful for performing work (i.e., not dissipated as heat) in fixed temperature. Again, this maximum is attained by a reversible process.

To demonstrate this point, let us consider the case where $\mathcal{E}(\boldsymbol{x})$ includes a term of a

potential energy that is given by the (scalar) product of a certain external force and the conjugate physical variable at which this force is exerted (e.g., pressure times volume, gravitational force times height, moment times angle, magnetic field times magnetic moment, electric field times charge, etc.), i.e.,

$$\mathcal{E}(\boldsymbol{x}) = \mathcal{E}_0(\boldsymbol{x}) - \lambda \cdot L(\boldsymbol{x}) \tag{37}$$

where λ is the force and $L(\boldsymbol{x})$ is the conjugate physical variable, which depends on (some coordinates of) the microstate. The partition function then depends on both β and λ and hence will be denoted $Z(\beta, \lambda)$. It is easy to see (similarly as before) that $\ln Z(\beta, \lambda)$ is convex in λ for fixed β. Also,

$$\langle L(\boldsymbol{X}) \rangle = kT \cdot \frac{\partial \ln Z(\beta, \lambda)}{\partial \lambda}. \tag{38}$$

The free energy is given by[6]

$$\begin{aligned} F &= E - TS \\ &= -kT \ln Z + \lambda \langle L(\boldsymbol{X}) \rangle \\ &= kT \left(\lambda \cdot \frac{\partial \ln Z}{\partial \lambda} - \ln Z \right). \end{aligned} \tag{39}$$

Now, let F_1 and F_2 be the equilibrium free energies pertaining to two values of λ, denoted λ_1 and λ_2. Then,

$$\begin{aligned} F_2 - F_1 &= \int_{\lambda_1}^{\lambda_2} \mathrm{d}\lambda \cdot \frac{\partial F}{\partial \lambda} \\ &= kT \cdot \int_{\lambda_1}^{\lambda_2} \mathrm{d}\lambda \cdot \lambda \cdot \frac{\partial^2 \ln Z}{\partial \lambda^2} \\ &= \int_{\lambda_1}^{\lambda_2} \mathrm{d}\lambda \cdot \lambda \cdot \frac{\partial \langle L(\boldsymbol{X}) \rangle}{\partial \lambda} \end{aligned}$$

[6] At this point, there is a distinction between the *Helmholtz free energy* and the *Gibbs free energy*. The former is defined as $F = E - TS$ in general, as mentioned earlier. The latter is defined as $G = E - TS - \lambda L = -kT \ln Z$, where L is shorthand notation for $\langle L(\boldsymbol{X}) \rangle$. The physical significance of the Gibbs free energy is similar to that of the Helmholtz free energy, except that it refers to the total work of all other external forces in the system (if there are any), except the work contributed by the force λ. The passage to the Gibbs ensemble, which replaces a fixed value of $L(\boldsymbol{x})$ (say, constant volume of a gas) by the control of the conjugate external force λ, (say, pressure in the example of a gas) can be carried out by another Legendre transform (see, e.g., [60, Sect. 1.14]).

$$= \int_{\langle L(X)\rangle_{\lambda_1}}^{\langle L(X)\rangle_{\lambda_2}} \lambda \cdot \mathrm{d}\,\langle L(\boldsymbol{X})\rangle \tag{40}$$

The product $\lambda \cdot \mathrm{d}\,\langle L(\boldsymbol{X})\rangle$ designates an infinitesimal amount of (average) work performed by the force λ on a small change in the average of the conjugate variable $\langle L(\boldsymbol{X})\rangle$, where the expectation is taken w.r.t. the actual value of λ. Thus, the last integral expresses the total work along a slow process of changing the force λ in small steps and letting the system adapt and equilibrate after this small change every time. On the other hand, it is easy to show (using the convexity of $\ln Z$ in λ), that if λ varies in large steps, the resulting amount of work will always be larger.

Returning to the definition of f, we see that the value ϵ^* of ϵ that minimizes f, dominates the partition function and hence captures most of the probability. As N grows without bound, the energy probability distribution becomes sharper and sharper around $N\epsilon^*$. Thus, we see that equilibrium in the canonical ensemble amounts to **minimum free energy.** This extends the second law of thermodynamics from the microcanonical ensemble of isolated systems, whose equilibrium obeys the maximum entropy principle. The maximum entropy principle is replaced, more generally, by the minimum free energy principle. Note that the Lagrange minimization problem that we formalized before, i.e.,

$$\min \left\{ \sum_{\boldsymbol{x}} P(\boldsymbol{x})\mathcal{E}(\boldsymbol{x}) - \frac{H(\boldsymbol{X})}{\beta} \right\}, \tag{41}$$

is nothing but minimization of the free energy, provided that we identify H with the physical entropy S (to be done soon) and the Lagrange multiplier $1/\beta$ with kT. Thus, the B–G distribution minimizes the free energy for a given temperature.

We have not yet seen this explicitly, but there were already hints (and terminology suggests) that the thermodynamical entropy $S(E)$ is intimately related to the Shannon entropy $H(\boldsymbol{X})$. We will also see it shortly in a more formal manner. But what is the information–theoretic analogue of the (Helmholtz) free energy?

Here is a preliminary guess based on a very rough consideration: The last chain of equalities in eq. (35) reminds us what happens when we use the method of types and sum over prob-

abilities type–by–type in information–theoretic problems: The exponentials $\exp\{-\beta\mathcal{E}(\boldsymbol{x})\}$ are analogous (up to a normalization factor) to probabilities, which in the memoryless case, are given by $P(\boldsymbol{x}) = \exp\{-N[\hat{H} + D(\hat{P}\|P)]\}$, where $\hat{H}$ is the empirical entropy pertaining to $\boldsymbol{x}$ and $\hat{P}$ is the empirical distribution. Each such probability is weighted by the size of the type class, which as is known from the method of types, is exponentially $e^{N\hat{H}}$, and whose physical analogue is $\Omega(E) = e^{Ns(\epsilon)/k}$. The product gives $\exp\{-ND(\hat{P}\|P)\}$ in information theory and $\exp\{-N\beta f\}$ in statistical physics. This suggests that perhaps the free energy has some analogy with the divergence. We will see shortly a more rigorous argument that relates the Helmholtz free energy and the divergence. The Gibbs free energy can also be related to an informational divergence.

More formally, let us define

$$\phi(\beta) = \lim_{N\to\infty} \frac{\ln Z(\beta)}{N} \tag{42}$$

and, in order to avoid dragging the constant k, let us define

$$\Sigma(\epsilon) = \lim_{N\to\infty} \frac{\ln \Omega(N\epsilon)}{N} = \frac{s(\epsilon)}{k}. \tag{43}$$

Then, the chain of equalities (35), written slightly differently, gives

$$\begin{aligned}
\phi(\beta) &= \lim_{N\to\infty} \frac{\ln Z(\beta)}{N} \\
&= \lim_{N\to\infty} \frac{1}{N} \ln \left\{ \sum_{\epsilon} e^{N[\Sigma(\epsilon)-\beta\epsilon]} \right\} \\
&= \max_{\epsilon} [\Sigma(\epsilon) - \beta\epsilon].
\end{aligned}$$

Thus, $\phi(\beta)$ is (a certain variant of) the *Legendre transform*[7] of $\Sigma(\epsilon)$. As $\Sigma(\epsilon)$ is (normally) a concave function, then it can readily be shown that the inverse transform is:

$$\Sigma(\epsilon) = \min_{\beta} [\beta\epsilon + \phi(\beta)]. \tag{44}$$

[7]More precisely, the 1D Legendre transform of a real function $f(x)$ is defined as $g(y) = \sup_x [xy - f(x)]$. If f is convex, it can readily be shown that: (i) The inverse transform has the very same form, i.e., $f(x) = \sup_y [xy - g(y)]$, and (ii) The derivatives $f'(x)$ and $g'(y)$ are inverses of each other.

The achiever, $\epsilon^*(\beta)$, of $\phi(\beta)$ in the forward transform is obtained by equating the derivative to zero, i.e., it is the solution to the equation

$$\beta = \Sigma'(\epsilon), \tag{45}$$

where $\Sigma'(\epsilon)$ is the derivative of $\Sigma(\epsilon)$. In other words, $\epsilon^*(\beta)$ the inverse function of $\Sigma'(\cdot)$. By the same token, the achiever, $\beta^*(\epsilon)$, of $\Sigma(\epsilon)$ in the backward transform is obtained by equating the other derivative to zero, i.e., it is the solution to the equation

$$\epsilon = -\phi'(\beta) \tag{46}$$

or in other words, the inverse function of $-\phi'(\cdot)$.

This establishes a relationship between the typical per–particle energy ϵ and the inverse temperature β that gives rise to ϵ (cf. the Lagrange interpretation above, where we said that β controls the average energy). Now, observe that whenever β and ϵ are related as explained above, we have:

$$\Sigma(\epsilon) = \beta\epsilon + \phi(\beta) = \phi(\beta) - \beta \cdot \phi'(\beta). \tag{47}$$

On the other hand, if we look at the Shannon entropy pertaining to the B–G distribution, we get:

$$\begin{aligned}
\bar{H}(\boldsymbol{X}) &= \lim_{N\to\infty} \frac{1}{N} \boldsymbol{E}\left\{\ln \frac{Z(\beta)}{e^{-\beta\mathcal{E}(\boldsymbol{X})}}\right\} \\
&= \lim_{N\to\infty} \left[\frac{\ln Z(\beta)}{N} + \frac{\beta \boldsymbol{E}\{\mathcal{E}(\boldsymbol{X})\}}{N}\right] \\
&= \phi(\beta) - \beta \cdot \phi'(\beta).
\end{aligned}$$

which is exactly the same expression as in (47), and so, $\Sigma(\epsilon)$ and $\bar{H}$ are identical whenever β and ϵ are related accordingly. The former, as we recall, we defined as the normalized logarithm of the number of microstates with per–particle energy ϵ. Thus, we have learned that the number of such microstates with energy E is exponentially $e^{N\bar{H}}$, a result that is well known from the method of types [16], where combinatorial arguments for finite–alphabet sequences are used. Here we obtained the same result from substantially different considerations, which are applicable in situations far more general than those of finite alphabets

(continuous alphabets included). Another look at this relation is the following:

$$\begin{aligned} 1 \geq & \sum_{\boldsymbol{x}:\ \mathcal{E}(\boldsymbol{x})\approx N\epsilon} P(\boldsymbol{x}) = \sum_{\boldsymbol{x}:\ \mathcal{E}(\boldsymbol{x})\approx N\epsilon} \frac{\exp\{-\beta\sum_i \mathcal{E}(x_i)\}}{Z^N(\beta)} \\ & \dot{=} \sum_{\boldsymbol{x}:\ \mathcal{E}(\boldsymbol{x})\approx N\epsilon} \exp\{-\beta N\epsilon - N\phi(\beta)\} \\ & = \Omega(N\epsilon)\cdot\exp\{-N[\beta\epsilon+\phi(\beta)]\} \end{aligned} \tag{48}$$

which means that

$$\Omega(N\epsilon) \leq \exp\{N[\beta\epsilon+\phi(\beta)]\} \tag{49}$$

for all β, and so,

$$\Omega(N\epsilon) \leq \exp\{N\min_{\beta}[\beta\epsilon+\phi(\beta)]\} = e^{N\Sigma(\epsilon)} = e^{N\bar{H}}. \tag{50}$$

A compatible lower bound is obtained by observing that the minimizing β gives rise to $\langle\mathcal{E}(X_1)\rangle = \epsilon$, which makes the event $\{\boldsymbol{x}:\ \mathcal{E}(\boldsymbol{x}) \approx N\epsilon\}$ a high–probability event, by the weak law of large numbers. A good reference for further study, and from a more general perspective, is the article by Hall [38]. See also [30].

Note also that eq. (47), which we will rewrite, with a slight abuse of notation as

$$\phi(\beta) - \beta\phi'(\beta) = \Sigma(\beta) \tag{51}$$

can be viewed as a first order differential equation in $\phi(\beta)$, whose solution is easily found to be

$$\phi(\beta) = -\beta\epsilon_{GS} + \beta\cdot\int_{\beta}^{\infty}\frac{\mathrm{d}\hat{\beta}\Sigma(\hat{\beta})}{\hat{\beta}^2}, \tag{52}$$

where $\epsilon_{GS} = \lim_{N\to\infty} E_{GS}/N$. Equivalently,

$$Z(\beta) \dot{=} \exp\left\{-\beta E_{GS} + N\beta\cdot\int_{\beta}^{\infty}\frac{\mathrm{d}\hat{\beta}\Sigma(\hat{\beta})}{\hat{\beta}^2}\right\}, \tag{53}$$

namely, the partition function at a certain temperature can be expressed as a functional of the entropy pertaining to all temperatures lower than that temperature. Changing the integration variable from β to T, this readily gives the relation

$$F = E_{GS} - \int_0^T S(T')\mathrm{d}T'. \tag{54}$$

Since $F = E - ST$, we have

$$E = E_{GS} + ST - \int_0^T S(T')\mathrm{d}T' = E_{GS} + \int_0^S T(S')\mathrm{d}S', \tag{55}$$

where the second term amounts to the heat Q that accumulates in the system, as the temperature is raised from 0 to T. This is a special case of the first law of thermodynamics. The more general form takes into account also possible work performed on (or by) the system.

Having established the identity between the Shannon–theoretic entropy and the thermodynamical entropy, we now move on, as promised, to the free energy and seek its information–theoretic counterpart. More precisely, we will look at the difference between the free energies of two different probability distributions, one of which is the B–G distribution. Consider first, the following chain of equalities concerning the B–G distribution:

$$\begin{aligned} P(\boldsymbol{x}) &= \frac{\exp\{-\beta\mathcal{E}(\boldsymbol{x})\}}{Z(\beta)} \\ &= \exp\{-\ln Z(\beta) - \beta\mathcal{E}(\boldsymbol{x})\} \\ &= \exp\{\beta[F(\beta) - \mathcal{E}(\boldsymbol{x})]\}. \end{aligned} \tag{56}$$

Consider next another probability distribution Q, different in general from P, and hence corresponding to non–equilibrium. Let us now look at the divergence:

$$\begin{aligned} D(Q\|P) &= \sum_{\boldsymbol{x}} Q(\boldsymbol{x}) \ln \frac{Q(\boldsymbol{x})}{P(\boldsymbol{x})} \\ &= -H_Q - \sum_{\boldsymbol{x}} Q(\boldsymbol{x}) \ln P(\boldsymbol{x}) \\ &= -H_Q - \beta \sum_{\boldsymbol{x}} Q(\boldsymbol{x})[F_P - \mathcal{E}(\boldsymbol{x})] \\ &= -H_Q - \beta F_P + \beta\langle\mathcal{E}\rangle_Q \\ &= \beta(F_Q - F_P) \end{aligned}$$

or equivalently,

$$F_Q = F_P + kT \cdot D(Q\|P). \tag{57}$$

Thus, the free energy difference is indeed related to the the divergence. For a given temperature, the free energy away from equilibrium is always larger than the free energy at

equilibrium. Since the system "wants" to minimize the free energy, it eventually converges to the B–G distribution. More details on these relations can be found in [2] and [94].

There is room for criticism, however, about the last derivation: the thermodynamical entropy of the system is known to be given by the Shannon (or, more precisely, the Gibbs) entropy expression only in equilibrium, whereas for non–equilibrium, it is not clear whether it continues to be valid. In fact, there is dispute among physicists on how to define the entropy in non–equilibrium situations, and there are many variations on the theme (see, for example, [7],[14],[40]). Nonetheless, we will accept this derivation of the relation between the divergence and the free energy difference for our purposes.

Another interesting relation between the divergence and physical quantities is that the divergence is proportional to the dissipated work (=average work minus free–energy difference) between two equilibrium states at the same temperature but corresponding to two different values of some external control parameter (see, e.g., [56]). We will elaborate on this in Subsection 3.3.

Let us now summarize the main properties of the partition function that we have seen thus far:

1. $Z(\beta)$ is a continuous function. $Z(0) = |\mathcal{X}^n|$ and $\lim_{\beta\to\infty} \frac{\ln Z(\beta)}{\beta} = -E_{GS}$.
2. Generating cumulants: $\langle \mathcal{E}(\boldsymbol{X}) \rangle = -\mathrm{d} \ln Z/\mathrm{d}\beta$, $\mathrm{Var}\{\mathcal{E}(\boldsymbol{X})\} = \mathrm{d}^2 \ln Z/\mathrm{d}\beta^2$, which implies convexity of $\ln Z$, and hence also of $\phi(\beta)$.
3. ϕ and Σ are a Legendre–transform pair. Σ is concave.
4. $\Sigma(\epsilon)$ coincides with the Shannon entropy of the B-G distribution.
5. $F_Q = F_P + kT \cdot D(Q\|P)$.

Comment: Consider $Z(\beta)$ for an *imaginary temperature* $\beta = j\omega$, where $j = \sqrt{-1}$, and define $z(E)$ as the inverse Fourier transform of $Z(j\omega)$. It can readily be seen that $z(E) = \Omega(E)$ is the density of states, i.e., for $E_1 < E_2$, the number of states with energy between E_1

and E_2 is given by $\int_{E_1}^{E_2} z(E)\mathrm{d}E$. Thus, $Z(\cdot)$ can be related to energy enumeration in two different ways: one is by the Legendre transform of $\ln Z(\beta)$ for real β, and the other is by the inverse Fourier transform of $Z(\beta)$ for imaginary β. It should be kept in mind, however, that while the latter relation holds for every system size N, the former is true only in the thermodynamic limit, as mentioned. This double connection between $Z(\beta)$ and $\Omega(E)$ is no coincidence, as we shall see later in Chapter 4. The reader might wonder about the usefulness and the meaning of complex temperature. It turns out to be a very useful tool in examining the analyticity of the partition function (as a complex function of the complex temperature) in the vicinity of the real axis, which is related to phase transitions. This is the basis for the Yang–Lee theory [63],[125].

Example – A two level system. Similarly to the earlier example of Schottky defects, which was previously given in the context of the microcanonical ensemble, consider now a system of N independent particles, each having two possible states: state 0 of zero energy and state 1, whose energy is ϵ_0, i.e., $\mathcal{E}(x) = \epsilon_0 x$, $x \in \{0, 1\}$. The x_i's are independent, each having a marginal:

$$P(x) = \frac{e^{-\beta\epsilon_0 x}}{1 + e^{-\beta\epsilon_0}} \quad x \in \{0, 1\}. \tag{58}$$

In this case,

$$\phi(\beta) = \ln(1 + e^{-\beta\epsilon_0}) \tag{59}$$

and

$$\Sigma(\epsilon) = \min_{\beta \geq 0}[\beta\epsilon + \ln(1 + e^{-\beta\epsilon_0})]. \tag{60}$$

To find $\beta^*(\epsilon)$, we take the derivative and equate to zero:

$$\epsilon - \frac{\epsilon_0 e^{-\beta\epsilon_0}}{1 + e^{-\beta\epsilon_0}} = 0 \tag{61}$$

which gives

$$\beta^*(\epsilon) = \frac{\ln(\epsilon/\epsilon_0 - 1)}{\epsilon_0}. \tag{62}$$

On substituting this back into the above expression of $\Sigma(\epsilon)$, we get:

$$\Sigma(\epsilon) = \frac{\epsilon}{\epsilon_0} \ln\left(\frac{\epsilon}{\epsilon_0} - 1\right) + \ln\left[1 + \exp\left\{-\ln\left(\frac{\epsilon}{\epsilon_0} - 1\right)\right\}\right], \tag{63}$$

which after a short algebraic manipulation, becomes

$$\Sigma(\epsilon) = h_2\left(\frac{\epsilon}{\epsilon_0}\right), \tag{64}$$

just like in the Schottky example. In the other direction:

$$\phi(\beta) = \max_{\epsilon}\left[h_2\left(\frac{\epsilon}{\epsilon_0}\right) - \beta\epsilon\right], \tag{65}$$

whose achiever $\epsilon^*(\beta)$ solves the zero–derivative equation:

$$\frac{1}{\epsilon_0}\ln\left[\frac{1-\epsilon/\epsilon_0}{\epsilon/\epsilon_0}\right] = \beta \tag{66}$$

or equivalently,

$$\epsilon^*(\beta) = \frac{\epsilon_0}{1+e^{-\beta\epsilon_0}}, \tag{67}$$

which is exactly the inverse function of $\beta^*(\epsilon)$ above, and which when plugged back into the expression of $\phi(\beta)$, indeed gives

$$\phi(\beta) = \ln(1+e^{-\beta\epsilon_0}). \quad \square \tag{68}$$

Comment: A very similar model (and hence with similar results) pertains to non–interacting spins (magnetic moments), where the only difference is that $x \in \{-1,+1\}$ rather than $x \in \{0,1\}$. Here, the meaning of the parameter ϵ_0 becomes that of a magnetic field, which is more customarily denoted by B (or H), and which is either parallel or anti-parallel to that of the spin, and so the potential energy (in the appropriate physical units), $\vec{B}\cdot\vec{x}$, is either Bx or $-Bx$. Thus,

$$P(x) = \frac{e^{\beta Bx}}{2\cosh(\beta B)}; \quad Z(\beta) = 2\cosh(\beta B). \tag{69}$$

The net *magnetization* per–spin is defined as

$$m \stackrel{\Delta}{=} \left\langle \frac{1}{N}\sum_{i=1}^{N} X_i \right\rangle = \langle X_1\rangle = \frac{\partial\phi}{\partial(\beta B)} = \frac{\partial}{\partial(\beta B)}\ln[2\cosh(\beta B)] = \tanh(\beta B). \tag{70}$$

This is the paramagnetic characteristic of the magnetization as a function of the magnetic field: As $B \to \pm\infty$, the magnetization $m \to \pm 1$ accordingly. When the magnetic field is removed ($B = 0$), the magnetization vanishes too. We will get back to this model and its extensions in Chapter 5. $\square$

2.5 The Energy Equipartition Theorem

It turns out that in the case of a quadratic Hamiltonian, $\mathcal{E}(x) = \frac{1}{2}\alpha x^2$ (e.g., $\frac{1}{2}mv^2$, $\frac{1}{2}kx^2$, $\frac{1}{2}I\omega^2$, etc.), which means that x is Gaussian with variance kT/α, the average per–particle energy, is always given by

$$\left\langle \frac{1}{2}\alpha X_i^2 \right\rangle = \frac{1}{2}\alpha \cdot \frac{kT}{\alpha} = \frac{kT}{2}, \tag{71}$$

independently of α. If we have N such additive quadratic terms, then of course, we end up with $NkT/2$. In the case of the ideal gas, we have three such terms (one for each dimension) per particle, thus a total of $3N$ terms, and so, $E = 3NkT/2$, which is exactly the expression we obtained also from the microcanonical ensemble in eq. (6). In fact, we observe that in the canonical ensemble, whenever we have an Hamiltonian of the form $\frac{\alpha}{2}x_i^2$ plus some arbitrary terms that do not depend on x_i, then x_i is Gaussian (with variance kT/α) and independent of the other variables, i.e., $p(x_i) \propto e^{-\alpha x_i^2/(2kT)}$. Hence it contributes an amount of $kT/2$ to the total average energy, again, independently of α. It is more precise to refer to this x_i as a *degree of freedom* rather than a particle. This is because in the three–dimensional world, the kinetic energy, for example, is given by $p_x^2/(2m) + p_y^2/(2m) + p_z^2/(2m)$, that is, each particle contributes *three* additive quadratic terms rather than one (just like three independent one–dimensional particles) and so, it contributes $3kT/2$. This principle is called the *the energy equipartition theorem*. In Subsection 3.2, we will see that it is quite intimately related to rate–distortion theory for quadratic distortion measures.

Below is a direct derivation of the equipartition theorem:

$$\begin{aligned}
\left\langle \frac{1}{2}aX^2 \right\rangle &= \frac{\int_{-\infty}^{\infty} \mathrm{d}x(\alpha x^2/2)e^{-\beta\alpha x^2/2}}{\int_{-\infty}^{\infty} \mathrm{d}x e^{-\beta\alpha x^2/2)}} \\
&= -\frac{\partial}{\partial\beta}\ln\left[\int_{-\infty}^{\infty} \mathrm{d}x e^{-\beta\alpha x^2/2}\right] \\
&= -\frac{\partial}{\partial\beta}\ln\left[\frac{1}{\sqrt{\beta}}\int_{-\infty}^{\infty} \mathrm{d}(\sqrt{\beta}x) e^{-\alpha(\sqrt{\beta}x)^2/2}\right] \\
&= -\frac{\partial}{\partial\beta}\ln\left[\frac{1}{\sqrt{\beta}}\int_{-\infty}^{\infty} \mathrm{d}u e^{-\alpha u^2/2}\right] \\
&= \frac{1}{2}\frac{\mathrm{d}\ln\beta}{\mathrm{d}\beta} = \frac{1}{2\beta} = \frac{kT}{2}.
\end{aligned}$$

Note that although we could have used closed–form expressions for both the numerator and the denominator of the first line, we have deliberately taken a somewhat different route in the second line, where we have presented it as the derivative of the denominator of the first line. Also, rather than calculating the Gaussian integral explicitly, we only figured out how it scales with β, because this is the only thing that matters after taking the derivative relative to β. The reason for using this trick of bypassing the need to calculate integrals, is that it can easily be extended in two directions at least:

1. Let $\boldsymbol{x} \in \mathbb{R}^N$ and let $\mathcal{E}(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T A \boldsymbol{x}$, where A is a $N \times N$ positive definite matrix. This corresponds to a physical system with a quadratic Hamiltonian, which includes also interactions between pairs (e.g., Harmonic oscillators or springs, which are coupled because they are tied to one another). It turns out that here, regardless of A, we get:

$$\langle \mathcal{E}(\boldsymbol{X}) \rangle = \left\langle \frac{1}{2}\boldsymbol{X}^T A \boldsymbol{X} \right\rangle = N \cdot \frac{kT}{2}. \tag{72}$$

2. Back to the case of a scalar x, but suppose now a more general power–law Hamiltonian, $\mathcal{E}(x) = \alpha|x|^\theta$. In this case, we get

$$\langle \mathcal{E}(X) \rangle = \left\langle \alpha|X|^\theta \right\rangle = \frac{kT}{\theta}. \tag{73}$$

Moreover, if $\lim_{x\to\pm\infty} xe^{-\beta\mathcal{E}(x)} = 0$ for all $\beta > 0$, and we denote $\mathcal{E}'(x) \stackrel{\Delta}{=} \mathrm{d}\mathcal{E}(x)/\mathrm{d}x$, then

$$\langle X \cdot \mathcal{E}'(X) \rangle = kT. \tag{74}$$

It is easy to see that the earlier power–law result is obtained as a special case of this, as $\mathcal{E}'(x) = \alpha\theta|x|^{\theta-1}\mathrm{sgn}(x)$ in this case.

Example – Ideal gas with gravitation [13, p. 424, Exercise 12.18]: Let

$$\mathcal{E}(x) = \frac{p_x^2 + p_y^2 + p_z^2}{2m} + mgz. \tag{75}$$

The average kinetic energy of each particle is $3kT/2$, as said before. The contribution of the average potential energy is kT (one degree of freedom with $\theta = 1$). Thus, the total is $5kT/2$, where 60% come from kinetic energy and 40% come from potential energy, universally, that is, independent of T, m, and g. $\square$

2.6 The Grand–Canonical Ensemble

A brief summary of what we have done thus far, is the following: we started with the microcanonical ensemble, which was very restrictive in the sense that the energy was held strictly fixed to the value of E, the number of particles was held strictly fixed to the value of N, and at least in the example of a gas, the volume was also held strictly fixed to a certain value V. In the passage from the microcanonical ensemble to the canonical one, we slightly relaxed the first of these parameters, E: Rather than insisting on a fixed value of E, we allowed energy to be exchanged back and forth with the environment, and thereby to slightly fluctuate (for large N) around a certain average value, which was controlled by temperature, or equivalently, by the choice of β. This was done while keeping in mind that the total energy of both system and heat bath must be kept fixed, by the law of energy conservation, which allowed us to look at the combined system as an isolated one, thus obeying the microcanonical ensemble. We then had a one–to–one correspondence between the extensive quantity E and the intensive variable β, that adjusted its average value. But the other extensive variables, like N and V were still kept strictly fixed.

It turns out, that we can continue in this spirit, and 'relax' also either one of the other variables N or V (but not both at the same time), allowing it to fluctuate around a typical average value, and controlling it by a corresponding intensive variable. Like E, both N and V are also subjected to conservation laws when the combined system is considered. Each one of these relaxations, leads to a new ensemble in addition to the microcanonical and the canonical ensembles that we have already seen. In the case where it is the variable N that is allowed to be flexible, this ensemble is called the *grand–canonical ensemble*. In the case where it is the variable V, this a certain instance of the Gibbs ensemble, that was mentioned earlier in a footnote. There are, of course, additional ensembles based on this principle, depending on the kind of the physical system. We will describe here, in some level of detail, only the grand–canonical ensemble.

The fundamental idea is essentially the very same as the one we used to derive the

canonical ensemble: Let us get back to our (relatively small) subsystem, which is in contact with a heat bath, and this time, let us allow this subsystem to exchange with the heat bath, not only energy, but also matter, i.e., particles. The heat bath consists of a huge reservoir of energy and particles. The total energy is E_0 and the total number of particles is N_0. Suppose that we can calculate the density of states of the heat bath as function of both its energy E' and amount of particles N', call it $\Omega_B(E', N')$. A microstate is now a combination $(\boldsymbol{x}, N)$, where N is the (variable) number of particles in our subsystem and $\boldsymbol{x}$ is as before for a given N. From the same considerations as before, whenever our subsystem is in state $(\boldsymbol{x}, N)$, the heat bath can be in any one of $\Omega_B(E_0 - \mathcal{E}(\boldsymbol{x}), N_0 - N)$ microstates of its own. Thus, owing to the microcanonical ensemble,

$$\begin{aligned} P(\boldsymbol{x}, N) &\propto \Omega_B(E_0 - \mathcal{E}(\boldsymbol{x}), N_0 - N) \\ &= \exp\{S_B(E_0 - \mathcal{E}(\boldsymbol{x}), N_0 - N)/k\} \\ &\approx \exp\left\{\frac{S_B(E_0, N_0)}{k} - \frac{1}{k}\frac{\partial S_B}{\partial E} \cdot \mathcal{E}(\boldsymbol{x}) - \frac{1}{k}\frac{\partial S_B}{\partial N} \cdot N\right\} \\ &\propto \exp\left\{-\frac{\mathcal{E}(\boldsymbol{x})}{kT} + \frac{\mu N}{kT}\right\} \end{aligned} \tag{76}$$

where we have now defined the *chemical potential* μ (of the heat bath) as:

$$\mu \triangleq -T \cdot \left.\frac{\partial S_B(E', N')}{\partial N'}\right|_{E'=E_0, N'=N_0}. \tag{77}$$

Thus, we now have the grand–canonical distribution:

$$P(\boldsymbol{x}, N) = \frac{e^{\beta[\mu N - \mathcal{E}(\boldsymbol{x})]}}{\Xi(\beta, \mu)}, \tag{78}$$

where the denominator is called the *grand partition function*:

$$\Xi(\beta, \mu) \triangleq \sum_{N=0}^{\infty} e^{\beta\mu N} \sum_{\boldsymbol{x}} e^{-\beta\mathcal{E}(\boldsymbol{x})} \triangleq \sum_{N=0}^{\infty} e^{\beta\mu N} Z_N(\beta). \tag{79}$$

It is sometimes convenient to change variables and to define $z = e^{\beta\mu}$ (which is called the *fugacity*) and then, define

$$\tilde{\Xi}(\beta, z) = \sum_{N=0}^{\infty} z^N Z_N(\beta). \tag{80}$$

This notation emphasizes the fact that for a given β, $\tilde{\Xi}(z)$ is actually the z–transform of the sequence Z_N. A natural way to think about $P(\boldsymbol{x}, N)$ is as $P(N) \cdot P(\boldsymbol{x}|N)$, where $P(N)$ is proportional to $z^N Z_N(\beta)$ and $P(\boldsymbol{x}|N)$ corresponds to the canonical ensemble as before.

Using the grand partition function, it is now easy to obtain moments of the random variable N. For example, the first moment is:

$$\langle N \rangle = \frac{\sum_N N z^N Z_N(\beta)}{\sum_N z^N Z_N(\beta)} = z \cdot \frac{\partial \ln \tilde{\Xi}(\beta, z)}{\partial z}. \tag{81}$$

Thus, we have replaced the fixed number of particles N by a random number of particles, which concentrates around an average controlled by the parameter μ, or equivalently, z. The dominant value of N is the one that maximizes the product $z^N Z_N(\beta)$, or equivalently, $\ln \tilde{\Xi} \sim \max_N [\beta \mu N + \ln Z_N(\beta)]$. Thus, $\ln \tilde{\Xi}$ is related to $\ln Z_N$ by another kind of a Legendre transform.

When two systems, with total energy E_0 and a total number of particles N_0, are brought into contact, allowing both energy and matter exchange, then the dominant combined states are those for which $\Omega_1(E_1, N_1) \cdot \Omega_2(E_0 - E_1, N_0 - N_1)$, or equivalently, $S_1(E_1, N_1) + S_2(E_0 - E_1, N_0 - N_1)$, is maximum. By equating to zero the partial derivatives w.r.t. both E_1 and N_1, we find that in equilibrium both the temperatures T_1 and T_2 are the same and the chemical potentials μ_1 and μ_2 are the same.

Finally, it should be pointed out that beyond the obvious physical significance of the grand–canonical ensemble, sometimes it proves useful to work with it from the reason of pure mathematical convenience. This is shown in the following example.

Example – Quantum Statistics. Consider an ensemble of indistinguishable particles, each one of which may be in a certain quantum state labeled by $1, 2, \ldots, r, \ldots$. Associated with quantum state number r, there is an energy ϵ_r. Thus, if there are N_r particles in each state r, the total energy is $\sum_r N_r \epsilon_r$, and so, the canonical partition function is:

$$Z_N(\beta) = \sum_{\boldsymbol{N}:\ \sum_r N_r = N} \exp\{-\beta \sum_r N_r \epsilon_r\}, \tag{82}$$

where $\boldsymbol{N}$ denotes the set of occupation numbers $(N_1, N_2, \ldots)$. The constraint $\sum_r N_r = N$, which accounts for the fact that the total number of particles must be N, causes considerable difficulties in the calculation. However, if we pass to the grand–canonical ensemble, things become extremely easy:

$$\begin{aligned}\tilde{\Xi}(\beta, z) &= \sum_{N\geq 0} z^N \sum_{\boldsymbol{N}:\ \sum_r N_r = N} \exp\{-\beta \sum_r N_r \epsilon_r\} \\ &= \sum_{N_1\geq 0}\sum_{N_2\geq 0}\cdots z^{\sum_r N_r} \exp\{-\beta \sum_r N_r \epsilon_r\} \\ &= \sum_{N_1\geq 0}\sum_{N_2\geq 0}\cdots \prod_{r\geq 1} z^{N_r} \exp\{-\beta N_r \epsilon_r\} \\ &= \prod_{r\geq 1}\sum_{N_r\geq 0} [ze^{-\beta\epsilon_r}]^{N_r} \end{aligned} \tag{83}$$

In the case where N_r is unlimited (*Bose–Einstein* particles, or *Bosons*), each factor indexed by r is clearly a geometric series, resulting in

$$\tilde{\Xi}(\beta, z) = \prod_r [1/(1 - ze^{-\beta\epsilon_r})]. \tag{84}$$

In the case where no quantum state can be populated by more than one particle, owing to Pauli's exclusion principle (*Fermi–Dirac* particles, or *Fermions*), each factor in the product contains two terms only, pertaining to $N_r = 0, 1$, and the result is

$$\tilde{\Xi}(\beta, z) = \prod_r (1 + ze^{-\beta\epsilon_r}). \tag{85}$$

In both cases, this is fairly simple. Having computed $\tilde{\Xi}(\beta, z)$, we can in principle, return to $Z_N(\beta)$ by applying the inverse z–transform. We will get back to this point in Chapter 4.

3 Physical Interpretations of Information Measures

In this chapter, we make use of the elementary background that was established in previous chapter, and we draw certain analogies between statistical physics and information theory, most notably, with Shannon theory. These analogies set the stage for physical interpretations of information measures and their properties, as well as fundamental coding theorems. As we shall see, these physical interpretations often prove useful in gaining new insights and perspectives, which may be beneficial for deriving new analysis tools.

In the first section, we begin from a simple correspondence between the maximum entropy principle and optimum assignment of probabilities of messages with given durations, or equivalently, optimum duration assignment for given message probabilities. In the second section, we use large deviations theory (most notably, the Legendre transform) to bridge between information theory and statistical mechanics. This will provide statistical–mechanical interpretations of the rate–distortion function and channel capacity. In this context, the statistical–mechanical perspective will be shown to yield a parametric representation of the rate–distortion function (and channel capacity) as an integral of the minimum mean square error (MMSE) of the distortion given the source symbol, which can be used for deriving bounds. In the third and the fourth sections, we discuss relationships between the second law of thermodynamics and the data processing theorem of information theory, from two completely different aspects. Finally, in the last section, we provide a relationship between the Fisher information and a generalized notion of temperature.

3.1 Statistical Physics of Optimum Message Distributions

To warm up, we begin with a very simple paradigm, studied by Reiss [96] and Reiss and Huang [98]. The analogy and the parallelism to the basic concepts of statistical mechanics, that were introduced in the previous chapter, will be quite evident from the choice of the notation, which is deliberately chosen to correspond to that of analogous physical quantities.

Consider a continuous–time communication system that includes a noiseless channel,

with capacity

$$C = \lim_{E \to \infty} \frac{\log M(E)}{E}, \tag{86}$$

where $M(E)$ is the number of distinct messages that can be transmitted over a time interval of E seconds. The channel is fed by the output of an encoder, which in turn is fed by an information source. Over a duration of E seconds, L source symbols are conveyed, so that the average transmission time per symbol is $\sigma = E/L$ seconds per symbol. In the absence of any constraints on the structure of the encoded messages, $M(E) = r^L = r^{E/\sigma}$, where r is the channel input–output alphabet size. Thus, $C = (\log r)/\sigma$ bits per second.

Consider now the thermodynamic limit of $L \to \infty$. Suppose that the L symbols of duration E form N *words*, where by 'word', we mean a certain variable–length string of channel symbols. The average transmission time per word is then $\epsilon = E/N$. Suppose further that the channel code defines a certain set of word transmission times: Word number i takes ϵ_i seconds to transmit. What is the optimum allocation of word probabilities $\{P_i\}$ that would support full utilization of the channel capacity? Equivalently, given the probabilities $\{P_i\}$, what are the optimum transmission times $\{\epsilon_i\}$? For simplicity, we will assume that $\{\epsilon_i\}$ are all distinct. Suppose that each word appears N_i times in the entire message. Denoting $\boldsymbol{N} = (N_1, N_2, \ldots)$, $P_i = N_i/N$, and $\boldsymbol{P} = (P_1, P_2, \ldots)$, the total number of messages pertaining to a given $\boldsymbol{N}$ is

$$\Omega(\boldsymbol{N}) = \frac{N!}{\prod_i N_i!} \doteq \exp\{N \cdot H(\boldsymbol{P})\} \tag{87}$$

where $H(\boldsymbol{P})$ is the Shannon entropy pertaining to the probability distribution $\boldsymbol{P}$. Now,

$$M(E) = \sum_{\boldsymbol{N}:\ \sum_i N_i \epsilon_i = E} \Omega(\boldsymbol{N}). \tag{88}$$

This sum is dominated by the maximum term, namely, the maximum–entropy assignment of relative frequencies

$$P_i = \frac{e^{-\beta \epsilon_i}}{Z(\beta)} \tag{89}$$

where $\beta > 0$ is a Lagrange multiplier chosen such that $\sum_i P_i \epsilon_i = \epsilon$, which gives

$$\epsilon_i = -\frac{\ln[P_i Z(\beta)]}{\beta}. \tag{90}$$

For $\beta = 1$, this is similar to the classical result that the optimum message length assignment in variable–length lossless data compression is according to the negative logarithm of the probability. For other values of β, this corresponds to the tilting required for optimum variable–length coding in the large deviations regime, when the minimization of the buffer overflow probability is required, see, e.g., [42],[46],[66],[124] and references therein.

Suppose now that $\{\epsilon_i\}$ are kept fixed and consider a small perturbation in P_i, denoted $\mathrm{d}P_i$. Then

$$\begin{aligned}
\mathrm{d}\epsilon &= \sum_i \epsilon_i \mathrm{d}P_i \\
&= -\frac{1}{\beta} \sum_i (\mathrm{d}P_i) \ln P_i \\
&= \frac{1}{k\beta} \mathrm{d} \left(-k \sum_i P_i \ln P_i \right) \\
&\triangleq T \mathrm{d}s,
\end{aligned} \tag{91}$$

where we have defined $T = 1/(k\beta)$ and $s = -k \sum_i P_i \ln P_i$. The free energy per particle is given by

$$f = \epsilon - Ts = -kT \ln Z, \tag{92}$$

which is related to the redundancy of the code, as both the free energy and the redundancy are characterized by the Kullback–Leibler divergence.

In [96], there is also an extension of this setting to the case where N is not fixed, with correspondence to the grand—canonical ensemble. However, we will not include it here.

3.2 Large Deviations and Physics of Coding Theorems

Deeper perspectives are offered via the large–deviations point of view. As said in the Introduction, large deviations theory, the branch of probability theory that deals with exponential decay rates of probabilities of rare events, has strong relations to information theory, which can easily be seen from the viewpoint of the method of types and Sanov's theorem. On the other hand, large deviations theory has also a strong connection to statistical mechanics,

as we are going to see shortly. Therefore, one of the links between information theory and statistical mechanics goes via rate functions of large deviations theory, or more concretely, Legendre transforms. This topic is based on [67].

Let us begin with a very simple question: We have a set of i.i.d. random variables $X_1, X_2, \ldots$ and a certain real function $\mathcal{E}(x)$. How fast does the probability of the event $\sum_{i=1}^{N} \mathcal{E}(X_i) \leq NE_0$ decay as N grows without bound, assuming that $E_0 < \langle \mathcal{E}(X) \rangle$? One way to handle this problem, at least in the finite alphabet case, is the method of types. Another method is the Chernoff bound: Denoting the indicator function of an event by $\mathcal{I}(\cdot)$, we have

$$\begin{aligned}
\Pr\left\{\sum_{i=1}^{N} \mathcal{E}(X_i) \leq NE_0\right\} &= \boldsymbol{E}\mathcal{I}\left\{\sum_{i=1}^{N} \mathcal{E}(X_i) \leq NE_0\right\} \\
&\leq \boldsymbol{E} \exp\left\{\beta\left[NE_0 - \sum_{i=1}^{N} \mathcal{E}(X_i)\right]\right\} \\
&= e^{\beta NE_0} \boldsymbol{E} \exp\left\{-\beta \sum_{i=1}^{N} \mathcal{E}(X_i)\right\} \\
&= e^{\beta NE_0} \boldsymbol{E}\left\{\prod_{i=1}^{N} \exp\{-\beta \mathcal{E}(X_i)\}\right\} \\
&= e^{\beta NE_0} \left[\boldsymbol{E} \exp\{-\beta \mathcal{E}(X_1)\}\right]^N \\
&= \exp\left\{N\left[\beta E_0 + \ln \boldsymbol{E} \exp\{-\beta \mathcal{E}(X_1)\}\right]\right\}
\end{aligned}$$

As this bound applies for every $\beta \geq 0$, the tightest bound of this family is obtained by minimizing the r.h.s. over β, which yields the exponential rate function:

$$\Sigma(E_0) = \min_{\beta \geq 0}[\beta E_0 + \phi(\beta)], \tag{93}$$

where

$$\phi(\beta) = \ln Z(\beta) \tag{94}$$

and

$$Z(\beta) = \boldsymbol{E} e^{-\beta \mathcal{E}(X)} = \sum_x p(x) e^{-\beta \mathcal{E}(x)}. \tag{95}$$

This is, obviously, very similar to the relation between the entropy function and the partition function, which we have seen in the previous chapter. Note that $Z(\beta)$ here differs from the partition function that we have encountered thus far only slightly: the Boltzmann exponentials are weighed by $\{p(x)\}$ which are independent of β. But this is not a crucial difference: one can imagine a physical system where each microstate x is actually a representative of a bunch of more refined microstates $\{x'\}$, whose number is proportional to $p(x)$ and which all have the same energy as x, that is, $\mathcal{E}(x') = \mathcal{E}(x)$. In the domain of the more refined system, $Z(\beta)$ is (up to a constant) a non–weighted sum of exponentials, as it should be. More precisely, if $p(x)$ is (or can be approximated by) a rational number $M(x)/M$, where M is independent of x, then imagine that each x gives rise to $M(x)$ microstates $\{x'\}$ with the same energy as x, so that

$$Z(\beta) = \frac{1}{M}\sum_{x} M(x)e^{-\beta\mathcal{E}(x)} = \frac{1}{M}\sum_{x'} e^{-\beta\mathcal{E}(x')}, \tag{96}$$

and we are back to an ordinary, non–weighted partition function, up to the constant $1/M$, which is absolutely immaterial.

We observe that the exponential rate function is given by the Legendre transform of the log–moment generating function. The Chernoff parameter β to be optimized plays the role of the equilibrium temperature pertaining to energy E_0.

Consider next what happens when $p(x)$ is itself a B–G distribution with Hamiltonian $\mathcal{E}(x)$ at a certain inverse temperature β_1, that is

$$p(x) = \frac{e^{-\beta_1\mathcal{E}(x)}}{\zeta(\beta_1)} \tag{97}$$

with

$$\zeta(\beta_1) \stackrel{\Delta}{=} \sum_{x} e^{-\beta_1\mathcal{E}(x)}. \tag{98}$$

In this case, we have

$$Z(\beta) = \sum_{x} p(x)e^{-\beta\mathcal{E}(x)} = \frac{\sum_x e^{-(\beta_1+\beta)\mathcal{E}(x)}}{\zeta(\beta_1)} = \frac{\zeta(\beta_1+\beta)}{\zeta(\beta_1)}. \tag{99}$$

Thus,

$$\begin{aligned}\Sigma(E_0) &= \min_{\beta\geq 0}[\beta E_0 + \ln\zeta(\beta_1+\beta)] - \ln\zeta(\beta_1)\\ &= \min_{\beta\geq 0}[(\beta+\beta_1)E_0 + \ln\zeta(\beta_1+\beta)] - \ln\zeta(\beta_1) - \beta_1 E_0\\ &= \min_{\beta\geq \beta_1}[\beta E_0 + \ln\zeta(\beta)] - \ln\zeta(\beta_1) - \beta_1 E_0\\ &= \min_{\beta\geq \beta_1}[\beta E_0 + \ln\zeta(\beta)] - [\ln\zeta(\beta_1) + \beta_1 E_1] + \beta_1(E_1 - E_0)\end{aligned}$$

where E_1 is the energy corresponding to β_1, i.e., E_1 is such that

$$\sigma(E_1) \stackrel{\Delta}{=} \min_{\beta\geq 0}[\beta E_1 + \ln\zeta(\beta)] \tag{100}$$

is achieved by $\beta = \beta_1$. Thus, the second bracketed term of the right–most side of the last chain is exactly $\sigma(E_1)$, as defined. If we now assume that $E_0 < E_1$, which is reasonable, because E_1 is the average of $\mathcal{E}(X)$ under β_1, and we are assuming that we are dealing with a rare event where $E_0 < \langle\mathcal{E}(X)\rangle$. In this case, the achiever β_0 of $\sigma(E_0)$ must be larger than β_1 anyway, and so, the first bracketed term on the right–most side of the last chain agrees with $\sigma(E_0)$. We have obtained then that the exponential decay rate (the rate function) is given by

$$I = -\Sigma(E_0) = \sigma(E_1) - \sigma(E_0) - \beta_1(E_1 - E_0). \tag{101}$$

Note that $I \geq 0$, due to the fact that $\sigma(\cdot)$ is concave. It has a simple graphical interpretation as the height difference, as seen at the point $E = E_0$, between the tangent to the curve $\sigma(E)$ at $E = E_1$ and the function $\sigma(E)$ itself (see Fig. 2).

Another point of view is the following:

$$\begin{aligned}I &= \beta_1\left[\left(E_0 - \frac{\sigma(E_0)}{\beta_1}\right) - \left(E_1 - \frac{\sigma(E_1)}{\beta_1}\right)\right]\\ &= \beta_1(F_0 - F_1)\\ &= D(P_{\beta_0}\|P_{\beta_1})\\ &= \min\{D(Q\|P_{\beta_1}):\ \boldsymbol{E}_Q\mathcal{E}(X) \leq E_0\}\end{aligned} \tag{102}$$

where the last line (which is easy to obtain) is exactly what we would have obtained using the method of types. This means that the dominant instance of the large deviations event

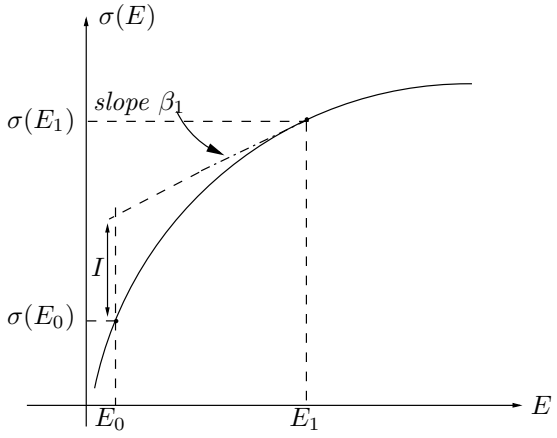


Figure 2: Graphical interpretation of the large deviations rate function I.

under discussion pertains to thermal equilibrium (minimum free energy) complying with the constraint(s) dictated by this event. This will also be the motive of the forthcoming results.

Let us now see how this discussion relates to very fundamental information measures, like the rate–distortion function and channel capacity. To this end, let us first slightly extend the above Chernoff bound. Assume that in addition to the random variables $X_1, \ldots, X_N$, there is also a deterministic sequence of the same length, $y_1, \ldots, y_N$, where each y_i takes on values in a finite alphabet $\mathcal{Y}$. Suppose also that the asymptotic regime is such that as N grows without bound, the relative frequencies $\{\frac{1}{N}\sum_{i=1}^{N} 1\{y_i = y\}\}_{y\in\mathcal{Y}}$ converge to certain probabilities $\{q(y)\}_{y\in\mathcal{Y}}$. Furthermore, the X_i's are still independent, but they are no longer necessarily identically distributed: each one of them is governed by $p(x_i|y_i)$, that is, $p(\boldsymbol{x}|\boldsymbol{y}) = \prod_{i=1}^{N} p(x_i|y_i)$. Now, the question is how does the exponential rate function behave if we look at the event

$$\sum_{i=1}^{N} \mathcal{E}(X_i, y_i) \leq NE_0 \tag{103}$$

where $\mathcal{E}(x, y)$ is a given 'Hamiltonian'. The natural questions that now arise are: what is the motivation for this question and where do we encounter such a problem?

The answer is that there are many examples (see [67]), but here are two very classical ones, where rate functions of large deviations events are directly related to very important information measures. In both examples, the distributions $p(\cdot|y)$ are actually the same for

all $y \in \mathcal{Y}$ (namely, $\{X_i\}$ are again i.i.d.).

Rate–distortion coding. Consider the good old problem of lossy compression with a randomly selected code. Let $\boldsymbol{y} = (y_1, \ldots, y_N)$ be a given source sequence, typical to $Q = \{q(y),\ y \in \mathcal{Y}\}$ (non–typical sequences are not important). Now, let us randomly select e^{NR} codebook vectors $\{\boldsymbol{X}(i)\}$ according to $p(\boldsymbol{x}) = \prod_{i=1}^{N} p(x_i)$. Here is how the direct part of the source coding theorem essentially works: We first ask ourselves what is the probability that a single randomly selected codeword $\boldsymbol{X} = (X_1, \ldots, X_N)$ would happen to fall at distance less than or equal to ND from $\boldsymbol{y}$, i.e., what is the exponential rate of the probability of the event $\sum_{i=1}^{N} d(X_i, y_i) \leq ND$? The answer is that it is exponentially about $e^{-NR(D)}$, and this the reason why we need slightly more than the reciprocal of this number, namely, $e^{+NR(D)}$ times to repeat this 'experiment' in order to see at least one 'success', which means being able to encode $\boldsymbol{y}$ within distortion D. So this is clearly an instance of the above problem, where $\mathcal{E} = d$ and $E_0 = D$.

Channel coding. In complete duality, consider the classical channel coding problem, for a discrete memoryless channel (DMC), using a randomly selected code. Again, we have a code of size e^{NR}, where each codeword is chosen independently according to $p(\boldsymbol{x}) = \prod_{i=1}^{N} p(x_i)$. Let $\boldsymbol{y}$ the channel output vector, which is (with very high probability), typical to $Q = \{q(y),\ y \in \mathcal{Y}\}$, where $q(y) = \sum_x p(x)W(y|x)$, W being the single–letter transition probability matrix of the DMC. Consider a (capacity–achieving) threshold decoder which selects the *unique* codeword that obeys

$$\sum_{i=1}^{N} [-\ln W(y_i|X_i)] \leq N[H(Y|X) + \epsilon] \quad \epsilon > 0 \tag{104}$$

and declares an error whenever no such codeword exists or when there is more than one such codeword. Now, in the classical proof of the direct part of the channel coding problem, we first ask ourselves: what is the probability that an independently selected codeword (and hence not the one transmitted) $\boldsymbol{X}$ will pass this threshold? The answer turns out to be exponentially e^{-NC}, and hence we can randomly select up to slightly less than the reciprocal

of this number, namely, e^{+NC} codewords, before we start to see incorrect codewords that pass the threshold. Again, this is clearly an instance of our problem with $\mathcal{E}(x,y) = -\ln W(y|x)$ and $E_0 = H(Y|X) + \epsilon$.

Equipped with these two motivating examples, let us get back to the generic problem we formalized. Once this has been done, we shall return to the examples. There are (at least) two different ways to address the problem using Chernoff bounds, and they lead to two *seemingly* different expressions, but since the Chernoff bounding technique gives the correct exponential behavior, these two expressions must agree. This identity between the two expressions will have a physical interpretation, as we shall see.

The first approach is a direct extension of the previous derivation:

$$\begin{aligned}
&\Pr\left\{\sum_{i=1}^{N}\mathcal{E}(X_i,y_i) \le nE_0\right\} \\
&= \boldsymbol{E}\mathcal{I}\left\{\sum_{i=1}^{N}\mathcal{E}(X_i,y_i) \le nE_0\right\} \\
&\le \boldsymbol{E}\exp\left\{\beta\left[NE_0 - \sum_{i=1}^{N}\mathcal{E}(X_i,y_i)\right]\right\} \\
&= e^{N\beta E_0}\prod_{y\in\mathcal{Y}}\boldsymbol{E}_y\exp\left\{-\beta\sum_{i:y_i=y}\mathcal{E}(X_i,y)\right\} \\
&= e^{\beta NE_0}\prod_{y\in\mathcal{Y}}\left[\boldsymbol{E}_y\exp\{-\beta\mathcal{E}(X,y)\}\right]^{N(y)} \\
&= \exp\left\{N\left[\beta E_0 + \sum_{y\in\mathcal{Y}}q(y)\ln\sum_{x\in\mathcal{X}}p(x|y)\exp\{-\beta\mathcal{E}(x,y)\}\right]\right\}
\end{aligned}$$

where $\boldsymbol{E}_y\{\cdot\}$ denotes expectation w.r.t. $p(\cdot|y)$ and $N(y)$ is the number of occurrences of $y_i = y$ in $(y_1,\ldots,y_n)$. The resulting rate function is given by

$$\Sigma(E_0) = \min_{\beta\ge 0}\left[\beta E_0 + \sum_{y\in\mathcal{Y}}q(y)\ln Z_y(\beta)\right] \tag{105}$$

where

$$Z_y(\beta) \stackrel{\Delta}{=} \sum_{x\in\mathcal{X}}p(x|y)\exp\{-\beta\mathcal{E}(x,y)\}. \tag{106}$$

In the rate–distortion example, this tells us that

$$R(D) = -\min_{\beta \geq 0} \left[\beta D + \sum_{y \in \mathcal{Y}} q(y) \ln \left(\sum_{x \in \mathcal{X}} p(x) e^{-\beta d(x,y)} \right) \right]. \tag{107}$$

This is a well–known parametric representation of $R(D)$, which can be obtained via a different route (see [36, p. 90, Corollary 4.2.3]), where the minimizing β is known to have the graphical interpretation of the negative local slope (i.e., the derivative) of the curve of $R(D)$. In the case of channel capacity, we obtain in a similar manner:

$$\begin{aligned} C &= -\min_{\beta \geq 0} \left[\beta H(Y|X) + \sum_{y \in \mathcal{Y}} q(y) \ln \left(\sum_{x \in \mathcal{X}} p(x) e^{-\beta[-\ln W(y|x)]} \right) \right] \\ &= -\min_{\beta \geq 0} \left[\beta H(Y|X) + \sum_{y \in \mathcal{Y}} q(y) \ln \left(\sum_{x \in \mathcal{X}} p(x) W^{\beta}(y|x) \right) \right]. \end{aligned}$$

Here, it is easy to see that the minimizing β is always $\beta^* = 1$.

The other route is to handle each $y \in \mathcal{Y}$ separately: First, observe that

$$\sum_{i=1}^{N} \mathcal{E}(X_i, y_i) = \sum_{y \in \mathcal{Y}} \sum_{i:\ y_i = y} \mathcal{E}(X_i, y), \tag{108}$$

where now, in each partial sum over $\{i :\ y_i = y\}$, we have i.i.d. random variables. The event $\sum_{i=1}^{N} \mathcal{E}(X_i, y_i) \leq NE_0$ can then be thought of as the union of all intersections

$$\bigcap_{y \in \mathcal{Y}} \left\{ \sum_{i:\ y_i = y} \mathcal{E}(X_i, y) \leq N(y) E_y \right\} \tag{109}$$

where the union is over all "possible partial energy allocations" $\{E_y\}$ which satisfy $\sum_y q(y) E_y \leq E_0$. Note that at least when $\{X_i\}$ take values on a finite alphabet, each partial sum $\sum_{i:\ y_i = y} \mathcal{E}(X_i, y)$ can take only a polynomial number of values in $N(y)$, and so, it is sufficient to 'sample' the space of $\{E_y\}$ by polynomially many vectors in order to cover all possible instances of the event under discussion (see more details in the paper). Thus,

$$\Pr \left\{ \sum_{i=1}^{N} \mathcal{E}(X_i, y_i) \leq NE_0 \right\}$$

$$
\begin{aligned}
&= \Pr \bigcup_{\{E_y:\ \sum_y q(y)E_y \le E_0\}} \bigcap_{y\in\mathcal{Y}} \left\{ \sum_{i:\ y_i=y} \mathcal{E}(X_i,y) \le N(y)E_y \right\} \\
&\dot{=} \max_{\{E_y:\ \sum_y q(y)E_y \le E_0\}} \prod_{y\in\mathcal{Y}} \Pr \left\{ \sum_{i:\ y_i=y} \mathcal{E}(X_i,y) \le N(y)E_y \right\} \\
&\dot{=} \max_{\{E_y:\ \sum_y q(y)E_y \le E_0\}} \prod_{y\in\mathcal{Y}} \exp \left\{ N(y) \min_{\beta_y\ge 0} [\beta_y E_y + \ln Z_y(\beta)] \right\} \\
&= \exp \left\{ N \cdot \max_{\{E_y:\ \sum_y q(y)E_y \le E_0\}} \sum_{y\in\mathcal{Y}} q(y)\Sigma_y(E_y) \right\}
\end{aligned}
$$

where we have defined

$$
\Sigma_y(E_y) \stackrel{\Delta}{=} \min_{\beta_y\ge 0} \left[\beta_y E_y + \ln Z_y(\beta_y)\right]. \tag{110}
$$

We therefore arrived at an alternative expression of the rate function, which is

$$
\max_{\{E_y:\ \sum_y q(y)E_y \le E_0\}} \sum_{y\in\mathcal{Y}} q(y)\Sigma_y(E_y). \tag{111}
$$

Since the two expressions must agree, we obtain following identity:

$$
\Sigma(E_0) = \max_{\{E_y:\ \sum_y q(y)E_y \le E_0\}} \sum_{y\in\mathcal{Y}} q(y)\Sigma_y(E_y). \tag{112}
$$

A few comments are now in order:

1. In [67], there is also a direct proof of this identity, without relying on Chernoff bound considerations.

2. This identity accounts for a certain generalized concavity property of the entropy function. Had all the $\Sigma_y(\cdot)$'s been the same function, then this would have been the ordinary concavity property. The interesting point here is that it continues to hold for different $\Sigma_y(\cdot)$'s too.

3. The l.h.s. of this identity is defined by minimization over one parameter only – the inverse temperature β. On the other hand, on the r.h.s. we have a separate inverse temperature

for every y, because each $\Sigma_y(\cdot)$ is defined as a separate minimization problem with its own β_y. Stated differently, the l.h.s. is the minimum of a sum, whereas in the r.h.s., for given $\{E_y\}$, we have the sum of minima. When do these two things agree? The answer is that it happens if all minimizers $\{\beta_y^*\}$ happen to be the *same*. But β_y^* depends on E_y. So what happens is that the $\{E_y\}$ (of the outer maximization problem) are such that the β_y^* would all be the same, and would agree also with the β^* of $\Sigma(E_0)$. To see why this is true, consider the following chain of inequalities:

$$\begin{aligned}
&\max_{\{E_y:\ \sum_y q(y)E_y \le E_0\}} \sum_y q(y)\Sigma_y(E_y) \\
&= \max_{\{E_y:\ \sum_y q(y)E_y \le E_0\}} \sum_y q(y) \min_{\beta_y}[\beta_y E_y + \ln Z_y(\beta_y)] \\
&\le \max_{\{E_y:\ \sum_y q(y)E_y \le E_0\}} \sum_y q(y)[\beta^* E_y + \ln Z_y(\beta^*)] \\
&\le \max_{\{E_y:\ \sum_y q(y)E_y \le E_0\}} [\beta^* E_0 + \sum_y q(y) \ln Z_y(\beta^*)] \\
&= \beta^* E_0 + \sum_y q(y) \ln Z_y(\beta^*) \\
&= \Sigma(E_0), \qquad (113)
\end{aligned}$$

where β^* achieves $\Sigma(E_0)$, the last inequality is because $\sum_y q(y)E_y \le E_0$, and the last equality is because the bracketed expression no longer depends on $\{E_y\}$. Both inequalities become equalities if $\{E_y\}$ are allocated such that: (i) $\sum_y q(y)E_y = E_0$ and (ii) $\beta_y^*(E_y) = \beta^*$ for all y. Since the β's have the meaning of inverse temperatures, what we have here is **thermal equilibrium**: Consider a set of $|\mathcal{Y}|$ subsystems, each one of $N(y)$ particles and Hamiltonian $\mathcal{E}(x, y)$, indexed by y. If all these subsystems are thermally separated, each one with energy E_y, then the total entropy per particle is $\sum_y q(y)\Sigma_y(E_y)$. The above identity tells us then what happens when all these systems are brought into thermal contact with one another: The total energy per particle E_0 is split among the different subsystems in a way that all temperatures become the same – thermal equilibrium. It follows then that the dominant instance of the large deviations event is the one where the contributions of each y, to the partial sum of energies, would correspond to equilibrium. In the rate–distortion example,

this characterizes how much distortion each source symbol contributes typically.

Let us now focus more closely on the rate–distortion function:

$$R(D) = -\min_{\beta \geq 0} \left[\beta D + \sum_{y \in \mathcal{Y}} q(y) \ln \left(\sum_{x \in \mathcal{X}} p(x) e^{-\beta d(x,y)} \right) \right]. \tag{114}$$

As said, the Chernoff parameter β has the meaning of inverse temperature. The inverse temperature β required to 'tune' the expected distortion (internal energy) to D, is the solution to the equation

$$D = -\frac{\partial}{\partial \beta} \sum_{y} q(y) \ln \left[\sum_{x} p(x) e^{-\beta d(x,y)} \right] \tag{115}$$

or equivalently,

$$D = \sum_{y} q(y) \cdot \frac{\sum_x p(x) d(x,y) e^{-\beta d(x,y)}}{\sum_x p(x) \cdot e^{-\beta d(x,y)}}. \tag{116}$$

The Legendre transform relation between the log–partition function and $R(D)$ induces a one–to–one mapping between D and β which is defined by the above equation. To emphasize this dependency, we henceforth denote the value of D, corresponding to a given β, by D_β. This expected distortion is defined w.r.t. the probability distribution:

$$P_\beta(x,y) = q(y) \cdot P_\beta(x|y) = q(y) \cdot \frac{p(x) e^{-\beta d(x,y)}}{\sum_{x'} p(x') e^{-\beta d(x',y)}}. \tag{117}$$

On substituting D_β instead of D in the expression of $R(D)$, we have

$$-R(D_\beta) = \beta D_\beta + \sum_{y} q(y) \ln \left[\sum_{x} p(x) e^{-\beta d(x,y)} \right]. \tag{118}$$

Note that $R(D_\beta)$ can be represented in an integral form as follows:

$$\begin{aligned} R(D_\beta) &= -\int_0^\beta \mathrm{d}\hat{\beta} \cdot \left(D_{\hat{\beta}} + \hat{\beta} \cdot \frac{\mathrm{d}D_{\hat{\beta}}}{\mathrm{d}\hat{\beta}} - D_{\hat{\beta}} \right) \\ &= -\int_{D_0}^{D_\beta} \hat{\beta} \cdot \mathrm{d}D_{\hat{\beta}}, \end{aligned} \tag{119}$$

where $D_0 = \sum_{x,y} p(x) q(y) d(x,y)$ is the value of D corresponding to $\beta = 0$, and for which $R(D) = 0$, This is exactly analogous to the thermodynamic equation $S = \int \mathrm{d}Q/T$ (following

from $1/T = \mathrm{d}S/\mathrm{d}Q$), that builds up the entropy from the cumulative heat. Note that the last equation, in its differential form, reads $\mathrm{d}R(D_\beta) = -\beta \mathrm{d}D_\beta$, or $\beta = -R'(D_\beta)$, which means that β is indeed the negative local slope of the rate–distortion curve $R(D)$. Returning to the integration variable $\hat{\beta}$, we have:

$$\begin{aligned} R(D_\beta) &= -\int_0^\beta \mathrm{d}\hat{\beta} \cdot \hat{\beta} \cdot \frac{\mathrm{d}D_{\hat{\beta}}}{\mathrm{d}\hat{\beta}} \\ &= \sum_y q(y) \int_0^\beta \mathrm{d}\hat{\beta} \cdot \hat{\beta} \cdot \mathrm{Var}_{\hat{\beta}}\{d(X,y)|Y=y\} \\ &= \int_0^\beta \mathrm{d}\hat{\beta} \cdot \hat{\beta} \cdot \mathrm{mmse}_{\hat{\beta}}\{d(X,Y)|Y\} \end{aligned}$$

where $\mathrm{Var}_{\hat{\beta}}\{d(X,y)|Y=y\}$ is the conditional variance of $d(X,y)$ given $Y=y$, whose expectation is the minimum mean square error (MMSE) of $d(X,Y)$ based on Y, with (X,Y) being distributed according to $P_{\hat{\beta}}(x,y)$. In this representation, the expected conditional variance, which is the minimum mean square error plays a role that is intimately related to the heat capacity of the analogous physical system (cf. eq. (34). In a similar manner, the distortion (which is analogous to the internal energy) is given by

$$D_\beta = D_0 - \int_0^\beta \mathrm{d}\hat{\beta} \cdot \mathrm{mmse}_{\hat{\beta}}\{d(X,Y)|Y\}. \tag{120}$$

We have therefore introduced an integral representation for $R(D)$ based on the MMSE in estimating the distortion variable $d(X,Y)$ based on Y. It should be kept in mind that in the above representation, $q(y)$ is kept fixed and it is the optimum output distribution corresponding to $D = D_\beta$, not to the distortion levels pertaining to values of the integration $\hat{\beta}$. Alternatively, q can be any fixed output distribution, not necessarily the optimum distribution, and then the above integrals correspond to $R_q(D_\beta)$ and D_β, where $R_q(D)$ is the minimum of $I(X;Y)$ subject to the distortion constraint and the additional constraint that the output distribution is q.

Example. Consider the binary symmetric source $\{q(y)\}$, where the coding distribution $\{p(x)\}$ is binary and symmetric as well. Let d be the Hamming distortion measure. Then,

$$P_\beta(x|y) = \frac{e^{-\beta d(x,y)}}{1+e^{-\beta}},$$

and since this channel is symmetric, it is clear that

$$\Pr\{d(X,Y)=1|Y=y\}=\frac{e^{-\beta}}{1+e^{-\beta}}=D,$$

independently of y, in other words, $d(X,Y)$ is independent of Y. It follows then that $\mathrm{mmse}_\beta\{d(X,Y)|Y\}$, which is also the expected conditional variance of $d(X,Y)$ given Y, is the same as the unconditional variance of $d(X,Y)$, which is easily found to be $e^{-\beta}/(1+e^{-\beta})^2$. Thus,

$$\begin{aligned}R(D)&=\int_0^\beta\frac{\tilde{\beta}e^{-\tilde{\beta}}\mathrm{d}\tilde{\beta}}{(1+e^{-\tilde{\beta}})^2}\\&=\ln 2+\frac{\beta e^{-\beta}}{1+e^{\beta}}-\ln(1+e^{\beta})\\&=\ln 2-h_2\left(\frac{e^{-\beta}}{1+e^{-\beta}}\right)\\&=\ln 2-h_2(D),\end{aligned}\tag{121}$$

as expected.

More often than not, an exact closed–form expression of $R(D)$ is hard to obtain, and one must resort to bounds. The MMSE representation opens the door to the derivation of families of upper and lower bounds on $R_q(D)$, which are based on upper and lower bounds on the MMSE, offered by the plethora of bounds available in estimation theory. This line of thought was exemplified and further developed in [72], where it was demonstrated that MMSE–based bounds may sometimes be significantly tighter than traditional bounds, like the Shannon lower bound. This demonstrates the point that the physical point of view may inspire a new perspective that leads to new results.

A few words about the high–resolution regime are in order. The partition function of each y is

$$Z_y(\beta)=\sum_x p(x)e^{-\beta d(x,y)},\tag{122}$$

or, in the continuous case,

$$Z_y(\beta)=\int_{\mathbb{R}}\mathrm{d}x p(x)e^{-\beta d(x,y)}.\tag{123}$$

Consider the L^θ distortion measure $d(x,y) = |x-y|^\theta$, where $\theta > 0$ and consider a uniform random coding distribution over the interval $[-A, A]$, supposing that it is the optimal (or close to optimal) one. Suppose further that we wish to work at a very small distortion level D (high resolution), which means a large value of β. Then,

$$\begin{aligned} Z_y(\beta) &= \frac{1}{2A}\int_{-A}^{+A} \mathrm{d}x e^{-\beta|x-y|^\theta} \\ &\approx \frac{1}{2A}\int_{-\infty}^{+\infty} \mathrm{d}x e^{-\beta|x-y|^\theta} \\ &= \frac{1}{2A}\int_{-\infty}^{+\infty} \mathrm{d}x e^{-\beta|x|^\theta} \end{aligned} \tag{124}$$

Thus, returning to the expression of $R(D)$, let us minimize over β by writing the zero–derivative equation, which yields:

$$D = -\frac{\partial}{\partial\beta}\ln\left[\frac{1}{2A}\int_{-\infty}^{+\infty}\mathrm{d}x e^{-\beta|x|^\theta}\right] \tag{125}$$

but this is exactly the calculation of the (generalized) equipartition theorem, which gives $1/(\beta\theta) = kT/\theta$. Now, we already said that $\beta = -R'(D)$, and so, $1/\beta = -D'(R)$. It follows then that the function $D(R)$, at this high res. limit, obeys a simple differential equation:

$$D(R) = -\frac{D'(R)}{\theta} \tag{126}$$

whose solution is

$$D(R) = D_0 e^{-\theta R}. \tag{127}$$

In the case where $\theta = 2$ (quadratic distortion), we obtain that $D(R)$ is proportional to e^{-2R}, which is a well–known result in high resolution quantization theory. For the Gaussian source, this is true for all R.

We have interpreted the Legendre representation of the rate–distortion function (107) in the spirit of the Legendre relation between the log–partition function and the entropy of a physical system, where the parameter β plays the role of inverse temperature. An alternative physical interpretation was given in [71],[72], in the spirit of the discussion around eqs. (38)–(40), where the temperature was assumed fixed, the notation of β was changed to another

symbol, say λ, with an interpretation of a generalized force acting on the system (e.g., pressure or magnetic field), and the distortion variable was the conjugate physical quantity influenced by this force (e.g., volume in the case of pressure, or magnetization in the case of a magnetic field). In this case, the minimizing λ means the equal force that each one of the various subsystems is applying on the others when they are brought into contact and they equilibrate (e.g., equal pressures between two volumes of a gas separated by piston which is free to move). In this case, $R_q(D)$ is interpreted as the free energy of the system, and the MMSE formulas are intimately related to the fluctuation–dissipation theorem (see, e.g., [5, Chap. 6], [39, Sect. 5.7], [59, part 2], [61, Chap. XII], [80, Sect. 2.3], [95, Chap. 15],[104, Chap. 10]) an important result statistical mechanics, which establishes a relationship between the linear response of a system to a small perturbation from equilibrium, and the equilibrium fluctuations of this system.

More concretely, it was shown in [71] and [75], that given a source distribution and a distortion measure, we can describe (at least conceptually) a concrete physical system[8] that emulates the rate–distortion problem in the following manner (see Fig. 3): When no force is applied to the system, its total length is nD_0, where n is the number of particles in the system (and also the block length in the rate–distortion problem), and D_0 is as defined above. If one applies to the system a contracting force, that increases from zero to some final value λ, such that the length of the system shrinks to nD, where $D < D_0$ is analogous to a prescribed distortion level, then the following two facts hold true: (i) An *achievable lower bound* on the total amount of mechanical work that must be carried out by the contracting force in order to shrink the system to length nD, is given by

$$W \geq nkTR_q(D). \tag{128}$$

(ii) The final force λ is related to D according to

$$\lambda = kTR_q'(D), \tag{129}$$

[8]In particular, this system consists of a long chain of particles, e.g., a polymer.

where $R_q'(\cdot)$ is the derivative of $R_q(\cdot)$. Thus, the rate–distortion function plays the role of a fundamental limit, not only in information theory, but in a certain way, in physics as well.

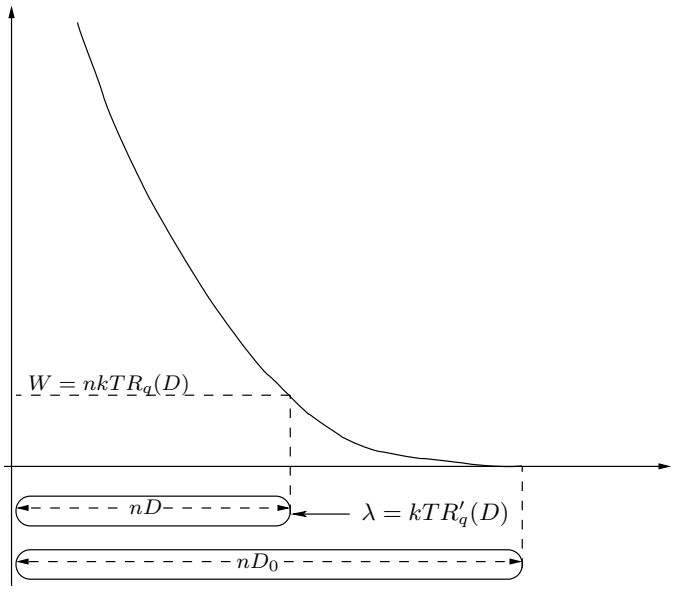


Figure 3: Emulation of the rate–distortion function by a physical system.

3.3 Gibbs' Inequality and the Second Law

Another aspect of the physical interpretation of information measures is associated with the following question: While the laws of physics draw the boundaries between the possible and the impossible in Nature, the coding theorems of information theory, or more precisely, their converses, draw the boundaries between the possible and the impossible in coded communication systems and data processing. Are there any relationships between these two facts?

We next demonstrate that there are some indications that the answer to this question is affirmative. In particular, we shall see that there is an intimate relationship between the second law of thermodynamics and the data processing theorem (DPT), asserting that if $X \to U \to V$ is a Markov chain, then $I(X;U) \geq I(X;V)$. The reason for focusing our attention on the DPT is that it is actually the most fundamental inequality that supports

most (if not all) proofs of converse theorems in information theory. Here are just a few points that make this quite clear.

Lossy/lossless source coding: Consider a source vector $U^N = (U_1, \ldots U_N)$ compressed into a bit-stream $X^n = (X_1, \ldots, X_n)$ from which the decoder generates a reproduction $V^N = (V_1, \ldots, V_N)$ with distortion $\sum_{i=1}^N \boldsymbol{E}\{d(U_i, V_i)\} \leq ND$. Then, by the DPT,

$$I(U^N; V^N) \leq I(X^n; X^n) = H(X^n), \tag{130}$$

where $I(U^N; V^N)$ is further lower bounded by $NR(D)$ and $H(X^n) \leq n$, which together lead to the converse to the lossy data compression theorem, asserting that the compression ratio n/N cannot be less than $R(D)$. The case of lossless compression is obtained as a special case where $D = 0$.

Channel coding under bit error probability: Let $U^N = (U_1, \ldots U_N)$ be drawn from the binary symmetric source (BSS), designating $M = 2^N$ equiprobable messages of length N. The encoder maps U^N into a channel input vector X^n, which in turn, is sent across the channel. The receiver observes Y^n, a noisy version of X^n, and decodes the message as V^N. Let

$$P_b = \frac{1}{N} \sum_{i=1}^N \Pr\{V_i \neq U_i\} \tag{131}$$

designate the bit error probability. Then, by the DPT, $I(U^N; V^N) \leq I(X^n; Y^n)$, where $I(X^n; Y^n)$ is further upper bounded by nC, C being the channel capacity, and

$$\begin{aligned} I(U^N; V^N) &= H(U^N) - H(U^N|V^N) \\ &\geq N - \sum_{i=1}^N H(U_i|V_i) \\ &\geq N - \sum_i h_2(\Pr\{V_i \neq U_i\}) \\ &\geq N[1 - h_2(P_b)]. \end{aligned} \tag{132}$$

Thus, for P_b to vanish, the coding rate, N/n should not exceed C.

Channel coding under block error probability – Fano's inequality: Same as in the previous item, except that the error performance is the block error probability $P_B = \Pr\{V^N \neq U^N\}$. This, time $H(U^N|V^N)$, which is identical to $H(U^N, E|V^N)$, with $E \equiv \mathcal{I}\{V^N \neq U^N\}$, is decomposed as $H(E|V^N) + H(U^N|V^N, E)$, where the first term is upper bounded by 1 and the second term is upper bounded by $P_B \log(2^N - 1) < NP_B$, owing to the fact that the maximum of $H(U^N|V^N, E = 1)$ is obtained when U^N is distributed uniformly over all $V^N \neq U^N$. Putting these facts all together, we obtain Fano's inequality $P_B \geq 1 - 1/n - C/R$, where $R = N/n$ is the coding rate. Thus, the DPT directly supports Fano's inequality, which in turn is the main tool for proving converses to channel coding theorems in a large variety of communication situations, including network configurations.

Joint source–channel coding and the separation principle: In a joint source–channel situation, where the source vector U^N is mapped to a channel input vector X^n and the channel output vector Y^n is decoded into a reconstruction V^N, the DPT gives rise to the chain of inequalities

$$NR(D) \leq I(U^N; V^N) \leq I(X^n; Y^n) \leq nC, \tag{133}$$

which is the converse to the joint source–channel coding theorem, whose direct part can be achieved by separate source- and channel coding. Items 1 and 2 above are special cases of this.

Conditioning reduces entropy: Perhaps even more often than the term "data processing theorem" can be found as part of a proof of a converse theorem, one encounters an equivalent of this theorem under the slogan "conditioning reduces entropy". This in turn is part of virtually every converse proof in the literature. Indeed, if (X, U, V) is a triple of random variables, then this statement means that $H(X|V) \geq H(X|U, V)$. If, in addition, $X \to U \to V$ is a Markov chain, then $H(X|U, V) = H(X|U)$, and so, $H(X|V) \geq H(X|U)$, which in turn is equivalent to the more customary form of the DPT, $I(X; U) \geq I(X; V)$, obtained by subtracting $H(X)$ from both sides of the entropy inequality. In fact, as we shall see shortly, it is this entropy inequality that lends itself more naturally to a physical

interpretation. Moreover, we can think of the conditioning–reduces–entropy inequality as another form of the DPT even in the absence of the aforementioned Markov condition, because $X \to (U, V) \to V$ is always a Markov chain.

Turning now to the physics point of view, consider a system which may have two possible Hamiltonians – $\mathcal{E}_0(\boldsymbol{x})$ and $\mathcal{E}_1(\boldsymbol{x})$. Let $Z_i(\beta)$, denote the partition function pertaining to $\mathcal{E}_i(\cdot)$, that is

$$Z_i(\beta) = \sum_{\boldsymbol{x}} e^{-\beta \mathcal{E}_i(\boldsymbol{x})}, \quad i = 0, 1. \tag{134}$$

The *Gibbs' inequality* asserts that

$$\ln Z_1(\beta) \geq \ln Z_0(\beta) + \beta \langle \mathcal{E}_0(\boldsymbol{X}) - \mathcal{E}_1(\boldsymbol{X}) \rangle_0 \tag{135}$$

where $\langle \cdot \rangle_0$ denotes averaging w.r.t. P_0 – the canonical distribution pertaining the Hamiltonian $\mathcal{E}_0(\cdot)$. Equivalently, this inequality can be presented as follows:

$$\langle \mathcal{E}_1(\boldsymbol{X}) - \mathcal{E}_0(\boldsymbol{X}) \rangle_0 \geq \left[-\frac{\ln Z_1(\beta)}{\beta} \right] - \left[-\frac{\ln Z_0(\beta)}{\beta} \right] \equiv F_1 - F_0, \tag{136}$$

where F_i is the free energy pertaining to the canonical ensemble of $\mathcal{E}_i$, $i = 0, 1$.

This inequality is easily proved by defining an Hamiltonian

$$\mathcal{E}_\lambda(\boldsymbol{x}) = (1 - \lambda)\mathcal{E}_0(\boldsymbol{x}) + \lambda \mathcal{E}_1(\boldsymbol{x}) = \mathcal{E}_0(\boldsymbol{x}) + \lambda[\mathcal{E}_1(\boldsymbol{x}) - \mathcal{E}_0(\boldsymbol{x})] \tag{137}$$

and using the convexity of the corresponding log–partition function w.r.t. λ. Specifically, let us define the partition function:

$$Z_\lambda(\beta) = \sum_{\boldsymbol{x}} e^{-\beta \mathcal{E}_\lambda(\boldsymbol{x})}. \tag{138}$$

Now, since $\mathcal{E}_\lambda(\boldsymbol{x})$ is affine in λ, then it is easy to see that $\mathrm{d}^2 \ln Z_\lambda / \mathrm{d}\lambda^2 \geq 0$ (for the same reason that $\mathrm{d}^2 \ln Z(\beta)/\mathrm{d}\beta^2 \geq 0$, as was shown earlier) and so $\ln Z_\lambda(\beta)$ is convex in λ for fixed β. It follows then that the curve of the function $\ln Z_\lambda(\beta)$, as a function of λ, must lie above the straight line that is tangent to this curve at $\lambda = 0$ (see Fig. 4), that is, the graph corresponding to the affine function

$$\ln Z_0(\beta) + \lambda \cdot \left[\frac{\partial \ln Z_\lambda(\beta)}{\partial \lambda} \right]_{\lambda=0}.$$

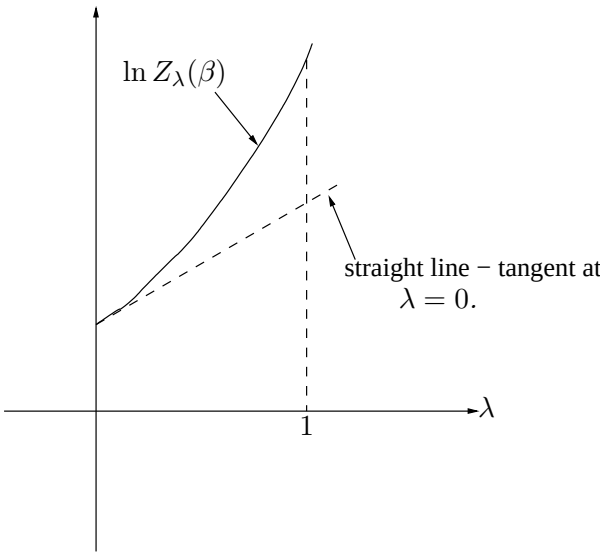


Figure 4: The function $\ln Z_\lambda(\beta)$ is convex in λ and hence lies above its tangent at the origin.

In particular, setting $\lambda = 1$, we get:

$$\ln Z_1(\lambda) \geq \ln Z_0(\beta) + \left.\frac{\partial \ln Z_\lambda(\beta)}{\partial \lambda}\right|_{\lambda=0}, \tag{139}$$

and the second term is:

$$\begin{aligned}\left.\frac{\partial \ln Z_\lambda(\beta)}{\partial \lambda}\right|_{\lambda=0} &= \frac{\beta \sum_{\boldsymbol{x}} [\mathcal{E}_0(\boldsymbol{x}) - \mathcal{E}_1(\boldsymbol{x})] e^{-\beta \mathcal{E}_0(\boldsymbol{x})}}{\sum_{\boldsymbol{x}} e^{-\beta \mathcal{E}_0(\boldsymbol{x})}} \\ &\stackrel{\Delta}{=} \beta \left\langle \mathcal{E}_0(\boldsymbol{X}) - \mathcal{E}_1(\boldsymbol{X}) \right\rangle_0 , \end{aligned} \tag{140}$$

Thus, we have obtained

$$\ln \left[\sum_{\boldsymbol{x}} e^{-\beta \mathcal{E}_1(\boldsymbol{x})} \right] \geq \ln \left[\sum_{\boldsymbol{x}} e^{-\beta \mathcal{E}_0(\boldsymbol{x})} \right] + \beta \left\langle \mathcal{E}_0(\boldsymbol{X}) - \mathcal{E}_1(\boldsymbol{X}) \right\rangle_0 , \tag{141}$$

and the proof is complete. In fact, the difference between the l.h.s. and the r.h.s. is exactly $D(P_0 \| P_1)$, where P_i is the B–G distribution pertaining to $\mathcal{E}_i(\cdot)$, $i = 0, 1$.

We now offer a possible physical interpretation to the Gibbs' inequality: Imagine that a system with Hamiltonian $\mathcal{E}_0(\boldsymbol{x})$ is in equilibrium for all $t < 0$, but then, at time $t = 0$, the Hamiltonian changes *abruptly* from the $\mathcal{E}_0(\boldsymbol{x})$ to $\mathcal{E}_1(\boldsymbol{x})$ (e.g., by suddenly applying a force on the system), which means that if the system is found at state $\boldsymbol{x}$ at time $t = 0$, additional

energy of $W = \mathcal{E}_1(\boldsymbol{x}) - \mathcal{E}_0(\boldsymbol{x})$ is suddenly 'injected' into the system. This additional energy can be thought of as work performed on the system, or as supplementary potential energy. Since this passage between $\mathcal{E}_0$ and $\mathcal{E}_1$ is abrupt, the average of W should be taken w.r.t. P_0, as the state $\boldsymbol{x}$ does not change instantaneously. This average is exactly what we have at the left–hand side eq. (136). The Gibbs inequality tells us then that this average work is at least as large as $\Delta F = F_1 - F_0$, the increase in free energy.[9] The difference $\langle W \rangle_0 - \Delta F$ is due to the irreversible nature of the abrupt energy injection, and this irreversibility means an increase of the total entropy of the system and its environment, and so, the Gibbs' inequality is, in fact, a version of the second law of thermodynamics. This excess work beyond the free–energy increase, $\langle W \rangle_0 - \Delta F$, which can be thought of as the "dissipated work," can easily shown to be equal to $kT \cdot D(P_0 \| P_1)$, where P_0 and P_1 are the canonical distributions pertaining to $\mathcal{E}_0$ and $\mathcal{E}_1$, respectively. Thus, the divergence is given yet another physical significance.

Now, let us see how the Gibbs' inequality is related to the DPT. Consider a triple of random variables $(\boldsymbol{X}, \boldsymbol{U}, \boldsymbol{V})$ which form a Markov chain $\boldsymbol{X} \to \boldsymbol{U} \to \boldsymbol{V}$. The DPT asserts that $I(\boldsymbol{X}; \boldsymbol{U}) \geq I(\boldsymbol{X}; \boldsymbol{V})$. We can obtain the DPT as a special case of the Gibbs inequality as follows: For a given realization $(\boldsymbol{u}, \boldsymbol{v})$ of the random variables $(\boldsymbol{U}, \boldsymbol{V})$, consider the Hamiltonians

$$\mathcal{E}_0(\boldsymbol{x}) = -\ln P(\boldsymbol{x}|\boldsymbol{u}) = -\ln P(\boldsymbol{x}|\boldsymbol{u}, \boldsymbol{v}) \tag{142}$$

and

$$\mathcal{E}_1(\boldsymbol{x}) = -\ln P(\boldsymbol{x}|\boldsymbol{v}). \tag{143}$$

Let us also set $\beta = 1$. Thus, for a given $(\boldsymbol{u}, \boldsymbol{v})$:

$$\begin{aligned} \langle W \rangle_0 &= \langle \mathcal{E}_1(\boldsymbol{X}) - \mathcal{E}_0(\boldsymbol{X}) \rangle_0 \\ &= \sum_{\boldsymbol{x}} P(\boldsymbol{x}|\boldsymbol{u}, \boldsymbol{v}) [\ln P(\boldsymbol{x}|\boldsymbol{u}) - \ln P(\boldsymbol{x}|\boldsymbol{v})] \\ &= H(\boldsymbol{X}|\boldsymbol{V} = \boldsymbol{v}) - H(\boldsymbol{X}|\boldsymbol{U} = \boldsymbol{u}) \end{aligned} \tag{144}$$

[9] This is related to the interpretation of the free–energy difference $\Delta F = F_1 - F_0$ as being the maximum amount of work in an isothermal process.

and after further averaging w.r.t. $(\boldsymbol{U}, \boldsymbol{V})$, the average work becomes

$$H(\boldsymbol{X}|\boldsymbol{V}) - H(\boldsymbol{X}|\boldsymbol{U}) = I(\boldsymbol{X};\boldsymbol{U}) - I(\boldsymbol{X};\boldsymbol{V}). \tag{145}$$

Concerning the free energies, we have

$$Z_0(\beta = 1) = \sum_{\boldsymbol{x}} \exp\{-1 \cdot [-\ln P(\boldsymbol{x}|\boldsymbol{u}, \boldsymbol{v})]\} = \sum_{\boldsymbol{x}} P(\boldsymbol{x}|\boldsymbol{u}, \boldsymbol{v}) = 1 \tag{146}$$

and similarly,

$$Z_1(\beta = 1) = \sum_{\boldsymbol{x}} P(\boldsymbol{x}|\boldsymbol{v}) = 1 \tag{147}$$

which means that $F_0 = F_1 = 0$, and so $\Delta F = 0$ as well. So by the Gibbs inequality, the average work $I(\boldsymbol{X};\boldsymbol{U}) - I(\boldsymbol{X};\boldsymbol{V})$ cannot be smaller than the free–energy difference, which in this case vanishes, namely,

$$I(\boldsymbol{X};\boldsymbol{U}) - I(\boldsymbol{X};\boldsymbol{V}) \geq 0, \tag{148}$$

which is the DPT. Note that in this case, there is a maximum degree of irreversibility: The identity

$$I(\boldsymbol{X};\boldsymbol{U}) - I(\boldsymbol{X};\boldsymbol{V}) = H(\boldsymbol{X}|\boldsymbol{V}) - H(\boldsymbol{X}|\boldsymbol{U}) \tag{149}$$

means that whole work

$$W = I(\boldsymbol{X};\boldsymbol{U}) - I(\boldsymbol{X};\boldsymbol{V}) \tag{150}$$

goes for entropy increase

$$S_1 T - S_0 T = H(\boldsymbol{X}|\boldsymbol{V}) \cdot 1 - H(\boldsymbol{X}|\boldsymbol{U}) \cdot 1, \tag{151}$$

whereas the free energy remains unchanged, as mentioned earlier.

The difference between $I(\boldsymbol{X};\boldsymbol{U})$ and $I(\boldsymbol{X};\boldsymbol{V})$, which accounts for the rate loss in any suboptimal coded communication system, is then given the meaning of irreversibility and entropy production in the corresponding physical system. Optimum (or nearly optimum) communication systems are corresponding to reversible isothermal processes, where the full free energy is exploited and no work is dissipated (or no work is carried out at all, in the first

place). In other words, had there been a communication system that violated the converse to the source/channel coding theorem, one could have created (at least conceptually) a corresponding physical system that violates the second law of thermodynamics, and this, of course, cannot be true.

For a more general physical perspective, let us consider again aforementioned parametric family of Hamiltonians

$$\mathcal{E}_\lambda(\boldsymbol{x}) = \mathcal{E}_0(\boldsymbol{x}) + \lambda[\mathcal{E}_1(\boldsymbol{x}) - \mathcal{E}_0(\boldsymbol{x})] \tag{152}$$

that interpolates linearly between $\mathcal{E}_0(\boldsymbol{x})$ and $\mathcal{E}_1(\boldsymbol{x})$. Here, the control parameter λ can be considered a generalized force. The *Jarzynski equality* [43] (see also [92] and references therein) asserts that under certain assumptions on the system and the environment, and given any protocol for a temporal change in λ, designated by $\{\lambda_t\}$, for which $\lambda_t = 0$ for all $t < 0$, and $\lambda_t = 1$ for all $t \geq \tau$ ($\tau \geq 0$), the work W applied to the system is a random variable that satisfies

$$\boldsymbol{E}\{e^{-\beta W}\} = e^{-\beta \Delta F}. \tag{153}$$

By Jensen's inequality,

$$\boldsymbol{E}\{e^{-\beta W}\} \geq \exp(-\beta \boldsymbol{E}\{W\}), \tag{154}$$

which then gives $\boldsymbol{E}\{W\} \geq \Delta F$, for an arbitrary protocol $\{\lambda_t\}$. The Gibbs inequality is then a special case, where λ_t is given by the unit step function, but it applies regardless of the conditions listed in [92]. At the other extreme, when λ_t changes very slowly, corresponding to a reversible process, W approaches determinism, and then Jensen's inequality becomes tight. In the limit of an arbitrarily slow process, this yields $W = \Delta F$, with no increase in entropy.

Returning now to the realm of information theory, the natural questions are: What is the information–theoretic analogue of Jarzynski's equality? Does it lead to a new generalized version of the information inequality? In other words, if the Gibbs inequality is obtained from Jarzynski's equality for the special case where the protocol $\{\lambda_t\}$ is the unit step function (i.e., $\lambda_t = u(t)$), then what would be the generalized information inequality corresponding

to a general protocol $\{\lambda_t\}$? We next make an attempt to answer these questions at least partially.

First, observe that for $\mathcal{E}_i(x) = -\ln P_i(x)$, $i = 0, 1$, and $\beta = 1$, Jarzynski's equality, for the case $\lambda_t = u(t)$, holds in a straightforward manner:

$$\begin{aligned} \boldsymbol{E}_0\{e^{-W(X)}\} &= \sum_x P_0(x) e^{\ln P_1(x) - \ln P_0(x)} \\ &= 1 = e^{-\Delta F}. \end{aligned} \tag{155}$$

How does this extend to a general protocol $\{\lambda_t\}$? Considering the family of linear interpolations between these two Hamiltonians, let P_0 and P_1 be two probability distributions, and for $\lambda \in [0, 1]$, define

$$P_\lambda(x) = \frac{P_0^{1-\lambda}(x) P_1^\lambda(x)}{Z(\lambda)}, \tag{156}$$

where

$$Z(\lambda) = \sum_x P_0^{1-\lambda}(x) P_1^\lambda(x). \tag{157}$$

This is the Boltzmann distribution pertaining to the Hamiltonian

$$\mathcal{E}_\lambda(x) = (1-\lambda)[-\ln P_0(x)] + \lambda[-\ln P_1(x)]. \tag{158}$$

Now, consider an arbitrary (not necessarily monotonically increasing) sequence of values of λ: $0 \equiv \lambda^0, \lambda^1, \ldots, \lambda^{n-1}, \lambda^n \equiv 1$, and let $\{X_i\}_{i=0}^{n-1}$ be independent random variables, $X_i \sim P_{\lambda^i}$, $i = 0, 1, \ldots, (n-1)$. This corresponds to a protocol where λ_t is a staircase function with jumps of sizes $(\lambda^{i+1} - \lambda^i)$, and where it is also assumed that the plateau segments of λ_t are long enough to let the system equilibrate for each λ^i. Then, we can easily prove a Jarzynski–like equality as follows:

$$\begin{aligned} \boldsymbol{E}\{e^{-W}\} &= \boldsymbol{E}\left\{\exp\left[\sum_{i=0}^{n-1} (\lambda^{i+1} - \lambda^i) \ln \frac{P_1(X_i)}{P_0(X_i)}\right]\right\} \\ &= \prod_{i=0}^{n-1} \boldsymbol{E}_{\lambda_i}\left\{\exp\left[(\lambda^{i+1} - \lambda^i) \ln \frac{P_1(X_i)}{P_0(X_i)}\right]\right\} \\ &= \prod_{i=0}^{n-1} \left(\sum_x P_{\lambda^i}(x) \left[\frac{P_1(x)}{P_0(x)}\right]^{\lambda^{i+1} - \lambda^i}\right) \end{aligned}$$

$$= \prod_{i=0}^{n-1} \frac{Z(\lambda^{i+1})}{Z(\lambda^i)} = \frac{Z(\lambda^n)}{Z(\lambda^0)} = \frac{Z(1)}{Z(0)} = \frac{1}{1} = 1. \tag{159}$$

In the limit of large n, if the density of $\{\lambda^i\}$ grows without bound across the entire unit interval, we get the following information–theoretic version of Jarzynski's equality:

$$\boldsymbol{E}\left\{\exp\left[-\int_0^1 \mathrm{d}\lambda_t \ln \frac{P_0(X_t)}{P_1(X_t)}\right]\right\} = 1, \tag{160}$$

where, again $\{\lambda_t\}$ is an arbitrary protocol, starting at $\lambda_0 = 0$ and ending at $\lambda_\tau = 1$. Applying Jensen's inequality to the left–hand side, we obtain the following generalization of the information inequality for a general protocol:

$$\boldsymbol{E}\{W\} \equiv \int_0^1 \mathrm{d}\lambda_t \cdot \boldsymbol{E}_{\lambda_t}\left\{\ln \frac{P_0(X)}{P_1(X)}\right\} \geq 0, \tag{161}$$

with equality in the case where λ_t is differentiable everywhere, which corresponds to a reversible process. Returning to the simpler case, of finitely many steps, this becomes

$$\boldsymbol{E}\{W\} \equiv \sum_{i=0}^{n-1} (\lambda^{i+1} - \lambda^i) \boldsymbol{E}_{\lambda_i}\left\{\ln \frac{P_0(X)}{P_1(X)}\right\} \geq 0. \tag{162}$$

In this sense, the left–hand side can be thought of as a generalized relative entropy pertaining to an arbitrary protocol.

This inequality has a direct relationship to the behavior of error exponents of hypothesis testing and the Neyman–Pearson lemma: Let P_0 and P_1 be two probability distributions of a random variable X taking values in an alphabet $\mathcal{X}$. Given an observation $x \in \mathcal{X}$, one would like to decide whether it emerged from P_0 or P_1. A decision rule is a partition of $\mathcal{X}$ into two complementary regions $\mathcal{X}_0$ and $\mathcal{X}_1$, such that whenever $X \in \mathcal{X}_i$ one decides in favor of the hypothesis that X has emerged from P_i, $i = 0, 1$. Associated with any decision rule, there are two kinds of error probabilities: $P_0(\mathcal{X}_1)$ is the probability of deciding in favor of P_1 while x has actually been generated by P_0, and $P_1(\mathcal{X}_0)$ is the opposite kind of error. The Neyman–Pearson lemma asserts that the optimum trade-off is given by the likelihood ratio test (LRT) $\mathcal{X}_0^* = (\mathcal{X}_1^*)^c = \{x : \ P_0(x)/P_1(x) \geq \mu\}$, where μ is a parameter that controls the trade-off. Assume now that instead of one observation x, we have a vector $\boldsymbol{x}$ of n i.i.d. observations

$(x_1, \ldots, x_n)$, emerging either all from P_0, or all from P_1. In this case, the error probabilities of the two kinds, pertaining to the LRT, $P_0(\boldsymbol{x})/P_1(\boldsymbol{x}) \geq \mu_n \equiv e^{\theta n}$, can decay asymptotically exponentially, provided that θ is chosen properly, and the asymptotic exponents,

$$e_0 = \lim_{n\to\infty} \left[-\frac{1}{n} \ln P_0(\mathcal{X}_1^*) \right] \tag{163}$$

and

$$e_1 = \lim_{n\to\infty} \left[-\frac{1}{n} \ln P_1(\mathcal{X}_0^*) \right] \tag{164}$$

can be easily found (e.g., by using the method of types) to be

$$e_i(\lambda) = D(P_\lambda \| P_i) = \sum_{x\in\mathcal{X}} P_\lambda(x) \ln \frac{P_\lambda(x)}{P_i(x)}; \quad i = 0, 1 \tag{165}$$

where $P_\lambda(x)$ is defined as before and $\lambda \in [0, 1]$ is determined by θ according to the relation

$$\theta = e_1(\lambda) - e_0(\lambda). \tag{166}$$

Now, the average work W is easily related to the error exponents:

$$\boldsymbol{E}\{W\} = \int_0^1 \mathrm{d}\lambda_t [e_1(\lambda_t) - e_0(\lambda_t)]. \tag{167}$$

Thus, we see that

$$\int_0^1 \mathrm{d}\lambda_t e_1(\lambda_t) \geq \int_0^1 \mathrm{d}\lambda_t e_0(\lambda_t) \tag{168}$$

with equality in the reversible case. Indeed, the last inequality can be also shown to hold using a direct derivation, and the equality is easily shown to hold whenever λ_t is differentiable for every $t \in [0, \tau]$, in which case, it becomes:

$$\int_0^\tau \mathrm{d}t \dot{\lambda}_t e_0(\lambda_t) = \int_0^\tau \mathrm{d}t \dot{\lambda}_t e_1(\lambda_t). \tag{169}$$

The left– (resp. right–) hand side is simply $\int_0^1 \mathrm{d}\lambda e_0(\lambda)$ (resp. $\int_0^1 \mathrm{d}\lambda e_1(\lambda)$) which means that the areas under the graphs of the functions e_0 and e_1 are always the same. This, of course, also means that

$$\int_0^1 \mathrm{d}\lambda \theta(\lambda) = 0, \tag{170}$$

where $\ln \mu(\lambda)$ is defined according to eq. (166). While these integral relations between the error exponent functions could have been derived without recourse to any physical considerations, it is the physical point of view that gives the trigger to point out these relations.

It would be interesting to find additional meanings and utilities for the generalized information inequality proposed here, as well as to figure out whether this is a Shannon-type inequality or a non–Shannon–type inequality [128, Chaps. 13,14]. These questions are beyond the scope of this work, but they are currently under study.

A more detailed exposition of the results of this section, as well as their implications, is provided in [73].

The Gibbs' Inequality and the Log–Sum Inequality

We now wish to take another look at the Gibbs' inequality, from a completely different perspective, namely, as a tool for generating useful bounds on the free energy, in situations where the exact calculation is difficult (see [54, p. 145]). As we show in this part, this inequality is nothing but the *log–sum inequality*, which is used in Information Theory, mostly for proving certain *qualitative* properties of information measures, like the data processing inequality of the divergence [13, Sect. 2.7]. But this equivalence now suggests that the log–sum inequality can perhaps be used in a similar way that it is used in physics, and then it could yield useful bounds on certain information measures. We try to demonstrate this point, first in physics, and then in a problem related to information theory.

Suppose we have an Hamiltonian $\mathcal{E}(\boldsymbol{x})$ for which we wish to know the partition function

$$Z(\beta) = \sum_{\boldsymbol{x}} e^{-\beta \mathcal{E}(\boldsymbol{x})} \tag{171}$$

but it is hard, if not impossible, to calculate in closed–form. Suppose further that for another, somewhat different Hamiltonian, $\mathcal{E}_0(\boldsymbol{x})$, it is rather easy to make calculations. The Gibbs' inequality can be presented as a lower bound on $\ln Z(\beta)$ in terms of B–G statistics pertaining

to $\mathcal{E}_0$.

$$\ln\left[\sum_{\boldsymbol{x}} e^{-\beta\mathcal{E}(\boldsymbol{x})}\right] \geq \ln\left[\sum_{\boldsymbol{x}} e^{-\beta\mathcal{E}_0(\boldsymbol{x})}\right] + \beta\left\langle\mathcal{E}_0(\boldsymbol{X}) - \mathcal{E}(\boldsymbol{X})\right\rangle_0, \tag{172}$$

The idea now is that we can obtain pretty good bounds thanks to the fact that we may have some freedom in the choice of $\mathcal{E}_0$. For example, one can define a parametric family of functions $\mathcal{E}_0$ and maximize the r.h.s. w.r.t. the parameter(s) of this family, thus obtaining the tightest lower bound within the family. Consider the following example.

Example – Non–harmonic oscillator. Consider the potential function

$$V(z) = Az^4 \tag{173}$$

and so

$$\mathcal{E}(x) = \frac{p^2}{2m} + Az^4, \tag{174}$$

where we approximate the second term by

$$V_0(z) = \begin{cases} 0 & |z| \leq \frac{L}{2} \\ +\infty & |z| > \frac{L}{2} \end{cases} \tag{175}$$

where L is a parameter to be optimized. Thus,

$$\begin{aligned} Z_0 &= \frac{1}{h}\int_{-\infty}^{+\infty}\mathrm{d}p\int_{-\infty}^{+\infty}\mathrm{d}z e^{-\beta[V_0(z)+p^2/(2m)]} \\ &= \frac{1}{h}\int_{-\infty}^{+\infty}\mathrm{d}p\cdot e^{-\beta p^2/(2m)}\int_{-L/2}^{+L/2}\mathrm{d}z \\ &= \frac{\sqrt{2\pi mkT}}{h}\cdot L \end{aligned}$$

and so, by the Gibbs inequality:

$$\begin{aligned} \ln Z &\geq \ln Z_0 + \beta\langle\mathcal{E}_0(\boldsymbol{X}) - \mathcal{E}(\boldsymbol{X})\rangle_0 \\ &\geq \ln Z_0 - \frac{1}{kT}\cdot\frac{1}{L}\int_{-L/2}^{+L/2}\mathrm{d}z\cdot Az^4 \\ &\geq \ln\left[\frac{L\sqrt{2\pi mkT}}{h}\right] - \frac{AL^4}{80kT} \\ &\triangleq f(L) \end{aligned}$$

To maximize $f(L)$ we equate its derivative to zero:

$$0 = \frac{\mathrm{d}f}{\mathrm{d}L} \equiv \frac{1}{L} - \frac{AL^3}{20kT} \implies L^* = \left(\frac{20kT}{A}\right)^{1/4}. \tag{176}$$

On substituting this back into the Gibbs lower bound and comparing to the *exact* value of Z (which is computable in this example), we find that $Z_{\text{approx}} \approx 0.91 Z_{\text{exact}}$, which is fairly good, considering the fact that the infinite potential well seems to be quite a poor approximation to the fourth order power law potential $V(z) = Az^4$.

As somewhat better approximation is the harmonic one:

$$V_0(z) = \frac{m\omega_0^2}{2} \cdot z^2 \tag{177}$$

where now ω_0 is the free parameter to be optimized. This gives

$$Z_0 = \frac{1}{h}\int_{-\infty}^{+\infty} \mathrm{d}p \int_{-\infty}^{+\infty} \mathrm{d}z e^{-\beta[m\omega_0^2 z^2/2 + p^2/(2m)]} = \frac{kT}{\hbar\omega_0} \tag{178}$$

and this time, we get:

$$\begin{aligned} \ln Z &\geq \ln\left(\frac{kT}{\hbar\omega_0}\right) + \frac{1}{kT}\left\langle \frac{m\omega_0^2 Z^2}{2} - AZ^4 \right\rangle_0 \\ &= \ln\left(\frac{kT}{\hbar\omega_0}\right) + \frac{1}{2} - \frac{3AkT}{m^2\omega_0^4} \\ &\triangleq f(\omega_0) \end{aligned}$$

To find the maximizing f, we have

$$0 = \frac{\mathrm{d}f}{\mathrm{d}\omega_0} \equiv -\frac{1}{\omega_0} + \frac{12AkT}{m^2\omega_0^5} \implies \omega_0^* = \frac{(12AkT)^{1/4}}{\sqrt{m}}. \tag{179}$$

This time, we get $Z_{\text{approx}} \approx 0.95 Z_{\text{exact}}$, i.e., this approximation is even better. □

Returning from physics to information theory, let us now look at the Gibbs inequality slightly differently. What we actually did, in different notation, is the following: Consider the function:

$$Z(\lambda) = \sum_{i=1}^{n} a_i^{1-\lambda} b_i^{\lambda} = \sum_{i=1}^{n} a_i e^{-\lambda \ln(a_i/b_i)}, \tag{180}$$

where $\{a_i\}$ and $\{b_i\}$ are positive reals. Since $\ln Z(\lambda)$ is convex (as before), we have:

$$\begin{aligned}\ln\left(\sum_{i=1}^{n} b_i\right) &\equiv \ln Z(1) \\ &\geq \ln Z(0) + 1 \cdot \left.\frac{\mathrm{d}\ln Z(\lambda)}{\mathrm{d}\lambda}\right|_{\lambda=0} \\ &= \ln\left(\sum_{i=1}^{n} a_i\right) + \frac{\sum_{i=1}^{n} a_i \ln(b_i/a_i)}{\sum_{i=1}^{n} a_i}\end{aligned}$$

which is nothing but the log–sum inequality. Once again, the idea is to lower bound an expression $\ln(\sum_{i=1}^{n} b_i)$, which may be hard to calculate, by the expression on the l.h.s. which is hopefully easier, and allows a degree of freedom concerning the choice of $\{a_i\}$, at least in accordance to some structure, and depending on a limited set of parameters.

Consider, for example, a hidden Markov model (HMM), which is the output of a DMC $W(\boldsymbol{y}|\boldsymbol{x}) = \prod_{t=1}^{n} W(y_t|x_t)$ fed by a first–order Markov process $\boldsymbol{X}$, governed by $Q(\boldsymbol{x}) = \prod_{t=1}^{n} Q(x_t|x_{t-1})$. The entropy rate of the hidden Markov process $\{Y_t\}$ does not admit a closed–form expression, so we would like to have at least good bounds. The importance of such bounds is self–evident, for example, in the derivation of upper bounds on the capacity of finite–state channels [34, Sect. 4.6]. Here, we propose an upper bound that stems from the Gibbs' inequality, or the log–sum inequality.

The probability distribution of $\boldsymbol{y}$ is

$$P(\boldsymbol{y}) = \sum_{\boldsymbol{x}} \prod_{t=1}^{n} [W(y_t|x_t)Q(x_t|x_{t-1})]. \tag{181}$$

This summation does not lend itself to a nice closed–form expression, but if the t–th factor depended only on t, this would have been easy and simple, as the sum of products would have boiled down to a product of sums. This motivates the following use of the log–sum inequality: For a given $\boldsymbol{y}$, let us think of $\boldsymbol{x}$ as the index i of the log–sum inequality and then

$$b(\boldsymbol{x}) = \prod_{t=1}^{n} [W(y_t|x_t)Q(x_t|x_{t-1})]. \tag{182}$$

Let us now define

$$a(\boldsymbol{x}) = \prod_{t=1}^{n} P_0(x_t, y_t), \tag{183}$$

where P_0 is an arbitrary joint distribution over $\mathcal{X} \times \mathcal{Y}$, to be optimized eventually. Thus, applying the log–sum inequality, we get:

$$\begin{aligned}
\ln P(\boldsymbol{y}) &= \ln \left(\sum_{\boldsymbol{x}} b(\boldsymbol{x}) \right) \\
&\geq \ln \left(\sum_{\boldsymbol{x}} a(\boldsymbol{x}) \right) + \frac{\sum_{\boldsymbol{x}} a(\boldsymbol{x}) \ln[b(\boldsymbol{x})/a(\boldsymbol{x})]}{\sum_{\boldsymbol{x}} a(\boldsymbol{x})} \\
&= \ln \left(\sum_{\boldsymbol{x}} \prod_{t=1}^{n} P_0(x_t, y_t) \right) + \frac{1}{\sum_{\boldsymbol{x}} \prod_{t=1}^{n} P_0(x_t, y_t)} \times \\
&\sum_{\boldsymbol{x}} \left[\prod_{t=1}^{n} P_0(x_t, y_t) \right] \cdot \ln \left[\prod_{t=1}^{n} Q(x_t|x_{t-1}) \frac{W(y_t|x_t)}{P_0(x_t, y_t)} \right] \qquad (184)
\end{aligned}$$

Now, let us denote $P_0(y) = \sum_{x \in \mathcal{X}} P_0(x, y)$, which is the marginal of y under P_0. Then, the first term is simply $\sum_{t=1}^{n} \ln P_0(y_t)$. As for the second term, we have:

$$\begin{aligned}
&\frac{\sum_{\boldsymbol{x}} [\prod_{t=1}^{n} P_0(x_t, y_t)] \cdot \ln[\prod_{t=1}^{n} [Q(x_t|x_{t-1}) W(y_t|x_t)/P_0(x_t, y_t)]}{\sum_{\boldsymbol{x}} \prod_{t=1}^{n} P_0(x_t, y_t)} \\
&= \sum_{t=1}^{n} \sum_{\boldsymbol{x}} \frac{\prod_{t=1}^{n} P_0(x_t, y_t) \ln[Q(x_t|x_{t-1}) W(y_t|x_t)/P_0(x_t, y_t)]}{\prod_{t=1}^{n} P_0(y_t)} \\
&= \sum_{t=1}^{n} \frac{\prod_{t' \neq t-1, t} P_0(y_{t'})}{\prod_{t=1}^{n} P_0(y_t)} \cdot \sum_{x_{t-1}, x_t} P_0(x_{t-1}, y_{t-1}) \times \\
&P_0(x_t, y_t) \cdot \ln \left[\frac{Q(x_t|x_{t-1}) W(y_t|x_t)}{P_0(x_t, y_t)} \right] \\
&= \sum_{t=1}^{n} \sum_{x_{t-1}, x_t} \frac{P_0(x_{t-1}, y_{t-1}) P_0(x_t, y_t)}{P_0(y_{t-1}) P_0(y_t)} \cdot \ln \left[\frac{Q(x_t|x_{t-1}) W(y_t|x_t)}{P_0(x_t, y_t)} \right] \\
&= \sum_{t=1}^{n} \sum_{x_{t-1}, x_t} P_0(x_{t-1}|y_{t-1}) P_0(x_t|y_t) \cdot \ln \left[\frac{Q(x_t|x_{t-1}) W(y_t|x_t)}{P_0(x_t, y_t)} \right] \\
&\stackrel{\Delta}{=} \sum_{t=1}^{n} \boldsymbol{E}_0 \left\{ \ln \left[\frac{Q(X_t|X_{t-1}) W(y_t|X_t)}{P_0(X_t, y_t)} \right] \middle| Y_{t-1} = y_{t-1}, Y_t = y_t \right\}
\end{aligned}$$

where $\boldsymbol{E}_0$ denotes expectation w.r.t. the product measure of P_0. Adding now the first term of the r.h.s. of the log–sum inequality, $\sum_{t=1}^{n} \ln P_0(y_t)$, we end up with the lower bound:

$$\ln P(\boldsymbol{y}) \geq \sum_{t=1}^{n} \boldsymbol{E}_0 \left\{ \ln \left[\frac{Q(X_t|X_{t-1}) W(y_t|X_t)}{P_0(X_t|y_t)} \right] \middle| Y_{t-1} = y_{t-1}, Y_t = y_t \right\}$$

$$\stackrel{\Delta}{=} \sum_{t=1}^{n} \Delta(y_{t-1}, y_t; P_0). \tag{185}$$

At this stage, we can perform the optimization over P_0 for each $\boldsymbol{y}$ individually, and then derive the bound on the expectation of $\ln P(\boldsymbol{y})$ to get a bound on the entropy. Note, however, that $\sum_t \Delta(y_{t-1}, y_t; P_0)$ depends on $\boldsymbol{y}$ only via its Markov statistics, i.e., the relative frequencies of transitions $y \Longrightarrow y'$ for all $y, y' \in \mathcal{Y}$. Thus, the optimum P_0 depends on $\boldsymbol{y}$ also via these statistics. Now, the expectation of $\sum_t \Delta(y_{t-1}, y_t; P_0)$ is dominated by the typical $\{\boldsymbol{y}\}$ for which these transition counts converge to the respective joint probabilities of $\{Y_{t-1} = y,\ Y_t = y\}$. So, it is expected that for large n, nothing will essentially be lost if we first take the expectation over both sides of the log–sum inequality and only then optimize over P_0. This would yield

$$H(Y^n) \leq -n \cdot \max_{P_0} \boldsymbol{E}\{\Delta(Y_0, Y_1; P_0)\}. \tag{186}$$

where the expectation on the r.h.s. is now under the *real* joint distribution of two consecutive samples of $\{Y_n\}$, i.e.,

$$P(y_0, y_1) = \sum_{x_0, x_1} \pi(x_0) Q(x_1|x_0) P(y_0|x_0) P(y_1|x_1), \tag{187}$$

where $\pi(\cdot)$ is the stationary distribution of the underlying Markov process $\{x_t\}$.

We have not pursued this derivation any further from this point, to see if it may yield upper bounds that are tighter than existing ones, for example, the straightforward bound $H(Y^n) \leq nH(Y_1|Y_0)$. This is certainly a question that deserves further study, and if this approach will turn out to be successful, it would be another example how the physical point of view may be beneficial for obtaining new results in information theory.

3.4 Boltzmann's H–Theorem and the DPT

In the previous section, we saw a relationship between the second law of thermodynamics and the DPT, and this relationship was associated with the equilibrium (stationary) probability distribution at any given time instant, namely, the B–G distribution. This is, in a certain

sense, the static point of view. In this section, we explore the relationship between the second law and the DPT from a dynamical point of view.

As said, the B–G distribution is the stationary state distribution over the microstates, but this only a small part of the physical picture. What is missing is the temporal probabilistic behavior, or in other words, the laws that underlie the evolution of the system microstate with time. These are dictated by dynamical properties of the system, which constitute the underlying physical laws in the microscopic level. It is customary then to model the microstate at time t as a random process $\{X_t\}$, where t may denote either discrete time or continuous time, and among the various models, one of the most common ones is the Markov model. This Markov assumption stems from the fact that many physical dynamical systems are modeled by stochastic differential equations, e.g., the Langevin equation and other diffusion equations, whose solutions are Markov processes.

In this section, we discuss a few properties of these processes as well as the evolution of information measures associated with them, like entropy, divergence (and more), and we shall see that these are also intimately related to data processing inequalities. More concretely, we shall see that the second law of the thermodynamics, this time, in its dynamical version, or more precisely, the *Boltzmann H–theorem*, and the data processing inequality, even in the most general form known, are both special cases of a more general principle, which we shall establish here (see also [76] for more details).

We begin with an isolated system in continuous time, which is not necessarily assumed to have reached yet its stationary distribution pertaining to equilibrium. Let us suppose that the state X_t may take on values in a finite set $\mathcal{X}$. For $x, x' \in \mathcal{X}$, let us define the state transition rates

$$W_{xx'} = \lim_{\delta \to 0} \frac{\Pr\{X_{t+\delta} = x' | X_t = x\}}{\delta} \qquad x' \neq x \tag{188}$$

which means, in other words,

$$\Pr\{X_{t+\delta} = x' | X_t = x\} = W_{xx'} \cdot \delta + o(\delta). \tag{189}$$

This is in the spirit (and an extension) of the definition of a Poisson process. Denoting

$$P_t(x) = \Pr\{X_t = x\}, \tag{190}$$

it is easy to see that

$$P_{t+dt}(x) = \sum_{x' \neq x} P_t(x') W_{x'x} \mathrm{d}t + P_t(x) \left(1 - \sum_{x' \neq x} W_{xx'} \mathrm{d}t\right), \tag{191}$$

where the first sum describes the probabilities of all possible transitions from other states to state x and the second term describes the probability of not leaving state x. Subtracting $P_t(x)$ from both sides and dividing by $\mathrm{d}t$, we immediately obtain the following set of differential equations:

$$\frac{\mathrm{d}P_t(x)}{\mathrm{d}t} = \sum_{x'} [P_t(x') W_{x'x} - P_t(x) W_{xx'}], \quad x \in \mathcal{X}, \tag{192}$$

where W_{xx} is defined in an arbitrary manner, e.g., $W_{xx} \equiv 0$ for all $x \in \mathcal{X}$. In the physics terminology (see, e.g., [59],[95]), these equations are called the *master equations*.[10] If we define the incoming probability flux into state x as $J_t^+(x) = \sum_{x'} P_t(x') W_{x'x}$ and the outgoing probability flux from state x as $J_t^-(x) = P_t(x) \sum_{x'} W_{xx'}$, then the master equations tell us that

$$\frac{\mathrm{d}P_t(x)}{\mathrm{d}t} = J_t^+(x) - J_t^-(x),$$

i.e., the rate of change in $P_t(x)$ is given by the net incoming probability flux, which is the difference $J_t^+(x) - J_t^-(x)$. One can think of the total probability as some mass that must be conserved, i.e., whatever is reduced at one place must be added back elsewhere. When the process reaches stationarity, i.e., for all $x \in \mathcal{X}$, $P_t(x)$ converge to some $P(x)$ that is time–invariant, then

$$\sum_{x'} [P(x') W_{x'x} - P(x) W_{xx'}] = 0, \quad \forall\, x \in \mathcal{X}. \tag{193}$$

This situation is called *global balance* or *steady state*. When the physical system under discussion is isolated, namely, no energy flows into the system or out, the steady state

[10] Note that the master equations apply in discrete time too, provided that the derivative at the l.h.s. is replaced by a simple difference, $P_{t+1}(x) - P_t(x)$, and $\{W_{xx'}\}$ are replaced one–step state transition probabilities.

distribution must be uniform across all states, because all accessible states must be of the same energy and the equilibrium probability of each state depends solely on its energy. Thus, in the case of an isolated system, $P(x) = 1/|\mathcal{X}|$ for all $x \in \mathcal{X}$. From quantum mechanical considerations, as well as considerations pertaining to time reversibility in the microscopic level,[11] it is customary to assume $W_{xx'} = W_{x'x}$ for all pairs $\{x, x'\}$. We then observe that, not only do $\sum_{x'}[P(x')W_{x'x} - P(x)W_{xx'}]$ all vanish, but moreover, each individual term in this sum vanishes, as

$$P(x')W_{x'x} - P(x)W_{xx'} = \frac{1}{|\mathcal{X}|}(W_{x'x} - W_{xx'}) = 0. \tag{194}$$

This property is called *detailed balance*, which is stronger than global balance, and it means equilibrium, which is stronger than steady state. While both steady–state and equilibrium refer to situations of time–invariant state probabilities $\{P(x)\}$, a steady–state still allows cyclic "flows of probability." For example, a Markov process with cyclic deterministic transitions $1 \to 2 \to 3 \to 1 \to 2 \to 3 \to \cdots$ is in steady state provided that the probability distribution of the initial state is uniform $(1/3, 1/3, 1/3)$, however, the cyclic flow among the states is in one direction. On the other hand, in detailed balance ($W_{xx'} = W_{x'x}$ for an isolated system), which is equilibrium, there is no net flow in any cycle of states. All the net cyclic probability fluxes vanish, and therefore, time reversal would not change the probability law, that is, $\{X_{-t}\}$ has the same probability law as $\{X_t\}$ (see [57, Sect. 1.2]). Thus, equivalent names for detailed balance are *reversibility* and *time reversal symmetry*. For example, if $\{Y_t\}$ is a Bernoulli process, taking values equiprobably in $\{-1, +1\}$, then X_t defined recursively by

$$X_{t+1} = (X_t + Y_t) \mathrm{mod} K, \tag{195}$$

has a symmetric state–transition probability matrix W, a uniform stationary state distribution, and it satisfies detailed balance. The equivalence between detailed balance and

[11] Consider, for example, an isolated system of moving particles of mass m and position vectors $\{\boldsymbol{r}_i(t)\}$, obeying the differential equations $m\mathrm{d}^2\boldsymbol{r}_i(t)/\mathrm{d}t^2 = \sum_{j\neq i} F(\boldsymbol{r}_j(t) - \boldsymbol{r}_i(t))$, $i = 1, 2, \ldots, n$, ($F(\boldsymbol{r}_j(t) - \boldsymbol{r}_i(t))$ being mutual interaction forces), which remain valid if the time variable t is replaced by $-t$ since $\mathrm{d}^2\boldsymbol{r}_i(t)/\mathrm{d}t^2 = \mathrm{d}^2\boldsymbol{r}_i(-t)/\mathrm{d}(-t)^2$.

time–reversal symmetry is very easy to see in the discrete–time case, where the Markov process is defined in terms of a matrix of one–step transition probabilities $\{P(x'|x)\}$:

$$\begin{aligned} P(x_1, \ldots, x_n) &= P(x_1)P(x_2|x_1)P(x_3|x_2) \cdots P(x_n|x_{n-1}) \\ &= P(x_1|x_2)P(x_2)P(x_3|x_2) \cdots P(x_n|x_{n-1}) \\ &= P(x_1|x_2)P(x_2|x_3)P(x_3)P(x_4|x_3) \cdots P(x_n|x_{n-1}) \\ &= \cdots \\ &= P(x_1|x_2)P(x_2|x_3) \cdots P(x_{n-1}|x_n) \\ &= P(x_n, x_{n-1}, \ldots, x_1) \end{aligned} \tag{196}$$

and the continuous–time analogue is only slightly more complicated, but very similar.

In the general case of detailed balance, i.e., when $P(x)W_{xx'} = P(x')W_{x'x}$, i.e., even if P is not necessarily the uniform distribution, there is a nice physical interpretation of the master equations [57, p. 20]: We can write the master equations as follows:

$$\frac{\mathrm{d}P_t(x)}{\mathrm{d}t} = \sum_{x'} \frac{1}{R_{xx'}} \left[\frac{P_t(x')}{P(x')} - \frac{P_t(x)}{P(x)} \right], \tag{197}$$

where $R_{xx'} = [P(x')W_{x'x}]^{-1} = [P(x)W_{xx'}]^{-1}$. Imagine now an electric circuit where the indices $\{x\}$ designate the various nodes. Nodes x and x' are connected by a wire with resistance $R_{xx'}$ and every node x is grounded via a capacitor with capacitance $P(x)$ (see Fig. 5). If $P_t(x)$ is the charge at node x at time t, then the master equations are the Kirchoff equations of the currents at each node in the circuit. Thus, the way in which probability spreads across the states is analogous to the way that electrical charge spreads across the circuit and probability fluxes are now analogous to electrical currents. In equilibrium, all nodes have the same potential, $P_t(x)/P(x) = 1$, and hence detailed balance corresponds to the situation where all individual currents vanish (not only their algebraic sum).

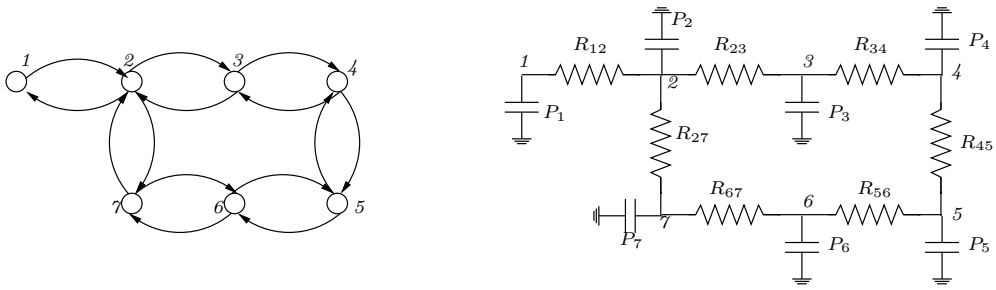


Figure 5: State transition diagram of a Markov chain (left part) and the electric circuit that emulates the dynamics of $\{P_t(x)\}$ (right part).

3.4.1 Monotonicity of Information Measures

Returning to the case where the process $\{X_t\}$ is pertaining to an isolated and the system has not necessarily reached equilibrium, let us take a look at the entropy of the state

$$H(X_t) = -\sum_{x \in \mathcal{X}} P_t(x) \log P_t(x). \tag{198}$$

The Boltzmann H–theorem (see, e.g., [5, Chap. 7], [54, Sect. 3.5], [59, pp. 171–173] [95, pp. 624–626]) asserts that $H(X_t)$ is monotonically non–decreasing. It is important to stress that while this result has the general spirit of the second law of thermodynamics, it is *not* quite the same statement, because $H(X_t)$ is not really the physical entropy of the system outside the regime of equilibrium. The second law simply states that if a system is thermally isolated, then for any process that begins in equilibrium and ends in (possibly, another) equilibrium, the entropy of the *final* state is never smaller than the entropy of the *initial* state, but there is no statement concerning monotonic evolution of the entropy (whatever its definition may be) along the process itself, when the system is out of equilibrium.

To see why the H–theorem is true, we next show that detailed balance implies

$$\frac{\mathrm{d}H(X_t)}{\mathrm{d}t} \geq 0, \tag{199}$$

where for convenience, we denote $\mathrm{d}P_t(x)/\mathrm{d}t$ by $\dot{P}_t(x)$. Now,

$$\frac{\mathrm{d}H(X_t)}{\mathrm{d}t} = -\sum_x [\dot{P}_t(x) \log P_t(x) + \dot{P}_t(x)]$$

$$
\begin{aligned}
&= -\sum_x \dot{P}_t(x) \log P_t(x) \\
&= -\sum_x \sum_{x'} W_{x'x} [P_t(x') - P_t(x)] \log P_t(x)) \\
&= -\frac{1}{2} \sum_{x,x'} W_{x'x} [P_t(x') - P_t(x)] \log P_t(x) - \\
&\frac{1}{2} \sum_{x,x'} W_{x'x} [P_t(x) - P_t(x')] \log P_t(x') \\
&= \frac{1}{2} \sum_{x,x'} W_{x'x} [P_t(x') - P_t(x)] \cdot [\log P_t(x') - \log P_t(x)] \\
&\geq 0, \qquad (200)
\end{aligned}
$$

where in the second line we used the fact that $\sum_x \dot{P}_t(x) = 0$, in the third line we used detailed balance ($W_{xx'} = W_{x'x}$), and the last inequality is due to the increasing monotonicity of the logarithmic function: the product $[P_t(x') - P_t(x)] \cdot [\log P_t(x') - \log P_t(x)]$ cannot be negative for any pair (x, x'), as the two factors of this product are either both negative, both zero, or both positive. Thus, $H(X_t)$ cannot decrease with time. The H–theorem has a discrete–time analogue: If a finite–state Markov process has a symmetric transition probability matrix (which is the discrete–time counterpart of the above detailed balance property), which means that the stationary state distribution is uniform, then $H(X_t)$ is a monotonically non–decreasing sequence.

A well–known paradox, in this context, is associated with the notion of the *arrow of time.* On the one hand, we are talking about time–reversible processes, obeying detailed balance, but on the other hand, the increase of entropy suggests that there is asymmetry between the two possible directions that the time axis can be exhausted, the forward direction and the backward direction. If we go back in time, the entropy would decrease. So is there an arrow of time? This paradox was resolved, by Boltzmann himself, once he made the clear distinction between equilibrium and non–equilibrium situations: The notion of time reversibility is associated with equilibrium, where the process $\{X_t\}$ is stationary. On the other hand, the increase of entropy is a result that belongs to the non–stationary regime, where the process is on its way to stationarity and equilibrium. In the latter case, the system

has been initially prepared in a non–equilibrium situation. Of course, when the process is stationary, $H(X_t)$ is fixed and there is no contradiction.

As mentioned earlier, for a general Markov process, whose steady state–distribution is not necessarily uniform, the condition of detailed balance, which means time–reversibility [57], reads

$$P(x)W_{xx'} = P(x')W_{x'x}, \tag{201}$$

in the continuous–time case. In the discrete–time case (where t takes on positive integer values only), it is defined by a similar equation, except that $W_{xx'}$ and $W_{x'x}$ are replaced by the corresponding one–step state transition probabilities, i.e.,

$$P(x)P(x'|x) = P(x')P(x|x'), \tag{202}$$

where

$$P(x'|x) \stackrel{\Delta}{=} \Pr\{X_{t+1} = x'|X_t = x\}. \tag{203}$$

The physical interpretation is that now our system is (a small) part of a much larger isolated system, which obeys detailed balance w.r.t. the uniform equilibrium distribution, as before. A well known example of a process that obeys detailed balance in its more general form is the M/M/1 queue with an arrival rate λ and service rate μ ($\lambda < \mu$). Here, since all states are arranged along a line, with bidirectional transitions between neighboring states only (see Fig. 6), there cannot be any cyclic probability flux. The steady–state distribution is well–known to be geometric

$$P(x) = \left(1 - \frac{\lambda}{\mu}\right) \cdot \left(\frac{\lambda}{\mu}\right)^x, \qquad x = 0, 1, 2, \ldots, \tag{204}$$

which indeed satisfies the detailed balance $P(x)\lambda = P(x+1)\mu$ for all x. Thus, the Markov process $\{X_t\}$, designating the number of customers in the queue at time t, is time–reversible.

For the sake of simplicity, from this point onward, our discussion will focus almost exclusively on discrete–time Markov processes, but the results to be stated, will hold for continuous–time Markov processes as well. We will continue to denote by $P_t(x)$ the probability of $X_t = x$, except that now t will be limited to take on integer values only. The one–step state transition probabilities will be denoted by $\{P(x'|x)\}$, as mentioned earlier.

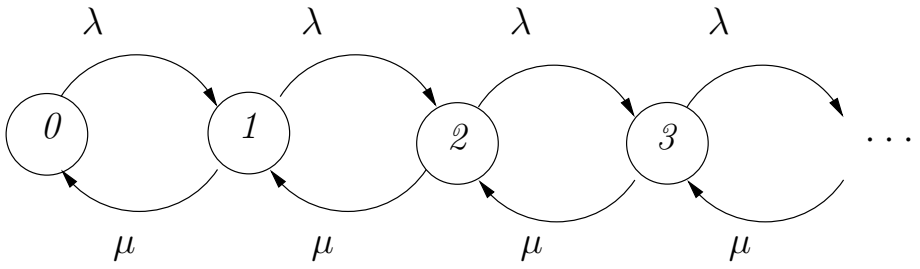


Figure 6: State transition diagram of an M/M/1 queue.

How does the H–theorem extend to situations where the stationary state distribution is not uniform? In [13, p. 82], it is shown (among other things) that the divergence,

$$D(P_t\|P) = \sum_{x\in\mathcal{X}} P_t(x) \log \frac{P_t(x)}{P(x)}, \tag{205}$$

where $P = \{P(x),\ x \in \mathcal{X}\}$ is a stationary state distribution, is a monotonically non–increasing function of t. Does this result have a physical interpretation, like the H–theorem and its connotation with the second law of thermodynamics? When it comes to non–isolated systems, where the steady state distribution is non–uniform, the extension of the second law of thermodynamics, replaces the principle of increase of entropy by the principle of decrease of free energy, or equivalently, the decrease of the difference between the free energy at time t and the free energy in equilibrium. The information–theoretic counterpart of this free energy difference is the divergence $D(P_t\|P)$, as we have seen earlier. Thus, the monotonic decrease of $D(P_t\|P)$ has a simple physical interpretation in the spirit of of free energy decrease, which is the natural extension of the entropy increase.[12] Indeed, particularizing this to the case where P is the uniform distribution (as in an isolated system), then

$$D(P_t\|P) = \log|\mathcal{X}| - H(X_t), \tag{206}$$

which means that the decrease of the divergence is equivalent to the increase of entropy, as before. However, here the result is more general than the H–theorem from an additional aspect: It does not require detailed balance. It only requires the existence of the stationary state distribution. Note that even in the earlier case, of an isolated system, detailed balance,

[12]Once again, we reiterate that a similar digression as before applies here too: The free energy is defined only in equilibrium, thus $D(P_t\|P)$ is not really the free energy out of equilibrium.

which means symmetry of the state transition probability matrix ($P(x'|x) = P(x|x')$), is a stronger requirement than uniformity of the stationary state distribution, as the latter requires merely that the matrix $\{P(x'|x)\}$ would be doubly stochastic, i.e., $\sum_x P(x|x') = \sum_x P(x'|x) = 1$ for all $x' \in \mathcal{X}$, which is weaker than symmetry of the matrix itself. The results shown in [13] are, in fact, somewhat more general: Let $P_t = \{P_t(x)\}$ and $P'_t = \{P'_t(x)\}$ be two time–varying state–distributions pertaining to the same Markov chain, but induced by two different initial state distributions, $\{P_0(x)\}$ and $\{P'_0(x)\}$, respectively. Then $D(P_t\|P'_t)$ is monotonically non–increasing. This is easily seen as follows:

$$\begin{aligned} D(P_t\|P'_t) &= \sum_x P_t(x) \log \frac{P_t(x)}{P'_t(x)} \\ &= \sum_{x,x'} P_t(x)P(x'|x) \log \frac{P_t(x)P(x'|x)}{P'_t(x)P(x'|x)} \\ &= \sum_{x,x'} P(X_t = x,\ X_{t+1} = x') \log \frac{P(X_t = x,\ X_{t+1} = x')}{P'(X_t = x,\ X_{t+1} = x')} \\ &\geq D(P_{t+1}\|P'_{t+1}) \end{aligned} \tag{207}$$

where the last inequality follows from the data processing theorem of the divergence: the divergence between two joint distributions of (X_t, X_{t+1}) is never smaller than the divergence between corresponding marginal distributions of X_{t+1}. Another interesting special case of this result is obtained if we now take the first argument of the divergence to the a stationary state distribution: This will mean that $D(P\|P_t)$ is also monotonically non–increasing.

In [57, Theorem 1.6], there is a further extension of all the above monotonicity results, where the ordinary divergence is actually replaced by the so called f–divergence, a generalized divergence which was invented by Csiszár [15] (though the relation to the f–divergence is not mentioned in [57]): In a nutshell, the f–divergence between two probability distributions P_1 and P_2 is defined similarly as the ordinary divergence, $D(P_1\|P_2) = \sum_x P_1(x)\left[-\log(P_2(x)/P_1(x))\right]$, except that the negative logarithm function is replaced by a general convex function f, that is

$$D_f(P_1\|P_2) = \sum_x P_1(x) f\left(\frac{P_2(x)}{P_1(x)}\right), \tag{208}$$

but throughout the sequel, we will denote the general convex function by Q rather than f. For a pair of correlated random variables, if P_1 is taken to be their joint distribution and P_2 is the product of their marginals, then D_f amounts to a generalized mutual information measure, which is well known to admit a data processing inequality [15], [131].

Returning now to [57, Theorem 1.6], if $\{X_t\}$ is a Markov process with a given state transition probability matrix $\{P(x'|x)\}$, then the function

$$U(t) = D_Q(P\|P_t) = \sum_{x\in\mathcal{X}} P(x)\cdot Q\left(\frac{P_t(x)}{P(x)}\right) \tag{209}$$

is monotonically non–increasing, provided that Q is convex. Moreover, $U(t)$ monotonically strictly decreasing if Q is strictly convex and $\{P_t(x)\}$ is not identical to $\{P(x)\}$. To see why this is true, define the backward transition probability matrix by

$$\tilde{P}(x|x') = \frac{P(x)P(x'|x)}{P(x')}. \tag{210}$$

Obviously,

$$\sum_x \tilde{P}(x|x') = 1 \tag{211}$$

for all $x'\in\mathcal{X}$, and so,

$$\frac{P_{t+1}(x)}{P(x)} = \sum_{x'} \frac{P_t(x')P(x|x')}{P(x)} = \sum_{x'} \frac{\tilde{P}(x'|x)P_t(x')}{P(x')}. \tag{212}$$

By the convexity of Q:

$$\begin{aligned}
U(t+1) &= \sum_x P(x)\cdot Q\left(\frac{P_{t+1}(x)}{P(x)}\right)\\
&= \sum_x P(x)\cdot Q\left(\sum_{x'}\tilde{P}(x'|x)\frac{P_t(x')}{P(x')}\right)\\
&\le \sum_x\sum_{x'} P(x)\tilde{P}(x'|x)\cdot Q\left(\frac{P_t(x')}{P(x')}\right)\\
&= \sum_x\sum_{x'} P(x')P(x|x')\cdot Q\left(\frac{P_t(x')}{P(x')}\right)\\
&= \sum_{x'} P(x')\cdot Q\left(\frac{P_t(x')}{P(x')}\right) = U(t).
\end{aligned} \tag{213}$$

Now, a few interesting choices of the function Q may be considered: As proposed in [57, p. 19], for $Q(u) = u \ln u$, we have $U(t) = D(P_t \| P)$, and we are back to the aforementioned result in [13]. Another interesting choice is $Q(u) = -\ln u$, which gives $U(t) = D(P \| P_t)$. Thus, the monotonicity of $D(P \| P_t)$ is also obtained as a special case.[13] Yet another choice is $Q(u) = -u^s$, where $s \in [0, 1]$ is a parameter. This would yield the increasing monotonicity of $\sum_x P^{1-s}(x) P_t^s(x)$, a 'metric' that plays a role in the theory of asymptotic exponents of error probabilities pertaining to the optimum likelihood ratio test between two probability distributions [119, Chapter 3]. In particular, the choice $s = 1/2$ yields balance between the two kinds of error and it is intimately related to the Bhattacharyya distance. Returning, for a moment, to the realm of continuous–time Markov processes in detailed balance, with the above–described electrical circuit analogue, then if we now choose $Q(u) = \frac{1}{2}u^2$, then

$$U(t) = \frac{1}{2} \sum_x \frac{P_t^2(x)}{P(x)}, \tag{214}$$

which means that the energy stored in the capacitors dissipates as heat in the wires until the system reaches equilibrium.

We have seen, in the above examples, that various choices of the function Q yield various f–divergences, or 'metrics', between $\{P(x))\}$ and $\{P_t(x)\}$, which are both marginal distributions of a single symbol x. What about joint distributions of two or more symbols? Consider, for example, the function

$$J(t) = \sum_{x,x'} P(X_0 = x, X_t = x') \cdot Q\left(\frac{P(X_0 = x) P(X_t = x')}{P(X_0 = x, X_t = x')} \right), \tag{215}$$

where Q is convex as before. Here, by the same token, $J(t)$ is the f–divergence between the joint probability distribution $\{P(X_0 = x, X_t = x')\}$ and the product of marginals $\{P(X_0 = x) P(X_t = x')\}$, namely, it is the generalized mutual information of [15] and [131], as mentioned earlier. Now, using a similar chain of inequalities as before, we get the

[13] We are not yet in a position to obtain the monotonicity of $D(P_t \| P'_t)$ as a special case of the monotonicity of $D_Q(P \| P_t)$. This will require a slight further extension of this information measure, to be carried out later on.

non–decreasing monotonicity of $J(t)$ as follows:

$$
\begin{aligned}
J(t) &= \sum_{x,x',x''} P(X_0 = x, X_t = x', X_{t+1} = x'') \times \\
&\quad Q\left(\frac{P(X_0 = x) P(X_t = x')}{P(X_0 = x, X_t = x')} \cdot \frac{P(X_{t+1} = x'' | X_t = x')}{P(X_{t+1} = x'' | X_t = x')} \right) \\
&= \sum_{x,x''} P(X_0 = x, X_{t+1} = x'') \sum_{x'} P(X_t = x' | X_0 = x, X_{t+1} = x'') \times \\
&\quad Q\left(\frac{P(X_0 = x) P(X_t = x', X_{t+1} = x'')}{P(X_0 = x, X_t = x', X_{t+1} = x'')} \right) \\
&\leq \sum_{x,x''} P(X_0 = x, X_{t+1} = x'') \times \\
&\quad Q\left(\sum_{x'} P(X_t = x' | X_0 = x, X_{t+1} = x'') \times \right. \\
&\quad \left. \frac{P(X_0 = x) P(X_t = x', X_{t+1} = x'')}{P(X_0 = x, X_t = x', X_{t+1} = x'')} \right) \\
&= \sum_{x,x''} P(X_0 = x, X_{t+1} = x'') \times \\
&\quad Q\left(\sum_{x'} \frac{P(X_0 = x) P(X_t = x', X_{t+1} = x'')}{P(X_0 = x, X_{t+1} = x'')} \right) \\
&= \sum_{x,x''} P(X_0 = x, X_{t+1} = x'') \cdot Q\left(\frac{P(X_0 = x) P(X_{t+1} = x'')}{P(X_0 = x, X_{t+1} = x'')} \right) \\
&= J(t+1). \qquad (216)
\end{aligned}
$$

This time, we assumed only the Markov property of (X_0, X_t, X_{t+1}) (not even homogeneity). This is, in fact, nothing but the generalized data processing theorem of Ziv and Zakai [131]. This data processing theorem sometimes gives tighter lower bounds on the distortion of simple joint source–channel codes. For example, consider a binary symmetric source and an erasure channel with erasure probaliblity p. Using the convex function $Q(t) = 1/t$, one obtains the lower bound $D \geq (1 - \sqrt{1-p})/2$, which is tighter (larger) than the ordinary data processing bound of $D \geq h^{-1}(p)$ (see [131] for more details on this example as well as a few other examples).

3.4.2 A Unified Framework

In spite of the general resemblance (via the notion of the f–divergence), the last monotonicity result, concerning $J(t)$, and the monotonicity of $D(P_t\|P'_t)$, do not seem, at first glance, to fall in the framework of the monotonicity of the f–divergence $D_Q(P\|P_t)$. This is because in the latter, there is an additional dependence on a stationary state distribution that appears neither in $D(P_t\|P'_t)$ nor in $J(t)$. However, two simple observations can put them both in the framework of the monotonicity of $D_Q(P\|P_t)$.

The first observation is that the monotonicity of $U(t) = D_Q(P\|P_t)$ continues to hold (with a straightforward extension of the proof) if $P_t(x)$ is extended to be a vector of time varying state distributions $(P_t^1(x), P_t^2(x), \ldots, P_t^k(x))$, and Q is taken to be a convex function of k variables. Moreover, each component $P_t^i(x)$ does not have to be necessarily a probability distribution. It can be any function $\mu_t^i(x)$ that satisfies the recursion

$$\mu_{t+1}^i(x) = \sum_{x'} \mu_t^i(x')P(x|x'), \qquad 1 \le i \le k. \tag{217}$$

Let us then denote $\boldsymbol{\mu}_t(x) = (\mu_t^1(x), \mu_t^2(x), \ldots, \mu_t^k(x))$ and assume that Q is jointly convex in all its k arguments. Then the redefined function

$$\begin{aligned} U(t) &= \sum_{x\in\mathcal{X}} P(x) \cdot Q\left(\frac{\boldsymbol{\mu}_t(x)}{P(x)}\right) \\ &= \sum_{x\in\mathcal{X}} P(x) \cdot Q\left(\frac{\mu_t^1(x)}{P(x)}, \ldots, \frac{\mu_t^k(x)}{P(x)}\right) \end{aligned} \tag{218}$$

is monotonically non–increasing with t.

The second observation is rooted in convex analysis, and it is related to the notion of the perspective of a convex function and its convexity property [8]. Here, a few words of background are in order. Let $Q(\boldsymbol{u})$ be a convex function of the vector $\boldsymbol{u} = (u_1, \ldots, u_k)$ and let $v > 0$ be an additional variable. Then, the function

$$\tilde{Q}(v, u_1, u_2, \ldots, u_k) \stackrel{\Delta}{=} v \cdot Q\left(\frac{u_1}{v}, \frac{u_2}{v}, \ldots, \frac{u_k}{v}\right) \tag{219}$$

is called the *perspective function* of Q. A well–known property of the perspective operation is conservation of convexity, in other words, if Q is convex in $\boldsymbol{u}$, then $\tilde{Q}$ is convex in $(v, \boldsymbol{u})$.

The proof of this fact, which is straightforward, can be found, for example, in [8, p. 89, Subsection 3.2.6] (see also [17]) and it is brought here for the sake of completeness: Letting λ_1 and λ_2 be two non–negative numbers summing to unity and letting $(v_1, \boldsymbol{u}_1)$ and $(v_2, \boldsymbol{u}_2)$ be given, then

$$\begin{aligned}
&\tilde{Q}(\lambda_1(v_1, \boldsymbol{u}_1) + \lambda_2(v_2, \boldsymbol{u}_2)) \\
&= (\lambda_1 v_1 + \lambda_2 v_2) \cdot Q\left(\frac{\lambda_1 \boldsymbol{u}_1 + \lambda_2 \boldsymbol{u}_2}{\lambda_1 v_1 + \lambda_2 v_2}\right) \\
&= (\lambda_1 v_1 + \lambda_2 v_2) \cdot Q\left(\frac{\lambda_1 v_1}{\lambda_1 v_1 + \lambda_2 v_2} \cdot \frac{\boldsymbol{u}_1}{v_1} + \frac{\lambda_2 v_2}{\lambda_1 v_1 + \lambda_2 v_2} \cdot \frac{\boldsymbol{u}_2}{v_2}\right) \\
&\leq \lambda_1 v_1 Q\left(\frac{\boldsymbol{u}_1}{v_1}\right) + \lambda_2 v_2 Q\left(\frac{\boldsymbol{u}_2}{v_2}\right) \\
&= \lambda_1 \tilde{Q}(v_1, \boldsymbol{u}_1) + \lambda_2 \tilde{Q}(v_2, \boldsymbol{u}_2). \qquad (220)
\end{aligned}$$

Putting these two observations together, we can now state the following result:

Theorem 1 *Let*

$$V(t) = \sum_x \mu_t^0(x) Q\left(\frac{\mu_t^1(x)}{\mu_t^0(x)}, \frac{\mu_t^2(x)}{\mu_t^0(x)}, \ldots, \frac{\mu_t^k(x)}{\mu_t^0(x)}\right), \qquad (221)$$

where Q is a convex function of k variables and $\{\mu_t^i(x)\}_{i=0}^k$ are arbitrary functions that satisfy the recursion

$$\mu_{t+1}^i(x) = \sum_{x'} \mu_t^i(x') P(x|x'), \qquad i = 0, 1, 2, \ldots, k, \qquad (222)$$

and where $\mu_t^0(x)$ is moreover strictly positive. Then, $V(t)$ is a monotonically non–increasing function of t.

Using the above mentioned observations, the proof of Theorem 1 is straightforward: Letting P be a stationary state distribution of $\{X_t\}$, we have:

$$\begin{aligned}
V(t) &= \sum_x \mu_t^0(x) Q\left(\frac{\mu_t^1(x)}{\mu_t^0(x)}, \frac{\mu_t^2(x)}{\mu_t^0(x)}, \ldots, \frac{\mu_t^k(x)}{\mu_t^0(x)}\right) \\
&= \sum_x P(x) \cdot \frac{\mu_t^0(x)}{P(x)} Q\left(\frac{\mu_t^1(x)/P(x)}{\mu_t^0(x)/P(x)}, \ldots, \frac{\mu_t^k(x)/P(x)}{\mu_t^0(x)/P(x)}\right) \\
&= \sum_x P(x) \tilde{Q}\left(\frac{\mu_t^0(x)}{P(x)}, \frac{\mu_t^1(x)}{P(x)}, \ldots, \frac{\mu_t^k(x)}{P(x)}\right). \qquad (223)
\end{aligned}$$

Since $\tilde{Q}$ is the perspective of the convex function Q, then it is convex as well, and so, the monotonicity of $V(t)$ follows from the first observation above. It is now readily seen that both $D(P_t\|P_t')$ and $J(t)$ are special cases of $V(t)$ and hence we have covered all special cases under the umbrella of the more general information functional $V(t)$.

In analogy to the link between $J(t)$ and the generalized information measure of [15] and [131], a similar link can be drawn between $V(t)$ and an even more general mutual information measure [129] that also admits a data processing inequality. For a given pair of random variables, this generalized information measure is defined by the same expression as $V(t)$, where μ_t^0 plays the role of the joint probability distribution of this pair of random variables, and $\mu_t^1, \ldots, \mu_t^k$ are arbitrary measures associated with this pair. This gives rise to a new look at the generalized data processing theorem, which suggests to exploit certain degrees of freedom that may lead to better bounds, for a given choice of the convex function that defines the generalized mutual information. This point is demonstrated in detail in [76].

3.5 Generalized Temperature and Fisher Information

The last information measure to be discussed in this chapter is the Fisher information. As argued in [85], the Fisher information turns out to be intimately related to a generalized notion of temperature, pertaining to non–equilibrium situations. In this section, we summarize the main derivations and findings of [85], where the main mathematical tool is an interesting modified version of de Bruijn's identity (see, e.g., [13, Sect. 17.7]. Specifically, the ordinary version of de Bruin's identity relates the Fisher information to the derivative of differential entropy of a random variable $X + \sqrt{\delta}Z$ w.r.t. δ, where X and Z are independent and Z is Gaussian. On the other hand, the modified de Bruijn identity of [85] is more general in the sense of allowing Z to have any density with zero mean, but less general in the sense of limiting the applicability of the identity to the vicinity of $\delta = 0$.

We begin by recalling the definition of temperature according to

$$\frac{1}{T} = \left(\frac{\partial S}{\partial E}\right)_V . \tag{224}$$

This definition corresponds to equilibrium. As we know, when the Hamiltonian is quadratic, i.e., $\mathcal{E}(x) = \frac{\alpha}{2}x^2$, the Boltzmann distribution, which is the equilibrium distribution, is Gaussian:

$$P(\boldsymbol{x}) = \frac{1}{Z(\beta)} \exp\left\{-\beta \cdot \frac{\alpha}{2}\sum_{i=1}^{N} x_i^2\right\} \tag{225}$$

and by the energy equipartition theorem, the average internal energy is given by

$$\bar{E}(P) \stackrel{\Delta}{=} \left\langle \frac{\alpha}{2}\sum_{i=1}^{N} X_i^2 \right\rangle_P = \frac{NkT}{2}. \tag{226}$$

In Chapter 2, we also computed the entropy, which is nothing but the entropy of a Gaussian vector $S(P) = \frac{Nk}{2}\ln(\frac{2\pi e}{\alpha\beta})$.

Consider now another probability density function $Q(\boldsymbol{x})$, which means a non–equilibrium probability law if it differs from P. Consider now the energy and the entropy pertaining to Q:

$$\bar{E}(Q) = \left\langle \frac{\alpha}{2}\sum_{i=1}^{N} X_i^2 \right\rangle_Q = \int \mathrm{d}\boldsymbol{x} Q(\boldsymbol{x}) \cdot \left[\frac{\alpha}{2}\sum_{i=1}^{N} x_i^2\right] \tag{227}$$

$$S(Q) = k \cdot \langle -\ln Q(\boldsymbol{X})\rangle_Q = -k\int \mathrm{d}\boldsymbol{x} Q(\boldsymbol{x}) \ln Q(\boldsymbol{x}), \tag{228}$$

where again, we remind the reader that this definition of entropy may be questionable in absence of equilibrium, as was mentioned in Chapter 2. In order to define a notion of generalized temperature, we have to establish a notion of a derivative of $S(Q)$ w.r.t. $\bar{E}(Q)$. Such a definition may make sense if it turns out that the ratio between the response of S to perturbations in Q and the response of $\bar{E}$ to the same perturbations, is independent of the "direction" of this perturbation, as long as it is "small" in some reasonable sense. It turns out the de Bruijn identity, in its modified version described above, helps us here.

Consider the perturbation of $\boldsymbol{X}$ by $\sqrt{\delta}\boldsymbol{Z}$ thus defining the perturbed version of $\boldsymbol{X}$ as $\boldsymbol{X}_\delta = \boldsymbol{X} + \sqrt{\delta}\boldsymbol{Z}$, where $\delta > 0$ is small and $\boldsymbol{Z}$ is an *arbitrary* i.i.d. zero–mean random vector, *not necessarily Gaussian*, whose components all have unit variance. Let Q_δ denote the density of $\boldsymbol{X}_\delta$, which is, of course, the convolution between Q and the density of Z,

scaled by $\sqrt{\delta}$. The proposed generalized definition of temperature is:

$$\frac{1}{T} \stackrel{\Delta}{=} \lim_{\delta \to 0} \frac{S(Q_\delta) - S(Q)}{\bar{E}(Q_\delta) - \bar{E}(Q)}. \tag{229}$$

The denominator is easy to handle since

$$\boldsymbol{E}\|\boldsymbol{X} + \sqrt{\delta}\boldsymbol{Z}\|^2 - \boldsymbol{E}\|\boldsymbol{X}\|^2 = 2\sqrt{\delta}\boldsymbol{E}\boldsymbol{X}^T\boldsymbol{Z} + N\delta = N\delta \tag{230}$$

and so, $\bar{E}(Q_\delta) - \bar{E}(Q) = N\alpha\delta/2$. In view of the above, our new definition of temperature becomes:

$$\begin{aligned} \frac{1}{T} &\stackrel{\Delta}{=} \frac{2k}{N\alpha} \cdot \lim_{\delta \to 0} \frac{h(\boldsymbol{X} + \sqrt{\delta}\boldsymbol{Z}) - h(\boldsymbol{X})}{\delta} \\ &= \frac{2k}{N\alpha} \cdot \frac{\partial h(\boldsymbol{X} + \sqrt{\delta}\boldsymbol{Z})}{\partial \delta}\bigg|_{\delta=0}. \end{aligned} \tag{231}$$

First, it is important to understand that the numerator of the middle expression is positive (and hence so is T) since

$$S(Q_\delta) = kh(\boldsymbol{X} + \sqrt{\delta}\boldsymbol{Z}) \geq kh(\boldsymbol{X} + \sqrt{\delta}\boldsymbol{Z}|\boldsymbol{Z}) = kh(\boldsymbol{X}) = S(Q). \tag{232}$$

In order to move forward from this point, we will need the aforementioned modified version of de Bruijn's identity. Suppose we have a family of pdf's $\{Q_\theta(x)\}$ where θ is a continuous valued parameter. The Fisher information is defined as

$$J(\theta) = \boldsymbol{E}_\theta \left\{ \left[\frac{\partial \ln Q_\theta(X)}{\partial \theta} \right]^2 \right\} = \int_{-\infty}^{+\infty} \frac{\mathrm{d}x}{Q_\theta(x)} \left[\frac{\partial}{\partial \theta} Q_\theta(x) \right]^2, \tag{233}$$

where $\boldsymbol{E}_\theta\{\cdot\}$ denotes expectation w.r.t. Q_θ. Consider now the special case where θ is a shift parameter (a.k.a. location parameter), i.e., $Q_\theta(x) = Q(x - \theta)$, then

$$\begin{aligned} J(\theta) &= \int_{-\infty}^{+\infty} \frac{\mathrm{d}x}{Q(x-\theta)} \left[\frac{\partial}{\partial \theta} Q(x-\theta) \right]^2 \\ &= \int_{-\infty}^{+\infty} \frac{\mathrm{d}x}{Q(x-\theta)} \left[\frac{\partial}{\partial x} Q(x-\theta) \right]^2 \\ &= \int_{-\infty}^{+\infty} \frac{\mathrm{d}x}{Q(x)} \left[\frac{\partial}{\partial x} Q(x) \right]^2 \\ &\stackrel{\Delta}{=} J \end{aligned} \tag{234}$$

independently of θ. As J is merely a functional of Q, we will henceforth denote it as $J(Q)$, with a slight abuse of notation. For the vector case, we define the Fisher information matrix, whose elements are

$$J_{ij}(Q) = \int_{\mathbb{R}^N} \frac{\mathrm{d}\boldsymbol{x}}{Q(\boldsymbol{x})} \left[\frac{\partial Q(\boldsymbol{x})}{\partial x_i} \cdot \frac{\partial Q(\boldsymbol{x})}{\partial x_j}\right] \quad i,j = 1,\ldots,N. \tag{235}$$

Shortly, we will relate T with the trace of this matrix.

To this end, we will need the following result, which, as described before, is a variant of the well–known *de Bruijn identity*, beginning with the scalar case: Let Q be the pdf of a scalar random variable X of finite variance. Let Z be a zero–mean, unit variance random variable, independent of X, and let $X_\delta = X + \sqrt{\delta}Z$. Then, it is proved in [85] that

$$\left.\frac{\partial h(X+\sqrt{\delta}Z)}{\partial \delta}\right|_{\delta=0} = \frac{J(Q)}{2}. \tag{236}$$

As explained in the introductory paragraph above, the original de Bruijn identity allows only a Gaussian perturbation Z, but it holds for any δ. Here, on the other hand, we allow an arbitrary density $M(z)$ of Z, but we insist on $\delta \to 0$. Consider the characteristic functions:

$$\Phi_X(s) = \int_{-\infty}^{+\infty} \mathrm{d}x e^{sx} Q(x) \tag{237}$$

and

$$\Phi_Z(s) = \int_{-\infty}^{+\infty} \mathrm{d}z e^{sz} M(z). \tag{238}$$

Due to the independence

$$\begin{aligned}
\Phi_{X_\delta}(s) &= \Phi_X(s) \cdot \Phi_{\sqrt{\delta}Z}(s) \\
&= \Phi_X(s) \cdot \Phi_Z(\sqrt{\delta}s) \\
&= \Phi_X(s) \cdot \int_{-\infty}^{+\infty} \mathrm{d}z e^{\sqrt{\delta}sz} M(z) \\
&= \Phi_X(s) \cdot \sum_{i=0}^{\infty} \frac{(\sqrt{\delta}s)^i}{i!} \mu_i(M) \\
&= \Phi_X(s) \cdot \left(1 + \frac{\delta s^2}{2} + \cdots\right)
\end{aligned} \tag{239}$$

where $\mu_i(M)$ denotes the i–th moment pertaining to the density M, and where we have used the assumption that $\mu_1(M) = 0$. Applying the inverse Fourier transform, we get:

$$Q_\delta(x) = Q(x) + \frac{\delta}{2} \cdot \frac{\partial^2 Q(x)}{\partial x^2} + o(\delta), \tag{240}$$

and so,

$$\left.\frac{\partial Q_\delta(x)}{\partial \delta}\right|_{\delta=0} = \frac{1}{2} \cdot \frac{\partial^2 Q(x)}{\partial x^2} \sim \frac{1}{2} \cdot \frac{\partial^2 Q_\delta(x)}{\partial x^2}. \tag{241}$$

Now, let us look at the differential entropy:

$$h(X_\delta) = -\int_{-\infty}^{+\infty} \mathrm{d}x Q_\delta(x) \ln Q_\delta(x). \tag{242}$$

Taking the derivative w.r.t. δ, we get:

$$\begin{aligned}
\frac{\partial h(X_\delta)}{\partial \delta} &= -\int_{-\infty}^{+\infty} \mathrm{d}x \left[\frac{\partial Q_\delta(x)}{\partial \delta} + \frac{\partial Q_\delta(x)}{\partial \delta} \cdot \ln Q_\delta(x)\right] \\
&= -\frac{\partial}{\partial \delta}\int_{-\infty}^{+\infty} \mathrm{d}x Q_\delta(x) - \int_{-\infty}^{+\infty} \mathrm{d}x \frac{\partial Q_\delta(x)}{\partial \delta} \cdot \ln Q_\delta(x) \\
&= -\frac{\partial}{\partial \delta} 1 - \int_{-\infty}^{+\infty} \mathrm{d}x \frac{\partial Q_\delta(x)}{\partial \delta} \cdot \ln Q_\delta(x) \\
&= -\int_{-\infty}^{+\infty} \mathrm{d}x \frac{\partial Q_\delta(x)}{\partial \delta} \cdot \ln Q_\delta(x)
\end{aligned} \tag{243}$$

and so,

$$\begin{aligned}
\left.\frac{\partial h(X_\delta)}{\partial \delta}\right|_{\delta=0} &= -\int_{-\infty}^{+\infty} \mathrm{d}x \cdot \left.\frac{\partial Q_\delta(x)}{\partial \delta}\right|_{\delta=0} \cdot \ln Q(x) \\
&= -\int_{-\infty}^{+\infty} \mathrm{d}x \cdot \frac{1}{2}\frac{\mathrm{d}^2 Q(x)}{\mathrm{d}^2 x} \cdot \ln Q(x).
\end{aligned} \tag{244}$$

Integrating by parts, we obtain:

$$\begin{aligned}
\left.\frac{\partial h(X_\delta)}{\partial \delta}\right|_{\delta=0} = &\left[-\frac{1}{2} \cdot \frac{\mathrm{d}Q(x)}{\mathrm{d}x} \cdot \ln Q(x)\right]_{-\infty}^{+\infty} + \\
&\frac{1}{2}\int_{-\infty}^{+\infty} \frac{\mathrm{d}x}{Q(x)} \left[\frac{\partial Q(x)}{\partial x}\right]^2.
\end{aligned} \tag{245}$$

The first term can be shown to vanish since

$$\dot{Q}(x) \ln \sqrt{Q(x)} = \frac{\dot{Q}(x)}{\sqrt{Q(x)}} \cdot \sqrt{Q(x)} \ln \sqrt{Q(x)},$$

where the second factor obviously tends to zero as $|x| \to \infty$ (as $Q(x)$ must tend to zero) and the first factor must be bounded since the integral of $[\dot{Q}(x)/\sqrt{Q(x)}]^2$, which is the Fisher information, is finite. The second term in (245) is exactly $J(Q)/2$. This completes the proof of the modified de Bruijn identity.

This result can be extended straightforwardly to the vector case, showing that for a vector $\boldsymbol{Z}$ with i.i.d. components, all with zero mean:

$$\begin{aligned}\frac{\partial h(\boldsymbol{X}+\sqrt{\delta}\boldsymbol{Z})}{\partial \delta} &= \frac{1}{2}\sum_{i=1}^{N}\int_{\mathbb{R}^N}\frac{\mathrm{d}\boldsymbol{x}}{Q(\boldsymbol{x})}\left[\frac{\partial Q(\boldsymbol{x})}{\partial x_i}\right]^2 \\ &= \frac{1}{2}\sum_{i=1}^{N} J_{ii}(Q) \\ &= \frac{1}{2}\mathrm{tr}\{J(Q)\}. \end{aligned} \tag{246}$$

Putting all this together, we end up with the following generalized definition of temperature:

$$\frac{1}{T} = \frac{k}{N\alpha}\cdot \mathrm{tr}\{J(Q)\}. \tag{247}$$

In the case where Q is symmetric w.r.t. all components of $\boldsymbol{x}$, $\{J_{ii}\}$ are all the same, call it $J(Q)$, and then

$$\frac{1}{T} = \frac{k}{\alpha}\cdot J(Q) \tag{248}$$

or, equivalently,

$$T = \frac{\alpha}{kJ(Q)} = \frac{\alpha}{k}\cdot \mathrm{CRB} \tag{249}$$

where CRB is the Cramér–Rao bound. High temperature means strong noise and this in turn means that it is hard to estimate the mean of X. In the Boltzmann case,

$$J(Q) = \frac{1}{\mathrm{Var}\{X\}} = \alpha\beta = \frac{\alpha}{kT} \tag{250}$$

and we are back to the ordinary definition of temperature.

Another way to look at this result is as an extension of the energy equipartition theorem: As we recall, in the ordinary case of a quadratic Hamiltonian and in equilibrium, we have:

$$\langle \mathcal{E}(X)\rangle = \left\langle \frac{\alpha}{2}X^2\right\rangle = \frac{kT}{2} \tag{251}$$

or

$$\frac{\alpha}{2}\sigma^2 \triangleq \frac{\alpha}{2}\langle X^2\rangle = \frac{kT}{2}. \tag{252}$$

In the passage to the more general case, σ^2 should be replaced by $1/J(Q) = \text{CRB}$. Thus, the induced generalized equipartition function, doesn't talk about average energy but about the CRB:

$$\frac{\alpha}{2} \cdot \text{CRB} = \frac{kT}{2}. \tag{253}$$

Now, the CRB is a lower bound to the estimation error which, in this case, is a translation parameter. For example, let x denote the location of a mass m tied to a spring of strength $m\omega_0^2$ and equilibrium location θ. Then,

$$\mathcal{E}(x) = \frac{m\omega_0^2}{2}(x-\theta)^2. \tag{254}$$

In this case, $\alpha = m\omega_0^2$, and we get:

$$\text{estimation error energy} = \frac{m\omega_0^2}{2} \cdot \boldsymbol{E}(\hat{\theta}(X) - \theta)^2 \geq \frac{kT}{2} \tag{255}$$

where $\hat{\theta}(X)$ is any unbiased estimator of θ based on a measurement of X. This is to say that the generalized equipartition theorem talks about the estimation error energy in the general case. Again, in the Gaussian case, the best estimator is $\hat{\theta}(x) = x$ and we are back to ordinary energy and the ordinary equipartition theorem.

In a follow–up paper [86], Narayanan and Srinivasa have slightly modified their definition of generalized temperature and applied it to the paradigm of a system that obeys a first order stochastic differential equation driven by white noise (Langevin dynamics). It was demonstrated that, away from equilibrium, the corresponding Cramér–Rao inequality can be interpreted as a statement of the second law.

4 Analysis Tools and Asymptotic Methods

4.1 Introduction

So far we have dealt mostly with relatively simple situations, where the Hamiltonian is additive, and then the resulting B–G distribution is i.i.d. But this is seldom the case in reality. Most models in physics, including those that will prove relevant for information theory, as we shall see in the sequel, are much more complicated, more difficult, but also more interesting. More often than not, they are so complicated and difficult, that they do not lend themselves to closed–form analysis at all. In some cases, analysis is still possible, but it requires some more powerful mathematical tools and techniques, which suggest at least some asymptotic approximations. The purpose of this chapter is to prepare these tools, before we can go on to the more challenging settings.

Before diving into the technical material, let us first try to give the flavor of the kind of calculations that we are now addressing. The best way of doing this is by example. We have seen in Subsection 2.6 the example of quantum particles, whose partition function is given by

$$Z_N(\beta) = \sum_{\boldsymbol{N}:\ \sum_r N_r = N} \exp\left\{-\beta \sum_r N_r \epsilon_r\right\}. \tag{256}$$

As mentioned already in Subsection 2.6, this partition function is hard to calculate in closed form due to the constraint $\sum_r N_r = N$. However, we have also defined therein the grand partition function, which may play the role of z–transform, or a generating function

$$\Xi(\beta, z) = \sum_{N \geq 0} z^N Z_N(\beta), \tag{257}$$

and we saw that $\Xi(\beta, z)$ has an easy closed–form expression

$$\Xi(\beta, z) = \prod_r \left[\sum_{N_r} \left(z e^{-\beta \epsilon_r}\right)^{N_r}\right]. \tag{258}$$

Splendid, but how can we get back from $\Xi(\beta, z)$ to $Z_N(\beta)$?

The general idea is to apply the inverse z–transform:

$$Z_N(\beta) = \frac{1}{2\pi j} \oint_{\mathcal{C}} \frac{\Xi(\beta, z)\mathrm{d}z}{z^{N+1}} = \frac{1}{2\pi j} \oint_{\mathcal{C}} \Xi(\beta, z) e^{-(N+1)\ln z}\mathrm{d}z, \tag{259}$$

where z is a complex variable, $j = \sqrt{-1}$, and $\mathcal{C}$ is any clockwise closed path encircling the origin and entirely in the region of convergence. An exact calculation of integrals of this type might still be difficult, in general, but often, we would be happy enough if at least we could identify how they behave in the thermodynamic limit of large N.

Similar needs are frequently encountered in information–theoretic problems. One example is in universal source coding (see, e.g., [13], Chapter 13, in particular, Sect. 13.2, and references therein). Suppose we have a family of information sources indexed by some parameter θ, say, Bernoulli with parameter $\theta \in [0, 1]$, i.e.,

$$P_\theta(\boldsymbol{x}) = (1-\theta)^{N-n}\theta^n, \tag{260}$$

where $\boldsymbol{x} \in \{0, 1\}^N$ and $n \leq N$ is the number of 1's in $\boldsymbol{x}$. When θ is unknown, it is customary to construct a universal code as the Shannon code w.r.t. a certain mixture of these sources

$$P(\boldsymbol{x}) = \int_0^1 \mathrm{d}\theta w(\theta) P_\theta(\boldsymbol{x}) = \int_0^1 \mathrm{d}\theta w(\theta) e^{Nf(\theta)} \tag{261}$$

where

$$f(\theta) = \ln(1-\theta) + q\ln\left(\frac{\theta}{1-\theta}\right); \quad q = \frac{n}{N}. \tag{262}$$

So here again, we need to evaluate an integral of an exponential function of N (this time, on the real line), in order to assess the performance of this universal code.

These are exactly the points where the first tool that we are going to study, namely, the *saddle point method* (a.k.a. the *steepest descent method*) enters into the picture: it gives us a way to assess how integrals of this kind scale as exponential functions of N, for large N. More generally, the saddle point method is a tool for evaluating the exponential order (and also the second order behavior) of an integral of the form

$$\int_{\mathcal{P}} g(z) e^{Nf(z)}\mathrm{d}z \quad \mathcal{P} \text{ is a path in the complex plane.} \tag{263}$$

We begin with the simpler case where the integration is over the real line (or a subset of the real line), whose corresponding asymptotic approximation method is called the *Laplace method of integration*. The exposition of the material in this chapter follows, to a large extent, the book by de Bruijn [20, Chapters 4,5], but with several modifications.

4.2 The Laplace Method

Consider first an integral of the form:

$$F_N \triangleq \int_{-\infty}^{+\infty} e^{Nh(x)}\mathrm{d}x, \tag{264}$$

where the function $h(\cdot)$ is independent of N. How does this integral behave exponentially for large N? Clearly, if it was a sum, like $\sum_i e^{Nh_i}$, rather than an integral, and the number of terms was finite and independent of N, then the dominant term, $e^{N\max_i h_i}$, would have dictated the exponential behavior. This continues to be true even if the sum contains infinitely many terms, provided that the tail of this series decays sufficiently rapidly. Since the integral is a limit of sums, it is conceivable to expect, at least when $h(\cdot)$ is "sufficiently nice", that something of the same spirit would happen with F_N, namely, that its exponential order would be, in analogy, $e^{N\max h(x)}$. In what follows, we are going to show this more rigorously, and as a bonus, we will also be able to say something about the second order behavior. In the above example of universal coding, this gives rise to redundancy analysis.

We will make the following assumptions on h:

1. h is real and continuous.
2. h is maximum at $x = 0$ and $h(0) = 0$ (w.l.o.g).
3. $h(x) < 0 \;\; \forall x \neq 0$, and $\exists b > 0,\; c > 0$ s.t. $|x| \geq c$ implies $h(x) \leq -b$.
4. The integral defining F_N converges for all sufficiently large N. Without loss of generality, let this sufficiently large N be $N = 1$, i.e., $\int_{-\infty}^{+\infty} e^{h(x)}\mathrm{d}x < \infty$.

5. The derivative $h'(x)$ exists at a certain neighborhood of $x = 0$, and $h''(0) < 0$. Thus, $h'(0) = 0$.

From these assumptions, it follows that for all $\delta > 0$, there is a positive number $\eta(\delta)$ s.t. for all $|x| \geq \delta$, we have $h(x) \leq -\eta(\delta)$. For $\delta \geq c$, this is obvious from assumption 3. If $\delta < c$, then the maximum of the continuous function h across the interval $[\delta, c]$ is strictly negative. A similar argument applies to the interval $[-c, -\delta]$. Consider first the tails of the integral under discussion:

$$\begin{aligned}\int_{|x|\geq\delta} e^{Nh(x)}\mathrm{d}x &= \int_{|x|\geq\delta} \mathrm{d}x e^{(N-1)h(x)+h(x)} \\ &\leq \int_{|x|\geq\delta} \mathrm{d}x e^{-(N-1)\eta(\delta)+h(x)} \\ &\leq e^{-(N-1)\eta(\delta)} \cdot \int_{-\infty}^{+\infty} e^{h(x)}\mathrm{d}x \to 0 \end{aligned} \tag{265}$$

and the convergence to zero is exponentially fast. In other words, the tails' contribution is vanishingly small. It remains to examine the integral from $-\delta$ to $+\delta$, that is, the neighborhood of $x = 0$. In this neighborhood, we shall use the Taylor series expansion of h. Since $h(0) = h'(0) = 0$, then $h(x) \approx \frac{1}{2}h''(0)x^2$. More precisely, for all $\epsilon > 0$, there is $\delta > 0$ s.t.

$$\left|h(x) - \frac{1}{2}h''(0)x^2\right| \leq \epsilon x^2 \quad \forall |x| \leq \delta. \tag{266}$$

Thus, this integral is sandwiched as follows:

$$\begin{aligned}\int_{-\delta}^{+\delta} \exp\left\{\frac{N}{2}(h''(0) - \epsilon)x^2\right\}\mathrm{d}x &\leq \int_{-\delta}^{+\delta} e^{Nh(x)}\mathrm{d}x \\ &\leq \int_{-\delta}^{+\delta} \exp\left\{\frac{N}{2}(h''(0) + \epsilon)x^2\right\}\mathrm{d}x. \end{aligned} \tag{267}$$

The right–most side is further upper bounded by

$$\int_{-\infty}^{+\infty} \exp\left\{\frac{N}{2}(h''(0) + \epsilon)x^2\right\}\mathrm{d}x \tag{268}$$

and since $h''(0) < 0$, then $h''(0) + \epsilon = -(|h''(0)| - \epsilon)$, and so, the latter is a Gaussian integral given by

$$\sqrt{\frac{2\pi}{(|h''(0)| - \epsilon)N}}. \tag{269}$$

The left–most side of eq. (267) is further lower bounded by

$$\begin{aligned}
&\int_{-\delta}^{+\delta} \exp\left\{-\frac{N}{2}(|h''(0)|+\epsilon)x^2\right\}\mathrm{d}x \\
&= \int_{-\infty}^{+\infty} \exp\left\{-\frac{N}{2}(|h''(0)|+\epsilon)x^2\right\}\mathrm{d}x - \\
&\int_{|x|\geq\delta} \exp\left\{-\frac{N}{2}(|h''(0)|+\epsilon)x^2\right\}\mathrm{d}x \\
&= \sqrt{\frac{2\pi}{(|h''(0)|+\epsilon)N}} - 2Q(\delta\sqrt{n(|h''(0)|+\epsilon)}) \\
&\geq \sqrt{\frac{2\pi}{(|h''(0)|+\epsilon)N}} - O\left(\exp\left\{-\frac{N}{2}(|h''(0)|+\epsilon)\delta^2\right\}\right) \\
&\sim \sqrt{\frac{2\pi}{(|h''(0)|+\epsilon)N}}
\end{aligned} \tag{270}$$

where the notation $A_N \sim B_N$ means that $\lim_{N\to\infty} A_N/B_N = 1$. Since ϵ and hence δ can be made arbitrary small, we find that

$$\int_{-\delta}^{+\delta} e^{Nh(x)}\mathrm{d}x \sim \sqrt{\frac{2\pi}{|h''(0)|N}}. \tag{271}$$

Finally, since the tails contribute an exponentially small term, which is negligible compared to the contribution of $O(1/\sqrt{N})$ order of the integral across $[-\delta, +\delta]$, we get:

$$\int_{-\infty}^{+\infty} e^{Nh(x)}\mathrm{d}x \sim \sqrt{\frac{2\pi}{|h''(0)|N}}. \tag{272}$$

Slightly more generally, if h is maximized at an arbitrary point $x = x_0$ this is completely immaterial because an integral over the entire real line is invariant under translation of the integration variable. If, furthermore, the maximum $h(x_0)$ is not necessarily zero, we can make it zero by decomposing h according to $h(x) = h(x_0) + [h(x) - h(x_0)]$ and moving the first term as a constant factor of $e^{Nh(x_0)}$ outside of the integral. The result would then be

$$\int_{-\infty}^{+\infty} e^{Nh(x)}\mathrm{d}x \sim e^{Nh(x_0)} \cdot \sqrt{\frac{2\pi}{|h''(x_0)|N}} \tag{273}$$

Of course, the same considerations continue to apply if F_N is defined over any finite or half–infinite interval that contains the maximizer $x = 0$, or more generally $x = x_0$ as an internal

point. It should be noted, however, that if F_N is defined over a finite or semi–infinite interval and the maximum of h is obtained at an edge of this interval, then the derivative of h at that point does not necessarily vanish, and the Gaussian integration would not apply anymore. In this case, the local behavior around the maximum would be approximated by an exponential $\exp\{-N|h'(0)|x\}$ or $\exp\{-N|h'(x_0)|x\}$ instead, which gives a somewhat different expression. However, the factor $e^{Nh(x_0)}$, which is the most important factor, would continue to appear. Normally, this will be the only term that will interest us, whereas the other factor, which provides the second order behavior will not be important for us. A further extension in the case where the maximizer is an internal point at which the derivative vanishes, is this:

$$\int_{-\infty}^{+\infty} g(x)e^{Nh(x)}\mathrm{d}x \sim g(x_0)e^{Nh(x_0)} \cdot \sqrt{\frac{2\pi}{|h''(x_0)|N}} \tag{274}$$

where g is another function that does not depend on N. This technique, of approximating an integral of a function, which is exponential in some large parameter N, by neglecting the tails and approximating it by a Gaussian integral around the maximum, is called the *Laplace method of integration*.

Example: Stirling's formula. Beginning from the identity $\int_0^\infty \mathrm{d}x e^{-sx} = 1/s$, and differentiating N times both sides w.r.t. s, we get, from the l.h.s., $(-1)^N \int_0^\infty x^N e^{-sx}\mathrm{d}x$, and from the r.h.s. $(-1)^N N!/s^{N+1}$, which together yield the identity

$$N! = s^{N+1}\int_0^\infty x^N e^{-sx}\mathrm{d}x,$$

which holds true for every $s > 0$. On substituting $s = N$, we get

$$N! = N^{N+1}\int_0^\infty x^N e^{-Nx}\mathrm{d}x.$$

Assessing this integral using the Laplace method, we have $h(x) = \ln x - x$, which is maximized at $x_0 = 1$, with $h(x_0) = h''(x_0) = -1$. Thus,

$$N! \sim N^{N+1}e^{-N\cdot 1}\sqrt{\frac{2\pi}{N\cdot 1}} = \left(\frac{N}{e}\right)^N \sqrt{2\pi N},$$

which is the well–known Stirling formula for approximating $N!$.

4.3 The Saddle Point Method

We now expand the scope to integrals along paths in the complex plane, which are also encountered and even more often than one would expect (cf. the earlier example). As said, the extension of the Laplace integration technique to the complex case is called the saddle–point method or the steepest descent method, for reasons that will become apparent shortly. Specifically, we are now interested in an integral of the form

$$F_N = \int_{\mathcal{P}} e^{Nh(z)} \mathrm{d}z \quad \text{or more generally} \quad F_N = \int_{\mathcal{P}} g(z) e^{Nh(z)} \mathrm{d}z \tag{275}$$

where $z = x + jy$ is a complex variable ($j = \sqrt{-1}$), and $\mathcal{P}$ is a certain path in the complex plane, starting at some point A and ending at point B. We will focus first on the former integral, without the factor g. We will assume that $\mathcal{P}$ is fully contained in a region where h is analytic.

The first observation is that the value of the integral depends only on A and B, and not on the details of $\mathcal{P}$: Consider any alternate path $\mathcal{P}'$ from A to B such that h has no singularities in the region surrounded by $\mathcal{P} \bigcup \mathcal{P}'$. Then, the integral of $e^{Nh(z)}$ over the closed path $\mathcal{P} \bigcup \mathcal{P}'$ (going from A to B via $\mathcal{P}$ and returning to A via $\mathcal{P}'$) vanishes, which means that the integrals from A to B via $\mathcal{P}$ and via $\mathcal{P}'$ are the same. This means that we actually have the freedom to select the integration path, as long as we do not go too far, to the other side of some singularity point, if there is any. This point will be important in our forthcoming considerations.

An additional important observation has to do with yet another basic property of analytic functions: the *maximum modulus theorem*, which basically tells that the modulus of an analytic function has no maxima. We will not prove here this theorem, but in a nutshell, the point is this: Let

$$h(z) = u(z) + jv(z) = u(x, y) + jv(x, y), \tag{276}$$

where u and v are real functions. If h is analytic, the following relationships (a.k.a. the

Cauchy–Riemann conditions)[14] between the partial derivatives of u and v must hold:

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}; \qquad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}. \tag{277}$$

Taking the second order partial derivative of u:

$$\frac{\partial^2 u}{\partial x^2} = \frac{\partial^2 v}{\partial x \partial y} = \frac{\partial^2 v}{\partial y \partial x} = -\frac{\partial^2 u}{\partial y^2} \tag{278}$$

where the first equality is due to the first Cauchy–Riemann condition and the third equality is due to the second Cauchy–Riemann condition. Equivalently,

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0, \tag{279}$$

which is the *Laplace equation*. This means, among other things, that no point at which $\partial u/\partial x = \partial u/\partial y = 0$ can be a local maximum (or a local minimum) of u, because if it is a local maximum in the x–direction, in which case, $\partial^2 u/\partial x^2 < 0$, then $\partial^2 u/\partial y^2$ must be positive, which makes it a local minimum in the y–direction, and vice versa. In other words, every point of zero partial derivatives of u must be a *saddle point*. This discussion applies now to the modulus of the integrand $e^{Nh(z)}$ because

$$\left| \exp\{Nh(z)\} \right| = \exp[N\mathrm{Re}\{h(z)\}] = e^{Nu(z)}. \tag{280}$$

Of course, if $h'(z) = 0$ at some $z = z_0$, then the derivatives both in the vertical and the horizontal directions vanish, that is

$$\left.\frac{\partial h(z)}{\partial x}\right|_{z=z_0} = \left.\frac{\partial u}{\partial x}\right|_{z=z_0} + j\left.\frac{\partial v}{\partial x}\right|_{z=z_0} = 0$$

and

$$\left.\frac{\partial h(z)}{\partial (jy)}\right|_{z=z_0} = -j\left.\frac{\partial u}{\partial y}\right|_{z=z_0} + \left.\frac{\partial v}{\partial y}\right|_{z=z_0} = 0$$

which mean, among other things, that then $\partial u/\partial x|_{z=z_0} = \partial u/\partial y|_{z=z_0} = 0$ too, and then z_0 is a saddle point of $|e^{Nh(z)}| = e^{Nu(z)}$. Thus, zero–derivative points of h are saddle points of $|e^{Nh(z)}|$.

[14] This is related to the fact that for the derivative $f'(z)$ to exist, it should be independent of the direction at which z is perturbed, whether it is, e.g., the horizontal or the vertical direction, i.e., $f'(z) = \lim_{\delta\to 0}[f(z+\delta) - f(z)]/\delta = \lim_{\delta\to 0}[f(z+j\delta) - f(z)]/(j\delta)$, where δ goes to zero along the reals.

Another way to understand the maximum modulus principle is the following (without relying just on second derivatives, which might vanish too): Given a complex analytic function $f(z)$, we argue that the average of f over a circle always agrees with its value at the center of this circle. Specifically, consider the circle of radius r centered at z_0, i.e., $z = z_0 + re^{j\theta}$. Then,

$$\begin{aligned}\frac{1}{2\pi}\int_{-\pi}^{\pi} f\left(z_0 + re^{j\theta}\right)\mathrm{d}\theta &= \frac{1}{2\pi j}\int_{-\pi}^{\pi}\frac{f\left(z_0 + re^{j\theta}\right) jre^{j\theta}\mathrm{d}\theta}{re^{j\theta}} \\ &= \frac{1}{2\pi j}\oint_{z=z_0+re^{j\theta}}\frac{f\left(z_0 + re^{j\theta}\right)\mathrm{d}\left(z_0 + re^{j\theta}\right)}{re^{j\theta}} \\ &= \frac{1}{2\pi j}\oint_{z=z_0+re^{j\theta}}\frac{f(z)\mathrm{d}z}{z - z_0} = f(z_0). \end{aligned} \tag{281}$$

and so,

$$|f(z_0)| \leq \frac{1}{2\pi}\int_{-\pi}^{\pi}\left|f\left(z_0 + re^{j\theta}\right)\right|\mathrm{d}\theta \leq \max_{\theta}\left|f\left(z_0 + re^{j\theta}\right)\right|, \tag{282}$$

which means that $|f(z_0)|$ cannot be strictly larger than *all* $|f(z)|$ in any neighborhood (an arbitrary radius r) of z_0. Now, apply this fact to $f(z) = e^{Nh(z)}$.

Equipped with this background, let us return to our integral F_N. Since we have the freedom to choose the path $\mathcal{P}$, suppose that we can find one which passes through a saddle point z_0 (hence the name of the method) and that $\max_{z\in\mathcal{P}}|e^{Nh(z)}|$ is attained at z_0. We expect then, that similarly as in the Laplace method, the integral would be dominated by $e^{Nh(z_0)}$. Of course, such a path would be fine only if it crosses the saddle point z_0 at a direction w.r.t. which z_0 is a local maximum of $|e^{Nh(z)}|$, or equivalently, of $u(z)$. Moreover, in order to apply our earlier results of the Laplace method, we will find it convenient to draw $\mathcal{P}$ such that any point z in the vicinity of z_0, where in the Taylor expansion is (by the fact that $h'(z_0) = 0$)

$$h(z) \approx h(z_0) + \frac{1}{2}h''(z_0)(z - z_0)^2 \tag{283}$$

the second term, $\frac{1}{2}h''(z_0)(z - z_0)^2$, is purely **real and negative**, and then it behaves locally as a negative parabola, just like in the Laplace method. This means that

$$\arg\{h''(z_0)(z - z_0)^2\} \equiv \arg\{h''(z_0)\} + 2\arg(z - z_0) = \pi \tag{284}$$

or equivalently

$$\arg(z - z_0) = \frac{\pi - \arg\{h''(z_0)\}}{2} \triangleq \theta. \tag{285}$$

Namely, $\mathcal{P}$ should cross z_0 in the direction θ. This direction is called the *axis* of z_0, and it can be shown to be the direction of **steepest descent** from the peak at z_0 (hence the name).[15]

So pictorially, we are going to select a path $\mathcal{P}$ from A to B, which will be composed of three parts (see Fig. 7): The parts $A \to A'$ and $B' \to B$ are quite arbitrary as they constitute the tail of the integral. The part from A' to B', in the vicinity of z_0, is a straight line on the axis of z_0. Now, let us decompose F_N into its three parts:

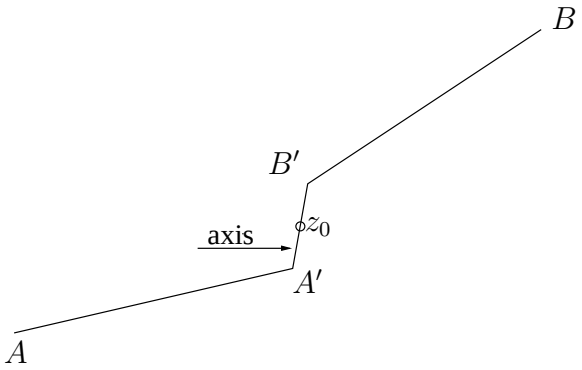


Figure 7: A path $\mathcal{P}$ from A to B, passing via z_0 along the axis.

$$F_N = \int_A^{A'} e^{Nh(z)}\mathrm{d}z + \int_{A'}^{B'} e^{Nh(z)}\mathrm{d}z + \int_{B'}^{B} e^{Nh(z)}\mathrm{d}z. \tag{286}$$

As for the first and the third terms,

$$\begin{aligned}\left|\left(\int_A^{A'} + \int_{B'}^{B}\right)\mathrm{d}z e^{Nh(z)}\right| &\le \left(\int_A^{A'} + \int_{B'}^{B}\right)\mathrm{d}z|e^{Nh(z)}| \\ &= \left(\int_A^{A'} + \int_{B'}^{B}\right)\mathrm{d}z e^{Nu(z)}\end{aligned} \tag{287}$$

[15]Note that in the direction $\theta - \pi/2$, which is perpendicular to the axis, $\arg[h''(z_0)(z - z_0)^2] = \pi - \pi = 0$, which means that $h''(z_0)(z - z_0)^2$ is real and positive (i.e., it behaves like a positive parabola). Therefore, in this direction, z_0 is a local minimum.

whose contribution is negligible compared to $e^{Nu(z_0)}$, just like the tails in the Laplace method. As for the middle integral,

$$\int_{A'}^{B'} e^{Nh(z)}\mathrm{d}z \approx e^{Nh(z_0)} \int_{A'}^{B'} \exp\{Nh''(z_0)(z-z_0)^2/2\}\mathrm{d}z. \tag{288}$$

By changing from the complex integration variable z to the real variable x, running from $-\delta$ to $+\delta$, with $z = z_0 + xe^{j\theta}$ (moving along the axis), we get exactly the Gaussian integral of the Laplace method, leading to

$$\int_{A'}^{B'} \exp\{Nh''(z_0)(z-z_0)^2/2\}\mathrm{d}z = e^{j\theta}\sqrt{\frac{2\pi}{N|h''(z_0)|}} \tag{289}$$

where the factor $e^{j\theta}$ is due to the change of variable ($\mathrm{d}z = e^{j\theta}\mathrm{d}x$). Thus,

$$F_N \sim e^{j\theta} \cdot e^{Nh(z_0)}\sqrt{\frac{2\pi}{N|h''(z_0)|}}, \tag{290}$$

and slightly more generally,

$$\int_{\mathcal{P}} g(z)e^{Nh(z)}\mathrm{d}z \sim e^{j\theta}g(z_0)e^{Nh(z_0)}\sqrt{\frac{2\pi}{N|h''(z_0)|}} \tag{291}$$

The idea of integration along the axis is that along this direction, the 'phase' of $e^{Nh(z)}$ is locally constant, and only the modulus varies. Had the integration been along another direction with an imaginary component $j\phi(z)$, the function $e^{Nh(z)}$ would have undergone 'modulation', i.e., it would have oscillated with a complex exponential $e^{Nj\phi(z)}$ of a very high 'frequency' (proportional to N) and then $e^{Nu(z_0)}$ would not have guaranteed to dictate the modulus and to dominate the integral.

Now, an important comment is in order: What happens if there is more than one saddle point? Suppose we have two saddle points, z_1 and z_2. On a first thought, one may be concerned by the following consideration: We can construct two paths from A to B, path $\mathcal{P}_1$ crossing z_1, and path $\mathcal{P}_2$ crossing z_2. Now, if z_i is the highest point along $\mathcal{P}_i$ for both $i = 1$ and $i = 2$, then F_N is exponentially both $e^{Nh(z_1)}$ and $e^{Nh(z_2)}$ at the same time. If $h(z_1) \neq h(z_2)$, this is a contradiction. But the following consideration shows that this cannot happen as

long as $h(z)$ is analytic within the region $\mathcal{C}$ surrounded by $\mathcal{P}_1 \cup \mathcal{P}_2$. Suppose conversely, that the scenario described above happens. Then either z_1 or z_2 maximize $|e^{Nh(z)}|$ along the closed path $\mathcal{P}_1 \cup \mathcal{P}_2$. Let us say that it is z_1. We claim that then z_1 cannot be a saddle point, for the following reason: No point in the interior of $\mathcal{C}$ can be higher than z_1, because if there was such a point, say, z_3, then we had

$$\max_{z\in\mathcal{C}} |e^{Nh(z)}| \geq |e^{Nh(z_3)}| > |e^{Nh(z_1)}| = \max_{z\in\mathcal{P}_1\cup\mathcal{P}_2} |e^{Nh(z)}| \tag{292}$$

which contradicts the maximum modulus principle. This then means, among other things, that in every neighborhood of z_1, all points in $\mathcal{C}$ are lower than z_1, including points found in a direction perpendicular to the direction of the axis through z_1. But this contradicts the fact that z_1 is a saddle point: Had it been a saddle point, it would be a local maximum along the axis and a local minimum along the perpendicular direction. Since z_1 was assumed a saddle point, then it cannot be the highest point on $\mathcal{P}_1$, which means that it doesn't dominate the integral.

One might now be concerned by the thought that the integral along $\mathcal{P}_1$ is then dominated by an even higher contribution, which still seems to contradict the lower exponential order of $e^{Nh(z_2)}$ attained by the path $\mathcal{P}_2$. However, this is not the case. The highest point on the path is guaranteed to dominate the integral only if it is a saddle point. Consider, for example, the integral $F_N = \int_{a+j0}^{a+j2\pi} e^{Nz}\mathrm{d}z$. Along the vertical line from $a+j0$ to $a+j2\pi$, the modulus (or attitude) is e^{Na} everywhere. If the attitude alone had been whatever counts (regardless of whether it is a saddle point or not), the exponential order of (the modulus of) this integral would be e^{Na}. However, the true value of this integral is zero! The reason for this disagreement is that there is no saddle point along this path.

What about a path $\mathcal{P}$ that crosses both z_1 and z_2? This cannot be a good path for the saddle point method, for the following reason: Consider two slightly perturbed versions of $\mathcal{P}$: path $\mathcal{P}_1$, which is very close to $\mathcal{P}$, it crosses z_1, but it makes a tiny detour that bypasses z_2, and similarly path $\mathcal{P}_2$, passing via z_2, but with a small deformation near z_1. Path $\mathcal{P}_2$ includes z_2 as saddle point, but it is not the highest point on the path, since $\mathcal{P}_2$ passes near

z_1, which is higher. Path $\mathcal{P}_1$ includes z_1 as saddle point, but it cannot be the highest point on the path because we are back to the same situation we were two paragraphs ago. Since both $\mathcal{P}_1$ and $\mathcal{P}_2$ are bad choices, and since they are both arbitrarily close to $\mathcal{P}$, then $\mathcal{P}$ cannot be good either.

To summarize: if we have multiple saddle points, we should find the one with the *lowest* attitude and then we have a chance to find a path through this saddle point (and only this one) along which this saddle point is dominant.

Consider next a few simple examples.

Example 1 – relation between $\Omega(E)$ and $Z(\beta)$ revisited. Assuming, without essential loss of generality, that the ground–state energy of the system is zero, we have seen before the relation

$$Z(\beta) = \int_0^\infty \mathrm{d}E\Omega(E)e^{-\beta E}, \tag{293}$$

which actually means that $Z(\beta)$ is the Laplace transform of $\Omega(E)$. Consequently, this means that $\Omega(E)$ is the inverse Laplace transform of $Z(\beta)$, i.e.,

$$\Omega(E) = \frac{1}{2\pi j}\int_{\gamma-j\infty}^{\gamma+j\infty} e^{\beta E}Z(\beta)\mathrm{d}\beta, \tag{294}$$

where the integration in the complex plane is along the vertical line $\mathrm{Re}(\beta) = \gamma$, which is chosen to the right of all singularity points of $Z(\beta)$. In the large N limit, this becomes

$$\Omega(E) \doteq \frac{1}{2\pi j}\int_{\gamma-j\infty}^{\gamma+j\infty} e^{N[\beta\epsilon+\phi(\beta)]}\mathrm{d}\beta, \tag{295}$$

which can now be assessed using the saddle point method. The derivative of the bracketed term at the exponent vanishes at the value of β that solves the equation $\phi'(\beta) = -\epsilon$, which is $\beta^*(\epsilon) \in \mathbb{R}$, thus we will choose $\gamma = \beta^*(\epsilon)$ (assuming that this is a possible choice) and thereby let the integration path pass through this saddle point. At $\beta = \beta^*(\epsilon)$, $|\exp\{N[\beta\epsilon + \phi(\beta)]\}|$ has its maximum along the vertical direction, $\beta = \beta^*(\epsilon) + j\omega$, $-\infty < \omega < +\infty$ (and hence it dominates the integral), but since it is a saddle point, it *minimizes* $|\exp\{N[\beta\epsilon + \phi(\beta)]\}| = \exp\{N[\beta\epsilon + \phi(\beta)]\}$, in the horizontal direction (the real line). Thus,

$$\Omega(E) \doteq \exp\{N\min_{\beta\in\mathbb{R}}[\beta\epsilon + \phi(\beta)]\} = e^{N\Sigma(\epsilon)}, \tag{296}$$

as we have seen in Chapter 2.

Example 2 – size of a type class. This is a question that can very easily be answered using simple combinatorics (the method of types). Among all binary sequences of length N, how many have n 1's and $(N-n)$ 0's? Let us calculate this number, M_n, using the saddle point method (see also [80, Sect. 4.7] for a more general calculation):

$$\begin{aligned}
M_n &= \sum_{\boldsymbol{x}\in\{0,1\}^N} \mathcal{I}\left\{\sum_{i=1}^N x_i = n\right\} \\
&= \sum_{x_1=0}^{1} \cdots \sum_{x_N=0}^{1} \mathcal{I}\left\{\sum_{i=1}^N x_i = n\right\} \\
&= \sum_{x_1=0}^{1} \cdots \sum_{x_N=0}^{1} \frac{1}{2\pi}\int_0^{2\pi} \mathrm{d}\omega \exp\left\{j\omega\left(n - \sum_{i=1}^N x_i\right)\right\} \\
&= \int_0^{2\pi} \frac{\mathrm{d}\omega}{2\pi} \sum_{x_1=0}^{1} \cdots \sum_{x_N=0}^{1} \exp\left\{j\omega\left(n - \sum_{i=1}^N x_i\right)\right\} \\
&= \int_0^{2\pi} \frac{\mathrm{d}\omega}{2\pi} e^{j\omega n} \prod_{i=1}^N \left[\sum_{x_i=0}^{1} e^{-j\omega x_i}\right] \\
&= \int_0^{2\pi} \frac{\mathrm{d}\omega}{2\pi} e^{j\omega n}(1+e^{-j\omega})^N \\
&= \int_0^{2\pi} \frac{\mathrm{d}\omega}{2\pi} \exp\{N[j\omega\alpha + \ln(1+e^{-j\omega})]\} \\
&= \int_0^{2\pi j} \frac{\mathrm{d}z}{2\pi j} \exp\{N[z\alpha + \ln(1+e^{-z})]\} \qquad (297)
\end{aligned}$$

where we have denoted $\alpha = n/N$ and in the last step we changed the integration variable according to $z = j\omega$. This is an integral with a starting point A at the origin and an ending point B at $2\pi j$. Here, $h(z) = z\alpha + \ln(1+e^{-z})$, and the saddle point, where $h'(z) = 0$, is on the real axis: $z_0 = \ln\frac{1-\alpha}{\alpha}$, where $h(z_0)$ gives the binary entropy of α, as expected. Thus, the integration path must be deformed to pass through this point on the real axis, and then to approach back the imaginary axis, so as to arrive at B. There is one caveat here, however: The points A and B are both higher than z_0: While $u(z_0) = -\alpha\ln(1-\alpha) - (1-\alpha)\ln(1-\alpha)$, at the edges we have $u(A) = u(B) = \ln 2$. So this is not a good saddle–point integral to

work with.

Two small modifications can, however, fix the problem: The first is to define the integration interval of ω to be $[-\pi,\pi]$ rather than $[0,2\pi]$ (which is, of course, legitimate), and then z would run from $-j\pi$ to $+j\pi$. The second is the following: Consider again the first line of the expression of M_n above, but before doing anything else, let us multiply the whole expression (outside the summation) by $e^{\theta n}$ (θ an arbitrary real), whereas the summand will be multiplied by $e^{-\theta\sum_i x_i}$, which exactly cancels the factor of $e^{\theta n}$ for every non–zero term of this sum. We can now repeat exactly the same calculation as above, but this time we get:

$$M_n = \int_{\theta-j\pi}^{\theta+j\pi} \frac{\mathrm{d}z}{2\pi j}\exp\{N[z\alpha+\ln(1+e^{-z})]\}, \tag{298}$$

namely, we moved the integration path to a parallel vertical line and shifted it by the amount of π to the south. Now, we have the freedom to choose θ. The obvious choice is to set $\theta=\ln\frac{1-\alpha}{\alpha}$, so that we cross the saddle point z_0. Now z_0 is the highest point on the path. Moreover, the vertical direction of the integration is also the direction of the axis of z_0 as it should be. Also, the second order factor of $O(1/\sqrt{N})$ of the saddle point integration agrees with the same factor that we can see from the Sterling approximation in the more refined formula.

A slightly different look at this example is as follows. Consider the Schottky example and the partition function

$$Z(\beta)=\sum_{\boldsymbol{x}} e^{-\beta\epsilon_0\sum_i x_i}, \tag{299}$$

which, on the one hand, is given by $\sum_{n=0}^{N} M_n e^{-\beta\epsilon_0 n}$, and on the other hand, is given also by $(1+e^{-\beta\epsilon_0})^N$. Thus, defining $s=e^{-\beta\epsilon_0}$, we have

$$Z(s)=\sum_{n=0}^{N} M_n s^n, \tag{300}$$

and so, $Z(s)=(1+s)^N$ is the z–transform of the finite sequence $\{M_n\}_{n=0}^{N}$. Consequently, M_n is given by the inverse z–transform of $Z(s)=(1+s)^N$, i.e.,

$$M_n=\frac{1}{2\pi j}\oint (1+s)^N s^{-n-1}\mathrm{d}s$$

$$= \frac{1}{2\pi j} \oint \exp\{N[\ln(1+s) - \alpha \ln s]\} \mathrm{d}s \tag{301}$$

This time, the integration path is any closed path that surrounds the origin, the saddle point is $s_0 = \alpha/(1-\alpha)$, so we take the path to be a circle whose radius is $r = \frac{\alpha}{1-\alpha}$. The rest of the calculation is essentially the same as before, and of course, so is the result. Note that this is actually the very same integral as before up to a change of the integration variable from z to s, according to $s = e^{-z}$, which maps the vertical straight line between $\theta - \pi j$ and $\theta + \pi j$ onto a circle of radius $e^{-\theta}$, centered at the origin. $\square$

Example 3 – surface area of a sphere. This is largely a continuous analogue of Example 2, which is given in order to show that this method, unlike the combinatorial method of types, extends to the continuous alphabet case. Let us compute the surface area of an N–dimensional sphere with radius NR:

$$\begin{aligned}
S_N &= \int_{\mathbb{R}^N} \mathrm{d}\boldsymbol{x} \cdot \delta\left(NR - \sum_{i=1}^N x_i^2\right) \\
&= e^{N\theta R} \int_{\mathbb{R}^N} \mathrm{d}\boldsymbol{x} e^{-\theta \sum_i x_i^2} \cdot \delta\left(NR - \sum_{i=1}^N x_i^2\right) \\
&= e^{N\theta R} \int_{\mathbb{R}^N} \mathrm{d}\boldsymbol{x} e^{-\theta \sum_i x_i^2} \int_{-\infty}^{+\infty} \frac{\mathrm{d}\omega}{2\pi} e^{j\omega(NR - \sum_i x_i^2)} \\
&= e^{N\theta R} \int_{-\infty}^{+\infty} \frac{\mathrm{d}\omega}{2\pi} e^{j\omega NR} \int_{\mathbb{R}^N} \mathrm{d}\boldsymbol{x} e^{-(\theta + j\omega)\sum_i x_i^2} \\
&= e^{N\theta R} \int_{-\infty}^{+\infty} \frac{\mathrm{d}\omega}{2\pi} e^{j\omega NR} \left[\int_{\mathbb{R}} \mathrm{d}x e^{-(\theta+j\omega)x^2}\right]^N \\
&= e^{N\theta R} \int_{-\infty}^{+\infty} \frac{\mathrm{d}\omega}{2\pi} e^{j\omega NR} \left(\frac{\pi}{\theta + j\omega}\right)^{N/2} \\
&= \frac{\pi^{N/2}}{2\pi} \int_{-\infty}^{+\infty} \mathrm{d}\omega \exp\left\{N\left[(\theta + j\omega)R - \frac{1}{2}\ln(\theta + j\omega)\right]\right\} \\
&= \frac{\pi^{N/2}}{2\pi j} \int_{\theta - j\infty}^{\theta + j\infty} \mathrm{d}z \exp\left\{N\left[zR - \frac{1}{2}\ln z\right]\right\}.
\end{aligned} \tag{302}$$

where $\delta(\cdot)$ denotes the Dirac delta function. So here

$$h(z) = zR - \frac{1}{2}\ln z \tag{303}$$

and the integration is along an arbitrary vertical straight line parametrized by θ. We will select this straight line to pass via the saddle point $z_0 = \frac{1}{2R}$. Now,

$$h(z_0) = \frac{1}{2}\ln(2\pi eR), \tag{304}$$

which is exactly the differential entropy of a Gaussian random variable, as expected. $\square$

Comment: In these examples, we have used an additional trick: whenever we had to deal with an problematic non–analytic function like the δ function, we presented it as the inverse Fourier transform of a 'nice' function, and then changed the order of integrations and summations. This idea will be repeated in the sequel. It is used very frequently in physics literature.

4.4 Extended Example: Capacity of a Disordered System

To summarize the analysis tools that we have seen thus far, we provide here an extended example of analyzing the capacity of a certain model of a disordered magnetic material, namely, the Sherrington–Kirkpatrick spin glass. We will elaborate more on models of this kind in the next chapter, but for now, this is merely brought here as an extensive exercise. The derivation and the results here follow the work of Shental and Kanter [105] (see also [21]).

For a given positive integer N, consider a set of $N(N-1)/2$ i.i.d., zero–mean Gaussian random variables, $\{J_{i\ell},\ 1 \leq i < \ell \leq N\}$ all with variance J^2/N, where $J > 0$ is fixed. Now form a symmetric $N \times N$ zero–diagonal matrix from $\{J_{i\ell}\}$, thus extending the definition of J_{ij} for all pairs (i, ℓ) in the range $\{1, 2, \ldots, N\}$. The problem addressed in [105] is the following: For a typical realization of $\{J_{i\ell}\}$, how many binary sequences $\boldsymbol{s} = (s_1, \ldots, s_N) \in \{-1, +1\}^N$ can be found such that the equations

$$s_i = \mathrm{sgn}\left(\sum_{\ell} J_{i\ell} s_\ell\right), \quad i = 1, \ldots, N, \tag{305}$$

are all satisfied simultaneously?

In a nutshell, the motivation for this question is that in the Sherrington–Kirkpatrick spin glass model, each such solution is a meta–stable state, which can be used to store information. The evaluation of the number of meta–stable states determines then the amount of memory, or the capacity of this system.

Let $M(N)$ denote the number of solutions of these equations in $\{-1,+1\}^N$ and define the capacity of the system as

$$C = \lim_{N\to\infty} \frac{\ln \bar{M}(N)\}}{N}, \tag{306}$$

where $\bar{M}(N)$ is the expectation of $M(N)$, taken w.r.t. the randomness of $\{J_{i\ell}\}$ and where it is assumed that the limit exists. Note that this is different, and in general larger, than $\lim_{N\to\infty} \frac{1}{N}\boldsymbol{E}\{\ln M(N)\}$, which captures the capacity for a *typical* realization of the system. These two different definitions will be discussed later, in Subsection 5.7.

The main result of [105] is the following single–letter expression for the capacity:

$$C = \ln[2(1-Q(t))] - \frac{t^2}{2} \tag{307}$$

where

$$Q(t) \stackrel{\Delta}{=} \frac{1}{2\pi}\int_t^\infty \mathrm{d}u \cdot e^{-u^2/2} \tag{308}$$

and t is the solution to the equation

$$t = \frac{e^{-t^2/2}}{\sqrt{2\pi}[1-Q(t)]}. \tag{309}$$

In fact, Shental and Kanter address a slightly more general question: Quite obviously, the meta–stability condition is that for every i there exists $\lambda_i > 0$ such that

$$\lambda_i s_i = \sum_j J_{ij} s_j. \tag{310}$$

The question addressed is then the following: Given a constant K, what is the expected number of states $\boldsymbol{s}$ for which there is $\lambda_i > K$ for each i such that $\lambda_i s_i = \sum_j J_{ij} s_j$? For $K \to -\infty$, one expects $C \to \ln 2$, and for $K \to \infty$, one expects $C \to 0$. The case of interest is $K = 0$.

Turning now to the analysis, we first observe that for each such state,

$$\int_K^\infty \cdots \int_K^\infty \prod_{i=1}^N \left[\mathrm{d}\lambda_i \delta \left(\sum_\ell J_{i\ell} s_\ell - \lambda_i s_i \right) \right] = 1 \tag{311}$$

thus

$$M(N) = \int_K^\infty \cdots \int_K^\infty \prod_{i=1}^N \mathrm{d}\lambda_i \sum_{\boldsymbol{s}} \left\langle \prod_{i=1}^N \delta \left(\sum_\ell J_{i\ell} s_\ell - \lambda_i s_i \right) \right\rangle_{\boldsymbol{J}}, \tag{312}$$

where $\langle \cdot \rangle_J$ denotes expectation w.r.t. the randomness of $\{J_{i\ell}\}$. Now, since $\{J_{i\ell}\}$ are $N(N-1)/2$ i.i.d., zero–mean Gaussian random variables with variance J^2/n, the expected number of solutions $\bar{M}(N)$ is given by

$$\bar{M}(N) = \left(\frac{N}{2\pi J^2} \right)^{N(N-1)/4} \int_{\mathbb{R}^{N(N-1)/2}} \mathrm{d}\boldsymbol{J} \exp \left\{ -\frac{N}{2J^2} \sum_{i>\ell} J_{i\ell}^2 \right\} \times$$
$$\sum_{\boldsymbol{s}} \int_K^\infty \cdots \int_K^\infty \mathrm{d}\boldsymbol{\lambda} \cdot \prod_{i=1}^N \delta \left(\sum_\ell J_{i\ell} s_\ell - \lambda_i s_i \right). \tag{313}$$

The next step is to represent each Dirac function as an inverse Fourier transform as we did earlier, i.e.,

$$\delta(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \mathrm{d}\omega e^{j\omega x} \qquad j = \sqrt{-1} \tag{314}$$

and then

$$\begin{aligned}
\bar{M}(N) = & \left(\frac{N}{2\pi J^2} \right)^{N(N-1)/4} \int_{\mathbb{R}^{N(N-1)/2}} \mathrm{d}\boldsymbol{J} \times \\
& \exp \left\{ -\frac{N}{2J^2} \sum_{i>\ell} J_{i\ell}^2 \right\} \cdot \sum_{\boldsymbol{s}} \int_K^\infty \cdots \int_K^\infty \mathrm{d}\boldsymbol{\lambda} \times \\
& \int_{\mathbb{R}^N} \frac{\mathrm{d}\boldsymbol{\omega}}{(2\pi)^N} \prod_{i=1}^N \exp \left\{ j\omega_i \left(\sum_\ell J_{i\ell} s_\ell - \lambda_i s_i \right) \right\} \\
= & \left(\frac{N}{2\pi J^2} \right)^{N(N-1)/4} \int_{\mathbb{R}^{N(N-1)/2}} \mathrm{d}\boldsymbol{J} \sum_{\boldsymbol{s}} \int_K^\infty \cdots \int_K^\infty \mathrm{d}\boldsymbol{\lambda} \times \\
& \int_{\mathbb{R}^N} \frac{\mathrm{d}\boldsymbol{\omega}}{(2\pi)^N} \exp \left\{ -\frac{N}{2J^2} \sum_{i>\ell} J_{i\ell}^2 + \right. \\
& \left. j \sum_{i>\ell} J_{i\ell} (\omega_i s_\ell + \omega_\ell s_i) - j \sum_i \omega_i s_i \lambda_i \right\}
\end{aligned} \tag{315}$$

We now use the so called Hubbard–Stratonovich transform, which is nothing but the identity

$$\int_{\mathbb{R}} \mathrm{d}x e^{ax^2+bx} \equiv \sqrt{\frac{\pi}{a}} e^{b^2/(4a)}, \tag{316}$$

with the assignments $a = n/(2J^2)$ and $b = \omega_i s_\ell + \omega_\ell s_i$ to obtain

$$\bar{M}(N) = \sum_{\boldsymbol{s}} \int_K^\infty \cdots \int_K^\infty \mathrm{d}\boldsymbol{\lambda} \int_{\mathbb{R}^N} \frac{\mathrm{d}\boldsymbol{\omega}}{(2\pi)^n} \times$$
$$\prod_{i=1}^N e^{-j\omega_i s_i \lambda_i} \prod_{i>\ell} \exp\{-(\omega_i s_\ell + \omega_\ell s_i)^2 J^2/(2N)\}. \tag{317}$$

Next observe that the summand does not actually depend on $\boldsymbol{s}$ because each s_i is multiplied by an integration variable that runs over $\mathbb{R}$ and thus the sign of s_i may be absorbed by this integration variable anyhow. Thus, all 2^N contributions are the same as that of $\boldsymbol{s} = (+1, \ldots, +1)$:

$$\bar{M}(N) = 2^N \int_K^\infty \cdots \int_K^\infty \mathrm{d}\boldsymbol{\lambda} \int_{\mathbb{R}^N} \frac{\mathrm{d}\boldsymbol{\omega}}{(2\pi)^N} \times$$
$$\prod_{i=1}^N e^{-j\omega_i \lambda_i} \prod_{i>\ell} \exp\{-(\omega_i + \omega_\ell)^2 J^2/(2N)\}. \tag{318}$$

Next use the following identity:

$$\frac{J^2}{2N} \sum_{i>\ell} (\omega_i + \omega_\ell)^2 = J^2 \frac{(N-1)}{2N} \sum_i \omega_i^2 + \frac{J^2}{N} \sum_{i>\ell} \omega_i \omega_\ell, \tag{319}$$

and so for large N,

$$\frac{J^2}{2N} \sum_{i>\ell} (\omega_i + \omega_\ell)^2 \approx \frac{J^2}{2} \sum_i \omega_i^2 + \frac{J^2}{N} \sum_{i>\ell} \omega_i \omega_\ell$$
$$\approx \frac{J^2}{2} \sum_i \omega_i^2 + \frac{J^2}{2N} \left(\sum_{i=1}^N \omega_i \right)^2. \tag{320}$$

Thus

$$\bar{M}(N) \approx 2^n \int_K^\infty \cdots \int_K^\infty \mathrm{d}\boldsymbol{\lambda} \int_{\mathbb{R}^n} \frac{\mathrm{d}\boldsymbol{\omega}}{(2\pi)^n} \times$$
$$\prod_{i=1}^n \exp\left\{ -j\omega_i \lambda_i - \frac{J^2}{2} \sum_{i=1}^n \omega_i^2 - \frac{J^2}{2n} \left(\sum_{i=1}^n \omega_i \right)^2 \right\}. \tag{321}$$

We now use again the Hubbard–Stratonovich transform

$$e^{a^2} \equiv \int_{\mathbb{R}} \frac{\mathrm{d}t}{2\pi} e^{j\sqrt{2}at - t^2/2} \tag{322}$$

and then, after changing variables $\lambda_i \to J\lambda_i$ and $J\omega_i \to \omega_i$, we get:

$$\begin{aligned}\bar{M}(N) \approx \frac{1}{\pi^n} \cdot \frac{1}{\sqrt{2\pi}} \int_{K/J}^{\infty} \cdots \int_{K/J}^{\infty} \mathrm{d}\boldsymbol{\lambda} \int_{\mathbb{R}} \mathrm{d}t e^{-t^2/2} \times \\ \prod_{i=1}^{n} \left[\int_{\mathbb{R}} \mathrm{d}\omega_i \exp\left\{ j\omega_i \left(-\lambda_i + \frac{t}{\sqrt{n}} \right) - \frac{1}{2} \sum_{i=1}^{n} \omega_i^2 \right\} \right]\end{aligned} \tag{323}$$

Changing the integration variable from $t/\sqrt{n}$ to t, this becomes

$$\begin{aligned}\bar{M}(N) &\approx \frac{1}{\pi^N} \cdot \frac{N}{\sqrt{2\pi}} \int_{\mathbb{R}} \mathrm{d}t e^{-Nt^2/2} \left[\int_{K/\lambda}^{\infty} \mathrm{d}\lambda \int_{\mathbb{R}} \mathrm{d}\omega e^{j\omega(t-\lambda) - \omega^2/2} \right]^N \\ &= \frac{1}{\pi^N} \cdot \frac{N}{\sqrt{2\pi}} \int_{\mathbb{R}} \mathrm{d}t e^{-Nt^2/2} \left[\sqrt{2\pi} \int_{K/\lambda}^{\infty} \mathrm{d}\lambda e^{-(t-\lambda)^2/2} \right]^N \\ &= \frac{1}{\pi^N} \cdot \frac{N}{\sqrt{2\pi}} \int_{\mathbb{R}} \mathrm{d}t e^{-N(t+K/J)^2/2} \left[\sqrt{2\pi} \int_{-\infty}^{t} \mathrm{d}\lambda e^{-\lambda^2/2} \right]^N \\ &= \frac{1}{\pi^N} \cdot \frac{N}{\sqrt{2\pi}} \int_{\mathbb{R}} \mathrm{d}t e^{-N(t+K/J)^2/2} \cdot [2\pi(1-Q(t))]^N \\ &= \frac{N}{\sqrt{2\pi}} \int_{\mathbb{R}} \mathrm{d}t \exp\left\{ -\frac{N}{2}(t+K/J)^2 + \ln[2(1-Q(t))] \right\} \\ &\approx \exp\left\{ N \cdot \max_t \left[\ln(2(1-Q(t)) - \frac{(t+K/J)^2}{2} \right] \right\}\end{aligned} \tag{324}$$

where in the last step, we used the saddle point method. The maximizing t zeroes out the derivative, i.e., it solves the equation

$$\frac{e^{-t^2/2}}{\sqrt{2\pi}[1-Q(t)]} = t + \frac{K}{J} \tag{325}$$

which for $K = 0$, gives exactly the asserted result about the capacity.

4.5 The Replica Method

The replica method is one of the most useful tools, which originally comes from statistical physics, but it finds its use in a variety of other fields, with communications and information theory included (e.g., multiuser detection). As we shall see, there are many models in

statistical physics, where the partition function Z depends, among other things, on some *random* parameters (to model disorder), and then Z, or $\ln Z$, becomes a random variable as well. Furthermore, it turns out that more often than not, the random variable $\frac{1}{N}\ln Z$ exhibits a concentration property, or in the jargon of physicists, a *self–averaging* property: in the thermodynamic limit of $N \to \infty$, it falls in the vicinity of its expectation $\frac{1}{N}\langle \ln Z \rangle$, with very high probability. Therefore, the computation of the per–particle free energy (and hence also many other physical quantities), for a typical realization of these random parameters, is associated with the computation of $\langle \ln Z \rangle$. The problem is that in most of the interesting cases, the exact closed form calculation of this expectation is extremely difficult if not altogether impossible. This is the point where the replica method enters into the picture.

Before diving into the description of the replica method, it is important to make a certain digression: This is a non–rigorous method, and it is not quite clear (yet) what are exactly the conditions under which it gives the correct result. Physicists tend to believe in it very strongly, because in many situations it gives results that make sense, agree with intuition, or make good fit to experimental results and/or simulation results. Moreover, in many cases, predictions made by the replica theory were confirmed by other, rigorous mathematical methods. The problem is that when there are no other means to test its validity, there is no certainty that it is credible and reliable. In such cases, it is believed that the correct approach would be to refer to the results it provides, as a certain educated guess or as a conjecture, rather than a solid scientific truth. Indeed, some reservations concerning the replica method have been raised in the literature [29], [118], [130]. Another, related method is the cavity method (see, e.g, [80, Chap. 19]), but it will not be covered in this work.

As we shall see shortly, the problematics of the replica method is not just that it depends on a certain interchangeability between a limit and an integral, but more severely, that the procedure that it proposes, is actually not even well–defined. In spite of these reservations, which are well known, the replica method has become highly popular and it is used extremely widely. A very partial list of (mostly relatively recent) articles that make use of this method includes [10], [11], [37], [41], [50], [51], [53], [58], [49],[52], [84], [83], [113],[114],[115], [107],

[126], and [127]. It is then appropriate to devote to the replica method some attention, and so, we shall indeed present its mean ideas, in the general level, up to a certain point. We shall not use, however, the replica method elsewhere in this work.

Consider then the calculation of $\boldsymbol{E}\ln Z$. The problem is that Z is a sum, and it is not easy to say something intelligent on the logarithm of a sum of many terms, let alone the expectation of this log–sum. If, instead, we had to deal with integer moments of Z, i.e., $\boldsymbol{E}Z^m$, this would have been much easier, because integer moments of sums, are sums of products. The idea is therefore to seek a way to relate moments $\boldsymbol{E}Z^m$ to $\boldsymbol{E}\ln Z$. This can be done if **real**, rather than just integer, moments are allowed. These could be related via the simple relation

$$\boldsymbol{E}\ln Z = \lim_{m\to 0}\frac{\boldsymbol{E}Z^m - 1}{m} = \lim_{m\to 0}\frac{\ln \boldsymbol{E}Z^m}{m} \tag{326}$$

provided that the expectation operator and the limit over m can be interchanged. But we know how to deal only with integer moments of m. The first courageous idea of the replica method, at this point, is to offer the following recipe: Compute $\boldsymbol{E}Z^m$, for a general positive integer m, and obtain an expression which is a function of m. Once this has been done, *forget* that m is an integer, and consider it as a *real* variable. Finally, use the above identity, taking the limit of $m \to 0$.

Beyond the technicality of interchanging the expectation operator with the limit, which applies in most conceivable cases, there is a more serious concern here, and this is that the above procedure is not well–defined, as mentioned earlier: We derive an expression $f(m) \stackrel{\Delta}{=} \boldsymbol{E}Z^m$, which is originally meant for m integer only, and then 'interpolate' in between integers by using the same expression, in other words, we take the analytic continuation. Actually, the right–most side of the above identity is $f'(0)$ where f' is the derivative of f. But there are infinitely many functions of a continuous variable m that match given values for integer values of m: If $f(m)$ is such, then $\tilde{f}(m) = f(m) + g(m)$ is good as well, for every g that vanishes on the integers, for example, take $g(m) = A\sin(\pi m)$. Nonetheless, $\tilde{f}'(0)$ might be different from $f'(0)$, and this is indeed the case with the example where g is sinusoidal. So in this step of the procedure there is some weakness, but this point is simply ignored.

After this introduction, let us now present the replica method on a concrete example, which is essentially taken from [80, Sect. 8.1]. In this example, $Z = \sum_{i=1}^{2^N} e^{-\beta E_i}$, where $\{E_i\}_{i=1}^{2^N}$ are i.i.d. random variables. In the sequel, we will work with this model quite a lot, after we see how it is relevant. It is called the *random energy model* (REM). For now, however, this is just a technical example on which we demonstrate the replica method. As the replica method suggests, let us first look at the integer moments. First, what we have is:

$$Z^m = \left[\sum_{i=1}^{2^N} e^{-\beta E_i}\right]^m = \sum_{i_1=1}^{2^N} \ldots \sum_{i_m=1}^{2^N} \exp\{-\beta \sum_{a=1}^{m} E_{i_a}\}. \tag{327}$$

The right–most side can be thought of as the partition function pertaining to a new system, consisting of m independent replicas (hence the name of the method) of the original system. Each configuration of the new system is indexed by an m–tuple $\boldsymbol{i} = (i_1, \ldots, i_m)$, where each i_a runs from 1 to 2^N, and the energy is $\sum_a E_{i_a}$. Let us now rewrite Z^m slightly differently:

$$\begin{aligned}
Z^m &= \sum_{i_1=1}^{2^N} \ldots \sum_{i_m=1}^{2^N} \exp\left\{-\beta \sum_{a=1}^{m} E_{i_a}\right\} \\
&= \sum_{\boldsymbol{i}} \exp\left\{-\beta \sum_{a=1}^{m} \sum_{j=1}^{2^N} \mathcal{I}(i_a = j) E_j\right\} \\
&= \sum_{\boldsymbol{i}} \exp\left\{-\beta \sum_{j=1}^{2^N} \sum_{a=1}^{m} \mathcal{I}(i_a = j) E_j\right\} \\
&= \sum_{\boldsymbol{i}} \prod_{j=1}^{2^N} \exp\left\{-\beta \sum_{a=1}^{m} \mathcal{I}(i_a = j) E_j\right\}
\end{aligned}$$

Let us now further suppose that each E_j is $\mathcal{N}(0, NJ^2/2)$, as is customary in the REM, for reasons that we shall see later on. Then, taking expectations w.r.t. this distribution, we get:

$$\begin{aligned}
\boldsymbol{E} Z^m &= \sum_{\boldsymbol{i}} \boldsymbol{E} \prod_{j=1}^{2^N} \exp\left\{-\beta \sum_{a=1}^{m} \mathcal{I}(i_a = j) E_j\right\} \\
&= \sum_{\boldsymbol{i}} \prod_{j=1}^{2^N} \exp\left\{\frac{\beta^2 N J^2}{4} \sum_{a,b=1}^{m} \mathcal{I}(i_a = j)\mathcal{I}(i_b = j)\right\}
\end{aligned}$$

$$= \sum_{\boldsymbol{i}} \exp\left\{\frac{\beta^2 N J^2}{4} \sum_{a,b=1}^{m} \sum_{j=1}^{2^N} \mathcal{I}(i_a = j)\mathcal{I}(i_b = j)\right\}$$
$$= \sum_{\boldsymbol{i}} \exp\left\{\frac{\beta^2 N J^2}{4} \sum_{a,b=1}^{m} \mathcal{I}(i_a = i_b)\right\}.$$

We now define an $m \times m$ binary matrix Q, called the *overlap matrix*, whose entries are $Q_{ab} = \mathcal{I}(i_a = i_b)$. Note that the summand in the last expression depends on $\boldsymbol{i}$ only via Q. Let $M_N(Q)$ denote the number of configurations $\{\boldsymbol{i}\}$ whose overlap matrix is Q. We have to exhaust all possible overlap matrices, which are all binary symmetric matrices with 1's on the main diagonal. Observe that the number of such matrices is $2^{m(m-1)/2}$ whereas the number of configurations is 2^{Nm}. Thus we are dividing the exponentially large number of configurations into a relatively small number (independent of N) of equivalence classes, something that rings the bell of the method of types. Let us suppose, for now, that there is some function $s(Q)$ such that $M_N(Q) \doteq e^{Ns(Q)}$, and so

$$\boldsymbol{E}Z^m \doteq \sum_{Q} e^{Ng(Q)} \tag{328}$$

with:

$$g(Q) = \frac{\beta^2 J^2}{4} \sum_{a,b=1}^{m} Q_{ab} + s(Q). \tag{329}$$

From this point onward, the strategy is to use the saddle point method. Note that the function $g(Q)$ is symmetric under replica permutations: let π be a permutation operator of m objects and let Q^π be the overlap matrix with entries $Q^\pi_{ab} = Q_{\pi(a)\pi(b)}$. Then, $g(Q^\pi) = g(Q)$. This property is called *replica symmetry* (RS), and this property is inherent to the replica method. In light of this, the first natural idea that comes to our mind is to postulate that the saddle point is symmetric too, in other words, to assume that the saddle–point Q has 1's on its main diagonal and all other entries are taken to be the same (binary) value, call it q_0. Now, there are only two possibilities:

- $q_0 = 0$ and then $M_N(Q) = 2^N(2^N-1)\cdots(2^N-m+1)$, which implies that $s(Q) = m\ln 2$, and then $g(Q) = g_0(Q) \stackrel{\Delta}{=} m(\beta^2 J^2/4 + \ln 2)$, thus $(\ln \boldsymbol{E}Z^m)/m = \beta^2 J^2/4 + \ln 2$, and so

is the limit as $m \to 0$. Later on, we will compare this with the result obtained from a more rigorous derivation.

- $q_0 = 1$, which means that all components of $\boldsymbol{i}$ are the same, and then $M_N(Q) = 2^N$, which means that $s(Q) = \ln 2$ and so, $g(Q) = g_1(Q) \stackrel{\Delta}{=} m^2\beta^2 J^2/4 + \ln 2$.

Now, one should check which one of these saddle points is the dominant one, depending on β and m. For $m \geq 1$, the behavior is dominated by $\max\{g_0(Q), g_1(Q)\}$, which is $g_1(Q)$ for $\beta \geq \beta_c(m) \stackrel{\Delta}{=} \frac{2}{J}\sqrt{\ln 2/m}$, and $g_0(Q)$ otherwise. For $m < 1$ (which is, in fact, the relevant case for $m \to 0$), one should look at $\min\{g_0(Q), g_1(Q)\}$, which is $g_0(Q)$ in the high–temperature range. As it turns out, in certain regions in the β–m plane, we must back off from the 'belief' that dominant configurations are *purely* symmetric, and resort to the quest for dominant configurations with a lower level of symmetry. The first step, after having exploited the purely symmetric case above, is called *one–step replica symmetry breaking* (1RSB), and this means some partition of the set $\{1, 2, \ldots, m\}$ into two complementary subsets (say, of equal size) and postulating a saddle point Q of the following structure:

$$Q_{ab} = \begin{cases} 1 & a = b \\ q_0 & a \text{ and } b \text{ are in the same subset} \\ q_1 & a \text{ and } b \text{ are in different subsets} \end{cases} \tag{330}$$

In further steps of symmetry breaking, one may split $\{1, 2, \ldots, m\}$ to a larger number of subsets or even introduce certain hierarchical structures. The replica method includes a variety of heuristic guidelines in this context. We will not delve into them any further in the framework of this monograph, but the interested reader can easily find more details in the literature, specifically, in [80].

5 Interacting Particles and Phase Transitions

In this chapter, we introduce additional physics background pertaining to systems with interacting particles. When the interactions among the particles are sufficiently significant, the system exhibits a certain collective behavior that, in the thermodynamic limit, may be subjected to *phase transitions*, i.e., abrupt changes in the behavior and the properties of the system in the presence of a gradual change in an external control parameter, like temperature, pressure, or magnetic field.

Analogous abrupt transitions, in the asymptotic behavior, are familiar to us also in information theory. For example, when the signal–to–noise (SNR) of a coded communication system crosses the value for which the capacity meets the coding rate, there is an abrupt transition between reliable communication and unreliable communication, where the probability of error essentially jumps between zero and one. Are there any relationships between the phase transitions in physics and those in information theory? It turns out that the answer is affirmative to a large extent. By mapping the mathematical formalism of the coded communication problem to that of an analogous physical system with interacting particles, some insights on these relations can be obtained. We will see later on that these insights can also be harnessed for sharper analysis tools.

5.1 Introduction – Sources of Interaction

As already mentioned in the introductory part of the previous chapter, so far, we have dealt almost exclusively with systems that have additive Hamiltonians, $\mathcal{E}(\boldsymbol{x}) = \sum_i \mathcal{E}(x_i)$, which means that the particles are i.i.d. and there is no interaction: each particle behaves as if it was alone in the world. In Nature, of course, this is seldom really the case. Sometimes this is still a reasonably good approximation, but in many other cases, the interactions are appreciably strong and cannot be neglected. Among the different particles there could be many sorts of mutual forces, such as mechanical, electrical, or magnetic forces, etc. There could also be interactions that stem from quantum–mechanical effects: As described at the end of Chapter

2, Pauli's exclusion principle asserts that for Fermions (e.g., electrons), no quantum state can be populated by more than one particle. This gives rise to a certain mutual influence between particles. Another type of interaction stems from the fact that the particles are indistinguishable, so permutations between them are not considered as distinct states. For example, referring again to the example of quantum statistics, at the end of Chapter 2, had the N particles been statistically independent, the resulting partition function would be

$$\begin{aligned} Z_N(\beta) &= \left[\sum_r e^{-\beta\epsilon_r}\right]^N \\ &= \sum_{\boldsymbol{N}:\ \sum_r N_r = N} \frac{N!}{\prod_r N_r!} \cdot \exp\left\{-\beta \sum_r N_r \epsilon_r\right\} \end{aligned} \tag{331}$$

whereas in eq. (82), the combinatorial factor, $N!/\prod_r N_r!$, that distinguishes between the various permutations among the particles, is absent. This introduces dependency, which physically means interaction.[16]

5.2 Models of Interacting Particles

The simplest forms of deviation from the purely additive Hamiltonian structure are those that consists, in addition to the individual energy terms, $\{\mathcal{E}(x_i)\}$, also terms that depend on pairs, and/or triples, and/or even larger cliques of particles. In the case of purely pairwise interactions, this means a structure like the following:

$$\mathcal{E}(\boldsymbol{x}) = \sum_{i=1}^{N} \mathcal{E}(x_i) + \sum_{(i,j)} \varepsilon(x_i, x_j) \tag{332}$$

where the summation over pairs can be defined over all pairs $i \neq j$, or over some of the pairs, according to a given rule, e.g., depending on the distance between particle i and particle j, and according to the geometry of the system, or according to a certain graph whose edges connect the relevant pairs of variables (that in turn, are designated as nodes).

[16] Indeed, in the case of the boson gas, there is a well–known effect referred to as *Bose–Einstein condensation*, which is actually a phase transition, but phase transitions can occur only in systems of interacting particles, as will be discussed in this set of lectures.

For example, in a one–dimensional array (a lattice) of particles, a customary model accounts for interactions between neighboring pairs only, neglecting more remote ones, thus the second term above would be $\sum_i \varepsilon(x_i, x_{i+1})$. A well known special case of this is that of a polymer [28], or a solid with crystal lattice structure, where in the one–dimensional version of the model, atoms are thought of as a chain of masses connected by springs (see left part of Fig. 8), i.e., an array of coupled harmonic oscillators. In this case, $\varepsilon(x_i, x_{i+1}) = \frac{1}{2}K(u_{i+1} - u_i)^2$, where K is a constant and u_i is the displacement of the i-th atom from its equilibrium location, i.e., the potential energies of the springs. This model has an easy analytical solution (by applying a Fourier transform on the sequence $\{u_i\}$), where by "solution", we mean a closed–form, computable formula for the log–partition function, at least in the thermodynamic limit. In higher dimensional arrays (or lattices), similar interactions apply, there are

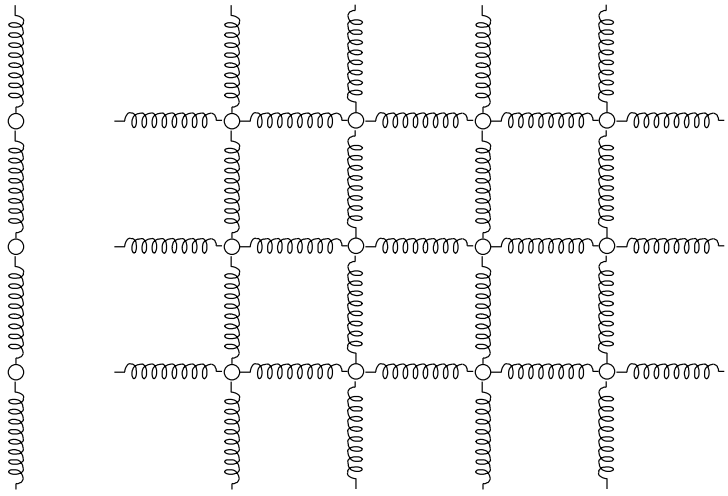


Figure 8: Elastic interaction forces between adjacent atoms in a one–dimensional lattice (left part of the figure) and in a two–dimensional lattice (right part).

just more neighbors to each site, from the various directions (see right part of Fig. 8). In a system where the particles are mobile and hence their locations vary and have no geometrical structure, like in a gas, the interaction terms are also potential energies pertaining to the mutual forces (see Fig. 9), and these normally depend solely on the distances $\|\vec{r}_i - \vec{r}_j\|$. For

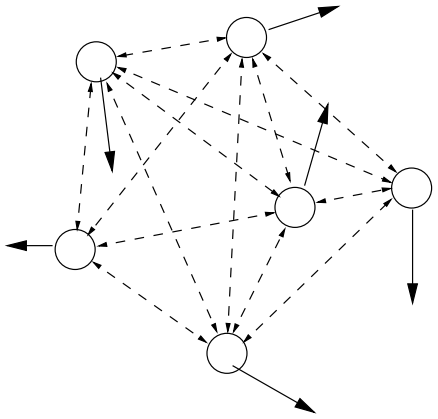


Figure 9: Mobile particles and mutual forces between them.

example, in a non–ideal gas,

$$\mathcal{E}(\boldsymbol{x}) = \sum_{i=1}^{N} \frac{\|\vec{p}_i\|^2}{2m} + \sum_{i \neq j} V(\|\vec{r}_i - \vec{r}_j\|). \tag{333}$$

A very simple special case is that of hard spheres (Billiard balls), without any forces, where

$$V(\|\vec{r}_i - \vec{r}_j\|) = \begin{cases} \infty & \|\vec{r}_i - \vec{r}_j\| < 2R \\ 0 & \|\vec{r}_i - \vec{r}_j\| \geq 2R \end{cases} \tag{334}$$

which expresses the simple fact that balls cannot physically overlap. This model can (and indeed is) being used to obtain bounds on sphere–packing problems, which are very relevant to channel coding theory. This model is not solvable in general and its solution is an open challenge. The interested reader can find more details on this line of research, in several articles, such as [18], [64], [99], [97], on the physical aspects, as well as [91] and [93] on the application to coding theory.

Yet another example of a model, or more precisely, a very large class of models with interactions, are those of magnetic materials. These models will closely accompany our discussions from this point onward, because as described in the Introduction, some of them lend themselves to mathematical formalisms that are analogous to those of coding problems, as we shall see. Few of these models are solvable, most of them are not. For the purpose of our discussion, a magnetic material is one for which the relevant property of each particle is its *magnetic moment*. The magnetic moment is a vector proportional to the angular

momentum of a revolving charged particle (like a rotating electron, or a current loop), or the *spin*, and it designates the intensity of its response to the net magnetic field that this particle 'feels'. This magnetic field may be the superposition of an externally applied magnetic field and the magnetic fields generated by the neighboring spins.

Quantum mechanical considerations dictate that each spin, which will be denoted by s_i, is quantized, that is, it may take only one out of finitely many values. In the simplest case to be adopted in our study – two values only. These will be designated by $s_i = +1$ ("spin up") and $s_i = -1$ ("spin down"), corresponding to the same intensity, but in two opposite directions, one parallel to the magnetic field, and the other – anti-parallel (see Fig. 10). The

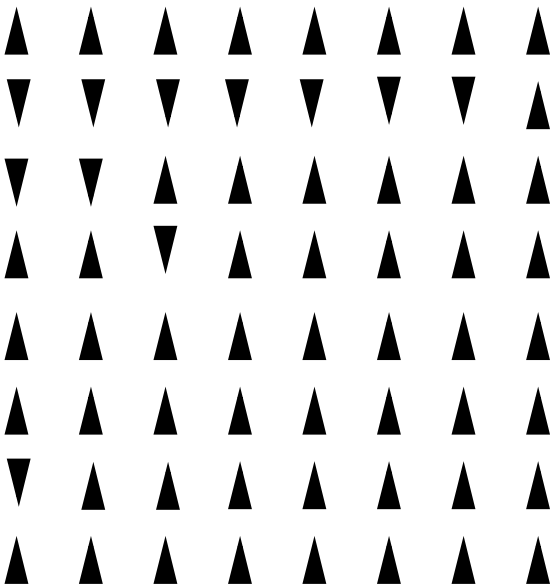


Figure 10: Illustration of a spin array on a square lattice.

Hamiltonian associated with an array of spins $\boldsymbol{s} = (s_1, \ldots, s_N)$ is customarily modeled (up to certain constants that, among other things, accommodate for the physical units) with a structure like this:

$$\mathcal{E}(\boldsymbol{s}) = -B \cdot \sum_{i=1}^{N} s_i - \sum_{(i,j)} J_{ij} s_i s_j, \tag{335}$$

where B is the externally applied magnetic field and $\{J_{ij}\}$ are the coupling constants that designate the levels of interaction between spin pairs, and they depend on properties of

the magnetic material and on the geometry of the system. The first term accounts for the contributions of potential energies of all spins due to the magnetic field, which in general, are given by the inner product $\vec{B} \cdot \vec{s}_i$, but since each $\vec{s}_i$ is either parallel or anti-parallel to $\vec{B}$, as said, these boil down to simple products, where only the sign of each s_i counts. Since $P(\boldsymbol{s})$ is proportional to $e^{-\beta \mathcal{E}(\boldsymbol{s})}$, the spins 'prefer' to be parallel, rather than anti-parallel to the magnetic field. The second term in the above Hamiltonian accounts for the interaction energy. If J_{ij} are all positive, they also prefer to be parallel to one another (the probability for this is larger), which is the case where the material is called *ferromagnetic* (like iron and nickel). If they are all negative, the material is *antiferromagnetic*. In the mixed case, it is called a *spin glass*. In the latter, the behavior is rather complicated, as we shall see later on.

Of course, the above model for the Hamiltonian can (and, in fact, is being) generalized to include interactions formed also, by triples, quadruples, or any fixed size p (that does not grow with N) of spin–cliques. At this point, it is instructive to see the relation between spin–array models (especially, those that involve large cliques of spins) to linear channel codes, which was first identified by Sourlas [111],[112]. Consider a linear code defined by a set of m parity–check equations (in $GF(2)$), each involving the modulo–2 sum of some subset of the components of the codeword $\boldsymbol{x}$. I.e., the ℓ–th parity–check equation is:

$$x_{i_1^\ell} \oplus x_{i_2^\ell} \oplus \cdots \oplus x_{i_{k_\ell}^\ell} = 0, \quad \ell = 1, \ldots, m, \tag{336}$$

where i_j^ℓ is the index of the j–th bit that takes part in the ℓ–th parity–check equation and k_ℓ is the number of bits involved in that equation. Transforming from $x_i \in \{0, 1\}$ to $s_i \in \{-1, +1\}$ via $s_i = 1 - 2x_i$, this is equivalent to

$$s_{i_1^\ell} s_{i_2^\ell} \cdots s_{i_{k_\ell}^\ell} = 1, \quad \ell = 1, \ldots, m. \tag{337}$$

The maximum a–posteriori (MAP) decoder estimates $\boldsymbol{s}$ based on the posterior

$$P(\boldsymbol{s}|\boldsymbol{y}) = \frac{P(\boldsymbol{s})P(\boldsymbol{y}|\boldsymbol{s})}{Z(\boldsymbol{y})}; \quad Z(\boldsymbol{y}) = \sum_{\boldsymbol{s}} P(\boldsymbol{s})P(\boldsymbol{y}|\boldsymbol{s}) = P(\boldsymbol{y}), \tag{338}$$

where $P(\boldsymbol{s})$ is normally assumed uniform over the codewords (we will elaborate on this posterior distribution function later on). Assuming, e.g., a binary symmetric channel (BSC)

or a Gaussian channel $P(\boldsymbol{y}|\boldsymbol{s})$, the relevant distance between the codeword $\boldsymbol{s} = (s_1, \ldots, s_N)$ and the channel output $\boldsymbol{y} = (y_1, \ldots, y_N)$ is proportional to $\|\boldsymbol{s} - \boldsymbol{y}\|^2 = \text{const.} - 2\sum_i s_i y_i$. Thus, $P(\boldsymbol{s}|\boldsymbol{y})$ can be thought of as a B–G distribution with Hamiltonian

$$\mathcal{E}(\boldsymbol{s}|\boldsymbol{y}) = -J \sum_{i=1}^{N} s_i y_i + \sum_{\ell=1}^{m} \phi(s_{i_1^\ell} s_{i_2^\ell} \cdots s_{i_{k_\ell}^\ell}) \tag{339}$$

where J is some constant (depending on the channel parameters), the function $\phi(u)$ vanishes for $u = 1$ and becomes infinite for $u \neq 1$, and the partition function given by the denominator of $P(\boldsymbol{s}|\boldsymbol{y})$. The first term plays the analogous role to that of the contribution of the magnetic field in a spin system model, where each 'spin' s_i 'feels' a different magnetic field proportional to y_i, and the second term accounts for the interactions among cliques of spins. In the case of low–density parity check (LDPC) codes, where each parity check equation involves only a small number of bits $\{s_i\}$, these interaction terms amount to cliques of relatively small sizes.[17] For a general code, the second term is replaced by $\phi_{\mathcal{C}}(\boldsymbol{s})$, which is zero for $\boldsymbol{s} \in \mathcal{C}$ and infinite otherwise.

Another aspect of this model of a coded communication system pertains to calculations of mutual information and capacity. The mutual information between $\boldsymbol{S}$ and $\boldsymbol{Y}$ is, of course, given by

$$I(\boldsymbol{S}; \boldsymbol{Y}) = H(\boldsymbol{Y}) - H(\boldsymbol{Y}|\boldsymbol{S}). \tag{340}$$

The second term is easy to calculate for every additive channel – it is simply the entropy of the additive noise. The first term is harder to calculate:

$$H(\boldsymbol{Y}) = -\boldsymbol{E}\{\ln P(\boldsymbol{Y})\} = -\boldsymbol{E}\{\ln Z(\boldsymbol{Y})\}. \tag{341}$$

Thus, we are facing a problem of calculating the free energy of a spin system with random magnetic fields designated by the components of $\boldsymbol{Y}$. This is the kind of calculations we

[17] Error correction codes can be represented by bipartite graphs with two types of nodes: variable nodes corresponding to the various s_i and function nodes corresponding to cliques. There is an edge between variable node i and function node j if s_i is a member in clique j. Of course each s_i may belong to more than one clique. When all cliques are of size 2, there is no need for the function nodes, as edges between nodes i and j simply correspond to parity check equations involving s_i and s_j.

mentioned earlier in the context of the replica method. Indeed, the replica method is used extensively in this context.

As we shall see in the sequel, it is also customary to introduce an inverse temperature parameter β, by defining

$$P_\beta(\boldsymbol{s}|\boldsymbol{y}) = \frac{P^\beta(\boldsymbol{s})P^\beta(\boldsymbol{y}|\boldsymbol{s})}{Z(\beta|\boldsymbol{y})} = \frac{e^{-\beta\mathcal{E}(\boldsymbol{s}|\boldsymbol{y})}}{Z(\beta|\boldsymbol{y})} \tag{342}$$

where β controls the sharpness of the posterior distribution and

$$Z(\beta|\boldsymbol{y}) = \sum_{\boldsymbol{s}} e^{-\beta\mathcal{E}(\boldsymbol{s}|\boldsymbol{y})}. \tag{343}$$

The motivations of this will be discussed extensively later on.

Finally, it should be pointed out that the analogies between models and magnetic materials and models of communications and signal processing are not limited to the application described above. Consider, for example, the very common signal model

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{s} + \boldsymbol{w}, \tag{344}$$

where $\boldsymbol{H}$ is a matrix (with either deterministic or random entries) and $\boldsymbol{w}$ is a Gaussian noise vector, with i.i.d. components, independent of $\boldsymbol{s}$ (and $\boldsymbol{H}$). In this case, the posterior, $P_\beta(\boldsymbol{s}|\boldsymbol{y})$, is proportional to

$$\exp\left\{-\frac{\beta}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{s}\|^2\right\}, \tag{345}$$

where the exponent (after expansion of the norm), clearly includes an "external–field term," proportional to $\boldsymbol{y}^T\boldsymbol{H}\boldsymbol{s}$, and a "pairwise spin–spin interaction term," proportional to $\boldsymbol{s}^T\boldsymbol{R}\boldsymbol{s}$, where $\boldsymbol{R} = \boldsymbol{H}^T\boldsymbol{H}$.

We will get back to this important class of models, as well as its many extensions, shortly. But before that, we discuss a very important effect that exists in some systems with strong interactions (both in magnetic materials and in other models): the effect of *phase transitions*.

5.3 A Qualitative Discussion on Phase Transitions

As was mentioned in the introductory paragraph of this chapter, a phase transition means an abrupt change in the collective behavior of a physical system, as we change gradually one

of the externally controlled parameters, like the temperature, pressure, or magnetic field. The most common example of a phase transition in our everyday life is the water that we boil in the kettle when we make coffee, or when it turns into ice as we put it in the freezer.

What exactly these phase transitions are? Before we refer to this question, it should be noted that there are also "phase transitions" in the behavior of communication systems: We already mentioned the phase transition that occurs as the SNR passes a certain limit (for which capacity crosses the coding rate), where there is a sharp transition between reliable and unreliable communication, i.e., the error probability (almost) 'jumps' from 0 to 1 or vice versa. Another example is the phenomenon of threshold effects in highly non–linear communication systems (e.g., FM, PPM, FPM, etc., see [123, Chap. 8]).

Are there any relationships between these phase transitions and those of physics? We will see shortly that the answer is generally affirmative. In physics, phase transitions can occur only if the system has interactions. Consider, the above example of an array of spins with $B = 0$, and let us suppose that all $J_{ij} > 0$ are equal, and thus will be denoted commonly by J. Then,

$$P(\boldsymbol{s}) = \frac{\exp\left\{\beta J \sum_{(i,j)} s_i s_j\right\}}{Z(\beta)} \tag{346}$$

and, as mentioned earlier, this is a ferromagnetic model, where all spins 'like' to be in the same direction, especially when β and/or J is large. In other words, the interactions, in this case, tend to introduce *order* into the system. On the other hand, the second law talks about maximum entropy, which tends to increase the *disorder*. So there are two conflicting effects here. Which one of them prevails?

The answer turns out to depend on temperature. Recall that in the canonical ensemble, equilibrium is attained at the point of minimum free energy $f = \epsilon - Ts(\epsilon)$. Now, T plays the role of a weighting factor for the entropy. At low temperatures, the weight of the second term of f is small, and minimizing f is approximately equivalent to minimizing ϵ, which is obtained by states with a high level of order, as $\mathcal{E}(\boldsymbol{s}) = -J \sum_{(i,j)} s_i s_j$, in this example. As T grows, however, the weight of the term $-Ts(\epsilon)$ increases, and $\min f$, becomes more and

more equivalent to $\max s(\epsilon)$, which is achieved by states with a high level of disorder (see Fig. 11). Thus, the order–disorder characteristics depend primarily on temperature. It turns

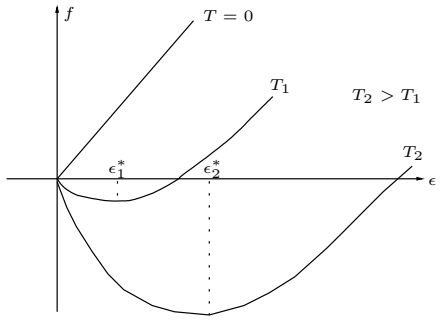


Figure 11: Qualitative graphs of $f(\epsilon)$ at various temperatures. The minimizing ϵ increases with T.

out that for some magnetic systems of this kind, this transition between order and disorder may be abrupt, in which case, we call it a *phase transition*. At a certain critical temperature, called the *Curie temperature*, there is a sudden transition between order and disorder. In the ordered phase, a considerable fraction of the spins align in the same direction, which means that the system is spontaneously magnetized (even without an external magnetic field), whereas in the disordered phase, about half of the spins are in either direction, and then the net magnetization vanishes. This happens if the interactions, or more precisely, their dimension in some sense, is strong enough.

What is the mathematical significance of a phase transition? If we look at the partition function, $Z(\beta)$, which is the key to all physical quantities of interest, then for every finite N, this is simply the sum of finitely many exponentials in β and therefore it is continuous and differentiable infinitely many times. So what kind of abrupt changes could there possibly be in the behavior of this function? It turns out that while this is true for all finite N, it is no longer necessarily true if we look at the thermodynamical limit, i.e., if we look at the behavior of

$$\phi(\beta) = \lim_{N\to\infty} \frac{\ln Z(\beta)}{N}. \tag{347}$$

While $\phi(\beta)$ must be continuous for all $\beta > 0$ (since it is convex), it need not necessarily have continuous derivatives. Thus, a phase transition, if exists, is fundamentally an asymptotic

property, it may exist in the thermodynamical limit only. While a physical system is, after all finite, it is nevertheless well approximated by the thermodynamical limit when it is very large. By the same token, if we look at the analogy with a coded communication system: for any finite block–length n, the error probability is a smooth function of the SNR, but in the limit of large n, it behaves like a step function that jumps between 0 and 1 at the critical SNR. As said earlier, we shall see that the two things are related.

Back to the physical aspects, the above discussion explains also why a system without interactions, where all $\{x_i\}$ are i.i.d., cannot have phase transitions. In this case, $Z_N(\beta) = [Z_1(\beta)]^N$, and so, $\phi(\beta) = \ln Z_1(\beta)$, which is always a smooth function without any irregularities. For a phase transition to occur, the particles must behave in some collective manner, which is the case only if interactions take place.

There is a distinction between two types of phase transitions:

- If $\phi(\beta)$ has a discontinuous first order derivative, then this is called a *first order phase transition*.
- If $\phi(\beta)$ has a continuous first order derivative, but a discontinuous second order derivative then this is called a *second order phase transition*, or a *continuous phase transition*.

We can talk, of course, about phase transitions w.r.t. additional parameters other than temperature. In the above magnetic example, if we introduce back the magnetic field B into the picture, then Z, and hence also ϕ, become functions of B too. If we then look at derivative of

$$\begin{aligned}\phi(\beta, B) &= \lim_{N\to\infty} \frac{\ln Z(\beta, B)}{N} \\ &= \lim_{N\to\infty} \frac{1}{N} \ln \left[\sum_{\boldsymbol{s}} \exp \left\{ \beta B \sum_{i=1}^{N} s_i + \beta J \sum_{(i,j)} s_i s_j \right\} \right] \end{aligned} \tag{348}$$

w.r.t. the product (βB), which multiplies the magnetization, $\sum_i s_i$, at the exponent, this

would give exactly the average magnetization per spin

$$m(\beta, B) = \left\langle \frac{1}{N} \sum_{i=1}^{N} S_i \right\rangle, \tag{349}$$

and this quantity might not always be continuous. Indeed, as mentioned earlier, below the Curie temperature there might be a spontaneous magnetization. If $B \downarrow 0$, then this magnetization is positive, and if $B \uparrow 0$, it is negative, so there is a discontinuity at $B = 0$. We shall see this more concretely later on.

5.4 Phase Transitions of the Rate–Distortion Function

We already mentioned that in the realm of information theory, there are phenomena that resemble phase transitions, like threshold effects in non–linear communication systems and the abrupt transitions from reliable to unreliable communication. Before we describe (in the next sections) in detail, models of interacting spins, with and without phase transitions, we pause, in this section, to discuss yet another type of a quite non–trivial phase transition, which occurs in the world of information theory: a phase transition in the behavior of the rate–distortion function. This type of a phase transition is very different from the class of phase transitions we discussed thus far: it stems from the optimization of the distribution of the reproduction variable, which turns out to have a rather irregular behavior, as we shall see. The derivation and the results, in this section, follow the paper by Rose [101]. This section can be skipped without loss of continuity.

We have seen in Chapter 2 that the rate–distortion function of a source $P = \{p(x),\ x \in \mathcal{X}\}$ can be expressed as

$$R(D) = -\min_{\beta \geq 0} \left[\beta D + \sum_{x} p(x) \ln \left(\sum_{y} q(y) e^{-\beta d(x,y)} \right) \right] \tag{350}$$

where $Q = \{q(y), y \in \mathcal{Y}\}$ is the output marginal of the test channel, which is also the one that minimizes this expression. We are now going to take a closer look at this function in the context of the quadratic distortion function $d(x, y) = (x - y)^2$. As said, the optimum Q

is the one that minimizes the above expression, or equivalently, the free energy

$$f(Q) = -\frac{1}{\beta}\sum_x p(x)\ln\left(\sum_y q(y)e^{-\beta d(x,y)}\right) \tag{351}$$

and in the continuous case, summations should be replaced by integrals:

$$f(Q) = -\frac{1}{\beta}\int_{-\infty}^{+\infty}\mathrm{d}x p(x)\ln\left(\int_{-\infty}^{+\infty}\mathrm{d}y q(y)e^{-\beta d(x,y)}\right). \tag{352}$$

Consider the representation of the random variable Y as a function of $U \sim \mathrm{unif}[0,1]$, and then, instead of optimizing Q, one should optimize the function $y(u)$ in:

$$f(y(\cdot)) = -\frac{1}{\beta}\int_{-\infty}^{+\infty}\mathrm{d}x p(x)\ln\left(\int_0^1 \mathrm{d}\mu(u)e^{-\beta d(x,y(u))}\right), \tag{353}$$

where $\mu(\cdot)$ is the Lebesgue measure (the uniform measure). A necessary condition for optimality,[18] which must hold for almost every u is:

$$\int_{-\infty}^{+\infty}\mathrm{d}x p(x)\cdot\left[\frac{e^{-\beta d(x,y(u))}}{\int_0^1 \mathrm{d}\mu(u')e^{-\beta d(x,y(u'))}}\right]\cdot\frac{\partial d(x,y(u))}{\partial y(u)} = 0. \tag{354}$$

Now, let us define the *support* of y as the set of values that y may possibly take on. Thus, this support is a subset of the set of all points $\{y_0 = y(u_0)\}$ for which:

$$\int_{-\infty}^{+\infty}\mathrm{d}x p(x)\cdot\left[\frac{e^{-\beta d(x,y_0)}}{\int_0^1 \mathrm{d}\mu(u')e^{-\beta d(x,y(u'))}}\right]\cdot\left.\frac{\partial d(x,y(u))}{\partial y(u)}\right|_{y(u)=y_0} = 0. \tag{355}$$

This is because y_0 must be a point that is obtained as $y(u)$ for some u. Let us define now the posterior:

$$q(u|x) = \frac{e^{-\beta d(x,y(u))}}{\int_0^1 \mathrm{d}\mu(u')e^{-\beta d(x,y(u'))}}. \tag{356}$$

Then,

$$\int_{-\infty}^{+\infty}\mathrm{d}x p(x)q(u|x)\cdot\frac{\partial d(x,y(u))}{\partial y(u)} = 0. \tag{357}$$

But $p(x)q(u|x)$ is a joint distribution $p(x,u)$, which can also be thought of as $\mu(u)p(x|u)$. So, if we divide the last equation by $\mu(u)$, we get, for almost all u:

$$\int_{-\infty}^{+\infty}\mathrm{d}x p(x|u)\frac{\partial d(x,y(u))}{\partial y(u)} = 0. \tag{358}$$

[18] The details are in [101], but intuitively, instead of a function $y(u)$ of a continuous variable u, think of a vector $\boldsymbol{y}$ whose components are indexed by u, which take on values in some grid of $[0,1]$. In other words, think of the argument of the logarithmic function as $\sum_{u=0}^{1} e^{-\beta d(x,y_u)}$.

Now, let us see what happens in the case of the quadratic distortion, $d(x,y) = (x-y)^2$. Suppose that the support of Y includes some interval $\mathcal{I}_0$ as a subset. For a given u, $y(u)$ is nothing other than a number, and so the optimality condition must hold for every $y \in \mathcal{I}_0$. In the case of the quadratic distortion, this optimality criterion means

$$\int_{-\infty}^{+\infty} \mathrm{d}x p(x)\lambda(x)(x-y)e^{-\beta(x-y)^2} = 0, \quad \forall y \in \mathcal{I}_0 \tag{359}$$

with

$$\lambda(x) \stackrel{\Delta}{=} \frac{1}{\int_0^1 \mathrm{d}\mu(u)e^{-\beta d(x,y(u))}} = \frac{1}{\int_{-\infty}^{+\infty} \mathrm{d}y q(y)e^{-\beta d(x,y)}}, \tag{360}$$

or, equivalently,

$$\int_{-\infty}^{+\infty} \mathrm{d}x p(x)\lambda(x)\frac{\partial}{\partial y}\left[e^{-\beta(x-y)^2}\right] = 0. \tag{361}$$

Since this must hold for all $y \in \mathcal{I}_0$, then all derivatives of the l.h.s. must vanish within $\mathcal{I}_0$, i.e.,

$$\int_{-\infty}^{+\infty} \mathrm{d}x p(x)\lambda(x)\frac{\partial^n}{\partial y^n}\left[e^{-\beta(x-y)^2}\right] = 0. \tag{362}$$

Now, considering the Hermitian polynomials

$$H_n(z) \stackrel{\Delta}{=} e^{\beta z^2}\frac{\mathrm{d}^n}{\mathrm{d}z^n}(e^{-\beta z^2}) \tag{363}$$

this requirement means

$$\int_{-\infty}^{+\infty} \mathrm{d}x p(x)\lambda(x)H_n(x-y)e^{-\beta(x-y)^2} = 0. \tag{364}$$

In words: $\lambda(x)p(x)$ is orthogonal to all Hermitian polynomials of order ≥ 1 w.r.t. the weight function $e^{-\beta z^2}$. Now, as is argued in the paper, since these polynomials are complete in $L^2(e^{-\beta z^2})$, we get

$$p(x)\lambda(x) = \text{const.} \tag{365}$$

because $H_0(z) \equiv 1$ is the only basis function orthogonal to all $H_n(z)$, $n \geq 1$. This yields, after normalization:

$$p(x) = \sqrt{\frac{\beta}{\pi}}\int_0^1 \mathrm{d}\mu(u)e^{-\beta(x-y(u))^2}$$

$$= \sqrt{\frac{\beta}{\pi}} \int_{-\infty}^{+\infty} \mathrm{d}y q(y) e^{-\beta(x-y)^2}$$
$$= Q \star \mathcal{N}\left(0, \frac{1}{2\beta}\right), \tag{366}$$

where $\star$ denotes convolution. The interpretation of the last equation is simple: the marginal of X is given by the convolution between the marginal of Y and the zero–mean Gaussian distribution with variance $D = 1/(2\beta)$ ($= kT/2$ of the equipartition theorem, as we already saw in Chapter 2). This means that X must be representable as

$$X = Y + Z \tag{367}$$

where $Z \sim \mathcal{N}\left(0, \frac{1}{2\beta}\right)$ and independent of Y. As is well known, this is exactly what happens when $R(D)$ coincides with its Gaussian lower bound, a.k.a. the Shannon lower bound (SLB). Here is a reminder of this:

$$\begin{aligned}
R(D) &= h(X) - \max_{\boldsymbol{E}(X-Y)^2 \leq D} h(X|Y) \\
&= h(X) - \max_{\boldsymbol{E}(X-Y)^2 \leq D} h(X-Y|Y) \\
&\geq h(X) - \max_{\boldsymbol{E}(X-Y)^2 \leq D} h(X-Y) \\
&= h(X) - \max_{\boldsymbol{E}Z^2 \leq D} h(Z) \quad Z \stackrel{\Delta}{=} X - Y \\
&\geq h(X) - \frac{1}{2}\ln(2\pi e D) \\
&\stackrel{\Delta}{=} R_S(D),
\end{aligned} \tag{368}$$

where $R_S(D)$ designates the SLB. The conclusion then is that if the support of Y includes an interval (no matter how small) then $R(D)$ coincides with $R_S(D)$. This implies that in all those cases that $R_S(D)$ is not attained, the support of the optimum test channel output distribution must be singular, i.e., it cannot contain an interval. It can be, for example, a set of isolated points.

But we also know that whenever $R(D)$ meets the SLB for some $D = D_0$, then it must also coincide with it for all $D < D_0$. This follows from the following simple consideration: If X can be represented as $Y + Z$, where $Z \sim \mathcal{N}(0, D_0)$ is independent of Y, then for every

$D < D_0$, we can always decompose Z as $Z_1 + Z_2$, where Z_1 and Z_2 are both zero–mean independent Gaussian random variables with variances $D_0 - D$ and D, respectively. Thus,

$$X = Y + Z = (Y + Z_1) + Z_2 \stackrel{\Delta}{=} Y' + Z_2 \tag{369}$$

and we have represented X as a noisy version of Y' with noise variance D. Whenever X can be thought of as a mixture of Gaussians, $R(D)$ agrees with its SLB for all D up to the variance of the narrowest Gaussian in this mixture. Thus, in these cases:

$$R(D) \quad \begin{cases} = R_S(D) & D \le D_0 \\ > R_S(D) & D > D_0 \end{cases} \tag{370}$$

It follows then that in all these cases, the optimum output marginal contains intervals for all $D \le D_0$ and then becomes abruptly singular as D exceeds D_0.

From the viewpoint of statistical mechanics, this has the flavor of a phase transition. Consider first an infinite temperature, i.e., $\beta = 0$, which means unlimited distortion. In this case, the optimum output marginal puts all its mass on one point: $y = \boldsymbol{E}(X)$, so it is definitely singular. This remains true even if we increase β to be the inverse temperature that corresponds to $D_{\max}$, the smallest distortion for which $R(D) = 0$. If we further increase β, the support of Y begins to change. In the next step it can include two points, then three points, etc. Then, if there is D_0 below which the SLB is met, then the support of Y abruptly becomes one that contains one interval at least. This point is also demonstrated numerically in [101].

5.5 The One–Dimensional Ising Model

As promised, we now return to models of interacting spins. The most familiar one is the one–dimensional Ising model, according to which

$$\mathcal{E}(\boldsymbol{s}) = -B \sum_{i=1}^{N} s_i - J \sum_{i=1}^{N} s_i s_{i+1} \tag{371}$$

with the periodic boundary condition $s_{N+1} = s_1$. Thus,

$$Z(\beta, B) = \sum_{\boldsymbol{s}} \exp \left\{ \beta B \sum_{i=1}^{N} s_i + \beta J \sum_{i=1}^{N} s_i s_{i+1} \right\}$$

$$= \sum_{\boldsymbol{s}} \exp\left\{ h \sum_{i=1}^{N} s_i + K \sum_{i=1}^{N} s_i s_{i+1} \right\} \qquad h \stackrel{\Delta}{=} \beta B, \quad K \stackrel{\Delta}{=} \beta J$$

$$= \sum_{\boldsymbol{s}} \exp\left\{ \frac{h}{2} \sum_{i=1}^{N} (s_i + s_{i+1}) + K \sum_{i=1}^{N} s_i s_{i+1} \right\}. \tag{372}$$

Consider now the 2×2 matrix P whose entries are $\exp\{\frac{h}{2}(s+s') + Kss'\}$, $s, s \in \{-1, +1\}$, i.e.,

$$P = \begin{pmatrix} e^{K+h} & e^{-K} \\ e^{-K} & e^{K-h} \end{pmatrix}. \tag{373}$$

Also, $s_i = +1$ will be represented by the column vector $\sigma_i = (1,0)^T$ and $s_i = -1$ will be represented by $\sigma_i = (0,1)^T$. Thus,

$$\begin{aligned} Z(\beta, B) &= \sum_{\sigma_1} \cdots \sum_{\sigma_N} (\sigma_1^T P \sigma_2) \cdot (\sigma_2^T P \sigma_3) \cdots (\sigma_N^T P \sigma_1) \\ &= \sum_{\sigma_1} \sigma_1^T P \left(\sum_{\sigma_2} \sigma_2 \sigma_2^T \right) P \left(\sum_{\sigma_3} \sigma_3 \sigma_3^T \right) P \cdots P \left(\sum_{\sigma_N} \sigma_n \sigma_N^T \right) P \sigma_1 \\ &= \sum_{\sigma_1} \sigma_1^T P \cdot I \cdot P \cdot I \cdots I \cdot P \sigma_1 \\ &= \sum_{\sigma_1} \sigma_1^T P^N \sigma_1 \\ &= \mathrm{tr}\{P^N\} \\ &= \lambda_1^N + \lambda_2^N \end{aligned} \tag{374}$$

where λ_1 and λ_2 are the eigenvalues of P, which are

$$\lambda_{1,2} = e^K \cosh(h) \pm \sqrt{e^{-2K} + e^{2K} \sinh^2(h)}. \tag{375}$$

Letting λ_1 denote the larger (the dominant) eigenvalue, i.e.,

$$\lambda_1 = e^K \cosh(h) + \sqrt{e^{-2K} + e^{2K} \sinh^2(h)}, \tag{376}$$

then clearly,

$$\phi(h, K) = \lim_{N \to \infty} \frac{\ln Z}{N} = \ln \lambda_1. \tag{377}$$

The average magnetization is

$$M(h, K) = \left\langle \sum_{i=1}^{N} S_i \right\rangle$$

$$= \frac{\sum_{\boldsymbol{s}} (\sum_{i=1}^{N} s_i) \exp\{h \sum_{i=1}^{N} s_i + K \sum_{i=1}^{N} s_i s_{i+1}\}}{\sum_{\boldsymbol{s}} \exp\{h \sum_{i=1}^{N} s_i + K \sum_{i=1}^{N} s_i s_{i+1}\}}$$
$$= \frac{\partial \ln Z(h,K)}{\partial h} \tag{378}$$

and so, the per–spin magnetization is:

$$m(h,K) \stackrel{\Delta}{=} \lim_{N\to\infty} \frac{M(h,K)}{N} = \frac{\partial \phi(h,K)}{\partial h} = \frac{\sinh(h)}{\sqrt{e^{-4K} + \sinh^2(h)}} \tag{379}$$

or, returning to the original parametrization:

$$m(\beta,B) = \frac{\sinh(\beta B)}{\sqrt{e^{-4\beta J} + \sinh^2(\beta B)}}. \tag{380}$$

For $\beta > 0$ and $B > 0$ this is a smooth function, and so, there is are no phase transitions and no spontaneous magnetization at any finite temperature.[19] However, at the absolute zero ($\beta \to \infty$), we get

$$\lim_{B\downarrow 0} \lim_{\beta\to\infty} m(\beta,B) = +1; \quad \lim_{B\uparrow 0} \lim_{\beta\to\infty} m(\beta,B) = -1, \tag{381}$$

thus m is discontinuous w.r.t. B at $\beta \to \infty$, which means that there is a phase transition at $T = 0$. In other words, the Curie temperature is $T_c = 0$.

We see then that one–dimensional Ising model is easy to handle, but it is not very interesting in the sense that there is actually no phase transition. The extension to the two–dimensional Ising model on the square lattice is surprisingly more difficult, but it is still solvable, albeit without a magnetic field. It was first solved in 1944 by Onsager [89], who has shown that it exhibits a phase transition with Curie temperature given by

$$T_c = \frac{2J}{k \ln(\sqrt{2}+1)}. \tag{382}$$

For lattice dimension larger than two, the problem is still open.

It turns out then that whatever counts for the existence of phase transitions, is not only the intensity of the interactions (designated by the magnitude of J), but more importantly,

[19] Note, in particular, that for $J = 0$ (i.i.d. spins) we get paramagnetic characteristics $m(\beta,B) = \tanh(\beta B)$, in agreement with the result pointed out in the example of two–level systems, in one of our earlier discussions.

the "dimensionality" of the structure of the pairwise interactions. If we denote by n_ℓ the number of ℓ–th order neighbors of every given site, namely, the number of sites that can be reached within ℓ steps from the given site, then whatever counts is how fast does the sequence $\{n_\ell\}$ grow, or more precisely, what is the value of $d \stackrel{\Delta}{=} \lim_{\ell\to\infty} \frac{1}{\ell} \ln n_\ell$, which is exactly the ordinary dimensionality for hyper-cubic lattices. Loosely speaking, this dimension must be sufficiently large for a phase transition to exist.

To demonstrate this point, we next discuss an extreme case of a model where this dimensionality is actually infinite. In this model "everybody is a neighbor of everybody else" and to the same extent, so it definitely has the highest connectivity possible. This is not quite a physically realistic model, but the nice thing about it is that it is easy to solve and that it exhibits a phase transition that is fairly similar to those that exist in real systems. It is also intimately related to a very popular approximation method in statistical mechanics, called the *mean field approximation*. Hence it is sometimes called the *mean field model*. It is also known as the *Curie–Weiss model* or the *infinite range model*.

Finally, we should comment that there are other "infinite–dimensional" Ising models, like the one defined on the Bethe lattice (an infinite tree without a root and without leaves), which is also easily solvable (by recursion) and it also exhibits phase transitions [4], but we will not discuss it here.

5.6 The Curie–Weiss Model

According to the Curie–Weiss (C–W) model,

$$\mathcal{E}(\boldsymbol{s}) = -B\sum_{i=1}^{N} s_i - \frac{J}{2N}\sum_{i\neq j} s_i s_j. \tag{383}$$

Here, all pairs $\{(s_i, s_j)\}$ communicate to the same extent, and without any geometry. The $1/N$ factor here is responsible for keeping the energy of the system extensive (linear in N), as the number of interaction terms is quadratic in N. The factor $1/2$ compensates for the fact that the summation over $i \neq j$ counts each pair twice. The first observation is the trivial

fact that

$$\left(\sum_i s_i\right)^2 = \sum_i s_i^2 + \sum_{i\neq j} s_i s_j = N + \sum_{i\neq j} s_i s_j \tag{384}$$

where the second equality holds since $s_i^2 \equiv 1$. It follows then, that our Hamiltonian is, up to a(n immaterial) constant, equivalent to

$$\begin{aligned}\mathcal{E}(\boldsymbol{s}) &= -B\sum_{i=1}^N s_i - \frac{J}{2N}\left(\sum_{i=1}^N s_i\right)^2 \\ &= -N\left[B\cdot\left(\frac{1}{N}\sum_{i=1}^N s_i\right) + \frac{J}{2}\left(\frac{1}{N}\sum_{i=1}^N s_i\right)^2\right],\end{aligned} \tag{385}$$

thus $\mathcal{E}(\boldsymbol{s})$ depends on $\boldsymbol{s}$ only via the magnetization $m(\boldsymbol{s}) = \frac{1}{N}\sum_i s_i$. This fact makes the C–W model very easy to handle similarly as in the method of types:

$$\begin{aligned}Z_N(\beta, B) &= \sum_{\boldsymbol{s}} \exp\left\{N\beta\left[B\cdot m(\boldsymbol{s}) + \frac{J}{2}m^2(\boldsymbol{s})\right]\right\} \\ &= \sum_{m=-1}^{+1} \Omega(m)\cdot e^{N\beta(Bm+Jm^2/2)} \\ &\doteq \sum_{m=-1}^{+1} e^{Nh_2((1+m)/2)}\cdot e^{N\beta(Bm+Jm^2/2)} \\ &\doteq \exp\left\{N\cdot\max_{|m|\leq 1}\left[h_2\left(\frac{1+m}{2}\right) + \beta Bm + \frac{\beta m^2 J}{2}\right]\right\}\end{aligned} \tag{386}$$

and so,

$$\phi(\beta, B) = \max_{|m|\leq 1}\left[h_2\left(\frac{1+m}{2}\right) + \beta Bm + \frac{\beta m^2 J}{2}\right]. \tag{387}$$

The maximum is found by equating the derivative to zero, i.e.,

$$0 = \frac{1}{2}\ln\left(\frac{1-m}{1+m}\right) + \beta B + \beta Jm \equiv -\tanh^{-1}(m) + \beta B + \beta Jm \tag{388}$$

or equivalently, the maximizing (and hence the dominant) m is a solution m^* to the equation[20]

$$m = \tanh(\beta B + \beta Jm).$$

[20] Once again, for $J = 0$, we are back to non–interacting spins and then this equation gives the paramagnetic behavior $m = \tanh(\beta B)$.

Consider first the case $B = 0$, where the equation boils down to

$$m = \tanh(\beta J m). \tag{389}$$

It is instructive to look at this equation graphically. Referring to Fig. 12, we have to make a distinction between two cases: If $\beta J < 1$, namely, $T > T_c \stackrel{\Delta}{=} J/k$, the slope of the function $y = \tanh(\beta J m)$ at the origin, βJ, is smaller than the slope of the linear function $y = m$, which is 1, thus these two graphs intersect only at the origin. It is easy to check that in this case, the second derivative of

$$\psi(m) \stackrel{\Delta}{=} h_2\left(\frac{1+m}{2}\right) + \frac{\beta J m^2}{2} \tag{390}$$

at $m = 0$ is negative, and therefore it is indeed the maximum (see Fig. 13, left part). Thus, the dominant magnetization is $m^* = 0$, which means disorder and hence no spontaneous magnetization for $T > T_c$. On the other hand, when $\beta J > 1$, which means temperatures

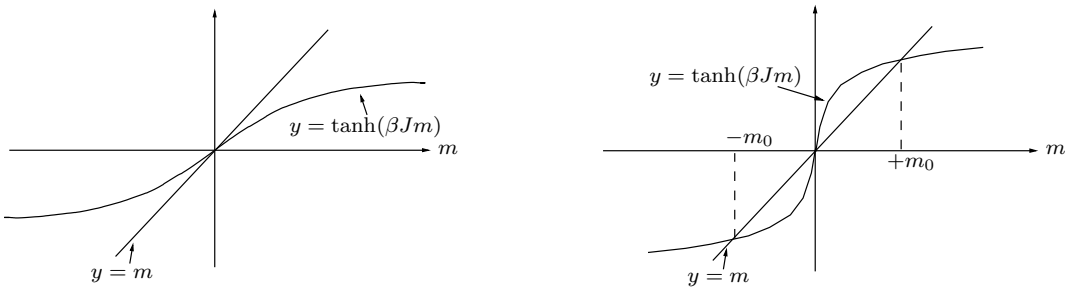


Figure 12: Graphical solutions of equation $m = \tanh(\beta J m)$: The left part corresponds to the case $\beta J < 1$, where there is one solution only, $m^* = 0$. The right part corresponds to the case $\beta J > 1$, where in addition to the zero solution, there are two non–zero solutions $m^* = \pm m_0$.

lower than T_c, the initial slope of the tanh function is larger than that of the linear function, but since the tanh function cannot take values outside the interval $(-1, +1)$, the two functions must intersect also at two additional, symmetric, non–zero points, which we denote by $+m_0$ and $-m_0$ (see Fig. 12, right part). In this case, it can readily be shown that the second derivative of $\psi(m)$ is positive at the origin (i.e., there is a local minimum at $m = 0$) and negative at $m = \pm m_0$, which means that there are maxima at these two points (see Fig. 13, right part). Thus, the dominant magnetizations are $\pm m_0$, each capturing about half of the probability.

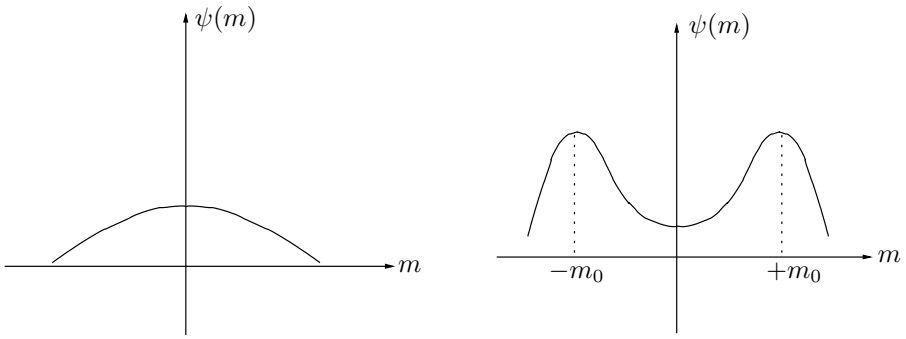


Figure 13: The function $\psi(m) = h_2((1+m)/2) + \beta J m^2/2$ has a unique maximum at $m = 0$ when $\beta J < 1$ (left graph) and two local maxima at $\pm m_0$, in addition to a local minimum at $m = 0$, when $\beta J > 1$ (right graph).

Consider now the case $\beta J > 1$, where the magnetic field B is brought back into the picture. This will break the symmetry of the right graph of Fig. 13 and the corresponding graphs of $\psi(m)$ would be as in Fig. 14, where now the higher local maximum (which is also the global one) is at $m_0(B)$ whose sign is as that of B. But as $B \to 0$, $m_0(B) \to m_0$ of Fig. 13. Thus, we see the spontaneous magnetization here. Even after removing the magnetic

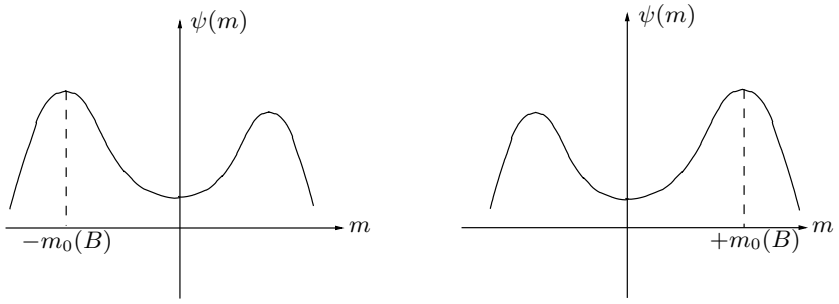


Figure 14: The case $\beta J > 1$ with a magnetic field B. The left graph corresponds to $B < 0$ and the right graph – to $B > 0$.

field, the system remains magnetized to the level of m_0, depending on the direction (the sign) of B before its removal. Obviously, the magnetization $m(\beta, B)$ has a discontinuity at $B = 0$ for $T < T_c$, which is a first order phase transition w.r.t. B (see Fig. 15). We note that the point $T = T_c$ is the boundary between the region of existence and the region of non–existence of a phase transition w.r.t. B. Such a point is called a *critical point*. The phase transition w.r.t. β is of the second order.

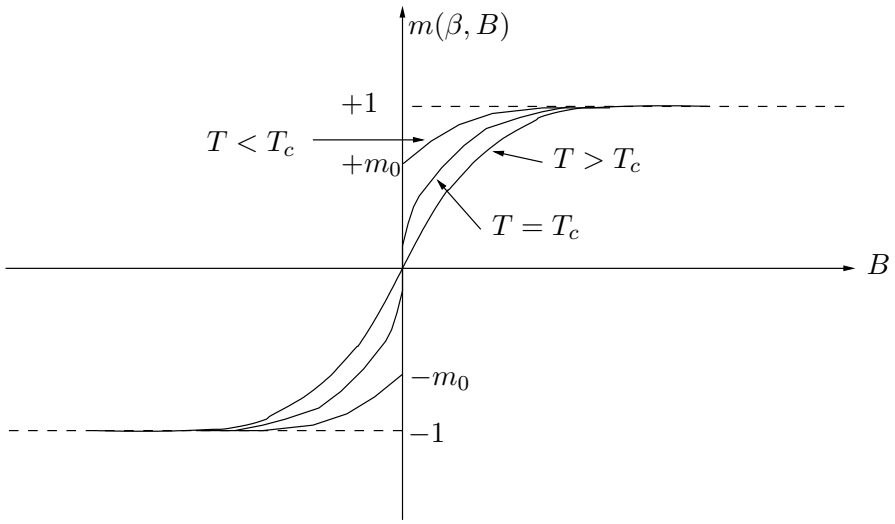


Figure 15: Magnetization vs. magnetic field: For $T < T_c$ there is spontaneous magnetization: $\lim_{B\downarrow 0} m(\beta, B) = +m_0$ and $\lim_{B\uparrow 0} m(\beta, B) = -m_0$, and so there is a discontinuity at $B = 0$.

Finally, we should mention here an alternative technique that can be used to analyze this model, which is based on the Hubbard–Stratonovich transform and the saddle point method. Specifically, we have the following chain of equalities:

$$\begin{aligned}
Z(h,K) &= \sum_{\boldsymbol{s}} \exp\left\{h\sum_{i=1}^N s_i + \frac{K}{2N}\left(\sum_{i=1}^N s_i\right)^2\right\} \qquad h \stackrel{\Delta}{=} \beta B,\ K \stackrel{\Delta}{=} \beta J \\
&= \sum_{\boldsymbol{s}} \exp\left\{h\sum_{i=1}^N s_i\right\} \cdot \exp\left\{\frac{K}{2N}\left(\sum_{i=1}^N s_i\right)^2\right\} \\
&= \sum_{\boldsymbol{s}} \exp\left\{h\sum_{i=1}^N s_i\right\} \cdot \sqrt{\frac{N}{2\pi K}} \int_{\mathbb{R}} \mathrm{d}z \exp\left\{-\frac{Nz^2}{2K} + z\cdot\sum_{i=1}^N s_i\right\} \\
&= \sqrt{\frac{N}{2\pi K}} \int_{\mathbb{R}} \mathrm{d}z e^{-Nz^2/(2K)} \sum_{\boldsymbol{s}} \exp\left\{(h+z)\sum_{i=1}^N s_i\right\} \\
&= \sqrt{\frac{N}{2\pi K}} \int_{\mathbb{R}} \mathrm{d}z e^{-Nz^2/(2K)} \left[\sum_{s=-1}^{1} e^{(h+z)s}\right]^N \\
&= \sqrt{\frac{N}{2\pi K}} \int_{\mathbb{R}} \mathrm{d}z e^{-Nz^2/(2K)} [2\cosh(h+z)]^N \\
&= 2^N \cdot \sqrt{\frac{N}{2\pi K}} \int_{\mathbb{R}} \mathrm{d}z \exp\{N[\ln\cosh(h+z) - z^2/(2K)]\} \qquad (391)
\end{aligned}$$

Using the the saddle point method (or the Laplace method), this integral is dominated by the maximum of the function in the square brackets at the exponent of the integrand, or equivalently, the minimum of the function

$$\gamma(z) = \frac{z^2}{2K} - \ln\cosh(h+z). \tag{392}$$

by equating its derivative to zero, we get the very same equation as $m = \tanh(\beta B + \beta J m)$ by setting $z = \beta J m$. The function $\gamma(z)$ is different from the function ψ that we maximized earlier, but the extremum is the same. This function is called the *Landau free energy*.

5.7 Spin Glasses and Random Code Ensembles

So far we discussed only models where the non–zero coupling coefficients, $\boldsymbol{J} = \{J_{ij}\}$ are equal, thus they are either all positive (ferromagnetic models) or all negative (antiferromagnetic models). As mentioned earlier, there are also models where the signs of these coefficients are mixed, which are called *spin glass* models.

Spin glass models have a much more complicated and more interesting behavior than ferromagnets, because there might be metastable states due to the fact that not necessarily all spin pairs $\{(s_i, s_j)\}$ can be in their preferred mutual polarization. It might be the case that some of these pairs are "frustrated." In order to model situations of amorphism and disorder in such systems, it is customary to model the coupling coefficients as random variables.

Some models allow, in addition to the random coupling coefficients, also random local fields, i.e., the term $-B\sum_i s_i$ in the Hamiltonian, is replaced by $-\sum_i B_i s_i$, where $\{B_i\}$ are random variables, similarly as in the representation of $P(\boldsymbol{s}|\boldsymbol{y})$ pertaining to a coded communication system, as discussed earlier, where $\{y_i\}$ play the role of local magnetic fields. The difference, however, is that here the $\{B_i\}$ are normally assumed i.i.d., whereas in the communication system model $P(\boldsymbol{y})$ exhibits memory (even if the channel is memoryless) due to memory in $P(\boldsymbol{s})$. Another difference is that in the physics model, the distribution of $\{B_i\}$ is assumed to be independent of temperature, whereas in coding, if we introduce a temperature parameter by exponentiating (i.e., $P_\beta(\boldsymbol{s}|\boldsymbol{y}) \propto P^\beta(s)P^\beta(\boldsymbol{y}|\boldsymbol{s})$), the induced

marginal of $\boldsymbol{y}$ will depend on β.

In the following discussion, let us refer to the case where only the coupling coefficients $\boldsymbol{J}$ are random variables (similar things can be said in the more general case, discussed in the last paragraph). This model with random parameters means that there are now two levels of randomness:

- Randomness of the coupling coefficients $\boldsymbol{J}$.
- Randomness of the spin configuration $\boldsymbol{s}$ given $\boldsymbol{J}$, according to the Boltzmann distribution, i.e.,

$$P(\boldsymbol{s}|\boldsymbol{J}) = \frac{\exp\left\{\beta\left[B\sum_{i=1}^{N} s_i + \sum_{(i,j)} J_{ij}s_is_j\right]\right\}}{Z(\beta, B|\boldsymbol{J})}. \tag{393}$$

However, these two sets of random variables have a rather different stature. The underlying setting is normally such that $\boldsymbol{J}$ is considered to be randomly drawn once and for all, and then remain fixed, whereas $\boldsymbol{s}$ keeps varying all the time (according to the dynamics of the system). At any rate, the time scale along which $\boldsymbol{s}$ varies is much smaller than that of $\boldsymbol{J}$. Another difference is that $\boldsymbol{J}$ is normally not assumed to depend on temperature, whereas $\boldsymbol{s}$, of course, does. In the terminology of physicists, $\boldsymbol{s}$ is considered an *annealed* random variable, whereas $\boldsymbol{J}$ is considered a *quenched* random variable. Accordingly, there is a corresponding distinction between *annealed averages* and *quenched averages*.

Actually, there is (or, more precisely, should be) a parallel distinction when we consider ensembles of randomly chosen codes in Information Theory. When we talk about random coding, we normally think of the randomly chosen code as being drawn once and for all, we do not reselect it after each transmission (unless there are security reasons to do so), and so, a random code should be thought of us a quenched entity, whereas the source(s) and channel(s) are more naturally thought of as annealed entities. Nonetheless, this is not what we usually do in Information Theory. We normally take double expectations of some performance measure w.r.t. both source/channel and the randomness of the code, on the

same footing.[21] We will elaborate on this point later on.

Returning to spin glass models, let is see what is exactly the difference between the quenched averaging and the annealed one. If we examine, for instance, the free energy, or the log–partition function, $\ln Z(\beta|\boldsymbol{J})$, this is now a random variable, of course, because it depends on the random $\boldsymbol{J}$. If we denote by $\langle\cdot\rangle_{\boldsymbol{J}}$ the expectation w.r.t. the randomness of $\boldsymbol{J}$, then quenched averaging means $\langle\ln Z(\beta|\boldsymbol{J})\rangle_{\boldsymbol{J}}$ (with the motivation of the self–averaging property of the random variable $\ln Z(\beta|\boldsymbol{J})$ in many cases), whereas annealed averaging means $\ln\langle Z(\beta|\boldsymbol{J})\rangle_{\boldsymbol{J}}$. Normally, the relevant average is the quenched one, but it is typically also much harder to calculate (and it is customary to apply the replica method then). Clearly, the annealed average is never smaller than the quenched one because of Jensen's inequality, but they sometimes coincide at high temperatures. The difference between them is that in quenched averaging, the dominant realizations of $\boldsymbol{J}$ are the typical ones, whereas in annealed averaging, this is not necessarily the case. This follows from the following sketchy consideration. As for the annealed average, we have:

$$\begin{aligned}\langle Z(\beta|\boldsymbol{J})\rangle &= \sum_{\boldsymbol{J}} P(\boldsymbol{J})Z(\beta|\boldsymbol{J}) \\ &\approx \sum_{\alpha} \Pr\{\boldsymbol{J}:\ Z(\beta|\boldsymbol{J}) \doteq e^{N\alpha}\}\cdot e^{N\alpha} \\ &\approx \sum_{\alpha} e^{-NE(\alpha)}\cdot e^{N\alpha} \qquad \text{(assuming exponential probabilities)} \\ &\doteq e^{N\max_{\alpha}[\alpha - E(\alpha)]}\end{aligned} \tag{394}$$

which means that the annealed average is dominated by realizations of the system with

$$\frac{\ln Z(\beta|\boldsymbol{J})}{N} \approx \alpha^* \stackrel{\Delta}{=} \arg\max_{\alpha}[\alpha - E(\alpha)], \tag{395}$$

which may differ from the typical value of α, which is

$$\alpha = \phi(\beta) \equiv \lim_{N\to\infty}\frac{1}{N}\left\langle \ln Z(\beta|\boldsymbol{J})\right\rangle. \tag{396}$$

On the other hand, when it comes to quenched averaging, the random variable $\ln Z(\beta|\boldsymbol{J})$

[21] There are few exceptions to this rule, like the work of Barg and Forney [3].

behaves linearly in N, and concentrates strongly around the typical value $N\phi(\beta)$, whereas other values are weighted by (exponentially) decaying probabilities.

In the coded communication setting, there is a strong parallelism. Here, there is a distinction between the exponent of the average error probability, $\ln \boldsymbol{E} P_e(\mathcal{C})$ (annealed) and the average exponent of the error probability $\boldsymbol{E} \ln P_e(\mathcal{C})$ (quenched), where $P_e(\mathcal{C})$ is the error probability of a randomly selected code $\mathcal{C}$. Very similar things can be said here too.

The literature on spin glasses includes many models for the randomness of the coupling coefficients. We end this part by listing just a few.

- The *Edwards–Anderson* (E–A) model, where $\{J_{ij}\}$ are non–zero for nearest–neighbor pairs only (e.g., $j = i \pm 1$ in one–dimensional model). According to this model, these J_{ij}'s are i.i.d. random variables, which are normally modeled to have a zero–mean Gaussian pdf, or binary symmetric with levels $\pm J_0$. It is customary to work with a zero–mean distribution if we have a pure spin glass in mind. If the mean is nonzero, the model has either a ferromagnetic or an anti-ferromagnetic bias, according to the sign of the mean.

- The *Sherrington–Kirkpatrick* (S–K) model, which is similar to the E–A model, except that the support of $\{J_{ij}\}$ is extended to include all $N(N-1)/2$ pairs, and not only nearest–neighbor pairs. This can be thought of as a stochastic version of the C–W model in the sense that here too, there is no geometry, and every spin 'talks' to every other spin to the same extent, but here the coefficients are random, as said.

- The *p–spin* model, which is similar to the S–K model, but now the interaction term consists, not only of pairs, but also triples, quadruples, and so on, up to cliques of size p, i.e., products $s_{i_1} s_{i_2} \cdots s_{i_p}$, where $(i_1, \ldots, i_p)$ exhaust all possible subsets of p spins out of N. Each such term has a Gaussian coefficient $J_{i_1,\ldots,i_p}$ with an appropriate variance.

Considering the p–spin model, it turns out that if we look at the extreme case of $p \to \infty$ (taken after the thermodynamic limit $N \to \infty$), the resulting behavior turns out to be ex-

tremely erratic: all energy levels $\{\mathcal{E}(\boldsymbol{s})\}_{\boldsymbol{s}\in\{-1,+1\}^N}$ become i.i.d. Gaussian random variables. This is, of course, a toy model, which has very little to do with reality (if any), but it is surprisingly interesting and easy to work with. It is called the *random energy model* (REM). We have already mentioned it as an example on which we demonstrated the replica method. We are next going to discuss it extensively because it turns out to be very relevant for random coding models.

6 The Random Energy Model and Random Coding

In this chapter, we first focus on the REM and its properties, and then relate it to random code ensembles. The first two sections are inspired by the exposition in [80, Chapters 5 and 6], but with a slightly different flavor and somewhat more detail.

6.1 REM Without a Magnetic Field

The REM was proposed by Derrida in the early eighties of the previous century in a series of papers [22], [23], [24]. As mentioned at the end of the previous chapter, the REM is inspired by the limit $p \to \infty$ in the p–spin model. More specifically, Derrida showed that the correlations between the random energies of two configurations, $\boldsymbol{s}$ and $\boldsymbol{s}'$, in the p–spin model are given by

$$\left(\frac{1}{N}\sum_{i=1}^{N} s_i s_i'\right)^p, \tag{397}$$

and since $|\frac{1}{N}\sum_{i=1}^{N} s_i s_i'| < 1$, these correlations vanish as $p \to \infty$. This has motivated Derrida to propose a model according to which the configurational energies $\{\mathcal{E}(\boldsymbol{s})\}$, in the absence of a magnetic field, are simply i.i.d. zero–mean Gaussian random variables with a variance that grows linearly with N (again, for reasons of extensivity). More concretely, this variance is taken to be $NJ^2/2$, where J is a constant parameter. This means that we actually forget that the spin array has any structure of the kind that we have seen before, and we simply randomly draw an independent Gaussian random variable $\mathcal{E}(\boldsymbol{s}) \sim \mathcal{N}(0, NJ^2/2)$ (other distributions are also possible) for every configuration $\boldsymbol{s}$. Thus, the partition function $Z(\beta) = \sum_{\boldsymbol{s}} e^{-\beta\mathcal{E}(\boldsymbol{s})}$ is a random variable as well, of course.

This is a toy model that does not describe faithfully any realistic physical system, but we will devote to it some considerable time, for several reasons:

- It is simple and easy to analyze.
- In spite of its simplicity, it is rich enough to exhibit phase transitions, and therefore it is interesting.

- Most importantly, it will prove very relevant to the analogy with coded communication systems with randomly selected codes.

As we shall see quite shortly, there is an intimate relationship between phase transitions of the REM and phase transitions in the behavior of coded communication systems, most notably, transitions between reliable and unreliable communication, but others as well.

What is the basic idea that stands behind the analysis of the REM? As said,

$$Z(\beta) = \sum_{\boldsymbol{s}} e^{-\beta \mathcal{E}(\boldsymbol{s})} \tag{398}$$

where $\mathcal{E}(\boldsymbol{s}) \sim \mathcal{N}(0, NJ^2/2)$ are i.i.d. Consider the density of states $\Omega(E)$, which is now a random variable: $\Omega(E)\mathrm{d}E$ is the number of configurations $\{\boldsymbol{s}\}$ whose randomly selected energy $\mathcal{E}(\boldsymbol{s})$ happens to fall between E and $E + \mathrm{d}E$, and of course,

$$Z(\beta) = \int_{-\infty}^{+\infty} \mathrm{d}E \Omega(E) e^{-\beta E}. \tag{399}$$

How does the random variable $\Omega(E)\mathrm{d}E$ behave like? First, observe that, ignoring non–exponential factors:

$$\Pr\{E \leq \mathcal{E}(\boldsymbol{s}) \leq E + \mathrm{d}E\} \approx f(E)\mathrm{d}E \doteq e^{-E^2/(NJ^2)}\mathrm{d}E, \tag{400}$$

and so,

$$\langle \Omega(E)\mathrm{d}E \rangle \doteq 2^N \cdot e^{-E^2/(NJ^2)} = \exp\left\{ N \left[\ln 2 - \left(\frac{E}{NJ} \right)^2 \right] \right\}. \tag{401}$$

We have reached the pivotal point behind the analysis of the REM, which is based on a fundamental principle that goes far beyond the analysis of the first moment of $\Omega(E)\mathrm{d}E$. In fact, this principle is frequently used in random coding arguments in information theory: Suppose that we have e^{NA} ($A > 0$, independent of N) independent events $\{\mathcal{E}_i\}$, each one with probability $\Pr\{\mathcal{E}_i\} = e^{-NB}$ ($B > 0$, independent of N). What is the probability that at least one of the $\mathcal{E}_i$'s would occur? Intuitively, we expect that in order to see at least one or a few successes, the number of experiments should be at least about $1/\Pr\{\mathcal{E}_i\} = e^{NB}$. If

$A > B$ then this is the case. On the other hand, for $A < B$, the number of trials is probably insufficient for seeing even one success. Indeed, a more rigorous argument gives:

$$\begin{aligned}
\Pr\left\{\bigcup_{i=1}^{e^{NA}} \mathcal{E}_i\right\} &= 1 - \Pr\left\{\bigcap_{i=1}^{e^{NA}} \mathcal{E}_i^c\right\} \\
&= 1 - \left(1 - e^{-NB}\right)^{e^{NA}} \\
&= 1 - \exp\{e^{NA}\ln(1 - e^{-NB})\} \\
&\approx 1 - \exp\{-e^{NA}e^{-NB}\} \\
&= 1 - \exp\{-e^{N(A-B)}\} \\
&\to \begin{cases} 1 & A > B \\ 0 & A < B \end{cases}
\end{aligned} \tag{402}$$

Now, to another question: For $A > B$, how many of the $\mathcal{E}_i$'s would occur in a typical realization of this set of experiments? The number Ω_N of 'successes' is given by $\sum_{i=1}^{e^{NA}} \mathcal{I}\{\mathcal{E}_i\}$, namely, it is the sum of e^{NA} i.i.d. binary random variables whose expectation is $\boldsymbol{E}\{\Omega_N\} = e^{N(A-B)}$. Therefore, its probability distribution concentrates very rapidly around its mean. In fact, the events $\{\Omega_N \geq e^{N(A-B+\epsilon)}\}$ ($\epsilon > 0$, independent of N) and $\{\Omega_N \leq e^{N(A-B-\epsilon)}\}$ are large deviations events whose probabilities decay exponentially in the number of experiments, e^{NA}, and hence *double–exponentially* in N.[22] Thus, for $A > B$, the number of successes is "almost deterministically" about $e^{N(A-B)}$.

Now, back to the REM: For E whose absolute value is less than

$$E_0 \triangleq NJ\sqrt{\ln 2} \tag{403}$$

the exponential increase rate, $A = \ln 2$, of the number $2^N = e^{N\ln 2}$ of configurations (which is the number of independent trials in randomly drawing energies $\{\mathcal{E}(\boldsymbol{s})\}$) is faster than the exponential decay rate of the probability, $e^{-N[E/(nJ)]^2)} = e^{-N(\epsilon/J)^2}$ (i.e., $B = (\epsilon/J)^2$) that $\mathcal{E}(\boldsymbol{s})$ would happen to fall around E. In other words, the number of these trials is way larger than the reciprocal of this probability and in view of the earlier discussion, the probability

[22] This will be shown rigorously later on.

that

$$\Omega(E)\mathrm{d}E = \sum_{\boldsymbol{s}} \mathcal{I}\{E \leq \mathcal{E}(\boldsymbol{s}) \leq E + \mathrm{d}E\}. \tag{404}$$

would deviate from its mean, which is exponentially $\exp\{N[\ln 2 - (E/(NJ))^2]\}$, by a multiplicative factor that falls out of the interval $[e^{-N\epsilon}, e^{+N\epsilon}]$, decays double–exponentially with N. In other words, we argue that for $-E_0 < E < +E_0$, the event

$$e^{-N\epsilon} \cdot \exp\left\{N\left[\ln 2 - \left(\frac{E}{NJ}\right)^2\right]\right\} \leq \Omega(E)\mathrm{d}E \leq e^{+N\epsilon} \cdot \exp\left\{N\left[\ln 2 - \left(\frac{E}{NJ}\right)^2\right]\right\}$$

happens with probability that tends to unity in a double–exponential rate. As discussed, $-E_0 < E < +E_0$ is exactly the condition for the expression in the square brackets at the exponent $[\ln 2 - (\frac{E}{NJ})^2]$ to be positive, thus $\Omega(E)\mathrm{d}E$ is exponentially large. On the other hand, if $|E| > E_0$, the number of trials 2^N is way smaller than the reciprocal of the probability of falling around E, and so, most of the chances are that we will see no configurations at all with energy about E. In other words, for these large values of $|E|$, $\Omega(E) = 0$ for typical realizations of the REM (see Fig. 16, left part). It follows then that for such a typical realization,

$$\begin{aligned}
Z(\beta) &\approx \int_{-E_0}^{+E_0} \langle \mathrm{d}E \cdot \Omega(E) \rangle \, e^{-\beta E} \\
&\doteq \int_{-E_0}^{+E_0} \mathrm{d}E \cdot \exp\left\{N\left[\ln 2 - \left(\frac{E}{NJ}\right)^2\right]\right\} \cdot e^{-\beta E} \\
&= \int_{-E_0}^{+E_0} \mathrm{d}E \cdot \exp\left\{N\left[\ln 2 - \left(\frac{E}{NJ}\right)^2 - \beta \cdot \left(\frac{E}{N}\right)\right]\right\} \\
&= N \cdot \int_{-\epsilon_0}^{+\epsilon_0} \mathrm{d}\epsilon \cdot \exp\left\{N\left[\ln 2 - \left(\frac{\epsilon}{J}\right)^2 - \beta\epsilon\right]\right\} \\
&\doteq \exp\left\{N \cdot \max_{|\epsilon| \leq \epsilon_0}\left[\ln 2 - \left(\frac{\epsilon}{J}\right)^2 - \beta\epsilon\right]\right\},
\end{aligned} \tag{405}$$

where we have defined $\epsilon = E/N$ and $\epsilon_0 = E_0/N$, and where in the last step we have used Laplace integration. The maximization problem at the exponent is very simple: it is that of

a quadratic function across an interval. The solution is of either one of two types, depending on whether the maximum is attained at a zero–derivative internal point in $(-\epsilon_0, +\epsilon_0)$ or at an edge-point. The choice between the two depends on β. Specifically, we obtain the following:

$$\phi(\beta) = \lim_{N\to\infty} \frac{\ln Z(\beta)}{N} = \begin{cases} \ln 2 + \frac{\beta^2 J^2}{4} & \beta \leq \beta_c \\ \beta J \sqrt{\ln 2} & \beta > \beta_c \end{cases} \tag{406}$$

where $\beta_c = \frac{2}{J}\sqrt{\ln 2}$. This dichotomy between two types of behavior means a phase transition. The function $\phi(\beta)$ changes its behavior abruptly at $\beta = \beta_c$, from being quadratic in β to being linear in β (see also Fig. 16, right part). The function ϕ is continuous (as always), and so is its first derivative, but the second derivative is not. Thus, it is a second order phase transition. Note that in the quadratic range, this expression is precisely the same as we got using the replica method, when we hypothesized that the dominant configuration is fully symmetric and is given by $Q = I_{m\times m}$. Thus, the replica symmetric solution indeed gives the correct result in the high temperature regime, but the low temperature regime seems to require symmetry breaking. What is the significance of each one of these phases? Let us

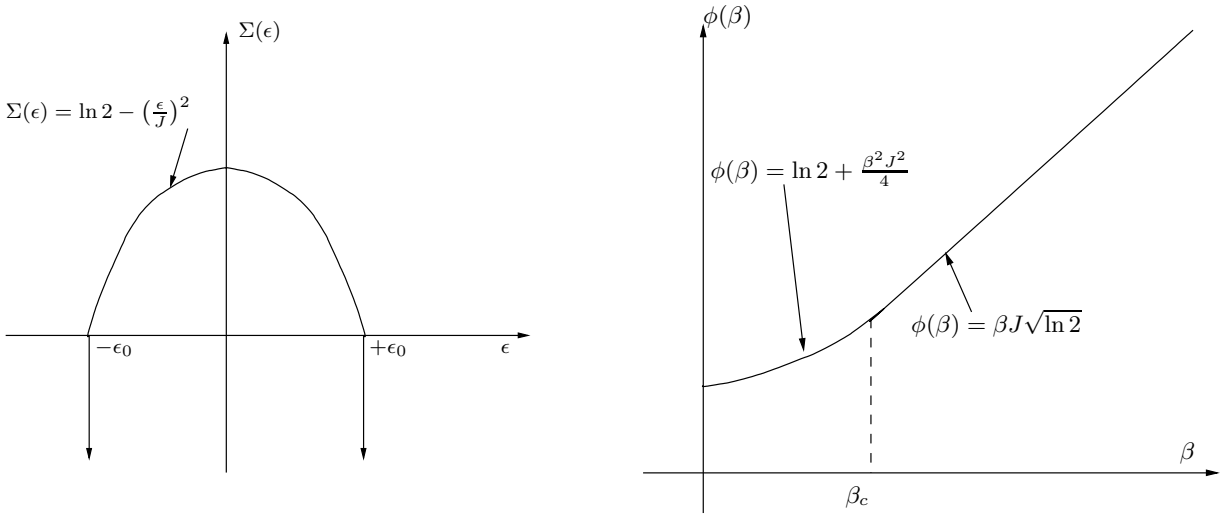


Figure 16: The entropy function and the normalized log–partition function of the REM.

begin with the second line of the above expression of $\phi(\beta)$, which is $\phi(\beta) = \beta J\sqrt{\ln 2} \equiv \beta\epsilon_0$ for $\beta > \beta_c$. What is the meaning of linear dependency of ϕ in β? Recall that the entropy Σ

is given by

$$\Sigma(\beta) = \phi(\eta) + \beta\epsilon = \phi(\beta) - \beta \cdot \phi'(\beta), \tag{407}$$

which in the case where ϕ is linear, simply vanishes. Zero entropy means that the partition function is dominated by a subexponential number of ground–state configurations (with per–particle energy about ϵ_0), just like when it is frozen (see also Fig. 16, left part: $\Sigma(-\epsilon_0) = 0$). For this reason, we refer to this phase as the *frozen phase* or the *glassy phase*.[23] In the high–temperature range, on the other hand, the entropy is strictly positive and the dominant per–particle energy level is $\epsilon^* = -\frac{1}{2}\beta J^2$, which is the point of zero–derivative of the function $[\ln 2 - (\epsilon/J)^2 - \beta\epsilon]$. Here the partition is dominated by exponentially many configurations whose energy is $E^* = n\epsilon^* = -\frac{N}{2}\beta J^2$. As we shall see later on, in this range, the behavior of the system is essentially paramagnetic (like in a system of i.i.d. spins), and so it is called the *paramagnetic phase*.

We therefore observe that the type of phase transition here is different from the one in the Curie–Weiss model. Here, there is is no spontaneous magnetization transition, but rather on a glass transition. In fact, we will not see here a spontaneous magnetization even when a magnetic field is applied (see Sect. 7.1).

From $\phi(\beta)$, one can go ahead and calculate other physical quantities, but we will not do this now. As a final note in this context, it should be emphasized that since the calculation of Z was carried out for the typical realizations of the quenched random variables $\{\mathcal{E}(\boldsymbol{s})\}$, we have actually calculated the quenched average of $\lim_{N\to\infty}(\ln Z)/N$. As for the annealed average, we have

$$\begin{aligned}
\lim_{N\to\infty}\frac{\ln\langle Z(\beta)\rangle}{N} &= \lim_{N\to\infty}\frac{1}{N}\ln\left[\int_{\mathbb{R}}\langle\Omega(E)\mathrm{d}\epsilon\rangle e^{-\beta N\epsilon}\right] \\
&= \lim_{N\to\infty}\frac{1}{N}\ln\left[\int_{\mathbb{R}}\exp\left\{N\left[\ln 2 - \left(\frac{\epsilon}{J}\right)^2 - \beta\epsilon\right]\right\}\right] \\
&= \max_{\epsilon\in\mathbb{R}}\left[\ln 2 - \left(\frac{\epsilon}{J}\right)^2 - \beta\epsilon\right]
\end{aligned}$$

[23] In this phase, the system behaves like a glass: on the one hand, it is frozen (so it consolidates), but on the other hand, it remains disordered and amorphous, like a liquid.

$$= \ln 2 + \frac{\beta^2 J^2}{4}, \tag{408}$$

which is the paramagnetic expression, without any phase transition since the maximization over ϵ is not constrained. This demonstrates the point that the annealed approximation is fine for high temperatures, but probably not for low temperatures, and it may fail to detect phase transitions.

6.2 Random Code Ensembles and the REM

Let us now see how does the REM relate to random code ensembles. This relationship was first observed in [82], but it is presented more comprehensively in [80]. The discussion in this section is based on [80], as well as on [69], but with more emphasis on the error exponent analysis technique that stems from this relation to the REM, which will be outlined in Section 6.3. Other relevant references, in this context, are [3] and [33].

Consider a discrete memoryless channel (DMC), defined by the conditional probability distribution

$$P(\boldsymbol{y}|\boldsymbol{x}) = \prod_{i=1}^{n} p(y_i|x_i), \tag{409}$$

where the input n–vector $\boldsymbol{x} = (x_1, \ldots, x_n)$ belongs to a codebook $\mathcal{C} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_M\}$, $M = e^{nR}$, with uniform priors, and where R is the coding rate in nats per channel use. The induced posterior, for $\boldsymbol{x} \in \mathcal{C}$, is then:

$$\begin{aligned} P(\boldsymbol{x}|\boldsymbol{y}) &= \frac{P(\boldsymbol{y}|\boldsymbol{x})}{\sum_{\boldsymbol{x}' \in \mathcal{C}} P(\boldsymbol{y}|\boldsymbol{x}')} \\ &= \frac{e^{-\ln[1/P(\boldsymbol{y}|\boldsymbol{x})]}}{\sum_{\boldsymbol{x}' \in \mathcal{C}} e^{-\ln[1/P(\boldsymbol{y}|\boldsymbol{x}')]}}. \end{aligned} \tag{410}$$

Here, the second line is deliberately written in a form that resembles the Boltzmann distribution, which naturally suggests to consider, more generally, the posterior distribution parametrized by β, that is

$$P_\beta(\boldsymbol{x}|\boldsymbol{y}) = \frac{P^\beta(\boldsymbol{y}|\boldsymbol{x})}{\sum_{\boldsymbol{x}' \in \mathcal{C}} P^\beta(\boldsymbol{y}|\boldsymbol{x}')}$$

$$= \frac{e^{-\beta \ln[1/P(\boldsymbol{y}|\boldsymbol{x})]}}{\sum_{\boldsymbol{x}' \in \mathcal{C}} e^{-\beta \ln[1/P(\boldsymbol{y}|\boldsymbol{x}')]}}$$
$$\triangleq \frac{e^{-\beta \ln[1/P(\boldsymbol{y}|\boldsymbol{x})]}}{Z(\beta|\boldsymbol{y})} \tag{411}$$

There are a few motivations for introducing the temperature parameter:

- It allows a degree of freedom in case there is some uncertainty regarding the channel noise level (small β corresponds to high noise level).

- It is inspired by the ideas behind simulated annealing techniques: by sampling from P_β while gradually increasing β (cooling the system), the minima of the energy function (ground states) can be found.

- By applying symbol-wise maximum a-posteriori (MAP) decoding, i.e., decoding the ℓ–th symbol of $\boldsymbol{x}$ as $\arg\max_a P_\beta(x_\ell = a|\boldsymbol{y})$, where

$$P_\beta(x_\ell = a|\boldsymbol{y}) = \sum_{\boldsymbol{x} \in \mathcal{C}:\ x_\ell = a} P_\beta(\boldsymbol{x}|\boldsymbol{y}), \tag{412}$$

  we obtain a family of *finite–temperature decoders* (originally proposed by Ruján [102]) parametrized by β, where $\beta = 1$ corresponds to minimum symbol error probability (with respect to the real underlying channel $P(\boldsymbol{y}|\boldsymbol{x})$) and $\beta \to \infty$ corresponds to minimum block error probability.

- This is one of our main motivations: the corresponding partition function, $Z(\beta|\boldsymbol{y})$, namely, the sum of (conditional) probabilities raised to some power β, is an expression frequently encountered in Rényi information measures as well as in the analysis of random coding exponents using Gallager's techniques. Since the partition function plays a key role in statistical mechanics, as many physical quantities can be derived from it, then it is natural to ask if it can also be used to gain some insights regarding the behavior of random codes at various temperatures and coding rates.

For the sake of simplicity, let us suppose further now that we are dealing with the binary

symmetric channel (BSC) with crossover probability p, and so,

$$P(\boldsymbol{y}|\boldsymbol{x}) = p^{d(\boldsymbol{x},\boldsymbol{y})}(1-p)^{n-d(\boldsymbol{x},\boldsymbol{y})} = (1-p)^n e^{-Jd(\boldsymbol{x},\boldsymbol{y})}, \tag{413}$$

where $J = \ln\frac{1-p}{p}$ and $d(\boldsymbol{x},\boldsymbol{y})$ is the Hamming distance. Thus, the partition function can be presented as follows:

$$Z(\beta|\boldsymbol{y}) = (1-p)^{\beta n}\sum_{\boldsymbol{x}\in\mathcal{C}} e^{-\beta Jd(\boldsymbol{x},\boldsymbol{y})}. \tag{414}$$

Now consider the fact that the codebook $\mathcal{C}$ is selected at random: Every codeword is randomly chosen independently of all other codewords. At this point, the analogy to the REM, and hence also its relevance, become apparent: If each codeword is selected independently, then the 'energies' $\{Jd(\boldsymbol{x},\boldsymbol{y})\}$ pertaining to the partition function

$$Z(\beta|\boldsymbol{y}) = (1-p)^{\beta n}\sum_{\boldsymbol{x}\in\mathcal{C}} e^{-\beta Jd(\boldsymbol{x},\boldsymbol{y})}, \tag{415}$$

(or, in the case of a more general channel, the energies $\{-\ln[1/P(\boldsymbol{y}|\boldsymbol{x})]\}$ pertaining to the partition function $Z(\beta|\boldsymbol{y}) = \sum_{\boldsymbol{x}\in\mathcal{C}} e^{-\beta\ln[1/P(\boldsymbol{y}|\boldsymbol{x})]}$), are i.i.d. random variables for all codewords in $\mathcal{C}$, with the exception of the codeword $\boldsymbol{x}_0$ that was actually transmitted and generated $\boldsymbol{y}$.[24] Since we have seen phase transitions in the REM, it is conceivable to expect them also in the statistical physics of the random code ensemble, and indeed we will see them shortly.

Further, we assume that each symbol of each codeword is drawn by fair coin tossing, i.e., independently and with equal probabilities for '0' and '1'. As said, we have to distinguish now between the contribution of the correct codeword $\boldsymbol{x}_0$, which is

$$Z_c(\beta|\boldsymbol{y}) \stackrel{\Delta}{=} (1-p)^{\beta n} e^{-Jd(\boldsymbol{x}_0,\boldsymbol{y})} \tag{416}$$

and the contribution of all other (incorrect) codewords:

$$Z_e(\beta|\boldsymbol{y}) \stackrel{\Delta}{=} (1-p)^{\beta n}\sum_{\boldsymbol{x}\in\mathcal{C}\setminus\{\boldsymbol{x}_0\}} e^{-Jd(\boldsymbol{x},\boldsymbol{y})}. \tag{417}$$

[24] This one is still independent, but it has a different distribution, and hence will be handled separately.

Concerning the former, things are very simple: Typically, the channel flips about np bits out the n transmissions, which means that with high probability, $d(\boldsymbol{x}_0, \boldsymbol{y})$ is about np, and so $Z_c(\beta|\boldsymbol{y})$ is expected to take values around $(1-p)^{\beta n} e^{-\beta J n p}$. The more complicated and more interesting question is how does $Z_e(\beta|\boldsymbol{y})$ behave, and here the treatment will be very similar to that of the REM.

Given $\boldsymbol{y}$, define $\Omega_{\boldsymbol{y}}(d)$ as the number of incorrect codewords whose Hamming distance from $\boldsymbol{y}$ is exactly d. Thus,

$$Z_e(\beta|\boldsymbol{y}) = (1-p)^{\beta n} \sum_{d=0}^{n} \Omega_{\boldsymbol{y}}(d) \cdot e^{-\beta J d}. \tag{418}$$

Just like in the REM, here too, the enumerator $\Omega_{\boldsymbol{y}}(d)$ is the sum of an exponential number, e^{nR}, of binary i.i.d. random variables:

$$\Omega_{\boldsymbol{y}}(d) = \sum_{\boldsymbol{x} \in \mathcal{C} \setminus \{\boldsymbol{x}_0\}} \mathcal{I}\{d(\boldsymbol{x}, \boldsymbol{y}) = d\}. \tag{419}$$

According to the method of types, the probability of a single 'success' $\{d(\boldsymbol{X}, \boldsymbol{y}) = n\delta\}$ is given by

$$\Pr\{d(\boldsymbol{X}, \boldsymbol{y}) = n\delta\} \doteq \frac{e^{n h_2(\delta)}}{2^n} = \exp\{-n[\ln 2 - h_2(\delta)]\}. \tag{420}$$

Thus, as in the REM, we have an exponential number of trials, e^{nR}, each one with an exponentially decaying probability of success, $e^{-n[\ln 2 - h_2(\delta)]}$. We already know how does this experiment behave: It depends which exponent is faster. If $R > \ln 2 - h_2(\delta)$, we will typically see about $\exp\{n[R + h_2(\delta) - \ln 2]\}$ codewords at distance $d = n\delta$ from $\boldsymbol{y}$. Otherwise, we see none. So the critical value of δ is the solution to the equation

$$R + h_2(\delta) - \ln 2 = 0. \tag{421}$$

There are two solutions to this equation, which are symmetric about 1/2. The smaller one is called the Gilbert–Varshamov (G–V) distance[25] and it will be denoted by $\delta_{GV}(R)$ (see Fig. 17). The other solution is, of course, $\delta = 1 - \delta_{GV}(R)$. Thus, the condition $R > \ln 2 - h_2(\delta)$

[25]The G–V distance was originally defined and used in coding theory for the BSC.

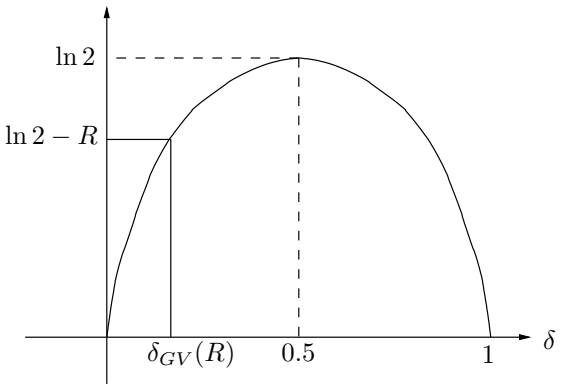


Figure 17: The Gilbert–Varshamov distance as the smaller solution to the equation $R + h_2(\delta) - \ln 2 = 0$.

is equivalent to $\delta_{GV}(R) < \delta < 1 - \delta_{GV}(R)$, and so, for a typical code in the ensemble:

$$\begin{aligned} Z_e(\beta|\boldsymbol{y}) &\approx (1-p)^{\beta n} \sum_{\delta=\delta_{GV}(R)}^{1-\delta_{GV}(R)} \exp\{n[R + h_2(\delta) - \ln 2]\} \cdot e^{-\beta J n \delta} \\ &= (1-p)^{\beta n} e^{n(R-\ln 2)} \cdot \sum_{\delta=\delta_{GV}(R)}^{1-\delta_{GV}(R)} \exp\{n[h_2(\delta) - \beta J \delta]\} \\ &= (1-p)^{\beta n} e^{n(R-\ln 2)} \times \\ &\exp\left\{ n \cdot \max_{\delta_{GV}(R) \le \delta \le 1-\delta_{GV}(R)} [h_2(\delta) - \beta J \delta] \right\}. \end{aligned} \tag{422}$$

Now, similarly as in the REM, we have to maximize a certain function within a limited interval. And again, there are two phases, corresponding to whether the maximizer falls at an edge–point (glassy phase) or at an internal point with zero derivative (paramagnetic phase). It is easy to show that in the paramagnetic phase, the maximum is attained at

$$\delta^* = p_\beta \stackrel{\Delta}{=} \frac{p^\beta}{p^\beta + (1-p)^\beta} \tag{423}$$

and then

$$\phi(\beta) = R - \ln 2 + \ln[p^\beta + (1-p)^\beta]. \tag{424}$$

In the glassy phase, $\delta^* = \delta_{GV}(R)$ and then

$$\phi(\beta) = \beta[\delta_{GV}(R) \ln p + (1 - \delta_{GV}(R)) \ln(1-p)], \tag{425}$$

which is again, linear in β and hence corresponds to zero entropy. The boundary between the two phases occurs when β is such that $\delta_{GV}(R) = p_\beta$, which is equivalent to

$$\beta = \beta_c(R) \stackrel{\Delta}{=} \frac{\ln[(1-\delta_{GV}(R))/\delta_{GV}(R)]}{\ln[(1-p)/p]}. \tag{426}$$

Thus, $\beta < \beta_c(R)$ is the paramagnetic phase of Z_e and $\beta > \beta_c(R)$ is its glassy phase.

But now we should remember that Z_e is only part of the partition function and it is now the time to put the contribution of Z_c back into the picture. Checking the dominant contribution of $Z = Z_e + Z_c$ as a function of β and R, we can draw a phase diagram, where we find that there are actually three phases, two contributed by Z_e, as we have already seen (paramagnetic and glassy), plus a third phase – contributed by Z_c, namely, the *ordered* or the *ferromagnetic* phase, where Z_c dominates (cf. Fig. 18), which means reliable communication, as the correct codeword $\boldsymbol{x}_0$ dominates the partition function and hence the posterior distribution. The boundaries of the ferromagnetic phase designate phase transitions from reliable to unreliable decoding.

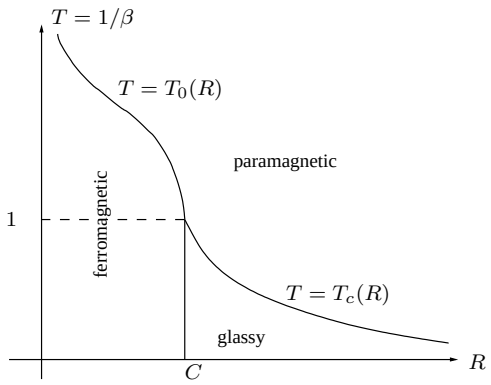


Figure 18: Phase diagram of the finite–temperature MAP decoder.

Both the glassy phase and the paramagnetic phase correspond to unreliable communication. What is the essential difference between them? As in the REM, the difference is that in the glassy phase, Z is dominated by a subexponential number of codewords at the 'ground–state energy', namely, that minimum seen distance of $n\delta_{GV}(R)$, whereas in the paramagnetic phase, the dominant contribution comes from an exponential number of codewords

at distance np_β. In the glassy phase, there is *seemingly* a smaller degree of uncertainty since $H(\boldsymbol{X}|\boldsymbol{Y})$ that is induced from the finite–temperature posterior has zero entropy. But this is fictitious since the main support of the posterior belongs to incorrect codewords. This is to say that we may have the illusion that we know quite a lot about the transmitted codeword, but what we know is wrong! This is like an event of an undetected error. In both glassy and paramagnetic phases, above capacity, the ranking of the correct codeword, in the list of decreasing $P_\beta(\boldsymbol{x}|\boldsymbol{y})$, is about $e^{n(R-C)}$.

6.3 Random Coding Exponents

It turns out that these findings are relevant to ensemble performance analysis of codes. This is because many of the bounds on code performance include summations of $P^\beta(\boldsymbol{y}|\boldsymbol{x})$ (for some β), which are exactly the partition functions that we worked with in the derivations of the previous subsection. These considerations can sometimes even help to get tighter bounds. We will first demonstrate this point in the context of the analysis of the probability of correct decoding above capacity.

We begin with the following well known expression of the probability of correct decoding:

$$\begin{aligned} P_c &= \frac{1}{M}\sum_{\boldsymbol{y}} \max_{\boldsymbol{x}\in\mathcal{C}} P(\boldsymbol{y}|\boldsymbol{x}) \\ &= \lim_{\beta\to\infty} \frac{1}{M}\sum_{\boldsymbol{y}} \left[\sum_{\boldsymbol{x}\in\mathcal{C}} P^\beta(\boldsymbol{y}|\boldsymbol{x})\right]^{1/\beta} \end{aligned} \tag{427}$$

The expression in the square brackets is readily identified with the partition function, and we note that the combination of $R > C$ and $\beta \to \infty$ takes us deep into the glassy phase. Taking the ensemble average, we get:

$$\bar{P}_c = \lim_{\beta\to\infty} \frac{1}{M}\sum_{\boldsymbol{y}} \boldsymbol{E}\left\{\left[\sum_{\boldsymbol{x}\in\mathcal{C}} P^\beta(\boldsymbol{y}|\boldsymbol{x})\right]^{1/\beta}\right\}. \tag{428}$$

At this point, the traditional approach would be to insert the expectation into the square brackets by applying Jensen's inequality (for $\beta \geq 1$), which would give us an upper bound.

Instead, our previous treatment of random code ensembles as a REM–like model can give us a hand on exponentially tight evaluation of the last expression, with Jensen's inequality being avoided. Consider the following chain:

$$\begin{aligned}
\boldsymbol{E}\left\{\left[\sum_{\boldsymbol{x}\in\mathcal{C}} P^{\beta}(\boldsymbol{y}|\boldsymbol{x})\right]^{1/\beta}\right\} &= (1-p)^n \boldsymbol{E}\left\{\left[\sum_{d=0}^{n} \Omega_{\boldsymbol{y}}(d) e^{-\beta J d}\right]^{1/\beta}\right\} \\
&\doteq (1-p)^n \boldsymbol{E}\left\{\left[\max_{0\le d\le n} \Omega_{\boldsymbol{y}}(d) e^{-\beta J d}\right]^{1/\beta}\right\} \\
&= (1-p)^n \boldsymbol{E}\left\{\max_{0\le d\le n} [\Omega_{\boldsymbol{y}}(d)]^{1/\beta} \cdot e^{-J d}\right\} \\
&\doteq (1-p)^n \boldsymbol{E}\left\{\sum_{d=0}^{n} [\Omega_{\boldsymbol{y}}(d)]^{1/\beta} \cdot e^{-J d}\right\} \\
&= (1-p)^n \sum_{d=0}^{n} \boldsymbol{E}\left\{[\Omega_{\boldsymbol{y}}(d)]^{1/\beta}\right\} \cdot e^{-J d}
\end{aligned}$$

Thus, it boils down to the calculation of (non–integer) moments of $\Omega_{\boldsymbol{y}}(d)$. At this point, we adopt the main ideas of the treatment of the REM, distinguishing between the values of δ below the G–V distance, and those that are above it. Before we actually assess the moments of $\Omega_{\boldsymbol{y}}(d)$, we take a closer look at the asymptotic behavior of these random variables. This will also rigorize our earlier discussion on the Gaussian REM.

For two numbers a and b in $[0,1]$, let us define the binary divergence as

$$D(a\|b) = a\ln\frac{a}{b} + (1-a)\ln\frac{1-a}{1-b}. \tag{429}$$

Using the inequality

$$\ln(1+x) = -\ln\left(1-\frac{x}{1+x}\right) \ge \frac{x}{1+x},$$

we get the following lower bound to $D(a\|b)$:

$$\begin{aligned}
D(a\|b) &= a\ln\frac{a}{b} + (1-a)\ln\frac{1-a}{1-b} \\
&= a\ln\frac{a}{b} + (1-a)\ln\left(1+\frac{b-a}{1-b}\right) \\
&\ge a\ln\frac{a}{b} + (1-a)\cdot\frac{(b-a)/(1-b)}{1+(b-a)/(1-b)}
\end{aligned}$$

$$
\begin{aligned}
&= a \ln \frac{a}{b} + b - a \\
&> a \left(\ln \frac{a}{b} - 1 \right)
\end{aligned}
$$

Now, as mentioned earlier, $\Omega_{\boldsymbol{y}}(d)$ is the sum of e^{nR} i.i.d. binary random variables, i.e., Bernoulli random variables with parameter $e^{-n[\ln 2 - h_2(\delta)]}$. Consider the event $\Omega_{\boldsymbol{y}}(d) \geq e^{nA}$, $A \geq 0$, which means that the relative frequency of 'successes' exceeds $\frac{e^{nA}}{e^{nR}} = e^{-n(R-A)}$. Then this is a large deviations event if $e^{-n(R-A)} > e^{-n[\ln 2 - h_2(\delta)]}$, that is,

$$
A > R + h_2(\delta) - \ln 2. \tag{430}
$$

Using the Chernoff bound,[26] one can easily show that

$$
\Pr\{\Omega_{\boldsymbol{y}}(d) \geq e^{nA}\} \leq \exp\{-e^{nR} D(e^{-n(R-A)} \| e^{-n[\ln 2 - h_2(\delta)]})\}. \tag{431}
$$

Now, by applying the above lower bound to the binary divergence, we can further upper bound the last expression as

$$
\begin{aligned}
\Pr\{\Omega_{\boldsymbol{y}}(d) \geq e^{nA}\} \leq \exp\{ & -e^{nR} \cdot e^{-n(R-A)} \times \\
& (n[\ln 2 - R - h_2(\delta) + A] - 1)\} \\
& = \exp\{-e^{nA} \cdot (n[\ln 2 - R - h_2(\delta) + A] - 1)\}
\end{aligned}
$$

Now, suppose first that $\delta_{GV}(R) < \delta < 1 - \delta_{GV}(R)$, and take $A = R + h_2(\delta) - \ln 2 + \epsilon$, where $\epsilon > 0$ may not necessarily be small. In this case, the term in the square brackets is ϵ, which means that the right–most side decays doubly–exponentially rapidly. Thus, for $\delta_{GV}(R) < \delta < 1 - \delta_{GV}(R)$, the probability that $\Omega_{\boldsymbol{y}}(d)$ exceeds $\boldsymbol{E}\{\Omega_{\boldsymbol{y}}(d)\} \cdot e^{n\epsilon}$ decays double–exponentially fast with n. One can show in a similar manner[27] that $\Pr\{\Omega_{\boldsymbol{y}}(d) < \boldsymbol{E}\{\Omega_{\boldsymbol{y}}(d)\} \cdot e^{-n\epsilon}\}$ decays in a double exponential rate as well. Finally, consider the case where $\delta < \delta_{GV}(R)$ or $\delta > 1 - \delta_{GV}(R)$, and let $A = 0$. This is also a large deviations event, and hence the above bound continues to be valid. Here, by setting $A = 0$, we get an ordinary exponential decay:

$$
\Pr\{\Omega_{\boldsymbol{y}}(d) \geq 1\} \doteq e^{-n[\ln 2 - R - h_2(\delta)]}. \tag{432}
$$

[26] We emphasize the use of the Chernoff bound, as opposed to the method of types, since the method of types would introduce the factor of the number of type classes, which is in this case $(e^{nR} + 1)$.

[27] This requires a slightly different lower bound to the binary divergence.

After having prepared these results, let us get back to the evaluation of the moments of $\Omega_{\boldsymbol{y}}(d)$. Once again, we separate between the two ranges of δ. For $\delta < \delta_{GV}(R)$ or $\delta > 1 - \delta_{GV}(R)$, we have the following:

$$\begin{aligned}\boldsymbol{E}\{[\Omega_{\boldsymbol{y}}(d)]^{1/\beta}\} &\doteq 0^{1/\beta} \cdot \Pr\{\Omega_{\boldsymbol{y}}(d) = 0\} + e^{n\cdot 0/\beta} \cdot \Pr\{1 \leq \Omega_{\boldsymbol{y}}(d) \leq e^{n\epsilon}\} \\ &\quad + \text{double–exponentially small terms} \\ &\doteq e^{n\cdot 0/\beta} \cdot \Pr\{\Omega_{\boldsymbol{y}}(d) \geq 1\} \\ &\doteq e^{-n[\ln 2 - R - h_2(\delta)]} \end{aligned} \tag{433}$$

Thus, in this range, $\boldsymbol{E}\{[\Omega_{\boldsymbol{y}}(d)]^{1/\beta}\} \doteq e^{-n[\ln 2 - R - h_2(\delta)]}$ independently of β. On the other hand in the range $\delta_{GV}(R) < \delta < 1 - \delta_{GV}(R)$,

$$\begin{aligned}\boldsymbol{E}\{[\Omega_{\boldsymbol{y}}(d)]^{1/\beta}\} &\doteq (e^{n[R + h_2(\delta) - \ln 2]})^{1/\beta} \times \\ &\quad \Pr\{e^{n[R + h_2(\delta) - \ln 2 - \epsilon]} \leq \Omega_{\boldsymbol{y}}(d) \leq e^{n[R + h_2(\delta) - \ln 2 + \epsilon]}\} + \\ &\quad + \text{double–exponentially small terms} \\ &\doteq e^{n[R + h_2(\delta) - \ln 2]/\beta} \end{aligned} \tag{434}$$

since the probability $\Pr\{e^{n[R + h_2(\delta) - \ln 2 - \epsilon]} \leq \Omega_{\boldsymbol{y}}(d) \leq e^{n[R + h_2(\delta) - \ln 2 + \epsilon]}\}$ tends to unity double–exponentially rapidly. So to summarize, we have shown that the moment of $\Omega_{\boldsymbol{y}}(d)$ undergoes a phase transition, as it behaves as follows:

$$\boldsymbol{E}\{[\Omega_{\boldsymbol{y}}(d)]^{1/\beta}\} \doteq \begin{cases} e^{n[R + h_2(\delta) - \ln 2]} & \delta < \delta_{GV}(R) \ \text{ or } \ \delta > 1 - \delta_{GV}(R) \\ e^{n[R + h_2(\delta) - \ln 2]/\beta} & \delta_{GV}(R) < \delta < 1 - \delta_{GV}(R) \end{cases}$$

Finally, on substituting these moments back into the expression of $\bar{P}_c$, and taking the limit $\beta \to \infty$, we eventually get:

$$\lim_{\beta \to \infty} \boldsymbol{E} \left\{ \left[\sum_{\boldsymbol{x} \in \mathcal{C}} P^{\beta}(\boldsymbol{y}|\boldsymbol{x}) \right]^{1/\beta} \right\} \doteq e^{-nF_g} \tag{435}$$

where F_g is the free energy of the glassy phase, i.e.,

$$F_g = \delta_{GV}(R) \ln \frac{1}{p} + (1 - \delta_{GV}(R)) \ln \frac{1}{1-p} \tag{436}$$

and so, we obtain a very simple relation between the exponent of $\bar{P}_c$ and the free energy of the glassy phase:

$$\begin{aligned}
\bar{P}_c &\doteq \frac{1}{M}\sum_{\boldsymbol{y}} e^{-nF_g} \\
&= \exp\{n(\ln 2 - R - F_g)\} \\
&= \exp\{n[\ln 2 - R + \delta_{GV}(R)\ln p + (1-\delta_{GV}(R))\ln(1-p)]\} \\
&= \exp\{n[h_2(\delta_{GV}(R)) + \delta_{GV}(R)\ln p + (1-\delta_{GV}(R))\ln(1-p)]\} \\
&= e^{-nD(\delta_{GV}(R)\|p)}
\end{aligned} \tag{437}$$

The last expression has an intuitive interpretation. It answers the following question: what is the probability that the channel would flip less than $n\delta_{GV}(R)$ bits although $p > \delta_{GV}(R)$? This is exactly the relevant question for correct decoding in the glassy phase, because in that phase, there is a "belt" of codewords "surrounding" $\boldsymbol{y}$ at radius $n\delta_{GV}(R)$ – these are the codewords that dominate the partition function in the glassy phase and there are no codewords closer to $\boldsymbol{y}$. The event of correct decoding happens if the channel flips less than $n\delta_{GV}(R)$ bits and then $\boldsymbol{x}_0$ is closer to $\boldsymbol{y}$ more than all belt–codewords. Thus, $\boldsymbol{x}_0$ is decoded correctly.

It is interesting, at this point, to compare the above calculation to the bound obtained using Jensen's inequality. There we have

$$\begin{aligned}
\boldsymbol{E}\left[\sum_{\boldsymbol{x}\in\mathcal{C}} P^\beta(\boldsymbol{y}|\boldsymbol{x})\right]^{1/\beta} &\le \left[\boldsymbol{E}\sum_{\boldsymbol{x}\in\mathcal{C}} P^\beta(\boldsymbol{y}|\boldsymbol{x})\right]^{1/\beta} \\
&= \left[e^{nR}\boldsymbol{E}P^\beta(\boldsymbol{y}|\boldsymbol{x})\right]^{1/\beta} \\
&= e^{nR/\beta}\left[\sum_{\boldsymbol{x}\in\{0,1\}^n} 2^{-n}\prod_{i=1}^n P^\beta(y_i|x_i)\right]^{1/\beta} \\
&= e^{nR/\beta}\left[\prod_{i=1}^n\sum_{x_i=0}^1\left(\frac{1}{2}P^\beta(y_i|x_i)\right)\right]^{1/\beta} \\
&= e^{nR/\beta}\left[\prod_{i=1}^n\left(\frac{1}{2}p^\beta + \frac{1}{2}(1-p)^\beta\right)\right]^{1/\beta}
\end{aligned}$$

$$= e^{nR/\beta}\left[\frac{1}{2}p^{\beta}+\frac{1}{2}(1-p)^{\beta}\right]^{n/\beta}, \tag{438}$$

which, after taking the limit $\beta\to\infty$, becomes $[\max\{p,1-p\}]^n = e^{n\max\{\ln p,\ln(1-p)\}}$. The exponential term $\max\{\ln p,\ln(1-p)\}$ should be compared to the weighted average of $\ln p$ and $\ln(1-p)$, with weights $\delta_{GV}(R)$ and $1-\delta_{GV}(R)$ that we obtained before, and so, the difference is quite clear.

One can also derive an upper bound on the error probability at $R<C$. The partition function $Z(\beta|\boldsymbol{y})$ plays a role there too according to Gallager's classical bounds. We will not delve now into it, but we only comment that in that case, the calculation is performed in the paramagnetic regime rather than the glassy regime that we have seen in the calculation of $\bar{P}_c$. The basic technique, however, is essentially the same.

We will now demonstrate the usefulness of this technique of assessing moments of distance enumerators in a certain problem of decoding with an erasure option. Consider the BSC with a crossover probability $p<1/2$, which is unknown and one employs a universal detector that operates according to the following decision rule: Select the message m if

$$\frac{e^{-n\beta\hat{h}(\boldsymbol{x}_m\oplus\boldsymbol{y})}}{\sum_{m'\neq m}e^{-n\beta\hat{h}(\boldsymbol{x}_{m'}\oplus\boldsymbol{y})}}\geq e^{nT} \tag{439}$$

where $\beta>0$ is an inverse temperature parameter and $\hat{h}(\boldsymbol{x}\oplus\boldsymbol{y})$ is the binary entropy pertaining to the relative number of 1's in the vector resulting from bit–by–bit XOR of $\boldsymbol{x}$ and $\boldsymbol{y}$, namely, the binary entropy function computed at the normalized Hamming distance between $\boldsymbol{x}$ and $\boldsymbol{y}$. If no message m satisfies (439), then an erasure is declared.

We have no optimality claims regarding this decision rule, but arguably, it is a reasonable decision rule (and hence there is motivation to analyze its performance): It is a universal version of the optimum decision rule:

$$\text{Decide } m \text{ if } \frac{P(\boldsymbol{y}|\boldsymbol{x}_m)}{\sum_{m'\neq m}P(\boldsymbol{y}|\boldsymbol{x}_{m'})}\geq e^{nT} \text{ and erase otherwise.} \tag{440}$$

The minimization of $\hat{h}(\boldsymbol{x}_m\oplus\boldsymbol{y})$ among all code-vectors $\{\boldsymbol{x}_m\}$, namely, the *minimum conditional entropy decoder* is a well–known universal decoding rule in the ordinary decoding

regime, without erasures, which in the simple case of the BSC, is equivalent to the *maximum mutual information* (MMI) decoder [16] and to the *generalized likelihood ratio test* (GLRT) decoder, which jointly maximizes the likelihood over both the message and the unknown parameter.

Here, we adapt the minimum conditional entropy decoder to the structure proposed by the optimum decoder with erasures, where the (unknown) likelihood of each codeword $\boldsymbol{x}_m$ is basically replaced by its maximum $e^{-n\hat{h}(\boldsymbol{x}_m \oplus \boldsymbol{y})}$, but with an additional degree of freedom of scaling the exponent by β. The parameter β controls the relative importance of the codeword with the second highest score. For example, when $\beta \to \infty$,[28] only the first and the second highest scores count in the decision, whereas if $\beta \to 0$, the differences between the scores of all codewords are washed out.

To demonstrate the advantage of the proposed analysis technique, we will now apply it in comparison to the traditional approach of using Jensen's inequality and supplementing an additional parameter ρ in the bound so as to monitor the loss of tightness due to the use of Jensen's inequality (see also [78]). Let us analyze the probability of the event $\mathcal{E}_1$ that the transmitted codeword $\boldsymbol{x}_m$ does not satisfy (439). We then have the following chain of inequalities, where the first few steps are common to the two analysis methods to be compared:

$$\begin{aligned}
\Pr\{\mathcal{E}_1\} &= \frac{1}{M}\sum_{m=1}^{M}\sum_{\boldsymbol{y}} P(\boldsymbol{y}|\boldsymbol{x}_m)\cdot 1\left\{\frac{e^{nT}\sum_{m'\neq m} e^{-n\beta\hat{h}(\boldsymbol{x}_{m'}\oplus\boldsymbol{y})}}{e^{-n\beta\hat{h}(\boldsymbol{x}_m\oplus\boldsymbol{y})}} \geq 1\right\} \\
&\leq \frac{1}{M}\sum_{m=1}^{M}\sum_{\boldsymbol{y}} P(\boldsymbol{y}|\boldsymbol{x}_m)\cdot\left[\frac{e^{nT}\sum_{m'\neq m} e^{-n\beta\hat{h}(\boldsymbol{x}_{m'}\oplus\boldsymbol{y})}}{e^{-n\beta\hat{h}(\boldsymbol{x}_m\oplus\boldsymbol{y})}}\right]^s \\
&= \frac{e^{nsT}}{M}\sum_{m=1}^{M}\sum_{\boldsymbol{y}} P(\boldsymbol{y}|\boldsymbol{x}_m)\cdot e^{n\beta s\hat{h}(\boldsymbol{x}_m\oplus\boldsymbol{y})}\times \\
&\quad\left[\sum_{m'\neq m} e^{-n\beta\hat{h}(\boldsymbol{x}_{m'}\oplus\boldsymbol{y})}\right]^s . \qquad (441)
\end{aligned}$$

Considering now the ensemble of codewords drawn independently by fair coin tossing, we

[28] As β varies it is plausible to let T scale linearly with β.

have:

$$\begin{aligned}\overline{\Pr}\{\mathcal{E}_1\} &\le e^{nsT}\sum_{\boldsymbol{y}} \boldsymbol{E}\left\{P(\boldsymbol{y}|\boldsymbol{X}_1)\cdot\exp[n\beta s\hat{h}(\boldsymbol{X}_1\oplus\boldsymbol{y})]\right\}\times\\ &\quad \boldsymbol{E}\left\{\left[\sum_{m>1}\exp[-n\beta\hat{h}(\boldsymbol{X}_m\oplus\boldsymbol{y})]\right]^s\right\}\\ &\stackrel{\Delta}{=} e^{nsT}\sum_{\boldsymbol{y}} A(\boldsymbol{y})\cdot B(\boldsymbol{y}) \end{aligned} \tag{442}$$

The computation of $A(\boldsymbol{y})$ is as follows: Denoting the Hamming weight of a binary sequence $\boldsymbol{z}$ by $w(\boldsymbol{z})$, we have:

$$\begin{aligned}A(\boldsymbol{y}) &= \sum_{\boldsymbol{x}} 2^{-n}(1-p)^n\cdot\left(\frac{p}{1-p}\right)^{w(\boldsymbol{x}\oplus\boldsymbol{y})}\exp[n\beta s\hat{h}(\boldsymbol{x}\oplus\boldsymbol{y})]\\ &= \left(\frac{1-p}{2}\right)^n\sum_{\boldsymbol{z}}\exp\left[n\left(w(\boldsymbol{z})\ln\frac{p}{1-p}+\beta s\hat{h}(\boldsymbol{z})\right)\right]\\ &\stackrel{\cdot}{=} \left(\frac{1-p}{2}\right)^n\sum_{\delta} e^{nh(\delta)}\cdot\exp\left[n\left(\beta s h(\delta)-\delta\ln\frac{1-p}{p}\right)\right]\\ &\stackrel{\cdot}{=} \left(\frac{1-p}{2}\right)^n\exp\left[n\max_{\delta}\left((1+\beta s)h(\delta)-\delta\ln\frac{1-p}{p}\right)\right].\end{aligned}$$

It is readily seen by ordinary optimization that

$$\begin{aligned}&\max_{\delta}\left[(1+\beta s)h(\delta)-\delta\ln\frac{1-p}{p}\right]\\ &= (1+\beta s)\ln\left[p^{1/(1+\beta s)}+(1-p)^{1/(1+\beta s)}\right]-\ln(1-p)\end{aligned} \tag{443}$$

and so upon substituting back into the the bound on $\overline{\Pr}\{\mathcal{E}_1\}$, we get:

$$\begin{aligned}\overline{\Pr}\{\mathcal{E}_1\} \le \exp\big[n\,(&sT+(1+\beta s)\times\\ &\ln\left[p^{1/(1+\beta s)}+(1-p)^{1/(1+\beta s)}\right]-\ln 2)\big]\cdot\sum_{\boldsymbol{y}} B(\boldsymbol{y}).\end{aligned} \tag{444}$$

It remains then to assess the exponential order of $B(\boldsymbol{y})$ and this will now be done in two different ways. The first is Forney's way [32] of using Jensen's inequality and introducing the additional parameter ρ, i.e.,

$$B(\boldsymbol{y}) = \boldsymbol{E}\left\{\left(\left[\sum_{m>1}\exp[-n\beta\hat{h}(\boldsymbol{X}_m\oplus\boldsymbol{y})\right]^{s/\rho}\right)^{\rho}\right\}$$

$$\leq \boldsymbol{E}\left\{\left(\sum_{m>1}\exp[-n\beta s\hat{h}(\boldsymbol{X}_m\oplus\boldsymbol{y})/\rho]\right)^{\rho}\right\} \qquad 0\leq s/\rho\leq 1$$
$$\leq e^{n\rho R}\left(\boldsymbol{E}\left\{\exp[-n\beta s\hat{h}(\boldsymbol{X}_m\oplus\boldsymbol{y})/\rho]\right\}\right)^{\rho}, \qquad \rho\leq 1 \tag{445}$$

where in the second line we have used the following inequality[29] for non–negative $\{a_i\}$ and $\theta\in[0,1]$:

$$\left(\sum_i a_i\right)^{\theta}\leq\sum_i a_i^{\theta}. \tag{446}$$

Now,

$$\begin{aligned}\boldsymbol{E}\left\{\exp[-n\beta s\hat{h}(\boldsymbol{X}_m\oplus\boldsymbol{y})/\rho]\right\} &= 2^{-n}\sum_{\boldsymbol{z}}\exp[-n\beta s\hat{h}(\boldsymbol{z})/\rho]\\ &\doteq 2^{-n}\sum_{\delta}e^{nh(\delta)}\cdot e^{-n\beta sh(\delta)/\rho}\\ &= \exp[n([1-\beta s/\rho]_+-1)\ln 2],\end{aligned}$$

where $[u]_+\stackrel{\Delta}{=}\max\{u,0\}$. Thus, we get

$$B(\boldsymbol{y})\leq\exp(n[\rho(R-\ln 2)+[\rho-\beta s]_+]), \tag{447}$$

which when substituted back into the bound on $\overline{\Pr}\{\mathcal{E}_1\}$, yields an exponential rate of

$$\begin{aligned}\tilde{E}_1(R,T) = \max_{0\leq s\leq\rho\leq 1}\{&(\rho-[\rho-\beta s]_+)\ln 2-\\ &-(1+\beta s)\ln\left[p^{1/(1+\beta s)}+(1-p)^{1/(1+\beta s)}\right]-\rho R-sT\}.\end{aligned}$$

On the other hand, estimating $B(\boldsymbol{y})$ by the new method, we have:

$$\begin{aligned}B(\boldsymbol{y}) &= \boldsymbol{E}\left\{\left[\sum_{m>1}\exp[-n\beta\hat{h}(\boldsymbol{X}_m\oplus\boldsymbol{y})]\right]^s\right\}\\ &= \boldsymbol{E}\left\{\left[\sum_{\delta}\Omega_{\boldsymbol{y}}(n\delta)\exp[-n\beta h(\delta)]\right]^s\right\}\end{aligned}$$

[29]To see why this is true, think of $p_i=a_i/(\sum_i a_i)$ as probabilities, and then $p_i^{\theta}\geq p_i$, which implies $\sum_i p_i^{\theta}\geq\sum_i p_i=1$. The idea behind the introduction of the new parameter ρ is to monitor the possible loss of exponential tightness due to the use of Jensen's inequality. If $\rho=1$, there is no loss at all due to Jensen, but there is maximum loss in the second line of the chain. If $\rho=s$, it is the other way around. Hopefully, after optimization over ρ, the overall loss in tightness is minimized.

$$
\begin{aligned}
&\dot{=} \sum_{\delta} \boldsymbol{E}\{\Omega_{\boldsymbol{y}}^{s}(n\delta)\} \cdot \exp(-n\beta s h(\delta)) \\
&\dot{=} \sum_{\delta \in \mathcal{G}_R^c} e^{n[R+h(\delta)-\ln 2]} \cdot \exp[-n\beta s h(\delta)] + \\
&\sum_{\delta \in \mathcal{G}_R} e^{ns[R+h(\delta)-\ln 2]} \cdot \exp[-n\beta s h(\delta)] \\
&\stackrel{\Delta}{=} U + V, \qquad (448)
\end{aligned}
$$

where $\mathcal{G}_R = \{\delta : \ \delta_{GV}(R) \leq \delta \leq 1-\delta_{GV}(R)\}$. Now, U is dominated by the term $\delta = 0$ if $\beta s > 1$ and $\delta = \delta_{GV}(R)$ if $\beta s < 1$. It is then easy to see that $U \dot{=} \exp[-n(\ln 2 - R)(1-[1-\beta s]_+)]$. Similarly, V is dominated by the term $\delta = 1/2$ if $\beta < 1$ and $\delta = \delta_{GV}(R)$ if $\beta \geq 1$. Thus, $V \dot{=} \exp[-ns(\beta[\ln 2 - R] - R[1-\beta]_+)]$. Therefore, defining

$$
\phi(R, \beta, s) = \min\{(\ln 2 - R)(1 - [1 - \beta s]_+), s(\beta[\ln 2 - R] - R[1-\beta]_+)\},
$$

the resulting exponent is

$$
\begin{aligned}
\hat{E}_1(R, T) = \max_{s \geq 0} \{&\phi(R, \beta, s) - (1+\beta s) \times \\
&\ln\left[p^{1/(1+\beta s)} + (1-p)^{1/(1+\beta s)}\right] - sT\Big\}. \qquad (449)
\end{aligned}
$$

Numerical comparisons show that while there are many quadruples (p, β, R, T) for which the two exponents coincide, there are also situations where $\hat{E}_1(R, T)$ exceeds $\tilde{E}_1(R, T)$. To demonstrate these situations, consider the values $p = 0.1$, $\beta = 0.5$, $T = 0.001$, and let R vary in steps of 0.01. Table 1 summarizes numerical values of both exponents, where the optimizations over ρ and s were conducted by an exhaustive search with a step size of 0.005 in each parameter. In the case of $\hat{E}_1(R, T)$, where $s \geq 0$ is not limited to the interval $[0, 1]$ (since Jensen's inequality is not used), the numerical search over s was limited to the interval $[0, 5]$.[30]

As can be seen (see also Fig. 19), the numerical values of the exponent $\hat{E}_1(R, T)$ are considerably larger than those of $\tilde{E}_1(R, T)$ in this example, which means that the analysis

[30] It is interesting to note that for some values of R, the optimum value s^* of the parameter s was indeed larger than 1. For example, at rate $R = 0$, we have $s^* = 2$ in the above search resolution.

	$R = 0.00$	$R = 0.01$	$R = 0.02$	$R = 0.03$	$R = 0.04$
$\tilde{E}_1(R,T)$	0.1390	0.1290	0.1190	0.1090	0.0990
$\hat{E}_1(R,T)$	0.2211	0.2027	0.1838	0.1642	0.1441

Table 1: Numerical values of $\tilde{E}_1(R,T)$ and $\hat{E}_1(R,T)$ as functions of R for $p = 0.1$, $\beta = 0.5$, and $T = 0.001$.

technique proposed here, not only simplifies exponential error bounds, but sometimes leads also to significantly tighter bounds.

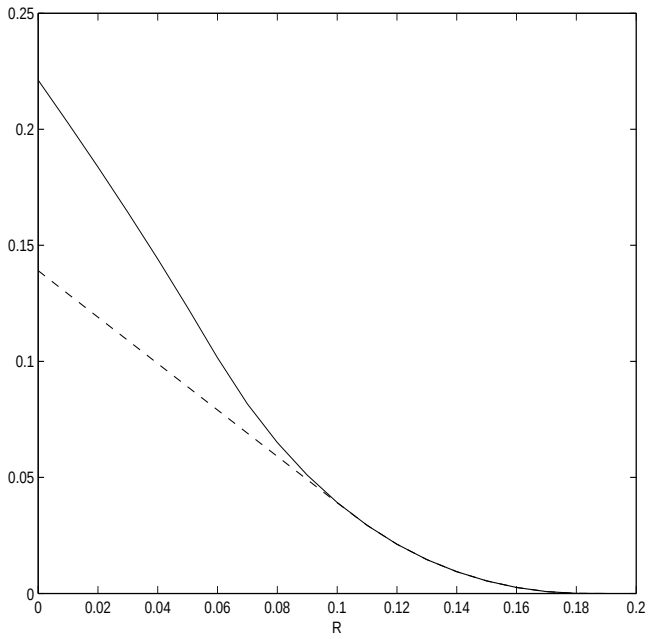


Figure 19: Graphs of $\hat{E}_1(R,T)$ (solid line) and $\tilde{E}_1(R,T)$ (dashed line) as functions of R for $p = 0.1$, $T = 0.001$ and $\beta = 0.5$.

There are other examples where these techniques are used in more involved situations, and in some of them they yield better performance bounds compared to traditional methods. These includes analysis of error exponents for interference channels [31], broadcast channels [55], and exact expressions for error exponents for decoding with erasures [106].

Beside error exponents, these techniques were used successfully in problems of signal estimation [79] and parameter estimation [77], where threshold effects in the estimation problems were induced by phase transitions in the analogous physical systems.

7 Extensions of the REM

In this chapter, we introduce a few extensions of the REM. The first extension, discussed in Section 7.1, incorporates an external magnetic field. In the realm of coded communication, the analogue of a magnetic field turns out to be related to possible non–uniformity in the probability distribution of the input messages, namely, it corresponds to joint source–channel coding. In Section 7.2, we discuss the generalized REM (GREM), which introduces correlations between the energies of the various configurations in an hierarchical structure. This will help us to gain some insights on the behavior certain structures of hierarchical codes. It should be pointed out that the GREM was extended and studied in probability theory under the name of "Ruelle probability cascades," first by Ruelle and then by Bolthausen and Sznitman (see, e.g., [1] and references therein). Finally, in Section 7.3, this hierarchical structure of the GREM will be pushed to the extreme of the largest possible depth, leading to another customary model of disordered system, called *directed polymer in a random medium* (DPRM), which has an intimate relation to tree coding. The DPRM was already used earlier in other information–theoretic studies, such as [81, Appendix B].

7.1 REM Under Magnetic Field and Source–Channel Coding

We begin with the physics background of extending the analysis of the REM so as to include an external magnetic field.

7.1.1 Magnetic Properties of the REM

Earlier, we studied the REM in the absence of an external magnetic field. The Gaussian randomly drawn energies that we discussed were a caricature of the interaction energies in the p–spin glass model for an extremely large level of disorder, in the absence of a magnetic field.

As said, we now expand the analysis of the REM so as to incorporate also an external magnetic field B. This will turn out to be relevant to a more general communication setting,

namely, that of joint source–channel coding, where as we shall see, the possible skewedness of the probability distribution of the source (when it is not symmetric) plays a role that is analogous to that of a magnetic field.

The Hamiltonian in the presence of the magnetic field is

$$\mathcal{E}(\boldsymbol{s}) = -B\sum_{i=1}^{N} s_i + \mathcal{E}_I(\boldsymbol{s}) \tag{450}$$

where $\mathcal{E}_I(\boldsymbol{s})$ stands for the interaction energy, previously modeled to be $\mathcal{N}(0, \frac{1}{2}nJ^2)$ according to the REM. Thus, the partition function is now

$$\begin{aligned}
Z(\beta, B) &= \sum_{\boldsymbol{s}} e^{-\beta\mathcal{E}(\boldsymbol{s})} \\
&= \sum_{\boldsymbol{s}} e^{-\beta\mathcal{E}_I(\boldsymbol{s})+\beta B\sum_{i=1}^{N} s_i} \\
&= \sum_{\boldsymbol{s}} e^{-\beta\mathcal{E}_I(\boldsymbol{s})+n\beta Bm(\boldsymbol{s})} \\
&= \sum_{m}\left[\sum_{\boldsymbol{s}:\ m(\boldsymbol{s})=m} e^{-\beta\mathcal{E}_I(\boldsymbol{s})}\right]\cdot e^{+N\beta Bm} \\
&\stackrel{\Delta}{=} \sum_{m} Z_0(\beta, m)\cdot e^{+N\beta Bm}
\end{aligned} \tag{451}$$

where we have introduced the notation $m(\boldsymbol{s}) = \frac{1}{N}\sum_i s_i$ and where $Z_0(\beta, m)$ is defined to be the expression in the square brackets in the second to the last line.[31] Now, observe that $Z_0(\beta, m)$ is just like the partition function of the REM without magnetic field, except that it has a smaller number of configurations – only those with magnetization m, namely, about $\exp\{Nh_2((1+m)/2)\}$ configurations. Thus, the analysis of $Z_0(\beta, m)$ is precisely the same as in the REM except that every occurrence of the term $\ln 2$ should be replaced by $h_2((1+m)/2)$. Accordingly,

$$Z_0(\beta, m) \doteq e^{N\psi(\beta,m)} \tag{452}$$

[31]Note that the relation between $Z_0(\beta, m)$ to $Z(\beta, B)$ is similar to the relation between $\Omega(E)$ of the microcanonical ensemble to $Z(\beta)$ of the canonical one (a Legendre relation in the log domain): we are replacing the fixed magnetization m, which is an extensive quantity, by an intensive variable B that controls its average.

with

$$\psi(\beta,m) = \max_{|\epsilon|\le J\sqrt{h_2((1+m)/2)}} \left[h_2\left(\frac{1+m}{2}\right) - \left(\frac{\epsilon}{J}\right)^2 - \beta\epsilon\right]$$
$$= \begin{cases} h_2\left(\frac{1+m}{2}\right) + \frac{\beta^2 J^2}{4} & \beta \le \beta_m \stackrel{\Delta}{=} \frac{2}{J}\sqrt{h_2\left(\frac{1+m}{2}\right)} \\ \beta J\sqrt{h_2\left(\frac{1+m}{2}\right)} & \beta > \beta_m \end{cases}$$

and from the above relation between Z and Z_0, we readily have the Legendre relation

$$\phi(\beta,B) = \max_m[\psi(\beta,m) + \beta m B]. \tag{453}$$

For small β (high temperature), the maximizing (dominant) m is attained with zero–derivative:

$$\frac{\partial}{\partial m}\left[h_2\left(\frac{1+m}{2}\right) + \frac{\beta^2 J^2}{4} + \beta m B\right] = 0 \tag{454}$$

that is

$$\frac{1}{2}\ln\frac{1-m}{1+m} + \beta B = 0 \tag{455}$$

which yields

$$m^* = m_p(\beta,B) \stackrel{\Delta}{=} \tanh(\beta B) \tag{456}$$

which is exactly the paramagnetic characteristic of magnetization vs. magnetic field (like that of i.i.d. spins), hence the name "paramagnetic phase." Thus, on substituting $m^* = \tanh(\beta B)$ back into the expression of ϕ, we get:

$$\phi(\beta,B) = h_2\left(\frac{1+\tanh(\beta B)}{2}\right) + \frac{\beta^2 J^2}{4} + \beta B\tanh(\beta B). \tag{457}$$

This solution is valid as long as the condition

$$\beta \le \beta_{m^*} = \frac{2}{J}\sqrt{h_2\left(\frac{1+\tanh(\beta B)}{2}\right)} \tag{458}$$

holds, or equivalently, the condition

$$\frac{\beta^2 J^2}{4} \le h_2\left(\frac{1+\tanh(\beta B)}{2}\right). \tag{459}$$

Now, let us denote by $\beta_c(B)$ the solution β to the equation:

$$\frac{\beta^2 J^2}{4} = h_2\left(\frac{1+\tanh(\beta B)}{2}\right). \tag{460}$$

As can be seen from the graphical illustration (Fig. 20), $\beta_c(B)$ is a decreasing function and hence $T_c(B) \stackrel{\Delta}{=} 1/\beta_c(B)$ is increasing. Thus, the phase transition temperature is increasing with $|B|$ (see Fig. 21). Below $\beta = \beta_c(B)$, we are in the glassy phase, where ϕ is given by:

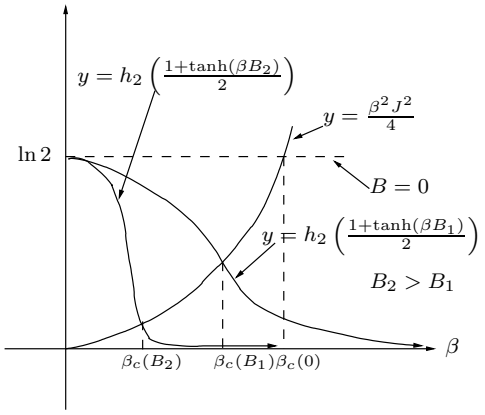


Figure 20: Graphical presentation of the solution $\beta_c(B)$ to the equation $\frac{1}{4}\beta^2 J^2 = h_2((1 + \tanh(\beta B))/2)$ for various values of B.

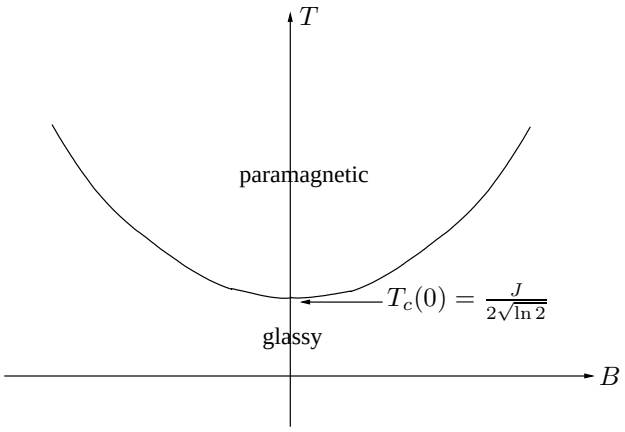


Figure 21: Phase diagram in the B–T plane.

$$\begin{aligned}\phi(\beta, B) &= \max_m \left[\beta J\sqrt{h_2\left(\frac{1+m}{2}\right)} + \beta m B\right] \\ &= \beta \cdot \max_m \left[J\sqrt{h_2\left(\frac{1+m}{2}\right)} + mB\right] \end{aligned} \tag{461}$$

thus, the maximizing m does not depend on β, only on B. On the other hand, it should be

the same solution that we get on the boundary $\beta = \beta_c(B)$, and so, it must be:

$$m^* = m_g(B) \stackrel{\Delta}{=} \tanh(B\beta_c(B)). \tag{462}$$

Thus, in summary

$$\phi(\beta, B) = \begin{cases} h_2\left(\frac{1+m_p(\beta,B)}{2}\right) + \frac{\beta^2 J^2}{4} + \beta B m_p(\beta, B) & \beta \leq \beta_c(B) \\ \beta J \sqrt{h_2\left(\frac{1+m_g(B)}{2}\right)} + \beta B m_g(B) & \beta > \beta_c(B) \end{cases}$$

In both phases $B \to 0$ implies $m^* \to 0$, therefore the REM does not exhibit spontaneous magnetization, only a glass transition, as described in the previous chapter.

Finally, we mention an important parameter in the physics of magnetic materials – the weak–field *magnetic susceptibility*, which is defined as $\chi \stackrel{\Delta}{=} \frac{\partial m^*}{\partial B}|_{B=0}$. It can readily be shown that in the REM case

$$\chi = \begin{cases} \frac{1}{T} & T \geq T_c(0) \\ \frac{1}{T_c(0)} & T < T_c(0) \end{cases} \tag{463}$$

The graphical illustration of this function is depicted in Fig. 22. The $1/T$ behavior for high temperature is known as *Curie's law*. As we heat a magnetic material up, it becomes more and more difficult to magnetize. The fact that here χ has an upper limit of $1/T_c(0)$ follows from the random interactions between spins, which make the magnetization more difficult too.

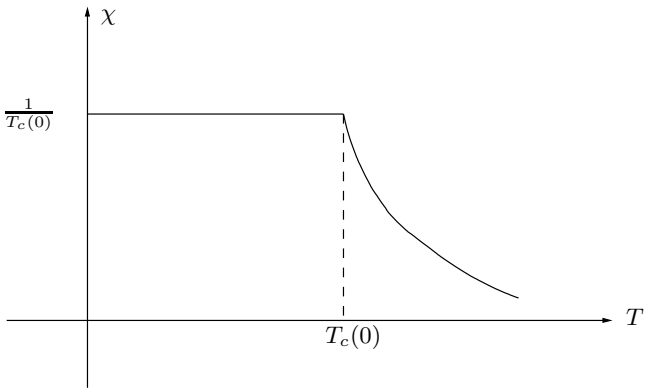


Figure 22: Magnetic susceptibility vs. temperature for a typical realization of the REM.

7.1.2 Relation to Joint Source–Channel Coding

We now relate these derivations to the behavior of joint source–channel coding systems. This subsection is a summary of the derivations and the results of [68].

Consider again our coded communication system with a few slight modifications (cf. Fig. 23). Instead of having e^{nR} equiprobable messages for channel coding, we are now in a joint source–channel coding scenario, where the message probabilities are skewed by the source probability distribution, which may not be symmetric. In particular, we consider the following: Suppose we have a vector $\boldsymbol{s} \in \{-1,+1\}^N$ emitted from a binary memoryless source with symbol probabilities $q = \Pr\{S_i = +1\} = 1 - \Pr\{S_i = -1\}$. The channel is still a BSC with crossover p. For every N–tuple emitted by the source, the channel conveys n channel binary symbols, which are the components of a codeword $\boldsymbol{x} \in \{0,1\}^n$, such that the ratio $\theta = n/N$, the *bandwidth expansion factor*, remains fixed. The mapping from $\boldsymbol{s}$ to $\boldsymbol{x}$ is the encoder. As before, we shall concern ourselves with random codes, namely, for every $\boldsymbol{s} \in \{-1,+1\}^N$, we randomly select an independent code-vector $\boldsymbol{x}(\boldsymbol{s}) \in \{0,1\}^n$ by fair coin tossing, as before. Thus, we randomly select 2^N code-vectors, each one of length $n = N\theta$. As in the case of pure channel coding, we consider the finite–temperature posterior:

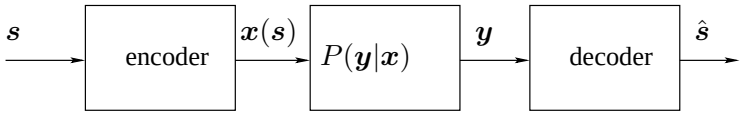


Figure 23: Block diagram of joint source–channel communication system.

$$P_\beta(\boldsymbol{s}|\boldsymbol{y}) = \frac{[P(\boldsymbol{s})P(\boldsymbol{y}|\boldsymbol{x}(\boldsymbol{s}))]^\beta}{Z(\beta|\boldsymbol{y})} \tag{464}$$

with

$$Z(\beta|\boldsymbol{y}) = \sum_{\boldsymbol{s}} [P(\boldsymbol{s})P(\boldsymbol{y}|\boldsymbol{x}(\boldsymbol{s}))]^\beta, \tag{465}$$

corresponding to the finite–temperature decoder:

$$\hat{s}_i = \arg\max_{s=\pm 1} \sum_{\boldsymbol{s}:\ s_i = s} [P(\boldsymbol{s})P(\boldsymbol{y}|\boldsymbol{x}(\boldsymbol{s}))]^\beta. \tag{466}$$

Once again, we separate the contributions of

$$Z_c(\beta|\boldsymbol{y}) = [P(\boldsymbol{s}_0)P(\boldsymbol{y}|\boldsymbol{x}(\boldsymbol{s}_0))]^\beta, \tag{467}$$

$\boldsymbol{s}_0$ being the true source message, and

$$Z_e(\beta|\boldsymbol{y}) = \sum_{\boldsymbol{s}\neq\boldsymbol{s}_0} [P(\boldsymbol{s})P(\boldsymbol{y}|\boldsymbol{x}(\boldsymbol{s}))]^\beta. \tag{468}$$

As we shall see quite shortly, Z_e behaves like the REM in a magnetic field given by

$$B = \frac{1}{2}\ln\frac{q}{1-q}. \tag{469}$$

Accordingly, we will henceforth denote $Z_e(\beta)$ also by $Z_e(\beta, B)$, to emphasize the analogy to the REM in a magnetic field.

To see that $Z_e(\beta, B)$ behaves like the REM in a magnetic field, consider the following: first, denote by $N_1(\boldsymbol{s})$ the number of $+1$'s in $\boldsymbol{s}$, so that the magnetization,

$$m(\boldsymbol{s}) \stackrel{\Delta}{=} \frac{1}{N}\left[\sum_{i=1}^N 1\{s_i = +1\} - \sum_{i=1}^N 1\{s_i = -1\}\right], \tag{470}$$

pertaining to spin configuration $\boldsymbol{s}$, is given by

$$m(\boldsymbol{s}) = \frac{2N_1(\boldsymbol{s})}{N} - 1. \tag{471}$$

Equivalently,

$$N_1(\boldsymbol{s}) = \frac{N[1+m(\boldsymbol{s})]}{2}, \tag{472}$$

and then

$$\begin{aligned} P(\boldsymbol{s}) &= q^{N_1(\boldsymbol{s})}(1-q)^{N-N_1(\boldsymbol{s})} \\ &= (1-q)^N\left(\frac{q}{1-q}\right)^{N(1+m(\boldsymbol{s}))/2} \\ &= [q(1-q)]^{N/2}\left(\frac{q}{1-q}\right)^{Nm(\boldsymbol{s}))/2} \\ &= [q(1-q)]^{N/2}e^{Nm(\boldsymbol{s})B} \end{aligned} \tag{473}$$

where B is defined as above. By the same token, for the binary symmetric channel we have:

$$P(\boldsymbol{y}|\boldsymbol{x}) = p^{d_H(\boldsymbol{x},\boldsymbol{y})}(1-p)^{n-d_H(\boldsymbol{x},\boldsymbol{y})} = (1-p)^n e^{-Jd_H(\boldsymbol{x},\boldsymbol{y})} \tag{474}$$

where $J = \ln\frac{1-p}{p}$ and $d_H(\boldsymbol{x},\boldsymbol{y})$ is the Hamming distance, as defined earlier. Thus,

$$\begin{aligned} Z_e(\beta,B) &= [q(1-q)]^{N\beta/2}\times \\ &\quad \sum_m \left[\sum_{\boldsymbol{x}(\boldsymbol{s}):\ m(\boldsymbol{s})=m} e^{-\beta\ln[1/P(\boldsymbol{y}|\boldsymbol{x}(\boldsymbol{s}))]}\right] e^{N\beta mB} \\ &= [q(1-q)]^{\beta N/2}(1-p)^{n\beta}\times \\ &\quad \sum_m \left[\sum_{\boldsymbol{x}(\boldsymbol{s}):\ m(\boldsymbol{s})=m} e^{-\beta J d_H(\boldsymbol{x}(\boldsymbol{s}),\boldsymbol{y})}\right] e^{\beta NmB} \\ &\stackrel{\Delta}{=} [q(1-q)]^{N\beta/2}(1-p)^{n\beta}\sum_m Z_0(\beta,m|\boldsymbol{y})e^{\beta NmB} \end{aligned} \tag{475}$$

The resemblance to the REM in a magnetic field is now self–evident. In analogy to the above analysis of the REM, $Z_0(\beta,m)$ here behaves like in the REM without a magnetic field, namely, it contains exponentially $e^{Nh((1+m)/2)} = e^{nh((1+m)/2)/\theta}$ terms, with the random energy levels of the REM being replaced now by random Hamming distances $\{d_H(\boldsymbol{x}(\boldsymbol{s}),\boldsymbol{y})\}$ that are induced by the random selection of the code $\{\boldsymbol{x}(\boldsymbol{s})\}$. Using the same considerations as with the REM in channel coding, we now get:

$$\begin{aligned} \psi(\beta,m) &\stackrel{\Delta}{=} \lim_{n\to\infty}\frac{\ln Z_0(\beta,m|\boldsymbol{y})}{n} \\ &= \max_{\delta_m\le\delta\le 1-\delta_m}\left[\frac{1}{\theta}h_2\left(\frac{1+m}{2}\right)+h_2(\delta)-\ln 2-\beta J\delta\right] \\ &= \begin{cases} \frac{1}{\theta}h_2\left(\frac{1+m}{2}\right)+h_2(p_\beta)-\ln 2-\beta J p_\beta & p_\beta\ge\delta_m \\ -\beta J\delta_m & p_\beta<\delta_m \end{cases} \end{aligned} \tag{476}$$

where we have defined

$$\delta_m \stackrel{\Delta}{=} \delta_{GV}\left(\frac{1}{\theta}h_2\left(\frac{1+m}{2}\right)\right) \tag{477}$$

and where again,

$$p_\beta = \frac{p^\beta}{p^\beta+(1-p)^\beta}. \tag{478}$$

The condition $p_\beta \geq \delta_m$ is equivalent to

$$\beta \leq \beta_0(m) \triangleq \frac{1}{J} \ln \frac{1-\delta_m}{\delta_m}. \tag{479}$$

Finally, back to the full partition function:

$$\begin{aligned} \phi(\beta, B) &= \lim_{n\to\infty} \frac{1}{N} \ln \left[\sum_m Z_0(\beta, m|\boldsymbol{y}) e^{N\beta B m} \right] \\ &= \max_m [\theta\psi(\beta, m) + \beta m B]. \end{aligned} \tag{480}$$

For small enough β, the dominant m is the one that maximizes

$$h_2\left(\frac{1+m}{2}\right) + \beta m B,$$

which is again the paramagnetic magnetization

$$m^* = m_p(\beta, B) = \tanh(\beta B). \tag{481}$$

Thus, in high decoding temperatures, the source vectors $\{\boldsymbol{s}\}$ that dominate the posterior $P_\beta(\boldsymbol{s}|\boldsymbol{y})$ behave like a paramagnet under a magnetic field defined by the prior $B = \frac{1}{2}\ln\frac{q}{1-q}$. In the glassy regime, similarly as before, we get:

$$m^* = m_g(B) \triangleq \tanh(B\beta_c(B)) \tag{482}$$

where this time, $\beta_c(B)$, the glassy–paramagnetic boundary, is defined as the solution to the equation

$$\ln 2 - h_2(p_\beta) = \frac{1}{\theta} h_2\left(\frac{1+\tanh(\beta B)}{2}\right). \tag{483}$$

The full details are in [68]. Taking now into account also Z_c, we get a phase diagram as depicted in Fig. 24. Here,

$$B_0 \triangleq \frac{1}{2} \ln \frac{q^*}{1-q^*} \tag{484}$$

where q^* is the solution to the equation

$$h_2(q) = \theta[\ln 2 - h_2(p)], \tag{485}$$

namely, it is the boundary between reliable and unreliable communication.

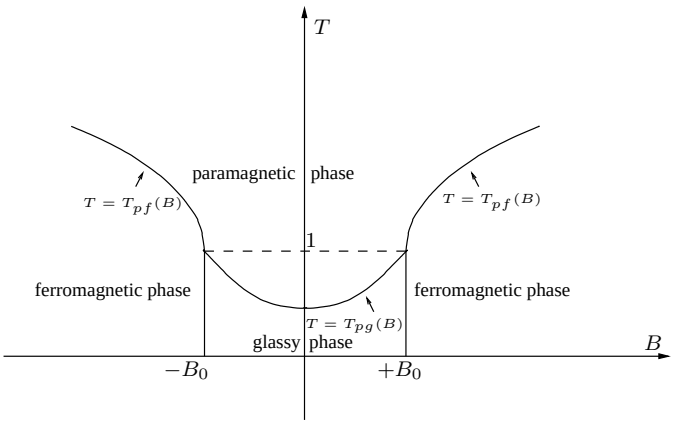


Figure 24: Phase diagram of joint source–channel communication system.

7.2 Generalized REM (GREM) and Hierarchical Coding

In the mid–eighties of the previous century, Derrida extended the REM to the generalized REM (GREM) [25], [26], which has an hierarchical tree structure to accommodate possible correlations between energy levels of various configurations (and hence is somewhat closer to reality). It turns out to have direct relevance to performance analysis of codes with a parallel hierarchical structure. Hierarchical structured codes are frequently encountered in many contexts, e.g., tree codes, multi–stage codes for progressive coding and successive refinement, codes for the degraded broadcast channel, codes with a binning structure (like in G–P and W–Z coding and coding for the wiretap channel), and so on. The material in this subsection is based on [70].

We begin from the physics of the GREM. For simplicity, we limit ourselves to two stages, but the discussion and the results extend to any fixed, finite number of stages. The GREM is defined by a few parameters: (i) a number $0 < R_1 < \ln 2$ and $R_2 = \ln 2 - R_1$. (ii) a number $0 < a_1 < 1$ and $a_2 = 1 - a_1$. Given these parameters, we now partition the set of 2^N configurations into e^{NR_1} groups, each having e^{NR_2} configurations.[32] The easiest way to describe it is with a tree (see Fig. 25), each leaf of which represents one spin configuration. Now, for each branch in this tree, we randomly draw an independent random variable, which

[32]Later, we will see that in the analogy to hierarchical codes, R_1 and R_2 will have the meaning of coding rates at two stages of a two–stage code.

will be referred to as an *energy component*: First, for every branch outgoing from the root, we randomly draw $\epsilon_i \sim \mathcal{N}(0, a_1 NJ^2/2)$, $1 \leq i \leq e^{NR_1}$. Then, for each branch $1 \leq j \leq e^{NR_2}$, emanating from node no. i, $1 \leq i \leq e^{NR_1}$, we randomly draw $\epsilon_{i,j} \sim \mathcal{N}(0, a_2 NJ^2/2)$. Finally, we define the energy associated with each configuration, or equivalently, each leaf indexed by (i,j), as $E_{i,j} = \epsilon_i + \epsilon_{i,j}$, $1 \leq i \leq e^{NR_1}$, $1 \leq j \leq e^{NR_2}$.

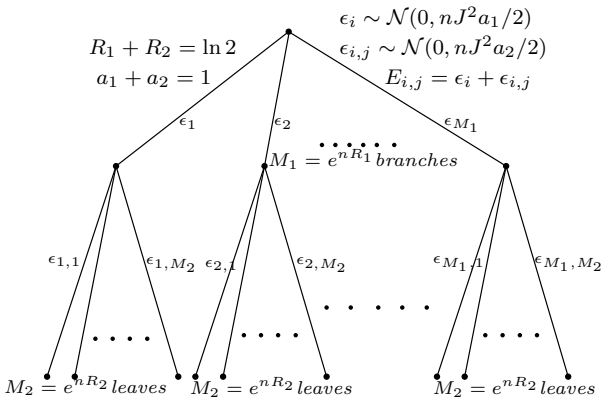


Figure 25: The GREM with $K = 2$ stages.

Obviously, the marginal pdf of each $E_{i,j}$ is $\mathcal{N}(0, NJ^2/2)$, just like in the ordinary REM. However, unlike in the ordinary REM, here the configurational energies $\{E_{i,j}\}$ are correlated: Every two leaves with a common parent node i have an energy component ϵ_i in common and hence their total energies are correlated.

An extension of the GREM to K stages is parametrized by $\sum_{\ell=1}^{K} R_\ell = \ln 2$ and $\sum_{\ell=1}^{K} a_\ell = 1$, where one first divides the entirety of 2^n configurations into e^{NR_1} groups, then each such group is subdivided into e^{nR_2} subgroups, and so on. For each branch of generation no. ℓ, an independent energy component is drawn according to $\mathcal{N}(0, a_\ell NJ^2/2)$ and the total energy pertaining to each configuration, or a leaf, is the sum of energy components along the path from the root to that leaf. An extreme case of the GREM is where $K = N$, which is referred to as the *directed polymer on a tree* or a *directed polymer in a random medium*. We will it later in Subsection 7.3, although it has a different asymptotic regime than the GREM, because in the GREM, K is assumed fixed while N grows without bound in the thermodynamic limit.

Returning back to the case of $K = 2$ stages, the analysis of the GREM is conceptually a simple extension of that of the REM: First, we ask ourselves what is the typical number of branches emanating from the root whose first–generation energy component, ϵ_i, is about ϵ? The answer is very similar to that of the REM: Since we have e^{NR_1} independent trials of an experiment for which the probability of a single success is exponentially $e^{-\epsilon^2/(NJ^2a_1)}$, then for a typical realization:

$$\Omega_1(\epsilon) \approx \begin{cases} 0 & |\epsilon| > NJ\sqrt{a_1 R_1} \\ \exp\left\{N\left[R_1 - \frac{1}{a_1}\left(\frac{\epsilon}{NJ}\right)^2\right]\right\} & |\epsilon| < NJ\sqrt{a_1 R_1} \end{cases} \tag{486}$$

Next, we ask ourselves what is the typical number $\Omega_2(E)$ of configurations with total energy about E? Obviously, each such configuration should have a first–generation energy component ϵ and second–generation energy component $E - \epsilon$, for some ϵ. Thus,

$$\Omega_2(\epsilon) \approx \int_{-NJ\sqrt{a_1 R_1}}^{+NJ\sqrt{a_1 R_1}} \mathrm{d}\epsilon \Omega_1(\epsilon) \cdot \exp\left\{N\left[R_2 - \frac{1}{a_2}\left(\frac{E-\epsilon}{NJ}\right)^2\right]\right\}. \tag{487}$$

It is important to understand here the following point: Here, we *no longer* zero–out the factor

$$\exp\left\{N\left[R_2 - \frac{1}{a_2}\left(\frac{E-\epsilon}{NJ}\right)^2\right]\right\} \tag{488}$$

when the expression in the square brackets at the exponent becomes negative, as we did in the first stage and in the REM. The reason is simple: Given ϵ, we are conducting $\Omega_1(\epsilon) \cdot e^{NR_1}$ independent trials of an experiment whose success rate is

$$\exp\left\{-\frac{N}{a_2}\left(\frac{E-\epsilon}{NJ}\right)^2\right\}. \tag{489}$$

Thus, whatever counts is whether the *entire* integrand has a positive exponent or not.

Consider next the entropy. The entropy behaves as follows:

$$\Sigma(E) = \lim_{N\to\infty} \frac{\ln \Omega_2(E)}{N} = \begin{cases} \Sigma_0(E) & \Sigma_0(E) \geq 0 \\ -\infty & \Sigma_0(E) < 0 \end{cases} \tag{490}$$

where $\Sigma_0(E)$ is the exponential rate of the above integral, which after applying the Laplace method, is shown to be:

$$\Sigma_0(E) = \max_{|\epsilon| \leq +NJ\sqrt{a_1 R_1}} \left[R_1 - \frac{1}{a_1}\left(\frac{\epsilon}{NJ}\right)^2 + R_2 - \frac{1}{a_2}\left(\frac{E-\epsilon}{NJ}\right)^2\right].$$

How does the function $\Sigma(E)$ behave?

It turns out that to answer this question, we will have to distinguish between two cases: (i) $R_1/a_1 < R_2/a_2$ and (ii) $R_1/a_1 \geq R_2/a_2$.[33] First, observe that $\Sigma_0(E)$ is an even function, i.e., it depends on E only via $|E|$, and it is monotonically non–increasing in $|E|$. Solving the optimization problem pertaining to Σ_0, we readily find:

$$\Sigma_0(E) = \begin{cases} \ln 2 - \left(\frac{E}{NJ}\right)^2 & |E| \leq E_1 \\ R_2 - \frac{1}{a_2}\left(\frac{E}{NJ} - \sqrt{a_1 R_1}\right)^2 & |E| > E_1 \end{cases}$$

where $E_1 \stackrel{\Delta}{=} NJ\sqrt{R_1/a_1}$. This is a phase transition due to the fact that the maximizing ϵ becomes an edge-point of its allowed interval. Imagine now that we gradually increase $|E|$ from zero upward. Now the question is what is encountered first: The energy level $\hat{E}$, where $\Sigma(E)$ jumps to $-\infty$, or E_1 where this phase transition happens? In other words, is $\hat{E} < E_1$ or $\hat{E} > E_1$? In the former case, the phase transition at E_1 will not be apparent because $\Sigma(E)$ jumps to $-\infty$ before, and that's it. In this case, according to the first line of $\Sigma_0(E)$, $\ln 2 - (E/NJ)^2$ vanishes at $\hat{E} = NJ\sqrt{\ln 2}$ and we get:

$$\Sigma(E) = \begin{cases} \ln 2 - \left(\frac{E}{NJ}\right)^2 & |E| \leq \hat{E} \\ -\infty & |E| > \hat{E} \end{cases} \tag{491}$$

exactly like in the ordinary REM. It follows then that in this case, $\phi(\beta)$ which is the Legendre transform of $\Sigma(E)$ will also be like in the ordinary REM, that is:

$$\phi(\beta) = \begin{cases} \ln 2 + \frac{\beta^2 J^2}{4} & \beta \leq \beta_0 \stackrel{\Delta}{=} \frac{2}{J}\sqrt{\ln 2} \\ \beta J\sqrt{\ln 2} & \beta > \beta_0 \end{cases} \tag{492}$$

As said, the condition for this is:

$$NJ\sqrt{\ln 2} \equiv \hat{E} \leq E_1 \equiv NJ\sqrt{\frac{R_1}{a_1}} \tag{493}$$

or, equivalently,

$$\frac{R_1}{a_1} \geq \ln 2. \tag{494}$$

[33] Accordingly, in coding, this will mean a distinction between two cases of the relative coding rates at the two stages.

On the other hand, in the opposite case, $\hat{E} > E_1$, the phase transition at E_1 is apparent, and so, there are now *two* phase transitions:

$$\Sigma(E) = \begin{cases} \ln 2 - \left(\frac{E}{NJ}\right)^2 & |E| \leq E_1 \\ R_2 - \frac{1}{a_2}\left(\frac{E}{NJ} - \sqrt{a_1 R_1}\right)^2 & E_1 < |E| \leq \hat{E} \\ -\infty & |E| > \hat{E} \end{cases} \tag{495}$$

and accordingly:

$$\phi(\beta) = \begin{cases} \ln 2 + \frac{\beta^2 J^2}{4} & \beta \leq \beta_1 \stackrel{\Delta}{=} \frac{2}{J}\sqrt{\frac{R_1}{a_1}} \\ \beta J\sqrt{a_1 R_1} + R_2 + \frac{a_2 \beta^2 J^2}{4} & \beta_1 \leq \beta < \beta_2 \stackrel{\Delta}{=} \frac{2}{J}\sqrt{\frac{R_2}{a_2}} \\ \beta J(\sqrt{a_1 R_1} + \sqrt{a_2 R_2}) & \beta \geq \beta_2 \end{cases} \tag{496}$$

The first line is a purely paramagnetic phase. In the second line, the first–generation branches are glassy (there is a subexponential number of dominant ones) but the second–generation is still paramagnetic. In the third line, both generations are glassy, i.e., a subexponential number of dominant first–level branches, each followed by a subexponential number of second–level ones, thus a total of a subexponential number of dominant configurations overall.

Now, there is a small technical question: what is it that guarantees that $\beta_1 < \beta_2$ whenever $R_1/a_1 < \ln 2$? We now argue that these two inequalities are, in fact, equivalent. In [12], the following inequality is proved for two positive vectors $(a_1, \ldots, a_k)$ and $(b_1, \ldots, b_k)$:

$$\min_i \frac{a_i}{b_i} \leq \frac{\sum_{i=1}^k a_i}{\sum_{i=1}^k b_i} \leq \max_i \frac{a_i}{b_i}. \tag{497}$$

Thus,

$$\min_{i \in \{1,2\}} \frac{R_i}{a_i} \leq \frac{R_1 + R_2}{a_1 + a_2} \leq \max_{i \in \{1,2\}} \frac{R_i}{a_i}, \tag{498}$$

but in the middle expression the numerator is $R_1 + R_2 = \ln 2$ and the denominator is $a_1 + a_2 = 1$, thus it is exactly $\ln 2$. In other words, $\ln 2$ is always in between R_1/a_1 and R_2/a_2. So $R_1/a_1 < \ln 2$ iff $R_1/a_1 < R_2/a_2$, which is the case where $\beta_1 < \beta_2$. To summarize our findings thus far, we have the following:

Case A: $R_1/a_1 < R_2/a_2$ – two phase transitions:

$$\phi(\beta) = \begin{cases} \ln 2 + \frac{\beta^2 J^2}{4} & \beta \le \beta_1 \\ \beta J\sqrt{a_1 R_1} + R_2 + \frac{a_2 \beta^2 J^2}{4} & \beta_1 \le \beta < \beta_2 \\ \beta J(\sqrt{a_1 R_1} + \sqrt{a_2 R_2}) & \beta \ge \beta_2 \end{cases} \tag{499}$$

Case B: $R_1/a_1 \ge R_2/a_2$ – one phase transition, like in the REM:

$$\phi(\beta) = \begin{cases} \ln 2 + \frac{\beta^2 J^2}{4} & \beta \le \beta_0 \\ \beta J\sqrt{\ln 2} & \beta > \beta_0 \end{cases} \tag{500}$$

We now move on to our coding problem, this time, it is about source coding with a fidelity criterion. For simplicity, we assume a binary symmetric source (BSS) and the Hamming distortion measure. Consider the following hierarchical structure of a code: Given a block length n, we break it into two segments of lengths n_1 and $n_2 = n - n_1$. For the first segment, we randomly select (by fair coin tossing) a codebook $\hat{\mathcal{C}} = \{\hat{\boldsymbol{x}}_i,\ 1 \le i \le e^{n_1 R_1}\}$. For the second segment, we do the following: For each $1 \le i \le e^{n_1 R_1}$, we randomly select (again, by fair coin tossing) a codebook $\tilde{\mathcal{C}}_i = \{\tilde{\boldsymbol{x}}_{i,j},\ 1 \le j \le e^{n_2 R_2}\}$. Now, given a source vector $\boldsymbol{x} \in \{0,1\}^n$, segmentized as $(\boldsymbol{x}', \boldsymbol{x}'')$, the encoder seeks a pair (i,j), $1 \le i \le e^{n_1 R_1}$, $1 \le j \le e^{n_2 R_2}$, such that $d(\boldsymbol{x}', \hat{\boldsymbol{x}}_i) + d(\boldsymbol{x}'', \tilde{\boldsymbol{x}}_{i,j})$ is minimum, and then transmits i using $n_1 R_1$ nats and j – using $n_2 R_2$ nats, thus a total of $(n_1 R_1 + n_2 R_2)$ nats, which means an average rate of $R = \lambda R_1 + (1-\lambda) R_2$ nats per symbol, where $\lambda = n_1/n$. Now, there are a few questions that naturally arise:

What is the motivation for codes of this structure? The decoder has a reduced delay. It can decode the first n_1 symbols after having received the first $n_1 R_1$ nats, and does not have to wait until the entire transmission of length $(n_1 R_1 + n_2 R_2)$ has been received. Extending this idea to K even segments of length n/K, the decoding delay is reduced from n to n/K. In the limit of $K = n$, in which case it is a tree code, the decoder is actually delayless.

What is the relation to the GREM? The hierarchical structure of the code is that of a tree, exactly like the GREM. The role of the energy components at each branch is now played by the segmental distortions $d(\boldsymbol{x}', \hat{\boldsymbol{x}}_i)$ and $d(\boldsymbol{x}'', \tilde{\boldsymbol{x}}_{i,j})$. The parameters R_1 and R_2 here are similar to those of the GREM.

Given an overall rate R, suppose we have the freedom to choose λ, R_1 and R_2, such that $R = \lambda R_1 + (1-\lambda)R_2$, are some choice better than others in some sense? This is exactly what we are going to figure out next.

As for the performance criterion, here, we choose to examine performance in terms of the characteristic function of the overall distortion,

$$\boldsymbol{E}[\exp\{-s \cdot \text{distortion}\}].$$

This is, of course, a much more informative figure of merit than the average distortion, because in principle, it gives information on the *entire probability distribution* of the distortion. In particular, it generates all the moments of the distortion by taking derivatives at $s = 0$, and it is useful in deriving Chernoff bounds on probabilities of large deviations events concerning the distortion. More formally, we make the following definitions: Given a code $\mathcal{C}$ (any block code, not necessarily of the class we defined), and a source vector $\boldsymbol{x}$, we define

$$\Delta(\boldsymbol{x}) = \min_{\hat{\boldsymbol{x}} \in \mathcal{C}} d(\boldsymbol{x}, \hat{\boldsymbol{x}}), \tag{501}$$

and we will be interested in the exponential rate of

$$\Psi(s) \stackrel{\Delta}{=} \boldsymbol{E}\{\exp[-s\Delta(\boldsymbol{X})]\}. \tag{502}$$

This quantity can be easily related to the "partition function":

$$Z(\beta|\boldsymbol{x}) \stackrel{\Delta}{=} \sum_{\hat{\boldsymbol{x}} \in \mathcal{C}} e^{-\beta d(\boldsymbol{x}, \hat{\boldsymbol{x}})}. \tag{503}$$

In particular,

$$\boldsymbol{E}\{\exp[-s\Delta(\boldsymbol{X})]\} = \lim_{\theta \to \infty} \boldsymbol{E}\left\{[Z(s \cdot \theta|\boldsymbol{X})]^{1/\theta}\right\}. \tag{504}$$

Thus, to analyze the characteristic function of the distortion, we have to assess (non-integer) moments of the partition function.

Let us first see what happens with ordinary random block codes, without any structure. This calculation is very similar the one we did earlier in the context of channel coding:

$$\boldsymbol{E}\left\{[Z(s \cdot \theta|\boldsymbol{X})]^{1/\theta}\right\} = \boldsymbol{E}\left\{\left[\sum_{\hat{\boldsymbol{x}} \in \mathcal{C}} e^{-s\theta d(\boldsymbol{x}, \hat{\boldsymbol{x}})}\right]^{1/\theta}\right\}$$

$$= \boldsymbol{E}\left\{\left[\sum_{d=0}^{n}\Omega(d)e^{-s\theta d}\right]^{1/\theta}\right\}$$
$$\doteq \sum_{d=0}^{n}\boldsymbol{E}\left\{[\Omega(d)]^{1/\theta}\right\}\cdot e^{-sd} \tag{505}$$

where, as we have already shown in the previous chapter that

$$\boldsymbol{E}\left\{[\Omega(d)]^{1/\theta}\right\} \doteq \begin{cases} e^{n[R+h_2(\delta)-\ln 2]} & \delta \le \delta_{GV}(R) \text{ or } \delta \ge 1-\delta_{GV}(R) \\ e^{n[R+h_2(\delta)-\ln 2]/\theta} & \delta_{GV}(R) \le \delta \le 1-\delta_{GV}(R) \end{cases}$$

Note that $\delta_{GV}(R)$ is exactly the distortion–rate function of the BSS w.r.t. the Hamming distortion. By substituting the expression of $\boldsymbol{E}\{[\Omega(d)]^{1/\theta}\}$ back into that of $\boldsymbol{E}\{[Z(s\cdot\theta|\boldsymbol{X})]^{1/\theta}\}$ and carrying out the maximization pertaining to the dominant contribution, we eventually obtain:

$$\Psi(s) \doteq e^{-nu(s,R)} \tag{506}$$

where

$$u(s,R) = \ln 2 - R - \max_{\delta\le\delta_{GV}(R)}[h_2(\delta)-s\delta]$$
$$= \begin{cases} s\delta_{GV}(R) & s \le s_R \\ v(s,R) & s > s_R \end{cases} \tag{507}$$

with

$$s_R \stackrel{\Delta}{=} \ln\left[\frac{1-\delta_{GV}(R)}{\delta_{GV}(R)}\right] \tag{508}$$

and

$$v(s,R) \stackrel{\Delta}{=} \ln 2 - R + s - \ln(1+e^s). \tag{509}$$

The function $u(s,R)$ is depicted qualitatively in Fig. 26.

Let us now move on to the hierarchical codes. The analogy with the GREM is fairly clear. Given $\boldsymbol{x}$, there are about $\Omega_1(\delta_1) \doteq e^{n_1[R_1+h_2(\delta_1)-\ln 2]}$ first–segment codewords $\{\hat{\boldsymbol{x}}_i\}$ in $\hat{\mathcal{C}}$ at distance $n_1\delta_1$ from the first segment $\boldsymbol{x}'$ of $\boldsymbol{x}$, provided that $R_1+h_2(\delta_1)-\ln 2 > 0$ and $\Omega_1(\delta_1) = 0$ otherwise. For each such first–segment codeword, there are about $e^{n_2[R_2+h_2(\delta_2)-\ln 2]}$ second–segment codewords $\{\tilde{\boldsymbol{x}}_{i,j}\}$ at distance $n_2\delta_2$ from the second segment $\boldsymbol{x}''$ of $\boldsymbol{x}$. Therefore, for

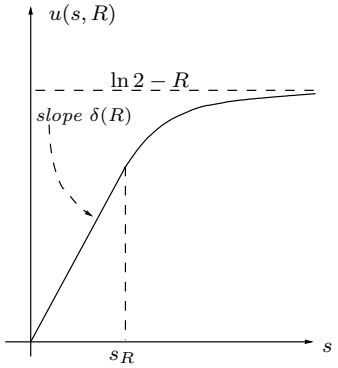


Figure 26: Qualitative graph of the function $u(s,R)$ as a function of s for fixed R.

$\delta = \lambda\delta_1 + (1-\lambda)\delta_2$,

$$\begin{aligned}\Omega_2(\delta) &= \sum_{\delta_1=\delta_{GV}(R_1)}^{1-\delta_{GV}(R_1)} e^{n_1[R_1+h_2(\delta_1)-\ln 2]} \cdot e^{n_2[R_2+h_2((\delta-\lambda\delta_1)/(1-\lambda))-\ln 2]} \\ &\doteq \exp\left\{n \cdot \max_{\delta_1\in[\delta_{GV}(R_1),1-\delta_{GV}(R_1)]} \times \right. \\ &\left.\left[R+\lambda h_2(\delta_1)+(1-\lambda)h_2\left(\frac{\delta-\lambda\delta_1}{1-\lambda}\right)\right]\right\} \end{aligned} \tag{510}$$

In analogy to the analysis of the GREM, here too, there is a distinction between two cases: $R_1 \geq R \geq R_2$ and $R_1 < R < R_2$. In the first case, the behavior is just like in the REM:

$$\Sigma(\delta) = \begin{cases} R + h_2(\delta) - \ln 2 & \delta \in [\delta_{GV}(R), 1-\delta_{GV}(R)] \\ -\infty & \text{elsewhere} \end{cases} \tag{511}$$

and then, of course, $\phi(\beta) = -u(\beta,R)$ behaves exactly like that of a general random code, in spite of the hierarchical structure. In the other case, we have two phase transitions:

$$\phi(\beta,R) = \begin{cases} -v(\beta,R) & \beta < \beta(R_1) \\ -\lambda\beta\delta_{GV}(R_1) - (1-\lambda)v(\beta,R_2) & \beta(R_1) < \beta < \beta(R_2) \\ -\beta[\lambda\delta_{GV}(R_1)+(1-\lambda)\delta_{GV}(R_2)] & \beta > \beta(R_2) \end{cases}$$

The last line is the purely glassy phase and this is the relevant phase because of the limit $\theta \to 0$ that we take in order to calculate $\Psi(s)$. Note that at this phase the slope is $\lambda\delta_{GV}(R_1)+(1-\lambda)\delta_{GV}(R_2)$ which means that code behaves as if the two segments were coded *separately*, which is worse that $\delta_{GV}(R)$ due to convexity arguments. Let us see this more concretely on the characteristic function: This time, it will prove convenient to define $\Omega(d_1,d_2)$ as an

enumerator of codewords whose distance is d_1 at the first segment and d_2 – on the second one. Now,

$$\boldsymbol{E}\left\{Z^{1/\theta}(s\cdot\theta)\right\} = \boldsymbol{E}\left\{\left[\sum_{d_1=0}^{n}\sum_{d_2=0}^{n}\Omega(d_1,d_2)\cdot e^{-s\theta(d_1+d_2)}\right]^{1/\theta}\right\}$$
$$\doteq \sum_{d_1=0}^{n}\sum_{d_2=0}^{n}\boldsymbol{E}\left\{\Omega^{1/\theta}(d_1,d_2)\right\}\cdot e^{-s(d_1+d_2)}. \tag{512}$$

Here, we should distinguish between four types of terms depending on whether or not $\delta_1 \in [\delta_{GV}(R_1), 1-\delta_{GV}(R_1)]$ and whether or not $\delta_2 \in [\delta_{GV}(R_2), 1-\delta_{GV}(R_2)]$. In each one of these combinations, the behavior is different (the details are in [70]). The final results are as follows:

For $R_1 < R_2$,

$$\lim_{n\to\infty}\left[-\frac{1}{n}\ln\boldsymbol{E}\exp\{-s\Delta(\boldsymbol{X})\}\right] = \lambda u(s,R_1) + (1-\lambda)u(s,R_2) \tag{513}$$

which means the behavior of two independent, decoupled codes for the two segments, which is bad, of course.

For $R_1 \geq R_2$,

$$\lim_{n\to\infty}\left[-\frac{1}{n}\ln\boldsymbol{E}\exp\{-s\Delta(\boldsymbol{X})\}\right] = u(s,R) \quad \forall s \leq s_0 \tag{514}$$

where s_0 is some positive constant. This means that the code behaves like an unstructured code (with delay) for all s up to a certain s_0 and the reduced decoding delay is obtained for free. Note that the domain of small s is relevant for moments of the distortion. For $R_1 = R_2$, s_0 is unlimited. Thus, the conclusion is that if we must work at different rates, it is better to use the higher rate first. A more elaborate discussion can be found in [70].

7.3 Directed Polymers in a Random Medium and Tree Codes

Finally, we discuss a related model that we mentioned earlier, which can be thought of as an extreme case of the GREM with $K = N$. This is the *directed polymer in a random medium* (DPRM): Consider a *Cayley tree*, namely, a full balanced tree with branching ratio

d and depth n (cf. Fig. 27, where $d = 2$ and $N = 3$). Let us index the branches by a pair of integers (i, j), where $1 \leq i \leq N$ describes the generation (with $i = 1$ corresponding to the d branches that emanate from the root), and $0 \leq j \leq d^i - 1$ enumerates the branches of the i–th generation, say, from left to right (again, see Fig. 27). For each branch (i, j), $1 \leq j \leq d^i$, $1 \leq i \leq N$, we randomly draw an independent random variable $\varepsilon_{i,j}$ according to a fixed probability function $q(\varepsilon)$ (i.e., a probability mass function in the discrete case, or probability density function in the continuous case). As explained earlier, the asymptotic regime here is different from that of the GREM: In the GREM we had a fixed number of stages K that did not grow with N and exponentially many branches emanating from each internal node. Here, we have $K = N$ and a fixed number d of branches outgoing from each note.

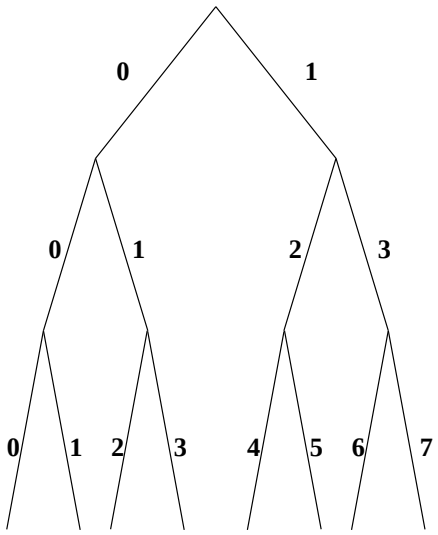


Figure 27: A Cayley tree with branching factor $d = 2$ and depth $N = 3$.

A *walk* $\boldsymbol{w}$, from the root of the tree to one of its leaves, is described by a finite sequence $\{(i, j_i)\}_{i=1}^N$, where $0 \leq j_1 \leq d - 1$ and $dj_i \leq j_{i+1} \leq dj_i + d - 1$, $i = 1, 2, \ldots, (N - 1)$.[34] For a given realization of the random variables $\{\varepsilon_{i,j} : \ i = 1, 2, \ldots, N, \ j = 0, 1, \ldots, d^i - 1\}$, we define the Hamiltonian associated with $\boldsymbol{w}$ as $\mathcal{E}(\boldsymbol{w}) = \sum_{i=1}^N \varepsilon_{i,j_i}$, and then the partition

[34] In fact, for a given N, the number j_N alone dictates the entire walk.

function as:

$$Z_N(\beta) = \sum_{\boldsymbol{w}} \exp\{-\beta \mathcal{E}(\boldsymbol{w})\}. \tag{515}$$

It turns out that this model is exactly solvable (in many ways) and one can show (see e.g., [9]) that it admits a glassy phase transition:

$$\phi(\beta) = \lim_{N\to\infty} \frac{\ln Z_N(\beta)}{n} = \begin{cases} \phi_0(\beta) & \beta < \beta_c \\ \phi_0(\beta_c) & \beta \geq \beta_c \end{cases} \qquad \text{almost surely} \tag{516}$$

where

$$\phi_0(\beta) \stackrel{\Delta}{=} \frac{\ln[d \cdot \boldsymbol{E} e^{-\beta\rho(\epsilon)}]}{\beta} \tag{517}$$

and β_c is the value of β that minimizes $\phi_0(\beta)$.

In analogy to the hierarchical codes inspired by the GREM, consider now an ensemble of tree codes for encoding source n–tuples, $\boldsymbol{x} = (x_1, \ldots, x_n)$, which is defined as follows: Given a coding rate R (in nats/source–symbol), which is assumed to be the natural logarithm of some positive integer d, and given a probability distribution on the reproduction alphabet, $q = \{q(y),\ y \in \mathcal{Y}\}$, let us draw $d = e^R$ independent copies of Y under q, and denote them by $Y_1, Y_2, \ldots, Y_d$. We shall refer to the randomly chosen set, $\mathcal{C}_1 = \{Y_1, Y_2, \ldots, Y_d\}$, as our 'codebook' for the first source symbol, X_1. Next, for each $1 \leq j_1 \leq d$, we randomly select another such codebook under q, $\mathcal{C}_{2,j_1} = \{Y_{j_1,1}, Y_{j_1,2}, \ldots, Y_{j_1,d}\}$, for the second symbol, X_2. Then, for each $1 \leq j_1 \leq d$ and $1 \leq j_2 \leq d$, we again draw under q yet another codebook $\mathcal{C}_{3,j_1,j_2} = \{Y_{j_1,j_2,1}, Y_{j_1,j_2,2}, \ldots, Y_{j_1,j_2,d}\}$, for X_3, and so on. In general, for each $t \leq n$, we randomly draw d^{t-1} codebooks under q, which are indexed by $(j_1, j_2, \ldots, j_{t-1})$, $1 \leq j_k \leq d$, $1 \leq k \leq t-1$.

Once the above described random code selection process is complete, the resulting set of codebooks $\{\mathcal{C}_1, \mathcal{C}_{t,j_1,\ldots,j_{t-1}},\ 2 \leq t \leq n,\ 1 \leq j_k \leq d,\ 1 \leq k \leq t-1\}$ is revealed to both the encoder and decoder, and the encoding–decoding system works as follows:

- *Encoding:* Given a source n–tuple X^n, find a vector of indices $(j_1^*, j_2^*, \ldots, j_n^*)$ that minimizes the overall distortion $\sum_{t=1}^n \rho(X_t, Y_{j_1,\ldots,j_t})$. Represent each component j_t^* (based on j_{t-1}^*) by $R = \ln d$ nats (that is, $\log_2 d$ bits), thus a total of nR nats.

- *Decoding:* At each time t ($1 \leq t \leq n$), after having decoded $(j_1^*, \ldots, j_t^*)$, output the reproduction symbol $Y_{j_1^*,\ldots,j_t^*}$.

In order to analyze the rate–distortion performance of this ensemble of codes, we now make the following assumption:

The random coding distribution q is such that the distribution of the random variable $\rho(x, Y)$ is the same for all $x \in \mathcal{X}$.

It turns out that this assumption is fulfilled quite often – it is the case whenever the random coding distribution together with distortion function exhibit a sufficiently high degree of symmetry. For example, if q is the uniform distribution over $\mathcal{Y}$ and the rows of the distortion matrix $\{\rho(x, y)\}$ are permutations of each other, which is in turn the case, for example, when $\mathcal{X} = \mathcal{Y}$ is a group and $\rho(x, y) = \gamma(x - y)$ is a difference distortion function w.r.t. the group difference operation. Somewhat more generally, this assumption still holds when the different rows of the distortion matrix are formed by permutations of each other subject to the following rule: $\rho(x, y)$ can be swapped with $\rho(x, y')$ provided that $q(y') = q(y)$.

For a given $\boldsymbol{x}$ and a given realization of the set of codebooks, define the partition function in analogy to that of the DPRM:

$$Z_n(\beta) = \sum_{\boldsymbol{w}} \exp\{-\beta \sum_{t=1}^{n} \rho(x_t, Y_{j_1,\ldots,j_t})\}, \tag{518}$$

where the summation extends over all d^n possible walks, $\boldsymbol{w} = (j_1, \ldots, j_n)$, along the Cayley tree. Clearly, considering our symmetry assumption, this falls exactly under the umbrella of the DPRM, with the distortions $\{\rho(x_t, Y_{j_1,\ldots,j_t})\}$ playing the role of the branch energies $\{\varepsilon_{i.j}\}$. Therefore, $\frac{1}{n\beta} \ln Z_n(\beta)$ converges almost surely, as n grows without bound, to $\phi(\beta)$, now defined as

$$\phi(\beta) = \begin{cases} \phi_0(\beta) & \beta \leq \beta_c \\ \phi_0(\beta_c) & \beta > \beta_c \end{cases} \tag{519}$$

where now

$$\phi_0(\beta) \triangleq \frac{\ln[d \cdot \boldsymbol{E}\{e^{-\beta\rho(x,Y)}\}]}{\beta}$$

$$= \frac{\ln[e^R \cdot \boldsymbol{E}\{e^{-\beta\rho(x,Y)}\}]}{\beta}$$
$$= \frac{R + \ln[\boldsymbol{E}\{e^{-\beta\rho(x,Y)}\}]}{\beta},$$

Thus, for every $(x_1, x_2, \ldots)$, the distortion is given by

$$\begin{aligned}
\limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \rho(x_t, Y_{j_1^*,\ldots,j_t^*}) &\triangleq \limsup_{n\to\infty} \frac{1}{n} \min_{\boldsymbol{w}} \left[\sum_{t=1}^{n} \rho(x_t, Y_{j_1,\ldots,j_t})\right] \\
&= \limsup_{n\to\infty} \limsup_{\ell\to\infty} \left[-\frac{\ln Z_n(\beta_\ell)}{n\beta_\ell}\right] \\
&\leq \limsup_{\ell\to\infty} \limsup_{n\to\infty} \left[-\frac{\ln Z_n(\beta_\ell)}{n\beta_\ell}\right] \\
&\overset{\text{a.s.}}{=} -\liminf_{\ell\to\infty} \phi(\beta_\ell) \\
&= -\phi_0(\beta_c) \\
&= \max_{\beta\geq 0} \left[-\frac{\ln[\boldsymbol{E}\{e^{-\beta\rho(x,Y)}\}] + R}{\beta}\right] \\
&= D(R),
\end{aligned}$$

where: (i) $\{\beta_\ell\}_{\ell\geq 1}$ is an arbitrary sequence tending to infinity, (ii) the almost–sure equality in the above mentioned paper, and (iii) the justification of the inequality can be found in [74]. The last equation is easily obtained (see [74]) by inverting the function $R(D)$ in its Legendre representation

$$R(D) = -\min_{\beta\geq 0} \min_{q} \left\{\beta D + \sum_{x\in\mathcal{X}} p(x) \ln \left[\sum_{y\in\mathcal{Y}} q(y) e^{-\beta\rho(x,y)}\right]\right\}. \tag{520}$$

Thus, the ensemble of tree codes achieves $R(D)$ almost surely. This strengthens well known coding theorems for tree codes, which make assertions on the achievability of $D(R)$ in expectation only [19],[27],[35],[47],[48].

8 Summary and Outlook

In these lecture notes, we have focused on relationships and analogies between general principles, as well as mathematical formalisms, that are common to both information theory and statistical physics. The emphasis was not merely on the analogies themselves, which may be interesting on their own right, but more importantly, also on new insights and analysis tools that the physical perspectives may contribute to problems in information theory.

We have seen quite a few examples for these insights and tools along the paper. To name a few: (i) The physical point of view on the rate–distortion function in Section 3.2, has lead to the MMSE representation, which sets the stage for the derivation of new bounds. (ii) Gibbs' inequality, which is used in physics to obtain bounds on free energies of complicated systems, may be used in information theory as well, and we have seen the example of the HMM (Subsection 3.3.1). (iii) The dynamical version of the second law was shown to be related to the generalized data processing theorems of Ziv and Zakai and to lead to an even more general data processing theorem (Section 3.4). (iv) The REM and its extensions has been proved useful to inspire a new alternative analysis technique for the derivation of error exponents in ensembles of channel codes, characteristic functions of distortion in source coding, and more (Chapters 6 and 7). It appears then that whatever the field of statistical mechanics can offer to us, in terms of ideas and tools, goes beyond the application of the replica method, which has been the main tool borrowed from statistical mechanics until now, in order to analyze the performance of large communication systems.[35]

We have made an attempt to explore a fairly wide spectrum of aspects of the parallelism and the analogy between the two fields, but as we emphasized already in the Introduction, this is merely a very small fraction of the many existing meeting points between them, and there was no attempt to provide a full comprehensive coverage. This was just a drop in the ocean. It is strongly believed that many additional ideas, insights and analysis tools in

[35] Of course, other statistical–mechanical tools have also been used, and this includes the cavity method, the Bethe–Peierls approximation, series expansions, duality, and others, but still, the replica method has been, by far, the most popular tool that is borrowed from statistical mechanics.

physics are still waiting to find their ways to penetrate into the world of information theory and to help us develop new concepts and techniques for facing the many existing challenges in our field.

It is also believed that the reverse direction, of generating a flow of ideas from information theory to physics, should be at least as fascinating and fruitful. This direction, however, may naturally belong more to the courtyard of physicists with background in information theory than to that of information theorists with background in physics. Indeed, quite many physicists have been active on this front. At the time of writing these lines, however, the author of this paper has not been involved in this line of work, at least not yet.

References

[1] L.–P. Arguin, "Spin glass computations and Ruelle's probability cascades," arXiv:math-ph/0608045v1, August 17, 2006.

[2] G. B. Bağci, "The physical meaning of Rényi relative entropies," arXiv:cond-mat/0703008v1, March 1, 2007.

[3] A. Barg and G. D. Forney, Jr., "Random codes: minimum distances and error exponents," *IEEE Trans. Inform. Theory*, vol. 48, no. 9, pp. 2568–2573, September 2002.

[4] R. J. Baxter, *Exactly Solved Models in Statistical Mechanics*, Academic Press, 1982.

[5] A. H. W. Beck, *Statistical Mechanics, Fluctuations and Noise*, Edward Arnold Publishers, 1976.

[6] C. H. Bennett, "Notes on Landauer's principle, reversible computation and Maxwell's demon," *Studies in History and Philosophy of Modern Physics*, vol. 34 pp. 501–510, 2003.

[7] G. P. Beretta and E. Zanchini, "Rigorous and general definition of thermodynamic entropy," arXiv:1010.0813v1 [quant-ph] 5 Oct 2010.

[8] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[9] E. Buffet, A. Patrick, and J. V. Pulé, "Directed polymers on trees: a martingale approach," *J. Phys. A: Math. Gen.*, vol. 26, pp. 1823–1834, 1993.

[10] F. Cousseau, K. Mimura, and M. Okada, "Statistical mechanics of lossy compression for non–monotonic multilayer perceptron," *Proc. ISIT 2008*, pp. 509–513, Toronto, Canada, July 2008.

[11] F. Cousseau, K. Mimura, T. Omori, and M. Okada, "Statistical mechanics of lossy compression for non–monotonic multilayer perceptrons," Phys. Rev. E 78, 021124 2008; arXiv:0807.4009v1 [cond-mat.stat-mech] 25 Jul 2008.

[12] T. M. Cover and E. Ordentlich, "Universal portfolios with side information," *IEEE Trans. Inform. Theory*, vol. IT–42, no. 2, pp. 348–363, March 1996.

[13] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, second edition, John Wiley & Sons, 2006.

[14] G. E. Crooks, "Beyond Boltzmann–Gibbs statistics: maximum entropy hyperensembles out of equilibrium," *Phys. Rev. E*, vol. 75, 041119, 2007.

[15] I. Csiszár, "A class of measures of informativity of observation channels," *Periodica Mathematica Hungarica*, vol. 22, no. 1–4, pp. 191–213, 1972.

[16] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete earlier Memoryless Systems*, Academic Press, New York, 1981.

[17] B. Dacorogna and P. Maréchal, "The role of perspective functions in convexity, polyconvexity, rank–one convexity and separate convexity," http://caa.epfl.ch/publications/2008-The_role_of_perspective_functions_in_convexity.pdf

[18] W.-S. Dai and M. Xie, "The explicit expression of the fugacity for hard–sphere Bose and Fermi gases," arXiv:0906.0952v1 [cond-mat.stat-mech] 4 June 2009.

[19] C. R. Davis and M. E. Hellman, "On tree coding with a fidelity criterion," *IEEE Trans. Inform. Theory*, vol. IT–21, no. 4, pp. 373–378, July 1975.

[20] N. G. de Bruijn, *Asymptotic Methods in Analysis*, Dover Publications, 1981.

[21] C. De Dominicis, M. Gabay, T. Garel, and H. Orland, "White and weighted averages over solutions of Thouless Anderson Palmer equations for the Sherrington Kirkpatrick spin glass," *J. Physique*, vol. 41, pp. 923–930, 1980.

[22] B. Derrida, "Random–energy model: limit of a family of disordered models," *Phys. Rev. Lett.*, vol. 45, no. 2, pp. 79–82, July 1980.

[23] B. Derrida, "The random energy model," *Physics Reports* (Review Section of Physics Letters), vol. 67, no. 1, pp. 29–35, 1980.

[24] B. Derrida, "Random–energy model: an exactly solvable model for disordered systems," *Phys. Rev. B*, vol. 24, no. 5, pp. 2613–2626, September 1981.

[25] B. Derrida, "A generalization of the random energy model which includes correlations between energies," *J. de Physique – Lettres*, vol. 46, L–401-107, May 1985.

[26] B. Derrida and E. Gardner, "Solution of the generalised random energy model," *J. Phys. C: Solid State Phys.*, vol. 19, pp. 2253–2274, 1986.

[27] R. J. Dick, T. Berger, and F. Jelinek, "Tree encoding of Gaussian sources," *IEEE Trans. Inform. Theory*, vol. IT–20, no. 3, pp. 332–336, May 1974.

[28] M. Doi and S. F. Edwards, *The Theory of Polymer Dynamics*, Oxford University Press, 1986.

[29] V. Dotsenko, "One more discussion on the replica trick: the examples of exact solutions," arXiv:1010.3913v1 [cond-mat.stat-mech] 19 Oct 2010.

[30] R. S. Ellis, *Entropy, large deviations, and statistical mechanics*, Springer–Verlag, NY, 1985.

[31] R. Etkin, N. Merhav and E. Ordentlich, "Error exponents of optimum decoding for the interference channel," *IEEE Trans. Inform. Theory*, vol. 56, no. 1, pp. 40–56, January 2010.

[32] G. D. Forney, Jr., "Exponential error bounds for erasure, list, and decision feedback schemes," *IEEE Trans. Inform. Theory*, vol. IT–14, no. 2, pp. 206–220, March 1968.

[33] G. D. Forney, Jr. and A. Montanari, "On exponential error bounds for random codes on the DMC," manuscript, 2001.
http://www.stanford.edu/∼montanar/PAPERS/FILEPAP/dmc.ps

[34] R. G. Gallager, *Information Theory and Reliable Communication*, John Wiley & Sons, 1968.

[35] R. G. Gallager, "Tree encoding for symmetric sources with a distortion measure," *IEEE Trans. Inform. Theory*, vol. IT–20, no. 1, pp. 65–76, January 1974.

[36] R. M. Gray, *Source Coding Theory*, Kluwer Academic Publishers, 1990.

[37] D. Guo and S. Verdú, "Randomly spread CDMA: asymptotics via statistical physics," *IEEE Trans. Inform. Theory*, vol. 51, no. 6, pp. 1982–2010, June 2005.

[38] M. J. W. Hall, "Universal geometric approach to uncertainty, entropy, and information," *Phys. Rev. A*, vol. 59, no. 4, pp. 2602–2615, April 1999.

[39] J. Honerkamp, *Statistical Physics – An Advanced Approach with Applications*, 2nd edition, Springer–Verlag, 2002.

[40] Wm. G. Hoover and C. G. Hoover, "Nonequilibrium temperature and thermometry in heat–conducting ϕ^4 models," *Phys. Rev. E*, vol. 77, 041104, 2008.

[41] T. Hosaka and Y. Kabashima, "Statistical mechanical approach to error exponents of lossy data compression," *J. Physical Society of Japan*, vol. 74, no. 1, pp. 488–497, January 2005.

[42] P. A. Humblet, "Generalization of Huffman coding to minimize the probability of buffer overflow," *IEEE Transactions on Information Theory*,

[43] C. Jarzynski, "Nonequilibrium equality for free energy differences," *Phys. Rev. Lett.*, vol. 78, no. 14, pp. 2690–2693, 7 April, 1997.

[44] E. T. Jaynes, "Information theory and statistical mechanics," *Phys. Rev. A*, vol. 106, pp. 620–630, May 1957.

[45] E. T. Jaynes, "Information theory and statistical mechanics - II," *Phys. Rev. A*, vol. 108, pp. 171–190, October 1957.

[46] F. Jelinek, "Buffer overflow in variable length coding of fixed rate sources," *IEEE Transactions on Information Theory*, vol. IT–14, no. 3, pp. 490–501, May 1968.

[47] F. Jelinek, "Tree encoding of memoryless time–discrete sources with a fidelity criterion," *IEEE Trans. Inform. Theory*, vol. IT–15, no. 5, pp. 584–590, September 1969.

[48] F. Jelinek and J. B. Anderson, "Instrumentable tree encoding of information sources," *IEEE Trans. Inform. Theory*, vol. IT–17, no. 1, pp. 118–119, January 1971.

[49] Y. Kabashima, "How could the replica method improve accuracy of performance assessment of channel coding?" *Proc. Int. Workshop on Statistical–Mechanical Informatics*, Sept. 14–17, 2008, Sendai, Japan. arXiv:0808.0548v1 [cs.IT] 5 Aug. 2008.

[50] Y. Kabashima and T. Hosaka, "Statistical mechanics for source coding with a fidelity criterion," *Progress of Theoretical Physics*, Supplement no. 157, pp. 197–204, 2005.

[51] Y. Kabashima, K. Nakamura, and J. van Mourik, "Statistical mechanics of typical set decoding," *Physical Review E*, vol. 66, 2002.

[52] Y. Kabashima and D. Saad, "Statistical mechanics of error correcting codes," *Europhysics Letters*, vol. 45, no. 1, pp. 97–103, 1999.

[53] Y. Kabashima and D. Saad, "Statistical mechanics of low–density parity check codes," *J. Phys. A: Math. Gen.*, vol. 37, pp. R1–R43, 2004.

[54] M. Kardar, *Statistical Physics of Particles*, Cambridge University Press, 2007.

[55] Y. Kaspi and N. Merhav, "Error exponents of optimum decoding for the degraded broadcast channel using moments of type class enumerators," *Proc. ISIT 2009*, pp. 2507–2511, Seoul, South Korea, June–July 2009. Full version: available in arXiv:0906.1339.

[56] R. Kawai, J. M. R. Parrondo, and C. Van den Broeck, "Dissipation: the phase–space perspective," *Phys. Rev. Lett.*, vol. 98, 080602, 2007.

[57] F. P. Kelly, *Reversibility and Stochastic Networks*, J. Wiley & Sons, 1979.

[58] K. Kitagawa and T. Tanaka, "Optimal spreading sequences in large CDMA systems: a *Proc. ISIT 2008*, pp. 1373–1377, Toronto, Canada, July 2008.

[59] C. Kittel, *Elementary Statistical Physics*, John Wiley & Sons, 1958.

[60] R. Kubo, *Statistical Mechanics*, North–Holland, 1961.

[61] L. D. Landau and E. M. Lifshitz, *Course of Theoretical Physics – volume 5: Statistical Physics, Part 1*, 3rd edition, Elsevier, 1980.

[62] R. Landauer, "Irreversibility and heat generation in the computing process," *IBM Journal of Research and Development*, vol. 5, pp. 183–191, 1961.

[63] T. D. Lee and C. N. Yang, "Statistical theory of equations of state and phase transitions. II. Lattice gas and Ising model," *Physical Review*, vol. 87, no. 3, pp. 410–419, August 1952.

[64] H. Löwen, "Fun with hard spheres," in *Spatial Statistics and Statistical Physics* edited by K. Mecke and D. Stoyan, Springer Lecture Notes in Physics, vol. 554, pp. 295–331, Berlin, 2000.

[65] F. Mandl, *Statistical Physics*, John Wiley & Sons, 1971.

[66] N. Merhav, "Universal coding with minimum probability of code word length overflow," *IEEE Trans. Inform. Theory*, vol. 37, no. 3, pp. 556–563, May 1991.

[67] N. Merhav, "An identity of Chernoff bounds with an interpretation in statistical physics and applications in information theory," *IEEE Trans. Inform. Theory*, vol. 54, no. 8, pp. 3710–3721, August 2008.

[68] N. Merhav, "The random energy model in a magnetic field and joint source–channel coding," *Physica A: Statistical Mechanics and Its Applications*, vol. 387, issue 22, pp. 5662–5674, September 15, 2008.

[69] N. Merhav, "Relations between random coding exponents and the statistical physics of random codes," *IEEE Trans. Inform. Theory*, vol. 55, no. 1, pp. 83–92, January 2009.

[70] N. Merhav, "The generalized random energy model and its application to the statistical physics of ensembles of hierarchical codes," *IEEE Trans. Inform. Theory*, vol. 55, no. 3, pp. 1250–1268, March 2009.

[71] N. Merhav, "Another look at the physics of large deviations with application to rate–distortion theory,"
http://arxiv.org/PS_cache/arxiv/pdf/0908/0908.3562v1.pdf.

[72] N. Merhav, "Rate–distortion function via minimum mean square error estimation," submitted to *IEEE Trans. Inform. Theory*, April 2010.

[73] N. Merhav, "Physics of the Shannon limits," to appear in *IEEE Trans. Inform. Theory*, September 2010.

[74] N. Merhav, "On the statistical physics of directed polymers in a random medium and their relation to tree codes," *IEEE Trans. Inform. Theory*, vol. 56, no. 3, pp. 1345–1350, March 2010.

[75] N. Merhav, "On the physics of rate–distortion theory," *Proc. ISIT 2010*, pp. 71–75, Austin, Texas, U.S.A., June 2010.

[76] N. Merhav, "Data processing theorems and the second law of thermodynamics," submitted to *IEEE Trans. Inform. Theory*, 2010.

[77] N. Merhav, "Threshold effects in parameter estimation as phase transitions in statistical mechanics," submitted to *IEEE Trans. Inform. Theory*, 2010.

[78] N. Merhav, "Error exponents of erasure/list decoding revisited via moments of distance enumerators," *IEEE Trans. Inform. Theory*, vol. 54, no. 10, pp. 4439–4447, October 2008.

[79] N. Merhav, D. Guo, and S. Shamai (Shitz), "Statistical physics of signal estimation in Gaussian noise: theory and examples of phase transitions," *IEEE Trans. Inform. Theory*, vol. 56, no. 3, pp. 1400–1416, March 2010.

[80] M. Mézard and A. Montanari, *Information, Physics and Computation*, Oxford University Press, 2009.

[81] A. Montanari, "Turbo codes: the phase transition," *The European Physical Journal B* vol. 18, p. 121, 2000. E–print: cond-mat/0003218.

[82] A. Montanari, "The glassy phase of Gallager codes," *The European Physical Journal B – Condensed Matter and Complex Systems*, Vol. 23, no. 1, pp. 121–136, 2001. E–print: cond-mat/0104079.

[83] T. Mora and O. Rivoire, "Statistical mechanics of error exponents for error–correcting codes," arXiv:cond-mat/0606696, June 2006.

[84] T. Murayama, "Statistical mechanics of the data compression theorem," *J. Phys. A: Math. Gen.*, vol. 35, pp. L95–L100, 2002.

[85] K. R. Narayanan and A. R. Srinivasa, "On the thermodynamic temperature of a general distribution," arXiv:0711.1460v2 [cond-mat.stat-mech], Nov. 10, 2007.

[86] K. R. Narayanan and A. R. Srinivasa, "A Shannon entropy–based non–equilibrium temperature of a general distribution with application to Langevin dynamics," preprint, May 2009.

[87] H. Nishimori, *Statistical Physics of Spin Glasses and Information Processing: an Introduction*, (International Series of Monographs on Physics, no. 111), Oxford University Press, 2001.

[88] C. R. de Oliveira and T. Werlang, "Ergodic hypothesis in classical statistical mechanics," *Revista Brasileira de Ensino de Fisíca*, vol. 29, no. 2, pp. 189–201, 2007. Also available

on–line:
http://www.sbfisica.org.br/rbef/pdf/060601.pdf

[89] L. Onsager, "Crystal statistics. I. A two-dimensional model with an order–disorder transition," *Phys. Rev.*, vol. 65, no. 2, pp. 117–149, 1944.

[90] D. P. Palomar and S. Verdú, "Lautum information," *IEEE Trans. Inform. Theory*, vol. 54, no. 3, pp. 964–975, March 2008.

[91] G. Parisi and F. Zamponi, "Mean field theory of the glass transition and jamming hard spheres," arXiv:0802.2180v2 [cond-mat.dis-nn] 18 Dec. 2008.

[92] P. Pradhan, Y. Kafri, and D. Levine, "Non–equilibrium fluctuation theorems in the presence of local heating," arXiv:0712.0339v2 [cond-mat.stat-mech] 3 Apr 2008.

[93] A. Procacci and B. Scoppola, "Statistical mechanics approach to coding theory," *J. of Statistical Physics*, vol. 96, nos. 3/4, pp. 907–912, 1999.

[94] H. Qian, "Relative entropy: free energy associated with equilibrium fluctuations and nonequilibrium deviations," *Phys. Rev. E*, vol. 63, 042103, 2001.

[95] F. Reif, *Fundamentals of Statistical and Thermal Physics*, McGraw–Hill, 1965.

[96] H. Reiss, "Thermodynamic–like transformations in Information Theory," *Journal of Statistical Physics*, vol. 1, no. 1, pp. 107–131, 1969.

[97] H. Reiss, H. L. Frisch, and J. L. Lebowitz, "Statistical mechanics of rigid spheres," *J. Chem. Phys.*, vol. 31, no. 2, pp. 369–380, August 1959.

[98] H. Reiss and C. Huang, "Statistical thermodynamic formalism in the solution of Information Theory problems," *Journal of Statistical Physics*, vol. 3, no. 2, pp. 191–211, 1971.

[99] H. Reiss and P. Schaaf, "Hard spheres: thermodynamics and geometry," *J. Chem. Phys.*, vol. 91, no. 4, pp. 2514–2524, 15 August 1989.

[100] T. Richardson and R. Urbanke, *Modern Coding Theory*, Cambridge University Press, 2008.

[101] K. Rose, "A mapping approach to rate-distortion computation and analysis," *IEEE Trans. Inform. Theory*, vol. 40, no. 6, pp. 1939–1952, November 1994.

[102] P. Ruján, "Finite temperature error–correcting codes," *Phys. Rev. Let.*, vol. 70, no. 19, pp. 2968–2971, May 1993.

[103] F. W. Sears, M. W. Zemansky and H. D. Young, *University Physics*, Addison–Wesley, 1976.

[104] J. P. Sethna, *Statistical Mechanics: Entropy, Order Parameters, and Complexity*, Oxford University Press, 2007.

[105] O. Shental and I. Kanter, "Shannon capacity of infinite–range spin–glasses," technical report, Bar Ilan University, 2005.

[106] A. Somekh–Baruch and N. Merhav, "Exact random coding exponents for erasure decoding," *Proc. ISIT 2010*, pp. 260–264, June 2010, Austin, Texas, U.S.A.

[107] O. Shental, N. Shental, S. Shamai (Shitz), I. Kanter, A. J. Weiss, and Y. Weiss, "Discrete input two–dimensional Gaussian channels with memory: estimation and information rates via graphical models and statistical mechanics," *IEEE Trans. Inform. Theory*, vol. 54, no. 4, April 2008.

[108] D. Ruelle, *Statistical Mechanics: Rigorous Results*, Addison–Wesley, 1989.

[109] J. E. Shore and R. W. Johnson, "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy," *IEEE Trans. Inform. Theory*, vol. IT–26, no. 1, pp. 26–37, January 1980.

[110] N. Simanyi, "The Boltzmann–Sinai ergodic hypothesis in full generality," arXiv:1007.1206v2 [math.DS] 10 Aug. 2010.

[111] N. Sourlas, "Spin–glass models as error–correcting codes," *Nature*, pp. 693–695, vol. 339, June 1989.

[112] N. Sourlas, "Spin glasses, error–correcting codes and finite–temperature decoding," *Europhysics Letters*, vol. 25, pp. 159–164, 1994.

[113] K. Takeuchi, M. Vehpaperä, T. Tanaka, and R. R. Müller, "Replica analysis of general multiuser detection in MIMO DS–CDMA channels with imperfect CSI," *Proc. ISIT 2008*, pp. 514–518, Toronto, Canada, July 2008.

[114] T. Tanaka, "Statistical mechanics of CDMA multiuser demodulation," *Europhysics Letters*, vol. 54, no. 4, pp. 540–546, 2001.

[115] T. Tanaka, "A statistical–mechanics approach to large–system analysis of CDMA multiuser detectors," *IEEE Trans. Inform. Theory*, vol. 48, no. 11, pp. 2888–2910, November 2002.

[116] H. Touchette, "Methods for calculating nonconcave entropies," arXiv:1003.0382v1 [cond-mat.stat-mech] 1 Mar 2010.

[117] A. M. Tulino and S. Verdú, "Random matrix theory and wireless communications," *Foundations and Trends in Communications and Information Theory*, vol. 1, issue 1, 2004.

[118] J. J. M. Varbaarschot and M. R. Zirnbauer, "Critique of the replica trick," *J. Phys. A: Math. Gen.*, vol. 17, pp. 1093–1109, 1985.

[119] A. J. Viterbi and J. K. Omura, *Principles of Digital Communication and Coding*, McGraw–Hill, 1979.

[120] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference," *Foundations and Trends in Machine Learning*, vol. 1, nos. 1–2, 2008.

[121] E. Weinstein and A. J. Weiss, "Lower bounds on the mean square estimation error," *Proc. of the IEEE*, vol. 73, no. 9, pp. 1433–1434, September 1985.

[122] A. J. Weiss, *Fundamental Bounds in Parameter Estimation*, Ph.D. dissertation, Tel Aviv University, Tel Aviv, Israel, June 1985.

[123] J. M. Wozencraft and I. M. Jacobs, *Principles of Communication Engineering*, John Wiley & Sons, 1965. Reissued by Waveland Press, 1990.

[124] A. D. Wyner, "On the probability of buffer overflow under an arbitrary bounded input-output distribution," *SIAM Journal on Applied Mathematics*, vol. 27, no. 4, pp. 544–570, December 1974.

[125] C. N. Yang and T. D. Lee, "Statistical theory of equations of state and phase transitions. I. Theory of condensation," *Physical Review*, vol. 87, no. 3, pp. 404–409, August 1952.

[126] J. S. Yedidia, "Quenched disorder: understanding glasses using a variational principle and the replica method," lectures delivered at the Santa Fe Summer School on Complex Systems, June 15–19, 1992.

[127] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Constructing free energy approximations and generalized belief propagation algorithms," *IEEE Trans. Inform. Theory*, vol. 51, no. 7, pp. 2282–2312, July 2005.

[128] R. W. Yeung, *A First Course in Information Theory*, Kluwer Academic/Plenum Publishers, New York, 2002. See also R. W. Yeung, *Information Theory and Network Coding*, Springer, 2008.

[129] M. Zakai and J. Ziv, "A generalization of the rate-distortion theory and applications," in: *Information Theory New Trends and Open Problems*, edited by G. Longo, Springer-Verlag, 1975, pp. 87–123.

[130] M. R. Zirnbauer, "Another critique of the replica trick," arXiv:cond-mat/9903338v1 [cond-mat.mes-hall] 22 Mar 1999.

[131] J. Ziv and M. Zakai, "On functionals satisfying a data–processing theorem," *IEEE Trans. Inform. Theory*, vol. IT–19, no. 3, pp. 275–283, May 1973.