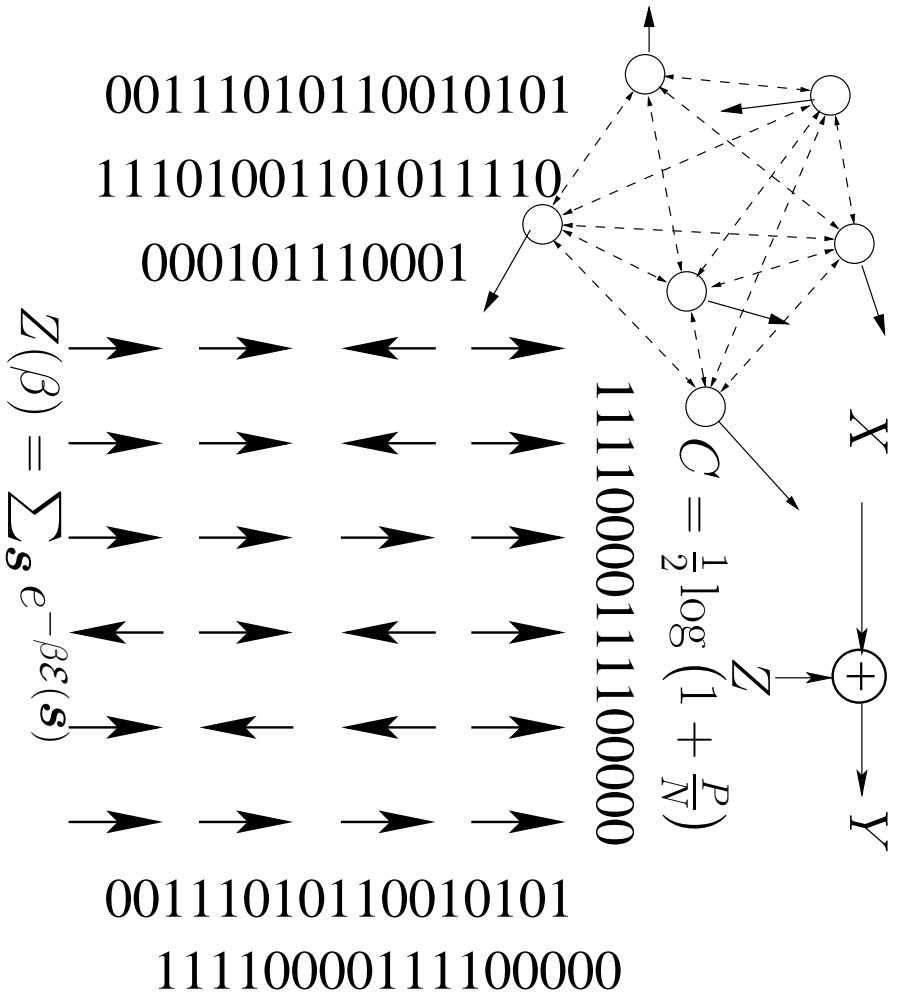


Information Theory and Statistical Physics –
Lecture Notes

Neri Merhav



Information Theory and Statistical Physics – Lecture Notes

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

Abstract

This document consists 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.

A short outline of the course is as follows: Introduction; Elementary Statistical Physics and its Relation to Information Theory; Analysis Tools in Statistical Physics; Systems of Interacting Particles and Phase Transitions; The Random Energy Model (REM) and Random Channel Coding; Additional Topics (optional).

Contents

1	Introduction	5
2	Elementary Stat. Physics and Its Relation to IT	10
2.1	What is Statistical Physics?	10
2.2	Basic Postulates and the Microcanonical Ensemble	11
2.3	The Canonical Ensemble	17
2.4	Properties of the Partition Function and the Free Energy	20
2.5	The Energy Equipartition Theorem	29
2.6	The Grand–Canonical Ensemble (Optional)	31
2.7	Gibbs’ Inequality, the 2nd Law, and the Data Processing Thm	34
2.8	Large Deviations Theory and Physics of Information Measures	40
3	Analysis Tools and Asymptotic Methods	53
3.1	Introduction	53
3.2	The Laplace Method	55
3.3	The Saddle Point Method	58
3.4	The Replica Method	68
4	Interacting Particles and Phase Transitions	74
4.1	Introduction – Origins of Interactions	74
4.2	A Few Models That Will be Discussed in This Subsection Only	74
4.3	Models of Magnetic Materials – General	76
4.4	Phase Transitions – A Qualitative Discussion	80
4.5	The One–Dimensional Ising Model	83

4.6	The Curie–Weiss Model	86
4.7	Spin Glass Models With Random Parameters and Random Code Ensembles	91
5	The Random Energy Model and Random Coding	96
5.1	The REM in the Absence of a Magnetic Field	96
5.2	The Random Code Ensemble and its Relation to the REM	102
5.3	Random Coding Exponents	108
6	Additional Topics (Optional)	119
6.1	The REM With a Magnetic Field and Joint Source–Channel Coding	119
6.1.1	Magnetic Properties of the REM	119
6.1.2	Relation to Joint Source–Channel Coding	122
6.2	The Generalized Random Energy Model (GREM) and Hierarchical Coding .	127
6.3	Phase Transitions of the Rate–Distortion Function	141
6.4	Capacity of the Sherrington–Kirkpatrick Spin Glass	145
6.5	Generalized Temperature, de Bruijn’s Identity, and Fisher Information . . .	149
6.6	The Gibbs Inequality and the Log–Sum Inequality	155
6.7	Dynamics, Evolution of Info Measures, and Simulation	161
6.7.1	Markovian Dynamics, Global Balance and Detailed Balance	161
6.7.2	Evolution of Information Measures	163
6.7.3	Monte Carlo Simulation	169

1 Introduction

This course is intended to EE graduate students in the field of Communications and Information Theory, and also to graduates of the Physics Department (in particular, graduates of the EE–Physics program) who have basic background in Information Theory, which is a prerequisite to this course. As its name suggests, this course focuses on relationships and interplay between Information Theory and *Statistical Physics* – a branch of physics that deals with many–particle systems using probabilistic/statistical methods in the microscopic level.

The relationships between Information Theory and Statistical Physics (+ thermodynamics) are by no means new, and many researchers have been exploiting them for many years. Perhaps the first relation, or analogy, that crosses our minds is that in both fields, there is a fundamental notion of *entropy*. Actually, in Information Theory, the term entropy was coined after the thermodynamic entropy. The thermodynamic entropy was first introduced by Clausius (around 1850), whereas its probabilistic–statistical interpretation is due to Boltzmann (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, and we will not cover all of them in this course, but just to give the idea 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, image processing, and the like. It is about 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 widespread 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 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 relevance, 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, and further advocated by Shore and Johnson afterwards. In the book by Cover and Thomas, there is a very nice chapter on this, but we will not delve into this any further in this course.

- *Landauer's Erasure Principle.* Another aspect of these relations has to do with a piece of theory whose underlying guiding principle is that *information is a physical entity*. In every information bit in the universe there is a certain amount of energy. Specifically, Landauer's erasure principle (from the early sixties of the previous century), which is based on a 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. It is my personal opinion that these kind of theories should be taken with a grain of salt, but this is only my opinion. At any rate, this is not going to be included in the course either.
- *Large Deviations Theory as a Bridge Between Information Theory and Statistical Physics.* Both Information Theory and Statistical Physics have an intimate relation to *large de-*

viations theory, a branch of probability theory which focuses on the assessment of the exponential rates of decay of probabilities of rare events, where the most fundamental mathematical tool is the *Chernoff bound*. This is a topic that will be covered in the course and quite soon.

- *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 hot area in probability theory with many applications, both in Statistical Physics and in Information Theory, especially in modern theories of wireless communication (e.g., MIMO systems). This is again outside the scope of this course, but whoever is interested to ‘taste’ it, is invited to read the 2004 paper by Tulino and Verdú in *Foundations and Trends in Communications and Information Theory*, a relatively new journal for tutorial papers.
- *Spin Glasses and Coding Theory*. 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* – amorphous 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.

We can go on and on with this list and add more items in the context of these very fascinating meeting points between Information Theory and Statistical Physics, but for now, we stop here. We just mention that the last item will form the main core of the course. We will 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.

At this point, let us have a brief review of the syllabus of this course, where as can be seen, the physics and the Information Theory 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 seen more readily. The detailed structure of the remaining part of this course is as follows:

1. *Elementary Statistical Physics and its Relation to Information Theory*: What is statistical physics? Basic postulates and the micro–canonical ensemble; the canonical ensemble: the Boltzmann–Gibbs law, the partition function, thermodynamical potentials and their relations to information measures; the equipartition theorem; generalized ensembles (optional); Chernoff bounds and the Boltzmann–Gibbs law: rate functions in Information Theory and thermal equilibrium; physics of the Shannon limits.
2. *Analysis Tools in Statistical Physics*: The Laplace method of integration; the saddle–point method; transform methods for counting and for representing non–analytic functions; examples; the replica method – overview.
3. *Systems of Interacting Particles and Phase Transitions*: Models of many–particle systems with interactions (general) and examples; a qualitative explanation for the existence of phase transitions in physics and in information theory; ferromagnets and Ising models: the 1D Ising model, the Curie–Weiss model; randomized spin–glass models: annealed vs. quenched randomness, and their relevance to coded communication systems.
4. *The Random Energy Model (REM) and Random Channel Coding*: Basic derivation and phase transitions – the glassy phase and the paramagnetic phase; random channel codes

and the REM: the posterior distribution as an instance of the Boltzmann distribution, analysis and phase diagrams, implications on code ensemble performance analysis.

5. *Additional Topics (optional)*: The REM in a magnetic field and joint source–channel coding; the generalized REM (GREM) and hierarchical ensembles of codes; phase transitions in the rate–distortion function; Shannon capacity of infinite–range spin–glasses; relation between temperature, de Bruijn’s identity, and Fisher information; the Gibbs inequality in Statistical Physics and its relation to the log–sum inequality of Information Theory.

As already said, there are also plenty of additional subjects that fall under the umbrella of relations between Information Theory and Statistical Physics, which will not be covered in this course. One very hot topic is that of codes on graphs, iterative decoding, belief propagation, and density evolution. The main reason for not including these topics is that they are already covered in the course of Dr. Igal Sason: “Codes on graphs.”

I would like to emphasize that prior basic background in Information Theory will be assumed, therefore, Information Theory is a prerequisite for this course. As for the physics part, prior background in statistical mechanics could be helpful, but it is not compulsory. The course is intended to be self–contained as far as the physics background goes. The bibliographical list includes, in addition to a few well known books in Information Theory, also several very good books in elementary Statistical Physics, as well as two books on the relations between these two fields.

As a final note, I feel compelled to clarify that the material of this course is by no means intended to be presented from a very comprehensive perspective and to consist of a full account of methods, problem areas and results. Like in many advanced graduate courses in our department, here too, the choice of topics, the approach, and the style strongly reflect the personal bias of the lecturer and his/her perspective on research interests in the field. This is also the reason that a considerable fraction of the topics and results that will be covered, are taken from articles in which I have been involved.

2 Elementary Stat. Physics and Its Relation to IT

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), e.g., of the order of magnitude of *Avogadro's number*, that is, about 10^{23} particles. 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 (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 bunch 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. Rather, 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 – the theory of statistical physics, in the limit of an infinite number of particles, that is, the *thermodynamic limit*. As we shall see throughout this course, 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 $\vec{r}_i(t)$ and the momentum $\vec{p}_i(t)$ 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 $\vec{x}(t) = \{(\vec{r}_i(t), \vec{p}_i(t)), i = 1, 2, \dots, n\}$, 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 at which $\vec{x}(t)$ 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 assumption of *ergodicity*.

What are then the probabilities of these cells? We would like to derive these probabilities from first principles, based on as few as possible basic postulates. Our first such postulate is that for an isolated system (i.e., whose energy is fixed) all microscopic states $\{\vec{x}(t)\}$ 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 course, but in a nutshell, it asserts that the phase trajectories must lie along hypersurfaces of constant probability density.¹

¹This 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.

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 $\mathbf{x} = (x_1, x_2, \dots, x_n)$, 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 \mathbf{x} , there is a certain *Hamiltonian* (i.e., energy function) that assigns to \mathbf{x} a certain energy $\mathcal{E}(\mathbf{x})$.² Now, let us denote by $\Omega(E)$ the *density-of-states* function, i.e., the volume of the shell $\{\mathbf{x} : \mathcal{E}(\mathbf{x}) = E\}$, or, slightly more precisely, $\Omega(E)dE = \text{Vol}\{\mathbf{x} : E \leq \mathcal{E}(\mathbf{x}) \leq E + dE\}$, which will be denoted also as $\text{Vol}\{\mathbf{x} : \mathcal{E}(\mathbf{x}) \approx E\}$, where the dependence on dE will normally be ignored since $\Omega(E)$ is typically exponential in n and dE 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 *microcanonical ensemble*, is that the probability density $P(\mathbf{x})$ is given by

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

In the discrete case, things are, of course, a lot easier: Then, $\Omega(E)$ would be the number of microstates with $\mathcal{E}(\mathbf{x}) = E$ (exactly) and $P(\mathbf{x})$ would be the uniform probability mass function across this set of states. In this case, $\Omega(E)$ is analogous to the size of a *type class* in Information Theory, and $P(\mathbf{x})$ is the uniform distribution across 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 get rid of these physical units because very soon we are going to apply non-linear functions on $\Omega(E)$, like the logarithmic function. Thus, we must normalize this volume by an elementary reference volume. In the gas example, this reference volume is taken to be h^{3n} , where h is *Planck's constant* $\approx 6.62 \times 10^{-34}$ Joules-sec. Informally, the intuition comes from the fact that h is our best available “resolution” in the plane spanned by each component of

²For example, in the case of an *ideal gas*, $\mathcal{E}(\mathbf{x}) = \sum_{i=1}^n \frac{\|\vec{p}_i\|^2}{2m}$, independently of the positions $\{\vec{r}_i\}$, 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.

\vec{r}_i and the corresponding component of \vec{p}_i , owing to the *uncertainty principle* in quantum mechanics, which tells us 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

$$S(E) = k \ln \Omega(E), \quad (2)$$

where k is *Boltzmann's constant* $\approx 1.38 \times 10^{-23}$ Joule/degree. We will soon see what is the relationship between $S(E)$ and the information-theoretic entropy.

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 $\mathcal{E}(\mathbf{x}) = \sum_{i=1}^n \frac{\|\vec{p}_i\|^2}{2m}$, then $\Omega(E)$ is the volume of a shell or surface of a $(3n)$ -dimensional sphere with radius $\sqrt{2mE}$, which is proportional to $(2mE)^{3n/2}V^n$, but we should divide this by $n!$ to account for the fact that the particles are indistinguishable and we don't count permutations as distinct physical states in this case.³ More precisely, one obtains:

$$S(E) = k \ln \left[\left(\frac{4\pi m E}{3n} \right)^{3n/2} \cdot \frac{V^n}{n! h^{3n}} \right] + \frac{3}{2}nk \approx nk \ln \left[\left(\frac{4\pi m E}{3n} \right)^{3/2} \cdot \frac{V}{nh^3} \right] + \frac{5}{2}nk. \quad (3)$$

Assuming $E \propto n$ and $V \propto n$, we get $S(E) \propto n$. A physical quantity like this, that has a linear scaling with the size of the system n , is called an *extensive quantity*. So, energy, volume and entropy are 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

³Since 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.

quantities of the system, e.g., temperature, pressure, and so on. The *temperature* T of the system is defined according to:

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

where $(\cdot)_V$ means that the derivative is taken in constant volume.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 $dE/dT < 0$). 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 we get $1/T \equiv \partial S/\partial E = 3nk/(2E)$, which means $E = 3nkT/2$. Pressure is obtained by $P = T \cdot \partial S/\partial V$, which in our example, gives rise to the state equation of the ideal gas, $P = nkT/V$.

How can we also see *mathematically* 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?

As both E and S are extensive quantities, let us define $E = n\epsilon$ and

$$s(\epsilon) = \lim_{n \rightarrow \infty} \frac{S(n\epsilon)}{n}, \quad (5)$$

i.e., the per-particle entropy as a function of the per-particle energy. Consider the case where the Hamiltonian is additive, i.e.,

$$\mathcal{E}(\mathbf{x}) = \sum_{i=1}^n \mathcal{E}(x_i) \quad (6)$$

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

$$\Omega(n_1\epsilon_1 + n_2\epsilon_2) \geq \Omega(n_1\epsilon_1) \cdot \Omega(n_2\epsilon_2), \quad (7)$$

⁴ This definition of temperature is related to the classical thermodynamical definition of entropy as $dS = dQ/T$, where Q is heat, as in the absence of external work, when the volume V is fixed, all the energy comes from heat and so, $dE = dQ$.

and so, we get:

$$\begin{aligned} \frac{k \ln \Omega(n_1 \epsilon_1 + n_2 \epsilon_2)}{n_1 + n_2} &\geq \frac{k \ln \Omega(n_1 \epsilon_1)}{n_1 + n_2} + \frac{k \ln \Omega(n_2 \epsilon_2)}{n_1 + n_2} \\ &= \frac{n_1}{n_1 + n_2} \cdot \frac{k \ln \Omega(n_1 \epsilon_1)}{n_1} + \frac{n_2}{n_1 + n_2} \cdot \frac{k \ln \Omega(n_2 \epsilon_2)}{n_2}. \end{aligned} \quad (8)$$

and so, by taking n_1 and n_2 to ∞ , with $n_1/(n_1 + n_2) \rightarrow \lambda \in (0, 1)$, we get:

$$s(\lambda \epsilon_1 + (1 - \lambda) \epsilon_2) \geq \lambda s(\epsilon_1) + (1 - \lambda) s(\epsilon_2), \quad (9)$$

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 $\mathcal{E}(\mathbf{x})$ includes also a limited degree of interactions (short range interactions), e.g., $\mathcal{E}(\mathbf{x}) = \sum_{i=1}^n \mathcal{E}(x_i, x_{i+1})$, 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 $\mathcal{E}(\mathbf{x})$ depend on a linear subset of particles. Simple examples can be found in: H. Touchette, “Methods for calculating nonconcave entropies,” arXiv:1003.0382v1 [cond-mat.stat-mech] 1 Mar 2010.

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,

$$\Omega(E) = \binom{N}{n} = \frac{N!}{n!(N-n)!}, \quad (10)$$

or, equivalently,

$$\begin{aligned} S(E) &= k \ln \Omega(E) = k \ln \left[\frac{N!}{n!(N-n)!} \right] \\ &\approx k [N \ln N - n \ln n - (N-n) \ln(N-n)] \quad \text{by the Stirling approximation} \end{aligned}$$

Thus,

$$\frac{1}{T} = \frac{\partial S}{\partial E} = \frac{dS}{dn} \cdot \frac{dn}{dE} = \frac{1}{\epsilon_0} \cdot k \ln \frac{N-n}{n}, \quad (11)$$

which gives the number of defects as

$$n = \frac{N}{\exp(\epsilon_0/kT) + 1}. \quad (12)$$

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

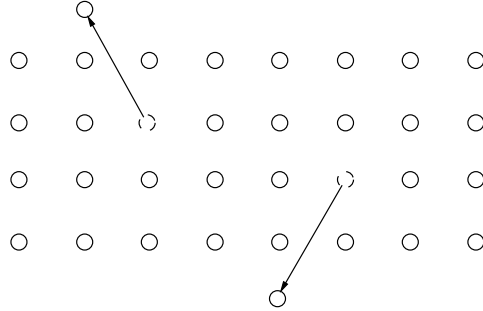


Figure 1: Schottky defects in a crystal lattice.

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

$$S(E) = k \ln \binom{N}{n} \approx kN h_2 \left(\frac{n}{N} \right) = kN h_2 \left(\frac{E}{N\epsilon_0} \right) = kN h_2 \left(\frac{\epsilon}{\epsilon_0} \right), \quad (13)$$

where

$$h_2(x) \triangleq -x \ln x - (1-x) \ln(1-x).$$

Thus, the thermodynamical entropy is intimately related to the Shannon entropy. We will see shortly that this is no coincidence. Note also that $S(E)$ is indeed concave in this example.

□

What happens if we have two independent systems with total energy E , which lie in equilibrium with each other. What is the temperature T ? How does the energy split between them? 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.,

$$S'_1(E_1) - S'_2(E - E_1) = 0 \tag{14}$$

or

$$S'_1(E_1) - S'_2(E_2) = 0 \tag{15}$$

which means

$$\frac{1}{T_1} \equiv S'_1(E_1) = S'_2(E_2) \equiv \frac{1}{T_2}. \tag{16}$$

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

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 . So what happens now?

Similarly as before, since the combined system is isolated, it is governed by the micro-canonical 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 \mathbf{x} (for whatever definition of the microstate vector) with energy $\mathcal{E}(\mathbf{x})$, then the heat bath must have energy $E_0 - \mathcal{E}(\mathbf{x})$ to complement the total energy to E_0 . The number of ways that the heat bath may have energy $E_0 - \mathcal{E}(\mathbf{x})$ is $\Omega_{HB}(E_0 - \mathcal{E}(\mathbf{x}))$, where $\Omega_{HB}(\cdot)$ 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 \mathbf{x} is $\Omega_{HB}(E_0 - \mathcal{E}(\mathbf{x}))$. Since the combined system is governed by the microcanonical ensemble, the probability of this is proportional to $\Omega_{HB}(E_0 - \mathcal{E}(\mathbf{x}))$. More precisely:

$$P(\mathbf{x}) = \frac{\Omega_{HB}(E_0 - \mathcal{E}(\mathbf{x}))}{\sum_{\mathbf{x}'} \Omega_{HB}(E_0 - \mathcal{E}(\mathbf{x}'))}. \quad (17)$$

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

$$\begin{aligned} P(\mathbf{x}) &\propto \Omega_{HB}(E_0 - \mathcal{E}(\mathbf{x})) \\ &= \exp\{S_{HB}(E_0 - \mathcal{E}(\mathbf{x}))/k\} \\ &\approx \exp\left\{\left.\frac{S_{HB}(E_0)}{k} - \frac{1}{k} \frac{\partial S_{HB}(E)}{\partial E}\right|_{E=E_0} \cdot \mathcal{E}(\mathbf{x})\right\} \\ &= \exp\left\{\frac{S_{HB}(E_0)}{k} - \frac{1}{kT} \cdot \mathcal{E}(\mathbf{x})\right\} \\ &\propto \exp\{-\mathcal{E}(\mathbf{x})/(kT)\}. \end{aligned} \quad (18)$$

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

$$\beta = \frac{1}{kT} \quad (19)$$

and so,

$$P(\mathbf{x}) \propto e^{-\beta\mathcal{E}(\mathbf{x})}. \quad (20)$$

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:

$$\boxed{P(\mathbf{x}) = \frac{\exp\{-\beta\mathcal{E}(\mathbf{x})\}}{Z(\beta)}}$$

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

$$Z(\beta) = \sum_{\mathbf{x}} \exp\{-\beta\mathcal{E}(\mathbf{x})\} \quad (21)$$

in the discrete case, or

$$Z(\beta) = \int d\mathbf{x} \exp\{-\beta\mathcal{E}(\mathbf{x})\} \quad (22)$$

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.

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 whenever the Hamiltonian is additive, that is, $\mathcal{E}(\mathbf{x}) = \sum_{i=1}^n \mathcal{E}(x_i)$, 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. And so, by the law of large numbers $\frac{1}{n} \sum_{i=1}^n \mathcal{E}(x_i)$ will tend (almost surely) to $\epsilon = \mathbf{E}\{\mathcal{E}(X_i)\}$. 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 Information Theory is as follows: The microcanonical ensemble is parallel to the uniform distribution over a type class and the canonical ensemble is parallel to a memoryless source.

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:

$$\begin{aligned} & \max H(\mathbf{X}) \\ & \text{s.t. } \sum_{\mathbf{x}} P(\mathbf{x}) \mathcal{E}(\mathbf{x}) = E \quad [\text{or in physicists' notation: } \langle \mathcal{E}(\mathbf{X}) \rangle = E] \end{aligned} \quad (23)$$

By formalizing the equivalent Lagrange problem, where β now plays the role of a Lagrange multiplier:

$$\max \left\{ H(\mathbf{X}) + \beta \left[E - \sum_{\mathbf{x}} P(\mathbf{x}) \mathcal{E}(\mathbf{x}) \right] \right\}, \quad (24)$$

or equivalently,

$$\min \left\{ \sum_{\mathbf{x}} P(\mathbf{x}) \mathcal{E}(\mathbf{x}) - \frac{H(\mathbf{X})}{\beta} \right\} \quad (25)$$

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. Note also that indeed, we have already seen in the Information Theory course that the Gaussian density maximizes the (differential) entropy s.t. 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 \mathbf{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:

$$\lim_{\beta \rightarrow \infty} \frac{\ln Z(\beta)}{\beta} = - \min_{\mathbf{x}} \mathcal{E}(\mathbf{x}) \triangleq -E_{GS} \quad (26)$$

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 log-moment generating function: By taking derivatives of $\ln Z(\beta)$, we can obtain moments (or

cumulants) of $\mathcal{E}(\mathbf{X})$. For the first moment, we have

$$\mathbf{E}\{\mathcal{E}(\mathbf{X})\} \equiv \langle \mathcal{E}(\mathbf{X}) \rangle = \frac{\sum_{\mathbf{x}} \mathcal{E}(\mathbf{x}) e^{-\beta \mathcal{E}(\mathbf{x})}}{\sum_{\mathbf{x}} e^{-\beta \mathcal{E}(\mathbf{x})}} = -\frac{d \ln Z(\beta)}{d\beta}. \quad (27)$$

Similarly, it is easy to show (exercise) that

$$\text{Var}\{\mathcal{E}(\mathbf{X})\} = \langle \mathcal{E}^2(\mathbf{X}) \rangle - \langle \mathcal{E}(\mathbf{X}) \rangle^2 = \frac{d^2 \ln Z(\beta)}{d\beta^2}. \quad (28)$$

This in turn implies that $\frac{d^2 \ln Z(\beta)}{d\beta^2} \geq 0$, which means that $\ln Z(\beta)$ must always be a convex function. Higher order derivatives provide higher order moments.

Next, we look at Z slightly differently than before. Instead of summing $e^{-\beta \mathcal{E}(\mathbf{x})}$ across all states, we go by energy levels (similarly as in the method of types). This amounts to:

$$\begin{aligned} Z(\beta) &= \sum_{\mathbf{x}} e^{-\beta \mathcal{E}(\mathbf{x})} \\ &= \sum_E \Omega(E) e^{-\beta E} \\ &\approx \sum_{\epsilon} e^{ns(\epsilon)/k} \cdot e^{-\beta n\epsilon} \quad \text{recall that } S(n\epsilon) \approx ns(\epsilon) \\ &= \sum_{\epsilon} \exp\{-n\beta[\epsilon - Ts(\epsilon)]\} \\ &\doteq \max_{\epsilon} \exp\{-n\beta[\epsilon - Ts(\epsilon)]\} \\ &= \exp\{-n\beta \min_{\epsilon} [\epsilon - Ts(\epsilon)]\} \\ &\triangleq \exp\{-n\beta[\epsilon^* - Ts(\epsilon^*)]\} \\ &\triangleq e^{-\beta F} \end{aligned} \quad (29)$$

The quantity $f \triangleq \epsilon - Ts(\epsilon)$ is the (per-particle) *free energy*. Similarly, the entire free energy, F , is defined as

$$F = E - TS = -\frac{\ln Z(\beta)}{\beta}. \quad (30)$$

The physical meaning of the free energy is this: A change, or a difference, $\Delta F = F_2 - F_1$, in the free energy means the minimum amount of work it takes to transfer the system from equilibrium state 1 to another equilibrium state 2 in an isothermal (fixed temperature) process. And this minimum is achieved when the process is *quasistatic*, i.e., so slow that

the system is always almost in equilibrium. Equivalently, $-\Delta F$ is the maximum amount of work that that can be exploited from the system, namely, the part of the energy that is *free* for doing work (i.e., not dissipated as heat) in fixed temperature. Again, this maximum is attained by a quasistatic process.

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.,

$$\min \left\{ \sum_{\mathbf{x}} P(\mathbf{x}) \mathcal{E}(\mathbf{x}) - \frac{H(\mathbf{X})}{\beta} \right\}, \quad (31)$$

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

Although 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 $H(\mathbf{X})$. We will also see it shortly in a more formal manner. But what is the information-theoretic analogue of the free energy?

Here is a preliminary guess based on a very rough consideration: The last chain of equalities reminds us what happens when we sum over probabilities type-by-type in IT problems: The exponentials $\exp\{-\beta\mathcal{E}(\mathbf{x})\}$ are analogous (up to a normalization factor) to probabilities, which in the memoryless case, are given by $P(\mathbf{x}) = \exp\{-n[\hat{H} + D(\hat{P}||P)]\}$. Each such probability is weighted by the size of the type class, which as is known from the method of types, is exponentially $e^{n\hat{H}}$, whose physical analogue is $\Omega(E) = e^{ns(\epsilon)/k}$. The product gives $\exp\{-nD(\hat{P}||P)\}$ in IT and $\exp\{-n\beta f\}$ in statistical physics. This suggests

that perhaps the free energy has some analogy with the divergence. Is this true? We will see shortly a somewhat more rigorous argument.

More formally, let us define

$$\phi(\beta) = \lim_{n \rightarrow \infty} \frac{\ln Z(\beta)}{n} \quad (32)$$

and, in order to avoid dragging the constant k , let us define $\Sigma(\epsilon) = \lim_{n \rightarrow \infty} \frac{1}{n} \ln \Omega(n\epsilon) = s(\epsilon)/k$. Then, the above chain of equalities, written slightly differently, gives

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

Thus, $\phi(\beta)$ is (a certain variant of) the *Legendre transform*⁵ of $\Sigma(\epsilon)$. As $\Sigma(\epsilon)$ is (normally) a concave function, then it can readily be shown (exercise) that the inverse transform is:

$$\Sigma(\epsilon) = \min_{\beta} [\beta\epsilon + \phi(\beta)]. \quad (33)$$

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

$$\beta = \Sigma'(\epsilon), \quad (34)$$

or in other words, 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

$$\epsilon = -\phi'(\beta) \quad (35)$$

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

Exercise: Show that the functions $\Sigma'(\cdot)$ and $-\phi'(\cdot)$ are inverses of one another. \square

This establishes a relationship between the typical per-particle energy ϵ and the inverse

⁵More 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.

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:

$$\Sigma(\epsilon) = \beta\epsilon + \phi(\beta) = \phi(\beta) - \beta \cdot \phi'(\beta). \quad (36)$$

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

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

which is exactly the same expression as before, 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 is exponentially $e^{n\bar{H}}$, a result that looks familiar to what we learned from the method of types in IT, using combinatorial arguments for finite-alphabet sequences. Here we got 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:

$$\begin{aligned} 1 &\geq \sum_{\mathbf{x}: \mathcal{E}(\mathbf{x}) \approx n\epsilon} P(\mathbf{x}) = \sum_{\mathbf{x}: \mathcal{E}(\mathbf{x}) \approx n\epsilon} \frac{\exp\{-\beta \sum_i \mathcal{E}(x_i)\}}{Z^n(\beta)} \\ &\approx \sum_{\mathbf{x}: \mathcal{E}(\mathbf{x}) \approx n\epsilon} \exp\{-\beta n\epsilon - n\phi(\beta)\} = \Omega(n\epsilon) \cdot \exp\{-n[\beta\epsilon + \phi(\beta)]\} \end{aligned} \quad (37)$$

which means that $\Omega(n\epsilon) \leq \exp\{n[\beta\epsilon + \phi(\beta)]\}$ for all β , and so,

$$\Omega(n\epsilon) \leq \exp\{n \min_{\beta} [\beta\epsilon + \phi(\beta)]\} = e^{n\Sigma(\epsilon)} = e^{n\bar{H}}. \quad (38)$$

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 $\{\mathbf{x} : \mathcal{E}(\mathbf{x}) \approx n\epsilon\}$ a high-probability event, by the weak law of large numbers. A good reference for further study and from a more general

perspective is:

M. J. W. Hall, “Universal geometric approach to uncertainty, entropy, and information,” *Phys. Rev. A*, vol. 59, no. 4, pp. 2602–2615, April 1999.

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:

$$\begin{aligned}
 P(\mathbf{x}) &= \frac{\exp\{-\beta\mathcal{E}(\mathbf{x})\}}{Z(\beta)} \\
 &= \exp\{-\ln Z(\beta) - \beta\mathcal{E}(\mathbf{x})\} \\
 &= \exp\{\beta[F(\beta) - \mathcal{E}(\mathbf{x})]\}.
 \end{aligned} \tag{39}$$

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

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

or equivalently,

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

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 this can be found in:

1. H. Qian, “Relative entropy: free energy ...,” *Phys. Rev. E*, vol. 63, 042103, 2001.
2. G. B. Bağcı, arXiv:cond-mat/070300v1, 1 Mar. 2007.

Another interesting relation between the divergence and physical quantities is that the divergence is proportional to the dissipated work (=average work – free energy difference) between two equilibrium states at the same temperature but corresponding to two different values of some external control parameter. Details can be found in: R. Kawai, J. M. R. Parrondo, and C. Van den Broeck, “Dissipation: the phase–space perspective,” *Phys. Rev. Lett.*, vol. 98, 080602, 2007.

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 $\lim_{\beta \rightarrow \infty} \frac{\ln Z(\beta)}{\beta} = -E_{GS}$.
2. Generating moments: $\langle \mathcal{E} \rangle = -d \ln Z / d\beta$, $\text{Var}\{\mathcal{E}(\mathbf{X})\} = d^2 \ln Z / d\beta^2 \rightarrow$ 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. $F_Q = F_P + kT \cdot D(Q||P)$.

Exercise: Consider $Z(\beta)$ for an *imaginary temperature* $\beta = j\omega$, where $j = \sqrt{-1}$, and define $z(E)$ as the inverse Fourier transform of $Z(j\omega)$. Show 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 $\int_{E_1}^{E_2} z(E) dE$. \square

Thus, $Z(\cdot)$ can be related to energy enumeration in two different ways: one is by the Legendre transform of $\ln Z$ for real β , and the other is by the inverse Fourier transform of Z for imaginary β . This double connection between Z and Ω is no coincidence, as we shall see later on.

Example – A two level system. Similarly to the earlier example of Schottky defets, 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:

$$P(x) = \frac{e^{-\beta\epsilon_0 x}}{1 + e^{-\beta\epsilon_0}} \quad x \in \{0, 1\}. \quad (40)$$

In this case,

$$\phi(\beta) = \ln(1 + e^{-\beta\epsilon_0}) \quad (41)$$

and

$$\Sigma(\epsilon) = \min_{\beta \geq 0} [\beta\epsilon + \ln(1 + e^{-\beta\epsilon_0})]. \quad (42)$$

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

$$\epsilon - \frac{\epsilon_0 e^{-\beta\epsilon_0}}{1 + e^{-\beta\epsilon_0}} = 0 \quad (43)$$

which gives

$$\beta^*(\epsilon) = \frac{\ln(\epsilon/\epsilon_0 - 1)}{\epsilon_0}. \quad (44)$$

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

$$\Sigma(\epsilon) = \frac{\epsilon}{\epsilon_0} \ln \left(\frac{\epsilon}{\epsilon_0} - 1 \right) + \ln \left[1 + \exp \left\{ -\ln \left(\frac{\epsilon}{\epsilon_0} - 1 \right) \right\} \right], \quad (45)$$

which after a short algebraic manipulation, becomes

$$\Sigma(\epsilon) = h_2 \left(\frac{\epsilon}{\epsilon_0} \right), \quad (46)$$

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

$$\phi(\beta) = \max_{\epsilon} \left[h_2 \left(\frac{\epsilon}{\epsilon_0} \right) - \beta\epsilon \right], \quad (47)$$

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

$$\frac{1}{\epsilon_0} \ln \left[\frac{1 - \epsilon/\epsilon_0}{\epsilon/\epsilon_0} \right] = \beta \quad (48)$$

or equivalently,

$$\epsilon^*(\beta) = \frac{\epsilon_0}{1 + e^{-\beta\epsilon_0}}, \quad (49)$$

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

$$\phi(\beta) = \ln(1 + e^{-\beta\epsilon_0}). \quad \square \quad (50)$$

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 antiparallel 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,

$$P(x) = \frac{e^{\beta Bx}}{2 \cosh(\beta B)}; \quad Z(\beta) = 2 \cosh(\beta B). \quad (51)$$

The net *magnetization* per-spin is defined as

$$m \triangleq \left\langle \frac{1}{n} \sum_{i=1}^n X_i \right\rangle = \langle X_1 \rangle = \frac{\partial \phi}{\partial(\beta B)} = \tanh(\beta B). \quad (52)$$

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

Exercise: Consider a system of n non-interacting particles, each having a quadratic Hamiltonian, $\mathcal{E}(x) = \frac{1}{2}\alpha x^2$, $x \in \mathbb{R}$. Show that here,

$$\Sigma(\epsilon) = \frac{1}{2} \ln \left(\frac{4\pi e\epsilon}{\alpha} \right) \quad (53)$$

and

$$\phi(\beta) = \frac{1}{2} \ln \left(\frac{2\pi}{\alpha\beta} \right). \quad (54)$$

Show that $\beta^*(\epsilon) = 1/(2\epsilon)$ and hence $\epsilon^*(\beta) = 1/(2\beta)$.

2.5 The Energy Equipartition Theorem

From the last exercise, we have learned that for a quadratic Hamiltonian, $\mathcal{E}(x) = \frac{1}{2}\alpha x^2$, we have $\epsilon^*(\beta)$, namely, the average per-particle energy, is given $1/(2\beta) = kT/2$, independently of α . If we have n such quadratic terms, then of course, we end up with $nkT/2$. In the case of the ideal gas, we have 3 such terms (one for each dimension) per particle, thus a total of $3n$ terms, and so, $E = 3nkT/2$, which is exactly what we obtained also in the microcanonical ensemble, which is equivalent (recall that this was obtained then by equating $1/T$ to the derivative of $S(E) = k \ln[\text{const} \times E^{3n/2}]$). In fact, we observe that in the canonical ensemble, whenever we have an Hamiltonian of the form $\frac{\alpha}{2}x_i^2 +$ some arbitrary terms that do not depend on x_i , then x_i is Gaussian (with variance kT/α) and independent of the other guys, i.e., $p(x_i) \propto e^{-\alpha x_i^2/(2kT)}$. Hence it contributes an amount of

$$\left\langle \frac{1}{2}\alpha X_i^2 \right\rangle = \frac{1}{2}\alpha \cdot \frac{kT}{\alpha} = \frac{kT}{2} \quad (55)$$

to the total average energy, 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 3D 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 the sequel, 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:

$$\begin{aligned}
\left\langle \frac{1}{2} a X^2 \right\rangle &= \frac{\int_{-\infty}^{\infty} dx (\alpha x^2 / 2) e^{-\beta \alpha x^2 / 2}}{\int_{-\infty}^{\infty} dx e^{-\beta \alpha x^2 / 2}} \quad \text{num. \& den. have closed forms, but we use another way:} \\
&= -\frac{\partial}{\partial \beta} \ln \left[\int_{-\infty}^{\infty} dx e^{-\beta \alpha x^2 / 2} \right] \\
&= -\frac{\partial}{\partial \beta} \ln \left[\frac{1}{\sqrt{\beta}} \int_{-\infty}^{\infty} d(\sqrt{\beta} x) e^{-\alpha (\sqrt{\beta} x)^2 / 2} \right] \\
&= -\frac{\partial}{\partial \beta} \ln \left[\frac{1}{\sqrt{\beta}} \int_{-\infty}^{\infty} du e^{-\alpha u^2 / 2} \right] \quad \text{The integral is now a constant, independent of } \beta. \\
&= \frac{1}{2} \frac{d \ln \beta}{d \beta} = \frac{1}{2\beta} = \frac{kT}{2}.
\end{aligned}$$

This simple trick, that bypasses the need to calculate integrals, can easily be extended in two directions at least (exercise):

- Let $\mathbf{x} \in \mathbb{R}^n$ and let $\mathcal{E}(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T A \mathbf{x}$, 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:

$$\langle \mathcal{E}(\mathbf{X}) \rangle = \left\langle \frac{1}{2} \mathbf{X}^T A \mathbf{X} \right\rangle = n \cdot \frac{kT}{2}. \quad (56)$$

- 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

$$\langle \mathcal{E}(X) \rangle = \langle \alpha |X|^\theta \rangle = \frac{kT}{\theta}. \quad (57)$$

Moreover, if $\lim_{x \rightarrow \pm\infty} x e^{-\beta \mathcal{E}(x)} = 0$ for all $\beta > 0$, and we denote $\mathcal{E}'(x) \triangleq d\mathcal{E}(x)/dx$, then

$$\langle X \cdot \mathcal{E}'(X) \rangle = kT. \quad (58)$$

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

Example/Exercise – Ideal gas with gravitation: Let

$$\mathcal{E}(x) = \frac{p_x^2 + p_y^2 + p_z^2}{2m} + mgz. \quad (59)$$

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 (Optional)

Looking a bit back, then a brief summary of what we have done thus far, is the following: we started off 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 is called the *Gibbs ensemble*. And there are, of course, additional ensembles based on this principle, depending on what kind of the physical system is under discussion. We will not delve into all of them here because this not a course in physics, after all. We will describe, however, in some level of detail the grand-canonical ensemble.

The fundamental idea is essentially the very same as the one we used to derive the canonical ensemble, we just extend it a little bit: 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 $\Omega_{HB}(E', n')$. A microstate now is a combination (\mathbf{x}, n) , where n is the (variable) number of particles in our subsystem and \mathbf{x} is as before for a given n . From the same considerations as before, whenever our subsystem is in state (\mathbf{x}, n) , the heat bath can be in any one of $\Omega_{HB}(E_0 - \mathcal{E}(\mathbf{x}), n_0 - n)$ microstates of its own. Thus, owing to the microcanonical ensemble,

$$\begin{aligned}
P(\mathbf{x}, n) &\propto \Omega_{HB}(E_0 - \mathcal{E}(\mathbf{x}), n_0 - n) \\
&= \exp\{S_{HB}(E_0 - \mathcal{E}(\mathbf{x}), n_0 - n)/k\} \\
&\approx \exp\left\{\frac{S_{HB}(E_0, n_0)}{k} - \frac{1}{k} \frac{\partial S_{HB}}{\partial E} \cdot \mathcal{E}(\mathbf{x}) - \frac{1}{k} \frac{\partial S_{HB}}{\partial n} \cdot n\right\} \\
&\propto \exp\left\{-\frac{\mathcal{E}(\mathbf{x})}{kT} + \frac{\mu n}{kT}\right\}
\end{aligned} \tag{60}$$

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

$$\mu \triangleq -T \cdot \left. \frac{\partial S_{HB}(E', n')}{\partial n'} \right|_{E'=E_0, n'=n_0} . \tag{61}$$

Thus, we now have the grand-canonical distribution:

$$P(\mathbf{x}, n) = \frac{e^{\beta[\mu n - \mathcal{E}(\mathbf{x})]}}{\Xi(\beta, \mu)}, \tag{62}$$

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

$$\Xi(\beta, \mu) \triangleq \sum_{n=0}^{\infty} e^{\beta\mu n} \sum_{\mathbf{x}} e^{-\beta\mathcal{E}(\mathbf{x})} \triangleq \sum_{n=0}^{\infty} e^{\beta\mu n} Z_n(\beta). \quad (63)$$

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

$$\tilde{\Xi}(\beta, z) = \sum_{n=0}^{\infty} z^n Z_n(\beta). \quad (64)$$

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 $P(\mathbf{x}, n)$ is as $P(n) \cdot P(\mathbf{x}|n)$, where $P(n)$ is proportional to $z^n Z_n(\beta)$ and $P(\mathbf{x}|n)$ corresponds to the canonical ensemble as before.

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

$$\langle n \rangle = \frac{\sum_n n z^n Z_n(\beta)}{\sum_n z^n Z_n(\beta)} = z \cdot \frac{\partial \ln \tilde{\Xi}(\beta, z)}{\partial z}. \quad (65)$$

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, $\beta\mu n + \ln Z_n(\beta)$. 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, I would like to point 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 $1, 2, \dots, r, \dots$. 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:

$$Z_n(\beta) = \sum_{\mathbf{n}: \sum_r n_r = n} \exp\{-\beta \sum_r n_r \epsilon_r\}. \quad (66)$$

The constraint $\sum_r n_r = n$, which accounts for the fact that the total number of particles must be n , causes an extremely severe headache in the calculation. However, if we pass to the grand-canonical ensemble, things becomes extremely easy:

$$\begin{aligned} \tilde{\Xi}(\beta, z) &= \sum_{n \geq 0} z^n \sum_{\mathbf{n}: \sum_r n_r = n} \exp\{-\beta \sum_r n_r \epsilon_r\} \\ &= \sum_{n_1 \geq 0} \sum_{n_2 \geq 0} \dots z^{\sum_r n_r} \exp\{-\beta \sum_r n_r \epsilon_r\} \\ &= \sum_{n_1 \geq 0} \sum_{n_2 \geq 0} \dots \prod_{r \geq 1} z^{n_r} \exp\{-\beta n_r \epsilon_r\} \\ &= \prod_{r \geq 1} \sum_{n_r \geq 0} [z e^{-\beta \epsilon_r}]^{n_r} \end{aligned} \quad (67)$$

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 $\tilde{\Xi} = \prod_r [1/(1 - z e^{-\beta \epsilon_r})]$. 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 $\tilde{\Xi} = \prod_r (1 + z e^{-\beta \epsilon_r})$. 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 in the sequel.

2.7 Gibbs' Inequality, the 2nd Law, and the Data Processing Thm

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 are now going to demonstrate that there are some indications that the answer to this question is affirmative. In particular, we are going to see that there is an intimate

relationship between the second law of thermodynamics and the data processing theorem (DPT), asserting that if $X \rightarrow U \rightarrow V$ 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 IT. Here are just a few points that make this quite clear.

1. *Lossy/lossless source coding:* Consider a source vector $U^N = (U_1, \dots, U_N)$ compressed into a bitstream $X^n = (X_1, \dots, X_n)$ from which the decoder generates a reproduction $V^N = (V_1, \dots, V_N)$ with distortion $\sum_{i=1}^N \mathbf{E}\{d(U_i, V_i)\} \leq ND$. Then, by the DPT, $I(U^N; V^N) \leq I(X^n; X^n) = H(X^n)$, 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 $D = 0$.
2. *Channel coding under bit error probability:* Let $U^N = (U_1, \dots, U_N)$ 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 $P_b = \frac{1}{N} \sum_{i=1}^N \Pr\{V_i \neq U_i\}$ 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 $I(U^N; V^N) = H(U^N) - H(U^N|V^N) \geq N - \sum_{i=1}^N H(U_i|V_i) \geq N - \sum_i h_2(\Pr\{V_i \neq U_i\}) \geq N[1 - h_2(P_b)]$. Thus, for P_b to vanish, the coding rate, N/n should not exceed C .
3. *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 dis-

tributed 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.

4. *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 $NR(D) \leq I(U^N; V^N) \leq I(X^n; Y^n) \leq nC$, 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.
5. *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 RV's, then this statement means that $H(X|V) \geq H(X|U, V)$. If, in addition, $X \rightarrow U \rightarrow V$ 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 $X \rightarrow (U, V) \rightarrow V$ is always a Markov chain.

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

that is

$$Z_i(\beta) = \sum_{\mathbf{x}} e^{-\beta \mathcal{E}_i(\mathbf{x})}, \quad i = 0, 1. \quad (68)$$

The *Gibbs' inequality* asserts that

$$\ln Z_1(\beta) \geq \ln Z_0(\beta) + \beta \langle \mathcal{E}_0(\mathbf{X}) - \mathcal{E}_1(\mathbf{X}) \rangle_0 \quad (69)$$

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:

$$\langle \mathcal{E}_1(\mathbf{X}) - \mathcal{E}_0(\mathbf{X}) \rangle_0 \geq \left[-\frac{\ln Z_1(\beta)}{\beta} \right] - \left[-\frac{\ln Z_0(\beta)}{\beta} \right] \equiv F_1 - F_0, \quad (*) \quad (70)$$

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 $\mathcal{E}_\lambda(\mathbf{x}) = (1 - \lambda)\mathcal{E}_0(\mathbf{x}) + \lambda\mathcal{E}_1(\mathbf{x}) = \mathcal{E}_0(\mathbf{x}) + \lambda[\mathcal{E}_1(\mathbf{x}) - \mathcal{E}_0(\mathbf{x})]$ and using the convexity of the corresponding log-partition function w.r.t. λ . Specifically, let us define the partition function:

$$Z_\lambda(\beta) = \sum_{\mathbf{x}} e^{-\beta \mathcal{E}_\lambda(\mathbf{x})}. \quad (71)$$

Now, since $\mathcal{E}_\lambda(\mathbf{x})$ is affine in λ , then it is easy to show that $d^2 \ln Z_\lambda / d\lambda^2 \geq 0$ (just like this was done with $d^2 \ln Z(\beta) / d\beta^2 \geq 0$ before) and so $\ln Z_\lambda(\beta)$ is convex in λ for fixed β . It follows then that the curve of $\ln Z_\lambda(\beta)$, as a function of λ , must lie above the straight line that is tangent to this curve at $\lambda = 0$ (see Fig. 2), that is, the graph corresponding to the affine function $\ln Z_0(\beta) + \lambda \cdot \left[\frac{\partial \ln Z_\lambda(\beta)}{\partial \lambda} \right]_{\lambda=0}$. In particular, setting $\lambda = 1$, we get:

$$\ln Z_1(\beta) \geq \ln Z_0(\beta) + \left. \frac{\partial \ln Z_\lambda(\beta)}{\partial \lambda} \right|_{\lambda=0}. \quad (72)$$

and the second term is:

$$\left. \frac{\partial \ln Z_\lambda(\beta)}{\partial \lambda} \right|_{\lambda=0} = \frac{\beta \sum_{\mathbf{x}} [\mathcal{E}_0(\mathbf{x}) - \mathcal{E}_1(\mathbf{x})] e^{-\beta \mathcal{E}_0(\mathbf{x})}}{\sum_{\mathbf{x}} e^{-\beta \mathcal{E}_0(\mathbf{x})}} \triangleq \beta \langle \mathcal{E}_0(\mathbf{X}) - \mathcal{E}_1(\mathbf{X}) \rangle_0, \quad (73)$$

Thus, we have obtained

$$\ln \left[\sum_{\mathbf{x}} e^{-\beta \mathcal{E}_1(\mathbf{x})} \right] \geq \ln \left[\sum_{\mathbf{x}} e^{-\beta \mathcal{E}_0(\mathbf{x})} \right] + \beta \langle \mathcal{E}_0(\mathbf{X}) - \mathcal{E}_1(\mathbf{X}) \rangle_0, \quad (74)$$

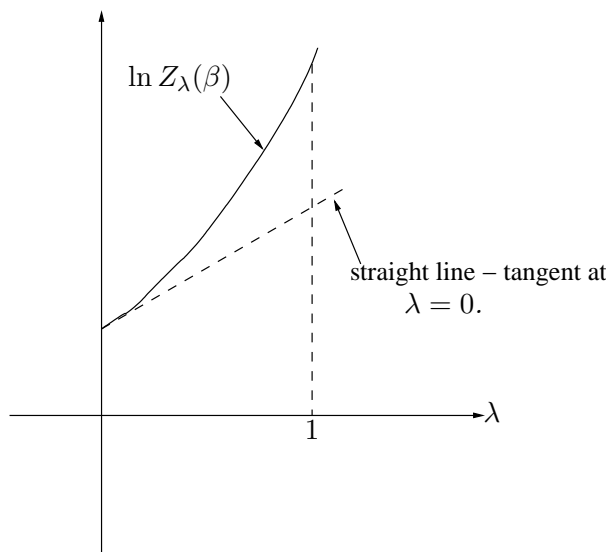


Figure 2: The function $\ln Z_\lambda(\beta)$ is convex in λ and hence lies above its tangent at the origin.

and the proof is complete. In fact, the l.h.s. minus the r.h.s. is nothing but $D(P_0||P_1)$, 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 $\mathcal{E}_0(\mathbf{x})$ is in equilibrium for all $t < 0$, but then, at time $t = 0$, the Hamiltonian changes *abruptly* from the $\mathcal{E}_0(\mathbf{x})$ to $\mathcal{E}_1(\mathbf{x})$ (e.g., by suddenly applying a force on the system), which means that if the system is found at state \mathbf{x} at time $t = 0$, additional energy of $W = \mathcal{E}_1(\mathbf{x}) - \mathcal{E}_0(\mathbf{x})$ 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 \mathbf{x} does not change instantaneously. This average is exactly what we have at the left–hand side eq. (*). 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.⁶ 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

⁶This 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.

is, in fact, a version of the second law of thermodynamics.⁷ 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 (exercise) 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 $(\mathbf{X}, \mathbf{U}, \mathbf{V})$ which form a Markov chain $\mathbf{X} \rightarrow \mathbf{U} \rightarrow \mathbf{V}$. The DPT asserts that $I(\mathbf{X}; \mathbf{U}) \geq I(\mathbf{X}; \mathbf{V})$. We can obtain the DPT as a special case of the Gibbs inequality as follows: For a given realization (\mathbf{u}, \mathbf{v}) of the random variables (\mathbf{U}, \mathbf{V}) , consider the Hamiltonians

$$\mathcal{E}_0(\mathbf{x}) = -\ln P(\mathbf{x}|\mathbf{u}) = -\ln P(\mathbf{x}|\mathbf{u}, \mathbf{v}) \quad (75)$$

and

$$\mathcal{E}_1(\mathbf{x}) = -\ln P(\mathbf{x}|\mathbf{v}). \quad (76)$$

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

$$\langle W \rangle_0 = \langle \mathcal{E}_1(\mathbf{X}) - \mathcal{E}_0(\mathbf{X}) \rangle_0 = \sum_{\mathbf{x}} P(\mathbf{x}|\mathbf{u}, \mathbf{v}) [\ln P(\mathbf{x}|\mathbf{u}) - \ln P(\mathbf{x}|\mathbf{v})] = H(\mathbf{X}|\mathbf{V} = \mathbf{v}) - H(\mathbf{X}|\mathbf{U} = \mathbf{u}) \quad (77)$$

and after further averaging w.r.t. (\mathbf{U}, \mathbf{V}) , the average work becomes $H(\mathbf{X}|\mathbf{V}) - H(\mathbf{X}|\mathbf{U}) = I(\mathbf{X}; \mathbf{U}) - I(\mathbf{X}; \mathbf{V})$. Concerning the free energies, we have

$$Z_0(\beta = 1) = \sum_{\mathbf{x}} \exp\{-1 \cdot [-\ln P(\mathbf{x}|\mathbf{u}, \mathbf{v})]\} = \sum_{\mathbf{x}} P(\mathbf{x}|\mathbf{u}, \mathbf{v}) = 1 \quad (78)$$

and similarly,

$$Z_1(\beta = 1) = \sum_{\mathbf{x}} P(\mathbf{x}|\mathbf{v}) = 1 \quad (79)$$

⁷ From a more general physical perspective, the Jarzynski equality tells that under certain conditions on the test system and the heat bath, and given any protocol $\{\lambda(t)\}$ of changing the control variable λ (of $\mathcal{E}_\lambda(\mathbf{x})$), the work W applied to the system is a RV which satisfies $\langle e^{-\beta W} \rangle = e^{-\beta \Delta F}$. By Jensen’s inequality, $\langle e^{-\beta W} \rangle$ is lower bounded by $e^{-\beta \langle W \rangle}$, and so, we obtain $\langle W \rangle \geq \Delta F$ (which is known as the minimum work principle), now in more generality than in the Gibbs’ inequality, which is limited to the case where $\lambda(t)$ is a step function. At the other extreme, when $\lambda(t)$ changes very slowly, corresponding to a reversible process, W approaches determinism, and then Jensen’s inequality becomes tight, which then gives (in the limit) $W = \Delta F$ with no increase in entropy.

which means that $F_0 = F_1 = 0$, and so $\Delta F = 0$ as well. So by the Gibbs inequality, the average work $I(\mathbf{X}; \mathbf{U}) - I(\mathbf{X}; \mathbf{V})$ cannot be smaller than the free-energy difference, which in this case vanishes, namely, $I(\mathbf{X}; \mathbf{U}) - I(\mathbf{X}; \mathbf{V}) \geq 0$, which is the DPT. Note that in this case, there is a maximum degree of irreversibility: The identity $I(\mathbf{X}; \mathbf{U}) - I(\mathbf{X}; \mathbf{V}) = H(\mathbf{X}|\mathbf{V}) - H(\mathbf{X}|\mathbf{U})$ means that whole work $W = I(\mathbf{X}; \mathbf{U}) - I(\mathbf{X}; \mathbf{V})$ goes for entropy increase $S_1T - S_0T = H(\mathbf{X}|\mathbf{V}) \cdot 1 - H(\mathbf{X}|\mathbf{U}) \cdot 1$, whereas the free energy remains unchanged, as mentioned earlier. Note that the Jarzynski formula (cf. last footnote) holds in this special case, i.e., $\langle e^{-1 \cdot W} \rangle = e^{-1 \cdot \Delta F} = 1$.

The difference between $I(\mathbf{X}; \mathbf{U})$ and $I(\mathbf{X}; \mathbf{V})$, 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 quasistatic 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 a corresponding physical system that violates the second law of thermodynamics, and this, of course, cannot be true.

2.8 Large Deviations Theory and Physics of Information Measures

As I said in the Intro, large deviations theory, the branch of probability theory that deals with exponential decay rates of probabilities of rare events, has strong relations to IT, which we have already seen in the IT course through the eye glasses 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 IT and statistical mechanics goes through rate functions of large deviations theory, or more concretely, Chernoff bounds. This topic is based on the paper: 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.

Let us begin with a very simple question: We have a bunch of i.i.d. RV's X_1, X_2, \dots 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 $E_0 < \langle \mathcal{E}(X) \rangle$ (so that this would be a rare event)? One way to handle this problem, at least in the finite alphabet case, is the method of types. Another method is the Chernoff bound:

$$\begin{aligned} \Pr \left\{ \sum_{i=1}^n \mathcal{E}(X_i) \leq nE_0 \right\} &= \mathbf{E} \mathcal{I} \left\{ \sum_{i=1}^n \mathcal{E}(X_i) \leq nE_0 \right\} \quad \mathcal{I}(\cdot) \text{ denoting the indicator function} \\ &\leq \mathbf{E} \exp \left\{ \beta \left[nE_0 - \sum_{i=1}^n \mathcal{E}(X_i) \right] \right\} \quad \leftarrow \quad \forall \beta \geq 0 : \mathcal{I}\{Z < a\} \leq e^{\beta(a-Z)} \\ &= e^{\beta n E_0} \mathbf{E} \exp \left\{ -\beta \sum_{i=1}^n \mathcal{E}(X_i) \right\} \\ &= e^{\beta n E_0} \mathbf{E} \left\{ \prod_{i=1}^n \exp\{-\beta \mathcal{E}(X_i)\} \right\} \\ &= e^{\beta n E_0} [\mathbf{E} \exp\{-\beta \mathcal{E}(X_1)\}]^n \\ &= \exp \{ n [\beta E_0 + \ln \mathbf{E} \exp\{-\beta \mathcal{E}(X_1)\}] \} \end{aligned}$$

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:

$$\Sigma(E_0) = \min_{\beta \geq 0} [\beta E_0 + \phi(\beta)], \quad (80)$$

where

$$\phi(\beta) = \ln Z(\beta) \quad (81)$$

and

$$Z(\beta) = \mathbf{E} e^{-\beta \mathcal{E}(X)} = \sum_x p(x) e^{-\beta \mathcal{E}(x)}. \quad (82)$$

Rings a bell? 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 $N(x)/N$, where N is independent of x , then imagine that each x gives rise to $N(x)$ microstates $\{x'\}$ with the same energy as x , so that

$$Z(\beta) = \frac{1}{N} \sum_x N(x) e^{-\beta \mathcal{E}(x)} = \frac{1}{N} \sum_{x'} e^{-\beta \mathcal{E}(x')}, \quad (83)$$

and we are back to an ordinary, non-weighted partition function, upto the constant $1/N$, which is absolutely immaterial.

To summarize what we have seen thus far: 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

$$p(x) = \frac{e^{-\beta_1 \mathcal{E}(x)}}{\zeta(\beta_1)} \quad (84)$$

with

$$\zeta(\beta_1) \triangleq \sum_x e^{-\beta_1 \mathcal{E}(x)}. \quad (85)$$

In this case, we have

$$Z(\beta) = \sum_x p(x) e^{-\beta \mathcal{E}(x)} = \frac{\sum_x e^{-(\beta_1 + \beta) \mathcal{E}(x)}}{\zeta(\beta_1)} = \frac{\zeta(\beta_1 + \beta)}{\zeta(\beta_1)}. \quad (86)$$

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

$$\sigma(E_1) \triangleq \min_{\beta \geq 0} [\beta E_1 + \ln \zeta(\beta)] \quad (87)$$

is achieved by $\beta = \beta_1$. Thus, the second bracketted 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 bracketted 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

$$I = -\Sigma(E_0) = \sigma(E_1) - \sigma(E_0) - \beta_1(E_1 - E_0). \quad (88)$$

Note that $I \geq 0$ thanks 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. 3).

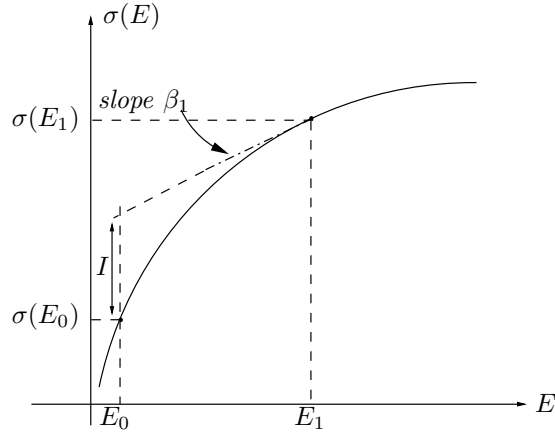


Figure 3: Graphical interpretation of the LD rate function I .

Another look is the following:

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

The last line is exactly what we would have obtained using the method of types. This means that the dominant instance of the large deviations event 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.

Exercise: What happens if $p(x)$ is B–G with an Hamiltonian $\hat{\mathcal{E}}(\cdot)$, different from the one of the LD event? \square

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 RV’s X_1, \dots, X_n , there is also a deterministic sequence of the same length, y_1, \dots, 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, $p(\mathbf{x}|\mathbf{y}) = \prod_{i=1}^n p(x_i|y_i)$. Now, the question is how does the exponential rate function behave if we look at the event

$$\sum_{i=1}^n \mathcal{E}(X_i, y_i) \leq nE_0 \tag{89}$$

where $\mathcal{E}(x, y)$ is a given ‘Hamiltonian’. What is the motivation for this question? Where and when do we encounter such a problem?

Well, there are many examples (cf. the above mentioned paper), but here are two very classical ones, where rate functions of LD 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 $\mathbf{y} = (y_1, \dots, y_n)$ 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 $\{\mathbf{X}(i)\}$ according to $p(\mathbf{x}) = \prod_{i=1}^n p(x_i)$. 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 $\mathbf{X} = (X_1, \dots, X_n)$ would happen to fall at distance $\leq nD$ from \mathbf{y} , i.e., what is the exponential rate of the probability of the event

$$\sum_{i=1}^n d(X_i, y_i) \leq nD? \quad (90)$$

The answer is that it is exponentially about $e^{-nR(D)}$, and that’s why we need slightly more than one over 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 \mathbf{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 $p(\mathbf{x}) = \prod_{i=1}^n p(x_i)$. Let \mathbf{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

$$\sum_{i=1}^n [-\ln W(y_i|X_i)] \leq n[H(Y|X) + \epsilon] \quad \epsilon > 0 \quad (91)$$

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) \mathbf{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 one over 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, and see what happens. 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 what we did before:

$$\begin{aligned}
& \Pr \left\{ \sum_{i=1}^n \mathcal{E}(X_i, y_i) \leq nE_0 \right\} \\
&= \mathbf{E}\mathcal{I} \left\{ \sum_{i=1}^n \mathcal{E}(X_i, y_i) \leq nE_0 \right\} \\
&\leq \mathbf{E} \exp \left\{ \beta \left[nE_0 - \sum_{i=1}^n \mathcal{E}(X_i, y_i) \right] \right\} \\
&= e^{n\beta E_0} \prod_{y \in \mathcal{Y}} \mathbf{E}_y \exp \left\{ -\beta \sum_{i: y_i=y} \mathcal{E}(X_i, y) \right\} \quad \mathbf{E}_y \triangleq \text{expectation under } p(\cdot|y) \\
&= e^{\beta n E_0} \prod_{y \in \mathcal{Y}} [\mathbf{E}_y \exp\{-\beta \mathcal{E}(X, y)\}]^{n(y)} \quad n(y) \triangleq \text{num. of } \{y_i = y\} \\
&= \exp \left\{ n \left[\beta E_0 + \sum_{y \in \mathcal{Y}} q(y) \ln \sum_{x \in \mathcal{X}} p(x|y) \exp\{-\beta \mathcal{E}(x, y)\} \right] \right\}
\end{aligned}$$

and so, the resulting rate function is given by

$$\Sigma(E_0) = \min_{\beta \geq 0} \left[\beta E_0 + \sum_{y \in \mathcal{Y}} q(y) \ln Z_y(\beta) \right] \tag{92}$$

where

$$Z_y(\beta) \triangleq \sum_{x \in \mathcal{X}} p(x|y) \exp\{-\beta \mathcal{E}(x, y)\}. \tag{93}$$

In the rate–distortion example, this tells us that

$$R(D) = - \min_{\beta \geq 0} \left[\beta D + \sum_{y \in \mathcal{Y}} q(y) \ln \sum_{x \in \mathcal{X}} p(x) e^{-\beta d(x,y)} \right]. \quad (94)$$

This is a well–known parametric representation of $R(D)$, which can be obtained via a different route (see, e.g., Gray’s book *Source Coding Theory*), where the minimizing β is known to have the graphical interpretation of the local negative slope (or derivative) of the curve of $R(D)$. In the case of channel capacity, we obtain in a similar manner:

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

Exercise: Show that for channel capacity, the minimizing β is always $\beta^* = 1$. \square

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

$$\sum_{i=1}^n \mathcal{E}(X_i, y_i) = \sum_{y \in \mathcal{Y}} \sum_{i: y_i=y} \mathcal{E}(X_i, y), \quad (95)$$

where now, in each partial sum over $\{i : y_i = y\}$, we have i.i.d. RV’s. 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

$$\bigcap_{y \in \mathcal{Y}} \left\{ \sum_{i: y_i=y} \mathcal{E}(X_i, y) \leq n(y)E_y \right\} \quad (96)$$

where the union is across all “possible partial energy allocations” $\{E_y\}$ which satisfy $\sum_y q(y)E_y \leq E_0$. Note that at least when the X_i ’s 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 $n(y)$ (why?), 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,

$$\begin{aligned}
& \Pr \left\{ \sum_{i=1}^n \mathcal{E}(X_i, y_i) \leq nE_0 \right\} \\
&= \Pr \bigcup_{\{E_y: \sum_y q(y)E_y \leq E_0\}} \bigcap_{y \in \mathcal{Y}} \left\{ \sum_{i: y_i=y} \mathcal{E}(X_i, y) \leq n(y)E_y \right\} \\
&\doteq \max_{\{E_y: \sum_y q(y)E_y \leq E_0\}} \prod_{y \in \mathcal{Y}} \Pr \left\{ \sum_{i: y_i=y} \mathcal{E}(X_i, y) \leq n(y)E_y \right\} \\
&\doteq \max_{\{E_y: \sum_y q(y)E_y \leq E_0\}} \prod_{y \in \mathcal{Y}} \exp \left\{ n(y) \min_{\beta_y \geq 0} [\beta_y E_y + \ln Z_y(\beta)] \right\} \\
&= \exp \left\{ n \cdot \max_{\{E_y: \sum_y q(y)E_y \leq E_0\}} \sum_{y \in \mathcal{Y}} q(y) \Sigma_y(E_y) \right\}
\end{aligned}$$

where we have defined

$$\Sigma_y(E_y) \triangleq \min_{\beta_y \geq 0} [\beta_y E_y + \ln Z_y(\beta)]. \quad (97)$$

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

$$\max_{\{E_y: \sum_y q(y)E_y \leq E_0\}} \sum_{y \in \mathcal{Y}} q(y) \Sigma_y(E_y). \quad (98)$$

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

$$\boxed{\Sigma(E_0) = \max_{\{E_y: \sum_y q(y)E_y \leq E_0\}} \sum_{y \in \mathcal{Y}} q(y) \Sigma_y(E_y)}$$

A few comments:

1. In the paper 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. What makes it interesting 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:

$$\begin{aligned}
& \max_{\{E_y: \sum_y q(y)E_y \leq E_0\}} \sum_y q(y) \Sigma_y(E_y) \\
= & \max_{\{E_y: \sum_y q(y)E_y \leq E_0\}} \sum_y q(y) \min_{\beta_y} [\beta_y E_y + \ln Z_y(\beta_y)] \\
\leq & \max_{\{E_y: \sum_y q(y)E_y \leq E_0\}} \sum_y q(y) [\beta^* E_y + \ln Z_y(\beta^*)] \quad \text{where } \beta^* \text{ achieves } \Sigma(E_0) \\
\leq & \max_{\{E_y: \sum_y q(y)E_y \leq E_0\}} [\beta^* E_0 + \sum_y q(y) \ln Z_y(\beta^*)] \quad \text{because } \sum_y q(y)E_y \leq E_0 \\
= & \beta^* E_0 + \sum_y q(y) \ln Z_y(\beta^*) \quad \text{the bracketted expression no longer depends on } \{E_y\} \\
= & \Sigma(E_0).
\end{aligned}$$

Both inequalities become equalities if $\{E_y\}$ would be allocated such that:⁸ (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 bunch 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 LD event is the one where the contributions

⁸Exercise: show that there exists an energy allocation $\{E_y\}$ that satisfies both (i) and (ii) at the same time.

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.

Now, let us look a bit more closely on the rate–distortion function:

$$R(D) = - \min_{\beta \geq 0} \left[\beta D + \sum_{y \in \mathcal{Y}} q(y) \ln \sum_{x \in \mathcal{X}} p(x) e^{-\beta d(x,y)} \right]. \quad (99)$$

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

$$D = - \frac{\partial}{\partial \beta} \sum_y q(y) \ln \sum_x p(x) e^{-\beta d(x,y)} \quad (100)$$

or equivalently,

$$D = \sum_y q(y) \cdot \frac{\sum_x p(x) d(x,y) e^{-\beta d(x,y)}}{\sum_x p(x) \cdot e^{-\beta d(x,y)}}. \quad (101)$$

The Legendre transform relation between the log–partition function and $R(D)$ induces a one–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:

$$P_\beta(x, y) = q(y) \cdot P_\beta(x|y) = q(y) \cdot \frac{p(x) e^{-\beta d(x,y)}}{\sum_{x'} p(x') e^{-\beta d(x',y)}}. \quad (102)$$

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

$$- R(D_\beta) = \beta D_\beta + \sum_y q(y) \ln \sum_x p(x) e^{-\beta d(x,y)}. \quad (103)$$

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

$$\begin{aligned} R(D_\beta) &= - \int_0^\beta d\hat{\beta} \cdot \left(D_{\hat{\beta}} + \hat{\beta} \cdot \frac{dD_{\hat{\beta}}}{d\hat{\beta}} - D_{\hat{\beta}} \right) \\ &= - \int_{D_0}^{D_\beta} \hat{\beta} \cdot dD_{\hat{\beta}}, \end{aligned} \quad (104)$$

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 $R_Q(D) = 0$. This is exactly analogous to the thermodynamic equation $S = \int dQ/T$ (following from $1/T = dS/dQ$), that builds up the entropy from the cumulative heat. Note that the last equation, in its differential form, reads $dR(D_\beta) = -\beta dD_\beta$, 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:

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

where $\text{Var}_{\hat{\beta}}\{\cdot\}$ and $\text{mmse}_{\hat{\beta}}\{\cdot\}$ are taken w.r.t. $P_{\hat{\beta}}(x,y)$. 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 . In those cases where an exact expression for $R(D)$ is hard to obtain, this opens the door to upper and lower bounds on $R(D)$, which are based on upper and lower bounds on the MMSE, offered by the plethora of bounds available in estimation theory.

Exercise: Show that $D_\beta = D_0 - \int_0^\beta d\hat{\beta} \cdot \text{mmse}_{\hat{\beta}}\{d(X,Y)|Y\}$.

Finally, a word about the high–resolution regime. The partition function of each y is

$$Z_y(\beta) = \sum_x p(x) e^{-\beta d(x,y)}, \quad (105)$$

or, in the continuous case,

$$Z_y(\beta) = \int_{\mathbb{R}} dx p(x) e^{-\beta d(x,y)}. \quad (106)$$

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 res), which means a large value of β (why?). Then,

$$\begin{aligned} Z_y(\beta) &= \frac{1}{2A} \int_{-A}^{+A} dx e^{-\beta|x-y|^\theta} \\ &\approx \frac{1}{2A} \int_{-\infty}^{+\infty} dx e^{-\beta|x-y|^\theta} \quad (\text{large } \beta) \\ &= \frac{1}{2A} \int_{-\infty}^{+\infty} dx e^{-\beta|x|^\theta} \quad (\text{the integral is independent of } y) \end{aligned}$$

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

$$D = -\frac{\partial}{\partial \beta} \ln \left[\frac{1}{2A} \int_{-\infty}^{+\infty} dx e^{-\beta|x|^\theta} \right] \quad (107)$$

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:

$$D(R) = -\frac{D'(R)}{\theta} \quad (108)$$

whose solution is

$$D(R) = D_0 e^{-\theta R}. \quad (109)$$

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

3 Analysis Tools and Asymptotic Methods

3.1 Introduction

So far we have dealt with relatively simple situations where the Hamiltonian is additive, the resulting B–G distribution is then i.i.d., and everything is very nice, easy, and simple. But this is seldom the case in reality. Most models in physics, including those that will prove relevant for IT, as we shall see in the sequel, are way 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 other cases, analysis is possible, but it requires some more powerful mathematical tools and techniques, which suggest at least some asymptotic approximations. These are tools and techniques that we must acquaint ourselves with. So the purpose of this part of the course is to prepare these tools, before we can go on to the more challenging settings that are waiting for us.

Before diving into the technical stuff, I’ll first try to give the flavor of the things I am going to talk about, and I believe the best way to do this is through an example. In quantum mechanics, as its name suggests, several physical quantities do not really take on values in the continuum of real numbers, but only values in a discrete set, depending on the conditions of the system. One such quantized physical quantity is energy (for example, the energy of light comes in quanta of $h\nu$, where ν is frequency). Suppose we have a system of n mobile particles (gas), whose energies take on discrete values, denoted $\epsilon_0 < \epsilon_1 < \epsilon_2 < \dots$. If the particles were not interacting, then the partition function would have been given by

$$\left[\sum_{r \geq 0} e^{-\beta \epsilon_r} \right]^n = \sum_{r_1 \geq 0} \sum_{r_2 \geq 0} \dots \sum_{r_n \geq 0} \exp \left\{ -\beta \sum_{i=1}^n \epsilon_{r_i} \right\} = \sum_{\mathbf{n}: \sum_r n_r = n} \frac{n!}{\prod_r n_r!} \cdot \exp \left\{ -\beta \sum_r n_r \epsilon_r \right\}. \quad (110)$$

However, since the particles are indistinguishable, then permutations among them are not considered distinct physical states (see earlier discussion on the ideal gas), and so, the combinatorial factor $n! / \prod_r n_r!$, that counts these permutations, should be eliminated. In other

words, the correct partition function should be

$$Z_n(\beta) = \sum_{\mathbf{n}: \sum_r n_r = n} \exp \left\{ -\beta \sum_r n_r \epsilon_r \right\}. \quad (111)$$

The problem is that this partition function is hard to calculate in closed form: the headache is caused mostly because of the constraint $\sum_r n_r = n$. However, if we define a corresponding generating function

$$\Xi(\beta, z) = \sum_{n \geq 0} z^n Z_n(\beta), \quad (112)$$

which is like the z -transform of $\{Z_n(\beta)\}$, this is easy to work with, because

$$\Xi(\beta, z) = \sum_{n_1 \geq 0} \sum_{n_2 \geq 0} \dots z^{\sum_r n_r} \exp \left\{ -\beta \sum_r n_r \epsilon_r \right\} = \prod_r \left[\sum_{n_r} (ze^{-\beta \epsilon_r})^{n_r} \right]. \quad (113)$$

Splendid, but we still want to obtain $Z_n(\beta)$...

The idea is to apply the inverse z -transform:

$$Z_n(\beta) = \frac{1}{2\pi j} \oint_{\mathcal{C}} \frac{\Xi(\beta, z) dz}{z^{n+1}} = \frac{1}{2\pi j} \oint_{\mathcal{C}} \Xi(\beta, z) e^{-(n+1) \ln z} dz, \quad (114)$$

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 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: Suppose we have a family of sources indexed by some parameter θ , say, Bernoulli with parameter $\theta \in [0, 1]$, i.e.,

$$P_\theta(\mathbf{x}) = (1 - \theta)^{N-n} \theta^n, \quad \mathbf{x} \in \{0, 1\}^N; \quad n = \# \text{ of 1's} \quad (115)$$

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

$$P(\mathbf{x}) = \int_0^1 d\theta w(\theta) P_\theta(\mathbf{x}) = \int_0^1 d\theta w(\theta) e^{Nh(\theta)} \quad (116)$$

where

$$h(\theta) = \ln(1 - \theta) + q \ln \left(\frac{\theta}{1 - \theta} \right); \quad q = \frac{n}{N}. \quad (117)$$

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.

This is exactly the point 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 (plus 2nd order behavior) of an integral of the form

$$\int_{\mathcal{P}} g(z) e^{nf(z)} dz \quad \mathcal{P} \text{ is a path in the complex plane.} \quad (118)$$

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*. The material here is taken mostly from de Bruijn's book, which appears in the bibliographical list.

3.2 The Laplace Method

Consider first an integral of the form:

$$F_n \triangleq \int_{-\infty}^{+\infty} e^{nh(x)} dx, \quad (119)$$

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 even infinitely many terms provided that the tail of this series decays sufficiently rapidly. Since the integral is, after all, 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. $h(x) < 0 \quad \forall x \neq 0$, 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 . W.l.o.g., let this sufficiently large n be $n = 1$, i.e., $\int_{-\infty}^{+\infty} e^{h(x)} dx < \infty$.
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:

$$\begin{aligned} \int_{|x| \geq \delta} e^{nh(x)} dx &= \int_{|x| \geq \delta} dx e^{(n-1)h(x)+h(x)} \\ &\leq \int_{|x| \geq \delta} dx e^{-(n-1)\eta(\delta)+h(x)} \\ &\leq e^{-(n-1)\eta(\delta)} \cdot \int_{-\infty}^{+\infty} e^{h(x)} dx \rightarrow 0 \quad \text{exponentially fast} \end{aligned}$$

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 take 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.

$$\left| h(x) - \frac{1}{2}h''(0)x^2 \right| \leq \epsilon x^2 \quad \forall |x| \leq \delta. \quad (120)$$

Thus, this integral is sandwiched as follows:

$$\int_{-\delta}^{+\delta} \exp \left\{ \frac{n}{2} (h''(0) - \epsilon) x^2 \right\} dx \leq \int_{-\delta}^{+\delta} e^{nh(x)} dx \leq \int_{-\delta}^{+\delta} \exp \left\{ \frac{n}{2} (h''(0) + \epsilon) x^2 \right\} dx. \quad (121)$$

The right-most side is further upper bounded by

$$\int_{-\infty}^{+\infty} \exp \left\{ \frac{n}{2} (h''(0) + \epsilon) x^2 \right\} dx \quad (122)$$

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

$$\sqrt{\frac{2\pi}{(|h''(0)| - \epsilon)n}}. \quad (123)$$

The left-most side of the earlier sandwich is further lower bounded by

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

where the notation $A_n \sim B_n$ means that $\lim_{n \rightarrow \infty} A_n/B_n = 1$. Since ϵ and hence δ can be made arbitrary small, we find that

$$\int_{-\delta}^{+\delta} e^{nh(x)} dx \sim \sqrt{\frac{2\pi}{|h''(0)|n}}. \quad (124)$$

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:

$$\int_{-\infty}^{+\infty} e^{nh(x)} dx \sim \sqrt{\frac{2\pi}{|h''(0)|n}}. \quad (125)$$

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

$$\int_{-\infty}^{+\infty} e^{nh(x)} dx \sim e^{nh(x_0)} \cdot \sqrt{\frac{2\pi}{|h''(x_0)|n}} \quad (126)$$

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:

$$\boxed{\int_{-\infty}^{+\infty} g(x)e^{nh(x)} dx \sim g(x_0)e^{nh(x_0)} \cdot \sqrt{\frac{2\pi}{|h''(x_0)|n}}$$

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*.

3.3 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

$$F_n = \int_{\mathcal{P}} e^{nh(z)} dz \quad \text{or more generally} \quad F_n = \int_{\mathcal{P}} g(z) e^{nh(z)} dz \quad (127)$$

where $z = x + jy$ is a complex variable ($j = \sqrt{-1}$), and \mathcal{P} is a certain path (or curve) 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 (differentiable as many times as we want).

The first observation, in this case, is that the value of the integral depends actually 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 $\mathcal{P} \cup \mathcal{P}'$. Then, the integral of $e^{nh(z)}$ over the closed path $\mathcal{P} \cup \mathcal{P}'$ (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

$$h(z) = u(z) + jv(z) = u(x, y) + jv(x, y), \quad (128)$$

where u and v are real functions. If h is analytic, the following relationships (a.k.a. the Cauchy–Riemann conditions)⁹ between the partial derivatives of u and v must hold:

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}; \quad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}. \quad (129)$$

⁹This 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., $f'(z) = \lim_{\delta \rightarrow 0} [f(z + \delta) - f(z)]/\delta = \lim_{\delta \rightarrow 0} [f(z + j\delta) - f(z)]/(j\delta)$, where δ goes to zero along the reals.

Taking the second order partial derivative of u :

$$\frac{\partial^2 u}{\partial x^2} = \frac{\partial^2 v}{\partial x \partial y} = \frac{\partial^2 v}{\partial y \partial x} = -\frac{\partial^2 u}{\partial y^2} \quad (130)$$

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,

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0, \quad (131)$$

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

$$\left| \exp\{nh(z)\} \right| = \exp[n\operatorname{Re}\{h(z)\}] = e^{nu(z)}. \quad (132)$$

Of course, if $h'(z) = 0$ at some $z = z_0$, then $\overline{u}'(z_0) = 0$ too, and then z_0 is a saddle point of $|e^{nh(z)}|$. Thus, zero-derivative points of h are saddle points.

Another way to see this is the following: 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,

$$\begin{aligned} \frac{1}{2\pi} \int_{-\pi}^{\pi} f(z_0 + Re^{j\theta}) d\theta &= \frac{1}{2\pi j} \int_{-\pi}^{\pi} \frac{f(z_0 + Re^{j\theta}) j Re^{j\theta} d\theta}{Re^{j\theta}} \\ &= \frac{1}{2\pi j} \oint_{z=z_0+Re^{j\theta}} \frac{f(z_0 + Re^{j\theta}) d(z_0 + Re^{j\theta})}{Re^{j\theta}} \\ &= \frac{1}{2\pi j} \oint_{z=z_0+Re^{j\theta}} \frac{f(z) dz}{z - z_0} = f(z_0). \end{aligned} \quad (133)$$

and so,

$$|f(z_0)| \leq \frac{1}{2\pi} \int_{-\pi}^{\pi} \left| f(z_0 + Re^{j\theta}) \right| d\theta \quad (134)$$

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:

$$h(z) \approx h(z_0) + \frac{1}{2}h''(z_0)(z - z_0)^2 \quad (\text{recall that } h'(z_0) = 0.) \quad (135)$$

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 case. This means that

$$\arg\{h''(z_0)\} + 2\arg(z - z_0) = \pi \quad (136)$$

or equivalently

$$\arg(z - z_0) = \frac{\pi - \arg\{h''(z_0)\}}{2} \triangleq \theta. \quad (137)$$

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).¹⁰

So pictorially, what we are going to do is choose a path \mathcal{P} from A to B , which will be composed of three parts (see Fig. 4): The parts $A \rightarrow A'$ and $B' \rightarrow B$ 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:

$$F_n = \int_A^{A'} e^{nh(z)} dz + \int_{A'}^{B'} e^{nh(z)} dz + \int_{B'}^B e^{nh(z)} dz. \quad (138)$$

¹⁰Note 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.

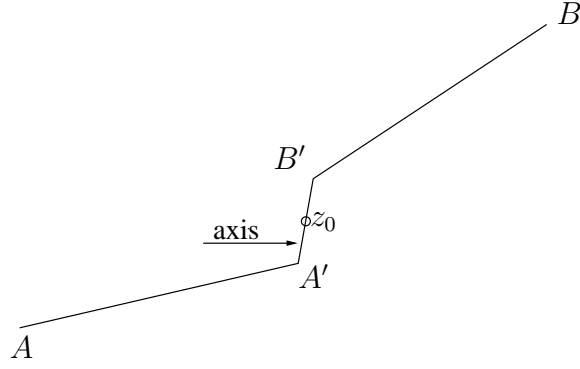


Figure 4: A path \mathcal{P} from A to B , passing via z_0 along the axis.

As for the first and the third terms,

$$\left| \left(\int_A^{A'} + \int_{B'}^B \right) dz e^{nh(z)} \right| \leq \left(\int_A^{A'} + \int_{B'}^B \right) dz |e^{nh(z)}| = \left(\int_A^{A'} + \int_{B'}^B \right) dz e^{nu(z)} \quad (139)$$

whose contribution is negligible compared to $e^{nu(z_0)}$, just like the tails in the Laplace method.

As for the middle integral,

$$\int_{A'}^{B'} e^{nh(z)} dz \approx e^{nh(z_0)} \int_{A'}^{B'} \exp\{nh''(z_0)(z - z_0)^2/2\} dz. \quad (140)$$

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

$$\int_{A'}^{B'} \exp\{nh''(z_0)(z - z_0)^2/2\} dz = e^{j\theta} \sqrt{\frac{2\pi}{n|h''(z_0)|}} \quad (141)$$

where the factor $e^{j\theta}$ is due to the change of variable ($dz = e^{j\theta} dx$). Thus,

$$F_n \sim e^{j\theta} \cdot e^{nh(z_0)} \sqrt{\frac{2\pi}{n|h''(z_0)|}}, \quad (142)$$

and slightly more generally,

$$\boxed{\int_{\mathcal{P}} g(z) e^{nh(z)} dz \sim e^{j\theta} g(z_0) e^{nh(z_0)} \sqrt{\frac{2\pi}{n|h''(z_0)|}}}$$

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} surround 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

$$\max_{z \in \mathcal{C}} |e^{nh(z)}| \geq |e^{nh(z_3)}| > |e^{nh(z_1)}| = \max_{z \in \mathcal{P}_1 \cup \mathcal{P}_2} |e^{nh(z)}| \quad (143)$$

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 saddlepoint. Consider, for example, the integral $F_n = \int_{a+j0}^{a+j2\pi} e^{nz} dz$. 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 saddlepoint (and only this one) along which this saddle point is dominant.

Let us look now at a few 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 $Z(\beta) = \int_0^\infty dE \Omega(E) e^{-\beta E}$, 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.,

$$\Omega(E) = \frac{1}{2\pi j} \int_{\gamma-j\infty}^{\gamma+j\infty} e^{\beta E} Z(\beta) d\beta, \quad (144)$$

where the integration in the complex plane is along the vertical line $\text{Re}(\beta) = \gamma$, which is

chosen to the right of all singularity points of $Z(\beta)$. In the large n limit, this becomes

$$\Omega(E) = \frac{1}{2\pi j} \int_{\gamma-j\infty}^{\gamma+j\infty} e^{n[\beta\epsilon + \phi(\beta)]} d\beta, \quad (145)$$

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, $\Omega(E) \doteq \exp\{n \min_{\beta \in \mathbb{R}}[\beta\epsilon + \phi(\beta)]\} = e^{n\Sigma(\epsilon)}$, as we have seen before.

Example 2 – size of a type class. Here is a question which we know how to answer using the method of types. Among all binary sequences of length N , how many have n 1's and $(N - n)$ 0's?

$$\begin{aligned} M_n &= \sum_{\mathbf{x} \in \{0,1\}^N} \mathcal{I} \left\{ \sum_{i=1}^N x_i = n \right\} \\ &= \sum_{x_1=0}^1 \dots \sum_{x_N=0}^1 \mathcal{I} \left\{ \sum_{i=1}^N x_i = n \right\} \\ &= \sum_{x_1=0}^1 \dots \sum_{x_N=0}^1 \frac{1}{2\pi} \int_0^{2\pi} d\omega \exp \left\{ j\omega \left(n - \sum_{i=1}^N x_i \right) \right\} \\ &= \int_0^{2\pi} \frac{d\omega}{2\pi} \sum_{x_1=0}^1 \dots \sum_{x_N=0}^1 \exp \left\{ j\omega \left(n - \sum_{i=1}^N x_i \right) \right\} \\ &= \int_0^{2\pi} \frac{d\omega}{2\pi} e^{j\omega n} \prod_{i=1}^N \left[\sum_{x_i=0}^1 e^{-j\omega x_i} \right] \\ &= \int_0^{2\pi} \frac{d\omega}{2\pi} e^{j\omega n} (1 + e^{-j\omega})^N \\ &= \int_0^{2\pi} \frac{d\omega}{2\pi} \exp \{ N[j\omega\alpha + \ln(1 + e^{-j\omega})] \} \quad \alpha \triangleq \frac{n}{N} \\ &= \int_0^{2\pi j} \frac{dz}{2\pi j} \exp \{ N[z\alpha + \ln(1 + e^{-z})] \} \quad j\omega \longrightarrow z \end{aligned} \quad (146)$$

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 serious 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 we do 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 (exercise), but this time we get:

$$M_n = \int_{\theta - j\pi}^{\theta + j\pi} \frac{dz}{2\pi j} \exp\{N[z\alpha + \ln(1 + e^{-z})]\}, \quad (147)$$

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 (exercise: please verify). Moreover, the vertical direction of the integration is also the direction of the axis of z_0 (exercise: verify this too), so now everything is fine. 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 Stirling approximation in the more refined formula.

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

$$Z(\beta) = \sum_{\mathbf{x}} e^{-\beta \epsilon_0 \sum_i x_i}, \quad (148)$$

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

$$\begin{aligned} M_n &= \frac{1}{2\pi j} \oint (1 + s)^N s^{-n-1} ds \\ &= \frac{1}{2\pi j} \oint \exp\{N[\ln(1 + s) - \alpha \ln s]\} ds \end{aligned} \quad (149)$$

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. Let us compute the surface area of an n -dimensional sphere with radius nR :

$$\begin{aligned} S_n &= \int_{\mathbb{R}^n} d\mathbf{x} \delta \left(nR - \sum_{i=1}^n x_i^2 \right) \\ &= e^{n\alpha R} \int_{\mathbb{R}^n} d\mathbf{x} e^{-\alpha \sum_i x_i^2} \cdot \delta \left(nR - \sum_{i=1}^n x_i^2 \right) \quad (\alpha > 0 \text{ to be chosen later.}) \\ &= e^{n\alpha R} \int_{\mathbb{R}^n} d\mathbf{x} e^{-\alpha \sum_i x_i^2} \int_{-\infty}^{+\infty} \frac{d\theta}{2\pi} e^{j\theta(nR - \sum_i x_i^2)} \\ &= e^{n\alpha R} \int_{-\infty}^{+\infty} \frac{d\theta}{2\pi} e^{j\theta nR} \int_{\mathbb{R}^n} d\mathbf{x} e^{-(\alpha + j\theta) \sum_i x_i^2} \\ &= e^{n\alpha R} \int_{-\infty}^{+\infty} \frac{d\theta}{2\pi} e^{j\theta nR} \left[\int_{\mathbb{R}} dx e^{-(\alpha + j\theta)x^2} \right]^n \\ &= e^{n\alpha R} \int_{-\infty}^{+\infty} \frac{d\theta}{2\pi} e^{j\theta nR} \left(\frac{\pi}{\alpha + j\theta} \right)^{n/2} \\ &= \frac{\pi^{n/2}}{2\pi} \int_{-\infty}^{+\infty} d\theta \exp \left\{ n \left[(\alpha + j\theta)R - \frac{1}{2} \ln(\alpha + j\theta) \right] \right\} \\ &= \frac{\pi^{n/2}}{2\pi} \int_{\alpha - j\infty}^{\alpha + j\infty} dz \exp \left\{ n \left[zR - \frac{1}{2} \ln z \right] \right\}. \end{aligned} \quad (150)$$

So here $h(z) = zR - \frac{1}{2} \ln z$ and the integration is along an arbitrary vertical straight line parametrized by α . We will choose this straight line to pass thru the saddle point $z_0 = \frac{1}{2R}$ (exercise: show that this is indeed the highest point on the path). Now, $h(z_0) = \frac{1}{2} \ln(2\pi eR)$, just like the differential entropy of a Gaussian RV (is this a coincidence?). \square

Comment: In these examples, we used an additional trick: whenever we had to deal with an ‘ugly’ function like the δ function, we presented it as an inverse transform of a ‘nice’ function, and then changed the order of integrations/summations. This idea will be repeated in the sequel. It is used very frequently by physicists.

3.4 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 a bunch of *random* parameters (to model disorder), and then Z , or $\ln Z$, becomes, of course, a random variable as well. Further, it turns out that more often than not, the RV $\frac{1}{n} \ln Z$ exhibits a concentration property, or in the jargon of physicists, a *self-averaging* property: in the thermodynamic limit of $n \rightarrow \infty$, 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, heuristic 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, live in harmony

with intuition, or make good fit to experimental results and/or simulation results. 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, I believe 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. 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 all this, since this method is so widely used, it would be inappropriate to completely ignore it in a course of this kind, and therefore, we will devote to the replica method at least a short period of time, presenting it in the general level, up to a certain point. However, we will not use the replica method elsewhere in this course.

Consider then the calculation of $\mathbf{E} \ln Z$. 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 , $\mathbf{E} Z^m$, life would have been much easier, because integer moments of sums, are sums of products. Is there a way then that we can relate moments $\mathbf{E} Z^m$ to $\mathbf{E} \ln Z$? The answer is, in principle, affirmative if **real**, rather than just integer, moments are allowed. These could be related via the simple relation

$$\mathbf{E} \ln Z = \lim_{m \rightarrow 0} \frac{\mathbf{E} Z^m - 1}{m} = \lim_{m \rightarrow 0} \frac{\ln \mathbf{E} Z^m}{m} \quad (151)$$

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 $\mathbf{E} Z^m$, for positive integer m , and obtain an expression which is a function of m . Once this has been done, now *forget* that m is an integer, and think of it as a *real* variable. Finally, use the above identity, taking the limit of $m \rightarrow 0$.

Beyond the technicality of interchanging the expectation operator with the limit, which is, after all, OK 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 $f(m) \triangleq \mathbf{E}Z^m$, 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 pass through given points at 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 is simply ignored...

After this introduction, let us now present the replica method on a concrete example, which is essentially taken from the book by Mézard and Montanari. In this example, $Z = \sum_{i=1}^{2^n} e^{-\beta E_i}$, where $\{E_i\}_{i=1}^{2^n}$ are i.i.d. RV's. In the sequel, we will work with this model quite a lot, after we see why, when and where it is relevant. It is called the *random energy model* (REM). But for now, this is just a technical example on which we demonstrate the replica method. As the replica method suggests, let's first look at the integer moments. First, what we have is:

$$Z^m = \left[\sum_{i=1}^{2^n} e^{-\beta E_i} \right]^m = \sum_{i_1=1}^{2^n} \dots \sum_{i_m=1}^{2^n} \exp\left\{-\beta \sum_{a=1}^m E_{i_a}\right\}. \quad (152)$$

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 $\mathbf{i} = (i_1, \dots, i_m)$, 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:

$$\begin{aligned}
Z^m &= \sum_{i_1=1}^{2^n} \dots \sum_{i_m=1}^{2^n} \exp \left\{ -\beta \sum_{a=1}^m E_{i_a} \right\} \\
&= \sum_{\mathbf{i}} \exp \left\{ -\beta \sum_{a=1}^m \sum_{j=1}^{2^n} \mathcal{I}(i_a = j) E_j \right\} \quad \mathcal{I}(\cdot) = \text{indicator function} \\
&= \sum_{\mathbf{i}} \exp \left\{ -\beta \sum_{j=1}^{2^n} \sum_{a=1}^m \mathcal{I}(i_a = j) E_j \right\} \\
&= \sum_{\mathbf{i}} \prod_{j=1}^{2^n} \exp \left\{ -\beta \sum_{a=1}^m \mathcal{I}(i_a = j) E_j \right\}
\end{aligned}$$

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:

$$\begin{aligned}
\mathbf{E} Z^m &= \sum_{\mathbf{i}} \mathbf{E} \prod_{j=1}^{2^n} \exp \left\{ -\beta \sum_{a=1}^m \mathcal{I}(i_a = j) E_j \right\} \\
&= \sum_{\mathbf{i}} \prod_{j=1}^{2^n} \exp \left\{ \frac{\beta^2 n J^2}{4} \sum_{a,b=1}^m \mathcal{I}(i_a = j) \mathcal{I}(i_b = j) \right\} \quad \text{using independence and Gaussianity} \\
&= \sum_{\mathbf{i}} \exp \left\{ \frac{\beta^2 n J^2}{4} \sum_{a,b=1}^m \sum_{j=1}^{2^n} \mathcal{I}(i_a = j) \mathcal{I}(i_b = j) \right\} \\
&= \sum_{\mathbf{i}} \exp \left\{ \frac{\beta^2 n J^2}{4} \sum_{a,b=1}^m \mathcal{I}(i_a = i_b) \right\}.
\end{aligned}$$

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 \mathbf{i} only via Q . Let $N_n(Q)$ denote the number of configurations $\{\mathbf{i}\}$ 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 $N_n(Q) \doteq e^{ns(Q)}$, and so

$$\mathbf{E}Z^m \doteq \sum_Q e^{ng(Q)} \quad (153)$$

with:

$$g(Q) = \frac{\beta^2 J^2}{4} \sum_{a,b=1}^m Q_{ab} + s(Q). \quad (154)$$

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 $Q_{ab}^\pi = Q_{\pi(a)\pi(b)}$. 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 $N_n(Q) = 2^n(2^n - 1) \cdots (2^n - m + 1)$, which implies that $s(Q) = m \ln 2$, and then $g(Q) = g_0(Q) \triangleq m(\beta^2 J^2/4 + \ln 2)$, thus $(\ln \mathbf{E}Z^m)/m = \beta^2 J^2/4 + \ln 2$, and so is the limit as $m \rightarrow 0$. Later on, we will compare this with the result obtained from a more rigorous derivation.
- $q_0 = 1$, which means that all components of \mathbf{i} are the same, and then $N_n(Q) = 2^n$, which means that $s(Q) = \ln 2$ and so, $g(Q) = g_1(Q) \triangleq m^2 \beta^2 J^2/4 + \ln 2$.

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 $\beta \geq \beta_c(m) \triangleq \frac{2}{J} \sqrt{\ln 2/m}$, and $g_0(Q)$ otherwise. For $m < 1$ (which is, in fact, the relevant case for $m \rightarrow 0$), 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 $\{1, 2, \dots, m\}$ into two complementary subsets (say, of equal size) and postulating a saddle point Q of the following structure:

$$Q_{ab} = \begin{cases} 1 & a = b \\ q_0 & a \text{ and } b \text{ are in the same subset} \\ q_1 & a \text{ and } b \text{ are in different subsets} \end{cases} \quad (155)$$

In further steps of symmetry breaking, one may split $\{1, 2, \dots, m\}$ 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 course, but the interested student/reader can easily find more details in the literature, specifically, in the book by Mézard and Montanari.

4 Interacting Particles and Phase Transitions

4.1 Introduction – Origins of Interactions

As I said already in the introductory part on the analysis tools and asymptotic methods, until now, we have dealt almost exclusively with systems that have additive Hamiltonians, $\mathcal{E}(\mathbf{x}) = \sum_i \mathcal{E}(x_i)$, 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 others the interactions are appreciably strong and cannot be neglected. Among the different particles there could be many sorts of mutual forces, e.g., mechanical, electrical, magnetic, etc. There could also be interactions that stem from quantum–mechanical effects: Pauli’s exclusion principle asserts that for a certain type of particles, called 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. We have already seen this as an example at the beginning of the previous set of lecture notes: In a quantum gas, as we eliminated the combinatorial factor (that counted indistinguishable states as distinguishable ones), we created statistical dependence, which physically means interactions.¹¹

4.2 A Few Models That Will be Discussed in This Subsection Only

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

¹¹Indeed, 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.

interactions, this means a structure like the following:

$$\mathcal{E}(\mathbf{x}) = \sum_{i=1}^n \mathcal{E}(x_i) + \sum_{(i,j)} \varepsilon(x_i, x_j) \quad (156)$$

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). 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 solid, i.e., a crystal lattice, 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. 5), 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

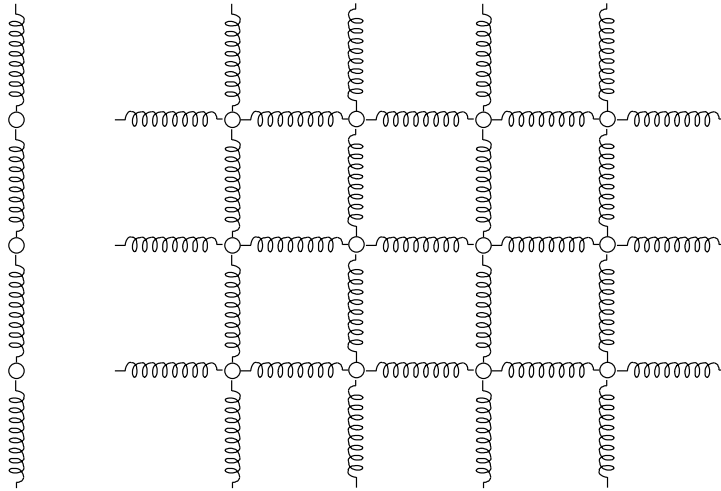


Figure 5: Elastic interaction forces between adjacent atoms in a one-dimensional lattice (left part of the figure) and in a two-dimensional lattice (right part).

arrays (or lattices), similar interactions apply, there are just more neighbors to each site, from the various directions (see right part of Fig. 5). 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. 6), and these normally depend solely on the distances $\|\vec{r}_i - \vec{r}_j\|$. For example, in a non-ideal gas,

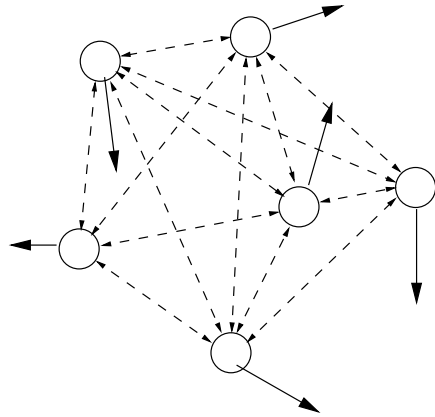


Figure 6: Mobile particles and mutual forces between them.

$$\mathcal{E}(\mathbf{x}) = \sum_{i=1}^n \frac{\|\vec{p}_i\|^2}{2m} + \sum_{i \neq j} V(\|\vec{r}_i - \vec{r}_j\|). \quad (157)$$

A very simple special case is that of hard spheres (Billiard balls), without any forces, where

$$V(\|\vec{r}_i - \vec{r}_j\|) = \begin{cases} \infty & \|\vec{r}_i - \vec{r}_j\| < 2R \\ 0 & \|\vec{r}_i - \vec{r}_j\| \geq 2R \end{cases} \quad (158)$$

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 also solvable, but this is beyond the scope of this course.

4.3 Models of Magnetic Materials – General

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 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, but most of them are not. For the purpose of our discussion, a magnetic material is one for which the important 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 – it may take only one out of finitely many values. In the simplest case to be adopted in our study – only two values. 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 – antiparallel (see Fig. 7). The

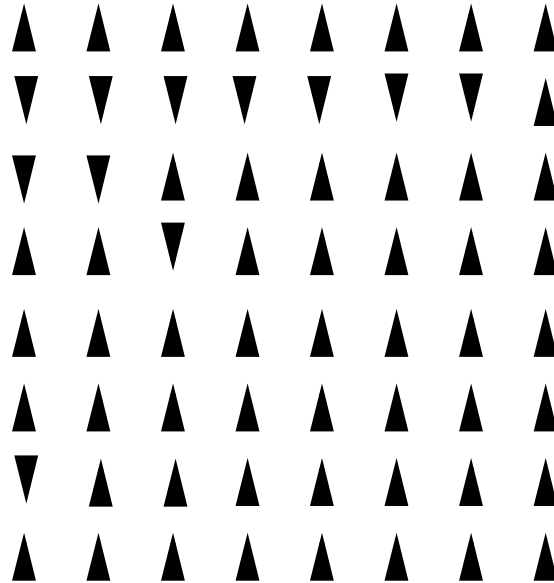


Figure 7: Illustration of a spin array on a square lattice.

Hamiltonian associated with an array of spins $\mathbf{s} = (s_1, \dots, s_n)$ is customarily modeled (up to certain constants that, among other things, accommodate for the physical units) with a

structure like this:

$$\mathcal{E}(\mathbf{s}) = -B \cdot \sum_{i=1}^n s_i - \sum_{(i,j)} J_{ij} s_i s_j, \quad (159)$$

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 antiparallel to \vec{B} , as said, these boil down to simple products, where only the sign of each s_i counts. Since $P(\mathbf{s})$ is proportional to $e^{-\beta\mathcal{E}(\mathbf{s})}$, the spins ‘prefer’ to be parallel, rather than antiparallel 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 channel codes, in particular, linear codes. 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 \mathbf{x} . I.e., the ℓ -th equation is: $x_{i_1}^\ell \oplus x_{i_2}^\ell \oplus \dots \oplus x_{i_{k_\ell}}^\ell = 0$, $\ell = 1, \dots, m$. Transforming from $x_i \in \{0, 1\}$ to $s_i \in \{-1, +1\}$ via $s_i = 1 - 2x_i$, this is equivalent to $s_{i_1}^\ell s_{i_2}^\ell \dots s_{i_{k_\ell}}^\ell = 1$. The MAP decoder would estimate \mathbf{s} based on the posterior

$$P(\mathbf{s}|\mathbf{y}) = \frac{P(\mathbf{s})P(\mathbf{y}|\mathbf{s})}{Z(\mathbf{y})}; \quad Z(\mathbf{y}) = \sum_{\mathbf{s}} P(\mathbf{s})P(\mathbf{y}|\mathbf{s}) = P(\mathbf{y}), \quad (160)$$

where $P(\mathbf{s})$ is normally assumed uniform over the codewords (we will elaborate on this posterior later). Assuming, e.g., a BSC or a Gaussian channel $P(\mathbf{y}|\mathbf{s})$, the relevant distance between the codeword $\mathbf{s} = (s_1, \dots, s_n)$ and the channel output $\mathbf{y} = (y_1, \dots, y_n)$ is propor-

tional to $\|\mathbf{s} - \mathbf{y}\|^2 = \text{const.} - 2 \sum_i s_i y_i$. Thus, $P(\mathbf{s}|\mathbf{y})$ can be thought of as a B–G distribution with Hamiltonian

$$\mathcal{E}(\mathbf{s}|\mathbf{y}) = -J \sum_{i=1}^n s_i y_i + \sum_{\ell=1}^m \phi(s_{i_1^\ell} s_{i_2^\ell} \cdots s_{i_{k_\ell}^\ell}) \quad (161)$$

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 $P(\mathbf{s}|\mathbf{y})$. 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 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.¹² For a general code, the second term is replaced by $\phi_{\mathcal{C}}(\mathbf{s})$, which is zero for $\mathbf{s} \in \mathcal{C}$ 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 \mathbf{S} and \mathbf{Y} is, of course, given by

$$I(\mathbf{S}; \mathbf{Y}) = H(\mathbf{Y}) - H(\mathbf{Y}|\mathbf{S}). \quad (162)$$

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:

$$H(\mathbf{Y}) = -\mathbf{E}\{\ln P(\mathbf{Y})\} = -\mathbf{E}\{\ln Z(\mathbf{Y})\}. \quad (163)$$

Thus, we are facing a problem of calculating the free energy of a spin system with random magnetic fields designated by the components of \mathbf{Y} . This is the kind of calculations we mentioned earlier in the context of the replica method. Indeed, the replica method is used extensively in this context.

¹²Error 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 .

As we will see in the sequel, it is also customary to introduce an inverse temperature parameter β , by defining

$$P_\beta(\mathbf{s}|\mathbf{y}) = \frac{P^\beta(\mathbf{s})P^\beta(\mathbf{y}|\mathbf{s})}{Z(\beta|\mathbf{y})} = \frac{e^{-\beta\mathcal{E}(\mathbf{s}|\mathbf{y})}}{Z(\beta|\mathbf{y})} \quad (164)$$

where β controls the sharpness of the posterior distribution and

$$Z(\beta|\mathbf{y}) = \sum_{\mathbf{s}} e^{-\beta\mathcal{E}(\mathbf{s}|\mathbf{y})}. \quad (165)$$

The motivations of this will be discussed extensively later on.

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*.

4.4 Phase Transitions – A Qualitative Discussion

Loosely speaking, 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, and so on. 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 are these phase transitions? Before we refer to this question, it should be noted that there are also “phase transitions” in the behavior of communication systems: As the SNR passes a certain limit (for which capacity crosses the coding rate), there is a sharp transition between reliable and unreliable communication, where the error probability (almost) ‘jumps’ from 0 to 1 or vice versa. We also know about certain threshold effects in highly non-linear communication systems. 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,

$$P(\mathbf{s}) = \frac{\exp \left\{ \beta J \sum_{(i,j)} s_i s_j \right\}}{Z(\beta)} \quad (166)$$

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 (and for $T = 0$, this is exact) equivalent to minimizing ϵ , which is obtained by states with a high level of order, as $\mathcal{E}(\mathbf{s}) = -J \sum_{(i,j)} s_i s_j$, 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. 8). Thus, the order–disorder characteristics depend primarily

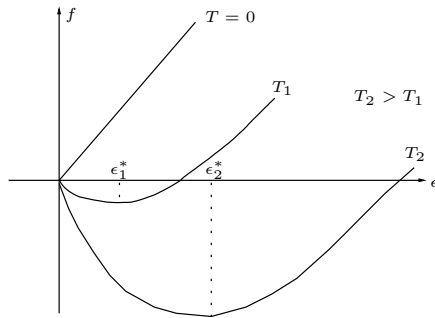


Figure 8: Qualitative graphs of $f(\epsilon)$ at various temperatures. The minimizing ϵ increases with T .

on temperature. It turns 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 a bunch of exponentials in β and therefore it is continuous and differentiable as many times as we want. 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 $\phi(\beta) = \lim_{n \rightarrow \infty} \frac{\ln Z(\beta)}{n}$. 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 ‘nice’ and 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. We will 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 ‘nice’ 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

$$\phi(\beta, B) = \lim_{n \rightarrow \infty} \frac{\ln Z(\beta, B)}{n} = \lim_{n \rightarrow \infty} \frac{1}{n} \ln \left[\sum_{\mathbf{s}} \exp \left\{ \beta B \sum_{i=1}^n s_i + \beta J \sum_{(i,j)} s_i s_j \right\} \right] \quad (167)$$

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

$$m(\beta, B) = \left\langle \frac{1}{n} \sum_{i=1}^n S_i \right\rangle, \quad (168)$$

and this quantity might not always be continuous. Indeed, as I 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 will see this more concretely later on. We next discuss a few solvable models of spin arrays, with and without phase transitions.

4.5 The One-Dimensional Ising Model

According to this model,

$$\mathcal{E}(\mathbf{s}) = -B \sum_{i=1}^n s_i - J \sum_{i=1}^n s_i s_{i+1} \quad (169)$$

with the periodic boundary condition $s_{n+1} = s_1$. Thus,

$$\begin{aligned} Z(\beta, B) &= \sum_{\mathbf{s}} \exp \left\{ \beta B \sum_{i=1}^n s_i + \beta J \sum_{i=1}^n s_i s_{i+1} \right\} \quad \text{Note: the kind of sums encountered in Markov chains} \\ &= \sum_{\mathbf{s}} \exp \left\{ h \sum_{i=1}^n s_i + K \sum_{i=1}^n s_i s_{i+1} \right\} \quad h \triangleq \beta B, \quad K \triangleq \beta J \\ &= \sum_{\mathbf{s}} \exp \left\{ \frac{h}{2} \sum_{i=1}^n (s_i + s_{i+1}) + K \sum_{i=1}^n s_i s_{i+1} \right\} \quad (\text{just to symmetrize the expression}) \end{aligned}$$

Consider now the 2×2 matrix P whose entries are $\exp\{\frac{h}{2}(s + s') + Kss'\}$, $s, s' \in \{-1, +1\}$, i.e.,

$$P = \begin{pmatrix} e^{K+h} & e^{-K} \\ e^{-K} & e^{K-h} \end{pmatrix}. \quad (170)$$

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,

$$\begin{aligned} Z(\beta, B) &= \sum_{\sigma_1} \cdots \sum_{\sigma_n} (\sigma_1^T P \sigma_2) \cdot (\sigma_2^T P \sigma_3) \cdots (\sigma_n^T P \sigma_1) \\ &= \sum_{\sigma_1} \sigma_1^T P \left(\sum_{\sigma_2} \sigma_2 \sigma_2^T \right) P \left(\sum_{\sigma_3} \sigma_3 \sigma_3^T \right) P \cdots P \left(\sum_{\sigma_n} \sigma_n \sigma_n^T \right) P \sigma_1 \\ &= \sum_{\sigma_1} \sigma_1^T P \cdot I \cdot P \cdot I \cdots I \cdot P \sigma_1 \\ &= \sum_{\sigma_1} \sigma_1^T P^n \sigma_1 \\ &= \text{tr}\{P^n\} \\ &= \lambda_1^n + \lambda_2^n \end{aligned} \quad (171)$$

where λ_1 and λ_2 are the eigenvalues of P , which are

$$\lambda_{1,2} = e^K \cosh(h) \pm \sqrt{e^{-2K} + e^{2K} \sinh^2(h)}. \quad (172)$$

Letting λ_1 denote the larger (the dominant) eigenvalue, i.e.,

$$\lambda_1 = e^K \cosh(h) + \sqrt{e^{-2K} + e^{2K} \sinh^2(h)}, \quad (173)$$

then clearly,

$$\phi(h, K) = \lim_{n \rightarrow \infty} \frac{\ln Z}{n} = \ln \lambda_1. \quad (174)$$

The average magnetization is

$$\begin{aligned} M(h, K) &= \left\langle \sum_{i=1}^n S_i \right\rangle \\ &= \frac{\sum_{\mathbf{s}} (\sum_{i=1}^n s_i) \exp\{h \sum_{i=1}^n s_i + K \sum_{i=1}^n s_i s_{i+1}\}}{\sum_{\mathbf{s}} \exp\{h \sum_{i=1}^n s_i + K \sum_{i=1}^n s_i s_{i+1}\}} \\ &= \frac{\partial \ln Z(h, K)}{\partial h} \end{aligned} \quad (175)$$

and so, the per-spin magnetization is:

$$m(h, K) \triangleq \lim_{n \rightarrow \infty} \frac{M(h, K)}{n} = \frac{\partial \phi(h, K)}{\partial h} = \frac{\sinh(h)}{\sqrt{e^{-4K} + \sinh^2(h)}} \quad (176)$$

or, returning to the original parametrization:

$$m(\beta, B) = \frac{\sinh(\beta B)}{\sqrt{e^{-4\beta J} + \sinh^2(\beta B)}}. \quad (177)$$

For $\beta > 0$ and $B > 0$ this is a nice function, and so, there is no phase transitions and no spontaneous magnetization at any finite temperature.¹³ However, at the absolute zero ($\beta \rightarrow \infty$), we get

$$\lim_{B \downarrow 0} \lim_{\beta \rightarrow \infty} m(\beta, B) = +1; \quad \lim_{B \uparrow 0} \lim_{\beta \rightarrow \infty} m(\beta, B) = -1, \quad (178)$$

thus m is discontinuous w.r.t. B at $\beta \rightarrow \infty$, 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 by Onsager in 1944, who has shown that it exhibits a phase transition with Curie temperature given by

$$T_c = \frac{2J}{k \ln(\sqrt{2} + 1)}, \quad (179)$$

where k is Boltzmann's constant. For lattice dimension ≥ 3 , the problem is still open.

It turns out then that whatever counts for the existence of phase transitions, is not the intensity of the interactions (designated by the magnitude of J), but rather 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,

¹³Note, 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.

or more precisely, what is the value of $d \triangleq \lim_{\ell \rightarrow \infty} \frac{1}{\ell} \ln n_\ell$, which is exactly the ordinary dimensionality for hypercubic 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, I 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 (see Baxter’s book), but we will not discuss it here.

4.6 The Curie–Weiss Model

According to the Curie–Weiss (C–W) model,

$$\mathcal{E}(\mathbf{s}) = -B \sum_{i=1}^n s_i - \frac{J}{2n} \sum_{i \neq j} s_i s_j. \quad (180)$$

Here, all pairs $\{(s_i, s_j)\}$ “talk to each other” with the same “voice intensity”, $J/(2n)$, 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

$$\left(\sum_i s_i \right)^2 = \sum_i s_i^2 + \sum_{i \neq j} s_i s_j = n + \sum_{i \neq j} s_i s_j \quad (181)$$

where the second equality holds since $s_i^2 \equiv 1$. It follows then, that our Hamiltonian is, upto a(n immaterial) constant, equivalent to

$$\mathcal{E}(\mathbf{s}) = -B \sum_{i=1}^n s_i - \frac{J}{2n} \left(\sum_{i=1}^n s_i \right)^2 = -n \left[B \cdot \left(\frac{1}{n} \sum_{i=1}^n s_i \right) + \frac{J}{2} \left(\frac{1}{n} \sum_{i=1}^n s_i \right)^2 \right], \quad (182)$$

thus $\mathcal{E}(\mathbf{s})$ depends on \mathbf{s} only via the magnetization $m(\mathbf{s}) = \frac{1}{n} \sum_i s_i$. This fact makes the C–W model very easy to handle similarly as in the method of types:

$$\begin{aligned} Z_n(\beta, B) &= \sum_{\mathbf{s}} \exp \left\{ n\beta \left[B \cdot m(\mathbf{s}) + \frac{J}{2} m^2(\mathbf{s}) \right] \right\} \\ &= \sum_{m=-1}^{+1} \Omega(m) \cdot e^{n\beta(Bm + Jm^2/2)} \\ &\doteq \sum_{m=-1}^{+1} e^{nh_2((1+m)/2)} \cdot e^{n\beta(Bm + Jm^2/2)} \\ &\doteq \exp \left\{ n \cdot \max_{|m| \leq 1} \left[h_2 \left(\frac{1+m}{2} \right) + \beta Bm + \frac{\beta m^2 J}{2} \right] \right\} \end{aligned}$$

and so,

$$\phi(\beta, B) = \max_{|m| \leq 1} \left[h_2 \left(\frac{1+m}{2} \right) + \beta Bm + \frac{\beta m^2 J}{2} \right]. \quad (183)$$

The maximum is found by equating the derivative to zero, i.e.,

$$0 = \frac{1}{2} \ln \left(\frac{1-m}{1+m} \right) + \beta B + \beta Jm \equiv -\tanh^{-1}(m) + \beta B + \beta Jm \quad (184)$$

or equivalently, the maximizing (and hence the dominant) m is a solution m^* to the equation¹⁴

$$m = \tanh(\beta B + \beta Jm).$$

Consider first the case $B = 0$, where the equation boils down to

$$m = \tanh(\beta Jm). \quad (185)$$

It is instructive to look at this equation graphically. Referring to Fig. 9, we have to make a distinction between two cases: If $\beta J < 1$, namely, $T > T_c \triangleq J/k$, the slope of the function

¹⁴Once again, for $J = 0$, we are back to non–interacting spins and then this equation gives the paramagnetic behavior $m = \tanh(\beta B)$.

$y = \tanh(\beta J m)$ 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 $\psi(m) \triangleq h_2((1+m)/2) + \beta J m^2/2$ at $m = 0$ is negative, and therefore it is indeed the maximum (see Fig. 10, left part). Thus, the dominant magnetization is $m^* = 0$, which means disorder and hence no spontaneous magnetization for $T > T_c$. On

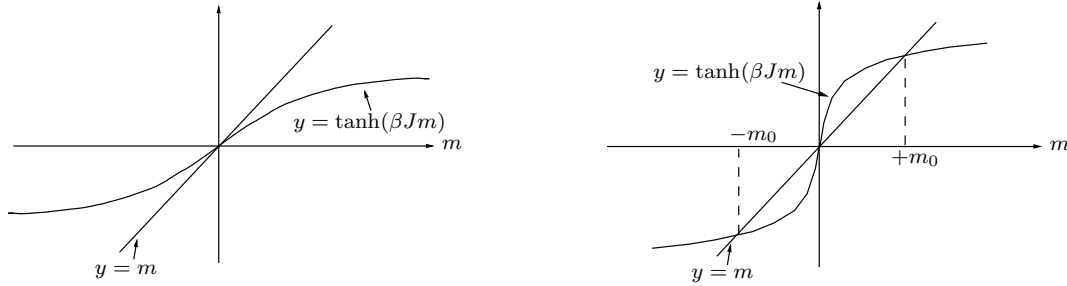


Figure 9: Graphical solutions of equation $m = \tanh(\beta J m)$: The left part corresponds to the case $\beta J < 1$, where there is one solution only, $m^* = 0$. The right part corresponds to the case $\beta J > 1$, where in addition to the zero solution, there are two non-zero solutions $m^* = \pm m_0$.

the other hand, when $\beta J > 1$, which means temperatures lower than T_c , the initial slope of the tanh function is larger than that of the linear function, but since the tanh 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. 9, 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. 10, right part). Thus, the dominant magnetizations are $\pm m_0$, each capturing about half of the probability.

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. 10 and the corresponding graphs of $\psi(m)$ would be as in Fig. 11, 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 $B \rightarrow 0$, $m_0(B) \rightarrow m_0$ of Fig. 10. Thus, we see the spontaneous magnetization here. Even after removing the magnetic 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

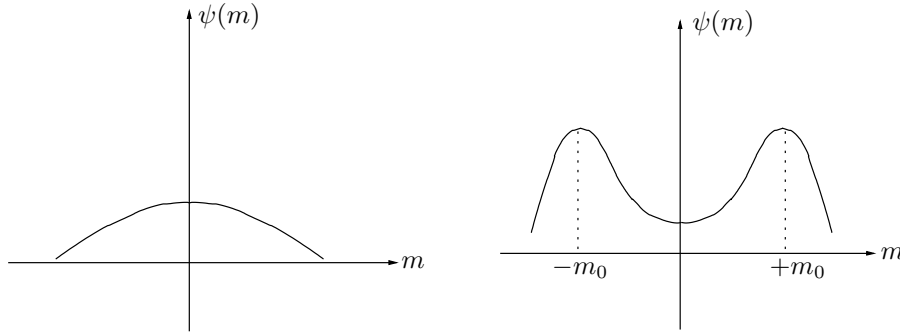


Figure 10: The function $\psi(m) = h_2((1+m)/2) + \beta J m^2/2$ has a unique maximum at $m = 0$ when $\beta J < 1$ (left graph) and two local maxima at $\pm m_0$, in addition to a local minimum at $m = 0$, when $\beta J > 1$ (right graph).

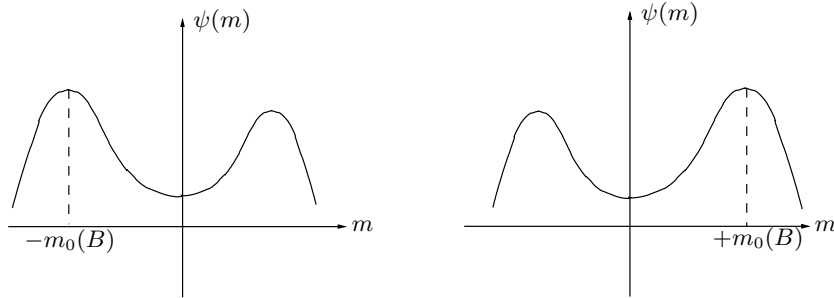


Figure 11: The case $\beta J > 1$ with a magnetic field B . The left graph corresponds to $B < 0$ and the right graph – to $B > 0$.

$B = 0$ for $T < T_c$, which is a first order phase transition w.r.t. B (see Fig. 12). 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.

Finally, we should mention here an alternative technique that can be used to analyze this model, which is useful in many other contexts as well. It is based on the idea of using a transform integral, in this case, the *Hubbard–Stratonovich transform*, and then the saddle

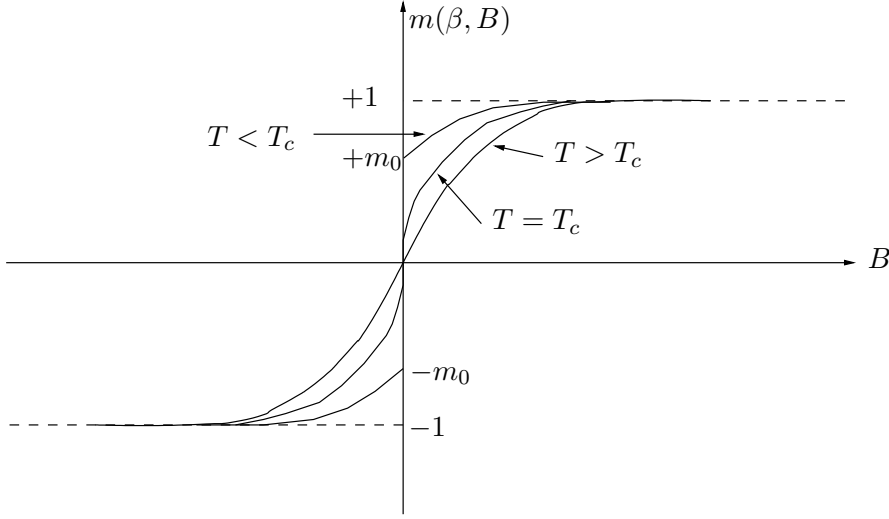


Figure 12: Magnetization vs. magnetic field: For $T < T_c$ there is spontaneous magnetization: $\lim_{B \downarrow 0} m(\beta, B) = +m_0$ and $\lim_{B \uparrow 0} m(\beta, B) = -m_0$, and so there is a discontinuity at $B = 0$.

point method. Specifically, we have the following chain of equalities:

$$\begin{aligned}
Z(h, K) &= \sum_{\mathbf{s}} \exp \left\{ h \sum_{i=1}^n s_i + \frac{K}{2n} \left(\sum_{i=1}^n s_i \right)^2 \right\} \quad h \triangleq \beta B, \quad K \triangleq \beta J \\
&= \sum_{\mathbf{s}} \exp \left\{ h \sum_{i=1}^n s_i \right\} \cdot \exp \left\{ \frac{K}{2n} \left(\sum_{i=1}^n s_i \right)^2 \right\} \\
&= \sum_{\mathbf{s}} \exp \left\{ h \sum_{i=1}^n s_i \right\} \cdot \sqrt{\frac{n}{2\pi K}} \int_{\mathbb{R}} dz \exp \left\{ -\frac{nz^2}{2K} + z \cdot \sum_{i=1}^n s_i \right\} \\
&= \sqrt{\frac{n}{2\pi K}} \int_{\mathbb{R}} dz e^{-nz^2/(2K)} \sum_{\mathbf{s}} \exp \left\{ (h+z) \sum_{i=1}^n s_i \right\} \\
&= \sqrt{\frac{n}{2\pi K}} \int_{\mathbb{R}} dz e^{-nz^2/(2K)} \left[\sum_{s=-1}^1 e^{(h+z)s} \right]^n \\
&= \sqrt{\frac{n}{2\pi K}} \int_{\mathbb{R}} dz e^{-nz^2/(2K)} [2 \cosh(h+z)]^n \\
&= 2^n \cdot \sqrt{\frac{n}{2\pi K}} \int_{\mathbb{R}} dz \exp \{ n [\ln \cosh(h+z) - z^2/(2K)] \}
\end{aligned}$$

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

$$\gamma(z) = \frac{z^2}{2K} - \ln \cosh(h + z). \quad (186)$$

by equating its derivative to zero, we get the very same equation as $m = \tanh(\beta B + \beta Jm)$ by setting $z = \beta Jm$. 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*.

4.7 Spin Glass Models With Random Parameters and Random Code Ensembles

So far we discussed only models where the non-zero coupling coefficients, $\mathbf{J} = \{J_{ij}\}$ 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 $P(\mathbf{s}|\mathbf{y})$ 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 $P(\mathbf{y})$ exhibits memory (even if the channel is memoryless) due to memory in $P(\mathbf{s})$. 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., $P_\beta(\mathbf{s}|\mathbf{y}) \propto P^\beta(\mathbf{s})P^\beta(\mathbf{y}|\mathbf{s})$), the induced marginal of \mathbf{y} will depend on β .

In the following discussion, let us refer to the case where only the coupling coefficients \mathbf{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 \mathbf{J} .
- Randomness of the spin configuration \mathbf{s} given \mathbf{J} , according to the Boltzmann distribution, i.e.,

$$P(\mathbf{s}|\mathbf{J}) = \frac{\exp\left\{\beta\left[B\sum_{i=1}^n s_i + \sum_{(i,j)} J_{ij}s_i s_j\right]\right\}}{Z(\beta, B|\mathbf{J})}. \quad (187)$$

However, these two sets of RV's have a rather different stature. The underlying setting is normally such that \mathbf{J} is considered to be randomly drawn once and for all, and then remain fixed, whereas \mathbf{s} keeps varying all the time (according to the dynamics of the system). At any rate, the time scale along which \mathbf{s} varies is much smaller than that of \mathbf{J} . Another difference is that \mathbf{J} is normally not assumed to depend on temperature, whereas \mathbf{s} , of course, does. In the terminology of physicists, \mathbf{s} is considered an *annealed* RV, whereas \mathbf{J} is considered a *quenched* RV. 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 don't 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.¹⁵ We will elaborate on this point later on.

¹⁵ There are few exceptions to this rule, e.g., a paper by Barg and Forney, IEEE Trans. on IT, Sept. 2002, and several follow-ups.

Returning to spin glass models, let's 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, $\ln Z(\beta|\mathbf{J})$, this is now a RV, of course, because it depends on the random \mathbf{J} . If we denote by $\langle \cdot \rangle_{\mathbf{J}}$ the expectation w.r.t. the randomness of \mathbf{J} , then quenched averaging means $\langle \ln Z(\beta|\mathbf{J}) \rangle_{\mathbf{J}}$ (with the motivation of the self-averaging property of the RV $\ln Z(\beta|\mathbf{J})$ in many cases), whereas annealed averaging means $\ln \langle Z(\beta|\mathbf{J}) \rangle_{\mathbf{J}}$. 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 \mathbf{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:

$$\begin{aligned}
\langle Z(\beta|\mathbf{J}) \rangle &= \sum_{\mathbf{J}} P(\mathbf{J}) Z(\beta|\mathbf{J}) \\
&\approx \sum_{\alpha} \Pr\{\mathbf{J} : Z(\beta|\mathbf{J}) \doteq e^{n\alpha}\} \cdot e^{n\alpha} \\
&\approx \sum_{\alpha} e^{-nE(\alpha)} \cdot e^{n\alpha} \quad (\text{assuming exponential probabilities}) \\
&\doteq e^{n \max_{\alpha} [\alpha - E(\alpha)]}
\end{aligned} \tag{188}$$

which means that the annealed average is dominated by realizations of the system with

$$\frac{\ln Z(\beta|\mathbf{J})}{n} \approx \alpha^* \triangleq \arg \max_{\alpha} [\alpha - E(\alpha)], \tag{189}$$

which may differ from the typical value of α , which is

$$\alpha = \phi(\beta) \equiv \lim_{n \rightarrow \infty} \frac{1}{n} \langle \ln Z(\beta|\mathbf{J}) \rangle. \tag{190}$$

On the other hand, when it comes to quenched averaging, the RV $\ln Z(\beta|\mathbf{J})$ 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, $\ln \mathbf{E}P_e(\mathcal{C})$ (annealed) and the average exponent of the error probability $\mathbf{E} \ln P_e(\mathcal{C})$ (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. RV's, 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, \dots, i_p) exhaust all possible subsets of p spins out of n . Each such term has a Gaussian coefficient J_{i_1, \dots, i_p} with an appropriate variance.

Considering the p -spin model, it turns out that if we look at the extreme case of $p \rightarrow \infty$ (taken after the thermodynamic limit $n \rightarrow \infty$), the resulting behavior turns out to be extremely erratic: all energy levels $\{\mathcal{E}(\mathbf{s})\}_{\mathbf{s} \in \{-1, +1\}^n}$ become i.i.d. Gaussian RV's. 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 talk about it extensively because it turns out to be very relevant for random coding models.

5 The Random Energy Model and Random Coding

5.1 The REM in the Absence of a Magnetic Field

The REM was proposed by the French physicist Bernard Derrida in the early eighties of the previous century in a series of papers:

1. B. Derrida, “Random–energy model: limit of a family of disordered models,” *Phys. Rev. Lett.*, vol. 45, no. 2, pp. 79–82, July 1980.
2. B. Derrida, “The random energy model,” *Physics Reports* (Review Section of Physics Letters), vol. 67, no. 1, pp. 29–35, 1980.
3. B. Derrida, “Random–energy model: an exactly solvable model for disordered systems,” *Phys. Rev. B*, vol. 24, no. 5, pp. 2613–2626, September 1981.

Derrida showed in one of his papers that, since the correlations between the random energies of two configurations, \mathbf{s} and \mathbf{s}' in the p –spin model are given by

$$\left(\frac{1}{n} \sum_{i=1}^n s_i s'_i \right)^p, \quad (191)$$

and since $|\frac{1}{n} \sum_{i=1}^n s_i s'_i| < 1$, these correlations vanish as $p \rightarrow \infty$. This has motivated him to propose a model according to which the configurational energies $\{\mathcal{E}(\mathbf{s})\}$, in the absence of a magnetic field, are simply i.i.d. zero–mean Gaussian RV’s 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 forget that the spin array has any structure of the kind that we have seen before, and we simply randomly draw an independent RV $\mathcal{E}(\mathbf{s}) \sim \mathcal{N}(0, nJ^2/2)$ (and other distributions are also possible) for every configuration \mathbf{s} . Thus, the partition function $Z(\beta) = \sum_{\mathbf{s}} e^{-\beta \mathcal{E}(\mathbf{s})}$ 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.
- Last but not least, 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,

$$Z(\beta) = \sum_{\mathbf{s}} e^{-\beta \mathcal{E}(\mathbf{s})} \quad (192)$$

where $\mathcal{E}(\mathbf{s}) \sim \mathcal{N}(0, nJ^2/2)$ are i.i.d. Consider the density of states $\Omega(E)$, which is now a RV: $\Omega(E)dE$ is the number of configurations $\{\mathbf{s}\}$ whose randomly selected energy $\mathcal{E}(\mathbf{s})$ happens to fall between E and $E + dE$, and of course,

$$Z(\beta) = \int_{-\infty}^{+\infty} dE \Omega(E) e^{-\beta E}. \quad (193)$$

How does the RV $\Omega(E)dE$ behave like? First, observe that, ignoring non-exponential factors:

$$\Pr\{E \leq \mathcal{E}(\mathbf{s}) \leq E + dE\} \approx f(E)dE \doteq e^{-E^2/(nJ^2)} dE, \quad (194)$$

and so,

$$\langle \Omega(E)dE \rangle \doteq 2^n \cdot e^{-E^2/(nJ^2)} = \exp \left\{ n \left[\ln 2 - \left(\frac{E}{nJ} \right)^2 \right] \right\}. \quad (195)$$

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 $\Omega(E)dE$. In fact, this principle is frequently used in random coding arguments in IT:

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:

$$\begin{aligned}
\Pr\left\{\bigcup_{i=1}^{e^{nA}} \mathcal{E}_i\right\} &= 1 - \Pr\left\{\bigcap_{i=1}^{e^{nA}} \mathcal{E}_i^c\right\} \\
&= 1 - (1 - e^{-nB})^{e^{nA}} \\
&= 1 - \left[e^{\ln(1-e^{-nB})}\right]^{e^{nA}} \\
&= 1 - \exp\{e^{nA} \ln(1 - e^{-nB})\} \\
&\approx 1 - \exp\{-e^{nA} e^{-nB}\} \\
&= 1 - \exp\{-e^{n(A-B)}\} \\
&\rightarrow \begin{cases} 1 & A > B \\ 0 & A < B \end{cases} \tag{196}
\end{aligned}$$

BTW, the 2nd line could have been shown also by the union bound, as $\sum_i \Pr\{\mathcal{E}_i\} = e^{nA} e^{-nB} \rightarrow 0$. Exercise: What happens when $A = B$?

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 RV’s whose expectation is $\mathbf{E}\{\Omega_n\} = e^{n(A-B)}$. 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} , i.e., *double-exponentially* (!) in n .¹⁶ 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

$$E_0 \triangleq nJ\sqrt{\ln 2} \tag{197}$$

¹⁶This will be shown rigorously later on.

the exponential increase rate, $A = \ln 2$, of the number $2^n = e^{n \ln 2}$ of configurations, = the number of independent trials in randomly drawing energies $\{\mathcal{E}(\mathbf{s})\}$, 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 $\mathcal{E}(\mathbf{s})$ would happen to fall around E . In other words, the number of these trials is way larger than one over this probability and in view of the earlier discussion, the probability that

$$\Omega(E)dE = \sum_{\mathbf{s}} \mathcal{I}\{E \leq \mathcal{E}(\mathbf{s}) \leq E + dE\}. \quad (198)$$

would deviate from its mean $\dot{=} \exp\{n[\ln 2 - (E/(nJ))^2]\}$, 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

$$e^{-n\epsilon} \cdot \exp\left\{n \left[\ln 2 - \left(\frac{E}{nJ}\right)^2 \right]\right\} \leq \Omega(E)dE \leq e^{+n\epsilon} \cdot \exp\left\{n \left[\ln 2 - \left(\frac{E}{nJ}\right)^2 \right]\right\} \quad (199)$$

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 $\Omega(E)dE$ is exponentially large. On the other hand, if $|E| > E_0$, the number of trials 2^n is way smaller than one over 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. It follows then that for such a typical realization,

$$\begin{aligned} Z(\beta) &\approx \int_{-E_0}^{+E_0} \langle dE \cdot \Omega(E) \rangle e^{-\beta E} \\ &\doteq \int_{-E_0}^{+E_0} dE \cdot \exp\left\{n \left[\ln 2 - \left(\frac{E}{nJ}\right)^2 \right]\right\} \cdot e^{-\beta E} \\ &= \int_{-E_0}^{+E_0} dE \cdot \exp\left\{n \left[\ln 2 - \left(\frac{E}{nJ}\right)^2 - \beta \cdot \left(\frac{E}{n}\right) \right]\right\} \\ &= n \cdot \int_{-\epsilon_0}^{+\epsilon_0} d\epsilon \cdot \exp\left\{n \left[\ln 2 - \left(\frac{\epsilon}{J}\right)^2 - \beta\epsilon \right]\right\} \quad \epsilon \triangleq \frac{E}{n}, \quad \epsilon_0 \triangleq \frac{E_0}{n} = J\sqrt{\ln 2}, \\ &\doteq \exp\left\{n \cdot \max_{|\epsilon| \leq \epsilon_0} \left[\ln 2 - \left(\frac{\epsilon}{J}\right)^2 - \beta\epsilon \right]\right\} \quad \text{by Laplace integration} \end{aligned}$$

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 endpoint. The choice between the two depends on β . Specifically, we obtain the following:

$$\phi(\beta) = \lim_{n \rightarrow \infty} \frac{\ln Z(\beta)}{n} = \begin{cases} \ln 2 + \frac{\beta^2 J^2}{4} & \beta \leq \beta_c \\ \beta J \sqrt{\ln 2} & \beta > \beta_c \end{cases} \quad (200)$$

where $\beta_c = \frac{2}{J} \sqrt{\ln 2}$. What we see here is 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. 13, 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. Thus, the condition $R > \ln 2 - h_2(\delta)$ is equivalent to

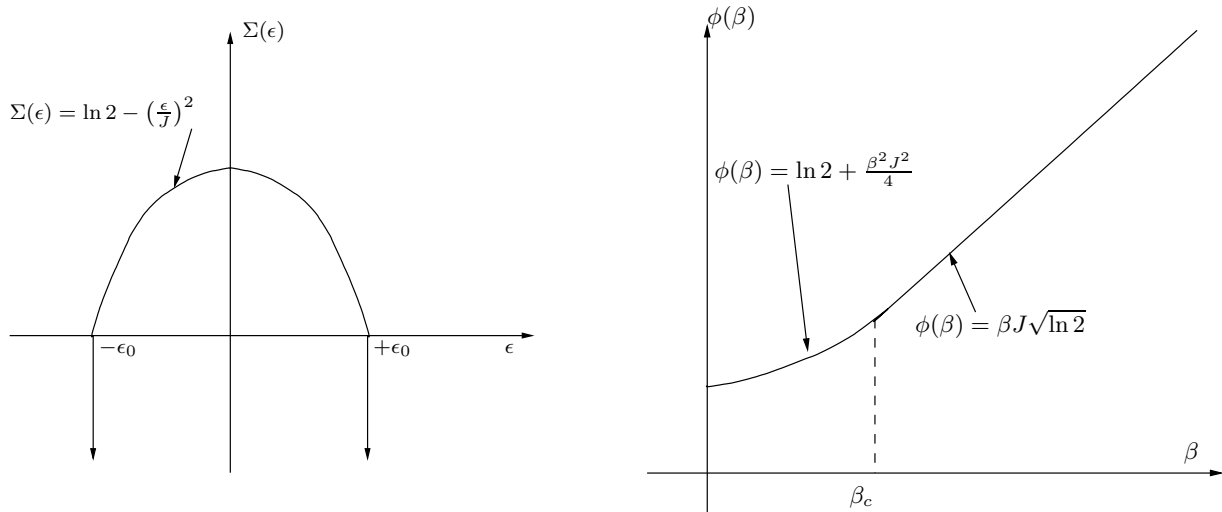


Figure 13: The entropy function and the normalized log-partition function of the REM.

What is the significance of each one of these phases? Let's 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

$$\Sigma(\beta) = \phi(\beta) - \beta \cdot \phi'(\beta),$$

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. 13, left part: $\Sigma(-\epsilon_0) = 0$). This is why we will refer to this phase as the *frozen phase* or the *glassy phase*.¹⁷ 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 (exercise: what is the exponent?) 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 than in the Curie–Weiss model. We are not talking here about spontaneous magnetization transition, but rather on a glass transition. In fact, we will not see here a spontaneous magnetization even if we add a magnetic field (time permits, this will be seen later on).

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, I wish to emphasize that since the calculation of Z was carried out for the typical realizations of the quenched RV’s $\{\mathcal{E}(\mathbf{s})\}$, we have actually calculated the quenched average of $\lim_n (\ln Z)/n$. As for the annealed average, we have

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{\ln \langle Z(\beta) \rangle}{n} &= \lim_{n \rightarrow \infty} \frac{1}{n} \ln \left[\int_{\mathbb{R}} \langle \Omega(E) d\epsilon \rangle e^{-\beta n \epsilon} \right] \\ &= \lim_{n \rightarrow \infty} \frac{1}{n} \ln \left[\int_{\mathbb{R}} \exp \left\{ n \left[\ln 2 - \left(\frac{\epsilon}{J} \right)^2 - \beta \epsilon \right] \right\} \right] \\ &= \max_{\epsilon \in \mathbb{R}} \left[\ln 2 - \left(\frac{\epsilon}{J} \right)^2 - \beta \epsilon \right] \quad \text{Laplace integration} \\ &= \ln 2 + \frac{\beta^2 J^2}{4}, \end{aligned} \tag{201}$$

¹⁷In 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.

which is the paramagnetic expression, without any phase transition since the maximization over ϵ is not constrained.

5.2 The Random Code Ensemble and its Relation to the REM

Let us now see how does the REM relate to random code ensembles. The discussion in this part is based on Mézard and Montanari’s book, as well as on the paper: 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. Another relevant paper is: 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.

Consider a DMC, $P(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^n p(y_i|x_i)$, fed by an input n -vector that belongs to a codebook $\mathcal{C} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M\}$, $M = e^{nR}$, with uniform priors, where R is the coding rate in nats per channel use. The induced posterior, for $\mathbf{x} \in \mathcal{C}$, is then:

$$\begin{aligned} P(\mathbf{x}|\mathbf{y}) &= \frac{P(\mathbf{y}|\mathbf{x})}{\sum_{\mathbf{x}' \in \mathcal{C}} P(\mathbf{y}|\mathbf{x}')} \\ &= \frac{e^{-\ln[1/P(\mathbf{y}|\mathbf{x})]}}{\sum_{\mathbf{x}' \in \mathcal{C}} e^{-\ln[1/P(\mathbf{y}|\mathbf{x}')]}}, \end{aligned} \tag{202}$$

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

$$\begin{aligned} P_\beta(\mathbf{x}|\mathbf{y}) &= \frac{P^\beta(\mathbf{y}|\mathbf{x})}{\sum_{\mathbf{x}' \in \mathcal{C}} P^\beta(\mathbf{y}|\mathbf{x}')} \\ &= \frac{e^{-\beta \ln[1/P(\mathbf{y}|\mathbf{x})]}}{\sum_{\mathbf{x}' \in \mathcal{C}} e^{-\beta \ln[1/P(\mathbf{y}|\mathbf{x}')]} \\ &\triangleq \frac{e^{-\beta \ln[1/P(\mathbf{y}|\mathbf{x})]}}{Z(\beta|\mathbf{y})} \end{aligned}$$

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 symbolwise maximum a-posteriori (MAP) decoding, i.e., decoding the ℓ -th symbol of \mathbf{x} as $\arg \max_a P_\beta(x_\ell = a|\mathbf{y})$, where

$$P_\beta(x_\ell = a|\mathbf{y}) = \sum_{\mathbf{x} \in \mathcal{C}: x_\ell = a} P_\beta(\mathbf{x}|\mathbf{y}), \quad (203)$$

we obtain a family of *finite-temperature decoders* (originally proposed by Ruján in 1993) parametrized by β , where $\beta = 1$ corresponds to minimum symbol error probability (with respect to the real underlying channel $P(\mathbf{y}|\mathbf{x})$) and $\beta \rightarrow \infty$ corresponds to minimum block error probability.

- This is one of our main motivations: the corresponding partition function, $Z(\beta|\mathbf{y})$, 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,

$$P(\mathbf{y}|\mathbf{x}) = p^{d(\mathbf{x},\mathbf{y})}(1-p)^{n-d(\mathbf{x},\mathbf{y})} = (1-p)^n e^{-Jd(\mathbf{x},\mathbf{y})}, \quad (204)$$

where $J = \ln \frac{1-p}{p}$ and $d(\mathbf{x}, \mathbf{y})$ is the Hamming distance. Thus, the partition function can be presented as follows:

$$Z(\beta|\mathbf{y}) = (1-p)^{\beta n} \sum_{\mathbf{x} \in \mathcal{C}} e^{-\beta J d(\mathbf{x},\mathbf{y})}. \quad (205)$$

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’ $\{Jd(\mathbf{x}, \mathbf{y})\}$ pertaining to the partition function

$$Z(\beta|\mathbf{y}) = (1 - p)^{\beta n} \sum_{\mathbf{x} \in \mathcal{C}} e^{-\beta Jd(\mathbf{x}, \mathbf{y})}, \quad (206)$$

(or, in the case of a more general channel, the energies $\{-\ln[1/P(\mathbf{y}|\mathbf{x})]\}$ pertaining to the partition function $Z(\beta|\mathbf{y}) = \sum_{\mathbf{x} \in \mathcal{C}} e^{-\beta \ln[1/P(\mathbf{y}|\mathbf{x})]}$), are i.i.d. random variables for all codewords in \mathcal{C} , with the exception of the codeword \mathbf{x}_0 that was actually transmitted and generated \mathbf{y} .¹⁸ 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 \mathbf{x}_0 , which is

$$Z_c(\beta|\mathbf{y}) \triangleq (1 - p)^{\beta n} e^{-Jd(\mathbf{x}_0, \mathbf{y})} \quad (207)$$

and the contribution of all other (incorrect) codewords:

$$Z_e(\beta|\mathbf{y}) \triangleq (1 - p)^{\beta n} \sum_{\mathbf{x} \in \mathcal{C} \setminus \{\mathbf{x}_0\}} e^{-Jd(\mathbf{x}, \mathbf{y})}. \quad (208)$$

Concerning the former, things are very simple: Typically, the channel flips about np bits out the n transmissions, which means that with high probability, $d(\mathbf{x}_0, \mathbf{y})$ is about np , and so $Z_c(\beta|\mathbf{y})$ is expected to take values around $(1 - p)^{\beta n} e^{-\beta Jnp}$. The more complicated and more interesting question is how does $Z_e(\beta|\mathbf{y})$ behave, and here the treatment will be very similar to that of the REM.

Given \mathbf{y} , define $\Omega_{\mathbf{y}}(d)$ as the number of incorrect codewords whose Hamming distance from \mathbf{y} is exactly d . Thus,

$$Z_e(\beta|\mathbf{y}) = (1 - p)^{\beta n} \sum_{d=0}^n \Omega_{\mathbf{y}}(d) \cdot e^{-\beta Jd}. \quad (209)$$

¹⁸This one is still independent, but it has a different distribution, and hence will be handled separately.

Just like in the REM, here too the enumerator $\Omega_{\mathbf{y}}(d)$ is the sum of an exponential number, e^{nR} , of binary i.i.d. RV's:

$$\Omega_{\mathbf{y}}(d) = \sum_{\mathbf{x} \in \mathcal{C} \setminus \{\mathbf{x}_0\}} \mathcal{I}\{d(\mathbf{x}, \mathbf{y}) = d\}. \quad (210)$$

According to the method of types, the probability of a single ‘success’ $\{d(\mathbf{X}, \mathbf{y}) = n\delta\}$ is given by

$$\Pr\{d(\mathbf{X}, \mathbf{y}) = n\delta\} = \frac{e^{nh_2(\delta)}}{2^n} = \exp\{-n[\ln 2 - h_2(\delta)]\}. \quad (211)$$

So, just like 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 \mathbf{y} . Otherwise, we see none. So the critical value of δ is the solution to the equation

$$R + h_2(\delta) - \ln 2 = 0. \quad (212)$$

There are two solutions to this equation, which are symmetric about $1/2$. The smaller one is called the Gilbert–Varshamov (G–V) distance¹⁹ and it will be denoted by $\delta_{GV}(R)$ (see Fig. 14). The other solution is, of course, $\delta = 1 - \delta_{GV}(R)$. Thus, the condition $R > \ln 2 - h_2(\delta)$

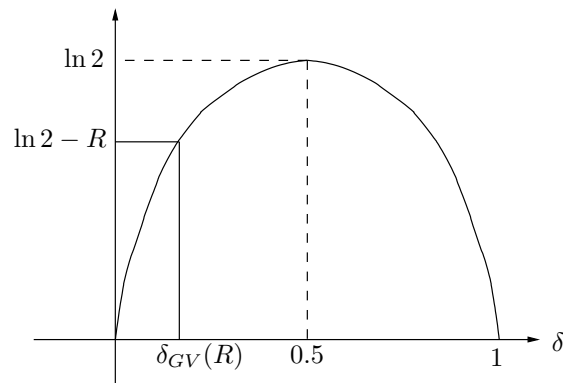


Figure 14: The Gilbert–Varshamov distance as the smaller solution to the equation $R + h_2(\delta) - \ln 2 = 0$.

¹⁹The G–V distance was originally defined and used in coding theory for the BSC.

is equivalent to $\delta_{GV}(R) < \delta < 1 - \delta_{GV}(R)$, and so, for a typical code in the ensemble:

$$\begin{aligned}
Z_e(\beta|\mathbf{y}) &\approx (1-p)^{\beta n} \sum_{\delta=\delta_{GV}(R)}^{1-\delta_{GV}(R)} \exp\{n[R + h_2(\delta) - \ln 2]\} \cdot e^{-\beta J n \delta} \\
&= (1-p)^{\beta n} e^{n(R-\ln 2)} \cdot \sum_{\delta=\delta_{GV}(R)}^{1-\delta_{GV}(R)} \exp\{n[h_2(\delta) - \beta J \delta]\} \\
&= (1-p)^{\beta n} e^{n(R-\ln 2)} \cdot \exp\left\{n \cdot \max_{\delta_{GV}(R) \leq \delta \leq 1-\delta_{GV}(R)} [h_2(\delta) - \beta J \delta]\right\}
\end{aligned}$$

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 endpoint (glassy phase) or at an internal point with zero derivative (paramagnetic phase). It is easy to show (exercise: fill in the details) that in the paramagnetic phase, the maximum is attained at

$$\delta^* = p_\beta \triangleq \frac{p^\beta}{p^\beta + (1-p)^\beta} \quad (213)$$

and then

$$\phi(\beta) = R - \ln 2 + \ln[p^\beta + (1-p)^\beta]. \quad (214)$$

In the glassy phase, $\delta^* = \delta_{GV}(R)$ and then

$$\phi(\beta) = \beta[\delta_{GV}(R) \ln p + (1 - \delta_{GV}(R)) \ln(1-p)], \quad (215)$$

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

$$\beta = \beta_c(R) \triangleq \frac{\ln[(1 - \delta_{GV}(R))/\delta_{GV}(R)]}{\ln[(1-p)/p]}. \quad (216)$$

So $\beta < \beta_c(R)$ 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 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. 15), which means reliable communication, as the correct codeword \mathbf{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.

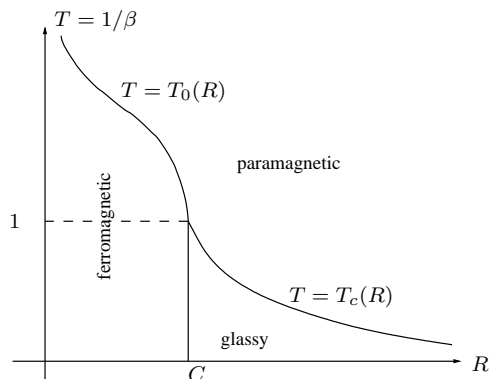


Figure 15: Phase diagram of the finite-temperature MAP decoder.

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 $H(\mathbf{X}|\mathbf{Y})$ 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 $P_\beta(\mathbf{x}|\mathbf{y})$, is about $e^{n(R-C)}$.

Exercise: convince yourself that the phase diagram is as depicted in Fig. 15 and find the equations of the boundaries between phases. Note that the triple point is $(C, 1)$ where

$C = \ln 2 - h_2(p)$ is the channel capacity. Also, the ferro-glassy boundary is the vertical straight line $R = C$. What does this mean? \square

5.3 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 $P^\beta(\mathbf{y}|\mathbf{x})$ (for some β), which are exactly the partition functions that we work with in the foregoing discussion. These considerations can sometimes even help to get tighter bounds. We will now demonstrate this point in the context of the analysis of the probability of correct decoding above capacity.

First, we have

$$\begin{aligned} P_c &= \frac{1}{M} \sum_{\mathbf{y}} \max_{\mathbf{x} \in \mathcal{C}} P(\mathbf{y}|\mathbf{x}) \quad M \triangleq e^{nR} \\ &= \lim_{\beta \rightarrow \infty} \frac{1}{M} \sum_{\mathbf{y}} \left[\sum_{\mathbf{x} \in \mathcal{C}} P^\beta(\mathbf{y}|\mathbf{x}) \right]^{1/\beta} \end{aligned}$$

The expression in the square brackets is readily identified with the partition function, and we note that the combination of $R > C$ and $\beta \rightarrow \infty$ takes us deep into the glassy phase. Taking the ensemble average, we get:

$$\bar{P}_c = \lim_{\beta \rightarrow \infty} \frac{1}{M} \sum_{\mathbf{y}} \mathbf{E} \left\{ \left[\sum_{\mathbf{x} \in \mathcal{C}} P^\beta(\mathbf{y}|\mathbf{x}) \right]^{1/\beta} \right\}. \quad (217)$$

At this point, the traditional approach would be to insert the expectation into the square brackets by applying Jensen's inequality (for $\beta > 1$), 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:

$$\begin{aligned}
\mathbf{E} \left\{ \left[\sum_{\mathbf{x} \in \mathcal{C}} P^\beta(\mathbf{y}|\mathbf{x}) \right]^{1/\beta} \right\} &= (1-p)^n \mathbf{E} \left\{ \left[\sum_{d=0}^n \Omega_{\mathbf{y}}(d) e^{-\beta J d} \right]^{1/\beta} \right\} \\
&\doteq (1-p)^n \mathbf{E} \left\{ \left[\max_{0 \leq d \leq n} \Omega_{\mathbf{y}}(d) e^{-\beta J d} \right]^{1/\beta} \right\} \\
&= (1-p)^n \mathbf{E} \left\{ \max_{0 \leq d \leq n} [\Omega_{\mathbf{y}}(d)]^{1/\beta} \cdot e^{-J d} \right\} \\
&\doteq (1-p)^n \mathbf{E} \left\{ \sum_{d=0}^n [\Omega_{\mathbf{y}}(d)]^{1/\beta} \cdot e^{-J d} \right\} \\
&= (1-p)^n \sum_{d=0}^n \mathbf{E} \{ [\Omega_{\mathbf{y}}(d)]^{1/\beta} \} \cdot e^{-J d} \\
&\doteq (1-p)^n \max_{0 \leq d \leq n} \mathbf{E} \{ [\Omega_{\mathbf{y}}(d)]^{1/\beta} \} \cdot e^{-J d}
\end{aligned}$$

Thus, it boils down to the calculation of (non-integer) moments of $\Omega_{\mathbf{y}}(d)$. 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 $\Omega_{\mathbf{y}}(d)$, we take a closer look at the asymptotic behavior of these RV's. 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

$$D(a||b) = a \ln \frac{a}{b} + (1-a) \ln \frac{1-a}{1-b}. \quad (218)$$

Using the inequality

$$\ln(1+x) = -\ln \left(1 - \frac{x}{1+x} \right) \geq \frac{x}{1+x},$$

we get the following lower bound to $D(a\|b)$:

$$\begin{aligned}
D(a\|b) &= a \ln \frac{a}{b} + (1-a) \ln \frac{1-a}{1-b} \\
&= a \ln \frac{a}{b} + (1-a) \ln \left(1 + \frac{b-a}{1-b} \right) \\
&\geq a \ln \frac{a}{b} + (1-a) \cdot \frac{(b-a)/(1-b)}{1 + (b-a)/(1-b)} \\
&= a \ln \frac{a}{b} + b - a \\
&> a \left(\ln \frac{a}{b} - 1 \right)
\end{aligned}$$

Now, as mentioned earlier, $\Omega_{\mathbf{y}}(d)$ is the sum of e^{nR} i.i.d. binary RV's, i.e., Bernoulli RV's with parameter $e^{-n[\ln 2 - h_2(\delta)]}$. Consider the event $\Omega_{\mathbf{y}}(d) \geq e^{nA}$, $A \geq 0$, 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,

$$A > R + h_2(\delta) - \ln 2. \quad (219)$$

Using the **Chernoff bound** (exercise: fill in the details), one can easily show that

$$\Pr\{\Omega_{\mathbf{y}}(d) \geq e^{nA}\} \leq \exp\{-e^{nR} D(e^{-n(R-A)} \| e^{-n[\ln 2 - h_2(\delta)]})\}. \quad (220)$$

Note: we have emphasized 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 $(e^{nR} + 1)$. Now, by applying the above lower bound to the binary divergence, we can further upper bound the last expression as

$$\begin{aligned}
\Pr\{\Omega_{\mathbf{y}}(d) \geq e^{nA}\} &\leq \exp\{-e^{nR} \cdot e^{-n(R-A)} \cdot (n[\ln 2 - R - h_2(\delta) + A] - 1)\} \\
&= \exp\{-e^{nA} \cdot (n[\ln 2 - R - h_2(\delta) + A] - 1)\}
\end{aligned}$$

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 $\Omega_{\mathbf{y}}(d)$ exceeds $\mathbf{E}\{\Omega_{\mathbf{y}}(d)\} \cdot e^{n\epsilon}$ decays double-exponentially fast with n . One can show in a similar manner (exercise: please do)²⁰ that $\Pr\{\Omega_{\mathbf{y}}(d) < \mathbf{E}\{\Omega_{\mathbf{y}}(d)\} \cdot e^{-n\epsilon}\}$ 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:

$$\Pr\{\Omega_{\mathbf{y}}(d) \geq 1\} \leq e^{-n[\ln 2 - R - h_2(\delta)]}. \quad (221)$$

Now, after having prepared these results, let's get back to the evaluation of the moments of $\Omega_{\mathbf{y}}(d)$. 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:

$$\begin{aligned} \mathbf{E}\{[\Omega_{\mathbf{y}}(d)]^{1/\beta}\} &\doteq 0^{1/\beta} \cdot \Pr\{\Omega_{\mathbf{y}}(d) = 0\} + e^{n \cdot 0/\beta} \cdot \Pr\{1 \leq \Omega_{\mathbf{y}}(d) \leq e^{n\epsilon}\} + \text{double-exp. terms} \\ &\doteq e^{n \cdot 0/\beta} \cdot \Pr\{\Omega_{\mathbf{y}}(d) \geq 1\} \\ &\doteq e^{-n[\ln 2 - R - h_2(\delta)]} \end{aligned}$$

Thus, in this range, $\mathbf{E}\{[\Omega_{\mathbf{y}}(d)]^{1/\beta}\} \doteq e^{-n[\ln 2 - R - h_2(\delta)]}$ independently of β . On the other hand in the range $\delta_{GV}(R) < \delta < 1 - \delta_{GV}(R)$,

$$\begin{aligned} \mathbf{E}\{[\Omega_{\mathbf{y}}(d)]^{1/\beta}\} &\doteq (e^{n[R+h_2(\delta)-\ln 2]})^{1/\beta} \cdot \Pr\{e^{n[R+h_2(\delta)-\ln 2-\epsilon]} \leq \Omega_{\mathbf{y}}(d) \leq e^{n[R+h_2(\delta)-\ln 2+\epsilon]}\} + \\ &\quad + \text{double-exp. terms} \\ &\doteq e^{n[R+h_2(\delta)-\ln 2]/\beta} \end{aligned}$$

since the probability $\Pr\{e^{n[R+h_2(\delta)-\ln 2-\epsilon]} \leq \Omega_{\mathbf{y}}(d) \leq e^{n[R+h_2(\delta)-\ln 2+\epsilon]}\}$ tends to unity double-exponentially rapidly. So to summarize, we have shown that the moment of $\Omega_{\mathbf{y}}(d)$ undergoes a phase transition, as it behaves as follows:

$$\mathbf{E}\{[\Omega_{\mathbf{y}}(d)]^{1/\beta}\} \doteq \begin{cases} e^{n[R+h_2(\delta)-\ln 2]/\beta} & \delta < \delta_{GV}(R) \text{ or } \delta > 1 - \delta_{GV}(R) \\ e^{n[R+h_2(\delta)-\ln 2]/\beta} & \delta_{GV}(R) < \delta < 1 - \delta_{GV}(R) \end{cases} \quad (222)$$

²⁰This requires a slightly different lower bound to the binary divergence.

Finally, by plugging these moments back into the expression of \bar{P}_c (exercise: fill in the details), and taking the limit $\beta \rightarrow \infty$, we eventually get:

$$\lim_{\beta \rightarrow \infty} \mathbf{E} \left\{ \left[\sum_{\mathbf{x} \in \mathcal{C}} P^\beta(\mathbf{y}|\mathbf{x}) \right]^{1/\beta} \right\} \doteq e^{-nF_g} \quad (223)$$

where F_g is the free energy of the glassy phase, i.e.,

$$F_g = \delta_{GV}(R) \ln \frac{1}{p} + (1 - \delta_{GV}(R)) \ln \frac{1}{1-p} \quad (224)$$

and so, we obtain a very simple relation between the exponent of \bar{P}_c and the free energy of the glassy phase:

$$\begin{aligned} \bar{P}_c &\doteq \frac{1}{M} \sum_{\mathbf{y}} e^{-nF_g} \\ &= \exp\{n(\ln 2 - R - F_g)\} \\ &= \exp\{n[\ln 2 - R + \delta_{GV}(R) \ln p + (1 - \delta_{GV}(R)) \ln(1-p)]\} \\ &= \exp\{n[h_2(\delta_{GV}(R)) + \delta_{GV}(R) \ln p + (1 - \delta_{GV}(R)) \ln(1-p)]\} \\ &= e^{-nD(\delta_{GV}(R)||p)} \end{aligned}$$

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” \mathbf{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 \mathbf{y} . The event of correct decoding happens if the channel flips less than $n\delta_{GV}(R)$ bits and then \mathbf{x}_0 is closer to \mathbf{y} more than all belt-codewords. Thus, \mathbf{x}_0 is decoded correctly.

One can also derive an upper bound on the error probability at $R < C$. The partition function $Z(\beta|\mathbf{y})$ 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

$$\frac{e^{-n\beta\hat{h}(\mathbf{x}_m\oplus\mathbf{y})}}{\sum_{m'\neq m} e^{-n\beta\hat{h}(\mathbf{x}_{m'}\oplus\mathbf{y})}} \geq e^{nT} \quad (225)$$

where $\beta > 0$ is an inverse temperature parameter and $\hat{h}(\mathbf{x} \oplus \mathbf{y})$ is the binary entropy pertaining to the relative number of 1's in the vector resulting from bit-by-bit XOR of \mathbf{x} and \mathbf{y} , namely, the binary entropy function computed at the normalized Hamming distance between \mathbf{x} and \mathbf{y} . If no message m satisfies (225), 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:

$$\text{Decide on } m \text{ if } \frac{P(\mathbf{y}|\mathbf{x}_m)}{\sum_{m'\neq m} P(\mathbf{y}|\mathbf{x}_{m'})} \geq e^{nT} \text{ and erase otherwise.} \quad (226)$$

The minimization of $\hat{h}(\mathbf{x}_m \oplus \mathbf{y})$ among all codewords $\{\mathbf{x}_m\}$, 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 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 \mathbf{x}_m is basically replaced by its maximum $e^{-n\hat{h}(\mathbf{x}_m\oplus\mathbf{y})}$, 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 $\beta \rightarrow \infty$,²¹ only the first and the second highest scores count in the decision, whereas if $\beta \rightarrow 0$, the differences between the scores of all codewords are washed out.

²¹As β varies it is plausible to let T scale linearly with β .

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. Let us analyze the probability of the event \mathcal{E}_1 that the transmitted codeword \mathbf{x}_m does not satisfy (225). We then have the following chain of inequalities, where the first few steps are common to the two analysis methods to be compared:

$$\begin{aligned}
\Pr\{\mathcal{E}_1\} &= \frac{1}{M} \sum_{m=1}^M \sum_{\mathbf{y}} P(\mathbf{y}|\mathbf{x}_m) \cdot \mathbf{1} \left\{ \frac{e^{nT} \sum_{m' \neq m} e^{-n\beta \hat{h}(\mathbf{x}_{m'} \oplus \mathbf{y})}}{e^{-n\beta \hat{h}(\mathbf{x}_m \oplus \mathbf{y})}} \geq 1 \right\} \\
&\leq \frac{1}{M} \sum_{m=1}^M \sum_{\mathbf{y}} P(\mathbf{y}|\mathbf{x}_m) \cdot \left[\frac{e^{nT} \sum_{m' \neq m} e^{-n\beta \hat{h}(\mathbf{x}_{m'} \oplus \mathbf{y})}}{e^{-n\beta \hat{h}(\mathbf{x}_m \oplus \mathbf{y})}} \right]^s \\
&= \frac{e^{nsT}}{M} \sum_{m=1}^M \sum_{\mathbf{y}} P(\mathbf{y}|\mathbf{x}_m) \cdot e^{n\beta s \hat{h}(\mathbf{x}_m \oplus \mathbf{y})} \cdot \left[\sum_{m' \neq m} e^{-n\beta \hat{h}(\mathbf{x}_{m'} \oplus \mathbf{y})} \right]^s \quad (227)
\end{aligned}$$

Considering now the ensemble of codewords drawn independently by fair coin tossing, we have:

$$\begin{aligned}
\overline{\Pr}\{\mathcal{E}_1\} &\leq e^{nsT} \sum_{\mathbf{y}} \mathbf{E} \left\{ P(\mathbf{y}|\mathbf{X}_1) \cdot \exp[n\beta s \hat{h}(\mathbf{X}_1 \oplus \mathbf{y})] \right\} \cdot \mathbf{E} \left\{ \left[\sum_{m>1} \exp[-n\beta \hat{h}(\mathbf{X}_m \oplus \mathbf{y})] \right]^s \right\} \\
&\triangleq e^{nsT} \sum_{\mathbf{y}} A(\mathbf{y}) \cdot B(\mathbf{y}) \quad (228)
\end{aligned}$$

The computation of $A(\mathbf{y})$ is as follows: Denoting the Hamming weight of a binary sequence \mathbf{z} by $w(\mathbf{z})$, we have:

$$\begin{aligned}
A(\mathbf{y}) &= \sum_{\mathbf{x}} 2^{-n} (1-p)^n \cdot \left(\frac{p}{1-p} \right)^{w(\mathbf{x} \oplus \mathbf{y})} \exp[n\beta s \hat{h}(\mathbf{x} \oplus \mathbf{y})] \\
&= \left(\frac{1-p}{2} \right)^n \sum_{\mathbf{z}} \exp \left[n \left(w(\mathbf{z}) \ln \frac{p}{1-p} + \beta s \hat{h}(\mathbf{z}) \right) \right] \\
&\doteq \left(\frac{1-p}{2} \right)^n \sum_{\delta} e^{nh(\delta)} \cdot \exp \left[n \left(\beta s h(\delta) - \delta \ln \frac{1-p}{p} \right) \right] \\
&\doteq \left(\frac{1-p}{2} \right)^n \exp \left[n \max_{\delta} \left((1 + \beta s) h(\delta) - \delta \ln \frac{1-p}{p} \right) \right]. \quad (229)
\end{aligned}$$

It is readily seen by ordinary optimization that

$$\max_{\delta} \left[(1 + \beta s) h(\delta) - \delta \ln \frac{1-p}{p} \right] = (1 + \beta s) \ln \left[p^{1/(1+\beta s)} + (1-p)^{1/(1+\beta s)} \right] - \ln(1-p) \quad (230)$$

and so upon substituting back into the the bound on $\overline{\Pr}\{\mathcal{E}_1\}$, we get:

$$\overline{\Pr}\{\mathcal{E}_1\} \leq \exp \left[n \left(sT + (1 + \beta s) \ln \left[p^{1/(1+\beta s)} + (1 - p)^{1/(1+\beta s)} \right] - \ln 2 \right) \right] \cdot \sum_{\mathbf{y}} B(\mathbf{y}). \quad (231)$$

It remains then to assess the exponential order of $B(\mathbf{y})$ and this will now be done in two different ways. The first is Forney's way of using Jensen's inequality and introducing the additional parameter ρ , i.e.,

$$\begin{aligned} B(\mathbf{y}) &= \mathbf{E} \left\{ \left(\left[\sum_{m>1} \exp[-n\beta\hat{h}(\mathbf{X}_m \oplus \mathbf{y})] \right]^{s/\rho} \right)^\rho \right\} \\ &\leq \mathbf{E} \left\{ \left(\sum_{m>1} \exp[-n\beta s \hat{h}(\mathbf{X}_m \oplus \mathbf{y}) / \rho] \right)^\rho \right\} \quad 0 \leq s/\rho \leq 1 \\ &\leq e^{n\rho R} \left(\mathbf{E} \left\{ \exp[-n\beta s \hat{h}(\mathbf{X}_m \oplus \mathbf{y}) / \rho] \right\} \right)^\rho, \quad \rho \leq 1 \end{aligned} \quad (232)$$

where in the second line we have used the following inequality²² for non-negative $\{a_i\}$ and $\theta \in [0, 1]$:

$$\left(\sum_i a_i \right)^\theta \leq \sum_i a_i^\theta. \quad (233)$$

Now,

$$\begin{aligned} \mathbf{E} \left\{ \exp[-n\beta s \hat{h}(\mathbf{X}_m \oplus \mathbf{y}) / \rho] \right\} &= 2^{-n} \sum_{\mathbf{z}} \exp[-n\beta s \hat{h}(\mathbf{z}) / \rho] \\ &\doteq 2^{-n} \sum_{\delta} e^{nh(\delta)} \cdot e^{-n\beta s h(\delta) / \rho} \\ &= \exp[n([1 - \beta s / \rho]_+ - 1) \ln 2], \end{aligned} \quad (234)$$

where $[u]_+ \triangleq \max\{u, 0\}$. Thus, we get

$$B(\mathbf{y}) \leq \exp(n[\rho(R - \ln 2) + [\rho - \beta s]_+]), \quad (235)$$

²²To 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.

which when substituted back into the bound on $\overline{\text{Pr}}\{\mathcal{E}_1\}$, yields an exponential rate of

$$\begin{aligned} \tilde{E}_1(R, T) &= \max_{0 \leq s \leq \rho \leq 1} \{(\rho - [\rho - \beta s]_+) \ln 2 - \\ &\quad - (1 + \beta s) \ln [p^{1/(1+\beta s)} + (1-p)^{1/(1+\beta s)}] - \rho R - sT\}. \end{aligned} \quad (236)$$

On the other hand, estimating $B(\mathbf{y})$ by the new method, we have:

$$\begin{aligned} B(\mathbf{y}) &= \mathbf{E} \left\{ \left[\sum_{m>1} \exp[-n\beta \hat{h}(\mathbf{X}_m \oplus \mathbf{y})] \right]^s \right\} \\ &= \mathbf{E} \left\{ \left[\sum_{\delta} \Omega_{\mathbf{y}}(n\delta) \exp[-n\beta h(\delta)] \right]^s \right\} \\ &\doteq \sum_{\delta} \mathbf{E}\{\Omega_{\mathbf{y}}^s(n\delta)\} \cdot \exp(-n\beta s h(\delta)) \\ &\doteq \sum_{\delta \in \mathcal{G}_R^c} e^{n[R+h(\delta)-\ln 2]} \cdot \exp[-n\beta s h(\delta)] + \sum_{\delta \in \mathcal{G}_R} e^{ns[R+h(\delta)-\ln 2]} \cdot \exp[-n\beta s h(\delta)] \\ &\triangleq U + V, \end{aligned} \quad (237)$$

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 $U \doteq \exp[-n(\ln 2 - R)(1 - [1 - \beta s]_+)]$. Similarly, V is dominated by the term $\delta = 1/2$ if $\beta < 1$ and $\delta = \delta_{GV}(R)$ if $\beta \geq 1$. Thus, $V \doteq \exp[-ns(\beta[\ln 2 - R] - R[1 - \beta]_+)]$. Therefore, defining

$$\phi(R, \beta, s) = \min\{(\ln 2 - R)(1 - [1 - \beta s]_+), s(\beta[\ln 2 - R] - R[1 - \beta]_+)\}, \quad (238)$$

the resulting exponent is

$$\hat{E}_1(R, T) = \max_{s \geq 0} \{\phi(R, \beta, s) - (1 + \beta s) \ln [p^{1/(1+\beta s)} + (1-p)^{1/(1+\beta s)}] - sT\}. \quad (239)$$

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 from 0 to 0.06 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

	$R = 0.00$	$R = 0.01$	$R = 0.02$	$R = 0.03$	$R = 0.04$	$R = 0.05$	$R = 0.06$
$\tilde{E}_1(R, T)$	0.1390	0.1290	0.1190	0.1090	0.0990	0.0890	0.0790
$\hat{E}_1(R, T)$	0.2211	0.2027	0.1838	0.1642	0.1441	0.1231	0.1015

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 $T = 0.001$.

interval $[0, 1]$ (since Jensen's inequality is not used), the numerical search over s was limited to the interval $[0, 5]$.²³

As can be seen (see also Fig. 16), 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 technique proposed here, not only simplifies exponential error bounds, but sometimes leads also to significantly tighter bounds.

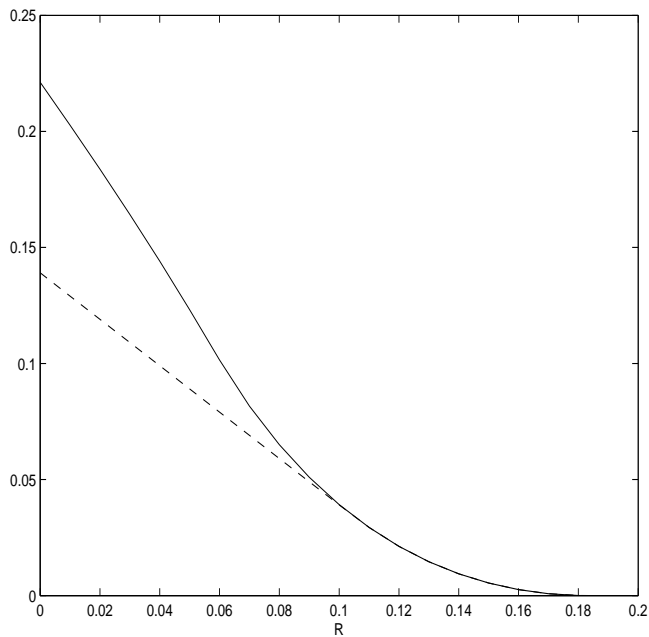


Figure 16: Graphs of $\hat{E}_1(R, T)$ (solid line) and $\tilde{E}_1(R, T)$ (dashed line) as functions of R for $p = 0.1$, $T = 0.001$ and $\beta = 0.5$.

There are other examples where these techniques are used in more involved situations,

²³ 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.

and in some of them they yield better performance bounds compared to traditional methods. Here is a partial list of papers:

- 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.
- 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.
- A. Somekh–Baruch and N. Merhav, “Exact random coding exponents for erasure decoding,” to appear in *Proc. ISIT 2010*, June 2010, Austin, Texas, U.S.A.

6 Additional Topics (Optional)

6.1 The REM With a Magnetic Field and Joint Source–Channel Coding

6.1.1 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.

We are now going to 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

$$\mathcal{E}(\mathbf{s}) = -B \sum_{i=1}^n s_i + \mathcal{E}_I(\mathbf{s}) \quad (240)$$

where $\mathcal{E}_I(\mathbf{s})$ 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

$$\begin{aligned} Z(\beta, B) &= \sum_{\mathbf{s}} e^{-\beta \mathcal{E}(\mathbf{s})} \\ &= \sum_{\mathbf{s}} e^{-\beta \mathcal{E}_I(\mathbf{s}) + \beta B \sum_{i=1}^n s_i} \\ &= \sum_{\mathbf{s}} e^{-\beta \mathcal{E}_I(\mathbf{s}) + n\beta B m(\mathbf{s})} \quad m(\mathbf{s}) = \frac{1}{n} \sum_i s_i \\ &= \sum_m \left[\sum_{\mathbf{s}: m(\mathbf{s})=m} e^{-\beta \mathcal{E}_I(\mathbf{s})} \right] \cdot e^{+n\beta B m} \\ &\triangleq \sum_m Z_0(\beta, m) \cdot e^{+n\beta B m} \end{aligned}$$

where $Z_0(\beta, m)$ is the *partial partition function*, defined to be the expression in the square

brackets in the second to the last line.²⁴ 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,

$$Z_0(\beta, m) \doteq e^{n\psi(\beta, m)} \quad (241)$$

with

$$\begin{aligned} \psi(\beta, m) &= \max_{|\epsilon| \leq J \sqrt{h_2((1+m)/2)}} \left[h_2\left(\frac{1+m}{2}\right) - \left(\frac{\epsilon}{J}\right)^2 - \beta\epsilon \right] \\ &= \begin{cases} h_2\left(\frac{1+m}{2}\right) + \frac{\beta^2 J^2}{4} & \beta \leq \beta_m \triangleq \frac{2}{J} \sqrt{h_2\left(\frac{1+m}{2}\right)} \\ \beta J \sqrt{h_2\left(\frac{1+m}{2}\right)} & \beta > \beta_m \end{cases} \end{aligned}$$

and from the above relation between Z and Z_0 , we readily have the Legendre relation

$$\phi(\beta, B) = \max_m [\psi(\beta, m) + \beta m B]. \quad (242)$$

For small β (high temperature), the maximizing (dominant) m is attained with zero-derivative:

$$\frac{\partial}{\partial m} \left[h_2\left(\frac{1+m}{2}\right) + \frac{\beta^2 J^2}{4} + \beta m B \right] = 0 \quad (243)$$

that is

$$\frac{1}{2} \ln \frac{1-m}{1+m} + \beta B = 0 \quad (244)$$

which yields

$$m^* = m_p(\beta, B) \triangleq \tanh(\beta B) \quad (245)$$

which is exactly the paramagnetic characteristic of magnetization vs. magnetic field (like that of i.i.d. spins), hence the name “paramagnetic phase.” Thus, plugging $m^* = \tanh(\beta B)$ back into the expression of ϕ , we get:

$$\phi(\beta, B) = h_2\left(\frac{1 + \tanh(\beta B)}{2}\right) + \frac{\beta^2 J^2}{4} + \beta B \tanh(\beta B). \quad (246)$$

²⁴Note 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.

This solution is valid as long as the condition

$$\beta \leq \beta_{m^*} = \frac{2}{J} \sqrt{h_2 \left(\frac{1 + \tanh(\beta B)}{2} \right)} \quad (247)$$

holds, or equivalently, the condition

$$\frac{\beta^2 J^2}{4} \leq h_2 \left(\frac{1 + \tanh(\beta B)}{2} \right). \quad (248)$$

Now, let us denote by $\beta_c(B)$ the solution β to the equation:

$$\frac{\beta^2 J^2}{4} = h_2 \left(\frac{1 + \tanh(\beta B)}{2} \right). \quad (249)$$

As can be seen from the graphical illustration (Fig. 17), $\beta_c(B)$ is a decreasing function and hence $T_c(B) \triangleq 1/\beta_c(B)$ is increasing. Thus, the phase transition temperature is increasing with $|B|$ (see Fig. 18). Below $\beta = \beta_c(B)$, we are in the glassy phase, where ϕ is given by:

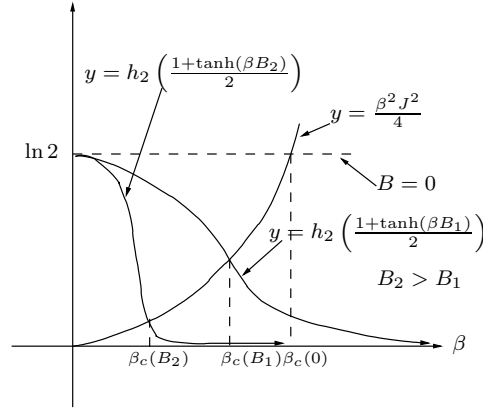


Figure 17: Graphical presentation of the solution $\beta_c(B)$ to the equation $\frac{1}{4}\beta^2 J^2 = h_2((1 + \tanh(\beta B))/2)$ for various values of B .

$$\phi(\beta, B) = \max_m \left[\beta J \sqrt{h_2 \left(\frac{1 + m}{2} \right)} + \beta m B \right] = \beta \cdot \max_m \left[J \sqrt{h_2 \left(\frac{1 + m}{2} \right)} + m B \right] \quad (250)$$

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:

$$m^* = m_g(B) \triangleq \tanh(B\beta_c(B)). \quad (251)$$

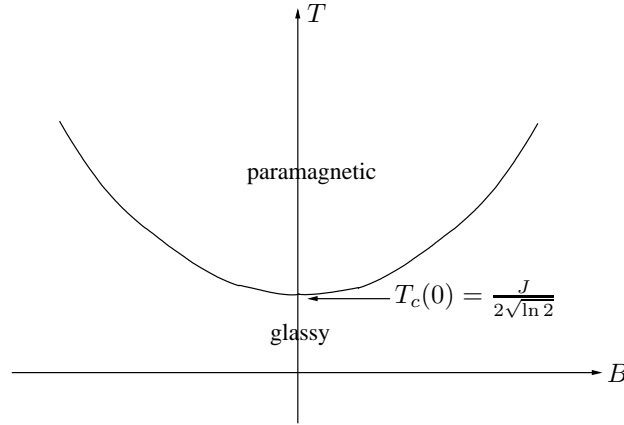


Figure 18: Phase diagram in the B - T plane.

Thus, in summary

$$\phi(\beta, B) = \begin{cases} h_2 \left(\frac{1+m_p(\beta, B)}{2} \right) + \frac{\beta^2 J^2}{4} + \beta B m_p(\beta, B) & \beta \leq \beta_c(B) \\ \beta J \sqrt{h_2 \left(\frac{1+m_g(B)}{2} \right)} + \beta B m_g(B) & \beta > \beta_c(B) \end{cases} \quad (252)$$

In both phases $B \rightarrow 0$ implies $m^* \rightarrow 0$, therefore the REM does not exhibit spontaneous magnetization, only a glass transition, as described.

Finally, we mention an important parameter in the physics of magnetic materials – the weak-field *magnetic susceptibility*, which is defined as $\chi \triangleq \frac{\partial m^*}{\partial B} |_{B=0}$. It can readily be shown that in the REM case

$$\chi = \begin{cases} \frac{1}{T} & T \geq T_c(0) \\ \frac{1}{T_c(0)} & T < T_c(0) \end{cases} \quad (253)$$

The graphical illustration of this function is depicted in Fig. 19. 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.

6.1.2 Relation to Joint Source–Channel Coding

We now relate these derivations to the behavior of joint source–channel coding systems. The full details of this part are in: N. Merhav, “The random energy model in a magnetic field

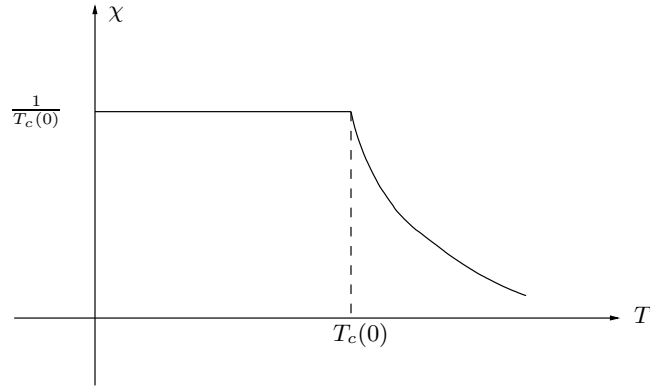


Figure 19: χ vs. T .

and joint source–channel coding,” *Physica A: Statistical Mechanics and Its Applications*, vol. 387, issue 22, pp. 5662–5674, September 15, 2008.

Consider again our coded communication system with a few slight modifications (cf. Fig. 20). Rather than e^{nR} equiprobable messages for channel coding, we are now talking about joint source–channel coding 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 $\mathbf{s} \in \{-1, +1\}^N$ 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 $\mathbf{x} \in \{0, 1\}^n$, such that the ratio $\theta = n/N$, the *bandwidth expansion factor*, remains fixed. The mapping from \mathbf{s} to \mathbf{x} is the encoder. As before, we shall concern ourselves with random codes, namely, for every $\mathbf{s} \in \{-1, +1\}^N$, we randomly select an independent codevector $\mathbf{x}(\mathbf{s}) \in \{0, 1\}^n$ by fair coin tossing, as before. Thus, we randomly select 2^N codevectors, each one of length $n = N\theta$. As in the case of pure

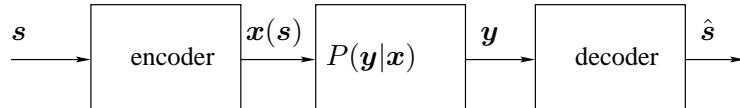


Figure 20: Block diagram of joint source–channel communication system.

channel coding, we consider the finite-temperature posterior:

$$P_\beta(\mathbf{s}|\mathbf{y}) = \frac{[P(\mathbf{s})P(\mathbf{y}|\mathbf{x}(\mathbf{s}))]^\beta}{Z(\beta|\mathbf{y})} \quad (254)$$

with

$$Z(\beta|\mathbf{y}) = \sum_{\mathbf{s}} [P(\mathbf{s})P(\mathbf{y}|\mathbf{x}(\mathbf{s}))]^\beta, \quad (255)$$

corresponding to the finite-temperature decoder:

$$\hat{s}_i = \arg \max_{s=\pm 1} \sum_{\mathbf{s}: s_i=s} [P(\mathbf{s})P(\mathbf{y}|\mathbf{x}(\mathbf{s}))]^\beta. \quad (256)$$

Once again, we separate the contributions of $Z_c(\beta|\mathbf{y}) = [P(\mathbf{s}_0)P(\mathbf{y}|\mathbf{x}(\mathbf{s}_0))]^\beta$, \mathbf{s}_0 being the true source message, and

$$Z_e(\beta|\mathbf{y}) = \sum_{\mathbf{s} \neq \mathbf{s}_0} [P(\mathbf{s})P(\mathbf{y}|\mathbf{x}(\mathbf{s}))]^\beta. \quad (257)$$

As we shall see quite shortly, Z_e behaves like the REM in a magnetic field given by $B = \frac{1}{2} \ln \frac{q}{1-q}$. 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 $N_1(\mathbf{s})$ the number of +1's in \mathbf{s} , so that the magnetization, $m(\mathbf{s}) \triangleq \frac{1}{N} [\sum_{i=1}^N 1\{s_i = +1\} - \sum_{i=1}^N 1\{s_i = -1\}]$, pertaining to spin configuration \mathbf{s} , is given by $m(\mathbf{s}) = 2N_1(\mathbf{s})/N - 1$. Equivalently, $N_1(\mathbf{s}) = N(1 + m(\mathbf{s}))/2$, and then

$$\begin{aligned} P(\mathbf{s}) &= q^{N_1(\mathbf{s})} (1-q)^{N-N_1(\mathbf{s})} \\ &= (1-q)^N \left(\frac{q}{1-q} \right)^{N(1+m(\mathbf{s}))/2} \\ &= [q(1-q)]^{N/2} \left(\frac{q}{1-q} \right)^{Nm(\mathbf{s})/2} \\ &= [q(1-q)]^{N/2} e^{Nm(\mathbf{s})B} \end{aligned}$$

where B is defined as above. By the same token, for the binary symmetric channel we have:

$$P(\mathbf{y}|\mathbf{x}) = p^{d_H(\mathbf{x}, \mathbf{y})} (1-p)^{n-d_H(\mathbf{x}, \mathbf{y})} = (1-p)^n e^{-Jd_H(\mathbf{x}, \mathbf{y})} \quad (258)$$

where $J = \ln \frac{1-p}{p}$ and $d_H(\mathbf{x}, \mathbf{y})$ is the Hamming distance, as defined earlier. Thus,

$$\begin{aligned}
Z_e(\beta, B) &= [q(1-q)]^{N\beta/2} \sum_m \left[\sum_{\mathbf{x}(\mathbf{s}): m(\mathbf{s})=m} e^{-\beta \ln[1/P(\mathbf{y}|\mathbf{x}(\mathbf{s}))]} \right] e^{N\beta m B} \\
&= [q(1-q)]^{\beta N/2} (1-p)^{n\beta} \sum_m \left[\sum_{\mathbf{x}(\mathbf{s}): m(\mathbf{s})=m} e^{-\beta J d_H(\mathbf{x}(\mathbf{s}), \mathbf{y})} \right] e^{\beta N m B} \\
&\triangleq [q(1-q)]^{N\beta/2} (1-p)^{n\beta} \sum_m Z_0(\beta, m|\mathbf{y}) e^{\beta N m B}
\end{aligned}$$

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 $\{d_H(\mathbf{x}(\mathbf{s}), \mathbf{y})\}$ that are induced by the random selection of the code $\{\mathbf{x}(\mathbf{s})\}$. Using the same considerations as with the REM in channel coding, we now get (exercise: fill in the details):

$$\begin{aligned}
\psi(\beta, m) &\triangleq \lim_{n \rightarrow \infty} \frac{\ln Z_0(\beta, m|\mathbf{y})}{n} \\
&= \max_{\delta_m \leq \delta \leq 1-\delta_m} \left[\frac{1}{\theta} h_2\left(\frac{1+m}{2}\right) + h_2(\delta) - \ln 2 - \beta J \delta \right] \quad \delta_m \triangleq \delta_{GV}\left(\frac{1}{\theta} h_2\left(\frac{1+m}{2}\right)\right) \\
&= \begin{cases} \frac{1}{\theta} h_2\left(\frac{1+m}{2}\right) + h_2(p_\beta) - \ln 2 - \beta J p_\beta & p_\beta \geq \delta_m \\ -\beta J \delta_m & p_\beta < \delta_m \end{cases}
\end{aligned}$$

where again,

$$p_\beta = \frac{p^\beta}{p^\beta + (1-p)^\beta}. \quad (259)$$

The condition $p_\beta \geq \delta_m$ is equivalent to

$$\beta \leq \beta_0(m) \triangleq \frac{1}{J} \ln \frac{1-\delta_m}{\delta_m}. \quad (260)$$

Finally, back to the full partition function:

$$\phi(\beta, B) = \lim_{n \rightarrow \infty} \frac{1}{N} \ln \left[\sum_m Z_0(\beta, m|\mathbf{y}) e^{N\beta m B} \right] = \max_m [\theta \psi(\beta, m) + \beta m B]. \quad (261)$$

For small enough β , the dominant m is the one that maximizes $[h_2((1+m)/2) + \beta m B]$, which is again the paramagnetic magnetization

$$m^* = m_p(\beta, B) = \tanh(\beta B). \quad (262)$$

Thus, in high decoding temperatures, the source vectors $\{\mathbf{s}\}$ that dominate the posterior $P_\beta(\mathbf{s}|\mathbf{y})$ 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:

$$m^* = m_g(B) \triangleq \tanh(B\beta_c(B)) \quad (263)$$

where this time, $\beta_c(B)$, the glassy–paramagnetic boundary, is defined as the solution to the equation

$$\ln 2 - h_2(p_\beta) = \frac{1}{\theta} h_2 \left(\frac{1 + \tanh(\beta B)}{2} \right). \quad (264)$$

The full details are in the paper. Taking now into account also Z_c , we get a phase diagram as depicted in Fig. 21. Here,

$$B_0 \triangleq \frac{1}{2} \ln \frac{q^*}{1-q^*} \quad (265)$$

where q^* is the solution to the equation

$$h_2(q) = \theta[\ln 2 - h_2(p)], \quad (266)$$

namely, it is the boundary between reliable and unreliable communication.

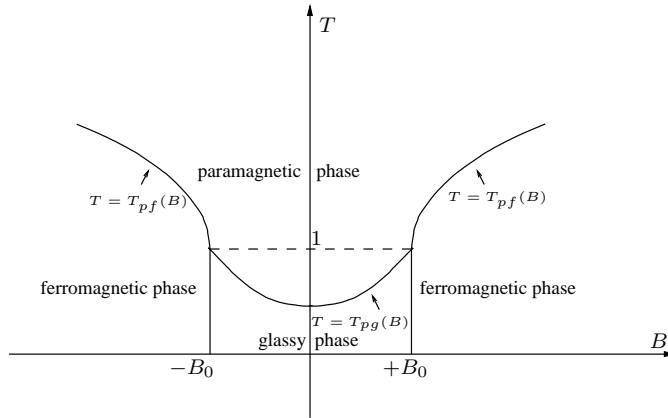


Figure 21: Phase diagram of joint source–channel communication system.

6.2 The Generalized Random Energy Model (GREM) and Hierarchical Coding

In the mid-eighties of the previous century, Derrida extended the REM to the generalized REM (GREM), 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. This part is based on the following papers:

- 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.
- B. Derrida and E. Gardner, “Solution of the generalised random energy model,” *J. Phys. C: Solid State Phys.*, vol. 19, pp. 2253–2274, 1986.
- 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.

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.²⁵ The easiest way to describe it is with a tree (see Fig. 22), each leaf of which represents one spin configuration. Now, for each branch in this tree, we randomly draw an independent random variable, which

²⁵ 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 $\epsilon_i \sim \mathcal{N}(0, a_1 n J^2 / 2)$, $1 \leq i \leq e^{nR_1}$. 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 $\epsilon_{i,j} \sim \mathcal{N}(0, a_2 n J^2 / 2)$. Finally, we define the energy associated with each configuration, or equivalently, each leaf indexed by (i, j) , as $E_{i,j} = \epsilon_i + \epsilon_{i,j}$, $1 \leq i \leq e^{nR_1}$, $1 \leq j \leq e^{nR_2}$.

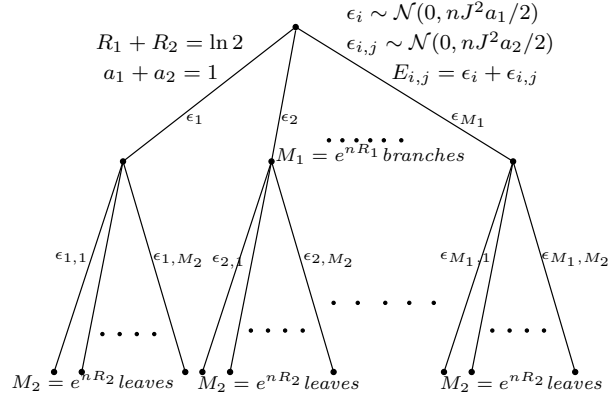


Figure 22: The GREM with $K = 2$ stages.

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 $\mathcal{N}(0, a_\ell n J^2 / 2)$ 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 say a few words about it later, 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:

$$\Omega_1(\epsilon) \approx \begin{cases} 0 & |\epsilon| > nJ\sqrt{a_1R_1} \\ \exp \left\{ n \left[R_1 - \frac{1}{a_1} \left(\frac{\epsilon}{nJ} \right)^2 \right] \right\} & |\epsilon| < nJ\sqrt{a_1R_1} \end{cases} \quad (267)$$

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,

$$\Omega_2(E) \approx \int_{-nJ\sqrt{a_1R_1}}^{+nJ\sqrt{a_1R_1}} d\epsilon \Omega_1(\epsilon) \cdot \exp \left\{ n \left[R_2 - \frac{1}{a_2} \left(\frac{E - \epsilon}{nJ} \right)^2 \right] \right\}. \quad (268)$$

It is important to understand here the following point: Here, we *no longer* zero-out the factor

$$\exp \left\{ n \left[R_2 - \frac{1}{a_2} \left(\frac{E - \epsilon}{nJ} \right)^2 \right] \right\} \quad (269)$$

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

$$\exp \left\{ -\frac{n}{a_2} \left(\frac{E - \epsilon}{nJ} \right)^2 \right\}. \quad (270)$$

Thus, whatever counts is whether the *entire* integrand has a positive exponent or not.

Consider next the entropy. The entropy behaves as follows:

$$\Sigma(E) = \lim_{n \rightarrow \infty} \frac{\ln \Omega_2(E)}{n} = \begin{cases} \Sigma_0(E) & \Sigma_0(E) \geq 0 \\ -\infty & \Sigma_0(E) < 0 \end{cases} \quad (271)$$

where $\Sigma_0(E)$ is the exponential rate of the above integral, which after applying the Laplace method, is shown to be:

$$\Sigma_0(E) = \max_{|\epsilon| \leq nJ\sqrt{a_1R_1}} \left[R_1 - \frac{1}{a_1} \left(\frac{\epsilon}{nJ} \right)^2 + R_2 - \frac{1}{a_2} \left(\frac{E - \epsilon}{nJ} \right)^2 \right]. \quad (272)$$

How does the function $\Sigma(E)$ behave like?

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) $R_1/a_1 \geq R_2/a_2$.²⁶ 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 $E_1 \triangleq nJ\sqrt{R_1/a_1}$. This is a phase transition due to the fact that the maximizing ϵ becomes an edgepoint 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:

$$\Sigma(E) = \begin{cases} \ln 2 - \left(\frac{E}{nJ}\right)^2 & |E| \leq \hat{E} \\ -\infty & |E| > \hat{E} \end{cases} \quad (273)$$

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:

$$\phi(\beta) = \begin{cases} \ln 2 + \frac{\beta^2 J^2}{4} & \beta \leq \beta_0 \triangleq \frac{2}{J}\sqrt{\ln 2} \\ \beta J\sqrt{\ln 2} & \beta > \beta_0 \end{cases} \quad (274)$$

As said, the condition for this is:

$$nJ\sqrt{\ln 2} \equiv \hat{E} \leq E_1 \equiv nJ\sqrt{\frac{R_1}{a_1}} \quad (275)$$

or, equivalently,

$$\frac{R_1}{a_1} \geq \ln 2. \quad (276)$$

²⁶Accordingly, 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:

$$\Sigma(E) = \begin{cases} \ln 2 - \left(\frac{E}{nJ}\right)^2 & |E| \leq E_1 \\ R_2 - \frac{1}{a_2} \left(\frac{E}{nJ} - \sqrt{a_1 R_1}\right)^2 & E_1 < |E| \leq \hat{E} \\ -\infty & |E| > \hat{E} \end{cases} \quad (277)$$

and accordingly (exercise: please show this):

$$\phi(\beta) = \begin{cases} \ln 2 + \frac{\beta^2 J^2}{4} & \beta \leq \beta_1 \triangleq \frac{2}{J} \sqrt{\frac{R_1}{a_1}} \\ \beta J \sqrt{a_1 R_1} + R_2 + \frac{a_2 \beta^2 J^2}{4} & \beta_1 \leq \beta < \beta_2 \triangleq \frac{2}{J} \sqrt{\frac{R_2}{a_2}} \\ \beta J (\sqrt{a_1 R_1} + \sqrt{a_2 R_2}) & \beta \geq \beta_2 \end{cases} \quad (278)$$

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 a paper by Cover and Ordentlich (IT Transactions, March 1996), the following inequality is proved for two positive vectors (a_1, \dots, a_n) and (b_1, \dots, b_n) :

$$\min_i \frac{a_i}{b_i} \leq \frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i} \leq \max_i \frac{a_i}{b_i}. \quad (279)$$

Thus,

$$\min_{i \in \{1,2\}} \frac{R_i}{a_i} \leq \frac{R_1 + R_2}{a_1 + a_2} \leq \max_{i \in \{1,2\}} \frac{R_i}{a_i}, \quad (280)$$

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 shown that:

Case A: $R_1/a_1 < R_2/a_2$ – two phase transitions:

$$\phi(\beta) = \begin{cases} \ln 2 + \frac{\beta^2 J^2}{4} & \beta \leq \beta_1 \\ \beta J \sqrt{a_1 R_1} + R_2 + \frac{a_2 \beta^2 J^2}{4} & \beta_1 \leq \beta < \beta_2 \\ \beta J (\sqrt{a_1 R_1} + \sqrt{a_2 R_2}) & \beta \geq \beta_2 \end{cases} \quad (281)$$

Case B: $R_1/a_1 \geq R_2/a_2$ – one phase transition, like in the REM:

$$\phi(\beta) = \begin{cases} \ln 2 + \frac{\beta^2 J^2}{4} & \beta \leq \beta_0 \\ \beta J \sqrt{\ln 2} & \beta > \beta_0 \end{cases} \quad (282)$$

We now move on to our coding problem, this time it is about source coding with a fidelity criterion. For simplicity, we will assume a binary symmetric source (BSS) and the Hamming distortion. 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 $\hat{\mathcal{C}} = \{\hat{\mathbf{x}}_i, 1 \leq i \leq e^{n_1 R_1}\}$. For the second segment, we do the following: For each $1 \leq i \leq e^{n_1 R_1}$, we randomly select (again, by fair coin tossing) a codebook $\tilde{\mathcal{C}}_i = \{\tilde{\mathbf{x}}_{i,j}, 1 \leq j \leq e^{n_2 R_2}\}$. Now, given a source vector $\mathbf{x} \in \{0, 1\}^n$, segmented as $(\mathbf{x}', \mathbf{x}'')$, the encoder seeks a pair (i, j) , $1 \leq i \leq e^{n_1 R_1}$, $1 \leq j \leq e^{n_2 R_2}$, such that $d(\mathbf{x}', \hat{\mathbf{x}}_i) + d(\mathbf{x}'', \tilde{\mathbf{x}}_{i,j})$ 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 $d(\mathbf{x}', \hat{\mathbf{x}}_i)$ and $d(\mathbf{x}'', \tilde{\mathbf{x}}_{i,j})$. 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 check out..*

As for the performance criterion, here, we choose to examine performance in terms of the characteristic function of the overall distortion, $\mathbf{E}[\exp\{-s \cdot \text{distortion}\}]$. 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 \mathbf{x} , we define

$$\Delta(\mathbf{x}) = \min_{\hat{\mathbf{x}} \in \mathcal{C}} d(\mathbf{x}, \hat{\mathbf{x}}), \quad (283)$$

and we will be interested in the exponential rate of

$$\Psi(s) \triangleq \mathbf{E}\{\exp[-s\Delta(\mathbf{X})]\}. \quad (284)$$

This quantity can be easily related to the “partition function”:

$$Z(\beta|\mathbf{x}) \triangleq \sum_{\hat{\mathbf{x}} \in \mathcal{C}} e^{-\beta d(\mathbf{x}, \hat{\mathbf{x}})}. \quad (285)$$

In particular,

$$\mathbf{E}\{\exp[-s\Delta(\mathbf{X})]\} = \lim_{\theta \rightarrow \infty} \mathbf{E}\{[Z(s \cdot \theta|\mathbf{X})]^{1/\theta}\}. \quad (286)$$

Thus, to analyze the characteristic function of the distortion, we have to assess (noninteger) moments of the partition function.

Let's first see what happens with ordinary random block codes, without any structure. This calculation is very similar the one we did before in the context of channel coding:

$$\begin{aligned}
\mathbf{E} \{ [Z(s \cdot \theta | \mathbf{X})]^{1/\theta} \} &= \mathbf{E} \left\{ \left[\sum_{\hat{\mathbf{x}} \in \mathcal{C}} e^{-s\theta d(\mathbf{x}, \hat{\mathbf{x}})} \right]^{1/\theta} \right\} \\
&= \mathbf{E} \left\{ \left[\sum_{d=0}^n \Omega(d) e^{-s\theta d} \right]^{1/\theta} \right\} \\
&\doteq \sum_{d=0}^n \mathbf{E} \{ [\Omega(d)]^{1/\theta} \} \cdot e^{-sd}
\end{aligned}$$

where, as we have already shown in the past:

$$\mathbf{E} \{ [\Omega(d)]^{1/\theta} \} \doteq \begin{cases} e^{n[R+h_2(\delta)-\ln 2]} & \delta \leq \delta_{GV}(R) \text{ or } \delta \geq 1 - \delta_{GV}(R) \\ e^{n[R+h_2(\delta)-\ln 2]/\theta} & \delta_{GV}(R) \leq \delta \leq 1 - \delta_{GV}(R) \end{cases} \quad (287)$$

Note that $\delta_{GV}(R)$ is exactly the distortion-rate function of the BSS w.r.t. the Hamming distortion. By plugging the expression of $\mathbf{E}\{[\Omega(d)]^{1/\theta}\}$ back into that of $\mathbf{E}\{[Z(s \cdot \theta | \mathbf{X})]^{1/\theta}\}$ and carrying out the maximization pertaining to the dominant contribution, we eventually (exercise: please show that) obtain:

$$\Psi(s) \doteq e^{-nu(s,R)} \quad (288)$$

where

$$\begin{aligned}
u(s, R) &= \ln 2 - R - \max_{\delta \leq \delta_{GV}(R)} [h_2(\delta) - s\delta] \\
&= \begin{cases} s\delta_{GV}(R) & s \leq s_R \\ v(s, R) & s > s_R \end{cases} \quad (289)
\end{aligned}$$

with

$$s_R \triangleq \ln \left[\frac{1 - \delta_{GV}(R)}{\delta_{GV}(R)} \right] \quad (290)$$

and

$$v(s, R) \triangleq \ln 2 - R + s - \ln(1 + e^s). \quad (291)$$

The function $u(s, R)$ is depicted qualitatively in Fig. 23.

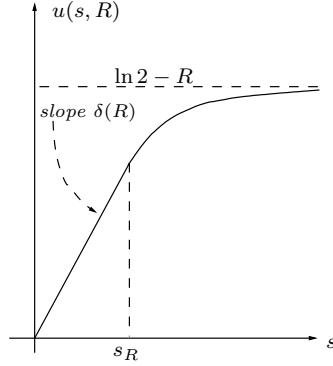


Figure 23: Qualitative graph of the function $u(s, R)$ as a function of s for fixed R .

Let's now move on to the hierarchical codes. The analogy with the GREM is fairly clear. Given \mathbf{x} , there are about $\Omega_1(\delta_1) \doteq e^{n_1[R_1+h_2(\delta_1)-\ln 2]}$ first-segment codewords $\{\hat{\mathbf{x}}_i\}$ in $\hat{\mathcal{C}}$ at distance $n_1\delta_1$ from the first segment \mathbf{x}' of \mathbf{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 $\{\tilde{\mathbf{x}}_{i,j}\}$ at distance $n_2\delta_2$ from the second segment \mathbf{x}'' of \mathbf{x} . Therefore, for $\delta = \lambda\delta_1 + (1-\lambda)\delta_2$,

$$\begin{aligned} \Omega_2(\delta) &= \sum_{\delta_1=\delta_{GV}(R_1)}^{1-\delta_{GV}(R_1)} e^{n_1[R_1+h_2(\delta_1)-\ln 2]} \cdot e^{n_2[R_2+h_2((\delta-\lambda\delta_1)/(1-\lambda))-\ln 2]} \\ &\doteq \exp \left\{ n \cdot \max_{\delta_1 \in [\delta_{GV}(R_1), 1-\delta_{GV}(R_1)]} \left[R + \lambda h_2(\delta_1) + (1-\lambda)h_2 \left(\frac{\delta - \lambda\delta_1}{1-\lambda} \right) \right] \right\} \end{aligned}$$

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:

$$\Sigma(\delta) = \begin{cases} R + h_2(\delta) - \ln 2 & \delta \in [\delta_{GV}(R), 1 - \delta_{GV}(R)] \\ -\infty & \text{elsewhere} \end{cases} \quad (292)$$

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:

$$\phi(\beta, R) = \begin{cases} -v(\beta, R) & \beta < \beta(R_1) \\ -\lambda\beta\delta_{GV}(R_1) - (1-\lambda)v(\beta, R_2) & \beta(R_1) < \beta < \beta(R_2) \\ -\beta[\lambda\delta_{GV}(R_1) + (1-\lambda)\delta_{GV}(R_2)] & \beta > \beta(R_2) \end{cases} \quad (293)$$

The last line is the purely glassy phase and this is the relevant phase because of the limit $\theta \rightarrow 0$ 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 than $\delta_{GV}(R)$ due to convexity arguments. Let's 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,

$$\mathbf{E} \{ Z^{1/\theta}(s \cdot \theta) \} = \mathbf{E} \left\{ \left[\sum_{d_1=0}^n \sum_{d_2=0}^n \Omega(d_1, d_2) \cdot e^{-s\theta(d_1+d_2)} \right]^{1/\theta} \right\} = \sum_{d_1=0}^n \sum_{d_2=0}^n \mathbf{E} \{ \Omega^{1/\theta}(d_1, d_2) \} \cdot e^{-s(d_1+d_2)}. \quad (294)$$

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 the paper). The final results are as follows:

- For $R_1 < R_2$,

$$\lim_{n \rightarrow \infty} \left[-\frac{1}{n} \ln \mathbf{E} \exp\{-s\Delta(\mathbf{X})\} \right] = \lambda u(s, R_1) + (1 - \lambda)u(s, R_2) \quad (295)$$

which means the behavior of two independent, decoupled codes for the two segments, which is bad, of course.

- For $R_1 \geq R_2$,

$$\lim_{n \rightarrow \infty} \left[-\frac{1}{n} \ln \mathbf{E} \exp\{-s\Delta(\mathbf{X})\} \right] = u(s, R) \quad \forall s \leq s_0 \quad (296)$$

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 given realization of the RV's $\{\varepsilon_{i,j} : i = 1, 2, \dots, n, j = 0, 1, \dots, d^i - 1\}$, we define the Hamiltonian associated with \mathbf{w} as $\mathcal{E}(\mathbf{w}) = \sum_{i=1}^n \varepsilon_{i,j_i}$, and then the partition function as:

$$Z_n(\beta) = \sum_{\mathbf{w}} \exp\{-\beta \mathcal{E}(\mathbf{w})\}. \quad (297)$$

It turns out that this model is exactly solvable (in many ways) and one can show (see e.g., 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) that it admits a glassy phase transition:

$$\phi(\beta) = \lim_{n \rightarrow \infty} \frac{\ln Z_n(\beta)}{n} = \begin{cases} \phi_0(\beta) & \beta < \beta_c \\ \phi_0(\beta_c) & \beta \geq \beta_c \end{cases} \quad \text{almost surely} \quad (298)$$

where

$$\phi_0(\beta) \triangleq \frac{\ln[d \cdot \mathbf{E} e^{-\beta \rho(\epsilon)}]}{\beta} \quad (299)$$

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, $\mathbf{x} = (x_1, \dots, x_n)$, 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, \dots, Y_d . We shall refer to the randomly chosen set, $\mathcal{C}_1 = \{Y_1, Y_2, \dots, Y_d\}$, 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 , $\mathcal{C}_{2,j_1} = \{Y_{j_1,1}, Y_{j_1,2}, \dots, Y_{j_1,d}\}$, 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 $\mathcal{C}_{3,j_1,j_2} = \{Y_{j_1,j_2,1}, Y_{j_1,j_2,2}, \dots, Y_{j_1,j_2,d}\}$, 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 $(j_1, j_2, \dots, j_{t-1})$, $1 \leq j_k \leq d$, $1 \leq k \leq t-1$.

Once the above described random code selection process is complete, the resulting set of codebooks $\{\mathcal{C}_1, \mathcal{C}_{t,j_1, \dots, j_{t-1}}, 2 \leq t \leq n, 1 \leq j_k \leq d, 1 \leq k \leq t-1\}$ 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 $(j_1^*, j_2^*, \dots, j_n^*)$ that minimizes the overall distortion $\sum_{t=1}^n \rho(X_t, Y_{j_1, \dots, j_t})$. 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^*, \dots, j_t^*) , output the reproduction symbol $Y_{j_1^*, \dots, j_t^*}$.

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 RV $\rho(x, Y)$ 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 \mathbf{x} and a given realization of the set of codebooks, define the partition function in analogy to that of the DPRM:

$$Z_n(\beta) = \sum_{\mathbf{w}} \exp\left\{-\beta \sum_{t=1}^n \rho(x_t, Y_{j_1, \dots, j_t})\right\}, \quad (300)$$

where the summation extends over all d^n possible walks, $\mathbf{w} = (j_1, \dots, j_n)$, along the Cayley tree. Clearly, considering our symmetry assumption, this falls exactly under the umbrella of the DPRM, with the distortions $\{\rho(x_t, Y_{j_1, \dots, j_t})\}$ 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

$$\phi(\beta) = \begin{cases} \phi_0(\beta) & \beta \leq \beta_c \\ \phi_0(\beta_c) & \beta > \beta_c \end{cases} \quad (301)$$

where now

$$\begin{aligned}
\phi_0(\beta) &\triangleq \frac{\ln[d \cdot \mathbf{E}\{e^{-\beta\rho(x,Y)}\}]}{\beta} \\
&= \frac{\ln[e^R \cdot \mathbf{E}\{e^{-\beta\rho(x,Y)}\}]}{\beta} \\
&= \frac{R + \ln[\mathbf{E}\{e^{-\beta\rho(x,Y)}\}]}{\beta},
\end{aligned}$$

Thus, for every (x_1, x_2, \dots) , the distortion is given by

$$\begin{aligned}
\limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{t=1}^n \rho(x_t, Y_{j_1^*, \dots, j_t^*}) &\triangleq \limsup_{n \rightarrow \infty} \frac{1}{n} \min_{\mathbf{w}} \left[\sum_{t=1}^n \rho(x_t, Y_{j_1, \dots, j_t}) \right] \\
&= \limsup_{n \rightarrow \infty} \limsup_{\ell \rightarrow \infty} \left[-\frac{\ln Z_n(\beta_\ell)}{n\beta_\ell} \right] \\
&\leq \limsup_{\ell \rightarrow \infty} \limsup_{n \rightarrow \infty} \left[-\frac{\ln Z_n(\beta_\ell)}{n\beta_\ell} \right] \\
&\stackrel{\text{a.s.}}{=} -\liminf_{\ell \rightarrow \infty} \phi(\beta_\ell) \\
&= -\phi_0(\beta_c) \\
&= \max_{\beta \geq 0} \left[-\frac{\ln[\mathbf{E}\{e^{-\beta\rho(x,Y)}\}] + R}{\beta} \right] \\
&= D(R),
\end{aligned}$$

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 at the third line is left as an exercise. The last equation is easily obtained by inverting the function $R(D)$ in its parametric representation that we have seen earlier:

$$R(D) = -\min_{\beta \geq 0} \min_Q \left\{ \beta D + \sum_{x \in \mathcal{X}} p(x) \ln \left[\sum_{y \in \mathcal{Y}} q(y) e^{-\beta\rho(x,y)} \right] \right\}. \quad (302)$$

Thus, the ensemble of tree codes achieves $R(D)$ almost surely.

6.3 Phase Transitions of the Rate–Distortion Function

The material in this part is based on the paper: K. Rose, “A mapping approach to rate–distortion computation and analysis,” *IEEE Trans. Inform. Theory*, vol. 40, no. 6, pp. 1939–1952, November 1994.

We have seen in one of the earlier meetings that the rate–distortion function of a source $P = \{p(x), x \in \mathcal{X}\}$ can be expressed as

$$R(D) = - \min_{\beta \geq 0} \left[\beta D + \sum_x p(x) \ln \left(\sum_y q(y) e^{-\beta d(x,y)} \right) \right] \quad (303)$$

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

$$f(Q) = -\frac{1}{\beta} \sum_x p(x) \ln \left(\sum_y q(y) e^{-\beta d(x,y)} \right) \quad (304)$$

and in the continuous case, summations should be replaced by integrals:

$$f(Q) = -\frac{1}{\beta} \int_{-\infty}^{+\infty} dx p(x) \ln \left(\int_{-\infty}^{+\infty} dy q(y) e^{-\beta d(x,y)} \right). \quad (305)$$

Rose suggests to represent the RV Y as a function of $U \sim \text{unif}[0, 1]$, and then, instead of optimizing Q , one should optimize the function $y(u)$ in:

$$f(y(\cdot)) = -\frac{1}{\beta} \int_{-\infty}^{+\infty} dx p(x) \ln \left(\int_0^1 d\mu(u) e^{-\beta d(x, y(u))} \right), \quad (306)$$

where $\mu(\cdot)$ is the Lebesgue measure (the uniform measure). A necessary condition for optimality,²⁸ which must hold for almost every u is:

$$\int_{-\infty}^{+\infty} dx p(x) \cdot \left[\frac{e^{-\beta d(x, y(u))}}{\int_0^1 d\mu(u') e^{-\beta d(x, y(u'))}} \right] \cdot \frac{\partial d(x, y(u))}{\partial y(u)} = 0. \quad (307)$$

²⁸The details are in the paper, but intuitively, instead of a function $y(u)$ of a continuous variable u , think of a vector \mathbf{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 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:

$$\int_{-\infty}^{+\infty} dx p(x) \cdot \left[\frac{e^{-\beta d(x, y_0)}}{\int_0^1 d\mu(u') e^{-\beta d(x, y(u'))}} \right] \cdot \left. \frac{\partial d(x, y(u))}{\partial y(u)} \right|_{y(u)=y_0} = 0. \quad (308)$$

This is because y_0 must be a point that is obtained as $y(u)$ for some u . Let us define now the posterior:

$$q(u|x) = \frac{e^{-\beta d(x, y(u))}}{\int_0^1 d\mu(u') e^{-\beta d(x, y(u'))}}. \quad (309)$$

Then,

$$\int_{-\infty}^{+\infty} dx p(x) q(u|x) \cdot \frac{\partial d(x, y(u))}{\partial y(u)} = 0. \quad (310)$$

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 :

$$\int_{-\infty}^{+\infty} dx p(x|u) \frac{\partial d(x, y(u))}{\partial y(u)} = 0. \quad (311)$$

Now, let's see what happens in the case of the quadratic distortion, $d(x, y) = (x - y)^2$. Let us 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

$$\int_{-\infty}^{+\infty} dx p(x) \lambda(x) (x - y) e^{-\beta(x-y)^2} = 0, \quad \forall y \in \mathcal{I}_0 \quad (312)$$

with

$$\lambda(x) \triangleq \frac{1}{\int_0^1 d\mu(u) e^{-\beta d(x, y(u))}} = \frac{1}{\int_{-\infty}^{+\infty} dy q(y) e^{-\beta d(x, y)}}, \quad (313)$$

or, equivalently,

$$\int_{-\infty}^{+\infty} dx p(x) \lambda(x) \frac{\partial}{\partial y} \left[e^{-\beta(x-y)^2} \right] = 0. \quad (314)$$

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.,

$$\int_{-\infty}^{+\infty} dx p(x) \lambda(x) \frac{\partial^n}{\partial y^n} \left[e^{-\beta(x-y)^2} \right] = 0. \quad (315)$$

Now, considering the Hermitian polynomials

$$H_n(z) \triangleq e^{\beta z^2} \frac{d^n}{dz^n} (e^{-\beta z^2}) \quad (316)$$

this requirement means

$$\int_{-\infty}^{+\infty} dx p(x) \lambda(x) H_n(x-y) e^{-\beta(x-y)^2} = 0. \quad (317)$$

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

$$p(x)\lambda(x) = \text{const.} \quad (318)$$

because $H_0(z) \equiv 1$ is the only basis function orthogonal to all $H_n(z)$, $n \geq 1$. This yields, after normalization:

$$p(x) = \sqrt{\frac{\beta}{\pi}} \int_0^1 d\mu(u) e^{-\beta(x-y(u))^2} = \sqrt{\frac{\beta}{\pi}} \int_{-\infty}^{+\infty} dy q(y) e^{-\beta(x-y)^2} = Q \star \mathcal{N}\left(0, \frac{1}{2\beta}\right). \quad (319)$$

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). This means that X must be representable as

$$X = Y + Z \quad (320)$$

where $Z \sim \mathcal{N}\left(0, \frac{1}{2\beta}\right)$ and independent of Y . From the Information Theory course we know that this is exactly what happens when $R(D)$ coincides with its Gaussian lower bound, a.k.a. the *Shannon lower bound*. Here is a reminder of this:

$$\begin{aligned} R(D) &= h(X) - \max_{\mathbf{E}(X-Y)^2 \leq D} h(X|Y) \\ &= h(X) - \max_{\mathbf{E}(X-Y)^2 \leq D} h(X-Y|Y) \\ &\geq h(X) - \max_{\mathbf{E}(X-Y)^2 \leq D} h(X-Y) \quad \text{equality if } (X-Y) \perp Y \\ &= h(X) - \max_{\mathbf{E}Z^2 \leq D} h(Z) \quad Z \triangleq X-Y \\ &\geq h(X) - \frac{1}{2} \ln(2\pi e D) \quad \text{equality if } Z \sim \mathcal{N}(0, D) \\ &\triangleq R_{\text{SLB}}(D) \end{aligned}$$

The conclusion then is that if the support of Y includes an interval (no matter how small) then $R(D)$ coincides with $R_{\text{SLB}}(D)$. This implies that in all those cases that $R_{\text{SLB}}(D)$ 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 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 RV's with variances $D_0 - D$ and D , respectively. Thus,

$$X = Y + Z = (Y + Z_1) + Z_2 \triangleq Y' + Z_2 \quad (321)$$

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:

$$R(D) \begin{cases} = R_{\text{SLB}}(D) & D \leq D_0 \\ > R_{\text{SLB}}(D) & D > D_0 \end{cases} \quad (322)$$

It follows then that in all these cases, the optimum output marginal contains intervals for all $D \leq D_0$ and then becomes abruptly singular as D exceeds D_0 . From the viewpoint of statistical mechanics, this looks like a phase transition, then. 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: $y = \mathbf{E}(X)$, so it is definitely singular. This remains true even if we increase β to 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 2 points, then 3 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 the paper.

An interesting topic for research evolves around possible extensions of these results to more general distortion measures, other than the quadratic distortion measure.

6.4 Capacity of the Sherrington–Kirkpatrick Spin Glass

This part is based on the paper: O. Shental and I. Kanter, “Shannon capacity of infinite-range spin-glasses,” Technical Report, Bar Ilan University, 2005. In this work, the authors consider the S–K model with independent Gaussian coupling coefficients, and they count the number $N(n)$ of meta-stable states in the absence of magnetic field. A meta-stable state means that each spin is in its preferred polarization according to the net field that it ‘feels’. i.e.,

$$s_i = \text{sgn} \left(\sum_j J_{ij} s_j \right), \quad i = 1, \dots, n. \quad (323)$$

They refer to the limit $\lim_{n \rightarrow \infty} [\ln N(n)]/n$ as the capacity C of the S–K model. However, they take an annealed rather than a quenched average, thus the resulting capacity is somewhat optimistic. The reason that this work is brought here is that many of the mathematical tools we have been exposed to are used here. The main result in this work is that

$$C = \ln[2(1 - Q(t))] - \frac{t^2}{2} \quad (324)$$

where

$$Q(t) \triangleq \frac{1}{2\pi} \int_t^\infty du \cdot e^{-u^2/2} \quad (325)$$

and t is the solution to the equation

$$t = \frac{e^{-t^2/2}}{\sqrt{2\pi}[1 - Q(t)]}. \quad (326)$$

The authors even address a slightly more general question: Quite obviously, the metastability condition is that for every i there exists $\lambda_i > 0$ such that

$$\lambda_i s_i = \sum_j J_{ij} s_j. \quad (327)$$

But they actually answer the following question: Given a constant K , what is the expected number of states for which there is $\lambda_i > K$ for each i such that $\lambda_i s_i = \sum_j J_{ij} s_j$? For $K \rightarrow -\infty$, one expects $C \rightarrow \ln 2$, and for $K \rightarrow \infty$, one expects $C \rightarrow 0$. The case of interest is exactly in the middle, where $K = 0$.

Moving on to the analysis, we first observe that for each such state,

$$\int_K^\infty \cdots \int_K^\infty \prod_{i=1}^n \left[d\lambda_i \delta \left(\sum_\ell J_{i\ell} s_\ell - \lambda_i s_i \right) \right] = 1 \quad (328)$$

thus

$$N(n) = \int_K^\infty \cdots \int_K^\infty \prod_{i=1}^n d\lambda_i \sum_{\mathbf{s}} \left\langle \prod_{i=1}^n \delta \left(\sum_\ell J_{i\ell} s_\ell - \lambda_i s_i \right) \right\rangle_{\mathbf{J}}. \quad (329)$$

Now, according to the S–K model, $\{J_{i\ell}\}$ are $n(n-1)/2$ i.i.d. zero–mean Gaussian RV’s with variance J^2/n . Thus,

$$\bar{N}(n) = \left(\frac{n}{2\pi J^2} \right)^{n(n-1)/4} \int_{\mathbb{R}^{n(n-1)/2}} d\mathbf{J} \exp \left\{ -\frac{n}{2J^2} \sum_{i>\ell} J_{i\ell}^2 \right\} \cdot \sum_{\mathbf{s}} \int_K^\infty \cdots \int_K^\infty d\boldsymbol{\lambda} \cdot \prod_{i=1}^n \delta \left(\sum_\ell J_{i\ell} s_\ell - \lambda_i s_i \right). \quad (330)$$

The next step is to represent each Dirac as an inverse Fourier transform of an exponent

$$\delta(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} d\omega e^{j\omega x} \quad j = \sqrt{-1} \quad (331)$$

which then becomes:

$$\begin{aligned} \bar{N}(n) &= \left(\frac{n}{2\pi J^2} \right)^{n(n-1)/4} \int_{\mathbb{R}^{n(n-1)/2}} d\mathbf{J} \exp \left\{ -\frac{n}{2J^2} \sum_{i>\ell} J_{i\ell}^2 \right\} \cdot \sum_{\mathbf{s}} \int_K^\infty \cdots \int_K^\infty d\boldsymbol{\lambda} \times \\ &\quad \int_{\mathbb{R}^n} \frac{d\boldsymbol{\omega}}{(2\pi)^n} \prod_{i=1}^n \exp \left\{ j\omega_i \left(\sum_\ell J_{i\ell} s_\ell - \lambda_i s_i \right) \right\} \\ &= \left(\frac{n}{2\pi J^2} \right)^{n(n-1)/4} \int_{\mathbb{R}^{n(n-1)/2}} d\mathbf{J} \sum_{\mathbf{s}} \int_K^\infty \cdots \int_K^\infty d\boldsymbol{\lambda} \times \\ &\quad \int_{\mathbb{R}^n} \frac{d\boldsymbol{\omega}}{(2\pi)^n} \exp \left\{ -\frac{n}{2J^2} \sum_{i>\ell} J_{i\ell}^2 + j \sum_{i>\ell} J_{i\ell} (\omega_i s_\ell + \omega_\ell s_i) - j \sum_i \omega_i s_i \lambda_i \right\} \end{aligned} \quad (332)$$

We now use the Hubbard–Stratonovich transform:

$$\int_{\mathbb{R}} dx e^{ax^2+bx} \equiv \sqrt{\frac{\pi}{a}} e^{b^2/(4a)} \quad (333)$$

with $a = n/(2J^2)$ and $b = \omega_i s_\ell + \omega_\ell s_i$:

$$\bar{N}(n) = \sum_{\mathbf{s}} \int_K^\infty \cdots \int_K^\infty d\boldsymbol{\lambda} \int_{\mathbb{R}^n} \frac{d\boldsymbol{\omega}}{(2\pi)^n} \prod_{i=1}^n e^{-j\omega_i s_i \lambda_i} \prod_{i>\ell} \exp \{ -(\omega_i s_\ell + \omega_\ell s_i)^2 J^2 / (2n) \}. \quad (334)$$

Next observe that the summand doesn't actually depend on \mathbf{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 (exercise: convince yourself). Thus, all 2^n contributions are the same as that of $\mathbf{s} = (+1, \dots, +1)$:

$$\bar{N}(n) = 2^n \int_K^\infty \cdots \int_K^\infty d\boldsymbol{\lambda} \int_{\mathbb{R}^n} \frac{d\boldsymbol{\omega}}{(2\pi)^n} \prod_{i=1}^n e^{-j\omega_i \lambda_i} \prod_{i>\ell} \exp\{-(\omega_i + \omega_\ell)^2 J^2 / (2n)\}. \quad (335)$$

Now, consider the following identity (exercise: prove it):

$$\frac{J^2}{2n} \sum_{i>\ell} (\omega_i + \omega_\ell)^2 = J^2 \frac{(n-1)}{2n} \sum_i \omega_i^2 + \frac{J^2}{n} \sum_{i>\ell} \omega_i \omega_\ell, \quad (336)$$

and so for large n ,

$$\frac{J^2}{2n} \sum_{i>\ell} (\omega_i + \omega_\ell)^2 \approx \frac{J^2}{2} \sum_i \omega_i^2 + \frac{J^2}{n} \sum_{i>\ell} \omega_i \omega_\ell \approx \frac{J^2}{2} \sum_i \omega_i^2 + \frac{J^2}{2n} \left(\sum_{i=1}^n \omega_i \right)^2. \quad (337)$$

thus

$$\bar{N}(n) \approx 2^n \int_K^\infty \cdots \int_K^\infty d\boldsymbol{\lambda} \int_{\mathbb{R}^n} \frac{d\boldsymbol{\omega}}{(2\pi)^n} \prod_{i=1}^n \exp \left\{ -j\omega_i \lambda_i - \frac{J^2}{2} \sum_{i=1}^n \omega_i^2 - \frac{J^2}{2n} \left(\sum_{i=1}^n \omega_i \right)^2 \right\}. \quad (338)$$

We now use again the Hubbard–Stratonovich transform

$$e^{a^2} \equiv \int_{\mathbb{R}} \frac{dt}{2\pi} e^{j\sqrt{2}at - t^2/2} \quad (339)$$

and then, after changing variables $\lambda_i \rightarrow J\lambda_i$ and $J\omega_i \rightarrow \omega_i$ (exercise: show that), we get:

$$\bar{N}(n) \approx \frac{1}{\pi^n} \cdot \frac{1}{\sqrt{2\pi}} \int_{K/J}^\infty \cdots \int_{K/J}^\infty d\boldsymbol{\lambda} \int_{\mathbb{R}} dt e^{-t^2/2} \prod_{i=1}^n \left[\int_{\mathbb{R}} d\omega_i \exp \left\{ j\omega_i \left(-\lambda_i + \frac{t}{\sqrt{n}} \right) - \frac{1}{2} \sum_{i=1}^n \omega_i^2 \right\} \right] \quad (340)$$

which after changing $t/\sqrt{n} \rightarrow t$, becomes

$$\begin{aligned}
\bar{N}(n) &\approx \frac{1}{\pi^n} \cdot \frac{n}{\sqrt{2\pi}} \int_{\mathbb{R}} dt e^{-nt^2/2} \left[\int_{K/\lambda}^{\infty} d\lambda \int_{\mathbb{R}} d\omega e^{j\omega(t-\lambda) - \omega^2/2} \right]^n \\
&= \frac{1}{\pi^n} \cdot \frac{n}{\sqrt{2\pi}} \int_{\mathbb{R}} dt e^{-nt^2/2} \left[\sqrt{2\pi} \int_{K/\lambda}^{\infty} d\lambda e^{-(t-\lambda)^2/2} \right]^n \quad (\text{again, the H-S identity}) \\
&= \frac{1}{\pi^n} \cdot \frac{n}{\sqrt{2\pi}} \int_{\mathbb{R}} dt e^{-n(t+K/J)^2/2} \left[\sqrt{2\pi} \int_{-\infty}^t d\lambda e^{-\lambda^2/2} \right]^n \quad t \rightarrow t + K/J, \lambda \rightarrow -\lambda + t + K/J \\
&= \frac{1}{\pi^n} \cdot \frac{n}{\sqrt{2\pi}} \int_{\mathbb{R}} dt e^{-n(t+K/J)^2/2} \cdot [2\pi(1 - Q(t))]^n \\
&= \frac{n}{\sqrt{2\pi}} \int_{\mathbb{R}} dt \exp \left\{ -\frac{n}{2}(t + K/J)^2 + \ln[2(1 - Q(t))] \right\} \\
&\doteq \exp \left\{ n \cdot \max_t \left[\ln(2(1 - Q(t))) - \frac{(t + K/J)^2}{2} \right] \right\} \quad \text{Laplace integration}
\end{aligned}$$

The maximizing t zeroes out the derivative, i.e., it solves the equation

$$\frac{e^{-t^2/2}}{\sqrt{2\pi}[1 - Q(t)]} = t + \frac{K}{J} \quad (341)$$

which for $K = 0$, gives exactly the asserted result about the capacity.

6.5 Generalized Temperature, de Bruijn's Identity, and Fisher Information

Earlier, we defined temperature by

$$\frac{1}{T} = \left(\frac{\partial S}{\partial E} \right)_V. \quad (342)$$

This definition corresponds to equilibrium. We now describe a generalized definition that is valid also for non-equilibrium situations, and see how it relates to concepts in information theory and estimation theory, like the Fisher information. The derivations here follow the paper: 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.

As we know, when the Hamiltonian is quadratic $\mathcal{E}(x) = \frac{\alpha}{2}x^2$, the Boltzmann distribution is Gaussian:

$$P(\mathbf{x}) = \frac{1}{Z} \exp \left\{ -\beta \cdot \frac{\alpha}{2} \sum_{i=1}^n x_i^2 \right\} \quad (343)$$

and by the equipartition theorem:

$$\bar{E}(P) \triangleq \left\langle \frac{\alpha}{2} \sum_{i=1}^n X_i^2 \right\rangle_P = n \frac{kT}{2}. \quad (344)$$

We also computed the entropy, which is nothing but the entropy of a Gaussian vector $S(P) = \frac{nk}{2} \ln\left(\frac{2\pi e}{\alpha\beta}\right)$. Consider now another probability density function $Q(\mathbf{x})$, which means a non-equilibrium probability law if it differs from P , and let's look also at the energy and the entropy pertaining to Q :

$$\bar{E}(Q) = \left\langle \frac{\alpha}{2} \sum_{i=1}^n X_i^2 \right\rangle_Q = \int d\mathbf{x} Q(\mathbf{x}) \cdot \left[\frac{\alpha}{2} \sum_{i=1}^n x_i^2 \right] \quad (345)$$

$$S(Q) = k \cdot \langle -\ln Q(\mathbf{X}) \rangle_Q = -k \int d\mathbf{x} Q(\mathbf{x}) \ln Q(\mathbf{x}). \quad (346)$$

In order to define a notion of generalized temperature, we have to define some sort of derivative of $S(Q)$ w.r.t. $\bar{E}(Q)$. This definition could 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 helps us here.

Consider now the perturbation of \mathbf{X} by $\sqrt{\delta}\mathbf{Z}$ thus defining the perturbed version of \mathbf{X} as $\mathbf{X}_\delta = \mathbf{X} + \sqrt{\delta}\mathbf{Z}$, where $\delta > 0$ is small and \mathbf{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 \mathbf{X}_δ (which is, of course, the convolution between Q and the density of Z , scaled by $\sqrt{\delta}$). The proposed generalized definition of temperature is:

$$\frac{1}{T} \triangleq \lim_{\delta \rightarrow 0} \frac{S(Q_\delta) - S(Q)}{\bar{E}(Q_\delta) - \bar{E}(Q)}. \quad (347)$$

The denominator is easy since

$$\mathbf{E}\|\mathbf{X} + \sqrt{\delta}\mathbf{Z}\|^2 - \mathbf{E}\|\mathbf{X}\|^2 = 2\sqrt{\delta}\mathbf{E}\mathbf{X}^T\mathbf{Z} + n\delta = n\delta \quad (348)$$

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

$$\frac{1}{T} \triangleq \frac{2k}{n\alpha} \cdot \lim_{\delta \rightarrow 0} \frac{h(\mathbf{X} + \sqrt{\delta}\mathbf{Z}) - h(\mathbf{X})}{\delta} = \frac{2k}{n\alpha} \cdot \left. \frac{\partial h(\mathbf{X} + \sqrt{\delta}\mathbf{Z})}{\partial \delta} \right|_{\delta=0}. \quad (349)$$

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

$$S(Q_\delta) = kh(\mathbf{X} + \sqrt{\delta}\mathbf{Z}) \geq kh(\mathbf{X} + \sqrt{\delta}\mathbf{Z}|\mathbf{Z}) = kh(\mathbf{X}) = S(Q). \quad (350)$$

In order to move forward from this point, we will need a piece of background. A well-known notion from estimation theory is the *Fisher information*, which is the basis for the Cramér–Rao bound for unbiased parameter estimators: Suppose we have a family of pdf’s $\{Q_\theta(x)\}$ where θ is a continuous valued parameter. The Fisher info is defined as

$$J(\theta) = \mathbf{E}_\theta \left\{ \left[\frac{\partial \ln Q_\theta(X)}{\partial \theta} \right]^2 \right\} = \int_{-\infty}^{+\infty} \frac{dx}{Q_\theta(x)} \left[\frac{\partial}{\partial \theta} Q_\theta(x) \right]^2. \quad (351)$$

Consider now the special case where θ is a translation parameter, i.e., $Q_\theta(x) = Q(x - \theta)$,

then

$$\begin{aligned}
J(\theta) &= \int_{-\infty}^{+\infty} \frac{dx}{Q(x-\theta)} \left[\frac{\partial}{\partial \theta} Q(x-\theta) \right]^2 \\
&= \int_{-\infty}^{+\infty} \frac{dx}{Q(x-\theta)} \left[\frac{\partial}{\partial x} Q(x-\theta) \right]^2 \quad \frac{\partial Q(x-\theta)}{\partial x} = -\frac{\partial Q(x-\theta)}{\partial \theta} \\
&= \int_{-\infty}^{+\infty} \frac{dx}{Q(x)} \left[\frac{\partial}{\partial x} Q(x) \right]^2 \\
&\triangleq J(Q) \quad \text{with a slight abuse of notation.}
\end{aligned}$$

independently of θ . For the vector case, we define the Fisher info matrix, whose elements are

$$J_{ij}(Q) = \int_{\mathbb{R}^n} \frac{d\mathbf{x}}{Q(\mathbf{x})} \left[\frac{\partial Q(\mathbf{x})}{\partial x_i} \cdot \frac{\partial Q(\mathbf{x})}{\partial x_j} \right] \quad i, j = 1, \dots, n. \quad (352)$$

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

To this end, we will need the following result, which is a variant of the well-known *de Bruijn identity*, first for the scalar case: Let Q be the pdf of a scalar RV X of finite variance. Let Z be a unit variance RV which is symmetric around zero, and let $X_\delta = X + \sqrt{\delta}Z$. Then,

$$\left. \frac{\partial h(X + \sqrt{\delta}Z)}{\partial \delta} \right|_{\delta=0} = \frac{J(Q)}{2}. \quad (353)$$

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 $\delta \rightarrow 0$. The proof of this result is essentially similar to the proof of the original result, which can be found, for example, in the book by Cover and Thomas: Consider the characteristic functions:

$$\Phi_X(s) = \int_{-\infty}^{+\infty} dx e^{sx} Q(x) \quad (354)$$

and

$$\Phi_Z(s) = \int_{-\infty}^{+\infty} dz e^{sz} M(z). \quad (355)$$

Due to the independence

$$\begin{aligned}
\Phi_{X_\delta}(s) &= \Phi_X(s) \cdot \Phi_{\sqrt{\delta}Z}(s) \\
&= \Phi_X(s) \cdot \Phi_Z(\sqrt{\delta}s) \\
&= \Phi_X(s) \cdot \int_{-\infty}^{+\infty} dz e^{\sqrt{\delta}sz} M(z) \\
&= \Phi_X(s) \cdot \sum_{i=0}^{\infty} \frac{(\sqrt{\delta}s)^i}{i!} \mu_i(M) \quad \mu_i(M) \text{ being the } i\text{-th moment of } Z \\
&= \Phi_X(s) \cdot \left(1 + \frac{\delta s^2}{2} + \dots\right) \quad \text{odd moments vanish due to symmetry}
\end{aligned}$$

Applying the inverse Fourier transform, we get:

$$Q_\delta(x) = Q(x) + \frac{\delta}{2} \cdot \frac{\partial^2 Q(x)}{\partial x^2} + o(\delta), \quad (356)$$

and so,

$$\left. \frac{\partial Q_\delta(x)}{\partial \delta} \right|_{\delta=0} = \frac{1}{2} \cdot \frac{\partial^2 Q(x)}{\partial x^2} \sim \frac{1}{2} \cdot \frac{\partial^2 Q_\delta(x)}{\partial x^2}. \quad (357)$$

Now, let's look at the entropy:

$$h(X_\delta) = - \int_{-\infty}^{+\infty} dx Q_\delta(x) \ln Q_\delta(x). \quad (358)$$

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

$$\begin{aligned}
\frac{\partial h(X_\delta)}{\partial \delta} &= - \int_{-\infty}^{+\infty} dx \left[\frac{\partial Q_\delta(x)}{\partial \delta} + \frac{\partial Q_\delta(x)}{\partial \delta} \cdot \ln Q_\delta(x) \right] \\
&= - \frac{\partial}{\partial \delta} \int_{-\infty}^{+\infty} dx Q_\delta(x) - \int_{-\infty}^{+\infty} dx \frac{\partial Q_\delta(x)}{\partial \delta} \cdot \ln Q_\delta(x) \\
&= - \frac{\partial}{\partial \delta} 1 - \int_{-\infty}^{+\infty} dx \frac{\partial Q_\delta(x)}{\partial \delta} \cdot \ln Q_\delta(x) \\
&= - \int_{-\infty}^{+\infty} dx \frac{\partial Q_\delta(x)}{\partial \delta} \cdot \ln Q_\delta(x)
\end{aligned} \quad (359)$$

and so,

$$\left. \frac{\partial h(X_\delta)}{\partial \delta} \right|_{\delta=0} = - \int_{-\infty}^{+\infty} dx \cdot \left. \frac{\partial Q_\delta(x)}{\partial \delta} \right|_{\delta=0} \cdot \ln Q(x) = - \int_{-\infty}^{+\infty} dx \cdot \frac{1}{2} \frac{d^2 Q(x)}{d^2 x} \cdot \ln Q(x). \quad (360)$$

Integrating by parts, we obtain:

$$\left. \frac{\partial h(X_\delta)}{\partial \delta} \right|_{\delta=0} = \left[-\frac{1}{2} \cdot \frac{dQ(x)}{dx} \cdot \ln Q(x) \right]_{-\infty}^{+\infty} + \frac{1}{2} \int_{-\infty}^{+\infty} \frac{dx}{Q(x)} \left[\frac{\partial Q(x)}{\partial x} \right]^2. \quad (361)$$

The first term can be shown to vanish (see paper and/or C&T) and the second term is exactly $J(Q)/2$. This completes the proof of the (modified) de Bruijn identity.

Exercise: Extend this to the vector case, showing that for a vector \mathbf{Z} with i.i.d. components, all symmetric around the origin:

$$\frac{\partial h(\mathbf{X} + \sqrt{\delta}\mathbf{Z})}{\partial \delta} = \frac{1}{2} \sum_{i=1}^n \int_{\mathbb{R}^n} \frac{d\mathbf{x}}{Q(\mathbf{x})} \left[\frac{\partial Q(\mathbf{x})}{\partial x_i} \right]^2 = \frac{1}{2} \sum_{i=1}^n J_{ii}(Q) = \frac{1}{2} \text{tr}\{J(Q)\}. \quad \square \quad (362)$$

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

$$\frac{1}{T} = \frac{k}{n\alpha} \cdot \text{tr}\{J(Q)\}. \quad (363)$$

In the ‘stationary’ case, where Q is symmetric w.r.t. all components of \mathbf{x} , $\{J_{ii}\}$ are all the same quantity, call it $J(Q)$, and then

$$\frac{1}{T} = \frac{k}{\alpha} \cdot J(Q) \quad (364)$$

or, equivalently,

$$T = \frac{\alpha}{kJ(Q)} = \frac{\alpha}{k} \cdot \text{CRB} \quad (365)$$

where CRB is the Cramér–Rao bound. High temperature means a lot of noise and this in turn means that it is hard to estimate the mean of X . In the Boltzmann case, $J(Q) = 1/\text{Var}\{X\} = \alpha\beta = \alpha/(kT)$ and we are back to the ordinary definition of temperature.

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

$$\langle \mathcal{E}(X) \rangle = \left\langle \frac{\alpha}{2} X^2 \right\rangle = \frac{kT}{2} \quad (366)$$

or

$$\frac{\alpha}{2} \sigma^2 \triangleq \frac{\alpha}{2} \langle X^2 \rangle = \frac{kT}{2}. \quad (367)$$

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:

$$\frac{\alpha}{2} \cdot \text{CRB} = \frac{kT}{2}. \quad (368)$$

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,

$$\mathcal{E}(x) = \frac{m\omega_0^2}{2}(x - \theta)^2. \quad (369)$$

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

$$\text{estimation error energy} = \frac{m\omega_0^2}{2} \cdot \mathbf{E}(\hat{\theta}(X) - \theta)^2 \geq \frac{kT}{2} \quad (370)$$

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.

6.6 The Gibbs Inequality and the Log–Sum Inequality

In one of our earlier meetings, we have seen the Gibbs’ inequality, its physical significance, and related it to the second law and the DPT. 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 Kardar’s book, p. 145). As we show in this part, this inequality is nothing else than the *log–sum inequality*, which is used in Information Theory, mostly for proving certain *qualitative* properties of information measures, like the data processing theorem of the divergence, etc. 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 perhaps yields useful bounds on certain information measures. We try to demonstrate this point here.

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

$$Z(\beta) = \sum_{\mathbf{x}} e^{-\beta\mathcal{E}(\mathbf{x})} \quad (371)$$

but it is hard, if not impossible, to calculate in closed–form. Suppose further that for another, somewhat different Hamiltonian, $\mathcal{E}_0(\mathbf{x})$, 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 .

$$\ln \left[\sum_{\mathbf{x}} e^{-\beta\mathcal{E}(\mathbf{x})} \right] \geq \ln \left[\sum_{\mathbf{x}} e^{-\beta\mathcal{E}_0(\mathbf{x})} \right] + \beta \langle \mathcal{E}_0(\mathbf{X}) - \mathcal{E}(\mathbf{X}) \rangle_0, \quad (372)$$

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. We next demonstrate this with an example:

Example – Non–harmonic oscillator. Consider the potential function

$$V(z) = Az^4 \quad (373)$$

and so

$$\mathcal{E}(x) = \frac{p^2}{2m} + Az^4, \quad (374)$$

where we approximate the second term by

$$V_0(z) = \begin{cases} 0 & |z| \leq \frac{L}{2} \\ +\infty & |z| > \frac{L}{2} \end{cases} \quad (375)$$

where L is a parameter to be optimized. Thus,

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

and so, by the Gibbs inequality:

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

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

$$0 = \frac{df}{dL} \equiv \frac{1}{L} - \frac{AL^3}{20kT} \implies L^* = \left(\frac{20kT}{A} \right)^{1/4}. \quad (376)$$

Plugging this back into the Gibbs lower bound and comparing to the *exact* value of Z (which is still computable in this example), we find that $Z_{\text{approx}} \approx 0.91Z_{\text{exact}}$, which is not that bad 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:

$$V_0(z) = \frac{m\omega_0^2}{2} \cdot z^2 \quad (377)$$

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

$$Z_0 = \frac{1}{h} \int_{-\infty}^{+\infty} dp \int_{-\infty}^{+\infty} dz e^{-\beta[m\omega_0^2 z^2/2 + p^2/(2m)]} = \frac{kT}{\hbar\omega_0} \quad \hbar = \frac{h}{2\pi} \quad (378)$$

and this time, we get:

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

Maximizing f :

$$0 = \frac{df}{d\omega_0} \equiv -\frac{1}{\omega_0} + \frac{12AkT}{m^2\omega_0^5} \implies \omega_0^* = \frac{(12AkT)^{1/4}}{\sqrt{m}}. \quad (379)$$

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

So much for physics. Let's look now at the Gibbs inequality slightly differently. What we actually did, in a nutshell, and in different notation, is the following: Consider the function:

$$Z(\lambda) = \sum_{i=1}^n a_i^{1-\lambda} b_i^\lambda = \sum_{i=1}^n a_i e^{-\lambda \ln(a_i/b_i)}, \quad (380)$$

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

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

which is nothing but the log-sum inequality, which in IT, is more customarily written as:

$$\sum_{i=1}^n a_i \ln \frac{a_i}{b_i} \geq \left(\sum_{i=1}^n a_i \right) \cdot \ln \frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i}. \quad (381)$$

Returning to the form:

$$\ln \left(\sum_{i=1}^n b_i \right) \geq \ln \left(\sum_{i=1}^n a_i \right) + \frac{\sum_{i=1}^n a_i \ln(b_i/a_i)}{\sum_{i=1}^n a_i}, \quad (382)$$

the idea now is, once again, 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 $W(\mathbf{y}|\mathbf{x}) = \prod_{t=1}^n W(y_t|x_t)$ fed by a first-order Markov process \mathbf{X} , governed by $Q(\mathbf{x}) = \prod_{t=1}^n Q(x_t|x_{t-1})$. 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. Here, we propose an upper bound that stems from the Gibbs inequality, or the log-sum inequality.

The probability distribution of \mathbf{y} is

$$P(\mathbf{y}) = \sum_{\mathbf{x}} \prod_{t=1}^n [W(y_t|x_t)Q(x_t|x_{t-1})]. \quad (383)$$

This summation does not lend itself to a nice closed-form expression, but if the t -th factor depended only on t (and not also on $t-1$) life would have been easy and simple as the sum of products would have boiled down to a product of sums. So this motivates the following use of the log-sum inequality: For a given \mathbf{y} , let's think of \mathbf{x} as the index i of the log-sum inequality and then

$$b(\mathbf{x}) = \prod_{t=1}^n [W(y_t|x_t)Q(x_t|x_{t-1})]. \quad (384)$$

Let us now define

$$a(\mathbf{x}) = \prod_{t=1}^n P_0(x_t, y_t), \quad (385)$$

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:

$$\begin{aligned}
\ln P(\mathbf{y}) &= \ln \left(\sum_{\mathbf{x}} b(\mathbf{x}) \right) \\
&\geq \ln \left(\sum_{\mathbf{x}} a(\mathbf{x}) \right) + \frac{\sum_{\mathbf{x}} a(\mathbf{x}) \ln[b(\mathbf{x})/a(\mathbf{x})]}{\sum_{\mathbf{x}} a(\mathbf{x})} \\
&= \ln \left(\sum_{\mathbf{x}} \prod_{t=1}^n P_0(x_t, y_t) \right) + \\
&\quad + \frac{\sum_{\mathbf{x}} [\prod_{t=1}^n P_0(x_t, y_t)] \cdot \ln[\prod_{t=1}^n [Q(x_t|x_{t-1})W(y_t|x_t)/P_0(x_t, y_t)]]}{\sum_{\mathbf{x}} \prod_{t=1}^n P_0(x_t, y_t)}. \quad (386)
\end{aligned}$$

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:

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

where \mathbf{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:

$$\ln P(\mathbf{y}) \geq \sum_{t=1}^n \mathbf{E}_0 \left\{ \ln \left[\frac{Q(X_t|X_{t-1})W(y_t|X_t)}{P_0(X_t|y_t)} \right] \middle| Y_{t-1} = y_{t-1}, Y_t = y_t \right\} \triangleq \sum_{t=1}^n \Delta(y_{t-1}, y_t; P_0). \quad (387)$$

At this stage, we can perform the optimization over P_0 for each \mathbf{y} individually, and then derive the bound on the expectation of $\ln P(\mathbf{y})$ to get a bound on the entropy. Note, however, that

$\sum_t \Delta(y_{t-1}, y_t; P_0)$ depends on \mathbf{y} only via its Markov statistics, i.e., the relative frequencies of transitions $y \implies y'$ for all $y, y' \in \mathcal{Y}$. Thus, the optimum P_0 depends on \mathbf{y} also via these statistics. Now, the expectation of $\sum_t \Delta(y_{t-1}, y_t; P_0)$ is going to be dominated by the typical $\{\mathbf{y}\}$ 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 give, assuming stationarity:

$$H(Y^n) \leq -n \cdot \max_{P_0} \mathbf{E}\{\Delta(Y_0, Y_1; P_0)\}. \quad (388)$$

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

$$P(y_0, y_1) = \sum_{x_0, x_1} \pi(x_0) Q(x_1|x_0) P(y_0|x_0) P(y_1|x_1), \quad (389)$$

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

6.7 Dynamics, Evolution of Info Measures, and Simulation

The material here is taken mainly from the books by Reif, Kittel, and F. P. Kelly, *Reversibility and Stochastic Networks*, (Chaps 1–3), J. Wiley & Sons, 1979.

6.7.1 Markovian Dynamics, Global Balance and Detailed Balance

So far we discussed only physical systems in equilibrium. For these systems, the Boltzmann–Gibbs distribution is nothing but the stationary distribution of the microstate x at every given time instant t . However, this is merely one part of the picture. What is missing is the temporal probabilistic behavior, or in other words, the laws that underly the evolution of the 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. In this section, we discuss a few of the properties of these processes as well as the evolution of information measures, like entropy, divergence (and more) associated with them.

We begin with an isolated system in continuous time, which is not necessarily assumed to have reached (yet) equilibrium. Let us suppose that X_t , the microstate at time t , can take on values in a discrete set \mathcal{X} . For $r, s \in \mathcal{X}$, let

$$W_{rs} = \lim_{\delta \rightarrow 0} \frac{\Pr\{X_{t+\delta} = s | X_t = r\}}{\delta} \quad r \neq s \quad (390)$$

in other words, $\Pr\{X_{t+\delta} = s | X_t = r\} = W_{rs} \cdot \delta + o(\delta)$. Letting $P_r(t) = \Pr\{X_t = r\}$, it is easy to see that

$$P_r(t + dt) = \sum_{s \neq r} P_s(t) W_{sr} dt + P_r(t) \left(1 - \sum_{s \neq r} W_{rs} dt \right), \quad (391)$$

where the first sum describes the probabilities of all possible transitions from other states to state r and the second term describes the probability of not leaving state r . Subtracting $P_r(t)$

from both sides and dividing by dt , we immediately obtain the following set of differential equations:

$$\frac{dP_r(t)}{dt} = \sum_s [P_s(t)W_{sr} - P_r(t)W_{rs}], \quad r \in \mathcal{X}, \quad (392)$$

where W_{rr} is defined in an arbitrary manner, e.g., $W_{rr} = 0$ for all r . These equations are called the *master equations*.²⁹ When the process reaches stationarity, i.e., for all $r \in \mathcal{X}$, $P_r(t)$ converge to some P_r that is time-invariant, then

$$\sum_s [P_s W_{sr} - P_r W_{rs}] = 0, \quad \forall r \in \mathcal{X}. \quad (393)$$

This is called *global balance* or *steady state*. When the system is isolated (microcanonical ensemble), the steady-state distribution must be uniform, i.e., $P_r = 1/|\mathcal{X}|$ for all $r \in \mathcal{X}$. From quantum mechanical considerations, as well as considerations pertaining to time reversibility in the microscopic level,³⁰ it is customary to assume $W_{rs} = W_{sr}$ for all pairs $\{r, s\}$. We then observe that, not only, $\sum_s [P_s W_{sr} - P_r W_{rs}] = 0$, but moreover, each individual term in the sum vanishes, as

$$P_s W_{sr} - P_r W_{rs} = \frac{1}{|\mathcal{X}|} (W_{sr} - W_{rs}) = 0. \quad (394)$$

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 a situation of time-invariant state probabilities $\{P_r\}$, a steady-state still allows cyclic flows of probability. For example, a Markov process with cyclic deterministic transitions $1 \rightarrow 2 \rightarrow 3 \rightarrow 1 \rightarrow 2 \rightarrow 3 \rightarrow \dots$ 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 ($W_{rs} = W_{sr}$ 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,

²⁹Note that the master equations apply in discrete time too, provided that the derivative at the l.h.s. is replaced by a simple difference, $P_r(t+1) - P_r(t)$, and $\{W_{rs}\}$ designate one-step state transition probabilities.

³⁰Think, for example, of an isolated system of moving particles, obeying the differential equations $m d^2 \mathbf{r}_i(t) / dt^2 = \sum_{j \neq i} F(\mathbf{r}_j(t) - \mathbf{r}_i(t))$, $i = 1, 2, \dots, n$, which remain valid if the time variable t is replaced by $-t$ since $d^2 \mathbf{r}_i(t) / dt^2 = d^2 \mathbf{r}_i(-t) / d(-t)^2$.

$\{X_{-t}\}$ has the same probability law as $\{X_t\}$. For example, if $\{Y_t\}$ is a Bernoulli process, taking values equiprobably in $\{-1, +1\}$, then X_t defined recursively by

$$X_{t+1} = (X_t + Y_t) \bmod K, \quad (395)$$

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

6.7.2 Evolution of Information Measures

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

$$H(X_t) = - \sum_r P_r(t) \log P_r(t). \quad (396)$$

We argue that $H(X_t)$ is monotonically non–decreasing, which is in agreement with the second law (a.k.a. the H–Theorem). To this end, we next show that

$$\frac{dH(X_t)}{dt} \geq 0, \quad (397)$$

where for convenience, we denote $dP_r(t)/dt$ by $\dot{P}_r(t)$.

$$\begin{aligned} \frac{dH(X_t)}{dt} &= - \sum_r [\dot{P}_r(t) \log P_r(t) + \dot{P}_r(t)] \\ &= - \sum_r \dot{P}_r(t) \log P_r(t) \quad \sum_r \dot{P}_r(t) = 0 \\ &= - \sum_r \sum_s W_{sr} [P_s(t) - P_r(t)] \log P_r(t) \quad W_{sr} = W_{rs} \\ &= - \frac{1}{2} \sum_{r,s} W_{sr} [P_s(t) - P_r(t)] \log P_r(t) - \\ &\quad \frac{1}{2} \sum_{s,r} W_{sr} [P_r(t) - P_s(t)] \log P_s(t) \\ &= \frac{1}{2} \sum_{r,s} W_{sr} [P_s(t) - P_r(t)] \cdot [\log P_s(t) - \log P_r(t)] \\ &\geq 0. \end{aligned} \quad (398)$$

where the last inequality is due to the increasing monotonicity of the logarithmic function: the product $[P_s(t) - P_r(t)] \cdot [\log P_s(t) - \log P_r(t)]$ cannot be negative for any pair (r, s) , as the two factors of this product are either both negative, both zero, or both positive. Thus, $H(X_t)$ cannot decrease with time.

This result has a discrete-time analogue: If a finite-state Markov process has a symmetric transition probability matrix, and so, the stationary state distribution is uniform, then $H(X_t)$ is a monotonically non-decreasing sequence.

A considerably more general result is the following: If $\{X_t\}$ is a Markov process with a given state transition probability matrix $W = \{W_{rs}\}$ (not necessarily symmetric) and $\{P_r\}$ is a stationary state distribution, then the function

$$U(t) = \sum_r P_r \cdot V\left(\frac{P_r(t)}{P_r}\right) \quad (399)$$

is monotonically strictly increasing provided that $V(\cdot)$ is strictly concave. To see why this is true, we use the fact that $P_s = \sum_r P_r W_{rs}$ and define $\tilde{W}_{sr} = P_r W_{rs} / P_s$. Obviously, $\sum_r \tilde{W}_{sr} = 1$ for all s , and so,

$$\frac{P_r(t+1)}{P_r} = \sum_s \frac{P_s(t) W_{sr}}{P_r} = \sum_s \frac{\tilde{W}_{sr} P_s(t)}{P_s} \quad (400)$$

and so, by the concavity of $V(\cdot)$:

$$\begin{aligned} U(t+1) &= \sum_r P_r \cdot V\left(\frac{P_r(t+1)}{P_r}\right) \\ &= \sum_r P_r \cdot V\left(\sum_s \tilde{W}_{sr} \frac{P_s(t)}{P_s}\right) \\ &> \sum_r \sum_s P_r \tilde{W}_{sr} \cdot V\left(\frac{P_s(t)}{P_s}\right) \\ &= \sum_r \sum_s P_s W_{sr} \cdot V\left(\frac{P_s(t)}{P_s}\right) \\ &= \sum_s P_s \cdot V\left(\frac{P_s(t)}{P_s}\right) = U(t). \end{aligned} \quad (401)$$

Here we required nothing except the existence of a stationary distribution. Of course in the above derivation $t+1$ can be replaced by $t+\tau$ for any positive real τ with the appropriate

transition probabilities, so the monotonicity of $U(t)$ applies to continuous-time Markov processes as well.

Now, a few interesting choices of the function V may be considered:

- For $V(x) = -x \ln x$, we have $U(t) = -D(P(t)||P)$. This means that the divergence between $\{P_r(t)\}$ and the steady state distribution $\{P_r\}$ is monotonically strictly decreasing, whose physical interpretation could be the decrease of the free energy, since we have already seen that the free energy is the physical counterpart of the divergence. This is a more general rule, that governs not only isolated systems, but any Markov process with a stationary limiting distribution (e.g., any Markov process whose distribution converges to that of the Boltzmann–Gibbs distribution). Having said that, if we now particularize this result to the case where $\{P_r\}$ is the uniform distribution (as in an isolated system), then

$$D(P(t)||P) = \log |\mathcal{X}| - H(X_t), \tag{402}$$

which means that the decrease of divergence is equivalent to the increase in entropy, as before. The difference, however, is that here it is more general as we only required a uniform steady-state distribution, not necessarily detailed balance.³¹

- Another interesting choice of V is $V(x) = \ln x$, which gives $U(t) = -D(P||P(t))$. Thus, $D(P||P(t))$ is also monotonically decreasing. In fact, both this and the monotonicity result of the previous item, are in turn, special cases of a more general result concerning the divergence (see also the book by Cover and Thomas, Section 4.4). Let $\{P_r(t)\}$ and $\{P'_r(t)\}$ be two time-varying state-distributions pertaining to the same Markov chain, but induced by two different initial state distributions, $\{P_r(0)\}$ and $\{P'_r(0)\}$. Then

³¹For the uniform distribution to be a stationary distribution, it is sufficient (and necessary) that W would be a doubly stochastic matrix, namely, $\sum_r W_{rs} = \sum_r W_{sr} = 1$. This condition is, of course, weaker than detailed balance, which means that W is moreover symmetric.

$D(P(t)\|P'(t))$ is monotonically non-increasing. This happens because

$$\begin{aligned}
D(P(t)\|P'(t)) &= \sum_r P_r(t) \log \frac{P_r(t)}{P'_r(t)} \\
&= \sum_{r,s} P_r(t) P(X_{t+\tau} = s | X_t = r) \log \frac{P_r(t) P(X_{t+\tau} = s | X_t = r)}{P'_r(t) P(X_{t+\tau} = s | X_t = r)} \\
&= \sum_{r,s} P(X_t = r, X_{t+\tau} = s) \log \frac{P(X_t = r, X_{t+\tau} = s)}{P'(X_t = r, X_{t+\tau} = s)} \\
&\geq D(P(t+\tau)\|P'(t+\tau))
\end{aligned} \tag{403}$$

where the last inequality follows from the data processing theorem of the divergence: the divergence between two joint distributions of $(X_t, X_{t+\tau})$ is never smaller than the divergence between corresponding marginal distributions of $X_{t+\tau}$.

- Yet another choice is $V(x) = x^s$, where $s \in [0, 1]$ is a parameter. This would yield the increasing monotonicity of $\sum_r P_r^{1-s} P_r^s(t)$, 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. In particular, the choice $s = 1/2$ yields balance between the two kinds of error and it is intimately related to the Bhattacharyya distance. Thus, we obtained some sorts of generalizations of the second law to information measures other than entropy.

For a general Markov process, whose steady state-distribution is not necessarily uniform, the condition of detailed balance, which means time-reversibility, reads

$$P_s W_{sr} = P_r W_{rs}, \tag{404}$$

both in discrete time and continuous time (with the corresponding meaning of $\{W_{rs}\}$). The physical interpretation is that now our system is (a small) part of a large 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 an 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. 25),

there cannot be any cyclic probability flux. The steady-state distribution is well-known to be geometric

$$P_r = \left(1 - \frac{\lambda}{\mu}\right) \cdot \left(\frac{\lambda}{\mu}\right)^r, \quad r = 0, 1, 2, \dots, \quad (405)$$

which indeed satisfies the detailed balance $P_r \lambda = P_{r+1} \mu$ for all r . Thus, the Markov process $\{X_t\}$, designating the number of customers in the queue at time t , is time-reversible.

It is interesting to point out that in order to check for the detailed balance property, one does not necessarily have to know the equilibrium distribution $\{P_r\}$ as above. Applying detailed balance to any k pairs of states in a cycle, $(s_1, s_2), (s_2, s_3), \dots, (s_k, s_1)$, and multiplying the respective detailed balance equations, the steady state probabilities cancel out and one easily obtains

$$W_{s_1 s_2} W_{s_2 s_3} \cdots W_{s_{k-1} s_k} W_{s_k s_1} = W_{s_k s_{k-1}} W_{s_{k-1} s_{k-2}} \cdots W_{s_2 s_1} W_{s_1 s_k}, \quad (406)$$

so this is clearly a necessary condition for detailed balance. One can show conversely, that if this equation applies to any finite cycle of states, then the chain satisfies detailed balance, and so this is also a sufficient condition. This is true both in discrete time and continuous time, with the corresponding meanings of $\{W_{rs}\}$ (see Kelly's book, pp. 22–23).

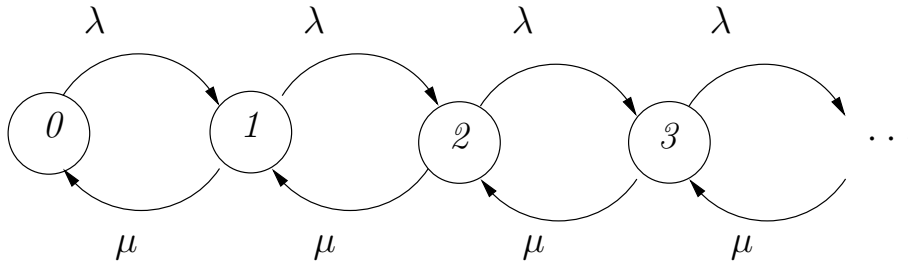


Figure 25: State transition diagram of an M/M/1 queue.

In the case of detailed balance, there is another interpretation of the approach to equilibrium and the growth of $U(t)$. We can write the master equations as follows:

$$\frac{dP_r(t)}{dt} = \sum_s \frac{1}{R_{sr}} \left(\frac{P_s(t)}{P_s} - \frac{P_r(t)}{P_r} \right) \quad (407)$$

where $R_{sr} = (P_s W_{sr})^{-1} = (P_r W_{rs})^{-1}$. Imagine now an electrical circuit where the indices $\{r\}$ designate the nodes. Nodes r and s are connected by a wire with resistance R_{sr} and every

node r is grounded via a capacitor with capacitance P_r (see Fig. 26). If $P_r(t)$ is the charge at node r 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 circuit is analogous to the way charge spreads across the circuit and probability fluxes are now analogous to electrical currents. If we now choose $V(x) = -\frac{1}{2}x^2$, then $-U(t) = \frac{1}{2} \sum_r \frac{P_r^2(t)}{P_r}$, which means that the energy stored in the capacitors dissipates as heat in the wires until the system reaches equilibrium, where all nodes have the same potential, $P_r(t)/P_r = 1$, and hence detailed balance corresponds to the situation where all individual currents vanish (not only their algebraic sum).

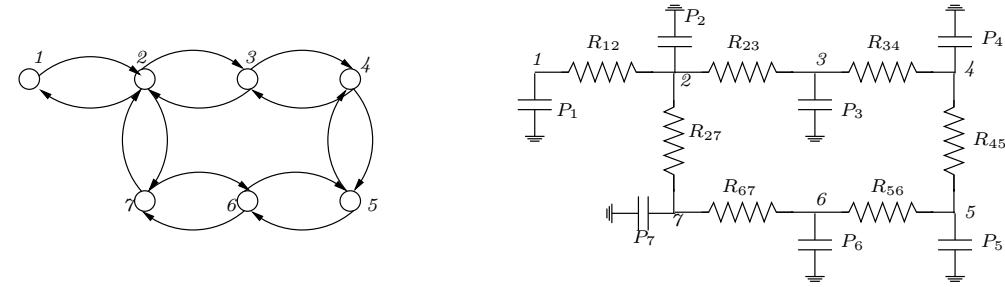


Figure 26: State transition diagram of a Markov chain (left part) and the electric circuit that emulates the dynamics of $\{P_r(t)\}$ (right part).

We have seen, in the above examples, that various choices of the function V yield various ‘metrics’ between $\{P_r(t)\}$ and $\{P_r\}$, which are both marginal distributions of a single symbol. What about joint distributions of two or more symbols? Consider, for example, the function

$$J(t) = \sum_{r,s} P(X_0 = r, X_t = s) \cdot V \left(\frac{P(X_0 = r)P(X_t = s)}{P(X_0 = r, X_t = s)} \right), \quad (408)$$

where V is concave as before. Here, by the same token, $J(t)$ is a ‘metric’ between the joint probability distribution $\{P(X_0 = r, X_t = s)\}$ and the product of marginals $\{P(X_0 = r)P(X_t = s)\}$, namely, it a measure of the amount of statistical dependence between X_0 and X_t . For $V(x) = \ln x$, we have, of course, $J(t) = -I(X_0; X_t)$. Now, using a similar chain of

inequalities as before, we get the non-decreasing monotonicity of $J(t)$ as follows:

$$\begin{aligned}
J(t) &= \sum_{r,s,u} P(X_0 = r, X_t = s, X_{t+\tau} = u) \cdot V \left(\frac{P(X_0 = r)P(X_t = s)}{P(X_0 = r, X_t = s)} \cdot \frac{P(X_{t+\tau} = u|X_t = s)}{P(X_{t+\tau} = u|X_t = s)} \right) \\
&= \sum_{r,u} P(X_0 = r, X_{t+\tau} = u) \sum_s P(X_t = s|X_0 = r, X_{t+\tau} = u) \times \\
&\quad V \left(\frac{P(X_0 = r)P(X_t = s, X_{t+\tau} = u)}{P(X_0 = r, X_t = s, X_{t+\tau} = u)} \right) \\
&\leq \sum_{r,u} P(X_0 = r, X_{t+\tau} = u) \times \\
&\quad V \left(\sum_s P(X_t = s|X_0 = r, X_{t+\tau} = u) \cdot \frac{P(X_0 = r)P(X_t = s, X_{t+\tau} = u)}{P(X_0 = r, X_t = s, X_{t+\tau} = u)} \right) \\
&= \sum_{r,u} P(X_0 = r, X_{t+\tau} = u) \cdot V \left(\sum_s \frac{P(X_0 = r)P(X_t = s, X_{t+\tau} = u)}{P(X_0 = r, X_{t+\tau} = u)} \right) \\
&= \sum_{r,u} P(X_0 = r, X_{t+\tau} = u) \cdot V \left(\frac{P(X_0 = r)P(X_{t+\tau} = u)}{P(X_0 = r, X_{t+\tau} = u)} \right) = J(t + \tau). \tag{409}
\end{aligned}$$

This time, we assumed nothing beyond Markovity (not even homogeneity). This is exactly the generalized data processing theorem of Ziv and Zakai (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), which yields the ordinary data processing theorem (of the mutual information) as a special case. Thus, we see that the second law of thermodynamics is (at least indirectly) related to the data processing theorem via the fact that they both stem from some more general principle concerning monotonic evolution of ‘metrics’ between probability distributions defined using convex functions. In a very similar manner, one can easily show that the generalized conditional entropy

$$\sum_{r,s} P(X_0 = r, X_t = s) \cdot V \left(\frac{1}{P(X_0 = r|X_t = s)} \right) \tag{410}$$

is monotonically non-decreasing with t for any concave V .

6.7.3 Monte Carlo Simulation

Returning to the realm of Markov processes with the detailed balance property, suppose we want to simulate a physical system, namely, to sample from the Boltzmann–Gibbs distribu-

tion

$$P_r = \frac{e^{-\beta E_r}}{Z(\beta)}. \quad (411)$$

In other words, we wish to generate a discrete-time Markov process $\{X_t\}$, possessing the detailed balance property, whose marginal converges to the Boltzmann–Gibbs distribution. This approach is called *dynamic Monte Carlo* or *Markov chain Monte Carlo* (MCMC). How should we select the state transition probability matrix W to this end? Substituting $P_r = e^{-\beta E_r}/Z(\beta)$ into the detailed balance equation, we readily see that a necessary condition is

$$\frac{W_{rs}}{W_{sr}} = e^{-\beta(E_s - E_r)}. \quad (412)$$

The *Metropolis algorithm* is one popular way to implement such a Markov process in a rather efficient manner. It is based on the concept of factoring W_{rs} as a product $W_{rs} = C_{rs}A_{rs}$, where C_{rs} is the conditional probability of selecting $X_{t+1} = s$ as a *candidate* for the next state, and A_{rs} designates the probability of *acceptance*. In other words, we first choose a candidate according to C , and then make a final decision whether we accept this candidate or stay in state r . The Metropolis algorithm picks C to implement a uniform distribution among n states ‘close’ to r (e.g., flipping one spin of a n -spin configuration). Thus, $W_{rs}/W_{sr} = A_{rs}/A_{sr}$, and so, it remains to choose A such that

$$\frac{A_{rs}}{A_{sr}} = e^{-\beta(E_s - E_r)}. \quad (413)$$

The Metropolis algorithm defines

$$A_{rs} = \begin{cases} e^{-\beta(E_s - E_r)} & E_s > E_r \\ 1 & \text{otherwise} \end{cases} \quad (414)$$

In simple words, the algorithm works as follows: Given that $X_t = r$, first randomly select one candidate s for X_{t+1} among n possible (neighboring) states. If $E_s < E_r$ always accept $X_{t+1} = s$ as the next state. If $E_s \geq E_r$, then randomly draw a RV $Y \in \text{Unif}[0, 1]$. If $Y < e^{-\beta(E_s - E_r)}$, then again, accept $X_{t+1} = s$ as the next state. Otherwise, stay in state r , i.e., $X_{t+1} = r$. To see why this choice of A works, observe that

$$\frac{A_{rs}}{A_{sr}} = \begin{cases} e^{-\beta(E_s - E_r)} & E_s > E_r \\ \frac{1}{e^{-\beta(E_r - E_s)}} & E_s \leq E_r \end{cases} = e^{-\beta(E_s - E_r)}. \quad (415)$$

There are a few nice things about this algorithm:

- Energy differences between neighboring states, $E_s - E_r$, are normally easy to calculate. If r and s differ by a single component of the microstate \mathbf{x} , and the if the Hamiltonian structure consists of short-range interactions only, then most terms of the Hamiltonian are the same for r and s , and only a local calculation is required for evaluating the energy difference.
- Calculation of $Z(\beta)$ is not required, and
- Chances are that you don't get stuck in the same state for too long.

The drawback, however, is that aperiodicity is not guaranteed. This depends on the Hamiltonian.

The *heat bath* algorithm (a.k.a. *Glauber dynamics*) alleviates this shortcoming and although somewhat slower than Metropolis to equilibrate, it guarantees all the good properties of a Markov chain: irreducibility, aperiodicity, and convergence to stationarity. The only difference is that instead of the above choice of A_{rs} , it is redefined as

$$\begin{aligned}
 A_{rs} &= \frac{1}{2} \left[1 - \tanh \left(\frac{\beta(E_s - E_r)}{2} \right) \right] \\
 &= \frac{e^{-\beta(E_s - E_r)}}{1 + e^{-\beta(E_s - E_r)}} \\
 &= \frac{P_s}{P_s + P_r}, \tag{416}
 \end{aligned}$$

which is also easily shown to satisfy the detailed balance condition. The heat bath algorithm generalizes easily to sample from any distribution $P(\mathbf{x})$ whose configuration space is of the form \mathcal{X}^n . The algorithm can be described by the following pseudocode:

1. Select \mathbf{X}_0 uniformly at random across \mathcal{X}^n .
2. **For** $t = 1$ **to** $t = T$:
3. Draw an integer i at random with uniform distribution across $\{1, 2, \dots, n\}$.

4. For each $x \in \mathcal{X}$, calculate

$$P(X^i = x | \mathbf{X}^{\sim i} = \mathbf{x}_t^{\sim i}) = \frac{P(X^i = x, \mathbf{X}^{\sim i} = \mathbf{x}_t^{\sim i})}{\sum_{x' \in \mathcal{X}} P(X^i = x', \mathbf{X}^{\sim i} = \mathbf{x}_t^{\sim i})}. \quad (417)$$

5. Set $x_{t+1}^j = x_t^j$ for all $j \neq i$ and $x_t^i = X^i$, where X^i is drawn according to

$$P(X^i = x | \mathbf{X}^{\sim i} = \mathbf{x}_t^{\sim i}).$$

6. **end**

7. **Return** the sequence $\mathbf{X}_t, t = 1, 2, \dots, T$.

It can be easily seen that the resulting Markov chain satisfies detailed balance and that in the case of binary alphabet (spin array) it implements the above expression of A_{rs} . One can also easily generalize the Metropolis algorithm, in the same spirit, as $e^{-\beta(E_s - E_r)}$ is nothing but the ratio P_s/P_r .

References

- [1] G. B. Bağcı, “The physical meaning of Rényi relative entropies,” arXiv:cond-mat/0703008v1, March 1, 2007.
- [2] 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.
- [3] A. H. W. Beck, *Statistical Mechanics, Fluctuations and Noise*, Edward Arnold Publishers, 1976.
- [4] 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.
- [5] 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.
- [6] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, second edition, John Wiley & Sons, 2006.
- [7] N. G. de Bruijn, *Asymptotic Methods in Analysis*, Dover Publications, 1981.
- [8] B. Derrida, “Random–energy model: limit of a family of disordered models,” *Phys. Rev. Lett.*, vol. 45, no. 2, pp. 79–82, July 1980.
- [9] B. Derrida, “The random energy model,” *Physics Reports* (Review Section of Physics Letters), vol. 67, no. 1, pp. 29–35, 1980.
- [10] B. Derrida, “Random–energy model: an exactly solvable model for disordered systems,” *Phys. Rev. B*, vol. 24, no. 5, pp. 2613–2626, September 1981.
- [11] 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.

- [12] B. Derrida and E. Gardner, “Solution of the generalised random energy model,” *J. Phys. C: Solid State Phys.*, vol. 19, pp. 2253–2274, 1986.
- [13] 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.
- [14] R. G. Gallager, *Information Theory and Reliable Communication*, John Wiley & Sons, 1968.
- [15] M. J. W. Hall, “Universal geometric approach to uncertainty, entropy, and information,” *Phys. Rev. A*, vol. 59, no. 4, pp. 2602–2615, April 1999.
- [16] J. Honerkamp, *Statistical Physics – An Advanced Approach with Applications*, 2nd edition, Springer–Verlag, 2002.
- [17] M. Kardar, *Statistical Physics of Particles*, Cambridge University Press, 2007.
- [18] Y. Kaspı 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.
- [19] R. Kawai, J. M. R. Parrondo, and C. Van den Broeck, “Dissipation: the phase–space perspective,” *Phys. Rev. Lett.*, vol. 98, 080602, 2007.
- [20] F. P. Kelly, *Reversibility and Stochastic Networks*, (Chaps 1–3), J. Wiley & Sons, 1979.
- [21] C. Kittel, *Elementary Statistical Physics*, John Wiley & Sons, 1958.
- [22] L. D. Landau and E. M. Lifshitz, *Course of Theoretical Physics – volume 5: Statistical Physics, Part 1*, 3rd edition, Elsevier, 1980.
- [23] F. Mandl, *Statistical Physics*, John Wiley & Sons, 1971.

- [24] 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.
- [25] 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.
- [26] 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.
- [27] 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.
- [28] M. Mézard and A. Montanari, *Information, Physics and Computation*, Oxford University Press, 2009.
- [29] 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.
- [30] H. Nishimori, *Statistical Physics of Spin Glasses and Information Processing: an Introduction*, (International Series of Monographs on Physics, no. 111), Oxford University Press, 2001.
- [31] H. Qian, “Relative entropy: free energy associated with equilibrium fluctuations and nonequilibrium deviations,” *Phys. Rev. E*, vol. 63, 042103, 2001.
- [32] F. Reif, *Fundamentals of Statistical and Thermal Physics*, McGraw–Hill, 1965.
- [33] K. Rose, “A mapping approach to rate-distortion computation and analysis,” *IEEE Trans. Inform. Theory*, vol. 40, no. 6, pp. 1939–1952, November 1994.

- [34] P. Ruján, “Finite temperature error–correcting codes,” *Phys. Rev. Let.*, vol. 70, no. 19, pp. 2968–2971, May 1993.
- [35] A. Somekh–Baruch and N. Merhav, “Exact random coding exponents for erasure decoding,” to appear in *Proc. ISIT 2010*, June 2010, Austin, Texas, U.S.A.
- [36] J. P. Sethna, *Statistical Mechanics: Entropy, Order Parameters, and Complexity*, Oxford University Press, 2007.
- [37] O. Shental and I. Kanter, “Shannon capacity of infinite–range spin–glasses,” technical report, Bar Ilan University, 2005.
- [38] H. Touchette, “Methods for calculating nonconcave entropies,” arXiv:1003.0382v1 [cond-mat.stat-mech] 1 Mar 2010.
- [39] 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.