LIMBURGS UNIVERSITAIR CENTRUM
IN HET CENTRUM VAN DE KENNIS

Departement Natuurkunde

Unsupervised Learning of Binary Vectors

Proefschrift voorgelegd tot het behalen van de graad van doctor in de wetenschappen groep natuurkunde aan het Limburgs Universitair Centrum te verdedigen door

Mauro Copelli Lopes da Silva

Promotor: Prof. Dr. C. Van den Broeck

1999

To Anna Cecília.

According to the guidelines of the Limburgs Universitair Centrum, a copy of this publication has been filed in the Royal Belgian Library Albert I, Brussels, as publication D/1999/2451/122.

Acknowledgements

First I would like to thank my advisor, Prof. Christian Van den Broeck, for the opportunity he gave me when he invited me to come to Belgium. His patient guidance, availability and experience were essential for this work. I am also deeply indebted to Prof. Marc Bouten for his invaluable contributions to my education during these last years. I am honored to have published a paper with him before his retirement. I would also like to thank the other members of the Theoretical Physics group, whom I could always count on for discussions and suggestions: Prof. Carlo Vanderzande, Prof. Roger Serneels, Dr. Bart Van Rompaey, Ioana Bena, Frank Daerden, Peter Leoni and Jef Hooyberghs. Thanks also to colleagues and friends at the LUC for making daily life more enjoyable: Dr. Geert Jan Bex, Martine Van Gastel, Dr. Luc Vandeurzen, Kristien Meykens, Isabelle Vanschoonbeek, Frank Neven, Javier Buceta and Dörthe Malzahn.

Two among my colleagues at the LUC deserve a special mention: Dr. Bart Van Rompaey, my office-mate, for his friendship during all these years and for keeping the spirit up, and Dr. Geert Jan Bex, who helped me with much more than computer problems. I am also thankful to the members of the LUC staff, for their constant assistance with administrative issues: Conny Wijnants, Hilde Wijnen, Viviane Mebis, Martine Machiels and Annemie Hermans, at the WNI Secretary. I acknowledge financial support from FWO Vlaanderen, the Belgian IUAP program (Prime Minister's Office) and, during the first year of my PhD, FAPESP (Brazil).

I greatly benefited from discussions and collaborations with several researchers/friends, among others: Prof. Nestor Caticha, Dr. Manfred Opper, Dr. Peter Reimann, Dr. Mirta Gordon, Dr. Arnaud Buhot, Dr. Michael Biehl, Dr. Peter Riegler, Prof. Wolfgang Kinzel, Dr. Osame Kinouchi and Renato Vicente. I am specially grateful to Prof. Nestor Caticha for his supervision while I was in São Paulo, as well as his friendship and collaboration ever since. I would also like to thank the opportunities I was given to visit and learn from the groups in Würzburg, King's College (London) and CEA/Grenoble.

Last but not least, thanks are due to my family and friends, who supported me through the difficult moments and/or shared the good ones. I am happy to realize the list is long, yet far from complete. To all of you, my thankfulness and love: Anna Cecília Copelli, Cristina Frigo, Dionízio A. B. Souza, Eduardo S. Cypriano, Ian Ribas, Ivan C. L. da Silva, Joana L. Baracuhy, João F. H. Jornada, Jörn "Pilli" Spiller, Juliana A. Pinto, Luís Felipe G. Pinto, Marcelo C. Pereira, Mariana R. Joffily, Nara C. Guisoni,

Pablo I. Rovira and Valéria Reginatto.

Finally, my deepest gratitude goes to Gêisa Fernandes, without whose love, encouragement and support none of this would have been possible.

Abstract

In this thesis, unsupervised learning of binary vectors from data is studied using methods from Statistical Mechanics of disordered systems. In the model, data vectors are distributed according to a single symmetry breaking direction. The aim of unsupervised learning is to provide a good approximation to this direction. The difference with respect to previous studies is the knowledge that this preferential direction has binary components.

It is shown that sampling from the posterior distribution (Gibbs learning) leads, for general smooth distributions, to an exponentially fast approach to perfect learning in the asymptotic limit of large number of examples. If the distribution is non-smooth, then first order phase transitions to perfect learning are expected. In the limit of poor performance, at the other end of the asymptotics, the binary nature of the preferential direction is irrelevant and the results are the same as for the spherical case: a second order phase transition ("retarded learning") is predicted to occur if the data distribution is not biased or, if the distribution is biased, learning starts off immediately.

Using concepts from Bayesian inference, the center of mass of the Gibbs ensemble is shown to have maximal average (Bayes-optimal) performance. This upper bound for continuous vectors is extended to a discrete space, resulting in the *clipped* center of mass of the Gibbs ensemble having maximal average performance among the *binary* vectors.

In order to calculate the performance of this *best binary* vector, the geometric properties of the center of mass of binary vectors are first studied. The surprising result is found that the center of mass of infinite binary vectors which obey some simple constraints, is again a binary vector. When disorder is taken into account in the calculation, however, the properties of the Bayes-optimal center of mass change completely, leading to a vector with continuous components. The performance of the best binary vector is calculated and shown to always lie above that of Gibbs learning and below the Bayes-optimal performance.

Making use of a variational approach under the replica symmetric *ansatz*, an optimal potential is constructed in the limits of zero temperature and mutual overlap 1. Under these assumptions, minimization of this potential in the binary space is shown not to saturate the best binary bound, except asymptotically and for a special case. The alternative technique of transforming the components of a continuous vector is studied, showing that, asymptotically and for the same special case, saturation of both bounds can occur.

Nederlandstalige samenvatting

In dit werk wordt het leren zonder begeleiding van binaire vectoren bestudeerd aan de hand van technieken uit de statistische mechanica van systemen met wanorde. Meer bepaald worden leervoorbeelden gegeven die isotroop zijn verdeeld met uitzondering van één enkele voorkeursrichting. De bedoeling van het leerproces is deze richting in goede benadering terug te vinden. Het verschil met vorige studies is het extra gegeven dat deze richting binaire componenten bezit.

We bewijzen dat de Gibbs leerregel voor distributies zonder discontinuiteiten leidt tot een exponentieel snelle daling van de orientatiefout in de limiet van een groot aantal leervoorbeelden. In het geval van discontinuiteiten in de distributie kan men eerste orde faseovergangen verwachten naar een perfecte herkenning van de voorkeursorientatie. Indien het aantal leervoorbeelden klein is zijn de resultaten identiek aan deze van een voorkeursrichting met continue componenten. In het bijzonder vindt men leervertraging gevolgd door een tweede orde faseovergang wanneer de gemiddelde projectie van de voorbeelden op de voorkeursrichting nul is.

Op basis van de Bayes regel kan men bewijzen dat het massacentrum van de studenten gevormd aan de hand van de Gibbs regel de beste gemiddelde schatting van de voorkeursrichting geeft. Verder vindt men de beste binaire vector door het teken te nemen van de componenten van dit massacentrum.

Teneinde de performantie van deze beste binaire vector te berekenen worden eerst de eigenschappen bestudeerd van het massacentrum van binaire vectoren, die aan eenvoudige geometrische condities gehoorzamen. We leiden het verrassend resultaat af dat dit massacentrum opnieuw een binaire vector is. Dit resultaat vervalt echter voor de Gibbs vectoren omdat de wanorde inherent in de keuze van de leervoorbeelden aanleiding geeft tot meer gecompliceerde beperkingen. Uit een expliciete replica berekening blijkt dat het massacentrum wel degelijk continue componenten bezit. Hieruit wordt de performantie van de beste binaire vector berekend. Deze ligt tussen die van de Gibbs vectoren en die van het massacentrum.

Aan de hand van een variationele berekening wordt een optimale potentiaal geconstrueerd. Onder een aantal technische voorwaarden (replica symmetrie, temperatuur 0 en overlap 1) kan men bewijzen dat de performantie van de beste binaire vector niet wordt bereikt, behalve in een triviaal geval. Gebruik makende van een alternatieve benadering, gebaseerd op een transformatie van de componenten van een continue vector kan men de optimale performantie wel bereiken in de limiet van een groot aantal leervoorbeelden.

Contents

1 Introduction **1**

1.1 Unsupervised Learning 1

1.1.1 The model 3

1.1.2 The prior distribution 5

1.2 Statistical Mechanics 6

1.2.1 The thermodynamic limit 7

1.2.2 The free energy 8

1.2.3 The entropy and the mutual information 12

1.3 Supervised learning 13

1.4 Overview 15

2 Gibbs learning **17**

2.1 Introduction 17

2.2 General results 18

2.2.1 The saddle point equations 18

2.2.2 Asymptotics 20

2.2.3 The entropy and the mutual information 22

2.3 A test case: supervised learning 24

2.4 A case study: the Gaussian scenario 25

2.4.1 The saddle point equations 26

2.4.2 The entropy 27

2.4.3 Asymptotics 27

2.4.4 The biased case 28

2.4.5 The unbiased case 30

3 Optimal learning: an upper bound **35**

3.1 Introduction 35

3.2 Watkin's reasoning 35

3.2.1 The best binary 38

3.3 Clipping 39

3.3.1 The original formulation 39

3.3.2 Extension to the best binary problem 41

4 The best binary: a geometric approach **45**
4.1 Introduction 45
4.2 The Maximum-Entropy formalism 46
4.2.1 Theory 47
4.2.2 Application to the best binary problem 48
4.3 The limit of infinite number of samples 50
4.4 Finite number of samples 52
4.4.1 Simulations 54
4.5 On the validity of the geometric approach 56

5 The center of mass of the Gibbs ensemble **59**
5.1 Introduction 59
5.2 $P(y)$ calculated via thermal averages 60
5.2.1 Simulations 62
5.3 The best binary vector 63
5.3.1 Asymptotics 67
5.4 The overlap between $\boldsymbol{J}_B$ and $\boldsymbol{J}_{bb}$ 68
5.5 Simulations 70
5.6 Discussion 71

6 Attempts to construct an optimal cost function **75**
6.1 Introduction 75
6.2 The *ansatz* of a unique minimum 76
6.2.1 Clipped Hebbian supervised learning 78
6.2.2 Minimal/Maximal Variance 80
6.2.3 Variationally optimal potentials 83
6.2.4 Discussion 88
6.3 Optimal potentials in the hypersphere 90
6.3.1 Transforming $\boldsymbol{J}^s_{opt}$ 92

7 Conclusions and perspectives **99**

A Notation **107**
A.1 General remarks 107
A.2 Functions and associated properties 108
A.3 Orthogonal transformation 109
A.4 Asymptotics and Series 110

B The replica calculation **111**
B.1 General results: the free energy . . . 111
B.2 The replica symmetric *ansatz* . . . 114
B.2.1 The energy term . . . 115
B.2.2 The Ising measure . . . 116
B.2.3 The spherical measure . . . 116
B.3 The entropy for a binary measure . . . 118

C Technical details of Gibbs learning **121**
C.1 Rewriting the free energy . . . 121
C.1.1 The saddle point equations . . . 122
C.2 The entropy . . . 125

D $P(y)$ without the ME formalism **127**
D.1 The natural rise of an extremum principle . . . 127
D.2 The RS *ansatz* . . . 128
D.3 Recovering the peaked distribution . . . 131

E Quenched moments **133**
E.1 General results . . . 133
E.2 The binary measure . . . 134
E.3 The spherical measure . . . 136
E.4 Proof of eq. 5.21 . . . 137

F The limit of zero temperature **139**
F.1 The free energy . . . 139
F.1.1 The energy term . . . 139
F.1.2 The Ising measure . . . 140
F.1.3 The spherical measure . . . 141
F.2 Proof that $\partial\mathcal{F}/\partial R \geq 0$ . . . 142

List of publications **143**

Bibliography **145**

Chapter 1

Introduction

1.1 Unsupervised Learning

The goal of unsupervised learning is to find structure in high-dimensional data. Loose as this definition may seem, it can be a good way to summarize the very many different techniques used in different scientific fields. Given N-dimensional vectors $\{\boldsymbol{\xi}^{\mu}\}$, $\mu = 1, \ldots, p$, how can one compress information about them and describe the set through some relevant numbers other than the vectors themselves? Clearly there is more than one answer to this question. Clustering, Principal Component Analysis and Independent Component Analysis are just a few examples of methods currently employed in applications of every sort.

The focus of the present work, however, is on the theoretical aspects of a particular model of unsupervised learning, where the "structure" of the data is represented simply by a symmetry breaking direction in the distribution of the vectors. Unsupervised learning is thus reduced to learning this high-dimensional preferential direction. This type of model has been studied before under the assumption that the symmetry breaking vector has real components [BM93, BM94, WN94, BG98]. The emphasis here will be instead on results concerning unsupervised learning of binary, or Ising, vectors. The discrete nature of the search space in this type of problem is the novelty of the research, which apart from that relies on well stablished techniques borrowed from the Physics of disordered systems. The calculations to be presented are closely related to those introduced by the seminal work of Gardner [Gar88, GD88], for the study of the Statistical Mechanics of Artificial Neural Networks.

Although the problem of learning a discrete preferential direction has been studied before for specific scenarios [Gyö90, SST92, WN94], the approach

here is to consider a more general formulation which encompasses specific models and sheds some light on the connections between them.

The problem of learning a binary direction shows two main differences with respect to its continuous counterpart: the number of states of the system is countable; and the absence of differentiability makes it extremely difficult to devise practical algorithms which lead to a desired configuration. These aspects can be regarded as complementary: the same discreteness that makes the number of available states infinitely smaller than for a continuous model, also generally prevents these very states to be found.

One of the possible connections with this work comes from the equivalence between supervised and unsupervised learning which was stablished in [RVdB96]. It allows to relate some results of this work with the important problem, in the Neural Network community, of supervised learning in the Ising perceptron [Gyö90, SST92, dM97]. As opposed to most of the situations which will be discussed, the classical model of supervised learning consists in finding an Ising vector which satisfies some *hard constraints* which are determined by the data. This problem belongs to the class of NP problems [PV88]. NP basically means that some sufficiently malicious data distributions exist, such that no algorithm is known to solve the problem in a time which scales with a polynomial of the system size. In other words, the problem of finding the Ising direction which satisfies the hard constraints *can* be extremely difficult for *some* data distributions, taking an exponential time to be solved with known algorithms (for instance, just by checking state by state). This kind of *worst case* analysis is very different from, and hard to compare with, the results to be shown in the following, since the methods of Statistical Physics render results for the *typical case* instead. The connection between the two approaches is therefore far from obvious, but nonetheless very interesting (if it exists at all!). The difference between *typical case* and *worst case* could be of great importance, since practical problems usually deal with the typical cases (the reader is referred to [Hay97] for an interesting discussion). But one should not be misled: the present work is not intended to directly address the difficult NP issue (which belongs rather to computational complexity theory), neither to focus on practical problems. It stands in between, addressing theoretical questions which may help understand either, neither or both of them. The theoretical study of unsupervised learning of binary vectors is a sufficiently difficult subject to be studied on its own.

Among the topics to be discussed, the following ones will be the most prominent:

1. Given the data and the knowledge of the discreteness of the preferential

direction, what is the best approximation one can provide?

2. How is this result affected in case the approximation is required to be discrete as well?

3. Is it possible to find this optimal approximation(s) by making use of some carefully devised cost function?

The first and second items can be cast in the framework of Bayesian inference. Addressing the problem of supervised perceptron learning, Opper and Haussler showed in [OH91] what the Bayes-optimal generalization is. Watkin [Wat93] then showed that the Bayes-optimal performance can be achieved by a properly constructed perceptron (i.e. a machine with the same architecture). These results were extended to an unsupervised scenario by Watkin and Nadal [WN94]. Watkin's reasoning is reproduced and extended here, in order to account for the extra constraints of a binary vector (a problem which is also briefly examined in [WRB93]).

The third item is inspired by Kinouchi and Caticha's scheme [KC92, KC96] to obtain variationally optimal potentials. Generalized to an unsupervised scenario by Van den Broeck and Reimann [VdBR96], this class of potentials has also been investigated by Buhot, Torres Moreno and Gordon [BTMG97, BG98, GB98]. Specifically related to the binary case, only the reference [dM97] could be found.

This work will rely on the so-called *inferential formalism* to be described in section 1.1.1, where the model is introduced. More specifically, Statistical Mechanics (SM) techniques borrowed from the study of disordered systems will be employed in that framework. These are described in section 1.2. The connection between unsupervised and supervised learning is explained in section 1.3, while section 1.4 contains an overview of the organization of the remaining chapters.

1.1.1 The model

The inferential formalism basically consists of building up a *model* of the data under study via an assignment of probabilities. In real world problems, the pathway leading to this assignment may require intuition, pre-processing of data, estimations and countless other possible techniques. However, these issues shall not be addressed here. The interested reader is referred to [Bis95] (for a more practical approach) and [Rei97] (for a more theoretical one). Here all probability functions are assumed to be known from the beginning. They are *parametrized* by other quantities which are unknown, and these are the goal of the search in the learning procedure, as will be seen below.

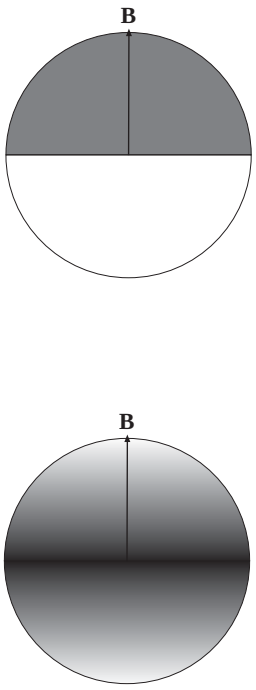

Figure 1.1: Pictorial representation of the N-dimensional hypersphere and the problem of unsupervised learning. The patterns can have either continuous (bottom figure) or discontinuous (top figure) distributions, as long as the only symmetry breaking direction of the problem is $\boldsymbol{B}$.

Note however that the functional form of the probabilities is assumed to be known.

The model under study in this work is as follows: the data D is a set of N-dimensional vectors (also called patterns, or examples) $\{\boldsymbol{\xi}^\mu\}_{\mu=1,\ldots,p}$,

$$D = \{\boldsymbol{\xi}^\mu\}, \qquad \mu = 1, \ldots, p \ , \tag{1.1}$$

which are generated by *independently* drawing from a *known* distribution $P(\boldsymbol{\xi}^\mu|\boldsymbol{B})$,

$$P(\{\boldsymbol{\xi}^\mu\}|\boldsymbol{B}) = \prod_{\mu=1}^{p} P(\boldsymbol{\xi}^\mu|\boldsymbol{B}) \ . \tag{1.2}$$

$\boldsymbol{B}$ is an N-dimensional *parameter vector*. It represents the *only* symmetry

breaking direction in the N-dimensional space where the data D is drawn from[1]. That means that the distribution of the *projections* of $\boldsymbol{\xi}$ along *all* $N-1$ directions orthogonal to $\boldsymbol{B}$ is the same. Only the projections on the $\boldsymbol{B}$ direction have a different distribution (see fig. 1.1), and this fact will be used to extract information from the data. A distribution $P(\boldsymbol{\xi}^\mu|\boldsymbol{B})$ that has these properties can always be written [RVdB96] in the form below:

$$P(\boldsymbol{\xi}^\mu|\boldsymbol{B}) = \frac{\delta\left(\boldsymbol{\xi}^\mu \cdot \boldsymbol{\xi}^\mu - N\right) \exp\left[-U\left(\frac{\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu}{\sqrt{N}}\right)\right]}{\int d\boldsymbol{\xi}'\, \delta\left(\boldsymbol{\xi}' \cdot \boldsymbol{\xi}' - N\right) \exp\left[-U\left(\frac{\boldsymbol{B}\cdot\boldsymbol{\xi}'}{\sqrt{N}}\right)\right]}\,, \tag{1.3}$$

where U is an *arbitrary known* function whose only requirement is that the denominator in eq. 1.3 is non-zero and non-divergent (i.e. the distribution must be properly normalized). The δ-distribution constrains the patterns to the hypersphere $\boldsymbol{\xi}^\mu \cdot \boldsymbol{\xi}^\mu = N$, which is a matter of choice. It is a reasonable choice though, just like the choice to deal with a normalized projection $\boldsymbol{B} \cdot \boldsymbol{\xi}^\mu/\sqrt{N}$, both of which will prove to be very convenient for the calculations in section 1.2.

In this model, *the goal of unsupervised learning is to find the best possible approximation for the vector* $\boldsymbol{B}$. The available information is the set D of p patterns, the knowledge of U and the knowledge of the *prior distribution* of $\boldsymbol{B}$.

1.1.2 The prior distribution

The prior distribution of $\boldsymbol{B}$ (hereafter also referred to simply as "prior") is denoted by $P(\boldsymbol{B})$. It characterizes previous, *a priori* knowledge (or ignorance) one has about vector $\boldsymbol{B}$. Once more, a warning is recommended at this point: in a problem with real data, choosing the most appropriate prior is not a simple task. This choice can be studied on its own as a separate subject, which will however not be the case in this work (the reader is again referred to [Bis95]). The interest here is understanding unsupervised learning from a theoretical point of view. Therefore priors are assumed to be known just as the other probability distributions.

A special subclass of priors will receive special attention, namely priors that represent *constraints*. In this scenario, all vectors $\boldsymbol{B}$ satisfying a given constraint are equiprobable, as shall be seen in the two cases addressed below.

Since the aim of learning is finding a *direction*, the size of $\boldsymbol{B}$ can be kept constant without loss of generality. A distribution that imposes this

[1]One could of course think of more difficult cases, with more than one symmetry breaking direction. Here only the simplest case will be studied.

condition is the *spherical prior* (probably the most studied prior in the literature [BM94, WN94, RVdB96, RVdBB96, VdBR96, BG98]), which only constrains $\boldsymbol{B}$ to a constant size without privileging any particular direction on the hypersphere:

$$P(\boldsymbol{B}) \overset{\text{Spherical}}{\equiv} P_s(\boldsymbol{B}) \sim \delta(\boldsymbol{B}\cdot\boldsymbol{B} - N)\ , \tag{1.4}$$

where the proportionality constant guarantees the proper normalization $\int d\boldsymbol{B}\ P_s(\boldsymbol{B}) = 1$. By using such a distribution[2], the underlying assumption is that one does not have any extra *a priori* information about the vector $\boldsymbol{B}$, apart from its size. All directions in the hypersphere are equally probable.

The main topic of this work, however, is the *binary (or Ising) prior* [WN94]:

$$P(\boldsymbol{B}) \overset{\text{Ising}}{\equiv} P_b(\boldsymbol{B}) = \prod_{j=1}^{N}\left[\frac{1}{2}\delta(B_j - 1) + \frac{1}{2}\delta(B_j + 1)\right]\ . \tag{1.5}$$

Some basic properties can immediately be read from eq. 1.5. First, the set of all vectors satisfying the Ising constraint is a subset of the set containing the vectors that satisfy the spherical constraint, since eq. 1.5 also implies $\boldsymbol{B}\cdot\boldsymbol{B} = N$. Second, the Ising constraint is "componentwise", in the sense that it enforces a constraint on *each* of the components of $\boldsymbol{B}$ (as opposed to eq. 1.4). Third, and perhaps the most important observation, this constraint is discrete. An Ising vector has components which can only take the values ± 1. Note that this is in strong contrast with the spherical constraint because it does imply preferential directions in the N-dimensional hypersphere. Binary vectors $\boldsymbol{B} \in \{-1, +1\}^N$ are said to lie on the corners of the N-dimensional *hypercube*.

1.2 Statistical Mechanics

How can one find a good approximation to vector $\boldsymbol{B}$, having as available information the data D and the knowledge of both U and $P(\boldsymbol{B})$? There are several possible procedures (or algorithms), depending on U and $P(\boldsymbol{B})$. They give as a result a *candidate vector* which approximates $\boldsymbol{B}$ to some extent, given the algorithm efficiency. Such a vector will be denoted by $\boldsymbol{J}$.

Equilibrium Statistical Mechanics (SM) is a very useful tool for analyzing the efficiency of a class of algorithms when the dimension of the problem

[2] In the limit of large N this is equivalent to choosing $P(B_j) = e^{-B_j^2/2}/\sqrt{2\pi}$, $j = 1, \ldots, N$.

grows very large. The thermodynamic limit (TL) $N \to \infty$ leads to major simplifications, allowing the calculation of averages and the invocation of thermodynamic postulates, the latter being expected to hold even though the system under study is not a physical object per se.

1.2.1 The thermodynamic limit

One of the simplifications obtained in the TL is the possibility of switching to the continuum limit during the calculations. Instead of tackling the multi-dimensional integrals over $P(\boldsymbol{\xi}|\boldsymbol{B})$, eq. 1.3 has, by construction, a symmetry which allows one to deal only with the probability density of the projection

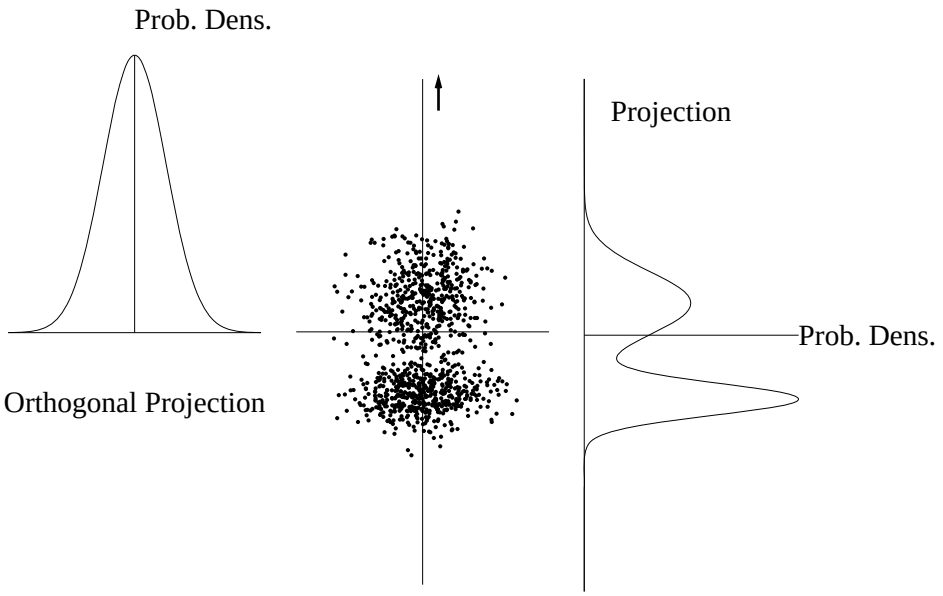

Figure 1.2: Two dimensional projections of 1000 patterns, for $N = 100$. The arrow is a unit vector in the direction of $\boldsymbol{B}$, and curves represent the theoretical prediction. In this picture $\mathcal{P}(b)$ was chosen to be a sum of Gaussians, the first with mean -2 and standard deviation 1/2, the second with mean 1 and standard deviation 1. Independently of this choice, the projection on the orthogonal direction is always normally distributed (eq. 1.8).

$$b \equiv \frac{\boldsymbol{B} \cdot \boldsymbol{\xi}}{\sqrt{N}} \ . \tag{1.6}$$

When $N \to \infty$ the Central Limit Theorem (CLT) holds and one obtains the distribution

$$\mathcal{P}(b) = \frac{\mathcal{N}}{\sqrt{2\pi}} \exp\left\{-\frac{b^2}{2} - U(b)\right\} , \tag{1.7}$$

where $\mathcal{N} = (2\pi)^{1/2}\left[\int dt\ \exp\{-t^2/2 - U(t)\}\right]^{-1}$ is a normalization constant. Likewise, and due to the chosen normalization of the patterns, the scaled projections of $\boldsymbol{\xi}$ on *any* direction $\boldsymbol{B}'$ orthogonal to $\boldsymbol{B}$ will have a Gaussian distribution with zero mean and unit variance.

$$t_2 \equiv \frac{\boldsymbol{B}' \cdot \boldsymbol{\xi}}{\sqrt{N}} , \qquad \boldsymbol{B}' \cdot \boldsymbol{B} = 0 \quad \Rightarrow \quad P(t_2) = \frac{e^{-t_2^2/2}}{\sqrt{2\pi}} \tag{1.8}$$

These properties are illustrated on fig. 1.2, where some computer generated patterns are projected onto the $(\boldsymbol{B}, \boldsymbol{B}')$ plane.

1.2.2 The free energy

Standard SM techniques of disordered systems can be applied to this problem if the candidate vector $\boldsymbol{J}$ is obtained by evolving in an energy landscape defined by a *cost function* $\mathcal{H}(\boldsymbol{J}, D)$. With the system kept at inverse temperature $\beta = 1/T$, one can think of the evolution of $\boldsymbol{J}$ as being governed by a Langevin-like stochastic process, or equivalently some Monte-Carlo-like spin flip dynamics, for Ising vectors. After a sufficiently long time, the system is supposed to reach equilibrium, which is characterized by the Boltzmann distribution

$$P(\boldsymbol{J}|D) = \frac{P(\boldsymbol{J})}{Z} \exp -\beta\mathcal{H}(\boldsymbol{J}, D) , \tag{1.9}$$

where Z is the partition function

$$Z\,(D) = \int d\boldsymbol{J}\ P(\boldsymbol{J})\ \exp -\beta\mathcal{H}\,(\boldsymbol{J}, D) \tag{1.10}$$

and $P(\boldsymbol{J})$ incorporates the constraints on $\boldsymbol{J}$, see below. This general procedure is called *off-line* or *batch* learning. It should be stressed that, from the Statistical Mechanics point of view, the equilibrium distribution 1.9 is the main assumption upon which all the calculations are based. Therefore the details of the microscopic dynamics which leads to this distribution are not discussed here. In this respect, the word "algorithm" used on page 6 should be understood in a loose sense (since no actual *procedure* is prescribed), its meaning being attached to a given cost function $\mathcal{H}$ and inverse temperature

β. One can thus regard the Statistical Mechanics approach from both sides: on one hand, it is limited because it does not necessarily provide a practical implementation of an algorithm; on the other hand, it is very powerful because its results are valid for *whichever* algorithm leads to a given Boltzmann distribution. Put in another way, learning is defined as sampling from the Bolztmann distribution, eq. 1.9.

$P(\boldsymbol{J})$ is a measure in $\boldsymbol{J}$ space. If one chooses $P(\boldsymbol{J})$ such that it enforces a constraint, then eq. 1.9 expresses the equilibrium probability density for a $\boldsymbol{J}$ that belongs to the subspace of vectors satisfying that constraint. Since $\boldsymbol{J}$ is to approximate $\boldsymbol{B}$, it would be wise to choose $P(\boldsymbol{J})$ such that this subspace contains $\boldsymbol{B}$ (for example, if $\boldsymbol{B}$ is known to be a binary vector, one could choose $\boldsymbol{J}$ to lie on the hypersphere – but not the other way round). Even though the consequences of not following this prescription are quite interesting (see for instance [VdBB93]), this shall not be pursued here. As a matter of fact, mostly the case $P(\boldsymbol{J}) = P(\boldsymbol{B})$ will be studied, for methodological reasons. The idea in this case is to understand the difference between searching on the (continuous differentiable) hypersphere and the (discrete exponentially many) corners of the hypercube, knowing in both cases that all $\boldsymbol{J}$'s are acceptable candidates for $\boldsymbol{B}$, in principle.

The space of possible cost functions $\mathcal{H}(\boldsymbol{J}, D)$ is clearly enormous, so one should restrict oneself to the study of a subclass thereof. Since eqs. 1.3 and 1.5 are invariant under permutation of the axes, the choice made here is to consider cost functions which respect the same symmetry. Therefore $\mathcal{H}$ should be a function of

$$\lambda_\mu \equiv \frac{\boldsymbol{J} \cdot \boldsymbol{\xi}^\mu}{\sqrt{N}}, \qquad \mu = 1, \ldots, p. \tag{1.11}$$

Moreover, the dependence of $\mathcal{H}$ on λ_μ should not depend on μ (since the order in which the data set is built up should be irrelevant in equilibrium). One chooses then cost functions of the form

$$\mathcal{H} = \sum_\mu^p V(\lambda_\mu) \,, \tag{1.12}$$

where V will be referred to as the *potential*.

It should be intuitively clear that the larger the dimension N of the problem, the harder it is to find a good approximation for $\boldsymbol{B}$. Therefore the number of examples p should scale with N if any learning is to occur, i.e. $p \to \infty$ in the thermodynamic limit with

$$\alpha \equiv \frac{p}{N} \tag{1.13}$$

constant (for a justification of this specific scaling, see appendix B, page 115).

Expression 1.10 is in general very hard to calculate due to the dependence on the randomness of the data. However, this can be averaged upon by noting that, in the TL, the free energy per component is a self-averaging quantity:

$$f = -\lim_{N\to\infty} \frac{1}{\beta N} \ln Z = -\lim_{N\to\infty} \frac{1}{\beta N} \langle \ln Z(D) \rangle_{D|\boldsymbol{B}} \ , \tag{1.14}$$

where $\langle(\cdots)\rangle_{D|\boldsymbol{B}}$ denotes an integration over $P(D|\boldsymbol{B})$ (see eq. 1.2). That means that even though one can only calculate an averaged free energy, its corresponding thermodynamic properties are identical to those of the non-averaged system.

This so-called "quenched" average can be performed by means of the replica trick, which makes use of the identity $\ln Z = \lim_{n\to 0}(Z^n - 1)/n$. Writing Z^n as the product of n replicated systems, $Z^n = \prod_a^n Z_a$, one is left with the problem of calculating $\langle \prod_a^n Z_a \rangle_{D|\boldsymbol{B}}$. The average over the disorder of the examples couples the different replicas and the limit $n \to 0$ is thereafter taken, assuming an analytic continuation. While the full calculation can be found in appendix B, here it suffices to say that f, which conveys all the thermodynamic information about the system, is written as a function of a set of self-averaging *order parameters*. Among all the configurations in $\boldsymbol{J}$ space, only a subset of them give a contribution to the free energy in the TL. This subset is characterized by the order parameters, which reduce the dimensionality of the problem. In the replica symmetric (RS) *ansatz* (see section B.2), they are

$$q_{ab} \equiv \frac{\boldsymbol{J}^a \cdot \boldsymbol{J}^b}{N} \overset{RS}{=} q \ , \tag{1.15}$$

the typical overlap between two different replicas and

$$R_a \equiv \frac{\boldsymbol{J}^a \cdot \boldsymbol{B}}{N} \overset{RS}{=} R \ , \tag{1.16}$$

which measures the proximity between $\boldsymbol{J}$ and $\boldsymbol{B}$. All samples of the distribution 1.9 must obey eqs. 1.15 and 1.16 in the TL. The role played by the function U, the inverse temperature β and the potential V is to determine the value of R and q (see below). This point will be discussed again in section 4.5.

It is also interesting to note that the emergence of these order parameters defines a "natural" measure for the efficiency of a given cost function: *in the following, the absolute value of R will be the main quantity of interest, accounting for the success of $\boldsymbol{J}$ in approximating $\boldsymbol{B}$.*

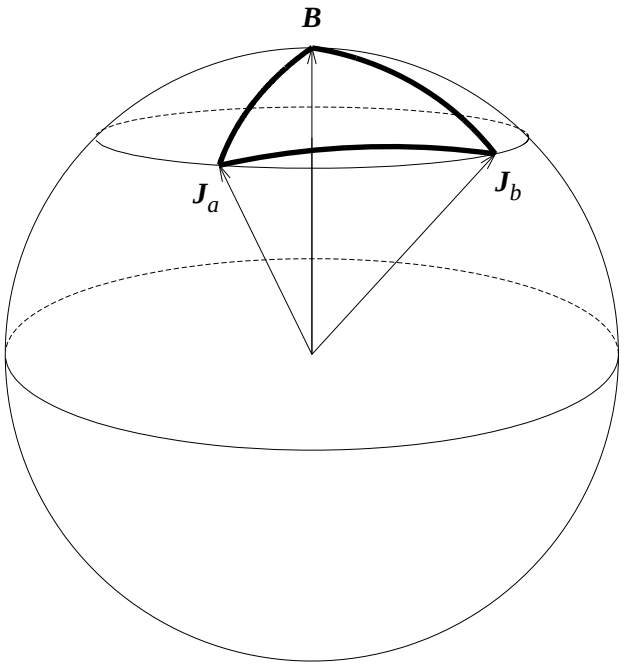

Figure 1.3: Pictorial representation of the constraints given by eqs. 1.15 and 1.16. In the thermodynamic limit *all* vectors $\boldsymbol{J}$ sampled from eq. 1.9 lie on a hypercone, defined by the two upper angles in the picture (which are both equal to $\arccos(R)$). The angle between the samples is also constrained (lower angle in the picture) to $\arccos(q)$.

For the binary constraint $P(\boldsymbol{J}) = P_b(\boldsymbol{J})$ and under a replica symmetric *ansatz*, the free energy is written as the extremum of a function $\hat{f}$, as can be seen in appendix B:

$$\begin{aligned} f &= \operatorname*{Extr}_{q,R,\hat{q},\hat{R}} \hat{f}(q,R,\hat{q},\hat{R};\beta,[U,V]) \\ &= \frac{1}{\beta} \operatorname*{Extr}_{R,q,\hat{R},\hat{q}} \left\{ \frac{1}{2}(1-q)\hat{q} + \hat{R}R - \int \mathcal{D}z \ \ln\cosh\left(z\sqrt{\hat{q}} + \hat{R}\right) \right. \\ &\quad -\alpha \int \mathcal{D}^* b \int \mathcal{D}t_2 \ \ln \int \frac{d\lambda}{\sqrt{2\pi(1-q)}} \exp\Big(-\beta V(\lambda) \\ &\quad \left. -\frac{(\lambda - t_2\sqrt{q-R^2} - bR)^2}{2(1-q)} \Big) \right\} , \end{aligned} \tag{1.17}$$

where $\mathcal{D}^* b = db\, \mathcal{P}(b) = \mathcal{D}b\, \mathcal{N} \exp -U(b)$ and $\mathcal{D}t_2 = dt_2\, (2\pi)^{-1/2} \exp(-t_2^2/2)$.

The extremum operation renders saddle point equations which determine the values of $\{R, q, \hat{R}, \hat{q}\}$. Given β, U and V, the physically relevant results (in a replica symmetric *ansatz*) are the functions $R(\alpha)$ and $q(\alpha)$ which extremize $\hat{f}$. No special notation will be systematically used in order to distinguish between the variables $\{R, q, \hat{R}, \hat{q}\}$ as such and their equilibrium values, but this should be clear from the context.

The free energy corresponding to a spherical constraint can be found in section B.2.3.

1.2.3 The entropy and the mutual information

In the case of a binary constraint, the $\boldsymbol{J}$-space is discrete and it is possible to define an entropy per component

$$s \equiv -\frac{\partial}{\partial T}\left[f - \frac{\ln 2}{\beta}\right] = \beta^2 \frac{\partial}{\partial \beta}\left[f - \frac{\ln 2}{\beta}\right] \tag{1.18}$$

which coincides with the physical entropy. The addition of an extra term $-(\ln 2)/\beta$ is justified because of the definition of the binary measure, recall eq. 1.5. Because of the factors $1/2$ in eq. 1.5, an integral $\int d\boldsymbol{J}\, P_b(\boldsymbol{J})(\cdots)$ differs from a sum $\sum_{\{J_j=\pm 1\}}(\cdots)$ by a factor 2^N. While the integral formulation allows one to discuss the present problem in the framework of Bayesian Statistics and Information Theory (thus requiring a normalized distribution $P_b(\boldsymbol{J})$, see below), the use of the summation is the standard in Statistical Mechanics calculation. The resulting difference is an additive constant $(\ln 2)/\beta$ in f, which is irrelevant as far as the equilibrium values of the order parameters are concerned. This, however, translates into an additive constant $\ln 2$ in s, which *is relevant*. As defined in eq. 1.18, s corresponds to the physical entropy per degree of freedom and should be positive, since it is a measure of the *number* (as opposed to the density) of vectors $\boldsymbol{J}$ which satisfy the equilibrium constraints. It is expected to decrease with increasing α and if it ever reaches zero, this means that there is a sub-exponential number of vectors compatible with the constraints.

The entropy is an interesting quantity to be studied because it can be used as a check of the stability of the RS *ansatz*. One could of course check the (local) stability of the RS solution by performing the AT calculation based on [dAT78]. In previous works, however, the positivity of the entropy has shown to be a stronger criterion than the AT condition (see [Gyö90] for an example of the supervised case and [GS90] for several variants of the capacity problem). As a matter of fact, in the capacity problem the condition of zero entropy was first used to bound the region where the RS solution is locally stable. This bound was then shown to exactly match the

RSB1 solution [KM89, ISB95]. This justifies the use of the positivity of the entropy as a criterion to check the stability of the RS solution, a procedure which will be adopted here and can be regarded as an educated guess. A general expression for s as a function of the order parameters can be found in section B.3.

Concepts from Information Theory [Sha48, CT91] are another possible tool to shed light on the problem of unsupervised learning. Defining the *pattern entropy* as

$$I_D \equiv -\int dD\, P(D) \ln P(D) \;, \tag{1.19}$$

where $dD = \prod_{\mu}^{p} d\boldsymbol{\xi}^{\mu}$, and the so-called *equivocation*

$$I_{D|\boldsymbol{B}} \equiv -\int d\boldsymbol{B}\, P(\boldsymbol{B}) \int dD\, P(D|\boldsymbol{B}) \ln P(D|\boldsymbol{B}) \;, \tag{1.20}$$

the *mutual information* between the patterns and the symmetry breaking direction $\boldsymbol{B}$ is given by

$$I(D;\boldsymbol{B}) \equiv I_D - I_{D|\boldsymbol{B}} \;. \tag{1.21}$$

The intensive quantity $i \equiv I(D;\boldsymbol{B})/N$ measures the mean amount of information (per component) about $\boldsymbol{B}$ which is conveyed by the data D. It is an absolute quantity in the sense that it is independent of the particular estimator $\boldsymbol{J}$ one might construct, having rather a functional dependence on the probability distributions which define the problem. In chapter 2 it will be shown how i is related to physical concepts such as the entropy and the free energy for Gibbs learning. Among the several exact bounds for i recently obtained [HN99], one of them will be applied here as another check of the stability of the RS *ansatz*.

1.3 Supervised learning

In [RVdB96], Reimann and Van den Broeck established an interesting connection between unsupervised learning and two neural network problems, namely supervised learning and the capacity problem in the perceptron. In the following, the connection with supervised learning will be explored.

In the above scenarios, the data is not only characterized by N-dimensional vectors $\{\boldsymbol{\xi}^{\mu}\}, \mu = 1, \ldots, p$, but also by the *dichotomic classification* of those vectors, here denoted by the labels $\tau^{\mu} \equiv \tau(\boldsymbol{\xi}^{\mu}) \in \{-1, +1\}$. This input-output relation is determined by a "teacher" vector in the supervised scenario and randomly in the capacity case. The equivalence between

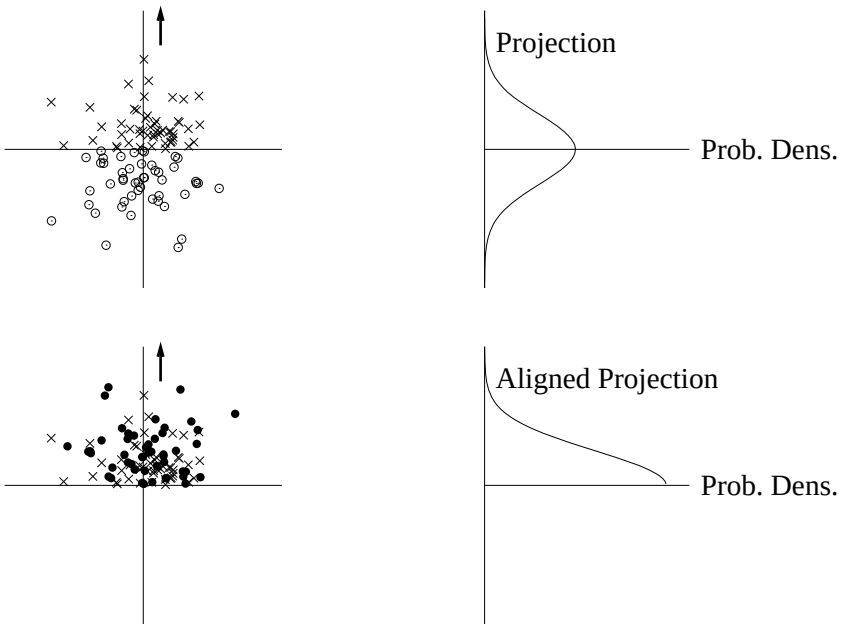

Figure 1.4: Arrow and projections like in Figure 1.2. In the upper left plot, patterns with a positive projection are classified with $\tau = 1$ (crosses) while those with a negative projection have $\tau = -1$ (white circles). The lower left plot shows the aligned patterns defined by eq. 1.23 (the different symbols were used only to clarify the alignment).

these problems and that of unsupervised learning can be established if the following conditions are satisfied:

$$\begin{aligned} P(\boldsymbol{\xi}^\mu|\boldsymbol{B}) &= \hat{P}(\boldsymbol{\xi}^\mu|\boldsymbol{B}) = \hat{P}\left(\frac{\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu}{\sqrt{N}}\right) \\ P(\tau^\mu|\boldsymbol{\xi}^\mu) &= \tilde{P}(\tau^\mu|\boldsymbol{\xi}^\mu) = \tilde{P}\left(\tau^\mu f_{odd}\left(\frac{\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu}{\sqrt{N}}\right)\right) \ , \end{aligned} \tag{1.22}$$

where f_{odd} is an odd function and $\hat{P}(\boldsymbol{\xi}^\mu|\boldsymbol{B})$ should be of the form of eq. 1.3. The first equation means that the input vectors can have at most one symmetry breaking direction[3], which on its turn must coincide with the "teacher"

[3]Note that this condition is satisfied by the vast majority of perceptron models studied in the literature, since input vectors are usually taken from a uniform distribution on the hypersphere. Of course, models where a different symmetry breaking direction structures

that determines (eventually under noisy conditions) the output (second equation).

If conditions 1.22 are met, then one can immediately calculate the probability distribution for the *aligned patterns* $\{\boldsymbol{\xi}'^{\mu}\}$

$$\boldsymbol{\xi}'^{\mu} \equiv \tau^{\mu}\boldsymbol{\xi}^{\mu} \ , \tag{1.23}$$

which is given by

$$P(\boldsymbol{\xi}'^{\mu}|\boldsymbol{B}) = \left[\hat{P}(\boldsymbol{\xi}'^{\mu}|\boldsymbol{B}) + \hat{P}(-\boldsymbol{\xi}'^{\mu}|\boldsymbol{B})\right] \tilde{P}\left(f_{odd}\left(\frac{\boldsymbol{B}\cdot\boldsymbol{\xi}'^{\mu}}{\sqrt{N}}\right)\right) \ . \tag{1.24}$$

Note that eq. 1.24 can always be written under the form of eq. 1.3, so that *the problem of supervised learning with patterns* $\{\boldsymbol{\xi}^{\mu}\}$ *is mapped into a problem of unsupervised learning with aligned patterns* $\{\boldsymbol{\xi}'^{\mu}\}$.

In order to give a more intuitive illustration of this equivalence, fig. 1.4 shows projections of patterns in the simplest case: the patterns are uniformly drawn from the hypersphere and the "teacher" vector $\boldsymbol{B}$ provides a noiseless classification $\tau = \text{sign}\left(\boldsymbol{B}\cdot\boldsymbol{\xi}/\sqrt{N}\right)$. Or, equivalently, $\hat{P}(\boldsymbol{\xi}^{\mu}|\boldsymbol{B}) \sim \delta(\boldsymbol{\xi}^{\mu}\cdot\boldsymbol{\xi}^{\mu} - N)$, $f_{odd}(x) = x$ and $\tilde{P}(x) = \Theta(x)$, where $\Theta(x)$ is the usual Heaviside function (eq. A.1). In this case the probability distribution of the aligned patterns has

$$\begin{aligned} U(b) &\begin{cases} = 0, & b \geq 0 \\ \to \infty, & b < 0 \ , \end{cases} \\ \Rightarrow \quad \mathcal{P}(b) &= 2\Theta(b)\frac{e^{-b^2/2}}{\sqrt{2\pi}} \ . \end{aligned} \tag{1.25}$$

This kind of discontinuity will prove to have interesting consequences (see chapter 2).

1.4 Overview

The organization of the next chapters is the following: in chapter 2 a particular (but very important) case of unsupervised learning is described: Gibbs learning. General asymptotic results are presented, as well as non-asymptotic

the input space, cannot be mapped to this unsupervised scenario. This is the case, for instance, of refs. [MBS95] and [MSBR96].

results for specific models. The technical details of the calculations regarding Gibbs learning are found in appendix C.

Chapter 3 describes what is called in the literature Optimal Learning [Wat93, WN94]. The theory is used to obtain an upper bound on the performance of any estimate $\boldsymbol{J}$ and shows how the center of mass of the Gibbs ensemble is optimal in the Bayesian sense. The problem of bounding the performance under the restriction that $\boldsymbol{J}$ is an Ising vector, is also addressed by a simple extension of the theory. The *best binary* vector is shown to be the clipped center of mass of the Gibbs ensemble, and its properties are studied in the following chapters.

Chapter 4 presents a study on the geometry of the center of mass of Ising vectors. Using the Maximum-Entropy formalism, it is shown that the center of mass of Ising vectors which obey some simple constraints, is again an Ising vector. The same result is derived in appendix D without making use of the Maximum-Entropy formalism.

In chapter 5 the properties of the center of mass of Ising vectors are again studied, but now taking into account the effects of the disorder of the examples (the calculations are in appendix E). The center of mass of the Gibbs ensemble is shown to be a continuous vector and the properties of its clipped counterpart (the best binary) are derived.

Results concerning the limit of zero temperature are shown in chapter 6. In particular, variational techniques are used as an attempt to construct a cost function which leads to the upper bound described in chapter 5. Alternatively, approximations to the bounds are also obtained by the technique of transforming the components of a continuous vector.

Chapter 7 contains the conclusions and perspectives for future work.

Appendix A contains a summary of the notation, formulas and expansions used in this thesis. It can be used as a quick reference in case of doubts.

Chapter 2

Gibbs learning

2.1 Introduction

The Bayes inversion formula can be applied to eqs. 1.2 and 1.3 in order to obtain the probability distribution of $\boldsymbol{B}$ given the data D:

$$P(\boldsymbol{B}|D) = \frac{P(D|\boldsymbol{B})\,P(\boldsymbol{B})}{P(D)}\ . \tag{2.1}$$

This so-called *posterior* distribution of $\boldsymbol{B}$ can be regarded as the knowledge about $\boldsymbol{B}$ which comes from the data D *and* the knowledge of all the probability distributions. Note that the initial ignorance about $\boldsymbol{B}$, expressed by the *prior* distribution $P(\boldsymbol{B})$, plays an important role in eq. 2.1. Replacing $\boldsymbol{B}$ with $\boldsymbol{J}$ in this formula gives the probability distribution that the guess $\boldsymbol{J}$ is the "true" direction $\boldsymbol{B}$, given the data. Gibbs learning is *defined* as sampling the vectors $\boldsymbol{J}$ from this distribution [WN94]. Explicitly inserting eqs. 1.2 and 1.3, one obtains the expression

$$\begin{aligned} P(\boldsymbol{J}|D) &= \frac{P(D|\boldsymbol{J})\,P(\boldsymbol{J})}{P(D)} \\ &= \frac{P(\boldsymbol{J})\prod_\mu^p \exp -U\left(\boldsymbol{J}\cdot\boldsymbol{\xi}^\mu/\sqrt{N}\right)}{\int P(\boldsymbol{J}')\,d\boldsymbol{J}'\prod_\mu^p \exp -U\left(\boldsymbol{J}'\cdot\boldsymbol{\xi}^\mu/\sqrt{N}\right)}\ . \end{aligned} \tag{2.2}$$

A comparison with eqs. 1.9 and 1.10 reveals that the thermodynamic properties of such a process can be described by the free energy 1.17 with

$$\begin{aligned} \beta &= 1 \\ V(\lambda) &= U(\lambda)\ . \end{aligned} \tag{2.3}$$

Gibbs learning can therefore be interpreted as sampling the vectors $\boldsymbol{J}$ according to their probability of being the "true" direction $\boldsymbol{B}$, given the data. This is why the prior information about $\boldsymbol{B}$ becomes clearly an important quantity: if $\boldsymbol{B}$ is known to be binary, then so must (and will) be all the Gibbsian candidates, hereafter referred to as $\boldsymbol{J}_G$.

2.2 General results

As explained in section 1.2.2, one should obtain the order parameters as functions of α, given $U = V$ and $\beta = 1$. The equilibrium values of the order parameters are obtained by calculating the saddle point of $\hat{f}_G = \hat{f}(q, R, \hat{q}, \hat{R}; \beta = 1; V = U)$. The explicit derivation of the saddle point equations is left to appendix C due to some lengthy technical questions which must be addressed. The main point concerning Gibbs learning is that it can be proven that

$$\begin{aligned} R_G \equiv \frac{\boldsymbol{J}_G^a \cdot \boldsymbol{B}}{N} &= q_G \equiv \frac{\boldsymbol{J}_G^a \cdot \boldsymbol{J}_G^b}{N} \\ \hat{R}_G &= \hat{q}_G \end{aligned} \tag{2.4}$$

is always a consistent *ansatz*, where the subscript $_G$ will be used to denote results concerning Gibbs learning. The equalities 2.4 had already been noted for both supervised and unsupervised learning in binary [Gyö90, WN94] and spherical [GT90, VdBR96] vectors, reflecting the symmetric role played by $\boldsymbol{J}_G$ and $\boldsymbol{B}$ in the calculation: note that, since the thermodynamic properties depend only on the order parameters, any vector $\boldsymbol{J}_G$ sampled from eq. 2.2 is as good a candidate as $\boldsymbol{B}$ itself, given αN examples.

2.2.1 The saddle point equations

The symmetry presented in eqs. 2.4 allows the reduction of the four original saddle point equations to only one (see appendix C for details):

$$R_G = F_B^2 \left(\mathcal{F} \left(\sqrt{R_G} \right) \right) \ , \tag{2.5}$$

where[1]

[1]The apparently unnecessary squares and square roots in eqs. 2.5-2.7 will be justified in chapter 6 (eq. 6.33), where they become convenient.

$$F_B(x) = \sqrt{\int \mathcal{D}z \, \tanh(zx + x^2)} \tag{2.6}$$

is a potential-independent function coming from the entropic term of the free energy, while

$$\mathcal{F}(R) = \sqrt{\alpha \int \mathcal{D}t \, \frac{Y^2(t;R)}{X(t;R)}} \tag{2.7}$$

is a function completely determined by U and α:

$$\begin{aligned} X(t;R) &= \mathcal{N} \int \mathcal{D}t' \, e^{-U(Rt+\sqrt{1-R^2}t')} \\ Y(t;R) &= \frac{1}{\sqrt{1-R^2}} \mathcal{N} \int \mathcal{D}t' \, t' e^{-U(Rt+\sqrt{1-R^2}t')} \, . \end{aligned} \tag{2.8}$$

The function $X(t;R)$ will be particularly important in the study of phase transitions, since the free energy at its minima is given by the expression (see appendix C for details)

$$\begin{aligned} f_G(R_G, \hat{R}_G) &= \frac{(1+R_G)\hat{R}_G}{2} - \int \mathcal{D}z \, \ln\cosh\left(z\sqrt{\hat{R}_G} + \hat{R}_G\right) \\ &\quad -\alpha \int \mathcal{D}t \, X\left(t;\sqrt{R_G}\right) \ln\left[\frac{X(t;\sqrt{R_G})}{\mathcal{N}}\right] \, , \end{aligned} \tag{2.9}$$

where R_G is the solution of eq. 2.5 while the equilibrium value of the conjugate parameter is given by

$$\hat{R}_G = \mathcal{F}^2\left(\sqrt{R_G}\right) \, . \tag{2.10}$$

It is then interesting to give a meaningful interpretation to $X(t;R)$. This can be done by first generalizing the result 1.8, which gives the distribution of an orthogonal projection $\boldsymbol{B'} \cdot \boldsymbol{\xi}/\sqrt{N}$. For any vector $\boldsymbol{J}$ satisfying $\boldsymbol{J} \cdot \boldsymbol{J} = N$ and $\boldsymbol{J} \cdot \boldsymbol{B} = NR$, the joint distribution of $b = \boldsymbol{B} \cdot \boldsymbol{\xi}/\sqrt{N}$ and $t \equiv \boldsymbol{J} \cdot \boldsymbol{\xi}/\sqrt{N}$ is given by

$$P(t,b) = \overbrace{\frac{\mathcal{N}}{\sqrt{2\pi}} \exp\left(-\frac{b^2}{2} - U(b)\right)}^{P(b) = \mathcal{P}(b)} \times \overbrace{\frac{1}{\sqrt{2\pi(1-R^2)}} \exp\left[\frac{-(t-Rb)^2}{2(1-R^2)}\right]}^{\equiv P(t|b)} \, . \tag{2.11}$$

Integrating over b, one obtains $P(t) = \int \mathcal{D}^* b\, P(t|b)$. Note that by setting $R = 0$ ($\boldsymbol{J}$ orthogonal to $\boldsymbol{B}$), one recovers eq. 1.8. By applying the change of variables $b = Rt + \sqrt{1-R^2}t'$ in eqs. 2.8, one obtains

$$X(t;R) = \frac{P(t)}{P_n(t)} \,, \tag{2.12}$$

where $P_n(t) \equiv \exp[-t^2/2]/\sqrt{2\pi}$ is the normal distribution. Therefore X measures the deviation of $P(t)$ from a normal distribution. The term with X in eq. 2.9 finally becomes

$$\int \mathcal{D}t\, X(t;R) \ln X(t;R) = \int dt\, P(t) \ln\left(\frac{P(t)}{P_n(t)}\right) \,, \tag{2.13}$$

which is the so-called[2] Kullback-Leibler distance [CT91] of $P(t)$ relative to the Gaussian $P_n(t)$. Since it is always a non-negative quantity, reaching zero only at $R = 0$, the above term gives a negative contribution to the free energy 2.9.

The problem is thus solved in principle for any distribution $\mathcal{P}(b)$. Given U, one has to calculate $\mathcal{F}$ (eq. 2.7) and then solve eq. 2.5. In case there is more than one solution, the one which minimizes the free energy 2.9 should be chosen as the thermodynamically stable one. In general one expects $R_G(\alpha)$ to be an increasing function, but in order to gain more insight about its behavior one can carry out asymptotic expansions.

2.2.2 Asymptotics

The limit $\alpha \to \infty$

In the limit of a large number of examples one can expand eqs. 2.5 and 2.10 as follows:

$$\begin{aligned} R_G &\overset{\hat{R}_G \to \infty}{\simeq} 1 - \sqrt{\frac{\pi}{2\hat{R}_G}} \exp\left(-\frac{\hat{R}_G}{2}\right) \left[1 + \mathcal{O}\left(\hat{R}_G^{-1}\right)\right] \\ \hat{R}_G &\overset{\alpha \to \infty}{\simeq} \alpha \left\langle (U')^2 \right\rangle_* \,, \end{aligned} \tag{2.14}$$

where $U' \equiv \frac{dU(b)}{db}$, $\langle(\cdots)\rangle_* \equiv \int \mathcal{D}^* b(\cdots)$ and $H(x) = \int_x^\infty \mathcal{D}t$. One then arrives at

[2] It is not really a distance, since it is not symmetric with respect to its two arguments, neither does it respect the triangular relation.

$$1 - R_G(\alpha) \overset{\alpha \to \infty}{\simeq} \sqrt{\frac{\pi}{2\alpha \left\langle (U')^2 \right\rangle_*}} \exp\left(\frac{-\alpha \left\langle (U')^2 \right\rangle_*}{2} \right) . \tag{2.15}$$

This expression should be compared with the result for a spherical prior [VdBR96], where the overlap approaches unity with a power law. Gibbs sampling in the binary space leads to an exponentially fast learning, a general result which is reminiscent of those of clipped Hebb learning in the supervised scenario [VdBB93, BS95] and generalizes the results obtained in [WN94].

The limit $R_G \to 0$

For R_G close to 0, one can expand

$$R_G \simeq \hat{R}_G - \hat{R}_G^2 + \mathcal{O}\left(\hat{R}_G^4\right) \tag{2.16}$$

$$\int \mathcal{D}t \frac{Y^2(t;\sqrt{R_G})}{X(t;\sqrt{R_G})} \simeq \begin{cases} \langle b \rangle_*^2 + \mathcal{O}(R_G), & \text{if } \langle b \rangle_* \neq 0 \\ \left(1 - \langle b^2 \rangle_*\right)^2 R_G + \mathcal{O}(R_G^2), & \text{if } \begin{cases} \langle b \rangle_* = 0 \\ \langle b^2 \rangle_* \neq 1 \, . \end{cases} \end{cases} \tag{2.17}$$

The two expansions in eq. 2.17 depend on the difficulty of learning direction $\boldsymbol{B}$. Two qualitatively different behaviors appear depending on whether the mean of the field b is zero or not. Assuming a smooth behavior[3] for $R_G(\alpha)$, one can solve eqs. 2.16 and 2.17 to obtain

$$\langle b \rangle_* \neq 0 \quad \Rightarrow \quad R_G \simeq \alpha \langle b \rangle_*^2 \tag{2.18}$$

$$\langle b \rangle_* = 0 \quad \Rightarrow \quad R_G \begin{cases} = 0, & \alpha \leq \alpha_G \\ \simeq C(\alpha - \alpha_G), & \alpha \geq \alpha_G \end{cases} \tag{2.19}$$

where $\alpha_G = \left(1 - \langle b^2 \rangle_*\right)^{-2}$ has the same value that has been obtained for spherical vectors [VdBR96] and C may depend on higher moments of b. If the distribution has a non-zero mean, then learning starts off as soon as α is non-zero. If it has a zero mean, then what has been called *retarded learning* occurs. The task is much harder in this case, and a non-zero overlap shows up only after a critical number of patterns which scales with the deviation from unity of the variance. One can in principle think of more and more difficult situations. For instance, if the mean is zero and the variance is one, then one needs to keep track of the next terms in expansions 2.16 and 2.17, and so on.

[3] First order transitions can appear and then the solution 2.19 is no longer valid. See section 2.4 for an example.

Discussion

The asymptotic results 2.15, 2.18 and 2.19 are valid in general, upon the following conditions: eq. 2.15 relies on the assumption that $\langle (U')^2 \rangle_*$ is finite, while 2.18 and 2.19 depend on the validity of the expansion around $R_G \simeq 0$. It is interesting to note, however, that these results also suggest what the consequences of the violation of these hypotheses would be. On mapping the problem of perceptron supervised learning onto the current framework of unsupervised learning [RVdB96], for instance, one is left with a function U which is discontinuous at the origin (at least in the cases of noiseless labels or output multiplicative noise – see section 1.3, particularly eq. 1.25). Eq. 2.15 then suggests that a first order phase transition to $R_G = 1$ should occur for some finite α, which is confirmed by Györgyi's results [Gyö90] (see section 2.3). For small α, on the other hand, nothing prevents different solutions from those of eqs. 2.18 and 2.19 to occur. For example, there can be first order phase transitions in which the overlap jumps from $R = 0$ to a finite value, at a value of α not necessarily equal to α_G (as has been observed for the spherical constraint [BG98]). In this case, a divergence of C can be used to bound the region where these phase transitions occur.

2.2.3 The entropy and the mutual information

While the explicit expression for the entropy in case of general β and V can be found on section B.3, Gibbs learning introduces some further simplifications. The expression for s can be rewritten (see section C.2) as

$$\begin{aligned} s_G(\alpha) &= \ln 2 - f(R_G, \hat{R}_G) + \alpha \langle U(b) \rangle_* \\ &= -\frac{(1+R_G)\hat{R}_G}{2} + \int \mathcal{D}z \, \ln 2 \cosh\left(z\sqrt{\hat{R}_G} + \hat{R}_G \right) \\ &\quad + \alpha \int \mathcal{D}t \, X\left(t; \sqrt{R_G}\right) \ln\left[\frac{X(t;\sqrt{R_G})}{\mathcal{N}} \right] + \alpha \langle U(b) \rangle_* \quad , \end{aligned} \tag{2.20}$$

where the order parameters should always be taken at their equilibrium value. On physical grounds this quantity should remain positive, and the results of section 2.2.2 can be useful to shed some light on its asymptotic behavior. In the limit $\alpha \to \infty$, for instance, the dominant term behaves like

$$s_G(\alpha) \overset{\alpha\to\infty}{\simeq} \sqrt{\frac{\pi\alpha \langle (U')^2 \rangle_*}{8}} \exp\left(\frac{-\alpha \langle (U')^2 \rangle_*}{2} \right) \to 0 \ , \tag{2.21}$$

whereas corrections are $\mathcal{O}(\hat{R}_G^{-1/2}\exp(-\hat{R}_G/2))$. This reasonable result reflects the shrinking of the solution space, until the perfect match $\boldsymbol{J} = \boldsymbol{B}$ is asymptotically reached.

In the poor performance regime, one can expand eq. 2.20 for $R_G \to 0$ and, assuming the expansion is valid, one gets

$$s_G(\alpha) \overset{R_G\to 0}{\simeq} \ln 2 + \alpha\left[\langle U(b)\rangle_* - \ln\mathcal{N}\right] + \mathcal{O}(R_G^2)\ . \tag{2.22}$$

It should be noted that, in the "difficult" cases (namely when $\langle U(b)\rangle_* = 0$, see eq. 2.19), the above expression is exact before retarded learning takes off, since then $R_G = 0$ identically.

The mutual information

The mutual information per degree of freedom i (eq. 1.21) for the family of models defined by eqs. 1.1, 1.2 and 1.3, is easily shown to be

$$i = -\alpha\langle U\rangle_* - \left\langle \ln\int d\boldsymbol{B}\, P(\boldsymbol{B})\, \exp -\sum_\mu^p U\left(\frac{\boldsymbol{B}\cdot\xi^\mu}{\sqrt{N}}\right)\right\rangle_{D|\boldsymbol{B}}\ , \tag{2.23}$$

a result which is valid for any value of N. A comparison with eqs. 1.10 and 2.20 shows that in the TL it can be simply rewritten in terms of the entropy (or the free energy) of Gibbs learning:

$$\begin{aligned} i &= -\alpha\langle U(b)\rangle_* + f_G \\ &= \ln 2 - s_G\ . \end{aligned} \tag{2.24}$$

Relying on the idea that the information of p examples cannot be larger than p times the information of one example, Herschkowitz and Nadal provide in [HN99] a proof of the upper bound $I_{D|\boldsymbol{B}} \le -p\left[\langle U\rangle_* - \ln\mathcal{N}\right]$. Making use of eq. 2.24, this translates into the equivalent bounds

$$\begin{aligned} f_G - \alpha\ln\mathcal{N} &\le 0 \\ s_G &\ge \ln 2 + \alpha\left[\langle U(b)\rangle_* - \ln\mathcal{N}\right]\ . \end{aligned} \tag{2.25}$$

As a general remark, one should note that the link between i and f_G provides an upper bound which gives meaning to the absolute value of the free energy. The entropy, on the other hand, is bounded from below and cannot decrease faster than linearly with α. A comparison between eqs. 2.25 and 2.22 shows that this information-theoretical bound is saturated *exactly* before the retarded learning phase transition occurs ($R_G = 0$).

2.3 A test case: supervised learning

Turning to specific models, this section recovers the results of perceptron noiseless supervised learning studied by Györgyi in ref. [Gyö90]. This model was chosen as a test case since it clarifies the (non-)validity of the asymptotic results obtained in section 2.2.2, apart of course from yielding very interesting results. The set of vectors $\{\boldsymbol{J}_G | P(\boldsymbol{J}_G|D) \neq 0\}$ has been called the *version space*, since any "student" $\boldsymbol{J}_G$ therein is a candidate to be the "teacher" $\boldsymbol{B}$.

Perceptron supervised learning from examples without noise corresponds to the example given in section 1.3, fig. 1.4. The pattern distribution of the mapped unsupervised problem is determined by the function U given by eq. 1.25. Noting that $X(t;R) = 2H(-tR/\sqrt{1-R^2})$, where $H(x) = \int_x^\infty \mathcal{D}t$, the equations for R_G and $\hat{R}_G$ are given by 2.5 and 2.10 with

$$\mathcal{F}^2\left(\sqrt{R_G}\right) = \frac{\alpha}{\pi\sqrt{1-R_G}} \int \mathcal{D}t \, \frac{e^{-t^2 R_G/2}}{H(t\sqrt{R_G})} \, , \tag{2.26}$$

while the entropy reads

$$\begin{aligned} s_G &= -\frac{(1+R_G)\hat{R}_G}{2} + \int \mathcal{D}z \, \ln 2\cosh\left(z\sqrt{\hat{R}_G} + \hat{R}_G\right) \\ &\quad +2\alpha \int \mathcal{D}t \, H\left(t\sqrt{R/(1-R)}\right) \ln H\left(t\sqrt{R/(1-R)}\right) \, . \end{aligned} \tag{2.27}$$

What can one expect from the asymptotics of such a model? In the small α regime, eqs. 2.18 and 2.22 predict a linear behavior for both R_G and s_G, since $\langle b \rangle_* = \sqrt{2/\pi}$. But the large α behavior certainly cannot be read from eqs. 2.15 and 2.21, since in this case $\langle (U')^2 \rangle_* \to \infty$ and the equations are not valid.

The answer comes from the numerical solution or eqs. 2.26, shown in fig. 2.1. First, one can immediately see from the saddle point eqs. 2.5 and 2.26 that $R_G = 1$ is *always* a solution for *any* α (definitely an idiosyncrasy of this model). But, for $\alpha < \alpha_{AT} = 1.493$, there are two other solutions. The criterion to determine which of the three solutions is thermodynamically stable, is straightforward: the free energy should be at its minimum. In this case, *the free energy and the entropy are the same*[4], apart from a change in sign and an additive constant ($\ln 2$). That means that among the three solutions,

[4] This is again an idiosyncrasy of this model (see eq. 2.20), due to the diverging potential which yields $\langle U(b) \rangle_* = 0$. In a way, this is what allows supervised Gibbs learning to be mapped to a zero temperature process, since $\beta \to \infty$ and $V(\lambda) = \Theta(-\lambda)$ together lead to the same definition of U, eq. 1.25.

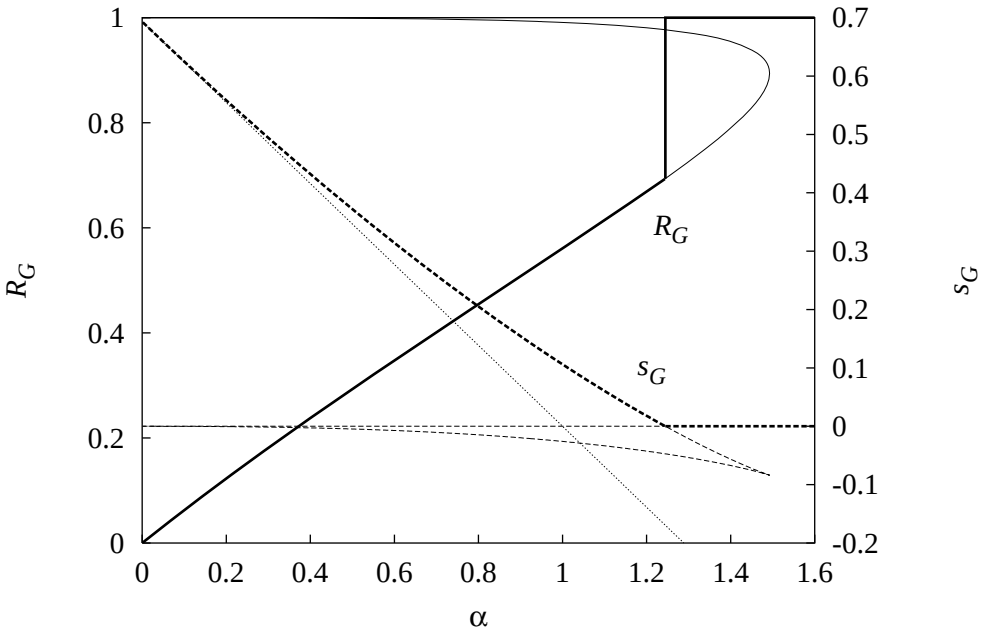

Figure 2.1: Numerical solution of eqs. 2.5 and 2.26 (solid lines, left axis) and the corresponding entropy (dashed lines, right axis). The thermodynamically stable solution is indicated with a thick line in both cases (see text for details). The linear bound of eq. 2.25 is depicted with the dotted line (right axis).

the one with maximal entropy (minimal free energy) is exponentially more probable. As can be seen on fig. 2.1, however, not all branches are physically acceptable. For $\alpha < \alpha_{GD} = 1.245$, the branch with positive entropy is the global minimum of the free energy, while $R_G = 1$ with $s_G = 0$ is metastable. For $\alpha \geq \alpha_{GD}$, $R_G = 1$ is the only acceptable solution and therefore the globally stable one, since the other branches have negative entropy. The system thus jumps discontinuously from an overlap $R \simeq 0.69$ to the perfect match $R = 1$ at $\alpha = \alpha_{GD}$. One should note on fig. 2.1 that the bound 2.25 for the linear decrease of the entropy is always satisfied.

2.4 A case study: the Gaussian scenario

The Gaussian scenario for unsupervised learning was introduced by Reimann *et. al.* [RVdBB96] as a model which allows easier calculations while preserving the richness of behavior presented by previous, more complicated pro-

posals [BM94, WN94]. It consists in studying a Gaussian distribution $\mathcal{P}(b)$, which amounts to a quadratic form in $U(b)$ (see eq. 1.7):

$$\begin{aligned} U(b) &= \frac{\hat{a}}{2}b^2 - \hat{b}b \\ \hat{a} &> -1 \\ \hat{b} &\geq 0 \ . \end{aligned} \tag{2.28}$$

The restriction on the parameter $\hat{a}$ assures the concavity of $\mathcal{P}(b)$, while the positivity of $\hat{b}$ is just a matter of choice, since it determines the sign of the bias of the patterns along the true direction $\boldsymbol{B}$. The original model still proposes a quadratic form for the potential $V(\lambda)$, but this will not be addressed here. According to eq. 2.3, Gibbs learning fixes $V = U$ and $\beta = 1$ (thus rendering indeed a quadratic V, but not with an extra set of parameters). Even though the calculations are more difficult in the case of binary vectors, the Gaussian scenario is nonetheless worth studying due to the wealth of behavior it generates.

Before proceeding, a short remark about notation should be made. It is very useful to express the results in terms of the mean $\langle b \rangle_* = \hat{b}/(1+\hat{a})$ and the variance $\langle b^2 \rangle_* - \langle b \rangle_*^2 = 1/(1+\hat{a})$, therefore one defines

$$\left.\begin{aligned} A &\equiv \tfrac{\hat{a}}{1+\hat{a}} = 1 - (\langle b^2 \rangle_* - \langle b \rangle_*^2) \\ B &\equiv \tfrac{\hat{b}}{1+\hat{a}} = \langle b \rangle_* \end{aligned}\right\} \Rightarrow \left\{\begin{aligned} \hat{a} &= \tfrac{A}{1-A} \\ \hat{b} &= \tfrac{B}{1-A} \ . \end{aligned}\right. \tag{2.29}$$

The use of the parameters $\hat{a}$, $\hat{b}$, A and B will be interchanged according to convenience.

2.4.1 The saddle point equations

The functions $X(t;R)$ and $Y(t;R)$ (see section 2.2.1) can be immediately calculated in the Gaussian scenario, yielding

$$\begin{aligned} X(t;R) &= \frac{1}{\sqrt{1-AR^2}} \exp\left[\frac{-B^2R^2}{2(1-AR^2)}\right] \exp\left[\frac{-AR^2t^2 + 2BRt}{2(1-AR^2)}\right] \\ Y(t;R) &= \left(\frac{B - ARt}{1-AR^2}\right) X(t;R) \ . \end{aligned} \tag{2.30}$$

The saddle point equations 2.5 and 2.10 are governed by

$$\mathcal{F}^2\left(\sqrt{R_G}\right) = \alpha\left[\frac{B^2 + AR_G(A-B^2)}{1-AR_G}\right] , \tag{2.31}$$

while the free energy at its minima (see eq. 2.9) reads

$$\begin{aligned} f_G &= \frac{(1+R_G)\hat{R}_G}{2} - \int \mathcal{D}z \, \ln\cosh\left(z\sqrt{\hat{R}_G} + \hat{R}_G\right) \\ &\quad + \frac{\alpha}{2}\left[\ln\left(\frac{1-AR_G}{1-A}\right) + R_G(A-B^2) - \frac{B^2}{1-A}\right] , \end{aligned} \tag{2.32}$$

since $\ln\mathcal{N} = -(\ln(1-A) + B^2/(1-A))/2$.

2.4.2 The entropy

According to eq. 2.20, and taking into account the result 2.32, one needs to calculate $\langle U(b)\rangle_*$ in order to write down an expression for the entropy. In the Gaussian scenario, this can be done exactly and the result is

$$\langle U(b)\rangle_* \overset{(2.3)}{=} \frac{A}{2} - \frac{B^2}{2}\left(\frac{2-A}{1-A}\right) , \tag{2.33}$$

so that

$$\begin{aligned} s_G &= -\frac{(1+R_G)\hat{R}_G}{2} + \int \mathcal{D}z \, \ln 2\cosh\left(z\sqrt{\hat{R}_G} + \hat{R}_G\right) \\ &\quad - \frac{\alpha}{2}\left[\ln\left(\frac{1-AR_G}{1-A}\right) + (B^2-A)(1-R_G)\right] \end{aligned} \tag{2.34}$$

2.4.3 Asymptotics

It is again useful to begin the analysis of the saddle point equations with the asymptotic behavior of the system.

When $\alpha \to \infty$, the overlap R_G tends to one exponentially, according to eq. 2.15. In the Gaussian scenario, this reads

$$1 - R_G \overset{\alpha\to\infty}{\simeq} \sqrt{\frac{\pi(1-A)}{2\alpha(B^2(1-A)+A^2)}} \exp\left[-\frac{\alpha}{2}\left(B^2 + \frac{A^2}{1-A}\right)\right] . \tag{2.35}$$

Note that unless the limit $A \to 1$, $A \to -\infty$ or $B \to \infty$ is taken, no phase transition to $R = 1$ ("perfect learning") is possible in this class of distributions.

In the vicinity of $R_G = 0$, the predictions for the Gaussian scenario are:

$$B \neq 0 \;\Rightarrow\; R_G \simeq \alpha B^2 \tag{2.36}$$

$$B = 0 \;\Rightarrow\; R_G \begin{cases} = 0, & \alpha \leq \alpha_G \\ \simeq C_G(\alpha - \alpha_G), & \alpha \geq \alpha_G \,, \end{cases} \tag{2.37}$$

where now

$$\begin{aligned} \alpha_G &\equiv \frac{1}{A^2} \\ C_G &\equiv \frac{A^2}{1-A} \,. \end{aligned} \tag{2.38}$$

One can see that in the so called *biased case* $B \neq 0$, it is much easier to learn. The *unbiased case* $B = 0$ presents much more difficulties for information about vector $\boldsymbol{B}$ to be extracted. In this case, *retarded learning* occurs, meaning that a non-zero macroscopic overlap R_G will be obtained only after a critical number of examples $\alpha_G N$ is presented. Since the average value of b is zero, information about its distribution comes via higher moments. While the projections[5] of $\boldsymbol{\xi}$ on the directions orthogonal to $\boldsymbol{B}$ have all zero mean and unit variance, eq. 2.37 assumes that $\langle b^2 \rangle_* - \langle b \rangle_*^2 \neq 1$. Note that α_G depends precisely on the difference between the variance of b and 1, in agreement with the discussions of page 21. These two qualitatively different situations will be studied away from the asymptotic regime in the next sections.

2.4.4 The biased case

The first case to be studied is $A = 0$ with $B \neq 0$. The non-zero bias makes sure learning starts off as soon as $\alpha \geq 0$, while $A = 0$ eliminates the dependence of $\hat{R}_G$ on R_G (see eqs. 2.10 and 2.31), simplifying immensely the solution of the saddle point equations. The behavior of R_G is seen to be completely determined by the rescaled variable

$$\alpha' \equiv \alpha B^2 \,, \tag{2.39}$$

namely $R_G = F_B^2\left(\sqrt{\alpha'}\right)$. In order to plot R_G as a function of α' one just has to perform the numerical integration on the r.h.s. of eq. 2.6. This function can be seen in fig. 2.2. It shows a linear increase for small α' and an exponential

[5] That is, the normalized projections $\boldsymbol{B}' \cdot \boldsymbol{\xi}/\sqrt{N}$, where $\boldsymbol{B}' \cdot \boldsymbol{B} = 0$.

behavior for $\alpha' \to \infty$. The entropy saturates the linear bound 2.25 only in the limit $\alpha' \to 0$, approaching zero exponentially when $\alpha' \to \infty$ but remaining otherwise strictly positive.

Note that $A = 0$ means that b has unit variance. The patterns can thus be pictured as being distributed in an N-dimensional *spherically symmetric* cloud, whose displacement B from the origin conveys the information about $\boldsymbol{B}$.

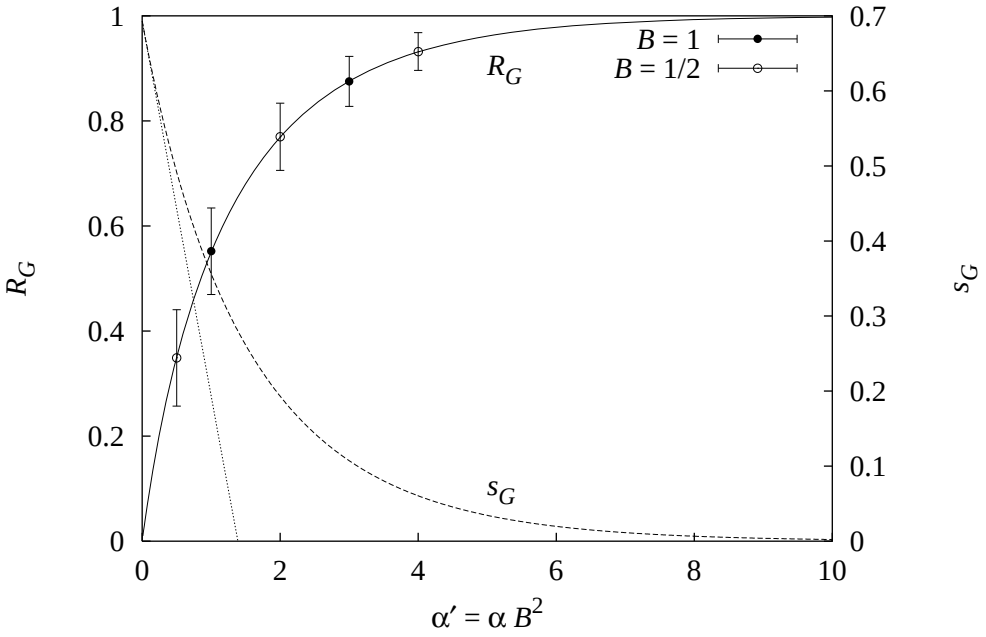

Figure 2.2: Overlap R_G (left axis) as a function of α' for $A = 0$ (see eq. 2.39): theory (solid line) and simulations with $N = 100$ (symbols; error bars represent one standard deviation, see text for details). The dashed line represents the entropy (right axis) while the dotted line shows the linear bound (eq. 2.25, right axis).

Simulations

Binary disordered systems are known to be very hard to simulate due to the existence of very many local minima. A noisy dynamics with unity temperature and general cost function U will typically get stuck in one of these minima, preventing a proper sampling of the posterior distribution 2.2 in an acceptable time. The Gaussian scenario with $A = 0$ provides an exception to

this rule, allowing Gibbs learning to be very easily implemented with a simple Metropolis algorithm: a component J_j is selected at random and flipped. The flip is accepted if it decreases the energy $\mathcal{H}$; otherwise it is accepted with a probability $\exp -\beta\Delta\mathcal{H}$. After repeating this procedure N times, one counts 1 Monte Carlo step per site (MCS/site). Since $A = 0$ implies a linear function $U(\lambda)$, the changes in energy can be very quickly calculated because it depends only on $\boldsymbol{J} \cdot \sum_\mu \boldsymbol{\xi}^\mu$.

Fig. 2.2 shows the results for simulations with $N = 100$ (the smallest system size simulated) and two values of B, checking the relevance of the variable α'. For each pattern set D, 10 samples of R_G and q_G were measured, after a random initialization of the system and a warming up of the dynamics (see further details below). The whole procedure was repeated for 1000 pattern sets and the standard deviation was calculated over these 10000 samples.

The measurement of q_G during simulations is an interesting tool which allows one to check both the property $q_G = R_G$ and the correctness of the RS *ansatz*. Fig. 2.3 focuses on the second simulated point of fig. 2.2 ($\alpha' = 1$). It shows histograms for both R_G and q_G (measured only between pairs of consecutive samples) which are virtually indistinguishable on the scale of the figure, with a mean value in excellent agreement with the theoretical prediction. The upper inset gives a glimpse of the Metropolis dynamics: the system is initialized randomly at $t = 0$ and evolves up to $t = 50$ MCS/site, where a different pattern set is drawn. The system reaches thermal equilibrium after $\mathcal{O}(10)$ MCS/site, which motivated the choice of safely waiting 100 MCS/site during the simulations before any measurement was made. The system was reinitialized after every measurement of the overlaps. Note that some pattern sets yield time-averaged values of R_G which deviate from theory (notably the first one for $N = 100$ and the second one for $N = 1000$) and only a second average over the pattern sets gives the right results. This reflects the property of self-averaging, which only holds in the thermodynamic limit (note that deviations from theory are smaller for larger N). The lower inset shows the typical scaling with $1/\sqrt{N}$ of the width of the distribution of overlaps.

2.4.5 The unbiased case

When $B = 0$ retarded learning is expected to occur, according to eq. 2.37. Fig. 2.4 shows the solution of the R_G saddle point equation for two values of A, namely 0.6 and -0.6. In both cases, a second order phase transition occurs at the critical value α_G predicted by eq. 2.37 and the entropy saturates exactly the linear bound of eq. 2.25 before the phase transition. Based on the relation between s_G and i (see section 2.2.3), the retarded learning phase

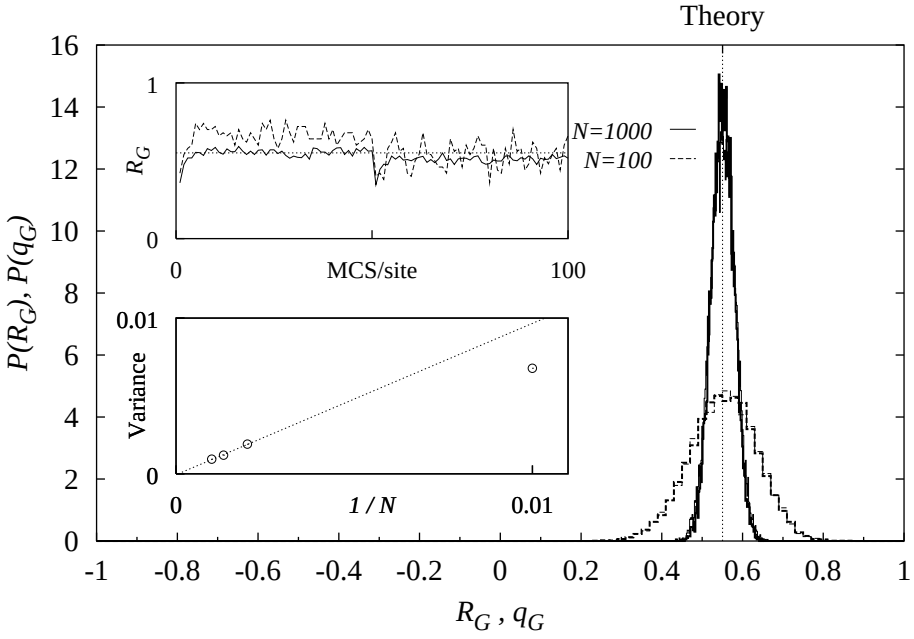

Figure 2.3: Gaussian scenario with $A = 0$, $B = 1$ and $\alpha = 1$. Histograms of q_G (thick lines) and R_G (thin lines) for $N = 100$ (dashed) and $N = 1000$ (solid); the vertical line is the theoretical prediction. The upper inset shows the Metropolis R_G-dynamics for two pattern sets (same legend, see text for details). The lower inset presents the variance of the distribution for R_G (symbols) as a function of $1/N$ for $N = 100, 500, 750$ and 1000; the dotted line is a linear fit of the three leftmost points.

transition can be interpreted as follows: for $\alpha \leq \alpha_G$, the system extracts maximal information from each pattern but is nonetheless unable to obtain a non-zero alignment R_G with the preferential direction $\boldsymbol{B}$. Only at $\alpha = \alpha_G$ does R_G depart from zero, which on its turn immediately gives an increasing degree of redundancy (measured by the deviation of s_G from its linear bound) to the patterns coming thereafter. Fig. 2.4 also shows the effect of a small bias $B = 0.1$ in an otherwise symmetric distribution: for sufficiently large α (say, $\alpha \gg \alpha_G$), the curve comes very close to that of $B = 0$ (as would be expected from continuity of eq. 2.35, for instance); but for small α the bias qualitatively changes the behavior of $R_G(\alpha)$, since the second order phase transition disappears due to the broken symmetry.

It is interesting to note in fig. 2.4 that even though the phase transition for $A = -0.6$ and $A = 0.6$ occurs at the same critical value, the overlap

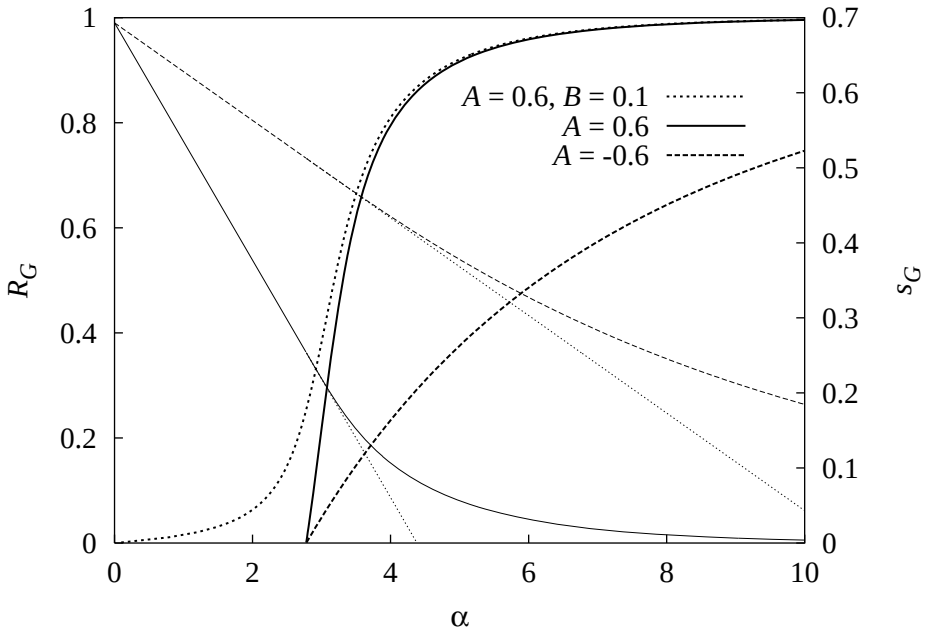

Figure 2.4: R_G (thick lines, left axis) and s_G (thin lines, right axis) as functions of α for $A = \pm 0.6$ and $B = 0$. The thin dotted lines correspond to the linear bound 2.25 which is saturated up to the second order phase transition. A small bias $B = 0.1$ (with $A = 0.6$) breaks the symmetry and destroys the second order phase transition (thick dotted line, left axis, entropy not shown).

increases much slower in the former case than in the latter. Recalling the definition of A (eq. 2.29), this means that prolate Gaussian distributions (N-dimensional "cigars" [BM94]) convey less information about the preferential direction than oblate distributions (N-dimensional "pancakes" [BM94]) for the same absolute value of A.

However, the second order phase transition at $\alpha_G = A^{-2}$ is not the only interesting phenomenon for this model. First order phase transitions are also possible, depending on the value of A. They can occur in two situations: either for $\alpha > \alpha_G$, in which case two consecutive phase transitions take place during learning (a second-order one followed by a first-order one), or $\alpha < \alpha_G$, in which case the asymptotic result 2.37 is overridden. The first order phase transition appears when there is more than one solution to the saddle point equation. In such cases the solution with minimal free energy has maximal probability of occurrence, being thus the thermodynamically stable state.

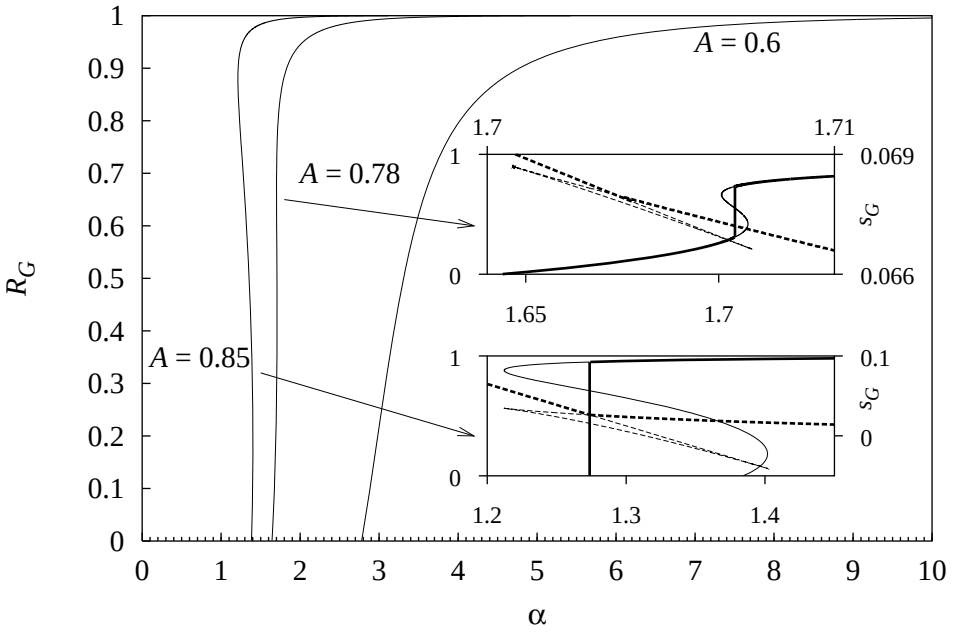

Figure 2.5: Solutions R_G of the saddle point equations 2.5 and 2.31 as a function of α for $B = 0$ and three values of A. s_G is plotted with dashed lines and the thermodynamically stable solutions are plotted with thick lines. $A = 0.78$ (upper inset): R_G (left axis) vs. α (bottom axis) and s_G (right axis) vs. α (top axis — note the different α scale, which zooms in the first order phase transition); $A = 0.85$ (lower inset): R_G (left axis) and s_G (right axis) vs. α.

An overview of this phenomenology is presented in fig. 2.5. It shows the three typical behaviors that occur for $B = 0$. For comparison, the case $A = 0.6$ previously plotted in fig. 2.4 is shown again, as an example of a parameter region where there is only a second order phase transition (at $\alpha_G = 2.778$). For $A = 0.78$ the second order phase transition at $\alpha_G = 1.643$ is followed by a first order phase transition at $\alpha_G^{(f)} = 1.704$ (upper inset, lower axis), while for $A = 0.85$ only a first order phase transition takes place at $\alpha = 1.274$, overriding the second order phase transition at $\alpha = 1.384$ (lower inset) which was predicted on asymptotics and smoothness grounds. Note that none of these first order phase transitions can be predicted by the general results of section 2.2 nor the asymptotic expansions of section 2.2.2 or 2.4.3. It is also interesting to observe that some solutions of the saddle point equation may violate the linear bound 2.25 and/or the positivity of

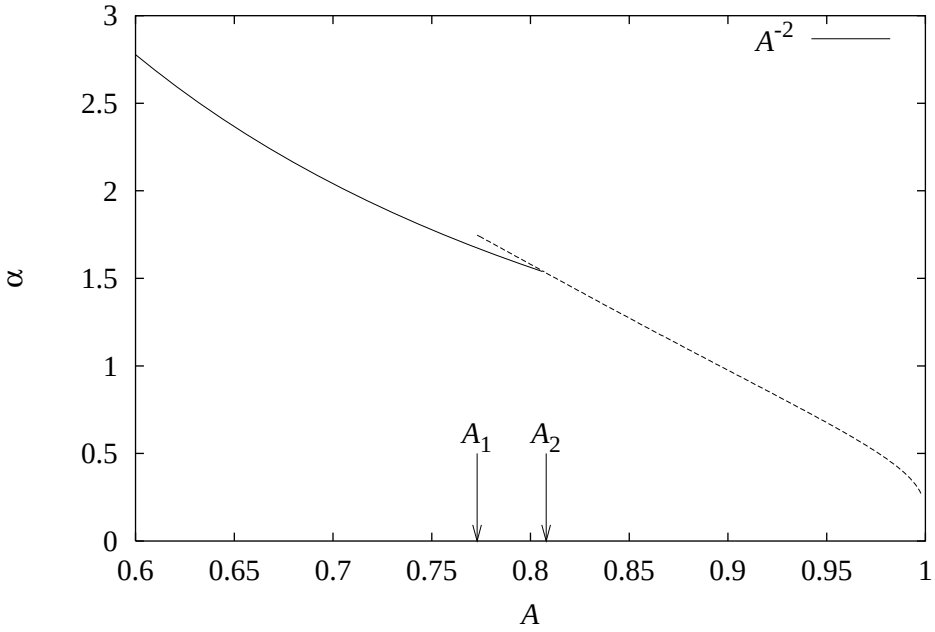

Figure 2.6: Phase diagram for the unbiased case ($B = 0$). Second order phase transitions occur at $\alpha = \alpha_G = A^{-2}$ (solid line), while first order phase transitions take place at $\alpha = \alpha_G^{(f)}$ (dashed line). See text for details.

the entropy (notably $A = 0.85$ in fig. 2.5). However, it turns out that these branches are always thermodynamically unstable, while the stable solutions satisfy all the requirements.

The whole phase diagram for $B = 0$ is shown in fig. 2.6. For $A > A_1 \simeq 0.773$, a first order phase transition takes place at the line $\alpha_G^{(f)}(A)$, after the second order one has already occurred. For increasing A, $\alpha_G^{(f)}(A)$ gets closer and closer to $\alpha_G(A)$, until there is finally a collapse at $A = A_2 \simeq 0.808$. For larger values of A, only the first order phase transition occurs.

Chapter 3

Optimal learning: an upper bound

3.1 Introduction

At first sight, Gibbs learning studied in chapter 2 seems to be no more than just one particular choice of learning parameters. This section is intended to show that this is not so, by addressing the next question one could ask in the present framework of inferential learning: given the αN data vectors and the prior information about $\boldsymbol{B}$, what is the *best* performance R one could possibly attain with a vector $\boldsymbol{J}$? The answer to this question will be given in terms of what has been called *optimal learning*, where the (Bayesian) meaning of "optimality" will be further explained in section 3.2. As it will turn out, optimal learning is closely related to Gibbs learning, which justifies a posteriori the interest in studying the latter. The reasoning was originally conceived by Watkin [Wat93, WN94], see also [OH91]. He showed that the center of mass of the Gibbs ensemble is a vector with optimal performance. In section 3.3 the reasoning is extended in order to account for the properties of the best *binary* vector.

3.2 Watkin's reasoning

Given a set of data $D \equiv \{\boldsymbol{\xi}^\mu\}$, $\mu = 1, \ldots, \alpha N$, an algorithm is any learning strategy which gives as result a vector $\boldsymbol{J}(D)$ which satisfies the constraint $\boldsymbol{J} \cdot \boldsymbol{J} = N$ (therefore not necessarily an Ising vector). There are several ways to account for the performance of such an algorithm. Watkin [Wat93, WN94] addressed this problem by defining an arbitrary *quality* measure $\mathcal{Q}(\boldsymbol{J}, \boldsymbol{B})$ which attains its maximum if $\boldsymbol{J} = \boldsymbol{B}$. This is formally a function of both

vectors, being thus an inaccessible quantity (since $\boldsymbol{B}$ is unknown). However, if one knows (or models) the form of the pattern distribution $P(D|\boldsymbol{B})$ and the prior distribution $P(\boldsymbol{B})$, it is possible to calculate the expected value of $\mathcal{Q}$ with respect to $\boldsymbol{B}$ given the data, namely

$$\tilde{\mathcal{Q}}(\boldsymbol{J}, D) \equiv \langle \mathcal{Q}(\boldsymbol{J}, \boldsymbol{B}) \rangle_{\boldsymbol{B}|D} = \int d\boldsymbol{B}\, P(\boldsymbol{B}|D) \mathcal{Q}(\boldsymbol{J}, \boldsymbol{B}) \; . \tag{3.1}$$

This way the dependence on the unknown $\boldsymbol{B}$ is averaged out, leaving $\tilde{\mathcal{Q}}$ as a *bona fide*, formally accessible measure of the algorithm performance. For a given data set, optimal learning is then *defined* as finding the vector $\boldsymbol{J}_B$ which maximizes $\tilde{\mathcal{Q}}$:

$$\boldsymbol{J}_B \equiv \operatorname*{Argmax}_{\boldsymbol{J}} \tilde{\mathcal{Q}}(\boldsymbol{J}, D) \; . \tag{3.2}$$

Because $\boldsymbol{J}_B$ maximizes $\mathcal{Q}$ averaged over the posterior distribution of $\boldsymbol{B}$, its performance is also referred to as "Bayes-optimal", or simply "Bayesian".

In general, optimal learning as posed above is not expected to be easy. Both the average over the posterior distribution which leads to $\tilde{\mathcal{Q}}$ and the high-dimensional maximization procedure which comes thereafter may be hard to compute. There are, however, favorable cases which allow some simplifications. Suppose, for instance, that one decides to measure the quality of learning with the overlap

$$r(\boldsymbol{J}, \boldsymbol{B}) = N^{-1} \boldsymbol{J} \cdot \boldsymbol{B} \; , \tag{3.3}$$

that is, $\mathcal{Q}(\boldsymbol{J}, \boldsymbol{B}) = r(\boldsymbol{J}, \boldsymbol{B})$. In this case the dependence on $\boldsymbol{J}$ can be factorized out of the average over the posterior:

$$\tilde{\mathcal{Q}}(\boldsymbol{J}, D) = N^{-1} \boldsymbol{J} \cdot \int d\boldsymbol{B}\, \boldsymbol{B}\, P(\boldsymbol{B}|D) \; . \tag{3.4}$$

The second argument of the dot product in eq. 3.4 is just the average of the Gibbsian vectors, that is, vectors sampled from the posterior distribution. Due to historical reasons in the development of the theory of supervised learning, this vector was called the "center of mass of the version space" (cf. page 24). Here and in the following it will be referred to as the center of mass of the Gibbs ensemble (or Gibbs space). Note that optimal learning is now trivial, at least formally: the vector $\boldsymbol{J}_B$ which maximizes $\tilde{\mathcal{Q}}(\boldsymbol{J}, D)$ is the center of mass of the Gibbs space itself,

$$\boldsymbol{J}_B = C \int d\boldsymbol{B}\, P(\boldsymbol{B}|D)\, \boldsymbol{B} \; , \tag{3.5}$$

where the constant C assures the required normalization[1]. Therefore

$$\tilde{\mathcal{Q}}(\boldsymbol{J}_B) - \tilde{\mathcal{Q}}(\boldsymbol{J}) \geq 0, \qquad \forall \boldsymbol{J}. \tag{3.6}$$

What still remains non-trivial is how to calculate the properties of such an optimal vector, since it depends on the data set. Again, an enormous simplification occurs when the system size becomes very large. In the thermodynamic limit, quantities like $r(\boldsymbol{J}, \boldsymbol{B})$ are expected to be self-averaging with respect to the disorder of the data. That means that *any* realization of the data D will yield the same value of $r(\boldsymbol{J}(D), \boldsymbol{B})$, for a given $\boldsymbol{B}$. Therefore, the equilibrium value of the order parameter $R = \langle r(\boldsymbol{J}, \boldsymbol{B}) \rangle_{D|\boldsymbol{B}}$, which one obtains from the Statistical Mechanics calculations, can be safely used as a substitute for $r(\boldsymbol{J}, \boldsymbol{B})$. Departing from eq. 3.6, one integrates over the data distribution to obtain [RVdB96]:

$$\begin{aligned} & \int dD\, P(D) \left[\tilde{\mathcal{Q}}(\boldsymbol{J}_B) - \tilde{\mathcal{Q}}(\boldsymbol{J}) \right] = \\ = & \int d\boldsymbol{B}\, P(\boldsymbol{B}) \int dD\, P(D|\boldsymbol{B}) \left[\frac{\boldsymbol{J}_B \cdot \boldsymbol{B}}{N} - \frac{\boldsymbol{J} \cdot \boldsymbol{B}}{N} \right] \\ \overset{N\to\infty}{=} & \int d\boldsymbol{B}\, P(\boldsymbol{B}) \left[R_B - R \right] \geq 0 \,, \end{aligned} \tag{3.7}$$

where $R_B \equiv \boldsymbol{J}_B \cdot \boldsymbol{B}/N$ and the integration over the patterns was bypassed because of the mentioned self-averaging. The Bayesian meaning of optimality becomes now clear: the overlap R_B cannot be beaten *on average*, where the average is with respect to the prior distribution of $\boldsymbol{B}$. It often turns out, however, that the dependence of R and R_B on $\boldsymbol{B}$ disappears. This is the case, for instance, of vectors uniformly distributed on the sphere or the corners of the hypercube in the thermodynamic limit. In these cases the integration left in eq. 3.7 can be performed immediately and the optimality is even stronger: R_B cannot be beaten *at all*, representing an upper bound on the performance of any vector.

[1] An intriguing issue arises if the symmetry $\boldsymbol{B} \to -\boldsymbol{B}$ is present. The integral in eq. 3.5 should vanish identically. In the thermodynamic limit, however, only part of the configuration space is available for a given realization of the disorder due to the spontaneous symmetry breaking (section 2.2), and the whole argument is saved. Formally, the result $\boldsymbol{J}_B = \boldsymbol{0}$ is similar to saying that the average magnetization of the Ising model in the ferromagnetic phase is zero, since it can take the values $\pm m$ ($\neq 0$) with equal probability. In the following it will be assumed that the symmetry (if it exists) is always broken, either spontaneously or by an infinitesimal field ϵb in $U(b)$ which can be taken to zero after the thermodynamic limit is taken.

The performance of $\boldsymbol{J}_B$ can be calculated via the "sample construction", as follows. Let $\boldsymbol{J}_G^a$ denote the a-th sample of Gibbs space, that is, a typical vector taken from the set of outcomes of Gibbs learning. Then a possible way to construct the optimal vector is by summing an infinitely large number of such samples, namely

$$\boldsymbol{J}_B = \lim_{n\to\infty} c_n \sum_a^n \boldsymbol{J}_G^a \tag{3.8}$$

$$c_n = \frac{1}{\sqrt{n + n(n-1)q_G}} , \tag{3.9}$$

where eq. 3.9 accounts for normalization. This construction allows one to easily verify that $R_B \equiv N^{-1}\boldsymbol{J}_B \cdot \boldsymbol{B}$ is related to $R_G = N^{-1}\boldsymbol{J}_G^a \cdot \boldsymbol{B}$ $(\forall a)$ in a very simple manner (remember that the mutual overlap $q_G = N^{-1}\boldsymbol{J}_G^a \cdot \boldsymbol{J}_G^b$ between different samples satisfies $q_G = R_G$):

$$R_B = \sqrt{R_G} . \tag{3.10}$$

This is the promised link between optimal and Gibbs learning. That means that all the results obtained in chapter 2 immediately translate into performance upper bounds.

3.2.1 The best binary

However, one should not forget that the focus here is on binary preferential directions. While Gibbs learning yields by definition binary candidate vectors, one expects $\boldsymbol{J}_B$ in eqs. 3.5 or 3.8 to have real components, in general. One would like to properly define an optimal *binary* vector whose performance could then be compared to that of e.g. a Gibbsian vector on fairer grounds. In other words, one is searching for a *binary* vector $\boldsymbol{J}$ which maximizes the r.h.s. of eq. 3.4. Fortunately, the answer to this question is simple (see for instance pages 519 and 530 of [WRB93]): *the binary vector* $\boldsymbol{J}_{bb}$ *which maximizes* $\tilde{\mathcal{Q}}$ *is obtained by clipping* $\boldsymbol{J}_B$, that is,

$$\boldsymbol{J}_{bb} = \mathrm{clip}(\boldsymbol{J}_B) , \tag{3.11}$$

where $\mathrm{clip}(\boldsymbol{V}) = \left\{ \tilde{\boldsymbol{V}} | \tilde{V}_j = \mathrm{sign}\,(V_j)\,, j = 1, \ldots, N \right\}$. In order to see this, notice that the quantity to be maximized (r.h.s. of eq. 3.4) is proportional to $\sum_j^N [\boldsymbol{J}_{bb}]_j [\boldsymbol{J}_B]_j$. Since $\boldsymbol{J}_{bb}$ is binary by definition, the sum will be maximized if all its terms are positive, which is accomplished by the clipping prescription 3.11. Note that the whole argument (eqs. 3.6 and 3.7) leading from

maximal quality to maximal overlap is unchanged, provided one restricts the space of allowed $\boldsymbol{J}$'s to the corners of the hypercube. Thus

$$\begin{aligned} & \tilde{\mathcal{Q}}(\boldsymbol{J}_{bb}) - \tilde{\mathcal{Q}}(\boldsymbol{J}) \geq 0 \ , \qquad \forall \boldsymbol{J} \in \{-1,+1\}^N \\ \overset{N\to\infty}{\Rightarrow} \quad & \int d\boldsymbol{B}\, P(\boldsymbol{B})\, [R_{bb} - R] \geq 0 \ , \end{aligned} \tag{3.12}$$

where R is the normalized overlap between $\boldsymbol{B}$ and any binary vector $\boldsymbol{J}$. Once more, if the values of the overlaps do not depend on the particular choice of $\boldsymbol{B}$, the bound is stronger and one obtains $R_{bb} \geq R$.

In the following, $\boldsymbol{J}_{bb}$ will be referred to as the *best binary* vector. Next section contains a general procedure for calculating the performance of clipped vectors, which will then be applied for this specific problem.

3.3 Clipping

3.3.1 The original formulation

The properties of clipped perceptrons have been studied before in some publications [VdBB93, GM93, BS95, SBVdB95]. The reasoning here will follow closely the approach by Schietse *et al.* [SBVdB95]. In that work, only a few hypotheses must be satisfied for some very general results to hold. Since some of these hypotheses are not satisfied in the problem under study here, the formalism will be extended accordingly.

Assuming that the preferential direction $\boldsymbol{B}$ has binary components ($B_j \in \{-1,+1\}, \forall j$) and a vector $\boldsymbol{J}$ with *continuous* components has been generated by evolving at temperature T in an energy landscape defined by some cost function $E(\boldsymbol{J}) = \sum_{\mu=1}^{\alpha N} V(N^{-1/2}\boldsymbol{J}\cdot\boldsymbol{\xi}^{\mu})$, Schietse *et al.* propose the following transformation $\boldsymbol{J} \to \tilde{\boldsymbol{J}}$:

$$\tilde{J}_j = \frac{\sqrt{N}\phi(J_j)}{\sqrt{\sum_i \phi^2(J_i)}} \ . \tag{3.13}$$

The only requirement on ϕ is that it must be an odd function. From this it immediately follows that the transformed overlap $\tilde{R} \equiv N^{-1}\tilde{\boldsymbol{J}}\cdot\boldsymbol{B}$ is given by

$$\tilde{R} = \frac{\sum_{i=1}^{N} \phi(J_i B_i)}{\sqrt{N\sum_{i=1}^{N} \phi^2(J_i B_i)}} \ . \tag{3.14}$$

The next step is to assume that $\tilde{R}$ is a self-averaging quantity. With this assumption they rewrite eq. 3.14 as

$$\tilde{R} = \frac{\int P(x)\,\phi(x)\,dx}{\left[\int P(x)\,\phi^2(x)\,dx\right]^{1/2}}\ , \tag{3.15}$$

where $x \equiv J_j B_j$ is supposed to have the same probability distribution regardless of the index j. This is a very reasonable assumption due to the permutation symmetry among the axes.

The last assumption concerns the geometric properties of vector $\boldsymbol{J}$. Schietse *et al.* assume that the algorithm which generated $\boldsymbol{J}$ does not make use of any extra prior information besides the spherical constraint $\boldsymbol{J}\cdot\boldsymbol{J} = N$. That is, despite the fact that one *knows* that $\boldsymbol{B}$ is a binary vector, this prior information is not taken into account for rendering $\boldsymbol{J}$. In this case one expects the outcomes of the algorithm to be uniformly distributed on the cone defined by $\boldsymbol{J}\cdot\boldsymbol{B} = NR$. $P(x)$ can then be calculated via the formula

$$P(x) = \frac{\int d\boldsymbol{J}\,\delta(J_1B_1 - x)\,\delta(\boldsymbol{J}\cdot\boldsymbol{J} - N)\,\delta(\boldsymbol{J}\cdot\boldsymbol{B} - NR)}{\int d\boldsymbol{J}\,\delta(\boldsymbol{J}\cdot\boldsymbol{J} - N)\,\delta(\boldsymbol{J}\cdot\boldsymbol{B} - NR)} \tag{3.16}$$

$$= \frac{1}{\sqrt{2\pi(1-R^2)}}\exp\left[\frac{-(x-R)^2}{2(1-R^2)}\right]\ , \tag{3.17}$$

a result which is valid *only* if $B_j \in \{-1,+1\}$.

Given eqs. 3.15 and 3.17, one can in principle calculate the function $\tilde{R}(R)$ for any odd function ϕ. Clipping corresponds to the particular case $\phi(x) = \mathrm{sign}(x)$, which leads to

$$R_{clip}(R) \overset{(3.17)}{=} \mathrm{erf}\left(\frac{R}{\sqrt{2(1-R^2)}}\right) = 1 - 2H\left(\frac{R}{\sqrt{1-R^2}}\right)\ . \tag{3.18}$$

Schietse *et al.* even proceed to find an optimal function ϕ^* which maximizes $\tilde{R}$ for given R. This nice result is obtained by developing eq. 3.15 using the fact that $\phi(x)$ is odd:

$$\tilde{R}^2 = \frac{\left(\int_0^\infty dx\,\phi(x)[P(x) - P(-x)]\right)^2}{\int_0^\infty dx\,\phi^2(x)[P(x) + P(-x)]}$$

$$= \langle\phi,\phi\rangle^{-1}\left\langle\phi, \frac{P(x)-P(-x)}{P(x)+P(-x)}\right\rangle^2\ , \tag{3.19}$$

where the internal product was conveniently defined as $\langle a(x), b(x)\rangle \equiv \int_0^\infty dx\,[P(x) + P(-x)]a(x)b(x)$. Now it is just a question of employing the

Schwarz inequality $|\langle a,b\rangle|^2 \leq \langle a,a\rangle\langle b,b\rangle$ to obtain the function ϕ^* which maximizes $\tilde{R}$:

$$\phi^*(x) = \frac{P(x)-P(-x)}{P(x)+P(-x)} \tag{3.20}$$

$$R^* \equiv \tilde{R}[\phi=\phi^*] = \left(\int_0^\infty dx\,\frac{[P(x)-P(-x)]^2}{P(x)+P(-x)}\right)^{1/2}. \tag{3.21}$$

From eqs. 3.13, 3.14, or 3.15, it is clear that $\tilde{R}$ does not change if ϕ is multiplied by a positive constant. Accordingly, ϕ^* is given by eq. 3.20 except for an irrelevant multiplicative constant. This result will be referred to in section 5.3.

Under the assumptions of Schietse *et al.*, eqs. 3.20 and 3.21 become

$$\phi^*(x) \overset{(3.17)}{=} \tanh\left(\frac{Rx}{1-R^2}\right) \tag{3.22}$$

$$R^* \overset{(3.17)}{=} \sqrt{\int \mathcal{D}t\,\tanh\left[\frac{R}{\sqrt{1-R^2}}\left(t+\frac{R}{\sqrt{1-R^2}}\right)\right]}, \tag{3.23}$$

a result which generalizes those obtained in [BS95] for the specific case of supervised Hebbian learning. Note that the *transformed vector $\tilde{\boldsymbol{J}}$ which best incorporates information*[2] *about the binary nature of $\boldsymbol{B}$, is itself not a binary vector*.

Fig. 3.1 shows the functions $R_{clip}(R)$ and $R^*(R)$ according to eqs. 3.18 and 3.23. For R sufficiently small, the difference between R^* and R is negligible, indicating that little can be done to improve the performance by transforming the components. For $R \to 1$, R^* approaches unity with an exponential behavior, $1-R^* \sim \exp[-R^2/(2(1-R^2))]$. Clipping, on the other hand, is a very drastic procedure, degrading the performance for $R < 0.77$. Above this value, clipping starts improving, and for $R \to 1$ R_{clip} also shows an exponential behavior. Note, however, that only at $R=1$ does clipping equal the optimal transformation. Asymptotically, $(1-R^*)/(1-R_{clip}) \to 1/2$.

3.3.2 Extension to the best binary problem

In order to obtain the properties of $\boldsymbol{J}_{bb} = \mathrm{clip}(\boldsymbol{J}_B)$, one would like simply to apply the result 3.18. That is, given the overlap $R_B = N^{-1}\boldsymbol{J}_B\cdot\boldsymbol{B}$, one would

[2]This statement should be carefully framed: the whole procedure of applying a function ϕ to the components of a spherical vector is not claimed to be optimal, but rather ϕ^* is supposed the best one of this class of transformations.

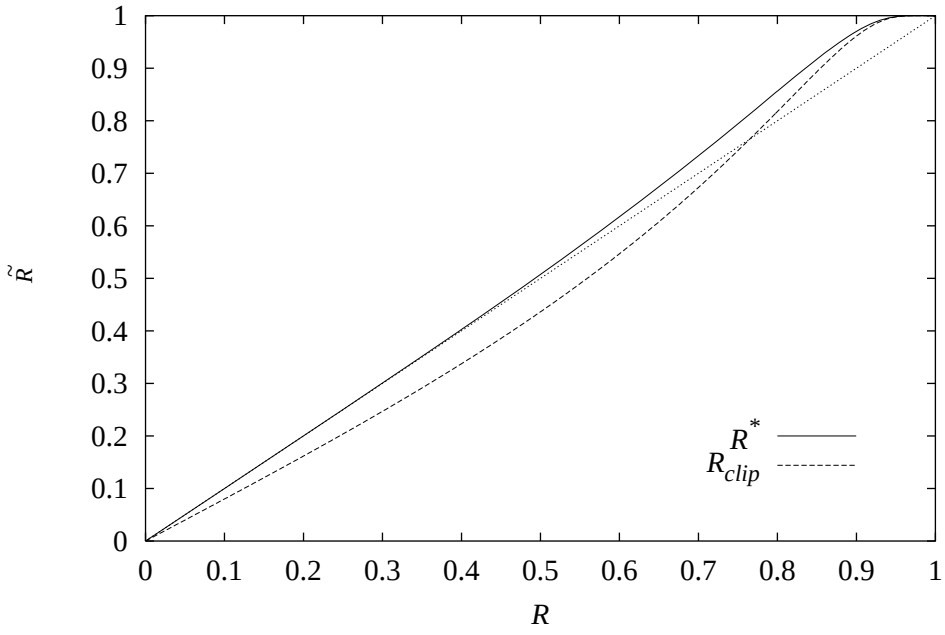

Figure 3.1: Transformed overlap according to ref. [SBVdB95]. The diagonal is plotted (dotted line) for comparison.

expect $R_{clip}(R_B)$ to be the desired $R_{bb} \equiv N^{-1}\boldsymbol{J}_{bb} \cdot \boldsymbol{B}$. But this would be a mistake, because one of the hypotheses taken for granted in [SBVdB95] is not satisfied in this case. The problem is in their last assumption of uniformity on the cone (page 40), which is reasonable in their context but not here. In this problem, the samples $\{\boldsymbol{J}_G^a\}$ are Ising vectors and $\boldsymbol{J}_B$ (which is the vector to be clipped) is a sum of these samples. Therefore one cannot expect different realizations of $\boldsymbol{J}_B$ to be *uniformly* distributed on the cone $\boldsymbol{J}_B \cdot \boldsymbol{B} = NR_B$. There should be a non-trivial structure in this distribution instead, due to the discrete nature of Gibbs space. Naively speaking, one would expect the vectors $\boldsymbol{J}_B$ to be more concentrated near the corners of the hypercube.

Therefore it is worthwhile going back to the equations of section 3.3.1 to make a clear distinction between the results which depend on that last assumption and those which do not. Eqs. 3.13 to 3.15 are clearly general, as long as the permutation symmetry among the axes holds (and there is no reason to believe it would not, in this case). Eqs. 3.16 and 3.17 do depend on the uniformity assumption, and so do all the equations which explicitly depend on them: these are eqs. 3.18, 3.22 and 3.23. The remaining ones, eqs. 3.19, 3.20 and 3.21, are again general.

One notices that the assumption of uniformity on the cone allows one to calculate $P(x)$ on geometric grounds only (see eq. 3.16). Without this assumption, $P(x)$ is the missing quantity which should be calculated in order for the properties of $\boldsymbol{J}_{bb}$ to be obtained. Note that this is not a simple problem: in [SBVdB95], the geometric properties of *one* vector $\boldsymbol{J}$ suffice to calculate the properties of the transformation 3.13. Here, $n \to \infty$ *correlated* vectors have to be taken into account.

Chapters 4 and 5 deal with the problem of calculating $P(x)$ in two different ways. The calculation in chapter 4 is inspired on the simplicity of eq. 3.16, based solely on geometry, while chapter 5 relies on a full replica calculation.

Chapter 4

The best binary: a geometric approach

4.1 Introduction

As mentioned in section 3.3.2, this chapter focuses on a purely geometric approach to obtain the properties of the clipped center of mass of the Gibbs ensemble. In fact, the calculations can be easily extended to treat the more general case of a vector

$$\boldsymbol{J}_{CM} \equiv \frac{1}{\sqrt{n+n(n-1)q}} \sum_{a=1}^{n} \boldsymbol{J}^a \ , \tag{4.1}$$

which is the (properly normalized) center of mass of n Ising vectors $\boldsymbol{J}^a$ satisfying the RS constraints, eqs. 1.15 and 1.16.

As discussed in section 3.3, the important quantity to be obtained is the probability distribution for $x = [\boldsymbol{J}_{CM}]_j B_j$, which in light of eq. 4.1 reads

$$x = \frac{B_1}{\sqrt{n+n(n-1)q}} \sum_{a=1}^{n} [\boldsymbol{J}^a]_1 \ . \tag{4.2}$$

Note that the choice of the first component is arbitrary due to the mentioned permutation symmetry among the axes. For convenience, however, calculations will be performed on the variable

$$y \equiv \frac{B_1}{n} \sum_{a=1}^{n} [\boldsymbol{J}^a]_1 \ , \tag{4.3}$$

which relates to eq. 4.2 by a simple multiplicative factor, $y = xn^{-1}\sqrt{n+n(n-1)q}$. One should remember that the sample construction

(eq. 4.1) is supposed to recover the results of an integral formulation of the center of mass (according to eq. 3.5, for instance), so ultimately the limit of interest is $n \to \infty$ (in which case $y \to x\sqrt{q}$).

The idea underlying the calculation that follows is to extend the argument behind eq. 3.16 given in [SBVdB95]. Namely, to calculate $P(y)$ by taking into account only the geometric constraints that the vectors are known to obey. In the case of [SBVdB95], the vector $\boldsymbol{J}$ was known to be spherically normalized and to have an overlap R with $\boldsymbol{B}$ (see eq. 3.16). The extension to the present problem is twofold: firstly, $\boldsymbol{J}_{CM}$ is composed of n vectors $\boldsymbol{J}^a$ (where $n \to \infty$ at the end); secondly, those vectors are Ising and obey the RS constraints given by eqs. 1.15 and 1.16. According to these extensions, a properly modified version of eq. 3.16 for the present case reads

$$\begin{aligned} P(y) &= \frac{1}{\mathcal{C}_n} \int \left[\prod_a^n d\boldsymbol{J}^a P_b(\boldsymbol{J}^a) \delta(\boldsymbol{J}^a \cdot \boldsymbol{B} - NR) \right] \\ &\times \left[\prod_{a<b} \delta(\boldsymbol{J}^a \cdot \boldsymbol{J}^b - Nq) \right] \delta\left(y - B_1 n^{-1} \sum_a^n J_1^a \right) , \end{aligned} \tag{4.4}$$

where $\mathcal{C}_n \equiv \int \left[\prod_a^n d\boldsymbol{J}^a P_b(\boldsymbol{J}^a)\delta(\boldsymbol{J}^a \cdot \boldsymbol{B} - NR)\right] \left[\prod_{a<b} \delta(\boldsymbol{J}^a \cdot \boldsymbol{J}^b - Nq)\right]$ is just a y-independent normalization constant and $P_b(\boldsymbol{J}^a)$ stands for the binary measure

$$P_b(\boldsymbol{J}) = \prod_{j=1}^N \left[\frac{1}{2}\delta(J_j - 1) + \frac{1}{2}\delta(J_j + 1) \right] . \tag{4.5}$$

The r.h.s. of eq. 4.4 resembles an ordinary replica calculation, except that the number of replicas tends to infinity rather than zero. This expression is calculated in appendix D. However, there is an equivalent but much more elegant way to calculate $P(y)$ on geometric grounds. It makes use of the Maximum-Entropy formalism (ME), which is dealt with in the next section.

4.2 The Maximum-Entropy formalism

This section is based upon the work of E. T. Jaynes [GMe93], who has developed a connection between information theory and statistical mechanics which proved very useful. The question addressed by Jaynes in his "Brandeis lectures" [Re83] is how to assign probabilities to a set of events given only a few physical or observable constraints. His prescription, the Maximum-Entropy formalism, acquired historical importance, among other reasons,

because it allows to postulate the usual probability distributions of Statistical Mechanics without the need to rely on ergodicity. Here it will be used as a powerful tool to simplify the calculations proposed in section 4.1. In the following, a brief overview of the Maximum-Entropy principle is given.

4.2.1 Theory

Given a quantity x, which can take on the values $(x_1, x_2, \ldots, x_n)$, and the average values of several functions $f_1(x), f_2(x), \ldots, f_m(x)$ (with $m < n$), how can one assign values to the probabilities $p_i \equiv p(x_i)$? The Maximum-Entropy formalism postulates that the set $\{p_i\}$ should maximize the Shannon [Sha48] information theory entropy $S_I \equiv -\sum_i p_i \log p_i$, subject to the constraints

$$\begin{aligned} p_i &\geq 0 \qquad i = 1, \ldots, n \\ \sum_{i=1}^{n} p_i &= 1 \\ \sum_{i=1}^{n} p_i f_k(x_i) &= \langle f_k(x) \rangle \qquad k = 1, \ldots, m \; . \end{aligned} \tag{4.6}$$

The solution to this variational problem can be found with Lagrange multipliers, which shall in the following be denoted by $\{\lambda_k\}$, $k = 1, \ldots, m$. The set $\{p_i^*\}$ which solves[1] the problem is given by

$$p_i^* = \frac{1}{Z(\lambda_1, \ldots, \lambda_n)} \exp\left(-\sum_{k=1}^{m} \lambda_k f_k(x_i) \right) \tag{4.7}$$

where

$$Z(\lambda_1, \ldots, \lambda_n) \equiv \sum_{i=1}^{n} \exp\left(-\sum_{k=1}^{m} \lambda_k f_k(x_i) \right) \; . \tag{4.8}$$

The last step then is to write the yet unspecified multipliers as functions of known quantities. One can readily verify that the following formula consistently satisfies the constraints 4.6:

$$\langle f_k \rangle = -\frac{\partial}{\partial \lambda_k} \ln Z(\lambda_1, \ldots, \lambda_n) \qquad k = 1, \ldots, m \; . \tag{4.9}$$

The solutions $\{\lambda_k^*\}$ to the above equations can then be used in the probabilities 4.7 and the problem is solved.

[1] For a formal proof that the set $\{p_i^*\}$ indeed yields the maximum of S_I, see [Re83].

4.2.2 Application to the best binary problem

This elegant approach is very convenient to deal with the problem of finding $P(y)$ according to the hypotheses of section 4.1. Instead of performing the calculations on the r.h.s. of eq. 4.4, one can make use in this case of the Maximum-Entropy formalism, which avoids the introduction of δ-distributions and the saddle point method. Both procedures give the same result, of course (see appendix D).

Noticing that the variables

$$x_a \equiv J_j^a B_j \quad a = 1, \ldots, n \tag{4.10}$$

can only take the values -1 or $+1$ and should not depend on j, the problem of finding $P(y)$ is almost solved if one can find the probabilities of the 2^n states of the vector $\boldsymbol{x} \equiv (x_1, \ldots, x_n)$. The constraints to be satisfied are

$$\begin{aligned} \langle x_a \rangle &= \frac{1}{N} \sum_{j=1}^{N} J_j^a B_j = R \qquad a = 1, \ldots, n \\ \langle x_a x_b \rangle &= \frac{1}{N} \sum_{j=1}^{N} J_j^a J_j^b = q \qquad \forall a < b \end{aligned} \tag{4.11}$$

and the corresponding "partition function" is simply given by

$$Z_{ME}\left(\{\hat{R}_a, \hat{q}_{ab}\}\right) = \int d\boldsymbol{x}\, P_b(\boldsymbol{x}) \exp\left[-\sum_a^n \hat{R}_a x_a - \sum_{a<b} \hat{q}_{ab} x_a x_b \right] . \tag{4.12}$$

The joint probability $P(\{x_a\})$ prescribed by the Maximum-Entropy formalism follows from 4.7,

$$P(\{x_a\}) = \frac{1}{Z_{ME}} \exp\left[-\sum_a^n \hat{R}_a x_a - \sum_{a<b} \hat{q}_{ab} x_a x_b \right] , \tag{4.13}$$

where, according to 4.9, the values of the Lagrange multipliers $\{\hat{R}_a\}$ and $\{\hat{q}_{ab}\}$ are to be determined by the relations $\langle x_a \rangle = -\partial \log Z_{ME}/\partial \hat{R}_a$ and $\langle x_a x_b \rangle = -\partial \log Z_{ME}/\partial \hat{q}_{ab}$. However, the dependence of Z_{ME} on $\hat{R}_a$ and $\hat{q}_{ab}$ must be independent of a and b due to the ("replica") symmetry present in the constraints 4.11. Therefore one concludes that the multipliers must satisfy

$$\begin{aligned} \hat{R}_a &= -\hat{R} \\ \hat{q}_{ab} &= -\hat{q} \end{aligned} \tag{4.14}$$

for some $\hat{R}$ and $\hat{q}$. Eqs. 4.14 can then be inserted back into eq. 4.12, rendering the evaluation of Z_{ME} very simple:

$$Z_{ME}(\hat{R},\hat{q}) = e^{-n\hat{q}/2} \int Dz \left[\cosh\left(\hat{R} + z\sqrt{\hat{q}}\right)\right]^n . \tag{4.15}$$

The last step is to obtain $\hat{R}$ and $\hat{q}$ as functions of R and q. The equations to be solved are

$$\begin{aligned} R &= \frac{1}{n}\frac{\partial}{\partial\hat{R}} \ln Z_{ME} \\ q &= \frac{2}{n(n-1)}\frac{\partial}{\partial\hat{q}} \ln Z_{ME} , \end{aligned} \tag{4.16}$$

where the n-dependent factors on the r.h.s. of the equations above come from the symmetry present in 4.14. One then immediately obtains

$$\begin{aligned} R &= \frac{\int du\, \exp\left[-(u-\hat{R})^2/2\hat{q}\right] (\cosh u)^n \tanh u}{\int du\, \exp\left[-(u-\hat{R})^2/2\hat{q}\right] (\cosh u)^n} \\ q &= \frac{\int du\, \exp\left[-(u-\hat{R})^2/2\hat{q}\right] (\cosh u)^n \tanh^2 u}{\int du\, \exp\left[-(u-\hat{R})^2/2\hat{q}\right] (\cosh u)^n} . \end{aligned} \tag{4.17}$$

Please note that these are equations for $\hat{R}$ and $\hat{q}$, where R and q are just parameters. Their solution $(\hat{R}_n(R,q), \hat{q}_n(R,q))$ can be inserted back into eq. 4.13 and the problem of finding $P(\{x_a\})$ is solved. Still, it remains to find the probability of $y = n^{-1}\sum_a^n x_a$. But a quick glimpse at eqs. 4.13 and 4.14 reveals that $P(\{x_a\})$ is already a function of y, since it only depends on the sum of the n variables[2]. Therefore one only needs to introduce a combinatorial factor,

$$P(y) = \frac{\exp\left[n\hat{R}_n y + n^2\hat{q}_n y^2/2\right] \binom{n}{\frac{n(1+y)}{2}}}{\sum_{y'} \exp\left[n\hat{R}_n y' + n^2\hat{q}_n y'^2/2\right] \binom{n}{\frac{n(1+y')}{2}}} , \tag{4.18}$$

[2]This should not be surprising, due to the symmetries present in the constraints 4.11.

where the sum on y' runs over its possible values $-1, -1+2/n, \ldots, 1-2/n, 1$.

The problem is thus solved, in principle. Unfortunately, an algebraic solution to eqs. 4.17 for general n could not be found, but they can always be solved numerically. This is done for finite n in section 4.4. The main interest here, however, is on reproducing the properties of the center of mass of infinite sample vectors.

4.3 The limit of infinite number of samples

In the limit $n \to \infty$ eqs. 4.17 can be solved exactly. The only requirement is that the Lagrange multipliers (or conjugate variables) $\hat{R}$ and $\hat{q}$ be properly rescaled with n:

$$\begin{aligned} \rho_n &\equiv n\,\hat{R}_n \\ \gamma_n &\equiv n\,\hat{q}_n\ . \end{aligned} \tag{4.19}$$

As will become clear soon, this scaling will turn out to be the correct one for the current purposes[3]. In terms of the new variables, eqs. 4.17 read

$$\begin{aligned} R &= \frac{\int du\, e^{-n\phi_n} \sinh(u\rho_n/\gamma_n) \tanh u}{\int du\, e^{-n\phi_n} \cosh(u\rho_n/\gamma_n)} \\ q &= \frac{\int du\, e^{-n\phi_n} \cosh(u\rho_n/\gamma_n) \tanh^2 u}{\int du\, e^{-n\phi_n} \cosh(u\rho_n/\gamma_n)}\ , \end{aligned} \tag{4.20}$$

where

$$\phi_n(u) \equiv \frac{u^2}{2\gamma_n} - \ln\cosh u\ . \tag{4.21}$$

Note that the only dependence of ϕ_n on n is through its dependence on γ_n. γ_n, on its turn, depends implicitly on n since its value is determined by the solution of eqs. 4.20. Let one assume (and check later) that γ_n reaches a finite value in the limit $n \to \infty$. Then the evaluation of the integrals on 4.20 via the saddle point method is straightforward. Note that, in each expression, the dominating exponential contributions in the numerator and denominator cancel, leaving just

[3] Eqs. 4.19 could also be justified as an *ansatz* since they arise naturally in a similar calculation for the spherical constraint.

$$\begin{aligned} R &\overset{n\to\infty}{=} \tanh(u_0\rho_\infty/\gamma_\infty)\tanh u_0 \\ q &\overset{n\to\infty}{=} \tanh^2 u_0 \ , \end{aligned} \tag{4.22}$$

where $u_0 \equiv \mathrm{Argmin}_u\, \phi_n$ satisfies[4]

$$u_0 = \gamma_\infty \tanh u_0 \ . \tag{4.23}$$

Eqs. 4.22 and 4.23 together finally yield

$$\begin{aligned} \gamma_\infty &= \frac{u_0}{\tanh u_0} = \frac{\mathrm{arctanh}\sqrt{q}}{\sqrt{q}} \\ \rho_\infty &= \frac{\mathrm{arctanh}(R/\sqrt{q})}{\sqrt{q}} \ . \end{aligned} \tag{4.24}$$

As previously announced, ρ_∞ and γ_∞ are finite[5], which gives consistency to the scaling *ansatz* 4.19. The solution 4.24 can now be inserted into eq. 4.18, the asymptotically dominant contribution from the combinatorial factor coming from Stirling's formula. Transforming the sum in the denominator of 4.18 into an integral, one arrives at the result

$$\begin{aligned} P(y) &\simeq \frac{\exp(\rho_\infty y)\exp n\left[\gamma_\infty y^2/2 - \ln\sqrt{1-y^2} - y\,\mathrm{arctanh}\, y\right]}{\int dy\ \exp(\rho_\infty y)\exp n\left[\gamma_\infty y^2/2 - \ln\sqrt{1-y^2} - y\,\mathrm{arctanh}\, y\right]} \\ &\overset{n\to\infty}{\to} \frac{1}{2}\left(1+\frac{R}{\sqrt{q}}\right)\delta(y-\sqrt{q}) + \frac{1}{2}\left(1-\frac{R}{\sqrt{q}}\right)\delta(y+\sqrt{q}) \ . \end{aligned} \tag{4.25}$$

Eq. 4.25 is the main result of this chapter. If one recalls the relation $x = [\boldsymbol{J}_{CM}]_1 B_1 \overset{n\to\infty}{\simeq} y/\sqrt{q}$ given on page 46, the conclusion is surprising: *in the limit of infinite number of samples, the center of mass of Ising vectors satisfying the constraints 4.11 is itself an Ising vector!*

One should contrast the delta-peaked distribution of eq. 4.25 with the Gaussian of eq. 3.17. Note that, according to eq. 4.25, the overlap $R_{CM} \equiv \boldsymbol{J}_{CM}\cdot\boldsymbol{B}/N$ is shown to satisfy $R_{CM} = \langle x\rangle = \langle y\rangle/\sqrt{q} = R/\sqrt{q}$ in the limit $n\to\infty$, which is reminiscent of eq. 3.10. A more extended discussion of the significance of these results will be left to section 4.5.

[4]Note that ϕ_n is even in u: if u_0 is a solution of 4.23, so is $-u_0$.

[5]Except for borderline cases such as $q\to 1$ or $q = R^2$, which are always singular.

4.4 Finite number of samples

Eq. 4.18 clearly shows that $\boldsymbol{J}_{CM}$ is a continuous vector if the limit $n \to \infty$ is not taken. This allows one to study the dependence of $P(y)$ on n to check how the delta-peaked shape is approached. For each triple (R, q, n), one has to solve eqs. 4.17 (or 4.20) to obtain the conjugate variables $\hat{R}_n$ and $\hat{q}_n$ (or γ_n and ρ_n) in order to plot eq. 4.18.

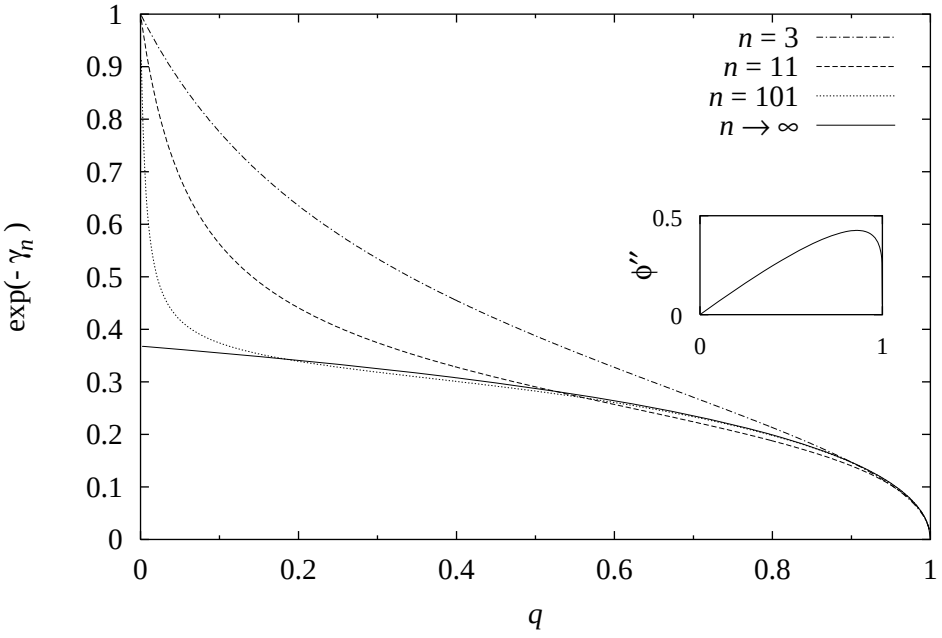

Figure 4.1: Numerical solution of the saddle point eqs. 4.20 for $q = R$ and $n = 3, 11, 101$. The inset shows that $\partial^2\phi_n/\partial u^2$ (as a function of q) remains always positive at $u = u_0$ for $n \to \infty$, which is a necessary condition for the validity of the saddle point method.

A few special cases can be solved exactly: for $R = 0$ one obtains $\hat{R}_n = 0$, while $\hat{q}_n(q)$ must still be solved for each value of n and q. For $q = R^2$, the Gaussians in eqs. 4.17 must become delta peaks for a solution to exist. This is attained by setting $\hat{q}_n = 0$, which implies $\hat{R}_n = \text{arctanh}R$, $\forall n$, and can be understood intuitively: if one fixes the overlap R *only*, the mutual overlap will *automatically* take the value $q = R^2$ with probability 1 in the TL. So there is no need for a second conjugate variable to enforce the q constraint. Precisely the same situation arises for $n = 1$, when it does not make sense

imposing the mutual overlap[6] q.

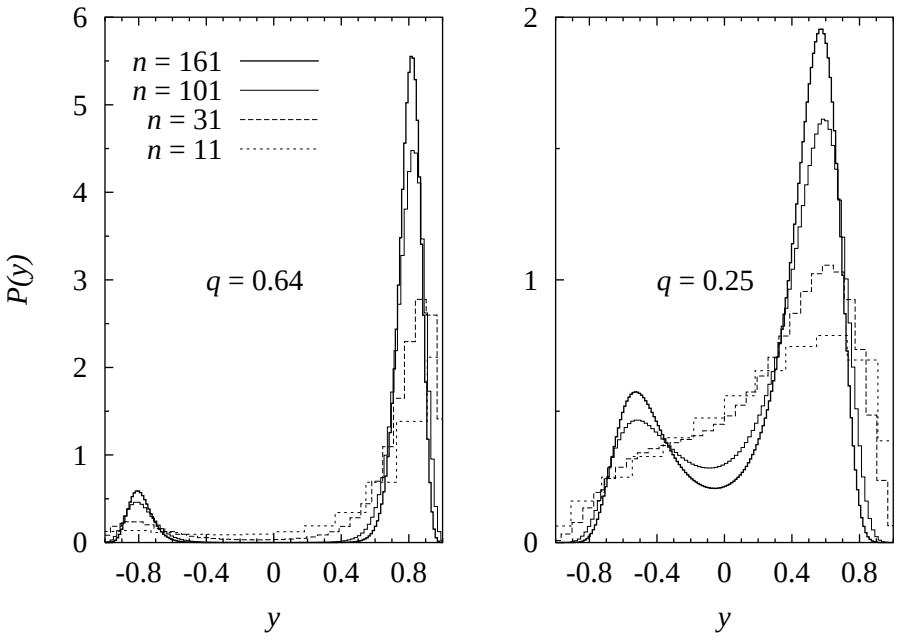

Figure 4.2: Probability density $P(y)$ for $q = R$ and $n = 11, 31, 101, 161$, according to eq. 4.18. Please note the different scales on the two plots.

The case $q = R$ was chosen to illustrate a scenario in which the solution of eqs. 4.20 had to be found numerically. In this situation the conjugate parameters can be immediately seen to satisfy $\rho_n = \gamma_n$, leaving just one equation to be solved. The behavior of γ_n as a function of q is similar to that of γ_∞ in the sense that it also has a logarithmic divergence when $q \to 1$ (see eqs. 4.24). For better visualization, fig. 4.1 plots $\exp(-\gamma_n)$ vs. q for some values of n. Note that nonuniform convergence seems to occur in the value of γ_n for $q \to 0$: $\lim_{q\to 0} \gamma_n = 0$ for any finite n, while $\lim_{q\to 0} \gamma_\infty = 1$.

At this stage, it is possible to visualize the convergence of $P(y)$ to the delta-peaked distribution predicted by eq. 4.25. Figure 4.2 shows that the value of n for which the peaks become distinct depends rather strongly on q. For instance, note that $n = 161$ is already large enough for the peaks to be clearly distinguished (around ± 0.8) for $q = 0.64$, but not for $q = 0.25$. From eq. 4.25, the width of the peaks should be approximately equal to $[(1-q)/(n(1-\gamma_\infty(1-q)))]^{1/2}$ for sufficiently large n, which means that small

[6] In this case on obtains the trivial result $P(y) = \left(\frac{1+R}{2}\right)\delta(y-1) + \left(\frac{1-R}{2}\right)\delta(y+1)$.

values of q require an extremely large number of samples for the peaked regime to be achieved. For $q \to 0$ and $n \to \infty$, the width of the peaks scales like $\sim (nq)^{-1/2}$.

4.4.1 Simulations

One would like to run simulations to check the theoretical result 4.18. One physical model which naturally fits in this framework is the ferromagnetic Ising model. In its mean field version (which is chosen for its simplicity), the probability of finding (in equilibrium) a spin configuration $\boldsymbol{S} = \{S_1, \ldots, S_N\} \in \{-1, +1\}^N$ is given by the Boltzmann factor, $P(\boldsymbol{S}) \sim \exp(-\mathcal{H}_{Ising}/kT)$, where

$$\frac{\mathcal{H}_{Ising}}{kT} = \frac{1}{\tau}\left\{ -\frac{1}{2N}\left(\sum_i^N S_i\right)^2 - \tilde{h}\sum_i^N S_i \right\} , \tag{4.26}$$

k is the Boltzmann constant and T is the temperature, while $\tau = kT/J$ and $\tilde{h} = h/J$ are the dimensionless temperature and magnetic field, $J > 0$ being the coupling constant.

In the TL, the distribution of the magnetization $m \equiv N^{-1}\sum_i S_i$ becomes peaked at the solution(s) of the equation

$$m = \tanh\left(\frac{m + \tilde{h}}{\tau}\right) . \tag{4.27}$$

Therefore one notices that, in the language of the current chapter, the realizations of $\boldsymbol{S}$ play the role of the samples $\boldsymbol{J}^a$, the preferential direction is $\boldsymbol{B} = (1, 1, \ldots, 1)$ and the magnetization m plays the role of R. Since the magnetization is the only constraint fixed by the spontaneous symmetry breaking, the mutual overlap between two different configurations obeys $\boldsymbol{S} \cdot \boldsymbol{S}'/N = m^2$, mapping the problem into the $q = R^2$ scenario which was mentioned on page 52. In this case, the center of mass (in the limit $n \to \infty$) is clearly the vector $\boldsymbol{B}$.

The simulations were performed as follows. Using the Metropolis algorithm (basically as described in section 2.4.4), the configurations were sampled every 5 MCS/site to allow sufficient decorrelation between consecutive samples. After n vectors were kept, their center of mass was constructed and the relevant variable x was measured for all the N components of the vector, the whole procedure being repeated until a good histogram could be achieved.

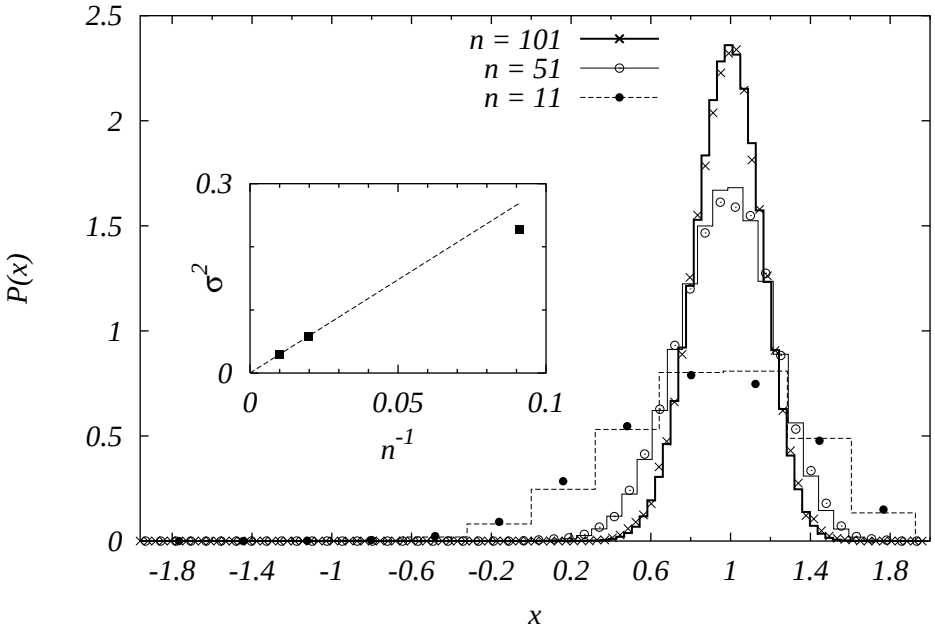

Figure 4.3: Probability density $P(x)$ for $q = R^2$ and $n = 11, 51$ and 101. The lines represent the theoretical curve (obtained by a trivial change of variables $x = y/\sqrt{q}$ in eq. 4.18), while points represent simulations with $N = 100$, $\tau = 1.09$ and $\tilde{h} = 0.1$ (amounting to a magnetization $m = R \simeq 0.5$). Inset: variance (σ^2) of the distribution as a function of $1/n$. Simulations (points) and the theoretical asymptotics (dashed line) $\sigma^2 \sim (1 - R^2)/(nR^2)$.

Results are shown in figure 4.3 for a system size $N = 100$ and model parameters $\tau = 1.09$ and $\tilde{h} = 0.1$. The agreement with theory is excellent, and deviations are due to finite size effects (for $N = 1000$, the results are nearly indistinguishable from the theoretical results on the scale of the figure). Note the $1/\sqrt{n}$ behavior of the width of the peaks, for large n. It should be stressed that, for finite N, the required constraints are satisfied only *on average*, the magnetization R being distributed with a width typically scaling with $1/\sqrt{N}$ (see, for instance, fig. 2.3). Nonetheless the behavior of $P(y)$ seems not to be significantly affected by these fluctuations.

4.5 On the validity of the geometric approach

At this point, an important discussion is in order. It concerns the validity of results 4.18 and 4.25 and the hypotheses involved in their derivation. Notice that both in the case of a single sample and several samples (eqs. 3.16 and 4.4 respectively – the starting points of the calculations), no mention whatsoever is made about the algorithm which gave rise to the student vector(s). Indeed, the proposed calculations do not involve the data $\{\boldsymbol{\xi}^\mu\}$, nor the cost function $\mathcal{H}$, nor the temperature T the system is immersed in. Actually, the r.h.s. of those equations can be interpreted as if the probability density of the samples was determined exclusively by the geometrical constraints, that is $P(\{\boldsymbol{J}^a\}) = c \times \prod_a P_b(\boldsymbol{J}^a)\delta(\boldsymbol{J}^a \cdot \boldsymbol{B} - NR) \prod_{a<b} \delta(\boldsymbol{J}^a \cdot \boldsymbol{J}^b - Nq)$. But how could this be compatible with the hypotheses underlying the replica calculation, namely that the equilibrium distribution of a candidate vector for a given data set is $P(\boldsymbol{J}|D) = Z(D)^{-1} \exp(-\beta\mathcal{H}(\boldsymbol{J}))$? Do the calculations of section 4.2.2 account for the properties of the solutions of a given learning algorithm or do they just reflect a geometrical property of infinite-dimensional Ising vectors?

This problem can be better understood if one goes back to the general structure of the replica calculation. In order to calculate $\langle \ln Z(D) \rangle_{D|\boldsymbol{B}}$, one introduces n replicas $\boldsymbol{J}^a$, $a = 1, \ldots, n$ of the original system, replacing the average of the logarithm by $\lim_{n\to 0} n^{-1} \langle Z^n \rangle_{D|\boldsymbol{B}} = \lim_{n\to 0} n^{-1} \langle \prod_a^n Z_a \rangle_{D|\boldsymbol{B}}$. This trick (of avoiding the average of the logarithm) has the price of coupling different replicas, which can be seen in terms like $\boldsymbol{J}^a \cdot \boldsymbol{J}^b$ which show up in the "entropic term" G_0 (see appendix B and eq. 4.28 below). In order to be able to proceed with the calculation, one introduces δ-distributions for the order parameters q_{ab} and R_a which, together with an appropriate *ansatz* (e.g. with replica symmetry or m-step broken symmetry) completely fixes the geometric structure of the replicas $\{\boldsymbol{J}^a\}$. The values of these order parameters are then determined by minimizing the free energy:

$$-\beta f = \lim_{n\to 0} \frac{1}{n} \operatorname*{Extr}_{\{\hat{R}_a, \hat{q}_{ab}, R_a, q_{ab}\}} \left\{ \sum_a^n \hat{R}_a R_a + \sum_{a<b} \hat{q}_{ab} q_{ab} + G_0 \left(\{\hat{R}_a, \hat{q}_{ab}\} \right) + \alpha G_1 \left(\{R_a, q_{ab}\}; \beta, [V] \right) \right\}, \tag{4.28}$$

where it is important to remind that the so-called "energy term" G_1 involves no integrals on the $\{\boldsymbol{J}^a\}$ space. *Therefore the geometric structure of the vectors is completely determined by the constraints comprised in the entropic term.* Please note also that G_0 depends only on the conjugate variables $\{\hat{R}_a, \hat{q}_{ab}\}$, while G_1 depends on the order parameters $\{R_a, q_{ab}\}$ and contains the information about the learning process through its dependence on β and

V. Moreover, G_0 corresponds *exactly* to the logarithm of the previously defined Z_{ME}, for general n. Taking this remark into account, one rewrites 4.28 to obtain

$$\begin{aligned} -\beta f \;=\; & \lim_{n\to 0} \frac{1}{n} \operatorname*{Extr}_{\{R_a, q_{ab}\}} \Bigg\{ \\ & \operatorname*{Extr}_{\{\hat{R}_a, \hat{q}_{ab}\}} \Big\{ \sum_a^n \hat{R}_a R_a + \sum_{a<b} \hat{q}_{ab} q_{ab} + \ln Z_{ME}\left(\{\hat{R}_a, \hat{q}_{ab}\}\right) \Big\} \\ & + \alpha G_1\left(\{R_a, q_{ab}\}; \beta, [V]\right) \Bigg\} . \end{aligned} \tag{4.29}$$

The innermost extremum operator yields equations for $\{\hat{R}_a, \hat{q}_{ab}\}$ as functions of $\{R_a, q_{ab}\}$ which precisely obey the prescriptions of the Maximum-Entropy formalism[7], eq. 4.9 (or 4.16 in the specific case of replica symmetry). In other words, the conjugate parameters at their "equilibrium" values are such that the entropic term is indeed maximized. Only after the outermost extremum operator is evaluated does the free energy attain its minimum value. This last step fixes the order parameters $\{R_a, q_{ab}\}$ as functions of α, β, V etc.

Therefore there seems to be several arguments supporting the presented calculations. If one assumed that no physical meaning is attached to the conjugate parameters $\{\hat{R}_a, \hat{q}_{ab}\}$, the conclusion that would follow is that in the replica calculation they are just auxiliary variables which connect the order parameters $\{R_a, q_{ab}\}$ with α, β, V etc., being otherwise irrelevant. With this reasoning, one could claim that the results of section 4.2.2 do reproduce the properties of vectors obtained via a learning process. One would just have to replace (R, q) by $(R(\alpha; \beta, [V]), q(\alpha; \beta, [V]))$.

However, an apparently technical subtlety could be a sign of difficulties in the reasoning: in the replica calculation, the limit $n \to 0$ is taken at the end, while the formulation in terms of the sample construction presented in this chapter requires the limit $n \to \infty$ to be taken. The value of n clearly changes the equilibrium value of the conjugate parameters in both formulations, while the values of the order parameters R and q are assumed given in the ME formalism. If $\hat{R}$ and $\hat{q}$ had indeed no physical meaning, this should not matter. But does it or does it not? After all, the trick of taking the limit $n \to 0$ was introduced precisely to allow the calculation of the average of $\ln Z$ over the disorder. The answer will be given in the next chapter.

[7]The ME-like character of the replica calculation is not a novelty on itself, having already been noticed by Gardner and Derrida (see page 272 of [GD88]).

Chapter 5

The center of mass of the Gibbs ensemble

5.1 Introduction

In chapter 4 a calculation based solely on the geometric properties of the samples led to the conjecture that the center of mass of the Gibbs ensemble would be a binary vector. Even though section 4.5 provided plausible arguments supporting the underlying assumptions, the present chapter is intended to show that the conclusions drawn from those results do not hold for disordered systems[1].

If one would start by summarizing the new results that are about to be shown, the basic statement could be the following: in a disordered system, the geometric properties of the samples do *not* convey enough information to describe the statistics of the problem, as it was assumed. The simplest way to verify this is by performing an alternative calculation of $P(y)$. Instead of using a "sample construction" like in eq. 3.8 and relying only on geometry, one can rather depart from an integral formulation like in eq. 3.5. In other words, one describes the center of mass $\boldsymbol{J}_{CM}$ of the samples, for a given learning process, as the *thermal average*

$$\boldsymbol{J}_{CM} \stackrel{N\to\infty}{=} q^{-1/2} \int d\boldsymbol{J}\, \boldsymbol{J}\, P(\boldsymbol{J}|D) = q^{-1/2} \int d\boldsymbol{J}\, \boldsymbol{J}\, \exp -\beta\mathcal{H}(\boldsymbol{J}, D)\ , \quad (5.1)$$

where the factor $q^{-1/2}$ accounting for the normalization can be obtained with the same reasoning employed in eqs. 3.8 and 3.9. One should note that the

[1]This explains the good agreement between theory and simulations for the Ising model — a system without disorder.

above calculation automatically gives the results for an infinite number of samples. A more extended discussion about the validity of the geometric calculation is presented in section 5.6.

5.2 $P(y)$ calculated via thermal averages

Once more, the probability distribution for $x = B_1[\boldsymbol{J}_{CM}]_1$ or $y = B_1 \langle J_1 \rangle_{\boldsymbol{J}} = x\sqrt{q}$ is the relevant quantity. The idea in this chapter is to reconstruct $P(y)$ from its *quenched moments*, which are averages over the examples of powers of y. This should be the correct way to account for the effect of the disorder. The expression hereby obtained will be referred to as $P_{CM}(y)$. Apart from normalization constants, the m-th moment of the first component of the center of mass[2] is

$$\langle \langle J_1 \rangle_{\boldsymbol{J}}^m \rangle_{D|\boldsymbol{B}} = \left\langle Z^{-m} \left(\int d\boldsymbol{J}\, P(\boldsymbol{J})\, e^{-\beta \mathcal{H}(\boldsymbol{J},D)} J_1 \right)^m \right\rangle_{D|\boldsymbol{B}} , \qquad (5.2)$$

where m is an integer. The replica trick can once more be applied to this calculation, the details of which can be found in appendix E. Here it suffices to present the final result, which for a RS *ansatz* is

$$\langle \langle J_1 \rangle_{\boldsymbol{J}}^m \rangle_{D|\boldsymbol{B}} = \int \mathcal{D}z \, \left[\tanh \left(z\sqrt{\hat{q}(\alpha)} + \hat{R}(\alpha) B_1 \right) \right]^m , \qquad (5.3)$$

or, equivalently,

$$\langle y^m \rangle_{D|\boldsymbol{B}} = \langle \langle B_1 J_1 \rangle_{\boldsymbol{J}}^m \rangle_{D|\boldsymbol{B}} = \int \mathcal{D}z \, \left[\tanh \left(z\sqrt{\hat{q}(\alpha)} + \hat{R}(\alpha) \right) \right]^m , \qquad (5.4)$$

which holds only if $\boldsymbol{B} \in \{-1, +1\}^N$, as usual. In both equations, $\hat{q}(\alpha)$ and $\hat{R}(\alpha)$ denote the conjugate parameters taken at their equilibrium values (see appendix E for details).

With an explicit expression for all the moments, one can reconstruct the probability distribution $P_{CM}(y)$. The easiest and most elegant way is by a close inspection of eq. 5.4: on the l.h.s. the m-th power of y is averaged over the disorder, whereas on the r.h.s. the m-th power of $\tanh \left(z\sqrt{\hat{q}(\alpha)} + \hat{R}(\alpha) \right)$ is averaged over a Gaussian distribution for z. Therefore one can immediately identify a transformation of (stochastic) variables

[2] Like in chapter 4, the calculation is done for a general learning process: Gibbs learning can then be obtained by setting $\beta = 1, V = U \Rightarrow R = q$, and then $\boldsymbol{J}_{CM} = \boldsymbol{J}_B$.

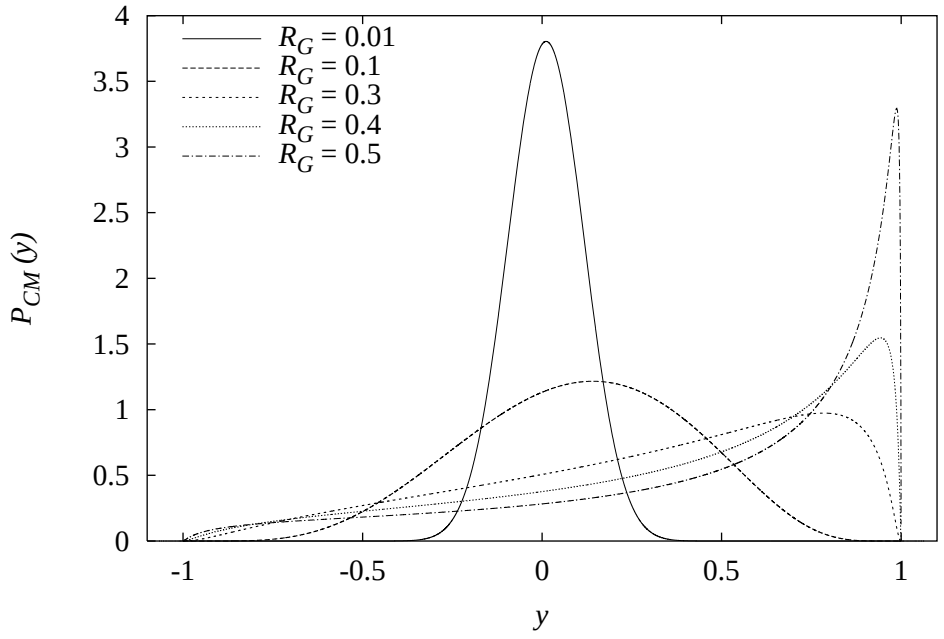

Figure 5.1: Probability distribution of $y = B_1 \langle J_1 \rangle_{\boldsymbol{J}}$ for different values of $R = q (\Rightarrow \hat{R} = \hat{q})$, according to eq. 5.6.

$$y = \tanh\left(z\sqrt{\hat{q}(\alpha)} + \hat{R}(\alpha) \right) \ , \tag{5.5}$$

where z is Gaussian distributed. By applying the identity $dy\, P_{CM}(y) = dz\, P_n(z)$ and noting that $dy/dz = \sqrt{\hat{q}(\alpha)}(1-y^2)$, one can write

$$P_{CM}(y) = \frac{1}{\sqrt{2\pi\hat{q}}(1-y^2)} \exp \frac{-1}{2\hat{q}} \left[\frac{1}{2} \ln\left(\frac{1+y}{1-y} \right) - \hat{R} \right]^2 \tag{5.6}$$

$$P_{CM}(x) = \frac{\sqrt{q}}{\sqrt{2\pi\hat{q}}(1-qx^2)} \exp \frac{-1}{2\hat{q}} \left[\frac{1}{2} \ln\left(\frac{1+\sqrt{q}x}{1-\sqrt{q}x} \right) - \hat{R} \right]^2 \ , \tag{5.7}$$

where the α dependence of q, $\hat{q}$ and $\hat{R}$ was omitted for clarity. From eq. 5.5 it becomes clear that the absolute value of y is bounded by one (as it should), and eq. 5.6 reflects this fact, since $P_{CM}(y)$ vanishes when $y \to \pm 1^{\mp}$.

In what follows, the special case of Gibbs learning will be emphasized. Not only it is the case of most interest here due to its connection with optimal learning, but it also presents the symmetry $\hat{q}_G(\alpha) = \hat{R}_G(\alpha)$, which simplifies

the calculations. If eqs. 2.5 and 2.10 (which specify R_G as a function of $\hat{R}_G$) can be formally inverted, then one can obtain the relation $\hat{R}_G(R_G)$. The distributions 5.6 or 5.7 can thus be plotted with R_G as a parameter, *regardless of* U *and* α.

The fact that α can be eliminated, leaving all the dependence of the results on the order parameter R_G, is analogous to the result obtained in [OK96] by Opper and Kinzel (see also [OH91]). Defining the volume of the "version space" for p patterns as V_p, they show that the probability density of the ratio V_{p+1}/V_p is completely parametrized by the order parameter q, the result holding for both spherical and binary problems.

$P_{CM}(y)$ is plotted in fig. 5.1. It is clearly *not* a sharply peaked distribution as those obtained in chapter 4, from which one can conclude that *the center of mass of the Gibbs ensemble is not a binary vector*, as opposed to what had been previously conjectured.

5.2.1 Simulations

In order to check the result 5.7, Gibbs learning was simulated for the Gaussian scenario in the biased region (see section 2.4.4), with parameters set to $A = 0$ and $B = 1$. The simulations were performed as described in section 2.4.4, with a system size $N = 500$. By tuning α, one determines the average value of R_G. For each training set of αN patterns, the following procedure was done:

1. After a random initialization of the $\boldsymbol{J}$ vector, 10 MCS/site were run in order for the system to reach equilibrium.

2. The current $\boldsymbol{J}$ sample was kept and the vector was reinitialized thereafter.

3. Items 1 and 2 were repeated until $n = 50$ samples had been collected.

4. After summing the 50 samples and normalizing, each of the N components provided one measure of x.

5. The procedure from 1 to 4 was repeated 100 times for each set of patterns.

6. The procedure from 1 to 5 was repeated for 100 different sets of patterns.

7. $P(x)$ was reconstructed from the histogram of $500 \times 100 \times 100$ measurements of x.

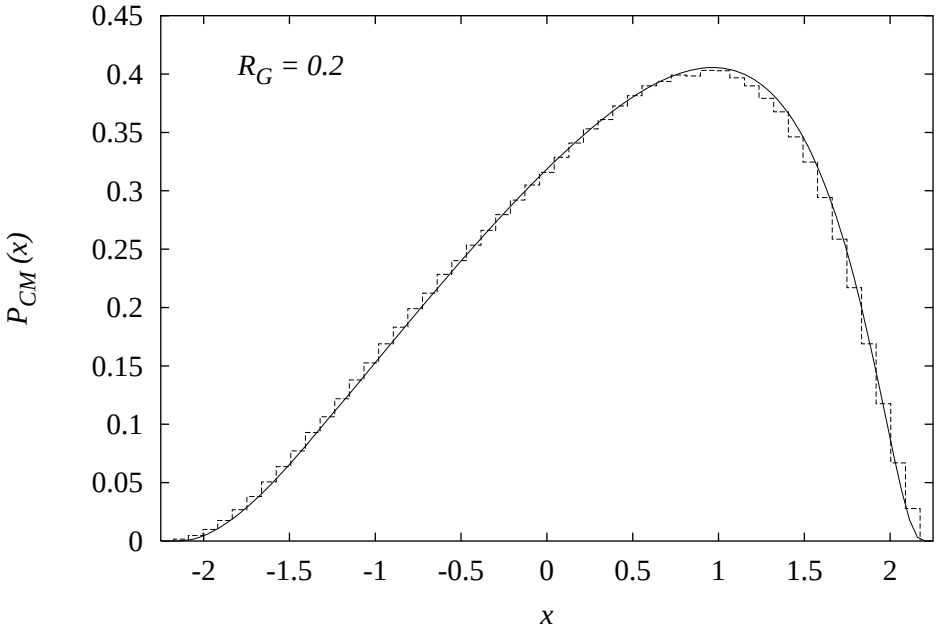

Figure 5.2: Probability distribution of $x = q^{-1/2} B_1 \langle J_1 \rangle_{\boldsymbol{J}}$ for $\alpha = 0.25 \Rightarrow R_G = q_G = 0.2$. Solid line: eq. 5.7. Dashed line: simulations.

The results of the simulations are presented in figs. 5.2 to 5.4. The agreement between theoretical results and simulations is very good, showing that indeed eq. 5.7 is the right expression for $P(x)$ in a quenched system, while eq. 4.25 properly describes a system without disorder.

5.3 The best binary vector

Since the center of mass $\boldsymbol{J}_B$ of the Gibbs ensemble is not binary, the best binary vector $\boldsymbol{J}_{bb}$ is obtained by clipping $\boldsymbol{J}_B$ (see section 3.3). In order to study the properties of this vector, the techniques described in section 3.3 can be employed.

First it is interesting to check the consistency of the results. If eq. 5.7 represents indeed the correct expression of $P(x)$, then *any* transformation $J_i' = \sqrt{N}\phi(J_i)/\sqrt{\sum_j \phi^2(J_j)}$ *necessarily* leads to a decrease in the overlap, for $\beta V = U$. Particularly, the optimal transformation ϕ^* should be such that the transformed overlap is unchanged. As can be seen from eq. 3.13, any linear transformation satisfies this requirement, implying that the following

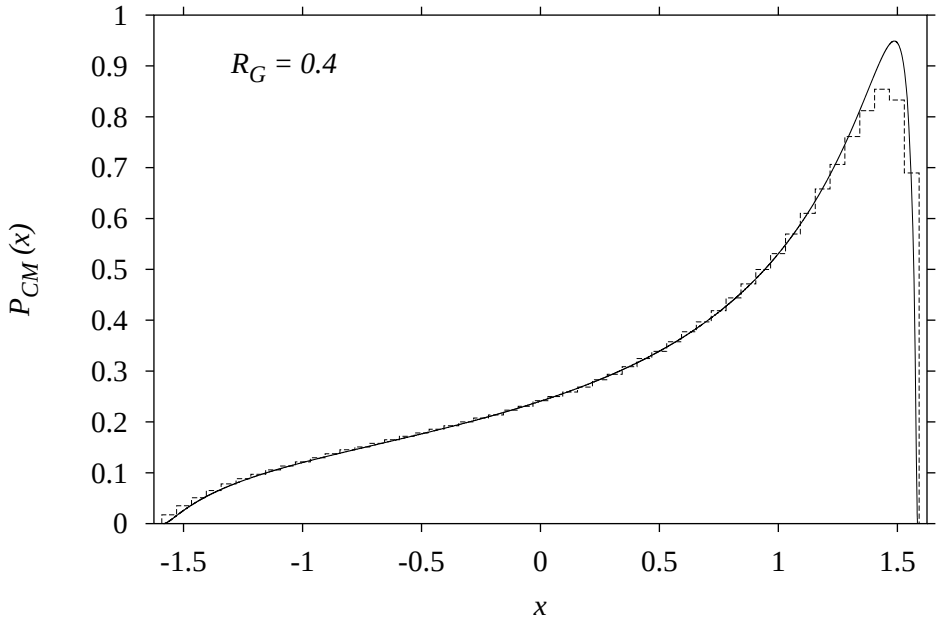

Figure 5.3: Probability distribution of $x = q^{-1/2} B_1 \langle J_1 \rangle_{\boldsymbol{J}}$ for $\alpha = 0.6 \Rightarrow R_G = q_G = 0.4$. Solid line: eq. 5.7. Dashed line: simulations.

condition should be obeyed:

$$\begin{aligned} \phi^*(x) &= \frac{P(x) - P(-x)}{P(x) + P(-x)} = cx \ , \qquad \forall c > 0 \\ \Rightarrow \frac{P(x)}{P(-x)} &= \frac{1 + cx}{1 - cx} \ . \end{aligned} \tag{5.8}$$

It is straightforward to verify that eq. 5.7 yields

$$\frac{P_{CM}(x)}{P_{CM}(-x)} \overset{(5.7)}{=} \exp\left[\frac{\hat{R}}{\hat{q}} \ln\left(\frac{1 + \sqrt{q}x}{1 - \sqrt{q}x}\right)\right] \tag{5.9}$$

$$\overset{\hat{R}=\hat{q}}{=} \frac{1 + \sqrt{q}x}{1 - \sqrt{q}x} \ . \tag{5.10}$$

Note that in order to get from eq. 5.9 to eq. 5.10, the condition $\hat{R} = \hat{q}$ must be imposed. Two important conclusions can be drawn from these equations: first, $P_{CM}(x)$ satisfies eq. 5.8, thus passing the consistency check; second,

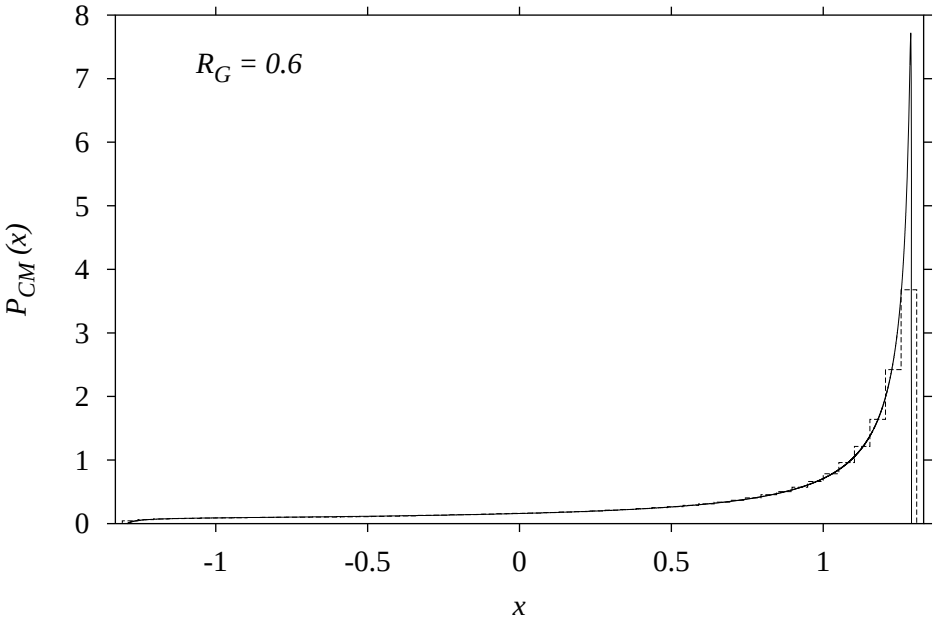

Figure 5.4: Probability distribution of $x = q^{-1/2} B_1 \langle J_1 \rangle_{\boldsymbol{J}}$ for $\alpha = 1.17 \Rightarrow R_G = q_G = 0.6$. Solid line: eq. 5.7. Dashed line: simulations.

this is true *only* for $\hat{R} = \hat{q}$, which means that the center of mass of the Gibbs ensemble is probably the only vector[3] which cannot be improved by a transformation (except, of course, for the cases where $q = 1$, when there is a single binary vector in the ensemble – binary vectors cannot be improved with a transformation like in eq. 3.13).

Now one should be able to calculate the performance of the vector resulting from clipping $\boldsymbol{J}_{CM}$, that is, for $\phi(x) = \text{sign}(x)$. Applying the change of variables 5.5 to eq. 3.15, one obtains the overlap $R_{CM}^{clip} = N^{-1} \sum_j^N \text{sign}([\boldsymbol{J}_{CM}]_j) B_j$:

$$\begin{aligned} R_{CM}^{clip} &= \int P_{CM}(x)\, \text{sign}(x)\, dx \\ &= \int \mathcal{D}z\, \text{sign}\left(z\sqrt{\hat{q}(\alpha)} + \hat{R}(\alpha) \right) \end{aligned}$$

[3]Unless one could be able to tune $\beta U \neq V$ such that $\hat{R}(\alpha) = \hat{q}(\alpha)$. Even though this seems to be possible in principle for a *fixed* value of α, the discussion in appendix C suggests that only Gibbs learning would give rise to this symmetry in general.

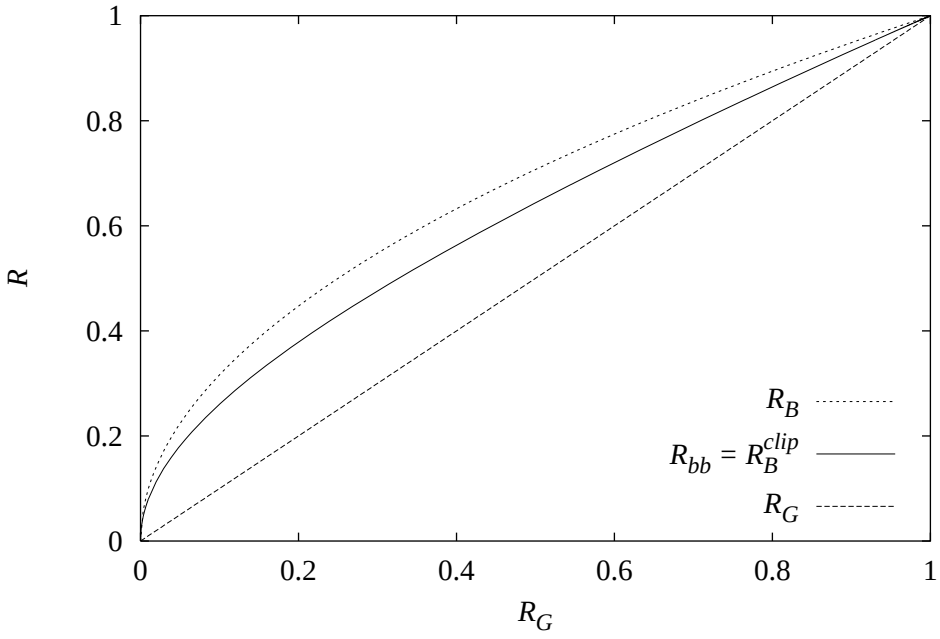

Figure 5.5: $R_{bb} = R_B^{clip}$ as a function of R_G, according to eq. 5.12. The diagonal is plotted for comparison.

$$= \quad 1 - 2H\left(\frac{\hat{R}(\alpha)}{\sqrt{\hat{q}(\alpha)}}\right) \ . \tag{5.11}$$

Since $\hat{q}_G(\alpha) = \hat{R}_G(\alpha)$ for Gibbs learning, and taking into account eqs. 2.5, 2.10 and 3.10 which connect $\hat{R}_G$ to $R_B = \sqrt{R_G}$, one concludes that *the performance of the best binary vector* $\boldsymbol{J}_{bb} = \mathrm{clip}(\boldsymbol{J}_B)$ *is given by*

$$R_{bb} = R_B^{clip} \quad = \quad 1 - 2H\left(F_B^{-1}(R_B)\right) = 1 - 2H\left(F_B^{-1}\left(\sqrt{R_G}\right)\right) \tag{5.12}$$

$$= \quad 1 - 2H\left(\mathcal{F}\left(R_B\right)\right) = 1 - 2H\left(\sqrt{\hat{R}_G}\right) \ , \tag{5.13}$$

where F_B^{-1} is the inverse function of F_B and the arguments of the functions above should be taken at their equilibrium values.

Eq. 5.12 is the main result of this chapter. It expresses R_{bb} as a function of R_B (or R_G) *regardless of the dependence of* R_G *on* U *and* α. The result is based purely on the binary nature of the samples and the RS solution for Gibbs learning. A plot of eq. 5.12 is shown in fig. 5.5.

5.3.1 Asymptotics

The performance of $\boldsymbol{J}_{bb}$ obviously depends on the performance of $\boldsymbol{J}_G$ (eq. 5.12), so it is as hard to describe the general behavior of the former as it is to describe the general behavior of the latter. For general α, little can be said when the function U is not specified. But for large and small α, the asymptotics of Gibbs learning (section 2.2.2) allows one to learn what the asymptotics of R_{bb} will be.

The limit $\alpha \to \infty$

Putting together eqs. 2.14 and 5.12, one concludes that the large α behavior of R_{bb} is given by

$$1 - R_{bb} \overset{\alpha\to\infty}{\simeq} \sqrt{\frac{2}{\pi\alpha \left\langle (U')^2 \right\rangle_*}} \exp\left(\frac{-\alpha \left\langle (U')^2 \right\rangle_*}{2}\right) . \tag{5.14}$$

This should be compared with the asymptotics of both Gibbs and Bayes learning (eq. 2.15). The dominant exponential term is the same in all three cases, with only a multiplicative factor differing among them. In order to have an idea of how good the performance of the best binary is, the following ratios seem appropriate:

$$\frac{1 - R_{bb}}{1 - R_G} \overset{\alpha\to\infty}{\to} \frac{2}{\pi} \simeq 0.64 \tag{5.15}$$

$$\frac{1 - R_B}{1 - R_{bb}} \overset{\alpha\to\infty}{\to} \frac{\pi}{4} \simeq 0.78 . \tag{5.16}$$

So even asymptotically, R_{bb} is larger than R_G and smaller than R_B (as it should).

The limit $R_G \to 0$

The series expansion of the H function in eq. 5.12 and the inclusion of eq. 2.16 immediately lead to

$$R_{bb} \overset{R_G\to 0}{\simeq} \sqrt{\frac{2R_G}{\pi}} . \tag{5.17}$$

This means that the phase transitions (if any) occur at exactly the same place as the ones for Gibbs learning. Moreover, in this poor performance regime the relation between R_{bb} and R_B is simply given by

$$\frac{R_{bb}}{R_B} \overset{R_G \to 0}{\to} \sqrt{\frac{2}{\pi}} \,. \tag{5.18}$$

The above ratio had been previously observed between the performances of supervised Hebbian[4] learning and its clipped counterpart [VdBB93, BS95].

5.4 The overlap between $\boldsymbol{J}_B$ and $\boldsymbol{J}_{bb}$

An interesting quantity to look at is the overlap between $\boldsymbol{J}_B$ and its clipped counterpart $\boldsymbol{J}_{bb}$. This new quantity will be called

$$\Gamma \equiv \frac{\boldsymbol{J}_B \cdot \boldsymbol{J}_{bb}}{N} \,. \tag{5.19}$$

If one takes randomly a vector $\boldsymbol{J}_{sph}$ from the hypersphere, the overlap between this vector and its clipped version will have a distribution peaked at $\sqrt{2/\pi}$ in the thermodynamic limit. $\boldsymbol{J}_B$, however, is not taken from such an isotropic distribution, and one can calculate from eqs. 5.5 and 5.6 what the value of Γ is:

$$\begin{aligned}
\Gamma &= \frac{1}{N}\sum_{i=1}^{N} [\boldsymbol{J}_B]_i [\boldsymbol{J}_{bb}]_i \\
&= \frac{1}{N}\sum_{i=1}^{N} |[\boldsymbol{J}_B]_i| \\
&= \langle |[\boldsymbol{J}_B]_1| \rangle \\
&= \frac{1}{\sqrt{q}} \int \mathcal{D}z \left| \tanh\left(z\sqrt{\hat{R}_G(\alpha)} + \hat{R}_G(\alpha) \right) \right| \\
&= \frac{R_{bb}}{\sqrt{q}} = \frac{R_{bb}}{R_B} \,.
\end{aligned} \tag{5.20}$$

The result of eq. 5.20 is plotted in fig. 5.6 as a function of R_G. Note that R_B is also plotted for comparison, lying always below Γ. As a matter of fact, the bound $\Gamma \geq R_B$ could have been imposed from the very beginning on geometric grounds, since by definition $\boldsymbol{J}_{bb}$ is the binary vector which is closest to $\boldsymbol{J}_B$ (therefore $\boldsymbol{B}$ could not be closer). It is not surprising either that

[4] It can be shown [KC96, VdBR96] that, for a spherical prior, Hebbian learning is Bayes-optimal in the limit $\alpha \to 0$. Eq. 5.18 states that, in the same regime, *clipped* Hebbian learning gives the best binary vector, for a *binary* prior.

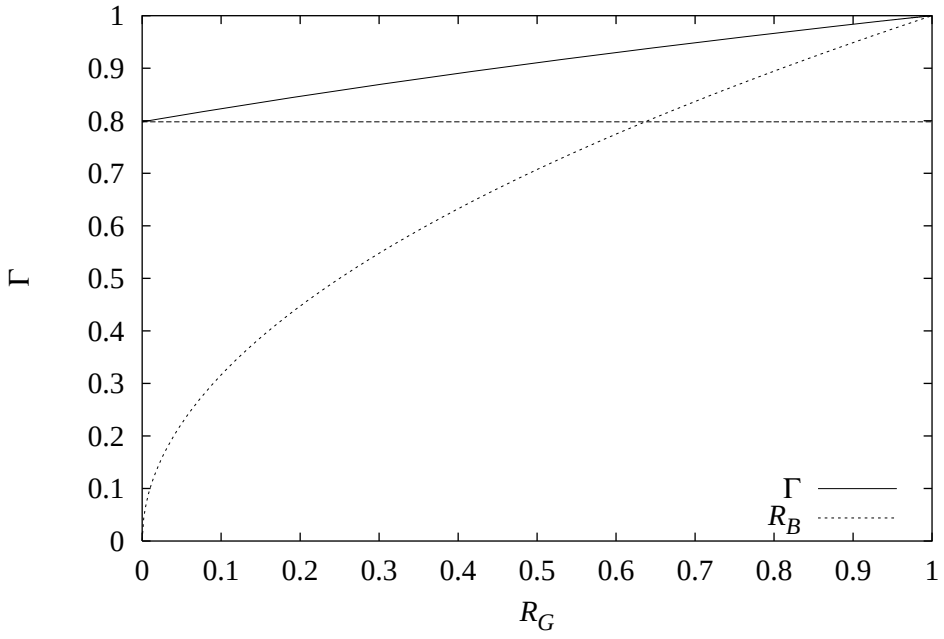

Figure 5.6: Γ as defined in eq. 5.19 as a function of R_G (R_B is also plotted for comparison). The horizontal line is $\sqrt{2/\pi}$.

$\Gamma \geq \sqrt{2/\pi}$, since $\boldsymbol{J}_B$ should indeed lie closer to the corners of the hypercube than a vector taken at random from the hypersphere.

Finally, a comment about eqs. 5.20 is in order. The attentive reader may have felt uncomfortable with the last passage, since it implies the somewhat counterintuitive equality

$$\int \mathcal{D}z \, \left|\tanh\left(za + a^2\right)\right| = \int \mathcal{D}z \, \mathrm{sign}\left(za + a^2\right) \ . \tag{5.21}$$

Indeed, the equality above holds true for any a and a proof is given in appendix E. Apart from that, the result $R_{bb} = \Gamma R_B$ has a simple geometric implication, which is derived below.

Denoting by $\boldsymbol{V}'$ or $\boldsymbol{V}''$ a properly normalized vector orthogonal to $\boldsymbol{V}$ (for any N-dimensional vector $\boldsymbol{V}$), one can write down the following three decompositions:

$$\boldsymbol{J}_{bb} = R_{bb}\boldsymbol{B} + \sqrt{1 - R_{bb}^2}\boldsymbol{B}' \tag{5.22}$$

$$\boldsymbol{J}_{bb} = \Gamma \boldsymbol{J}_B + \sqrt{1 - \Gamma^2}\boldsymbol{J}_B' \tag{5.23}$$

$$\boldsymbol{J}_B \;=\; R_B\boldsymbol{B} + \sqrt{1-R_B^2}\,\boldsymbol{B}'' \;. \tag{5.24}$$

Inserting eq. 5.24 into eq. 5.23, one obtains

$$\boldsymbol{J}_{bb} = \Gamma\left[R_B\boldsymbol{B} + \sqrt{1-R_B^2}\,\boldsymbol{B}''\right] + \sqrt{1-\Gamma^2}\,\boldsymbol{J}'_B \;. \tag{5.25}$$

Comparing now eq. 5.25 with eq. 5.22, one concludes that

$$\left[R_{bb} - \Gamma R_B\right]\boldsymbol{B} = \Gamma\sqrt{1-R_B^2}\,\boldsymbol{B}'' + \sqrt{1-\Gamma^2}\,\boldsymbol{J}'_B - \sqrt{1-R_{bb}^2}\,\boldsymbol{B}' \;. \tag{5.26}$$

Up to now the above equation is of course valid for any three vectors. According to the result 5.20, however, the l.h.s. of eq. 5.26 is zero, which leads to a vector equation relating the components of $\boldsymbol{J}_{bb}$ and $\boldsymbol{J}_B$ which are orthogonal to $\boldsymbol{J}_B$ and $\boldsymbol{B}$. Squaring the $\boldsymbol{B}'$ term and using again eq. 5.20, one obtains $\boldsymbol{B}'' \cdot \boldsymbol{J}'_B = 0$, or, in terms of the original vectors,

$$\underbrace{(\boldsymbol{J}_{bb} - \Gamma\boldsymbol{J}_B)}_{\text{orthogonal to } \boldsymbol{J}_B} \;\cdot\; \underbrace{(\boldsymbol{J}_B - R_B\boldsymbol{B})}_{\text{orthogonal to } \boldsymbol{B}} = 0 \;. \tag{5.27}$$

A deeper interpretation of this result, if there is any, is still lacking.

5.5 Simulations

In this section some simulations are shown to confirm the theoretical predictions about the values of R_G, R_B, R_{bb} and Γ. The simulations are precisely the same as those described on page 62. The only difference is that for each center of mass which was built, the overlaps R_B, R_{bb} and Γ were measured, as well as R_G (whose statistics are therefore more refined, with n times more samples).

The replica calculation predicts that each of those overlaps will have a well defined value in the thermodynamic limit $N \to \infty$. For finite systems, one expects them to be more or less peaked, with a variance which will scale with N^{-1}. In order to give an idea of the behavior for a system size with $N = 500$, Fig. 5.7 shows the histograms obtained at $\alpha = 1.17$. One can see that the histograms are relatively broad, but still well centered around the theoretical values. Also to be noticed is the much smaller variance of Γ, which can be explained by the same reasoning employed in page 68: $\boldsymbol{J}_B$ is typically much closer to the corners of the hypercube than the average spherical vector, so not only Γ is large, but also it cannot take that many different values.

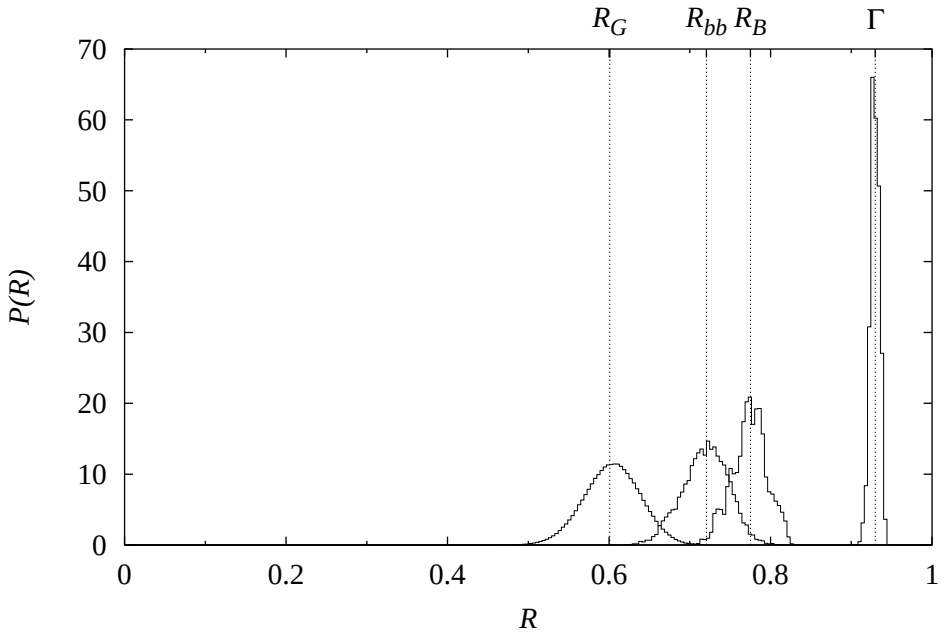

Figure 5.7: Histograms of the overlaps (from left to right) R_G, R_{bb}, R_B and Γ for $\alpha = 1.17$. The vertical lines show the theoretical predictions.

Finally, putting together figs. 5.5 and 5.6 and simulations for several values of α, one concludes from fig. 5.8 that the agreement is excellent. Note in particular that the Γ curve falls within one standard deviation of the simulated points, even though its distribution is very sharp.

5.6 Discussion

A comparison between eqs. 4.18 and 5.6 shows that the effect of the disorder is drastic. In the first case (a geometric calculation), the center of mass of Ising vectors is predicted to be an Ising vector as well, while in the second (disordered) case the same quantity is shown to have continuous components. The present section is intended to point out that the discrepancy between the two results is far from obvious.

In order to do so, results for the same calculations are presented in the following case: one assumes that $\boldsymbol{B}$ is *binary*, but the vectors $\boldsymbol{J}$ are nonetheless allowed to have continuous components, being sampled from a *spherically symmetric distribution*, $P(\boldsymbol{J}) = P_s(\boldsymbol{J}) \sim \delta(\boldsymbol{J} \cdot \boldsymbol{J} - N)$. In this case, the ex-

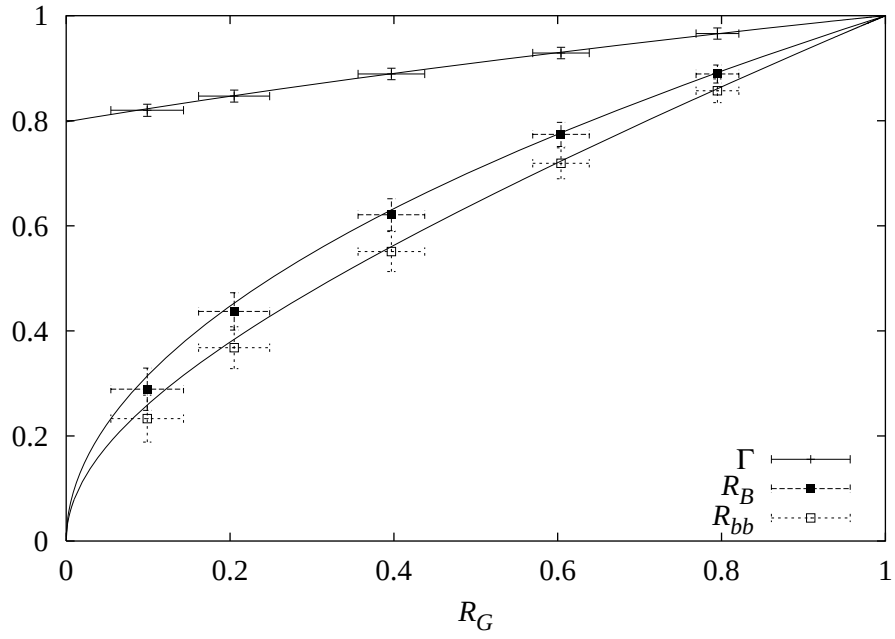

Figure 5.8: Comparison between simulations and theory, for different values of α. Error bars represent one standard deviation.

pression for $P(y)$ based solely on a geometric viewpoint is obtained by substituting $P_s(\boldsymbol{J}^a)$ for $P_b(\boldsymbol{J}^a)$ in eq. 4.4. By performing this calculation [Bou99], one obtains

$$P(x) = \frac{1}{\sqrt{2\pi(1-R_{CM}^2)}} \exp\left[\frac{-(x-R_{CM})^2}{2(1-R_{CM}^2)}\right] , \tag{5.28}$$

$$R_{CM} = \frac{nR}{\sqrt{n+n(n-1)q}} \tag{5.29}$$

a result which holds $\forall n$. It should also be stressed that eq. 5.28 is valid *only* if $B_j \in \{-1,+1\}$. Note in particular that in the limit $n \to \infty$ one obtains $R_{CM} \to R/\sqrt{q}$.

If one performs the replica calculation of the quenched moments instead, replacing now $P(\boldsymbol{J})$ by $P_s(\boldsymbol{J})$ in eq. 5.2, the result is (see section E.3)

$$P(x) = \frac{1}{\sqrt{2\pi(1-R^2/q)}} \exp\left[\frac{-(x-R/\sqrt{q})^2}{2(1-R^2/q)}\right] , \tag{5.30}$$

where once more $\boldsymbol{B}$ must be Ising. Comparing eqs. 5.28 and 5.30, it turns out that *the results are identical when the limit $n \to \infty$ is taken in eq. 5.29.*

One therefore verifies that the effect of the disorder is completely different in the two scenarios. For spherically distributed $\boldsymbol{J}$ vectors, a purely geometric calculation gives the *correct* result in the limit[5] $n \to \infty$, a fact which in hindsight could have been used to justify the geometric calculation of chapter 4 for $\boldsymbol{J} \in \{-1, +1\}^N$.

[5]Note however that spherical symmetry alone cannot account for this coincidence, since $\boldsymbol{B}$ must be Ising for both eq. 5.28 and 5.30 to hold.

Chapter 6

Attempts to construct an optimal cost function

6.1 Introduction

While chapter 3 describes how $\boldsymbol{J}_{bb}$ is formally defined (see eqs. 3.11 and 3.5), chapter 5 presents the results concerning its performance. The present chapter aims at tentatively answering the question that naturally follows the previous ones: is it possible to obtain $\boldsymbol{J}_{bb}$ by making use of some optimized potential?

Optimal potentials have been broadly explored in neural network problems with continuous vectors. In both the on-line [KC92, KC93, CC95, BRS95, KC95, CKC96, SC96, CEK+97, Cop97, VC97, SR98, VKC98, CC98] and off-line [KC96, VdBR96, BTMG97, BG98] scenarios, it has proven useful in providing upper bounds — when they did not exist — and saturating upper bounds — when they did. They have also given rise to insights on the implementation of efficient algorithms by automatically providing the relevant variables and features which enhance their performance. And last but not least they have theoretical significance *per se*, helping to understand the connections between the different approaches from Statistical Mechanics and Statistics to learning theory.

However, much less has been studied about optimal potentials in a discrete space [dM97]. The upper bound obtained in chapter 5 provides a convenient background to which the results of attempted optimizations can be compared. Two main paths arise: the first one, discussed in section 6.2, consists in trying to construct an optimal potential V_{opt} in the discrete space itself; the second one, discussed in section 6.3, consists in initially obtaining a potential V_{opt}^{sphere} which acts on the vectors lying on the hypersphere, and

finally clipping them.

6.2 The *ansatz* of a unique minimum

Following the discussion of section 6.1, one would like to obtain a potential V_{opt} such that the resulting overlap R_{opt} would saturate the upper bound R_{bb} given by eq. 5.12. In what follows, this search will be carried out in the limit of *zero temperature*, assuming that the potential has a *unique minimum*. The zero temperature limit ensures that the resulting vector is the ground state of the potential.

This rather strong (and dramatic, as shall be seen) assumption can be justified on two different basis. First, it is technically much simpler to obtain optimal potentials when the limits $\beta \to \infty$ and $q \to 1$ are simultaneously taken. Second, the results of chapter 3 can be used as a motivation: the Bayesian vector is the center of mass of the Gibbs ensemble, being therefore *unique* for a given set of examples. Its clipped counterpart $\boldsymbol{J}_{bb}$ is thus also a unique vector, motivating the $q \to 1$ *ansatz*.

The analysis of potentials with a unique minimum can be made by properly rescaling the order parameters and respective conjugate variables. In appendix F this is done in detail, only the results are shown here. In order to keep the free energy (eq. 1.17) finite (i.e. $\mathcal{O}(\beta^0)$) in the limit $\beta \to \infty$, the following variables should remain finite[1]:

$$\begin{aligned} \eta &\equiv \beta(1-q) \\ \hat{\eta} &\equiv \frac{\hat{q}}{\beta^2} \\ \hat{y} &\equiv \frac{\hat{R}}{\beta} \,. \end{aligned} \tag{6.1}$$

It is particularly important to check that the variable η remains finite and positive in equilibrium, otherwise the hypothesis that $q \to 1$ is violated. In terms of the new variables, the free energy reads, to leading order,

$$\begin{aligned} f \;=\; & \underset{R,\eta,\hat{\eta},\hat{y}}{\text{Extr}} \left\{ \frac{\eta\hat{\eta}}{2} + \hat{y}R - 2\sqrt{\hat{\eta}} P_n\left(\hat{y}/\sqrt{\hat{\eta}}\right) - \hat{y}\left[1 - 2H\left(\hat{y}/\sqrt{\hat{\eta}}\right)\right] \right. \\ & \left. + \alpha \int \mathcal{D}^* b \int \mathcal{D}t_2 \; \min_{\lambda} \left[V(\lambda) + \frac{(\lambda - t)^2}{2\eta} \right] \right\} , \end{aligned} \tag{6.2}$$

[1] Except for borderline cases, as already noted in footnote 5 on page 51.

where $P_n(x) = (2\pi)^{-1/2}\exp(-x^2/2)$ is a Gaussian and

$$t = t(b, t_2; R) \equiv bR + t_2\sqrt{1-R^2}\ . \tag{6.3}$$

The extremum operation of eq. 6.2 yields the following saddle point equations for the order parameters (see appendix F):

$$\begin{aligned} R &= 1 - 2H\left(\frac{\hat{y}}{\sqrt{\hat{\eta}}}\right) \\ \eta &= \frac{2}{\sqrt{\hat{\eta}}}P_n\left(\frac{\hat{y}}{\sqrt{\hat{\eta}}}\right) \\ \hat{\eta} &= \frac{\alpha}{\eta^2}\int \mathcal{D}t\, X(t;R)\left[\lambda_0(t,\eta) - t\right]^2 \\ \hat{y} &= \frac{\alpha}{\eta}\int \mathcal{D}t\, Y(t;R)\left[\lambda_0(t,\eta) - t\right]\ , \end{aligned} \tag{6.4}$$

where X and Y are as defined in eqs. 2.8,

$$\lambda_0(t,\eta) = \underset{\lambda}{\text{Argmin}}\left[V(\lambda) + \frac{(\lambda - t)^2}{2\eta}\right] \tag{6.5}$$

and one is implicitly assuming

$$V''\left(\lambda_0(t,\eta)\right) + \frac{1}{\eta} > 0\ , \qquad \forall t\ . \tag{6.6}$$

Given a potential $V(\lambda)$, one just has to calculate $\lambda_0(t,\eta)$ and insert the result back into eqs. 6.4. This procedure is exactly the same as the one which was introduced by Bouten *et al* [BSVdB95] for the supervised problem with a spherical prior, except that in the present case the conjugate variables cannot be eliminated algebraically.

The fact that the whole calculation was performed without any previous assumption about the form of the potential V, should not be underemphasized. After the function λ_0 is calculated, one should carefully check that the side condition 6.6 is satisfied and that η has a finite solution. Failure to meet these conditions could be the result of a major difficulty: the potential could have a degenerate minimum, implying that the RS *ansatz* is unstable. The necessity of these a posteriori checks seems to be the price paid for the simplicity of eqs. 6.4 and 6.5, through which the final performance $R(\alpha)$ of a given potential can be obtained with little algebra.

Interestingly, the extremum in eq. 6.2 can be applied to the variables η, $\hat{\eta}$ and $\hat{y}$ first. If the remaining minimization with respect to R is left to the last

stage, one can readily verify that the thermodynamically stable free energy can be simply written as

$$f = \min_R \alpha \int \mathcal{D}t\, X(t;R)\, V\left(\lambda_0(t,\eta)\right) \ , \tag{6.7}$$

an expression which has been shown to be formally identical in the case of spherical vectors [BG98].

6.2.1 Clipped Hebbian supervised learning

Clipped supervised learning is a problem which has been studied before in the literature [VdBB93, GM93, BS95] and the particular case of the Hebb rule can be used here as an example to check the general approach described in section 6.2.

The problem of supervised learning can be once again mapped to the present framework of unsupervised learning by using the *aligned* patterns described in section 1.3. The noiseless case to be studied has therefore $P(b) = 2\Theta(b)P_n(b)$, according to eq. 1.25, where $\Theta(x)$ is the Heaviside function.

Hebbian learning is defined as constructing the vector $\boldsymbol{J}_H \propto \sum_{\mu=1}^{\alpha N} \boldsymbol{\xi}^\mu$, where $\{\boldsymbol{\xi}^\mu\}$ are the aligned patterns, and its clipped version $\boldsymbol{J}_{cH}$ is obtained by taking the sign of each component of $\boldsymbol{J}_H$ (see page 38),

$$\boldsymbol{J}_{cH} \equiv \mathrm{clip}(\boldsymbol{J}_H) \ . \tag{6.8}$$

Because of the very definition of $\boldsymbol{J}_{cH}$, it obviously maximizes the dot product $\boldsymbol{J} \cdot \boldsymbol{J}_H \propto \boldsymbol{J} \cdot \sum_{\mu=1}^{\alpha N} \boldsymbol{\xi}^\mu \propto \sum_{\mu=1}^{\alpha N} \lambda_\mu$ in the binary space. Correspondingly, $\boldsymbol{J}_{cH}$ should be obtained — using the zero temperature calculation streamlined in section 6.2 — as the ground state of the potential

$$V(\lambda) = -d\ \lambda \ , \tag{6.9}$$

where $d > 0$ is a constant. Inserting eq. 6.9 into eq. 6.5, one obtains

$$\lambda_0(t,\eta) = d\eta + t \ . \tag{6.10}$$

Remembering that $X(t;R)$ for supervised learning is as in section 2.3 (see page 24) and $Y(t;R) = R^{-1}\partial X(t;R)/\partial t$, one inserts eq. 6.10 into the saddle point equations 6.4 to obtain

$$\begin{aligned} \hat{\eta}(\alpha) &= d^2\alpha \ , \\ \hat{y}(\alpha) &= d\,\alpha\sqrt{\frac{2}{\pi}} \ , \end{aligned}$$

$$\begin{aligned} \eta(\alpha) &= \frac{1}{d}\sqrt{\frac{2}{\alpha\pi}}\exp\left(\frac{-\alpha}{\pi}\right) \\ R(\alpha) &= 1 - 2H\left(\sqrt{\frac{2\alpha}{\pi}}\right) . \end{aligned} \tag{6.11}$$

Asymptotics

The result $R(\alpha)$ in eq. 6.11 is identical to the one obtained in [VdBB93], as it should. The asymptotics for small and large α are, respectively,

$$R \overset{\alpha\to 0}{\simeq} \frac{2}{\pi}\sqrt{\alpha} \tag{6.12}$$

$$R \overset{\alpha\to\infty}{\simeq} 1 - \frac{1}{\sqrt{\alpha}}\exp\left(\frac{-\alpha}{\pi}\right) . \tag{6.13}$$

Using the asymptotic results of Gibbs learning (section 2.3) together with the results for the best binary (see eq. 5.17), a comparison with eq. 6.12 shows that *clipped Hebbian learning is able to asymptotically saturate the R_{bb} bound for $\alpha \to 0$*. For $\alpha \to \infty$, the exponential behavior of eq. 6.13 shows a major improvement when compared with simple Hebbian learning, which behaves asymptotically as [Val89] $R \simeq 1 - \pi/(4\alpha)$. However, it obviously fails to saturate the R_{bb} bound for $\alpha \to \infty$, since it misses the first order transition at $\alpha = 1.245$. These comparisons can be seen in fig. 6.1.

Discussion

The performance $R(\alpha)$ for clipped Hebbian learning can also be obtained in a different way. Using the result $R(\alpha) = (1 + \pi/(2\alpha))^{-1/2}$ for Hebbian learning [Val89], one just has to note that the resulting vector is a continuous one, and uniformly distributed on the cone $R = \boldsymbol{J}\cdot\boldsymbol{B}/N$. Making use of the general formalism for clipping [SBVdB95] described in section 3.3.1, one inserts Vallet's result into eq. 3.18 and the result is identical to eq. 6.11. Another possible approach, based on a signal-to-noise analysis, is developed in [BS95], leading to the same results.

Despite the redundancy of the results, however, eqs. 6.11 remain interesting as a check of the validity of the general results of section 6.2. First, one notices that the rescaled variables do remain finite, as originally requested. In particular $\eta = \beta(1-q)$ is finite, giving consistency to the *ansatz* of a unique minimum[2]. Second, the side condition 6.6 reads, in this case, $1/\eta > 0$, which

[2]On page 227 of [VdBB93], a statement is made about an order parameter whose

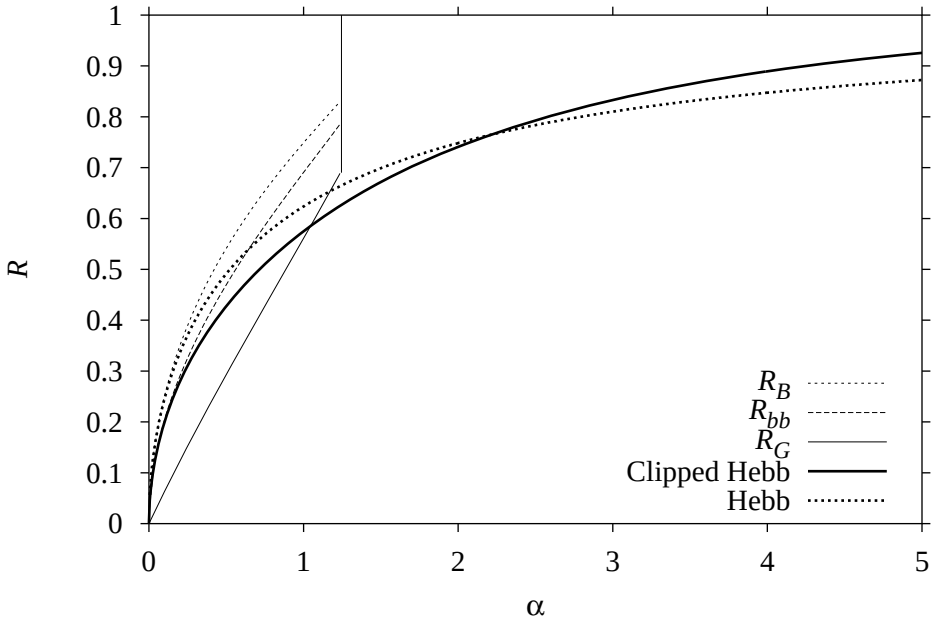

Figure 6.1: Overlaps as functions of α: R_B, R_{bb}, R_G (only the thermodynamically stable solutions are shown), clipped Hebb according to eq. 6.11 and Hebbian learning according to [Val89].

is also clearly satisfied. Therefore clipped Hebbian learning seems to be a case where eqs. 6.4-6.6 properly describe the system.

6.2.2 Minimal/Maximal Variance

The next test case to be studied is the Minimal/Maximal Variance (Min/Max) cost functions in the unbiased Gaussian scenario described in section 2.4. These cost functions are of the form

$$V(\lambda) = \frac{c}{2}\lambda^2 \,, \tag{6.14}$$

where c can be either positive (Minimal Variance) or negative (Maximal Variance). The cases of interest are those which have $\mathrm{sign}(c) = \mathrm{sign}(A)$,

meaning is "less clear". By comparing the equation obeyed by that order parameter with eqs. 6.11, one concludes that the mysterious quantity is proportional to η. This is consistent with the findings of [BS95], where η is shown to account for the signal-to-noise ratio in the field λ.

that is, one minimizes (maximizes) $\sum_{\mu}^{\alpha N} \lambda_{\mu}^2$ in $\boldsymbol{J}$ space when $\langle b^2 \rangle_*$ is indeed minimal (maximal) as compared to the variance of the projection of the examples $\boldsymbol{\xi}$ along the directions perpendicular to $\boldsymbol{B}$.

Following the streamlined procedure of section 6.2, one inserts the potential 6.14 into eq. 6.5 to obtain

$$\lambda_0(t,\eta) = \frac{t}{1+c\eta} \; . \tag{6.15}$$

Recalling the expressions 2.30 for the Gaussian scenario and combining eq. 6.15 with the saddle point equations 6.4, one obtains, for $B = 0$,

$$R = 1 - 2H\left(\frac{\sqrt{\alpha}|A|R}{\sqrt{1-AR^2}}\,\mathrm{sign}(1+c\eta)\right) \tag{6.16}$$

$$\frac{\eta}{|1+c\eta|} = \frac{2}{\sqrt{\alpha(1-AR^2)}} P_n\left(\frac{\sqrt{\alpha}|A|R}{\sqrt{1-AR^2}}\,\mathrm{sign}(1+c\eta)\right) \; . \tag{6.17}$$

The first equation accounts for the performance of the cost function, while the second allows to check whether η remains finite. Note that the argument $1 + c\eta = \eta(V'' + 1/\eta)$ in the above equations should be positive as a consequence of the side condition eq. 6.6. Therefore it can be handled like this: one assumes that $1 + c\eta$ is positive and then checks the consistency of the assumption.

Asymptotics

An asymptotic expansion of eq. 6.16 yields the following behavior: in the large α regime the performance is better than that of Gibbs learning (compare with eq. 2.35), *saturating the R_{bb} bound* (compare with eq. 5.15):

$$1 - R \overset{\alpha\to\infty}{\simeq} \sqrt{\frac{2(1-A)}{\pi\alpha A^2}} \exp\left[-\frac{\alpha}{2}\left(\frac{A^2}{1-A}\right)\right] \; . \tag{6.18}$$

Things are rather different, however, when the learning process has just started. The asymptotics for small R, assuming a smooth behavior for $R(\alpha)$, gives

$$R \begin{cases} = 0, & \alpha \le \alpha_m \\ \overset{R\to 0}{\simeq} \sqrt{C(\alpha-\alpha_m)}, & \alpha \ge \alpha_m \end{cases} \tag{6.19}$$

where C is a constant and the value of the critical load α_m is now

$$\alpha_m \equiv \frac{\pi}{2A^2} \tag{6.20}$$

One should note that the $\sim (\alpha - \alpha_m)^{1/2}$ increase in eq. 6.19 is much steeper than the linear one obtained for Gibbs learning (compare with eq. 2.19 or 2.37). However, this square root behavior — comparable in shape to both R_B and R_{bb} — occurs only after the presentation of a critical number of examples $N\alpha_m$ *which is larger than that of Gibbs learning* (see eq. 2.38). Eqs. 6.19 and 6.20 would imply that the same potential V (quadratic in λ) can lead to a better performance at finite temperature (Gibbs learning) than at zero temperature (at least for $\alpha_G \leq \alpha \leq \alpha_m$).

The intermediate non-asymptotic behavior can be seen in fig. 6.2, where a solution of eq. 6.16 can be compared to the performance of the best binary vector.

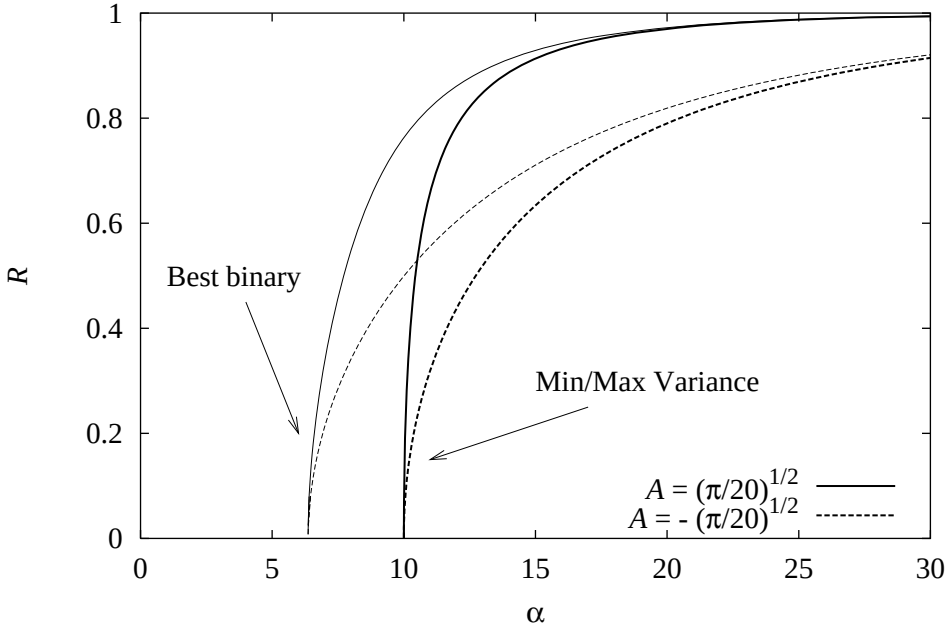

Figure 6.2: Solution of eq. 6.16 for two values of A (thick lines) and the R_{bb} bound (thin lines) based on the solution of eqs. 2.5, 2.31 and 5.12.

Consistency checks

One should check whether η has a finite solution for both Minimal and Maximal Variance cost functions. Starting with Maximal Variance ($c = -|c|$ and $A = -|A|$), one would like to verify an even more restrictive condition, namely $\eta < 1/|c| \Rightarrow \text{sign}(1 + c\eta) > 0$. Rewriting the saddle point

equation 6.17 for η in this special case, one gets

$$\frac{\eta}{1-|c|\eta} = \frac{2}{\sqrt{\alpha(1+|A|R^2)}} P_n\left(\frac{\sqrt{\alpha}|A|R}{\sqrt{1+|A|R^2}}\right) . \tag{6.21}$$

Since the function $\frac{\eta}{1-|c|\eta}$ maps the interval $[0, 1/|c|[$ into $[0, \infty[$ and the r.h.s. of eq. 6.21 is finite for all A, R and $\alpha \neq 0$, there is always a finite solution satisfying $\eta < 1/|c|$.

For Minimal Variance ($c = +|c|$ and $A = +|A|$), the analogous of eq. 6.21 is

$$\frac{\eta}{1+|c|\eta} = \frac{2}{\sqrt{\alpha(1-|A|R^2)}} P_n\left(\frac{\sqrt{\alpha}|A|R}{\sqrt{1-|A|R^2}}\right) . \tag{6.22}$$

Now one needs to check that a solution exists in the interval $0 < \eta < \infty$, which automatically satisfies the side condition $\eta > -1/|c|$. This is not so simple as in the previous case, since the function $\frac{\eta}{1+|c|\eta}$ now maps the interval $[0, \infty[$ into $[0, 1/|c|[$. Using this fact, one can always find a value $|c^*|$, for any A, R and α (which fix the r.h.s. of eq. 6.22), such that no satisfactory solution for η exists for $|c| > |c^*|$.

While Maximal Variance seems to pass the consistency checks, Minimal Variance does not. Even though one could argue that a sufficiently well-tuned value of c might lead to a satisfactory solution, the sheer fact that the solutions depend on the absolute value of c *at all* should rise suspicions in the first place[3]. After all, if a potential V is minimized for a given vector, so is a potential $\tilde{V} = \text{constant} \times V$. This kind of contradiction is a sign of problems with the equations developed in section 6.2 and a more thorough discussion will be carried out in section 6.2.4.

6.2.3 Variationally optimal potentials

Despite the problems found in section 6.2.2 while testing the general eqs. 6.4, an attempt to construct an optimal cost function in this framework will be made in this section. Optimality is defined here as maximizing the overlap R for a given value of α. By doing so from the equations of section 6.2, one is trying to find the best possible binary vector which, as shown in chapter 5, leads to the R_{bb} overlap described by eq. 5.12. One is thereby implicitly assuming that this can be done with a potential which has a single minimum. Whether this is possible or not is by no means obvious, in principle. But one

[3]Note that this is not the case of clipped Hebbian learning. Eqs. 6.11 show that $\eta > 0$ for whatever value of $d > 0$.

can always construct a variationally optimal potential V_{opt} and then check whether the corresponding $R_{opt}(\alpha)$ saturates the bound $R_{bb}(\alpha)$.

The optimization procedure to be described is inspired by Kinouchi and Caticha's work [KC96] on optimal supervised learning in the spherical perceptron (see also [BTMG97]), which was soon after extended to an unsupervised scenario [VdBR96, BG98]. One assumes that V is unknown and then variationally determines V_{opt} such that R is maximized for given α. From eqs. 6.4 it is clear that the equilibrium value of R is a monotonically increasing function of $\hat{y}^2/\hat{\eta}$, which on its turn is determined by the up to now unknown potential V. Defining

$$F(t,\eta) \equiv \lambda_0(t,\eta) - t = -\eta \left.\frac{\partial V}{\partial \lambda}\right|_{\lambda=\lambda_0(t,\eta)} , \tag{6.23}$$

one has

$$\frac{\hat{y}^2}{\hat{\eta}} = \alpha \frac{\left[\int \mathcal{D}t\, Y(t;R)\, F(t,\eta)\right]^2}{\int \mathcal{D}t\, X(t;R)\, F^2(t,\eta)} = \alpha \frac{\left[\int \mathcal{D}t\, \left(\frac{Y^2}{X}\right) \frac{FX}{Y}\right]^2}{\int \mathcal{D}t\, \left(\frac{Y^2}{X}\right) \left(\frac{FX}{Y}\right)^2} . \tag{6.24}$$

The variational optimization takes place at this point, where the Schwarz inequality can be invoked to show[4] that the r.h.s. of the above equation is maximized if $F(t,\eta)X(t;R)/Y(t;R) = k$, where k is any t-independent function. This determines the optimal function F_{opt}, namely,

$$F_{opt}(t,\eta;R) = k(R,\eta)\frac{Y(t;R)}{X(t;R)} . \tag{6.25}$$

The constant k has yet to be determined, but it is interesting to note that its value is irrelevant for the determination of R:

$$\left.\frac{\hat{y}^2}{\hat{\eta}}\right|_{F=F_{opt}} = \alpha \int \mathcal{D}t\, \frac{Y^2(t;R)}{X(t;R)} . \tag{6.26}$$

This equation can be inserted back into 6.4 and yields the final equation for the overlap,

$$R_{opt} = 1 - 2H\left(\sqrt{\alpha \int \mathcal{D}t\, \frac{Y^2(t;R_{opt})}{X(t;R_{opt})}}\right) . \tag{6.27}$$

[4] Defining the internal product between two functions $x_1(t)$ and $x_2(t)$ as $\langle x_1, x_2\rangle \equiv \int \mathcal{D}t\, \frac{Y^2(t;R)}{X(t;R)}\, x_1(t)\, x_2(t)$, one just has to recall the Schwarz inequality $|\langle x_1, x_2\rangle|^2 \leq \langle x_1, x_1\rangle \langle x_2, x_2\rangle$. The function X is strictly positive, according to its definition in eq. 2.8.

Eq. 6.27 is the main result of this section, representing the best performance that can be achieved by a potential with a unique minimum.

There are however remaining questions, in order to explicitly construct the optimal potential V_{opt}. These are briefly addressed here in order to support the discussion of section 6.2.4.

One should start by determining the constant $k(R,\eta)$, which is easily obtained by inserting eq. 6.25 into the saddle point equations for η and $\hat{\eta}$. This yields a self-consistent equation for $k(R,\eta) = k(R_{opt},\eta)$ while leaving η itself undetermined (a fact which occurs whenever $\lambda_0(t,\eta)$ does not depend on η):

$$k(R_{opt}) = \frac{2}{\sqrt{\alpha \int \mathcal{D}t \, \frac{Y^2(t;R_{opt})}{X(t;R_{opt})}}} P_n \left(\sqrt{\alpha \int \mathcal{D}t \, \frac{Y^2(t;R_{opt})}{X(t;R_{opt})}} \right) , \tag{6.28}$$

where $R_{opt} = R_{opt}(\alpha)$ is the solution of eq. 6.27. Note that k does not depend on η.

The second step to obtain V_{opt} is the integration of eq. 6.23 for $F = F_{opt}$. The procedure in this case is identical to the one presented in [KC96] and the details will be omitted. The result is

$$\begin{aligned} V_{opt}(\lambda;R_{opt}) &= \frac{1}{\eta} \Bigg\{ E_{opt} \left(\lambda_{opt}^{(-1)}(\lambda;R_{opt});R_{opt} \right) \\ &\quad - \frac{1}{2} F_{opt}^2 \left(\lambda_{opt}^{(-1)}(\lambda;R_{opt});R_{opt} \right) \Bigg\} , \end{aligned} \tag{6.29}$$

where

$$\begin{aligned} E_{opt}(t;R_{opt}) &\equiv -\frac{k(R_{opt})}{R_{opt}} \ln X(t;R_{opt}) \\ F_{opt}(t;R_{opt}) &\equiv -\frac{\partial}{\partial t} E_{opt}(t;R_{opt}) \\ \lambda_{opt}(t;R_{opt}) &\equiv F_{opt}(t;R_{opt}) + t \end{aligned} \tag{6.30}$$

and $\lambda_{opt}^{(-1)}$ is the inverse function of λ_{opt} (assuming the inverse exists).

Performance of the variationally optimal potential

The main issue at hand is whether $R_{opt}(\alpha)$, which is the solution of eq. 6.27, saturates or not the best binary bound $R_{bb}(\alpha)$. One should start by noting

the very similar form of the respective equations. From the definition 2.7 of $\mathcal{F}$, and recalling eqs. 5.13, 3.10 and 2.5, one rewrites

$$R_{opt} = 1 - 2H\left(\mathcal{F}(R_{opt})\right) \tag{6.31}$$
$$R_{bb} = 1 - 2H\left(\mathcal{F}(R_B)\right) \tag{6.32}$$
$$R_B = F_B\left(\mathcal{F}(R_B)\right) , \tag{6.33}$$

where the potential-independent F_B is defined in eq. 2.6. However, one should not be misled by the similarity of eqs. 6.31 and 6.32: eq. 6.31 should be *solved*, while eq. 6.32 just *maps the solution* of eq. 6.33, which should be solved as well. Keeping this picture in mind, one can try to compare $R_{opt}(\alpha)$ with $R_{bb}(\alpha)$.

$F_B(x)$ and $1 - 2H(x)$ are both monotonically increasing functions which do not depend on the potential, distribution of patterns, etc. They are determined only by the binary nature of the $\boldsymbol{J}$ space. One can check on fig. 6.3 that $F_B(x) \geq 1 - 2H(x)$, $\forall x$. Also, $F_B(x) \geq F_s(x) \equiv x/\sqrt{1+x^2}$, a relation that will be useful in section 6.3 for comparison with potentials on the hypersphere.

$\mathcal{F}(R)$ is the quantity that carries the information about the distribution of patterns. After a long but straightforward calculation (see section F.2), one can show that $\mathcal{F}$ is a monotonically increasing function of R as well, $\partial\mathcal{F}/\partial R \geq 0$. If one does not take into account the possibility of multiple solutions to eqs. 6.31 and 6.33 (first order phase transitions), the conclusion is that $R_B(\alpha) \geq R_{opt}(\alpha)$. Of course, the absence of multiple solutions cannot be guaranteed (much on the contrary, they can occur both for the binary case, as seen in chapter 2, as for continuous problems [BG98]). But as long as the solution is unique, for given α, one has[5] $R_B(\alpha) \geq R_{opt}(\alpha)$. Inserting this inequality on the r.h.s. of eqs. 6.31 and 6.32, one concludes that $R_{bb} \geq R_{opt}$. *Therefore R_{opt} does not saturate the bound R_{bb}, in general.*

A simple counterexample

There is a simple exception to the general failure of R_{opt} in saturating R_{bb}. If one studies the biased case (with $A = 0$) of the Gaussian scenario, eq. 2.31 yields

$$\mathcal{F}(R) = |B|\sqrt{\alpha} . \tag{6.34}$$

[5] $R_B(\alpha) \geq R_{opt}(\alpha)$ is a reasonable result, since the l.h.s. is attained by a continuous vector, while the r.h.s. corresponds to the performance of a binary vector.

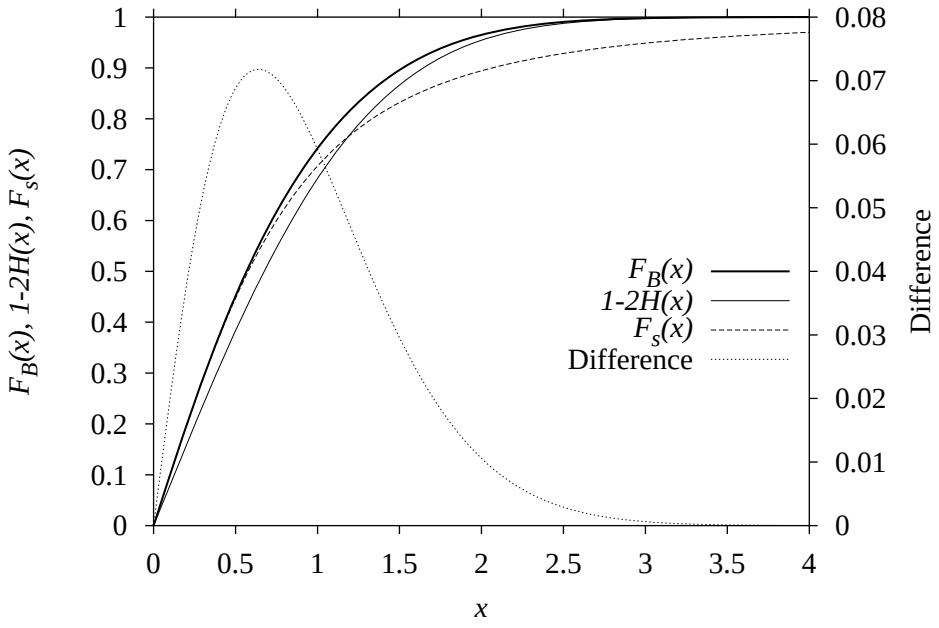

Figure 6.3: $F_B(x)$, as defined in eq. 2.6, $1-2H(x)$ and $F_s(x) = x/\sqrt{1+x^2}$ are plotted on the left axis; the difference $F_B - (1-2H(x))$ is plotted on the right axis.

Therefore $\mathcal{F}(R)$ does not depend on R and the arguments given above to justify the failure do not hold. Inserting eq. 6.34 into eqs. 6.31 and 6.32, one can easily verify that $R_{opt}(\alpha) = R_{bb}(\alpha) = 1-2H\,(|B|\sqrt{\alpha})$. This was the only case which could be found where the optimization procedure was successful. A justification will be given in section 6.2.4.

Asymptotics

Going back to the general case, the asymptotics of R_{opt} is interesting. For large α, one obtains

$$1 - R_{opt}(\alpha) \overset{\alpha\to\infty}{\simeq} \sqrt{\frac{2}{\pi\alpha\,\langle (U')^2\rangle_*}}\exp\left(\frac{-\alpha\,\langle (U')^2\rangle_*}{2}\right)\,, \qquad (6.35)$$

that is, R_{opt} manages to saturate R_{bb} only *asymptotically* (see eqs. 2.15 and 5.15).

For small values of R, one assumes a smooth behavior for $R_{opt}(\alpha)$ and the result is

$$\langle b \rangle_* \neq 0 \quad \Rightarrow \quad R_{opt} \simeq |\langle b \rangle_*| \sqrt{\frac{2\alpha}{\pi}} \tag{6.36}$$

$$\langle b \rangle_* = 0 \quad \Rightarrow \quad R \begin{cases} = 0, & \alpha \leq \alpha_c \\ \simeq \sqrt{C(\alpha - \alpha_c)}, & \alpha \geq \alpha_c \end{cases} , \tag{6.37}$$

where C is again a constant and the critical value $\alpha = \alpha_c$ is given by (compare with the results of page 21)

$$\begin{aligned} \alpha_c &\equiv \frac{\pi}{2} \frac{1}{(1 - \langle b \rangle_*^2)^2} \\ &= \frac{\pi}{2} \alpha_G \ . \end{aligned} \tag{6.38}$$

The two regimes which have been obtained for Gibbs learning occur again here: if $\langle b \rangle_* \neq 0$, R_{opt} starts increasing as soon as α is non-zero, eq. 6.36. In this case R_{opt} is again able to *asymptotically* saturate the bound R_{bb}, according to eqs. 2.18 and 5.17. But if $\langle b \rangle_* = 0$, then the performance of V_{opt} is disastrous, and the second order transition occurs at a critical value α_c which is larger than that of Gibbs learning. This second picture generalizes the results which were obtained in section 6.2.2 for Min/Max cost functions. This is no coincidence: for the Gaussian scenario with $B = 0$, the general result eq. 6.27 becomes exactly the saddle point eq. 6.16 for Min/Max. In other words, Min/Max cost functions coincide with the optimal potentials for the Gaussian scenario with $B = 0$.

6.2.4 Discussion

The following questions arise: why does the variationally-obtained potential V_{opt} generally fail to saturate the best binary bound? And which are the characteristics the exceptional cases have that allow V_{opt} to succeed?

In order to answer these questions, one should keep track of the hypotheses assumed in all the derivations of section 6.2. First, replica symmetry is assumed. Second, the limits $\beta \to \infty$ and $q \to 1$ are simultaneously taken, with $\eta = \beta(1 - q)$ finite.

These hypotheses were imposed to ensure the uniqueness of the ground state of the potential. This is a desirable property, because of the argument given on page 76: the center of mass $\boldsymbol{J}_B$ of the Gibbs ensemble is a unique vector for a given set of examples, and so should be its clipped counterpart $\boldsymbol{J}_{bb}$. Even though the discussion on page 102 will argue that $q \to 1$ is not necessarily the *only* acceptable result, it is certainly the *simplest* one.

The explanation for the failure of V_{opt} to be sustained here is that *the limit $q \to 1$ — corresponding to the minimum of the potential being attained by a single vector — can hardly be self-consistently imposed for binary vectors.* Or, in other words, the *ansatz* $\beta \to \infty$, $q \to 1$ with η finite is generally not correct, except for some special cases.

These statements are supported by preliminary calculations of the entropy for the Gaussian scenario. Without reproducing them here, the general idea is as follows: *if one assumes that $q \to 1$, the only physically acceptable value of s is zero*, since there is only one state available to the system. Therefore, one should calculate the value of the entropy for $\beta \to \infty$ and check whether it vanishes identically, $\forall \alpha$. If it does not, then the *ansatz* $q \to 1$ is clearly wrong. Note that the discussion of this zero-entropy problem is more general than the problem of finding an optimal potential. The question is whether potentials with a unique minimum can be constructed *at all*.

Therefore the reason for the failure of V_{opt} in saturating the R_{bb} bound seems not to lie in the optimization procedure, but rather on the sheer difficulty of constructing potentials which have a unique minimum in the binary space.

Indeed, a quadratic potential $V(\lambda) = c\lambda^2/2 - d\lambda$ (where c and d are constants) for the Gaussian scenario seems to have an entropy different from zero at zero temperature, a result which would be clearly incompatible with the *ansatz* $q \to 1$. There is an exception, however, if $c = 0$. Then the entropy is exactly zero, showing that the linear potential satisfies the requirement of a unique minimum.

This discussion helps to clarify the apparently contradictory results of section 6.2.2. In the case of Minimal/Maximal Variance, the problem of non-zero entropy is clearly present. This invalidates most of the results (but not all of them, as discussed below) of that section, which were nonetheless presented for didactic reasons.

If one goes back to the results of clipped Hebbian supervised learning, on the other hand, the explanation for the success of the zero temperature approach is now clear: the potential is *linear*, which guarantees the existence of a unique minimum. In fact, the linear potential was chosen, in section 6.2.1, exactly because it leads to the desired unique binary vector, as can be seen in the discussion of page 78. Therefore $q \to 1$ is a consistent *ansatz* for all values of α, as opposed to the previously discussed example.

The counterexample of page 86 is another case where the results are consistent with the zero entropy criterion. The biased Gaussian scenario has, however, an important difference with respect to clipped Hebbian supervised learning: the results are not only consistent, but also optimal, saturating the R_{bb} bound for all α. The explanation is again very simple: in the Gaussian

scenario with $A = 0$, the pattern distribution is governed by $U(b) = -Bb$, i.e. a linear function. This, in turn, translates into a linear optimal potential $V_{opt}(\lambda) = -(k_{opt}B/\eta)\lambda$, according to eqs. 2.30, 6.25 and 6.23. And linear potentials pass the test of zero entropy, as discussed in the previous paragraphs.

The last apparent paradox to be cleared is the successful asymptotic result eq. 6.36, valid for V_{opt} in the limit $R_{opt} \to 0$ for $\langle b \rangle_* \neq 0$. In this case, one can show that, if $\langle b \rangle_* \neq 0$, then $V_{opt}(\lambda) \simeq -(k_{opt} \langle b \rangle_* / \eta)\lambda$ for $R_{opt} \to 0$. In other words, *the optimal potential is asymptotically linear* for $\alpha \to 0$, consistently satisfying the required condition $s \to 0$ in that limit. A similar reasoning might be employed at the other end of the asymptotics, namely for $\alpha \to \infty$. Even though a more rigorous result could not be obtained, eq. 6.35 shows that the R_{bb} bound is asymptotically saturated in that regime. The argument in this case is that $R \to 1$ necessarily implies the shrinking of the solution space, asymptotically leading to $s \to 0$ and giving consistency to the *ansatz* $q \to 1$.

In light of these discussions, one can conclude that *the optimal potential V_{opt} is able to saturate the R_{bb} bound whenever the condition $s = 0$ is met, either exactly or asymptotically.*

6.3 Optimal potentials in the hypersphere

The second major approach to the problem of finding $\boldsymbol{J}_{bb}$ consists in finding $\boldsymbol{J}_B$ instead, and then clipping it (see eq. 3.11). An advantage one can immediately foresee for such a strategy is the fact that the search could be performed in a *continuous* space, thereby avoiding the difficulties of constructing potentials in a discrete space.

This kind of approach has been very successful in providing good approximations of the maximally stable binary perceptron in the capacity problem (see [VR99] and references therein). The so-called "precursor strategy" aims at obtaining a good "precursor" vector which, on clipping, renders a binary vector whose stability is as close as possible to the bound obtained previously by Gardner and Derrida [GD88] (see also [KM89]). The similarity with the present problem is clear, but one significant difference remains: in the problem of unsupervised learning, one knows that a single well-defined vector $\boldsymbol{J}_B$ exists, such that clipping it leads to the best binary vector. In this sense, an "optimal precursor" exists which leads to the desired solution with probability 1 (in the TL). The question remains, of course, how to find it. This is not an easy task, despite the advantage of $\boldsymbol{J}_B$ being a continuous vector. In order to put the problem into perspective, it is natural to stablish a comparison

with a similar problem already studied in the literature: the Bayes-optimal vector for a spherical prior.

Suppose that, instead of a binary prior $P_b(\boldsymbol{B})$ for the preferential direction $\boldsymbol{B}$, one has $P(\boldsymbol{B}) = P_s(\boldsymbol{B}) \sim \delta(\boldsymbol{B}\cdot\boldsymbol{B} - N)$ (see eq. 1.4), where the subscript s stands for *spherical*. Similar replica calculations can be performed for a given pattern distribution and cost function [BM93, BM94, WN94, BG98, Buh99], but here only the general results will be presented. For Gibbs learning — defined once more as sampling from the posterior distribution — the equation to be solved is [VdBR96]:

$$\frac{R_G^s}{1-R_G^s} = \mathcal{F}^2\left(\sqrt{R_G^s}\right) \ , \tag{6.39}$$

where $\mathcal{F}$ is as in eq. 2.7 and $R_G^s = \boldsymbol{B}\cdot\boldsymbol{J}_G^s/N$ is the order parameter measuring the alignment of the real valued Gibbsian vector $\boldsymbol{J}_G^s$ with $\boldsymbol{B}$. One can now repeat the arguments presented in chapter 3: the center of mass $\boldsymbol{J}_B^s$ of the (spherical) Gibbs ensemble is Bayes-optimal, and its performance $R_B^s = \sqrt{R_G^s}$ satisfies, in light of eq. 6.39,

$$R_B^s = F_s\left(\mathcal{F}(R_B^s)\right) \ , \tag{6.40}$$

where $F_s(x) = x/\sqrt{1+x^2}$ was defined on page 86 and plotted in fig. 6.3.

The thermodynamically stable solution of eq. 6.40 is thus an upper bound to the performance of any properly normalized vector $\boldsymbol{J}$ in approximating $\boldsymbol{B}$, given the data and the prior distribution $P_s(\boldsymbol{B})$. A comparison with eq. 6.33 shows that the spherical prior renders a slower asymptotic decay than its binary counterpart, yielding a power law[6] $1 - R_B^s \simeq (2\alpha\left\langle (U')^2\right\rangle_*)^{-1}$ for large α. The asymptotics for small R, on the other hand, are identical to those obtained in chapter 2, a result that can be explained by the fact that $F_B(x) \simeq F_s(x)$ for small x.

If one now tries to construct an optimal potential V_{opt}^s whose minimum $\boldsymbol{J}_{opt}^s$ has an overlap $R_{opt}^s = R_B^s$ with $\boldsymbol{B}$, the results are very different from the ones obtained in section 6.2.3. It turns out [VdBR96] that R_{opt}^s satisfies

$$R_{opt}^s = F_s\left(\mathcal{F}(R_{opt}^s)\right) \ , \tag{6.41}$$

which is *identical* in form to eq. 6.40. If the solution of eqs. 6.40 or 6.41 is unique (as it is, for instance, in the problem of supervised learning, see [KC96, BTMG97]), then the bound is clearly saturated for all α. There are cases, however, where first order phase transitions appear due to the occurrence

[6] The result is valid for smooth distributions only. If U is discontinuous, then an α^{-2} behavior appears, see [VdBR96].

of multiple solutions. In these cases, the solution which minimizes the free energy must be chosen. Note that the analysis of eqs. 6.40 and 6.41 then splits, since the first is governed by the free energy of Gibbs learning, while the free energy of the second is obtained by the zero temperature formalism, cf. eq. 6.7. This is a difficult problem which is not addressed here. The reader is referred to [BG98, GB98], where a particular distribution leads to first order transitions for R_{opt} which are numerically shown to violate the Bayesian bound, an issue which seems to be up to now unsettled.

With these results in mind, one can go back to the original problem, namely, how to obtain $\boldsymbol{J}_B$ (as opposed to $\boldsymbol{J}_B^s$) using an optimal potential? In the calculations performed for the spherical prior, the only restriction imposed on the vector $\boldsymbol{J}$ is the spherical constraint $\boldsymbol{J} \cdot \boldsymbol{J} = N$. This is very convenient, since $\boldsymbol{J}_B$ must satisfy the same constraint. But the approach suffers from a major drawback: *if the only constraint imposed on* $\boldsymbol{J}$ *is the spherical one, the information about the binary nature of* $\boldsymbol{B}$ *is lost*. This can be more clearly seen on the discussion of page 117, where the free energy for a spherical $\boldsymbol{J}$ is shown to depend only on $\boldsymbol{B} \cdot \boldsymbol{B}$. In other words, the results for a spherical $\boldsymbol{J}$ are the same, regardless of the preferential direction $\boldsymbol{B}$ being spherical or binary. Therefore, by relaxing the binary constraint on $\boldsymbol{J}$ in order to facilitate the search of $\boldsymbol{J}_B$, one ends up "throwing away the baby with the bathwater". Recalling eqs. 6.41 and 6.33, the fact that $F_B(x) \geq F_s(x)$, $\forall x$, implies that such a strategy fails in rendering a vector with an overlap R_B with $\boldsymbol{B}$. In particular, the resulting $R_{opt}^s(\alpha)$ approaches 1 with a power law, while $1 - R_B \sim \exp(-\alpha)$.

6.3.1 Transforming $\boldsymbol{J}_{opt}^s$

Despite the fact that in general $R_{opt}^s \leq R_B$, one can nonetheless make use of the resulting $\boldsymbol{J}_{opt}^s$ to produce a binary vector $\boldsymbol{J}_{clip}^s$ by clipping. This would be an approximation to $\boldsymbol{J}_{bb}$. It is ironic that the possibility of directly applying the general clipping results of [SBVdB95] is granted in this case *exactly because the information about the binary nature of* $\boldsymbol{B}$ *has been lost*. This is because those results depend on the uniformity of $\boldsymbol{J}$ on the cone $\boldsymbol{J} \cdot \boldsymbol{B}/N$, according with the discussion of section 3.3.

More specifically, one can directly apply eq. 3.18 to the solution of eq. 6.41, the result being

$$R_{clip}^s \equiv 1 - 2H\left(\frac{R_{opt}^s}{\sqrt{1-(R_{opt}^s)^2}}\right)$$

$$\stackrel{(6.41)}{=} \quad 1 - 2H\left(\mathcal{F}(R^s_{opt})\right) \ . \tag{6.42}$$

The above equation should be compared to eq. 6.32. One concludes that, since $\mathcal{F}$ is monotonically increasing with R (and apart from first order phase transitions), then $R^s_{clip} \leq R_{bb}$.

The above inequality is not exactly surprising, since it is just a consequence of $R^s_{opt} \leq R_B$. Maybe of greater importance is the fact that, despite the qualitatively different behavior of R^s_{opt} and R_B, R^s_{clip} and R_{bb} do share some common features. This is more clearly depicted in the asymptotics of R^s_{clip}. For large α, one has

$$\begin{aligned} 1 - R^s_{clip} &\simeq \sqrt{\frac{2}{\pi}} \frac{\sqrt{1-(R^s_{opt})^2}}{R^s_{opt}} \exp\left[-\frac{R^s_{opt}}{2(1-(R^s_{opt})^2)}\right] \\ &\simeq \sqrt{\frac{2}{\pi\,\alpha\,\langle (U')^2 \rangle_*}} \exp\left[\frac{-\alpha\,\langle (U')^2\rangle_*}{2}\right] \ , \end{aligned} \tag{6.43}$$

assuming $\langle (U')^2 \rangle_*$ is finite. Recalling eq. 5.14, one notices that R^s_{clip} asymptotically saturates the R_{bb} bound in the $\alpha \to \infty$ limit, showing an exponential behavior which was not originally present in R^s_{opt}. At the other end of the spectrum, expanding eq. 6.42 around $R^s_{opt} = 0$ yields

$$\begin{aligned} R^s_{clip} &\simeq \sqrt{\frac{2}{\pi}} R^s_{opt} \\ &\simeq \sqrt{\frac{2}{\pi}} R_B \ . \end{aligned} \tag{6.44}$$

A comparison with eq. 5.18 shows that in the poor performance regime the R_{bb} bound is also asymptotically saturated. For intermediate values of α, however, R^s_{clip} does not saturate R_{bb}, in general. This can be seen in the dashed curves of fig. 6.4, where results for the Gaussian scenario are shown in the two relevant cases: zero and non-zero $\langle b \rangle_*$.

The next question one could ask, using the techniques reproduced in section 3.3, is the following : when one makes use of the optimal function $\phi^*(x) = \tanh(Rx/(1-R^2))$ (see eqs. 3.22 and 3.23) to transform the components of the vector $\boldsymbol{J}^s_{opt}$, how close is the resulting $R^s_* \equiv R^*(R = R^s_{opt})$ to R_B? By rewriting eq. 3.23, one is once more left with a deceivingly familiar formula,

$$R^s_* = F_B\left(\mathcal{F}(R^s_{opt})\right) \ , \tag{6.45}$$

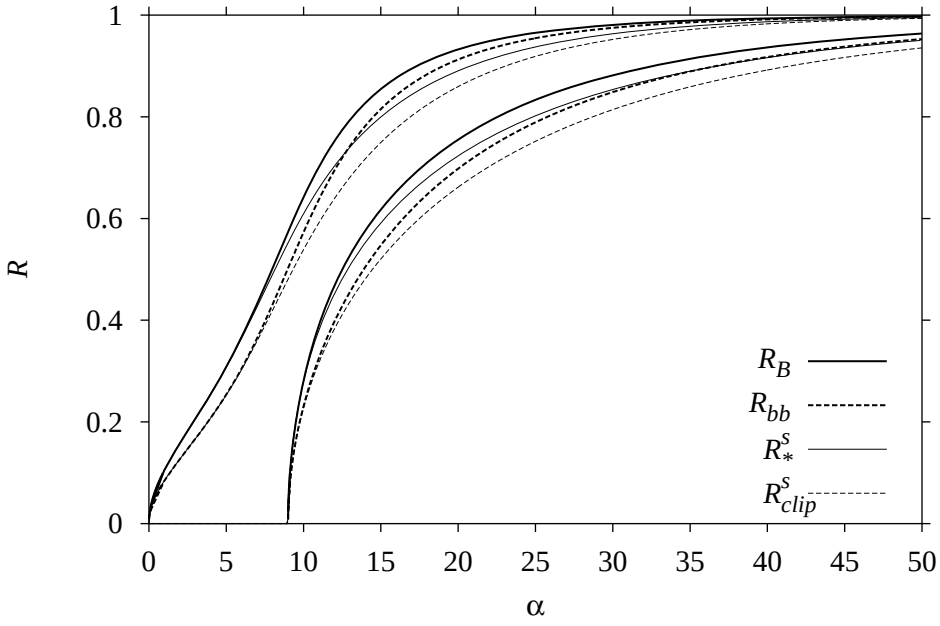

Figure 6.4: Overlaps as functions of α for two choices of parameters in the Gaussian scenario: $A = 1/3$ with $B = 0.1$ (upper curves) and $A = -1/3$ with $B = 0$ (lower curves). The upper bounds R_B (solid) and R_{bb} (dashed) are depicted with thick lines, while the approximations R^s_* (solid) and R^s_{clip} (dashed) are plotted with thin lines.

which should be compared with eq. 6.33. The monotonicity of $\mathcal{F}(R)$ plays again a central role, now forbidding $R^s_*(\alpha)$ to saturate the bound $R_B(\alpha)$, since in general $R^s_{opt} \leq R_B$. Saturation occurs only asymptotically, both for large α and for small R:

$$1 - R^s_* \overset{\alpha\to\infty}{\simeq} \sqrt{\frac{\pi}{8\alpha \left\langle (U')^2 \right\rangle_*}} \exp\left[\frac{-\alpha \left\langle (U')^2 \right\rangle_*}{2}\right]$$

$$\simeq \left(\frac{\pi}{4}\right)(1 - R^s_{clip}) \tag{6.46}$$

$$R^s_* \overset{R^s_{opt}\to 0}{\simeq} \sqrt{\frac{\pi}{2}} R^s_{clip} \simeq R_B \ . \tag{6.47}$$

Eqs. 6.46 and 6.47 should be compared to eqs. 5.16 and 5.18, respectively. The comparison between R^s_* and R_B is graphically very similar to that of R^s_{clip} and R_{bb}, as can be seen in the solid lines of fig. 6.4.

Another available measure of the success of the optimal transformation ϕ^* in rendering a good approximation for $\boldsymbol{J}_B$, is the probability distribution $P(x_*)$, where $x_* \equiv \phi^*(x)/R^*$ (see eqs. 3.13, 3.22 and 3.23). x_* reflects thus the structure of $\boldsymbol{J}_*^s \equiv \{\boldsymbol{J}_*^s | [\boldsymbol{J}_*^s]_i = \phi^*([\boldsymbol{J}_{opt}^s]_i)/R_*^s\}$. In order to obtain its probability distribution, one just has to recall eq. 3.17. It states that, for vectors uniformly distributed on the cone $\boldsymbol{J}\cdot\boldsymbol{B}/N$, x is Gaussian distributed, in this case with mean R_{opt}^s and variance $1-(R_{opt}^s)^2$. This is one of the signs that a spherical constraint imposed on the vectors $\boldsymbol{J}$ is not able to incorporate the information that $\boldsymbol{B}$ is binary: the Gaussian is much less structured than the distribution $P_{CM}(x)$ obtained in chapter 5 for $\boldsymbol{J}_B$, eq. 5.7, which "pushes" $\boldsymbol{J}_B$ closer to the corners of the hypercube (see figs. 5.1-5.4). The optimal transformation $x_* = \phi^*(x)/R^* = \tanh(Rx/(1-R^2))/R^*$ can be regarded as an attempt to fix this problem, attaching some structure to the transformed x_*. With a simple change of variables, $P(x_*)$ is readily seen to be

$$P(x_*) \overset{(3.17)}{=} \frac{R^*\sqrt{1-R^2}}{\sqrt{2\pi}R(1-(R^*x_*)^2)} \times \exp\left\{\frac{-(1-R^2)}{2R^2}\left[\frac{1}{2}\ln\left(\frac{1+R^*x_*}{1-R^*x_*}\right) - \frac{R^2}{1-R^2}\right]^2\right\}, \quad (6.48)$$

with $R = R_{opt}^s$ and $R^* = R_*^s$ as the present case of interest. A comparison with eq. 5.7 shows that the two equations are very similar, but not identical. Some similarity in shape should indeed be expected, mainly because $P(x_*)$, just like $P_{CM}(x)$, *must be* such that $P(x_*)/P(-x_*) = (1+cx_*)/(1-cx_*)$, for some constant c, in order to prevent the construction of a second optimal function after the first transformation[7].

The distributions $P_{CM}(x)$ (eq. 5.7) and $P(x_*)$ (eq. 6.48) can be compared in fig. 6.5. The curves correspond to the Gaussian scenario with $A = 1/3$ and $B = 0.1$ for two values of α. One can thus refer to the upper solid curves of fig. 6.4. Note that for $\alpha = 8$, the difference between R_B and R_*^s is very small in fig. 6.4, which is reflected in the solid curves of fig. 6.5 being very close to each other. Accordingly, the dashed curves in fig. 6.5 get further apart for $\alpha = 10$ as the mismatch between the overlaps increase in fig. 6.4.

The similar shape of eqs. 5.7 and 6.48 raises an immediate question: does a function $\phi(x)$ exist such that the transformation $x_* = \phi(x)$ leads to a distribution $P(x_*)$ *identical* to eq. 5.7? The answer is yes, but one should not be misled by the result. One can readily verify that, if x is Gaussian distributed with mean R and variance $1-R^2$, then the transformation

[7] As can be easily checked, $P(x_*)/P(-x_*) = (1+x_*)/(1-x_*)$.

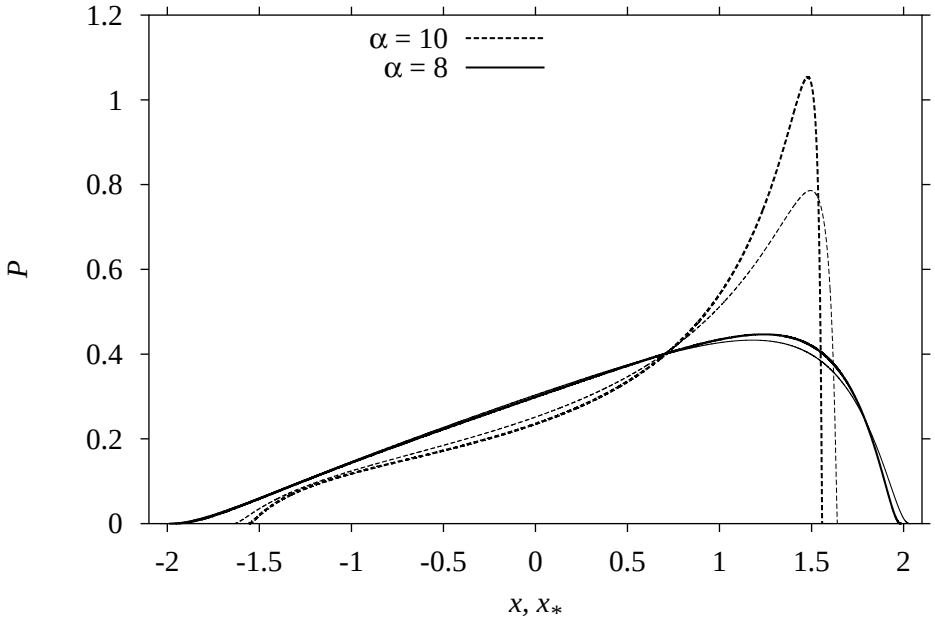

Figure 6.5: Distributions $P_{CM}(x)$ (thick line) and $P(x_*)$ (thin line) according to eqs. 5.7 and eq. 6.48, respectively. The values $\alpha = 8$ (solid) and $\alpha = 10$ (dashed) refer to the Gaussian scenario with $A = 1/3$ and $B = 0.1$.

$$x_* = \phi(x) = \frac{1}{\sqrt{q_{bin}}} \tanh\left(\sqrt{\frac{\hat{q}_{bin}}{(1-R^2)}}(x - R) + \hat{R}_{bin}\right) \tag{6.49}$$

formally leads to $P(x_*) = P_{CM}(x_*)$. The momentary change of notation is intended to stress that the variables with index *bin* are the equilibrium values obtained for Gibbs learning *in the binary space*, as discussed in chapter 5. But more importantly, one should focus on the word "formally" above. The problem is that, even though eq. 6.49 leads to the desired distribution, *it does not arise from any known transformation $\phi(J_i)$ on the components of the vector $\boldsymbol{J}^s_{opt}$*. The reason is that the whole technique of transforming the components is based on the hypothesis that the transformation function is odd (see page 39), which is not the case of eq. 6.49. If a transformation $J'_i \sim \varphi(J_i)$ could be found which would lead to eq. 6.49, the problem of finding both $\boldsymbol{J}_B$ and $\boldsymbol{J}_{bb}$ would have been solved. This would be quite remarkable, since eq. 6.49 would be applicable regardless of the value of R — and the desired Bayes-optimal solutions would be attainable with no more effort than

a simple transformation in the components, which seems unlikely. Note that $\phi(x)$ in eq. 6.49 becomes an odd function only if the equation $R/\sqrt{1-R^2} = \hat{R}_{bin}/\sqrt{\hat{q}_{bin}}$ is satisfied. On the other hand, the saddle point equations B.28 for the spherical case show that $R/\sqrt{1-R^2} = \hat{R}_s/\sqrt{\hat{q}_s}$, in the limit $q_s \to 1$. Once more the index s was introduced to stress that the parameters are the solution for the spherical case. Since the equilibrium value of the conjugate parameters are in general different for the spherical and binary cases, $\phi(x)$ can hardly be made an odd function.

A simple counterexample

As the discussions of section 6.2 attest, the case of a linear function $U(b) = -Bb$ is a special one. This is not different here. For a linear U, the monotonicity of $\mathcal{F}(R)$ attains its lowest with $d\mathcal{F}(R)/dR = 0$, identically. Therefore eqs. 6.42 and 6.45 immediately reveal that, for this very special case,

$$R^s_{clip}(\alpha) = R_{bb}(\alpha)\ , \qquad \forall\alpha \tag{6.50}$$

$$R^s_*(\alpha) = R_B(\alpha)\ , \qquad \forall\alpha\ , \tag{6.51}$$

which can be deduced by a comparison with eqs. 6.32 and 6.33, respectively. Note that the first result, eq. 6.50, could already have been expected from the previously obtained success of V_{opt} in finding $\boldsymbol{J}_{bb}$ in the binary space. This is because the resulting optimal potential is *linear*, which means that its minimization in the binary space is equivalent to its minimization in the continuous space followed by clipping. The second result, eq. 6.51, is more impressive, because it stablishes a result which could not be found elsewhere in the literature: *the optimal transformation manages to completely incorporate the information about the binary nature of the prefential direction, leading to* $\boldsymbol{J}_B$ *without the need of explicitly constructing the center of mass of the Gibbs ensemble*. In other words, the technique of non-linear transforming the components of the vectors, introduced by [BS95] and extended by [SBVdB95], is able to give a definitive answer to the problem it aims to solve.

Chapter 7

Conclusions and perspectives

In this thesis, techniques of Statistical Mechanics were used to derive theoretical results on unsupervised learning of a single N-dimensional direction $\boldsymbol{B}$, with emphasis on the case where $\boldsymbol{B} \in \{-1,+1\}^N$. Chapters 3, 4 and 5 focus on the calculation of an upper bound which is ultimately related to the general results of chapter 2. Chapter 6, on the other hand, contains a non-exhaustive discussion of some possibilities on how to saturate those bounds.

In chapter 2, Gibbs learning (which is sampling from the posterior distribution) is shown to have the same asymptotic behavior as for the smooth case in the limit $R_G \to 0$: for biased distribution of patterns, $R_G \propto \alpha$, while for non-biased distributions one obtains in general a second order phase transition (retarded learning) $R_G \sim (\alpha - \alpha_G)\Theta(\alpha - \alpha_G)$. In the limit $\alpha \to \infty$, the results are in sharp contrast with those of the smooth case, since the asymptotical approach to the perfect match $\boldsymbol{J} = \boldsymbol{B}$ occurs exponentially, $1 - R_G \simeq \sqrt{\pi/(2\alpha \langle (U')^2 \rangle_*)} \exp(-\alpha \langle (U')^2 \rangle_* /2)$. Non-asymptotic results for the Gaussian scenario also show that the discrete nature of the search space can lead to first order phase transitions as well.

Chapter 3 reproduces a well known reasoning which leads to the Bayes-optimal performance being formally attained by the center of mass $\boldsymbol{J}_B$ of the ensemble of Gibbsian vectors. This Bayes optimal performance obeys $R_B = \sqrt{R_G}$. The clipped version $\boldsymbol{J}_{bb}$ of $\boldsymbol{J}_B$ is shown to lead to the Bayes-optimal performance within the binary vectors. Using previous results on the theory of clipping, it is shown that in order to obtain its performance R_{bb}, one needs to calculate the probability distribution of the variable $x = B_1[\boldsymbol{J}_B]_1$.

In chapter 4 a reasoning is proposed, according to which the geometric constraints of the Gibbs ensemble would be enough to account for the statistical properties of its center of mass. This assumption is used to simplify the calculation of $P(x)$. Making use of the Maximum-Entropy formalism, it

is shown that the center of mass of binary vectors which obey some simple ("replica symmetric") constraints, is again a binary vector. Despite several arguments in favor of the reasoning, however, the assumption is not valid for a disordered system, which motivates an alternative calculation.

Chapter 5 presents the results of a replica calculation of $P(x)$ which properly accounts for the disorder of the examples, accurately reproducing the results of the simulations for a special data distribution. From the theoretical expression, the results of chapter 3 are used in order to obtain the upper bound R_{bb}, which obeys $R_G \leq R_{bb} \leq R_B$, equality being attained only in the limits $R_G \to 0$ and $R_G \to 1$. Asymptotically, $R_{bb} \simeq \sqrt{2R_G/\pi}$ for $R_G \to 0$, and $1 - R_{bb} \simeq \sqrt{2/(\pi\alpha \langle (U')^2 \rangle_*)} \exp(-\alpha \langle (U')^2 \rangle_* /2)$ for $\alpha \to \infty$. As a spin-off of the calculation, the overlap $\Gamma = \boldsymbol{J}_B \cdot \boldsymbol{J}_{bb}/N$ is obtained and shown to be always $\geq \sqrt{2/\pi}$.

Two approaches, to explicitly obtain vectors that could saturate the R_{bb} bound, are discussed in chapter 6. The first, which consists in minimizing — among the binary vectors — a variationally optimal potential $V_{opt}(\lambda)$ under the assumption that the order parameter $q \to 1$, is shown to generally fail. The conjecture is that this is due to the difficulty of constructing potentials with a unique minimum in the binary space. The second approach consists in obtaining a continuous vector $\boldsymbol{J}^s_{opt}$ by minimization of a different optimal potential V^s_{opt}. The components of $\boldsymbol{J}^s_{opt}$ are then either optimally transformed in order to approximate $\boldsymbol{J}_B$, or clipped to approximate $\boldsymbol{J}_{bb}$. In general, the approximations are successful only asymptotically.

In order to compare the results obtained in this thesis with results already known in the literature, it is convenient to group them together in tables. In table 7.1, results for upper bounds are grouped together and followed by results related to optimized cost functions. One should keep in mind that several results mentioned in this table were derived specifically in the context of supervised learning. In this respect, some of the questions posed on the leftmost column may not make much sense in the framework of unsupervised learning. They are nonetheless important issues which help to frame the results which have been obtained in the last chapters. The focus of the present work is clearly on the rightmost column.

From top to bottom, Opper and Haussler [OH91] calculated the *generalization error* (defined as $\pi^{-1} \arccos(R_B)$) of the so-called Bayes algorithm[1] for the perceptron problem. Watkin [Wat93], using the reasoning reproduced in chapter 3, then showed that this Bayes-optimal performance could be achieved by a machine with the same architecture (another perceptron), namely the center of mass $\boldsymbol{J}^s_B$ of the version space. The reasoning was ex-

[1] This rule consists in giving as binary output the "majority vote" of the version space.

tended to an unsupervised scenario in [WN94], where the optimality of the center of mass $\boldsymbol{J}_B$ of binary vectors was also mentioned. The fact that optimality *within* the class of binary vectors is obtained by $\boldsymbol{J}_{bb} = \text{clip}(\boldsymbol{J}_B)$ was mentioned in [WRB93]. But its performance R_{bb} could not be found in the literature. After attempting to solve this problem with the geometrical approach of chapter 4, the result is finally presented in chapter 5.

	Spherical $\boldsymbol{B}$		Binary $\boldsymbol{B}$	
How does the Bayes-optimal performance R_B relate to R_G?	$R_B^s = \sqrt{R_G^s}$ [OH91]		$R_B = \sqrt{R_G}$ [OH91]	
	Spherical $\boldsymbol{B}$		Binary $\boldsymbol{B}$	
	Spherical $\boldsymbol{J}$	Binary $\boldsymbol{J}$	Spherical $\boldsymbol{J}$	Binary $\boldsymbol{J}$
Can R_B be attained by an N-dimensional vector?	Yes, $\boldsymbol{J} = \boldsymbol{J}_B^s$ [Wat93]	No	Yes, $\boldsymbol{J} = \boldsymbol{J}_B$ [WN94]	No
What is the best *binary* vector?		$\text{clip}(\boldsymbol{J}_B^s)$		$\boldsymbol{J}_{bb} = \text{clip}(\boldsymbol{J}_B)$ [WRB93]
And its performance?		Bad[2]		R_{bb} (chapter 5)
$\exists V(\lambda)$ such that minimization with $q \to 1$ leads to the bound?	Yes, $V = V_{opt}^s$ [KC96], [VdBR96]	No	Still unknown	No, in general (chapter 6)

Table 7.1: Upper bounds and variationally optimal potentials.

Turning to potentials, the Bayes-optimal bound was shown to be at-

[2]If $\boldsymbol{B}$ is spherical and $\boldsymbol{J}$ is binary, the perfect match will not occur, on average. See [VdBB93] for an example.

tainable, in the spherical case, through the minimization with $q \to 1$ of a potential $V^s_{opt}(\lambda)$, by Kinouchi and Caticha [KC96], Van den Broeck and Reimann [VdBR96], and Buhot, Torres Moreno and Gordon [BTMG97, GB98]. The equivalent problem in the binary space was addressed in [dM97], for the supervised case, and in chapter 6, for the unsupervised case. In section 6.2, it is shown that the R_{bb} bound cannot be saturated (except for special cases) by minimizing a variationally optimal potential $V_{opt}(\lambda)$ under the assumption that $q \to 1$.

A discussion is in order at this point. It should be mentioned that the uniqueness of the center of mass of the Gibbs ensemble *given the examples*, does not necessarily enforces the limit $q \to 1$. This assumption was made based on previous results for the spherical case, where it was successful. Note that in the replica calculation the average over the examples is taken, while the Bayes-optimal vector remains a function of the examples (recall chapter 3, page 36). The connection between the two approaches is made thanks to the self-averaging property of the order parameters, brought by the thermodynamic limit. But in principle, it is possible to conceive that an optimized potential $\tilde{V}_{opt}$ could lead to $R = R_{bb}$ and a matrix $\{q_{ab}\}$ structured according to some step of replica symmetry breaking. At zero temperature, this could possibly imply a degenerate minimum, with each of the minima saturating the R_{bb} bound, *on average*. Once more assuming self-averaging, the only conditions to be verified would be: 1) that the center of mass of this new ensemble remains with an overlap with $\boldsymbol{B}$ less than R_B; 2) that the clipped center of mass remains with an overlap with $\boldsymbol{B}$ less than R_{bb}. This would require a new calculation, similar to the one presented in chapter 5.

Another possibility (and perhaps simpler to address) would be another attempt to construct a potential to saturate the performance of $\boldsymbol{J}_B$ which, despite being a continuous vector, incorporates optimally the information that $\boldsymbol{B}$ is binary. In this respect, the techniques of coupled systems [WRS92] could be useful. These are systems where the same set of patterns is used for an ensemble of binary vectors, say $\{\boldsymbol{J}^a\}$, and an ensemble of spherical vectors, say $\{\boldsymbol{W}^\nu\}$. The average over the patterns couples the two spaces, giving rise to a new order parameter $\boldsymbol{J}^a \cdot \boldsymbol{W}^\nu / N$. By tuning the potential in order to control this order parameter, one could perhaps conjugate the advantages of working in a continuous space with the need of addressing the discrete nature of $\boldsymbol{B}$ (see [VR99] and references therein). This could be a way of overcoming the possible limitations of working with a cost function of the form $\mathcal{H} = \sum_\mu V(\lambda_\mu)$, which alone might not be able to properly pinpoint the binary vectors. Alternatively, different techniques already used for the supervised problem, like the TAP equations [OW96], could shed light on the unsupervised problem as well.

Returning to the review of the results, section 6.3 describes an alternative technique which provides approximations to $\boldsymbol{J}_{bb}$ and $\boldsymbol{J}_B$. In table 7.2 below, a list of vectors is provided in order to facilitate the discussion.

Vector	Nature	Definition	Overlap with $\boldsymbol{B} \in \{-1,+1\}^N$
Gibbsian: $\boldsymbol{J}_G^a$	binary	a sample of $P(\boldsymbol{J}\|D) = \frac{P_b(\boldsymbol{J}) e^{-\sum_\mu^{\alpha N} U(\lambda_\mu)}}{Z}$	R_G
Bayesian: $\boldsymbol{J}_B$	spherical	CM of the $\{\boldsymbol{J}_G^a\}$: $= \frac{\int d\boldsymbol{J}\, P_b(\boldsymbol{J})\, \boldsymbol{J}\, e^{-\sum_\mu^{\alpha N} U(\lambda_\mu)}}{\sqrt{R_G}}$ $= \frac{1}{\sqrt{R_G}} \lim_{n\to\infty} \sum_a^n \boldsymbol{J}_G^a$	$R_B = \sqrt{R_G}$
Best Binary $\boldsymbol{J}_{bb}$	binary	$= \mathrm{clip}(\boldsymbol{J}_B)$	R_{bb}
"Spherical-Optimal" $\boldsymbol{J}_{opt}^s$	spherical	$= \underset{\boldsymbol{J}\in\mathbb{R}^N}{\mathrm{Argmin}} \sum_\mu^{\alpha N} V_{opt}^s \left(\frac{\boldsymbol{J}\cdot\boldsymbol{\xi}^\mu}{\sqrt{N}} \right)$	$R_{opt}^s = R_B^s$
$\boldsymbol{J}_*^s$	spherical	$[\boldsymbol{J}_*^s]_i = \frac{\phi^*\left([\boldsymbol{J}_{opt}^s]_i\right)}{R_*^s}$	R_*^s
$\boldsymbol{J}_{clip}^s$	binary	$= \mathrm{clip}(\boldsymbol{J}_{opt}^s)$	R_{clip}^s

Table 7.2: Definitions of some vectors.

The three first rows describe vectors that are easily defined in formal terms, but for which there are no obvious prescriptions as to how they can be constructed in practice. In particular, $\boldsymbol{J}_B$ and $\boldsymbol{J}_{bb}$ are associated with the upper bounds of table 7.1. The three last rows, on the other hand, describe vectors that are transformations of $\boldsymbol{J}_{opt}^s$, which is obtained by minimizing a

properly defined potential V^s_{opt} [KC96, BTMG97, VdBR96]. Such transformations, as described in chapter 3, aim at incorporating information about the binary nature of $\boldsymbol{B}$. The vectors $\boldsymbol{J}^s_*$ and $\boldsymbol{J}^s_{clip}$ can thus be regarded as approximations to $\boldsymbol{J}_B$ and $\boldsymbol{J}_{bb}$, respectively. Their success in accomplishing so is measured by comparing the associated overlaps. This analysis can be summarized as follows: the inequalities $R^s_* \leq R_B$ and $R^s_{clip} \leq R_{bb}$ are always satisfied, with equality always holding asymptotically for large and small α.

The only model that was found where the equalities above hold $\forall\alpha$, is the linear case $U(b) = -Bb$. Moreover, for a linear $U(b)$ one obtains a linear $V_{opt}(\lambda)$, which leads to the optimal potential being successful in yielding a binary vector with overlap R_{bb} at its ground state. These apparent coincidences motivate a more thorough discussion of the linear case. The formal reason behind the equalities is the fact that $\partial\mathcal{F}/\partial R = 0$ in this case. It immediately raises the need of a more careful functional analysis to determine whether this is true *only* for the linear case. One could argue that this might be true — and this is left here as a conjecture — based on the intuition gained with the difficulty in constructing potentials with a unique minimum in the binary space. The reasoning supporting the conjecture is: the optimal potential can only be successful if the ground state of the potential is unique, by definition; on the other hand, it generally fails, except when the condition $s \to 0$ is seen to be satisfied; therefore one could expect that, *if* a potential with a unique minimum *could* be found in general, the optimization would probably be able to find it.

Still related to this issue, another fact worth mentioning is that simulations become extremely difficult as soon as the smallest degree of non-linearity is introduced in $V(\lambda)$. This is true even for finite temperature, which is why simulations for Gibbs learning were presented only for $A = 0$. It is hard to tell whether sampling from the posterior distribution could have anything to do with the NP problem mentioned in the introductory chapter, in general. But if it has, one could conjecture that a possible signature of it might be the fact that V_{opt} fails in saturating the upper bound R_{bb}.

Conjectures apart, mean field techniques could perhaps be used, on the other hand, to study how the non-linearity induces a slowing down of the sampling dynamics, as done in refs. [Hor92a, Hor92b, Hor93]. But regardless of what these studies might reveal, the fact of the matter remains that, at least for the models studied here, *in general the optimal potential does not work, unless when it is trivial.* The word "trivial" is used here in the following sense: whenever V_{opt} is known to lead to a value of R_{opt} saturating the R_{bb} bound, the procedure of minimizing it is also extremely simple. This is because V_{opt} seems to accomplish its mission whenever clipping is optimal. This is always true asymptotically (for small and large α) and, if U is

linear, $\forall\alpha$. In any of these successful cases, minimizing V_{opt} boils down to clipping. Either clipping (for instance) the "spherical-optimal" $\boldsymbol{J}^s_{opt}$ in the general asymptotic case, or clipping the Hebbian vector $\boldsymbol{J}_H \propto \sum_\mu \boldsymbol{\xi}^\mu$ for any α in the linear case. In short, the optimization procedure succeeds in the easy cases.

	$\boldsymbol{J}$ spherical	$\boldsymbol{J}$ binary
Off-line "difficult"	$1-R \simeq C'\alpha^{-1}$	$1-R \sim e^{-\alpha}$
On-line "difficult"	$1-R \simeq C'\alpha^{-1}$	$R=0$
Off-line "easy"	$1-R \simeq \frac{C}{4}\alpha^{-2}$	First order phase transition
On-line "easy"	$1-R \simeq C\alpha^{-2}$	$R=0$

Table 7.3: Results for the optimal performance in on-line and off-line learning. C and C' are constants. See text for details.

It is interesting to compare the above scenario with some previous results, relating the optimal potential in the hypersphere V^s_{opt} with results from on-line learning (see table 7.3). On-line learning in the hypersphere consists in making use of a single different pattern $\boldsymbol{\xi}^\mu$ at each infinitesimal modification in the candidate vector $\boldsymbol{J}$. Thus after αN steps, one has made use of αN patterns, yielding as result a vector $\boldsymbol{J}^s_{ol}$ with an overlap $R^s_{ol}(\alpha)$. It is possible to variationally tune the incremental modifications in such a way that, on average, $R^s_{ol}(\alpha)$ is maximized (the procedure is nearly identical to the one described in chapter 6). In this case, $R^s_{ol}(\alpha)$ can be compared to $R^s_{opt}(\alpha)$, which clearly extracts more information from the patterns (since all patterns are used at all the infinite steps of the dynamics until equilibrium is reached). Surprisingly, the general asymptotic result is [VdBR96] $R^s_{opt}(\alpha) = R^s_{ol}(\alpha)$, i.e. on-line learning is able to yield the same performance as off-line learning in the limit $\alpha \to \infty$. However, the result is valid only if $\langle (U')^2 \rangle_*$ is finite, the "difficult" cases. If it diverges (the "easy" cases), then the result is $R^s_{opt}(\alpha) = R^s_{ol}(2\alpha)$, i.e., on-line learning takes twice as many examples to

yield the same performance. Or, seen from another standpoint, the off-line optimization is able to be twice as effective in the "easy" cases. Recently, Kinzel and Urbanczik showed that on-line learning in a binary space leads to $R = 0$, for the supervised case [KU98]. Since supervised learning maps into the "easy" cases of unsupervised learning, one would expect that on-line learning also fails in the "difficult" cases. Off-line learning, on the other hand, is theoretically able to take advantage of the reduced discrete number of states to render a better performance.

This discussion confirms thus a tendency of the optimal off-line strategies to successfully exploit the facilities present in a given problem (be it a diverging $\langle (U')^2 \rangle_*$, the discreteness of the search space, or both). Specifically in the problem of constructing an optimal potential V_{opt} in the binary space, this tendency apparently manifests itself in the narrow window where this construction is possible, namely, clipping.

Finally, going back to the technique of transforming the components of a conveniently constructed spherical vector, one should stress that, from the practical point of view, its merit is huge. After all, both clipping and the optimal transformation ϕ^* turn a power law into an exponential decay. Even away from the asymptotic regimes (where the bounds can be saturated), the procedure yields reasonable results (specially ϕ^*, which never decreases the performance). However, from the theoretical point of view, this approach is still not well understood, since it consists in an *ad hoc* procedure which therefore does not answer some deeper theoretical questions. One of such questions that immediately comes to mind refers to the possibility of constructing optimal cost functions: what is the best way (assuming there is one) to incorporate information about the binary nature of $\boldsymbol{B}$, other than by constraining $\boldsymbol{J}$ to the binary space in the partition function? It is actually remarkable that ϕ^* is able to *perfectly* do so in the linear case, yielding for the first time an explicit procedure (other than the center of mass recipe) for constructing a vector with Bayes-optimal performance for the binary prior. But it only does so in the linear case, exactly when constructing the center of mass is not difficult. Therefore a review of these alternatives seems to be still missing. The same type of comment can be applied to other *ad hoc* procedures which seek to incorporate information about the binary nature of $\boldsymbol{B}$, such as constraining $\boldsymbol{J}$ to the interior of the hypercube ($|J_i| \leq 1$), for instance. They can be very effective from the practical point of view but, more than that, they could perhaps suggest the path which would lead to a more systematic approach, based on first principles.

Appendix A

Notation

A.1 General remarks

- A bold letter like $\boldsymbol{J}$ is a vector whose j-th component may be denoted either by J_j or $[\boldsymbol{J}]_j$, according to convenience. The dot product is the usual one, thus $\boldsymbol{J} \cdot \boldsymbol{B} = \sum_{j=1}^{N} J_j B_j$, where N is the dimension of $\boldsymbol{J}$ and $\boldsymbol{B}$.

- For all integrals:
  - the integration interval is $(-\infty, +\infty)$ unless otherwise stated.
  - $d\boldsymbol{X}$ stands for $d^D\boldsymbol{X}$, where D is the dimension of vector $\boldsymbol{X}$.

- A δ-function with a single argument denotes a Dirac delta distribution, like $\delta(\boldsymbol{J} \cdot \boldsymbol{J} - N)$, for instance.

- Two arguments are used to denote a Kronecker delta, like $\delta(i, j)$, for instance.

- The typical abuse of notation for probabilities and probability densities is employed: if x and y are two variables with different probability distributions, $P(x)$ and $P(y)$ will eventually be used to describe them, even though the symbol "P" obviously stands for different functions in each case. No special notation will be employed to distinguish between probabilities and probability densities either. This should be clear from the context, the former usually being assigned to discrete variables and the latter to continuous variables.

A.2 Functions and associated properties

$\Theta(x)$ is the usual Heaviside function:

$$\Theta(x) = \begin{cases} 1 & , \text{if } x \geq 0 \\ 0 & , \text{if } x < 0 \end{cases} \tag{A.1}$$

$$\Theta(x) + \Theta(-x) = 1 \text{ (for } x \neq 0) \tag{A.2}$$

The Gaussian measure is:

$$\mathcal{D}x \equiv dx\, P_n(x) \tag{A.3}$$

$$P_n(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}} \tag{A.4}$$

$$\int \mathcal{D}x = 1 \tag{A.5}$$

The H function and the error function are:

$$H(x) \equiv \int_x^\infty \mathcal{D}t = \frac{1}{2}\left[1 - \text{erf}(x/\sqrt{2})\right] \tag{A.6}$$

$$H(x) + H(-x) = 1 \tag{A.7}$$

$$\text{erf}(x) \equiv \frac{2}{\sqrt{\pi}} \int_0^x dt\; e^{-t^2} = 1 - 2H\left(x\sqrt{2}\right) \tag{A.8}$$

More specifically related to the problem of unsupervised learning, one has the following integration measure and averages:

$$\begin{aligned} \mathcal{D}^* b &\equiv \mathcal{D}b\, \mathcal{N} \exp -U(b) \\ &= \frac{db\, e^{-b^2/2 - U(b)}}{\int db\, e^{-b^2/2 - U(b)}} \end{aligned} \tag{A.9}$$

$$\int \mathcal{D}^* b = 1\,, \tag{A.10}$$

where

$$\mathcal{N} = \frac{\sqrt{2\pi}}{\int dk\, \exp\left[-k^2/2 - U(k)\right]} = \frac{1}{\int \mathcal{D}k\, \exp -U(k)}\,. \tag{A.11}$$

$$\langle(\cdots)\rangle_* \equiv \int \mathcal{D}^* b\,(\cdots)\,. \tag{A.12}$$

The binary (Ising) and spherical integration measures are

$$\begin{aligned} dm(\boldsymbol{J}) &\overset{\text{Ising}}{=} d\boldsymbol{J}\,P_b(\boldsymbol{J}) \equiv \prod_{j=1}^{N} d^b J_j \\ &= \prod_{j=1}^{N} dJ_j \left[\frac{1}{2}\delta(J_j-1)+\frac{1}{2}\delta(J_j+1)\right] \\ dm(\boldsymbol{J}) &\overset{\text{Spherical}}{=} d\boldsymbol{J}\,P_s(\boldsymbol{J}) \sim d\boldsymbol{J}\,\delta(\boldsymbol{J}\cdot\boldsymbol{J}-N)\,. \end{aligned} \tag{A.13}$$

The following functions also appear quite frequently:

$$X(t;R) = \mathcal{N}\int \mathcal{D}t'\,e^{-U(Rt+\sqrt{1-R^2}t')} \tag{A.14}$$

$$\begin{aligned} Y(t;R) &= \frac{\mathcal{N}}{\sqrt{1-R^2}}\int \mathcal{D}t'\,t'\,e^{-U(Rt+\sqrt{1-R^2}t')} \\ &= \mathcal{N}\int \mathcal{D}t'\,e^{-U(Rt+\sqrt{1-R^2}t')}\left[-U'(Rt+\sqrt{1-R^2}t')\right] \\ &= \frac{1}{R}\frac{\partial}{\partial t}X(t;R)\,, \end{aligned} \tag{A.15}$$

where it should also be noted that

$$\int \mathcal{D}t\,X(t,R) = 1\,, \tag{A.16}$$

$$\begin{aligned} \frac{\partial X(t;R)}{\partial R} &= tY(t;R)+\frac{R}{(1-R^2)^{3/2}}\Bigg[X(t;R) \\ &\quad -\mathcal{N}\int \mathcal{D}t'\,(t')^2 e^{-U(Rt+\sqrt{1-R^2}t')}\Bigg]\,. \end{aligned} \tag{A.17}$$

A.3 Orthogonal transformation

The following orthogonal transformation is extensively used, particularly in appendix C:

$$\left.\begin{array}{l} b(u,v;R) = \frac{R}{\sqrt{q}}u + \sqrt{\frac{q-R^2}{q}}v \\ t_2(u,v;R) = \sqrt{\frac{q-R^2}{q}}u - \frac{R}{\sqrt{q}}v \end{array}\right\} \quad \Leftrightarrow \quad \left\{\begin{array}{l} u(b,t_2;R) = \frac{R}{\sqrt{q}}b + \sqrt{\frac{q-R^2}{q}}t_2 \\ v(b,t_2;R) = \sqrt{\frac{q-R^2}{q}}b - \frac{R}{\sqrt{q}}t_2 \end{array}\right.$$

$$\begin{aligned} db\, dt_2 &= du\, dv \\ b^2 + t_2^2 &= u^2 + v^2 \ , \end{aligned} \tag{A.18}$$

which implies

$$\begin{aligned} &\int \mathcal{D}b\, e^{-U(b)} \int \mathcal{D}t_2\, f(b,t_2) = \\ &\quad \int \mathcal{D}u \int \mathcal{D}v\, e^{-U\left(\frac{R}{\sqrt{q}}u + \sqrt{\frac{q-R^2}{q}}v\right)} f\left(\frac{R}{\sqrt{q}}u + \sqrt{\frac{q-R^2}{q}}v,\right. \\ &\quad \left.\sqrt{\frac{q-R^2}{q}}u - \frac{R}{\sqrt{q}}v\right) . \end{aligned} \tag{A.19}$$

A.4 Asymptotics and Series

$$\int \mathcal{D}z\ \ln 2\cosh\left(z\sqrt{x}+x\right) \left\{\begin{array}{ll} \overset{x\to\infty}{\simeq} & x + \sqrt{\frac{\pi}{2x}}e^{-x/2}\left[1+\mathcal{O}(x^{-1})\right] \simeq x \\ \overset{x\to 0}{\simeq} & \ln 2 + \frac{x}{2} + \frac{x^2}{4} - \frac{x^3}{6} + \mathcal{O}(x^4) \end{array}\right. \tag{A.20}$$

$$\int \mathcal{D}z\ \tanh\left(z\sqrt{x}+x\right) \left\{\begin{array}{ll} \overset{x\to\infty}{\simeq} & 1 - \sqrt{\frac{\pi}{2x}}e^{-x/2}\left[1+\mathcal{O}(x^{-1})\right] \simeq 1 \\ \overset{x\to 0}{\simeq} & x - x^2 + \mathcal{O}(x^4) \end{array}\right. \tag{A.21}$$

$$H(x) \left\{\begin{array}{ll} \overset{x\to\infty}{\simeq} & \frac{e^{-x^2/2}}{\sqrt{2\pi}}\left\{\frac{1}{x} - \frac{1}{x^3} + \frac{3}{x^5} - \frac{3.5}{x^7} + \ldots\right\} \\ \overset{x\to 0}{\simeq} & \frac{1}{2} - \frac{x}{\sqrt{2\pi}} + \frac{x^3}{6\sqrt{2\pi}} + \mathcal{O}(x^5) \end{array}\right. \tag{A.22}$$

$$\begin{aligned} \left.\frac{\partial X(t;R)}{\partial R}\right|_{R=1} &= 0 \\ \Rightarrow X(t;R) &\overset{R\to 1}{\simeq} \mathcal{N} e^{-U(t)} + \mathcal{O}\left((1-R)^2\right) \end{aligned} \tag{A.23}$$

$$\begin{aligned} \left.\frac{\partial X(t;R)}{\partial R}\right|_{R=0} &= t\,\langle b\rangle_* \\ \Rightarrow X(t;R) &\overset{R\to 0}{\simeq} 1 + t\,\langle b\rangle_*\, R + \mathcal{O}\left(R^2\right) \end{aligned} \tag{A.24}$$

Appendix B

The replica calculation

B.1 General results: the free energy

In the following, the preferential direction $\boldsymbol{B}$ is an Ising vector, $B_j \in \{-1,+1\}$, unless otherwise stated. Given a set of patterns (or examples) $D = \{\boldsymbol{\xi}^\mu\}$, $\mu = 1,\ldots,p$, the equilibrium properties of an ensemble of vectors $\boldsymbol{J}$ evolving in an energy landscape $\mathcal{H} = \sum_{\mu=1}^{p} V(\boldsymbol{J}\cdot\boldsymbol{\xi}^\mu/\sqrt{N})$ at inverse temperature β can be studied via the partition function

$$Z(\{\boldsymbol{\xi}^\mu\}) = \int dm(\boldsymbol{J}) \exp\left[-\beta \sum_{\mu=1}^{p} V\left(\frac{\boldsymbol{J}\cdot\boldsymbol{\xi}^\mu}{\sqrt{N}}\right)\right] , \tag{B.1}$$

where $dm(\boldsymbol{J})$ stands for the measure in $\boldsymbol{J}$ space. In this appendix, only two measures will be considered, namely the binary (Ising) measure

$$\begin{aligned} dm(\boldsymbol{J}) &\overset{\text{Ising}}{=} d\boldsymbol{J}\, P_b(\boldsymbol{J}) \equiv \prod_{j=1}^{N} d^b J_j \\ &= \prod_{j=1}^{N} dJ_j \left[\frac{1}{2}\delta(J_j-1) + \frac{1}{2}\delta(J_j+1)\right] \end{aligned} \tag{B.2}$$

and the spherical measure

$$dm(\boldsymbol{J}) \overset{\text{Spherical}}{=} d\boldsymbol{J}\, P_s(\boldsymbol{J}) = d\boldsymbol{J}\, \delta(\boldsymbol{J}\cdot\boldsymbol{J} - N) . \tag{B.3}$$

The patterns are random and will be here assumed to be uncorrelated, being independently drawn from a distribution with a single symmetry breaking direction:

$$P(\{\boldsymbol{\xi}^\mu\}|\boldsymbol{B}) = \prod_{\mu=1}^{p} P_u(\boldsymbol{\xi}^\mu|\boldsymbol{B}) = \prod_{\mu=1}^{p} \frac{\delta(\boldsymbol{\xi}^\mu\cdot\boldsymbol{\xi}^\mu - N)\,\exp\left[-U\left(\frac{\boldsymbol{\xi}^\mu\cdot\boldsymbol{B}}{\sqrt{N}}\right)\right]}{\int d\boldsymbol{\xi}'\,\delta(\boldsymbol{\xi}'\cdot\boldsymbol{\xi}' - N)\,\exp\left[-U\left(\frac{\boldsymbol{\xi}'\cdot\boldsymbol{B}}{\sqrt{N}}\right)\right]}\ . \tag{B.4}$$

In order to describe the average properties of the system, the following expression for the free energy is studied:

$$f = -\lim_{N\to\infty}\frac{1}{\beta N}\langle \ln Z(\{\boldsymbol{\xi}^\mu\})\rangle_{\{\boldsymbol{\xi}^\mu\}|\boldsymbol{B}}\ , \tag{B.5}$$

where $\langle(\cdots)\rangle_{\{\boldsymbol{\xi}^\mu\}|\boldsymbol{B}} = \int P(\{\boldsymbol{\xi}^\mu\}|\boldsymbol{B})\,(\cdots)\prod_{\mu=1}^{p} d\boldsymbol{\xi}^\mu$. The average of the logarithm in eq. B.5 is usually very hard to compute analytically, which motivates the use of the replica trick. Making use of the identity $\ln a = \lim_{n\to 0}(a^n-1)/n$, one rewrites

$$f = -\lim_{n\to 0}\lim_{N\to\infty}\frac{1}{\beta N n}\left(\langle Z^n\rangle_{\{\boldsymbol{\xi}^\mu\}|\boldsymbol{B}} - 1\right)\ . \tag{B.6}$$

The term Z^n can be rewritten as a product of the partition functions of n *replicated* systems, labeled by a: $Z^n = \prod_a^n Z_a$. For a given set of examples, the different replicas do not interact. But when the average over the disorder of the examples is performed, they get coupled, as will soon become evident.

First it is important to realize that the independence of the examples simplifies very much the calculation:

$$\begin{aligned}
\langle Z^n\rangle_{\{\boldsymbol{\xi}^\mu\}|\boldsymbol{B}} &= \left\langle \int\prod_a^n dm(\boldsymbol{J}^a)\exp\left[-\beta\sum_{\mu=1}^{p}\sum_{a=1}^{n} V\left(\frac{\boldsymbol{J}^a\cdot\boldsymbol{\xi}^\mu}{\sqrt{N}}\right)\right]\right\rangle_{\{\boldsymbol{\xi}^\mu\}|\boldsymbol{B}} \\
&= \int\prod_a^n dm(\boldsymbol{J}^a)\prod_{\mu=1}^{p}\int d\boldsymbol{\xi}^\mu\ P_u(\boldsymbol{\xi}^\mu|\boldsymbol{B}) e^{-\beta\sum_{a=1}^n V\left(\frac{\boldsymbol{J}^a\cdot\boldsymbol{\xi}^\mu}{\sqrt{N}}\right)} \\
&= \int\prod_a^n dm(\boldsymbol{J}^a)\left[\int d\boldsymbol{\xi}\ P_u(\boldsymbol{\xi}|\boldsymbol{B}) e^{-\beta\sum_{a=1}^n V\left(\frac{\boldsymbol{J}^a\cdot\boldsymbol{\xi}}{\sqrt{N}}\right)}\right]^p .
\end{aligned} \tag{B.7}$$

In order to perform the integration on $\boldsymbol{\xi}$, one introduces δ-distributions for the fields $\lambda_a \equiv N^{-1/2}\boldsymbol{J}^a\cdot\boldsymbol{\xi}$ through its well known Fourier representation:

$$\begin{aligned}
\langle Z^n\rangle_{\{\boldsymbol{\xi}^\mu\}|\boldsymbol{B}} &= \int\prod_a^n dm(\boldsymbol{J}^a)\left[\int\frac{d\lambda_a\,d\hat{\lambda}_a}{2\pi}\ e^{i\sum_a\hat{\lambda}_a\lambda_a}\ e^{-\beta\sum_a V(\lambda_a)}\right. \\
&\times \left.\int d\boldsymbol{\xi}\ P_u(\boldsymbol{\xi}|\boldsymbol{B})\exp -i\sum_a\hat{\lambda}_a\frac{\boldsymbol{J}^a\cdot\boldsymbol{\xi}}{\sqrt{N}}\right]^p\ .
\end{aligned} \tag{B.8}$$

One can now repeat this procedure to extract $\boldsymbol{\xi}$ from the argument of U (see B.4), the function that controls the non-uniformity of the patterns along the direction $\boldsymbol{B}$. Introducing a δ-distribution for $b \equiv N^{-1/2}\boldsymbol{B}\cdot\boldsymbol{\xi}$ one is left with Gaussian integrals which can be immediately performed. Defining the measure

$$\mathcal{D}^* b \equiv \frac{db\, e^{-b^2/2-U(b)}}{\int db'\, e^{-b'^2/2-U(b')}}\ , \tag{B.9}$$

the above described calculations yield[1]

$$\int d\boldsymbol{\xi}\, P_u(\boldsymbol{\xi}|\boldsymbol{B}) \exp -i\sum_a \hat{\lambda}_a \frac{\boldsymbol{J}^a\cdot\boldsymbol{\xi}}{\sqrt{N}} = \exp\left[\frac{1}{2}\left(\sum_a \hat{\lambda}_a \frac{\boldsymbol{J}^a\cdot\boldsymbol{B}}{N}\right)^2 - \frac{1}{2N}\sum_{a,b}\hat{\lambda}_a\hat{\lambda}_b \boldsymbol{J}^a\cdot\boldsymbol{J}^b\right] \times \int \mathcal{D}^* b \exp -ib\sum_a \hat{\lambda}_a \frac{\boldsymbol{J}^a\cdot\boldsymbol{B}}{N}. \tag{B.10}$$

Note that replicas are now coupled. In order to proceed with the integration over $dm(\boldsymbol{J})$ it should be stressed that for both measures under consideration (eqs. B.2 and B.3) the diagonal term in $\boldsymbol{J}^a\cdot\boldsymbol{J}^b$ equals N. One only has to introduce δ-distributions for the physically relevant order parameters

$$\begin{aligned} q_{ab} &\equiv \frac{\boldsymbol{J}^a\cdot\boldsymbol{J}^b}{N}, \qquad a<b \\ R_a &\equiv \frac{\boldsymbol{J}^a\cdot\boldsymbol{B}}{N}, \qquad a=1,\ldots,n\ , \end{aligned} \tag{B.11}$$

which leads to the following expression:

$$\begin{aligned} \langle Z^n\rangle_{\{\boldsymbol{\xi}^\mu\}|\boldsymbol{B}} &= \prod_{a=1}^{n}\left[\int dR_a \int_{-i\infty}^{i\infty}\frac{d\hat{R}_a}{2\pi i N}\right]\prod_{a<b}\left[\int dq_{ab}\int_{-i\infty}^{i\infty}\frac{d\hat{q}_{ab}}{2\pi i N}\right] \\ &\times \exp N\Big\{\sum_a \hat{R}_a R_a + \sum_{a<b}\hat{q}_{ab}q_{ab} + G_0(\{\hat{R}_a,\hat{q}_{ab}\}) \\ &+ \frac{p}{N}G_1(\{R_a,q_{ab}\};\beta,[V,U])\Big\}\ , \end{aligned} \tag{B.12}$$

where the functions

[1] In fact, the result B.10 is unchanged if the spherical constraint $\delta(\boldsymbol{\xi}\cdot\boldsymbol{\xi}-N)$ on the patterns is replaced by a binary one, $P_b(\boldsymbol{\xi})$, in eq. B.4.

$$G_0 \equiv \frac{1}{N}\ln\int\prod_a dm(\boldsymbol{J}^a)e^{-\sum_a \hat{R}_a \boldsymbol{J}^a\cdot\boldsymbol{B}-\sum_{a<b}\hat{q}_{ab}\boldsymbol{J}^a\cdot\boldsymbol{J}^b} \tag{B.13}$$

$$G_1 \equiv \ln\int\prod_a d\lambda_a\; e^{-\beta\sum_a V(\lambda_a)}\int\mathcal{D}^*b\prod_a\frac{d\hat{\lambda}_a}{2\pi}e^{i\sum_a\hat{\lambda}_a(\lambda_a-bR_a)-\sum_a\hat{\lambda}_a^2/2} \times e^{\left(\sum_a\hat{\lambda}_a R_a\right)^2/2-\sum_{a\neq b}q_{ab}\hat{\lambda}_a\hat{\lambda}_b/2} \tag{B.14}$$

are both $\mathcal{O}(1)$. G_0 will hereafter be referred to as the *entropy term*, while G_1 is called the *energy term*.

In the limit $N\to\infty$ eq. B.12 is exponentially dominated by the saddle point(s) of the expression between braces. At this point it becomes clear that the number of examples must scale with the dimensionality of the system, that is, $p=\alpha N$, otherwise no interesting trade-off between G_0 and G_1 is obtained. Recalling eq. B.6, one arrives at

$$f = -\lim_{n\to 0}\frac{1}{\beta n}\underset{\{\hat{R}_a,\hat{q}_{ab},R_a,q_{ab}\}}{\mathrm{Extr}}\Big\{\sum_a\hat{R}_aR_a+\sum_{a<b}\hat{q}_{ab}q_{ab}+G_0(\{\hat{R}_a,\hat{q}_{ab}\}) + \alpha\, G_1(\{R_a,q_{ab}\};\beta,[V,U])\Big\}\,. \tag{B.15}$$

B.2 The replica symmetric *ansatz*

In order to obtain the equilibrium values of the order parameters $\{R_a,q_{ab}\}$ as functions of α, one needs to assign some kind of structure to those matrices (as well as to the conjugate parameters $\{\hat{R}_a,\hat{q}_{ab}\}$) before proceeding with the calculation. This assignment amounts to an *ansatz*, whose marginal stability can be checked later. The simplest structure is that of the replica symmetric (RS) *ansatz*, namely

$$\begin{aligned} R_a &\overset{\mathrm{RS}}{=} R, \qquad a=1,\ldots,n \\ q_{ab} &\overset{\mathrm{RS}}{=} q, \qquad a<b \\ \hat{R}_a &\overset{\mathrm{RS}}{=} -\hat{R}, \qquad a=1,\ldots,n \\ \hat{q}_{ab} &\overset{\mathrm{RS}}{=} -\hat{q}, \qquad a<b\,. \end{aligned} \tag{B.16}$$

The physical meaning of R is the typical (normalized) overlap between $\boldsymbol{J}$ and $\boldsymbol{B}$, while q stands for the mutual overlap between different samples $\boldsymbol{J}$ and $\boldsymbol{J}'$.

B.2.1 The energy term

The energy term B.14 can be easily calculated under the RS assumption B.16. Making use of the identity

$$\int \mathcal{D}k\, e^{ixk} = e^{-x^2/2}\,, \tag{B.17}$$

where $\mathcal{D}k \equiv dk\,(2\pi)^{-1/2}\exp(-k^2/2)$ is a Gaussian measure, one can rewrite

$$\exp\left[-\frac{1}{2}(q-R^2)\left(\sum_a \hat{\lambda}_a\right)^2\right] = \int \mathcal{D}t_2 \exp\left[-it_2\sqrt{q-R^2}\sum_a \hat{\lambda}_a\right], \tag{B.18}$$

so that the integral on $\{\hat{\lambda}_a\}$ becomes

$$\int \mathcal{D}t_2 \int \prod_a \frac{d\hat{\lambda}_a}{2\pi} e^{-\sum_a \hat{\lambda}_a^2(1-q)/2 + i\sum_a \hat{\lambda}_a(\lambda_a - t_2\sqrt{q-R^2} - tR)} =$$
$$\int \mathcal{D}t_2 \prod_a \left[\frac{1}{\sqrt{2\pi(1-q)}}\exp -\frac{(\lambda_a - t_2\sqrt{q-R^2} - bR)^2}{2(1-q)}\right]. \tag{B.19}$$

At this point it should be noted that the RS *ansatz* requires $q \geq R^2$ for consistency. With the factorization of the replica index in eq. B.19, one arrives at

$$G_1 \overset{\text{RS}}{=} \ln \int \mathcal{D}^*b \int \mathcal{D}t_2 \left[\int \frac{d\lambda}{\sqrt{2\pi(1-q)}} \exp\Big(-\beta V(\lambda) - \frac{(\lambda - t_2\sqrt{q-R^2} - bR)^2}{2(1-q)}\Big)\right]^n, \tag{B.20}$$

for general n. Expanding $\langle x^n\rangle \overset{n\to 0}{\simeq} \langle 1 + n\ln x\rangle = 1 + n\langle \ln x\rangle \simeq \exp(n\langle \ln x\rangle) + \mathcal{O}(n^2)$, one finally gets

$$G_1(R,q;\beta,[V,U]) \overset{\text{RS}}{=} n \int \mathcal{D}^*b \int \mathcal{D}t_2 \ln \int \frac{d\lambda}{\sqrt{2\pi(1-q)}} \exp\Big(-\beta V(\lambda) - \frac{(\lambda - t_2\sqrt{q-R^2} - bR)^2}{2(1-q)}\Big). \tag{B.21}$$

B.2.2 The Ising measure

The entropy term G_0 is determined solely by the measure used in the $\boldsymbol{J}$ space. In this subsection the calculation is presented for the Ising constraint B.2:

$$
\begin{aligned}
e^{NG_0} &\overset{\text{RS}}{=} \prod_{j=1}^{N} \left\{ \int \prod_a d^b J_j^a \ \exp\left[\hat{R} \sum_a J_j^a B_j - \frac{\hat{q}}{2} \sum_a (J_j^a)^2 \right.\right. \\
&\qquad \left.\left. + \frac{\hat{q}}{2} \left(\sum_a J_j^a \right)^2 \right] \right\} \\
&\overset{(B.17)}{=} \prod_{j=1}^{N} \left\{ \int \mathcal{D}z \left[\int d^b J_j \ \exp\left(-J_j \left(z\sqrt{\hat{q}} - \hat{R} B_j \right) - \frac{\hat{q}}{2} J_j^2 \right) \right]^n \right\} \\
&= \prod_{j=1}^{N} \left\{ \int \mathcal{D}z \ \exp n \left(-\frac{\hat{q}}{2} + \ln\cosh\left(z\sqrt{\hat{q}} - \hat{R} B_j \right) \right) \right\} \\
&\overset{n\to 0}{\simeq} \prod_{j=1}^{N} \left\{ \exp n \ \left[-\frac{\hat{q}}{2} + \int \mathcal{D}z \ \ln\cosh\left(z\sqrt{\hat{q}} - \hat{R} B_j \right) \right] \right\} \\
&= \left\{ \exp n \ \left[-\frac{\hat{q}}{2} + \int \mathcal{D}z \ \ln\cosh\left(z\sqrt{\hat{q}} + \hat{R} \right) \right] \right\}^N , \qquad \text{(B.22)}
\end{aligned}
$$

where the last equality is justified by the invariance of the integral on z with respect to the choice of the jth component of $\boldsymbol{B}$ (as long as it remains binary, of course). The resulting expression for the entropy term is

$$
G_0 = n \left\{ -\frac{\hat{q}}{2} + \int \mathcal{D}z \ \ln\cosh\left(z\sqrt{\hat{q}} + \hat{R} \right) \right\} \qquad \text{(B.23)}
$$

which, together with eqs. B.15, B.16 and B.21, yields

$$
\begin{aligned}
f &= -\frac{1}{\beta} \underset{R,q,\hat{R},\hat{q}}{\text{Extr}} \left\{ -\frac{1}{2}(1-q)\hat{q} - \hat{R}R + \int \mathcal{D}z \ \ln\cosh\left(z\sqrt{\hat{q}} + \hat{R} \right) \right. \\
&\qquad \left. + \frac{\alpha}{n} G_1(R, q; \beta, [V, U]) \right\} . \qquad \text{(B.24)}
\end{aligned}
$$

B.2.3 The spherical measure

In order to calculate G_0 using eq. B.3, one needs to introduce Fourier representations of the δ-distributions for each replica:

$$
\begin{aligned}
e^{NG_0} &\overset{(B.16)}{=} \int \prod_a d\boldsymbol{J}^a \; \delta(\boldsymbol{J}^a \cdot \boldsymbol{J}^a - N) \; e^{\hat{R}B_j \sum_a J_j^a + \hat{q} \sum_{a<b} J_j^a J_j^b} \\
&= \int \prod_a \frac{dE_a}{2\pi} e^{-iN \sum_a E_a} \prod_{j=1}^{N} \Bigg\{ \int \prod_a dJ_j^a \; e^{i \sum_a E_a (J_j^a)^2} \\
&\qquad e^{\hat{R}B_j \sum_a J_j^a + (\hat{q}/2) \sum_{a \neq b} J_j^a J_j^b} \Bigg\}. \qquad \text{(B.25)}
\end{aligned}
$$

The set of conjugate variables $\{E_a\}$ plays an identical role to that of the variables $\{\hat{R}_a, \hat{q}_{ab}, R_a, q_{ab}\}$. The $\{E_a\}$ must also be extremized for the final expression of the free energy to be obtained. They will also be assumed to obey replica symmetry, that is, $E_a = E$, $a = 1, \ldots, n$, in which case the term between curly braces in eq. B.25 reads

$$
\begin{aligned}
&\int \mathcal{D}z \int \prod_a dJ_j^a \; e^{-(\hat{q}/2 - iE) \sum_a (J_j^a)^2 + (z\sqrt{\hat{q}} + \hat{R}B_j) \sum_a J_j^a} \\
&= \int \mathcal{D}z \left[\sqrt{\frac{2\pi}{\hat{q} - 2iE}} \exp \frac{(z\sqrt{\hat{q}} + \hat{R}B_j)^2}{2(\hat{q} - 2iE)} \right]^n \\
&\overset{n \to 0}{\simeq} \exp n \int \mathcal{D}z \; \ln \left[\sqrt{\frac{2\pi}{\hat{q} - 2iE}} \exp \frac{(z\sqrt{\hat{q}} + \hat{R}B_j)^2}{2(\hat{q} - 2iE)} \right] . \qquad \text{(B.26)}
\end{aligned}
$$

After the evaluation of the last Gaussian integral, one notes that the remaining dependence on the $\boldsymbol{B}$ components is only on $\sum_{j=1}^{N} B_j^2$, which equals N in every scenario considered in this work. Changing variables $\mathcal{E} = -2iE$, one may therefore rewrite the free energy as

$$
\begin{aligned}
f &= -\frac{1}{\beta} \underset{\hat{R}, \hat{q}, R, q, \mathcal{E}}{\text{Extr}} \Bigg\{ -\hat{R}R + \frac{1}{2}\hat{q}q + \frac{\mathcal{E}}{2} - \frac{1}{2} \ln(\hat{q} + \mathcal{E}) \\
&\qquad + \frac{\hat{q} + \hat{R}^2}{2(\hat{q} + \mathcal{E})} + \frac{\alpha}{n} G_1(R, q; \beta, [V, U]) \Bigg\} . \qquad \text{(B.27)}
\end{aligned}
$$

The extremization with respect to the conjugate variables $\hat{R}, \hat{q}$ and $\mathcal{E}$ can be carried out algebraically:

$$
\frac{\partial f}{\partial \hat{R}} = 0 \quad \Rightarrow \quad \hat{R} = R(\hat{q} + \mathcal{E})
$$

$$\begin{aligned}\frac{\partial f}{\partial \hat{q}} = 0 \quad &\Rightarrow \quad \frac{\hat{q}}{(\hat{q}+\mathcal{E})^2} = q - R^2 \\ \frac{\partial f}{\partial \mathcal{E}} = 0 \quad &\Rightarrow \quad \hat{q} + \mathcal{E} = \frac{1}{1-q} \; . \end{aligned} \tag{B.28}$$

Inserting the results B.28 back into B.27 one arrives at

$$f = -\frac{1}{\beta} \underset{R,q}{\text{Extr}} \left\{ \frac{1}{2}\ln(1-q) + \frac{q-R^2}{2(1-q)} + \frac{\alpha}{n} G_1(R,q;\beta,[V,U]) \right\} \; , \tag{B.29}$$

with G_1 given by B.21.

B.3 The entropy for a binary measure

In order to calculate eq. 1.18, one first notes that the free energy is written as an explicit function of β but also as an implicit function thereof, through its dependence on the order parameters (whose equilibrium values are also determined by β), that is, $f = f\left(\beta, R(\alpha;\beta), q(\alpha;\beta), \hat{R}(\alpha;\beta), \hat{q}(\alpha;\beta)\right)$. However, the implicit dependence does not contribute to the entropy since the order parameters (and their conjugate parameters) are located at saddle points of f:

$$\begin{aligned}\frac{df}{d\beta} \;=\; & \left.\frac{\partial f}{\partial \beta}\right|_{R,q,\hat{R},\hat{q}} \\ & + \overbrace{\left.\frac{\partial f}{\partial R}\right|_{\beta,q,\hat{R},\hat{q}}}^{=0} \frac{\partial}{\partial \beta} R(\alpha;\beta) + \overbrace{\left.\frac{\partial f}{\partial q}\right|_{\beta,R,\hat{R},\hat{q}}}^{=0} \frac{\partial}{\partial \beta} q(\alpha;\beta) \\ & + \overbrace{\left.\frac{\partial f}{\partial \hat{R}}\right|_{\beta,R,q,\hat{q}}}^{=0} \frac{\partial}{\partial \beta} \hat{R}(\alpha;\beta) + \overbrace{\left.\frac{\partial f}{\partial \hat{q}}\right|_{\beta,R,q,\hat{R}}}^{=0} \frac{\partial}{\partial \beta} \hat{q}(\alpha;\beta) \; . \end{aligned} \tag{B.30}$$

With this remark taken into account, one can proceed to calculate, according to eqs. B.24 and B.21,

$$\frac{\partial f}{\partial \beta} = -\frac{f}{\beta} + \frac{\alpha}{\beta} \int \mathcal{D}^* b \int \mathcal{D}t_2 \left\{ \frac{\int d\lambda \, V(\lambda) \exp\left[-\beta V(\lambda) - \frac{(\lambda - u\sqrt{q})^2}{2(1-q)}\right]}{\int d\lambda \, \exp\left[-\beta V(\lambda) - \frac{(\lambda - u\sqrt{q})^2}{2(1-q)}\right]} \right\} \; , \tag{B.31}$$

where $u = u(b, t_2) = (bR + t_2\sqrt{q - R^2})/\sqrt{q}$. Putting together eqs. 1.18 and B.31, one obtains:

$$s = \ln 2 - \beta f + \beta\alpha \int \mathcal{D}^* b \int \mathcal{D}t_2 \left\{ \frac{\int d\lambda\, V(\lambda) \exp\left[-\beta V(\lambda) - \frac{(\lambda - u\sqrt{q})^2}{2(1-q)}\right]}{\int d\lambda\, \exp\left[-\beta V(\lambda) - \frac{(\lambda - u\sqrt{q})^2}{2(1-q)}\right]} \right\}, \tag{B.32}$$

where one should not forget that all terms above must be evaluated at values $R(\alpha;\beta), q(\alpha;\beta), \hat{R}(\alpha;\beta), \hat{q}(\alpha;\beta)$ which extremize the free energy f.

Appendix C

Technical details of Gibbs learning

C.1 Rewriting the free energy

This appendix starts by rewriting the expression of the free energy B.24 (i.e. for the Ising measure) in a form which is more convenient for the study of Gibbs learning. An orthogonal transformation

$$\left.\begin{array}{l} b(u,v;R) = \frac{R}{\sqrt{q}}u + \sqrt{\frac{q-R^2}{q}}v \\ t_2(u,v;R) = \sqrt{\frac{q-R^2}{q}}u - \frac{R}{\sqrt{q}}v \end{array}\right\} \quad \Leftrightarrow \quad \left\{\begin{array}{l} u(b,t_2;R) = \frac{R}{\sqrt{q}}b + \sqrt{\frac{q-R^2}{q}}t_2 \\ v(b,t_2;R) = \sqrt{\frac{q-R^2}{q}}b - \frac{R}{\sqrt{q}}t_2 \end{array}\right.$$

$$\begin{aligned} db\, dt_2 &= du\, dv \\ b^2 + t_2^2 &= u^2 + v^2 \end{aligned} \tag{C.1}$$

implying

$$\begin{aligned} &\int \mathcal{D}b\, e^{-U(b)} \int \mathcal{D}t_2\, f(b,t_2) = \\ &\quad \int \mathcal{D}u \int \mathcal{D}v\, e^{-U\left(\frac{R}{\sqrt{q}}u + \sqrt{\frac{q-R^2}{q}}v\right)} f\left(\frac{R}{\sqrt{q}}u + \sqrt{\frac{q-R^2}{q}}v,\right. \\ &\quad \left.\sqrt{\frac{q-R^2}{q}}u - \frac{R}{\sqrt{q}}v\right), \end{aligned} \tag{C.2}$$

allows one to rewrite the general expression for the energy term B.21 as

$$n^{-1}G_1(R,q;\beta,[V,U]) = \int \mathcal{D}u\, X\left(u;\frac{R}{\sqrt{q}}\right) \ln \int \mathcal{D}z\, e^{-\beta V(z\sqrt{1-q}+u\sqrt{q})} \ , \quad \text{(C.3)}$$

where

$$X(t;R) = \mathcal{N}\int \mathcal{D}t'\, e^{-U(Rt+\sqrt{1-R^2}t')} \ . \quad \text{(C.4)}$$

In the specific case of Gibbs learning ($V = U$ and $\beta = 1$), eq. C.3 becomes

$$n^{-1}G_1(R,q;1,[U,U]) = \int \mathcal{D}u\, X\left(u;\frac{R}{\sqrt{q}}\right) \ln\left[\frac{X(u;\sqrt{q})}{\mathcal{N}}\right] \ . \quad \text{(C.5)}$$

Finally, one should note that the term above involving $\ln\mathcal{N}$ is irrelevant as far as the saddle point equations are concerned, since $\int \mathcal{D}u\, X(u,R) = 1$. However, this additive constant is important for the calculation of the entropy, so it must be kept. The free energy for Gibbs learning reads then

$$\begin{aligned} f \ = \ & \underset{R,q,\hat{R},\hat{q}}{\text{Extr}} \left\{ \frac{1}{2}(1-q)\hat{q} + \hat{R}R - \int \mathcal{D}z\ \ln\cosh\left(z\sqrt{\hat{q}}+\hat{R}\right) \right. \\ & \left. -\alpha \int \mathcal{D}u\, X\left(u;\frac{R}{\sqrt{q}}\right) \ln\left[\frac{X(u;\sqrt{q})}{\mathcal{N}}\right] \right\} . \end{aligned} \quad \text{(C.6)}$$

C.1.1 The saddle point equations

The extremum operator in eq. C.6 gives rise to saddle point equations for the variables $R, q, \hat{R}$ and $\hat{q}$. Here it is shown that the *ansatz* $R = q$, $\hat{R} = \hat{q}$ is consistent. Starting with the easier saddle point equations, eq. C.6 immediately leads to

$$\begin{aligned} \frac{\partial \hat{f}}{\partial \hat{R}} = 0 \ \ &\Rightarrow \ \ R = \int \mathcal{D}z\ \tanh\left(z\sqrt{\hat{q}}+\hat{R}\right) \\ \frac{\partial \hat{f}}{\partial \hat{q}} = 0 \ \ &\Rightarrow \ \ q = \int \mathcal{D}z\ \tanh^2\left(z\sqrt{\hat{q}}+\hat{R}\right) \ . \end{aligned} \quad \text{(C.7)}$$

The derivatives with respect to R and q require some extra steps and the calculations are much simplified if one makes extensive use of the orthogonal transformation C.1. The axes (u,v) and (b,t_2) will be interchanged often in the passages below, and special care has been taken with the notation: note

for instance that while t_2 is a variable in the system (b, t_2), $t_2(u, v; R)$ is a function in the system (u, v) (see C.1). It is also useful to make use of the function

$$\begin{aligned} Y(t;R) &= \frac{\mathcal{N}}{\sqrt{1-R^2}} \int \mathcal{D}t'\, t'\, e^{-U(Rt+\sqrt{1-R^2}t')} \\ &= \mathcal{N} \int \mathcal{D}t'\, e^{-U(Rt+\sqrt{1-R^2}t')} \left[-U'(Rt+\sqrt{1-R^2}t')\right] \ , \end{aligned} \tag{C.8}$$

which relates to $X(t;R)$ like

$$Y(t;R) = \frac{1}{R}\frac{\partial}{\partial t} X(t;R) \ . \tag{C.9}$$

Starting with the derivative of C.5 with respect to R, one has

$$\begin{aligned} n^{-1}\frac{\partial G_1}{\partial R} &= \frac{\mathcal{N}}{\sqrt{q-R^2}} \int \mathcal{D}u\, \ln X(u;\sqrt{q}) \int \mathcal{D}v\, e^{-U(b(u,v;R))} \\ &\quad \times [-U'(b(u,v;R))]\, t_2(u,v;R) \\ &= \frac{\mathcal{N}}{\sqrt{q-R^2}} \int \mathcal{D}b\, e^{-U(b)} [-U'(b)] \int \mathcal{D}t_2 \\ &\quad \times \frac{\partial}{\partial t_2} \ln X(u(b,t_2;R);\sqrt{q}) \\ &= \mathcal{N} \int \mathcal{D}u \int \mathcal{D}v\, e^{-U(b(u,v;R))} [-U'(b(u,v;R))] \frac{Y(u;\sqrt{q})}{X(u;\sqrt{q})} \\ &= \int \mathcal{D}u\, \frac{Y(u;R/\sqrt{q})\, Y(u;\sqrt{q})}{X(u;\sqrt{q})} \ . \end{aligned} \tag{C.10}$$

The derivative with respect to q has two terms,

$$n^{-1}\frac{\partial G_1}{\partial q} = \frac{-R}{2q}\left(n^{-1}\frac{\partial G_1}{\partial R}\right) + \int \mathcal{D}u\, \frac{X(u;R/\sqrt{q})}{X(u;q)} \frac{\partial}{\partial q} X(u;\sqrt{q}) \ , \tag{C.11}$$

where the second term above equals

$$\int \mathcal{D}u\, \frac{X(u;R/\sqrt{q})}{X(u;q)} \frac{\partial}{\partial q} X(u;\sqrt{q}) = \int \mathcal{D}u\, \frac{X(u;R/\sqrt{q})}{X(u;q)}$$

$$
\begin{aligned}
& \times \frac{\mathcal{N}}{(-2q^{3/2})} \int \mathcal{D}v \, e^{-U(b(u,v;\sqrt{q}))} \left[-U'(b(u,v;\sqrt{q}))\right] \frac{t_2(u,v;\sqrt{q})}{\sqrt{1-q}} \\
= \quad & \frac{\mathcal{N}}{(-2q^{3/2})\sqrt{1-q}} \int \mathcal{D}b \, e^{-U(b)} \left[-U'(b)\right] \\
& \times \int \mathcal{D}t_2 \, t_2 \frac{X(u(b,t_2;\sqrt{q});R/\sqrt{q})}{X(u(b,t_2;\sqrt{q});\sqrt{q})} \, . \qquad \text{(C.12)}
\end{aligned}
$$

At this point it is possible to check the consistency of the *ansatz*

$$
\begin{aligned}
R &= q \\
\hat{R} &= \hat{q} \, . \qquad \text{(C.13)}
\end{aligned}
$$

Notice the last integral on t_2 in eq. C.12. If one sets $R = q$, the ratio $X(u(b,t_2;\sqrt{q});R/\sqrt{q})/X(u(b,t_2;\sqrt{q});\sqrt{q})$ equals one and the integral $\int \mathcal{D}t_2 \, t_2$ vanishes identically. Therefore one obtains

$$
\begin{aligned}
\frac{\partial \hat{f}}{\partial R} = 0 \quad &\overset{R=q}{\Rightarrow} \quad \hat{R} = \alpha \int \mathcal{D}u \, \frac{Y^2(u;\sqrt{q})}{X(u;\sqrt{q})} \\
\frac{\partial \hat{f}}{\partial q} = 0 \quad &\overset{R=q}{\Rightarrow} \quad \hat{q} = \alpha \int \mathcal{D}u \, \frac{Y^2(u;\sqrt{q})}{X(u;\sqrt{q})} \, . \qquad \text{(C.14)}
\end{aligned}
$$

Having thus checked that $R = q \Rightarrow \hat{R} = \hat{q}$, it remains to prove the opposite, namely, that $\hat{R} = \hat{q} \Rightarrow R = q$. For that purpose, one should go back to eqs. C.7. Imposing $\hat{R} = \hat{q}$, note that it is sufficient to show that the difference below is null, a proof that I owe to Prof. M. Bouten:

$$
\begin{aligned}
& \int \mathcal{D}z \, \tanh(az+a^2) - \int \mathcal{D}z \, \tanh^2(az+a^2) = \\
& \qquad \int \mathcal{D}z \, \tanh(az+a^2) \left[1 - \tanh(az+a^2)\right] \\
= \quad & \int \mathcal{D}z \, \tanh(az+a^2) \frac{e^{-az-a^2}}{\cosh(az+a^2)} \\
= \quad & \int \frac{dz}{\sqrt{2\pi}} e^{-z^2/2-az-a^2} \frac{\sinh(az+a^2)}{\cosh^2(az+a^2)} \\
\overset{y=z+a}{=} \quad & \frac{e^{-a^2/2}}{\sqrt{2\pi}} \int dy \, e^{-y^2/2} \frac{\sinh(ay)}{\cosh^2(ay)} \\
= \quad & 0 \, . \qquad \text{(C.15)}
\end{aligned}
$$

With the consistency of the *ansatz* proven, the free energy C.6 can be rewritten as

$$f(R,\hat{R}) = \underset{R,\hat{R}}{\text{Extr}} \left\{ \frac{(1+R)\hat{R}}{2} - \int \mathcal{D}z \, \ln\cosh\left(z\sqrt{\hat{R}} + \hat{R}\right) - \alpha \int \mathcal{D}t \, X\left(t;\sqrt{R}\right) \ln\left[\frac{X(u;\sqrt{R})}{\mathcal{N}}\right] \right\} . \quad \text{(C.16)}$$

C.2 The entropy

In order to obtain the entropy for Gibbs learning, one should initially focus on the last term of the general expression, eq. B.32. It can be much simplified with the orthogonal transformations A.18 if the result C.13 is also taken into account:

$$\begin{aligned}
& \int \mathcal{D}^* b \int \mathcal{D}t_2 \left\{ \frac{\int d\lambda \, V(\lambda) \exp\left[-\beta V(\lambda) - \frac{(\lambda - u(b,t_2)\sqrt{q})^2}{2(1-q)}\right]}{\int d\lambda \, \exp\left[-\beta V(\lambda) - \frac{(\lambda - u(b,t_2)\sqrt{q})^2}{2(1-q)}\right]} \right\} \\
= & \int \mathcal{D}u \, X(u;R/\sqrt{q}) \left\{ \frac{\int \mathcal{D}z \, U\left(z\sqrt{1-q} + u\sqrt{q}\right) \, e^{-U(z\sqrt{1-q}+u\sqrt{q})}}{\int \mathcal{D}z \, e^{-U(z\sqrt{1-q}+u\sqrt{q})}} \right\} \\
= & \mathcal{N} \int \mathcal{D}u \int \mathcal{D}z \, U\left(z\sqrt{1-q} + u\sqrt{q}\right) \exp -U\left(z\sqrt{1-q} + u\sqrt{q}\right) \\
\overset{(A.18)}{=} & \mathcal{N} \int \mathcal{D}b \, e^{-U(b)} \int \mathcal{D}t_2 \, U(b) \\
= & \langle U(b) \rangle_* \ . \qquad \text{(C.17)}
\end{aligned}$$

Collecting all the terms in eq. B.32, one obtains

$$\begin{aligned}
s(\alpha) &= \ln 2 - f(R_G, \hat{R}_G) + \alpha \langle U(b) \rangle_* \\
&= -\frac{(1+R_G)\hat{R}_G}{2} + \int \mathcal{D}z \, \ln 2\cosh\left(z\sqrt{\hat{R}_G} + \hat{R}_G\right) \\
&\quad + \alpha \int \mathcal{D}t \, X\left(t;\sqrt{R_G}\right) \ln\left[\frac{X(t;\sqrt{R_G})}{\mathcal{N}}\right] + \alpha \langle U(b) \rangle_* \ , \quad \text{(C.18)}
\end{aligned}$$

where $R_G(\alpha)$ and $\hat{R}_G(\alpha)$ minimize the free energy C.16.

Appendix D

$P(y)$ without the ME formalism

D.1 The natural rise of an extremum principle

In this appendix the main result of chapter 4 (eq. 4.25) is obtained in a different way. Here, instead of making use of the ME formalism, expression 4.4 is calculated explicitly:

$$\begin{aligned} \mathcal{C}_n P(y) &= \int \left[\prod_a^n d\boldsymbol{J}^a P_b(\boldsymbol{J}^a) \delta(\boldsymbol{J}^a \cdot \boldsymbol{B} - NR) \right] \\ &\times \left[\prod_{a<b} \delta(\boldsymbol{J}^a \cdot \boldsymbol{J}^b - Nq) \right] \delta \left(y - B_1 n^{-1} \sum_a^n J_1^a \right) , \end{aligned} \tag{D.1}$$

where the normalization constant $\mathcal{C}_n$ is determined by imposing $\int P(y)\, dy = 1$ and P_b is again the binary measure.

After the Fourier representation of the δ-distributions are introduced (precisely the same way as done in appendix B), eq. D.1 reads (apart from the normalization constant)

$$\begin{aligned} P(y) \sim & \int \frac{d\hat{y}}{2\pi} e^{i\hat{y}y} \left[\int \prod_a \frac{d\hat{R}_a}{2\pi} \right] \left[\int \prod_{a<b} \frac{d\hat{q}_{ab}}{2\pi} \right] e^{-iNR\sum_a \hat{R}_a - iNq\sum_{a<b}\hat{q}_{ab}} \\ & \times \exp N\zeta \left(\left\{ \hat{R}_a, \hat{q}_{ab} \right\}, \hat{y} \right) , \end{aligned} \tag{D.2}$$

where

$$\zeta\left(\left\{\hat{R}_a,\hat{q}_{ab}\right\},\hat{y}\right) \equiv \frac{1}{N}\ln\int\left[\prod_a d\boldsymbol{J}^a\, P_b(\boldsymbol{J}^a)\right]\exp\left[i\sum_a \hat{R}_a\boldsymbol{J}^a\cdot\boldsymbol{B} + i\sum_{a<b}\hat{q}_{ab}\boldsymbol{J}^a\cdot\boldsymbol{J}^b - i(\hat{y}/n)B_1\sum_a J_1^a\right] . \quad \text{(D.3)}$$

The first thing one should note is that ζ also depends parametrically on the preferential direction $\boldsymbol{B}$. This explicit dependence was omitted for the sake of clarity, but will become important a few steps ahead. The second thing to be noted is that ζ should be $\mathcal{O}(1)$, since the vectors in the exponential are N-dimensional.

In the thermodynamic limit $N\to\infty$, eq. D.2 is governed by the saddle point of the exponentially dominant term, namely

$$P(y)\sim\int\frac{d\hat{y}}{2\pi}e^{i\hat{y}y+N\zeta\left(\left\{\hat{R}_a^*,\hat{q}_{ab}^*\right\},\hat{y}\right)} , \quad \text{(D.4)}$$

where

$$\hat{R}_a^* = \underset{\hat{R}_a}{\text{Arg Extr}}\left[-iR\sum_a\hat{R}_a - iq\sum_{a<b}\hat{q}_{ab} + \zeta\left(\left\{\hat{R}_a,\hat{q}_{ab}\right\},\hat{y}\right)\right]$$

$$\hat{q}_{ab}^* = \underset{\hat{q}_{ab}}{\text{Arg Extr}}\left[-iR\sum_a\hat{R}_a - iq\sum_{a<b}\hat{q}_{ab} + \zeta\left(\left\{\hat{R}_a,\hat{q}_{ab}\right\},\hat{y}\right)\right] . \quad \text{(D.5)}$$

Note that *without imposing the principle of Maximum-Entropy, one is again left with an extremization procedure.* In this sense, eqs. D.5 are analogous to eqs. 4.9 or 4.16. Here the extremum is imposed by the thermodynamic limit, while in the ME formalism it arises from the requirement that a number of constraints be satisfied (recall sections 4.2.1 and 4.2.2). Since the constraints 4.11 are satisfied *only* in the thermodynamic limit, there is no contradiction. Much on the contrary, this agreement shows how the ME formalism very elegantly accounts for the relevant features brought by the thermodynamic limit, giving the same results but in a much simplified way.

D.2 The RS *ansatz*

Noticing that the integrals in $e^{N\zeta}$ can be factorized into N components, one rewrites

$$\begin{aligned}\exp N\zeta &= \Psi_{n1}\left(\hat{R}_a - \frac{y}{n}, \hat{q}_{ab}\right)\prod_{i\neq 1}\Psi_{ni}\left(\hat{R}_a, \hat{q}_{ab}\right)\\ \Psi_{ni}\left(\hat{R}_a, \hat{q}_{ab}\right) &\equiv \int\left[\prod_{a=1}^{n} d^b J_i^a\right] e^{iB_i\sum_a \hat{R}_a J_i^a + i\sum_{a<b}\hat{q}_{ab}J_i^a J_i^b}\ .\end{aligned} \tag{D.6}$$

At this point it is possible to invoke the same reasoning given in section 4.2.2 and impose the replica symmetric (RS) *ansatz*

$$\begin{aligned}\hat{R}_a &\overset{RS}{=} -i\hat{R}\\ \hat{q}_{ab} &\overset{RS}{=} -i\hat{q}\ .\end{aligned} \tag{D.7}$$

The evaluation of $\Psi_{ni}\left(\hat{R}_a, \hat{q}_{ab}\right)$ becomes much easier and one finds

$$\Psi_{ni}\left(-i\hat{R}, -i\hat{q}\right) = e^{-n\hat{q}/2}\int \mathcal{D}z\ \left[\cosh\left(\hat{R}B_i + z\sqrt{\hat{q}}\right)\right]^n\ . \tag{D.8}$$

Notice now that $\Psi_{ni}\left(-i\hat{R}, -i\hat{q}\right)$ is invariant with respect to the transformation $B_i \to -B_i$, which means that *the dependence on the preferential direction* $\boldsymbol{B}$ *disappears as long as it is a binary vector.* Comparing now eq. D.8 with 4.15, one further concludes that

$$\Psi_{ni}\left(-i\hat{R}, -i\hat{q}\right) = Z_{ME}\left(\hat{R}, \hat{q}\right)\ . \tag{D.9}$$

It should not be surprising then that the extremum operations D.5 render exactly the same equations for the RS conjugate parameters as the ones obtained via the ME formalism. In order to show this, one first rewrites ζ, taking into account eq. D.8:

$$\zeta\left(-i\hat{R}, -i\hat{q}, \hat{y}\right) = \ln\Psi_{ni}\left(-i\hat{R}, -i\hat{q}\right) + \frac{1}{N}\ln\left[\frac{\Psi_{ni}\left(-i\hat{R} - \frac{\hat{y}}{n}, -i\hat{q}\right)}{\Psi_{ni}\left(-i\hat{R}, -i\hat{q}\right)}\right]\ . \tag{D.10}$$

When $N \to \infty$, only the first term above contributes to the extremization D.5, which now reads

$$\begin{aligned}\hat{R}^* &= \operatorname*{Arg\,Extr}_{\hat{R}}\left[-nR\hat{R} - \frac{n(n-1)}{2}q\hat{q} + \ln\Psi_{ni}\left(-i\hat{R}, -i\hat{q}\right)\right]\\ \hat{q}^* &= \operatorname*{Arg\,Extr}_{\hat{q}}\left[-nR\hat{R} - \frac{n(n-1)}{2}q\hat{q} + \ln\Psi_{ni}\left(-i\hat{R}, -i\hat{q}\right)\right]\ .\end{aligned} \tag{D.11}$$

Since $\Psi_{ni}\left(-i\hat{R},-i\hat{q}\right) = Z_{ME}\left(\hat{R},\hat{q}\right)$, it follows immediately that the extremum equations above are identical to 4.16, 4.17.

Going back to eq. D.4, the relevant contribution to $P(y)$ comes now from the *second* term of eq. D.10, since the first one is just a constant with respect to $\hat{y}$. Moreover, the denominator in the argument of the logarithm makes sure that $P(y)$ is properly normalized, leading to

$$\begin{aligned} P(y) &= \frac{1}{\Psi_{ni}\left(-i\hat{R},-i\hat{q}\right)} \int \frac{d\hat{y}}{2\pi}\, e^{iy\hat{y}}\, \Psi_{ni}\left(-i\hat{R}-\frac{\hat{y}}{n},-i\hat{q}\right) \\ &= \frac{e^{-n\hat{q}/2}}{Z_{ME}} \int \frac{d\hat{y}}{2\pi}\, e^{iy\hat{y}} \int \mathcal{D}z\, \left[\cosh\left(\hat{R}-\frac{i\hat{y}}{n}+z\sqrt{\hat{q}}\right)\right]^n \quad , \text{(D.12)} \end{aligned}$$

where for simplicity of notation $\hat{R}^*$ and $\hat{q}^*$ were replaced by $\hat{R}$ and $\hat{q}$ (but should be taken at their equilibrium value, determined by eq. D.11).

For the sake of completeness a proof is given below that eqs. D.12 and 4.18 are actually identical. Introducing dummy variables $\{x_a\} \in \{-1,+1\}^n$, $a = 1,\ldots,n$, one rewrites

$$\begin{aligned} \frac{Z_{ME}P(y)}{e^{-n\hat{q}/2}} &= \int \frac{d\hat{y}}{2\pi}\, e^{iy\hat{y}} \int \mathcal{D}z\, \left[\int d^b x\; e^{x\left(\hat{R}-\frac{i\hat{y}}{n}+z\sqrt{\hat{q}}\right)}\right]^n \\ &= \int \frac{d\hat{y}}{2\pi}\, e^{iy\hat{y}} \int \mathcal{D}z \int \left[\prod_{a=1}^{n} d^b x_a\right] e^{\sum_a x_a\left(\hat{R}-\frac{i\hat{y}}{n}+z\sqrt{\hat{q}}\right)} \\ &= \int \mathcal{D}z \int \left[\prod_{a=1}^{n} d^b x_a\right] e^{\sum_a x_a\left(\hat{R}+z\sqrt{\hat{q}}\right)} \delta\left(y-\frac{1}{n}\sum_a x_a\right) \\ &= \int \left[\prod_{a=1}^{n} d^b x_a\right] e^{\hat{R}\sum_a x_a+\frac{\hat{q}}{2}\left(\sum_a x_a\right)^2} \delta\left(y-\frac{1}{n}\sum_a x_a\right) \quad \text{(D.13)} \end{aligned}$$

At this stage one can interpret the term $e^{\hat{R}\sum_a x_a+\frac{\hat{q}}{2}\left(\sum_a x_a\right)^2}$ as $P(\{x_a\})$ (apart from the normalization constant $e^{-n\hat{q}/2}/Z_{ME}$). Since $P(\{x_a\})$ depends only on $\sum_a x_a = ny$, one just needs to weigh each occurrence of y with the number $n!/((n(1+y)/2)!((n(1-y)/2))!)$ of possible sums that can amount to it. This finally yields, apart from a normalization constant,

$$P(y) \sim e^{n\hat{R}y+n^2y^2\hat{q}/2} \left(\begin{array}{c} n \\ \frac{n(1+y)}{2} \end{array} \right) , \qquad \text{(D.14)}$$

which is identical to eq. 4.18.

D.3 Recovering the peaked distribution

Starting from eq. D.12 one can deduce in a slightly different way the delta-peaked distribution 4.25 which arises in the limit $n \to \infty$. The rescaled conjugate parameters

$$\begin{aligned} \rho_n &\equiv n\,\hat{R}_n \\ \gamma_n &\equiv n\,\hat{q}_n \end{aligned} \tag{D.15}$$

must be determined by the solution of the saddle point equations D.11, which read

$$\begin{aligned} R &= \frac{\int du\, e^{-n\phi_n} \sinh(u\rho_n/\gamma_n) \tanh u}{\int du\, e^{-n\phi_n} \cosh(u\rho_n/\gamma_n)} \\ q &= \frac{\int du\, e^{-n\phi_n} \cosh(u\rho_n/\gamma_n) \tanh^2 u}{\int du\, e^{-n\phi_n} \cosh(u\rho_n/\gamma_n)}\,, \end{aligned} \tag{D.16}$$

$$\phi_n(u) \equiv \frac{u^2}{2\gamma_n} - \ln\cosh u\,. \tag{D.17}$$

In the limit $n \to \infty$ the integrals above are determined by the minima of ϕ_n. Since ϕ_n is an even function, two solutions $\pm u_0 = \mathrm{Argmin}_u\, \phi_n(u)$ exist (depending on the value of γ_∞), satisfying $u_0 = \gamma_\infty \tanh u_0$. Using this result and applying the Laplace method, eqs. D.16 become

$$\begin{aligned} R &\overset{n\to\infty}{=} \tanh(u_0\rho_\infty/\gamma_\infty) \tanh u_0 \\ q &\overset{n\to\infty}{=} \tanh^2 u_0\,, \end{aligned} \tag{D.18}$$

which can be explicitly solved:

$$\begin{aligned} \gamma_\infty &= \frac{u_0}{\tanh u_0} = \frac{\mathrm{arctanh}\sqrt{q}}{\sqrt{q}} \\ \rho_\infty &= \frac{\mathrm{arctanh}(R/\sqrt{q})}{\sqrt{q}}\,. \end{aligned} \tag{D.19}$$

Note that the symmetry $u_0 \to -u_0$ is associated with the symmetry $R \to -R$, so that $\rho_\infty(-R,q) = -\rho_\infty(R,q)$.

With the problem of determining the equilibrium values of γ_∞ and ρ_∞ solved, one may return to eq. D.12, rewrite it in terms of the rescaled parameters and make a change of variables:

$$
\begin{aligned}
P(y) &= \frac{e^{-\gamma_n/2}}{Z_{ME}} \int \frac{d\hat{y}}{2\pi}\, e^{iy\hat{y}} \int \mathcal{D}z \left[\cosh\left(\frac{\rho_n - i\hat{y}}{n} + z\sqrt{\frac{\gamma_n}{n}}\right)\right]^n \\
&= \frac{e^{-\gamma_n/2}}{Z_{ME}} \int \frac{d\hat{y}}{2\pi}\, e^{iy\hat{y}} \int \frac{du}{\sqrt{2\pi}}\, (\cosh u)^n \exp -\frac{n}{2\gamma_n}\left[u - \left(\frac{\rho_n - i\hat{y}}{n}\right)\right]^2 \\
&= \frac{e^{-\gamma_n/2}}{Z_{ME}} \int \frac{d\hat{y}}{2\pi}\, e^{iy\hat{y}} \exp \frac{(\rho_n - i\hat{y})^2}{2\gamma_n n} \\
&\quad \times \int \frac{du}{\sqrt{2\pi}}\, e^{-n\phi_n} \exp \frac{u}{\gamma_n}(\rho_n - i\hat{y}) \,. \qquad \text{(D.20)}
\end{aligned}
$$

In the limit $n \to \infty$ the term $\exp(\rho_n - i\hat{y})^2/(2\gamma_n n)$ becomes negligible and the integral on u is again dominated by the minima of ϕ_n. Note that now it is important to keep track of both minima $\pm u_0$, since this will determine the correct normalization of $P(y)$:

$$
\begin{aligned}
P(y) &\overset{n\to\infty}{\simeq} \frac{\int du\, e^{-n\phi_n} \exp(u\rho_\infty/\gamma_\infty)\delta\left(y - \frac{u}{\gamma_\infty}\right)}{\int du\, e^{-n\phi_n} \exp(u\rho_\infty/\gamma_\infty)} \\
&\simeq \frac{\exp(u_0\gamma_\infty/\rho_\infty)\delta\left(y - \frac{u_0}{\gamma_\infty}\right) + \exp(-u_0\gamma_\infty/\rho_\infty)\delta\left(y + \frac{u_0}{\gamma_\infty}\right)}{2\cosh(u_0\gamma_\infty/\rho_\infty)} \\
&\overset{(D.19)}{=} \frac{1}{2}\left(1 + \frac{R}{\sqrt{q}}\right)\delta\left(y - \sqrt{q}\right) + \frac{1}{2}\left(1 - \frac{R}{\sqrt{q}}\right)\delta\left(y + \sqrt{q}\right) \,. \qquad \text{(D.21)}
\end{aligned}
$$

This result is identical to the one obtained via the ME formalism.

Appendix E

Quenched moments

E.1 General results

In this appendix the m-th *quenched moment* of a single component of the $\boldsymbol{J}$ vector will be calculated in detail. It is defined as the *average over the disorder* of the m-th power of the *thermal average* of J_1 (the result is expected to be independent of the index, due to permutation symmetry among the axes). In other words, the quantity to be calculated is

$$\langle \langle J_1 \rangle_{\boldsymbol{J}}^m \rangle_{D|\boldsymbol{B}} = \left\langle Z^{-m} \left(\int dm(\boldsymbol{J})\, e^{-\beta \mathcal{H}(\boldsymbol{J},D)} J_1 \right)^m \right\rangle_{D|\boldsymbol{B}} , \qquad \text{(E.1)}$$

where $Z = Z(D)$ is given by eq. B.1 and $dm(\boldsymbol{J}) = d\boldsymbol{J}\, P(\boldsymbol{J})$ determines the measure on $\boldsymbol{J}$ space. The innermost brackets denote the thermal average, while the outermost are the average over the disorder. This expression can be calculated using the same techniques as those described in section 1.2 and employed in appendix B. In this case, the replica trick is applied the following way:

$$\begin{aligned}
\langle \langle J_1 \rangle_{\boldsymbol{J}}^m \rangle_{D|\boldsymbol{B}} &= \lim_{n\to 0} \left\langle Z^{n-m} \left(\int dm(\boldsymbol{J})\, e^{-\beta \mathcal{H}(\boldsymbol{J},D)} J_1 \right)^m \right\rangle_{D|\boldsymbol{B}} \\
&= \lim_{n\to 0} \left\langle \int \left(\prod_{a=1}^{n} dm(\boldsymbol{J}^a) \right) e^{-\beta \sum_a^n \mathcal{H}(\boldsymbol{J}^a,D)} \left(\prod_{\nu=1}^{m} J_1^\nu \right) \right\rangle_{D|\boldsymbol{B}} \\
&= \lim_{n\to 0} \int \left(\prod_{a=1}^{n} dm(\boldsymbol{J}^a) \right) \left(\prod_{\nu=1}^{m} J_1^\nu \right) \\
&\quad \times \left\langle e^{-\beta \sum_a^n \mathcal{H}(\boldsymbol{J}^a,D)} \right\rangle_{D|\boldsymbol{B}} . \qquad \text{(E.2)}
\end{aligned}$$

The above equation is very similar to eq. B.7, and obviously reduces to it when $m = 0$. One can immediately recognize the average over the disorder in eq. E.2 as the energy term G_1 (see eq. B.14). The only difference is that the order parameters $\{R_a, q_{ab}\}$ have not yet been introduced, so the correct way to write it is

$$\left\langle e^{-\beta \sum_a^n \mathcal{H}(\boldsymbol{J}^a, D)} \right\rangle_{D|\boldsymbol{B}} = \exp\left[\alpha N G_1\left(\left\{\frac{\boldsymbol{J}^a \cdot \boldsymbol{B}}{N}, \frac{\boldsymbol{J}^a \cdot \boldsymbol{J}^b}{N}\right\}; \beta, [V, U]\right)\right] \quad \text{(E.3)}$$

The difference between eq. E.2 and eq. B.7 is that m out of n integrals have an extra term J_1^ν, which means that only the entropy term changes. Introducing delta functions for the order parameters, one obtains

$$\begin{aligned}
\langle\langle J_1\rangle_{\boldsymbol{J}}^m\rangle_{D|\boldsymbol{B}} \;=\; & \prod_{a=1}^{n}\left[\int dR_a \int_{-i\infty}^{i\infty} \frac{d\hat{R}_a}{2\pi i N}\right] \prod_{a<b}\left[\int dq_{ab} \int_{-i\infty}^{i\infty} \frac{d\hat{q}_{ab}}{2\pi i N}\right] \\
& \times \exp N\Big\{ \sum_a \hat{R}_a R_a + \sum_{a<b} \hat{q}_{ab} q_{ab} \\
& + \alpha\, G_1(\{R_a, q_{ab}\}; \beta, [V, U])\Big\} \\
& \times \int \left(\prod_a^n dm(\boldsymbol{J}^a)\right)\left(\prod_{\nu=1}^m J_1^\nu\right) \exp\Big[- \sum_a \hat{R}_a \boldsymbol{J}^a \cdot \boldsymbol{B} \\
& - \sum_{a<b} \hat{q}_{ab} \boldsymbol{J}^a \cdot \boldsymbol{J}^b\Big] \,. \qquad \text{(E.4)}
\end{aligned}$$

Renaming the integrals over $\{\boldsymbol{J}^a\}$ as $\exp N G_0^{(m)}(\{\hat{R}_a, \hat{q}_{ab}\})$, one concludes that the previously defined *entropy term* (eq. B.13) corresponds simply to the case $m = 0$.

E.2 The binary measure

Imposing the RS *ansatz* (eq. B.16) and employing the factorization of the binary measure, $dm(\boldsymbol{J}) = \prod_{j=1}^N dm(J_j)$, one writes

$$\begin{aligned}
e^{N G_0^{(m)}} \;\overset{RS}{=}\; & e^{(N-1) G_0^{(0)}} \int \left(\prod_a^n dm(J_1^a)\right)\left(\prod_{\nu=1}^m J_1^\nu\right) \exp\Big[\hat{R} B_1 \sum_a J_1^a \\
& + \frac{\hat{q}}{2}\left(\sum_a J_1^a\right)^2 - \frac{\hat{q}}{2}\sum_a (J_1^a)^2\Big]
\end{aligned}$$

$$= e^{(N-1)G_0^{(0)}} \int \mathcal{D}z \int \left(\prod_a^n dm(J_1^a)\right)\left(\prod_{\nu=1}^m J_1^\nu\right) \exp\Big[-\frac{\hat{q}}{2}\sum_a (J_1^a)^2 + \left(z\sqrt{\hat{q}}+\hat{R}B_1\right)\sum_a J_1^a\Big] \tag{E.5}$$

With the introduction of the z integral, the whole expression factorizes on the replica index a and one obtains

$$\frac{e^{NG_0^{(m)}}}{e^{(N-1)G_0^{(0)}}} = \int \mathcal{D}z \frac{\left(\int dm(J_j)\, J_j\, \exp\left[-\hat{q}J_j^2/2 + (z\sqrt{\hat{q}}+\hat{R}B_1)J_j\right]\right)^m}{\left(\int dm(J_j)\, \exp\left[-\hat{q}J_j^2/2 + (z\sqrt{\hat{q}}+\hat{R}B_1)J_j\right]\right)^{m-n}} . \tag{E.6}$$

Imposing now the binary measure in eq. E.6, one obtains

$$\frac{e^{NG_0^{(m)}}}{e^{(N-1)G_0^{(0)}}} = e^{-n\hat{q}/2} \int \mathcal{D}z \frac{\left(\sinh\left(z\sqrt{\hat{q}}+\hat{R}B_1\right)\right)^m}{\left(\cosh\left(z\sqrt{\hat{q}}+\hat{R}B_1\right)\right)^{m-n}} . \tag{E.7}$$

Therefore, in the limit $N \to \infty$ the expression for the m-th moment of J_1 becomes

$$\begin{aligned}\langle\langle J_1\rangle_{\boldsymbol{J}}^m\rangle_{D|\boldsymbol{B}} \;\simeq\; & \lim_{n\to 0}\lim_{N\to\infty} \int dR\, d\hat{R}\, dq\, d\hat{q}\, \exp\left[nN\hat{f}(q,R,\hat{q},\hat{R};\beta,[U,V])\right] \\ & \times e^{-n\hat{q}/2} \int \mathcal{D}z \frac{\left(\sinh\left(z\sqrt{\hat{q}}+\hat{R}B_1\right)\right)^m}{\left(\cosh\left(z\sqrt{\hat{q}}+\hat{R}B_1\right)\right)^{m-n}} .\end{aligned} \tag{E.8}$$

In the thermodynamic limit, the integral is dominated by the extremum of $\hat{f}$ (recall eq. 1.17), which becomes the free energy after the extremization. Therefore, the equilibrium values of q, R, $\hat{q}$, $\hat{R}$ are completely determined by $\hat{f}$ and can be inserted in the non-exponential terms. If one now takes the limit $n \to 0$, the exponential term vanishes[1] and one arrives at the result

$$\langle\langle J_1\rangle_{\boldsymbol{J}}^m\rangle_{D|\boldsymbol{B}} = \int \mathcal{D}z \left[\tanh\left(z\sqrt{\hat{q}(\alpha)}+\hat{R}(\alpha)B_1\right)\right]^m , \tag{E.9}$$

[1]Note the difference between this quantity and the free energy, eq. B.6: in the latter expression, a factor n^{-1} guarantees that the exponential contribution is the dominant one. Here the exponential only determines the value of the order parameters, vanishing afterwards.

or, equivalently,

$$\langle\langle B_1 J_1\rangle_{\boldsymbol{J}}^m\rangle_{D|\boldsymbol{B}} = \int \mathcal{D}z \, \left[\tanh\left(z\sqrt{\hat{q}(\alpha)} + \hat{R}(\alpha)\right)\right]^m \, , \tag{E.10}$$

which holds as long as $\boldsymbol{B} \in \{-1,+1\}^N$ (see the discussion on pages 116 or 129). In these equations, $\hat{q}(\alpha)$ and $\hat{R}(\alpha)$ denote the conjugate parameters taken at their equilibrium values.

E.3 The spherical measure

One can introduce the spherical measure $P(\boldsymbol{J}) \sim \delta(\boldsymbol{J}\cdot\boldsymbol{J} - N)$ in eq. E.4 via its Fourier representation, as done in section B.2.3. Using the same notation from that section, imposing the RS *ansatz* $E_a = i\mathcal{E}/2$ and taking the limit $n \to 0$, one arrives at

$$\frac{e^{NG_0^{(m)}}}{e^{(N-1)G_0^{(0)}}} = \int \mathcal{D}z \, \left(\frac{z\sqrt{\hat{q}} + \hat{R}B_1}{\hat{q} + \mathcal{E}}\right)^m \, , \tag{E.11}$$

where $\mathcal{E}$, $\hat{q}$ and $\hat{R}$ are now the conjugate parameters for the *spherical case*. With this result, the same reasoning that led from eq. E.7 to eq. E.10 applies: the values of the conjugate parameters are determined by the saddle point equations, the dependence on B_1 can be simplified[2] and one writes

$$\langle\langle B_1 J_1\rangle_{\boldsymbol{J}}^m\rangle_{D|\boldsymbol{B}} = \int \mathcal{D}z \, \left(\frac{z\sqrt{\hat{q}(\alpha)} + \hat{R}(\alpha)}{\hat{q}(\alpha) + \mathcal{E}(\alpha)}\right)^m \, . \tag{E.12}$$

Inserting the solutions $\mathcal{E}(\alpha)$, $\hat{R}(\alpha)$ and $\hat{q}(\alpha)$ from eqs. B.28, one obtains

$$\langle\langle B_1 J_1\rangle_{\boldsymbol{J}}^m\rangle_{D|\boldsymbol{B}} = \int \mathcal{D}z \, \left(z\sqrt{q - R^2} + R\right)^m \, . \tag{E.13}$$

Going back to the original quantity of interest, $P(x)$ can be obtained by a simple change of variables in eq. E.13, yielding finally

$$P(x) = \frac{1}{\sqrt{2\pi(1 - R^2/q)}} \exp\left[-\frac{(x - R/\sqrt{q})^2}{2(1 - R^2/q)}\right] \, . \tag{E.14}$$

[2] It is important to recall that the scenario under consideration here is that of a spherical $\boldsymbol{J}$ but still with a binary $\boldsymbol{B}$.

E.4 Proof of eq. 5.21

In order to show that eq. 5.21 is correct, one just has to show that the difference between the right and left hand sides is zero:

$$\begin{aligned}
& \int \mathcal{D}z \, \left[\mathrm{sign}(az+a^2) - |\tanh(az+a^2)|\right] \\
= & \int \mathcal{D}z \, \mathrm{sign}(az+a^2) \left[1 - \tanh(az+a^2)\right] \\
\overset{z=y-a}{=} & \int \frac{dy}{\sqrt{2\pi}} e^{-(y-a)^2/2} \, \mathrm{sign}(ay) \left[1 - \tanh(ay)\right] \\
= & \; e^{-a^2/2} \int \mathcal{D}y \, \mathrm{sign}(ay) e^{ay} \left[\frac{e^{-ay}}{\cosh(ay)}\right] \\
= & \; 0 \, , \qquad \text{(E.15)}
\end{aligned}$$

a proof that is clearly similar to that of eq. C.15.

Appendix F

The limit of zero temperature

F.1 The free energy

The special limit of interest $\beta \to \infty$ can be treated by introducing rescaled order parameters in order to achieve the proper scaling ($\mathcal{O}(\beta^0)$) of the free energy. In this regime the solutions of the saddle point equations correspond to vectors that minimize the cost function V. In particular, if this minimum is non-degenerate, then the order parameter q is expected to tend to one.

F.1.1 The energy term

The first term to be dealt with is the (replica symmetric) energy term G_1, given by eq. B.21. One needs to introduce a new variable

$$\eta \equiv \beta(1-q) \tag{F.1}$$

which must remain finite in the limits $q \to 1$ and $\beta \to \infty$. With this change of variables eq. B.21 reads

$$G_1(R,\eta) = n \int \mathcal{D}^* b \int \mathcal{D}t_2 \ \ln \sqrt{\frac{\beta}{\eta}} \int \frac{d\lambda}{\sqrt{2\pi}} \exp -\beta \left[V(\lambda) + \frac{(\lambda - t)^2}{2\eta} \right] , \tag{F.2}$$

where $t \equiv t_2\sqrt{q-R^2} + bR$. The integral on λ can be evaluated via the Laplace method, yielding:

$$\frac{G_1(R,\eta)}{n} \overset{\beta\to\infty}{\simeq} -\beta \int \mathcal{D}^* b \int \mathcal{D}t_2 \ \left[V(\lambda_0(t,\eta)) + \frac{(\lambda_0(t,\eta) - t)^2}{2\eta} \right]$$

$$\stackrel{(A.18)}{=} \quad -\beta \int \mathcal{D}t\, X(t;R) \left[V\left(\lambda_0(t,\eta)\right) + \frac{\left(\lambda_0(t,\eta)-t\right)^2}{2\eta} \right] \tag{F.3}$$

where

$$\lambda_0(t,\eta) = \underset{\lambda}{\text{Argmin}} \left[V(\lambda) + \frac{(\lambda - t)^2}{2\eta} \right] \tag{F.4}$$

and assuming

$$V''(\lambda_0(t,\eta)) + 1/\eta \geq 0, \qquad \forall t\ . \tag{F.5}$$

F.1.2 The Ising measure

The calculation of the entropic term for the Ising measure in the limit $\beta \to \infty$ requires an extra change of variables for the proper scaling to be achieved. By introducing eq. F.1 in the free energy B.24, one obtains

$$\begin{aligned} f \;=\; & -\underset{R,\eta,\hat{R},\hat{q}}{\text{Extr}} \Big\{ -\frac{\eta\hat{q}}{2\beta^2} - \frac{\hat{R}R}{\beta} + \frac{1}{\beta}\int \mathcal{D}z\ \ln\cosh\left(z\sqrt{\hat{q}} + \hat{R}\right) \\ & + \frac{\alpha}{n\beta}\, G_1(R,\eta) \Big\}\ , \end{aligned} \tag{F.6}$$

which clearly indicates the rescaling

$$\begin{aligned} \hat{\eta} \;&\equiv\; \frac{\hat{q}}{\beta^2} \\ \hat{y} \;&\equiv\; \frac{\hat{R}}{\beta} \end{aligned} \tag{F.7}$$

in order for the free energy to be $\mathcal{O}(\beta^0)$. Substituting the change of variables F.7 in eq. F.6, one is left with an integral on z which can be performed in the limit $\beta \to \infty$, rendering

$$\begin{aligned} & \frac{1}{\beta}\int \mathcal{D}z\ \ln 2\cosh\left[\beta\left(z\sqrt{\hat{\eta}} + \hat{y}\right)\right] \stackrel{\beta\to\infty}{\simeq} \\ & \hat{y}\left[1 - 2H\left(\frac{\hat{y}}{\sqrt{\hat{\eta}}}\right)\right] + 2\sqrt{\hat{\eta}}\, P_n\left(\frac{\hat{y}}{\sqrt{\hat{\eta}}}\right) + \mathcal{O}(\beta^{-2})\ . \end{aligned} \tag{F.8}$$

Therefore the free energy is given by, to dominant order,

$$f = \underset{R,\eta,\hat{\eta},\hat{y}}{\text{Extr}} \left\{ \frac{\eta\hat{\eta}}{2} + \hat{y}R - 2\sqrt{\hat{\eta}}P_n\left(\hat{y}/\sqrt{\hat{\eta}}\right) - \hat{y}\left[1 - 2H\left(\hat{y}/\sqrt{\hat{\eta}}\right)\right] + \alpha \int \mathcal{D}t\, X(t;R) \left[V\left(\lambda_0(t,\eta)\right) + \frac{\left(\lambda_0(t,\eta) - t\right)^2}{2\eta} \right] \right\} \quad \text{(F.9)}$$

where

$$\begin{aligned} P_n(t) &\equiv \frac{e^{-t^2/2}}{\sqrt{2\pi}} \\ H(x) &\equiv \int_x^\infty \mathcal{D}t = \int_x^\infty dt\, P_n(t)\,. \end{aligned} \quad \text{(F.10)}$$

The saddle point equations

The extremum operator renders the following saddle point equations for the order parameters:

$$\begin{aligned} \frac{\partial f}{\partial \hat{y}} = 0 &\Rightarrow R = 1 - 2H\left(\frac{\hat{y}}{\sqrt{\hat{\eta}}}\right) \\ \frac{\partial f}{\partial \hat{\eta}} = 0 &\Rightarrow \eta = \frac{2}{\sqrt{\hat{\eta}}} P_n\left(\frac{\hat{y}}{\sqrt{\hat{\eta}}}\right) \\ \frac{\partial f}{\partial \eta} = 0 &\Rightarrow \hat{\eta} = \frac{\alpha}{\eta^2} \int \mathcal{D}t\, X(t;R) \left[\lambda_0(t,\eta) - t\right]^2 \\ \frac{\partial f}{\partial R} = 0 &\Rightarrow \hat{y} = \frac{\alpha}{\eta} \int \mathcal{D}t\, Y(t;R) \left[\lambda_0(t,\eta) - t\right]\,. \end{aligned} \quad \text{(F.11)}$$

F.1.3 The spherical measure

The limit $\beta \to \infty$ in equation B.29 can be taken using again the rescaling F.1. Making use of eq. F.3, the free energy for vectors $\boldsymbol{J}$ constrained to the N-hypersphere reads, to leading order,

$$f = \underset{R,\eta}{\text{Extr}} \left\{ \frac{-(1-R^2)}{2\eta} + \alpha \int \mathcal{D}t\, X(t;R) \left[V\left(\lambda_0(t,\eta)\right) + \frac{\left(\lambda_0(t,\eta) - t\right)^2}{2\eta} \right] \right\} \quad \text{(F.12)}$$

F.2 Proof that $\partial\mathcal{F}/\partial R \geq 0$

In order to show that $\mathcal{F}(R)$ (defined in eq. 2.7) is a monotonically increasing function, one defines

$$\begin{aligned} g(R) &\equiv \frac{\mathcal{F}^2(R)}{\alpha} = \int \mathcal{D}t\, \frac{Y^2(t;R)}{X(t;R)} \\ &\equiv \frac{1}{1-R^2} \int \mathcal{D}t\, \frac{C_1^2(t;R)}{C_0(t;R)}\,, \end{aligned} \tag{F.13}$$

where

$$C_n(t;R) \equiv \mathcal{N} \int \mathcal{D}t'\, (t')^n e^{-U(Rt+\sqrt{1-R^2}t')} \tag{F.14}$$

and $\mathcal{N}$ is defined on page 8. It suffices thus to prove that $g(R)$ is monotonic in R. This is accomplished by noting that the functions C_n obey the following recursion relations:

$$\frac{\partial}{\partial t} C_n = \frac{R}{\sqrt{1-R^2}} \left(C_{n+1} - nC_{n-1}\right) \tag{F.15}$$

$$\begin{aligned} \frac{\partial}{\partial R} C_n &= \frac{t}{R}\frac{\partial}{\partial t} C_n - \frac{1}{\sqrt{1-R^2}} \frac{\partial}{\partial t} C_{n+1} \\ &= \frac{1}{\sqrt{1-R^2}} \left\{ t\left(C_{n+1} - nC_{n-1}\right) - \frac{\partial}{\partial t} C_{n+1} \right\}\,. \end{aligned} \tag{F.16}$$

To calculate the derivative $\partial g/\partial R$, one applies eqs. F.15 and F.16 to the r.h.s. of eq. F.13. Apart from the integration measure $\mathcal{D}t$, one obtains two kinds of terms: "pure terms", which only depend on $\{C_k\}$, and "pure terms" multiplied by t. The latter can be handled by noticing the equality $\int \mathcal{D}t\, t\, f(t) = \int \mathcal{D}t\, f'(t)$. Therefore one applies the recursion relations F.15 and F.16 once more and rearranges the terms. The final result is

$$\begin{aligned} \frac{\partial g}{\partial R} &= \frac{2R}{(1-R^2)^2} \int \mathcal{D}t\, \frac{\left[C_0^2 - (C_0 C_2 - C_1^2)\right]^2}{C_0^3} \\ &\geq 0\,, \end{aligned} \tag{F.17}$$

since $C_0 \geq 0$.

List of publications

- "On-line learning in the committee machine"
  M. Copelli and N. Caticha
  J. Phys. A: Math. Gen., **28** 1615-1625 (1995)

- "Optimal local learning in a multilayer machine"
  M. Copelli, O. Kinouchi and N. Caticha
  in Anais do II Congresso Brasileiro de Redes Neurais, 7 pages, Curitiba, Brazil (1995)

- "Equivalence between learning in noisy perceptrons and tree committee machines"
  M. Copelli, O. Kinouchi and N. Caticha
  Phys. Rev. E, **53** 6341-6352 (1996)

- "Noise robustness in multilayer neural networks"
  M. Copelli, R. Eichorn, O. Kinouchi, M. Biehl, R. Simonetti, P. Riegler and N. Caticha
  Europhys. Lett., **37** (6), 427-432 (1997)

- "Noise robustness in the perceptron"
  M. Copelli
  in Proceedings of the ESANN'97, ed. by M. Verleysen, D facto (Belgium), pp. 181-186 (1997)

- "Universal asymptotics in committee machines with tree architecture"
  M. Copelli and N. Caticha
  in On-Line Learning in Neural Networks, ed. by D. Saad, Cambridge University Press (UK), pp. 165-181 (1998)

- "On the center of mass of Ising vectors"
  M. Copelli, M. Bouten, C. Van den Broeck and B. Van Rompaey
  Europhys. Lett., **47** (2), 139-144 (1999)

- "Bayes-optimal performance in a discrete space"
  M. Copelli, C. Van den Broeck and M. Opper
  http://xxx.lanl.gov/abs/cond-mat/9906356
  (to appear in J. Phys. A: Math. Gen.)

- "Stokes' drift: a rocking ratchet"
  I. Bena, M. Copelli and C. Van den Broeck
  http://xxx.lanl.gov/abs/cond-mat/9908338
  (submitted)

- "Unsupervised learning of binary vectors: a Gaussian scenario"
  M. Copelli and C. Van den Broeck
  http://xxx.lanl.gov/abs/cond-mat/9910365
  (submitted)

Bibliography

[BG98] A. Buhot and M. B. Gordon. Phase transitions in optimal unsupervised learning. *Phys. Rev. E*, 57:3326–3333, 1998.

[Bis95] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford Univ. Press, 1995.

[BM93] M. Biehl and A. Mietzner. Statistical mechanics of unsupervised learning. *Europhys. Lett.*, 24:421–426, 1993.

[BM94] M. Biehl and A. Mietzner. Statistical mechanics of unsupervised structure recognition. *J. Phys. A: Math. Gen.*, 27:1885, 1994.

[Bou99] M. Bouten. Private communication, 1999.

[BRS95] M. Biehl, P. Riegler, and M. Stechert. Learning from noisy data: an exactly solvable model. *Phys. Rev. E*, 52:R4624–R4627, 1995.

[BS95] D. Bollé and G. M. Shim. Nonlinear Hebbian training of the perceptron. *Network: Computation in Neural Systems*, 6:619–633, 1995.

[BSVdB95] M. Bouten, J. Schietse, and C. Van den Broeck. Gradient descent learning in perceptrons: A review of its possibilities. *Phys. Rev. E*, 52:1958, 1995.

[BTMG97] A. Buhot, J.-M. Torres Moreno, and M. B. Gordon. Finite size scaling of the Bayesian perceptron. *Phys. Rev. E*, 55:7434–7440, 1997.

[Buh99] A. Buhot. *Étude de propriétés d'apprentissage supervisé et non supervisé par des méthodes de Physique Statistique*. PhD thesis, Université Grenoble I - Joseph Fourier, 1999. (in French).

[CC95] M. Copelli and N. Caticha. On-line learning in the committee machine. *J. Phys. A: Math. Gen.*, 28:1615, 1995.

[CC98] M. Copelli and N. Caticha. Universal asymptotics in committee machines with tree architecture. In D. Saad, editor, *On-line Learning in Neural Networks*, pages 165–181. Cambridge University Press (UK), 1998.

[CEK+97] M. Copelli, R. Eichorn, O. Kinouchi, M. Biehl, R. Simonetti, P. Riegler, and N. Caticha. Noise robustness in multilayer neural networks. *Europhys. Lett.*, 37:427, 1997.

[CKC96] M. Copelli, O. Kinouchi, and N. Caticha. Equivalence between learning in noisy perceptrons and tree committee machines. *Phys. Rev. E*, 53:6341, 1996.

[Cop97] M. Copelli. Noise robustness in the perceptron. In M. Verleysen, editor, *Proceedings of the ESANN'97*. Dfacto (Belgium), 1997.

[CT91] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 1991.

[dAT78] J. R. L. de Almeida and D. J. Thouless. Stability of the Sherrington-kirkpatrick solution of a spin glass model. *J. Phys. A: Math. Gen.*, 11, 1978.

[dM97] C. R. de Mattos. *Aplicações de Mecânica Estatística ao Perceptron Binário e ao Processamento de Imagens*. PhD thesis, Universidade de São Paulo, 1997. (in Portuguese).

[Gar88] E. Gardner. The space of interactions in neural networks models. *J. Phys. A: Math. Gen.*, 21:257–270, 1988.

[GB98] M. B. Gordon and A. Buhot. Bayesian learning versus optimal learning. *Physica A*, 257:85–98, 1998.

[GD88] E. Gardner and B. Derrida. Optimal storage properties of neural network models. *J. Phys A: Math. Gen.*, 21:271–284, 1988.

[GM93] M. Golea and M. Marchand. Learning curves of the clipped Hebb rule for networks with binary weights. *J. Phys. A: Math. Gen.*, 26:5751–5766, 1993.

[GMe93] W. T. Grandy and P. W. Milonni (eds.). *Physics and probability: essays in honor of Edwin T. Jaynes*. Cambridge University Press, 1993.

[GS90] H. Gutfreund and Y. Stein. Neural networks with discrete synaptic couplings. In W. K. Theumann and R. Köberle, editors, *Neural Networks and Spin Glasses*. World Scientific, 1990.

[GT90] G. Györgyi and N. Tishby. Statistical theory of learning a rule. In W. K. Theumann and R. Köberle, editors, *Neural Networks and Spin Glasses*. World Scientific, Singapore, 1990.

[Gyö90] G. Györgyi. First-order transition to perfect generalization in a neural network with binary synapses. *Phys. Rev. A*, 41:7097, 1990.

[Hay97] B. Hayes. Can't get no satisfaction. *American Scientist*, 85:108–112, 1997.

[HN99] D. Herschkowitz and J.-P. Nadal. Unsupervised and supervised learning: Mutual information between parameters and observations. *Phys. Rev. E*, 59:3344–3360, 1999.

[Hor92a] H. Horner. Dynamics of learning and generalization in a binary perceptron model. *Z. Phys. B*, 87:371–376, 1992.

[Hor92b] H. Horner. Dynamics of learning for the binary perceptron problem. *Z. Phys. B*, 86:291–308, 1992.

[Hor93] H. Horner. Dynamics of learning and generalization in perceptrons with constraints. *Physica A*, 200:552–562, 1993.

[ISB95] J. Iwanski, J. Schietse, and M. Bouten. Replica symmetry breaking in a diluted network with binary couplings. *Phys. Rev. E*, 52:888, 1995.

[KC92] O. Kinouchi and N. Caticha. Optimal generalization in perceptrons. *J. Phys. A: Math. Gen.*, 25:6243, 1992.

[KC93] O. Kinouchi and N. Caticha. Lower bounds on generalization errors for drifting rules. *J. Phys. A: Math. Gen.*, 26:6161, 1993.

[KC95] O. Kinouchi and N. Caticha. On-line versus off-line learning in the linear perceptron: A comparative study. *Phys. Rev. E*, 52:2878, 1995.

[KC96] O. Kinouchi and N. Caticha. Learning algorithm that gives the Bayes generalization limit for perceptrons. *Phys. Rev. E*, 54:R54, 1996.

[KM89] W. Krauth and M Mézard. Storage capacity of memory networks with binary couplings. *J. Phys. France*, 50:3057–3066, 1989.

[KU98] W. Kinzel and R. Urbanczik. On-line learning in a discrete state space. *J. Phys. A: Math. Gen.*, 31:L27–L30, 1998.

[MBS95] C. Marangi, M. Biehl, and S. A. Solla. Supervised learning from clustered input examples. *Europhys. Lett.*, 30:117, 1995.

[MSBR96] C. Marangi, S. A. Solla, M. Biehl, and P. Riegler. Off-line supervised learning from clustered input examples. In M. Marinaro and R. Tagliaferri, editors, *Proceedings of the 7th Italian Workshop on Neural Networks 1995*. World Scientific (Singapore), 1996.

[OH91] M. Opper and D. Haussler. Generalization performance of Bayes optimal classification algorithm for learning a perceptron. *Phys. Rev. Lett.*, 66:2677–2680, 1991.

[OK96] M. Opper and W. Kinzel. Statistical mechanics of generalization. In E. Domany, J. L. van Hemmen, and K. Schulten, editors, *Models of Neural Networks III*. Springer-Verlag, 1996.

[OW96] M. Opper and O. Winther. Mean field approach to Bayes learning in feed-forward neural networks. *Phys. Rev. Lett.*, 76:1964, 1996.

[PV88] L. Pitt and L. G. Valiant. Computational limitations on learning from examples. *Journal of the Association for Computer Machinery*, 35:965–984, 1988.

[Re83] R. D. Rosenkrantz (ed.). *E. T. Jaynes: Papers on Probability, Statistics and Statistical Physics*. D. Reidel Publishing Company, 1983. ISBN 90-277-1448-7.

[Rei97] P. Reimann. Unsupervised learning of distributions. *Europhys. Lett.*, 40:251–256, 1997.

[RVdB96] P. Reimann and C. Van den Broeck. Learning by examples from a nonuniform distribution. *Phys. Rev. E*, 53:3989, 1996.

[RVdBB96] P. Reimann, C. Van den Broeck, and G. J. Bex. A Gaussian scenario for unsupervised learning. *J. Phys. A: Math. Gen.*, 29:3521, 1996.

[SBVdB95] J. Schietse, M. Bouten, and C. Van den Broeck. Training binary perceptrons by clipping. *Europhys. Lett.*, 32:279–284, 1995.

[SC96] R. Simonetti and N. Caticha. On-line learning in parity machines. *J. Phys. A: Math. Gen.*, 29:4859, 1996.

[Sha48] C. E. Shannon. A mathematical theory of communication. *Bell Sys. Tech. J.*, 27:379–423,623–656, 1948. Reprinted in C. E. Shannon and W. Weaver, "The Mathematical Theory of Communication," Univ. Ill. Press, Urbana, Illinois, 1949.

[SR98] D. Saad and M. Rattray. Globally optimal parameters for online learning in multilayer neural networks. *Phys. Rev. Lett.*, 79:2578–2581, 1998.

[SST92] H. S. Seung, H. Sompolinsky, and N. Tishby. Statistical mechanics of learning from examples. *Phys. Rev. A*, 45:6056, 1992.

[Val89] F. Vallet. The Hebb rule for learning linearly separable boolean functions: Learning and generalization. *Europhys. Lett.*, 8:747–751, 1989.

[VC97] R. Vicente and N. Caticha. Functional optimization of online algorithms in multilayer neural networks. *J. Phys. A: Math. Gen.*, 30:L599, 1997.

[VdBB93] C. Van den Broeck and M. Bouten. Clipped-Hebbian training of the perceptron. *Europhys. Lett.*, 22:223, 1993.

[VdBR96] C. Van den Broeck and P. Reimann. Unsupervised learning by examples: On-line versus off-line. *Phys. Rev. Lett.*, 76:2188, 1996.

[VKC98] R. Vicente, O. Kinouchi, and N. Caticha. Statistical mechanics of online learning of drifting concepts: A variational approach. *Machine Learning Journal*, 32:179–201, 1998.

[VR99] Bart Van Rompaey. *Leerstrategieën voor het Binaire Perceptron.* PhD thesis, Limburgs Universitair Centrum, 1999. (in Dutch).

[Wat93] T. L. H. Watkin. Optimal learning with a neural network. *Europhys. Lett.*, 21:871, 1993.

[WN94] T. L. H. Watkin and J.-P. Nadal. Optimal unsupervised learning. *J. Phys. A: Math. Gen.*, 27:1899–1915, 1994.

[WRB93] T. L. H. Watkin, A. Rau, and M. Biehl. The statistical mechanics of learning a rule. *Rev. Mod. Phys.*, 65, 1993.

[WRS92] K. Y. M. Wong, A. Rau, and D. Sherrington. Weight space organization of optimized neural networks. *Europhys. Lett.*, 19:559–564, 1992.