FACOLTÀ DI SCIENZE E TECNOLOGIE
CORSO DI LAUREA MAGISTRALE IN FISICA

GENERALIZATION FROM CORRELATED INPUTS IN A SIMPLE MODEL OF SUPERVISED NEURAL NETWORK

Relatore:
Prof. Dr. Sergio Caracciolo

Francesco Borra
864813

Correlatori:
Dr. Marco Gherardi
Dr. Pietro Rotondo

Anno Accademico 2016-2017

Generalization from correlated inputs in a simple model of supervised neural network

Francesco Borra

January 25, 2018

Contents

Chapter 1

Neural networks and statistical mechanics: introduction and formalism

1.1 Introduction

1.1.1 Neural networks and statistical mechanics

A neural network, also referred to as machine, is a generic term to define a computing system inspired to biology. In this thesis, we will focus on a simple classifying neural network, called perceptron. In the neural network framework, the problem of classification can be stated as follows: is it possible to teach a machine to correctly classify a set of inputs, according to a given rule? The answer clearly depends on the input, the machine and the rule.

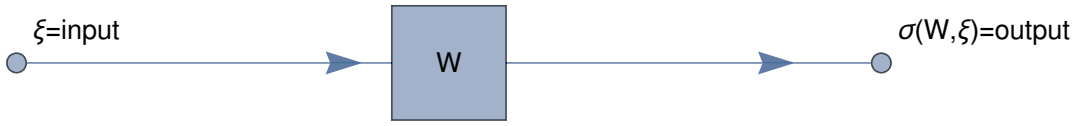


Figure 1.1: A schematic representation of a classifying neural network. Given an input ξ, the network yields an output σ which depends on the internal parameters (synaptic weights W) of the network.

The machine is, in general, an algorithm, which receives the **input data** (that will be called ξ) and yields an **output** (called $\sigma(\xi)$), which is a function of the input. The network is required to have a general structure and to be adaptable to the desired recognition task. To be more precise, the algorithm should depend on some variables or **configuration** W which can be fixed so that some specified input patterns are correctly classified. More precisely, if we want the input ξ to have output σ, we can choose some configuration W so that $\sigma_W(\xi) = \sigma$. In this sense, the network can be "taught" or "trained". This process is called **supervised learning**. The algorithm clearly works with any (compatible) input and not just those with which the machine has been taught. So, how are other patterns classified? Ideally, we would like the machine to be capable of assigning each input to the most similar pattern among those that it has learned. When this happens, the machine is said to **generalize**. Another condition - the machine only capable of classifying what it has learned - is said **overfitting**. Another relevant feature of the network is the amount of information, i.e. the size of W, which it has to store in order to perform correctly. The less the

information, the better. Ideally, the size of W should be much less than the whole information contained in the learned patterns themselves. In the case that will be treated, W will be a single vector and will be enough to classify a whole set of input vectors of the same size. Given this premise, the information should be "compressed", and the learning can be studied on a statistical basis. Not all sets of patterns will be teachable and, among those that can be taught, not all patterns will be equally easy to teach. More technically, it will not always be possible, or easy, to set the correct internal variables to perform a given classification task.

The input is, in general, an array of numerical data. In order to study the statistical behaviour of the machine, one needs to know the likelihood that a certain set of patterns will be chosen to be taught. As anticipated, from now on, the set of patterns will be named $\xi = \{\xi_\alpha\}$ and its distribution $P(\xi)$. Likewise $\sigma = \{\sigma_\alpha\}$ will denote the set of desired outputs. The two sets form answer-question pairs (ξ, σ). If $P(\xi)$ and $P(\sigma)$ are known, one can focus on the most likely scenarios. More precisely, it may happen: that it is possible to teach the machine with probability 1; that non-learnable sets of patterns exist, but the probability of any of them being picked is zero (or vanishing). An interesting quantity is therefore the typical behaviour of the network.

In order to study the statistical behaviour of a machine, the classification rule must be known. While the rules can be easy or very complicated, it is possible to define a cost function or energy H to study the problem. The cost function should be chosen according to the problem that is being study and to the machine under exam. It is important to point out that such energy is not an intrinsic property of the neural network, but rather an external tool employed to investigate the efficiency of the algorithm. If W is the set of parameters of the machine, the cost function is a quantity

$$H(\sigma, \xi; W)$$

which should quantify how well the setting W solves the classification problem of linking each ξ_α to its σ_α. A simple choice of energy is the number of errors, i.e. the number of misclassified patterns. For example

$$\chi(\sigma, \xi; W) = \sum_\alpha \delta(\sigma_W(\xi_\alpha), \sigma_\alpha)$$

where σ_W is the function which assign each input to an output, given the set of parameters W. In general, the lower the cost function, the better the performance. Let now the energy be some kind of error count. In this case, all zero energy states are solutions.

A relevant function which can be computed by the use of a cost function, is the number of solutions for a given (ξ, σ) problem:

$$Z(\xi, \sigma) = \int P(W)\, \mathbb{X}_{(\xi,\sigma)}(W) \tag{1.1.1}$$

where $\mathbb{X}_{(\xi,\sigma)}(W)$ is 1 if W is a solution and zero otherwise. A further step is to introduce some noise and allow all Ws in the count, weighting them according to their energy. For example:

$$Z(\xi, \sigma) = \int P(W)\, e^{-\beta\chi(\xi,\sigma;W)} \tag{1.1.2}$$

The parameter β, which quantifies the noise, is an inverse-temperature like parameter. Clearly, if $\beta \to \infty$, (1.1.2) reduces to (1.1.1). Therefore, no noise equals zero temperature.

A more general expression is

$$Z(\xi, \sigma) = \int P(W)\, e^{-\beta H(\xi,\sigma;W)} \tag{1.1.3}$$

This function can be called partition function as in statistical mechanics.

1.1.2 Average over input-disorder and the thermodynamic limit

As anticipated, to get statistically meaningful observable, it is necessary to average $Z(\xi, \sigma; W)$ over (σ, ξ), i.e. over all possible input output pairs we want to teach the network. However, the most obvious average turns out to be incorrect:

$$\bar{Z} = \int P(\xi, \sigma) Z(\xi, \sigma) \tag{1.1.4}$$

This average is called annealed. The correct average is called quenched: instead of integrating the partition function, one integrates the free energy:

$$\bar{f} = \int P(\xi, \sigma) \ln Z(\xi, \sigma) \tag{1.1.5}$$

In statistical mechanics, the annealed average is used when disorder fluctuates in time along with the degrees of freedom. In this case, the roles of the disorder and the degrees of freedom of the system are equivalent and the system effectively interacts with the average disorder. On the other hands, in the quenched average, disorder is assumed to be static, at least with respect to timescale of the system. Therefore, the system interacts with a single configuration of disorder, which is however random. In the case of neural network, the internal parameters of the machine constitute the system, while the different choices of learning sets represent the disorder. Since different sets of training input represent different scenarios, the quenched averaging is intuitively more appropriate.

There is a more compelling and formal reason: we are interested in a network with the potential to classify great amounts of information. This means that each set of patterns should be "big", i.e. it should contain p patterns, with $p \to \infty$. This scenario can be called thermodynamic limit, in accordance with statistical mechanic terminology. Hence, it is natural to choose models in which the energy is extensive, meaning that

$$H(\xi, \sigma; W) \sim p\, h(\xi, \sigma; W)$$

For instance, the aforementioned error-counting cost function clearly exhibits this property. It is known that, under this assumption, the partition function is not extensive

$$Z \sim \exp(pF)$$

for some $F \in \mathbb{R}$. On the other hand, the free energy

$$f = -\frac{1}{\beta} \log Z \sim -\frac{p}{\beta} \log z$$

is extensive. Therefore, only the free energy (divided by p) is a proper self-averaging observable, for any configuration of the disorder. Hence, the correct choice is

$$\boxed{\bar{f} = \frac{1}{p} \int P(\xi, \sigma) \ln Z(\xi, \sigma)} \tag{1.1.6}$$

1.1.3 The replica formalism

In order to overcome the difficulties introduced by the averaging logarithm in formula (1.1.6), some more sophisticated computation techniques are necessary. One of the best known techniques is the replica method. It was originally developed for fully

connected spin glass models. Spin glasses are Ising-like models in which the spin couplings are quenched random variables and are allowed to assume both positive and negative values (ferromagnetic and antiferromagnetic). A well known solvable example is the Sherrington-Kirkpatrick model: a binary spin is placed on each vertex of a fully connected graph and each coupling is a Gaussian random variables.

$$H_{SK} = -\sum_{i<j}^{N} J_{ij} s_i s_j$$

with

$$\langle J_{ij} \rangle = 0 \text{ and } \langle J_{ij}^2 \rangle = 1/N$$

Though the study of spin glasses is, in some sense, propaedeutic to neural networks, from now on, for consistency's sake, the focus will be kept on neural networks, as much as possible.

The replica method is based on the following mathematical identity

$$\int P(\xi) \ln Z = \lim_{m \to 0} \frac{1}{m} \ln \int P(\xi) Z^m$$

The trick consists in computing

$$\overline{Z^m} = \int P(\xi)\ Z^m$$

for integer m first. Then, one can take the analytic continuation of $\overline{Z^m}$ and finally take the limit $m \to 0$. Neither of these two passages is, in general, easy. Among the many reasons, one has to deal with matrices with a non-integer (and vanishing) number of dimensions m. This can have very counterintuitive consequences such as the presence of a negative number of off diagonal matrix elements. Nonetheless, everything remains consistent, as long as one remembers that everything should be considered as an analytical continuation.

Let us focus on the integer m case. Z^m can be seen as the partition function of m identical copies, the so-called **replicas**, of the same system. These replicas acquire an interaction due to the averaging over disorder:

$$\overline{Z^m} = \int P(\xi) \prod_{a=1}^{m} \int dW_a\ e^{-\beta H_\xi(W_a)} = \int \prod_{a=1}^{m} dW_a e^{-\beta \tilde{H}(\{W_b\}_{b=1,\dots,m})} \tag{1.1.7}$$

This trick can be used for evaluating the expected value of any observable, say $A(W)$. Then

$$\begin{aligned}
\overline{A} &= \int P(\xi) \frac{1}{Z_\xi} \int dW\ e^{-\beta H_\xi(W)}\ A(W) \\
&= \lim_{m \to 0} \int P(\xi) Z_\xi^{m-1} \int dW\ e^{-\beta H_\xi(W)}\ A(W) \\
&= \lim_{m \to 0} \int P(\xi) \int \prod_{a=1}^{m} dW_a\ e^{-\beta H_\xi(W_a)} A(W_1) \\
&= \lim_{m \to 0} \int P(\xi) \int \prod_{a=1}^{m} dW_a\ e^{-\beta \tilde{H}_\xi(\{W_b\}_{b=1,\dots,m})} A(W_1)
\end{aligned}$$

Before proceeding any further, it is worth specifying that, in the models that will be examined, the internal parameter called W is a vector of N components. It turns

out that the ratio p/N should be finite for the result to be of any interest. Each component W_i of W can be either discrete of continuous. This is not relevant here: one should just keep in mind that, in the discrete case, integrals should be replaced with sums, without any repercussions.

Let us now go back to equation (1.1.7). As already explained, the quenched average generates an interaction between replicas:

$$\sum_a H(W_a) \overset{disorder}{\longrightarrow} \tilde{H}(\{W_b\}_{b=1,\dots,m})$$

Let us now suppose that $\tilde{H}$ is only a function of the overlaps

$$W_a \cdot W_b/N$$

between different replicas:

$$\tilde{H}(\{W_a \cdot W_b/N\}_{a,b=1,\dots,m\ a\neq b}) \tag{1.1.8}$$

This is obviously a strong hypothesis and should be verified in the specific model. It will turn out to be correct for the purposes of this thesis.

In order to handle the scalar products appearing in (1.1.7) via $\tilde{H}$, it is convenient to introduce the auxiliary variable Q_{ab} with a resolution of identity.

$$1 = \int \prod_{a<b} dQ_{ab}\, \delta\left(Q_{ab} - W_a \cdot W_b/N\right)$$

Hence

$$\overline{Z^m} = \int \prod_{a<b} dQ_{ab} \int \prod_{a=1}^{m} dW_a e^{-\beta \tilde{H}(\{Q_{ab}\})} \delta\left(Q_{ab} - W_a \cdot W_b/N\right)$$

Let us summarize in one equation what has been achieved so far:

$$\boxed{\bar{f} = \lim_{m\to 0} \frac{1}{mN} \ln \int \prod_{a<b} dQ_{ab}\, e^{-NA(Q)}} \tag{1.1.9}$$

with

$$\boxed{A(Q) = \frac{1}{N} \ln \left[\int P(\xi) \prod_{a=1}^{m} \left(dW_a\, e^{-\beta H(W_a)} \right) \delta\left(Q_{ab} - W_a \cdot W_b/N\right) \right]} \tag{1.1.10}$$

It was assumed in the previous section that H is extensive in p. Since p/N is finite, this implies that H is extensive in N. Therefore, $A(Q)$ is finite even if $N \to \infty$.

Let us now focus on the quantities m and N. Though only the limit in m is explictly written, in our minds N should go to ∞. However, from a rigorous standpoint, the limit in m should be taken first. However, it turns out that the result is correct even if one takes $N \to \infty$ before $m \to 0$. This is very useful since (1.1.9) can be solved with the saddle point method. In other words, one just has to find the Q_{ab} that minimizes $A(Q)$. Let us call the variables $\{Q_{ab}\}$ collectively as Q. Then

$$Q^{SP} := \arg\min_Q A(Q)$$

and

$$\bar{f} = -\frac{1}{m} A(Q^{SP})$$

Clearly, as a necessary condition for Q^{SP} to be the correct solution, it must hold that

$$\frac{d}{dQ_{ab}} A(Q)\bigg|_{Q^{SP}} = 0 \tag{1.1.11}$$

There is one of these equations for every couple ab, $a \neq b$. This is not sufficient though.

There are indeed two necessary conditions which any solution $\hat{Q}$ must satisfy in order to be acceptable.

- The first necessary condition is that the solution found as a stationary point is a minimum (or a maximum depending on an arbitrary sign). In order to do that, one can look at the sign of the eigenvalues of the Hessian matrix. This is the condition of local stability.
- The second condition only holds for discrete systems. In this case, the solution should yield a positive entropy.

In general, a solution can be locally stable but have a negative entropy or the other way round.

1.1.4 Replica symmetry ansatz (RS) and its breaking (RSB)

The solution of equation (1.1.11) is a matrix and may not be easy to find unless some reasonable assumptions are made.

First of all, the matrix Q_{ab} must be symmetric. The diagonal elements can also be added for the sake of completeness, depending on the model. For the purposes of this thesis, they can be assumed to be 1. In spin glasses models, they are set to zero.

Aside for these simple remarks, the relevant observation is that every equation that has been written so far is symmetric with respect to the exchange of any couple of replica indices.

Therefore, the most natural ansatz, called replica symmetric ansatz (RS) is

$$Q_{ab} = Q \quad \forall a \neq b \tag{1.1.12}$$

One can check whether this solution is stable or not. In case it is not, the RS assumption must be dropped. The replica symmetry can be broken to different degrees until the solution is stable or just in order to approximate a stable solution. The RSB ansatz were proposed by Parisi and would. In this section, the procedure, which is nontrivial, will be briefly sketched.

The 1RSB goes as follows. Let us consider the symmetric matrix Q which is $m \times m$. Let m_1 be an integer such that $m/m_1 = n_1 \in \mathbb{N}$. Then $Q^{(1)}$ is chosen to be a $m \times m$ matrix with n_1 diagonal blocks of size $m_1 \times m_1$. Inside every diagonal block, all elements are set to a certain value Q_1 (except for the diagonal). All other elements are set to another value Q_0.

One can go further to 2RSB. In this case, let $Q^{(1)}$ be the 1RSB matrix. Let m_2 be an integer such that $n_2 = m_1/m_2 \in \mathbb{N}$. Inside each $m_1 \times m_1$ blocks one can identify n_2 diagonal sub-blocks and set all the elements they contain (except for the diagonal) to a certain value Q_2.

The procedure can be repeated indefinitely.

In the case of infinite iteration, the result should not depend on the values of $n_i = m_i/m_{i+1}$. However, if one stops after a finite number of step, the values n_i can be considered as simple variables in which the function A must be minimized. Therefore:

$$\frac{d}{dn_i} A(\{Q_i\}, \{n_i\}) = 0$$

This equation clearly does not make sense for integer m_is. Nonetheless, it should be reminded that the ultimate goal is to take the limit $m \to 0$. Therefore, all quantities should be thought of in the frame of an analytical continuation in all indices m_is. Hence n_is can be taken as real numbers.

1.1.5 Functions of overlaps between replicas

The replica trick is not just a mathematical shortcut, but yields some useful information about physics. In particular, the auxiliary parameters Q_{ab}^{SP}, i.e. the overlap between replicated W_as, are a relevant quantity.

Let us first consider a general function F of the overlap between two "real" replicas of the system, say

$$F(W_1 \cdot W_2/N)$$

The average over W_1 and W_2 is therefore an average over the overlaps between two configurations of the system. This is especially meaningful if the two systems are averaged over the same disorder ξ.

$$\begin{aligned}
\bar{F} &= \int P(d\xi) \int \frac{1}{Z_\xi^2} e^{-\beta H(W_1)} e^{-\beta H(W_2)} \; F(W_1 \cdot W_2/N) \\
&= \lim_{m\to 0} \int P(d\xi) \int e^{-\beta H(W_1)} e^{-\beta H(W_2)} Z_\xi^{m-2} \; F(W_1 \cdot W_2/N) \\
&= \lim_{m\to 0} \int P(d\xi) \int \prod_{a=1}^{m} e^{-\beta H(W_a)} \; F(W_1 \cdot W_2/N)
\end{aligned}$$

Let us now observe that this average depends only on the first two replicas, while the formula is symmetric with respect to replica permutation. This is no issue if the RS is intact. However, since it can be broken, it is best to symmetrize explicitly over the replica symmetry. This will yield, in the end, the correct result. Should this symmetrization not been done, the result would be correct only in the RS ansatz. To get the result, it is enough to perform some passages similar to those that led to (1.1.10).

$$
\begin{aligned}
\bar{F} &= \lim_{m\to 0} \int P(d\xi) \int \prod_{a=1}^{m} e^{-\beta H(W_a)} \sum_{\sigma\in S_m} \frac{F(W_{\sigma(1)}\cdot W_{\sigma(2)}/N)}{m(m-1)/2} \\
&= \lim_{m\to 0} \int P(d\xi) \int \prod_{a=1}^{m} e^{-\beta H(W_a)} \int \prod_{a<b} dQ_{ab}\, \delta(W_a\cdot W_b/N - Q_{ab}) \\
&\qquad\qquad \sum_{\sigma\in S_m} \frac{F(W_{\sigma(1)}\cdot W_{\sigma(2)}/N)}{m(m-1)/2} \\
&= \lim_{m\to 0} \int \prod_{a<b} dQ_{ab} \left[\int \prod_{a=1}^{m} dW_a\, \delta(W_a\cdot W_b/N - Q_{ab}) \int P(d\xi) \prod_{a=1}^{m} e^{-\beta H(W_a)} \right] \\
&\qquad\qquad \sum_{\sigma\in S_m} \frac{F(Q_{\sigma(1)\sigma(2)})}{m(m-1)/2} \\
&= \lim_{m\to 0} \int \prod_{a<b} dQ_{ab}\, e^{-NA[Q]} \sum_{\sigma\in S_m} \frac{F(Q_{\sigma(1)\sigma(2)})}{m(m-1)/2} \\
&= \lim_{m\to 0} \frac{2}{m(m-1)} \sum_{\sigma\in S_m} F(Q^{SP}_{\sigma(1)\sigma(2)}) \\
&= \lim_{m\to 0} \frac{2}{m(m-1)} \sum_{a<b} F(Q^{SP}_{ab})
\end{aligned}
$$

Let us first choose

$$F(W_1\cdot W_2/N) = \delta(W_1\cdot W_2/N - q)$$

Then, we can define the average overlap between two configurations of the system:

$$\bar{P}(q) := \overline{\langle \delta(W_1\cdot W_2/N - q)\rangle_{12}} \tag{1.1.13}$$

The previous computation implies that

$$\boxed{\bar{P}(q) = \lim_{m\to 0} \frac{2}{m(m-1)} \sum_{a<b} \delta(Q^{SP}_{ab} - q)} \tag{1.1.14}$$

Therefore, the fraction of configuration overlaps equal to q is the fraction of off diagonal elements of the SP solution matrix Q which equal q.

Other average quantity, which will be relevant later, are the average powers of the overlap:

$$\overline{\langle (W_1\cdot W_2/N)^k\rangle_{12}} = \lim_{m\to 0} \frac{2}{m(m-1)} \sum_{a<b} (Q^{SP}_{ab})^k \tag{1.1.15}$$

Since one can write

$$(W_1\cdot W_2/N)^k = \int dq\; q^k\; \delta(W_1\cdot W_2/N - q)$$

the average powers turn out to be the moments of $\bar{P}$

$$\boxed{\int \bar{P}(q)\; q^k = \lim_{m\to 0} \frac{2}{m(m-1)} \sum_{a<b} (Q^{SP}_{ab})^k} \tag{1.1.16}$$

1.1.6 Pure states and Gibbs states

A relevant consequence of the presence of disorder is that the energy landscape is far from being a simple. This is due to a phenomenon called frustration [23]. This is a feature of both spin glasses and neural networks.

In particular, we can focus on the minima of the energy. In the Ising model, one can know a priory what configurations represent the global and local minima of the energy: there will be just two global minima (with no external magnetic field), all spins up and all spins down. Below the critical temperature, the free energy is characterized by two minima as well, which are separated by an infinite barrier (in the thermodynamic limit) and said to be different pure states, and are characterized by a non-zero expected value of the magnetization. The convex combination of these equally weighted pure states is said to be a Gibbs state.

In disordered systems, there can be many minima of free energy. In a mean field framework, they can be thought of as separated by infinite potential barriers and identified as pure states.

Pure states can be characterized in terms of broken symmetries or in terms of correlations.

A measure of probability defines a pure state α if truncated correlations decay to zero. In systems with a nontrivial distance, the correlation decays to zero at great distance. On the other hands, in fully connected systems as spin glasses or neural networks, the truncated correlation are just required to vanish in the thermodynamic limit:

$$\langle W_i W_j \rangle_\alpha - \langle W_i \rangle_\alpha \langle W_j \rangle_\alpha \sim \frac{1}{N^\delta} \qquad \forall i, j = 1, ..., N$$

In general, **pure states** are defined by the **clustering property**

$$\left\langle \prod_{i=1}^{k} W_{j_i} \right\rangle_\alpha = \prod_{i=1}^{k} \langle W_{j_i} \rangle_\alpha \text{ with } j_i \neq j_{i'} \; \forall i, i' \tag{1.1.17}$$

This property, in particular, implies that intensive quantities do not fluctuate (i.e. are self-averaging. A simple example of self-averaging quantity is a generalized magnetization

$$M_v = \frac{v \cdot W}{N}$$

for some $v \in \mathbb{R}^N$.

A **Gibbs state** is defined as a convex linear combination of pure states.

$$\langle \ \rangle_G = \sum_\alpha w_\alpha \langle \ \rangle_\alpha \text{ with } \sum_\alpha w_\alpha = 1 \tag{1.1.18}$$

Coefficients w_α are called Gibbs weights and can be determined by looking at the expectev value of any observable:

$$\begin{aligned}
\langle F \rangle &= \frac{1}{Z} \int dW \; e^{-\beta H(W)} F(W) \\
&= \sum_\alpha \frac{\int_\alpha dW e^{-\beta H(W)}}{Z} \frac{1}{\int_\alpha dW e^{-\beta H(W)}} \int_\alpha dW e^{-\beta H(W)} F(W) \\
&= \sum_\alpha w_\alpha \frac{1}{Z_\alpha} \int_\alpha dW e^{-\beta H(W)} F(W) \\
&= \sum_\alpha w_\alpha \langle F \rangle_\alpha
\end{aligned}$$

Hence

$$w_\alpha = \frac{Z_\alpha}{Z} = \exp(-\beta(f_\alpha - f))$$

Gibbs state in general do not have the clustering property.

An **overlap between pure states** (α and β) can be defined

$$q_{\alpha\beta} = \frac{1}{N}\sum_{i=1}^{N}\langle W_i\rangle_\alpha\langle W_i\rangle_\beta \tag{1.1.19}$$

This is a consistent definition, since the $q_{\alpha\beta}$s are self-averaging thanks to the clustering property, as can be easily seen by observing that

$$\left\langle \frac{1}{N^2}\left(\sum_i W_i^\alpha W_i^\beta\right)^2\right\rangle_{\alpha\beta} = \frac{1}{N^2}\sum_i\sum_j \langle W_i^\alpha W_j^\beta\rangle_{\alpha\beta}\langle W_i^\alpha W_j^\beta\rangle_{\alpha\beta}$$
$$\overset{clustering}{=} \left(\frac{1}{N}\sum_i \langle W\rangle_\alpha\langle W\rangle_\beta\right)^2$$

Incidentally, the **self overlap** $q_{\alpha\alpha}$ is called **Edward Anderson parameter**.

The disorder averaged distribution of the overlaps between pure states $\bar{P}_s$ is a relevant quantity and can be used as an **order parameter** for disordered systems, since it reveals the structure of minima and, in particular, the breaking of the Gibbs state. To determine $\bar{P}_s$, it is enough to observe that, since $q_{\alpha\beta}$ is self-averaging, then, for any observable, it must hold

$$F(q_{\alpha\beta}) = \langle F(W_\alpha \cdot W_\beta/N)\rangle_{\alpha\beta}$$

which is manifestly a mean field condition. This implies that

$$\sum_{\alpha\beta} w_\alpha w_\beta F(q_{\alpha\beta}) = \sum_{\alpha\beta} w_\alpha w_\beta \langle F(W_\alpha \cdot W_\beta/N)\rangle_{\alpha\beta} = \langle F(W_1 \cdot W_2/N)\rangle_{12}$$

This means that averaging such observables over pure states is equivalent to averaging them over the Gibbs state. In particular, by choosing $F(q_{\alpha\beta}) = \delta(q - q_{\alpha\beta})$, it can be seen that (see (1.1.14))

$$\boxed{\bar{P}_s(q) = \bar{P}(q) = \lim_{m\to 0}\frac{2}{m(m-1)}\sum_{a<b}\delta(q - Q_{ab})} \tag{1.1.20}$$

This means that the overlap between pure states can be computed as an overlap between two real replicas in the replica formalism.

1.1.7 Replicas and physics

Now that the order parameter $\bar{P}$ has been identified, it is possible to study the physical implications of the RS and the k-RSB ansatz.

After a little reasoning, it can be concluded that the matrix Q has

$$\frac{1}{2}m(m_i - m_{i+1})$$

elements which equal q_i.

Hence, by using the RSB notation, equation (1.1.20) can be rewritten as

$$\bar{P}(q) = \lim_{m\to 0}\frac{1}{m-1}\sum_{i=0}^{k}(m_{i+1} - m_i)\,\delta(q - q_i) \tag{1.1.21}$$

Before the $m \to 0$ limit is taken, obviously $m_i > m_{i+1}$. However, once one takes the analytic continuation and m is close to 0, for $\bar{P}(q)$ to be positive, it must hold that

$$m_{i+1} \geq m_i$$

Since (1.1.21) becomes

$$\boxed{\bar{P}(q) = \sum_{i=0}^{k} (m_i - m_{i+1})\ \delta(q - q_i)} \tag{1.1.22}$$

Furthermore, $\bar{P}$ should be normalized to 1:

$$1 = \int dq\ \bar{P}(q) = \sum_{i=0}^{k} (m_{i+1} - m_i) = m_k - m_0$$

Since $m_0 = m = 0$, then

$$0 = m_0 \leq ... \leq m_i \leq m_{i+1} \leq ... \leq m_k = 1$$

If $k \to \infty$, let

$$\hat{q} : [0,1] \to \mathbb{R}^+ \quad \hat{q}(x) = q_i \text{ if } x \in (m_i, m_{i+1}] \tag{1.1.23}$$

In the $k \to \infty$ limit, $\hat{q}(x)$ becomes a continuous function. This function is invertible, since $q_{i+1} > q_i$ is a reasonable assumption. Its inverse is

$$x(q)$$

By integrating $\bar{P}$ up to a certain value q, it is clear that

$$\int_0^q dq'\ \bar{P}(q') = x(q) \tag{1.1.24}$$

Equivalently

$$\bar{P}(\tilde{q}) = \left.\frac{dx}{dq}\right|_{q=\tilde{q}} \tag{1.1.25}$$

Now it is possible to describe the structure implied by equation (1.1.21). In the RS ansatz

$$\bar{P}_{RS}(q) = \delta(q - q_0) \tag{1.1.26}$$

This means that all pure states, i.e. all minima, have the same overlap with each other. In other words, all pure states are equidistant from each other. If the RS symmetry is broken once, it means that

$$\bar{P}_{1RSB}(q) = m_1\delta(q - q_0) + (m_2 - m_1)\delta(q - q_1)$$

This distribution describes a cluster structure: there are equidistant clusters of minima in which all states are equidistant. The addition of further deltas indicates the presence of new mutually equidistant clusters inside each cluster. It can be concluded that the minima landscape is characterized by a hierarchy of concentric clusters. At each level of the hierarchy, clusters are equidistant. This structure is called **hypermetric** (see [22], [23] or [**?**] for details). If $k \to \infty$ the structure should be thought of as continuous.

Chapter 2

A brief review of some useful known results

2.1 An overview of simple perceptron

2.1.1 Simple perceptron: notation and classic results

The term neural network identifies a broad category of computing systems with very different features. Therefore, in this thesis, the focus will be kept on the feed-forward, one-layer, classifying machine called simple perceptron.

A perceptron is, in general, a system composed of nodes, also called neurons and links, called synapsis. A number, called synaptic weight, is associated to each link. The nodes are organized in layers. Each node first layer of synapsis receives an input signal, and elaborates it. Each node from the first layer produces an output. Each node from the following layers receives as an input the weighted sum (with synaptic weights) of the outputs of the previous nodes it is connected. This is iterated until the last layer yields an output. This network is feed-forward, since it does not allow back-propagation, i.e. the signal cannot be sent back to previous layers. Backpropagation is known to improve the performance.

In case of the single perceptron, there is just layer and one final node that represent the global output. The layer is described by an array of synaptic weights W of N components with

$$W_i = \pm 1 \tag{2.1.1}$$

The synaptic weights represent the links between each node from the layer and the final output node. An input pattern is a N sized vector as well. Two main models can be investigated with the replica formalism. In the **spherical** or continuous model, the patterns are vectors in $\mathbb{R}^N$ with the spherical constraint

$$\sum_i \xi_i^2 = N \text{ or } \xi \cdot \xi / N = 1 \tag{2.1.2}$$

In the Ising-spin or **discrete model**, each component is binary

$$\xi_i = \pm 1 \tag{2.1.3}$$

The output as a function of the input is, in both cases

$$\mathrm{sgn}(\xi \cdot W/N - K) \tag{2.1.4}$$

A set of patterns is a set $\xi = \{\xi_\alpha\}_{\alpha=1,\dots,p}$. From now on, a set of input patterns will de identified by the letter ξ without indices. Individual patterns in a set will be labelled

with a greek letter ξ_α. While the components of a single pattern will be referred to with a latin index (ξ_α^i).

Given a certain set of input-output pairs

$$\{(\sigma^\alpha, \xi^\alpha)\}_{\alpha=1,\dots,p}$$

the cost function is conventionally chosen as the error count

$$H_\xi(W) = \sum_\alpha \Theta(-\sigma_\alpha(\xi_\alpha \cdot W/N - K)) \tag{2.1.5}$$

A geometrical approach is possible. In this thesis, we will take the statistical approach based on the replica method. The earliest analysis of both spherical and continuous models in the replica formalism is due to Gardner and Derrida. It is worth mentioning that there exists an equivalent formalism called belief-propagation, which will not be used here.

The replica formalism allows producing phase diagrams for the perceptron. As already explained, different degrees of replica symmetry breaking correspond to different phases and a different hierarchic structure of pure states. These phase diagrams usually use K, β ("inverse temperature") and $\alpha = p/N$ as parameters. Throughout this thesis we will focus on the case $K = 0$ and $\beta = \infty$. Hence, we just refer to the literature for phase diagrams and shift the focus to a fundamental quantity, called **storage capacity**. The storage capacity, usually to as α_c. α_c is the maximum ratio p/N above which no more patterns can be typically learned. This is an intrinsically statistical quantity which depens on the statistics of the input output pairs (ξ, σ). It is indeed not true that learning is geometrically impossible above the capacity (see, for example [6]), just that it is impossible with probability 1, given a certain input statistics. Some main statistics have drawn attention in studies.

First, the maximally entropic statistics in which both ξ and σ are chosen randomly. The capacity associated to this statistic wa computed by Garner and Derrida in the spherical case. In particular, they found that the spherical capacity at $T = 0$ is

$$\alpha_c^{sph}(\beta = \infty) = 2 \tag{2.1.6}$$

This quantity is the result of an RS computation which was shown to be correct. They also computed the RS capacity for the discrete perceptron. The result was

$$\alpha_c^{dis,RS}(\beta = \infty) = 4/\pi \tag{2.1.7}$$

They also proved this number is incorrect, since the RS ansatz leads to a negative entropy in this region. Nontheless, the authors could almost guess the correct result with a simple geometric argument (also reported in [21]). Further studies, for example that by Krauth and Mezard [21], who showed that the correct capacity is

$$\alpha_c^{dis,1RSB}(\beta = \infty) = 0.833 \tag{2.1.8}$$

They also showed that the point in which the entropy becomes negative is very similar to the capacity. An important remark is that, **in the discrete perceptron, the RS solution is unstable at zero temperature**, while it becomes stable above a certain critical temperature (there is a phase transition).

All the previous results were obtained under the assumptions of random inputs with no imposed correlation. Correlated inputs can be studied as well. Early studies showed that, if inputs are correlated with each others, the capcity can go far beyond α_c [28][17][31]. More recently, Shinzato and Kabashima [28][17] showed that it is

possible to study the statistics (and the capacity) associated to correlated inputs by studying the greatest eigenvalues of the overlap matrix $\Xi_{\alpha\beta} = \xi_\alpha \cdot \xi_\beta / N$. These results are obtained through the replica and belief propagation formalisms, combined with the Haar integration. While, in this thesis, correlated inputs will be a central topic, another formalism will be used and, therefore, we refer to the aforementioned articles for these results.

A different statistics for inputs-outputs pairs is called teacher-student scenario. It is aimed at the study of generalization, and it will be the main subject of the next section.

2.1.2 Generalization: teacher-student

The perceptron is expected to generalize. In other words, it should be capable of correctly classifying, with a certain accuracy, a pattern that it has never seen before. This feature is studied in the teacher-student framework.

Let us suppose that there is a **teacher machine** with a certain synaptic configuration W_T. This means that, for any ξ the teacher produces an output $\sigma_T(\xi)$ which can be regarded as the "correct" answer. Aside from the implementation, these outputs can be essentially thought of as the results of a number of -noisy or not- measures aimed at probing a classification rule which is not known a priori.

A second machine can be called **student**. Let its synaptic configuration be W_S. The student has no direct access to the teacher's configuration, but it should learn to classify patterns from examples provided by the teacher. The student can adjust its configuration in order to simulate the behaviour of the teacher.

As already said, the teacher is just a model for an unknown source of information. In this sense, it does not have to be a simple perceptron, but it can be another kind of network as well, such as a multi-layer perceptron. In these cases, it is possible that the student cannot - not even in principle - simulate the behaviour of the teacher without committing mistakes (**unlearnable rules**).

The averaging over all possible examples constitutes the disorder, which is the core difference between the TS scenario and the simple case. In the latter setup, both inputs and outputs are chosen randomly and independently so that

$$P(\xi, \sigma) = P(\xi)\ P(\sigma)$$

In case there is a teacher, though, each output is deterministically given by W_T as $\sigma W_T(\xi)$ for each ξ. Consequently, the inputs ξ are the only quantity to average over. Subsequently, one has to integrate over all possible W_T, in order to obtain a result for the typical teacher

$$P_{TS}(\xi, \sigma) = \int_{W_T} dW_T\ P(W_T)\ \delta(\sigma_{W_T}(\xi) - \sigma)\ P(\xi)$$

The difference is crucial since, if T is a simple perceptron, then there will always be some W_S which replicate the exact behaviour of the teacher. Consequently, it may happen that $\alpha > \alpha_c$. Clearly, even though a solution $W_S = W_T$ always exists, it is not necessarily accessible, and, in case there is some noise, the student maybe unable to emulate the teacher beyond a certain accuracy.

These are several learning algorithms, with different features. There is, as always, a tradeoff between time efficiency and accuracy. Algorithms will not be discussed in this thesis. Nonetheless, it is worth mentioning the way learning rules are classified.

- **Unsupervised learning**: only input patterns are provided, with no teacher classifying them. The synaptic configuration is modelled after the input distribution.

- **Supervised learning**: the aforementioned teacher student scenario.
  - **Offline or batch learning**: the student optimizes its configuration in order to correctly associate a whole set of inputs with the outputs provided by the teacher. If new examples are provided, the old set of input-output pairs is updated and the optimization has to be repeated.
  - **Online learning**: the students modifies its configuration to adapt dynamically to the examples provided by the teacher, without optimizing with respect to the previous examples, which could be no more reproducible.

From now on, the focus will be kept on the case in which both the teacher and the student are simple perceptrons.

A key quantity is the **generalization function** E, which is the probability that, given a random input pattern ξ, the student and the teacher give a different answer. In other words, it is the likelihood that the student commits a mistake:

$$E := \int P(\xi)\ \Theta(-(\sigma_{W_T} \cdot \xi)(\xi \cdot W_S)) \tag{2.1.9}$$

As can be deduced by (2.1.9), the only quantities whose statistics is relevant are $t = \xi \cdot W_T/\sqrt{N}$, $s = \xi \cdot W_s/\sqrt{N}$, and

$$R = \frac{W_T \cdot W_S}{N} \tag{2.1.10}$$

so that

$$E = \int dt\ ds\ P(t,s)\ \sum_{\alpha=1}^{p} \Theta(-s_\alpha\, t_\alpha)$$

By the central limit theorem, s and t are distributed according to the law

$$P_R(s,t) = \frac{1}{\pi} \exp\left(-\frac{1}{2\sqrt{1-R^2}}\left[t^2 + s^2 - 2R\ st\right]\right) \tag{2.1.11}$$

which accounts for disorder. As a result

$$E(R) = \frac{1}{\pi} \arccos(R) \tag{2.1.12}$$

Given this tool, suppose that we have a teacher which provides a certain number of examples p to the teacher with a certain algorithm. Let us define $\alpha = p/N$. The more examples the students sees, the more it learns and the more R approaches 1. We can define therefore $R = R(\alpha)$ as a function of the number of examples provided. It turns out that R is **self-averaging** and can be computed with the replica approach. Then, the **learning curve of a learning algorithm** can be defined as

$$\epsilon_g(\alpha) = E(R(\alpha)) \tag{2.1.13}$$

The **perceptron generalizes because there exist algorithms for which $\epsilon_g(\alpha)$ is monotonically decreasing in α and, for big values of α, vanishes** (if there is no noise). In presence of noise, it is possible that the student cannot reproduce the teacher arbitrarily well.

It is worth mentioning that there exists an optimal, though not implementable (for a single perceptron), algorithm: the **Bayesian algorithm**. It goes as follows. Suppose p pairs $\{(\xi_\alpha, \sigma_\alpha)\}_p$ are provided by the teacher. The conditional probability of the synapsis W_S is given by Bayes theorem as

$$P(W_S|\{\sigma_\alpha\}_p) = \frac{P(\{\sigma_\alpha\}_p|W_S)\ P(W_S)}{Z} \tag{2.1.14}$$

If W_S, after p examples from the teachers, is chosen according to the probability (2.1.14) then, given a new output ξ, the likelihood of the outputs $\sigma_\pm$ can be inferred from previous examples as

$$V_\pm = P(\sigma_T(\xi) = \sigma_\pm | \{\sigma_\alpha\}_p) = \int dW_S \; P(W_S|\{\sigma_\alpha\}_p) \; \Theta(\pm\xi \cdot W_S)$$

The expected output is

$$\langle \sigma_\pm \rangle = V_+ - V_-$$

Therefore, the best possible guess is

$$\sigma = \mathrm{sgn}(V_+ - V_-)$$

The optimal Bayesian learning curve is, for big α

$$\epsilon_g(\alpha) \sim \frac{0.44}{\alpha} \tag{2.1.15}$$

2.2 The solutions' landscape

2.2.1 Local algorithms and isolated solutions

Knowing that a certain learning problem has some solutions is not enough: one has to find one. The most simple solution-seeking algorithms are local [16], such as Metropolis or gradient descent. The issue is that they do not seem to reach a global minimum in reasonable times, but they end up trapped in local minima.

Huang and Kabashima proposed a thermodynamic explanation for this computational hardness. Their investigation tool is the Franz-Parisi potential. A general introduction to the FP potential is beyond the scope of this thesis and can be found in [14]. Conversely, the formula will presented in a special case and its utility will be illustrated.

$$F(\beta, x) = \frac{1}{N} \int P(\xi) \left[\frac{1}{Z(\beta,\xi)} \sum_W e^{-\beta H_\xi(W)} \ln \sum_{\bar{W}} e^{-\beta H_\xi(\bar{W}) + xW\cdot\bar{W}} \right] \tag{2.2.1}$$

To understand the previous formula, let us first consider

$$f(x, W) = \ln \sum_{\bar{W}} e^{-\beta H_\xi(\bar{W}) + xW\cdot\bar{W}} \tag{2.2.2}$$

It can be shown that $f(x, W)$ is the free energy associated to the maximally entropic distribution with fixed average energy ($= E$) and fixed average overlap ($= p$) with a reference vector W. In other words, it is the free energy of the canonic ensemble distribution with an additional overlap constraint. To see this, one should minimize the entropy

$$S[P] = \sum_{\bar{W}} P(\bar{W}) \ln P(\bar{W})$$

with respect to $p(\bar{W})$, assuming the following constraints

$$\begin{cases} \sum_{\bar{W}} P(\bar{W}) = 1 & \text{with L. multiplier } \gamma \\ \sum_{\bar{W}} P(\bar{W}) \; H(\bar{W}) = E & \text{with L. multiplier } \beta \\ \sum_{\bar{W}} P(\bar{W}) \; \bar{W} \cdot W' = Np & \text{with L. multiplier } -x \end{cases}$$

The Lagrange equation is

$$0 = 1 + \ln P(\bar{W}) + \beta H(\bar{W}) + \gamma + x\, W \cdot \bar{W}$$

which is satisfied by the measure

$$P(\bar{W}) = \frac{1}{Z(x, \beta, W)} e^{-\beta H(\bar{W}) + x\, W \cdot \bar{W}}$$

It is immediate that the average overlap with the reference can be obtained back as

$$p = \frac{d}{dx} f(W, x) \tag{2.2.3}$$

Since $f(W, x)$ is self-averaging with respect to both the reference (W) statistics and the disorder (ξ), the FP potential $F(\beta, x)$ is just the average of the free energy $f(W, x)$ with respect to both those random variables.

To get a clearer and simpler picture, β can be set to ∞ (and so will be assumed from now on). In this way

$$e^{\beta H_\xi(W)} = \mathbb{X}_\xi(W)$$

where $\mathbb{X}_\xi(W)$ is 1 if W is a solution and 0 otherwise. In this way, only exact solutions count in the average:

$$F(\beta, x) = \frac{1}{N} \int P(\xi) \left[\frac{1}{Z_\xi} \sum_W \mathbb{X}_\xi(W) \ln \sum_{\bar{W}} \mathbb{X}_\xi(\bar{W}) e^{xW \cdot \bar{W}} \right] \tag{2.2.4}$$

Therefore, the FP potential becomes the average logarithm of the number of solutions at an average distance ($p = \frac{d}{dx} f(W, x)$) from a reference solution. In order to introduce an explicit dependence on p, the Legendre transformation can be used

$$\mathcal{V}(p) = F(x(p)) - p\, x(p) \tag{2.2.5}$$

The Legendre transformation is only defined for functions, which do not flip concavity. Consequently, if $\mathcal{V}$ at some point p assumes the "wrong" concavity, it can be deduced that there are on average no solution at overlap p. Huang and Kabashima proved that

$$\frac{d}{dp} \mathcal{V}(p) \sim \frac{\alpha C}{\sqrt{1-p}} \text{ with } p \to 1^- \text{ and } C > 0$$

which turns out to have the wrong concavity. This implies that there always exist, for any $p/N = \alpha$, a neighborhood of p where there are typically no solutions. To state this in an even clearer way, let us define the Hamming distance between two patterns

$$d(W, W') = \frac{1}{N}(W - W')^2 = \frac{1}{2}(1 - W \cdot W'/N)$$

This means that $p \sim 1$ implies $d \sim 0$. Hence, given a reference solution, there are on average no other solutions below a certain distance

$$d_{\min}(\alpha, N) = O(N) \tag{2.2.6}$$

This distance increases with both α and N. The relevant information is that typical solutions are isolated and their average distance scales linearly in N. The conclusion is that local searching algorithm have no hope to find a solution if $N \to \infty$.

A few technical notes on the computational procedure are needed. The FP potential is, as anticipated, evaluated at $\beta = \infty$. Hence

$$F(x) = \overline{\frac{1}{Z} \sum_{W} \prod_{\alpha=1}^{p} \Theta(W \cdot \xi_\alpha) \ln \sum_{\bar{W}} \prod_{\alpha=1}^{p} \Theta(\bar{W} \cdot \xi_\alpha)\, e^{xW \cdot \bar{W}}}$$

The upper line is a shorthand notation for the average over pattern disorder $\frac{1}{2^{Np}} \sum_{\{\xi\}}$. The following replica tricks are used

$$\lim_{m \to 0} \frac{dZ^m}{dm} = \ln Z$$

$$\lim_{n \to 0} Z^{n-1} = \frac{1}{Z}$$

This results in the potential becoming

$$F(x) = \lim_{m,n \to 0} \frac{d}{dm} \left\{ \sum_{\{W_a \bar{W}_b\}} \prod_{a=1}^{n} \prod_{b=1}^{m} \overline{\prod_{\alpha=1}^{p} \left[\Theta\left(W_a \cdot \xi_\alpha\right) \Theta\left(\bar{W}_b \cdot \xi_\alpha\right) \right]}\, e^{xW_1 \cdot \bar{W}_b} \right\}$$

This quantity is computed in the RS ansatz: one should fix the value of the self-averaging quantities

$$Q_{ab} = W_a \cdot W_b / \sqrt{N} \qquad P_{ab} = \bar{W}_a \cdot \bar{W}_b / \sqrt{N} \qquad R_{ab} = W_a \cdot \bar{W}_b / \sqrt{N}$$

with appropriate deltas, e.g.

$$1 = \int \prod_{ab} dQ_{ab}\, \delta(Q_{ab} - W_a \cdot W_b / \sqrt{N}) = \int \prod_{ab} dQ_{ab}\, d\hat{Q}_{ab}\, e^{i\hat{Q}_{ab}(Q_{ab} - W_a \cdot W_b / \sqrt{N})}$$

Then one assumes the RS ansatz for all variables $Q, \hat{Q}, R, \hat{R}, P, \hat{P}$ and follows the procedure. Finally, one has to find the saddle point with respect to the eight[1] parameters.

2.2.2 Clusters of subdominant minima

The computation of the previous paragraph is meaningful, but it overshadows some relevant features of the minima landscape. Zecchina et al pointed out [1][2] that, in spite of the rarity[2] of typical solutions, algorithms can be designed in order to find some minima efficiently. These accessible minima though, are not accounted for by the Franz-Parisi potential. This can be understood as follows. It is useful to recall that[3]

$$F(x) = \frac{1}{N} \int P(\xi) \left[\underbrace{\frac{1}{Z_\xi} \sum_{W} \mathbb{X}_\xi(W)}_{(!)} \ln \sum_{\bar{W}} \mathbb{X}_\xi(\bar{W}) e^{xW \cdot \bar{W}} \right] \tag{2.2.7}$$

The quenched average over the reference is highlighted with the (!) mark: since the reference is chosen according to the maximally entropic measure, any subset of minima, whose measure vanishes in the thermodynamic limit, does not contribute to the sum.

[1] Both R and $\hat{R}$ depend on two parameters in the RS ansatz: $R_{ab} = r\delta_{ab} + r'(1 - \delta_{ab})$

[2] A solution is rare if it cannot be typically found by a local search algorithm.

[3] $\mathbb{X}(W)$ is 1 if W is a solution and 0 otherwise.

In other words, only typical solutions count. Suppose for instance that the number of typical solutions (which are isolated) grows as

$$S_i \sim e^{N\Sigma_i}$$

Let us suppose that there is another set C of solutions which are not isolated, but rather organized in clusters. Suppose their number grows as

$$S_c \sim e^{N\Sigma_c}$$

with $\Sigma_c < \Sigma_i$. Then, they would be invisible to Huang and Kabashima's analysis. In other words, the FP potential fails to detect the presence of clustered solutions unless they are typical.

Zecchina et al introduced a new measure to probe the existence of clusters of subdominant solutions. Since the proposed measure differs from the natural Boltzmann weighting (notably, it is non-local), this analysis is referred to as **outside equilibrium** by the authors. The measure is

$$P(W; d, y) = \mathbb{X}_\xi(W)\mathcal{N}(W, d)^y \tag{2.2.8}$$

with

$$\mathcal{N}(W, d) = \sum_{\bar{W}} \mathbb{X}_\xi(\bar{W})\ \delta(W \cdot \bar{W}, N(1-2d)) \tag{2.2.9}$$

and y being an inverse-temperature like parameter. $P(W; d, y)$ is the y-weighted number of solutions $\bar{W}$ that lay apart from the reference W at a distance d. Form the expression for the free energy

$$\mathcal{F}(d, y) = -\frac{1}{yN} \int P(\xi)\ \ln \sum_W P(W; y, d) \tag{2.2.10}$$

it can be seen that, while the average over disorder is quenched, the sum over the references is more similar to an annealed average (it is an annealed average for $y = 1$).

Let us stop to compare the new quantity (2.2.10) with the FP potential from the previous section. The relevant parts of the two formulas are written below as F_2 and F_1 respectively.

$$\begin{aligned} F_1 =& \frac{1}{Z} \sum_W \mathbb{X}(W) \ln \sum_{\bar{W}} \mathbb{X}(\bar{W}) \delta(W \cdot \bar{W} - Nq) \\ F_2 =& \frac{1}{Z} \sum_W \mathbb{X}(W) \left[\sum_{\bar{W}} \mathbb{X}(\bar{W}) \delta(W \cdot \bar{W} - Nq) \right]^y \end{aligned}$$

To visualize how these two quantities are different, suppose that every isolated solution has on average $e^{-N\sigma_i(d)}$ other minima at distance d. This could even be a reasonable guess [4]. Then, suppose $d > \bar{d}$ ($\bar{d}$ is the typical minimal distance from a reference solution below which no other solutions can be found). Let us call $N\sigma_c(d)$

[4] The number of $\bar{W}$ with overlap q with the reference is

$$V(q) = \binom{qN}{N} \approx \exp\left[N\left(q \ln q + (1-q)\ln(1-q)\right)\right] =: e^{NS(q)}$$

The number of $\bar{W}$ with overlap $\geq q$ with the reference is

$$V(\geq q) = \int_q^1 dq'\ V(q') \approx e^{NS(q)}$$

the average number of solution surrounding, at distance d, a given reference, picked from a subdominant cluster

$$F_1 \approx \frac{e^{N\Sigma_i} \ln e^{-N\sigma_i(d)} + e^{N\Sigma_c} \ln e^{N\sigma_c(d)}}{e^{N\Sigma_i} + e^{N\Sigma_c}} \sim -N\sigma_i(d)$$

if $\Sigma_i > \Sigma_c$. On the other hand

$$F_2 \approx \frac{e^{N\Sigma_i} e^{-yN\sigma_i(d)} + e^{N\Sigma_c} e^{yN\sigma_c(d)}}{e^{N\Sigma_i} + e^{N\Sigma_c}} \sim e^{N(\Sigma_c - \Sigma_i) + yN\sigma_c(d)} \approx e^{yN\sigma_c(d)}$$

if y is big enough. The conclusion is that, in a non-equilibrium analysis, typical solutions can be redefined as those belonging to a cluster (if there is any), while isolated solution are not thermodynamically relevant.

Measure (2.2.8) can be written as a Boltzmann weight associated to the (non-local) energy

$$\mathcal{E}(W; d) = -\frac{1}{N} \ln \mathcal{N}(W; d) \tag{2.2.11}$$

What is formally equivalent to (minus) the mean energy can be considered a **internal local entropy**:

$$S_I(d) = -\langle \mathcal{E}(W; d) \rangle_{W,\xi} = \frac{1}{N} \langle \ln \mathcal{N}(W; d) \rangle_{W,\xi} = \partial_y (y\mathcal{F}(d, y)) \tag{2.2.12}$$

It is easy to see that, as long as $\mathcal{N}$ is an integer, then $\mathcal{E} \leq 0$. In particular, if every reference has just solution at distance d, then $\mathcal{E} = 0$. Therefore, the case $\langle \mathcal{E} \rangle = 0$ is a threshold between the average presence ($\langle \mathcal{E} \rangle < 0$) or absence of solution at a certain distance ($\langle \mathcal{E} \rangle > 0$). In terms of internal local entropy, a **dense cluster** is present if

$$\boxed{S_I(d, y) > 0 \ \ \forall d < \bar{d}} \tag{2.2.13}$$

for some $\bar{d}$. Moreover, the formal entropy of the system[5], which can be referred to as **external entropy**, must be positive since the system is discrete:

$$S_E(d, y) = -y[\mathcal{F} + S_I(d, y)] > 0 \tag{2.2.14}$$

The main results are the following.

- For big d, $S_I(d)$ encounters a second order transition, after which the S_I becomes the usual equilibrium entropy given by typical isolated solutions.

All polynomial contributions are being neglected. Let us part the "ball" of "radius" $1 - q$ into equal smaller "balls" each of which contains a single isolated solution. The volume of each of these balls will be something like

$$V_i \approx e^{Nc}$$

with $c > 0$. Thus, the reference's bubble contains

$$\frac{V(\geq q)}{V_i} \approx e^{N(S(q) - c)}$$

minima, which are mainly located, near its boundary

$$\frac{V(q)}{V_i} = \frac{d}{dq} \frac{V(\geq q)}{V_i} \approx \frac{V(\geq q)}{V_i}$$

Since typical minima are isolated, there will be a certain $\bar{q}$ beyond wich $S(q) < c$. Let us call $\sigma(q) = c - S(q) > 0$. Then, one can infer that each isolated solution is surrounded by an average of

$$e^{-N\sigma(q)}$$

typical solutions at distance $q > \bar{q}$.

[5] $F = \langle E \rangle - TS$. Set $T = 1/y$, $S = S_E$, $\langle E \rangle = \langle \mathcal{E} \rangle = -S_I$ and $F = \mathcal{F}$.

- For $\alpha < \alpha_c \approx 0.83$ there is always a neighborhood of $d = 0$ in which $S_I > 0$ implying the presence of clusters. These clusters shrink with the increase of α.
- For $\alpha < \alpha_U \approx 0.77$, S_I is monotonic in in d. The hypothesis is that below α_U the cluster is unique, allowing for an RS ansatz to be correct.
- S_I does not depend on α for small d, implying the existence of a very dense structure.

The practical consequence of the discovery of clusters is that, instead of searching for random minima with a local algorithm, it is possible to design efficient algorithms, which seek for the cluster. In order to do that, a new ensemble is defined: the **robust ensemble** which is a generalization of the previous measure

$$P(W;\beta,y,\gamma) = \frac{1}{Z(\beta,y,\gamma)} e^{y\Phi(W;\beta,\gamma)} \tag{2.2.15}$$

with the free local entropy

$$\Phi(W;\beta,\gamma) = \ln \sum_{\bar{W}} e^{-\beta H(W) - \gamma d(W,\bar{W})} \tag{2.2.16}$$

In this picture, most likely configurations are those which are close to many low energy configurations, but not necessarily low energy themselves. If $\gamma \to \infty$ the reference is relevant only if it very close to a low energy configuration, which imply it should have a low energy itself.

A nice feature of the robust ensemble is that the partition function has a natural interpretation in terms of real replicas. If y is an integer, then

$$\begin{aligned} Z(\beta,y,\gamma) &= \sum_{W} e^{y\Phi(W;\beta,\gamma)} \\ &= \sum_{W} \sum_{\bar{W}_a} \exp\left[-\beta \sum_{a=1}^{y} H(W_a) - \gamma \sum_{a=1}^{y} d(W,\bar{W}_a) \right] \end{aligned}$$

On can go further and trace the reference away

$$Z(\beta,y,\gamma) = \sum_{\bar{W}_a} \exp\left[-\beta \sum_{a=1}^{y} H(W_a) + A(\{W_a\};\beta,\gamma) \right]$$

$$A(\{W_a\};\beta,\gamma) = -\frac{1}{\beta} \ln \sum_{\bar{W}} e^{-\gamma \sum_{a=1}^{y} d(W,\bar{W}_a)}$$

The authors also propose some algorithms based on this formalism which can be found in [2].

Chapter 3

A different aspect of generalization: difference between outputs as a function of the difference between inputs

3.1 Notation summary: perceptron

The discrete Perceptron is a neural network which associates an output

$$\sigma = \pm 1$$

to a given input pattern

$$\xi = (\xi_1, ..., \xi_N)$$

with

$$\xi_i = \pm 1$$

The association rule is

$$\sigma = \Theta(\xi \cdot W)$$

The vectors

$$W = (W_1, ..., W_N)$$

are called synaptic weights and

$$W_i = \pm 1$$

The perceptron is said to be trained to identify a certain set of patterns

$$\xi = \{\xi^\alpha\}_{\alpha=1,...,n}$$

if, given a desired set of input-output pairs

$$\{(\xi^\alpha, \sigma^\alpha)\}$$

all patterns are correctly classified, so that

$$0 = \sum_\alpha \Theta(-\sigma^\alpha \; W \cdot \xi^\alpha)$$

In other words, a cost function or energy

$$H_\xi(W) = \sum_\alpha \Theta(-\sigma^\alpha \; W \cdot \xi^\alpha)$$

can be associated to any set of paths ξ. A synaptic weight vector W is a solution of the learning problem if $H_\xi(W) = 0$.

A conventional statistical quantity is

$$Z(\xi, \beta) = \sum_W e^{-\beta H_\xi(W)}$$

3.2 Difference between outputs as a function of the difference between inputs

3.2.1 Similarity between input patterns in terms of their memory representation

As already explained, one of the most remarkable features of the neural networks is the potential to handle more general information than the piece it is trained with. It is called generalization. As discussed in section 2.1.2, in the case of perceptron, generalization is studied in the teacher student scenario. In other words, one is concerned with the possibility to reproduce the results of a teacher machine with a student machine, which is given a number of examples of input-output pairs provided by the teacher.

In the following sections, a different approach on the generalization will be outlined. Firstly, a general argument about neural networks will be presented so that the forthcoming definitions could be applied to other situations and setups. Finally, these definitions will be declined to the perceptron case in a way that allows an analytical computation.

As explained in the introduction, a well performing machine, after being trained to correctly classify a certain input set $(\xi, \sigma(\xi))$, once presented new input set $\bar{\xi}$ "similar" to the previous one, should yield a similar output. Informally, suppose that a certain machine is trained to tell cats from "non-cats". This can be implemented by presenting the machine with a set of pictures of cats. When presented a new cat, maybe in a different position, or of a different race, the network should correctly yield the output "cat". However, will it happen? Furthermore, let us suppose that we give the machine the picture of a tiger. Will the machine consider the tiger to be enough cat-like to be classified as such? Moreover, the machine should not classify a dog as a cat. A dog, though, shares many non-superficial features with a cat. Therefore, we would like the machine not to generalize too much the "idea of cat" to the point that a dog could fit it. Finally, what will happen if we choose a new set of cats to train the machine? Will the "idea of cat" of the machine be the same as the old one? In addition, even more importantly, will the old set of cats fit it?

In order to proceed in a more formal way, we should define what it means for us that two patterns are similar. In other words, we should define our idea of "catness". This can be done by defining an external distance between any two set of patterns ξ and $\bar{\xi}$

$$d(\xi, \bar{\xi}) \tag{3.2.1}$$

If both ξ and $\bar{\xi}$ are sets of cats, their distance should be very small. Conversely, the distance should be bigger, and ideally beyond some threshold, between cats and both tigers and dogs.

A machine with a given synaptic configuration W defines a different kind of distance or "similarity" between all possible input patterns, namely:

$$\mathcal{D}_W(\xi, \bar{\xi}) = d(\sigma_W(\xi), \sigma_W(\bar{\xi})) \tag{3.2.2}$$

Suppose that the machine is trained with the set (ξ, σ) and let us call a certain solution as $W(\xi)$. Then, assuming some kind of distance is available for the outputs, a

interesting quantity is

$$\mathcal{D}(\xi, \bar{\xi} \,|\, W(\xi)) = d(\sigma_{W(\xi)}(\xi), \sigma_{W(\xi)}(\bar{\xi})) \tag{3.2.3}$$

Informally, if ξ are cats, this formula quantifies how much the machine considers the set $\bar{\xi}$ to be cat-like. The solution $W(\xi)$ is not unique. Therefore, it is reasonable to average over solutions

$$\mathcal{D}(\xi, \bar{\xi} \,|\, (\xi, \sigma)) = \int P(W(\xi))\ \mathcal{D}(\xi, \bar{\xi} \,|\, W) \tag{3.2.4}$$

The choice of $P(W(\xi, \sigma))$ is arbitrary. For example, one can use the robust ensemble to only explore robust and accessible solutions. In this thesis, the standard maximally entropic measure will be adopted:

$$P(W(\xi, \sigma)) = \frac{\delta(\sigma_W(\xi) - \sigma)}{\int P(W)\ \delta(\sigma_W(\xi) - \sigma)} \tag{3.2.5}$$

The previous formula con be loosened to allow for some noise:

$$P(W(\xi, \sigma)) = \frac{e^{-\beta H_{\xi,\sigma}(W)}}{\int dW\ e^{-\beta H_{\xi,\sigma}(W)}} \tag{3.2.6}$$

In order to compare the machine-inborn distance with the a priori distance, a comparison between the two distances can be done in probabilistic terms. Given two patterns with a distance d, what is the likelihood that the machine will attribute them a distance D, given a certain configuration W? The answer is

$$P_W(D \,|\, d) = \int P_d(\xi, \bar{\xi})\ \delta(\mathcal{D}_W(\xi, \bar{\xi}) - D) \tag{3.2.7}$$

with

$$P_d(\xi, \bar{\xi}) = \frac{\delta(d(\xi, \bar{\xi}) - d)}{\int P(\xi', \bar{\xi'})\ \delta(d(\xi', \bar{\xi'}) - d)} \tag{3.2.8}$$

Again, suppose that the machine was trained with one of the two sets, say (ξ, σ). As said before, for any learning problem there is, in general, more than one solution W, W is a random variable whose distribution depends on ξ and should be averaged upon. This quantity is

$$P(D \,|\, d) = \int P_d((\xi, \sigma), \bar{\xi})\ \ \frac{1}{Z_\xi} \int P(W(\xi, \sigma))\ \delta(\mathcal{D}_W(\xi, \bar{\xi}) - D) \tag{3.2.9}$$

and can be computed in the replica formalism. However, for consistency with literature about the perceptron, the following similar free energy will be studied:

$$\boxed{F(D \,|\, d) = \int P_d((\xi, \sigma), \bar{\xi})\ \ln \int P(W(\xi, \sigma))\ \delta(\mathcal{D}_W(\xi, \bar{\xi}) - D)} \tag{3.2.10}$$

The reason is that it gives direct information about the minima landscape of the perceptron, as will be mentioned later.

The previous definitions are quite general and no mention has been made about any specific neural network. In the following sections they will applied to the binary simple perceptron, trained with random patterns. A warning is needed: the cat-pictures'example cannot be carried any further, since it would not count as a random training set.

3.2.2 A distance for inputs in perceptron

In the previous section, the possibility to define a distance between sets of input patterns $d(\xi, \bar{\xi})$ was used. However, while to define a distance between a couple of patterns is easy, the definition of a **distance for sets of patterns**, in a strict or even loose sense, is less immediate.

The distance between patterns is called Hamming distance, and is closely related to the overlap q:

$$q(\xi_\alpha, \xi_\beta) = \frac{\xi_\alpha \cdot \xi_\beta}{N}$$

$$d(\xi_\alpha, \xi_\beta) = \frac{1}{2}\left(1 - \frac{\xi_\alpha \cdot \xi_\beta}{N}\right) = \frac{1}{2}\left(1 - q(\xi_\alpha, \xi_\beta)\right)$$

Clearly $q \in [-1, 1]$ and $d \in [0, 1]$.

Let us now consider two sets $\xi = \{\xi_\alpha\}_{\alpha=1,\dots,p}$ and $\bar{\xi} = \{\bar{\xi}_\alpha\}_{\alpha=1,\dots,p}$ containing the same number p of patterns. A possible way is to induce a distance from patterns to sets. The most natural approach would be to use the average distance among pairs

$$d_1(\xi, \bar{\xi}) = \frac{1}{p}\sum_\alpha d(\xi_\alpha, \bar{\xi}_\alpha) \tag{3.2.11}$$

This definition has a problem, though. The perceptron's energy is manifestly invariant under a permutation of the order of the input set. Basically, unless, we are interested in breaking that symmetry, it would be better if

$$d(\xi, \pi(\xi)) = 0 \text{ with } \pi \in S_p \tag{3.2.12}$$

$$d(\pi(\xi), \pi(\bar{\xi})) = d(\xi, \bar{\xi}) \tag{3.2.13}$$

where $\pi : \xi_\alpha \mapsto \xi_{\pi(\alpha)}$. (3.2.11) is not consistent with this request.

In order to write a more appropriate distance, it is worth considering what we would like two similar sets to look like. It is well known that two randomly-chosen patterns (when $N \to \infty$) are orthogonal with probability 1, as can clearly be deduced by the law of great numbers applied to the RV

$$q(\xi_\alpha, \xi_\beta) = \sum_{i=1}^{N} \frac{\xi_\alpha^i \, \xi_\beta^i}{N} \to 0$$

However, when a whole set of patterns of $p = O(N)$ is considered, the overlap distribution becomes nontrivial. This can be proved by looking at the distribution of the eigenvalues of the overlap matrix $q(\xi_\alpha, \xi_\beta)$, which is called Marchenko-Pastur law. If all patterns were orthogonal to each other, the spectral distribution would be $p(\lambda) = \delta(\lambda - 1)$, but it is not. However, as can be can be shown (see 7.0.5), as assured by the central limit theorem, most patterns still have overlaps close to zero. On the other hand, if a set ξ is identical to $\bar{\xi}$, then each pattern in ξ is identical to a pattern in $\bar{\xi}$. In terms of overlaps, we can require that almost all (up to subextensive subsets) patterns in ξ have overlap 1 with a pattern in $\bar{\xi}$. In other words, the patterns are identical pairwise. Each pair, though, will be about orthogonal to each other pair. To figure out what similar sets look like, just replace "identical" with "similar" in the previous picture. Clearly, two patterns are similar if their overlap is close to 1. A distance that describes this idea of similarity between sets is:

$$\boxed{d(\xi, \bar{\xi}) = \frac{1}{p}\min_{\pi \in S_p} \sum_{\alpha=1}^{p} d(\xi_\alpha, \bar{\xi}_{\pi(\alpha)})} \tag{3.2.14}$$

In other words, we can look for the permutation $\pi \in S_p$ which optimizes sum $\sum_{\alpha=1}^{p} d(\xi_\alpha, \bar{\xi}_{\pi(\alpha)})$ over the matchings and take that sum as a distance. This minimization is just a way to ensure the correct pairings.

This distance [1] satisfies (3.2.12), (3.2.13) and respects the triangular inequality too[2]. An overlap can be defined for set of patterns as well:

$$\boxed{q(\xi, \bar{\xi}) = \frac{1}{p} \max_{\pi \in S_p} \sum_{\alpha=1}^{p} q(\xi_\alpha, \bar{\xi}_{\pi(\alpha)})} \tag{3.2.15}$$

The distance and the overlap between sets still satisfy the property $q \in [-1, 1]$ and $d \in [0, 1]$ and the relationship

$$d(\xi, \bar{\xi}) = \frac{1}{2}(1 - q(\xi, \bar{\xi}))$$

A convention can be introduced to simplify the notation. The main concern is to keep track of the optimal couplings. For this purpose, given a pattern $\xi_\alpha \in \xi$, let us call $\bar{\alpha}$ the index of its counterpart $\bar{\xi}_{\bar{\alpha}} \in \bar{\xi}$. With this notation:

$$d(\xi, \bar{\xi}) = \frac{1}{p} \sum_{\alpha=1}^{p} \frac{\xi_\alpha \cdot \bar{\xi}_{\bar{\alpha}}}{N}$$

So far, it was argued that two sets, which are intuitively similar, have a small distance (3.2.15). Is the converse also true? Some delicate aspects should be mentioned. The matching could be unstable and rearrange significantly with to small input modifications. This could be an issue if the distance changes considerably. Furthermore, the rearrangement would imply that the intuitive argument behind pattern-pairing would fail. Unfortunately, this is statistically bound to happen. Nonetheless, it should only be a matter of concern if it affects a significant number of pairs. A rigorous analysis would certainly be required, however, in this thesis, just a heuristic argument will be given. A more careful evaluation would require the solution of an optimization probem, which is, in general, nontrivial.

Suppose that two sets has overlap $q(\xi, \bar{\xi}) \approx 1$. This can only be satisfied if a fraction ≈ 1 of pairs have overlap ≈ 1 (say $1 - O(\epsilon)$). In this case, the only possible ambiguity arises from the possibility that two pairs "meet", i.e. the overlap between the four of them is pairwise $1 - O(\epsilon)$. As explained before, this seems to be very unlikely, so that it should not change the average too much. The conclusion is that **the closer the distance is to zero, the more meaningful it is**. Consequently, the results obtained with this definition should not be relied on too much when the two sets are too distant.

To investigate this further, suppose that two input sets have overlap $q(\xi, \bar{\xi}) = q \gg 1$ and that the matchings are well defined and stable. How is the typical distribution of the overlaps between pairs? Let us call

$$q_\alpha = q(\xi_\alpha, \bar{\xi}_{\bar{\alpha}})$$

[1] Technically it is a pseudodistance unless one identifies patterns with overlap 1.

[2] Proof of the triangle inequality. Suppose $\pi, \eta \in S_p$:

$$\frac{1}{p} \sum_{\alpha} d(\xi_{\pi(\alpha)}, \bar{\xi}_{\eta(\alpha)}) \leq \frac{1}{p} \sum_{\alpha} d(\xi_{\pi(\alpha)}, \xi'_\alpha) + \frac{1}{p} \sum_{\alpha} d(\xi'_\alpha, \bar{\xi}_{\eta(\alpha)})$$

Let us minimize the r.h.s of the previous inequality with respect to η and π. Moreover, suppose the optimal permutations are $\bar{\eta}$ and $\bar{\pi}$ respectively. Then

$$d(\xi, \bar{\xi}) = \frac{1}{p} \min_{\tau \in S_p} \sum_{\alpha} d(\xi_{\tau(\alpha)}, \bar{\xi}_\alpha) \leq \frac{1}{p} \sum_{\alpha} d(\xi_{\bar{\pi}(\alpha)}, \bar{\xi}_{\bar{\eta}(\alpha)}) \leq d(\xi, \xi') + d(\xi', \bar{\xi})$$

so that

$$q = \frac{1}{p}\sum_{\alpha} q_\alpha$$

The goal is to compute the joint distribution of the overlaps. In absence of constraints, each overlap would be distributed normally (central limit theorem) as

$$P(q_\alpha) = \sqrt{\frac{N}{2\pi}}\exp(-Nq_\alpha^2/2)$$

Hence

$$P(\{q_\alpha\}|\sum_{\alpha=1}^{p} q_\alpha = q) = \frac{1}{P(q)}\prod_{\alpha=1}^{p} P(q_\alpha)\ \delta\left(\frac{1}{p}\sum_{\alpha=1}^{p} q_\alpha - q\right) \tag{3.2.16}$$

It can be shown that, in the thermodynamic limit, only the configuration

$$q_\alpha = q\ \ \forall\alpha = 1, ..., p \tag{3.2.17}$$

counts in the previous expression. To understand this, consider that $\prod_\alpha P(q_\alpha) \propto \exp(-N\sum_\alpha q_\alpha^2/2)$. Therefore, the only relevant configuration is the one (or the ones) that minimizes $\sum_\alpha q_\alpha^2$ with respect to $\{q_\alpha\}$, on the hypersurface $\sum_\alpha q_\alpha/p - q = 0$, which corresponds to the constraint imposed by the delta function. (3.2.17) is obtained from the minimization via Lagrange multipliers. It follows from the previous analysis that, up to matching ambiguities, typical configurations of $\{q_\alpha\}$ are given by (3.2.17). Therefore

$$P(\{q_\alpha\}|\sum_{\alpha=1}^{p} q_\alpha = q) \to \prod_{\alpha=1}^{p}\delta(q_\alpha - q) \tag{3.2.18}$$

All these results can be put together in order to compute $P(d|D)$ as given by (3.2.9). For this purpose, an expression for the probability $P_d(\xi, \bar{\xi})$ is needed. If pairing ambiguities are disregarded, with the aforementioned caveats, we replace $P_d(\xi, \bar{\xi})$ with an averaging over pairs of patterns $(\xi_\alpha, \bar{\xi}_{\bar{\alpha}})$ with given overlap q_α. In this framework, overlaps are distributing according to (3.2.16) which reduces to (3.2.18). Clearly, it is implied that $q = (1-d)/2$. Finally:

$$\boxed{P_d(\xi, \bar{\xi}) = \prod_{\alpha=1}^{p} P(\xi_\alpha)\ P(\bar{\xi}_{\bar{\alpha}})\ \delta(d - d(\xi_\alpha, \bar{\xi}_{\bar{\alpha}}))} \tag{3.2.19}$$

This formula has two good features. The first is that it is suitable for analytic computations. The second is that, even if d is not small and matching ambiguities may arise, it still has a clear meaning and can stand alone as a starting point for a reasonable computation. $P_d(\xi, \bar{\xi})$ as defined above is the probability distribution of two random input sets, in which each pattern from the first set has a fixed correlation with an output from the second set. For this reason, this definition makes sense even if $q < 0$ $(d > 1/2)$.

3.2.3 Output difference for fixed inputs' difference: perceptron

Distance (3.2.15) can be used to compute $P(D\,|\,d)$ (3.2.9) or $F(D|d)$ 3.2.10. The first step is to define a distance for outputs. The most obvious choice is to count the number of pairs (as defined by the optimization in (3.2.15)) whose output coincide. Since the output is either ± 1, a possible definition for output distance is

$$\mathcal{D}_W(\xi, \bar{\xi}) = \frac{1}{p}\sum_{\alpha=1}^{p}\Theta(-\sigma_W(\xi_\alpha)\,\sigma_W(\bar{\xi}_{\bar{\alpha}})) = \frac{1}{p}\sum_{\alpha=1}^{p}\Theta(-W\cdot\xi_\alpha\ W\cdot\bar{\xi}_{\bar{\alpha}}) \tag{3.2.20}$$

Before proceeding any further, it is convenient to recall that, since the energy is a function of the product $\sigma_\alpha\,\xi_\alpha$, when averaging over both input $\{\xi_\alpha\}$ and output $\{\sigma_\alpha\}$, all outputs can be set to $+1$:

$$\frac{1}{2^N}\sum_{\{\sigma_\alpha\}}\frac{1}{2^{pN}}\sum_{\{\xi_\alpha\}} f(\{\sigma_\alpha\,\xi_\alpha\}) = \frac{1}{2^{pN}}\sum_{\{\sigma_\alpha\,\xi_\alpha\}} f(\{\xi_\alpha\}) = \frac{1}{2^{pN}}\sum_{\{\xi_\alpha\}} f(\{\xi_\alpha\})$$

This can still be done when averaging over $P_d((\sigma,\xi)\,\bar{\xi})$. Consider, for instance

$$\bar{f} = \frac{1}{2^N}\sum_{\{\sigma_\alpha\}}\frac{1}{2^{2pN}}\sum_{\{\xi_\alpha\}}\sum_{\{\bar{\xi}_\alpha\}} f(\{\sigma_\alpha\,\xi_\alpha\},\{\sigma_\alpha\,\bar{\xi}_{\bar{\alpha}}\},\{\xi_\alpha\cdot\bar{\xi}_{\bar{\alpha}}\})$$

The function f from above only depends on the optimal couplings $\alpha \leftrightarrow \bar{\alpha}$. Let us now consider a partition of the space $\{(\bar{\xi},\xi)\}$ as

$$\{(\bar{\xi},\xi)\} = \bigcup_{\pi\in S_p} V_\pi$$

where V_π is the set of pairs of sets $(\xi,\bar{\xi})$ whose distance (3.2.15) is optimized by the permutation π. It should be pointed out that, $\forall\epsilon_\alpha = \pm 1$

$$(\{\xi_\alpha\},\{\bar{\xi}_\beta\}) \in V_\pi \implies (\{\epsilon_\alpha\xi_\alpha\},\{\epsilon_\alpha\bar{\xi}_{\bar{\alpha}}\}) \in V_\pi$$

since the distance (3.2.15) is a function of the products $\xi_\alpha\cdot\bar{\xi}_{\bar{\alpha}}$. Then

$$\begin{aligned}
\bar{f} &= \frac{1}{2^N}\sum_{\{\sigma_\alpha\}}\frac{1}{2^{2pN}}\sum_{(\xi,\bar{\xi})\in V_\pi} f(\{\sigma_\alpha\,\xi_\alpha\},\{\sigma_\alpha\,\bar{\xi}_{\bar{\alpha}}\},\{\xi_\alpha\cdot\bar{\xi}_{\bar{\alpha}}\}) \\
&= \frac{1}{2^N}\sum_{\{\sigma_\alpha\}}\frac{1}{2^{2pN}}\sum_{(\{\sigma_\alpha\xi_\alpha\},\{\sigma_\alpha\bar{\xi}_{\bar{\alpha}}\})\in V_\pi} f(\{\xi_\alpha\},\{\bar{\xi}_{\bar{\alpha}}\},\{\xi_\alpha\cdot\bar{\xi}_{\bar{\alpha}}\}) \\
&= \frac{1}{2^{2pN}}\sum_{(\xi,\bar{\xi})\in V_\pi} f(\{\xi_\alpha\},\{\bar{\xi}_{\bar{\alpha}}\},\{\xi_\alpha\cdot\bar{\xi}_{\bar{\alpha}}\}) \\
&= \frac{1}{2^{2pN}}\sum_{\{\xi_\alpha\}}\sum_{\{\bar{\xi}_\alpha\}} f(\{\xi_\alpha\},\{\bar{\xi}_{\bar{\alpha}}\},\{\xi_\alpha\cdot\bar{\xi}_{\bar{\alpha}}\})
\end{aligned}$$

The conclusion is that all σ_α can be set to 1, as expected. Therefore, $\mathcal{D}$ is the fraction of patterns in $\bar{\xi}$ which yield output -1. Consequently

$$\delta(D-\mathcal{D}_{W(\xi)}(\xi,\bar{\xi})) = \sum_{\{\eta_\mu=\pm1\}}\delta\left(\sum_\mu\eta_\mu, p(1-2D)\right)\prod_{\mu=1}^{p}\Theta\left(\xi_\mu\cdot W\right)\prod_{\bar{\mu}=1}^{p}\Theta\left(\eta_\mu\bar{\xi}_{\bar{\mu}}\cdot W\right) \tag{3.2.21}$$

The previous computations and definitions can be put together in order to write the expression for $F(d|D)$ (at zero temperature)

$$\boxed{F(d|D) = \frac{1}{N}\int P_d(\xi,\bar{\xi})\ \ln\int P(W)\ \prod_{\alpha=1}^{p}\Theta(\xi_\alpha\cdot W)}$$

$$\boxed{\sum_{\{\eta_\mu=\pm1\}}\delta\left(\sum_\mu\eta_\mu, p(1-2D)\right)\prod_{\mu=1}^{p}\Theta\left(\xi_\mu\cdot W\right)\prod_{\bar{\mu}=1}^{p}\Theta\left(\eta_\mu\bar{\xi}_{\bar{\mu}}\cdot W\right)} \tag{3.2.22}$$

As anticipated, this quantity can be interpreted in terms of configurations' landscape. **Given two random input sets ξ and $\bar{\xi}$ with distance d, $\exp NF(D|d)$ is the typical fraction of configurations which are solutions to both up to pD outputs.**

To introduce noise, one can substitute

$$\Theta(x) \mapsto \exp(-\beta\Theta(-x))$$

in the previous expression.

Let us focus on the special case $D = 0$. In this case:

$$F(0|d) = \int P_d(\xi, \bar{\xi}) \ln \int P(W) \prod_{\alpha=1}^{p} \Theta(\xi_\alpha \cdot W)\, \Theta(\bar{\xi}_{\bar{\alpha}} \cdot W) \tag{3.2.23}$$

The argument of logarithm can be generalized to define an **overlap of sets of input patterns in terms of memory representation**:

$$\boxed{\mathcal{Q}(\xi, \bar{\xi}) := \frac{\int dW\ \exp\left(-\beta[H_\xi(W) + H_{\bar{\xi}}(W)]\right)}{\sqrt{\int dW\ \exp\left(-\beta H_\xi(W)\right)\ \int dW\ \exp\left(-\beta H_{\bar{\xi}}(W)\right)}}} \tag{3.2.24}$$

The quantity $\mathcal{Q}$ is the **normalized number of common minima** (plus eventual noise) **shared by two sets** ξ and $\bar{\xi}$. $\mathcal{Q}$ can be used to define a distance

$$\boxed{\mathbb{D}(\xi, \bar{\xi}) := -\frac{1}{N} \ln \mathcal{Q}(\xi, \bar{\xi})} \tag{3.2.25}$$

$\mathbb{D}$ is positive and possesses the triangle inequality thanks to Cauchy-Swartz inequality, which can be applied to "vectors" $e^{-\beta H_\xi(W)}$ for which W plays the role of an index. It follows from the definition that $\mathbb{D}$ is a self-averaging quantity and

$$\bar{\mathbb{D}}(d) = \int P_d(\xi, \bar{\xi})\, \mathbb{D}(\xi, \bar{\xi}) = P(0|0) - P(0|d) = \frac{\overline{\ln Z}}{N} - P(0|d) \tag{3.2.26}$$

$\bar{\mathbb{D}}(d)$ yields the typical normalized number of common solutions associated to two input sets with fixed distance d.

3.2.4 Generalization and correlation between inputs in simple perceptron: the function $F(D|d)$.

In the previous section, the quantity $F(D|d)$ has been introduced. Since it is the fundamental quantity in this thesis and it will be the key object of the computation in the last chapters, it deserves a section alone. It is worth rewriting its expression here

$$\boxed{F_\alpha(d|D) = \frac{1}{N} \int P_d(\xi, \bar{\xi})\ \ln \int P(W) \prod_{\alpha=1}^{p} \Theta(\xi_\alpha \cdot W)}$$

$$\boxed{\sum_{\{\eta_\mu = \pm 1\}} \delta\left(\sum_\mu \eta_\mu, p(1-2D)\right) \prod_{\mu=1}^{p} \Theta\left(\xi_\mu \cdot W\right) \prod_{\bar{\mu}=1}^{p} \Theta\left(\eta_\mu \bar{\xi}_{\bar{\mu}} \cdot W\right)} \tag{3.2.27}$$

As anticipated, this quantity can be interpreted in terms of configurations' landscape. **Given two random input sets ξ and $\bar{\xi}$ whose patterns have distance d pairwise, $\exp NF(D|d)$ is the typical fraction of configurations which are solutions to both inputs up to a fraction D outputs.**

This quantity offers a double interpretation, which allows connecting the two ideas: correlation between inputs and generalization.

- **The solutions' landascape**. If $F(D|d)$ is finite, for any input pair $(\xi, \bar{\xi})$ so that $d(\xi_\mu, \bar{\xi}_{\bar{\mu}}) = d$, there exists a thermodynamically relevant[3] class of configurations which classifies a fraction $1 - D$ of patterns pairs in the same way. In particular, if we set $D = 0$, we can find out how many solutions are typically shared by two input sets which have distance d.

  For a given d, if $F(D|d) = -\infty$, it means that the network cannot be taught to associate a fraction $1 - D$ of pairwise identical outputs to two input sets with distance d.

  Consider now $d < 1/2$. It is possible to define

$$\boxed{D_{min}(d) = \min_D \{D : F(D|d) = -\infty\}} \tag{3.2.28}$$

  If generalization is the ability to yield similar outputs for similar inputs, then $D_{min}(d)$ is the threshold D below which the network cannot be taught to generalize for a given distance d.

  There is an equivalent point of view too. If $F(D|d) > -\infty$, then the network can be taught to resolve, with precision $1 - D$, two inputs that lies d apart. In this case, $D_{min}(d)$ represents the maximal precision with which two input sets, laying d apart, can be distinguished.

  The fact that some configuration exist and can be found, does not mean that they are possible to be found when picking a random solution in a certain class.

- **Conditional probability**. A further step is to consider the conditional probability of D with respect to d. While we do not have direct access to this quantity (we should compute (3.2.9)), we can study it in the following way. We can compute the fraction between the typical number of common minima with fixed D and the typical number of common minima with any value of D. Then, up to subexponential values in N:

$$\begin{aligned} \bar{P}(D|d) &\approx \frac{\exp\left(NF(D|d)\right)}{\int_0^1 dD' \ \exp\left(NF(D'|d)\right)} \\ &\approx \exp\left(N[F(D|d) - \min_D F(D|d)]\right) \end{aligned}$$

  Therefore, there are two possibilities:

$$\bar{P}(D|d) \begin{cases} = 0 & F(D|d) > \min_D F(D|d) \\ \neq 0 & F(D|d) = \min_D F(D|d) \end{cases} \tag{3.2.29}$$

  One can get information about the generalization by studying

$$\{D : P(D|d) \neq 0\} = \{D : F(D|d) > \min_D F(D|d)\} \tag{3.2.30}$$

  This would show what outcomes D are possible when a perceptron trained with a set ξ is presented a similar (distance d) input $\bar{\xi}$ set.

We must remark that there is a possible limitation about this approach. In the previous sections, the generalization performance have been studied as an average over configurations. However, a single perceptron possesses a single configuration. Therefore, it is possible that these optimal performances, like for the bayesian algorithm,

[3] That grows as e^{NF}.

cannot be achived by a single perceptron. Instead, a number of perceptrons could be trained to solve the same problem: then, for any single input, their outputs are averaged into one. This is called commitee machine in literature. In could be worth investigating this aspect in future works, i.e. whether single configurations possess, on average, ceratin generalization capabilities, or whether these performances only arise from averge over many solutions.

Finally, the capacity can be defined as a function of d and D as the minimal ratio p/N for which $F_\alpha(D|d) = -\infty$:

$$\alpha_c(D|d) = \min\{\alpha : \ F_\alpha(D|d) = -\infty\} \tag{3.2.31}$$

This quantity can be referred to as a **generalization capacity**. It indicates the boundaries, for fixed α, of the graph of $F(D|d)$. It will turn out that this quantity can be computed analytically in the RS assumption. Its expression and the conclusions that can be drawn from it are presented in section **??**.

3.2.5 Overview and outlook

Summary of the previous sections of this chapter:

- A possible approach to generalization based on the probability of the outputs difference, given the input difference, has been outlined and quantified by $P(D|d)$ (see (3.2.9)) and especially $F(D|d)$ (see (3.2.10)).
- In the case of perceptron, two distances for input sets of patterns, one "intrinsic" (d, see (3.2.15)) and one "machine-based" ($\mathcal{D}$), have been proposed. $F(D|d)$ has been written in this framework.
- An approximated and practical way to compute $F(D|d)$ with distance (3.2.15) has been explained (see (3.2.17) and (3.2.18)). It has been shown that this approximation has a meaning itself even in region where distance (3.2.15) is not well defined.
- An additional distance $\mathbb{D}$ (3.2.25) for inputs, counting the number of common solutions, has been shown to arise naturally from the expression of $F(0|d)$.
- In section 3.2.4, it has been explained how $F(D|d)$ can be employed as a tool to explore generalization properties in terms of correlated inputs.
- The relationship between the approach of this thesis and upper-bound results from the teacher-student scenario deserves a deeper exam. In particular, there exists an extensive literature which approaches the learning problem from a geometrical point of view.

In the following chapters, the quantities $F(D|d)$ and $\bar{\mathbb{D}}(d)$ will be studied with the replica formalism. In the spherical case, the RS computations should be correct, while, in the discrete case, the RS ansatz should only be regarded as a first step and it is not expected to yield physical results.

Aside from the general expression of $F(D|d)$, the focus will be kept on two limits. The first limit will be taken to ensure that known quantities (like the capacity) are recovered. The second limit will yield the generalization capacity as defined in section 3.2.4.

Finally, the quantity $\mathbb{D}$ could reveal some property of the clusters in the discrete perceptron. In particular, the solutions belonging to clusters are thought to be better at generalizing that isolate ones. In order to investigate this feature with the previous

tools, consider the following hypothesis. As before, let us assume that, for a given input set ξ, isolated solutions grow as $e^{\Sigma_i N}$, while clustered solutions as $e^{N\Sigma_c}$ with $\Sigma_i > \Sigma_c$. Now choose a similar input set, say $\bar{\xi}$. If $q(\xi, \bar{\xi}) = 1$, then they share all solutions. If $\bar{\xi}$ is slightly modified, then its solutions should change as well. In particular, clusters and isolated solutions should rearrange and/or shift. In case they shift, a small displacement should be enough to move isolated solutions away for their original sites. Therefore, the number $e^{N\Sigma_i(q)}$ of typical common minima between ξ and $\bar{\xi}$ should drop quickly as q decreases. On the other hand, slightly shifted clusters should still have a nonempty intersection with the non-shifted ones. Therefore, the number $e^{N\Sigma(q)}$ of common clustered minima should decrease as well, but at a different rate. It always holds that

$$\frac{1}{N}\ln(e^{N\Sigma_i(q)} + e^{N\Sigma_i(q)}) \sim \max(\Sigma_i(q), \Sigma_c(q))$$

Consequently, for fixed α, a possible point on non-analyticity could appear if at some q it happens that $\Sigma_c(q) = \Sigma_i(q)$. This could reveal a passage from a phase dominated by isolated solutions to a phase dominated by clusters. Conversely, for fixed q, we could investigate whether anything indicates the breaking of the cluster around α_U (or some $\alpha_U(q)$). A RSB approach is most likely needed to test these hypotheses.

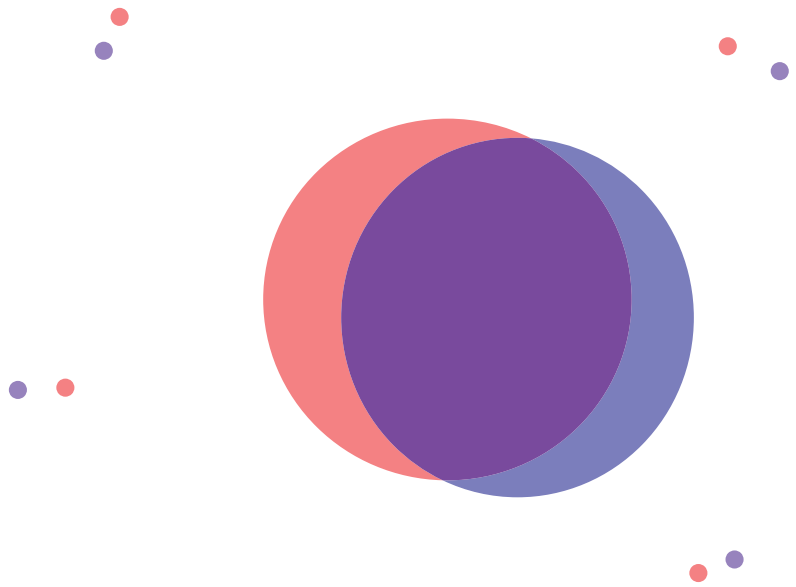


Figure 3.1: A qualitative picture. The solutions corresponding to ξ and $\bar{\xi}$ are painted in red and blue respectively. The same small shift is shown to separate isolated minima but not clusters.

Chapter 4

Computation of $F(0|d)$: the simple case subextensive inputs

4.1 Inputs of subextensive size for the discrete perceptron: $p/N \to 0$

4.1.1 Introduction and structure of the following sections

The goal of this chapter is to compute $F(D|d)$. The first step is to attempt the computation for input sets of subextensive clusters. For this purpose, the $N \to \infty$ limit will be taken, while p will be kept finite. As it will be shown, it is possible to obtain a result without using the replica trick with Gaussian variables. From a closer evaluation of the Gaussian procedure, it will be shown that it might be possible to use a similar analysis to compute the result for extensive inputs $p = O(N)$. Nonetheless, the most natural (approximated) generalization to the extensive case yields incorrect results and, therefore, further studies would be necessary to confirm or rule out such a possibility. In the following sections we will just deal with the $D = 0$ case for the discrete perceptron, since the more general scenario $D \neq 0$ will be accessible from the $p/N \to 0$ limit of the forthcoming replica computation. On the other hand, this special case ($D = 0$ and $p/N \to 0$) provides a useful check for the full results from the replica calculation.

Finally, it is worth mentioning that another reason why the $p/N \to 0$ case deserves a separate approach is that it is numerically accessible. Numerical simulations aimed at probing statistical properties of the perceptron may be very difficult, since they often involve some kind of solutions'counting. It has already been explained why, in the discrete model, finding typical solutions can be a very hard task: local algorithms cannot reach them and cluster-seeking algorithms ignore them. Even when studying clusters with dedicated algorithms, one has to be careful about comparing the sampling statistics with Boltzmann statistics.

Before proceeding to the actual computation it is worth pointing out that the $p/N \to 0$ case should be treated with a little extra care as long as the thermodynamic limit is involved. As it will be clear from the following computation, there is a finite fraction, say C, of the synaptic weights' space which is occupied by solutions. $\mathcal{Q}$ in this case would be something like

$$\frac{1}{N} \ln(2^N C) \to \ln 2$$

Therefore, in order to obtain a nontrivial result, it is very important to normalize to

1 the sum over synaptic weights

$$\frac{1}{2^N}\sum_W$$

while in the extensive case, this is not necessary.

The quantity that it is going to be computed is, therefore, the average of

$$\mathcal{Q}(\xi,\bar{\xi}) = \frac{1}{p}\ln\frac{1}{2^N}\sum_W\prod_\mu^p\Theta\left(\frac{1}{\sqrt{N}}W\cdot\xi_\mu\right)\prod_{\bar{\mu}}^p\Theta\left(\frac{1}{\sqrt{N}}W\cdot\bar{\xi}_{\bar{\mu}}\right) \tag{4.1.1}$$

which is $F(0|q) =: F(q)$

4.1.2 Analysis

The procedure to complete this computation consists in three steps:

- introducing an integration over Gaussian variables (with a quadratic form Ξ) in place of a sum over configurations W for fixed ξ
- taking a limit of the quadratic form Ξ
- computing the remaining Gaussian integral

Consider a given realization of the disorder $(\xi,\bar{\xi})_q$. A different notation for the sets ξ and $\bar{\xi}$ is not relevant now. Consequently, for now, we will consider a single set ξ with size $2p$.

The synaptic weights W only appear in the following scalar products:

$$W_\alpha = \frac{1}{\sqrt{N}}\xi_\alpha\cdot W$$

$\alpha = 1, ..., 2p$. Therefore, it makes sense to try to replace the sum over W with the integration of the joint probability of $\{W_\alpha\}$. Each W_α is a sum of independently and identically distributed random variables divided by the square root of their number. The central limit theorem implies that its marginal distribution is

$$P(W_\alpha)\sim\frac{1}{\sqrt{2\pi}}e^{-W_\alpha^2/2}$$

One could guess that the joint distribution is given by a Gaussian distribution whose quadratic form is given by fixing the correlations:

$$\langle W_\alpha W_\beta\rangle = \frac{\xi_\alpha\cdot\xi_\beta}{N}$$

It is immediate that the quadratic form would be the inverse of

$$\Xi = \begin{bmatrix} 1 & \xi_1\cdot\xi_2/N & \xi_1\cdot\xi_3/N & ... & \xi_1\cdot\xi_{2n}/N \\ \xi_2\cdot\xi_1/N & 1 & \xi_2\cdot\xi_3/N & ... & \xi_2\cdot\xi_{2n}/N \\ ... & ... & ... & ... & ... \\ \xi_{2n}\cdot\xi_1/N & \xi_{2n}\cdot\xi_2/N & \xi_{2n}\cdot\xi_3/N & ... & 1 \end{bmatrix}$$

i.e.

$$\Xi_{\mu\nu} = \frac{1}{N}\xi_\mu\cdot\xi_\nu$$

This can be shown to be correct as proven in section 4.2.1. The result is

$$P(\{W_\xi\})\sim\frac{1}{\sqrt{(2\pi)^p\det\Xi}}\exp\left(\sum_{\mu<\nu}\xi^\mu\Xi^{-1}_{\mu\nu}\xi^\nu\right)$$

As explained in section 4.2.1, the fact that the scalar products W_α only appear inside the theta functions would allow to apply this result even to extensive inputs. The reason is that the Heaviside theta provide a sort of scale invariance.

Let now consider the averaging over inputs $(\xi, \bar{\xi})$. In the Gaussian formalism, $(\xi, \bar{\xi})$ only appear in the overlap matrix $\Xi_{\mu\nu}$. Instead of integrating over input patterns, one can integrate over the overlap matrix. This passage resembles the procedure followed in [28][17] in which, however, the integration over the Haar measure and the Machenko-Pastur distribution is combined with the replica/belief propagation formalisms. In the present case, on the other hand, this is not needed.

Let us consider the factorized measure

$$P(\xi, \bar{\xi})_q = \prod_\mu \frac{1}{2^N} \sum_{\xi_\mu, \bar{\xi}_{\bar{\mu}}} \delta(q - \xi_\mu \cdot \bar{\xi}_{\bar{\mu}}/N)$$

as it was defined before. Then, from the law of great numbers it follows that

$$P(\Xi_{\mu\nu}) \to \delta_{\mu\nu}$$

$$P(\Xi_{\mu\bar{\nu}}) \to q\delta_{\mu\bar{\nu}}$$

$$P(\Xi_{\bar{\mu}\bar{\nu}}) \to \delta_{\bar{\mu}\bar{\nu}}$$

This means that it is legitimate to consider

$$\Xi \mapsto \delta_{\mu\nu} + \delta_{\bar{\mu}\bar{\nu}} + 2q\ \delta_{\mu\bar{\nu}}$$

since p is subextensive while the vanishing matrix elements decay as a function of N. In case $p = O(N)$ while single matrix elements of Ξ still vanish with N, their number grows as N^2 so that their overall contribution is not neglectable. This will be remarked in the next section as well.

4.1.3 Computing $F(0|d)$

By putting together the results from the previous paragraph, we get:

$$F(q) = \frac{1}{p} \int \prod_{\mu'} dq_{\mu'}\ \delta(q_{\mu'} - q)\ \ln \frac{1}{\sqrt{(2\pi)^p \det \Xi}} \int \prod_\mu dW_\mu d\bar{W}_\mu\ e^{-\frac{1}{2}\sum_\mu \sum_{\alpha\beta} W_\mu^\alpha G(q_\mu) W_\mu^\beta} \prod_\mu \Theta(W_\mu)\Theta(\bar{W}_\mu) \quad (4.1.2)$$

with

$$W_\mu^1 = W_\mu \qquad W_\mu^2 = \bar{W}_\mu$$

and

$$G(q) = \frac{1}{1-q^2} \begin{bmatrix} 1 & -q \\ -q & 1 \end{bmatrix}$$

The logarithm is therefore the sum of p terms

$$\ln \int dW_\mu d\bar{W}_\mu\ e^{-\frac{1}{2}\sum_{\alpha\beta} W_\mu^\alpha G(q_\mu) W_\mu^\beta} \Theta(W_\mu)\Theta(\bar{W}_\mu)$$

The eigenvectors of the matrix in the exponential are

$$W_\mu^\pm = \frac{1}{\sqrt{2}}(W_\mu \pm \bar{W}_\mu)$$

with eigenvalues

$$\lambda_{\pm}(q) = \frac{1}{1 \pm q}$$

Hence

$$\begin{aligned}
&\ln \frac{1}{2\pi\sqrt{(1-q^2)}} \int dW_+ \, dW_- \, e^{-\frac{1}{2}(\lambda_- W_-^2 + \lambda_+ W_+^2)} \Theta(W_+ + W_-)\Theta(W_+ - W_-) \\
&= \ln \frac{1}{2\pi\sqrt{(1-q^2)}} \int_{5\pi/4}^{7\pi/4} d\theta \int_0^{\infty} dr \; r \; e^{-\frac{1}{2}r^2(\lambda_+ \sin^2\theta + \lambda_- \cos^2\theta)} \\
&= \ln \frac{1}{2\pi\sqrt{(1-q^2)}} \int_{5\pi/4}^{7\pi/4} d\theta \; \frac{1}{\lambda_+ \sin^2\theta + \lambda_- \cos^2\theta} \\
&= \ln \left[1 - (2/\pi)\tan^{-1}(\sqrt{\lambda_+/\lambda_-})\right]
\end{aligned}$$

Hence

$$\bar{F}(q) = \ln \left[1 - (2/\pi)\tan^{-1}(\sqrt{(1-q)/(1+q)})\right]$$

for $N \to \infty$, $\forall p << N$.

It is worth noticing that as $q \sim 1$, then

$$F(q) \sim \sqrt{(1-q)/2}$$

As expected, there is no complicated structure of memories if $p/N \to 0$. Similar solutions share a finite fraction which drops abruptly from 1 (identical-pattern case) but is otherwise a smooth function. The shared fraction is

$$C(q) = 2^{-p}[1 - (2/\pi)\tan^{-1}(\sqrt{(1-q)/(1+q)})]^p \tag{4.1.3}$$

as shown in figure (4.1).

The interpretation in the following. If the patterns were identical pairwise, the space of solution would be halved by the addition of any new pair. On the other hand, pairs of patterns with overlap q reduce the space of solutions by a factor $C(q)/2 < 1/2$.

Finally, it can be explicitly shown that this result cannot be extended to finite p/N. If $q = 1$ we are left with a single set of patterns since ξ and $\bar{\xi}$ coincide. According to (4.1.3), the solutions for an input of size p are $1/2^p$ of the space of all configurations. The number of total configuration is 2^N. Therefore, there is less than one solution if $p > N$. That would imply that the discrete capacity is $\alpha_c^{dis} = 1$. While this is a better result that the RS capacity, it is not correct. The conclusion is that a naive approach, in which all patterns in a single input are assumed orthogonal, is incorrect.

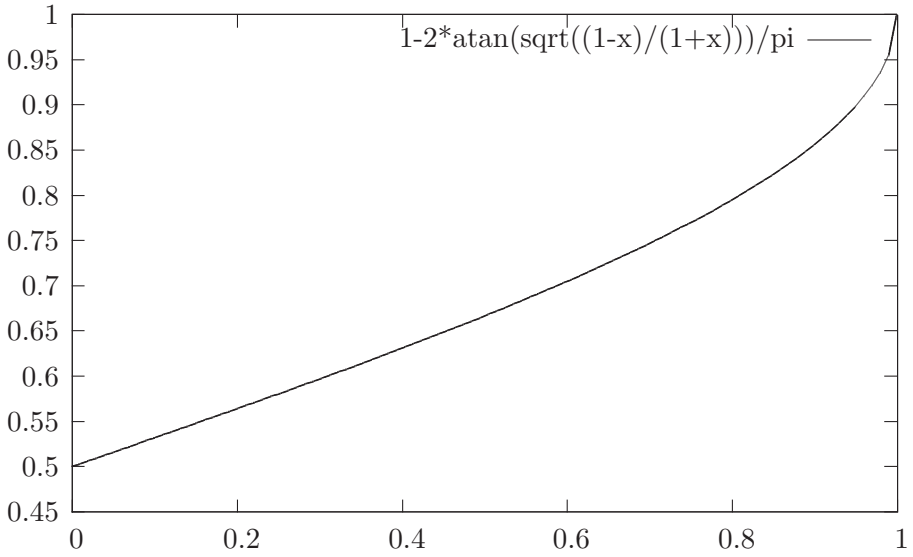
4.1.4 Numerical analysis

In this section, a small computational check of the previous result will be presented. The goal of the computation is to count the number of common minima of the discrete perceptron associated to two correlated patterns. We will be briefly discuss the Montecarlo algorithm that has been used.

The most difficult part is to generate input sets whose patterns are pairwise correlated. In order to achieve a meaningful result, the patterns must contain a big number N of elements in order to simulate the thermodynamical limit. This makes it difficult to generate random patterns, say ξ_α and $\bar{\xi}_{\bar{\alpha}}$ with a fixed correlation. This problem can be overcome by observing that, in the thermodynamical limit

$$\langle \xi_\alpha^i \; \bar{\xi}_{\bar{\alpha}}^j \rangle_q = q \; \delta_{ij}$$

Figure 4.1: The plot shows $[2 * C(q)]^{1/p}$ (see (4.1.3)). $C(q)$ is the shared fraction of solutions C as a function of q, in the $p/N \to 0$ case.



This is shown in section 7.0.8. Consequently, we can randomly generate $\bar{\xi}_{\bar{\alpha}}$ from another random ξ_α by flipping every spin in ξ_α with probability $(1 + q)/2$. This procedure allows creating the two correlated input sets but its drawback is that it introduces an error on the correlations.

Given this premise, the Montecarlo algorithm itself is simple. One should simply create random configurations W and find out the fraction of those which have positive overlaps with all patterns in both ξ and $\bar{\xi}$. The result is shown in figure 4.2

The algorithm is very slow and the computation time increases roughly exponentially in p, as it should. For this reason, several attempts at probing quantities for a greater p/N ratio have failed. It could be interesting to use actual learning algorithms to tests theoretical results; however, it was not possible within the scope of this thesis.

4.2 Derivation of the Gaussian distribution for synaptic weights

In this section it will be shown how the "scale invariance" of the arguments of the theta functions can be used to introduce a Gaussian integration in place of the sum over configurations.

The goal is to rewrite

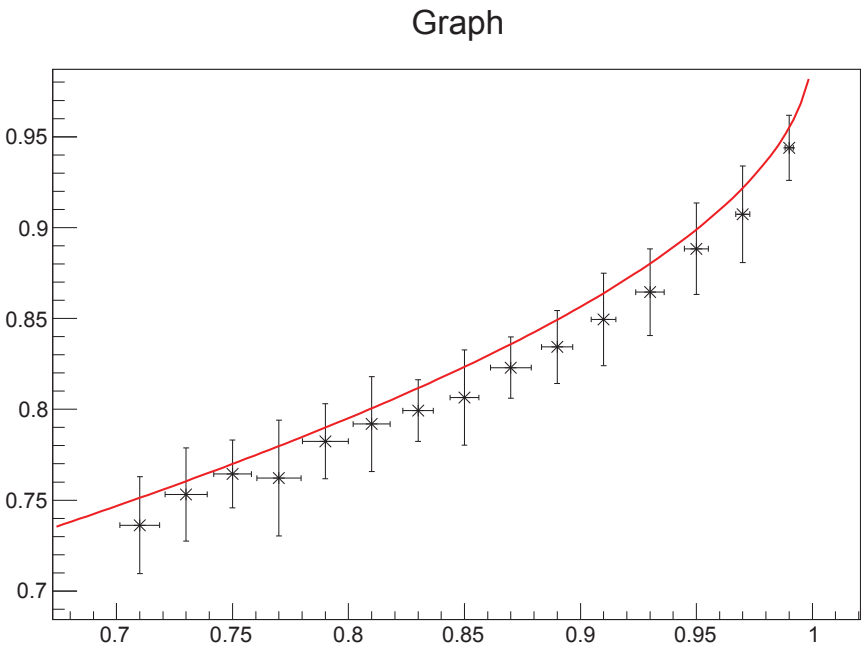
$$\frac{1}{2^N} \sum_W \prod_{\mu=1}^{p} \Theta(W^\mu)$$

in a more convenient way.

The first step is to find the joint distribution for

$$W_\mu = W \cdot \xi_\mu$$

Figure 4.2: The plot shows $[2 * C(q)]^{1/p}$. The theoretical curve appearing in figure 4.1 is compared with the numerical result. The result was obtained with $N = 500$ and $p = 10$. 20 correlated input pairs for each value of q have been generated. 1000000 configurations W have been tested in order to find solutions. The error bars show the standard deviations.



It is

$$\begin{aligned}
P(\{W_\mu\}) &= \frac{1}{2^N} \sum_{\{W^i=\pm 1\}} \prod_{\mu=1}^{p} \delta\left(W_\mu - W\cdot\xi_\mu\right) \\
&= \frac{1}{(2\pi)^p 2^N} \sum_{\{W^i=\pm 1\}} \int \prod_{\mu=1}^{p} dx_\mu \; e^{-ix_\mu(W_\mu - W\cdot\xi_\mu)} \\
&= \frac{1}{(2\pi)^p} \int \prod_{\mu=1}^{p} dx_\mu \; e^{-ix_\mu W_\mu} \prod_{i=1}^{N} \cos\left(\sum_\nu x_\nu \xi_\nu^i\right) \\
&= \frac{1}{(2\pi)^p} \int \prod_{\mu=1}^{p} dx_\mu \; \exp\left\{-ix_\mu W_\mu + \sum_i \ln\cos\left(\sum_{\nu=1}^{p} x_\nu \xi_\nu^i\right)\right\}
\end{aligned}$$

We can proceed further by observing that the variables W_μ only appear in theta functions. Since

$$\Theta(W_\mu) = \Theta(xW_\mu)$$

$\forall x > 0$, then it is possible to choose a prefactor $x = f(N)$. Let us now assume that $p = O(N)$ or smaller. Then

$$\left|\frac{1}{f(N)} \sum_{\nu=1}^{p} x_\nu \xi_\nu^i\right| \leq C(\{x_\nu\}) \frac{N}{f(N)}$$

$f(N)$ can be chosen arbitrarily diverging. Even $f(N) = N!$ or $f(N) = N^N$. Therefore, the sum above can always be set to vanish in the thermodynamic limit.

Hence, by substituting $\xi_\mu \mapsto \xi_\mu / f(N)$, we can proceed in the following way:

$$\begin{aligned}
P(\{W_\mu/N\}) &= \frac{1}{(2\pi)^p} \int \prod_{\mu=1}^{p} dx_\mu \; \exp\left\{-ix_\mu W_\mu + \sum_i \ln\cos\left(\sum_{\nu=1}^{p} x_\nu \xi_\nu^i / f(N)\right)\right\} \\
&\approx \frac{1}{(2\pi)^p} \int \prod_{\mu=1}^{p} dx_\mu \; \exp\left\{-ix_\mu W_\mu - \frac{1}{2} \sum_{\nu,\mu=1}^{p} x_\mu x_\nu \frac{\xi_\nu \cdot \xi_\mu}{f(N)^2}\right\}
\end{aligned}$$

After an integration

$$P(\{W_\mu/f(N)\}) \approx \sqrt{\frac{f(N)^p}{N^{p/2}(2\pi)^p}} \frac{1}{\sqrt{\det \Xi}} \exp\left\{-\frac{f(N)^2}{2N} \sum_{\mu,\nu=1}^{p} W_\mu W_\nu \Xi^{-1}_{\mu\nu}\right\}$$

with

$$\Xi_{\mu\nu} = \xi_\mu \cdot \xi_\nu / N$$

The variables can be rescaled as

$$W_\mu \mapsto \frac{\sqrt{N}}{f(N)} W_\mu$$

As a result, in the thermodynamical limit, the following substitution is possible

$$\frac{1}{2^N} \sum_W \prod_{\mu=1}^{p} \Theta(W^\mu) \mapsto \frac{1}{(2\pi)^p} \frac{1}{\sqrt{\det \Xi}} \int \prod_{\mu=1}^{p} \left[dW_\mu \; \Theta(W_\mu)\right] \exp\left(-\tfrac{1}{2} W^T \Xi^{-1} W\right) \quad (4.2.1)$$

This holds if $p/N \to 0$ as well, as a special case.

Chapter 5

General computation of $F(D|d)$: inputs of extensive size

5.1 Replica approach for extensive inputs

5.1.1 Introduction and structure of the forthcoming calculations

In order to compute the quantity $F(D|d)$ in the case of extensive ($p = O(N)$) inputs, the replica approach is needed. Throughout the following sections, the overlap q will be used instead of the distance d. For the sake of simplicity, before proceeding to the most general and complicated case, the tools and techniques will be introduced for the special case $F(0|q)$, i.e. the zero temperature computation with no output errors.

The spherical and discrete cases will be treated simultaneously, as it is often done in literature. For this purpose, the discrete notation will be used, unless otherwise specified. To obtain the discrete case, just replace

$$\sum_{\{W_i=\pm 1\}} \mapsto \int d^N W \; \delta\left(\sum_i W_i^2 - N\right)$$

Aside for RSB issues, this replacement has only one formal consequence. It will turn out that, in both cases

$$F \sim \frac{1}{m} \ln \int dQ e^{mp[\ln A + f + \ldots]}$$

The function called A is the same in both models, while f is not. The two different computations for $F(q) = F(0|q)$ will be presented in the general framework and completed in the RS ansatz, which should yield the correct result at least in the spherical case. In the process, it will be illustrated how a set of Gaussian variables (section 5.1.3) is suitable for the computation of the function A. This choice allows for a straightforward generalization in the RSB case, as it is preliminarily shown in chapter 7.

After the RS case $D = 0$ is completed, the computations will be extended in the RS ansatz to the $D \neq 0$ case. Finally, the results will be presented.

5.1.2 No error $D = 0$ for extensive inputs with correlation q: the replica approach

As stated in the introduction, we will set $D = 0$ and compute the following quantity with the replica approach:

$$F(q) = \int P_q(\xi,\bar{\xi}) \ln \sum_W \prod_\mu \Theta\left(\frac{\xi^\mu \cdot W}{\sqrt{N}}\right) \Theta\left(\frac{\bar{\xi}^{\bar{\mu}} \cdot W}{\sqrt{N}}\right)$$

It is worth recalling that the replica method is a procedure based on the following identity

$$\lim_{m\to 0} \frac{1}{m} \ln \int P(\xi)\ Z^m = \lim_{m\to 0} \frac{1}{m} \ln \int P(\xi)\ (1 + m \ln Z) = \int P(\xi) \ln Z$$

amd thta the trick consists in calculating the formula for integer m and then taking the limit $m \to 0$ of its analytical continuation.

Thus

$$F(q) = \lim_{m\to 0} \frac{1}{m} \ln \int P_q(\xi,\bar{\xi}) \sum_{\{W_a\}} \prod_{a,b=1}^{m} \prod_\mu \Theta\left(\frac{W^a \cdot \xi^\mu}{\sqrt{N}}\right) \Theta\left(\frac{W^b \cdot \bar{\xi}^{\bar{\mu}}}{\sqrt{N}}\right)$$

Now the following identity can be inserted in the integral

$$1 = \int dQ_{ab}\ \delta\left(Q_{ab} - \frac{W^a \cdot W^b}{N}\right)$$

so that

$$F(q) = \lim_{m\to 0} \frac{1}{m} \ln \int P_q(\xi,\bar{\xi}) \sum_{\{W_a\}} \int \prod_{a<b} dQ_{ab}\ \delta\left(Q_{ab} - \frac{W^a \cdot W^b}{N}\right) \prod_{a,b=1}^{m} \prod_\mu \Theta\left(\frac{W^a \cdot \xi^\mu}{\sqrt{N}}\right) \Theta\left(\frac{W^b \cdot \bar{\xi}^{\bar{\mu}}}{\sqrt{N}}\right)$$

The delta can be written as an exponential, using the auxiliary variable $N\hat{Q}_{ab}$:

$$F(q) = \lim_{m\to 0} \frac{1}{m} \ln N^{-m(m-1)} \int dQ_{ab}\ d\hat{Q}_{ab} \sum_{\{W_a\}} e^{-iN \sum_{a<b} \hat{Q}_{ab}\left(Q_{ab} - \frac{W^a \cdot W^b}{N}\right)} \int P(\xi,\bar{\xi}) \prod_{a,b=1}^{m} \prod_\mu \Theta\left(\frac{W^a \cdot \xi^\mu}{\sqrt{N}}\right) \Theta\left(\frac{W^b \cdot \bar{\xi}^{\bar{\mu}}}{\sqrt{N}}\right) \quad (5.1.1)$$

5.1.3 Introduction of Gaussian variables for the derivation of the saddle point equations

Let us focus on the quantity

$$\mathcal{A}(Q)_q = \int P(\xi,\bar{\xi}) \prod_{a,b=1}^{m} \prod_\mu \Theta\left(\frac{W^a \cdot \xi^\mu}{\sqrt{N}}\right) \Theta\left(\frac{W^b \cdot \bar{\xi}^{\bar{\mu}}}{\sqrt{N}}\right) \quad (5.1.2)$$

which appears in (5.1.1).

As explained in the introductory chapter, the distribution of correlated sets of patterns will be defined as the factorized distributions of p independent pairs with a certain correlation q.

$$P(\xi,\bar{\xi}) = \prod_\mu P_q(\xi^\mu,\bar{\xi}^{\bar{\mu}})$$

With this choice, two claims can be made

- Function $\mathcal{A}_q$ as given by (5.1.2) can be expressed as an integral over Gaussian variables in place of the sum over $\{W_a\}$. The expression is given by (5.1.3)
- Function $\mathcal{A}_q$ depends on $\{W_a\}$ only through $W_a \cdot W_b/N = Q_{ab}$ which appears in the quadratic form

The actual proof is a bit technical and can be found in 7.0.9 and 7.0.8. Nonetheless, this can be understood as follows. The marginal distribution of the variables $w_a = \xi \cdot W^a/\sqrt{N}$ and $\bar{w}_a = \bar{\xi} \cdot W^a/\sqrt{N}$ are distributed as Gaussian according to the central limit theorem. It is easy to prove that their correlation is

$$\langle w_a w_b \rangle = \frac{W_a \cdot W_b}{N} = Q_{ab}$$

Therefore, it could be inferred that the joint distribution is a multivariate Gaussian whose nondiagonal terms are obtained by fixing the correlations between pairs of w_as as indicated above.

The result is

$$\mathcal{A}(Q) = \left\{ \frac{1}{(2\pi)^m \sqrt{\det G}} \int \prod_c^m dw_c \, d\bar{w}_c \right. \\ \left. \exp\left(-\frac{1}{2} \sum_{u,v=\pm 1} \sum_{a,b=1}^m w_a^u [G^{-1}]_{ab}^{uv} w_b^v \right) \prod_{a,b} \Theta(w_a)\Theta(\bar{w}_b) \right\}^p \quad (5.1.3)$$

with

$$\bar{w}_a = w_a^{(-1)} \qquad w_a = w_a^{(1)}$$

and (Q is a $m \times m$ matrix with $Q_{aa} = 1$)

$$G = \begin{bmatrix} Q & qQ \\ qQ & Q \end{bmatrix}$$

$$G^{-1} = \frac{1}{(1-q^2)} \begin{bmatrix} Q^{-1} & -qQ^{-1} \\ -qQ^{-1} & Q^{-1} \end{bmatrix}$$

Equivalently

$$\mathcal{A}(Q) = A^p(Q) = \left\{ \frac{1}{(2\pi)^m \sqrt{\det G}} \int \prod_{c=1}^{2m} dw_c \, \exp\left(-\frac{1}{2} \sum_{a,b=1}^{2m} w_a G_{ab}^{-1} w_b \right) \prod_{a=1}^{2m} \Theta(w_a) \right\}^p$$

Thus, what is left is

$$F(q) = \lim_{m \to 0} \frac{1}{m} \ln N^{-m(m-1)} \int \prod_{a<b} dQ_{ab} \, d\hat{Q}_{ab} \, e^{-iN \sum_{a<b} Q_{ab} \hat{Q}_{ab}} \sum_{\{W^a\}} e^{i \sum_{a<b} \hat{Q}_{ab} W^a \cdot W^b} A^p(Q)$$

Now, let us define:

$$\begin{aligned} \sum_{\{W^a\}} e^{i \sum_{a<b} \hat{Q}_{ab} W^a \cdot W^b} &= \prod_j \sum_{\{W_j^a = \pm 1\}} e^{i \sum_{a<b} \hat{Q}_{ab} W_j^a W_j^b} \\ &= \left\{ \sum_{\{W^a = \pm 1\}} e^{i \sum_{a<b} (\hat{Q}_{ab}) W^a W^b} \right\}^N \\ &=: e^{N f(\hat{Q})} \end{aligned}$$

where $f(\hat{Q})$ is a sort of free energy of a fully connected Ising model with imaginary temperature, couplings $\hat{Q}_{ab}$ and m nodes.

Then

$$F(q) = \lim_{m\to 0} \frac{1}{m} \ln \int \prod_{a<b} dQ_{ab}\, d\hat{Q}_{ab}\, e^{-i\sum_{a<b} Q_{ab}\hat{Q}_{ab}} e^{Nf(\hat{Q})} A^p(Q)$$

or

$$F(q) = \lim_{m\to 0} \frac{1}{m} \ln \int \prod_{a<b} dQ_{ab}\, d\hat{Q}_{ab}\, \exp\left\{ N\left[-i\sum_{a<b} Q_{ab}\hat{Q}_{ab} + f(\hat{Q}) + \alpha \ln A(Q) \right] \right\} \tag{5.1.4}$$

The saddle point equations are obtained by the extremization of the exponent and are therefore

$$iQ_{ab} = \frac{d}{d\hat{Q}_{ab}} f(\hat{Q})$$

$$i\hat{Q}_{ab} = \alpha \frac{d}{dQ_{ab}} \ln A(Q)$$

with

$$f(\hat{Q}) = \ln \sum_{\{W^a=\pm 1\}} e^{i\sum_{a<b}(\hat{Q}_{ab})W^a W^b}$$

$$A(Q) = \frac{1}{(2\pi)^m \sqrt{\det G}} \int \prod_{c=1}^{2m} dw_c\, \exp\left(-\frac{1}{2} \sum_{a,b=1}^{2m} w_a G^{-1}_{ab} w_b \right) \prod_{a=1}^{2m} \Theta(w_a)$$

In order to solve them, the explicit expressions for A and f (the latter of which depends on the discrete/spherical model are needed. For this purpose, the RS ansatz should be introduced.

5.1.4 Replica symmetric equations

Let us suppose that $Q_{ab} = Q(1-\delta_{ab}) + \delta_{ab}$. Equivalently

$$Q = Q1_m + (1-Q)\mathbb{I}_m$$

with $[1_m]_{ab} = 1$. The matrices $\mathbb{I}_m$ and 1_m form a closed algebra since $[1_m]^2 = m1_m$. The same assumptions hold for $\hat{Q}$.

In this approximation, $f(\hat{Q})$ is a exactly proportional to the free energy of a fully connected Ising model with imaginary temperature, couplings $\hat{Q}_{ab}$ and m nodes. Hence

$$\frac{d}{d\hat{Q}} f(\hat{Q})$$

is the Ising spin-spin correlation.

As for A, it becomes:

$$A(Q) = \frac{1}{(2\pi)^m \sqrt{\det G}} \int \prod_{c=1}^{2m} dw_c\, \exp\left(-\frac{1}{2} \sum_{a,b=1}^{2m} w_a G^{-1}_{ab} w_b \right) \prod_{a=1}^{2m} \Theta(w_a)$$

The matrix G^{-1} is

$$G^{-1} = \frac{1}{1-q^2} \begin{bmatrix} [Q1_m + (1-Q)\mathbb{I}_m]^{-1} & -q[Q1_m + (1-Q)\mathbb{I}_m]^{-1} \\ -q[Q1_m + (1-Q)\mathbb{I}_m]^{-1} & [Q1_m + (1-Q)\mathbb{I}_m]^{-1} \end{bmatrix}$$

It can be shown that

$$[Q1_m + (1-Q)\mathbb{I}_m]^{-1} = -\frac{Q}{(1-Q)(mq-Q-1)}1_m + \frac{1}{1-Q}\mathbb{I}_m$$

Symbolically

$$G^{-1} = \begin{bmatrix} A\mathbb{I}_m - B\ 1_m & -q[A\mathbb{I}_m - B\ 1_m] \\ -q[A\mathbb{I}_m - B\ 1_m] & A\mathbb{I}_m - B\ 1_m \end{bmatrix}$$

with

$$A = \frac{1}{1-Q} > 0 \ \ \forall 0 < Q < 1$$

$$B = \frac{Q}{(1-Q)(1-Q+mQ)} > 0 \ \ \forall 0 < Q < 1$$

Further details about G can be found in the appendix 7.0.7.

The saddle point equations, with the RS ansatz, become

$$i\frac{1}{2}m(m-1)Q = \frac{d}{d\hat{Q}}f(\hat{Q}) \tag{5.1.5}$$

$$i\frac{1}{2}m(m-1)\hat{Q} = \alpha\frac{d}{dQ}\ln A(Q) \tag{5.1.6}$$

Before proceeding any further, it must be pointed out that, if Q is real, then $\ln A_q(Q)$ and $f(\hat{Q})$ must be real. Thus, according to (5.1.6), $\hat{Q}$ must be imaginary.

Furthermore, we can anticipate that

$$\frac{d}{dQ}\ln A_q < 0 \ \ \forall Q \in (0,1)$$

Thus, when the limit $m \to 0$ is taken, (5.1.6) implies that

$$i\hat{Q} > 0$$

In the following sections the detailed computations for $\ln A$ (see 5.1.7), for $f_{spherical}$ (see 5.1.5) and $f_{discrete}$ (see 5.1.6) are shown. The results are reported in section 5.2.1.

5.1.5 Spherical perceptron: computing $f(\hat{Q})$

The spherical computations are almost identical to the discrete ones. One wust has to make the replacement

$$\sum_W \mapsto \frac{1}{V_0}\int d\vec{W}\ \delta\left(W\cdot W - N\right)$$

V_0 is the area of a sphere S^N. While $\ln A_q$ is unchanged, f is not. In the RS ansatz, it becomes

$$\begin{aligned} e^{Nf} &= \frac{1}{V_0^m}\int\prod_{a=1}^m d\vec{W}_a\ \delta(W_a\cdot W_a - N)\ \exp\left(i\hat{Q}\sum_{a<b}W_a\cdot W_b\right) \\ &= \frac{1}{V_0^m}\int\prod_{a=1}^m \frac{dJ_a}{2\pi}\ d\vec{W}_a\ \exp\left(i\sum_a J_a(W_a\cdot W_a - N) + i\hat{Q}\sum_{a<b}W_a\cdot W_b\right) \\ &= \frac{1}{V_0^m}\int\prod_{a=1}^m \frac{dJ_a}{2\pi}\ \exp\left(-iN\sum_a J_a\right)\left\{\int\prod_{a=1}^m dW_a\ \exp\left(i\sum_a J_a W_a^2 + i\hat{Q}\sum_{a<b}W_aW_b\right)\right\}^N \end{aligned}$$

The variables J_a appear in the integral and can be minimized upon according to the saddle point method. This means that the RS ansatz can be applied to them as well:

$$J_a = J \tag{5.1.7}$$

Then

$$e^{Nf} \sim \frac{1}{V_0^m} \frac{e^{-iNmJ}}{(2\pi)^{m/2}} \left\{ \int \prod_{a=1}^{m} dW_a \, \exp\left(iJ \sum_a W_a^2 + i\hat{Q} \sum_{a<b} W_a W_b \right) \right\}^N$$

The quadratic form is of the usual kind $M_{ab} = (iJ - i\hat{Q}/2)\mathbb{I}_m + i\hat{Q}/2 \; 1_m$. Its eigenvalues are $(iJ - i\hat{Q}/2)$ $(\times m-1)$ and $(iJ + (m-1)i\hat{Q}/2)$ $(\times 1)$. Hence

$$e^{Nf} \sim \frac{e^{-iNmJ}}{V_0^m (2\pi)^{m/2}} \; (2\pi)^{mN/2} \; \left[(iJ + (m-1)i\hat{Q}/2)(iJ - i\hat{Q}/2)^{m-1} \right]^{-N/2}$$

Since

$$V_0 = \frac{(2\pi)^{N/2}}{\Gamma(N/2)} N^{(N-1)/2} \sim \frac{(2\pi)^{(N-1)/2}}{N/2-1} \left[\frac{e}{N/2-1} \right]^{N/2-1} N^{(N-1)/2}$$

then

$$V_0^{-m} = (2\pi)^{-m(N-1)/2} \; e^{o(N)}$$

Hence

$$f = -imJ - (1/2) \ln \left(\frac{iJ + (m-1)i\hat{Q}/2}{(iJ - i\hat{Q}/2)^{1-m}} \right)$$

For small m,

$$f = -m \left\{ iJ + \frac{1}{4} \frac{i\hat{Q}}{iJ - i\hat{Q}/2} \right\}$$

The stability with respect to J (see (5.1.5)) implies

$$\frac{d}{dJ} f = 0$$

and is satisfied by two points, one of which is a global minimum

$$iJ = \frac{1}{2}(\sqrt{i\hat{Q}} + i\hat{Q})$$

This solution can be inserted back into f_{sph}:

$$f_{sph}(\hat{Q}) = -m \left\{ \sqrt{i\hat{Q}} + i\hat{Q}/2 \right\} \tag{5.1.8}$$

and

$$\frac{d}{d\hat{Q}} f_{sph}(\hat{Q}) = -im \frac{1}{2} \left(1 + 1/\sqrt{i\hat{Q}} \right) \tag{5.1.9}$$

5.1.6 Discrete perceptron: computing $f(\hat{Q})$

In the following passages it will be assumed that $i\hat{Q} > 0$.

$$\begin{aligned} e^f = \sum_{\{W_a\}} e^{i\hat{Q}\sum_{a<b} W_a W_b} &= e^{-im\hat{Q}/2} \sum_{\{W_a\}} \exp\left(i\frac{\hat{Q}}{2} \sum_{ab} W_a W_b \right) \\ &= e^{-im\hat{Q}/2} \sum_{\{W_a\}} \exp\left(\frac{i\hat{Q}}{2} \left(\sum_a W_a \right)^2 \right) \\ &= e^{-im\hat{Q}/2} \sum_{\{W_a\}} \frac{1}{\sqrt{2\pi i\hat{Q}}} \int dx \ \exp\left(-\frac{1}{2i\hat{Q}} x^2 + x \sum_a W_a \right) \\ &= e^{-im\hat{Q}/2} \frac{1}{\sqrt{2\pi i\hat{Q}}} \int dx \ \exp\left(-\frac{1}{2i\hat{Q}} x^2 \right) \left[\sum_{W=\pm 1} e^{xW} \right]^m \\ &= 2^m e^{-im\hat{Q}/2} \frac{1}{\sqrt{2\pi i\hat{Q}}} \int dx \ \exp\left(-\frac{1}{2i\hat{Q}} x^2 \right) \cosh^m(x) \end{aligned} \tag{5.1.10}$$

So that, for small m,

$$\begin{aligned} f &= \ln \left\{ 2^m e^{-im\hat{Q}/2} \frac{1}{\sqrt{2\pi i\hat{Q}}} \int dx \ \exp\left(-\frac{1}{2i\hat{Q}} x^2 + m \ln\cosh(x) \right) \right\} \\ &= m \left\{ \ln 2 - i\hat{Q}/2 + \frac{1}{\sqrt{2\pi}} \int dx \ \exp\left(-\frac{1}{2} x^2 \right) \ln\cosh\left(x\sqrt{i\hat{Q}} \right) \right\} + o(m) \end{aligned}$$

The derivative of f with respect to $\hat{Q}$ is

$$\frac{d}{d\hat{Q}} f = im \left\{ -1/2 + \frac{1}{\sqrt{8\pi i\hat{Q}}} \int dx \ x \ \exp\left(-\frac{1}{2} x^2 \right) \tanh\left(x\sqrt{i\hat{Q}} \right) \right\} + o(m) \tag{5.1.11}$$

5.1.7 Discrete and spherical perceptron: computing $\ln A_q$

Given $Q_+ = \{x_i > 0, \bar{x}_i > 0\}$ and

$$A = \frac{1}{\sqrt{1-q^2}} \frac{1}{1-Q} > 0 \ \ \forall Q < 1$$

$$B = \frac{1}{\sqrt{1-q^2}} \frac{Q}{(1-Q)(1-Q+mQ)} > 0 \ \ \forall Q \in (0,1)$$

so that

$$G^{-1} = \frac{1}{\sqrt{1-q^2}} \begin{bmatrix} A\mathbb{I}_m - B\,1_m & -q(A\mathbb{I}_m - B\,1_m) \\ -q(A\mathbb{I}_m - B\,1_m) & A\mathbb{I}_m - B\,1_m \end{bmatrix}$$

then

$$\frac{1}{(2\pi)^m \sqrt{\det G}} \int_{Q_+} \prod_i dx_i \, d\bar{x}_i \ \exp\left(-\frac{A}{2} \sum_i (x_i^2 + \bar{x}_i^2 - 2q x_i \bar{x}_i) + \frac{B}{2} \sum_{ij} (x_i x_j + \bar{x}_i \bar{x}_j - 2q x_i \bar{x}_j) \right)$$

$$= \frac{1}{(2\pi)^m\sqrt{\det G}} \int_{Q_+} \prod_i dx_i\, d\bar{x}_i\, \exp\Bigg\{ -\frac{A}{2}\sum_i (x_i^2+\bar{x}_i^2-2qx_i\bar{x}_i) + \frac{1}{2}B(1+q)\left[\left(\sum_i x_i\right)^2+\left(\sum_i \bar{x}_i\right)^2\right] - \frac{qB}{2}\left(\sum_i (x_i+\bar{x}_i)\right)^2 \Bigg\}$$

The following identity can be used three times

$$e^{ax^2/2} = \sqrt{1/2a\pi}\int dy\, \exp(-y^2/2a+xy)$$

to replace the squares of the sums. Then

$$\frac{1}{(2\pi)^m\sqrt{\det G}}\sqrt{\frac{1}{8\pi^3B^3q(1+q)^2}}\int_{Q_+} dy\, d\bar{y}\, du\, \prod_i dx_i\, d\bar{x}_i\, \exp\Bigg\{-\frac{A}{2}\sum_i (x_i^2+\bar{x}_i^2-2qx_i\bar{x}_i) - \frac{1}{2B(1+q)}\left[y^2+\bar{y}^2\right] - \frac{1}{2qB}u^2 + iu\left(\sum_i (x_i+\bar{x}_i)\right) + y\sum_i x_i + \bar{y}\sum_i \bar{x}_i\Bigg\}$$

In order to decouple the m integrals in x_i, it is worth applying the following change of variables [1]

$$u \mapsto u + \frac{iq}{1+q}(y+\bar{y})$$

which, after some algebra, allows to rewrite the previous integral as

$$\frac{1}{(2\pi)^m\sqrt{\det G}}\sqrt{\frac{1}{8\pi^3B^3q(1+q)^2}}\int_{Q_+} dy\, d\bar{y}\, du\, \prod_i dx_i\, d\bar{x}_i$$
$$\exp\Bigg\{\sum_i\Bigg[-\frac{A}{2}\left(x_i-\frac{1}{(1+q)A}y-\frac{i}{(1-q)A}u\right)^2-\frac{A}{2}\left(\bar{x}_i-\frac{1}{(1+q)A}\bar{y}-\frac{i}{(1-q)A}u\right)^2 + Aq\left(x_i-\frac{1}{(1+q)A}y-\frac{i}{(1-q)A}u\right)\left(\bar{x}_i-\frac{1}{(1+q)A}\bar{y}-\frac{i}{(1-q)A}u\right)\Bigg] - \frac{1}{2}\left(\frac{1}{qB}+\frac{2m}{A(1-q)}\right)u^2 - \frac{1}{2}\left(\frac{1}{B(1+q)}-\frac{m}{A(1+q)^2}-\frac{q}{(1+q)^2B}\right)(y^2+\bar{y}^2) + \left(-\frac{i}{(1+q)B}+\frac{im}{A(1+q)}\right)u(y+\bar{y}) + \left(-\frac{mq}{A(1+q)^2}+\frac{q}{B(1+q)^2}\right)y\bar{y}\Bigg\}$$

Let us define a new function, which is a sort of two dimensional error function

$$E_q(y,\bar{y}) = \frac{1}{2\pi}\int_y^{+\infty} dx \int_{\bar{y}}^{+\infty} d\bar{x}\, \exp\left[-\frac{1}{2\sqrt{1-q^2}}\begin{bmatrix}x & \bar{x}\end{bmatrix}\begin{bmatrix}1 & -q\\ -q & 1\end{bmatrix}\begin{bmatrix}x\\ \bar{x}\end{bmatrix}\right] \quad (5.1.12)$$

[1] See 7.0.4 for the explaination of the imaginary translation

Therefore, the integral becomes

$$\frac{1}{\sqrt{\det G}}\sqrt{\frac{1}{8\pi^3 B^3 q(1+q)^2 A^{2m}}}\int dy\, d\bar{y}\, du$$
$$\left[E_q\left(\frac{(1-q^2)^{1/4}}{(1+q)\sqrt{A}}y+\frac{i(1-q^2)^{1/4}}{(1-q)\sqrt{A}}u,\frac{(1-q^2)^{1/4}}{(1+q)\sqrt{A}}\bar{y}+\frac{i(1-q^2)^{1/4}}{(1-q)\sqrt{A}}u\right)\right]^m$$
$$\exp\left\{-\frac{1}{2}\left(\frac{1}{qB}+\frac{2m}{A(1-q)}\right)u^2-\frac{1}{2}\left(\frac{1}{B(1+q)}-\frac{m}{A(1+q)^2}-\frac{q}{(1+q)^2B}\right)(y^2+\bar{y}^2)\right.$$
$$\left.+\left(-\frac{i}{(1+q)B}+\frac{im}{A(1+q)}\right)u(y+\bar{y})+\left(-\frac{mq}{A(1+q)^2}+\frac{q}{B(1+q)^2}\right)y\bar{y}\right\}$$

The y and $\bar{y}$ can be rescaled

$$y\to y-\frac{i(1+q)}{1-q}u$$

so that that the u can be integrated. The exponential can be written as

$$\exp\left\{-\tfrac{1}{2}M_{uu}u^2+iM_{yu}u(y+\bar{y})-\tfrac{1}{2}M_{yy}(y^2+\bar{y}^2)-M_{y\bar{y}}y\bar{y}\right\}$$

with

$$M_{uu}=\left(\frac{1}{qB}+\frac{2m}{A(1-q)}\right)-4\left(-\frac{1}{(1+q)B}+\frac{m}{A(1+q)}\right)\frac{(1+q)}{1-q}$$
$$-\left(\frac{2}{B(1+q)}-\frac{2m}{A(1+q)^2}-\frac{2q}{(1+q)^2B}+\frac{2mq}{A(1+q)^2}-\frac{2q}{B(1+q)^2}\right)\frac{(1+q)^2}{(1-q)^2}$$
$$=\frac{1+q}{q(1-q)B} \tag{5.1.13}$$

$$M_{uy}=\left(-\frac{1}{(1+q)B}+\frac{m}{A(1+q)}\right)+\frac{(1+q)}{1-q}\left[-\left(-\frac{mq}{A(1+q)^2}+\frac{q}{B(1+q)^2}\right)\right.$$
$$\left.+\left(\frac{1}{B(1+q)}-\frac{m}{A(1+q)^2}-\frac{q}{(1+q)^2B}\right)\right]=0$$

$$M_{yy}=\frac{1}{(1+q)^2}\left(\frac{1}{B}-\frac{m}{A}\right)$$

$$M_{y\bar{y}}=-\frac{1}{(1+q)^2}\left(-\frac{mq}{A}+\frac{q}{B}\right)=-qM_{yy}$$

It is worth noticing that $M_{yy}>0$ and $M_{uu}>0$ for $0<Q<1$, $\forall m\geq 0$. Hence, u can be integrated away. For simplicity's sake, a change of variable is convenient

$$y\to\frac{(1+q)}{(1-q^2)^{1/4}}\left(\frac{1}{B}-\frac{m}{A}\right)^{-1/2}y$$

By recalling that

$$\det G=(mQ+(1-Q))^2(1-Q)^{2m-2}(1-q^2)^m$$

and observing that all prefactors cancel each other

$$\left[\frac{(1+q)}{(1-q^2)^{1/4}}\left(\frac{1}{B}-\frac{m}{A}\right)^{-1/2}\right]^2 \frac{1}{\sqrt{\det G}}\sqrt{\frac{2\pi}{M_{uu}}}\sqrt{\frac{1}{8\pi^3(B^3)q(1+q)^2A^{2m}}} = \frac{1}{2\pi}$$

we are left with

$$\boxed{A_q(Q) = \frac{1}{2\pi}\int dy\, d\bar{y}\ \left[E_q\left(\sqrt{\frac{Q}{1-Q}}\,y, \sqrt{\frac{Q}{1-Q}}\,\bar{y}\right)\right]^m \exp(-Y^t G_q\, Y/2)} \quad (5.1.14)$$

The notation is

$$Y = \begin{bmatrix} y \\ \bar{y} \end{bmatrix}$$

$$\mathcal{G}_q = \frac{1}{\sqrt{1-q^2}}\begin{bmatrix} 1 & -q \\ -q & 1 \end{bmatrix}$$

The matrix $\mathcal{G}_q$ is the same as the one appearing in $E_q(x, \bar{x})$. Equation (5.1.14) is indeed quite similar to (5.2.7).

Finally, the small-m expansion is needed

$$\ln A_q(Q) = m\frac{1}{2\pi}\int dy\, d\bar{y}\ \exp(-Y^t G_q\, Y/2)\ \ln E_q\left(\sqrt{\frac{Q}{1-Q}}\,y, \sqrt{\frac{Q}{1-Q}}\,\bar{y}\right) + o(m)$$

5.2 Solving of the saddle point equations

5.2.1 RS saddle point equations for $D = 0$

The final step is to solve the RS saddle point equations. **The first SP equation** is, for both the discrete and the spherical model:

$$-\frac{i}{2}\hat{Q} = \frac{1}{2\pi}\frac{d}{dQ}\int dy\, d\bar{y}\ \exp(-Y^t G_q\, Y/2)\ \ln E_q\left(\sqrt{\frac{Q}{1-Q}}\,y, \sqrt{\frac{Q}{1-Q}}\,\bar{y}\right) \quad (5.2.1)$$

with

$$E_q(y,\bar{y}) = \frac{1}{2\pi}\int_y^{+\infty} dx \int_{\bar{y}}^{+\infty} d\bar{x}\ \exp\left[-\frac{1}{2\sqrt{1-q^2}}\begin{bmatrix} x & \bar{x} \end{bmatrix}\begin{bmatrix} 1 & -q \\ -q & 1 \end{bmatrix}\begin{bmatrix} x \\ \bar{x} \end{bmatrix}\right] \quad (5.2.2)$$

and

$$G_q = \frac{1}{\sqrt{1-q^2}}\begin{bmatrix} 1 & -q \\ -q & 1 \end{bmatrix}$$

The **second SP equation** is: for the discrete model

$$-\frac{1}{2}Q = -1/2 + \frac{1}{\sqrt{8\pi i\hat{Q}}}\int dx\, x\, \exp\left(-\frac{1}{2}x^2\right)\tanh\left(x\sqrt{i\hat{Q}}\right) \quad (5.2.3)$$

For the spherical model

$$-\frac{1}{2}Q = -\frac{1}{2}\left(1 + 1/\sqrt{i\hat{Q}}\right) \quad (5.2.4)$$

The most general equations should be solved numerically, and would yield

$$Q_{sol} = Q(q,\alpha) \text{ and } \hat{Q}_{sol} = \hat{Q}(Q(q,\alpha))$$

The result can be inserted back in (5.1.4) to give

$$\bar{F}(q,\alpha) = \frac{-imQ_{sol}\hat{Q}_{sol}/2 + f(\hat{Q}_{sol}) + \alpha \ln A(Q_{sol})}{m} \tag{5.2.5}$$

For computational purposes, it is convenient to rewrite $\ln A_q$ as

$$\begin{aligned}\ln A_q(Q) = m\frac{1}{2\pi}\int dS\ dD\ \exp\left(-\frac{1}{2}\sqrt{\frac{1-q}{1+q}}S^2 - \frac{1}{2}\sqrt{\frac{1+q}{1-q}}D^2\right) \\ \ln\frac{1}{\sqrt{2\pi}}\left[\frac{1-q}{1+q}\right]^{1/4}\int_{\sqrt{Q/(1-Q)}\,S}^{\infty} ds\ \exp\left(-\frac{1}{2}\sqrt{\frac{1-q}{1+q}}\,s^2\right) \\ \left\{\operatorname{erfc}\left(\left[\frac{1+q}{1-q}\right]^{1/4}\left[\sqrt{\frac{Q}{1-Q}}(D+S) - s\right]\right)\right. \\ \left. - \operatorname{erfc}\left(\left[\frac{1+q}{1-q}\right]^{1/4}\left[\sqrt{\frac{Q}{1-Q}}(D-S) + s\right]\right)\right\}\end{aligned} \tag{5.2.6}$$

These equations should be solved numerically. Unfortunately, this task has turned out to be harder than expected. Attempts with several simple techniques with both C++ and Mathematica have encountered different but equally fatal problems associated to the working precision in performing integrations with the available computation power. It is surely possible to find a solution with more sophisticated numerical tools. However, due to the difficulties encountered, it seemed more sensible to postpone such calculations for future works and focus on some analytically computable task.

For this reason, we have chosen to focus on the capacity $\alpha_c(D|d)$. Two main results will be presented in the following sections.

- A consistency check. It will be shown that the formalism under exam leads to classic know quantities. In particular the RS standard capacity $\alpha_c = 2$ (or $4/\pi$) will be recovered both by repeating the previous passages in the simplified case of a single input set and by taking the limit $q \to 1$. It is worth recalling that, in this limit, the two input sets become indistinguishable.
- The capacity $\alpha_c(D|d)$ will be computed in the RS framework, by taking the limit $Q \to 1$, which corresponds to assuming that all solutions become one (see the introduction of 5.2.2 for more details).

5.2.2 RS capacity

The perceptron capacity can be obtain from the Z of a simple learning problem of a set ξ of p random patterns of length N. The relevant quantity is

$$f = \frac{1}{N}\int P(\xi)\ln\sum_{\{W\}}\prod_{\alpha=1}^{p}\Theta(W\cdot\xi_\alpha)$$

It follows that f is the logarithm of the number of solutions divided by N, on average. The capacity is defined as the ratio $\alpha_c = p/N$ above which there are typically no solutions. The most intuitive way to obtain the capacity would to impose that $f \to -\infty$. However, there exists another possibility. As $p/N = \alpha$ approaches α_c, solutions become less and less and typically more similar to each other. The "similarity" is

represented by the overlap $W \cdot W'/N$. The overlap distribution $P(\tilde{Q})$ is an order parameter adopted in spin glass theories. Furthermore, in the replica formalism, $P(\tilde{Q})$ can be deduced from the distribution of the overlaps between replicated synaptic weights $Q_{ab} = W_a \cdot W_b/N$ (see 1.1.26). Hence, the capacity is reached when $Q_{ab} \to 1$. In the RS scheme, replicated weights W_a are assumed to have fixed mutual overlap Q. Hence, it is enough to take the limit $Q \to 1$. This is only correct if the RS ansatz is correct. In the case of the discrete perceptron the RS solution is unstable (negative entropy), and the actual RSB capacity $\alpha_c = 0.83$ is reached for a much lower value of $Q \approx 1/2$. In this section the unstable RS result $\alpha_c = 4/3$ for the discrete perceptron will be obtained as a consistency check.

The saddle point equations (5.1.5) and (5.1.6) clearly hold for the present purpose. It is enough to specify that the matrix G appearing in A (see (5.1.6)) is simply

$$G_{ab} = Q_{ab}$$

With the RS ansatz

$$Q_{ab} = (1-Q)\delta_{ab} + Q(1-\delta_{ab})$$

Hence

$$G = (1-Q)\mathbb{I}_m + Q\,1_m$$

and

$$G^{-1} = \frac{1}{1-Q}\mathbb{I}_m - \frac{Q}{(1-Q)(1-Q+mQ)}\,1_m =: A\mathbb{I}_m - B\,1_m$$
$$\det G = (1-Q)^{m-1}(1-Q+mQ)$$

A simpler expression for the quantity $\ln A$ is needed. Given $Q_+ = \{x_a > 0\}_{a=1,\dots,m}$ and observing that

$$A^m\, B\, \det G = \frac{1}{\frac{1}{B} - \frac{m}{A}}$$

then

$$\begin{aligned}
\ln A &= \frac{1}{\sqrt{(2\pi)^m \det G}} \int_{Q_+} \prod dx_a\, \exp\left(-\frac{A}{2}\sum_a x_a^2 + \frac{B}{2}\sum_{ab} x_a x_b\right) \\
&= \frac{1}{\sqrt{(2\pi)^m \det G}} \int_{Q_+} \prod dx_a\, \exp\left(-\frac{A}{2}\sum_a x_a^2 + \frac{B}{2}\left(\sum_a x_a\right)^2\right) \\
&= \frac{1}{\sqrt{(2\pi)^m \det G\, B}} \int_{Q_+} du \prod dx_a\, \exp\left(-\frac{A}{2}\sum_a x_a^2 - \frac{1}{2B}u^2 + u\sum_a\right) \\
&= \frac{1}{\sqrt{(2\pi)^{m+1} \det G\, B}} \int_{Q_+} du \prod dx_a\, \exp\left(-\frac{A}{2}\sum_a (x_a - u/A)^2 - \frac{1}{2}\left(\frac{1}{B} - \frac{m}{A}\right)u^2\right) \\
&= \frac{1}{\sqrt{2\pi\, A^m\, \det G\, B}} \int du\, [\mathrm{erfc}(u/\sqrt{A})]^m \exp\left(-\frac{1}{2}\left(\frac{1}{B} - \frac{m}{A}\right)u^2\right)
\end{aligned}$$

Hence

$$\ln A = \frac{1}{\sqrt{2\pi}} \int du\, [\mathrm{erfc}(\sqrt{Q/(1-Q)}\,u)]^m \exp\left(-u^2/2\right) \tag{5.2.7}$$

with[2]

$$\mathrm{erfc}(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty dt\, e^{-t^2/2}$$

[2]Warning: this is not the usual notation which would be

$$\mathrm{erfc}(x) = \frac{1}{\sqrt{\pi}} \int_x dt\, e^{-t^2}$$

The small m expansion of $\ln A$ is

$$\ln A = m\frac{1}{\sqrt{2\pi}}\int du\ \exp\left(-u^2/2\right)\ln\operatorname{erfc}\left(\sqrt{Q/(1-Q)}\,u\right)+o(m)$$

It is now necessary to take the derivative of this quantity with respect to Q:

$$\frac{d}{dQ}\ln A = m\frac{1}{2\pi\sqrt{Q(1-Q)^3}}\int du\ u\ \exp\left(-\frac{1}{2(1-Q)}u^2\right)\frac{1}{\operatorname{erfc}\left(\sqrt{Q/(1-Q)}\,u\right)}+o(m) \tag{5.2.8}$$

Thus, the RS equations (5.1.5) and (5.1.6) read

$$\begin{cases} -Q/2 = -1/2 + \frac{1}{\sqrt{8\pi i\hat{Q}}}\int dx\ x\ \exp\left(-\frac{1}{2}x^2\right)\tanh\left(x\sqrt{i\hat{Q}}\right) \\ -i\hat{Q}/2 = \alpha\frac{1}{2\pi\sqrt{Q(1-Q)^3}}\frac{1}{\sqrt{2\pi}}\int du\ u\ \exp\left(-\frac{1}{2(1-Q)}u^2\right)\frac{1}{\operatorname{erfc}\left(\sqrt{Q/(1-Q)}\,u\right)} \end{cases} \tag{5.2.9}$$

We are now interested in solving (5.2.9) in the $Q\to 1$ limit in which, however, the quantities in (5.2.9) are divergent. It is nonetheless possible to perform a $Q=1^-$ expansion and then impose that the divergences (or the infinitesimals) match.

Let us first consider (5.2.8). It is convenient to split the integral over u into two integrals over the positive and negative half lines:

$$\frac{1}{m}\frac{d}{dQ}\ln A = I_+ + I_-$$

Both I_+ and I_- are divergent, but I_+ is the dominant one. Since for $x<0$ $1/\operatorname{erfc}(x) < 2$

$$I_- \le \frac{1}{\pi\sqrt{Q(1-Q)}}$$

As for I_+, the asymptotic expansion[3] of erfc can be used. Hence

$$\operatorname{erfc}(\sqrt{Q/(1-Q)}u) \sim \frac{1}{u}\sqrt{\frac{1-Q}{2\pi Q}}\exp\left(-\frac{Q}{2(1-Q)}u^2\right)$$

Therefore

$$I_+ \sim -\frac{1}{2\sqrt{2\pi}(1-Q)^2}\int_0^\infty du\ u^2\ \exp(-u^2/2) = -\frac{1}{4(1-Q)^2} \tag{5.2.10}$$

One can conclude that $\frac{1}{m}\frac{d}{dQ}\ln A \sim I_+$. Furthermore, the previous result and (5.2.9) imply that

$$i\hat{Q} = \alpha\frac{1}{2(1-Q)^2}$$

[3]Let $f=\exp(-x^2/2)$. Then $f'=-xf$. Hence, integrating by parts:

$$\int f = -\int f'/x = -f/x - \int f/x^2$$

Thus

$$f(x)/x = \int_x^\infty dt\,(1+1/t^2)f$$

If $x\to\infty$

$$f(x)/x \sim \int_x^\infty dt\, f$$

It should be highlight that $i\hat{Q} > 0$, as anticipated (see (5.1.10)). In addition, it is now clear that in the $Q \to 1$ limit, $i\hat{Q} \to \infty$. This is relevant because the first saddle point equation (5.2.9) can be simplified by using the asymptotic approximation $\tanh(x) \sim \mathrm{sgn}(x)$ for $x \to \infty$:

$$\tanh\left(x\sqrt{i\hat{Q}}\right) = \tanh\left(\frac{\sqrt{\alpha}\,x}{\sqrt{2}(1-Q)}\right) \sim \mathrm{sgn}(x)$$

Hence, the first SP equation in (5.2.9) becomes

$$-Q/2 = -1/2 + \frac{(1-Q)}{\sqrt{2\pi\alpha}} + o(1-Q)$$

The equation is satisfied to the first order in $1-Q$ only if $\alpha = 4/\pi$. Hence, the RS capacity is

$$\alpha_c^{RS} = \frac{4}{\pi}$$

5.2.3 The simplest case: $\alpha \to 0$ in the discrete model

By setting $\alpha \to 0$, we should recover the results of section 4.1.3. This result matchs with what has been found in 4 and, therefore, the RS computation could be valid in this limit.

It is immediate from (5.1.6), that

$$i\hat{Q} = O(\alpha)$$

Moreover, by looking at the expansion around zero of $\tanh(x) = x + x^3/3 + o(x^3)$, one can conclude that

$$Q = O(\alpha)$$

as well. Furthermore, by looking at the expression of the function f

$$f = m\left\{\ln 2 - i\hat{Q}/2 + \frac{1}{\sqrt{2\pi}}\int dx\ \exp\left(-\frac{1}{2}x^2\right)\ln\cosh\left(x\sqrt{i\hat{Q}}\right)\right\}$$

and considering that $\ln\cosh(x) \sim x^2/2 - x^4/12 + o(x^4)$, we can deduce that

$$f = m\ln 2 + o(i\hat{Q}) = m\ln 2 + o(\alpha)$$

Analogously

$$i\hat{Q}Q = O(\alpha^2) = o(\alpha)$$

On the other hand

$$\ln A_q(Q) = m\frac{1}{2\pi}\int dy\ d\bar{y}\ \exp(-Y^t G_q\, Y/2)\ \ln E_q\,(0,0) + o(Q) = m\ln E_q\,(0,0) + o(Q)$$

is finite in Q. Hence

$$\alpha \ln A_q(0) = O(\alpha)$$

As explained in section 4.1.1, in order to obtain a nontrivial result, the sum over W must be divided by $1/2^N$. Hence, after the system is replicated, a prefactor

$$\frac{1}{2^{mN}}$$

appears. Therefore

$$F_q = \frac{1}{Nm}\lim_{N\to\infty}\ln N^{-m}\frac{1}{2^{mN}}e^{mN\ln 2 + Nm\alpha\ln E_q(0,0)} = p\ln E_q(0,0)$$

which is the expected result.

5.2.4 The $q \to 1$ limit: perceptron capacity

The limit $q \to 1$ can be used to recover the RS capacity in both the discrete and the spherical case. The reason is that q appears only in A and not in f. Therefore, f is left unchanged and it is enough to check that $\lim_{q\to 1} \ln A_q(Q)$ matches with Gardner's classic result as computed in 5.2.2.

In order to take the limit $q \to 1$, is useful to find a suitable set of variables in which the divergences are easy to treat:

$$(y, \bar{y}) \mapsto (S, D)$$

with

$$S = \frac{y + \bar{y}}{\sqrt{2}} \qquad S = \frac{y - \bar{y}}{\sqrt{2}}$$

and

$$(x, \bar{x}) \mapsto (s, d)$$

with

$$s = \frac{x + \bar{x}}{\sqrt{2}} \qquad d = \frac{x - \bar{x}}{\sqrt{2}}$$

The Jacobian is 1 in both cases. $\ln A_q$ becomes

$$\begin{aligned}
\ln A_q(Q) = m \frac{1}{2\pi} \int dS\, dD\ \exp\left(-\frac{1}{2}\sqrt{\frac{1-q}{1+q}} S^2 - \frac{1}{2}\sqrt{\frac{1+q}{1-q}} D^2\right) \\
\ln \frac{1}{\sqrt{2\pi}} \left[\frac{1-q}{1+q}\right]^{1/4} \int_{\sqrt{Q/(1-Q)}\, S}^{\infty} ds\ \exp\left(-\frac{1}{2}\sqrt{\frac{1-q}{1+q}}\, s^2\right) \\
\left\{ \operatorname{erfc}\left(\left[\frac{1+q}{1-q}\right]^{1/4} \left[\sqrt{\frac{Q}{1-Q}}(D+S) - s\right]\right) \right. \\
\left. - \operatorname{erfc}\left(\left[\frac{1+q}{1-q}\right]^{1/4} \left[\sqrt{\frac{Q}{1-Q}}(D-S) + s\right]\right) \right\}
\end{aligned}$$

The logarithm's argument decreases exponentially in the worst case. Consequently, ln diverges as a polynomial in S and D, in the worst case. Hence, it is the exponential in S and D which determines the relevant region for the S and D variables. Since the variance of S is $o((1-q)^{1/2})$ while the variance of D is $o((1-q)^{-1/2})$, it can be deduced that the relevant contribution only comes from

$$S \gg D$$

$$S = O((1-q)^{-1/4})$$

$$D = O((1-q)^{1/4})$$

By looking at the region of integration of s and its exponential, one can conclude that

$$s = O((1-q)^{-1/4})$$

$$s - \sqrt{Q/(1-Q)}\, S = ((1-q)^{-1/4})$$

Hence, D can be neglect in the arguments of the erfc functions. Consequently, D is decoupled and can be integrated away. Furthermore, the arguments of both erfc functions are both of order $O((1-q)^{-1/2})$ in the relevant region $s = O((1-q)^{-1/4})$. Therefore, if $u = s - \sqrt{Q/(1-Q)}\, S \to \infty$, then

$$\operatorname{erfc}(-u) - \operatorname{erfc}(u) \sim \operatorname{erfc}(-u) \sim \Theta(u)$$

But $\Theta(s - \sqrt{Q/(1-Q)}\,S) = 1$ since $s - \sqrt{Q/(1-Q)}\,S > 0$. Hence

$$\ln A_q(Q) = m\frac{1}{\sqrt{2\pi}}\left[\frac{1-q}{1+q}\right]^{1/4}\int dS\,dD\,\exp\left(-\frac{1}{2}\sqrt{\frac{1-q}{1+q}}S^2\right)$$
$$\ln\frac{1}{\sqrt{2\pi}}\left[\frac{1-q}{1+q}\right]^{1/4}\int_{\sqrt{Q/(1-Q)}\,S}^{\infty} ds\,\exp\left(-\frac{1}{2}\sqrt{\frac{1-q}{1+q}}\,s^2\right)$$

By rescaling both s and S, the special case of a single set of random patterns (see paragraph 5.2.2) is recovered:

$$\ln A_{q=1}(Q) = m\frac{1}{\sqrt{2\pi}}\int dS\,e^{-S^2/2}\ln\mathrm{erfc}(\sqrt{Q/(1-Q)}\,S)$$

This proves that, as expected, on average, the number of solutions to the learning problem of two maximally correlated sets of patterns is equivalent to that of a single set of the same size.

5.2.5 Computation of the RS generalization capacity. Part 1: $Q \to 1^-$; $D = 0$

Before starting over with the computation, let us call $\ln A$ as $\ln A_+$: this will be relevant later.

In this section the limit $Q \to 1$ will be taken, while keeping q and $1-q$ finite. As explained with more details in section 5.2.2, in the RS framework, the value $Q = 1$ signals the threshold of p/N above which no more patterns can be learned. This is exact in the spherical case, but not in the discrete case.

The first step is to compute the **generalization capacity** for $D = 0$. This will yield the function

$$\alpha_c^{RS}(q) = \alpha_c^{RS}(0|d(q)) \tag{5.2.11}$$

that can be referred to as a **correlation dependent capacity**. This quantity, multiplied by N, is the typical maximum size of two correlated sets of patterns which can be learned.

The core quantity is

$$\ln A_q^+(Q) = m\frac{1}{2\pi}\int dS\,dD\,\exp\left(-\frac{1}{2}\sqrt{\frac{1-q}{1+q}}S^2 - \frac{1}{2}\sqrt{\frac{1+q}{1-q}}D^2\right)$$
$$\ln\frac{1}{\sqrt{2\pi}}\left[\frac{1-q}{1+q}\right]^{1/4}\int_{\sqrt{Q/(1-Q)}\,S}^{\infty} ds\,\exp\left(-\frac{1}{2}\sqrt{\frac{1-q}{1+q}}\,s^2\right)$$
$$\left\{\mathrm{erfc}\left(\left[\frac{1+q}{1-q}\right]^{1/4}\left[\sqrt{\frac{Q}{1-Q}}(D+S) - s\right]\right)\right.$$
$$\left.-\,\mathrm{erfc}\left(\left[\frac{1+q}{1-q}\right]^{1/4}\left[\sqrt{\frac{Q}{1-Q}}(D-S) + s\right]\right)\right\}$$

It can be rewritten with a shift and a rescaling so that results from section 7.0.6 can

be used:

$$\ln A_q^+(Q) = m\frac{1}{2\pi}\int dS\ dD\ \exp\left(-\frac{1}{2}S^2 - \frac{1}{2}D^2\right) \tag{5.2.12}$$

$$\ln\frac{1}{\sqrt{2\pi}}\sqrt{\frac{1-Q}{Q}}\int_0^\infty ds\ \exp\left(-\frac{1}{2}\,\frac{Q}{1-Q}\,(s+S)^2\right) \tag{5.2.13}$$

$$\left\{\ \mathrm{erfc}\left(\sqrt{\frac{Q}{1-Q}}\left[D - \sqrt{\frac{1+q}{1-q}}s\right]\right)\right. \tag{5.2.14}$$

$$\left. -\ \mathrm{erfc}\left(\sqrt{\frac{Q}{1-Q}}\left[D + \sqrt{\frac{1+q}{1-q}}s\right]\right)\ \right\} \tag{5.2.15}$$

The limit $Q \to 1^-$ can be taken. Equation (7.0.10) with

$$M_{a,b}^{\pm}(x) = \Theta(\pm b - x)\ (x \mp b)^2\ + (x/R + a)^2 \tag{5.2.16}$$

and

$$R = \sqrt{\frac{1-q}{1+q}}$$

implies that the logarithm in the previous equation can be replaced with a quadratic form

$$\begin{aligned}
-\frac{1}{2}I_R(S,D) =& M_{S,D}^+((RD - R^2S)/[R^2+1])\ \Theta(RD+S)\ \Theta(D)\ \Theta(-S+DR)\\
&+ M_{S,D}^+(0)\ \Theta(RD+S)\ \Theta(D)\ \Theta(-D+SR)\\
&+ M_{S,D}^-((-RD - R^2S)/[R^2+1])\ \Theta(-D)\ \Theta(S-RD)\ \Theta(-RS-D)\\
&+ M_{S,D}^-(0)\ \Theta(-D)\ \Theta(S-DR)\ \Theta(RS+D)
\end{aligned}$$

in S and D whose prefactor is $Q/(1-Q)$. The integration of this quadratic form yields

$$1 + \frac{4}{\pi}\mathrm{atan}(R)$$

If $q = 1$, this result implies that $\ln A_q^+(Q \to 1^-) = -m\frac{1}{4(1-Q)}$ which is consistent with 5.2.2. However, if we choose $q = -1$, the result is $\ln A_{-1}^+(Q \to 1^-) = -m\frac{3}{4(1-Q)}$. This is physically inconsistent since, by proceeding further with the computation, it can be concluded that there exists a certain value of α below which the perceptron is capable of performing a contradictory task. In other words, it could give the same output for two opposite inputs ξ and $-\xi$. This is, however, not the case. If the limit $q \to -1^+$ is taken from equation (5.2.12), before the limit $Q \to 1^-$ is taken, it would follow, correctly, that

$$\lim_{q\to -1^+}\ln A_q^+(Q) = -\infty$$

since

$$\lim_{\epsilon\to 0^+}\int_0^\infty dx\ e^{-\frac{1}{2}\,(x-a)^2}\ \mathrm{erfc}(x/\epsilon + b) = 0$$

The reason for this discrepancy is that the limit $q \to -1^+$ and $Q \to 1^-$ do not commute. This means that there exists a singular point at $q = -1$ that ensures that $\alpha_c(-1) = 0$. The implications of this discontinuity will be discussed later. In the present section it is enough to correct the result of the integral with a delta function that introduces the missing divergence. The "ugly" term $\infty\,\delta(q+1)$ should be intended

as the result of a limit and is not problematic since it will not appear directly in physical quantities.

Hence

$$\ln A_q^+(Q) \sim -m\frac{Q}{1-Q}\frac{1}{4\pi}\int dS\, dD\, \exp\left(-\frac{1}{2}S^2-\frac{1}{2}D^2\right)\, I_R(S,D)$$

which is

$$\boxed{\ln A_q^+(Q\to 1^-) = -m\frac{1}{1-Q}\frac{1}{4}T_+(R)} \tag{5.2.17}$$

with

$$\boxed{T_+(q) = 1+\frac{4}{\pi}\text{atan}(R)+\infty\,\delta(1+q)} \tag{5.2.18}$$

Finally

$$\boxed{\frac{d}{dQ}\ln A_q^+(Q\to 1^-) = -m\frac{1}{(1-Q)^2}\frac{1}{4}T_+(R)} \tag{5.2.19}$$

5.2.6 RS correlation dependent capacity $\alpha_c(q)$

The previous results can be used to compute $\alpha_c(q) = \alpha_c(0|d(q))$. Let us begin with the spherical perceptron. To obtain this result, equations (5.2.19) and (5.1.9) can be combined to solve the SP equations.

The first SP equation is

$$-iQ/2 = \frac{d}{d\hat{Q}}f_{sph}$$

which is satisfied by

$$i\hat{Q} = \frac{1}{(1-Q)^2}$$

From (5.1.6) it holds that

$$-m\frac{Q}{2(1-Q)^2} = \alpha\frac{d}{dQ}\ln A_q \tag{5.2.20}$$

If $Q\to 1$, then

$$-m\frac{Q}{2(1-Q)^2} = -m\frac{\alpha}{4(1-Q^2)}T(q)$$

The following result is the **correlation dependent capacity** of the spherical perceptron and is plotted in figure 5.1.

$$\alpha_c^{sph}(q) = \frac{2}{T(q)} \tag{5.2.21}$$

$$\boxed{\alpha_c^{sph}(q) = \frac{2}{1+4/\pi\text{atan}(\sqrt{(1+q)/(1-q)})}-\frac{2}{3}\delta(1+q)} \tag{5.2.22}$$

This result is compatible with the literature: in the $q\to 1$ limit, the capacity reduces to 2. Furthermore, unlike in the discrete case, since the RS ansatz is correct in the spherical case even at $\beta=\infty$, this result should be exact.

The discrete case is almost identical, with the single difference that $4/\pi$ appears in the result instead of 2

$$\alpha_c^{dis}(q) = \frac{4/\pi}{1+4/\pi\text{atan}(\sqrt{(1+q)/(1-q)})}-\frac{4}{3\pi}\delta(1+q) \tag{5.2.23}$$

this result is not expected to be correct, though.

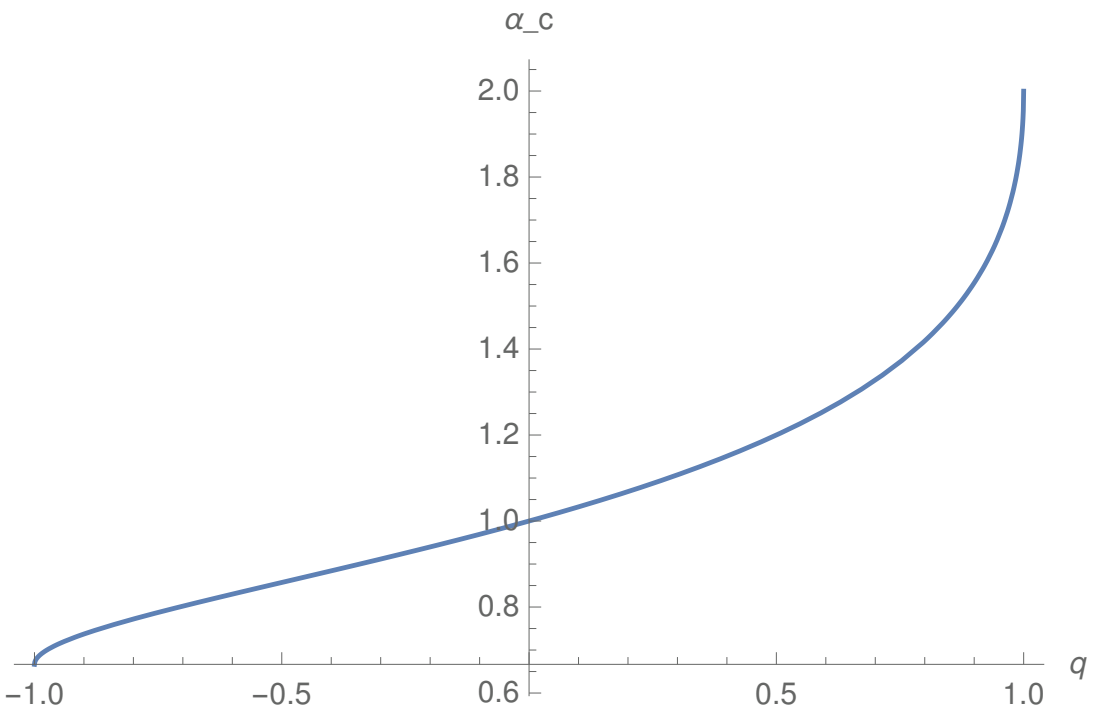


Figure 5.1: Correlation dependent capacity $\alpha_c(q)$ (5.2.21) for spherical perceptron.

5.2.7 Fixed error $D \neq 0$ for a given input overlap q in the RS framework: derivation of the SP equations

In this chapter, the results from the computation of $F(0|d)$ will be generalized to the $D \neq 0$ case. The steps are analogous to the simpler case $D = 0$, however the notation is a little heavier. The goal is to compute:

$$F(D|q) = \int P_q(\xi, \bar{\xi}) \ln \sum_W \prod_\mu \Theta\left(\frac{\xi^\mu \cdot W}{\sqrt{N}}\right) \sum_{\{\eta_\mu = \pm 1\}} \delta\left(\sum_\mu \eta_\mu, (1-2D)p\right) \prod_{\bar{\mu}} \Theta\left(\eta_\mu \frac{\bar{\xi^\mu} \cdot W}{\sqrt{N}}\right) \quad (5.2.24)$$

The replica method will be applied to this case too:

$$F(D|q) = \lim_{m\to 0} \frac{1}{m} \ln \int P(\xi, \bar{\xi}) \sum_{\{W^a\}} \prod_{a=1}^m \prod_\mu \Theta\left(\frac{\xi^\mu \cdot W^a}{\sqrt{N}}\right) \sum_{\{\eta_\mu^a = \pm 1\}} \delta\left(\sum_\mu \eta_\mu^a, (1-2D)p\right) \prod_{\bar{\mu}} \Theta\left(\eta_\mu^a \frac{\bar{\xi^\mu} \cdot W^a}{\sqrt{N}}\right)$$

The next step is to introduce the order parameter Q_{ab} and the auxiliary variable $N\hat{Q}_{ab}$:

$$F(D|q) = \lim_{m\to 0} \frac{1}{m} \ln N^{-m(m-1)} \int dQ_{ab}\, d\hat{Q}_{ab} \sum_{\{W_a\}} e^{-iN \sum_{a<b} \hat{Q}_{ab}\left(Q_{ab} - \frac{W^a \cdot W^b}{N}\right)}$$
$$\int P(\xi, \bar{\xi}) \prod_{a=1}^m \prod_\mu \Theta\left(\frac{\xi^\mu \cdot W^a}{\sqrt{N}}\right) \sum_{\{\eta_\mu^a = \pm 1\}} \delta\left(\sum_\mu \eta_\mu^a, (1-2D)p\right) \prod_{\bar{\mu}} \Theta\left(\eta_\mu^a \frac{\bar{\xi^\mu} \cdot W^a}{\sqrt{N}}\right)$$

The following quantity appears in the integral (the sums and products have been rearranged):

$$\mathcal{A}(Q) := \sum_{\{\eta_\mu^a=\pm 1\}} \delta\left(\sum_\mu \eta_\mu^a, (1-2D)p\right) \int P(\xi,\bar{\xi}) \prod_\mu \prod_{a=1}^m \Theta\left(\frac{\xi^\mu \cdot W^a}{\sqrt{N}}\right) \prod_{\bar{\mu}} \Theta\left(\eta_\mu^a \frac{\bar{\xi}^{\bar{\mu}} \cdot W^a}{\sqrt{N}}\right) \quad (5.2.25)$$

$\mathcal{A}$ can be rewritten, if we work under the assumption that $P(\xi,\bar{\xi}) = \prod_\mu P(\xi^\mu, \bar{\xi}^{\bar{\mu}})$ is factorized, as explained before. Therefore, calling $w_a^\mu = \xi^\mu \cdot W^a/\sqrt{N}$ and $\bar{w}_a^\mu = \bar{\xi} \cdot W^a/\sqrt{N}$

$$\mathcal{A}(Q) = \sum_{\{\eta_\mu^a=\pm 1\}} \left(\sum_\mu \eta_\mu^a, (1-2D)p\right) \prod_\mu \frac{1}{(2\pi)^m \sqrt{\det G}} \int \prod_c^m dw_c \, d\bar{w}_c \exp\left(-\frac{1}{2} \sum_{u,v=\pm 1} \sum_{a,b=1}^m w_a^{\mu u} [G^{-1}]_{ab}^{uv} w_b^{\mu v}\right) \prod_{a,b} \Theta(w_a^\mu)\Theta(\eta_\mu^b \, \bar{w}_b^\mu) \quad (5.2.26)$$

with

$$\bar{w}_a = w_a^{(-1)} \qquad w_a = w_a^{(1)}$$

and (Q is a $m \times m$ matrix with $Q_{aa} = 1$)

$$G = \begin{bmatrix} Q & qQ \\ qQ & Q \end{bmatrix}$$

$$G^{-1} = \frac{1}{(1-q^2)} \begin{bmatrix} Q^{-1} & -qQ^{-1} \\ -qQ^{-1} & Q^{-1} \end{bmatrix}$$

For any given μ, let us define

$$X_\mu := \frac{1}{2}\left(m - \sum_{a=1}^m \eta_\mu^a\right) \quad (5.2.27)$$

for which the following equation holds

$$\sum_{\mu=1}^p X_\mu = mpD \quad (5.2.28)$$

Let us rewrite equation (5.2.26) in the following way

$$\mathcal{A}(Q) = \int \prod_\mu dX_\mu \sum_{\{\eta_\mu^a=\pm 1\}} \left(\sum_\mu \eta_\mu^a, (1-2D)p\right) \delta\left(X_\mu - \frac{1}{2}\left(m - \sum_{a=1}^m \eta_\mu^a\right)\right) \prod_\mu A_\mu(X_\mu|q) \quad (5.2.29)$$

Let us now assume to be working in the RS ansatz. As can be deduced by looking at the passages in section 5.1.7, the functions $A_\mu(X_\mu|d)$ only depend on $\{\eta_\mu^a\}$ through

X_μ. It can be seen that

$$A_\mu(X_\mu|q) = \frac{1}{2\pi}\int dy\; d\bar{y}\; \exp(-Y^t G_q Y/2)$$
$$\left[E_q^+\left(\sqrt{\frac{Q}{1-Q}}y, \sqrt{\frac{Q}{1-Q}}\bar{y}\right)\right]^{m-X_\mu}\; \left[E_q^-\left(\sqrt{\frac{Q}{1-Q}}y, \sqrt{\frac{Q}{1-Q}}\bar{y}\right)\right]^{X_\mu} \quad (5.2.30)$$

with

$$E_q^+(y,\bar{y}) = \frac{1}{2\pi}\int_y^\infty dx \int_{\bar{y}}^\infty d\bar{x}\; \exp(-Y^t G_q Y/2) \quad (5.2.31)$$

$$E_q^-(y,\bar{y}) = \frac{1}{2\pi}\int_y^\infty dx \int_\infty^{\bar{y}} d\bar{x}\; \exp(-Y^t G_q Y/2) \quad (5.2.32)$$

If we expand A_μ for small m (and $X_\mu = 0(m)$), we find

$$A_\mu(D|q) = 1 + (m - X_\mu)\; A^+(q) + X_\mu\; A^-(q) + o(m)$$

with

$$A^\pm(Q)_q = \frac{1}{2\pi}\int dy\; d\bar{y}\; \exp(-Y^t G_q Y/2)\; \ln E_q^\pm\left(\sqrt{\frac{Q}{1-Q}}y, \sqrt{\frac{Q}{1-Q}}\bar{y}\right) \quad (5.2.33)$$

Hence

$$\prod_\mu A_\mu(X_\mu|q) = \prod_\mu e^{(m-X_\mu)\; A_q^+(Q) + X_\mu\; A_q^-(Q)}$$

It follows from (5.2.28) that

$$\prod_\mu A_\mu(X_\mu|q) = e^{mp[(1-D)\; A - q^+(Q) + D\; A_q^-(Q)]}$$

On the other hand

$$\int \prod_\mu dX_\mu \sum_{\{\eta_\mu^a = \pm 1\}} \left(\sum_\mu \eta_\mu^a, (1-2D)p\right) \delta\left(X_\mu - \frac{1}{2}\left(m - \sum_{a=1}^m \eta_\mu^a\right)\right) = \binom{p}{pD}^m$$

Therefore

$$\ln \mathcal{A}(D|q) \sim mp[(1-D)\; A^+(q) + D\; A^-(q) + (1-D)\ln(1-D) + D\ln D] \quad (5.2.34)$$

Then

$$F(D|q) = \lim_{m\to 0}\frac{1}{m}\ln\int dQ\; d\hat{Q}\; e^{-iN\frac{m(m-1)}{2}Q\hat{Q} + Nf(\hat{Q})}\mathcal{A}(D|q)$$

or

$$F(D|q) = \lim_{m\to 0}\frac{1}{m}\ln\int dQ\; d\hat{Q}\; e^{N\left[-i\frac{m(m-1)}{2}Q\hat{Q} + f(\hat{Q}) + m\alpha[(1-D)\; A_q^+(Q) + D\; A_q^-(Q)]\right]} \quad (5.2.35)$$

This last equation implies that, in order to allow for $D \neq 0$, it is enough to replace

$$\ln A_q(Q) \mapsto (1-D)\; A_q^+(Q) + D\; A_q^-(Q) \quad (5.2.36)$$

in the saddle point equations.

5.2.8 Computation of the generalization capacity. Part 2: $Q \to 1^-; D \neq 0$

In case $D \neq 0$, we should compute $\ln A_q^-$ as well. The same change of variables as before yields

$$\begin{aligned}\ln A_q^- = \frac{m}{2\pi}\int dS\, dD\, \exp\left(-\frac{1}{2}\sqrt{\frac{1-q}{1+q}}S^2 - \frac{1}{2}\sqrt{\frac{1+q}{1-q}}D^2\right) \\ \ln\frac{1}{\sqrt{2\pi}}\left[\frac{1-q}{1+q}\right]^{1/4}\left\{\int_{-\infty}^{\sqrt{Q/(1-Q)}S} ds\, \exp\left(-\frac{1}{2}\sqrt{\frac{1-q}{1+q}}s^2\right)\right. \\ \operatorname{erfc}\left(\left[\frac{1+q}{1-q}\right]^{1/4}\left[\sqrt{\frac{Q}{1-Q}}(D+S) - s\right]\right) \\ \left. + \int_{\sqrt{Q/(1-Q)}S}^{\infty} ds\, \exp\left(-\frac{1}{2}\sqrt{\frac{1-q}{1+q}}s^2\right)\right. \\ \left.\operatorname{erfc}\left(\left[\frac{1+q}{1-q}\right]^{1/4}\left[\sqrt{\frac{Q}{1-Q}}(D-S) + s\right]\right)\right\}\end{aligned}$$

This can conveniently rewritten as

$$\begin{aligned}\ln A_q^- = \frac{m}{2\pi}\int dS\, dD\, \exp\left(-\frac{1}{2}S^2 - \frac{1}{2}D^2\right) \\ \ln\frac{1}{\sqrt{2\pi}}\left\{\int_0^{\infty} ds\, \exp\left(-\frac{1}{2}\frac{Q}{1-Q}(s-S)^2\right)\right. \\ \operatorname{erfc}\left(\left[\sqrt{\frac{Q}{1-Q}}\left[D + \sqrt{\frac{1+q}{1-q}}s\right]\right]\right) \\ + \int_0^{\infty} ds\, \exp\left(-\frac{1}{2}\frac{Q}{1-Q}(s+S)^2\right) \\ \left.\operatorname{erfc}\left(\left[\sqrt{\frac{Q}{1-Q}}\left[D + \sqrt{\frac{1+q}{1-q}}s\right]\right]\right)\right\}\end{aligned}$$

so that results from section 7.0.6 (see 7.0.4) can be applied. We are interested in the limit $Q = 1^-$. Let

$$L_{a,b}^+(x) = \left[(x/R - b)^2\, \Theta(x/R - b) + (x - a)^2\,\right] \tag{5.2.37}$$

and

$$Y_R^+(a,b) = \min_{x>0} L_{a,b}^+(x) \tag{5.2.38}$$

then

$$\begin{aligned}\ln A_q^- \sim -\frac{m}{2\pi}\int dS\, dD\, \exp\left(-\frac{1}{2}S^2 - \frac{1}{2}D^2\right) \\ \frac{1}{2}\frac{1}{1-Q}\min[Y_R(-S,-D), Y_R(S,-D)]\end{aligned}$$

The integral can be done by splitting the region of integration in a proper way. It turns out that

$$\begin{aligned}
&\min[Y_R(-S,-D), Y_R(S,-D)] \\
&\quad = \Theta(-D)\ \Theta(S)\ \Theta(S+RD)\ L_{S,-D}((S-D/R)/(1+R^2)) \\
&\quad = \Theta(-D)\ \Theta(S)\ \Theta(-S+RD)\ L_{-S,-D}((-S-D/R)/(1+R^2)) \\
&\quad = \Theta(D)\ \Theta(-S)\ [\Theta(S+D/R)\ L_{-S,-D}(0) + \Theta(-S-D/R)\ L_{-S,-D}((-S-D/R)/(1+R^2))] \\
&\quad = \Theta(D)\ \Theta(S)\ [\Theta(S-D/R)\ L_{S,-D}((S-D/R)/(1+R^2)) + \Theta(-S+D/R)\ L_{S,-D}(0)]
\end{aligned}$$

As before, we should add an infinite delta contribution in the point $q = 1$ to the result. Finally

$$\ln A_q^- \sim -\frac{m}{4(1-Q)}\ T_-(q) \tag{5.2.39}$$

with

$$\boxed{T_-(q) = \left[1 + \frac{4}{\pi}\,\mathrm{acot}(R)\right] + \infty\ \delta(q-1)} \tag{5.2.40}$$

and

$$R = \sqrt{\frac{1-q}{1+q}}$$

The conclusion is that

$$\boxed{\frac{d}{dQ}\ln A_q^- = -\frac{m}{4(1-Q)^2}\ T_-(q)} \tag{5.2.41}$$

This result can be put together with the previous one. The complete RS saddle point equation involving $\ln A$ (see (5.2.20) and (5.2.36)) reads

$$-i\hat{Q}/2 = -\alpha\frac{1}{4(1-Q)^2}\ [(1-D)\ T_+(q) + D\ T_-(q)] \tag{5.2.42}$$

5.3 Analytical results

5.3.1 Result: the generalization capacity and its phase diagram

As already pointed out, in the RS framework, the expressions for the generalization capacity of the spherical and the discrete model are only different by a factor. While the former is correct, the second is not. Likewise, we can expect the RS generalization capacity to be correct only for the spherical model.

The **generalization capacity for the spherical perceptron** is

$$\boxed{\alpha_c^{sph}(D|d) = \alpha_c^{sph}\left\{\frac{1}{(1-D)\ T_+^{RS}(d) + D\ T_-^{RS}(d)}\right.}$$

$$\left.\boxed{-\frac{1}{3}[\delta(d)\,(1-\delta(D)) + \delta(d-1)\,(1-\delta(1-D))]}\right\} \tag{5.3.1}$$

with

$$\boxed{T_\pm^{RS}(q) = 1 + (4/\pi)\mathrm{atan}(R^{\pm 1})} \tag{5.3.2}$$

$$\boxed{R = \sqrt{(1-q)/(1+q)} = \sqrt{\frac{d}{1-d}}} \tag{5.3.3}$$

$$\alpha_c^{sph} = 2$$

The function is plotted in figure 5.2. While it behaves as one could intuitively expect, some less trivial conclusions can be drawn from this formula.

- The vertical tangent in $q = 1$ for fixed $D = 0$ implies that, when α is close to the capacity, there is typically no chance that any input is classified as the training set.

- The generalization capacity never vanishes but when at $(d = 1, D \neq 0)$ and $(d = 1, D \neq 1)$. This means that, for any choice of (d, D) there exists some α below which $F(D|d) > -\infty$

- There is a value of α

$$\boxed{\alpha_{low} = \frac{\alpha_c^{sph}}{3} = \frac{2}{3}} \tag{5.3.4}$$

  below which $F_\alpha(D|d) > -\infty$ (there are always solutions) except in the case in which a contradictory task is required i.e. expecting different outputs for the same input or expecting the same output for opposite inputs.

  Conversely, for $\alpha > 2/3$, for every $d > 1/2$, there exists some D_d so that if $D > D_d$ for which $F_\alpha(D|d) = -\infty$. Conversely, for every $d < 1/2$, there exists some D_d so that if $D < D_d$, then $F_\alpha(D|d) = -\infty$.

  If we consider $D = 0$, the existence of α_{low} is equivalent to saying that, below α_{low}, the perceptron can be successfully trained to fully discern two arbitrarily similar sets unless they are identical.

  Conversely, given $D = 1$, this means that, below α_{low}, the perceptron can be successfully trained to give the same output to two arbitrarily different sets unless they are opposite.

- The function $\alpha_c(D|d)$ is invariant with respect to the transformation

$$(d, D) \mapsto (1 - d, 1 - D)$$

  i.e. a reflection with respect to the point $(1/2, 1/2)$.

In the case of the discrete perceptron the computation is alike and the results is

$$\alpha_c^{discrete(RS)}(D|d) = \alpha_c^{RS} \left\{ \frac{1}{(1-D)\, T_+^{RS}(d) + D\, T_-^{RS}(d)} - \frac{1}{3}[\delta(d)\,(1-\delta(D)) + \delta(d-1)\,(1-\delta(1-D))] \right\} \tag{5.3.5}$$

with $\alpha_c^{RS} = 4/\pi$. While it is a first step, this result is surely not correct and probably not even a good qualitative approximation. This means that a RSB computation is needed.

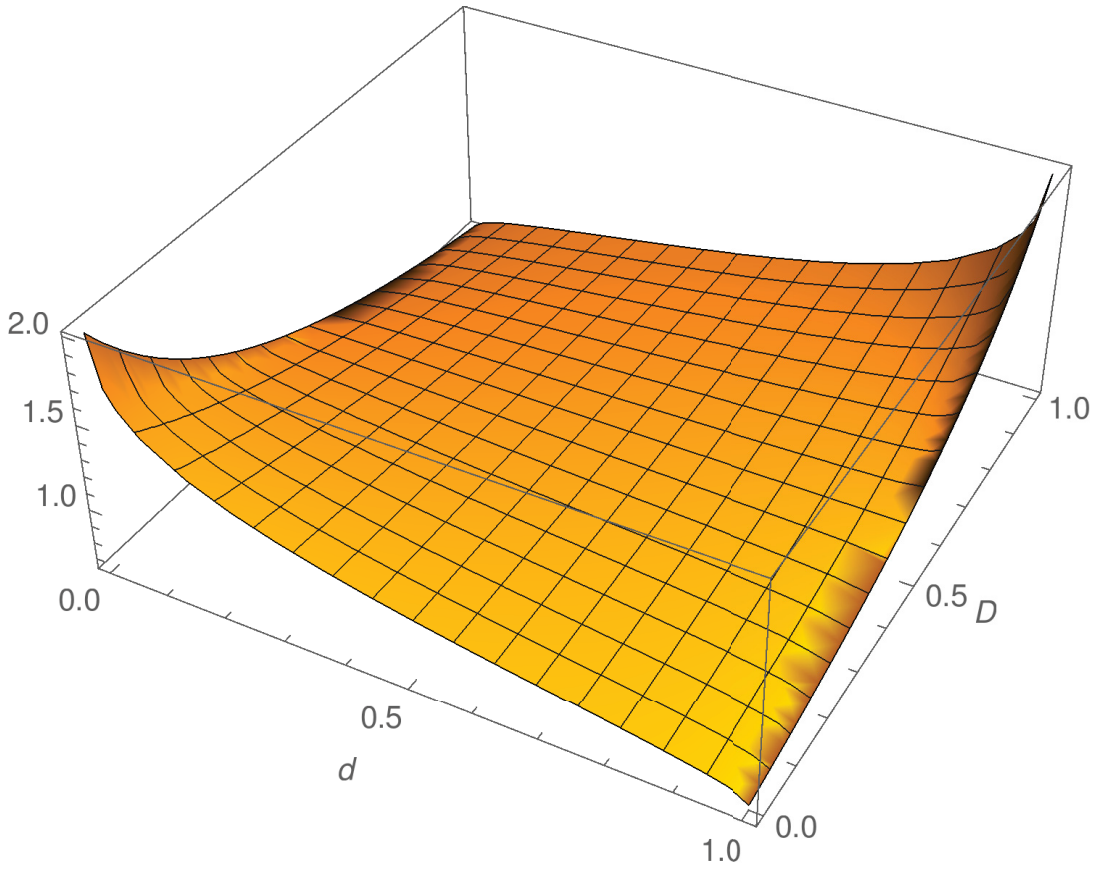


Figure 5.2: The generalization capacity $\alpha_c^{sph}(D|d)$ for spherical perceptron. The delta contribution is omitted here

Chapter 6

Outlook and possible future development

6.1 A list of open questions

In the previous chapters, it has been shown that generalization can be studied in terms of input correlations with the replica formalism. In this section we propose a list of some open questions.

- In chapter 3 we have presented distance (3.2.15) and we have presented an euristic argument if favor of the factorization of the probability $P_d(\xi, \bar{\xi})$ with this measure. A more rigorous approach could be followed to quantify the goodness of this assumption.
- In chapter 3 we have chosen to use distance (3.2.15). It could be interesting to study whether other distances can be defined to probe different properties of perceptron.
- The computations presented in this thesis are aimed at studying the quantity $F(D|d)$. The computations could be repeated for (3.2.9) and compared to the results presented here.
- Due to working precision issues, some numerical computations have been left for future works. It could be interesnting to complete them to get the full expression of $F(D|d)$.
- As it has been remarked multiple times, RSB results are incorrect for the discrete model. A natural developement of this thesis could be a study of the stability of the solutions (even the spherical result should be checked, for the sake completeness, in terms of stability) and the sign of the entropy. Moreover, an RSB computation is probabily needed in this case.

As a last remark, we close this thesis with a preliminary approach to the RSB computation. The purpose of the following sections is to show that the gaussian formalism which has been adopted can be generalized to the RSB case. This is shown in the simpler case of a single set of pattern as it would be required to compute 1RSB capacity. The aim is not to solve the SP equations, which would require more than a few sections, but to show that the fundamental quantities can be computed.

6.2 A preliminary gaussian approach to RSB computations

6.2.1 An algebra for RSB

The aim of the following section is to provide a preliminary approach to the computations in the RSB ansatz, as mentioned in the previous section.

The first step is to define the matrices

$$D(k,m) = \mathrm{diag}(\underbrace{1_k, 1_k, ..., 1_k}_{m/k=n}) \tag{6.2.1}$$

with $m/k \in \mathbb{N}$ and with 1_k being a $k \times k$ suare matrix with all entries equal to one. For a given m, these matrixes form a closed abelian algebra with the matrix product

$$D(k,m)D(k',m) = \min(k,k')\, D(\max(k,k'),m) \tag{6.2.2}$$

Let us observe that

$$D(1,m) = \mathbb{I}_m$$
$$D(m,m) = 1_m$$

For any given increasing sequence of integers

$$K = \{k_i\}_{i=0,1,...,h,h+1}$$

such that

$$n_i := k_{i+1}/k_i \in \mathbb{N}$$
$$k_{h+1} = m \qquad k_0 = 1$$

the matrix $Q \in M_m(\mathbb{R})$, according to the RSB ansatz, can be written as

$$Q_K = \sum_{i=0}^{h+1} C(i)\, D(k_i,m) \tag{6.2.3}$$

The coefficients $C(i)$ are chosen in the following way

$$C(0) = 1 - Q_1$$
$$C(h+1) = Q_{h+1}$$
$$C(i) = Q_i - Q_{i+1}$$

Given this premise, Q_K^{-1} can be expanded on the same algebra

$$Q_K^{-1} = \sum_{i=0}^{h+1} \bar{C}(i)\, D(k_i,m) \tag{6.2.4}$$

The coefficients $\bar{C}(i)$ satisfy the following equations

$$\begin{cases} 1 = \bar{C}(0)C(0) \\ 0 = \sum_{j<i}[k_j\, C(i)\bar{C}(j) + k_j\, \bar{C}(i)C(j)] + k_i\bar{C}(i)C(i) \quad i > 1 \end{cases} \tag{6.2.5}$$

which can be solved recursively.

6.2.2 1RSB Capacity

The first nontrivial case is 1RSB. The matrix Q will be assumed to be a $m \times m$ matrix. All elements on the diagonal are 1. There are $n = m/k$ diagonal blocks of size k, whose entries (except for the diagonal) are equal to Q_1. All other entries are Q_2. By using the previous notation

$$Q^{-1} = C(0)\mathbb{I}_m + C(1)D(k,m) + C(2)1_m$$

with

$$C(0) = 1 - Q_1 \qquad C(1) = Q_1 - Q_2 \qquad C(2) = Q_2$$

The recursive equations yield

$$\bar{C}(0) = \frac{1}{1-Q_1}$$

$$\bar{C}(1) = -\frac{Q_1 - Q_2}{(1-Q_1)(1-Q_1+k_2(Q_1-Q_2))}$$

$$\bar{C}(2) = -\frac{Q_2}{(1-Q_1+k_2(Q_1-Q_2))(1-Q_1+k_2(Q_1-Q_2)+mQ_2)}$$

Since $Q_1 > Q_2$, then

$$\bar{C}(0) > 0 \qquad \bar{C}(1) < 0 \qquad \bar{C}(2) < 0$$

For simplicity, we will use $-\bar{C}(1)$ ($\bar{C}(1) > 0$) instead of $\bar{C}(1)$ and the same for $\bar{C}(2)$.

In order to compute the 1RSB capacity, the next step is to compute $\ln A$ and f.

$$\ln A = \frac{1}{\sqrt{(2\pi)^m \det G}} \int_{Q_+} \prod dx_a$$
$$\exp\left(-\frac{\bar{C}(0)}{2} \sum_a x_a^2 + \frac{\bar{C}(1)}{2} \sum_{ab} x_a x_b + \frac{\bar{C}(2)}{2} \sum_{l=1}^{n} \sum_{a,b=(l-1)k}^{lk} x_a x_b \right)$$

Now, a set of auxiliary variables can be introduced. Each variable will have as many labels as the RBS level. Each index labels the position of the variables in a square submatrix. In this case, we will have

$$u \text{ and } \{u_l\} \text{ with } l = 1, ..., n$$

Hence

$$\ln A = \frac{1}{\sqrt{(2\pi)^m \det G}} \frac{1}{\sqrt{(2\pi)^{n+1} \bar{C}(1) \bar{C}(2)^n}} \int_{Q_+} du \prod du_l \prod dx_a$$
$$\exp\left(-\frac{\bar{C}(0)}{2} \sum_a x_a^2 + u \sum_a x_a + \sum_{l=1}^{n} u_l \sum_{a=(l-1)k}^{lk} x_a - \frac{1}{2\bar{C}(1)} u^2 - \frac{1}{2\bar{C}(2)} \sum_{l=1}^{n} u_l^2 \right)$$

The sum over a (general replica index) can be split into n subsums. After a little

algebra, one gets

$$\ln A = \frac{1}{\sqrt{(2\pi)^m \det G}} \frac{1}{\sqrt{(2\pi)^{n+1} \bar{C}(1) \bar{C}(2)^n \bar{C}(0)^m}} \int du \prod_{l=1}^{n} du_l \left[\mathrm{erfc}\left(x - u/\sqrt{\bar{C}(0)} - u_l/\sqrt{\bar{C}(0)} \right) \right]^k \exp\left(-\frac{1}{2}\left(\frac{1}{\bar{C}(2)} - \frac{m}{\bar{C}(0)} \right) u^2 - \frac{1}{2}\left(\frac{1}{\bar{C}(1)} - \frac{k}{\bar{C}(0)} \right) \sum_{l=1}^{n} u_l^2 - \frac{k}{\bar{C}(0)} u \sum_{l=1}^{n} u_l \right)$$

The final expression for $\ln A$, which depends explicitly on m, k and n, is

$$\ln A = \frac{1}{\sqrt{\det G}} \frac{1}{\sqrt{(2\pi)^{n+1} \bar{C}(1) \bar{C}(2)^n \bar{C}(0)^m}} \int du \prod_{l=1}^{n} du_l \exp\left(-\frac{\bar{C}(0)}{2}\left(\frac{1}{\bar{C}(1)} - \frac{m}{\bar{C}(0)} \right) u^2 \right) \left[\int du_1 \exp\left(-\frac{\bar{C}(0)}{2}\left(\frac{1}{\bar{C}(2)} - \frac{k}{\bar{C}(0)} \right) u_1^2 - k u u_1 \right) \left[\mathrm{erfc}\left(x - u\sqrt{\bar{C}(0)} - u_1\sqrt{\bar{C}(0)} \right) \right]^k \right]^n$$

Let us now consider f. The 1RSB matrix $\hat{Q}$ is

$$\hat{Q} = (\hat{Q}_1 - \hat{Q}_2) D(k, m) + \hat{Q}_2 \, 1_m$$

$$
\begin{aligned}
e^f &= \sum_{\{W_a\}} \exp\left(i(\hat{Q}_1 - \hat{Q}_2) \sum_{a<b} W_a W_b + i\hat{Q}_2 \sum_{l=1}^{n} \sum_{l(k-1)<a<b<lk} W_a W_b \right) \\
&= \sum_{\{W_a\}} \exp(-im\hat{Q}_1/2) \exp\left(i(\hat{Q}_1 - \hat{Q}_2) \sum_{a,b} W_a W_b/2 + i\hat{Q}_2 \sum_{l=1}^{n} \sum_{l(k-1)<a,b<lk} W_a W_b/2 \right) \\
&= \frac{1}{\sqrt{i\hat{Q}_2}\sqrt{i(\hat{Q}_1 - \hat{Q}_2)}} \sum_{\{W_a\}} \exp(-im\hat{Q}_1/2) \int dx \prod_{l=1}^{n} dx_l \\
&\qquad \exp\left(-\frac{1}{2iQ_2}x^2 - \frac{1}{2i(\hat{Q}_1 - \hat{Q}_2)} \sum_{l=1}^{n} x_l^2 + \sum_a W_a x + \sum_{l=1}^{n} x_l \sum_{l(k-1)<a<lk} W_a \right) \\
&= \frac{1}{\sqrt{i\hat{Q}_2}\sqrt{i(\hat{Q}_1 - \hat{Q}_2)}} 2^m \exp(-im\hat{Q}_1/2) \int dx \prod_{l=1}^{n} dx_l \\
&\qquad \exp\left(-\frac{1}{2iQ_2}x^2 - \frac{1}{2i(\hat{Q}_1 - \hat{Q}_2)} \sum_{l=1}^{n} x_l^2 \right) \prod_{l=1}^{n} \left[\cosh(x + x_l)\right]^k \\
&= \frac{1}{\sqrt{i\hat{Q}_2}\sqrt{i(\hat{Q}_1 - \hat{Q}_2)}} 2^m \exp(-im\hat{Q}_1/2) \int dx \ \exp\left(-\frac{1}{2iQ_2}x^2 \right) \\
&\qquad \left[\int dx_1 \ \exp\left(-\frac{1}{2i(\hat{Q}_1 - \hat{Q}_2)} x_1^2 \right) \left[\cosh(x + x_1)\right]^k \right]^n \\
&= 2^m \exp(-im\hat{Q}_1/2) \int dx \ \exp\left(-\frac{1}{2}x^2 \right) \left[\int dx_1 \right. \\
&\quad \left. \exp\left(-\frac{1}{2}x_1^2 \right) \left[\cosh\left(x\sqrt{i\hat{Q}_2} + x_1\sqrt{i(\hat{Q}_1 - \hat{Q}_2)} \right) \right]^k \right]^n
\end{aligned}
$$

Finally, the saddle point equations should be written.

$$
\begin{aligned}
\frac{m(k-1)}{2} i\hat{Q}_1 &= \frac{d}{dQ_1} \ln A \\
\frac{m(m-k-2)}{2} i\hat{Q}_2 &= \frac{d}{dQ_2} \ln A \\
\frac{m(k-1)}{2} iQ_1 &= \frac{d}{d\hat{Q}_1} f \\
\frac{m(m-k-2)}{2} iQ_2 &= \frac{d}{d\hat{Q}_2} f
\end{aligned}
$$

6.2.3 h-RSB capacity

The passages of the previous sections will be generalized to derive the capacity with arbitrary RS breaking.

The matrix Q will be taken as described in section 6.2.1: a $k_{h+1} \times k_{h+1} = m \times m$ matrix with $n_h = k_{h+1}/k_h$ diagonal blocks of size k_h. Each of these has n_{h-1} diagonal sub-blocks of size k_{h-1} and so on. Let us call I any set of integers which identify the

position of all diagonal blocks in which a certain diagonal element is contained. I is a string of $h+1$ ordered integers. The first integer identifies the positon of the element in the smallest set of blocks

$$I_h = (i_h, i_{h-1}, ..., i_1, i_0) \text{ with } 0 < i_0 \leq k_1 = n_0;\ 0 < i_2 \leq n_2; ...; 0 < i_h \leq n_h$$

Analogously, for any $r < h+1$, a set of $h-r$ integers I_r identifies the position a diagonal block within all super blocks in which it is contained. Furthermore, the set of such indices can be equipped with a partial-ordering relationship. Let us say that

$$I_t \prec I_{t'}$$

if $t > t'$ and I_t identifies a sub-block contained in a block identified by I'_t. Equivalently all the indices in I_t are the $h-t$ indices inside $I_{t'}$, on the left. Hence

$$(I_r, \prec) \tag{6.2.6}$$

describes the hierarchical structure of the diagonal block matrices. This notation will be relevant later and will allow for a compact manipulation of hierarchic variables.

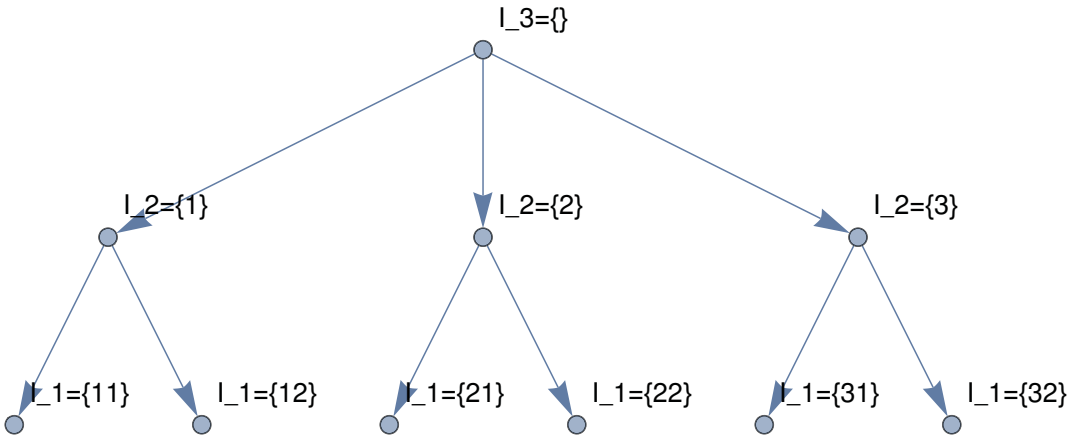


Figure 6.1: Example of tree, representing diagonal blocks hierarchy (6.2.6) at 2RSB. In the picture $m/k_2 = k_3/k_2 = 3$ and $k_2/k_1 = 2$. Each arrow $\rightarrow$ represents a $\succ$ relationship.

The first step is to solve the recursive equations for $\bar{C}(i)$. Clearly, it always holds that

$$\bar{C}(0) = \frac{1}{1-Q_1} > 0$$

The recursive equations are

$$\bar{C}(i) = -\frac{\sum_{j<i} k_j C(i) \bar{C}(j)}{\sum_{j \leq i} k_j C(j)} \tag{6.2.7}$$

and the solution is

$$\boxed{\bar{C}(i) = -\frac{C(i)}{\left[\sum_{j \leq i} k_j C(j)\right]\ \left[\sum_{j<i} k_j C(j)\right]} < 0} \tag{6.2.8}$$

As in the 1RSB computation, instead of negative $\bar{C}(i)$, we will use positive constants multiplied by a minus.

The next step is to compute $\ln A$. An auxiliary variable will be needed for any sub-block in Q^{-1}.

$$\begin{aligned}
\ln A =& \frac{1}{\sqrt{(2\pi)^m \det G}} \int_{Q_+} \prod_I dx_I \ \exp\left(-\frac{\bar{C}(0)}{2} \sum_I x_I^2 + \frac{1}{2} \sum_{j=1}^{h+1} \bar{C}(j) \sum_{I_{h+1-j}} \sum_{I, I' \prec I_{h+1-j}} x_I x_{I'} \right) \\
=& \frac{1}{\sqrt{(2\pi)^m \det G}} \frac{1}{\sqrt{\prod_j [2\pi \bar{C}(j)]^{k_j}}} \int_{Q_+} \prod_{j=1}^{h+1} \prod_{I_{h+1-j}} du_{I_{h+1-j}} \prod_I dx_I \\
& \exp\left(-\frac{\bar{C}(0)}{2} \sum_I x_I^2 + \sum_{j=1}^{h+1} \sum_{I_{h+1-j}} u_{I_{h+1-j}} \sum_{I \prec I_{h+1-j}} x_I - \frac{1}{2} \sum_{j=1}^{h+1} \sum_{I_{h+1-j}} \frac{1}{\bar{C}(j)} u_{I_{h+1-j}}^2 \right) \\
=& \frac{1}{\sqrt{\bar{C}(0)^m \det G}} \frac{1}{\sqrt{\prod_j [2\pi \bar{C}(j)]^{k_j}}} \int \prod_{j=1}^{h+1} \prod_{I_{h+1-j}} du_{I_{h+1-j}} \\
& \prod_I \operatorname{erfc}\left(x - \frac{1}{\sqrt{\bar{C}(0)}} \sum_{j=1}^{h+1} \sum_{I_{h+1-j} \succ I} u_{I_{h+1-j}} \right) \\
& \exp\left(k_1 \sum_{i=1}^{h+1} \sum_{I_{h+1-i}} u_{I_{h+1-i}} \sum_{j>i}^{h+1} \sum_{I_{h+1-j} \prec I_{h+1-i}} u_{I_{h+1-j}} \right) \\
& \exp\left(-\frac{1}{2} \sum_{j=1}^{h+1} \sum_{I_{h+1-j}} \left[\frac{1}{\bar{C}(j)} - \frac{k_j}{\bar{C}(0)} \right] u_{I_{h+1-j}}^2 \right)
\end{aligned}$$

A small rearrangement leads to

$$\begin{aligned}
\ln A =& \frac{1}{\sqrt{C(0)^m \det G}} \frac{1}{\sqrt{\prod_j [2\pi \bar{C}(j)]^{k_j}}} \int du_0 \exp\left(-\frac{1}{2} \left[\frac{1}{\bar{C}(1)} - \frac{k_h}{\bar{C}(0)} \right] u_0^2 \right) \\
& \prod_{I_1} \int du_{I_1} \exp\left(-\frac{1}{2} \left[\frac{1}{\bar{C}(2)} - \frac{k_{h-1}}{\bar{C}(0)} \right] u_{I_1}^2 + k_1 u_{I_1} u_0 \right) \\
& \prod_{I_2 \prec I_1} \int du_{I_2} \exp\left(-\frac{1}{2} \left[\frac{1}{\bar{C}(3)} - \frac{k_{h-2}}{\bar{C}(0)} \right] u_{I_2}^2 + k_1 u_{I_2} \sum_{J \succ I_2} u_J \right) \\
& \ldots \\
& \prod_{I_h \prec I_{h-1}} \int du_{I_h} \exp\left(-\frac{1}{2} \left[\frac{1}{\bar{C}(h+1)} - \frac{k_1}{\bar{C}(0)} \right] u_{I_h}^2 + k_1 u_{I_h} \sum_{J \succ I_h} u_J \right) \\
& \left[\operatorname{erfc}\left(x - \frac{1}{\sqrt{\bar{C}(1)}} \sum_{j=1}^{n} u_{I_{h-j}} \right) \right]^{k_1}
\end{aligned}$$

The index notation can be simplified

$$\begin{aligned}\ln A =& \frac{1}{\sqrt{\bar{C}(0)^m \det G}} \frac{1}{\sqrt{\prod_j [2\pi \bar{C}(j)]^{k_j}}} \int du_0 \exp\left(-\frac{1}{2}\left[\frac{1}{\bar{C}(1)} - \frac{k_h}{\bar{C}(0)}\right] u_0\right) \\ & \left[\int du_1 \exp\left(-\frac{1}{2}\left[\frac{1}{\bar{C}(2)} - \frac{k_{h-1}}{\bar{C}(0)}\right] u_1^2 + k_1 u_1 u\right)\right. \\ & \left[\int du_2 \exp\left(-\frac{1}{2}\left[\frac{1}{\bar{C}(2)} - \frac{k_{h-2}}{\bar{C}(1)}\right] u_2^2 + k_1 u_2 (u + u_1)\right)\right. \\ & ... \\ & \left[\int du_i \exp\left(-\frac{1}{2}\left[\frac{1}{\bar{C}(i)} - \frac{k_{h-1}}{\bar{C}(0)}\right] u_i^2 + k_1 u_i \sum_{j<i} u_j\right)\right. \\ & ... \\ & \left[\int du_h \exp\left(-\frac{1}{2}\left[\frac{1}{\bar{C}(h)} - \frac{k_1}{\bar{C}(1)}\right] u_h^2 + k_1 u_h \sum_{j<h} u_j\right)\right. \\ & \left.\left.\left[\operatorname{erfc}\left(x - \frac{1}{\sqrt{\bar{C}(1)}} \sum_{j=1}^{n} u_j\right)\right]^{k_1}\right]^{n_1} ...\right]^{n_{h-1}}\Bigg]^{n_h}\end{aligned}$$

This formula could be further manipulated. However, the goal was just to show that this procedure is possible. Hence, in this thesis we will just add that we speculte that, at the end of this procedure, with proper recursive translations, we could get:

$$\begin{aligned}\ln A =& \frac{1}{\sqrt{\bar{C}(0)^m \det G}} \frac{\sqrt{\prod_j \alpha^{(j)}(h+1-j)}}{\sqrt{\prod_j [2\pi \bar{C}(j)]^{k_j}}} \int du_0 e^{-\frac{1}{2}u_0^2} \left[\int du_1\, e^{-\frac{1}{2}u_1^2} \left[\int du_2\, e^{-\frac{1}{2}u_2^2} \dots\right.\right. \\ & \left.\left.\left[\int du_h\, e^{-\frac{1}{2}u_h^2} \left[\operatorname{erfc}\left(x - \frac{1}{\sqrt{\bar{C}(1)}} \sum_{j=1}^{n} \gamma_j u_j\right)\right]^{k_1}\right]^{n_1} ...\right]^{n_{h-1}}\right]^{n_h}\end{aligned}$$

Chapter 7

Appendix

7.0.4 Imaginary translation of a Gaussian variable

$$I_0 = \int_{\mathbb{R}} dx\ e^{-x^2/2} = \int_{\mathbb{R}+w} dx\ e^{-x^2/2} = I_w$$

with $w \in \mathbb{C}$, since

$$I_w = \int_{\mathbb{R}+w} dx\ e^{-x^2/2} = \int_{\mathbb{R}} dx\ e^{-(x-w)^2/2}$$

$\mathbb{R}$ and $\mathbb{R}+w$ can be seen as the limit case of the horizontal sides of the rectangle in $\mathbb{C}$

$$\mathcal{R} = \{(a,b) \in \mathbb{C} :\ a \in [-K, K]\ \ b \in [0, \Im(w)]\}$$

The integral on the path described by $\partial\mathcal{R}$ receives no contribution by the vertical sides, since

$$\int_{\text{v.s.}} |e^{-(x-w)^2}| \leq |\Im(w)| e^{-K^2/2 + \text{const}\, K} |e^{-w^2/2}| \underset{K\to\infty}{\longrightarrow} 0$$

Hence

$$0 = \int_{\partial R} e^{-x^2/2} = I_0 - I_w$$

for e^{-x^2} is an integer function.

7.0.5 Distribution of overlaps of random patterns

Given p binary patterns of size N, the following calculation yields the fraction $r(q)$ of overlaps which equal to q.

$$\begin{aligned} r(q) =& \frac{1}{2^{pN}} \sum_{\{\xi_\alpha\}} \frac{1}{p(p-1)/2} \sum_{\alpha<\beta} \delta\left(q - \frac{\xi_\alpha \cdot \xi_\beta}{N}\right) \\ =& \frac{1}{2^{pN}} \sum_{\{\xi_\alpha\}} \frac{1}{p(p-1)/2} \sum_{\alpha<\beta} \int \frac{dx_{\alpha\beta}}{2\pi}\ \exp\left(i x_{\alpha\beta}\left(q - \frac{\xi_\alpha \cdot \xi_\beta}{N}\right)\right) \end{aligned}$$

The order of the sums $\sum_{\{\xi_\alpha\}}$ and $\sum_{\alpha<\beta}$ can be inverted

$$\begin{aligned} r(q) =& \frac{1}{2^{pN}} \frac{1}{p(p-1)/2} \sum_{\alpha<\beta} \int \frac{dx_{\alpha\beta}}{2\pi}\ \exp\left(i x_{\alpha\beta} q\right) \sum_{\{\xi_\alpha\}} \exp\left(-i x_{\alpha\beta} \frac{\xi_\alpha \cdot \xi_\beta}{N}\right) \\ =& \frac{1}{2^{2N}} \frac{1}{p(p-1)/2} \sum_{\alpha<\beta} \int \frac{dx_{\alpha\beta}}{2\pi}\ \exp\left(i x_{\alpha\beta} q\right) \sum_{\{\xi_\alpha\},\{\xi_\beta\}} \exp\left(-i x_{\alpha\beta} \frac{\xi_\alpha \cdot \xi_\beta}{N}\right) \end{aligned}$$

All greek indeces can be dropped

$$\begin{aligned} r(q) =& \frac{1}{2^{2N}} \int \frac{dx}{2\pi} \exp(ixq) \sum_{\{\xi\},\{\bar{\xi}\}} \exp\left(-ix\frac{\bar{\xi}\cdot\xi}{N}\right) \\ &= \frac{1}{2^{2N}} \int \frac{dx}{2\pi} \exp(ixq) \left[\sum_{\xi=\pm 1}\sum_{\bar{\xi}=\pm 1} \exp\left(-ix\xi\bar{\xi}/N\right)\right]^N \\ &= \int \frac{dx}{2\pi} \exp(ixq) \left[\cos(x/N)\right]^N \end{aligned}$$

If $N \to \infty$ then, by using the usual expansion $\ln\cos(\epsilon) \sim -\frac{1}{2}\epsilon^2$, one gets

$$r(q) \sim \sqrt{\frac{N}{2\pi}} \exp\left(-\frac{N}{2q^2}\right)$$

Notably, the result does not depend on p.

The conclusion that can be drawn is that, in absence of constraints, the fraction of pairs whose overlap is above any finite $q > 0$ is neglectable, in the thermodynamical limit.

7.0.6 Important limits

In these sections, all limits which are needed for the computation of the correlation dependent capacity will be studied. Consider $\epsilon = \sqrt{1-Q}$ and, for simplicity $\mathrm{erfc}(x) = \frac{1}{\sqrt{\pi}} \int_x dy\ \exp(-y^2)$

- $$I^-(a,b) = \lim_{\epsilon\to 0^+} \epsilon^2 \ln \int dx\ e^{-x^2}\ \Theta(x - a/\epsilon)\ \mathrm{erfc}(b/\epsilon - x) \tag{7.0.1}$$

  Rescale x

  $$I^-(a,b) = \lim_{\epsilon\to 0^+} \epsilon^2 \ln \frac{1}{\epsilon} \int dx\ e^{-x^2/\epsilon^2}\ \Theta(x-a)\ \mathrm{erfc}((b-x)/\epsilon)$$

  The following can be evaluated with the saddle point method.

  Since if $x < b$

  $$\mathrm{erfc}((b-x)/\epsilon) = \exp\left(-(b-x)^2/\epsilon^2\right)\left[\epsilon/(b-x) + o(\epsilon)\right]$$

  then

  $$\ln\mathrm{erfc}((b-x)/\epsilon) + (b-x)^2/\epsilon^2 = o(\epsilon)$$

  Conversely, if $x > b$

  $$\ln\mathrm{erfc}((b-x)/\epsilon) = 1 + o(\epsilon)$$

  Therefore

  $$\begin{aligned} I^-(a,b) =& \lim_{\epsilon\to 0^+} \epsilon^2 \ln \frac{1}{\epsilon} \int_a^b dx\ e^{-\frac{1}{\epsilon^2}\left[x^2+(x-b)^2\right]}\ e^{\ln\mathrm{erfc}((b-x)/\epsilon)+(x-b)^2/\epsilon^2} \\ &+ \frac{1}{\epsilon}\int_b^{+\infty} dx\ e^{-\frac{1}{\epsilon^2}x^2}\ e^{\ln\mathrm{erfc}((b-x)/\epsilon)} \\ =& \lim_{\epsilon\to 0^+} \epsilon^2 \ln \frac{1}{\epsilon} \int_a^b dx\ e^{-\frac{1}{\epsilon^2}\left[x^2+(x-b)^2\right]} + \frac{1}{\epsilon}\int_b^{+\infty} dx\ e^{-\frac{1}{\epsilon^2}x^2} \\ =& \lim_{\epsilon\to 0^+} \epsilon^2 \ln \frac{1}{\epsilon} \int_a^{\infty} dx\ \exp\left(-\frac{1}{\epsilon^2}\left[\left(x^2+(x-b)^2\right)\ \Theta(b-x) + x^2\ \Theta(x-b)\right]\right) \end{aligned}$$

By studying the function

$$L_b^-(x) = \left[(x-b)^2\ \Theta(b-x) + x^2\right] \tag{7.0.2}$$

at the exponent of the integrand, it can be deduced that

$$I^-(a,b) = -L_b(\max[a,b/2])\Theta(b) - L_b(\max[a,0])\Theta(-b) \tag{7.0.3}$$

- $$I^+(a,b) = \lim_{\epsilon\to 0^+} \epsilon^2 \ln \int dx\ e^{-x^2}\ \Theta(x-a/\epsilon)\ \mathrm{erfc}(x-b/\epsilon) \tag{7.0.4}$$

  This limit can be evaluated in the same way as I^-. The result is

$$I^+(a,b) = -L_b^+(\max[a,0])\Theta(b) - L_b^+(\max[a,b/2])\Theta(-b) \tag{7.0.5}$$

  with

$$L_b^+(x) = \left[x^2 + (x-b)^2\ \Theta(x-b)\right] = L_{-b}^-(-x) \tag{7.0.6}$$

- Consider $x > 0$ and

$$M_b(x) = \lim_{\epsilon\to 0} \epsilon^2 \ln[\mathrm{erfc}((-x+b)/\epsilon) - \mathrm{erfc}(\mathrm{erfc}((x+b)/\epsilon))] \tag{7.0.7}$$

  Since erfc is monotonically decreasing and $x > 0$, it follows that the argument of the logarithm is well defined for every finite ϵ.

  Assume $x > 0$ $x < b$, then

$$\begin{aligned}
&\mathrm{erfc}((-x+b)/\epsilon) - \mathrm{erfc}(\mathrm{erfc}((x+b)/\epsilon)) \\
&\qquad = \frac{\epsilon}{-x+b} e^{-(x-b)^2/\epsilon^2}(1+o(\epsilon)) - \frac{\epsilon}{x+b} e^{-(x+b)^2/\epsilon^2}(1+o(\epsilon)) \\
&\qquad = \frac{\epsilon}{-x+b} e^{-(x-b)^2/\epsilon^2}\left[(1+o(\epsilon))\left(1 - \frac{-x+b}{x+b}\right) e^{-4bx/\epsilon^2}\right] \\
&\qquad = \frac{\epsilon}{-x+b} e^{-(x-b)^2/\epsilon^2}(1+o(\epsilon)) \\
&\qquad = \mathrm{erfc}(-(x+b)/\epsilon)(1+(\epsilon))
\end{aligned}$$

  Now assume $x > b$. It can easily shown that

$$\mathrm{erfc}((-x+b)/\epsilon) - \mathrm{erfc}(\mathrm{erfc}((x+b)/\epsilon)) = 1 + o(\epsilon)$$

  Hence, if $b > 0$

$$M_b(x) = \Theta(b-x)\ (x-b)^2$$

  Now assume $b < 0$. The property

$$\mathrm{erfc}(x) = 1 - \mathrm{erfc}(-x)$$

  implies that

$$\mathrm{erfc}((-x+b)/\epsilon) - \mathrm{erfc}(\mathrm{erfc}((x+b)/\epsilon)) = \mathrm{erfc}((-x-b)/\epsilon) - \mathrm{erfc}(\mathrm{erfc}((x-b)/\epsilon))$$

  Therefore, if $b < 0$, then

$$M_b(x) = \Theta(-b-x)\ (x+b)^2$$

  If the two previous cases are put together, it follows that

$$M_b(x) = \Theta(b)\ \Theta(b-x)\ (x-b)^2 + \Theta(-b)\ \Theta(-b-x)\ (x+b)^2 \tag{7.0.8}$$

- It follows from the previous points that

$$I(a,b) = \lim_{\epsilon\to 0^+} \epsilon^2 \ln \int_0^\infty dx\; e^{-(x+a)^2/\epsilon^2} \left[\operatorname{erfc}\left(\frac{-x+b}{\epsilon}\right) - \operatorname{erfc}\left(\frac{x+b}{\epsilon}\right)\right]$$
$$= \arg\min_{x>0} \left[(x+a)^2 + M_b(x)\right]$$

Let us call

$$M^{\pm}_{a,b}(x) = (x+a)^2 + M_b(x)\;\Theta(\pm b) = \Theta(\pm b - x)\;(x \mp b)^2 + (x+a)^2 \tag{7.0.9}$$

Then, after studying the minima of M, one can conclude that

$$\begin{aligned} I(a,b) =& M^+_{a,b}(\max[(b-a)/2, 0])\;\Theta(b+a)\;\Theta(b) \\ &+ M^-_{a,b}(\max[(-a-b)/2, 0])\;\Theta(a-b)\;\Theta(-b) \\ =& M^+_{a,b}((b-a)/2)\;\Theta(b+a)\;\Theta(b)\;\Theta(-a+b) \\ &+ M^+_{a,b}(0)\;\Theta(b+a)\;\Theta(b)\;\Theta(b-a) \\ &+ M^-_{a,b}((-b-a)/2)\;\Theta(a-b)\;\Theta(-b)\;\Theta(-a-b) \\ &+ M^-_{a,b}(0)\;\Theta(a-b)\;\Theta(-b)\;\Theta(a+b) \end{aligned} \tag{7.0.10}$$

7.0.7 Further details about A in the RS ansatz

In this section, as an integration, the eigenvectors of the quadratic form appearing in A will be briefly discussed. The knowledge of their structure could be functional to a geometric approach to the probem, which is not used in this thesis.

Let us call the vector space on which

$$G^{-1} = \begin{bmatrix} A\mathbb{I}_m - B\;1_m & -q[A\mathbb{I}_m - B\;1_m] \\ -q[A\mathbb{I}_m - B\;1_m] & A\mathbb{I}_m - B\;1_m \end{bmatrix}$$

is defined as $U = U_+ \oplus U_-$ with $\dim U_\pm = m$. Let $\{u^\pm_j\}$ be an orthonormal basis for $V_\pm$ so that $[u^-_j]_a = \delta_{ja}$ and $[u^+_j]_a = \delta_{j(m+a)}$. Then, eigenvalues and eigenvectors of G^{-1} are

$$(1-q)(A-mB) \longleftrightarrow v^+ = \frac{1}{\sqrt{2m}} \sum_{j=1}^m (u^+_j + u^-_j)$$

$$(1+q)(A-mB) \longleftrightarrow v^- = \frac{1}{\sqrt{2m}} \sum_{j=1}^m (u^+_j - u^-_j)$$

$$(1-q)A \longleftrightarrow V_+ = \left\{ v = \sum_{j=1}^m a_j (u^+_j + u^-_j) :\; \sum_{j=1}^m a_j = 0 \right\}$$

$$(1+q)A \longleftrightarrow V_- = \left\{ v = \sum_{j=1}^m a_j (u^+_j - u^-_j) :\; \sum_{j=1}^m a_j = 0 \right\}$$

An orthonormal basis for $V_\pm$ is

$$\left\{ v^\pm_a = \frac{1}{\sqrt{a^2+a}} \left[a(u^+_{a+1} \pm u^-_{a+1}) - \sum_{j=1}^a (u^+_j \pm u^-_j) \right] \right\}_{a=1,\dots,m-1}$$

Then, with this basis

$$u_a^{\pm} = \frac{1}{\sqrt{2m}}(v^+ \pm v^-) + \frac{(a-1)}{\sqrt{(a-1)^2+(a-1)}}(v_a^+ \pm v_a^-) - \sum_{j=a}^{m-1} \frac{1}{\sqrt{j^2+j}}(v_j^+ \pm v_j^-)$$

7.0.8 Correlation between elements of two patterns with fixed overlap

The purpose of this paragraph is to prove that, given two random patterns ξ and $\bar{\xi}$ with a fixed overlap q, then, in the thermodynamical limit

$$\langle \xi_i \bar{\xi}_j \rangle = q\delta_{ij} \tag{7.0.11}$$

Proof

$$\begin{aligned}
Z = \frac{1}{2^{2N}} \sum_{\xi,\bar{\xi}} \delta(\xi \cdot \bar{\xi}/N - q) &= \frac{1}{2^{2N}} \sum_{\xi,\bar{\xi}} \int dx \ \exp[-ix(\xi \cdot \bar{\xi}/N - q)] \\
&= \frac{1}{2^{2N}} \int dx \ e^{ixq} \prod_{j=1} \sum_{\xi_i=\pm 1} \sum_{\bar{\xi}_i=\pm 1}^{N} e^{-ix\xi_i\bar{\xi}_i} \\
&= \int dx \ e^{ixq} \cos^N(x/N) \\
&= \int dx \ e^{ixq + N \ln\cos(x/N)} \\
&\approx \int dx \ e^{ixq - \frac{1}{2N}x^2} \\
&= \sqrt{\frac{2\pi}{N}} e^{-\frac{1}{2}Nq^2}
\end{aligned}$$

Then, the correlation becomes

$$\langle \xi_i \bar{\xi}_j \rangle = \frac{1}{Z 2^{2N}} \sum_{\xi,\bar{\xi}} \delta(\xi \cdot \bar{\xi}/N - q) \xi_i \bar{\xi}_j = \frac{1}{Z 2^{2N}} \sum_{\xi,\bar{\xi}} \int dx \ \exp[-ix(\xi \cdot \bar{\xi}/N - q)] \xi_i \bar{\xi}_j$$

Let $i = j$

$$\begin{aligned}
\langle \xi_j \bar{\xi}_j \rangle &= \frac{1}{Z 2^{2N}} \int dx \ e^{ixq} \left[\prod_{k \neq j} \sum_{\xi_k=\pm 1} \sum_{\bar{\xi}_k=\pm 1}^{N} e^{-ix\xi_k\bar{\xi}_k} \right] \left[\sum_{\xi_j=\pm 1} \sum_{\bar{\xi}_j=\pm 1}^{N} \bar{\xi}_j \xi_j e^{-ix\xi_j\bar{\xi}_j} \right] \\
&= i \frac{1}{Z} \int dx \ e^{ixq} \cos^N(x/N) \tan(x/N) \\
&\approx -\frac{1}{NZ} \int dx \ e^{ixq - \frac{1}{2N}x^2} x \\
&= -\frac{1}{NZ} \frac{d}{dq} \int dx \ e^{ixq - \frac{1}{2N}x^2} = q
\end{aligned}$$

Let $i \neq j$

$$\begin{aligned}
\langle \xi_i \bar{\xi}_j \rangle = \frac{1}{Z 2^{2N}} \int dx \ e^{ixq} &\left[\prod_{k \neq i,j} \sum_{\xi_k=\pm 1} \sum_{\bar{\xi}_k=\pm 1}^{N} e^{-ix\xi_k\bar{\xi}_k} \right] \left[\sum_{\xi_j=\pm 1} \sum_{\bar{\xi}_j=\pm 1}^{N} \bar{\xi}_j e^{-ix\xi_j\bar{\xi}_j} \right] \\
&\left[\sum_{\xi_i=\pm 1} \sum_{\bar{\xi}_i=\pm 1}^{N} \xi_i e^{-ix\xi_i\bar{\xi}_i} \right]
\end{aligned}$$

But

$$\sum_{\xi_i=\pm 1}\sum_{\bar{\xi}_i=\pm 1}^{N}\xi_i e^{-ix\xi_i\bar{\xi}_i}=0$$

Hence

$$\langle \xi_i\bar{\xi}_j\rangle = 0$$

7.0.9 Introduction of Gaussian variables for the solution of the SP equations.

The purpose of this section is to introduce a Gaussian representation for equation (5.1.2).

$$\mathcal{A}(Q)=\int P_q(\xi,\bar{\xi})\prod_{a,b=1}^{m}\prod_{\mu}\Theta\left(\frac{W^a\cdot\xi^\mu}{\sqrt{N}}\right)\Theta\left(\frac{W^b\cdot\bar{\xi}^\mu}{\sqrt{N}}\right)$$

with

$$P_q(\xi,\bar{\xi})=\frac{1}{Z_q 2^N}\prod_{\alpha}\delta(q-\xi_\alpha\cdot\bar{\xi}_\alpha/N)$$

$$Z_q=\frac{1}{2^N}\sum_{\{\xi\}}\sum_{\{\bar{\xi}\}}\prod_{\alpha}\delta(q-\xi_\alpha\cdot\bar{\xi}_\alpha/N)$$

Since the patterns are assumed to be correlated pairwise, it is enough to consider (notation warning: here ξ and $\bar{\xi}$ are patterns, not sets)

$$P_q(\xi,\bar{\xi})=\frac{1}{Z_q 2^{2N}}\delta(q-\xi\cdot\bar{\xi}/N)$$

Then, let us consider the RVs

$$w_a=W_a\cdot\xi/\sqrt{N}$$

$$\bar{w}_a=W_a\cdot\bar{\xi}/\sqrt{N}$$

Then, by (7.0.11)

$$\langle w_a w_b\rangle=\langle\bar{w}_a\bar{w}_b\rangle=\frac{W_a\cdot W_b}{N}=Q_{ab}$$

$$\langle w_a\bar{w}_b\rangle=q\frac{W_a\cdot W_b}{N}=qQ_{ab}$$

The previous observation suggests that (5.1.2) can be rewritten as (5.1.3). In order to proof that, the joint distribution of the RVs

$$w_a=\xi\cdot W_a/\sqrt{N} \tag{7.0.12}$$

is needed. The prefactor $1/\sqrt{N}$ is arbitrary since the Θ functions are invariant under

$$w_a\mapsto x\,w_a$$

This choice though is convenient.

Then

$$
\begin{aligned}
Z_q\, P(\{w_a\},\{\bar w_a\}) &= \frac{1}{2^{2N}} \sum_{\{\xi\}\{\bar\xi\}} \delta(q-\xi\cdot\bar\xi/N) \prod_{a=1}^{m} \delta(w_a-\xi\cdot W_a/\sqrt{N})\; \delta(\bar w_a-\bar\xi\cdot W_a/\sqrt{N}) \\
&= \frac{1}{2^{2N}} \sum_{\{\xi\}\{\bar\xi\}} \int dy \prod_{a=1}^{m} dx_a\; d\bar x_a\; e^{-iy(q-\xi\cdot\bar\xi/N)} \prod_{a=1}^{m} e^{-ix_a[w_a-\xi\cdot W_a/\sqrt{N}]}\; e^{-i\bar x_a[\bar w_a-\bar\xi\cdot W_a/\sqrt{N}]} \\
&= \frac{1}{2^{2N}} \int dy \prod_{a=1}^{m} dx_a\; d\bar x_a\; \exp\left[-i(yq+\sum_a [x_a w_a+\bar x_a \bar w_a])\right] \\
&\qquad \prod_{j=1}^{N} \sum_{\bar\xi^j=\pm 1} \sum_{\xi^j=\pm 1} \exp\left[i\sum_a \frac{x_a\xi^j W_a^j + \bar x_a \bar\xi^j W_a^j}{\sqrt{N}} + iy\frac{\xi^j\bar\xi^j}{N}\right] \\
&= \int dy \prod_{a=1}^{m} dx_a\; d\bar x_a\; \exp\left[-i(yq+\sum_a [x_a w_a+\bar x_a \bar w_a])\right] \\
&\qquad \prod_{j=1}^{N} \left[\cos\left(\sum_a \frac{x_a W_a^j + \bar x_a W_a^j}{\sqrt{N}}\right) e^{iy/N} + \cos\left(\sum_a \frac{x_a W_a^j - \bar x_a W_a^j}{\sqrt{N}}\right) e^{-iy/N}\right] \\
&\approx \int dy \prod_{a=1}^{m} dx_a\; d\bar x_a\; \exp\left[-i(yq+\sum_a [x_a w_a+\bar x_a \bar w_a])\right] \\
&\qquad \prod_{j=1}^{m} \exp\left(-\frac{1}{2N}\sum_{a,b=1}^{m} (x_a W_a^j W_b^j x_b + \bar x_a W_a^j W_b^j \bar x_b)\right) \\
&\qquad \cosh\left(-\frac{1}{N}\sum_{a=1}^{m} x_a W_a^j W_b^j \bar x_b + iy/N\right) \\
&\approx \int dy \prod_{a=1}^{m} dx_a\; d\bar x_a\; \exp\left[-i(yq+\sum_a [x_a w_a+\bar x_a \bar w_a]) - \frac{1}{2}y^2\right] \\
&\qquad \exp\left(-\frac{1}{2N}\sum_{a,b=1}^{m} (x_a W_a\cdot W_b x_b + \bar x_a W_a\cdot W_b \bar x_b + 2iy\; \bar x_a W_a\cdot W_b x_b)\right) \\
&\approx e^{-\frac{1}{2}q^2} \int \prod_{a=1}^{m} dx_a\; d\bar x_a\; \exp\left[-i\sum_a [x_a w_a+\bar x_a \bar w_a]\right] \\
&\qquad \exp\left(-\frac{1}{2N}\sum_{a,b=1}^{m} (x_a W_a\cdot W_b x_b + \bar x_a W_a\cdot W_b \bar x_b + q\; \bar x_a W_a\cdot W_b x_b)\right) \\
&= e^{-\frac{1}{2}q^2} \frac{1}{(2\pi)^m\sqrt{\det G}} \exp\left(-\frac{1}{2}\sum_{a,b}\sum_{\alpha\beta} w_a^\alpha [G^{-1}]_{ab}^{\alpha\beta} w_b^\beta\right)
\end{aligned}
$$

Dividing by Z_q, one gets (5.1.3).

Acknowledgements

First, i would like to thank my supervisor, professor Sergio Caracciolo and my assistant supervisors Marco Gherardi and Pietro Rotondo for their precious help. I would also like to mention Marco Cosentino Lagomarsino, Andrea di Gioacchino and Enrico Malatesta for the support and the advices. My gratitude also goes to professor Bruno Bassetti, whose lessons in statistical mechanics inspired me to choose this field of study.

Second, a special thanks goes to my good friends Giulio Amato and Matteo Bertoli, who have always been supportive throughout these years. Finally, I want to thank my parents, Massimo and Flores, who have always encouraged and supported me.

Bibliography

[1] C. Baldassi, A. Ingrosso, C. Lucibello, L. Saglietti, R. Zecchina, *Subdominant Dense Clusters Allow for Simple Learning and High Computational Performance in Neural Networks with Discrete Synapses*, Phys. Rev. Lett. 115, 128101, 2015

[2] C. Baldassi, C. Borgs, J. Chayes, A. Ingrosso, C. Lucibello, L. Saglietti, R. Zecchina, *Unreasonable Effectiveness of Learning Neural Networks: From Accessible States and Robust Ensembles to Basic Algorithmic Schemes*, Proc. Natl. Acad. Sci. U.S.A. 113(48):E7655-E7662, 2016

[3] N. Brunel, J. Nadal, *Information capacity of a perceptron*,Journal of Physics A General Physics January 1999

[4] T. Castellani, A. Cavagna, *Spin-Glass Theory for Pedestrians*,J. Stat. Mech. (2005) P05012

[5] E. T. Copson, *Asymptotic Expansions*, Cambridge University Press, Cambridge 2004 2000, Vol. 28, No. 2, 725758

[6] P. Del Giudice, S. Franz, M. A. Virasoro, *Perceptron beyond the limit of capacity*,Journal de Physique, 1989, 50 (2), pp.121-134

[7] B. Derrida, R. B. Griffiths and A. Prgel-Bennett, *Finite-size effects and bounds for perceptron models*,J. Phys. A: Math. Gen., 24, 4907, 1991.

[8] R. Erichsen and W. K. Thuemann, *Optimal storage of a neural network model: a replica symmetry-breaking solution*,J. Phys. A: Math. Gen. 26, 61, 1993

[9] F Fontanari and W K Thenmann , *On the computational capability of a perceptron* ,J Phys. A: Math. Gen. 26 (1993) L1233-Ll238.

[10] F Fontanari , *Equilibrium properties of the linear perceptron* ,i 1993 J. Phys. A: Math. Gen. 26 6147

[11] E. Gardner, B. Derrida, *Optimal storage properties of neural network models* , J. Phys. A: Math. Gen. 21 (1988) 271-284.

[12] G. Gavallotti, *Statistical mechanics : a short treatise* , Springer, Berlin 1999.

[13] R. Gibbons, R. D. Bock, D. R. Hedeker, *Approximating Multivariate Normal Orthant Probabilities*, University of Illinois at Chicago (Paperback), 1990

[14] S. Franz, G. Parisi, *Recipes for metastable states in Spin Glasses* , Journal de Physique I, EDP Sciences, 1995, 5 (11), pp.1401-1415.

[15] J. A. Hertz et al, *Phase transitions in simple learning* , J. Phys. A: Math. Gen. 22 2133., 1989

[16] H. Huang, Y. Kabashima, *Origin of the computational hardness for learning with binary synapses*, Phys. Rev. E 90, 052813, 2014

[17] Y. Kabashima, *Inference from correlated patterns: a unified theory for perceptron learning and linear vector channels* , J. Phys.: Conf. Ser. 95 012001, 2008

[18] O. Kinouchi, N. Caticha, *Qn-line versus off-line learning in the linear perceptron: A comparative study* , Physical Review E.52, 3,1995

[19] O. Kinouchi, N. Caticha, *Learning algorithm that gives the Bayes generalization limit for perceptron* , Pyss. Rev. E, 54, 1, 1996

[20] J. M. Kosterlitz, D. J. Thouless, and Raymund C. Jones, *Spherical Model of a Spin-Glass* , Phys. Rev. Lett. 36, 1217, 1976

[21] W. Krauth, M. Mézard, *Storage capacity of memory networks with binary couplings* , Journal de Physique, 1989, 50 (20), pp.3057-3066

[22] M. Mézard, G. Parisi, N. Sourlas, G. Toulouse, M. Virasoro, *Replica symmetry breaking and the nature of the spin glass phase*, ournal de Physique, 1984, 45 (5), pp.843-854.

[23] M. Mézard, Giorgio Parisi, M. Virasoro, *Spin Glass Theory and Beyond*, World Scientific, Singapore 1986

[24] F. Morone, F. Caltagirone, E. Harrison, G. Parisi, *Replica Theory and Spin Glasses*, 2014arXiv1409.2722M, 2014

[25] H. Nishimori, *Statistical physics of spin glasses and information processing: an Introduction*, Oxford University press, Oxford 2001

[26] M. Opper M, D. Haussler, *Generalization performance of Bayes optimal classification algorithm for learning a perceptron.* , Phys. Rev. Lett. 66 (20) 2677-2680.

[27] J. Sethna , *Entropy, Order Parameters, and Complexity*, Oxford University press, Oxford 2006

[28] T. Shinzato and Y. Kabashima, *Learning from correlated patterns by simple perceptrons*, J. Phys. A: Math. Theor. 42 015005, 2009

[29] G. P. Steck, *Orthant Probabilities for the Equicorrelated Multivariate Normal Distribution* , Biometrika, Vol. 49, No. 3/4 (Dec., 1962), pp. 433-445

[30] M. Talagrand, *Intersecting random half-planes: toward the Gardner-Derrida formula*, The Annals of Probability 2000, Vol. 28, No. 2, 725758

[31] W. K. Theumann and R. Erichsen Jr., *Gardner-Derrida neural networks with correlated patterns*, J. Phys. A: Math. Gen. 24, 565, 1991

[32] J. Zinn-Justin., *Quantum field theory and critical phenomena*, Clarendon Press, Oxford, 2002