UNIVERSITÀ DEGLI STUDI DI MILANO

Scuola di Dottorato in Fisica, Astrofisica e Fisica Applicata
Dipartimento di Fisica

Corso di Dottorato in Fisica, Astrofisica e Fisica Applicata
Ciclo XXXIII

Replicas in complex systems: applications to large deviations and neural networks

Settore Scientifico Disciplinare FIS/02

Supervisore: Professor Sergio CARACCIOLO

Coordinatore: Professor Matteo PARIS

Tesi di Dottorato di:
Mauro Pastore

Anno Accademico 2020/2021

Thesis defended on October 27th, 2020

at Università degli Studi di Milano, Dipartimento di Fisica, Milano, Italy

Commission of the final examination:

External Members:
Professor Federico RICCI-TERSENGHI
Professor Riccardo ZECCHINA

Internal Member:
Professor Sergio CARACCIOLO

External referees:

Dottor Giacomo GRADENIGO
Dottor Pierfrancesco URBANI

MIUR subjects:

FIS/02

In the loving memory
of my parents

Acknowledgments

In the past three years I went through the most difficult days of my life. If I came out of them relatively unscathed, I owe it to the people who walked with me along the way, and to whom I would like to express my deepest gratitude.

I literally would not be here writing these lines without the support of my supervisor, professor Sergio Caracciolo: to start with, his trust in me has been my main motivation to take this path; more important, I have always felt his closeness even when far apart. The second pillar of my PhD is Pietro Rotondo, who granted me some free tickets for the Emotional Rollercoaster of Scientific Research: working with him was one of the most stimulating experience of my life. Other people took in turn a guiding role in my journey: I thank Fabrizio Palumbo for the beautiful days in Frascati, Luca Molinari for the passion he can convey with words (and his patience for my RMT-related delays), Marco Gherardi for his sincere commitment to the cause of widening the horizons of knowledge, Matteo Cardella for the massive amount of crazy stuff he tried to explain to me.

My most treasured memories are related to all the friends I have met during these years. I thank the members, effective and honorary, of the Lev Landau Appreciation Club, for all the fun I had working, living, gambling, playing volley together (and the best is yet to come): Riccardo Capelli, Andrea Di Gioacchino, Vittorio Erba, Riccardo Fabbricatore, Enrico Malatesta, Alessandro Montoli, Federica Simonetto.

On a personal level, I thank uncle Sandro and aunt Anna for being there when I needed them most. Last but not least, I must share all the love for my Snack Fellows: my life would be empty without you guys, you are my family.

Contents

Introduction

The replica method has a long history of setbacks and triumphs in Physics. Known by mathematicians at least since [HLP34] as an "obvious identity" (wording by P. W. Anderson, [And88]) to evaluate the average of a logarithm, the replica trick was introduced in statistical mechanics to study problems with a quenched disorder, i.e. systems where the disorder (such as impurities in a lattice) does not thermalize with the other microscopic degrees of freedom, remaining "frozen" with respect to their motion. The theoretical progenitor of these kind of systems is the Edwards-Anderson spin glass model [EA75], a version of the Ising model with random couplings between first-neighbors spins, whose difficulty lead to the formulation of others mean-field models [SK75].

However, for years after its first application, physicists could not overcome some inconsistencies [HP79] arising from this approach, casting a shadow of illegitimacy on its use in the context of disordered systems. It was only with the discovery of the mechanism of replica symmetry breaking [Par79] that the replica trick was not only fully redeemed, but also recognized as a valuable tool to investigate this, at the time new and relatively unexpected, phenomenon; integrated with a series of results needed to describe correctly this mechanism, nowadays it goes under the name of *replica theory* [Dot00], or method. Though not a mathematical well defined theory, the replica method has been proven rigorously to produce the correct results in many models of interest [Gue03].

Since its introduction, replica theory has been applied successfully in a lot of Physics and Mathematics subjects: random matrix theory [EJ76], machine learning and neural networks [Gar87; Gar88], thermodynamics of amorphous solids [MP99], combinatorial optimization problems [CS02a; Cap+18], high energy physics [MS16], etc.; this list is not exhaustive and could go on indefinitely. Its wide range of applicability makes the knowledge of replica methods a relevant point in the cultural background of a theoretical physicist.

Thesis overview

In this thesis, we apply methods from replica theory to deal with two contemporary problems in the study of complex systems.

In **Part I**, we discuss the behavior of the rare fluctuations of the observable of choice in most models of spin glasses: the free energy. Due to the quenched disorder, this and other thermodynamic quantities are self-averaging random variables, whose probability

distribution can be evaluated within replica theory. In particular, the probability of the free energy fluctuations above its typical value shows an anomalous scaling with the number of degrees of freedom, at variance with the ordinary exponential suppression for fluctuations below. We explain how the introduction of a small magnetic field can remove this anomalous behavior. This Part is organized as follows.

In Chap. 1, we present a brief introduction of Large Deviation Theory, the mathematical framework studying the probability of rare events. We introduce the concepts of scaled cumulant generating function and rate function. We discuss briefly the Gärtner-Ellis theorem, a result we will apply in the following chapters.

In Chap. 2, we explain how to apply Large Deviation Theory to spin glass models, in order to find the probability distribution describing the rare fluctuations of the free energy. Using the Random Energy Model as a first example, we give an essential treatment of the replica method and introduce the mechanism of replica symmetry breaking. We discuss the problem of the "very large" deviations, the anomalous scaling of the fluctuations above the typical value, with the aid of extreme value statistics.

In Chap. 3, we apply the formalism introduced before to approach the p-spin spherical model. While the behavior at zero external magnetic field is quite similar to the REM's one, we observe how the anomalous behavior of the free energy fluctuations disappears as soon as the field is switched on. This is our main original contribution to this Part, presented in [PDR19].

In Chap. 4, we evaluate analytically the rate function of the 2-spin spherical model and of the Sherrington-Kirkpatrick model in a magnetic field. In doing so, we explain how to implement a mechanism of full replica symmetry breaking in the context of a large deviation analysis. The results give us a better understanding on the fate of the very large deviations when the field is present.

In **Part II**, we apply the replica formalism to the problem of linear classification of objects with a geometrical structure, in the context of machine learning. In particular, using combinatorial techniques we evaluate the number of dichotomies (binary classifications) of a set of structured inputs achievable by a linear classifier, as a function of the number of inputs to classify. We prove that this number shows an additional critical point beyond the usual storage capacity for isolated points, at which the number of admissible dichotomies becomes zero in the thermodynamic limit; the associated phase transition present a certain degree of replica symmetry breaking. This behavior is due to a trade-off between the increasing number of points to classify, and the increasing volume excluded by their geometrical structure. This approach goes in the direction of finding bounds on the generalization error of certain simple neural network architectures, more stringent for structured data than the ones known from Statistical Learning Theory. This part, which is mostly drawn from [Pas+20; RPG20], is organized as follows.

In Chap. 5, we present the problem of classifications in machine learning, introducing some concepts of Statistical Learning Theory, the mathematical framework where bounds on the generalization performance of neural network architectures are formulated. We explain how this kind of problems can be analyzed from a statistical physics perspective, presenting the classical result of storage capacity by E. Gardner.

In Chap. 6, we introduce a model of data structure, originally proposed in [Bor+19; RLG20]: instead of isolated point, the architecture is required to classify simplexes of points with fixed geometrical interrelations. We evaluate the storage capacity associated to this problem.

In Chap. 7, we devise a combinatorial method to find the asymptotic behavior of the number of admissible dichotomies of simplexes that a linear classifier can realize. This

number is not monotonic, at variance with the unstructured case, and present a novel critical point beyond the storage capacity. This point of transition can be evaluated in a replica approach accounting for some level of replica symmetry breaking.

In Chap. 8, we analyze, in the replica approach, the problem of margin learning from the data structure point of view, finding the (qualitatively) same critical point we encountered in the case of simplexes. This fact leads us to the conclusion that the novel, data-driven phase transition is a general property associated to the geometrical structure of the objects to classify.

While the two parts are fairly independent in language and scope, the first presents the replica framework in a more pedagogical fashion, so the interested reader is encouraged to start from there. For a more methodical exposition, we introduce and conclude each part with chapters of motivation and discussion.

Part I

Large and very large deviations in spin glasses

Motivations

The theory of disordered systems has been mainly developed to describe the typical behavior of physical observables. However, as it has been argued since the early days of the subject, one can employ spin glass techniques in a more general setting, to estimate probability distributions [TD81] and fluctuations around the typical values [TFI89; Cri+90] of quantities of interest. More recently, Rivoire [Riv05], Parisi and Rizzo [PR08; PR09; PR10b; PR10a] and others [ABM04; NH08; NH09] followed this line of thought, providing a bridge between spin glasses (and disordered systems more in general, as in [MPS19]) and the theory of large deviations, that deals with rare events whose probability decays exponentially in the system size. This topic, which is the natural framework to set statistical mechanics in a mathematical perspective, has recently been the subject of a comprehensive and pedagogical review by Touchette [Tou09], as well as of intensive efforts in non-equilibrium statistical physics [PH19].

In this Part, we deal with the probability of rare fluctuations of the free energy in some models of spin glass, our mainly original contribution being a large deviation study of the p-spin spherical model [PDR19]. In zero external magnetic field, we show that a calculation at one-step of replica symmetry breaking produces a very peculiar form of the rate function describing these fluctuations, which is infinite for fluctuations of the free energy above its typical value. In practice, this means that this kind of events are more than exponentially suppressed in probability with the number of degrees of freedom, a situation anomalous in statistical physics, where the scaling of observables usually follows from simple extensivity arguments. This property, which is commonly described stating that the free energy has a "very-large" deviation behavior for positive fluctuations, is present in several other spin glass problems, as discussed for example in [PR10b], and, more generally, in other systems showing extreme value statistics [ABM04]. In some of the early literature [DFM94], this feature is also called "overfrustration".

The situation changes dramatically when a small external magnetic field is applied: the rate function is finite everywhere, although highly asymmetric around the typical value, and so the very-large deviation feature disappears. We explain intuitively the reason of this change of regime in light of the geometrical interpretation discussed for the case without magnetic field, and argue that the introduction of a magnetic field could act as procedure to regularize the anomalous scaling of the large deviation principle for this kind of systems.

CHAPTER 1

Notes on Large Deviation Theory

In this chapter we present a basic mathematical introduction on Large Deviation Theory, defining the language we will use in the rest of this Part. Our aim is to give a self-consistent treatment of the problem to the reader with only elementary notions of probability theory, without bothering him/her with details inessential to the present discussion. For further insights on the subject, we address the reader with a physical background to the nice review [Tou09], to the books [Ell06], [Vul+14] and to the corresponding chapter in [MM09]. A standard mathematical textbook for reference is [DZ10].

1.1 Large deviation principles and rate functions

Suppose, as is common in probability theory and in statistical mechanics, to deal with a sequence of random variables $\{X_i\}_{i=1}^n$, drawn, to fix ideas, independently from the same probability distribution, with mean μ and finite variance σ^2. Starting from this set, we can define random variables that depends collectively on the underling X_i, such as the empirical mean

$$S_n = \frac{1}{n}\sum_{i=1}^{n} X_i \,, \tag{1.1}$$

distributed with a law to be determined from the ones of the X_i. Varying the integer n, we thus obtain a sequence of probability measures for the variables S_n. Of course, in the limit $n \to \infty$ the empirical mean S_n converges to the mean μ, meaning that the probability measure of S_n becomes more and more peaked around μ as n grows: this is simply the statement of the law of large numbers (LLN),

$$\lim_{n\to\infty} P(|S_n - \mu| < \epsilon) = 1 \,. \tag{1.2}$$

Moreover, one of the most celebrated results in probability theory, the central limit theorem (CLT), states that the probability distribution of the variable $\sqrt{n}\,(S_n - \mu)$ converges for $n \to \infty$ to a normal distribution with zero mean and variance σ^2:

$$\lim_{n\to\infty} P\big(\sqrt{n}\,(S_n - \mu) \in [x, x + \mathrm{d}x]\big) = \frac{e^{-\frac{x^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}\,\mathrm{d}x \,. \tag{1.3}$$

This result is so important because it establish the normal distribution as the limiting distribution of a large universality class: not only it can be simply extended to deal with sums of independent random variables with different variances (as long as they are finite), but versions of it hold even when the X_i are dependent (weakly correlated).

Note that the CLT deals with deviations from the mean $\delta S_n = S_n - \mu$ that are "small", of order $1/\sqrt{n}$: how does the tails of the distribution, accounting for the probability of rare events to occur, behaves asymptotically for large n? *Large deviation theory* (LDT) is the mathematical framework to answer this question, in cases where this probability is exponentially suppressed in the limit of $n \to \infty$. At variance with the CLT, the results from LDT preserve some peculiarities of the underlying distributions of the X_i, so we illustrate an elementary case as an example: a Bernoulli process. Suppose to perform a classic coin-toss experiment; the random variables X_i can take two values, head (1) with probability p and tail (0) with probability $(1-p)$; clearly, the mean value of each X_i is equal to p, while the variance is $p(1-p)$. After n tosses, what is the probability for S_n, defined as in (1.1), to take the value k/n? Of course, the ways a fixed number of heads can occur in the sequence is counted by the corresponding binomial coefficient. Being each toss independent, the resulting probability is

$$P(S_n = k/n) = \frac{n!}{k!(n-k)!}\, p^k (1-p)^{n-k}\,. \tag{1.4}$$

To obtain the asymptotic behavior for large n of this law, we must send $n \to \infty$, $k \to \infty$, $x = k/n$ fixed. Using Stirling's approximation, the following formula easily follows:

$$P(S_n = x) \sim \exp\left\{-n\left[x \ln \frac{x}{p} + (1-x)\ln\frac{1-x}{1-p}\right]\right\}, \qquad 0 \le x \le 1. \tag{1.5}$$

This is a large deviation result: it says that this probability is exponentially suppressed in n with a rate given by the non-negative function

$$I(x) = \begin{cases} x \ln \frac{x}{p} + (1-x)\ln\frac{1-x}{1-p} & \text{if } 0 \le x \le 1\,, \\ +\infty & \text{otherwise}\,. \end{cases} \tag{1.6}$$

This is a simple example of a central quantity in LDT, the *rate function*. Moreover, this result implies the LLN, as the only values of x where the probability remains finite are the zeros of $I(x)$, in this case $x = p$. It implies also the CLT, as for $x = p + \delta x$

$$I(x) = \frac{\delta x^2}{2p(1-p)} + O(\delta x^3)\,, \tag{1.7}$$

reproducing in (1.5) the Gaussian function for deviations $\delta x = O(1/\sqrt{n})$. In this regarding, LDT is an extension of those classical results in probability theory. The final aim of a large deviation approach to a certain problem is usually to obtain the rate function of the corresponding probability distribution. As should be clear from our previous example, its specific form depends on the details of the asymptotic expansion of the corresponding distribution at finite n, and so it must be analyzed case by case. However, there are properties of the rate function, some of which we already mentioned, that are completely general. In the following we give a more precise overview of these properties.

Given a sequence of random variables $\{A_n\}$, we denote with $P(A_n \in B)$ the probability that A_n takes value in a set B. We say that A_n satisfies a *Large Deviation Principle* (LDP) with rate $I_B \ge 0$ if

$$\lim_{n\to\infty} -\frac{1}{n}\log P(A_n \in B) = I_B\,. \tag{1.8}$$

The continuous, probability-density version of the above statement is

$$p(A_n = a) \sim e^{-nI(a)}\,, \tag{1.9}$$

where $p(A_n = a)$ is defined by

$$P(A_n \in [a, a + \mathrm{d}a]) = p(A_n = a)\,\mathrm{d}a\,, \tag{1.10}$$

and the symbol $\sim$ is to be intended in the sense of Eq. (1.8). To define a *rate function*, $I : \mathcal{A} \to [0, +\infty]$ must be an extended real-valued function defined on a Hausdorff space $\mathcal{A}$, such that it is not identically $+\infty$ and the sub-level set

$$\{a \in \mathcal{A} | I(a) \leq c\} \quad \text{for } c \geq 0 \tag{1.11}$$

are close in $\mathcal{A}$ (*lower semi-continuity*). If these sets are also compact, then I is called a *good rate function*. As we already noted, the LDP implies the law of large numbers: $p(A_n = a)$ is exponentially small with n *except* where $I(a) = 0$, so the typical values a_{typ} of A_n (those values where the distribution concentrates) are identified with the zeros of the rate function. Moreover, because of the positivity of $I(a)$ (so that a_{typ} is a global minimum for I), an expansion (when possible) around a_{typ} gives

$$p\big(A_n = a_{\text{typ}} + \delta a\big) \sim e^{-\frac{nf''(a_{\text{typ}})}{2}\delta a^2} \tag{1.12}$$

which is the CLT, giving a finite large-n probability for fluctuations $\delta a = O(1/\sqrt{n})$ (the so-called Gaussian regime).

1.2 The scaled cumulant generating function

In most of the realistic cases in Physics, the direct evaluation of the rate function associated to a certain process is beyond the reach of simple asymptotic estimates as the argument we used for the previous example: more sophisticated methods are in need. For this reason, we introduce another key quantity that will recur often in the future discussion, which is easier to evaluate in a lot of cases of interest: the *scaled cumulant generating function* (SCGF, or *logarithmic moment generating function*) of A_n, defined by

$$\Lambda_n(k) = \frac{1}{n}\log \overline{e^{knA_n}}\,, \qquad \Lambda(k) = \lim_{n\to\infty} \Lambda_n(k)\,, \tag{1.13}$$

where the overline indicates the average with respect to the density $p_{A_n}(a) = p(A_n = a)$:

$$\overline{e^{knA_n}} = \int e^{kna} p_{A_n}(a)\,\mathrm{d}a\,. \tag{1.14}$$

This quantity has some remarkable properties:

(i) It is null at the origin, because of the normalization of the probability measure:

$$\Lambda(0) = \lim_{n\to\infty} \frac{1}{n}\log \overline{1} = 0\,. \tag{1.15}$$

(ii) The derivative of $\Lambda(k)$, evaluated at the origin, gives the typical value of A_n in the limit of large n:

$$\Lambda'(0) = \lim_{n\to\infty} \frac{\overline{A_n e^{knA_n}}}{\overline{e^{knA_n}}}\bigg|_{k=0} = \lim_{n\to\infty} \overline{A_n}\,. \tag{1.16}$$

Because of property (i), the derivative can also be evaluated as

$$\Lambda'(0) = \lim_{k\to 0} \frac{\Lambda(k)}{k} \,. \tag{1.17}$$

(iii) The second derivative of $\Lambda(k)$, evaluated at the origin, gives the (rescaled) variance of A_n in the limit of large n:

$$\Lambda''(0) = \lim_{n\to\infty} n \left[\frac{\overline{A_n^2 e^{knA_n}}}{\overline{e^{knA_n}}} - \frac{\left(\overline{A_n e^{knA_n}}\right)^2}{\left(\overline{e^{knA_n}}\right)^2} \right]_{k=0} = \lim_{n\to\infty} n \left[\overline{A_n^2} - \left(\overline{A_n}\right)^2 \right] . \tag{1.18}$$

Of course, this argument is true for all the successive moments of the distribution and is the reason why Λ is called a generating function.

(iv) The function $\Lambda_n(k)$ is convex for any finite n, as can be proven from Hölder inequality:

$$\overline{XY} \le \left(\overline{X^{1/p}}\right)^p \left(\overline{Y^{1/q}}\right)^q , \qquad 0 \le p, q \le 1, \quad p+q = 1 \,, \tag{1.19}$$

using $X = e^{pk_1 nA_n}$, $Y = e^{(1-p)k_2 nA_n}$, so that

$$\overline{e^{[pk_1+(1-p)k_2]nA_n}} \le \left(\overline{e^{k_1 nA_n}}\right)^p \left(\overline{e^{k_2 nA_n}}\right)^{1-p} \tag{1.20}$$

and taking the logarithm:

$$\Lambda_n[pk_1 + (1-p)k_2] \le p\Lambda_n(k_1) + (1-p)\Lambda_n(k_2) \,, \tag{1.21}$$

which is the definition of convexity.

(v) The function $\Lambda_n(k)/k$ is monotonic, as can be proven from another Hölder inequality: take (1.19), but now with $X = e^{kpnA_n}$, $Y = 1$. Indeed

$$\overline{e^{kpnA_n}} \le \left(\overline{e^{knA_n}}\right)^p \tag{1.22}$$

and so the logarithm

$$\Lambda_n(pk) \le p\Lambda_n(k) \,. \tag{1.23}$$

As p is an arbitrary number between 0 and 1, $k' = pk \le k$, so

$$\Lambda_n(k') \le \frac{k'}{k} \Lambda_n(k) \qquad \forall\, k' < k \,, \tag{1.24}$$

and the function $\Lambda_n(k)/k$ must be non-decreasing.

Note, however, that if $f(k)$ is a generic function defined on an interval, then $f(k)$ is convex iff the quantity $[f(k_1) - f(k_2)]/(k_1 - k_2)$ is monotonic in k_1 for every fixed k_2, and vice versa (and so also for $k_2 = 0$, if it is in the domain): the property (v) follows from (i) and (iv).

Why we introduced this quantity, and how is it related to the rate function? Whenever the probability distribution of A_n is given by (1.9), the limit for large n of the SCGF can be evaluated with the saddle point method:

$$\Lambda_n(k) = \frac{1}{n} \log \left[\int e^{kna} p_{A_n}(a) \, \mathrm{d}a \right] \sim \frac{1}{n} \log \left[\int e^{-n[I(a)-ka]} \, \mathrm{d}a \right] \tag{1.25}$$

and so

$$\Lambda(k) = \sup_a \left[ka - I(a) \right] . \tag{1.26}$$

This relation, which says that *the SCGF is the Legendre-Fenchel transformation of the rate function*, is always true whenever a LDP holds for A_n and $\Lambda(k)$ exists. Note that, thanks to Eq. (1.26), the property of convexity (iv) can be thought as a consequence of the SCGF coming from a Legendre-Fenchel transformation. Moreover, because of property (i), the positivity of the rate function follows:

$$0 = \Lambda(0) = \sup_a \left[-I(a) \right] = - \inf_a \left[I(a) \right] . \tag{1.27}$$

1.3 Gärtner-Ellis theorem

As we will see in practice, the direct evaluation of the SCGF is often feasible in statistical mechanics and, in particular, in the replica approach to the study of complex systems. For this reason, we can exploit its relation with the rate function to formulate criteria to understand whether a LDP holds case by case. In other words, we need to clarify under what hypothesis we can invert Eq. (1.26), obtaining I from Λ. Denoting as f^* the Legendre transformation of a function f, we know that $f^{**} = f$ (that is, the Legendre transformation is an involution) only if the original function f is convex and lower semi-continuous: this is the statement of the Fenchel-Moreau theorem. We already know that the rate function is guaranteed, by definition, to comply only with the second request. However, as the rate function is the unknown quantity in this approach, we would like to shift the hypothesis on I into hypothesis on Λ: when is it true that the inverse Legendre transformation of $\Lambda(k)$,

$$I_\Lambda(a) = \sup_k \left[ka - \Lambda(k) \right] , \tag{1.28}$$

is equal to the rate function, $I_\Lambda = I$?

The answer of this question is provided by the *Gärtner-Ellis theorem*, an extension to quite general sequences of probability distributions of the renowned Cramér's theorem, which in turn holds only for empirical means of i.i.d variables as in (1.1). Informally, under the hypothesis that the limit in Eq. (1.13) exists finite and $\Lambda(k)$ *is differentiable for any* $k \in \mathbb{R}$, this theorem states that A_n satisfies a LDP with a rate function given by

$$I(a) = I_\Lambda(a) = \sup_k \left[ka - \Lambda(k) \right] . \tag{1.29}$$

The reader with a background in thermodynamics and statistical mechanics can understand this argument recollecting some properties of the Legendre-Fenchel transformation: if I is a non-convex function, its Legendre transformation Λ has non differentiable points, so we can conclude that I_Λ is equal to the *convex hull* of I, and not to the rate function itself.

Of course, the above argument is not rigorous: to begin with, the theorem gives conditions for a LDP to hold, but to derive Eq. (1.26) we had to impose it a priori; still,

it is valuable to stress the importance of the request of differentiability on the SCGF. To formulate the theorem in mathematical terms, we only need the following definition.

Definition 1.1 (Exposed point). *$x \in \mathbb{R}$ is an **exposed point** of the function $\Lambda : \mathbb{R} \to \mathbb{R}$ if there is a $t \in \mathbb{R}$ such that*

$$ty - \Lambda(y) > tx - \Lambda(x) \qquad \forall\, y \neq x \tag{1.30}$$

In practice, a point x is exposed if the curve $\Lambda(y)$ lies strictly above the line of slope t passing through the point $(x, \Lambda(x))$. If Λ in convex, a sufficient condition for x to be an exposed point is that Λ is twice differentiable at x with $\Lambda''(x) > 0$. Given that, the Gärtner-Ellis theorem follows:

Theorem 1.1 (Gärtner-Ellis). *Consider a sequence of random variables A_n and assume that $\Lambda_n(k)$ defined in (1.13) exists and has a finite limit $\Lambda(k) = \lim_{n\to\infty} \Lambda_n(k)$ for any $k \in \mathbb{R}$. Define I_Λ as the inverse Legendre transform in (1.28) and let E be the set of exposed points of I_Λ. Then*

1. *For any closed set $C \in \mathbb{R}$,*

$$\limsup_{n\to\infty} \frac{1}{n} \log P(A_n \in C) \leq - \inf_{a\in C} I_\Lambda(a) \tag{1.31}$$

2. *For any open set $O \in \mathbb{R}$,*

$$\limsup_{n\to\infty} \frac{1}{n} \log P(A_n \in O) \geq - \inf_{a\in O\cap E} I_\Lambda(a) \tag{1.32}$$

3. *If $\Lambda(k)$ is differentiable for any $k \in \mathbb{R}$, then a LDP holds for A_n with the good rate function $I = I_\Lambda$.*

The last statement can be made weaker (requiring Λ to be essentially smooth in $\mathbb{R}$), but this is beyond our scope: in short, the differentiability of Λ in $\mathbb{R}$ ensures that the bounds (1.31) and (1.32) are strict.

In the next chapters we will try to apply this theorem to study the probability distribution of physical observables in some disordered models of interest, using the principles of convexity and differentiability of the SCGF we presented so far to evaluate the corresponding rate functions. We will also discuss cases where the hypothesis of differentiability does not hold.

CHAPTER 2

Spin glasses and large deviations

The large deviation theory we presented in the previous chapter can be thought as the mathematical framework to formulate statistical mechanics itself, as a probabilistic theory of a large number of degrees of freedom (for more on this point of view, see the book [Ell06]). For example, in equilibrium statistical mechanics the entropy and the free energy can be thought, respectively, as the rate function and the SCGF of the mean energy (energy per state) with respect to the probability distribution of the ensemble of microstates. Because of this broad scope, it is appropriate, whenever large deviations results are invoked to deal with a physical system, to start from clarifying to what random variable are applied, and to study the fluctuations with respect to what probability measure. In the following, we will focus on the fluctuations of the free energy of certain models of mean-field spin glasses at equilibrium, with respect to the probability distribution of the disorder.

In this chapter, we start sketching briefly what a spin glass is, and how the large deviation formalism can be quite naturally applied in this context. Then, we illustrate the main results of this approach with the aid of the Random Energy Model, a toy model of disordered system simple enough to grant a complete analytical control on various calculations, but still with a non-trivial behavior we will find also in other models.

2.1 Replicating the partition function

Since its introduction in the 1970s to model some magnetic properties experimentally observed in diluted magnetic alloys, spin glass theory has gone far beyond the horizon of solid-state physics, finding applications in neural networks, genetics, evolution, optimization problems and more. For a charming report on the history of the field, we refer to the series of papers [And88]; for classic textbooks on the subject, we address the reader to [MPV86; Dot95; Nis01].

For our purpose, we can say that spin glasses are systems described by Hamiltonians $\mathcal{H}_J[\sigma]$, depending on the configurations of n spins $\{\sigma_i\}_{i=1}^n$ coupled with some parameters $\{J\}$ that are random variables, modeling the disorder. The spins can be discrete variables, as in the Ising and Potts models, or real continuous variables subject to some global constraints, as in spherical models. They live on the vertices of a graph,[1] whose links, weighted by the couplings J, represent the interactions between them. Whenever this graph is *complete* (i.e. each vertex is first neighbor to any other), the model is called *fully-connected*, or *mean-field*. In the simplest models, the couplings are chosen to

[1] Technically, a hypergraph, to model systems with multiple-body interactions. A convenient way to represent these systems is via *factor graphs*, with two kind of vertices (i.e., bipartite), one for the spins and one for the couplings between them. See, for further details, [MM09].

be i.i.d. random variables. For example, in a model with only 2-body interactions, J can be thought as a matrix, whose entry J_{ij} is the coupling between the spins i and j, extracted from a normal distribution (Gaussian model)

$$p(J_{ij}) = \frac{1}{\sqrt{2\pi J^2}} \exp\left[-\frac{(J_{ij} - J_0)^2}{2J^2}\right], \tag{2.1}$$

where J_0 is the mean and J^2 the variance of the distribution, or from a Bernoulli distribution ($\pm J$ model)

$$p(J_{ij}) = \alpha\delta(J_{ij} - J) + (1-\alpha)\delta(J_{ij} + J), \qquad 0 \leq \alpha \leq 1. \tag{2.2}$$

Of course, we could think of more realistic models where the interaction between spins depends, for example, on the Euclidean distance between them, in such a way that the couplings become correlated; however, in the following we will not dwell on problems of this kind.

A main property of the disorder in spin glass systems is to be *quenched*: from the point of view of the thermal fluctuations of the elementary degrees of freedom (the spins), it is frozen in a way similar to what happens to impurities trapped in a material under rapid cooling, from which the term is borrowed. The reason behind this request is to describe systems where the disorder evolution is much slower than the time needed by the other variables to thermalize. The opposite situation, where the disordered parameters can fluctuate on the same time-scales as the other degrees of freedom, and so can be treated as dynamical variables, is called *annealed disorder*. In practice, in a quenched system we have to fix the couplings J to some values chosen from the corresponding distribution, and then evaluate all the thermodynamic observables without changing those values.

Quenched disorder is known to introduce *frustration* in the system: in general, if we follow a closed loop on the graph, we cannot orientate the spins in such a way to minimize the energy of every link of the path, due to the fact that both ferromagnetic and antiferromagnetic links can be met with a certain probability. For this reason, there is a great number of degenerate states that globally minimize, as well as possible, the energy, and the energy landscape of the system assumes a very complicated profile, with lots of valleys and spikes; in some cases, a large basin in the energy landscape, corresponding to a macroscopic equilibrium state, is broken up at a critical temperature in a fractal hierarchy of basins within basis, producing a very rough profile [Cha+14]. It can happens that at certain temperatures the system is trapped in some of these valleys, because the thermal fluctuations are not strong enough to overcome the energy barriers: the system can no more explore the full phase space, and so the phenomenon of *ergodicity breaking* occur.

Evaluating the thermodynamic quantities, such as the free energy, at a given instance of the disorder means that they inherit a stochastic nature from the couplings: for different instances, they assume different values. An interesting question would be: how can the probability distribution of the free energy of a given model be obtained from the one of the couplings? The (density of) free energy at a given instance of the disorder is

$$f_J = \lim_{n\to\infty} f_{J,n} = -\frac{1}{\beta} \lim_{n\to\infty} \frac{1}{n} \log \mathcal{Z}_{J,n}, \tag{2.3}$$

where the partition function is defined as

$$\mathcal{Z}_{J,n} = \mathrm{Tr}_\sigma\, e^{-\beta H_J[\sigma]}, \qquad \mathcal{Z}_J = \lim_{n\to\infty} \mathcal{Z}_{J,n}. \tag{2.4}$$

Denoting with an overline the average on the disorder, the typical value is

$$f_{\rm typ} = \overline{f_J} = -\frac{1}{\beta}\lim_{n\to\infty}\frac{1}{n}\overline{\log \mathcal{Z}_{J,n}}\ . \tag{2.5}$$

The free energy is known to be a *self-averaging* observable, meaning that the probability measure of $f_{J,n}$ concentrates, in the limit of large n, around the typical value. From the point of view of the discussion in Chap. 1, we can say that in this case a LLN holds:

$$\lim_{n\to\infty} P\big(|f_{J,n} - f_{\rm typ}| < \epsilon\big) = 1\,, \tag{2.6}$$

which means that, for example, $\lim_{n\to\infty}\overline{(f_{J,n})^2}\,/\overline{(f_{J,n})}^2 = 1$. In order to obtain general results independent from the specific realization of the disorder, it is important in statistical mechanics to identify the observables having this property. However, while this is true for the (rescaled) logarithm of $\mathcal{Z}_{J,n}$, which is indeed the free energy, the partition function itself in not a self-averaging quantity, meaning in particular that

$$\overline{\log \mathcal{Z}_J} \neq \log \overline{\mathcal{Z}_J}\ . \tag{2.7}$$

Note that on the RHS, the average over the disorder is performed at the same level of the summation over the spin configurations in (2.4): it is an annealed calculation, while the correct quenched one is on the LHS. As the partition function is, in statistical mechanics, the fundamental quantity to start with, the evaluation of the averaged logarithm in Eq. (2.7) poses a technical problem: while the annealed approach is usually easy, because in most of the models the parameters J enter linearly in the Hamiltonian and the average can be performed, the quenched one is a lot more challenging. A fundamental tool in this respect is the so-called *replica trick*:[2]

$$\overline{\log \mathcal{Z}_n} = \lim_{k\to 0}\frac{\overline{(\mathcal{Z}_n)^k} - 1}{k}\,. \tag{2.8}$$

The reason for the success of this formula, which is of course an elementary analytical identity for the logarithmic function, is that there is nothing easier than "replicating" the partition function:

$$(\mathcal{Z}_n)^k = \Big(\mathrm{Tr}_\sigma\, e^{-\beta H_J[\sigma]}\Big)^k = \left[\prod_{a=1}^{k}\mathrm{Tr}_{\sigma_a}\right] e^{-\beta\sum_{a=1}^{k} H_J[\sigma_a]}\,, \tag{2.9}$$

so that the spins from different replicas are now labeled by the index a and the average of this quantity poses no more problems than the simple annealed evaluation. The k independent copies of the theory becomes interacting once the average over J is performed, in a way we will see in a number of examples in the following. We stress here that the apparently simple trick (2.8) and its application to spin glass models hide a lot of subtleties, actually. For example, to perform the calculations the order of limits implicit in Eq. (2.8) (first the limit $k \to 0$, and then the thermodynamic limit in Eq. (2.5), i.e. $n \to \infty$) is reversed, a procedure usually justified a posteriori. However, the main issue with the replica trick, which has been historically a serious obstacle at the early days of spin glass theory, is that the recipe (2.9) to evaluate a power of the partition function makes sense only for k integer: there is no real hint on how to perform the continuation to $k \in \mathbb{R}$ that we need to obtain the limit $k \to 0$.

[2] In the following, we will omit the subscript J from the fixed-instance quantities.

Putting aside for a moment this difficulty, we observe that, once the moments $\overline{\mathcal{Z}^k}$ of the partition function are known, we also have the full SCGF of the free energy: given the quantity

$$G(k) = \lim_{n\to\infty} -\frac{1}{\beta n} \log \overline{(\mathcal{Z}_n)^k} = \lim_{n\to\infty} -\frac{\log \overline{e^{-k\beta n f_n}}}{\beta n} , \tag{2.10}$$

then the SCGF of the free energy is

$$\Lambda(k) = \lim_{n\to\infty} \frac{\log \overline{e^{knf_n}}}{n} = -\beta G(-k/\beta) . \tag{2.11}$$

In view of this equality, in the following we will use the term SCGF indifferently for G and Λ, being careful to use the correct one to evaluate the rate function.

From the disordered systems perspective, most of the standard results of spin glass theory obtained within the replica method concern only the very special limit $k \to 0$ in Eq. (2.8), to obtain the typical values, whereas to evaluate the full form of $I(x)$ that describes arbitrary rare fluctuations of the free-energy one needs to work out the SCGF for finite replica index k. This problem is clearly equivalent to the one we mentioned above: that is, to determine the full analytical continuation of the averaged replicated partition function from integer to real number of replicas k. It was extensively investigated in the early stage of the research in disordered systems in order to understand the manifestation of the (at that time surprising) mechanism of replica symmetry breaking [Par79]. Since these results are particularly interesting from the more modern large deviation viewpoint, we now briefly mention the main ones.

Van Hemmen and Palmer [HP79] were the first ones to observe that the expression in Eq. (2.10) must be a convex function of the replica index k, a property crucial in order to interpret this quantity as a SCGF. Shortly later, Rammal [Ram81] added that $\Lambda(k)/k$ must be monotonic, which is, as we know, a necessary condition for the convexity of $\Lambda(k)$. However, the replica symmetric (RS) ansatz, which provides the most obvious analytical continuation to real k of the replicated partition function, gives often a trial SCGF which is not convex, or such that $\Lambda(k)/k$ is not monotonic. This problem has been analyzed for the first time in the context of the Sherrington-Kirkpatrick (SK) model. After Parisi introduced his remarkable hierarchical scheme for replica symmetry breaking, Kondor [Kon83] argued that his full RSB solution was very likely to provide a good analytical continuation of Eq. (2.10), not only around $k = 0$.

These results may be considered nowadays as the initial stage of a work that attempted to give mathematical soundness to the replica method. Although this vast program is mostly unfinished, Parisi and Rizzo in a series of papers [PR08; PR09; PR10b; PR10a] realized that the original analysis presented by Kondor is fundamental to investigate the large deviations of the free-energy in the SK model. Large deviations have been examined only for a few other spin glass models: Gardner and Derrida discussed the form of the SCGF in the random energy model (REM) in a seminal paper [GD89], and many rigorous results have been established later on [FFM07]; Ogure and Kabashima [OK04; OK09a; OK09b] considered analyticity with respect to the replica number in more general REM-like models; Nakajima and Hukushima investigated the p-body SK model [NH08] and dilute finite-connectivity spin glasses [NH09] to specifically address the form of the SCGF for models where one-step replica symmetry breaking (1RSB) is exact; Andreanov, Barbieri and Martin [ABM04] considered the fluctuations of the ground-state energy of some models of spin glasses. The replica method has been applied to study the large deviations of observables in more general settings than spin

glasses, as in [MPS19]. A parallel approach, based on the cavity method, has been pioneered by Rivoire [Riv05].

To start going into detail, we review, in the rest of this chapter, some results from replica theory and large deviations in the REM.

2.2 The Random Energy Model: a simple model of disordered system

The random energy model (REM) was introduced by Derrida [Der80; Der81] as a toy model of spin glasses.[3] Though not a genuine spin model, it can be seen as limit of a family of fully-connected spin glass: a generalization of the SK model consisting of systems with n Ising spins with infinite-range random Gaussian p-body interactions, in the limit $p \to \infty$. For a proof of this property in the replica approach, see also [GM84]. In this respect, most of the techniques used to deal with spin glasses can be applied to study this model, with the remarkable advantage that, being a pure probabilistic model, the results from the replica method can be checked via formal mathematical approaches.

To introduce the model, we suppose that the energy levels E_i, with $i \in \{1, \cdots, 2^n\}$, of a given instance of our disordered system are i.i.d. Gaussian random variables, with zero mean and variance $n/2$. Therefore the partition function of a specific instance is

$$\mathcal{Z} = \sum_{i=1}^{2^n} e^{-\beta E_i}. \tag{2.12}$$

An instance of the problem is defined giving the set $\{E_1, E_2, \dots, E_{2^n}\} = \{\mathbf{E}\}$, and we are interested in quantities averaged on the disorder (that is, on the value of each energy level). We use the overline notation to identify this kind of average operation, that is

$$\overline{\mathcal{O}(\mathbf{E})} = \int \prod_{i=1}^{2^n} \frac{\mathrm{d}E_i}{\sqrt{\pi n}} e^{-E_i^2/n}\, \mathcal{O}(\mathbf{E}). \tag{2.13}$$

Our aim is the computation of the free energy density $f(\mathbf{E}) = -1/(n\beta) \log \mathcal{Z}$ in the limit $n \to \infty$ in the quenched case, and the large-deviation rate function of f, using a replica approach. We will mainly follow, from a large-deviations perspective, the exposition in [MM09]. Eventually, we will compare the results with the ones known from probabilistic methods.

We start writing our k-replicated partition function as

$$\mathcal{Z}^k = \sum_{i_1,\dots,i_k=1}^{2^n} e^{-\beta \sum_{a=1}^k E_{i_a}} = \sum_{i_1,\dots,i_k=1}^{2^n} \prod_{j=1}^{2^n} e^{-\beta E_j \sum_{a=1}^k \delta_{i_a,j}}. \tag{2.14}$$

Then, we perform the average over the disorder, obtaining, with a simple Gaussian integration,

$$\overline{\mathcal{Z}^k} = \sum_{i_1,\dots,i_k=1}^{2^n} \exp\left(\frac{\beta^2 n}{4} \sum_{a,b=1}^{k} \delta_{i_a,i_b} \right). \tag{2.15}$$

Notice that the k-replicated-system configuration is given by $(i_1, \dots, i_k)$, where $i_\ell = j$ means that the ℓ-th replica is in the j-th energy state. After the average over the energy

[3] A kind of random energy model was considered before by Nicola Cabibbo, who did not publish any result. See [Gue13].

levels, the only meaningful information remaining is which replicas are in the same state and which are not. In this sense, the disorder average introduces a correlation between different replicas, which initially are independent. Therefore, we can define a parameter which encodes in a natural way this information, the $k \times k$ *overlap matrix*

$$Q_{ab} = \delta_{i_a, i_b} \,. \tag{2.16}$$

Denoting now with $\mathcal{N}(\boldsymbol{Q})$ the number of configurations $(i_1, \dots, i_k)$ whose overlap matrix is a certain $\boldsymbol{Q}$ (multiplicity), we have

$$\overline{\mathcal{Z}^k} = \sum_{\boldsymbol{Q}} \mathcal{N}(\boldsymbol{Q}) \exp\left(\frac{\beta^2 n}{4} \sum_{a,b=1}^{k} Q_{ab} \right) , \tag{2.17}$$

where the sum runs over all the symmetric matrices with off-diagonal elements $\{0, 1\}$ and 1 on the diagonal. Defining a sort of entropy density function

$$s(\boldsymbol{Q}) = \frac{\log \mathcal{N}(\boldsymbol{Q})}{n} \tag{2.18}$$

for the multiplicity of the matrices $\boldsymbol{Q}$, assuming implicitly that this quantity is intensive, we obtain

$$\overline{Z^k} = \sum_{\boldsymbol{Q}} \exp\left[n g(\boldsymbol{Q}) \right] , \tag{2.19}$$

with

$$g(\boldsymbol{Q}; k, \beta) = \frac{\beta^2}{4} \sum_{a,b=1}^{k} Q_{ab} + s(\boldsymbol{Q}) \,. \tag{2.20}$$

Since n is large and the partition function is a sum of exponential terms in n, we can search the only relevant value(s) of $\boldsymbol{Q}$ as the extrema of $g(\boldsymbol{Q})$ (maxima or minima, depending on its sign). If the dominant extremum is $\boldsymbol{Q}^\star$, we have[4]

$$G(k, \beta) = \lim_{n \to \infty} g(\boldsymbol{Q}^\star; k, \beta) \,. \tag{2.21}$$

Note that, for any given values of k and β, we have to find the corresponding extremum, so $\boldsymbol{Q}^\star = \boldsymbol{Q}^\star(k, \beta)$ is a function of the number of replicas and of the temperature.

In the search for the extrema of $g(\boldsymbol{Q}; k, \beta)$, we observe that this function has a very important property: it is symmetric under permutation of any pair of replicas. Indeed, given a permutation of k objects $\pi \in S_k$ and denoted $Q^\pi_{ab} = Q_{\pi(a)\pi(b)}$, then $g(\boldsymbol{Q}^\pi; k, \beta) = g(\boldsymbol{Q}; k, \beta)$. This property is called *replica symmetry*. In principle, the dominant extremum of $g(Q; k, \beta)$ could break this symmetry: it is the same mechanism at work in systems exhibiting a spontaneous symmetry breaking, so this phenomenon is called *replica symmetry breaking* (RSB). Let us *assume* that for $k \in \mathbb{N}^+$ (that is, k positive integer) the replica symmetry is not broken.[5] Under this fundamental hypothesis, we have

$$Q_{ab} = \begin{cases} 1 & \text{if } a = b, \\ q & \text{if } a \neq b, \end{cases} \tag{2.22}$$

[4] For the REM, we are using for convenience a different convention from that Eq. (2.10), which can be obtained using $G_{\text{REM}} = -\beta G$.

[5] This assumption is not strictly true even for $k \in \mathbb{N}^+$, as explained in [Gue13]: RSB can occur for an integer number of replicas, although in this case the broken phase would present the same free energy, internal energy and entropy of the symmetric one. Very different is the case for $k \in [0, 1]$, which we will discuss in the following, where RSB must be postulate to obtain the correct thermodynamic observables.

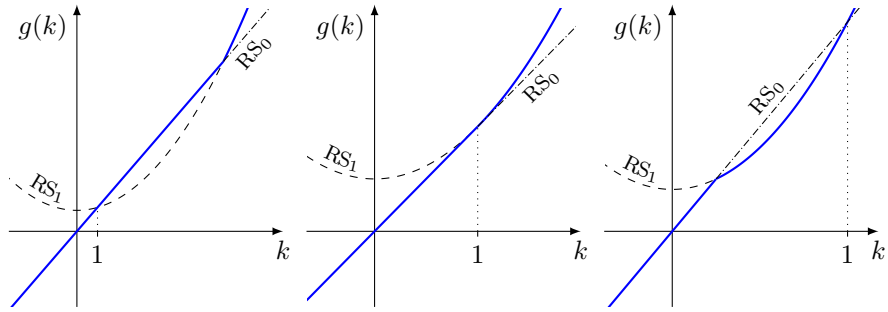

Figure 2.1: Trial SCGF for the REM: $\beta < \beta_c$ (left), $\beta = \beta_c$ (center), $\beta > \beta_c$ (right). The point where the blue curve changes from linear to parabolic is k_{textRS}.

with $q = 0$ or $q = 1$. With a replica matrix of this form, the function (2.20) becomes

$$g_{\mathrm{RS}}(q; k, \beta) = \frac{\beta^2}{4} k \left[(k-1)q + 1\right] + s(\boldsymbol{Q}) \,. \tag{2.23}$$

If $q = 0$, we have $\mathcal{N}(\boldsymbol{Q}) = 2^n(2^n - 1)\cdots(2^n - k + 1)$ (because all the replicas must be different: we can choose the first replica in one of the 2^n states, but then the second must be in one of the remaining $2^n - 1$, etc...), so, keeping only the leading terms in n,

$$g_{\mathrm{RS}_0}(k, \beta) = k\left(\frac{\beta^2}{4} + \log 2\right). \tag{2.24}$$

This quantity is the same we would obtain in the annealed approximation $\overline{\mathcal{Z}^k} \to \overline{\mathcal{Z}}^k$, as we can see averaging directly Eq. (2.12) and then taking the k-th power: indeed, $q = 0$ means that the replicas are not correlated and the product factorizes, leaving only a linear term in k at the exponent. We will sometimes call this solution the *paramagnetic line*. Otherwise, if $q = 1$ we have all the replicas in the same energy level and there are only $\mathcal{N}(\boldsymbol{Q}) = 2^n$ matrices (one for each state), so

$$g_{\mathrm{RS}_1}(k, \beta) = \frac{\beta^2}{4} k^2 + \log 2. \tag{2.25}$$

The two solutions coincide in $k = 1$, where the annealed and quenched calculation are trivially equivalent, and in the point $k_{\mathrm{RS}} = 4\log(2)/\beta^2$, which crosses the point $k = 1$ at the temperature

$$\beta_c = 2\sqrt{\log 2}\,. \tag{2.26}$$

With this definition,

$$k_{\mathrm{RS}} = \frac{\beta_c^2}{\beta^2}\,. \tag{2.27}$$

We are now in the position to perform the extremization needed to obtain G from g, Eq. (2.21). Of course, for $k \in \mathbb{N}^+$, the correct choice is

$$G(k, \beta) = \max\left\{g_{\mathrm{RS}_0}(k, \beta)\,, g_{\mathrm{RS}_1}(k, \beta)\right\} \quad \text{for } k \in \mathbb{N}^+, \tag{2.28}$$

as this is the term dominating the sum (2.19) for large n. Note however that, after our RS ansatz (2.22), the number of replicas k enter as a parameter in the functions g_{RS_0}, g_{RS_1},

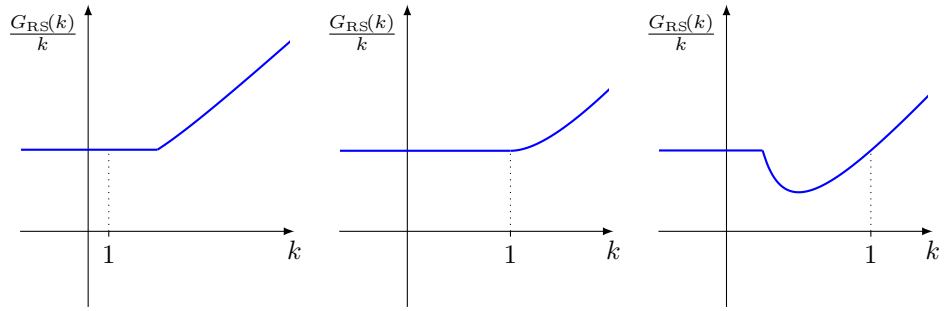

Figure 2.2: Trial $G(k)/k$ for the REM: $\beta < \beta_c$ (left), $\beta = \beta_c$ (center), $\beta > \beta_c$ (right). On the right, the curve $G_{\rm RS}(k)/k$ is monotonically non-increasing for any $k < k_m$, increasing for $k > k_m$.

so we would like to extend this prescription to find the analytic continuation defining the full SCGF of the free energy, promoting k to a real variable. Let us analyze first the high-temperature case, $\beta < \beta_c$ (see Fig. 2.1, left). For $k > k_{\rm RS}$, our "extended" maximization prescription selects $g_{\rm RS_1}$, the parabolic function, which dominates for large k as expected; for $1 < k < k_{\rm RS}$ the linear function, $g_{\rm RS_0}$, is the leading term; something very strange happens for $k < 1$. Indeed, if we stick with our maximization prescription, we should switch again functions at $k = 1$, remaining on the parabolic branch for all $k < 1$. However, this continuation is clearly inconsistent: the resulting trial SCGF does not even pass through the origin! We can overcome this obstacle with the following argument: there is no general reason to change branch in $k = 1$; however, in this point the functions $g_{\rm RS_0}$ and $g_{\rm RS_1}$ switch inevitably order; the only way to remain on the same branch we were on $k = 1^+$, is to change the extremization procedure from a $\max$ to a $\min$ for $k < 1$. This is indeed the continuation prescription to $k \in \mathbb{R}$ in spin glass theory:

$$G_{\rm RS}(k,\beta) = \begin{cases} \max\left\{g_{\rm RS_0}(k,\beta)\,, g_{\rm RS_1}(k,\beta)\right\} & \text{if } k \geq 1, \\ \min\left\{g_{\rm RS_0}(k,\beta)\,, g_{\rm RS_1}(k,\beta)\right\} & \text{if } k < 1. \end{cases} \tag{2.29}$$

We know of no better or formal argument to justify this prescription, which is applied everywhere in the replica approach of disordered system, in the present model. The more popular, hand-waving reasoning goes like this: in models where the replicated partition function has a term like $\sum_{a,b} Q_{ab}$ at the exponent, the RS ansatz gives $k(k-1)q$ from the off-diagonal elements; the sign of $k(k-1)$, which is the factor determining which is the dominant contribution in the sum over different replica matrices, changes in $k = 1$, forcing to switch the prescription to find the extremum (because the maximum above $k = 1$ becomes a minimum below). However, our large-deviation point of view, which drives us to interpret $G(k)$ as a SCGF, should be enough in the present context to understand why this prescription must be true: once we adopt it, we are able to define a legitimate cumulant interpolating the RS integer points. Of course, the resulting function is still not differentiable in $k_{\rm RS}$, meaning that in this point there is actually a phase transition and that we cannot hope to use the Gärtner-Ellis theorem to find the rate function, but only its convex hull.

Let us come back again to Fig. 2.1 (right), to study the low-temperature case of $\beta > \beta_c$. We now see that the point $k_{\rm RS}$ is less than 1, and our $G_{\rm RS}(k,\beta)$, obtained from Eq. (2.29), has a serious problem that invalidates its probabilistic interpretation: *it is not convex*, a property that, as we know from Sec. 1.2 (property (iv)), a legitimate SCGF must have by construction! Another way to understand the problem is to look at Fig. 2.2, where

we plot $G_{\mathrm{RS}(k)}/k$: this function must be monotonic, because of property (v), but in the low-temperature phase it loses monotonicity at the point

$$k_m = \frac{\beta_c}{\beta}\,. \tag{2.30}$$

The reason why this and other inconsistencies appear in this formulation has been extensively investigated during the 1970s: for a careful critical account of the RS replica approach and its problems, written before the solution to the puzzle was found, see [HP79]. Nowadays, we know that the problem is in the ansatz (2.22): replica symmetry is broken at low temperature, as we will see in the following.

2.3 SCGF and rate function of the REM free energy

After the first attempts to go beyond the RS ansatz by Bray and Moore [BM78] and Blandin [Bla78; BGG80], Parisi, in his celebrated series of papers [Par79], found the appropriate parametrization for the replica matrix Q_{ab} providing the correct continuation to $k \to 0$ for systems exhibiting RSB. In the Parisi's scheme, replica symmetry is broken via a hierarchical, step-wise procedure, in the limit of infinite number of step. However, for our current purposes, we only need to introduce the so-called *first step of replica symmetry breaking* (1RSB), which gives already the correct answer for the REM. We will return to the general scheme in Chap. 4.

In the 1RSB ansatz, the replica matrix has the following form:

$$Q_{ab} = \begin{cases} 1 & \text{if } a = b, \\ q_1 & \text{if } a \neq b \text{ are in a diagonal } m \times m \text{ block}, \\ q_0 & \text{otherwise.} \end{cases} \tag{2.31}$$

For example, in the case $k = 9$, $m = 3$ we have

$$Q = \begin{pmatrix} 1 & q_1 & q_1 & q_0 & q_0 & q_0 & q_0 & q_0 & q_0 \\ q_1 & 1 & q_1 & q_0 & q_0 & q_0 & q_0 & q_0 & q_0 \\ q_1 & q_1 & 1 & q_0 & q_0 & q_0 & q_0 & q_0 & q_0 \\ q_0 & q_0 & q_0 & 1 & q_1 & q_1 & q_0 & q_0 & q_0 \\ q_0 & q_0 & q_0 & q_1 & 1 & q_1 & q_0 & q_0 & q_0 \\ q_0 & q_0 & q_0 & q_1 & q_1 & 1 & q_0 & q_0 & q_0 \\ q_0 & q_0 & q_0 & q_0 & q_0 & q_0 & 1 & q_1 & q_1 \\ q_0 & q_0 & q_0 & q_0 & q_0 & q_0 & q_1 & 1 & q_1 \\ q_0 & q_0 & q_0 & q_0 & q_0 & q_0 & q_1 & q_1 & 1 \end{pmatrix} \tag{2.32}$$

In practice, we have subdivided the k replicas in k/m groups of size m, supposing that replicas inside the same group have overlap q_1, while replicas from different groups have overlap $q_0 \leq q_1$. In the REM, both q_1 and q_0 take values in $\{0, 1\}$, so the form of this matrix is different from the RS one (2.22) only when $q_1 = 1$, $q_0 = 0$. How many choices of the original indices produce a matrix of this form? We can always choose the first m replicas in the first group (so, in 2^n different ways), the second m in the second group (in the remaining 2^{n-1} ways) and so on until we run out of groups, so

$\mathcal{N}(\boldsymbol{Q}) = 2^n(2^n-1)\cdots(2^n-k/m+1)$. The sum over the matrix elements is easy, so we obtain from Eq. (2.20)

$$\begin{aligned} g_{\text{1RSB}}(m;k,\beta) &= \frac{\beta^2}{4}k\left[-q_1+m\left(q_1-q_0\right)+kq_0+1\right]+\frac{k}{m}\log 2\,, \\ &= k\left(\frac{\beta^2}{4}m+\frac{\log 2}{m}\right)\,. \end{aligned} \tag{2.33}$$

Note that, if we take $m=1$ (k trivial blocks with only the diagonal element equal to 1), we obtain again g_{RS_0}, while for $m=k$ (a single block coinciding with the matrix itself) we get g_{RS_1}.

The main advantage of Eq. (2.33) is that we can now treat m as a variational parameter, possibly finding new solutions for our extremization procedure. In what domain can we choose this parameter? Of course, as long as $k\in\mathbb{N}^+$, m must be a positive integer as well, between 1 and k by construction. When $k\in\mathbb{R}$, there is no problem to extend this domain to the real interval $[1,k]$, as long as $k\geq 1$. When $k<1$, in addition to the $\max\to\min$ prescription we explained in (2.29), we also have to reverse the order relation to $k\leq m\leq 1$; we will also take $m\geq 0$, for now simply as a continuation prescription, without further justification than keeping finite and positive $s(\boldsymbol{Q})$ (note that g_{1RSB} is non-analytic in $m=0$). We obtain

$$G_{\text{1RSB}}(k,\beta) = \begin{cases} \max\limits_{1\leq m\leq k} g_{\text{1RSB}}(m;k,\beta) & \text{if } k\geq 1, \\ \min\limits_{\substack{k\leq m\leq 1 \\ m\geq 0}} g_{\text{1RSB}}(m;k,\beta) & \text{if } k<1. \end{cases} \tag{2.34}$$

To find the optimal value $m^\star$, we start from the stationarity condition:

$$\frac{\partial g_{\text{1RSB}}}{\partial m}=0 \quad\Longrightarrow\quad m_\pm = \pm\frac{2\sqrt{\log 2}}{\beta}=\pm\frac{\beta_c}{\beta}\,, \tag{2.35}$$

where we used (2.26). The point m_+ is always a minimum of the function. At high temperature, $\beta<\beta_c$, m_+ is greater than 1: depending on the value of k, we have to select one of the extrema of the domain of m ($m^\star=1$ or $m^\star=k$); for $k<1$, $m\in[k,1]$ can never reach m_+, the function is always decreasing and the minimum is in $m^\star=1$ (paramagnetic solution); for $k>0$, the function is first decreasing and then increasing, so the maximum is in $m^\star=1$ as long as $k<k_{\text{RS}}$, otherwise it is in $m^\star=k$ (RS solution). So far, we obtain the same results as before.

Instead, in the case of low temperature, $\beta>\beta_c$, the point m_+ is less than 1: for $k>m_+$, m is always greater than m_+, the function is always increasing, so we have to select $m^\star=k$ (which is a minimum in m for $k\in[m_+,1]$, a maximum for $k>1$); otherwise, for $k<m_+$, this point is in the domain of m, in the region where we have to select a minimum, so we get $m^\star=m_+$. We report the low-temperature result, which is the only different from before, in Fig. 2.3. Note that, due to the fact that $m_+=k_m$ in Eq. (2.30), the 1RSB ansatz is equivalent to a simple, Maxwell-like construction starting from the RS: simply make marginally monotonic the curve $G_{\text{RS}}(k)/k$ drawing a straight line from k_m (or, drawing the tangent of G_{RS} intercepting the origin). We will say more

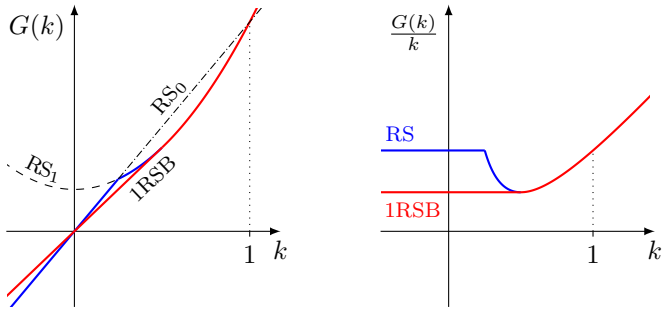

Figure 2.3: $G(k)$ (left) and $G(k)/k$ (right) for the REM in the low-temperature phase $\beta > \beta_c$, using RS and 1RSB ansätze: the 1RSB ansatz depart from the RS one from k_m.

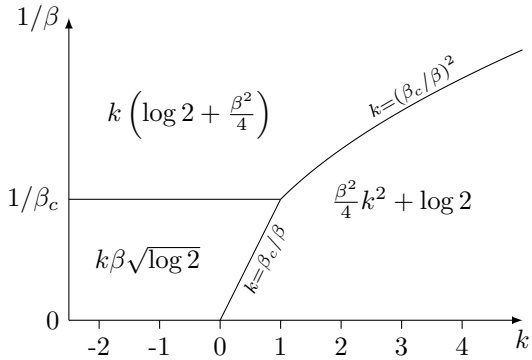

Figure 2.4: Phase diagram of the REM SCGF in the plane $k-1/\beta$ (reported from [GD89], with our notations).

on this fact, first noted by Rammal in [Ram81], in Chap. 3. Summarizing, we find

$$G_{\text{1RSB}}(k,\beta) = \begin{cases} \begin{cases} k\left(\dfrac{\beta^2}{4} + \log 2\right) & \text{if } k < \beta_c^2/\beta^2 \\ \dfrac{\beta^2}{4}k^2 + \log 2 & \text{if } k \geq \beta_c^2/\beta^2 \end{cases} & \text{if } \beta < \beta_c\,, \\ \begin{cases} k\dfrac{\beta\beta_c}{2} & \text{if } k < \beta_c/\beta \\ \dfrac{\beta^2}{4}k^2 + \log 2 & \text{if } k \geq \beta_c/\beta \end{cases} & \text{if } \beta \geq \beta_c\,. \end{cases} \tag{2.36}$$

Admittedly, the reader not accustomed with replica theory can find the continuation prescriptions we explained so far quite arbitrary. However, in many models they have been proven, with rigorous probabilistic methods, to give the correct result, in some cases at least in the limit $k \to 0$, for the SCGF looked for. In particular, the results from the REM were obtained in [GD89] via a direct evaluation of the asymptotic behaviour of the moments of the partition function, without using replica theory. For reference, we report this result, coinciding with Eq. (2.36), in Fig. 2.4.

We are now able to perform the Legendre transformation (1.28). From (2.11) (and

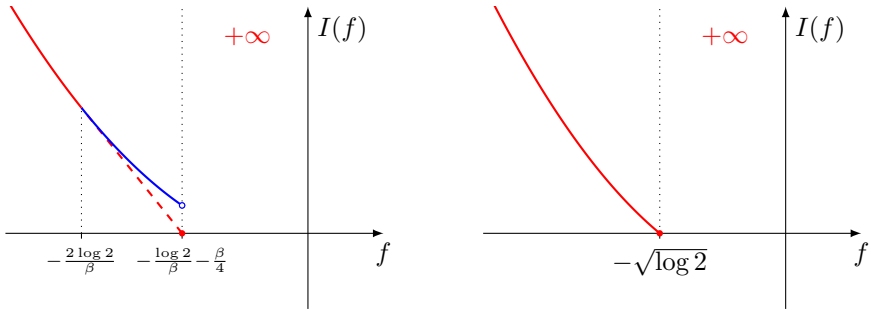

Figure 2.5: Rate function of the REM free energy. Left: high-temperature phase, $\beta < \beta_c$; the straight dashed line is the convex hull of the true rate function (continuous red and blue curve), obtained as the Legendre transformation of the non-differentiable SCGF. Right: low-temperature phase, $\beta > \beta_c$; the SCGF is differentiable and the Gärtner-Ellis theorem holds.

note 4), we can write

$$\Lambda(k,\beta) = \begin{cases} \begin{cases} \dfrac{k^2}{4} + \log 2 & \text{if } k < -4\log 2/\beta \\ -k\left(\dfrac{\beta}{4} + \dfrac{\log 2}{\beta}\right) & \text{if } k \geq -4\log 2/\beta \end{cases} & \text{if } \beta < \beta_c\,, \\ \begin{cases} \dfrac{k^2}{4} + \log 2 & \text{if } k < -2\sqrt{\log 2} \\ -k\sqrt{\log 2} & \text{if } k \geq -2\sqrt{\log 2} \end{cases} & \text{if } \beta \geq \beta_c\,. \end{cases} \tag{2.37}$$

We can perform the Legendre transformation analytically. In the high-temperature phase, due to the SCGF being non-differentiable in the point $k = -4\log 2/\beta$, we obtain a straight segment in an interval equal to the difference between the left and the right derivatives in the point. We find

$$I_\Lambda(f,\beta) = \begin{cases} f^2 - \log 2 & \text{if } f \leq -\frac{2\log 2}{\beta} \\ -\frac{4\log 2}{\beta} f - \frac{4(\log 2)^2}{\beta^2} - \log 2 & \text{if } -\frac{2\log 2}{\beta} < f < -\frac{\beta}{4} - \frac{\log 2}{\beta} \\ 0 & \text{if } f = -\frac{\beta}{4} - \frac{\log 2}{\beta} \\ +\infty & \text{if } f > -\frac{\beta}{4} - \frac{\log 2}{\beta} \end{cases} \qquad \boxed{\beta < \beta_c} \tag{2.38}$$

We report this result in Fig. 2.5 (left, red curve). We know that in this case I_Λ is only the convex hull of the true rate function, because the SCGF does not comply with the differentiability condition of the Gärtner-Ellis theorem. However, for the REM the true rate function can be obtain in a full probabilistic framework, devised in [FFM07]; it turns out that we must continue the parabolic branch up to the typical value (see Fig. 2.5, left, blue curve):

$$I(f,\beta) = \begin{cases} f^2 - \log 2 & \text{if } f < -\frac{\beta}{4} - \frac{\log 2}{\beta} \\ 0 & \text{if } f = -\frac{\beta}{4} - \frac{\log 2}{\beta} \\ +\infty & \text{if } f > -\frac{\beta}{4} - \frac{\log 2}{\beta} \end{cases} \qquad \boxed{\beta < \beta_c} \tag{2.39}$$

In the low-temperature phase, the conjunction between the RS and 1RSB branches is differentiable, so the transformation does not present marginally-convex segments. We can directly evaluate

$$I(f,\beta) = \begin{cases} f^2 - \log 2 & \text{if } f < -\sqrt{\log 2} \\ 0 & \text{if } f = -\sqrt{\log 2} \\ +\infty & \text{if } f > -\sqrt{\log 2} \end{cases} \qquad \boxed{\beta \geq \beta_c} \tag{2.40}$$

From this curves, the typical values of the free energy easily follow, as zeroes of the rate function:

$$f_{\text{typ}}(\beta) = \begin{cases} -\frac{\beta}{4} - \frac{\log 2}{\beta} & \text{if } \beta < \beta_c\,, \\ -\sqrt{\log 2} & \text{if } \beta \geq \beta_c\,. \end{cases} \tag{2.41}$$

2.4 Very large deviations and extreme value statistics

The rate functions we have written, though good rate functions in the proper mathematical definition, present a very interesting feature: both in the low- and high-temperature phases, they are equal to $+\infty$ for fluctuations above the typical value. This fact is unavoidable once the SCGF becomes linear, as the Legendre transformation of a straight line is infinite for all the values different from the slope of the line. This means that the corresponding fluctuations are *more than exponentially suppressed* in the limit of large n.

To understand why it is so, we will often resort, in the following, to an analogy from the mathematical theory of random matrices [Meh04], that we now briefly explain. Think of an ensemble of $n \times n$ matrices whose entries are i.i.d. (independent identically distributed) random variables chosen from a certain probability distribution; for example, real symmetric matrices whose independent elements are normal-distributed variables form the so-called Gaussian Orthogonal Ensemble (GOE). The term "orthogonal" is used because the probability distribution of a matrix M in the ensemble is invariant under orthogonal transformations:

$$\begin{aligned} P(M) &= \frac{1}{\mathcal{Z}_{\text{GOE}}} \prod_{i<j} \left[\mathrm{d}M_{ij}\, e^{-n M_{ij}^2} \right] \prod_i \left[\mathrm{d}M_{ii}\, e^{-\frac{n}{2} M_{ii}^2} \right] \\ &= \frac{1}{\mathcal{Z}_{\text{GOE}}} \prod_{i<j} [\mathrm{d}M_{ij}] \prod_i [\mathrm{d}M_{ii}]\, e^{-\frac{n}{2} \operatorname{tr} M^2} \\ &= P(OMO^{-1})\,, \end{aligned} \tag{2.42}$$

with O an orthogonal matrix. Though their entries are uncorrelated, the eigenvalues λ_i of these matrices are correlated random variables whose spacings follow a probability distribution that well approximates some universal properties (such as the "level repulsion") of spectra known, for example, from nuclear experiments, in such a way that the GOE matrices can be thought as the null model for the true, overly complicated and unknown nuclear Hamiltonians with the same symmetries in certain physical systems; this is the reason why this theory was initially studied by Wigner in the 1950ies. For the purposes of the current discussion, it is enough to know that the eigenvalues of the GOE matrices for large n are distributed according to the *Wigner semi-circle law*

$$p(\lambda) = \begin{cases} \dfrac{1}{n\pi} \sqrt{2n - \lambda^2} & \text{if } |\lambda| < \sqrt{2n}\,, \\ 0 & \text{if } |\lambda| > \sqrt{2n}\,, \end{cases} \tag{2.43}$$

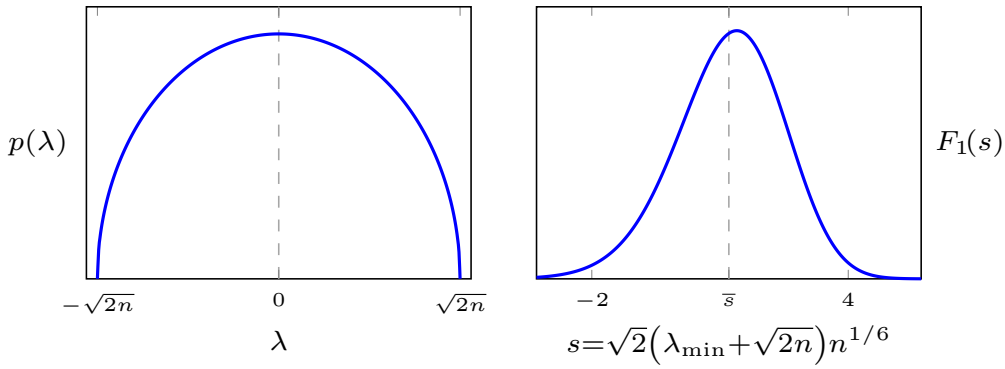

Figure 2.6: Left: Wigner semi-circle law, describing the eigenvalue distribution for the GOE ensemble. Right: Tracy-Widom distribution $F_1(s)$, describing the probability of small fluctuations (order $n^{-1/6}$) of the lowest eigenvalue $\lambda_{\min}$: the distribution is not centered in 0 due to $O(n^{-1/6})$ corrections, and drops down faster on the right.

which we plot in Fig. 2.6 (left).

However, what is the probability for the highest eigenvalue $\lambda_{\min}$, an *extreme value* of the spectrum, to fluctuate? Qualitatively speaking, there is no obstacle for $\lambda_{\min}$ to fluctuate below its typical value, so this event follows a standard LDP, with exponential suppression in n. However, in order to fluctuate above its typical value, a finite fraction of the nearest-lowest eigenvalues must move: $\lambda_{\min}$ cannot pass any other eigenvalue, because it is always defined as the minimal one. This *joint event* is heavily suppressed in probability, in such a way that a number of important consequences follows. To start with, the small fluctuations $\sqrt{2}\left(\lambda_{\min}+\sqrt{2n}\right)n^{1/6}$ does not comply with the CLT, being distributed according to the *Tracy-Widom law* we report in Fig. 2.6 (right). Note that not only the small deviation regime is $O(n^{-1/6})$, at variance with the usual Gaussian regime $O(n^{-1/2})$ for variables in the universality class of the CLT, but also their distribution is skewed, so that the probability of positive fluctuations drops down to zero faster. More relevant to the present discussion, the negative and positive large deviations are suppressed with different exponential speeds: while the negative ones present the usual scaling (1.9), the positive ones are distributed according to a LDP with an exponential speed of n^2 (see [DM06; DM08]):

$$P(\lambda_{\min}\geq t)\sim\exp\left[-n^2\Phi\left(\frac{\sqrt{2n}+t}{\sqrt{n}}\right)\right], \tag{2.44}$$

where $t\sim-O(\sqrt{n})\geq-\sqrt{2n}$ and $\Phi(y)=0$ for $y\leq 0$. In the physical literature, this feature is usually called a *very large deviation* behavior [PR10b], or *overfrustration* [DFM94].

The free energy is an example of observable distributed as an extreme value. The easiest way to see this is to take the limit $\beta\to\infty$, when it reduces to the ground-state energy. Therefore the asymmetric, anomalous form of its fluctuations can be explained from extreme value statistics. However, it is exceptionally difficult to resolve the anomalous scaling of the very large deviations in the framework of statistical physics, whose tools mainly works for ordinary extensive (or intensive) observables, with simple large-n scalings. Results from the REM come again from probability theory: for example, it is possible to prove (see [ABM04]) that the ground-state energy of the REM comply with

an asymmetric, anomalous LDP

$$p(E_0 = t) \sim \begin{cases} e^{-n(t^2 - \log 2)} & \text{if } t < -\sqrt{\log 2}\,, \\ e^{-\exp\left[n(\log 2 - t^2)\right]} & \text{if } -\sqrt{\log 2} < t < 0\,, \end{cases} \tag{2.45}$$

to be compared with Eq. (2.40). It is clear that, supposing a normal scaling on both side, the super-exponential behavior produces an infinity for fluctuations above the typical value. We will say more on this problem in the next chapter.

CHAPTER 3

The p-spin spherical model: large deviations in a magnetic field

In this chapter, we present our main original contribution on the subject, mostly drawn from [PDR19]. On the basis of what we have said so far, and following the path we traced for the REM, we apply Large Deviation Theory to evaluate the probability of fluctuations of the free energy of a celebrated model of spin glass, the p-spin spherical model. We see that the resulting LDP has a lot of features in common with the REM's one, anomalous asymmetric scaling included. Moreover, with a little effort we are able to explore the fate of the very large deviations whenever a magnetic field is switched on.

3.1 The model

The p-spin glass spherical model consists of a p-body interaction of n continuous spins with the following Hamiltonian:

$$H_p = - \sum_{1 \le i_1 < i_2 < \cdots < i_p \le n} J_{i_1 \cdots i_p} \sigma_{i_1} \sigma_{i_2} \cdots \sigma_{i_p} - h \sum_{i=1}^{n} \sigma_i \,, \tag{3.1}$$

where h represents an external magnetic field coupled with the spins, the J-couplings are independent quenched random variables normally distributed with zero mean and variance

$$\overline{J^2_{i_1 \cdots i_p}} = \frac{J^2 p!}{2n^{p-1}} \,, \tag{3.2}$$

while the spins are real variables with range in $(-\infty, \infty)$ subject to a global spherical constraint such that the measure is

$$\mathrm{Tr}_\sigma \equiv 2\sqrt{n} \int_{-\infty}^{\infty} \prod_{i=1}^{n} d\sigma_i \, \delta \left(\sum_{i=1}^{n} \sigma_i^2 - n \right) . \tag{3.3}$$

These scalings guarantee the extensivity of the free energy.

The thermodynamics of this model was studied in the seminal work [CS92] by Crisanti and Sommers (CS); the special case $p = 2$, which presents a rather different phenomenology, was introduced before in [KTJ76]. We dedicate the remaining of this section to report some of the knowledge on its non-trivial behavior, which motivated its fruitful applications as a toy model of supercooled liquids and structural glasses; for further references, see [KPA93; Bar97; CC05; Zam14].

As we will see explicitly in the following, for $p \ge 3$ the spherical model typically exhibits a one-step replica symmetry breaking phase transition for small values of temperature and magnetic field, similar to the one we observed in Chap. 2 for the REM, and

corresponding to the continuous line in the T-h plane we report in Fig. 3.1. At variance with the REM, however, the parameters representing the elements of the replica matrix

$$q_{ab} = \frac{1}{n}\sum_{i=1}^{n} \sigma_{i,a}\sigma_{i,b}\,, \tag{3.4}$$

which in the 1RSB ansatz (2.32) are q_0 and $q_1 \geq q_0$, are continuous variables in $[0,1]$. At zero external magnetic field, the critical temperature is given by

$$T_c(p; h=0) = 1/\beta_c(p; h=0) = y\sqrt{\frac{p}{2y}}(1-y)^{p/2-1}\,, \tag{3.5}$$

with y solution of the equation

$$\frac{2}{p} = -\frac{2y(1-y+\log y)}{(1-y)^2}\,. \tag{3.6}$$

At the transition $T_c = T_c(h)$, the combination $q_1 - q_0$ jumps from zero (in the paramagnetic phase) to a finite value discontinuously up to a certain critical value h_c of the external field; for $h > h_c$, the transition becomes continuous in $q_1 - q_0$, until it eventually disappears for higher values of the field.

What if we tried to solve the problem using the RS ansatz? In this model, where the parameters of the replica matrix are real variables, the extremization procedure we explained in Chap. 2 to select their optimal values becomes a saddle-point analysis: the integral

$$\int_{\mathbf{q}} \left[\prod_{a<b} \mathrm{d}q_{ab}\right] e^{-ng[\mathbf{q}]} \sim e^{-ng[\mathbf{q}^\star]} \tag{3.7}$$

for large n is dominated by the stationary point $\mathbf{q}^\star$,

$$\left.\frac{\partial g}{\partial q_{ab}}\right|_{\mathbf{q}=\mathbf{q}^\star} = 0\,, \tag{3.8}$$

corresponding to a minimum of the function g; this is simply the continuous version of Eq. (2.19). In order to test the legitimacy of an ansatz $\tilde{\mathbf{q}}$ on the form of $\mathbf{q}$, it is mandatory to verify its *stability* in the space of all the possible replica matrices, meaning that one has to be sure not only that the chosen restriction corresponds to a stationary point for generic variations of the parameters q_{ab}, but also that this point is a true minimum, for example verifying the Hessian $\partial^2 g/\partial q_{ab}\partial q_{cd}$ to be positive-definite. If this operator happens to have a negative eigenvalue in $\tilde{\mathbf{q}}^\star$, the corresponding ansatz $\tilde{\mathbf{q}}$ must be discarded as unstable. We call this kind of pathological behavior a *de Almeida-Thouless (dAT) instability* [AT78], from the names of the authors who first observes this phenomenon in the SK model. For $p \geq 3$, the RS ansatz becomes unstable at a certain temperature $T_{\mathrm{dAT}}(h)$, which in Fig. 3.1 is represented by a dashed line for $h < h_c$, and coincides with the line $T_c(h)$ for $h > h_c$. The interesting point to note is that there is a region between the lines $T_{\mathrm{dAT}}(h)$ and $T_c(h)$ where both the ansätze, the RS and the 1RSB, are stable in the de Almeida-Thouless sense, but still the 1RSB is the global minimum (the RS being only a local one). However, in our large deviation approach we will see that, in order to obtain the typical value in the RS ansatz, taking the limit (2.8), even in this range of temperatures we would have to follow a non-convex SCGF: the dAT stability is only a necessary condition that the continuation procedure we use to extrapolate the typical and non-typical behavior of the model has to comply with.

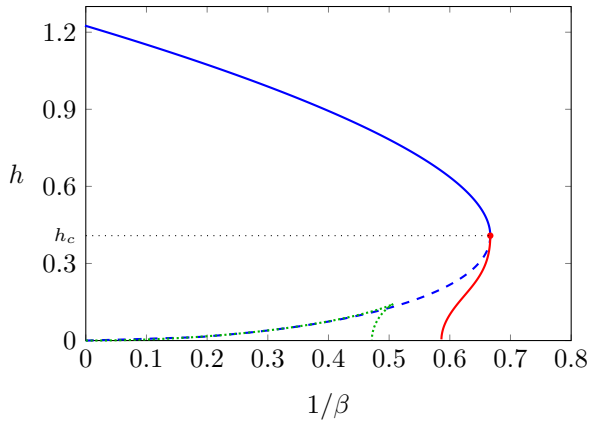

Figure 3.1: Phase diagram for the typical behavior of the spherical ($p = 3$)-spin, from [CS92]. At the blue line (continuous and dotted) the RS solution for the saddle point becomes unstable, in the dAT sense (at $h = 0$, the only RS solution which is stable for all the values of temperature is the paramagnetic one, with $q = 0$). The continuous blue line, for $h > h_c$, corresponds to a 1RSB phase transition with $q_1 - q_0$ starting from 0 and varying smoothly. The red line, for $h < h_c$, corresponds to a 1RSB phase transition with $q_1 - q_0$ jumping from 0 to a finite value discontinuously at the transition. The green dotted line bounds a region where other solutions of the RS saddle point equations appear discontinuously, one of which corresponds to a global minimum restricted space of RS matrices, but is unstable, because it is a saddle in the full space of replica matrices (the Hessian for generic variations of the parameters develops a negative eigenvalue). For $p = 3$, $T_c(3; h = 0) \approx 0.59$.

3.1.1 Complexity

So far, we only reported results from the typical equilibrium ("static") behavior of the p-spin spherical model. However, most of its success as a toy model for structural glasses and other complex systems is due to the very non-trivial phenomena that happens already above the critical temperature T_c, where one naively would expect to observe a simple paramagnet. These phenomena can be investigated with different approaches, such as

- the analysis of the *TAP equations* [TAP77; CC05], a system in the local magnetizations whose solutions correspond to the metastable states of the system;
- the study of the Langevin dynamics of the model, which can be highly non-trivial when many metastable states are present, exhibiting a regime where the system does not relax to its equilibrium state described by thermodynamic, remaining instead trapped in one of the many metastable states due to ergodicity breaking of the phase space.

We will not give the details of these approaches: the interested reader can find them in reviews like [CC05]. Here, we just describe Fig. 3.2, reporting the free energy of various states of the system as a function of temperature. Starting from the high-temperature phase and cooling the system, we find at least three significant values of T above the 1RSB static transition T_c:

- For $T > T_{\text{TAP}}$, the only solution of the saddle point and TAP equations is the paramagnetic one, with $q = 0$.

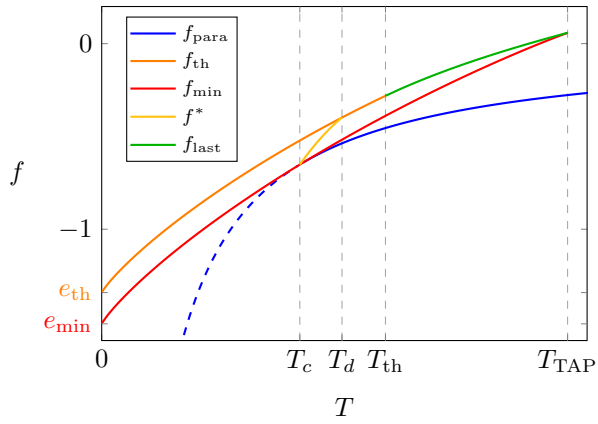

Figure 3.2: Free energies as functions of the temperature for the metastable states of the $(p = 10)$-spin spherical model for zero external magnetic field, from a TAP and a dynamical analyses. $f_{\rm para}$ is the paramagnetic ($q = 0$) free energy, which is the equilibrium value above the static phase transition T_c (continuous line), and a non-optimal stable solution of the RS saddle point equation below (dashed line). Below $T_{\rm TAP}$, well above the static critical temperature T_c, and so in the naively expected paramagnetic phase, non-trivial solutions of the TAP equations emerge: these solutions represent metastable states with free energy higher than the paramagnetic one, for all the values between the lines $f_{\rm min}$ and the threshold $f_{\rm th}$. These states are labeled by their ground-state energy $e = f(0) \in [e_{\rm min}, e_{\rm th}]$ and their free energy can be followed increasing T, tracing a line between $f_{\rm min}(T)$ and $f_{\rm th}(T)$ up to the green line $f_{\rm last}(T)$, where they disappear. The yellow line f^* between T_c and T_d is the free energy of the states whose exponential number makes their total free energy equal to the paramagnetic one. Note that this graph, at variance with the one in [Zam14], is quantitatively precise.

- For $T_d < T < T_{\rm TAP}$, solutions of the TAP equation with $q \neq 0$ appear. However, the dominant state in the thermodynamic limit is still the paramagnetic one, which has lower free energy.
- For $T_c < T < T_d$, there is a value of the free energy (f^* in Fig. 3.2) corresponding to a number of metastable states so large that, even though each one of them has free energy higher than the paramagnetic one, their presence generates an additional entropic contribution balancing this difference. This entropic contribution Σ (that is, the logarithm of the number of these metastable states) is called *complexity*. In this phase there is no thermodynamic phase transition yet, but due to the degeneracy between the paramagnetic free energy and the total free energy of the metastable states ($f - T\Sigma$), the dynamics does not converge to equilibrium and ergodicity is broken.
- For $T < T_c$, we are in the 1RSB phase we described above and the complexity vanishes.

The same rich phenomenology, with the insurgence a large number of metastable states producing extreme consequences on the dynamical properties, is known to occur in many other complex systems, making the p-spin spherical model a valuable tool of investigation even in more realistic settings.

In the following, we will only study the large deviations of the equilibrium free energy; however, the fluctuations associated to the complexity states and their relevance

for the dynamics have been recently investigated in [BKL19; FR20].

3.2 Replicated partition function

Our aim is to analyze the large deviations of the free energy of this model. Following CS for reference, we start writing the partition function

$$\mathcal{Z}_{J,n} = \mathrm{Tr}_\sigma \exp\left[-\beta H_p\right] . \tag{3.9}$$

Introducing k replicas and performing the Gaussian integration over the disorder, it is easy to obtain

$$\overline{(\mathcal{Z}_n)^k} = \mathrm{Tr}_\sigma \exp\left[n \frac{(\beta J)^2}{4} \frac{p!}{n^p} \sum_{a,b} \sum_{1\le i_1<i_2<\cdots<i_p\le n} \sigma_{i_1,a}\sigma_{i_1,b}\cdots\sigma_{i_p,a}\sigma_{i_p,b} + \beta h \sum_{i=1}^{n}\sum_{a=1}^{k}\sigma_{i,a}\right] . \tag{3.10}$$

The restricted sum over $i_1 < i_2 < \cdots < i_p$ can be written as

$$p! \sum_{1\le i_1<i_2<\cdots<i_p\le n} = \sum_{i_1\neq i_2\neq\cdots\neq i_p=1}^{n} = \sum_{i_1,i_2,\cdots,i_p=1}^{n} + \cdots , \tag{3.11}$$

where the $\cdots$ are for terms suppressed for large n with respect to the first. Indeed, the factorial accounts for the order, and to reconstruct the sum with indices all different we have to subtract the cases when some indices are repeated; for example, the first correction is for only 2 indices repeated:

$$\sum_{i_1\neq i_2\neq\cdots\neq i_p=1}^{n} = \sum_{i_1,i_2,\cdots,i_p=1}^{n} - \frac{p(p-1)}{2} \sum_{i_1=i_2\neq i_3,\cdots,i_p=1}^{n} + \cdots , \tag{3.12}$$

and the second term, which is of order n^{p-1}, is suppressed with respect to the first one, of order n^p. We can insert the positive-definite replica matrix $\mathbf{q}$, with elements q_{ab} defined in Eq. (3.4), using the delta-function identities

$$\prod_{a<b}\delta\left(nq_{ab} - \sum_{i=1}^{n}\sigma_{i,a}\sigma_{i,b}\right) = \int_{-i\infty}^{+i\infty}\left[\prod_{a<b}\frac{\mathrm{d}\lambda_{ab}}{2\pi i}\right] e^{-\frac{1}{2}\sum_{a\neq b}\lambda_{ab}\left(nq_{ab}-\sum_{i=1}^{n}\sigma_{i,a}\sigma_{i,b}\right)} , \tag{3.13}$$

$$\prod_{a}\delta\left(n - \sum_{i=1}^{n}\sigma_{i,a}^2\right) = \int_{-i\infty}^{+i\infty}\left[\prod_{a}\frac{\mathrm{d}\lambda_{aa}}{4\pi i}\right] e^{-\frac{1}{2}\sum_{a}\lambda_{aa}\left(nq_{aa}-\sum_{i=1}^{n}\sigma_{i,a}^2\right)} , \tag{3.14}$$

with $q_{aa} = 1$, obtaining

$$\overline{(\mathcal{Z}_n)^k} = \int_{\mathbf{q}>0}\left[\prod_{a<b}\mathrm{d}q_{ab}\right]\int_{-i\infty}^{+i\infty}\left[\prod_{a<b}\frac{n\,\mathrm{d}\lambda_{ab}}{2\pi i}\right]\int_{-i\infty}^{+i\infty}\left[\prod_{a}\frac{\sqrt{n}\,\mathrm{d}\lambda_{aa}}{2\pi i}\right] e^{-ng[\mathbf{q},\boldsymbol{\lambda}]} , \tag{3.15}$$

with

$$g[\mathbf{q},\boldsymbol{\lambda}] = -\frac{(\beta J)^2}{4}\sum_{a,b}q_{ab}^p + \frac{1}{2}\sum_{a,b}\lambda_{ab}q_{ab} - \log\int_{-\infty}^{+\infty}\left[\prod_{a}\mathrm{d}\sigma_a\right] e^{\frac{1}{2}\sum_{a,b}\lambda_{ab}\sigma_a\sigma_b+\beta h\sum_a\sigma_a} . \tag{3.16}$$

Note that λ_{ab} and λ_{aa} are the Lagrange multipliers enforcing, respectively, the definition of the replica matrix (3.4) and the spherical constraint (3.3). The remaining spin integration is Gaussian, so

$$g[\mathbf{q},\boldsymbol{\lambda}] = -\frac{(\beta J)^2}{4}\sum_{a,b=1}^{k} q_{ab}^p + \frac{1}{2}\sum_{a,b=1}^{k}\lambda_{ab}q_{ab} + \frac{1}{2}\log\det\left(-\boldsymbol{\lambda}\right) + \frac{(\beta h)^2}{2}\sum_{a,b=1}^{k}\left(\boldsymbol{\lambda}^{-1}\right)_{ab} - \frac{k}{2}\log(2\pi)\,. \tag{3.17}$$

In the following, we will depart from [CS92], from which the calculation so far are taken, by keeping finite the number of replicas k, in order to obtain the full SCGF of the free energy, instead of only its typical value.

3.3 From replicas to the scaled cumulant generating function

We start our analysis with the case of null magnetic field $h = 0$, when the calculations simplify a lot. In this case, the $\boldsymbol{\lambda}$ integration is easily obtained for large n, when the saddle-point simply gives $-\boldsymbol{\lambda}^{-1}_{ab} = q_{ab}$. Accordingly, the partition function is:

$$\overline{\mathcal{Z}_n^k} = \int_{\mathbf{q}>0}\prod_{a<b}\sqrt{\frac{n}{2\pi}}\,\mathrm{d}q_{ab}\,e^{-ng(\mathbf{q})}\,, \tag{3.18}$$

where

$$g(\mathbf{q}) = -\frac{(\beta J)^2}{4}\sum_{a,b=1}^{k} q_{ab}^p - \frac{1}{2}\log\det\mathbf{q} - ks(\infty)\,. \tag{3.19}$$

and $s(\infty) = [1+\log(2\pi)]/2$ is the entropy density in the infinite temperature limit. To evaluate the integrals on q_{ab} we use the saddle-point method together with the 1RSB ansatz, which is formulated in terms of the three parameters (q_1, q_0, m):

$$q_{ab} = (1-q_1)\delta_{ab} + (q_1-q_0)\epsilon_{ab} + q_0 \tag{3.20}$$

with ϵ_{ab} defined as

$$\epsilon_{ab} = \begin{cases} 1 & \text{if } a,\, ba \text{ are in a diagonal } m\times m \text{ block,} \\ 0 & \text{otherwise.}\end{cases} \tag{3.21}$$

The eigenvalues of $\mathbf{q}$, with the respective degeneracies, are

$$\begin{aligned} \eta_0 &= 1-q_1 && \text{deg.} = k(m-1)/m \\ \eta_1 &= 1-(1-m)q_1 - mq_0 && \text{deg.} = k/m-1 \\ \eta_2 &= 1-(1-m)q_1-(m-k)q_0 && \text{deg.} = 1 \end{aligned} \tag{3.22}$$

Using this and inserting the ansatz (3.20) in (3.19) we find

$$\begin{aligned} g(k;q_0,q_1,m) = &-\frac{(\beta J)^2}{4}k\left[1+(m-1)q_1^p+(k-m)q_0^p\right] - \frac{k(m-1)}{2m}\log\left(\eta_0\right) \\ &- \frac{k}{2m}\log\left(\eta_1\right) - \frac{1}{2}\log\left(1+\frac{kq_0}{\eta_1}\right) - ks(\infty)\,. \end{aligned} \tag{3.23}$$

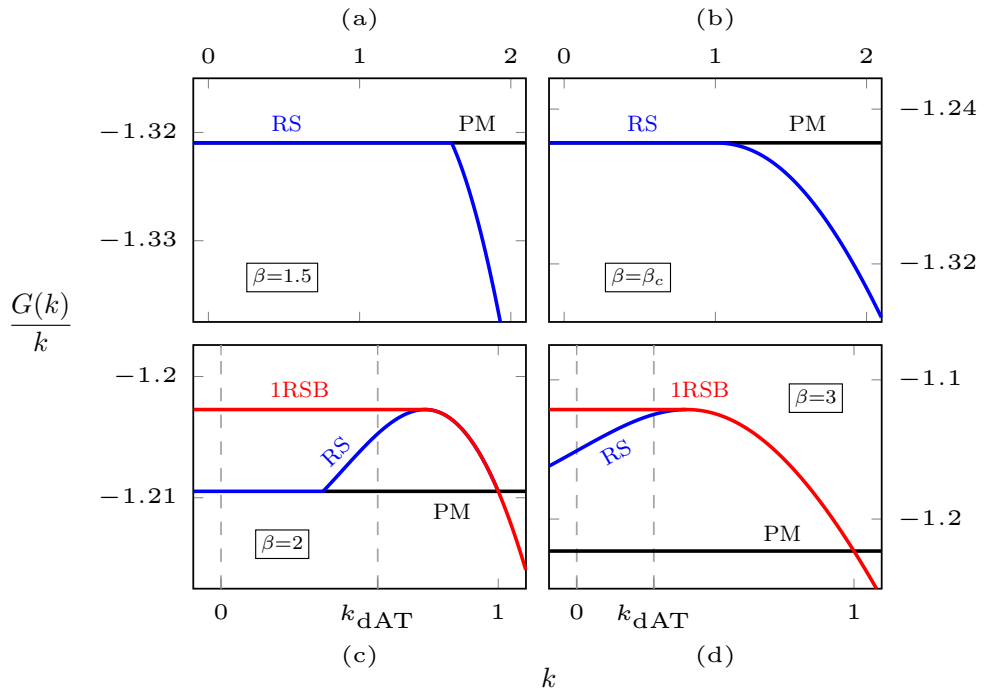

Figure 3.3: The function $G(k)/k$ for the ($p = 3$)-spin in zero external magnetic field, for different values of β. (a) At high temperature ($\beta = 1.5$) the 1RSB anstatz coincides with the RS one (blue curve); the solution joins the paramagnetic line (in black) in a point $k_c > 1$, where the function is not differentiable. (b) At $\beta = \beta_c \approx 1.706$, the junction is in $k_c = 1$ and becomes smooth. For $\beta = 2$ (c) and $\beta = 3$ (d), the 1RSB solution (red curve) departs from the RS one and becomes a straight line for all the $k < k_c$, which is the point where the RS function loses its monotonicity. The critical value k_c approaches zero for $\beta \to \infty$. (c)-(d) note that k_{dAT}, the point where the RS solution becomes unstable in the de Almeida-Thouless sense, does not coincides with k_c, and $k_{\text{dAT}} < k_c$. Figure from [PDR19].

This functional is evaluated numerically at the saddle-point $(q_1^\star, q_0^\star, m^\star)$ for the 1RSB parameters for each value of k. The three parameters take values in the domains $q_1 \in [0, 1]$, $q_0 \in [0, q_1]$, $m \in [1, k]$ (if $k > 1$) or $m \in [k, 1]$ (otherwise), and for $k < 1$ the saddle point is obtained with a maximization of the functional instead of a minimization, as usual in replica theory. Using Eq. (2.11), we obtain a SCGF $\Lambda(k)$ which becomes linear above a certain value $k = k_c$, depending on temperature. To ease the visualization of this feature, in Fig. 3.3 we plot the function $G(k)/k = g(k; q_1^\star, q_0^\star, m^\star)/(k\beta)$ which, when $\Lambda(k)$ is linear, intersects the vertical axis in f_{typ}. The figure does not change qualitatively for $p \geq 3$.

The $p = 2$ case at low temperature is different: the 1RSB ansatz reduces to the RS one (that is, $\bar{q}_1 = \bar{q}_0$) as long as $k \geq 0$, therefore the typical values of all the thermodynamic quantities are obtained under the RS ansatz [KTJ76]. On the opposite, for $k < 0$ we need to introduce again the 1RSB ansatz which, as in the $p \geq 3$ case, gives the linear behavior of the SCGF. In other words, $k_c = 0$ for the 2-spin spherical model for all $\beta > \beta_c$.

3.3.1 Rammal's construction

Before turning to the evaluation of the rate function, we discuss an interesting geometrical interpretation of the SCGF shape, which we mentioned briefly in the previous chapter. To this aim, let us consider the RS ansatz (that is, Eq. (3.23) with $q_1 = q_0 = q$ and $m = 1$). As we can see in Fig. 3.3, the RS solution (blue curve) is non-monotonic for $\beta > \beta_c$. On the other hand, we know that $G(k)/k$ has to be a monotonic quantity, therefore the RS solution can be ruled out. We can check that the 1RSB solution gives a perfectly fine monotonic $G(k)/k$ (red curve in Fig. 3.3), as one could expect due to the fact that this ansatz gives the correct typical free energy for this model. Interestingly, however, exactly the same monotonic curve can be obtained by using a much simpler geometric construction: just consider the RS solution, which is the right one for large k (remember, indeed, that the RS ansatz gives the correct solution for all the integer points $k \geq 1$), and when $G(k)/k$ starts to be non-monotonic continue with a straight horizontal line (in the $G(k)/k$ vs k plot).

This construction actually dates back to Rammal [Ram81] and can be found in [Kon83] (similar considerations in [OK04; NH08; NH09]). We dedicate the rest of this section to discuss it. We reproduce here the reasoning not only as an historical curiosity: first of all, we see it as an enlightening approach to the problem of the continuation of the replicated partition function to real number of replicas, particularly suitable for a finite k analysis. Moreover, we note that this interpretation, whenever it works, gives a flavor of "uniqueness" (though not in a strict mathematical sense) to the resulting solution, being based only on the properties of convexity and extremality that the function $\Lambda(k)$ must have. In this respect, a generalization of this result would be of great interest in order to better understand the necessity of Parisi hierarchical RSB procedure, which has been dubbed as "magic" even in relatively recent works, like [Dot11]; however, a true geometrical interpretation of the full machinery of RSB, beyond the simple case considered here, still lacks. Finally, in the context of this paper we are able to show a case where the construction gives the correct answer (the p-spin spherical model at zero external magnetic field) and a case where it fails (when the field is switched on).

We start claiming that $G(k)/k$ obtained by using the 1RSB ansatz or the Rammal construction are the same because of the following facts:

i) for $k > k_c$ the 1RSB and RS ansätze coincide ($q_1^\star = q_0^\star = q \neq 0$) and k_c is exactly the point where $G(k)/k$ is not monotonic anymore if one uses the RS ansatz;

ii) from the saddle point equations obtained by extremizing Eq. (3.23) when $k < k_c$, one obtains $q_0^\star = 0$;

iii) the remaining saddle point equations fix q_1 and m, and one can see that these equations are identical to those needed to perform the Rammal construction, which fix the point k_c and the parameter of the RS ansatz q.

In the following, we will prove this claim.

We already know, from Chap. 1, some important properties of the SCGF, derived in full generality using its definition only. In particular, we know that $\Lambda(k)$ must be a convex function of k (and so $G(k)$ a concave function, because of Eq. (2.11)) and that $G(k)/k$ must be monotonic (properties (iv) and (v)). Given that, the explicit evaluation of G is performed for each system within replica theory: an ansatz is imposed on the form of the replica overlap matrix, the number of replicas k is then continued from integer to real values, the corresponding $G(k)$ is evaluated with the saddle-point method for large n and finally a check is performed *a posteriori* to verify its validity. In the SK model,

the system originally considered by Rammal, at low temperatures the replica symmetric ansatz, which still gives the correct values of the positive integer momenta of the partition function, fails to produce a sensible solution for the SCGF at $k < 1$, in at least three way:

- it becomes unstable under variations around the saddle point (de Almeida-Thouless instability [AT78]) below $k = k_{\text{dAT}}$;
- it produces a $G(k)$ that is non-concave (and so a non-convex $\Lambda(k)$) around $k = k_{\text{conv}}$, meaning that $G''(k)$ changes sign at k_{conv};
- it produces a $G(k)/k$ that loses monotonicity a $k = k_m$.

In the SK model k_{dAT} is the largest ($k_{\text{dAT}} > k_m > k_{\text{conv}}$), and so it is the first problem one encounters in extrapolating the RS solution from integer values of k. However, from the point of view of convexity and monotonicity alone, Rammal proposed to build a marginally monotone $G(k)/k$ in a minimal way, starting from the RS and simply keeping it constant below k_m at the value $G(k_m)/k_m$. While the resulting function is not the correct one for the SK model, which needs a full RSB analysis to be solved, surprisingly enough for the spherical p-spin in zero magnetic field (and also for the REM, as we know) this approach reproduces the solution obtained with a 1RSB ansatz with $q_0 = 0$ (see Fig. 3.3). Notice that in the present model the RS solution suffers from the same inconsistencies as in the SK model, but now k_m is the largest of the three problematic points.

To convince the reader that the two approaches are actually equivalent we prove, as final part of this section, that without an external magnetic field the 1RSB solution of the spherical p-spin and the Rammal construction coincide. In order to obtain this result, we have to prove that:

- the 1RSB solution for $G(k)/k$ becomes a constant below $k = k_c$, which is defined as the point where the RS and 1RSB ansätze branch out, as we did above;
- this constant is the same as the one in the Rammal construction, that is $G(k_m)/k_m$;
- the points k_c and k_m are the same.

As k_c is the point where the RS solution is not optimal anymore, for $k < k_c$ we have $q_0^\star = 0$, as discussed in [CS92]. Let us now consider Eq. (3.23) with $q_0 = 0$: differentiating with respect to q_1 and m and setting the results equal to 0 we get the equations for $q_1^\star$ and $m^\star$, which read[1]

$$\begin{cases} \mu\, q_1^{\star p-2} - \dfrac{1}{(1-q_1^\star)[1-(1-m^\star)q_1^\star]} = 0 \\ \dfrac{\mu}{2} m^{\star 2} q_1^{\star p} - \dfrac{1}{2}\log\left(1 + \dfrac{m^\star\, q_1^\star}{1-q_1^\star}\right) + \dfrac{m^\star}{2}\dfrac{q_1^\star}{1-(1-m^\star)q_1^\star} = 0 \end{cases} \tag{3.24}$$

where $\mu = p(\beta J)^2/2$. These equations can be solved numerically (as we did to obtain the plots in the main text), but to show our point here we do not really need the explicit solution. Indeed it is enough to notice that $m^\star$ and $q_1^\star$ do not depend on k and therefore $g(k; 0, q_1^\star, m^\star)/k$ is a constant. Then, we need to check that it is the same constant as the

[1] In general, it is not true that the solution of the saddle point equations for a functional $g(q_0, q_1, m)$ with respect to all its arguments is equal to the solution of the saddle point equations for $g(q_0^\star, q_1, m)$ with respect to q_1, m. In this case it is true, as the reader can directly check.

one obtained by Rammal. Again starting from Eq. (3.23), by putting $q_1 = q_0 = q$ we obtain the RS solution, which is

$$\begin{aligned} g_0(k;q) = &- \frac{(\beta J)^2}{4}\left[k + k(k-1)q^p\right] - \frac{k-1}{2}\log(1-q) \\ &- \frac{1}{2}\log\left[1-(1-k)q\right] - ks(\infty). \end{aligned} \tag{3.25}$$

In this case, extremizing with respect to q, we have an equation which gives the RS solution on the saddle point, $q^\star$. To find k_m, we then require $\frac{\partial}{\partial k} g_0/k = 0$. The two resulting equations are:

$$\begin{cases} \mu\, q^{\star p-2} - \dfrac{1}{(1-q^\star)\left[1-(1-k_m)q^\star\right]} = 0 \\ \dfrac{\mu}{2} k_m^2 q^{\star p} - \dfrac{1}{2}\log\left(1 + \dfrac{k_m\, q^\star}{1-q^\star}\right) + \dfrac{k_m}{2}\dfrac{q^\star}{1-(1-k_m)q^\star} = 0 \end{cases} \tag{3.26}$$

that are exactly Eqs. (3.24) with k_m instead of $m^\star$ and $q^\star$ instead of $q_1^\star$. Therefore $k_m = m^\star$ and $q^\star = q_1^\star$ and one can check that

$$\frac{g(k;0,q^\star,k_m)}{k} = \frac{g_0(k_m,q)}{k_m}\,. \tag{3.27}$$

It only remains to prove that k_c and k_m, which in general can be different points, are actually the same. As the 1RSB ansatz gives the correct solution for the present model, the corresponding SCGF must be convex and thus, in particular, continuous. The only way to obtain a continuous function which is equal to the RS one above k_c and to the Rammal's constant below, is to take $k_c = k_m$, and so the two functions coincide everywhere.

3.4 Rate function and very large deviations

Starting from the SCGF evaluated in the last section, we perform a numerical Legendre transformation to obtain the rate function according to Eq. (1.28). The result is shown in Fig. 3.4 for different values of β. The rate function displays the following behavior:

- for $f = f_{\text{typ}}$, it is null as expected;
- for $f < f_{\text{typ}}$, $I(f)$ is finite, indicating that a regular large deviation principle holds for fluctuations below the typical value. When $\beta > \beta_c$ the SCGF is smooth, so we obtain the rate function via the Gärtner-Ellis theorem. On the other hand, when $\beta < \beta_c$ the SCGF is not differentiable in a point (see Fig. 3.3), so we are only able to obtain the convex hull of the rate function (see Fig. 3.4);
- for $f > f_{\text{typ}}$, $I(f) = +\infty$. This is due to the linear behavior of the SCGF below k_c discussed in the previous section and it is a signature (as for the REM) of an anomalous scaling with n of the rare fluctuations above the typical value.

An ambitious goal would be the identification of the correct behavior with n of these very large deviations. Indeed, a more general way of stating a large deviation principle is

$$P(f_n \in [x, x+\mathrm{d}x]) \sim \begin{cases} e^{-a_n I_-(x)}\,\mathrm{d}x & \text{if } x \le f_{\text{typ}}\,, \\ e^{-b_n I_+(x)}\,\mathrm{d}x & \text{if } x > f_{\text{typ}}\,, \end{cases} \tag{3.28}$$

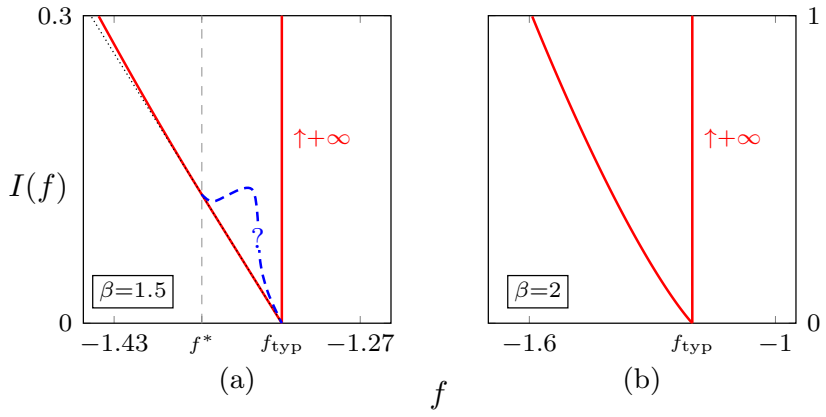

Figure 3.4: Rate function of the free energy for the ($p = 3$)-spin in zero external magnetic field, for different values of β. The fluctuations above the typical value correspond to the linear part of the SCGF, so that the Legendre transformation gives an infinite rate function. The fluctuations below the typical value are described by the branch in red. For $\beta = 1.5 < \beta_c$ (a), as the SCGF is not differentiable, we obtain only the convex-hull of the true rate function; in the interval $[f^*, f_{\rm typ}]$, where our result gives a straight segment (the part of the curve overlapping the dotted line), the true, unknown rate function is represented by the curve in blue. For $\beta = 2 > \beta_c$ (b) the SCGF is smooth and the Gärtner-Ellis theorem applies. Figure from [PDR19].

where $a_n, b_n \to \infty$ when $n \to \infty$. In other words, the fluctuations resulting in values of x lower than $f_{\rm typ}$ are given by the rate function $I_-(x)$, while those resulting in values larger than $f_{\rm typ}$ have rate function $I_+(x)$, but with different scalings a_n, b_n. In our case, we have $a_n \sim n$, then the rate function defined in Eq. (1.9), which assumes a fixed scaling n for both sides, can be written as

$$I(x) \sim \begin{cases} I_-(x) & \text{if } x \le f_{\rm typ}\,, \\ \frac{b_n}{n} I_+(x) & \text{if } x > f_{\rm typ}\,, \end{cases} \tag{3.29}$$

with $b_n/n \to \infty$. For this reason, fluctuations above the typical value are referred to as "very large deviations". The physical explanation of the substantial difference in scaling of the deviations of thermodynamic quantities below and above their typical values resides in the different number of elementary degrees of freedom involved to obtain the corresponding fluctuation: while in the first case it is sufficient that only one of the elementary variables assumes an anomalous value below its typical, the others being fixed, in the second case all the variables have to fluctuate, a joint event with probability heavily suppressed with respect to the first one.

This argument shows the importance of the resolution of the anomalous scaling behavior leading to the very large deviations we explained above. In general, however, although the Gärtner-Ellis theorem can be extended to find rate functions for large deviation principles with arbitrary speed a_n, b_n, we lack techniques to compute the asymptotic scaling of a_n and b_n for large n, because of additional inputs needed to calculate the corresponding SCGF with a saddle-point approximation (for some other systems this problem has been solved with ad-hoc methods [ABM04; DM08], while in [PR10b] a method is proposed in the context of the SK model).

In the next section we present the main result of our work, which could be useful to study this anomalous kind of fluctuations also in other problems: through an extension

of the replica calculation to the case with an external magnetic field, we are able to numerically check that the very large deviation effect disappears. More in detail, we obtain that with a magnetic field, no matter how small, not only $a_n \sim n$ as before, but also $b_n \sim n$.

3.5 Large deviations of the p-spin model in a magnetic field

In this section we generalize the previous discussion to the case of non-zero magnetic field. The computation of the SCGF at $h \neq 0$ goes beyond the approach of the work by Crisanti and Sommers, who only considered the typical case. In contrast to the problem with $h = 0$, where the finite-k calculation consists of a quite straightforward generalization of the standard one, here a more substantial effort is needed to extend the $k = 0$ result.

The starting point is Eq. (3.17). In the presence of a magnetic field, the saddle-point integration in the $\boldsymbol{\lambda}$-variables is not straightforward as to obtain (3.23). Derivation with respect to λ_{ab} leads to the following saddle-point equations:

$$q_{ab} + \left(\boldsymbol{\lambda}^{-1}\right)_{ab} - (\beta h)^2 \sum_{c,d=1}^{k} \left(\boldsymbol{\lambda}^{-1}\right)_{ca} \left(\boldsymbol{\lambda}^{-1}\right)_{bd} = 0\,, \tag{3.30}$$

where we have used the identity:

$$\frac{\partial \left(\boldsymbol{\lambda}^{-1}\right)_{cd}}{\partial \lambda_{ab}} = -\left(\boldsymbol{\lambda}^{-1}\right)_{ca} \left(\boldsymbol{\lambda}^{-1}\right)_{bd}\,. \tag{3.31}$$

Equations (3.30) are solved via successive contractions of the replica indices: a double summation over a, b leads to an equation for the scalar $\sum_{a,b} \left(\boldsymbol{\lambda}^{-1}\right)_{ab}$ with solutions:

$$\sum_{a,b=1}^{k} \left(\boldsymbol{\lambda}^{-1}\right)_{ab} = \frac{1 \pm \sqrt{1 + 4(\beta h)^2 q_s}}{2(\beta h)^2} \equiv l_\pm\,, \qquad q_s = \sum_{a,b=1}^{k} q_{ab}\,. \tag{3.32}$$

Similarly, a single contraction gives:

$$\sum_{a=1}^{k} \left(\boldsymbol{\lambda}^{-1}\right)_{ab} = -\frac{\sum_a q_{ab}}{1 - (\beta h)^2 l_\pm}\,, \tag{3.33}$$

and finally

$$\left(\boldsymbol{\lambda}^{-1}\right)_{ab} = -q_{ab} + \frac{(\beta h)^2 \sum_c q_{ca} \sum_d q_{db}}{\left[1 - (\beta h)^2 l_\pm\right]^2}\,. \tag{3.34}$$

Given the 1RSB ansatz (3.20), q_{ab} has k elements 1 on the diagonal, $m(m-1)k/m$ elements q_1 in the internal blocks, the remaining $k^2 - k - k(m-1)$ elements q_0, so

$$q_s = k + k(m-1)q_1 + k(k-m)q_0 = k\eta_2 \tag{3.35}$$

Every row (column) contains the same elements, so

$$q_r \equiv \sum_{b=1}^{k} q_{ab} = 1 + (m-1)q_1 + (k-m)q_0 = \eta_2 \qquad \forall\, a\,. \tag{3.36}$$

To find which of the parameters $l_\pm$ in Eq. (3.34) is the right one, we can perform the limit $k \to 0$:

$$q_s \to 0\,, \qquad q_r \to 1 + (m-1)q_1 - mq_0\,, \qquad l_\pm(q_s) \to l_\pm(0) = \begin{cases} 1/(\beta h)^2\,, \\ 0\,, \end{cases} \tag{3.37}$$

so that $\boldsymbol{\lambda}$ has a finite limit only with l_-, for which the saddle-point equations become

$$\left(\boldsymbol{\lambda}^{-1}\right)_{ab} = -q_{ab} + \hat{q}_-\,, \tag{3.38}$$

where

$$\hat{q}_- = \frac{4(\beta h)^2\eta_2^2}{\left[1+\sqrt{1+4(\beta h)^2 k\eta_2}\right]^2} \tag{3.39}$$

$$\underset{k\to 0}{\longrightarrow} (\beta h)^2\left[1+(m-1)q_1 - mq_0\right]^2 = (\beta h)^2\eta_1^2\,. \tag{3.40}$$

The structure is the same as the one of q_{ab}, with a constant added to each entry. Thus, the entries of $\boldsymbol{\lambda}^{-1}$ can be written as

$$\left(\boldsymbol{\lambda}^{-1}\right)_{ab} = (q_1-1)\delta_{ab} + (q_0-q_1)\epsilon_{ab} - q_0 + \hat{q}_-\,. \tag{3.41}$$

It is also easy to see, inverting a matrix with a 1RSB structure, that

$$\lambda_{ab} = -\frac{1}{\eta_0}\delta_{ab} + \frac{q_1-q_0}{\eta_0\eta_1}\epsilon_{ab} + \frac{q_0-\hat{q}_-}{\eta_1\left(\eta_2 - k\hat{q}_-\right)} \tag{3.42}$$

and that $\boldsymbol{\lambda}$ has eigenvalues

$$\begin{aligned} &\kappa_0 = -1/\eta_0 && \text{deg.} = n(m-1)/m\,, \\ &\kappa_1 = -1/\eta_1 && \text{deg.} = n/m - 1\,, \\ &\kappa_2 = -1/\left(\eta_2 - k\hat{q}_-\right) && \text{deg.} = 1\,. \end{aligned} \tag{3.43}$$

The next step is to evaluate the trace appearing in (3.17):

$$\mathrm{Tr}\left(\boldsymbol{\lambda}\times\mathbf{q}\right) = -k\left(1 + \frac{\hat{q}_-}{\eta_2 - n\hat{q}_-}\right)\,. \tag{3.44}$$

Using all these ingredients, we can write the functional $g(\mathbf{q})$ in the 1RSB ansatz for finite k:

$$\begin{aligned} g(k;q_0,q_1,m) = &-\frac{(\beta J)^2}{4}k\left[1+(m-1)q_1^p + (k-m)q_0^p\right] - \frac{k\hat{q}_-}{2(\eta_2 - k\hat{q}_-)} \\ &- \frac{k(m-1)}{2m}\log\left(\eta_0\right) - \frac{k}{2m}\log\left(\eta_1\right) - \frac{1}{2}\log\left(1 + \frac{k(q_0-\hat{q}_-)}{\eta_1}\right) \\ &- \frac{(\beta h)^2}{2}k\left(\eta_2 - k\hat{q}_-\right) - ks(+\infty)\,. \end{aligned} \tag{3.45}$$

As in the previous section, we numerically obtain and plot, in Fig. 3.5, $G(k)/k = g(k;q_1^\star,q_0^\star,m^\star)/(k\beta)$, where again $q_1^\star, q_0^\star, m^\star$ are the solutions of the saddle point equations, obtained by extremization of Eq. (3.45). The most striking feature of these plots

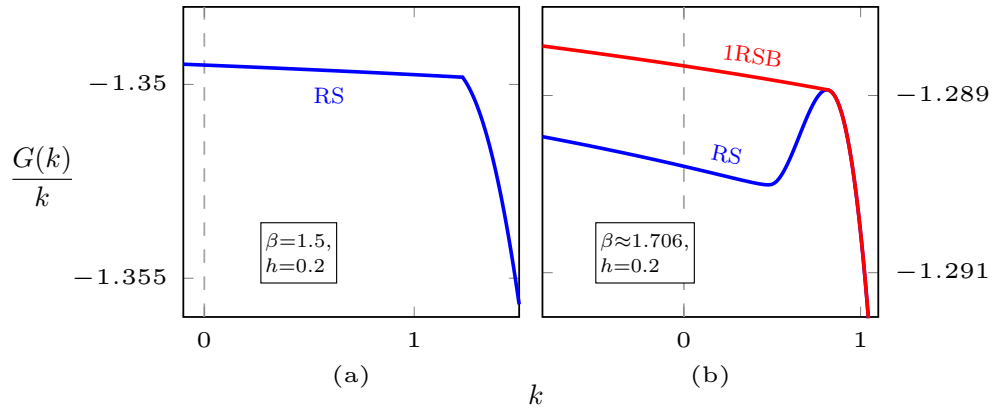

Figure 3.5: The function $G(k)/k$ for the $(p = 3)$-spin in a magnetic field $h = 0.2$, for different values of β: (a) $\beta = 1.5 < \beta_c(h)$, (b) $\beta = \beta_c(h = 0) > \beta_c(h)$. The application of a magnetic field washes out the linear behavior at small k observed in zero magnetic field. Figure from [PDR19].

is the difference from those represented in Fig. 3.3: the linear behavior is replaced by curves (again given by the 1RSB ansatz) with non-null derivative. Let us analyze more closely what is going on and why the external magnetic field is modifying the behavior of the system. As discussed in Sec. 3.3.1, one can apply the Rammal's construction to correct the non-monotonic behavior of the RS version of $G(k)/k$ (plotted as a blue curve in Fig. 3.5). Exactly as in the $h = 0$ case, the resulting function will be monotonic and linear, which is the smooth continuation of $G(k)/k$ from k_m, the point where it loses its monotonicity. However, as one can see from Fig. 3.5), the result will not be the 1RSB solution. This difference from the $h = 0$ case can be seen as a consequence of the saddle point equations: now the equation for q_0 is non-trivial and so either $q_1^\star, q_0^\star, m^\star$ depends on k also in the 1RSB phase, giving rise to the non-constant behavior of $G(k)/k$ also for $k < k_c$. It is worth mentioning another point: when $h = 0$, the critical point k_c where the 1RSB solution departs from the RS one, coincides with k_m, the point where $G(k)/k$ obtained by the RS ansatz loses its monotonicity. Differently, with $h \neq 0$, we have that $k_c > k_m$ for $\beta > \beta_c$, so that the 1RSB branch departs from the RS one above k_m. Finally, we numerically checked that the shape of $G(k)/k$ below k_c depends on p.

This change in the SCGF has an important effect, in turn, on the rate function: performing the numerical Legendre transformation of the SCGF we now obtain a continuous curve, meaning that very rare fluctuations are washed out, see Fig. 3.6. In other words, now the two quantities a_n and b_n introduced in Eq. (3.28) are such that $a_n \sim n$ and $b_n \sim n$. This effect is present also for very small magnetic field, even though the rate function is more and more asymmetrical around $f = f_{\text{typ}}$ as we decrease h.

Let us summarize and comment briefly our results. In this chapter, we analyzed the behavior of the large (and very large) deviations of the free energy for the spherical p-spin model, exploiting the Gärtner-Ellis theorem to obtain the rate function. Without external magnetic field, we are able to compute the rate function in the spin-glass phase, while in the paramagnetic phase we obtain its convex hull, due to the non-differentiability of the SCGF. As a result, we have a standard large deviation principle for fluctuations below the typical value of the free energy, depressed exponentially in the size of the system. On the other hand, fluctuations above the typical value have a different behavior, being suppressed more than exponentially, and the corresponding rate

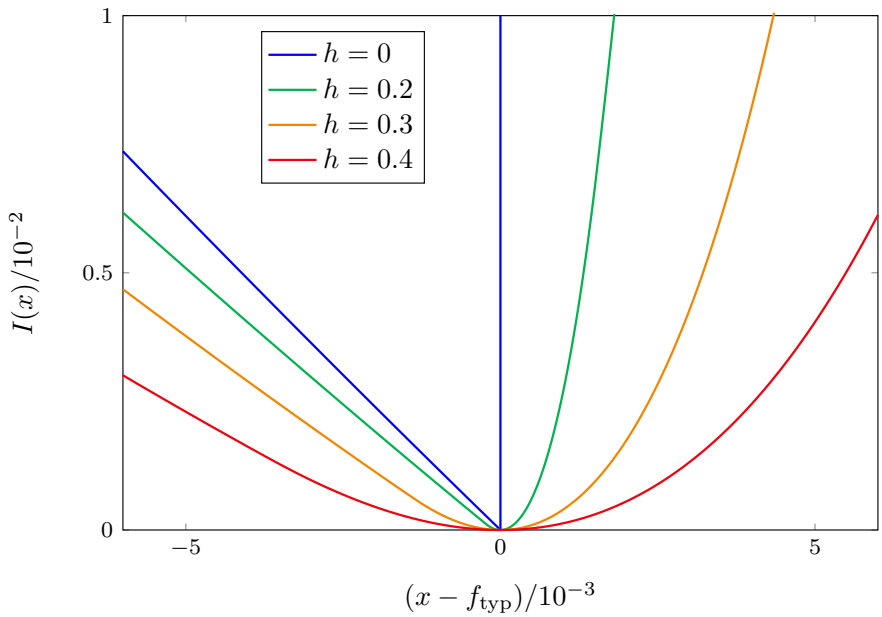

Figure 3.6: Rate function of the free energy for the $(p = 3)$-spin at $\beta = 3$, for different values of the external magnetic field. The infinite branch of the rate functions in Fig. 3.4 is replaced by a curve gradually less steep as the magnetic field is increased. Figure from [PDR19].

function is infinite. When a magnetic field is applied, this anomalous very large deviation disappears and the rate function is finite everywhere. Since this remains true even if the field is very small, an open question is whether this effect can be exploited to obtain insights on the very large fluctuations, by sending the magnetic field to zero carefully choosing its dependence on the system size. We will try to explore this possibility in the next chapter, studying analytically two models whose large deviation behavior has been already discussed in literature with similar or complementary approaches: the $p = 2$ case of the spherical model we analyzed so far, and the Sherrington-Kirkpatrick model.

To suggest the physical reason why the introduction of a magnetic field provides a regularization of the super-exponential large deviation principle for the free energy fluctuations, we can exploit the connection with extreme value statistics we explained in Sec. 2.4. As we wrote, the strong suppression in probability of the above-typical fluctuations can be ultimately traced back microscopically (i.e. with respect to the elementary degrees of freedom of the problems, the spins and their couplings) to the rarity of the joint events associated: in the random matrix theory analogy, a finite fraction of the lowest eigenvalues must fluctuate collectively above their typical values to "make room" for $\lambda_{\min}$ to move; in a similar way, at (inverse) temperature β a certain number of non-equilibrium spin configurations with free energy close to its equilibrium value (which is the "optimal" one, that is the lowest) must present non-typical patterns of the couplings making its value to fluctuate in the same way. An external magnetic field, on the other hand, produces a collective shift in energy that could possibly make a lot easier for these events to occur, to the extent that the corresponding anomalous scaling in n at zero magnetic field becomes a regular scaling when the field is applied. We will clarify and quantify this expectation in the next chapter, in cases where analytical evaluations can

be performed.

In addition, we provided a geometrical interpretation to support our numerical findings. Indeed we showed, as noticed previously in the literature for different models, that for $h = 0$ the Rammal construction is equivalent to the 1RSB ansatz. However, we also showed that this is due to the simple structure of the 1RSB ansatz without external magnetic field, where one can immediately fix one of the 1RSB parameters. When a magnetic field is applied, all the parameters have non-trivial values (which we obtained numerically by solving the saddle point equations) and the Rammal construction, which gives in turn the infinite-rate-function behavior, fails. Another interesting question is whether it is possible to generalize the geometrical construction by Rammal to correct in the right way the RS solution not only for $h = 0$, but also when $h \neq 0$.

CHAPTER 4

Remarks on the case of 2-spin models

In this chapter we explore, at a very preliminary level, the possibility to use a small magnetic field, which ultimately can be seen as a weak interaction between different replicas, as a regularizer to access informations on the anomalous very large deviation behavior of some models. While recovering the true exponential speed in n of the LDP associated to these kind of fluctuations seems to be a hopeless task in this approach, because it would require to know how to scale $h \to 0$ as n approaches infinity (something we cannot do without knowing already the speed and fine-tuning the result), it is legitimate to ask if at least the functional form of the rate function in a magnetic field does correspond with the one obtained for the anomalous branch with other approaches. Otherwise, we should conclude that the fluctuations induced by h are of a different kind with respect to the anomalous ones, which in turn are screened, and washed out, by the magnetic field on top.

To this aim, we will focus on models where we have some level of analytical control on the results, namely the $(p = 2)$-spin spherical model and the Sherrington-Kirkpatrick model. We will start analyzing the saddle-point equations of these models, in presence of a small magnetic field. Then, in Sec. 4.3 we will report briefly some state-of-the-art results known in literature, mostly from reference [PR10b], to check the validity of our approach.

4.1 The 2-spin spherical model

4.1.1 Saddle-point equations

To study the 2-spin spherical model, we can simply specialize to $p = 2$ the formulas we obtained in Chap. 3. In particular, the functional to minimize/maximize with respect to the replica parameters is, from Eq. (3.45),

$$\begin{aligned} g(k; q_0, q_1, m) = &- \frac{(\beta J)^2}{4} k \left[1 + (m-1)q_1^2 + (k-m)q_0^2\right] - \frac{k\hat{q}_-}{2(\eta_2 - k\hat{q}_-)} \\ &- \frac{k(m-1)}{2m} \log(\eta_0) - \frac{k}{2m} \log(\eta_1) - \frac{1}{2} \log\left(1 + \frac{k(q_0 - \hat{q}_-)}{\eta_1}\right) \\ &- \frac{(\beta h)^2}{2} k (\eta_2 - k\hat{q}_-) - ks(+\infty) \,. \end{aligned} \tag{4.1}$$

For small values of the magnetic field h, a limit we take in the spirit of considering the magnetic field as a regularizer (and to perform the calculations), the functional becomes

$$\begin{aligned} g(k;q_0,q_1,m) = &-\frac{(\beta J)^2}{4}k\left[1+(m-1)q_1^2+(k-m)q_0^2\right]-\frac{k(m-1)}{2m}\log\left(1-q_1\right) \\ &-\frac{k}{2m}\log\left[1-q_1+m(q_1-q_0)\right]-\frac{1}{2}\log\left(1+\frac{kq_0}{1-q_1+m(q_1-q_0)}\right) \\ &-ks(+\infty)-\frac{(\beta h)^2}{2}k\left[1+(m-1)q_1+(k-m)q_0\right]+O(h^4) \end{aligned} \tag{4.2}$$

which, in the RS ansatz, reduces to

$$\begin{aligned} g_{\text{RS}}(k;q) = &-\frac{(\beta J)^2}{4}k\left[1+(k-1)q^2\right]-\frac{k}{2}\log(1-q)-\frac{1}{2}\log\left(1+\frac{kq}{1-q}\right) \\ &-ks(+\infty)-\frac{(\beta h)^2}{2}k\left[1+(k-1)q\right]+O(h^4)\,. \end{aligned} \tag{4.3}$$

Requiring stationarity with respect to the parameters, we obtain the saddle-point equations, whose solutions are candidates for the optimal $q_1^\star, q_0^\star, m^\star$ (in the 1RSB case) and $q^\star$ (in the RS case):[1]

$$\begin{cases} \beta^2 q_0+(\beta h)^2-\dfrac{q_0}{\left[1-q_1+m(q_1-q_0)\right]\left[1-q_1+m(q_1-q_0)+kq_0\right]}=0\,, \\ \beta^2 q_1+(\beta h)^2-\dfrac{q_0}{\left[1-q_1+m(q_1-q_0)\right]\left[1-q_1+m(q_1-q_0)+kq_0\right]} \\ \qquad\qquad -\dfrac{q_1-q_0}{(1-q_1)\left[1-q_1+m(q_1-q_0)\right]}=0\,, \\ \dfrac{1}{2}\beta^2\left(q_0^2-q_1^2\right)-(q_1-q_0)(\beta h)^2+\dfrac{q_0(q_1-q_0)}{\left[1-q_1+m(q_1-q_0)\right]\left[1-q_1+m(q_1-q_0)+kq_0\right]} \\ \qquad -\dfrac{(q_1-q_0)}{m\left[1-q_1+m(q_1-q_0)\right]}+\dfrac{1}{m^2}\log\left[\dfrac{1-q_1+m(q_1-q_0)}{1-q_1}\right]=0\,. \end{cases} \tag{4.4}$$

$$\beta^2 q-\frac{q}{(1-q)\left[(k-1)q+1\right]}+(\beta h)^2=0\,. \tag{4.5}$$

These equations are correct to $O(h^4)$, because in the functional only even powers of h appear. Let us first review what happens for $h=0$, as the behavior is a bit different than the previously analyzed for $p>2$.

Null magnetic field In this case the RS saddle-point equation is

$$\beta^2 q-\frac{q}{(1-q)\left[1+(k-1)q\right]}=0\,, \tag{4.6}$$

with solutions

$$q_{\text{PM}}=0\,,\qquad q_\pm(k)=\frac{\beta(k-2)\pm\sqrt{\beta^2k^2-4k+4}}{2\beta\,(k-1)}\,. \tag{4.7}$$

[1] In the following, we take $J=1$.

The solutions $q_\pm$ exist for any k if $\beta > 1$, otherwise they are defined, and included in the interval $q \in [0,1]$, only for

$$k > k_+ = \frac{2 + 2\sqrt{1-\beta^2}}{\beta^2} \qquad \text{if } \beta < 1\,. \tag{4.8}$$

This means that at high temperature, $\beta < \beta_c = 1$, in the integer point $k = 2$ only the trivial solution $q_{\rm PM}$ exists: extrapolating to $k \to 0$, we can conclude that the system is in the paramagnetic phase. For higher values of k, there is another point $k_{\rm RS} > k_+$ where the RS solution evaluated at q_+ crosses the paramagnetic line and becomes the optimal curve, with a non-differentiable junction, as discussed in the previous chapter.

In the low-temperature phase, $\beta > \beta_c$, the solution included in the domain $q \in [0,1]$ is q_+ for any k, which is indeed the right extremum for the $\min/\max$ procedure of the replica method. The resulting $G_{\rm RS}(k)$ obtained continuing trivially the functional to real k has no problem as long as $k > 0$, meaning that also the typical properties of the systems are replica-symmetric. However, the function $G_{\rm RS}(k)/k$ loses monotonicity in $k_m = 0$: to find the solution for negative k, we need to use the 1RSB ansatz or, equivalently, the Rammal's construction. We find

$$\begin{aligned} &\boxed{k > 0} \qquad && q^\star = q_+(k) = \frac{\beta(k-2) + \sqrt{\beta^2 k^2 - 4k + 4}}{2\beta\,(k-1)}\,, \\ &\boxed{k < 0} \qquad && m^\star = k_m = 0\,, \qquad q_0^\star = 0\,, \qquad q_1^\star = q_+(k_m) = \frac{\beta - 1}{\beta}\,. \end{aligned} \tag{4.9}$$

We summarize this result in Fig. 4.1.

Small magnetic field Let us focus on the low-temperature phase $\beta > \beta_c$. The RS equation (4.5) is the third-grade equation

$$\beta^2\,(k-1)\,q^3 + \beta^2\left[h^2(k-1) - k + 2\right] q^2 - \left\{\beta^2\left[h^2(k-2)+1\right] - 1\right\} q - \beta^2 h^2 = 0\,, \tag{4.10}$$

admitting the solution

$$q^\star(k) = \frac{\beta(k-2) + \sqrt{\beta^2 k^2 - 4k + 4}}{2\beta(k-1)} - \frac{\beta(k-2) - \sqrt{\beta^2 k^2 - 4k + 4}}{2\,(\beta^2 - 1)\,\sqrt{\beta^2 k^2 - 4k + 4}}\,h^2 + O(h^4)\,, \tag{4.11}$$

which is in the right domain $q \in [0,1]$ as long as h is small. This result can be obtained supposing the solution to be analytic in h,

$$q = q^\star + q^{(1)} h + q^{(2)} h^2 + \cdots \tag{4.12}$$

and solving the resulting equation order by order. However, the complete solution can be obtained as well with the standard methods to deal with a third-grade equation.

In the 1RSB case, we can use the expansion

$$\begin{aligned} q_0 &= h q_0^{(1)} + h^2 q_0^{(2)} + h^3 q_0^{(3)} + \cdots\,, \\ q_1 &= q_1^\star + h q_1^{(1)} + h^2 q_1^{(2)} + h^3 q_1^{(3)} + \cdots\,, \\ m &= h m^{(1)} + h^2 m^{(2)} + h^3 m^{(3)} + \cdots \end{aligned} \tag{4.13}$$

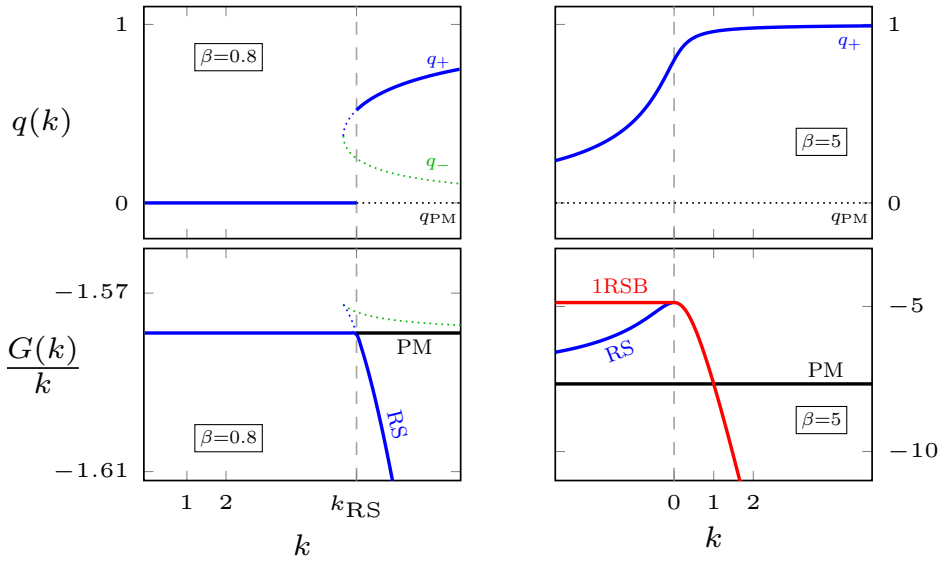

Figure 4.1: Solutions of the RS saddle point equations (top) and SCGF (bottom) for $\beta < \beta_c$ (left) and $\beta > \beta_c$ (right) as functions of the number of replicas k in zero external magnetic field. For $\beta < \beta_c$, the three solutions $q_\pm$ (upper and middle curve) and $q_{\rm PM}$ (the line $q = 0$) exist for $k > k_+$; q_- always corresponds to a maximum of the SCGF, q_+ is a minimum for $k > k_{\rm RS}$, $q_{\rm PM}$ becomes the minimum for $k < k_{\rm RS}$; the continuous blue line is the right solution of the saddle point equations. For $\beta > \beta_c$, q_+ is always the right solution of the RS saddle point, but the RS trial $G(k)/k$ loses monotonicity in $k = 0$; the true SCGF for $k < 0$ is obtained via a 1RSB construction (red line).

in Eq. (4.4) to get[2]

$$\begin{cases} q_0^\star = \sqrt{-\dfrac{1}{\beta k}}\, h - \dfrac{1}{2}\, h^2 + \dfrac{1}{8}\sqrt{-\beta k}\, h^3 + O(h^4)\,, \\ q_1^\star = \dfrac{\beta - 1}{\beta} + O(h^4)\,, \\ m^\star = O(h^4)\,. \end{cases} \tag{4.14}$$

Note that $q_0^\star$ diverges as $k^{-1/2}$ for $k \to 0^-$. However, for this solution to be sensible, it must be $q_0^\star < q_1^\star$: as the two parameters become the same when k is equal to

$$k_{\text{1RSB}} = -\frac{\beta h^2}{\beta - 1} + O(h^4)\,, \tag{4.15}$$

we can take the 1RSB anstaz in this form only of $k < k_{\text{1RSB}}$. Moreover, from the RS Eq. (4.11) we see that

$$q^\star(k_{\text{1RSB}}) = \frac{\beta - 1}{\beta} + O(h^4) = q_1^\star\,, \tag{4.16}$$

so the two ansätze coincides in this point. We can conclude that $k_{\text{1RSB}} < 0$ is indeed the point where the 1RSB solution branches from the RS for this model. Note that it is no more true that $k_{\text{1RSB}} = k_m = m^\star$, as in the Rammal's construction (the three values are all different, actually).

The solution for $q_0^\star$, as written in (4.14), has a problem also for $k \to -\infty$, as it does not remain finite. This is due to the expansion in powers of h, which appears in non-trivial combinations with k in the original expression (3.45). However, supposing that $q_1^\star$ and $m^\star$ do not change with h (the system (4.4) can be cast in a form where two of the equations do not depend on the magnetic field, as already noted in [CS92]), the full solution for $q_0^\star$ can be obtained:

$$q_0^\star = -\frac{\beta h^2 k + h\sqrt{\beta^2 h^2 k^2 - 4\beta k}}{2\beta k} \underset{k\to-\infty}{\longrightarrow} 0\,, \tag{4.17}$$

given the caveat in note 2.

4.1.2 SCGF and rate function in the low-temperature phase

Evaluating expression (4.2) at the solutions of the saddle-point equations, we find, for $\beta > \beta_c$,

$$G(k) = \begin{cases} -k\left[1 - \frac{\log\beta}{2\beta} + \frac{\log 2\pi}{2\beta} - \frac{1}{4\beta} + \frac{h^2}{2}\right] - \frac{1}{3}|k|^{3/2}\sqrt{\beta}h^3 & \text{if } k \le k_{\text{1RSB}}\,, \\ g_{\text{RS}}(k; q^\star(k))/\beta & \text{if } k > k_{\text{1RSB}}\,. \end{cases} \tag{4.18}$$

As expected the magnetic field, moving q_0 away from 0, changes the linear behavior of the SCGF in the 1RSB branch; however, the correction $O(h^2)$ is not enough to see this

[2] To obtain the coefficient of the q_0's $O(h^3)$ term, some of the saddle-point equations must be evaluated up to $O(h^4)$. Here we are forgetting all the successive powers in h of the original theory, supposing the system in (4.4) to be exact. A more careful evaluation, which keeps also the $O(h^4)$ terms in the equations, gives a numerical coefficient of $5/8$ instead of $1/8$ for the q_0's $O(h^3)$ term, and complicates a bit the successive discussion.

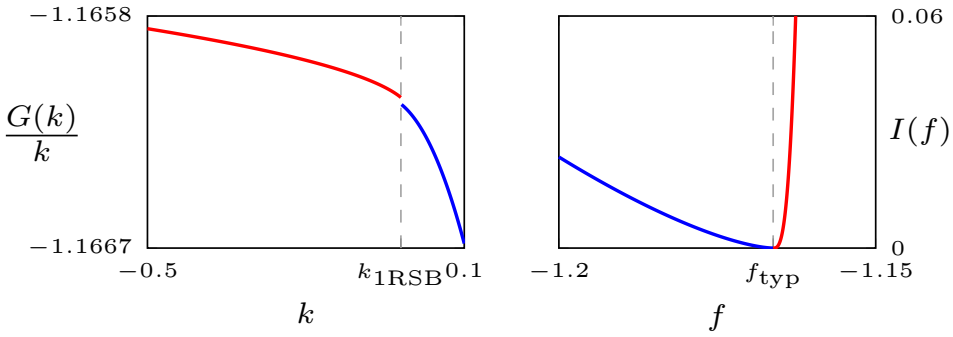

Figure 4.2: Left: SCGF for the 2-spin spherical model in a small magnetic field ($h = 0.1$) at low temperature ($\beta = 2$) from Eq. (4.18), with the two branches represented in different colors; the two lines do not meet precisely in $k_{\text{1RSB}} < 0$ due to higher order corrections in h. Right: rate function for the same values of temperature and magnetic field; the branch $I_-(f)$ (blue line) is obtained via a numerical Legendre transformation of the left branch of Eq. (4.19), the branch $I_+(f)$ (red line) is evaluate analytically from Eq. (4.22).

effect, and we have to go up to $O(h^3)$. From Eq. (2.11) we get

$$\Lambda(k) = \begin{cases} -g_{\text{RS}}(-k/\beta; q^\star(-k/\beta)) & \text{if } k \leq \frac{(\beta h)^2}{\beta - 1}\,, \\ -k\left[1 - \frac{\log\beta}{2\beta} + \frac{\log 2\pi}{2\beta} - \frac{1}{4\beta} + \frac{h^2}{2}\right] + \frac{1}{3}k^{3/2}h^3 & \text{if } k > \frac{(\beta h)^2}{\beta - 1}\,. \end{cases} \tag{4.19}$$

To find I_+, the rate function for the deviations above the typical value, we have to evaluate the Legendre transformation[3]

$$I_+(f) = \sup_{k > \frac{(\beta h)^2}{\beta - 1}} \left[kf - ka(h) - k^{3/2}b(h)\right], \tag{4.20}$$

where

$$a(h) = -\left[1 - \frac{\log\beta}{2\beta} + \frac{\log 2\pi}{2\beta} - \frac{1}{4\beta} + \frac{h^2}{2}\right] \approx f_{\text{typ}}(h)\,, \qquad b(h) = \frac{1}{3}h^3\,. \tag{4.21}$$

With a simple analytical calculation, we find

$$I_+(f) = \frac{4}{3}\frac{\left[f - f_{\text{typ}}(h)\right]^3}{h^6}\,. \tag{4.22}$$

Note that this quantity diverges for $h \to 0$, as expected. We report the main results of this section in Fig. 4.2.

4.2 The Sherrington-Kirkpatrick model

Originally proposed in [SK75] as the fully-connected version of the Edwards-Anderson model [EA75], the Sherrington-Kirkpatrick (SK) model, although quite optimistically

[3] Supposing $k_{\text{1RSB}} \approx 0$.

called "solvable" by its creators, presented some serious problems to the theoretical physics community of the time, as nicely summarized in [HP79]. In the attempt to overcome these difficulties, two parallel, complementary and very fruitful lines of research opened up: first, in [TAP77] Thouless, Anderson and Palmer wrote a system of self-consistent equations for the local magnetizations whose solutions (which are in an extensive number) represent the possible metastable states of the systems, as seen in [BM80]; then, in the groundbreaking series of works [Par79], Parisi proposed his mechanism of replica symmetry breaking which leads to the solution of the model using the replica method. Nowadays, we know indeed that the inconsistencies encountered by the early investigators were due to a replica approach based on the assumption of unbroken replica symmetry: in our present discussion, this is the first (and only) model we deal with that requires to go beyond the 1RSB ansatz we explained in Chap 2. Moreover, some of the ideas we reported to explain our "finite-k" approach, such as the property of convexity and the Rammal's construction, were first applied to the SK model, as in [Ram81; Kon83].

The model is defined by the Hamiltonian

$$H = -\sum_{i<j=1}^{n} J_{ij} S_i S_j - h \sum_{i=1}^{n} S_i \,, \tag{4.23}$$

where the couplings $J_{ij} = J_{ji}$ are i.i.d random variables such that

$$p(J_{ij}) = \frac{1}{J}\sqrt{\frac{n}{2\pi}} \exp\left[-\frac{n}{2J^2} J_{ij}^2\right] \,, \tag{4.24}$$

and the spins are binary (Ising) variables, $S_i \in \{-1, +1\}$. This is the only difference with respect to the 2-spin spherical model, where the spins are continuous variables under the global constraint (3.3). The averaged replicated partition function for k replicas is

$$\overline{\mathcal{Z}^k} = \exp\left[\frac{nk(\beta J)^2}{4}\right] [\mathrm{Tr}_S]^k \exp\left[\frac{(\beta J)^2}{2n} \sum_{a<b=1}^{k} \left(\sum_{i=1}^{n} S_i^a S_i^b\right)^2 - \beta h \sum_{a=1}^{k}\sum_{i=1}^{n} S_i\right] \,, \tag{4.25}$$

where

$$\mathrm{Tr}_S = \sum_{S_1=\pm 1} \cdots \sum_{S_n=\pm 1} \,. \tag{4.26}$$

As the theory is quadratic in the overlaps, we can use a Hubbard-Stratonovich (HS) transformation, via the simple Gaussian identity

$$e^{\frac{(\beta J)^2}{2n}\left(\sum_{i=1}^{n} S_i^a S_i^b\right)^2} = \sqrt{\frac{n(\beta J)^2}{2\pi}} \int_{-\infty}^{+\infty} \mathrm{d}Q_{ab}\, e^{-\frac{n(\beta J)^2}{2} Q_{ab}^2 + (\beta J)^2 Q_{ab} \sum_{i=1}^{n} S_i^a S_i^b} \,, \tag{4.27}$$

leading to (neglecting sub-exponential factors)[4]

$$\overline{\mathcal{Z}^k} = \int \prod_{a<b} \mathrm{d}Q_{ab} \, \exp\left[-n g(\mathbf{Q})\right] \,, \tag{4.28}$$

[4]Note that this procedure is equivalent to the one we used to get Eq. (3.16) for $p = 2$, once the auxiliary variables $\mathbf{q}$ are integrated out, leaving only the ones we called $\boldsymbol{\lambda}$.

with

$$g(\mathbf{Q}) = \frac{(\beta J)^2}{2} \sum_{a<b=1}^{k} Q_{ab}^2 - \frac{k(\beta J)^2}{4} - \log \left(\sum_{S^1,\cdots,S^k=\pm 1} e^{\beta h \sum_a S^a + (\beta J)^2 \sum_{a<b} S^a S^b Q_{ab}} \right). \tag{4.29}$$

The SCGF of the model, as usual, is given by

$$\beta G(k) = g\left(\mathbf{Q}^\star\right), \tag{4.30}$$

where $\mathbf{Q}^\star$ is the solution of the saddle-point equations

$$\frac{\partial g}{\partial Q_{ab}} = 0, \qquad 1 \le a, b \le k, \tag{4.31}$$

which corresponds to a maximum for g, to be evaluated at each value of k, β, h.

To proceed, we need to guess the form of the overlap matrix $\mathbf{Q}$. The simplest one, originally proposed in [SK75], is the RS ansatz:

$$Q_{ab} = q(1-\delta_{ab}) = \begin{cases} 0 & \text{if } a = b, \\ q & \text{if } a \neq b. \end{cases} \tag{4.32}$$

In this way, the remaining sums over the spin variables can be factorized via another HS transformation,

$$e^{q(\beta J)^2 \sum_{a<b} S^a S^b} = e^{-\frac{kq(\beta J)^2}{2} + \frac{q(\beta J)^2}{2} \sum_{a,b} S^a S^b} = e^{-\frac{kq(\beta J)^2}{2}} \int \mathrm{D}x\, e^{\beta J \sqrt{q} x \sum_a S^a}, \tag{4.33}$$

where

$$\mathrm{D}x = \frac{\mathrm{d}x}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right), \tag{4.34}$$

so that the spin can be summed, obtaining

$$g_{\mathrm{RS}}(q;k) = \frac{(\beta J)^2}{4} k\left[(k-1)q^2 + 2q - 1\right] - \log \int \mathrm{D}x \left\{2\cosh\left[\beta\left(h + J\sqrt{q}\,x\right)\right]\right\}^k. \tag{4.35}$$

The saddle-point system becomes the single equation for q

$$q = \frac{\int \mathrm{D}x \cosh^k\left[\beta(h+\sqrt{q}x)\right] \tanh^2\left[\beta(h+\sqrt{q}x)\right]}{\int \mathrm{D}x \cosh^k\left[\beta(h+\sqrt{q}x)\right]} = 1 - \frac{\int \mathrm{D}x \cosh^{k-2}\left[\beta(h+\sqrt{q}x)\right]}{\int \mathrm{D}x \cosh^k\left[\beta(h+\sqrt{q}x)\right]}. \tag{4.36}$$

The equation is in an implicit form and can be solved with different methods, analytical and numerical. For example, in the limit $\beta \to \infty$, where all the replicas are in the same state and so

$$q(k) = 1 - (\text{corrections suppressed in } \beta) \qquad \text{for } \beta \to \infty, \tag{4.37}$$

one can neglect the dependence on q in the RHS. However, in the following we will adopt a different strategy, studying the model near the critical temperature.

4.2.1 The truncated model

As noted in [BM79; Par79], near $\beta = \beta_c$ the entries of the matrix $Q_{\alpha\beta}$ for $\alpha \neq \beta$ are small, at least for $k < 1$, where the solution must be near the paramagnetic line. Therefore, we can expand the term still depending on the spin configurations in (4.29), inside the logarithm:

$$\begin{aligned} e^{\beta h \sum_a S^a + (\beta J)^2 \sum_{a<b} S^a S^b Q_{ab}} = 1 + (\beta J)^2 \sum_{a<b} S^a S^b Q_{ab} + \frac{(\beta J)^4}{2} \sum_{\substack{a<b,\\ c<d}} S^a S^b S^c S^d Q_{ab} Q_{cd} \\ + \beta^3 J^2 h \sum_{\substack{a,\\ b<c}} S^a S^b S^c Q_{bc} + \frac{(\beta J)^6}{6} \sum_{a<b} \sum_{c<d} \sum_{e<f} S^a S^b S^c S^d S^e S^f Q_{ab} Q_{cd} Q_{ef} \\ + \frac{\beta^4 J^2 h^2}{2} \sum_{\substack{a,b,\\ c<d}} S^a S^b S^c S^d Q_{cd} + \cdots , \end{aligned} \tag{4.38}$$

where we have neglected all the terms not containing $\mathbf{Q}$. Now, we can perform in a simple way the summation over the spin variables: whenever a replica index appears an odd number of times in a term, the sum is zero, because of the alternating signs, otherwise it produces a factor of 2. For example, the cubic term

$$\frac{\beta^4 J^2 h^2}{2} \sum_{\substack{a,b,\\ c<d}} S^a S^b S^c S^d Q_{cd} \quad \xrightarrow[\sum_{S=\pm 1}]{} \quad \beta^4 J^2 h^2 \sum_{a<b} Q_{ab} \,. \tag{4.39}$$

At the end, we can re-exponentiate the result, obtaining the *truncated model* with functional (see also [DG06] for reference)

$$\begin{aligned} g(\mathbf{Q}) \approx \frac{\beta^2}{2} \sum_{a<b} Q_{ab}^2 - \frac{\beta^4}{4} \operatorname{tr} \mathbf{Q}^2 - \frac{\beta^6}{6} \operatorname{tr} \mathbf{Q}^3 \\ - \beta^8 \left[\frac{1}{12} \sum_{a,b} Q_{ab}^4 + \frac{1}{8} \operatorname{tr} \mathbf{Q}^4 - \frac{1}{4} \sum_{a,b,c} Q_{ab}^2 Q_{ac}^2 \right] - \frac{\beta^4 h^2}{2} \sum_{a,b} Q_{ab} \,. \end{aligned} \tag{4.40}$$

Here and in the following, we take $J^2 = 1$. Rescaling $\mathbf{Q} \to \mathbf{Q}/\beta^2$,

$$\begin{aligned} g(\mathbf{Q}) = -\frac{1}{4} \left(1 - \frac{1}{\beta^2} \right) \operatorname{tr} \mathbf{Q}^2 - \frac{1}{6} \operatorname{tr} \mathbf{Q}^3 - \frac{1}{12} \sum_{a,b} \mathbf{Q}_{ab}^4 \\ - \frac{1}{8} \operatorname{tr} \mathbf{Q}^4 + \frac{1}{4} \sum_{a,b,c} Q_{ab}^2 Q_{ac}^2 - \frac{\beta^2 h^2}{2} \sum_{a,b} Q_{ab} \,. \end{aligned} \tag{4.41}$$

Near $\beta_c = 1$, we can write $(1 - 1/\beta^2) \approx 2(\beta - 1) \approx 2\tau = 2(1 - 1/\beta)$, so

$$g(\mathbf{Q}) = -\frac{1}{2} \left[\tau \operatorname{tr} \mathbf{Q}^2 + \frac{1}{3} \operatorname{tr} \mathbf{Q}^3 + \frac{1}{6} \sum_{a,b} Q_{ab}^4 + \frac{1}{4} \operatorname{tr} \mathbf{Q}^4 - \frac{1}{2} \sum_{a,b,c} Q_{ab}^2 Q_{ac}^2 + h^2 \sum_{a,b} Q_{ab} \right] . \tag{4.42}$$

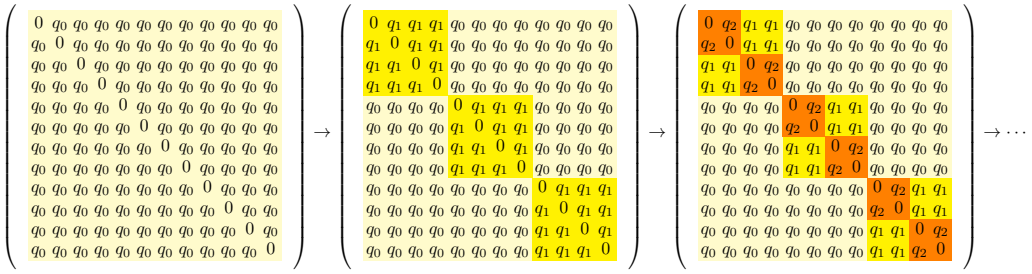

Figure 4.3: Parisi's hierarchical RSB scheme: at first step, replica symmetry is broken in diagonal blocks of dimension m_1, then each block is broken with other blocks of dimension m_2, and so on.

Usually, only the first of the quartic terms is retained, because is the one responsible of the RSB phenomenology:

$$g(\mathbf{Q}) = -\frac{1}{2}\left[\tau \operatorname{tr} \mathbf{Q}^2 + \frac{1}{3}\operatorname{tr} \mathbf{Q}^3 + \frac{1}{6}\sum_{a,b} Q_{ab}^4 + h^2 \sum_{a,b} Q_{ab}\right]. \tag{4.43}$$

This is the starting point of our finite-k analysis, which generalizes the reasoning in [Kon83] by keeping the magnetic field, following closely and reproducing extensively, for reference, the calculations performed in [DFM94].

4.2.2 Full replica symmetry breaking

To solve the model, we need to explain the full procedure of replica symmetry breaking introduced by Parisi in [Par79], whose 1RSB ansatz (2.32) is the first step. The idea is to break iteratively each diagonal block of the 1RSB replica matrix in the same way as the first step, as we see in Fig. 4.3. Therefore, denoting with N the number of step in this scheme, we know that a NRSB matrix $\mathbf{Q}$ is parametrized by

$$k \geq m_1 \geq m_2 \geq \cdots \geq m_N \geq 1\,, \qquad 0 \leq q_0 \leq q_1 \leq \cdots \leq q_N \leq 1\,. \tag{4.44}$$

For example, in Fig. 4.3 (right) we have a 2RSB matrix with $k = 12$, $m_1 = 4$, $m_2 = 2$. How can we write the various term in Eq. (4.43) for a NRSB matrix? For example,

$$\sum_{a,b} Q_{ab}^2 = k\left[(k-m_1)q_0^2 + (m_1 - m_2)q_1^2 + \cdots + (m_N - 1)q_N^2\right] = k\sum_{j=0}^{N}(m_j - m_{j+1})q_j^2\,, \tag{4.45}$$

where we set $m_0 = k$, $m_{N+1} = 1$. In general, for functions of a single element,

$$\sum_{a,b} f(Q_{ab}) = k\sum_{j=0}^{N}(m_j - m_{j+1})f(q_j)\,. \tag{4.46}$$

Products of NRSB matrices with the same structure also produce NRSB matrices (they form an algebra): if $\mathbf{A}$, $\mathbf{B}$ are matrices with diagonal elements, respectively, $\tilde{a}$, $\tilde{b}$ and off-diagonal elements a_i, b_i, $i = 0, \cdots, N$, then $\mathbf{C} = \mathbf{AB}$ has on the diagonal

$$\begin{aligned}\tilde{c} &= \tilde{a}\tilde{b} + (k-m_1)a_0 b_0 + (m_1 - m_2)a_1 b_1 + \cdots + (m_N - 1)a_N b_N\,,\\ &= \tilde{a}\tilde{b} + \sum_{j=0}^{N}(m_j - m_{j+1})a_j b_j\,,\end{aligned} \tag{4.47}$$

and off-diagonal elements

$$\begin{aligned} c_j = \tilde{a}b_j + a_j\tilde{b} + a_j \sum_{i=j+1}^{N}(m_i - m_{i+1})b_i + b_j \sum_{i=j+1}^{N}(m_i - m_{i+1})a_i \\ + \sum_{i=0}^{j}(m_i - m_{i+1})a_i b_i - m_{j+1}a_j b_j \,, \end{aligned} \tag{4.48}$$

as one can check recursively, starting from RS, 1RSB, 2RSB matrices.[5]

We are now in the position to perform two important steps of the Parisi's scheme. The first is the continuation of this anstaz to $k \in \mathbb{R}$, as we already discussed in the 1RSB approach: the m_i parameters are no more integer numbers. As we know, for $k < 1$ we have to reverse their order relation (4.44):

$$k \leq m_1 \leq m_2 \leq \cdots \leq m_N \leq 1 \qquad \text{if } k < 1 \,. \tag{4.49}$$

In the following we will focus on this region, as we expect that for $k > 1$ replica symmetry is unbroken (neglecting the subtlety we alluded in note 5 of Chap. 2). The second crucial step is to send N, the degree of replica symmetry breaking, to infinity. The parameters q_i, which for N finite can be represented by a stepwise function

$$q(x) = q_j \qquad \text{if } \; k \leq m_j \leq x \leq m_{j+1} \leq 1 \,, \tag{4.50}$$

become a continuous function, and the sums appearing in the formulas above become integrals: for example,

$$\sum_{j=0}^{N}(m_j - m_{j+1})f(q_j) = -\int_k^1 \mathrm{d}x\, q(x) f[q(x)] \,, \tag{4.51}$$

where the minus sign is due to the fact that in this range $(m_j - m_{j+1}) = -\,\mathrm{d}x$ is negative. In the limit $N \to \infty$, a NRSB matrix $\mathbf{Q}$ is parametrized by its diagonal element $\tilde{q}$ and by the off-diagonal function $q(x)$; following the standard notation (see for reference [Dot95]), we write $\mathbf{Q} \to (\tilde{q}, q(x))$. The formulas for the algebra, from Eq. (4.47) and (4.48), given $\mathbf{A} \to (\tilde{a}, a(x))$, $\mathbf{B} \to (\tilde{b}, b(x))$, $\mathbf{C} = \mathbf{AB} \to (\tilde{c}, c(x))$, are

$$\begin{aligned} \tilde{c} &= \tilde{a}\tilde{b} - \int_k^1 \mathrm{d}x\, a(x)b(x) \,, \\ c(x) &= \tilde{a}b(x) + \tilde{b}a(x) - b(x)\int_x^1 \mathrm{d}y\, a(y) - a(x)\int_x^1 \mathrm{d}y\, b(y) - \int_k^x \mathrm{d}y\, a(y)b(y) - xa(x)b(x) \,, \end{aligned} \tag{4.52}$$

while the Hadamard product $\mathbf{A} \cdot \mathbf{B}$ (such that $(\mathbf{A} \cdot \mathbf{B})_{ab} = A_{ab}B_{ab}$) in the limit is represented by $(\tilde{a}\tilde{b}, a(x)b(x))$. Moreover,

$$\begin{aligned} \operatorname{tr} \mathbf{A} &\longrightarrow k\tilde{a} \,, \\ \sum_{a,b=1}^{k}(A_{ab})^l &\longrightarrow k\tilde{a}^l - k\int_k^1 \mathrm{d}x a^l(x) \,. \end{aligned} \tag{4.53}$$

[5] For example, the component c_1 of a 2RSB matrix is in a block diagonal $m_1 \times m_1$ block, so the matrix product has simply $k - m_1$ terms equal to $a_0 b_0$ and the others equal to the ones of a 1RSB matrix (of dimension m_1) product.

To evaluate the terms in (4.43), we need

$$\begin{aligned}
\mathbf{Q}^2 &\longrightarrow \left(\tilde{q}^2 - \int_k^1 \mathrm{d}x\, q^2(x)\,,\ -xq^2(x) + 2\tilde{q}q(x) - 2q(x)\int_x^1 \mathrm{d}y\, q(y) - \int_k^x \mathrm{d}y\, q^2(y) \right), \\
\mathbf{Q}^3 &\longrightarrow \left(\tilde{q}^3 + \int_k^1 \mathrm{d}x \left\{ 3q^2(x) \left[\int_x^1 \mathrm{d}y\, q(y) - \tilde{q} \right] + xq^3(x) \right\},\ \cdots \right).
\end{aligned} \tag{4.54}$$

Taking $\tilde{q} = 0$, we find

$$\begin{aligned}
\operatorname{tr} \mathbf{Q}^2 &\longrightarrow -k \int_k^1 \mathrm{d}x\, q^2(x)\,, \qquad \operatorname{tr} \mathbf{Q}^3 \longrightarrow k \int_k^1 \mathrm{d}x \left[xq^3(x) + 3q^2(x) \int_x^1 \mathrm{d}y\, q(y) \right], \\
\sum_{a,b} Q_{ab} &\longrightarrow -k \int_k^1 \mathrm{d}x\, q(x)\,, \qquad \sum_{a,b} Q_{ab}^4 \longrightarrow -k \int_k^1 \mathrm{d}x\, q^4(x)\,.
\end{aligned} \tag{4.55}$$

Eventually, (4.43) becomes

$$g[q] = \frac{k}{2} \int_k^1 \mathrm{d}x \left\{ \tau q^2(x) - \frac{1}{3} \left[xq^3(x) + 3q^2(x) \int_x^1 \mathrm{d}y\, q(y) \right] + \frac{1}{6} q^4(x) + h^2 q(x) \right\}. \tag{4.56}$$

We see that $g[q]$ is now a *functional* of the function $q(x)$. The procedure we used to write it is called full-RSB (fRSB) anstaz.

4.2.3 Saddle-point equations

To find the saddle-point, we report here for reference and explain in full detail the evaluation in [DFM94]. Varying the functional (4.56) with respect to $q(x)$:

$$\frac{2}{k} \frac{\delta g[q]}{\delta q(x)} = 2\tau q(x) - xq^2(x) - 2q(x) \int_x^1 \mathrm{d}y\, q(y) - \int_k^x \mathrm{d}y\, q^2(y) + \frac{2}{3} q^3(x) + h^2 = 0\,, \tag{4.57}$$

where, of course, $x \in [k, 1]$. To solve this equation, we can derive it with respect to x:

$$q'(x) \left[\tau - xq(x) - \int_x^1 \mathrm{d}y\, q(y) + q^2(x) \right] = 0\,. \tag{4.58}$$

A solution is

$$q'(x) = 0\,. \tag{4.59}$$

Deriving again the quantity in the square brackets with respect to x, we get

$$q'(x) \left[-x + 2q(x) \right] = 0\,, \tag{4.60}$$

with solutions

$$q'(x) = 0\,, \qquad q(x) = \frac{x}{2}\,. \tag{4.61}$$

Joining the possible branches, we see that the general solution is

$$q(x) = \begin{cases} q_0 & \text{if } k \le x < x_0\,, \\ x/2 & \text{if } x_0 \le x < x_1\,, \\ q_1 & \text{if } x_1 \le x \le 1\,. \end{cases} \tag{4.62}$$

We have to use the other equations to find the values of the constants q_0, q_1, x_0, x_1. Inserting this expression in Eq. (4.58), we find, for different ranges of x,

$$\begin{cases} \tau - q_1 + q_1^2 = 0 & \text{if } x_1 \le x \le 1\,, \\ \dfrac{x_1^2}{4} - q_1 x_1 + (q_1 - \tau) = 0 & \text{if } x_0 \le x < x_1\,, \\ \dfrac{x_0^2}{4} - q_0 x_0 + q_0^2 = 0 & \text{if } k \le x < x_0\,, \end{cases} \tag{4.63}$$

with solution

$$q^\star_{1,\pm} = \frac{1 \pm \sqrt{1-4\tau}}{2}\,, \qquad x_1^\star = 2q_1\,, \qquad x_0^\star = 2q_0\,. \tag{4.64}$$

As we expect that at $\tau = 0$ (the critical temperature) q_1 start from 0 (the paramagnetic solution), we have to choose

$$q_1^\star = \frac{1 - \sqrt{1-4\tau}}{2}\,. \tag{4.65}$$

To determine q_0 we have to use the full saddle-point equation (4.57). With $k \le x < x_0$, we can use

$$\int_x^1 \mathrm{d}y\, q(y) = q_0(x_0 - x) + \int_{x_0}^{x_1} \mathrm{d}y\, \frac{y}{2} + q_1(1 - x_1) = q_0^2 - q_1^2 + q_1 - q_0 x \tag{4.66}$$

to write the cubic algebraic equation for q_0

$$-\frac{4}{3} q_0^3 + k q_0^2 + h^2 = 0\,. \tag{4.67}$$

The simplest cases are when $h = 0$ or $k = 0$:

$$\begin{aligned} &\boxed{h = 0} && q_0^\star = 0\,, \quad q_0^\star = \frac{3}{4}\,k\,, \\ &\boxed{k = 0} && q_0^\star = \sqrt[3]{\frac{3h^2}{4}}\,. \end{aligned} \tag{4.68}$$

In all the other cases, we have to solve the cubic equation. In general, the equation

$$a x^3 + b x^2 + c x + d = 0\,, \tag{4.69}$$

with determinant

$$\Delta = 18abcd - 4b^3 d + b^2 c^2 - 4ac^3 - 27a^2 d^2\,, \tag{4.70}$$

has 3 real roots if $\Delta > 0$, only 1 when $\Delta < 0$. In our case

$$a = -\frac{4}{3}\,, \quad b = k\,, \quad c = 0\,, \quad d = h^2 \quad \Longrightarrow \quad \Delta = -4(\beta h)^2 \left[12h^2 + k^3\right]\,, \tag{4.71}$$

so

$$\Delta > 0 \quad \Longleftrightarrow \quad k < -\sqrt[3]{12h^2}\,. \tag{4.72}$$

When $\Delta > 0$, we can write the three real solutions using Viète's formula. First we write the depressed cubic equation

$$t^3 + ut + v = 0\,. \tag{4.73}$$

If the coefficients are chosen to be

$$u = \frac{3ac - b^2}{3a^2} = -\frac{3k^2}{16}\,, \qquad v = \frac{2b^3 - 9abc + 27a^2 d}{27a^3} = -\frac{3h^2}{4} - \frac{k^3}{32}\,, \tag{4.74}$$

then its solutions are related to the ones of the original cubic by:

$$x_n = t_n - \frac{b}{3a} = t_n + \frac{k}{4}\,. \tag{4.75}$$

Moreover, the solutions can be expressed using trigonometric functions via

$$t_n = 2\sqrt{-\frac{u}{3}}\cos\left[\frac{1}{3}\arccos\left(\frac{3v}{2u}\sqrt{-\frac{3}{u}}\right) - \frac{2\pi n}{3}\right]\,, \quad n = 0, 1, 2\,. \tag{4.76}$$

When $\Delta < 0$, we can perform the same procedure, only with hyperbolic functions instead of trigonometric ones. As, in this regime,

$$u < 0\,, \qquad 4u^3 + 27v^2 = \frac{81}{64}\,h^2\left[12h^2 + k^3\right] > 0\,, \tag{4.77}$$

we obtain the representation

$$t_0 = -2\frac{|v|}{v}\sqrt{-\frac{u}{3}}\cosh\left[\frac{1}{3}\operatorname{arcosh}\left(-\frac{3|v|}{2u}\sqrt{-\frac{3}{u}}\right)\right]\,. \tag{4.78}$$

As v is given by (4.74), we find

$$v < 0 \quad \Longleftrightarrow \quad k > -\sqrt[3]{24h^2}\,. \tag{4.79}$$

However, as we are studying the regime where $\Delta < 0$, v is always negative. Eventually, the general solution for q_0 is

$$q_0^\star(k;h) = \begin{cases} \dfrac{k}{4} + \dfrac{k}{2}\cosh\left(\dfrac{1}{3}\operatorname{arcosh}\left\{\dfrac{-1}{2k^3}\left[-48h^2 - 2k^3\right]\right\}\right) & \text{if } 0 < k < k_{\text{RSB}} \\ \dfrac{k}{4} - \dfrac{k}{2}\cosh\left(\dfrac{1}{3}\operatorname{arcosh}\left\{\dfrac{1}{2k^3}\left[-48h^2 - 2k^3\right]\right\}\right) & \text{if } -\sqrt[3]{12h^2} < k < 0 \\ \dfrac{k}{4} - \dfrac{k}{2}\cos\left(\dfrac{1}{3}\arccos\left\{\dfrac{1}{2k^3}\left[-48h^2 - 2k^3\right]\right\}\right) & \text{if } k < -\sqrt[3]{12h^2} \end{cases} \tag{4.80}$$

where k_{RSB} is the point where $q_0 = q_1$ and the fRSB ansatz reduces to the RS one. For a compact reference on the solutions of a cubic equation see, for example, [McK84].

Now that we have the solution (4.62), we can insert it in the functional (4.56). We give some steps of the evaluation for $k < x_0$, as a reference:

$$\begin{aligned} \int_k^1 \mathrm{d}x\, q^2(x) \int_x^1 \mathrm{d}y\, q(y) &= q_0^2 \int_k^{x_0} \mathrm{d}x \left[(x_0 - x)q_0 + \frac{x_1^2}{4} - \frac{x_0^2}{4} + (1 - x_1)q_1\right] \\ &\quad + \int_{x_0}^{x_1} \mathrm{d}x\, \frac{x^2}{4}\left[\frac{x_1^2}{4} - \frac{x^2}{4} + (1 - x_1)q_1\right] + q_1^2 \int_{x_1}^1 \mathrm{d}x\, \left[(1 - x)q_1\right] \\ &= \frac{1}{2}k^2 q_0^3 - k q_0^4 + k q_1^2 q_0^2 - k q_1 q_0^2 + \frac{2q_0^5}{5} - \frac{4}{3}q_1^2 q_0^3 + \frac{4}{3}q_1 q_0^3 \\ &\quad + \frac{14q_1^5}{15} + \frac{q_1^3}{2} - \frac{4q_1^4}{3}\,, \end{aligned} \tag{4.81}$$

where we used the expression for x_0, x_1, and

$$\begin{aligned}
&\int_k^1 \mathrm{d}x \left\{ \tau q^2(x) - \frac{1}{3} x q^3(x) + \frac{1}{6} q^4(x) + h^2 q(x) \right\} \\
&\quad = (x_0 - k) \left[\tau q_0^2 + \frac{1}{6} q_0^4 + h^2 q_0 \right] - \frac{q_0^3}{6} \left(x_0^2 - k^2 \right) + \int_{x_0}^{x_1} \mathrm{d}x \left[\tau \frac{x^2}{4} - \frac{x^4}{32} + (\beta h)^2 \frac{x}{2} \right] \\
&\qquad + (1 - x_1) \left[\tau q_1^2 + \frac{1}{6} q_1^4 + h^2 q_1 \right] - \frac{q_1^3}{6} \left(1 - x_1^2 \right) \\
&\quad = -h^2 k q_0 + h^2 q_0^2 - h^2 q_1^2 + h^2 q_1 + \frac{1}{6} k^2 q_0^3 - k q_0^2 \tau - \frac{k q_0^4}{6} \\
&\qquad + \frac{4 q_0^3 \tau}{3} + q_1^2 \tau - \frac{4 q_1^3 \tau}{3} - \frac{1}{15} 2 q_0^5 + \frac{2 q_1^5}{15} + \frac{q_1^4}{6} - \frac{q_1^3}{6} \,.
\end{aligned} \tag{4.82}$$

Subtracting these contributions and substituting the expression (4.65) for q_1, we find

$$\begin{aligned}
G(k;h) = &\frac{k}{2} \left[-\frac{4}{15} \sqrt{1 - 4\tau} \tau^2 + \frac{\tau^2}{2} - \frac{1}{60} \sqrt{1 - 4\tau} + \frac{2}{15} \tau \sqrt{1 - 4\tau} - \frac{\tau}{6} + \frac{1}{60} \right] \\
&+ \frac{k}{2} \left[h^2 \left(-k q_0^\star + q_0^{\star 2} + \tau \right) - \frac{1}{3} k^2 q_0^{\star 3} + \frac{5}{6} k q_0^{\star 4} - \frac{}{15} q_0^{\star 5} \right] .
\end{aligned} \tag{4.83}$$

This is the form of the functional at the saddle-point using the fRSB ansatz, i.e. for low temperature in the domain $k < k_{\mathrm{RSB}}$.

4.2.4 SCGF and rate function in the low temperature phase

The first property of the functional (4.83) to note is that for $q_0^\star = 0$, which is the correct solution for $h = 0$ and $k < 0$, it is linear in k, as its spherical counterpart (4.18). This produces the same divergent behavior of the rate function for fluctuations of the free energy above its typical value, as we have already discussed at length in the other models analyzed.

However, for small values of the magnetic field we can expand the solution (4.80) in powers of h, obtaining

$$q_0^\star(k;h) = \begin{cases} \dfrac{3k}{4} + \dfrac{4}{3k^2} h^2 - \dfrac{128}{27k^5} h^4 + O(h^6) & \text{if } 0 < k < k_{\mathrm{RSB}} \,, \\ |k|^{-1/2} h - \dfrac{2}{3k^2} h^2 + \dfrac{10}{9} |k|^{-7/2} h^3 + \dfrac{64}{27k^5} h^4 + O(h^5) & \text{if } k < 0 \,. \end{cases} \tag{4.84}$$

The nonsensical behavior in $k = 0$ is due to the fact that we are expanding in powers of h at k fixed: for $|k| \ll h$ the correct expansion is the one with the order of limits reversed, which gives

$$q_0^\star(k;h) = \sqrt[3]{\frac{3h^2}{4}} + \frac{k}{4} + \frac{k^2}{8\sqrt[3]{6h^2}} - \frac{k^3}{24 \left(6h^2\right)^{2/3}} + O(k^4) \tag{4.85}$$

(see Eq. (4.68) for a comparison). The important thing to notice is the behavior $|k|^{-1/2}$ of the first term in the expansion in h for $k < 0$: inserted in Eq. (4.83), this gives

$$G(k, \beta) = -\frac{1}{3} h^3 |k|^{3/2} + k f_{\mathrm{typ}}(h) \,. \tag{4.86}$$

We can conclude that the $k < 0$ branch of the SCGF of the (truncated) SK model is qualitatively identical to the one of the 2-spin spherical model, Eq. (4.18). Accordingly, also the rate function has the same behavior (4.22).

4.3 Comparison with known results on very large deviations

In both the models we have analyzed in this chapter, we have verified how the magnetic field removes the infinity associated to the very large deviations in the free energy's rate function. Is it possible to access the still unknown super-exponentially suppressed regime by scaling opportunely $h \to 0$ as $n \to \infty$? We think that a necessary condition to answer this question positively is: the functional form of the rate function in a magnetic field must match the one associated to the super-exponential speed. For example, if

$$\begin{aligned} p\big(f > f_{\text{typ}}; h \neq 0\big) &\sim \exp\left[-n\frac{4(f-f_{\text{typ}})^3}{3h^6}\right], \\ p\big(f > f_{\text{typ}}; h = 0\big) &\sim \exp\left[-n^2 L(f)\right], \end{aligned} \tag{4.87}$$

we can hope to obtain $L(f)$ by scaling h (in this case as $h \sim n^{-1/6}$) only if

$$L(f) \stackrel{?}{=} O\big[(f-f_{\text{typ}})^3\big]. \tag{4.88}$$

Otherwise, we should conclude that the two rate functions are associated with different kind of fluctuations and that the resulting LDP is a non-trivial combination of the two laws we reported above. To verify which is the correct answer, in this section we compare our finite-h formulas with some results on the very large deviations of these models known in literature.

In reference [PR10b], Parisi and Rizzo (PR) proposed a method to resolve the infinity in the rate function due to the very large deviations for $h = 0$. The approach is first tested on the 2-spin spherical model, for which pure mathematical results from Random Matrix Theory [DM08] are known, and then applied to the SK model. The presentation is quite technical, so we only report here a qualitative discussion to give the reader a first idea on the procedure. The central point in the PR approach is to scale the number of replicas $k = \alpha n$ with the number of degrees of freedom, with $\alpha < 0$, and then to evaluate the anomalous fluctuations from the correction to the saddle-points fixing the values Q_{ab} of the replica matrix for large n. The procedure works as follows:

1. Search for a SCGF that scales as n^2:

$$\Phi(\alpha) = -\frac{1}{n^2}\log\overline{\mathcal{Z}^{\alpha n}} \tag{4.89}$$

   Now the number of replicas $k = \alpha n$ is taken extensive.

2. Introduce an overlap matrix $\tilde{Q}_{ab}$, with $a, b = 1, 2, \cdots, \alpha n$. For example, for the SK model, the averaged replicated partition function is, as usual (see Eq. (4.28)),

$$\begin{aligned} \overline{\mathcal{Z}^{\alpha n}} = e^{\frac{\alpha n^2 \beta^2}{4}} \int \prod_{a<b=1}^{\alpha n} \mathrm{d}\tilde{Q}_{ab} \, \exp\Bigg[& -n\frac{\beta^2}{2}\sum_{a<b=1}^{\alpha n}\tilde{Q}_{ab}^2 \\ & + n\log\left(\sum_{S^1,\cdots,S^{\alpha n}=\pm 1} e^{\beta^2 \sum_{a<b} S^a S^b \tilde{Q}_{ab}}\right)\Bigg]. \end{aligned} \tag{4.90}$$

3. As the number of integrals is extensive, a saddle-point analysis is not feasible immediately. To perform the saddle-point integrals, subdivide the large matrix $\tilde{Q}_{ab}$ into blocks of arbitrary dimension d. The number of blocks is, of course, $(\alpha n/d)^2$. The elements of a block in position (i,j) are Q^{ij}_{ab}, with $i,j = 1,\cdots,\alpha n/d$ identifying the block, and $a,b = 1,2,\cdots,d$ the elements inside.

4. Perform first the saddle-point method for the off-diagonal blocks. The corresponding equations admit a solution for $Q^{ij}_{ab} = 0$, $i \neq j$. However, expand to third order (second order is not enough to get a non-linear SCGF) around this saddle-point, leaving the integrals in the fluctuations to be evaluated.

5. Take all the diagonal blocks, which are $\alpha n/d$, equal to a certain matrix, $Q^{ii}_{ab} = Q_{ab}$. The partition function is now:

$$\overline{\mathcal{Z}^{\alpha n}} = \int \prod_{a<b=1}^{d} \mathrm{d}Q_{ab}\, \exp\left[-\frac{\alpha n^2\beta^2}{2d} \sum_{a<b=1}^{d} Q^2_{ab} + \frac{\alpha n^2\beta^2}{4} + \frac{\alpha n^2}{d} \log\left(\sum_{S^1,\cdots,S^d=\pm 1} e^{\beta^2 \sum_{a<b} S^a S^b Q_{ab}} \right) - n^2 S[\mathbf{Q},\alpha] \right], \quad (4.91)$$

where only the integrals in the diagonal block are written explicitly, while all the ones in the off-diagonal fluctuations are hidden in S. Calling

$$\alpha\beta F[\mathbf{Q}] = \frac{\alpha\beta^2}{2d} \sum_{a<b=1}^{d} Q^2_{ab} - \frac{\alpha\beta^2}{4} - \frac{\alpha}{d} \log\left(\sum_{S^1,\cdots,S^d=\pm 1} e^{\beta^2 \sum_{a<b} S^a S^b Q_{ab}} \right), \quad (4.92)$$

the result can be written compactly as

$$\overline{\mathcal{Z}^{\alpha n}} = \int \prod_{a<b=1}^{d} \mathrm{d}Q_{ab}\, \exp\left[-n^2 \left(\alpha\beta F[\mathbf{Q}] + S[\mathbf{Q},\alpha] \right) \right]. \quad (4.93)$$

6. Perform the saddle-point method for the elements in the diagonal blocks, *treating* $S[\mathbf{Q},\alpha]$ *as a small perturbation* (as long as α is small): this means that the saddle-point equations are

$$\left. \frac{\partial F[\mathbf{Q}]}{\partial Q_{ab}} \right|_{\mathbf{Q}=\mathbf{Q}^\star} = 0\,, \quad (4.94)$$

to be studied with the usual replica approach (in the limit $d \to 0$, the block dimension), and to first order S can be evaluated at $\mathbf{Q}^\star$, obtaining

$$\overline{\mathcal{Z}^{\alpha n}} = \exp\left[-n^2 \left(\alpha\beta F[\mathbf{Q}^\star] + S[\mathbf{Q}^\star,\alpha] \right) \right]. \quad (4.95)$$

7. Perform the integrals in the remaining off-diagonals fluctuation in $S[\mathbf{Q}^\star]$. As the terms up to third order have been retained, these integrals can be performed in perturbation theory for a cubic model, expanding in loop diagrams. Eventually, this series can be resummed.

With this method, PR find, for small α,

$$\begin{aligned} \Phi_{\text{spherical}}(\alpha) &= \alpha A + |\alpha\beta|^{3/2} B + o(\alpha^{3/2}) \,, \\ \Phi_{\text{SK}}(\alpha) &= \alpha A' + |\alpha\beta|^{12/7} B' + o(\alpha^{12/7}) \,, \end{aligned} \tag{4.96}$$

respectively, for the 2-spin spherical model and for the SK model (we use capital letters for constants in α). Accordingly, the anomalous LDP for these models is, via Gärtner-Ellis and Legendre,

$$\begin{aligned} p_{\text{spherical}}(\Delta f) &\sim e^{-n^2 C \Delta f^3} \,, \\ p_{\text{SK}}(\Delta f) &\sim e^{-n^2 C' \Delta f^{12/5}} \,. \end{aligned} \tag{4.97}$$

The result for the spherical model is confirmed by Random Matrix Theory: for $\beta \to \infty$, the 2-spin model becomes a pure matrix model for the Gaussian matrix J_{ij}, and the free energy fluctuations become the deviations of its lowest eigenvalue (see [DM08] and our Sec. 2.4). For the SK model, the exponent $12/5$ of the rate function is a bit controversial in literature: its value can be related to the scaling exponent of the small deviations (which in these cases are not Gaussian and are in universality classes different from the one provided by the CLT), whose value has been debated in the past. For reference, see [MG10], which studies the distribution of the ground-state energy of some spin glass models (to which the free energy reduces for $\beta \to \infty$): the other possibility is a SCGF of order $\alpha^{8/5}$, producing a rate function $O(\Delta f^{8/3})$. Here we will not address this problem; we just note that, while for the spherical model the exponent of the rate function associated to the very large deviations matches with the one we obtained for the ordinary rate function in a magnetic field, Eq. (4.22), this is not true for the SK model, in neither of the approaches we referred on. We postpone further comments to the conclusive chapter.

Discussion

In this Part we analyzed the behavior of the large (and very large) deviations of the free energy for some models of spin glass, mainly the p-spin spherical model, exploiting the Gärtner-Ellis theorem to obtain the rate function. Without external magnetic field, we were able to compute the rate function in the low-temperature phase, while in the paramagnetic phase we obtained its convex hull, due to the non-differentiability of the SCGF. As a result, we have a standard large deviation principle for fluctuations below the typical value of the free energy, that is they are depressed exponentially in the size of the system. On the other hand, fluctuations above the typical value have a different behavior, being suppressed more than exponentially, and the corresponding rate function is infinite. When a magnetic field is applied, this anomalous very large deviation disappears and the rate function is finite everywhere.

Since this remains true even if the field is very small, we questioned whether this effect can be exploited to obtain insights on the very large fluctuations, by sending the magnetic field to zero carefully choosing its dependence on the system size. Indeed, we know that the anomalous behavior is due to fluctuations associated to very rare joint events, in the context of extreme value statistics. A constant external magnetic field provides the type of collective correlations between the spins that makes easier for those events to occur. For example, it can shift globally the energy spectrum of a theory, in such a way that the probability of lowest eigenvalue to fluctuate around its typical value becomes of the same order on both sides. We tested this idea in the context of the 2-spin spherical model and the (truncated) SK model, where analytical results are quite easy to derive. While the ordinary, magnetic-field induced, fluctuations match in probability the anomalous ones in the spherical model, for a certain choice of the limit $h \to 0$, this is not true in the case of the SK model. We cannot say if this effect is due to the truncation we implemented, or due to a different nature of the magnetic fluctuations compared to the anomalous ones. Because of the preliminary nature of this last analysis, we cannot draw definitive conclusions on this subject, leaving open this point to future investigation. To start with, models with a better analytical control than the SK model, but still presenting most of its non-trivial features, can be studied in detail with the methods we extensively explained in this thesis. As an example, we mention the $s+p$ spherical model by Crisanti and Leuzzi [CL07] as a promising candidate to explore this possibility.

In addition, we provided a geometrical interpretation of the 1RSB ansatz. Indeed we showed, as noticed previously in the literature for different models, that for $h = 0$ the Rammal construction is an equivalent way to continue the RS ansatz for a real number of replicas. However, we also showed that this is due to the simple structure of the 1RSB ansatz without external magnetic field, where one can immediately fix one of the 1RSB

parameters. When a magnetic field is applied, all the parameters have non-trivial values (which we obtained numerically by solving the saddle point equations in the p-spin model, and analytically in the cases of REM and 2-spin) and the Rammal construction, which gives in turn the infinite-rate-function behavior, fails. Another interesting open question is whether it is possible to generalize the geometrical construction by Rammal to correct in the right way the RS solution not only for $h = 0$, but also when $h \neq 0$. Moreover, as we have seen in the SK model, the fRSB approach also produce a linear SCGF below a certain critical value of the number of replicas; however, this is not the point where the RS ansatz becomes non-monotonic, and is reached via the non-trivial fRSB branch.

Part II

Machine learning of geometrically structured data

Motivations

The success of deep learning has transformed data science profoundly in the last decade, within and outside physics [LBH15; GBC16; He+16]. In spite of the accomplishments in practical applications, we are currently facing a lack of fundamental theoretical understanding in the field [Mal16; Bal+16]. Outstanding open questions concern the surprising effectiveness of stochastic gradient descent, which is capable of finding good minima in complex energy landscapes, and the identification of informative metrics to predict the performances of deep (many small layers) and shallow (few large layers) neural networks [Rag+17; MMN18; CS18]. Particularly troublesome is the apparent incompatibility, within the accepted mathematical theories, between the expressive power and the generalization abilities of neural networks: ultimately, the reason why deep architectures with millions of parameters generalize well is mostly unknown [Zha+17; MM19; Cha+19; Ney+17; LS18].

The idea of investigating machine learning within the tools provided by the statistical physics of disordered system is more than thirty years old, starting with the seminal papers by Amit, Gutfreund and Sompolinsky [AGS85a; AGS85b] on the Hopfield model, and with Gardner's replica analysis of the Perceptron architecture [Gar87; GD88]. Many of the results produced in this field have been obtained under the restrictive and unrealistic hypothesis that the inputs of the training set were independent identically distributed random variables with no correlation with their labels. Only quite recently, physicists working in this field are starting to probe the impact of more realistic generative models of synthetic data on the available theoretical frameworks. Sompolinsky and collaborators investigated the problem of the linear classification of perceptual manifolds [CLS18; CLS16] and provided a first quantitative measurement of the ability to support the classification of object manifolds in deep neural networks [Coh+20]. Mézard suggested that hierarchical architectures with hidden layers naturally emerge in the context of Hopfield models, assuming that the training patterns are structured as superpositions of a given set of random features [Méz17], a common property of empirical data [Maz+18a; Maz+18b]. Zdeborová and collaborators provided exact results for the generalization error within the replica approach for two different scenarios of synthetic data: random features and the hidden manifold model [Gol+20; Ger+20]. One of the motivations behind these choices is the observation that many machine learning datasets, or their representations within deep networks, lie on the surface of low dimensional manifolds, as also verified often in practice by measuring their so-called *intrinsic dimension* [Coh+20; EGR19; Ans+19; Fac+17; Erb+20]. More in general, a generative model with a factorized joint probability distribution of the inputs and their corresponding labels is expected to be unrealistic, with respect to the benchmark datasets commonly used in

machine learning (e.g., MNIST, CIFAR-10, or Imagenet). Intuitively, one expects there to be a notion of similarity among inputs that constrains similar inputs to have the same label. This regularity is expected to be related to the problem of generalization, i.e., the ability of a classifier to correctly classify inputs beyond the data set used for training.

The results obtained in the statistical physics framework address the typical case performance. In contrast, statistical learning theory (SLT) [Vap13], a successful mathematical framework in the theory of machine learning, follows the tradition of computer science of establishing worst-case bounds. This difference in scope made it difficult, for physicists and computer scientists alike, to work towards inter-disciplinary results, and few examples of cross-fertilization are found in the literature. Statistical learning theory is the branch of mathematics and computer science that studies inference, or the problem of generating models starting from data [BBL04]. It provides formal definitions for words like "generalization or "overfitting", and it is ultimately designed to evaluate the performance of learning algorithms. As such, it represents the ideal framework to study the problem of generalization in deep learning. Unfortunately, in spite of its elegance, the insight it provides into the impressive generalization abilities of present deep learning models is poor. The main product of the theory in this setting is a set of upper bounds on the generalization error (which roughly counts the average number of errors made on the test set). These upper bounds in many cases turn out to be too loose to be useful [Zha+17; MM19; Bot15; CT92].

The main drawback of this class of bounds is generally recognized to be their being distribution independent, meaning that they hold for any probability distribution over inputs and labels of the data set, and for all models of the chosen hypothesis class. Substantial effort is being put, within statistical learning theory, to overcome these shortcomings and formulate rigorous data-dependent results [Bot15; Ant+03; KLL01; Sha+98]. The Vapnik-Chervonenkis (VC) entropy is a way to establish distribution dependent, and hopefully tighter, bounds to the generalization error [BBL04]. Informally, the VC entropy measures the number of different ways a given class of functions can classify the inputs of the training set. Unfortunately it is usually very difficult to compute explicitly. Linear classifiers and kernel architectures represent a notable exception. Their VC entropy has been evaluated analytically in a remarkable paper by Cover long ago [Cov65], under very mild hypotheses on the probability distribution of the inputs. The explicit calculation shows, however, that knowing the VC entropy does not improve significantly the standard bound obtained using the growth function [Vap99]. In fact, both quantities scale logarithmically with the size of the training set, and depend linearly on the VC dimension, a well known measure of model complexity.

Our goal here is to show how the concept of data structure, as it is emerging in the physics literature, can be addressed within statistical learning theory, thereby providing a bridge between the two viewpoints. This bridge immediately allows a quantification of the generalization capabilities of simple hypothesis classes, which shows how severely loose the classic rigorous bounds in SLT are. Concretely, we investigate the finite-size and asymptotic behavior of the VC entropy of linear classifiers by using both combinatorial and replica techniques. While replica theory is well established in the statistical mechanics of neural networks, combinatorial tools, though certainly not foreign to statistical mechanics [McC10; Car+18; CS02b], have been developed only very recently for what concerns the role of data structure in machine learning [RLG20; CLS18]. We concentrate on two simple models of data structure, or "object manifolds": (i) k-dimensional simplexes with prescribed geometric relations and (ii) spherical manifolds, which are equivalent to classify unstructured data points with margin (and are related to support vector machines [CV95]). These models are not new, and have already received atten-

tion for their being general enough to provide insight, but simple enough to allow full analytical treatment.

Therefore, the main objective of this Part is to investigate the effect that data structure has on the model complexity of simple architectures in machine learning. Previous research in the physics literature addressed this question via the traditional concept of storage capacity α_c, which measures the maximum load α (number of data points over number of parameters) that a model can learn with probability 1 in the thermodynamic limit. By viewing supervised learning as a constraint satisfaction problem, capacity corresponds to the transition between a satisfiable (SAT) and an unsatisfiable (UNSAT) phase, above which perfect training accuracy is achievable with probability 0. Here we show that the compact description of learning provided by the capacity hides important detail about the model, related to its expressive power on structured data. We show that the VC entropy is non-monotonic as a function of the load, and vanishes asymptotically, at variance with the data-agnostic setting. This also contrasts with the classic bounds in statistical learning theory, which are mostly obtained by upper bounding the VC entropy with quantities that grow polynomially in the size of the training set [BBL04; Vap99]. The hallmark of this non-monotonic behavior is an additional phase transition above the storage capacity. The new critical point signals the entrance into the UNSAT phase of another satisfiability problem, related to data structure.

The exposition is mostly drawn from [RPG20; Pas+20].

CHAPTER 5

Linear classification of points

In this chapter we present the problem of pattern recognition, one of the main applications of machine learning. Since this topic is not usually part of a physicist's background, we start spending a few words to set the problem in the more general framework of supervised learning. We then give a brief mathematical introduction on Statistical Learning Theory, which is a very general theoretical framework to study the problem of function estimation (inference) from a given set of data, defining properly the concepts introduced before. Then, we will focus on the problem of linear classification, reviewing some important result from combinatorics (the Cover's theorem) and from statistical mechanics (the Gardner volume). In this last approach, we will see a first example on how replica theory has been applied successfully in the past to find relevant properties of learning architectures.

5.1 Supervised learning and simple perceptron

Supervised learning is a branch of machine learning whose aim is to infer, from a given dataset consisting of input objects (usually represented as points in a suitable space, such as $\mathbb{R}^n$ for vectors, or the corresponding pixel and color space for images), each of them coming with a label identifying the feature to learn (for example, the color of a point, blue or red, or the species of an animal in a picture), the function associating to each object its label. Once this function is known, at the end of the so-called learning process, any new object of the same kind of the ones in the initial dataset (the training set) can be in principle identified with the correct label. The term "supervised" is used in analogy with learning systems where an instructor teaches an apprentice what to do by means of his example, and it is usually opposed to the concept of "unsupervised learning", where the trainee has to learn independently, exploiting similarities and differences of the objects in the training set, presented to him without any previous classification.

Usually, also a test set of object-label pairs is used to check whether the machine can perform correctly the required task, comparing its output with the labels in the test set. During the training process, not only one has to be sure to reduce the training error, that is the number of mistakenly classified objects in the training set, but also to maintain low the test (or *generalization*) error, in order to be able to classify objects not present in the training set. Indeed, in general a machine classifying "too well" the training set can perform poorly on the test set, because it is fine-tuned to identify the objects in the training set only. A behavior of this kind is called *overfitting*: indeed, fitting is different than predicting, and a well-trained machine is not required to perfectly fit known points, but to predict with a good rate of success the outcome of new submissions. For this reason, a good estimation of the generalization error is a central issue in machine learning, as we will see in the next section.

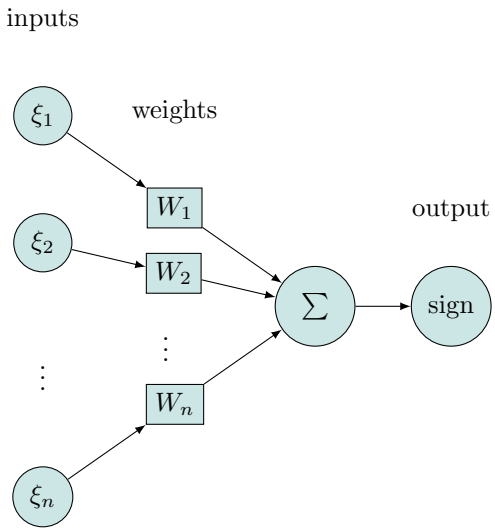

Figure 5.1: A simple perceptron, a linear classifier taking a vector $X = (\xi_1, \cdots, \xi_n) \in \mathbb{R}^n$ as input and returning the sign of the scalar product $W \cdot \xi$ as output, with W vector of weights.

To start fixing ideas, suppose to deal with a training set of p points X^μ in $\mathbb{R}^n$, $\mu = 1, \cdots, p$, with binary labels σ^μ (red/blue, ± 1, cat/dog, ...), and to hope that the dataset can be classified linearly, that is assuming that each label can be associated to the corresponding point via the function

$$\sigma^\mu = \text{sign}\left(\frac{1}{\sqrt{n}} \sum_{j=1}^{n} T_j \xi_j^\mu\right) , \tag{5.1}$$

where ξ_j^μ are the components of X^μ and T is an unknown opportunely normalized vector, which should be the target of the learning process.

We can face the problem of binary linear classification with a simple artificial neural network (ANN) architecture: the *perceptron*, represented in Fig. 5.1. Given a n-dimensional input $X^\mu = (\xi_1^\mu, \cdots, \xi_n^\mu)$ with label σ^μ in the training set, the perceptron gives on output

$$\tilde{\sigma}^\mu = \text{sign}\left(\frac{1}{\sqrt{n}} \sum_{j=1}^{n} W_j \xi_j^\mu\right) . \tag{5.2}$$

The weights are normalized according to

$$\sum_{j=1}^{n} W_j^2 = n \, , \tag{5.3}$$

a convenient choice for a large-n analysis. Of course, as long as the weights are chosen randomly, there is no reason for $\tilde{\sigma}^\mu$ to coincide to the true label σ^μ for any μ. The perceptron algorithm provides the recipe to update the weights $W_j = W_j(t)$ in order to find a

solution $W_j^\star$ such that

$$\sigma^\mu = \text{sign}\left(\frac{1}{\sqrt{n}}\sum_{j=1}^{n} W_j^\star \xi_j^\mu\right), \qquad \mu = 1, \cdots, p\,. \tag{5.4}$$

This procedure, which can be found in standard textbooks like [Nis01], is proven to converge provided that the original dataset Z_p were linearly separable. In the special case in which the labels in the training set are not simply given, but generated via Eq. (5.1), we are in the so-called *teacher-student* scenario, with the perceptron with weights T as the teacher, the one with weights W as the student [EB01]. In the following, we will focus instead on random generated dataset, with a very specific correlation between the points and their labels to implement data structure.

In general, for values of n (dimension of the input space) and p (number of points in the training set) in a certain range, the learning protocol on a typical instance of the training set cannot find any solution $W^\star$ of the linear classification problem. Indeed, the problem can be cast in the form of a *Constraint Satisfaction Problem* (CSP): find a vector $W^\star \in \mathbb{R}^n$ such that all the p constraints (5.4) are respected. As observed in other CSPs, varying the ratio $\alpha = p/n$, which takes the role of the temperature as the driving parameter, a transition from a phase where a solution is easy to a phase where a solution is hard or impossible to find can occur. This transition is very similar to the one from the paramagnetic to the spin glass phase for models we analyzed in Part I, and can be studied with the same tools, as we will see in this and in the next chapters. In the case of the glass transition, the complex form of the energy landscape induces ergodicity breaking and makes in general difficult to find the equilibrium state of the system, which is the one of minimal free energy; here, the space of solutions $W^\star$ is progressively broken into an exponential number of smaller clusters, up to the point that no possible solution remains [Krz+07]. The phase where a solution can still be found is called satisfiable (SAT), the one where there is no solution unsatisfiable (UNSAT). We will use again these concepts in Chap. 7.

In the next section we give a formalization of the inference problem introduced here using the language of Statistical Learning Theory.

5.2 Basic results in Statistical Learning Theory

Statistical Learning Theory (SLT) is the branch of mathematics and computer science that studies inference, or the problem of generating models starting from data. In this section we recall the basic facts of SLT, mostly following reference [BBL04]. The exposition is taken from [Pas+20].

We restrict to binary classification problems, in which the goal is to find a function g mapping the input space $\mathcal{X}$ to the output space $\mathcal{Y} = \{+1, -1\}$. Each pair $Z^\mu = (X^\mu, Y^\mu)$ (with $\mu = 1, \ldots, p$) in the training set $Z_p = (Z^1, \ldots, Z^p)$ is drawn by the unknown joint probability distribution $P_{\mathcal{X},\mathcal{Y}}(X, Y)$. A map g between the set of inputs $X_p = (X^1, \cdots, X^p)$ and $\{+1, -1\}$ is called a dichotomy of X_p. The criterion to choose g is the minimization of the risk

$$R(g) = \langle \mathbb{1}_{g(X)\neq Y} \rangle_P\,, \tag{5.5}$$

which is the probability of error. Ideally, we should look for $\inf_g R(g)$ over all the possible

g's. Since P is unknown, the best we can do is to consider the empirical risk

$$R_p(g) = \frac{1}{p}\sum_{\mu=1}^{p} \mathbb{1}_{g(X^\mu)\neq Y^\mu}, \tag{5.6}$$

and limit the search within a specific hypothesis class $\mathcal{G}$ (a "model") to prevent overfitting. A dichotomy g is called *realizable* if $g \in \mathcal{G}$. The output of a learning algorithm is a function g_p that depends on the data Z_p. The goodness of the choice of g_p can be measured by its generalization error $\epsilon_{\text{gen}}(g_p)$, where

$$\epsilon_{\text{gen}}(g) = R(g) - R_p(g). \tag{5.7}$$

Notice that $\epsilon_{\text{gen}}(g) \leq 1$. In practice, R_p is evaluated on the training set and R is estimated on a test set [Meh+19]. One of the primary goals of SLT is to establish rigorous bounds on the generalization error.

A complementary description of risk minimization within a class $\mathcal{G}$ is given through the definition of the loss class $\mathcal{L}$:

$$\mathcal{L} = \left\{\ell_g : (X,Y) \mapsto \mathbb{1}_{g(X)\neq Y},\, g \in \mathcal{G}\right\}. \tag{5.8}$$

To each $g \in \mathcal{G}$, we associate a function ℓ_g such that $\ell_g((x,y)) = 1$ if $g(x) \neq y$, and is zero otherwise. In this way, while elements of $\mathcal{G}$ take values in $\{+1,-1\}$, those of $\mathcal{L}$ have range $\{0,1\}$. ℓ_{g_p} can be used to count the number of errors made on the training set by the function g_p. Given a loss class $\mathcal{L}$, we can consider its projection on the sample Z_p, by defining

$$\mathcal{L}_{Z_p} = \left\{(\ell(Z^1), \ell(Z^2), \ldots, \ell(Z^p)) \,:\, \ell \in \mathcal{L}\right\}. \tag{5.9}$$

This is the set of all possible ways that the functions in $\mathcal{G}$ can classify, correctly or incorrectly, each sample in Z_p. For example, if a function $g \in \mathcal{G}$ classifies correctly all the elements of the training set but the first one, it corresponds to the string $(1,0,0,\cdots,0)$ in $\mathcal{L}_{Z_p}$, as any other functions in $\mathcal{G}$ with the same output. Importantly, $\mathcal{L}_{Z_p}$ can be interpreted as the set of all different classifications of the points in X_p that can be realized by the model, i.e., the set of all labels $(Y^1,\ldots,Y^p)$ such that there exists (at least) a $g \in \mathcal{G}$ such that $Y^\mu = g(X^\mu)$ for all μ. This representation reveals a useful bijection between $\mathcal{L}_{Z_p}$ and the set of realizable dichotomies.

A key quantity in SLT is the Vapnik-Chervonenkis (VC) entropy $\mathcal{H}_{\mathcal{L}}(Z_p)$, which measures the size of $\mathcal{L}_{Z_p}$:

$$\mathcal{H}_{\mathcal{L}}(Z_p) = \log\left|\mathcal{L}_{Z_p}\right|. \tag{5.10}$$

By virtue of the bijection discussed above, valid when $\mathcal{Y} = \{+1,-1\}$, $\mathcal{H}_{\mathcal{L}}(Z_p)$ can be defined equivalently as

$$\mathcal{H}_{\mathcal{L}}(Z_p) = \log \mathcal{N}_{\mathcal{G}}(X_p), \tag{5.11}$$

where $\mathcal{N}_{\mathcal{G}}(X_p)$ is the number of dichotomies of the set X_p realizable by $\mathcal{G}$. The VC entropy controls a rigorous upper bound to the generalization error:

Theorem 5.1. *For any $0 < \delta \leq \delta_{max} = \min(1, 2e^{\mathcal{H}_{\mathcal{L}}(2p)})$, with probability at least $1-\delta$,*

$$\forall g \in \mathcal{G}, \quad \epsilon_{\text{gen}}(g) \leq 2\sqrt{2\frac{\mathcal{H}_{\mathcal{L}}(2p) + \log\frac{2}{\delta}}{p}}, \tag{5.12}$$

where the annealed VC entropy $\mathcal{H}_{\mathcal{L}}(p)$ is defined as:

$$\mathcal{H}_{\mathcal{L}}(p) = \log\langle \mathcal{N}_{\mathcal{L}}(Z_p)\rangle \tag{5.13}$$

and $\langle\cdot\rangle$ is the average over the joint probability distribution $P_{\mathcal{X},\mathcal{Y}}$ of the training set.

Unfortunately, direct computation of the VC entropy is unfeasible in most cases. For this reason, a main goal of SLT is to construct more tractable upper bounds to the VC entropy. The classic example is based on the Vapnik-Chervonenkis dimension, which is a scalar metric of the expressivity of a given hypothesis class $\mathcal{G}$. More formally, the VC dimension $d_{\rm VC}$ of a class $\mathcal{G}$ is the largest integer such that there exists at least one set of $d_{\rm VC}$ inputs $X_{d_{\rm VC}}$ such that

$$\mathcal{N}_{\mathcal{G}}(X_{d_{\rm VC}}) = 2^{d_{\rm VC}} \tag{5.14}$$

(i.e., the class $\mathcal{G}$ realizes all possible dichotomies of the inputs). With this definition, it can be proved that

$$\mathcal{H}_{\mathcal{L}}(p) \leq d_{\rm VC} \log\left(\frac{ep}{d_{\rm VC}}\right). \tag{5.15}$$

Hence, a corollary of Theorem 5.1 is the well-known upper bound first obtained by Vapnik: if the class $\mathcal{G}$ has finite VC dimension $d_{\rm VC}$, then, with probability at least $1-\delta$,

$$\forall g \in \mathcal{G}, \quad \epsilon_{\rm gen}(g) \leq 2\sqrt{2\frac{d_{\rm VC}\log\left(\frac{2ep}{d_{\rm VC}}\right) + \log\frac{2}{\delta}}{p}}. \tag{5.16}$$

A crucial property of this elegant result is its being distribution independent, meaning that the bound is uniform in the function g, and does not depend on the particular problem at hand. Owing to its universality, the bound is often too loose for most practical applications [Bot15]. Let us consider for instance a deep neural network with a number of weights $w = 10^6$–10^9. In this case the VC dimension is of order $d_{\rm VC} \sim w \log w$ [Son98]. When the typical size of the dataset is $p = 10^4$–10^6, as is often the case in practice, is is evident that bounds such as the one in Eq. (5.16) do not offer any insight on the generalization performance of deep neural networks. Indeed, one of the main pursuits of contemporary SLT is to provide better results on the generalization error, going beyond distribution independent bounds. Several strategies have been proposed, advocating the importance of considering data-dependent hypothesis classes [Sha+98] and data-dependent measures of complexity (such as the Rademacher complexity [BM03], which was recently connected to the statistical mechanics of disordered systems [Abb+19]), also in relation to the original concept of VC entropy itself [Ang+14].

5.3 Vapnik-Chervonenkis entropy of linear classifiers

As mentioned above, in most cases it is not possible to compute the VC entropy directly. However, linear classifiers (such as the perceptron we introduced above) and kernel machines are a notable exception: their VC entropy was computed half century ago by Cover [Cov65]. Kernel architectures provide a special realization of one-hidden layer neural networks and are at the core of the idea of support vector machines. In these machines, one defines *a priori* a feature map $\phi : \mathbb{R}^n \to \mathbb{R}^d$, that maps n-dimensional inputs to a d-dimensional feature space. One of the simplest realizations of such maps is a quadratic polynomial kernel, such that each input X is mapped on a $d(d+1)/2$-dimensional feature space via a kernel $\phi^{(2)}$ with components $\phi^{(2)}_{ij} = X_i X_j, \forall i \leq j$. The map from feature space to the space of labels is realized by a linear separator:

$$Y = \text{sign}(W \cdot \phi(X)), \tag{5.17}$$

where the weight vector, $W \in \mathbb{R}^d$, is the set of learnable parameters.

Cover's theorem is a function counting theorem: it computes the number of dichotomies $\mathcal{N}_\phi(X_p)$ of this function class, the logarithm of which is the VC entropy. It is simpler to state Cover's theorem for linear separators, i.e., for $d = n$ and $\phi = \mathbb{1}$; the realizable dichotomies in this case are called linearly realizable. We comment below on the extension to general ϕ. The key idea behind the theorem is twofold:

(i) under a weak condition on the inputs X_p, the number of dichotomies $\mathcal{N}_\mathbb{1}(X_p)$ is a function solely of the dimension n and the number of points p;

(ii) it is possible to write a solvable recurrence relation, in n and p, for this function.

Following Cover's original paper, we denote the (data-independent) number of dichotomies $\mathcal{N}_\mathbb{1}(X_p)$ by $C_{n,p}$, and the corresponding VC entropy by $\mathcal{H}_{n,p} = \log C_{n,p}$.

Theorem 5.2 (Cover, 1965)**.** *Let X_p be a set of p points in $\mathbb{R}^n$. If the points are in general position, i.e., if the points in X' are linearly independent for all subsets $X' \subseteq X_p$ such that $|X'| \leq n$, then $\mathcal{N}_\mathbb{1}(X_p) = C_{n,p}$, where*

$$C_{n,p} = 2\sum_{j=0}^{n-1} \binom{p-1}{j}. \tag{5.18}$$

The proof of Theorem 5.2 is based on a simple recurrence relation for $C_{n,p}$:

$$C_{n,p+1} = C_{n,p} + C_{n-1,p}, \tag{5.19}$$

with boundary conditions

$$C_{n\geq 1,1} = 2, \quad C_{0,p} = 0. \tag{5.20}$$

Equation (5.19) states that adding the $(p+1)$th point X to X_p increases the number of dichotomies by $C_{n-1,p}$, which is the number of dichotomies of X_p that are realizable by a vector W such that $W \cdot X = 0$. Cover actually proved a more general statement. Informally, if one maps all elements of X_p by the non-linear kernel function ϕ from $\mathbb{R}^n$ to $\mathbb{R}^d$ with d larger than n, then, under mild assumptions on ϕ, Eq. (5.18) holds with d in place of n.

Notice that Eq. (5.18) implies that the VC entropy grows asymptotically as $\mathcal{H}_{n,p} \sim (n-1)\log p$ for large number of inputs p (see Sec. 7.1.2 for a derivation). This is the same behavior as that obtained by bounding the VC entropy as in Eq. (5.15).

Two remarks can be made, concerning the generality of Cover's theorem. First, the general position is a rather weak condition. For instance, we mention three examples of distributions of the points $\xi_\mu \in X_p$ under which the general position holds with probability 1:

(i) $\xi^\mu \in X_p$ are i.i.d. variables with the uniform measure on the sphere S^{n-1};

(ii) $\xi^\mu \in X_p$ are i.i.d. variables with marginal probability distribution $P(\xi)$, and the support of P is $\mathbb{R}^n$;

(iii) the coordinates of each $\xi^\mu \in X_p$ are i.i.d. variables, with discrete probability distribution $p(x) = (1+m)/2\delta_{x,1} + (1-m)/2\delta_{x,-1}$, for any $m \in [-1, 1]$.

Clearly, there are trivial ways to violate general position: for instance, if the probability distribution of (i) or (ii) above is conditioned to assigning the same value to a fixed subset of size $k < n$ of the coordinates of all inputs. Then Cover's theorem still applies in the subspace, with $n - k$ in place of n.

Second, the condition that ϕ must satisfy for the theorem to apply to the kernel machine specified by ϕ is essentially that the vectors $\phi(\xi^\mu)$ must be in general position in the feature space $\mathbb{R}^d$. This again is a very mild condition. Starting with a set of inputs X_p in general position in the original n-dimensional space, most interesting mappings satisfy the condition. This includes polynomial kernels, but also more complex functions, such as those of the form $\phi_i(\xi) = g\left(\sum_j W_{ij}^{(1)} \xi_j\right)$, where g is an activation function (e.g., ReLU or $\tanh$) and $W^{(1)}$ is any rectangular random matrix, whose entries are the weights associated to the hidden layer of the network. The latter case is relevant for the theory of extreme learning machines [HZS06].

5.4 Gardner volume and storage capacity

What is the relation between the Vapnik-Chervonenkis theory and the statistical mechanics setting we introduced at the end of Sec. 5.1? To understand it, we need to go back to analyze the solution of the linear classification problem (5.4). Clearly, a solution $W^\star$ complying with this equation is such that the quantity

$$\Delta_\mu = \frac{\sigma^\mu}{\sqrt{n}} \sum_{j=1}^{n} W_j^\star \xi_j^\mu > 0\,, \qquad \mu = 1, \cdots, p\,. \tag{5.21}$$

How many of these solutions of the classification problem exist? The answer is given by the Gardner volume [Gar87; Gar88; GD88]:

$$V_{\mathrm{G}} = \int \left[\prod_{j=1}^{n} \mathrm{d}W_j\right] \delta\left(\sum_{j=1}^{n} W_j^2 - n\right) \prod_{\mu=1}^{p} \theta\left(\frac{\sigma^\mu}{\sqrt{n}} \sum_{j=1}^{n} W_j \xi_j^\mu - \kappa\right). \tag{5.22}$$

The parameter $\kappa > 0$, that we introduced for future convenience (see Chap. 8), is called *margin*: from a geometrical perspective, it is the distance between the hyperplane realizing the dichotomy and the nearest points. The meaning of the volume V_{G} should be clear: the Heaviside theta function counts the number of configurations of the weights W_j producing, under the normalization constraint, an output concordant with the label σ^μ, for any μ (net of the margin). We can think that, if the size of the training set p, that is the number of points to classify, is small with respect of the dimension of the space n, on average it is always possible to find one or more solutions (an infinite number, in fact) of the classification problem, because typically the points are sparse in a space much larger than their number and many hyperplanes can separate them coherently with their labels. On the other side, when p becomes of the same order of n the volume of solutions shrinks until $\log V$ changes sign, so that $\log V \to -\infty$, meaning that the problem becomes unsolvable; we are interested in searching the critical value of the *load* $\alpha = p/n$ where this divergence occurs, in the limit of both n and p large.

Of course, this argument works on the *typical instance* of a random-generated dataset, meaning that we have to evaluate $\overline{\log V}$, where the average is with respect to the probability distribution of random points and labels, generated independently. In the following, we will take

$$\mathrm{d}P(\xi_j^\mu) = \frac{1}{2}\left[\delta(\xi_j^\mu - 1) + \delta(\xi_j^\mu + 1)\right] \mathrm{d}\xi_j^\mu \implies \int \prod_{\mu=1}^{p} \prod_{j=1}^{n} \mathrm{d}P(\xi_j^\mu) = \frac{1}{2^{pn}} \sum_{\{\xi_j^\mu = \pm 1\}} \tag{5.23}$$

and random labels with equal probability.

To evaluate the averaged logarithm of this volume, Gardner exploited the replica method; we retrace here her derivation for reference. Using integral identities for the theta and delta functions:

$$\begin{aligned}\theta\left(\frac{\sigma^\mu}{\sqrt{n}}\sum_{j=1}^{n} W_j\xi_j^\mu - \kappa\right) &= \int_\kappa^\infty \frac{\mathrm{d}\lambda^\mu}{2\pi}\int_{-\infty}^{+\infty}\mathrm{d}x^\mu\, e^{ix^\mu\left(\lambda^\mu - \frac{\sigma^\mu}{\sqrt{n}}\sum_j W_j\xi_j^\mu\right)},\\ \delta\left(\sum_{j=1}^{n} W_j^2 - n\right) &= \int_{-\infty}^{+\infty}\frac{\mathrm{d}E}{2\pi}e^{iE\left(\sum_j W_j^2 - n\right)},\end{aligned} \tag{5.24}$$

we can write the replicated volume as:

$$\begin{aligned}V_{\mathrm{G}}^t = \int\left[\prod_{a=1}^{t}\prod_{j=1}^{n}\mathrm{d}W_{j,a}\right]\int_\kappa^\infty\left[\prod_{a=1}^{t}\prod_{\mu=1}^{p}\frac{\mathrm{d}\lambda_a^\mu}{2\pi}\right]\int_{-\infty}^{+\infty}\left[\prod_{a=1}^{t}\prod_{\mu=1}^{p}\mathrm{d}x_a^\mu\right]\int_{-\infty}^{+\infty}\left[\prod_{a=1}^{t}\frac{\mathrm{d}E_a}{2\pi}\right]\\ \times e^{-in\sum_a E_a + i\sum_a E_a\sum_j W_{j,a}^2 + i\sum_{a,\mu}x_a^\mu\left(\lambda_a^\mu - \frac{\sigma^\mu}{\sqrt{n}}\sum_j W_{j,a}\xi_j^\mu\right)}\end{aligned} \tag{5.25}$$

The ensemble average over the binary inputs gives

$$\begin{aligned}\frac{1}{2^{pn}}\sum_{\{\xi_j^\mu=\pm1\}} e^{-i\sum_\mu\sum_j\xi_j^\mu\sum_a x_a^\mu\frac{\sigma^\mu}{\sqrt{n}}W_{j,a}} &= \prod_{\mu=1}^{p}\prod_{j=1}^{n}\cos\left(\frac{\sigma^\mu}{\sqrt{n}}\sum_{a=1}^{n}x_a^\mu W_{j,a}\right)\\ &\approx \prod_{\mu=1}^{p} e^{\sum_j\log\left(1-\frac{1}{2}\sum_{a,b}x_a^\mu x_b^\mu W_{j,a}W_{j,b}/n\right)}\\ &\approx \prod_{\mu=1}^{p} e^{-\frac{1}{2}\sum_a {x_a^\mu}^2 - \sum_{a<b}^{n}x_a^\mu x_b^\mu\frac{1}{n}\sum_j W_{j,a}W_{j,b}},\end{aligned} \tag{5.26}$$

where we used the fact that n is large, so that

$$\begin{aligned}&\frac{1}{2^{pn}}\sum_{\{\xi_j^\mu=\pm1\}}\int_\kappa^\infty\left[\prod_{a,\mu}\frac{\mathrm{d}\lambda_a^\mu}{2\pi}\right]\int_{-\infty}^{+\infty}\left[\prod_{a,\mu}\mathrm{d}x_a^\mu\right]e^{i\sum_{a,\mu}x_a^\mu\left(\lambda_a^\mu - \frac{\sigma^\mu}{\sqrt{n}}\sum_j W_{j,a}\xi_j^\mu\right)}\\ &\quad\approx\left\{\int_\kappa^\infty\left[\prod_{a=1}^{t}\frac{\mathrm{d}\lambda_a}{2\pi}\right]\int_{-\infty}^{+\infty}\left[\prod_{a=1}^{t}\mathrm{d}x_a\right]e^{i\sum_a x_a\lambda_a - \frac{1}{2}\sum_a x_a^2 - \sum_{a<b}x_a x_b\frac{1}{n}\sum_j W_{j,a}W_{j,b}}\right\}^p.\end{aligned} \tag{5.27}$$

Note that the labels σ^μ disappear in the calculation, so the average over their distribution is trivial. Introducing the replica matrix in the usual way:

$$Q_{ab} = \frac{1}{n}\sum_{j=1}^{n} W_{j,a}W_{j,b}\,, \tag{5.28}$$

we find

$$\begin{aligned}\overline{V_{\mathrm{G}}^{t}} = & \int \left[\prod_{a=1}^{t}\prod_{j=1}^{n} \mathrm{d}W_{j,a}\right] \int_{-\infty}^{+\infty} \left[\prod_{a=1}^{t} \frac{\mathrm{d}E_a}{2\pi}\right] \int_{-\infty}^{+\infty} \left[\prod_{a<b} \frac{\mathrm{d}F_{ab}\,\mathrm{d}Q_{ab}}{2\pi}\right] \\ & \times e^{-in\sum_a E_a - in\sum_{a<b} F_{ab}Q_{ab} + i\sum_a E_a \sum_j W_{j,a}^2 + i\sum_{a<b} F_{ab}\sum_j W_{j,a}W_{j,b}} \\ & \times \left\{\int_{\kappa}^{\infty} \left[\prod_{a=1}^{t} \frac{\mathrm{d}\lambda_a}{2\pi}\right] \int_{-\infty}^{+\infty} \left[\prod_{a=1}^{t} \mathrm{d}x_a\right] e^{i\sum_a x_a\lambda_a - \frac{1}{2}\sum_{a,b} x_a x_b Q_{ab}}\right\}^{p},\end{aligned} \tag{5.29}$$

were the auxiliary variables F_{ab} are Lagrange multipliers enforcing the definition (5.28). Now we can perform the Gaussian integration over the weights:

$$\int \left[\prod_{a=1}^{t}\prod_{j=1}^{n} \mathrm{d}W_{j,a}\right] e^{i\sum_a E_a \sum_j W_{j,a}^2 + i\sum_{a<b} F_{ab}\sum_j W_{j,a}W_{j,b}} = e^{-\frac{n}{2}\log\det(-i\boldsymbol{G}) + \frac{nt}{2}\log(2\pi)}, \tag{5.30}$$

where $\boldsymbol{G}$ is the symmetric matrix

$$G_{ab} = 2E_a\delta_{ab} - (1-\delta_{ab})F_{ab}\,. \tag{5.31}$$

The integral over the elements of $\boldsymbol{G}$ is performed via the saddle-point method. Ignoring all the factors suppressed in n,

$$\int_{-\infty}^{+\infty} \left[\prod_{a=1}^{t} \frac{\mathrm{d}G_{aa}}{4\pi}\right] \int_{-\infty}^{+\infty} \left[\prod_{a<b} \frac{\mathrm{d}G_{ab}}{2\pi}\right] e^{-\frac{n}{2}\sum_{a,b} iG_{ab}Q_{ab} - \frac{n}{2}\log\det(-i\boldsymbol{G})} \sim e^{\frac{nt}{2} + \frac{n}{2}\log\det(\boldsymbol{Q})} \tag{5.32}$$

where we used the saddle-point equations

$$\frac{\partial}{\partial G_{ab}} \left[\sum_{c,d} iG_{cd}Q_{cd} + \log\det\left(-i\boldsymbol{G}\right)\right] = iQ_{ab} + \left(\boldsymbol{G}^{-1}\right)_{ba} = 0\,. \tag{5.33}$$

Eventually, the replicated volume is:

$$\begin{aligned}\overline{V_{\mathrm{G}}^{t}} = & \int_{-\infty}^{+\infty} \left[\prod_{a<b} \mathrm{d}Q_{ab}\right] e^{\frac{nt}{2}[1+\log(2\pi)] + \frac{n}{2}\log\det(\boldsymbol{Q})} \\ & \times \left\{\int_{\kappa}^{\infty} \left[\prod_{a=1}^{t} \frac{\mathrm{d}\lambda_a}{2\pi}\right] \int_{-\infty}^{+\infty} \left[\prod_{a=1}^{t} \mathrm{d}x_a\right] e^{i\sum_a x_a\lambda_a - \frac{1}{2}\sum_{a,b} x_a x_b Q_{ab}}\right\}^{p}.\end{aligned} \tag{5.34}$$

Gardner also proved that the correct ansatz for the form of the replica matrix is the RS one:

$$Q_{ab} = (1-q)\delta_{ab} + q\,, \qquad 0 \le q \le 1\,. \tag{5.35}$$

The measure part gives

$$\log\det(\boldsymbol{Q}) = \log\left[(1-q)^{n-1}\,(1-q+nq)\right] = n\log(1-q) + \frac{nq}{1-q}\,, \tag{5.36}$$

while the term inside the curly brackets can be solved via a HS transformation. After some manipulations (see [Nis01]),

$$\int_{\kappa}^{\infty}\left[\prod_{a=1}^{t}\frac{\mathrm{d}\lambda_a}{2\pi}\right]\int_{-\infty}^{+\infty}\left[\prod_{a=1}^{t}\mathrm{d}x_a\right]e^{i\sum_a x_a\lambda_a-\frac{1}{2}\sum_{a,b}x_a x_b Q_{ab}}\propto\int\mathrm{D}y\left\{\mathrm{erfc}\left[\frac{\kappa+y\sqrt{q}}{\sqrt{2(1-q)}}\right]\right\}^t, \tag{5.37}$$

where y is the auxiliary HS variable to be integrated in the Gaussian measure $\mathrm{D}y$, and erfc is the complementary error function. Exponentiating this result and taking the limit $t\to 0$ of the replica trick, as we are interested in the typical properties of this observable, we find

$$p\log\int\mathrm{D}y\left\{\mathrm{erfc}\left[\frac{\kappa+y\sqrt{q}}{\sqrt{2(1-q)}}\right]\right\}^t\approx pt\int\mathrm{D}y\log\left\{\mathrm{erfc}\left[\frac{\kappa+y\sqrt{q}}{\sqrt{2(1-q)}}\right]\right\}. \tag{5.38}$$

We can proceed with the following argument, recalling that we are searching for the point α_c where the volume of solutions of the classification problem shrink to zero, in the setting of replica theory: as we are approaching this point, the different replicas become more and more correlated, because they can be chosen from a smaller and smaller set: at the transition, it must be $q\to 1$. In this way, we can exploit the asymptotic expansion of the complementary error function to sum this term with the measure part we evaluated before, and we can simply search for the point $\alpha=p/n$ where the most divergent part in $(1-q)^{-1}$ changes sign. The Gardner's result is

$$\alpha_c(\kappa)=\left[\int_{-\kappa}^{+\infty}\mathrm{D}y\,(\kappa+y)^2\right]^{-1}. \tag{5.39}$$

This critical value of the ratio p/n, for both p and n large, is called *storage capacity*. For $\kappa=0$, we find

$$\alpha_c(0)=2\,. \tag{5.40}$$

We will say more on the problem of margin learning in Chap. 8, where we will explain its connection with the emergence of the satisfiability phase transition driven by data structure we will explain in Chap. 7.

5.4.1 Connection with Cover's result

It is straightforward to see how the storage capacity α_c can be obtained from $C_{n,p}$ in Eq. (5.18), which is a quantity defined at finite n and p. The number of dichotomies is a combinatorial quantity, and is expected to scale exponentially in n, at least for small α. Thus, an intensive quantity can be defined by normalizing $C_{n,p}$ with the total number of dichotomies of p points. The fraction of dichotomies $c_{n,p}\equiv C_{n,p}/2^p$ is bounded, $0\le c_{n,p}\le 1$, and has a non-trivial thermodynamic limit $c_\infty(\alpha)$. The thermodynamic limit is defined by taking both $n,p\to\infty$, with fixed $\alpha=p/n$. It is not hard to see directly from Eq. (5.18) that

$$c_\infty(\alpha)=\theta\left(\alpha_\mathrm{c}-\alpha\right), \tag{5.41}$$

with $\alpha_\mathrm{c}=2$. The expression in Eq. (5.41) takes the value 1 for $\alpha<\alpha_\mathrm{c}$, the value 0 for $\alpha>\alpha_\mathrm{c}$, and the value $1/2$ for $\alpha=\alpha_\mathrm{c}$ (θ is the Heaviside step function). Qualitatively, $c_{n,\alpha n}$ as a function of α is a decreasing sigmoid, which is steeper for larger values of n

(see Fig. 7.1a). This allows the definition of a notion of capacity at finite dimension n, as the value $\tilde{\alpha}_{\rm c}(n)$ such that $c_{n,\tilde{\alpha}_{\rm c}(n)n} = 1/2$, or

$$C_{n,\tilde{\alpha}_{\rm c}(n)n} = 2^{p-1}. \tag{5.42}$$

Another notable value of p can be read off of $c_{n,p}$: it is the Vapnik-Chervonenkis dimension $d_{\rm VC}$, equal to the maximum p such that $c_{n,p} = 1$. For a linear separator, $d_{\rm VC} = n$. Notice that one cannot use the asymptotic form Eq. (5.41) to this aim, since the thermodynamic limit pushes $c_{n,\alpha n}$ to 1 for all values of α up to $\alpha_{\rm c}$.

CHAPTER 6

A geometrical model of data structure

The discussion in the previous chapter suggests that, in order to go beyond the prediction of Cover's theorem, one needs a way of introducing statistical dependence between the inputs X_p and their labels $Y_p = (Y^1, \dots, Y^p)$. This reflects a simple observation that can be made on empirical datasets of images: similar inputs tend to be classified similarly. For instance, one expects that there exists an (unknown) set of transformations on an input image X, possibly including some translations, dilations, and rotations, that leave the classification of X invariant. Such intuition agrees with the concepts, put forward in neuroscience and gaining momentum in physics, of invariant recognition (the similar neural representation of the same object in different conditions) and object manifolds (sets of input stimuli giving rise to the same neural representation) [Coh+20; CLS18; CLS16; Ans+16; SL00].

Integrating data structure within the framework of statistical mechanics is relatively straightforward and usually follows two steps: (i) define a generative model for the data, given in terms of a non-factorized joint probability distribution $P(X_p, Y_p)$; (ii) compute averages over the measure P (the "disorder"); this is what was done for instance in [Bor+19; CLS18; CLS18; Ger+20]. How to best address data dependence in the SLT formalism, instead, is a debated issue. Here we follow a simple strategy inspired by recent literature in statistical physics: we change the input space $\mathcal{X}$. Each input X^μ is now an object manifold, i.e., a (possibly countably or uncountably infinite) set of points that, by definition, are be classified coherently.

In this chapter we focus on a simple realizations of data structure, introduced in [Bor+19; RLG20], and motivated by the availability of analytical results.

6.1 Simplex learning

In the framework we present here, inputs are "multiplets" of k points with fixed geometric interrelations. The input set is $X_p = \{X^\mu\}_{\mu=1,\dots,p}$, where each $X^\mu = \{\xi_a^\mu\}_{a=1,\dots,k}$ is a set of k points defined in one of the following spaces:

(i) on the unit $(n-1)$-sphere, $\xi_a^\mu \in S^{n-1}$:

$$\xi_a^\mu \cdot \xi_a^\mu = 1 \qquad \text{for all } \; a = 1, \cdots, k \,; \mu = 1, \cdots, p \,, \tag{6.1}$$

where the dot indicates the usual scalar product in $\mathbb{R}^n$;

(ii) on the vertices of a n-dimensional hypercube of edge 2:

$$\xi_{a,j}^\mu \in \{\pm 1\} \qquad \text{for all } \; a = 1, \cdots, k \,; \quad \mu = 1, \cdots, p \,; \quad j = 1, \cdots, n \,, \tag{6.2}$$

a convention we will adopt in the replica calculations to mirror closely the original Gardner's approach we explained in section 5.4.

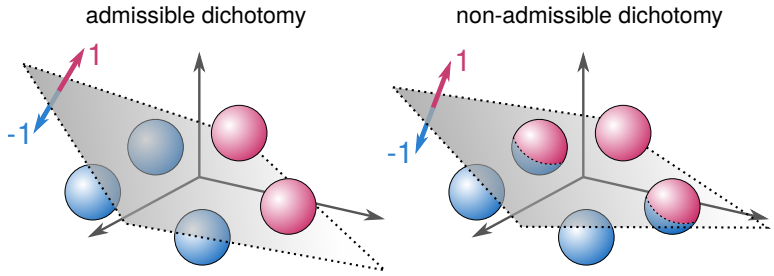

Figure 6.1: Input data are structured as groups of points sharing the same label (pink $= +1$, blue $= -1$). Each sphere denotes, in a stylized way, a group of points. Singly colored spheres contribute to admissible dichotomies; conversely, a dichotomy containing a doubly colored sphere is not admissible. Figure from [RPG20].

In both cases, the $k(k-1)/2$ overlaps within each multiplet are fixed:

$$\xi_a^\mu \cdot \xi_b^\mu = \rho_{ab} \qquad \text{for all } \mu = 1, \cdots, p\,. \tag{6.3}$$

Note that $-1 \leq \rho_{ab} \leq 1$. We assume the uniform probability measure on each point ξ_a^μ, conditioned on the constraint on the overlaps. The usual unconstrained ensemble of points is recovered for $k = 1$, or at any k if $\rho_{ab} = 1$ for all a, b. The name "simplex" is justified by the fact that, since linear classification is a projective problem, if $Y = g(X)$ for each X in a set of points X^μ, then $Y = g(X)$ for all X in the convex hull of X^μ. The input space $\mathcal{X}_S(\{\rho_{ab}\})$ depends on k and ρ_{ab}, and is the set of all multiplets with the given constraints.

Similar dataset have been recently proposed in literature. In [CLS18], a model of segments with fixed Euclidean length R and random Gaussian centers and orientations is introduced, resembling our $k = 2$ case once the endpoints of each segment are normalized to the unit sphere and their overlaps is averaged over their probability distribution. In [FHU19], the correlations between the points are fixed component-wise, instead that using their scalar product.

However, the ensemble introduced in this section is promising mainly because, very recently, the combinatorial approach introduced by Cover was extended to formulate a mean field theory of simplex learning [RLG20]. The definition we have given above specifies the ensemble of the sets X_p; it remains to define the hypothesis class $\mathcal{G}_M(\{\rho_{ab}\})$. This is straightforward: one starts from the class $\mathcal{G}$ of linear separators in $\mathbb{R}^n$ and restricts it to the class $\hat{\mathcal{G}}(\{\rho_{ab}\})$ of those functions $h \in \mathcal{G}$ that assign the same label to all points in each multiplet X^μ (i.e., those that are constant on each multiplet). Then the restricted hypothesis class is defined as

$$\mathcal{G}_M(\{\rho_{ab}\}) = \Big\{ g : \exists\, h \in \hat{\mathcal{G}}(\{\rho_{ab}\}) \quad \text{s.t.} \quad \forall\, X^\mu \in X_p,\, g(X^\mu) = h(\xi \in X^\mu) \Big\}. \tag{6.4}$$

The functions in $\hat{\mathcal{G}}(\{\rho_{ab}\})$ are called admissible (see Fig. 6.1). The mean-field combinatorial theory allows the computation of the average $\langle \mathcal{N}_{\mathbb{1}}(X_p) \rangle_{X_p}$, i.e., the average number of admissible dichotomies of simplexes that can be realized linearly. We will still denote this number with $C_{n,p}$, although it depends on the parameters k and $\{\rho_{ab}\}$ of the ensemble.

The quantities $C_{n,p}$ satisfy a recurrence relation

$$C_{n,p+1} = \sum_{l=0}^{k} \theta_l^k C_{n-l,p}, \tag{6.5}$$

where the constant coefficients θ_l^k are fixed in turn by the recurrence relation

$$\theta_l^k = \psi_k \theta_l^{k-1} + (1-\psi_k)\, \theta_{l-1}^{k-1}, \tag{6.6}$$

with boundary conditions

$$\theta_0^1 = \theta_1^1 = 1\,, \qquad \theta_{-1}^k = \theta_{k+1}^k = 0\,. \tag{6.7}$$

The boundary conditions for Eq. (6.5) are difficult to express precisely for generic k. Here we will assume the boundary conditions in Eq. (5.20) for all k. This approximation is expected to have a negligible effect for the asymptotic analysis presented in the next chapter; we checked the validity of this approximation numerically for the first non-trivial cases $k = 2$ and $k = 3$. Each coefficient θ_l^k in Eq. (6.5) depends on $k-1$ numbers $\{\psi_m\}_{m=2,\dots,k}$, with $0 \le \psi_m \le 1$, having the following geometric-probabilistic interpretation. Let $w \in S^{n-1}$ be a random vector with the flat measure on the unit sphere. Consider any multiplet X^μ, and a subset $X' \subseteq X^\mu$ of $m \le k$ points. Then ψ_m is the symmetrized probability that the scalar product $w \cdot \xi$ has the same sign for all $\xi \in X'$, conditioned on it having the same sign for all $\xi \in X' \setminus \{\xi_\star\}$:

$$\psi_m = 2 \left\langle \mathrm{P}\left[(w\cdot\xi_\star) > 0 \mid (w\cdot\xi) > 0 \;\forall \xi \in X' \setminus \{\xi_\star\}\right] \right\rangle_{\mathrm{sym}},$$

where the symmetrization $\langle\cdot\rangle_{\mathrm{sym}}$ is performed by averaging over all subsets X' and over all choices of $\xi_\star \in X'$. These quantities can be expressed in terms of the overlaps ρ_{ab}, e.g.,

$$\psi_2(\rho) = \frac{2}{\pi} \arctan \sqrt{\frac{1+\rho}{1-\rho}}\,. \tag{6.8}$$

The explicit solution of the recurrence (6.5) was given in [RLG20], to which we address the reader for more details on ψ_m and $C_{n,p}$ for generic k; in the next chapter, we will see how the methods of analytic combinatorics can be used to obtain the asymptotic behavior of $C_{n,p}$ starting from the recurrence relations (6.5) and (6.6). Here, we only report the solution for doublets ($k = 2$):

$$C_{n,p} = 2\sum_{i=0}^{n-2} K_{i,p} + 2\psi_2(\rho) K_{n-1,p} \qquad \boxed{k=2}\,, \tag{6.9}$$

with

$$K_{i,p} = \sum_{m=0}^{p-1} \binom{p-1}{m, i-2m} \psi_2(\rho)^{p-1-i+m} \left[1-\psi_2(\rho)\right]^m \qquad \boxed{k=2}\,, \tag{6.10}$$

and the multinomial coefficient is defined as

$$\binom{n}{m_1, m_2} = \frac{n!}{m_1! m_2! (n-m_1-m_2)!}\,. \tag{6.11}$$

Moreover, the notion of storage capacity α_c can be defined also for structured data, similarly to Cover's unstructured case we discussed in section 5.4.1, Eq. (5.42); the combinatorial theory yields

$$\alpha_c(k) = \left(k - \frac{1}{2} - \sum_{l=2}^{k} \psi_l \right)^{-1} . \tag{6.12}$$

In the rest of this chapter, we will explain how to obtain this quantity in the replica framework.

6.2 Storage capacity for multiplets

In this section we retrace, in a similar fashion of our future replica calculations in Chap. 7, the evaluation of α_c for the linear classification of simplexes, performed in [Bor+19]. We unify the notation and give full details on the steps of the derivation.

6.2.1 Replicated Gardner volume for multiplets

From the above discussion, it should be clear how to generalize the Gardner volume (5.22) to count the number of ways a linear classifier can assign correctly the labels on the multiplets in the training set. As each point in a multiplet must have the same label, this volume is

$$V_{\mathrm{G}}(X_p) = \int \left[\prod_{j=1}^{n} \mathrm{d}W_j \right] \delta \left(\sum_{j=1}^{n} W_j^2 - n \right) \prod_{\mu=1}^{p} \prod_{a=1}^{k} \theta \left(\frac{\sigma^\mu}{\sqrt{n}} \sum_{j=1}^{n} W_j \xi_{a,j}^\mu \right) , \tag{6.13}$$

where $\theta(\cdot)$ is the Heaviside theta, $\xi_{a,j}^\mu$ denotes the j-th component of the a-th element of the μ-th multiplet and the weights lie on the surface of a n-dimensional sphere of radius $\sqrt{n}$. Note indeed that data structure is implemented in Eq. (6.13) by asking that each point of the μth simplex be labeled by σ^μ. The inputs, constituting the set X_p, are chosen randomly according to the distribution

$$\mathrm{d}P(X_p) = \nu^{-1} \prod_{\mu=1}^{p} \prod_{a=1}^{k} \prod_{b=1}^{a-1} \delta \left(\rho_{ab} - \frac{1}{n} \sum_{j=1}^{n} \xi_{a,j}^\mu \xi_{b,j}^\mu \right) \prod_{j=1}^{n} \left[\delta(\xi_{a,j}^\mu - 1) + \delta(\xi_{a,j}^\mu + 1) \right] \mathrm{d}\xi_{a,j}^\mu , \tag{6.14}$$

where $-1 \leq \rho_{ab} \leq 1$ are the overlaps, ν is a normalization factor and the inputs lie on the vertices of a n-dimensional hypercube.

In the following, to make the calculations practicable, we focus on the case $k = 2$ (data in pairs, or "doublets"). We adopt the special notation $X^\mu = \{\xi^\mu, \bar{\xi}^\mu\}$, reserving the early latin indices for the replicas. The volume becomes

$$V_{\mathrm{G}} = \int \left[\prod_{j=1}^{n} \mathrm{d}W_j \right] \delta \left(\sum_{j=1}^{n} W_j^2 - n \right) \prod_{\mu=1}^{p} \theta \left(\frac{\sigma^\mu}{\sqrt{n}} \sum_{j=1}^{n} W_j \xi_j^\mu \right) \theta \left(\frac{\sigma^\mu}{\sqrt{n}} \sum_{j=1}^{n} W_j \bar{\xi}_j^\mu \right) . \tag{6.15}$$

Using the standard integral representations for the delta and theta functions (5.24), we

can write

$$V_{\rm G} = \int \left[\prod_{j=1}^{n} \mathrm{d}W_j\right] \int_0^{+\infty} \left[\prod_{\mu=1}^{p} \frac{\mathrm{d}\lambda^\mu \, \mathrm{d}\bar{\lambda}^\mu}{(2\pi)^2}\right] \int_{-\infty}^{+\infty} \left[\prod_{\mu=1}^{p} \mathrm{d}x^\mu \, \mathrm{d}\bar{x}^\mu\right] \int_{-\infty}^{+\infty} \frac{\mathrm{d}E}{2\pi} \\ \times \; e^{iE\left(\sum_j W_j^2 - n\right) + i\sum_\mu x^\mu \left(\lambda^\mu - \frac{\sigma^\mu}{\sqrt{n}} \sum_j W_j \xi_j^\mu\right) + i \sum_\mu \bar{x}^\mu \left(\bar{\lambda}^\mu - \frac{\sigma^\mu}{\sqrt{n}} \sum_j W_j \bar{\xi}_j^\mu\right)} \,, \tag{6.16}$$

where the auxiliary variable E enforces the spherical constraint, while the integral representation of the theta function is obtained via the auxiliary variables λ, x. Note that we can always redefine $\xi^\mu \to \sigma^\mu \xi^\mu$, $\bar{\xi}^\mu \to \sigma^\mu \bar{\xi}^\mu$ (the rest of the integral an the probability measure (6.14) are invariant) to get rid of the labels. Replicating t times this volume, we find

$$V_{\rm G}^t = \int \left[\prod_{a=1}^{t} \prod_{j=1}^{n} \mathrm{d}W_{j,a}\right] \int_{-\infty}^{+\infty} \left[\prod_{a=1}^{t} \frac{\mathrm{d}E_a}{2\pi}\right] \int_{-\infty}^{+\infty} \left[\prod_{a<b} \frac{\mathrm{d}F_{ab} \, \mathrm{d}Q_{ab}}{2\pi}\right] \\ \times \, e^{i\sum_a E_a \left(\sum_j W_{j,a}^2 - n\right) + i \sum_{a<b} F_{ab} \left(\sum_j W_{j,a} W_{j,b} - n Q_{ab}\right)} \int_0^{+\infty} \left[\prod_{a=1}^{t} \prod_{\mu=1}^{p} \frac{\mathrm{d}\lambda_a^\mu \, \mathrm{d}\bar{\lambda}_a^\mu}{(2\pi)^2}\right] \\ \times \int_{-\infty}^{+\infty} \left[\prod_{a=1}^{t} \prod_{\mu=1}^{p} \mathrm{d}x_a^\mu \, \mathrm{d}\bar{x}_a^\mu\right] e^{i\sum_{a,\mu} x_a^\mu \left(\lambda_a^\mu - \frac{1}{\sqrt{n}} \sum_j W_{j,a} \xi_j^\mu\right) + i \sum_{a,\mu} \bar{x}_a^\mu \left(\bar{\lambda}_a^\mu - \frac{1}{\sqrt{n}} \sum_j W_{j,a} \bar{\xi}_j^\mu\right)} \,, \tag{6.17}$$

where $1 \le a, b \le t$ are replica indices (not to be confused with the indices running inside the multiplets, a notation we abandoned at the beginning of this section, when we specialized our calculation to doublets), Q_{ab} is the replica matrix (with $Q_{aa} = 1$) and F_{ab} are the Lagrange multipliers enforcing the constraint

$$Q_{ab} = \frac{1}{n} \sum_{j=1}^{n} W_{j,a} W_{j,b} \,. \tag{6.18}$$

Eq. (6.17) is the starting point for the quenched computation of the storage capacity for doublets.

6.2.2 Averaging over the input distribution

To perform the average over the input ensemble, we observe from Eq. (6.14), specialized to $k = 2$, that at fixed overlap ρ, given $c, d \in \mathbb{N}$ the numbers of concordant and discordant signs of the components of the pair for each μ, then $c - d = \rho n$, $c + d = n$, so

$$c = (1 + \rho) n/2 \,, \qquad d = (1 - \rho) n/2 \,. \tag{6.19}$$

For each μ, we can freely choose in 2^n different ways the components of ξ^μ, but then for $\bar{\xi}^\mu$ we must take c components with the same sign of their counterparts and d with the opposite. We can do that in $\binom{n}{c}$ different ways, so the normalization factor is

$$\nu = 2^{pn} \binom{n}{\frac{(1+\rho)n}{2}}^p . \tag{6.20}$$

However, the order of the components of the vectors ξ, $\bar{\xi}$ is completely irrelevant, because they appear only in scalar products, among themselves (in the overlap constraint)

and with the same vector W, whose components again we are free to relabel. This means that we can choose as a representative of the vector $\bar{\xi}$, for example, the one with the concordant components at the beginning. We can write the ensemble measure as

$$\mathrm{d}P_\rho(\Xi) = \prod_{\mu=1}^{p} \left[\prod_{j=1}^{c} \mathrm{d}P(\xi_j^\mu) \delta(\xi_j^\mu - \bar{\xi}_j^\mu)\, \mathrm{d}\bar{\xi}_j^\mu \right] \left[\prod_{j=c+1}^{n} \mathrm{d}P(\xi_j^\mu) \delta(\xi_j^\mu + \bar{\xi}_j^\mu)\, \mathrm{d}\bar{\xi}_j^\mu \right], \tag{6.21}$$

where

$$\mathrm{d}P(\xi_j^\mu) = \frac{1}{2}\left[\delta(\xi_j^\mu - 1) + \delta(\xi_j^\mu + 1)\right] \mathrm{d}\xi_j^\mu \,. \tag{6.22}$$

Note that with the choice of a representative we are explicitly breaking the invariance of the original expression under permutation (relabeling) of the indices j, a symmetry we will reintroduce by hand in the following calculation.

We can now perform the averages of the volume (6.17). Isolating the only part depending on the inputs in the integrand, we find

$$\begin{aligned}
&\int \mathrm{d}P_\rho(\Xi)\, e^{-i\sum_{a,\mu} x_a^\mu \sum_j \frac{\xi_j^\mu W_{j,a}}{\sqrt{n}} - i\sum_{a,\mu} \bar{x}_a^\mu \sum_j \frac{\bar{\xi}_j^\mu W_{j,a}}{\sqrt{n}}} \\
&= \prod_{\mu=1}^{p} \prod_{j=1}^{c} \cos\left[\frac{1}{\sqrt{n}} \sum_a (x_a^\mu + \bar{x}_a^\mu) W_{j,a}\right] \prod_{j=c+1}^{n} \cos\left[\frac{1}{\sqrt{n}} \sum_a (x_a^\mu - \bar{x}_a^\mu) W_{j,a}\right] \\
&\approx \prod_{\mu=1}^{p} e^{-\frac{1}{2}\sum_{a,b}\left[x_a^\mu x_b^\mu \sum_{j=1}^{n} \frac{W_{j,a}W_{j,b}}{n} + \bar{x}_a^\mu \bar{x}_b^\mu \sum_{j=1}^{n} \frac{W_{j,a}W_{j,b}}{n} + 2x_a^\mu \bar{x}_b^\mu \left(\sum_{j=1}^{c} - \sum_{j=c+1}^{n}\right) \frac{W_{j,a}W_{j,b}}{n}\right]},
\end{aligned} \tag{6.23}$$

where, in the final step, a large n expansion is performed. The last term at the exponent, consisting in a sum over j that does not extend over all the n components, cannot be readily solved inserting the replica matrix, but we can write it as

$$\left(\sum_{j=1}^{c} - \sum_{j=c+1}^{n}\right) \frac{W_{j,a}W_{j,b}}{n} = \left(2\sum_{j=1}^{c} - \sum_{j=1}^{n}\right) \frac{W_{j,a}W_{j,b}}{n} \,. \tag{6.24}$$

Now, only the first sum is not invariant under permutations of the components. However, since the starting point was symmetric, we can also multiply this expression by similar ones obtained with other choices of the vector $\bar{\xi}^\mu$, and then take the corresponding root of the result, obtaining an equivalent formula. The trick to restore a complete sum over the n components, is to multiply by all the c-permutations of n, and then take the $n!/(n-c)!$-th root of the result. The only non-trivial term at the exponent during this procedure is indeed the partial sum, which reads:

$$\frac{(n-c)!}{c!} \sum_{j=1}^{c} \sum_{\substack{\pi_1 \neq \pi_2 \neq \cdots \neq \pi_c \\ \forall i,\, 1 \leq \pi_i \leq n}} \frac{W_{\pi_j,a} W_{\pi_j,b}}{n} = \frac{c}{n} \sum_{i=1}^{n} \frac{W_{i,a}W_{i,b}}{n} \,. \tag{6.25}$$

Now we can insert the replica matrix (6.18) in all terms. Using $(2c/n - 1) = \rho$, and factorizing the p integrals over the auxiliary variables x and λ, we obtain, for the x and λ integrals,

$$\left\{ \int_0^{+\infty} \left[\prod_{a=1}^{t} \frac{\mathrm{d}^2 \boldsymbol{\lambda}_a}{(2\pi)^2}\right] \int_{-\infty}^{+\infty} \left[\prod_{a=1}^{t} \mathrm{d}^2 \boldsymbol{x}_a\right] e^{-\frac{1}{2}\sum_{a,b} Q_{ab} \boldsymbol{x}_a^T \mathcal{R} \boldsymbol{x}_b + i\sum_a \boldsymbol{x}_a^T \boldsymbol{\lambda}_a} \right\}^p, \tag{6.26}$$

where we already inserted the replica matrix using (6.18) and we introduced the notation

$$\boldsymbol{x} = \begin{pmatrix} x \\ \bar{x} \end{pmatrix}, \quad \boldsymbol{\lambda} = \begin{pmatrix} \lambda \\ \bar{\lambda} \end{pmatrix}, \quad \mathcal{R} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \tag{6.27}$$

The remaining integrals can be performed in the same way as to write Eq. (5.34) in Gardner's unconstrained calculation, so the resulting averaged replicated volume to be evaluated is

$$\begin{aligned} \overline{V_{\mathrm{G}}^t} = & \int_{-\infty}^{+\infty} \left[\prod_{a<b} \mathrm{d}Q_{ab} \right] e^{\frac{nt}{2} + \frac{n}{2} \log\det(\boldsymbol{Q})} \\ & \times \left\{ \int_0^{+\infty} \left[\prod_{a=1}^{t} \frac{\mathrm{d}^2 \boldsymbol{\lambda}_a}{(2\pi)^2} \right] \int_{-\infty}^{+\infty} \left[\prod_{a=1}^{t} \mathrm{d}^2 \boldsymbol{x}_a \right] e^{-\frac{1}{2} \sum_{a,b} Q_{ab} \boldsymbol{x}_a^T \mathcal{R} \boldsymbol{x}_b + i \sum_a \boldsymbol{x}_a^T \boldsymbol{\lambda}_a} \right\}^p . \end{aligned} \tag{6.28}$$

We cannot proceed further, in taking the limit $t \to 0$ as prescribed by the replica approach, without making an ansatz on the form of the replica matrix Q_{ab}. Following Gardner, we impose the RS ansatz, in which the replica matrix has the form

$$Q_{ab} = (1-q)\delta_{ab} + q\,, \qquad 0 \le q \le 1\,, \tag{6.29}$$

so that

$$\log\det(\boldsymbol{Q}) \quad \underset{t\to 0}{\longrightarrow} \quad t\log(1-q) + \frac{tq}{1-q}\,. \tag{6.30}$$

The quadratic form at the exponent of Eq. (6.28) reads

$$\sum_{a,b} Q_{ab} \boldsymbol{x}_a^T \mathcal{R} \boldsymbol{x}_b = (1-q) \sum_a \boldsymbol{x}_a^T \mathcal{R} \boldsymbol{x}_a + q \left(\sum_a \boldsymbol{x}_a \right)^T \mathcal{R} \left(\sum_b \boldsymbol{x}_b \right) . \tag{6.31}$$

The last term can be linearized with a HS transformation:

$$e^{-\frac{q}{2} \left[\sum_a \boldsymbol{x}_a\right]^T \mathcal{R} \left[\sum_b \boldsymbol{x}_b\right]} = \int_{-\infty}^{+\infty} \frac{\mathrm{d}^2 \boldsymbol{y}}{2\pi\sqrt{1-\rho^2}}\, e^{-\frac{1}{2}\boldsymbol{y}^T \mathcal{R}^{-1} \boldsymbol{y} + i\sqrt{q} \sum_a \boldsymbol{x}_a^T \boldsymbol{y}}\,. \tag{6.32}$$

so that replica indices factorize, to get, after an integration over $\boldsymbol{x}$,

$$\begin{aligned} & \left\{ \int_{-\infty}^{+\infty} \frac{\mathrm{d}^2 \boldsymbol{y}}{2\pi\sqrt{1-\rho^2}}\, e^{-\frac{1}{2}\boldsymbol{y}^T \mathcal{R}^{-1} \boldsymbol{y}} \right. \\ & \qquad \left. \times \left[\frac{2\pi}{(1-q)\sqrt{1-\rho^2}} \int_0^{+\infty} \frac{\mathrm{d}^2 \boldsymbol{\lambda}}{(2\pi)^2} e^{-\frac{1}{2(1-q)} (\boldsymbol{\lambda} + \sqrt{q}\boldsymbol{y})^T \mathcal{R}^{-1} (\boldsymbol{\lambda} + \sqrt{q}\boldsymbol{y})} \right]^t \right\}^p . \end{aligned} \tag{6.33}$$

Defining $L_{\mathrm{G}}(\boldsymbol{y})$ the quantity in square brackets, the limit $t \to 0$ gives

$$\begin{aligned} p \log & \left\{ \int_{-\infty}^{+\infty} \frac{\mathrm{d}^2 \boldsymbol{y}}{2\pi\sqrt{1-\rho^2}}\, e^{-\frac{1}{2}\boldsymbol{y}^T \mathcal{R}^{-1} \boldsymbol{y}} \left[L_{\mathrm{G}}(\boldsymbol{y}) \right]^t \right\} \\ & \to pt \int_{-\infty}^{+\infty} \frac{\mathrm{d}^2 \boldsymbol{y}}{2\pi\sqrt{1-\rho^2}}\, e^{-\frac{1}{2}\boldsymbol{y}^T \mathcal{R}^{-1} \boldsymbol{y}} \log\left[L_{\mathrm{G}}(\boldsymbol{y}) \right] . \end{aligned} \tag{6.34}$$

We can now search for the critical value of the load $\alpha = p/n$ in the limit $q \to 1$, in the spirit of Gardner's argument we used to write Eq. (5.39). In this limit, we can evaluate the integral in $\mathrm{d}^2\boldsymbol{\lambda}$ via a saddle-point analysis: we can write it as

$$\int_0^{+\infty} \mathrm{d}^2\boldsymbol{\lambda}\, e^{-\frac{1}{2(1-q)} f(\boldsymbol{\lambda})}\,,$$

where

$$f(\boldsymbol{\lambda}) = \frac{1}{1-\rho^2}\left[(\lambda+y)^2 + (\bar{\lambda}+\bar{y})^2 - 2\rho(\lambda+y)(\bar{\lambda}+\bar{y})\right] \tag{6.35}$$

and $1/(1-q)$ is the large parameter of the asymptotic expansion. The stationary point of the function f is

$$\lambda + y = 0\,, \qquad \bar{\lambda} + \bar{y} = 0\,.$$

If this point is in the domain of integration, it is clearly the minimum of $f(\boldsymbol{\lambda})$, otherwise we must choose the lowest limit of integration ($\lambda = 0$ or $\bar{\lambda} = 0$). In practice:

- if both $y < 0$, $\bar{y} < 0$, then $\lambda^\star = -y$, $\bar{\lambda}^\star = -\bar{y}$ and

  $$f(\boldsymbol{\lambda}^\star) = 0\,.$$

- If $y > 0$, then $\lambda^\star = 0$. To find $\bar{\lambda}^\star$ we have to minimize the function

  $$f(0,\bar{\lambda}) = \frac{1}{1-\rho^2}\left[y^2 + (\bar{\lambda}+\bar{y})^2 - 2\rho y(\bar{\lambda}+\bar{y})\right]$$

  with respect to $\bar{\lambda}$, obtaining

  $$\bar{\lambda} + \bar{y} - \rho y = 0\,.$$

  This equation admits a solution in the domain $\bar{\lambda} > 0$ only if $\bar{y} < \rho y$, in which case $\bar{\lambda}^\star = \rho y - \bar{y}$ and

  $$f(\boldsymbol{\lambda}^\star) = y^2\,.$$

  Otherwise we must take $\bar{\lambda}^\star = 0$ and

  $$f(\boldsymbol{\lambda}^\star) = f(0,0) = \frac{1}{1-\rho^2}\left[y^2 + \bar{y}^2 - 2\rho y\bar{y}\right]\,.$$

- If $\bar{y} > 0$, then $\bar{\lambda}^\star = 0$. To find $\lambda^\star$ we have to minimize the functional

  $$f(\lambda,0) = \frac{1}{1-\rho^2}\left[(\lambda+y)^2 + \bar{y}^2 - 2\rho\bar{y}(\lambda+y)\right]$$

  with respect to λ, obtaining

  $$\lambda + y - \rho\bar{y} = 0$$

  This equation admits a solution in the domain $\lambda > 0$ only if $y < \rho\bar{y}$, in which case $\lambda^\star = \rho\bar{y} - y$ and

  $$f(\boldsymbol{\lambda}^\star) = \bar{y}^2\,.$$

  Otherwise we must take $\lambda^\star = 0$ and again $f(\boldsymbol{\lambda}^\star) = f(0,0)$.

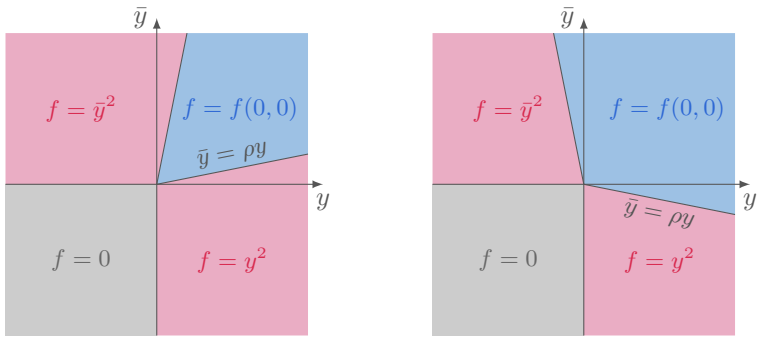

Figure 6.2: Value of the saddle-point of the function f in Eq. (6.35), for $\rho > 0$ (left) and for $\rho < 0$ (right).

These results are summarized in Fig 6.2. Eventually, we find

$$\begin{aligned}
&\frac{2\pi}{(1-q)\sqrt{1-\rho^2}} \int_0^{+\infty} \frac{\mathrm{d}^2\boldsymbol{\lambda}}{(2\pi)^2} e^{-\frac{1}{2(1-q)}(\boldsymbol{\lambda}+\sqrt{q}\boldsymbol{y})^T \mathcal{R}^{-1}(\boldsymbol{\lambda}+\sqrt{q}\boldsymbol{y})} \\
&\sim \theta(-y)\theta(-\bar{y}) + \theta(y)\theta(\rho y - \bar{y}) \frac{e^{-\frac{y^2}{2(1-q)}}}{y} \sqrt{\frac{1-q}{8\pi}} \\
&\qquad + \theta(\rho\bar{y} - y)\theta(\bar{y}) \frac{e^{-\frac{\bar{y}^2}{2(1-q)}}}{\bar{y}} \sqrt{\frac{1-q}{8\pi}} \\
&\qquad + \theta(y - \rho\bar{y})\theta(\bar{y} - \rho y) \frac{1}{2\pi} \frac{(1-q)(1-\rho^2)^{3/2}}{(\bar{y} - \rho y)(y - \rho\bar{y})} e^{-\frac{1}{2(1-q)}\boldsymbol{y}^T \mathcal{R}^{-1} \boldsymbol{y}} .
\end{aligned} \tag{6.36}$$

When we take the logarithm of this quantity, ignoring all the terms regular in $(1-q)$, we find

$$-\frac{1}{2(1-q)} \left[\theta(y)\theta(\rho y - \bar{y}) y^2 + \theta(\rho\bar{y} - y)\theta(\bar{y})\bar{y}^2 + \theta(y - \rho\bar{y})\theta(\bar{y} - \rho y)\boldsymbol{y}^T \mathcal{R}^{-1} \boldsymbol{y}\right] \tag{6.37}$$

Integrating in y, $\bar{y}$:

$$\begin{aligned}
-\frac{1}{2\pi(1-q)\sqrt{1-\rho^2}} \Bigg\{ &\int_0^{+\infty} \mathrm{d}y\, y^2 \int_{-\infty}^{\rho y} \mathrm{d}\bar{y}\, e^{-\frac{1}{2(1-\rho^2)}(y,\bar{y}) \left(\begin{smallmatrix} 1 & -\rho \\ -\rho & 1 \end{smallmatrix}\right) \left(\begin{smallmatrix} y \\ \bar{y} \end{smallmatrix}\right)} \\
&+ \frac{1}{2} \int_{\mathcal{D}} \mathrm{d}^2\boldsymbol{y}\, \boldsymbol{y}^T \mathcal{R}^{-1} \boldsymbol{y}\, e^{-\frac{1}{2}\boldsymbol{y}^T \mathcal{R}^{-1} \boldsymbol{y}} \Bigg\}
\end{aligned} \tag{6.38}$$

where $\mathcal{D}$ is the domain defined by the theta functions in the last term of Eq. (6.37) (in blue in Fig. 6.2). The first integral gives

$$\int_0^{+\infty} \mathrm{d}y\, y^2 \int_{-\infty}^{\rho y} \mathrm{d}\bar{y}\, e^{-\frac{1}{2(1-\rho^2)}(y,\bar{y}) \left(\begin{smallmatrix} 1 & -\rho \\ -\rho & 1 \end{smallmatrix}\right) \left(\begin{smallmatrix} y \\ \bar{y} \end{smallmatrix}\right)} = \frac{\pi}{2}\sqrt{1-\rho^2} . \tag{6.39}$$

To perform the second integral, we can change variables

$$\boldsymbol{z} = \mathcal{R}^{-1}\boldsymbol{y} , \tag{6.40}$$

so that

$$\begin{aligned}\frac{1}{2}\int_{\mathcal{D}} \mathrm{d}^2\boldsymbol{y}\,\boldsymbol{y}^T\mathcal{R}^{-1}\boldsymbol{y}\,e^{-\frac{1}{2}\boldsymbol{y}^T\mathcal{R}^{-1}\boldsymbol{y}} &= \frac{(1-\rho^2)}{2}\int_0^{+\infty}\mathrm{d}^2\boldsymbol{z}\,\boldsymbol{z}^T\mathcal{R}\boldsymbol{z}\,e^{-\frac{1}{2}\boldsymbol{z}^T\mathcal{R}\boldsymbol{z}} \\ &= -(1-\rho^2)\,\frac{\mathrm{d}}{\mathrm{d}\beta}\int_0^{+\infty}\mathrm{d}^2\boldsymbol{z}\,e^{-\frac{\beta}{2}\boldsymbol{z}^T\mathcal{R}\boldsymbol{z}}\Big|_{\beta=1} \\ &= \sqrt{1-\rho^2}\left[\frac{\pi}{2}-\arcsin\rho\right]\end{aligned} \tag{6.41}$$

where, in the last step, we used the known formula for the quadrant probability of a bivariate Gaussian distribution (see, for example, [Gup63]):

$$\int_0^{+\infty}\mathrm{d}^2\boldsymbol{z}\,e^{-\frac{\beta}{2}\boldsymbol{z}^T\mathcal{R}\boldsymbol{z}} = \frac{2\pi\sqrt{1-\rho^2}}{\beta(1-\rho^2)}\left[\frac{1}{4}-\frac{1}{2\pi}\arcsin\rho\right]. \tag{6.42}$$

Finally, the integral (6.38) becomes

$$-\frac{1}{4(1-q)}\left[2-\frac{2}{\pi}\arcsin(\rho)\right] = -\frac{1}{4(1-q)}\left[1+\frac{4}{\pi}\arctan\left(\sqrt{\frac{1-\rho}{1+\rho}}\right)\right] \tag{6.43}$$

The measure part has the same pole in $(1-q)$ as Gardner's: when we sum the two contributions and impose that the resulting quantity must be null, we find

$$1-\frac{p}{n}\left[\frac{1}{2}+\frac{2}{\pi}\arctan\left(\sqrt{\frac{1-\rho}{1+\rho}}\right)\right] = 0\,, \tag{6.44}$$

so that the critical value of the load is

$$\alpha_c(\rho) = \left[\frac{1}{2}+\frac{2}{\pi}\arctan\left(\sqrt{\frac{1-\rho}{1+\rho}}\right)\right]^{-1} = \left[\frac{3}{2}-\frac{2}{\pi}\arctan\left(\sqrt{\frac{1+\rho}{1-\rho}}\right)\right]^{-1}. \tag{6.45}$$

We point out here an interesting limit of Eq. (6.45), to convince the reader of the correctness of the result: when $\rho\to-1$, the points in a doublet are on opposite vertices of the unitary hypercube. The only way to classify all the doublets coherently with a plane, is to find a vector W which is perpendicular to all of them. This means that the problem reduces to a linear classification problem in dimension $n-p$ (because of the p constraints). As we know from Eq. (5.40) that the critical value of the capacity is 2, we find the equation

$$\frac{p}{n-p} = 2\,. \tag{6.46}$$

However, in term of the capacity of the original problem we know that $p_c = \alpha_c n$, so we find

$$\alpha_c = \frac{2}{3} = \lim_{\rho\to-1}\left[\frac{3}{2}-\frac{2}{\pi}\arctan\left(\sqrt{\frac{1+\rho}{1-\rho}}\right)\right]^{-1}. \tag{6.47}$$

For a similar reasoning, see [CLS18].

Please note finally that Eq. (6.45) is in agreement with Eq. (6.12) for $k=2$: since, to perform the replica calculations, we used a RS ansatz (6.29), while the combinatorial approach returns directly a quenched result, we can conclude that, as in the case of the classical Gardner's calculation for isolated points we explained in Sec. 5.4, replica symmetry breaking does not occur in the phase space whose partition function is represented by the volume (6.13), that is the space of the solutions of the linear classification problem we deal with in this section (see, for more details, [Bor+19]).

CHAPTER 7

Beyond the storage capacity: a data driven satisfiability transition

In this chapter, we set out to investigate the behavior of the VC entropy of linear classifiers for the data structures defined in Chap. 6, simplexes of k points in $\mathbb{R}^n$, in order to quantify how loose the logarithmic upper bounds from SLT are. We do so by means of two complementary approaches:

(i) we devise a combinatorial framework extending the original Cover's computation of Sec. 5.3, evaluating the asymptotic behavior of the number of dichotomies $C_{n,p}$ for generic k and overlaps between the points. This method allows us to access directly the VC entropy for the learning problem defined in section 6.1. Moreover, we find a novel transition beyond the Gardner's storage capacity: indeed, not only the fraction of dichotomies $c_{n,p} = C_{n,p}/2^p$ is bounded and complies with Eq. (5.41), but, in the case of structured data, $C_{n,p}$ itself is non-monotonic in the load, decreasing to zero in the thermodynamic limit after a point $\alpha_* > \alpha_c$;

(ii) we identify the appropriate synaptic volume to evaluate the point α_* within replica theory.

The transition identified by the load's novel critical value α_* is the point where, in the thermodynamic limit, no more solution of the linear classification problem of random multiplets, *regardless to the specific values of the multiplets' labels*, can be found: after the transition, even adjusting at will the labels, there is no hyperplane subdividing the multiplets coherently, i.e. without breaking any of them. Indeed, we clarify that this transition, which is not present in the case of linear separation (without margin) of isolated point, is due to the competition between an entropic contribution (as the number of multiplets increases, so does the number of ways to classify them) and an energetic contribution (which suppresses the dichotomies not respecting the structure of the multiplets: from a geometrical point of view, this term accounts for the excluded volume due to data structure). In the replica approach, we find that the synaptic volume associated to this transition, which is *different* from the generalized Gardner's one (6.13), exhibits some level of replica symmetry breaking that we are able to analyze heuristically in Sec. 7.2, exploiting the comparison with the combinatoric results. The exposition is mostly taken from [Pas+20].

7.1 Combinatorial approach

7.1.1 Asymptotic analysis via analytic combinatorics

In the case of unstructured data, we know that the growth of $C_{n,p}$ as a function of p is exponential up to the capacity $p_c = 2n$ and sub-exponential afterwards. Due to this

change of behavior, the fraction of linearly realizable dichotomies, $c_{n,p} = C_{n,p}/2^p$, has a discontinuous transition from 1 to 0 in the thermodynamic limit (see Fig. 7.1a). What is the asymptotic growth rate of $C_{n,p}$? This question can be answered by inspecting the explicit solution Eq. (5.18). However, we construct a different method here, based on the techniques of analytic combinatorics. Our method has the crucial advantage of being applicable to cases where (i) the solution $C_{n,p}$ is not known explicitly, and (ii) the recurrence equation is given implicitly, as a relation between its coefficients.

Let $g_n(z)$ be the ordinary generating function of $C_{n,p}$ with respect to the variable p:

$$g_n(z) = \sum_{p=1}^{\infty} C_{n,p} z^p. \tag{7.1}$$

Formally, the coefficient $C_{n,p}$ can be obtained by derivation as

$$C_{n,p} = \frac{1}{p!} \frac{\mathrm{d}^p}{\mathrm{d}z^p} g_n(z) \bigg|_{z=0}. \tag{7.2}$$

When it is unfeasible to compute the pth derivative explicitly, one can extract information on the asymptotic behavior of $C_{n,p}$ for large p by means of analytic techniques (see for instance [FS09]).

Whenever the generating function Eq. (7.1) is a rational function analytic in $z = 0$, it admits a partial fraction expansion

$$g_n(z) = Q_n(z) + \sum_s \sum_{r=1}^{r_s} \frac{a_{s,r}}{(z - z_s)^r}, \tag{7.3}$$

where Q_n is a polynomial, s ranges over the poles of g_n, and r_s is the multiplicity of the pole s. Then, the asymptotic form of the coefficients of $g_n(z)$ can be read off the series expansion of $(z - z_s)^{-r}$:

$$(z - z_s)^{-r} = \frac{(-1)^r}{z_s^r} \sum_{p=0}^{\infty} \binom{p + r - 1}{r - 1} z_s^{-p} z^p. \tag{7.4}$$

By substituting (7.4) in Eq. (7.3) one obtains r_s different contributions for each pole s. The overall leading term corresponds to the dominant singularity z_0 of $g_n(z)$, i.e., the one with smallest modulus $|z_0|$. This is due to the term z_s^{-p} in (7.4) that suppresses the sub-dominant poles exponentially. Among the contributions due to z_0, the leading one is that with $r = r_s$, because the binomial coefficient in (7.4) is a polynomial of degree $r - 1$ in p. Putting it all together, if the dominant singularity is a pole of order r, then

$$C_{n,p} \sim R z_0^{-p-r} \binom{p + r - 1}{r - 1}, \tag{7.5}$$

where the constant R can be obtained by factoring out the singularity:

$$R = \lim_{z \to z_0} (z_0 - z)^r g_n(z). \tag{7.6}$$

Equation (7.5) shows that if $|z_0| < 1$ (respectively, > 1), $C_{n,p}$ increases (respectively, decreases) exponentially with p at fixed n; if $|z_0| = 1$ then the asymptotic behavior is polynomial (of order $r - 1$).

In simple cases, when it is possible to obtain $g_n(z)$ in closed form, this method can be applied straightforwardly. However, this set up allows to probe the asymptotics of $C_{n,p}$ even in more complicated scenarios, where $g_n(z)$ cannot be solved for explicitly, or when even the recurrence relation for $g_n(z)$ is not specified completely. Section 7.1.5 shows how to tackle this more general problem. Before that, we consider the simpler cases $k = 1$ and $k = 2$.

7.1.2 Asymptotics for unstructured data

As a "warm-up exercise", we use the combinatorial method described above to explore the asymptotics of $C_{n,p}$ in the well-understood unstructured case of section 5.3.

By multiplying both sides of Eq. (5.19) by z^p and summing over p one obtains

$$\frac{1}{z} g_n(z) - 2 = g_n(z) + g_{n-1}(z), \tag{7.7}$$

where the constant term 2 comes from the initial condition (5.20). It is useful to rewrite the equation as

$$g_n(z) = \frac{z}{1-z} \left[g_{n-1}(z) + 2 \right]. \tag{7.8}$$

The boundary condition is $g_0(z) = 0$, due to every $C_{0,p}$ being zero. The relation (7.8) is a linear (non homogeneous) first-order recurrence with constant coefficients, whose solution is

$$g_n(z) = \frac{2z}{2z-1} \left[\left(\frac{z}{1-z} \right)^n - 1 \right]. \tag{7.9}$$

Equation (7.9) shows that $g_n(z)$ has a single pole at $z_0 = 1$, of order n, with finite part $R = 2$. Therefore, the corresponding asymptotic form has no exponential factor, and is purely polynomial:

$$C_{n,p} \sim 2 \binom{p+n-1}{n-1} = \frac{2}{(n-1)!} p^{n-1} + O\left(p^{n-2}\right). \tag{7.10}$$

Note that the right-hand side of Eq. (7.9) has a removable discontinuity in $z_1 = 1/2$, where the apparent pole in the first term gets canceled by a zero in the numerator (the term in square brackets). The corresponding exponential asymptotic growth, 2^p, is present in $C_{n,p}$ only transiently, for $p < n$.

7.1.3 Non-monotonicity of the VC entropy for structured data

The behavior of $C_{n,p}$, and therefore of the VC entropy, changes dramatically when data structure is present, already in the simplest case where the training data are structured as pairs of points, i.e., $k = 2$. Figure 7.1 shows the fraction of dichotomies, $C_{n,p}/2^p$, and the number of dichotomies, $C_{n,p}$, as functions of α for increasing values of the dimension n, for $k = 1$ and $k = 2$ with $\rho = 0.3$. The fraction of dichotomies is qualitatively similar in the two scenarios, the only apparent difference being the expected decrease in the storage capacity. A remarkable divergence appears instead in the asymptotic behavior of $C_{n,p}$. The absolute number of dichotomies is non-monotonic for simplex learning already in the simplest non-degenerate case $k = 2$ with $\rho < 1$. What is also evident in Fig. 7.1b is the fact that the storage capacity $\alpha_c(k)$ does not pinpoint any qualitatively special point for the unnormalized $C_{n,p}$, and therefore for the VC entropy.

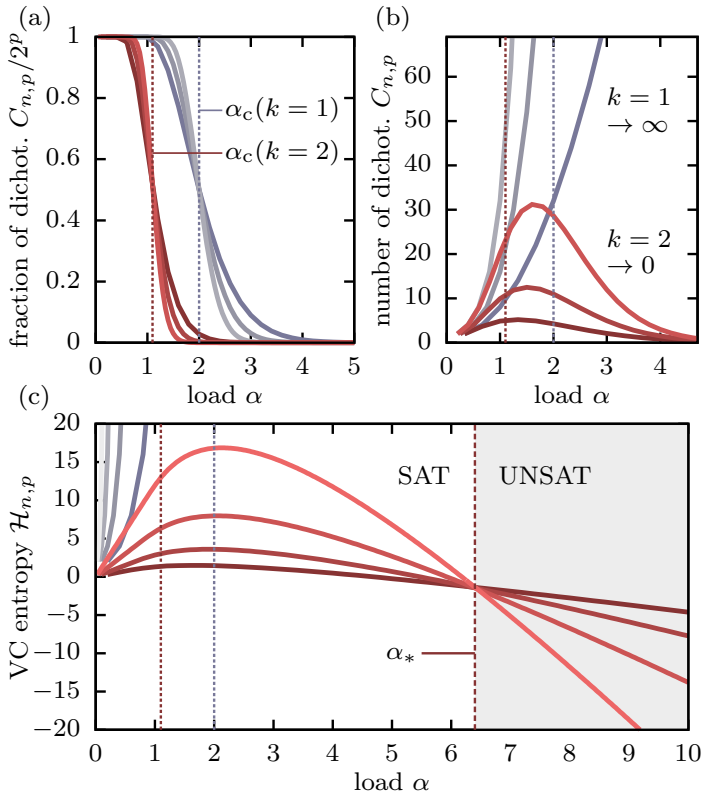

Figure 7.1: While the fraction of admissible dichotomies (a) has qualitatively similar behavior for unstructured (gray curves, $k = 1$) and structured (red curves, $k = 2$) data, the absolute number of dichotomies (b) has different limit behaviors. As a consequence, the VC entropy (c) diverges to $+\infty$ for unstructured data and to $-\infty$ for structured data. Curves of the VC entropy at different values of n intersect, for large n, at the same critical value α_* of the load. Vertical dotted lines in all panels are the storage capacities. The dashed line in (c) is the transition caused by data structure. The curves are obtained from the explicit solution (6.9), with $\mathcal{H}_{n,p} = \log C_{n,p}$ [$n = 5, 10, 20$ in (a), $n = 3, 4, 5$ in (b), $n = 5, 10, 20, 40$ in (c).] Figure from [Pas+20].

Since the two-point case $k = 2$ is the simplest case where the non-monotonicity of the VC entropy arises, we work it out in detail, before showing the general k-point case below. The geometry of the problem is fixed by the single quantity ψ_2. The recurrence equation reads

$$C_{n,p+1} = \psi_2 C_{n,p} + C_{n-1,p} + (1 - \psi_2)\, C_{n-2,p}, \tag{7.11}$$

with boundary conditions $C_{0,p} = 0, C_{n,1} = 2\{1 - [1 - \psi_2(d)]\delta_{n,1}\}$. In order to simplify the computations, we will use the same boundary conditions as for $k = 1$, i.e., $C_{0,p} = 0$ and $C_{n\geq 1,1} = 2$. This approximation has negligible effects in the large-n limit [RLG20].

Equation (7.11) fixes the recurrence relation satisfied by the generating function $g_n(z)$:

$$g_n(z) = \frac{z}{1 - \psi_2 z}\left[g_{n-1}(z) + (1 - \psi_2) g_{n-2}(z) + 2\right], \tag{7.12}$$

with boundary condition $g_{n\leq 0}(z) = 0$. The solution, which can be found by means of

the characteristic polynomial method, reads

$$g_n(z) = \left[\frac{z-\sqrt{\Delta(z)}}{2(1-\psi_2 z)}\right]^n \frac{z}{2z-1}\left(1+z\frac{2\psi_2-3}{\sqrt{\Delta(z)}}\right) + \left[\frac{z+\sqrt{\Delta(z)}}{2(1-\psi_2 z)}\right]^n \frac{z}{2z-1}\left(1-z\frac{2\psi_2-3}{\sqrt{\Delta(z)}}\right) - \frac{2z}{2z-1},$$

where $\Delta(z) = z[4(1-\psi_2)+z(1-2\psi_2)^2]$. The explicit solution has a pole of order n in $z_0 = 1/\psi_2$, with finite part

$$R = 2\psi_2^{-2n}. \tag{7.13}$$

Similarly to the unstructured case, the singularity in $z = 1/2$ is removable, which signals that the initial exponential increase of the number of dichotomies must be superseded eventually by the asymptotic behavior due to z_0. Altogether, the large-p form of $C_{n,p}$ is

$$C_{n,p} \sim 2\binom{p+n-1}{n-1}\psi_2^{p-n}. \tag{7.14}$$

The crucial difference between the results for $k = 1$, Eq. (7.10), and $k = 2$, Eq. (7.14), lies in the fact that while the first is asymptotically increasing, the second is exponentially decreasing whenever $\psi_2 < 1$, i.e., when the two partner points are distinct. Observe that $C_{n,p}$ always increases for small p; this is a consequence of the fact that the VC dimension of a linear classifier in n dimensions is $d_{\rm VC} = n$, therefore all dichotomies of kp points can be realized when $p \leq n/k$, meaning that $C_{n,p\leq n/k} = 2^p$. The decreasing asymptotic form then proves that the VC entropy $\mathcal{H}_{n,p}$ is non-monotonic in p (and therefore in α) for fixed n. Intuitively, the non-monotonicity is due to the competition of two opposing effects. On one hand, the addition of a new pair of points $\{\xi, \bar{\xi}\}$ to a set of p existing pairs entails a combinatorial increase in the total number of linearly-realizable dichotomies. On the other hand, some of the $C_{n,p}$ admissible dichotomies can become invalid if they are realizable only by hyperplanes intersecting the segment connecting ξ and $\bar{\xi}$.

7.1.4 Emergence of a data-driven satisfiability transition

A non-trivial consequence of the non-monotonic VC entropy can be observed in Fig. 7.1c. Consider the VC entropy $\mathcal{H}_{n,\alpha n}$ as a function of α. The curves $\mathcal{H}_{n,\alpha n}$ at different values of n intersect each other roughly around the same point α_*. More precisely, if $\mathcal{H}_{n,\alpha n}$ and $\mathcal{H}_{n-1,\alpha(n-1)}$ intersect at $\alpha_*(n)$, then $\alpha_* = \lim_{n\to\infty}\alpha_*(n)$. This empirical observation can be clarified analytically.

As a function of the load $\alpha = p/n$, Eq. (7.14) becomes

$$C_{n,\alpha n} \sim C(\alpha; n) \equiv 2\frac{\Gamma(\alpha n + n)}{\Gamma(n)\Gamma(\alpha n + 1)}\psi_2^{(\alpha-1)n} \tag{7.15}$$

(Γ is the Euler gamma function), or $\mathcal{H}_{n,\alpha n} \sim \mathcal{H}(\alpha; n)$ with

$$\mathcal{H}(\alpha; n) \equiv \log\left[2\frac{\Gamma(\alpha n + n)}{\Gamma(n)\Gamma(\alpha n + 1)}\right] + (\alpha - 1)n\log\psi_2. \tag{7.16}$$

In the non-degenerate case (whenever $\psi_2 < 1$) the second term in (7.16) is negative for $\alpha > 1$, while the first term is always positive. This competition gives rise to a transition

at $\alpha = \alpha_* > 1$, where the asymptotic limit of the VC entropy changes:

$$\lim_{n\to\infty} \mathcal{H}(\alpha; n) = \begin{cases} -\infty & \alpha < \alpha_* \\ \infty & \alpha > \alpha_*. \end{cases} \tag{7.17}$$

The transition point is pinpointed by the condition

$$\lim_{n\to\infty} \frac{\mathrm{d}}{\mathrm{d}n} \mathcal{H}(\alpha_*; n) = 0. \tag{7.18}$$

With $\mathcal{H}(\alpha; n)$ given by Eq. (7.16), the condition reads

$$\lim_{n\to\infty} \left[(\alpha_* - 1)\log\psi_2 + (\alpha_* + 1)\Psi(\alpha_* n + n) - \Psi(n) - \alpha_* \Psi(\alpha_* n + 1)\right] = 0, \tag{7.19}$$

where $\Psi(z) \equiv \partial_z \log\Gamma(z)$ is the poly-gamma function, whose asymptotic behavior is $\Psi(z) = \log(z) + O(1/z)$. Sending n to infinity then gives the transcendental equation

$$(\alpha_* + 1)\log(\alpha_* + 1) - \alpha_* \log\alpha_* + (\alpha_* - 1)\log\psi_2 = 0, \tag{7.20}$$

which has two solutions: α_* is the larger. As a function of ψ_2, the transition point α_* has limits

$$\lim_{\psi_2\to 0} \alpha_* = 1\,, \qquad \lim_{\psi_2\to 1} \alpha_* = \infty. \tag{7.21}$$

As expected, when ψ_2 goes to 1, the problem reduces to that of classifying unstructured data, and the transition runs to infinity.

The phase transition at α_* can be rationalized as the SAT-UNSAT transition of a random constraint satisfaction problem (CSP). First, we recall that the storage capacity α_c itself corresponds to the transition between the satisfiable and the unsatisfiable phase of an appropriate satisfiability problem, as we briefly mentioned in Chap. 5. The CSP relevant to α_c can be stated as follows:

Constraint satisfaction problem 1. *Given a set of kn input-label pairs $\{\xi_a^\mu, \sigma^\mu\}$ (with $a = 1, \dots, k$ and $\mu = 1, \dots, p$), find a vector w such that $\mathrm{sign}(w \cdot \xi_a^\mu) = \sigma^\mu$ for all μ and a.*

The input data of this problem satisfies the admissibility constraints by construction. A corresponding random constraint satisfaction problem (rCSP) is an ensemble of CSPs, specified by a probability measure on the input data. The rCSP is in the SAT (respectively UNSAT) phase when the satisfiability problem admits a solution with probability one (respectively zero) in the thermodynamic limit. The storage capacity (6.12) marks the transition between the SAT and the UNSAT phases of the rCSP corresponding to problem 1 with the probability measure of simplex learning described in Chap. 6

A different problem can be constructed by moving the admissibility property from the definition of the input data to the conditions defining the solution:

Constraint satisfaction problem 2. *Given a set of kn input points $\{\xi_a^\mu\}$, (with $a = 1, \dots, k$ and $\mu = 1, \dots, p$), find a set of labels $\{\sigma^\mu\}$ and a vector w such that $\mathrm{sign}(w \cdot \xi_a^\mu) = \sigma^\mu$ for all μ and a.*

Notice that this problem is trivially satisfiable for unstructured data, i.e., it is satisfied by almost all vectors w when the constraint of admissibility is irrelevant (i.e., when $k = 1$). A solution to problem 2 is given by specifying an admissible dichotomy $\{\sigma^\mu\}$ and a vector w. In this framework, the VC entropy counts the (logarithm of the) number of distinct dichotomies $\{\sigma^\mu\}$ that can appear in such a solution. This means that the corresponding rCSP is in the UNSAT phase when $\mathcal{H}(\alpha; n) \to -\infty$ and in the SAT phase otherwise.

7.1.5 Transition point for generic k

Now we address the more general case where the number of partners in a multiplet is k. The generating function $g_n(z)$ satisfies the recurrence equation

$$g_n(z) = \frac{z}{1 - z\theta_0^k}\left[2 + \sum_{l=1}^{k} \theta_l^k g_{n-l}(z)\right], \tag{7.22}$$

as can be obtained from Eq. (6.5). Solving for $g_n(z)$ from Eqs. (7.22) and (6.6) would be hopeless. However, the asymptotic analysis discussed above only needs three pieces of information about $g_n(z)$, namely (i) the location z_0 of the dominant singularity, (ii) its order r, and (iii) its finite part R. These can be extracted from the recurrence relations without solving them.

The right-hand side of Eq. (7.22) has a singularity in $z = 1/\theta_0^k$. The boundary condition is $g_{n\leq 0}(z) = 0$, therefore the first non-zero function is $g_1(z) = 2\sigma(z)$, where

$$\sigma(z) = \frac{z}{1 - z\theta_0^k} \tag{7.23}$$

encapsulates the singularity. Since the number of terms in the sum in Eq. (7.22) is finite, no other singularity can appear at finite n. Therefore

$$z_0 = \frac{1}{\theta_0^k}. \tag{7.24}$$

Now consider one iteration of Eq. (7.22): the singularity with largest order in the right-hand side comes from $g_{n-1}(z)$, and the singular term gets multiplied by $\theta_1^k \sigma(z)$. Indeed, it is easy to see by induction that the leading term $\hat{g}_n(z)$ in the Laurent expansion of $g_n(z)$ around z_0 is

$$\hat{g}_n(z) = 2\left(\theta_1^k\right)^{n-1} \sigma(z)^n. \tag{7.25}$$

Therefore, the order of the singularity is $r = n$. The constant R [Eq. (7.6)] can be obtained by multiplying Eq. (7.25) by $(1/\theta_0^k - z)^n$ and evaluating it at $z = 1/\theta_0^k$:

$$R = 2\left(\theta_1^k\right)^{n-1} \left(\theta_0^k\right)^{-2n}. \tag{7.26}$$

Finally, the asymptotic behavior of $C_{n,p}$ is

$$C_{n,p} \sim 2\binom{p+n-1}{n+1} \left(\theta_1^k\right)^{n-1} \left(\theta_0^k\right)^{p-n}, \tag{7.27}$$

from which one readily obtains the asymptotic form $C(\alpha; n)$ for the number of dichotomies,

$$C(\alpha; n) = 2\frac{\Gamma(\alpha n + n)}{\Gamma(n)\Gamma(\alpha n + 1)} \left(\theta_1^k\right)^{n-1} \left(\theta_0^k\right)^{(\alpha-1)n}, \tag{7.28}$$

and the corresponding one for the VC entropy,

$$\mathcal{H}(\alpha; n) = \log\left[2\frac{\Gamma(\alpha n + n)}{\Gamma(n)\Gamma(\alpha n + 1)}\right] + (n-1)\log\theta_1^k + (\alpha - 1)n\log\theta_0^k. \tag{7.29}$$

As above, the existence of a critical value α_* can be established by finding the zeros of the derivative of $\mathcal{H}(\alpha; n)$ with respect to n, in the large-n limit. One finds

$$(\alpha_* + 1)\log(\alpha_* + 1) - \alpha\log\alpha_* + (\alpha_* - 1)\log\theta_0^k + \log\theta_1^k = 0. \tag{7.30}$$

The two coefficients θ_0^k and θ_1^k can be obtained from Eq. (6.6) as functions of the ψ's. By solving the recurrence equation, specialized to $l = 0$, one has

$$\theta_0^k = \prod_{m=2}^{k} \psi_m. \tag{7.31}$$

Then, by substituting expression (7.31) into Eq. (6.6) with $l = 1$, one obtains the recurrence relation

$$\theta_1^k = \psi_k \theta_1^{k-1} + (1 - \psi_k) \prod_{m=2}^{k-1} \psi_m, \tag{7.32}$$

with boundary condition $\theta_1^1 = 1$. The solution is

$$\theta_1^k = \left(2 - k + \sum_{m=2}^{k} \frac{1}{\psi_m}\right) \prod_{m=2}^{k} \psi_m. \tag{7.33}$$

Specializing to $k = 3$, for instance, yields

$$\begin{aligned} \theta_0^3 &= \psi_3 \psi_2 \\ \theta_1^3 &= \psi_3 + \psi_2 - \psi_3 \psi_2. \end{aligned} \tag{7.34}$$

Because of the way θ_0^k and θ_1^k are constructed via the geometric quantities $\psi_m \in [0, 1]$, they are not independent. The range of θ_0^k is $[0, 1]$, as can be seen from Eq. (7.31). The sup and inf of θ_1^k at fixed θ_0^k can be obtained by considering the two extremal cases

$$\begin{aligned} &\text{(i)} \ \{\psi_m\}_m = \left\{1, \ldots, 1, \theta_0^k, 1, \ldots, 1\right\}, \\ &\text{(ii)} \ \{\psi_m\}_m = \{\left(\theta_0^k\right)^{1/(k-1)}, \ldots, \left(\theta_0^k\right)^{1/(k-1)}\}. \end{aligned} \tag{7.35}$$

The fact that the evaluation on the two extremal cases gives the appropriate bounds is not obvious: it can be proved by induction using Lagrange's theorem for constrained optimization (taking care to consider the boundary of the domain as well); see Appendix 1 of [Pas+20]. From (i) and (ii) respectively one gets

$$\begin{aligned} &\text{(i)} \ \sup \theta_1^k = 1, \\ &\text{(ii)} \ \inf \theta_1^k = (k-1)\left(\theta_0^k\right)^{1 - \frac{1}{k-1}} + (2-k)\theta_0^k. \end{aligned} \tag{7.36}$$

The inf is monotonically decreasing with k; therefore, by letting $k \to \infty$ one obtains a global lower bound independent of k:

$$\theta_1^k > \theta_1^\infty = \theta_0^k \left[1 - \log \theta_0^k\right]. \tag{7.37}$$

The upper bound (i) is already k-independent.

Figure 7.2 summarizes the results concerning the value of α_* for generic k. It also shows a comparison with numerical results obtained for $k = 3$, (with $\{\rho_{ab}\}$ given by the equilateral geometry). The theoretical bounds in the figure (dashed lines) are obtained by substituting the k-independent bounds above into Eq. (7.30).

We point out that there are two sources of approximation in the computations above, namely (i) the modified boundary conditions, and (ii) the perturbative nature of the asymptotic analysis. Concerning (i), we remark that the numerical results were obtained by using the correct boundary conditions. However, using the modified conditions does not change the numerical results appreciably. The small discrepancies apparent in the Fig. 7.2 are therefore due almost entirely to (ii).

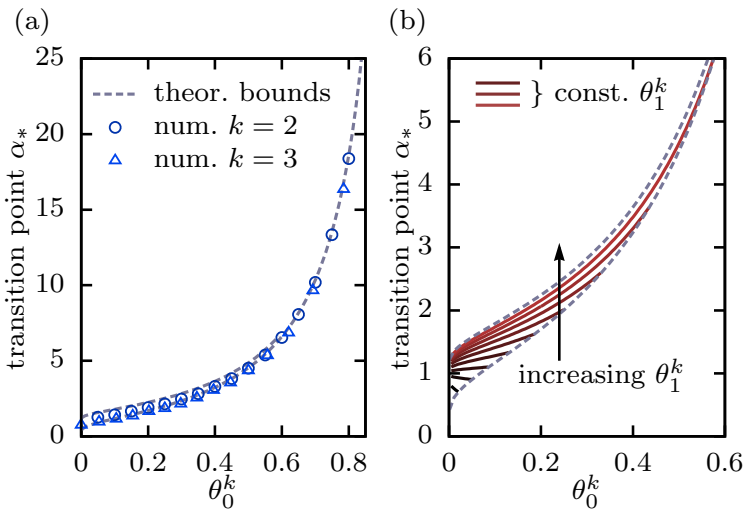

Figure 7.2: (a) Numerical estimates of α_* at varying θ_0^k for two different geometries: $k = 2$ (where θ_0^2 is just ψ_2) and $k = 3$. In the latter case we fix $\{\rho_{ab}\}$ by requiring that the three points in the simplex form an equilateral triangle of varying sizes. (b) Theoretical results (red curves) for α_* as a function of θ_0^k for increasing values of θ_1^k, within its allowed range given by Eqs. (7.36) and (7.37). Dashed lines in both panels are the k-independent upper and lower bounds for α_*. Figure from [Pas+20].

7.1.6 Finite-size scaling at the critical point

In the vicinity of the transition point α_*, the quantity $C(\alpha; n)$ satisfies finite-size scaling, as happens for other random satisfiability problems [KS94; LRZ01]. In this section we compute the scaling form and its critical exponents.

Let us define a scaling variable y as n times the reduced load $(\alpha - \alpha_*)/\alpha_*$ around α_*:

$$y = n\frac{\alpha - \alpha_*}{\alpha_*}. \tag{7.38}$$

By inserting $\alpha = \alpha_* y/n + \alpha_*$ in Eq. (7.28), and using the asymptotic expansion of the Γ function,

$$\Gamma(x) = e^{x\log x - x}\left[\sqrt{2\pi}x^{-1/2} + O\left(x^{-3/2}\right)\right], \tag{7.39}$$

one obtains in the large-n limit

$$C(\alpha; n) = e^{nA+B}\left[\frac{\sqrt{2/\pi}}{\sqrt{\alpha_*(1+\alpha_*)}}n^{-1/2} + O\left(n^{-3/2}\right)\right],$$

with

$$A = (\alpha_* + 1)\log(\alpha_* + 1) - \alpha_*\log\alpha_* + (\alpha_* - 1)\log\theta_0^k + \log\theta_1^k,$$
$$B = -\log\theta_1^k + \alpha_* y\log(\alpha_* + 1) - \alpha_* y\log\alpha_* + \alpha_* y\log\theta_0^k.$$

The linear term nA in the exponential vanishes by Eq. (7.30). Hence,

$$C(\alpha; n) = n^{-1/2}\frac{1}{\theta_1^k}\frac{\sqrt{2/\pi}}{\sqrt{\alpha_*(1+\alpha_*)}}\left(\frac{\alpha_* + 1}{\alpha_*}\theta_0^k\right)^{\alpha_* y}\left[1 + O\left(n^{-3/2}\right)\right], \tag{7.40}$$

which shows that in the thermodynamic limit $C(\alpha; n)$ obeys the scaling form

$$C(\alpha; n) = n^{-1/2} F\left(\frac{\alpha - \alpha_*}{\alpha_*} n\right) \tag{7.41}$$

with the exponential scaling function

$$F(y) = \frac{1}{\theta_1^k} \frac{\sqrt{2/\pi}}{\sqrt{\alpha_*(1+\alpha_*)}} \left(\frac{\alpha_* + 1}{\alpha_*} \theta_0^k\right)^{\alpha_* y}. \tag{7.42}$$

Equation (7.41) shows that, within the approximation of our asymptotic analysis, the number of dichotomies satisfies the finite-size scaling form

$$C_{n,\alpha n} \sim n^{-\beta/\nu} F\left(\frac{\alpha - \alpha_*}{\alpha_*} n^{1/\nu}\right) \tag{7.43}$$

(where F is regular), with critical exponents

$$\beta = 1/2, \quad \nu = 1. \tag{7.44}$$

Let $h(\alpha)$ be the VC entropy density in the thermodynamic limit:

$$h(\alpha) = \lim_{n\to\infty} \frac{1}{n} \mathcal{H}(\alpha; n). \tag{7.45}$$

The condition $h(\alpha) = 0$, satisfied by α_*, can be written from Eq. (7.42) as

$$(\alpha - \alpha_*) \log\left(\frac{\alpha_* + 1}{\alpha_*} \theta_0^k\right) = 0. \tag{7.46}$$

Curiously, Eq. (7.46) is satisfied identically in α if $\alpha_* = \theta_0^k/(1-\theta_0^k)$. By plugging this value of α_* into Eq. (7.30), one obtains the simple condition $\theta_1^k = \theta_0^k(1-\theta_0^k)$. For data structure with θ_0^k and θ_1^k satisfying this relation, one therefore expects that $C(\alpha; n)$ is constant in α in the large-n limit; equivalently, the VC entropy will be approximately independent of the load, $\mathcal{H}_{n,p} \sim \mathcal{H}_n$.

7.2 Replica approach

The discussion in the foregoing sections shows that (i) the VC entropy has non-monotonic behavior for simplex learning, (ii) the hallmark of the non-monotonicity is the existence of a phase transition, and (iii) the transition can be framed as the SAT-UNSAT transition of a constraint satisfaction problem, which is different from the one that defines the storage capacity. Since it is often challenging to deal with the combinatorics of complex data structures, our goal in this section is to identify an appropriate synaptic volume that provides access to the transition. Once this observable is identified, we will be able to pinpoint the existence of the phase transition without direct access to the VC entropy, in the same spirit of the original work by Gardner [Gar87] we reported in sections 5.4 and 6.2, by using disordered systems techniques.

We define the synaptic volume by leveraging on the definition of the CSP corresponding to the transition. As already noted, in looking for a solution to the constraint satisfaction problem 2 (defined in Sec. 7.1.4), we have the freedom to adjust both the synaptic

weights W and the outputs σ. This means that the outputs are promoted to be dynamical variables and should be treated at the same level of the synaptic weights. This suggests that the relevant synaptic volume for identifying the corresponding phase transition is the following:

$$V(X_p) = \sum_{\{\sigma^\mu=\pm 1\}} \int \left[\prod_{j=1}^{n} \mathrm{d}W_j\right] \delta\left(\sum_{j=1}^{n} W_j^2 - n\right) \prod_{\mu=1}^{p} \prod_{a=1}^{k} \theta\left(\frac{\sigma^\mu}{\sqrt{n}} \sum_{j=1}^{n} W_j \xi_{a,j}^\mu\right), \tag{7.47}$$

where $\theta(\cdot)$ is the Heaviside theta, $\xi_{a,j}^\mu$ denotes the j-th component of the a-th element of the μ-th multiplet and the weights lie on the surface of a n-dimensional sphere of radius $\sqrt{n}$ The inputs, constituting the set X_p, are chosen randomly according to the distribution (6.14).

This synaptic volume differs from the ordinary Gardner volume (6.15) by the integration over the labels σ, considered dynamical variables on the same foot of the weights W. Intuitively, an exponential growth of $V(X)$ with n at fixed load α means that, in the thermodynamic limit, at least one classification compatible with the input-label constraints can be expressed by the model; on the contrary, when $V(X)$ decreases exponentially in n then no such classification exists for $n \to \infty$. Thus, the logarithm of $V(X)$ is a suitable observable to assess the non-monotonic behavior of the VC entropy for a given data structure.

We will apply replica theory to compute the averaged (over the inputs positions) logarithm of the synaptic volume defined in Eq. (7.47), in order to identify the transition. The goal will be the evaluation of the critical value of α where this volume changes regime, as a function of the overlaps. In the following, we will restrict to the case $k = 2$, i.e. to data organized in doublets, so that the geometry of the simplex is fully specified by a single parameter ρ; as we did in Sec. 6.2, to lighten the notation, we will omit the index $a = 1, 2$, simply denoting the doublets as $(\xi, \bar{\xi})$. Using the standard integral representations for the delta and theta functions, Eq. (5.24), we can write the volume of interest as

$$\begin{aligned} V = \sum_{\{\sigma^\mu=\pm 1\}} \int \left[\prod_{j=1}^{n} \mathrm{d}W_j\right] \int_0^{+\infty} \left[\prod_{\mu=1}^{p} \frac{\mathrm{d}\lambda^\mu \, \mathrm{d}\bar{\lambda}^\mu}{(2\pi)^2}\right] \int_{-\infty}^{+\infty} \left[\prod_{\mu=1}^{p} \mathrm{d}x^\mu \, \mathrm{d}\bar{x}^\mu\right] \int_{-\infty}^{+\infty} \frac{\mathrm{d}E}{2\pi} \\ \times \; e^{iE\left(\sum_j W_j^2 - n\right) + i\sum_\mu x^\mu \left(\lambda^\mu - \frac{\sigma^\mu}{\sqrt{n}} \sum_j W_j \xi_j^\mu\right) + i\sum_\mu \bar{x}^\mu \left(\bar{\lambda}^\mu - \frac{\sigma^\mu}{\sqrt{n}} \sum_j W_j \bar{\xi}_j^\mu\right)}, \end{aligned} \tag{7.48}$$

where the auxiliary variable E enforces the spherical constraint, while the standard integral representation of the theta function is obtained via the auxiliary variables λ, x.

We dedicate the following sections to the calculation of the averaged logarithm of this volume in the annealed, replica symmetric (RS) and one-step replica symmetry breaking (1RSB) approximations. We start from the easiest one, the annealed approximation, because we have no a priori knowledge on the kind of transition, as the volume V in Eq. (7.48) is different from the Gardner volume (6.16), due to the summation over the labels. However, we find a posteriori that we need at least the 1RSB evaluation to observe quantitative accordance with the combinatorial result, suggesting that the transition present some level of replica symmetry breaking. The main results of this section, to which we address the reader not interested in the details, are Eq. (7.53), (7.67) and (7.79).

7.2.1 Annealed computation

The annealed calculation is based on the substitution $\overline{\log V} \to \log \overline{V}$, so we simply need to average the volume (7.48) with respect to the input distribution (indicated by the overline); the strategy is the same as the one devised in section 6.2.2. After a large-n expansion and the average over the inputs, the integrals in x and λ can be solved explicitly:

$$\begin{aligned}\left[\sum_{\{\sigma=\pm 1\}} \int_0^{+\infty} \frac{\mathrm{d}^2\boldsymbol{\lambda}}{(2\pi)^2} \int_{-\infty}^{+\infty} \mathrm{d}^2\boldsymbol{x}\, e^{-\frac{1}{2}\boldsymbol{x}^T\mathcal{R}\boldsymbol{x}+i\boldsymbol{x}^T\boldsymbol{\lambda}}\right]^p &= \left[2\int_0^{+\infty} \frac{\mathrm{d}^2\boldsymbol{\lambda}}{(2\pi)^2} \frac{2\pi}{\sqrt{1-\rho^2}} e^{-\frac{1}{2}\boldsymbol{\lambda}^T\mathcal{R}^{-1}\boldsymbol{\lambda}}\right]^p \\ &= \left[\frac{1}{2}+\frac{1}{\pi}\arcsin\rho\right]^p ,\end{aligned} \tag{7.49}$$

where we introduced the notation (6.27) and we used the known formula for the quadrant probability of a bivariate normal distribution, see [Gup63]. The remaining integrals can be performed: the one over the weights is Gaussian

$$\int \left[\prod_{j=1}^{n} \mathrm{d}W_j\right] e^{iE\sum_j W_j^2} = e^{n\left[\frac{1}{2}\log\pi - \frac{1}{2}\log(-iE)\right]} , \tag{7.50}$$

while the one over E can be performed via a saddle-point method for large n:

$$\int_{-\infty}^{+\infty} \frac{\mathrm{d}E}{2\pi}\, e^{-inE-\frac{n}{2}\log(-iE)} \sim \frac{1}{2\sqrt{\pi n}} e^{n\left[\frac{1}{2}+\frac{\log 2}{2}\right]} . \tag{7.51}$$

Assembling everything, and ignoring inessential factors, we find

$$\overline{V} = \exp\left\{n\left[\frac{p}{n}\log\left(\frac{1}{2}+\frac{1}{\pi}\arcsin\rho\right)+\frac{1+\log 2\pi}{2}\right]\right\} . \tag{7.52}$$

Defining the critical value of $\alpha = p/n$ as the one where the exponent changes sign, we find

$$\alpha_*^A(\rho) = -\frac{1+\log 2\pi}{2\log\left(\frac{1}{2}+\frac{1}{\pi}\arcsin\rho\right)} . \tag{7.53}$$

A comparison of the annealed approximation and of the result obtained with combinatorics in Eq. (7.20) is shown in Fig. 7.3. Although the annealed approximation fails in reproducing quantitatively the behavior of $\alpha_*(\rho)$, it bounds the combinatorial result from below, and qualitatively recovers the expected divergence for $\psi_2 \to 1$.

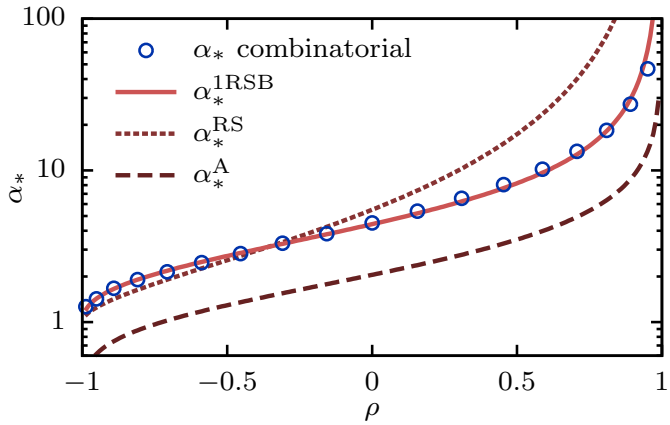

Figure 7.3: Critical value of the load α as a function of the overlap ρ for $k = 2$ (data in pairs). Circles represent the combinatorial result solution of Eq. (7.20), which is in agreement with numerical simulations [performed, as a crude check, in this way: fix a plane; launch randomly a set of p doublets and check if the plane realizes an admissible dichotomy; if not, go on launching sets of p doublets up to a certain large number; if no (at least one) launch has been separated correctly by the plane, we say that we are in the UNSAT (SAT) phase; change p to explore the region]. All the different approximation schemes used for the replica computations display the same qualitative shape. However the annealed and RS ansatz fail in reproducing quantitatively the combinatorial result. Using a 1RSB ansatz we obtain a one-parameter expression for α_* [Eq. (7.79)] that fits the combinatorial result tightly. Figure from [Pas+20].

7.2.2 Quenched computation

The quenched calculation of $\overline{\log V}$ is performed via the replica trick. First, we replicate t times the volume (7.48), obtaining

$$
\begin{aligned}
V^t = \sum_{\{\sigma_a^\mu = \pm 1\}} \int \left[\prod_{a=1}^{t} \prod_{j=1}^{n} \mathrm{d}W_{j,a} \right] \int_{-\infty}^{+\infty} \left[\prod_{a=1}^{t} \frac{\mathrm{d}E_a}{2\pi} \right] \int_{-\infty}^{+\infty} \left[\prod_{a<b} \frac{\mathrm{d}F_{ab}\, \mathrm{d}Q_{ab}}{2\pi} \right] \\
\times e^{i \sum_a E_a \left(\sum_j W_{j,a}^2 - n \right) + i \sum_{a<b} F_{ab} \left(\sum_j W_{j,a} W_{j,b} - n Q_{ab} \right)} \int_0^{+\infty} \left[\prod_{a=1}^{t} \prod_{\mu=1}^{p} \frac{\mathrm{d}\lambda_a^\mu\, \mathrm{d}\bar{\lambda}_a^\mu}{(2\pi)^2} \right] \\
\times \int_{-\infty}^{+\infty} \left[\prod_{a=1}^{t} \prod_{\mu=1}^{p} \mathrm{d}x_a^\mu\, \mathrm{d}\bar{x}_a^\mu \right] e^{i \sum_{a,\mu} x_a^\mu \left(\lambda_a^\mu - \frac{\sigma_a^\mu}{\sqrt{n}} \sum_j W_{j,a} \xi_j^\mu \right) + i \sum_{a,\mu} \bar{x}_a^\mu \left(\bar{\lambda}_a^\mu - \frac{\sigma_a^\mu}{\sqrt{n}} \sum_j W_{j,a} \bar{\xi}_j^\mu \right)} ,
\end{aligned}
\tag{7.54}
$$

where $1 \le a, b \le t$ are replica indices, Q_{ab} is the replica matrix (with $Q_{aa} = 1$) and F_{ab} are the Lagrange multipliers enforcing the constraint

$$
Q_{ab} = \frac{1}{n} \sum_{j=1}^{n} W_{j,a} W_{j,b} \, . \tag{7.55}
$$

Now we can perform the average over the input ensemble. With the same steps we used to get equation (7.49) (see section 6.2.2), we obtain, for the x and λ integrals,

$$\left\{ \sum_{\{\sigma_a=\pm 1\}} \int_0^{+\infty} \left[\prod_{a=1}^{t} \frac{\mathrm{d}^2 \boldsymbol{\lambda}_a}{(2\pi)^2} \right] \int_{-\infty}^{+\infty} \left[\prod_{a=1}^{t} \mathrm{d}^2 \boldsymbol{x}_a \right] e^{-\frac{1}{2}\sum_{a,b} Q_{ab} \boldsymbol{x}_a^T \mathcal{R} \boldsymbol{x}_b + i \sum_a \sigma_a \boldsymbol{x}_a^T \boldsymbol{\lambda}_a} \right\}^p , \quad (7.56)$$

where we already inserted the replica matrix using (7.55) and we isolated the outputs σ in the source term via the transformation $\boldsymbol{x} \to \sigma \boldsymbol{x}$. As in Gardner's calculation, the remaining integral over the weights is Gaussian:

$$\int \left[\prod_{a,j} \mathrm{d}W_{j,a} \right] e^{i \sum_a E_a \sum_j W_{j,a}^2 + i \sum_{a<b} F_{ab} \sum_j W_{j,a} W_{j,b}} = e^{-\frac{n}{2} \log\det(-i\boldsymbol{G}) + \frac{nt}{2} \log(2\pi)} , \quad (7.57)$$

where $\boldsymbol{G}$ is the symmetric matrix defined in Eq. (5.31); the integral over the elements of G is performed via a saddle-point. Finally, the resulting averaged replicated volume to be evaluated is

$$\begin{aligned} \overline{V^t} = & \int_{-\infty}^{+\infty} \left[\prod_{a<b} \mathrm{d}Q_{ab} \right] e^{\frac{nt}{2} + \frac{n}{2} \log\det(\boldsymbol{Q})} \\ & \times \left\{ \sum_{\{\sigma_a=\pm 1\}} \int_0^{+\infty} \left[\prod_{a=1}^{t} \frac{\mathrm{d}^2 \boldsymbol{\lambda}_a}{(2\pi)^2} \right] \int_{-\infty}^{+\infty} \left[\prod_{a=1}^{t} \mathrm{d}^2 \boldsymbol{x}_a \right] e^{-\frac{1}{2}\sum_{a,b} Q_{ab} \boldsymbol{x}_a^T \mathcal{R} \boldsymbol{x}_b + i \sum_a \sigma_a \boldsymbol{x}_a^T \boldsymbol{\lambda}_a} \right\}^p . \end{aligned} \quad (7.58)$$

We cannot proceed further, in taking the limit $t \to 0$ as prescribed by the replica approach, without making an ansatz on the form of the replica matrix Q_{ab}.

RS ansatz

In the RS ansatz, the replica matrix has the form

$$Q_{ab} = (1-q)\delta_{ab} + q , \quad 0 \le q \le 1 , \quad (7.59)$$

so that

$$\log\det(\boldsymbol{Q}) \underset{t\to 0}{\to} t \log(1-q) + \frac{tq}{1-q} . \quad (7.60)$$

The quadratic form at the exponent of Eq. (7.58) reads

$$\sum_{a,b} Q_{ab} \boldsymbol{x}_a^T \mathcal{R} \boldsymbol{x}_b = (1-q) \sum_a \boldsymbol{x}_a^T \mathcal{R} \boldsymbol{x}_a + q \left(\sum_a \boldsymbol{x}_a \right)^T \mathcal{R} \left(\sum_b \boldsymbol{x}_b \right) . \quad (7.61)$$

The last term can be linearized with a Hubbard-Stratonovich transformation:

$$e^{-\frac{q}{2}\left[\sum_a \boldsymbol{x}_a\right]^T \mathcal{R} \left[\sum_b \boldsymbol{x}_b\right]} = \int_{-\infty}^{+\infty} \frac{\mathrm{d}^2 \boldsymbol{y}}{2\pi\sqrt{1-\rho^2}} e^{-\frac{1}{2}\boldsymbol{y}^T \mathcal{R}^{-1} \boldsymbol{y} + i\sqrt{q} \sum_a \boldsymbol{x}_a^T \boldsymbol{y}} . \quad (7.62)$$

so that replica indices factorize, to get, after an integration over $\boldsymbol{x}$,

$$\begin{aligned} & \left\{ \int_{-\infty}^{+\infty} \frac{\mathrm{d}^2 \boldsymbol{y}}{2\pi\sqrt{1-\rho^2}} e^{-\frac{1}{2}\boldsymbol{y}^T \mathcal{R}^{-1} \boldsymbol{y}} \right. \\ & \left. \times \left[\frac{2\pi}{(1-q)\sqrt{1-\rho^2}} \sum_{\{\sigma=\pm 1\}} \int_0^{+\infty} \frac{\mathrm{d}^2 \boldsymbol{\lambda}}{(2\pi)^2} e^{-\frac{1}{2(1-q)} (\boldsymbol{\lambda} + \sigma\sqrt{q}\boldsymbol{y})^T \mathcal{R}^{-1} (\boldsymbol{\lambda} + \sigma\sqrt{q}\boldsymbol{y})} \right]^t \right\}^p . \end{aligned} \quad (7.63)$$

Defining $L_{\rm RS}(\boldsymbol{y})$ the quantity in square brackets, the limit $t \to 0$ gives

$$p \log\left\{ \int_{-\infty}^{+\infty} \frac{\mathrm{d}^2\boldsymbol{y}}{2\pi\sqrt{1-\rho^2}}\, e^{-\frac{1}{2}\boldsymbol{y}^T\mathcal{R}^{-1}\boldsymbol{y}} \left[L_{\rm RS}(\boldsymbol{y})\right]^t \right\} \\ \to pt \int_{-\infty}^{+\infty} \frac{\mathrm{d}^2\boldsymbol{y}}{2\pi\sqrt{1-\rho^2}}\, e^{-\frac{1}{2}\boldsymbol{y}^T\mathcal{R}^{-1}\boldsymbol{y}} \log\left[L_{\rm RS}(\boldsymbol{y})\right] . \quad (7.64)$$

Since we are looking for the critical value of α of the SAT-UNSAT transition of our CSP, we can just apply the standard argument by Gardner [Gar87]: starting with a load below the critical value and increasing the number of patterns, the set of solutions in the space of weights shrinks down to a single configuration at the transition (in the thermodynamic limit). This means that, approaching the critical point, the replicas of the vector W must be more and more correlated and therefore $q \to 1$ at the transition. In this limit, the factor $(1-q)^{-1}$ is large and the integrals in $L_{\rm RS}(\boldsymbol{y})$ can be evaluated with a saddle point: we need to find the stationary points of the exponent in the integrands as a function of $\boldsymbol{\lambda}$. According to the position of the vector $\boldsymbol{y}$ on the plane, the saddle is in one of the three following spots: (i) inside the region of integration over $\boldsymbol{\lambda}$; (ii) at one of its boundaries; (iii) at the origin (see section 6.2 for details). We obtain:

$$\int_0^{+\infty} \frac{\mathrm{d}^2\boldsymbol{\lambda}}{2\pi(1-q)\sqrt{1-\rho^2}}\, e^{-\frac{1}{2(1-q)}(\boldsymbol{\lambda}+\sigma\sqrt{q}\boldsymbol{y})^T\mathcal{R}^{-1}(\boldsymbol{\lambda}+\sigma\sqrt{q}\boldsymbol{y})} \\ \sim \theta(-\sigma y)\theta(-\sigma\bar{y}) + \theta(\sigma y)\theta[\sigma(\rho y-\bar{y})]\frac{e^{-\frac{y^2}{2(1-q)}}}{y}\sqrt{\frac{1-q}{8\pi}} \\ + \theta[\sigma(\rho\bar{y}-y)]\theta(\sigma\bar{y})\frac{e^{-\frac{\bar{y}^2}{2(1-q)}}}{\bar{y}}\sqrt{\frac{1-q}{8\pi}} \\ + \theta[\sigma(y-\rho\bar{y})]\theta[\sigma(\bar{y}-\rho y)]\frac{1}{2\pi}\frac{(1-q)(1-\rho^2)^{3/2}}{(\bar{y}-\rho y)(y-\rho\bar{y})}\, e^{-\frac{1}{2(1-q)}\boldsymbol{y}^T\mathcal{R}^{-1}\boldsymbol{y}} , \quad (7.65)$$

with the theta functions selecting in turn one of the above cases.

So far, the computation is a minor variation with respect to the one we reported in Sec. 6.2 to evaluate the storage capacity for doublets. Note however that in the definition of $L_{\rm RS}(\boldsymbol{y})$ in Eq. (7.63) we still have to perform the sum over the labels, which in the other case was performed quenched and not replicated. In the summation over $\sigma = \pm 1$, in each domain of $\boldsymbol{y}$ survives only the dominant addend in $(1-q)$: this is the finite term in the first and third quadrant, and the terms proportional to $\exp\{-y^2/[2(1-q)]\}$ or $\exp\{-\bar{y}^2/[2(1-q)]\}$ in the second and forth quadrant (the quadrants bisectors discriminating the larger). In the end, using the obvious symmetry between y and $\bar{y}$ as integration variables and ignoring suppressed factors in $(1-q)$, we get

$$\int_{-\infty}^{+\infty} \frac{\mathrm{d}^2\boldsymbol{y}}{2\pi\sqrt{1-\rho^2}}\, e^{-\frac{1}{2}\boldsymbol{y}^T\mathcal{R}^{-1}\boldsymbol{y}} \log\left[L_{\rm RS}(\boldsymbol{y})\right] \\ = \int_0^{+\infty} \frac{\mathrm{d}y}{\pi\sqrt{1-\rho^2}} \frac{-y^2}{1-q} \int_{-\infty}^{-y} \mathrm{d}\bar{y}\, e^{-\frac{1}{2}(y,\,\bar{y})\mathcal{R}^{-1}\binom{y}{\bar{y}}} \quad (7.66) \\ = \frac{1}{4(1-q)}\left(\frac{2}{\pi}\sqrt{1-\rho^2} - \frac{4}{\pi}\arctan\frac{\sqrt{1-\rho}}{\sqrt{1+\rho}}\right) .$$

Selecting only the most divergent terms in $(1-q)$ from (7.60) and (7.66), we have all the ingredients to evaluate the replica limit of $(\overline{V^t}-1)/t$ for $t\to 0$. The result is zero when the load α assumes the critical value

$$\alpha_*^{\mathrm{RS}}(\rho)=\frac{\pi}{2\arctan\sqrt{(1-\rho)/(1+\rho)}-\sqrt{1-\rho^2}}\,. \tag{7.67}$$

The result is reported in Fig. 7.3: the RS curve presents the expected limits (7.21), but again we do not observe quantitative agreement with the combinatorial curve. We are therefore led to conjecture that we need at least one step of replica symmetry breaking. We work out the derivation of α_* within the 1RSB ansatz in the next section. We stress here that, in order to prove that the RS ansatz is not the right one, we should have performed a stability check similar to the one we explained in Chap. 3 for the de Almeida-Thouless instability. We leave this calculation for future works, contenting us in this thesis with the heuristic comparison with the combinatorial result.

1RSB ansatz

In the 1RSB ansatz the replica matrix has the form

$$Q_{ab}=(1-q_1)\delta_{ab}+(q_1-q_0)\varepsilon_{ab}+q_0\,, \tag{7.68}$$

where $\varepsilon_{ab}=1$ if a,b belongs to a diagonal block $m\times m$, 0 otherwise, so that

$$\log\det(\boldsymbol{Q})\to t\left\{\frac{m-1}{m}\log(1-q_1)+\frac{1}{m}\log[1-q_1+m(q_1-q_0)]+\frac{q_0}{1-q_1+m(q_1-q_0)}\right\}. \tag{7.69}$$

From (7.58), we get

$$\begin{aligned}\sum_{a,b}Q_{ab}\boldsymbol{x}_a^T\mathcal{R}\boldsymbol{x}_b=(1-q_1)\sum_a\boldsymbol{x}_a^T\mathcal{R}\boldsymbol{x}_a+(q_1-q_0)\sum_{B=0}^{t/m-1}\left(\sum_{a=1}^{m}\boldsymbol{x}_{mB+a}\right)^T\mathcal{R}\left(\sum_{b=1}^{m}\boldsymbol{x}_{mB+b}\right)\\ +q_0\left(\sum_a\boldsymbol{x}_a\right)^T\mathcal{R}\left(\sum_b\boldsymbol{x}_b\right),\end{aligned} \tag{7.70}$$

where B is a block index. We now need $2(t/m+1)$ auxiliary HS variables to linearize the sums over replica indices: to get, after the usual factorizations and the integration over $\boldsymbol{x}$,

$$\left\{\int\frac{\mathrm{d}^2\boldsymbol{y}\,e^{-\frac{1}{2}\boldsymbol{y}^T\mathcal{R}^{-1}\boldsymbol{y}}}{2\pi\sqrt{1-\rho^2}}\left[\int\frac{\mathrm{d}^2\boldsymbol{z}\,e^{-\frac{1}{2}\boldsymbol{z}^T\mathcal{R}^{-1}\boldsymbol{z}}}{2\pi\sqrt{1-\rho^2}}\left(\sum_{\{\sigma=\pm1\}}\int_0^{+\infty}\frac{\mathrm{d}^2\boldsymbol{\lambda}}{2\pi(1-q_1)\sqrt{1-\rho^2}}\right.\right.\right.$$
$$\left.\left.\left.\times\,e^{-\frac{[\sigma(\sqrt{q_1-q_0}\boldsymbol{z}+\sqrt{q_0}\boldsymbol{y})+\boldsymbol{\lambda}]^T\mathcal{R}^{-1}[\sigma(\sqrt{q_1-q_0}\boldsymbol{z}+\sqrt{q_0}\boldsymbol{y})+\boldsymbol{\lambda}]}{2(1-q_1)}}\right)^m\right]^{\frac{t}{m}}\right\}^p. \tag{7.71}$$

Defining $L_{1\mathrm{RSB}}(\boldsymbol{y})$ the argument of the square brackets, we know that the logarithm of the above quantity for $t\to 0$ gives

$$\frac{pt}{m}\int\frac{\mathrm{d}^2\boldsymbol{y}\,e^{-\frac{1}{2}\boldsymbol{y}^T\mathcal{R}^{-1}\boldsymbol{y}}}{2\pi\sqrt{1-\rho^2}}\,\log\left[L_{1\mathrm{RSB}}(\boldsymbol{y})\right]\,. \tag{7.72}$$

To simplify $L_{\text{1RSB}}(\boldsymbol{y})$ and to get an expression similar to the one we studied before, we can shift the $\boldsymbol{z}$ variables to

$$\boldsymbol{z} \to \boldsymbol{z} - \frac{\sqrt{q_0}}{\sqrt{q_1 - q_0}}\boldsymbol{y}\,, \tag{7.73}$$

obtaining

$$L_{\text{1RSB}}(\boldsymbol{y}) = \int \frac{\mathrm{d}^2\boldsymbol{z}\, e^{-\frac{1}{2}\left(\boldsymbol{z}-\frac{\sqrt{q_0}}{\sqrt{q_1-q_0}}\boldsymbol{y}\right)^T \mathcal{R}^{-1}\left(\boldsymbol{z}-\frac{\sqrt{q_0}}{\sqrt{q_1-q_0}}\boldsymbol{y}\right)}}{2\pi\sqrt{1-\rho^2}} \times \left(\sum_{\{\sigma=\pm 1\}} \int_0^{+\infty} \frac{\mathrm{d}^2\boldsymbol{\lambda}\, e^{-\frac{\left[\sigma\sqrt{q_1-q_0}\boldsymbol{z}+\boldsymbol{\lambda}\right]^T \mathcal{R}^{-1}\left[\sigma\sqrt{q_1-q_0}\boldsymbol{z}+\boldsymbol{\lambda}\right]}{2(1-q_1)}}}{2\pi(1-q_1)\sqrt{1-\rho^2}}\right)^m . \tag{7.74}$$

In order to find the critical load, we investigate the behavior of the 1RSB parameters close to the transition: it turns out that q_1 has to be sent to one (in analogy with the RS case) and m to zero [BMZ19] as

$$q_1 \to 1\,, \qquad m \to (1-q_1)w\,, \tag{7.75}$$

with w a finite parameter. In this limit we can evaluate the integral over $\boldsymbol{\lambda}$ with a saddle point. We get

$$\theta(z)\theta(\bar{z}) + \theta(-z)\theta(-\bar{z}) + 4\theta(z)\theta(-\bar{z}-z)e^{-\frac{w(1-q_0)z^2}{2}}\,. \tag{7.76}$$

Analytical computations are rather cumbersome after this point. However, the result simplifies a lot if we take $q_0 = 0$. Then the integral over $\boldsymbol{y}$ decouples and simply gives 1, while the one over $\boldsymbol{z}$ breaks into the regions

$$\int_0^{+\infty} \frac{\mathrm{d}^2\boldsymbol{z}}{\pi\sqrt{1-\rho^2}}\, e^{-\frac{1}{2}\boldsymbol{z}^T\mathcal{R}\boldsymbol{z}} = \frac{1}{2} + \frac{1}{\pi}\arcsin(\rho) \tag{7.77}$$

and

$$\int_0^{+\infty} \frac{2\,\mathrm{d}z}{\pi\sqrt{1-\rho^2}} \int_{-\infty}^{-z} \mathrm{d}\bar{z}\, e^{-\frac{1}{2}\boldsymbol{z}^T\mathcal{R}\boldsymbol{z} - \frac{wz^2}{2}} = \frac{2\arctan\left(\sqrt{(1+w)\frac{1-\rho}{1+\rho}}\right)}{\pi\sqrt{1+w}}\,. \tag{7.78}$$

In the end, we find

$$\alpha_*^{\text{1RSB}}(\rho; q_0 = 0, w) = \frac{-\log[1+w]}{2\log\left[\frac{1}{2} + \frac{1}{\pi}\arcsin(\rho) + \frac{2\arctan\left(\sqrt{(1+w)\frac{1-\rho}{1+\rho}}\right)}{\pi\sqrt{1+w}}\right]}\,. \tag{7.79}$$

We stress that this last result is not the optimal 1RSB solution: in principle we should consider the full expression of $\alpha_*^{\text{1RSB}}(\rho; q_0, w)$ and optimize upon the remaining parameters q_0 and w. However, this is beyond the scope of this section: here, we simply verify that the functional form $\alpha_*^{\text{1RSB}}(\rho; q_0 = 0, w)$ allows to fit nicely the combinatorial result, by a adjusting the parameter w (see Fig. 7.3). This simple observation strongly supports our conjecture that this SAT-UNSAT transition exhibits at least one step of RSB, but it does not rule out a full-RSB scenario.

CHAPTER 8

Margin learning from data structure point of view

In this chapter, we reprise the problem of margin learning we briefly mentioned in Sec. 5.4, exploiting our new understanding of the phase transition driven by data structure we found in the problem of simplex learning, in the previous chapter. Indeed, learning with margin can be interpreted geometrically as the linear separation of spheres, as we explain below. In this context, we find the same additional satisfiability transition, concluding that this feature is a general property in the classification problems of extended objects. The exposition is again drawn from [Pas+20].

8.1 Margin learning

Given a kernel machine with feature map $\phi : \mathbb{R}^n \to \mathbb{R}^d$ and inputs $X \in \mathcal{X} = \mathbb{R}^n$, learning with margin κ is defined by the class $\mathcal{G}(\kappa)$ of all functions

$$g_\kappa(X) = \begin{cases} +1 & W \cdot \phi(X) > \kappa \\ -1 & W \cdot \phi(X) < -\kappa. \end{cases} \tag{8.1}$$

Cases falling within the margin $(-\kappa, \kappa)$ can be defined with a third value, for instance 0, or left undefined. Hence, the corresponding loss class projected on a sample (X_p, Y_p), Eqs. (5.8) and (5.9), contains all the dichotomies of X_p that can be realized by an element of $\mathcal{G}_{\mathrm{M}}(\kappa)$.

An alternative representation of margin learning can be given via the definition of appropriate object manifolds. In fact, linear separation of points with margin κ is equivalent to zero-margin linear separation of spherical object manifolds with radius κ [CLS18]: see Fig. 8.1. Thus, $Y^\mu = g_\kappa(X^\mu)$ for all μ if and only if $Y^\mu = g_0(Q^\mu)$ for all μ and all Q^μ such that $|Q^\mu - \phi(X^\mu)|^2 < \kappa^2$. The input space $\mathcal{X}_{\mathrm{M}}(\kappa)$ is the set of the preimages, via ϕ, of all spheres of radius κ in $\mathbb{R}^d$. Note that, while margin learning has a natural description in terms of the original space $\mathcal{X} = \mathbb{R}^n$, through the hypothesis class $\mathcal{G}_{\mathrm{M}}(\kappa)$, simplex learning does not have such a straightforward representation, and is defined directly by means of the object space $\mathcal{X}^{\mathrm{S}}(\{\rho_{ab}\})$.

The VC entropy for margin learning, $\mathcal{H}_\kappa$, can be bounded from above by means of the VC dimension $d_{\mathrm{VC}}(\kappa)$:

$$\mathcal{H}_\kappa \leq d_{\mathrm{VC}}(\kappa) \log p, \quad p > d_{\mathrm{VC}}(\kappa). \tag{8.2}$$

In turn, an upper bound of the VC dimension exists for points lying on the d-dimensional sphere of radius R [Vap99]:

$$d_{\mathrm{VC}}(\kappa) \leq \min \left[\frac{R^2}{\kappa^2}, d \right]. \tag{8.3}$$

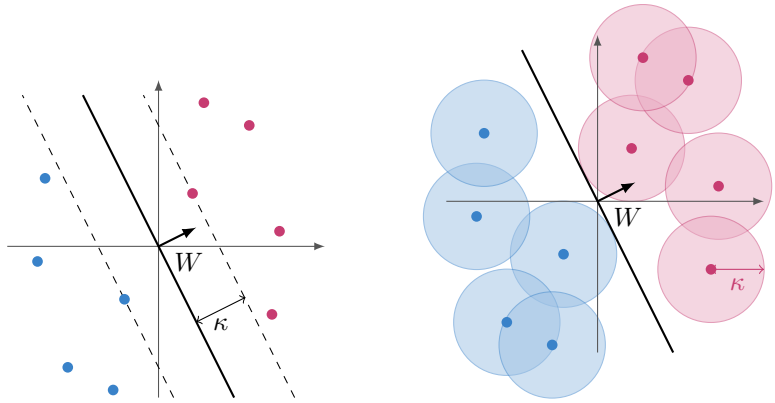

Figure 8.1: Classifying isolated points with margin κ is equivalent to classifying overlapping spheres with radius κ

The standard bound is therefore again logarithmic in the sample size p.

8.2 Satisfiability transition in margin learning

Replica theory turns out to be essential to explore the role of data structure whenever alternative, *ad hoc* methods (such as the combinatorial one) are not available. Here we apply it to identify the SAT-UNSAT transition occurring in margin learning. The synaptic volume relevant to this case is

$$V_\kappa = \sum_{\{\sigma^\mu=\pm 1\}} \int \left[\prod_{j=1}^{n} \mathrm{d}W_j\right] \delta\left(\sum_{j=1}^{n} W_j^2 - n\right) \prod_{\mu=1}^{p} \theta\left(\frac{\sigma^\mu}{\sqrt{n}} \sum_{j=1}^{n} W_j \xi_j^\mu - \kappa\right) , \tag{8.4}$$

where κ is the margin. Note again that here, as in the case of Eq. (7.47), the outputs σ^μ are dynamical variables, at variance with the usual Gardner's volume. We skip the details on the annealed and quenched calculations, which are in spirit very similar to those of the previous sections. Nonetheless, it is worth to point out that the tricky multivariate integrals in the auxiliary variable, are now replaced by Gaussian integrals, with the margin κ appearing as an integration limit. The annealed approximation leads to

$$\alpha_*^{\mathrm{A}}(\kappa) = -\frac{1+\log(2\pi)}{2\log[2\,\mathrm{erfc}(\kappa)]} . \tag{8.5}$$

In the quenched calculation, the RS ansatz is again implemented by requiring $q \to 1$; one obtains the critical threshold

$$\alpha_*^{\mathrm{RS}}(\kappa) = \frac{1}{2}\left[\int_0^\kappa \mathrm{D}y\,(\kappa - y)^2\right]^{-1} , \tag{8.6}$$

where $\mathrm{D}y$ is the Gaussian measure. Note the difference with Gardner's result (5.39) for the storage capacity. The one-step RSB ansatz again depends on the parameters q_0 and w, which should be investigated numerically. However, in the special case $q_0 = 0$ we

find the simpler expression

$$\alpha_*^{\text{1RSB}}(\kappa; q_0 = 0, w) = \frac{-\log[1+w]}{2\log\left\{2\left[\text{erfc}(\kappa) + \int_0^\kappa \text{D}z\, e^{-w\frac{(z-\kappa)^2}{2}}\right]\right\}}\,. \tag{8.7}$$

These results essentially share the same features of those for the simplexes computed above: in particular, at variance with the usual storage capacity (5.39), α_* computed in all the different approximation schemes diverges in the limit $\kappa \to 0^+$, when the problem reduces to a standard classification of points (or equivalently, in the object manifold description, when the radius of the spheres shrinks to zero). Even in absence of a closed expression for the VC entropy of margin classification, the existence of the phase transition at a finite load is a clear indication of its non-monotonicity.

Some of these facts were already pointed out in the seminal paper [Opp99].

Discussion

Understanding how data specificities impact the performance of machine learning models and algorithms can be considered one of the major challenges for contemporary statistical physics. Here we have shown how to deal with data structure, as it is being established in Physics, within the framework of the statistical theory of learning. The presence of input-output correlations in a dataset suggests constraints to be applied to the hypothesis class under consideration. As a result, the corresponding VC entropy, deeply connected to the generalization capabilities of the model, is considerably lower than in the unstructured case.

For simple models of data structure we have observed two striking phenomena that take place above the VC dimension. First, the VC entropy becomes non-monotonic. This is a strong indication that the rigorous bounds in SLT may be substantially improved by taking data structure into account. Second, a novel transition appears beyond the well-known storage capacity, at the onset of unsatisfiability for a data-related constraint satisfaction problem. When available, a combinatorial theory *à la Cover* allows one to compute the VC entropy of a finite-size system, and to reveal explicitly its non-monotonic behavior. However, this is not always feasible, such as for spherical object manifolds and margin learning. In these cases, we showed how the phase transition can be probed with the standard tools of statistical physics, thus allowing an indirect quantification of the data-dependent behavior.

The new satisfiability transition is due to a competition between the combinatorial expansion, with sample size, of the space of possible functions and the reduction due to the constraints given by the geometrical structure. We believe, as this observation suggests, that the emergence of the data-driven transition, as well as the non-monotonic VC entropy it entails, is not specific to the two models of data that we have studied here, but is more generally present whenever the constraints imposed on the hypothesis class by data structure are strong enough. On a more quantitative level, notice that the upper and lower bounds obtained for α_* in Sec. 7.1.5 are very close to one another. The bounds are independent of the particular choice of simplexes, i.e., they do not depend on k or on $\{\rho_{ab}\}$. This is a clue pointing to the robustness of the phenomenology for disparate data structures. We remark that the combinatoric analysis was done at leading order in α; thus, it remains to assess how much the bounds are affected by perturbative corrections.

An ambitious and pressing goal concerns the generalization of our results to other architectures, notably deep neural networks, in the same spirit of what was achieved in SLT regarding the VC dimension.

Bibliography

[Abb+19] A. Abbara et al. *Rademacher complexity and spin glasses: A link between the replica and statistical theories of learning*. 2019. arXiv: 1912.02729 [cond-mat.dis-nn].

[ABM04] A. Andreanov, F. Barbieri, and O. C. Martin. "Large deviations in spin-glass ground-state energies". In: *The European Physical Journal B - Condensed Matter and Complex Systems* 41.3 (Oct. 2004), pp. 365–375. ISSN: 1434-6036.

[AGS85a] D. J. Amit, H. Gutfreund, and H. Sompolinsky. "Spin-glass models of neural networks". In: *Phys. Rev. A* 32 (2 Aug. 1985), pp. 1007–1018.

[AGS85b] D. J. Amit, H. Gutfreund, and H. Sompolinsky. "Storing Infinite Numbers of Patterns in a Spin-Glass Model of Neural Networks". In: *Phys. Rev. Lett.* 55 (14 Sept. 1985), pp. 1530–1533.

[And88] P. W. Anderson. "Spin Glass I: A Scaling Law Rescued". In: *Physics Today* 41.1 (Jan. 1988), p. 9; "Spin Glass II: Is There a Phase Transition?" In: *Physics Today* 41.3 (Jan. 1988), p. 9; "Spin Glass III: Theory Raises its Head". In: *Physics Today* 41.6 (Jan. 1988), p. 9; "Spin Glass II: Is There a Phase Transition?" In: *Physics Today* 41.3 (Jan. 1988), p. 9; "Spin Glass V: Real Power Brought to Bear". In: *Physics Today* 42.7 (Jan. 1989), p. 9; "Spin Glass VI: Spin Glass As Cornucopia". In: *Physics Today* 42.9 (Jan. 1989), p. 9; "Spin Glass VII: Spin Glass as Paradigm". In: *Physics Today* 43.3 (Jan. 1990), p. 9.

[Ang+14] D. Anguita et al. "A Deep Connection Between the Vapnik–Chervonenkis Entropy and the Rademacher Complexity". In: *IEEE Transactions on Neural Networks and Learning Systems* 25.12 (Dec. 2014), pp. 2202–2211. ISSN: 2162-2388.

[Ans+16] F. Anselmi et al. "Unsupervised Learning of Invariant Representations". In: *Theor. Comput. Sci.* 633.C (2016), pp. 112–121. ISSN: 0304-3975.

[Ans+19] A. Ansuini et al. *Intrinsic dimension of data representations in deep neural networks*. 2019. arXiv: 1905.12784 [cs.LG].

[Ant+03] A. Antos et al. "Data-Dependent Margin-Based Generalization Bounds for Classification". In: *J. Mach. Learn. Res.* 3.null (Mar. 2003), pp. 73–98. ISSN: 1532-4435.

[AT78] J. R. L. de Almeida and D. J. Thouless. "Stability of the Sherrington-Kirkpatrick solution of a spin glass model". In: *Journal of Physics A: Mathematical and General* 11.5 (May 1978), pp. 983–990.

[Bal+16] C. Baldassi et al. "Unreasonable effectiveness of learning neural networks: From accessible states and robust ensembles to basic algorithmic schemes". In: *Proceedings of the National Academy of Sciences* 113.48 (2016), E7655–E7662. ISSN: 0027-8424.

[Bar97] A. Barrat. *The p-spin spherical spin glass model*. 1997. arXiv: cond-mat/9701031 [cond-mat.dis-nn].

[BBL04] O. Bousquet, S. Boucheron, and G. Lugosi. "Introduction to Statistical Learning Theory". In: *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2 - 14, 2003, Tübingen, Germany, August 4 - 16, 2003, Revised Lectures*. Ed. by O. Bousquet, U. von Luxburg, and G. Rätsch. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 169–207. ISBN: 978-3-540-28650-9.

[BGG80] A. Blandin, M. Gabay, and T. Garel. "On the mean-field theory of spin glasses". In: *Journal of Physics C: Solid State Physics* 13.3 (Jan. 1980), pp. 403–418.

[BKL19] H.-H. Boltz, J. Kurchan, and A. J. Liu. *Fluctuation Distributions of Energy Minima in Complex Landscapes*. 2019. arXiv: 1911.08943.

[Bla78] Blandin, A. "Theories versus experiments in the spin glass systems". In: *J. Phys. Colloques* 39 (1978),

[BM03] P. L. Bartlett and S. Mendelson. "Rademacher and Gaussian Complexities: Risk Bounds and Structural Results". In: *J. Mach. Learn. Res.* 3.null (Mar. 2003), pp. 463–482. ISSN: 1532-4435.

[BM78] A. J. Bray and M. A. Moore. "Replica-Symmetry Breaking in Spin-Glass Theories". In: *Phys. Rev. Lett.* 41 (15 Oct. 1978), pp. 1068–1072.

[BM79] A. J. Bray and M. A. Moore. "Replica symmetry and massless modes in the Ising spin glass". In: *Journal of Physics C: Solid State Physics* 12.1 (Jan. 1979), pp. 79–104.

[BM80] A. J. Bray and M. A. Moore. "Broken replica symmetry and metastable states in spin glasses". In: *Journal of Physics C: Solid State Physics* 13.31 (Nov. 1980), pp. L907–L912.

[BMZ19] C. Baldassi, E. M. Malatesta, and R. Zecchina. "Properties of the Geometry of Solutions and Capacity of Multilayer Neural Networks with Rectified Linear Unit Activations". In: *Phys. Rev. Lett.* 123 (17 Oct. 2019), p. 170602.

[Bor+19] F. Borra et al. "Generalization from correlated sets of patterns in the perceptron". In: *Journal of Physics A: Mathematical and Theoretical* 52.38 (Aug. 2019), p. 384004.

[Bot15] L. Bottou. "Making Vapnik–Chervonenkis Bounds Accurate". In: Sept. 2015, pp. 143–155. ISBN: 978-3-319-21851-9.

[Cap+18] R. Capelli et al. "Exact value for the average optimal cost of the bipartite traveling salesman and two-factor problems in two dimensions". In: *Phys. Rev. E* 98 (3 Sept. 2018), p. 030101.

[Car+18] S. Caracciolo et al. "Solution for a bipartite Euclidean traveling-salesman problem in one dimension". In: *Phys. Rev. E* 97 (5 May 2018), p. 052109.

[CC05] T. Castellani and A. Cavagna. "Spin-glass theory for pedestrians". In: *Journal of Statistical Mechanics: Theory and Experiment* 2005.05 (May 2005), P05012.

[Cha+14] P. Charbonneau et al. "Fractal free energy landscapes in structural glasses". In: *Nature Communications* 5.1 (Apr. 2014), p. 3725. ISSN: 2041-1723.

[Cha+19] P. Chaudhari et al. "Entropy-SGD: biasing gradient descent into wide valleys". In: *Journal of Statistical Mechanics: Theory and Experiment* 2019.12 (Dec. 2019), p. 124018.

[CL07] A. Crisanti and L. Leuzzi. "Amorphous-amorphous transition and the two-step replica symmetry breaking phase". In: *Phys. Rev. B* 76 (18 Nov. 2007), p. 184417.

[CLS16] S. Chung, D. D. Lee, and H. Sompolinsky. "Linear readout of object manifolds". In: *Phys. Rev. E* 93 (6 June 2016), p. 060301.

[CLS18] S. Chung, D. D. Lee, and H. Sompolinsky. "Classification and Geometry of General Perceptual Manifolds". In: *Phys. Rev. X* 8 (3 July 2018), p. 031003.

[Coh+20] U. Cohen et al. "Separability and geometry of object manifolds in deep neural networks". In: *Nature Communications* 11.1 (Feb. 2020), p. 746. ISSN: 2041-1723.

[Cov65] T. M. Cover. "Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition". In: *IEEE Transactions on Electronic Computers* EC-14.3 (June 1965), pp. 326–334. ISSN: 0367-7508.

[Cri+90] A. Crisanti et al. "Fluctuations of correlation functions in disordered spin systems". In: *Journal of Physics A: Mathematical and General* 23.13 (July 1990), pp. 3083–3093.

[CS02a] S. Caracciolo and A. Sportiello. "An exactly solvable random satisfiability problem". In: *Journal of Physics A: Mathematical and General* 35.36 (Aug. 2002), pp. 7661–7688.

[CS02b] S. Caracciolo and A. Sportiello. "An exactly solvable random satisfiability problem". In: *Journal of Physics A: Mathematical and General* 35.36 (Aug. 2002), pp. 7661–7688.

[CS18] P. Chaudhari and S. Soatto. "Stochastic Gradient Descent Performs Variational Inference, Converges to Limit Cycles for Deep Networks". In: *2018 Information Theory and Applications Workshop (ITA)*. 2018, pp. 1–10.

[CS92] A. Crisanti and H. J. Sommers. "The spherical p-spin interaction spin glass model: the statics". In: *Zeitschrift für Physik B Condensed Matter* 87.3 (Oct. 1992), pp. 341–354. ISSN: 1431-584X.

[CT92] D. Cohn and G. Tesauro. "How Tight Are the Vapnik-Chervonenkis Bounds?" In: *Neural Computation* 4.2 (Mar. 1992), pp. 249–269. ISSN: 0899-7667.

[CV95] C. Cortes and V. Vapnik. "Support-vector networks". In: *Machine Learning* 20.3 (Sept. 1995), pp. 273–297. ISSN: 1573-0565.

[Der80] B. Derrida. "Random-Energy Model: Limit of a Family of Disordered Models". In: *Phys. Rev. Lett.* 45 (2 July 1980), pp. 79–82.

[Der81] B. Derrida. "Random-energy model: An exactly solvable model of disordered systems". In: *Phys. Rev. B* 24 (5 Sept. 1981), pp. 2613–2626.

[DFM94] V. Dotsenko, S. Franz, and M. Mezard. "Partial annealing and overfrustration in disordered systems". In: *Journal of Physics A: Mathematical and General* 27.7 (Apr. 1994), pp. 2351–2365.

[DG06] C. De Dominicis and I. Giardina. *Random Fields and Spin Glasses: A Field Theory Approach*. Cambridge University Press, 2006.

[DM06] D. S. Dean and S. N. Majumdar. "Large Deviations of Extreme Eigenvalues of Random Matrices". In: *Phys. Rev. Lett.* 97 (16 Oct. 2006), p. 160201.

[DM08] D. S. Dean and S. N. Majumdar. "Extreme value statistics of eigenvalues of Gaussian random matrices". In: *Physical Review E* 77.4 (2008), p. 041108.

[Dot00] V. Dotsenko. *Introduction to the Replica Theory of Disordered Statistical Systems*. Collection Alea-Saclay: Monographs and Texts in Statistical Physics. Cambridge University Press, 2000.

[Dot11] V. Dotsenko. "Replica solution of the random energy model". In: *EPL (Europhysics Letters)* 95.5 (Aug. 2011), p. 50006.

[Dot95] V. Dotsenko. *An Introduction to the Theory of Spin Glasses and Neural Networks*. World Scientific, 1995. eprint: https://www.worldscientific.com/doi/pdf/10.1142/2460.

[DZ10] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*. Springer-Verlag, 2010.

[EA75] S. F. Edwards and P. W. Anderson. "Theory of spin glasses". In: *Journal of Physics F: Metal Physics* 5.5 (May 1975), pp. 965–974.

[EB01] A. Engel and C. P. L. V. d. Broeck. *Statistical Mechanics of Learning*. New York, NY, USA: Cambridge University Press, 2001. ISBN: 0521774799.

[EGR19] V. Erba, M. Gherardi, and P. Rotondo. "Intrinsic dimension estimation for locally undersampled data". In: *Scientific Reports* 9.1 (Nov. 2019), p. 17133. ISSN: 2045-2322.

[EJ76] S. F. Edwards and R. C. Jones. "The eigenvalue spectrum of a large symmetric random matrix". In: *Journal of Physics A: Mathematical and General* 9.10 (Oct. 1976), pp. 1595–1603.

[Ell06] R. Ellis. *Entropy, Large Deviations, and Statistical Mechanics*. Springer-Verlag, 2006.

[Erb+20] V. Erba et al. "Random geometric graphs in high dimension". In: *Phys. Rev. E* 102 (1 July 2020), p. 012306.

[Fac+17] E. Facco et al. "Estimating the intrinsic dimension of datasets by a minimal neighborhood information". In: *Scientific Reports* 7.1 (Sept. 2017).

[FFM07] M. Fedrigo, F. Flandoli, and F. Morandin. "A Large Deviation Principle for the free energy of random Gibbs measures with application to the REM". In: *Annali di Matematica Pura ed Applicata* 186.3 (July 2007), pp. 381–417. ISSN: 1618-1891.

[FHU19] S. Franz, S. Hwang, and P. Urbani. "Jamming in Multilayer Supervised Learning Models". In: *Phys. Rev. Lett.* 123 (16 Oct. 2019), p. 160602.

[FR20] S. Franz and J. Rocchi. "Large deviations of glassy effective potentials". In: *Journal of Physics A: Mathematical and Theoretical* (2020).

[FS09] P. Flajolet and R. Sedgewick. *Analytic combinatorics*. cambridge University press, 2009.

[Gar87] E. Gardner. "Maximum Storage Capacity in Neural Networks". In: *Europhysics Letters (EPL)* 4.4 (Aug. 1987), pp. 481–485.

[Gar88] E. Gardner. "The space of interactions in neural network models". In: *Journal of Physics A: Mathematical and General* 21.1 (Jan. 1988), pp. 257–270.

[GBC16] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. The MIT Press, 2016. ISBN: 0262035618, 9780262035613.

[GD88] E. Gardner and B. Derrida. "Optimal storage properties of neural network models". In: *Journal of Physics A: Mathematical and General* 21.1 (Jan. 1988), pp. 271–284.

[GD89] E. Gardner and B. Derrida. "The probability distribution of the partition function of the random energy model". In: *Journal of Physics A: Mathematical and General* 22.12 (June 1989), pp. 1975–1981.

[Ger+20] F. Gerace et al. *Generalisation error in learning with random features and the hidden manifold model*. 2020. arXiv: 2002.09339 [math.ST].

[GM84] D. Gross and M. Mezard. "The simplest spin glass". In: *Nuclear Physics B* 240.4 (1984), pp. 431–452. ISSN: 0550-3213.

[Gol+20] S. Goldt et al. "Modelling the influence of data structure on learning in neural networks: the hidden manifold model". In: (2020). arXiv: 1909.11500 [stat.ML].

[Gue03] F. Guerra. "Broken Replica Symmetry Bounds in the Mean Field Spin Glass Model". In: *Communications in Mathematical Physics* 233.1 (Feb. 2003), pp. 1–12. ISSN: 1432-0916.

[Gue13] F. Guerra. "Spontaneous Replica Symmetry Breaking and Interpolation Methods for Complex Statistical Mechanics Systems". In: *Correlated Random Systems: Five Different Methods*. Ed. by V. Gayrard and N. Kistler. Springer, 2013. Chap. 2, pp. 45–70.

[Gup63] S. S. Gupta. "Probability Integrals of Multivariate Normal and Multivariate t^1". In: *Ann. Math. Statist.* 34.3 (Sept. 1963), pp. 792–828.

[He+16] K. He et al. "Deep Residual Learning for Image Recognition". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2016.

[HLP34] G. Hardy, J. Littlewood, and G. Pólya. *Inequalities*. Cambridge University Press, 1934.

[HP79] J. L. van Hemmen and R. G. Palmer. "The replica method and solvable spin glass model". In: *Journal of Physics A: Mathematical and General* 12.4 (Apr. 1979), pp. 563–580.

[HZS06] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew. "Extreme learning machine: Theory and applications". In: *Neurocomputing* 70.1 (2006). Neural Networks, pp. 489–501. ISSN: 0925-2312.

[KLL01] B. Kégl, T. Linder, and G. Lugosi. "Data-Dependent Margin-Based Generalization Bounds for Classification". In: *Computational Learning Theory*. Ed. by D. Helmbold and B. Williamson. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001, pp. 368–384. ISBN: 978-3-540-44581-4.

[Kon83] I. Kondor. "Parisi's mean-field solution for spin glasses as an analytic continuation in the replica number". In: *Journal of Physics A: Mathematical and General* 16.4 (Mar. 1983), pp. L127–L131.

[KPA93] J. Kurchan, G. Parisi, and V. M. A. "Barriers and metastable states as saddle points in the replica approach". In: *J. Phys. I France* 3.8 (1993), pp. 1819–1838.

[Krz+07] F. Krzakala et al. "Gibbs states and the set of solutions of random constraint satisfaction problems". In: *Proceedings of the National Academy of Sciences* 104.25 (2007), pp. 10318–10323. ISSN: 0027-8424. eprint: https://www.pnas.org/content/104/25/10318.full.pdf.

[KS94] S. Kirkpatrick and B. Selman. "Critical Behavior in the Satisfiability of Random Boolean Expressions". In: *Science* 264.5163 (1994), pp. 1297–1301. ISSN: 0036-8075.

[KTJ76] J. M. Kosterlitz, D. J. Thouless, and R. C. Jones. "Spherical Model of a Spin-Glass". In: *Phys. Rev. Lett.* 36 (20 May 1976), pp. 1217–1220.

[LBH15] Y. LeCun, Y. Bengio, and G. Hinton. "Deep learning". In: *Nature* 521.7553 (May 2015), pp. 436–444. ISSN: 1476-4687.

[LRZ01] M. Leone, F. Ricci-Tersenghi, and R. Zecchina. "Phase coexistence and finite-size scaling in random combinatorial problems". In: *Journal of Physics A: Mathematical and General* 34.22 (May 2001), pp. 4615–4626.

[LS18] B. Li and D. Saad. "Exploring the Function Space of Deep-Learning Machines". In: *Phys. Rev. Lett.* 120 (24 June 2018), p. 248301.

[Mal16] S. Mallat. "Understanding deep convolutional networks". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374.2065 (2016), p. 20150203.

[Maz+18a] A. Mazzolini et al. "Statistics of Shared Components in Complex Component Systems". In: *Phys. Rev. X* 8 (2 Apr. 2018), p. 021023.

[Maz+18b] A. Mazzolini et al. "Zipf and Heaps laws from dependency structures in component systems". In: *Phys. Rev. E* 98 (1 July 2018), p. 012315.

[McC10] B. M. McCoy. *Advanced statistical mechanics*. Oxford: Oxford University Press, 2010. ISBN: 9780199556632 0199556636.

[McK84] J. P. McKelvey. "Simple transcendental expressions for the roots of cubic equations". In: *American Journal of Physics* 52.3 (1984), pp. 269–270. eprint: https://doi.org/10.1119/1.13706.

[Meh+19] P. Mehta et al. "A high-bias, low-variance introduction to Machine Learning for physicists". In: *Physics Reports* 810 (2019). A high-bias, low-variance introduction to Machine Learning for physicists, pp. 1–124. ISSN: 0370-1573.

[Meh04] M. L. Mehta. *Random Matrices*. Elsevier, 2004.

[Méz17] M. Mézard. "Mean-field message-passing equations in the Hopfield model and its generalizations". In: *Phys. Rev. E* 95 (2 Feb. 2017), p. 022117.

[MG10] C. Monthus and T. Garel. "Matching between typical fluctuations and large deviations in disordered systems: application to the statistics of the ground state energy in the SK spin-glass model". In: *Journal of Statistical Mechanics: Theory and Experiment* 2010.02 (Feb. 2010), P02023.

[MM09] M. Mézard and A. Montanari. *Information, Physics, and Computation*. Oxford University Press, 2009.

[MM19] C. H. Martin and M. W. Mahoney. *Rethinking generalization requires revisiting old ideas: statistical mechanics approaches and complex learning behavior*. 2019. arXiv: 1710.09553 [cs.LG].

[MMN18] S. Mei, A. Montanari, and P.-M. Nguyen. "A mean field view of the landscape of two-layer neural networks". In: *Proceedings of the National Academy of Sciences* 115.33 (2018), E7665–E7671. ISSN: 0027-8424. eprint: https://www.pnas.org/content/115/33/E7665.full.pdf.

[MP99] M. Mézard and G. Parisi. "Thermodynamics of glasses: a first principles computation". In: *Journal of Physics: Condensed Matter* 11.10A (Jan. 1999), A157–A165.

[MPS19] E. M. Malatesta, G. Parisi, and G. Sicuro. "Fluctuations in the random-link matching problem". In: *Phys. Rev. E* 100 (3 Sept. 2019), p. 032102.

[MPV86] M. Mézard, G. Parisi, and M. A. Virasoro. *Spin Glass Theory and Beyond*. World Scientific, 1986. eprint: https://www.worldscientific.com/doi/pdf/10.1142/0271.

[MS16] J. Maldacena and D. Stanford. "Remarks on the Sachdev-Ye-Kitaev model". In: *Phys. Rev. D* 94 (10 Nov. 2016), p. 106002.

[Ney+17] B. Neyshabur et al. "Exploring Generalization in Deep Learning". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., 2017, pp. 5949–5958. ISBN: 9781510860964.

[NH08] T. Nakajima and K. Hukushima. "Large Deviation Property of Free Energy in p-Body Sherrington–Kirkpatrick Model". In: *Journal of the Physical Society of Japan* 77.7 (2008), p. 074718. eprint: https://doi.org/10.1143/JPSJ.77.074718.

[NH09] T. Nakajima and K. Hukushima. "Thermodynamic construction of a one-step replica-symmetry-breaking solution in finite-connectivity spin glasses". In: *Phys. Rev. E* 80 (1 July 2009), p. 011103.

[Nis01] H. Nishimori. *Statistical Physics of Spin Glasses and Information Processing: An Introduction*. Oxford University Press, 2001.

[OK04] K. Ogure and Y. Kabashima. "Exact Analytic Continuation with Respect to the Replica Number in the Discrete Random Energy Model of Finite System Size". In: *Progress of Theoretical Physics* 111.5 (2004), pp. 661–688.

[OK09a] K. Ogure and Y. Kabashima. "On analyticity with respect to the replica number in random energy models: I. An exact expression for the moment of the partition function". In: *Journal of Statistical Mechanics: Theory and Experiment* 2009.03 (Mar. 2009), P03010.

[OK09b] K. Ogure and Y. Kabashima. "On analyticity with respect to the replica number in random energy models: II. Zeros on the complex plane". In: *Journal of Statistical Mechanics: Theory and Experiment* 2009.05 (May 2009), P05011.

[Opp99] M. Opper. *On the annealed VC entropy for margin classifiers: A statistical mechanics study*. Feb. 1999.

[Par79] G. Parisi. "Toward a mean field theory for spin glasses". In: *Physics Letters A* 73.3 (1979), pp. 203–205. ISSN: 0375-9601; "Infinite Number of Order Parameters for Spin-Glasses". In: *Phys. Rev. Lett.* 43 (23 Dec. 1979), pp. 1754–1756; G. Parisi. "The order parameter for spin glasses: a function on the interval 0-1". In: *Journal of Physics A: Mathematical and General* 13.3 (Mar. 1980), pp. 1101–1112; "A sequence of approximated solutions to the S-K model for spin glasses". In: *Journal of Physics A: Mathematical and General* 13.4 (Apr. 1980), pp. L115–L121.

[Pas+20] M. Pastore et al. "Statistical learning theory of structured data". In: *Phys. Rev. E* 102 (3 Sept. 2020), p. 032119.

[PDR19] M. Pastore, A. Di Gioacchino, and P. Rotondo. "Large deviations of the free energy in the p-spin glass spherical model". In: *Phys. Rev. Research* 1 (3 Nov. 2019), p. 033116.

[PH19] C. Pérez-Espigares and P. I. Hurtado. "Sampling rare events across dynamical phase transitions". In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 29.8 (2019), p. 083106. eprint: https://doi.org/10.1063/1.5091669.

[PR08] G. Parisi and T. Rizzo. "Large Deviations in the Free Energy of Mean-Field Spin Glasses". In: *Phys. Rev. Lett.* 101 (11 Sept. 2008), p. 117205.

[PR09] G. Parisi and T. Rizzo. "Phase diagram and large deviations in the free energy of mean-field spin glasses". In: *Phys. Rev. B* 79 (13 Apr. 2009), p. 134205.

[PR10a] G. Parisi and T. Rizzo. "Large deviations of the free energy in diluted mean-field spin-glass". In: *Journal of Physics A: Mathematical and Theoretical* 43.4 (2010), p. 045001.

[PR10b] G. Parisi and T. Rizzo. "Universality and deviations in disordered systems". In: *Phys. Rev. B* 81 (9 Mar. 2010), p. 094201.

[Rag+17] M. Raghu et al. "On the Expressive Power of Deep Neural Networks". In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. ICML'17. Sydney, NSW, Australia: JMLR.org, 2017, pp. 2847–2854.

[Ram81] R. Rammal. In: *PhD thesis (unpublished), Grenoble University* (1981).

[Riv05] O. Rivoire. "The cavity method for large deviations". In: *Journal of Statistical Mechanics: Theory and Experiment* 2005.07 (July 2005), P07004–P07004.

[RLG20] P. Rotondo, M. C. Lagomarsino, and M. Gherardi. "Counting the learnable functions of geometrically structured data". In: *Phys. Rev. Research* 2 (2 May 2020), p. 023169.

[RPG20] P. Rotondo, M. Pastore, and M. Gherardi. "Beyond the Storage Capacity: Data-Driven Satisfiability Transition". In: *Phys. Rev. Lett.* 125 (12 Sept. 2020), p. 120601.

[Sha+98] J. Shawe-Taylor et al. "Structural risk minimization over data-dependent hierarchies". In: *IEEE Transactions on Information Theory* 44.5 (Sept. 1998), pp. 1926–1940. ISSN: 1557-9654.

[SK75] D. Sherrington and S. Kirkpatrick. "Solvable Model of a Spin-Glass". In: *Phys. Rev. Lett.* 35 (26 Dec. 1975), pp. 1792–1796.

[SL00] H. S. Seung and D. D. Lee. "The Manifold Ways of Perception". In: *Science* 290.5500 (2000), pp. 2268–2269. ISSN: 0036-8075.

[Son98] E. D. Sontag. "VC dimension of neural networks". In: *NATO ASI Series F Computer and Systems Sciences* 168 (1998), pp. 69–96.

[TAP77] D. J. Thouless, P. W. Anderson, and R. G. Palmer. "Solution of 'Solvable model of a spin glass'". In: *The Philosophical Magazine: A Journal of Theoretical Experimental and Applied Physics* 35.3 (1977), pp. 593–601. eprint: https://doi.org/10.1080/14786437708235992.

[TD81] Toulouse, G. and Derrida, B. "Free energy probability distribution in the SK spin glass model". In: *Proceedings of the Sixth Symposium on Theoretical Physics*. Rio de Janeiro, Brazil, 1981, p. 217.

[TFI89] T. Tanaka, H. Fujisaka, and M. Inoue. "Free-energy fluctuations in a one-dimensional random Ising model". In: *Phys. Rev. A* 39 (6 Mar. 1989), pp. 3170–3172.

[Tou09] H. Touchette. "The large deviation approach to statistical mechanics". In: *Physics Reports* 478.1 (2009), pp. 1–69. ISSN: 0370-1573.

[Vap13] V. Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

[Vap99] V. N. Vapnik. "An overview of statistical learning theory". In: *IEEE Transactions on Neural Networks* 10.5 (Sept. 1999), pp. 988–999. ISSN: 1941-0093.

[Vul+14] A. Vulpiani et al. *Large Deviations in Physics*. Springer-Verlag, 2014.

[Zam14] F. Zamponi. *Mean field theory of spin glasses*. 2014. arXiv: 1008.4844 [cond-mat.stat-mech].

[Zha+17] C. Zhang et al. "Understanding deep learning requires rethinking generalization". In: *Proceedings of the International Conference on Learning Representations*. 2017.