LIMBURGS UNIVERSITAIR CENTRUM

Faculteit Wetenschappen

Towards Bayesian learning for the perceptron

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

Jan SCHIETSE

Promotor
Prof.dr.M.Bouten

1996

D. BOLLÉ (KULEUVEN)
A. ENGEL (MAGDEBURG)
~~A. KOMODA (NAGASAKI)~~
R. SEANGEES (LUC)
C. VAN DEN BROECK (LUC)
C. VAN DER VANDE (LUC)
M. BOUTEN (Nijmegen) LUC
J. MULLENS (Voorwitte) LUC

536.9

LIMBURGS UNIVERSITAIR CENTRUM
Faculteit Wetenschappen

Towards Bayesian learning for the perceptron

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

Jan SCHIETSE

Promotor
Prof.dr.M.Bouten

1996

Dankwoord

Eerst en vooral wil ik Marc Bouten danken voor zijn optimale begeleiding en steun.
Verder hebben alle leden van de groep theoretische natuurkunde (Roger Serneels, Chris Van den Broeck, Carlo Vanderzande, Geert Jan Bex, Eddy Lootens, Frank Daerden, Peter Heirman, Bart Van Rompaey) bijgedragen tot dit doctoraatswerk. Een speciale vermelding verdient Geert Jan Bex, die er steeds voor zorgde dat de computers deden wat ik wou en niet omgekeerd. De buitenlandse gasten (Jacek Iwanski, Peter Reimann, Lüder Reimers,...) wil ik danken voor de interessante discussies. Mijn dank gaat ook uit naar de dames van het secretariaat WNI, die altijd bereid waren om me te helpen bij administratieve taken.

Voor uiteenlopende redenen wil ik verder nog de volgende mensen vermelden : Horzak, Orwad, Torgal en de andere Astorianen, Luc V., Kristien B., Bert S., Martine V., Else D., Kristien M., Robert D., Denise V., Mike D., Brett E. .

Tenslotte wil ik Edith, Denis en Geert Schietse bedanken.

Contents

Dankwoord i

1 An introduction to learning with artificial neural nets. 1
- 1.1 Modelling the brain. . . . 1
- 1.2 The teacher-student scenario . . . 5
- 1.3 Learning rules. . . . 8
  - 1.3.1 The Hebb rule. . . . 8
  - 1.3.2 The Adaline rule. . . . 9
  - 1.3.3 The Gibbs rule. . . . 10
  - 1.3.4 The Adatron rule. . . . 11
  - 1.3.5 The Bayes rule. . . . 12

2 A unified approach : Gradient descent learning. 15
- 2.1 General theory . . . 15
- 2.2 Recovering previous results . . . 19
  - 2.2.1 The Hebb potential . . . 19
  - 2.2.2 The Adaline potential . . . 21
  - 2.2.3 The Adatron potential . . . 24
- 2.3 A class of cost functions which favour large overlaps within the version space. . . . 30
  - 2.3.1 The Hebb rule within the version space. . . . 30
  - 2.3.2 Bounds for the generalization error associated to a monotonous potential. . . . 33
  - 2.3.3 A class of repulsive potentials within the version space. . . . 34
  - 2.3.4 Numerical results. . . . 38
  - 2.3.5 Asymptotic behaviour. . . . 38

2.4 A class of potentials which favour small overlaps within the version space 41
2.5 Overlaps between student vectors 47

3 Learning from non-uniformly distributed examples. 55
3.1 The gaussian model. 55
3.2 Gibbs and Bayes learning. 57
3.3 Calculation of the overlap with the teacher using gradient descent learning. 59
3.4 Hebb and maximal stability learning. 61
3.4.1 Hebb learning. 61
3.4.2 Maximal stability learning. 61
3.4.3 Numerical results. 62
3.5 Gradient descent learning with a repulsive potential. . . 64

4 The Ising teacher problem. 71
4.1 Introduction 71
4.2 Clipping. 74
4.3 Partial clipping. 81
4.4 Optimal transformation 83

A Replica calculation for a general cost function. 89

B Convexity of the cost function associated with the potential $V_s^+(\lambda)$. 97

C Large α behaviour for the generalisation error corresponding with the repulsive potential $V_s^+(\lambda)$. 99

D Replica calculation of the overlap S. 105

E Clipping : the replica calculation. 111

Nederlandstalige samenvatting 117

Bibliography 130

Publikatielijst 135

Chapter 1

An introduction to learning with artificial neural nets.

1.1 Modelling the brain.

One of the main reasons to mimic the brain by creating networks of artificial neurons (nerve cells) is the fact that it can only be outperformed by a computer in tasks based on simple arithmetic. Indeed, the human brain, which learns without any explicit instructions, interpretes imprecise information from the senses at an incredible rapid rate and it can deal with information that is fuzzy, noisy or inconsistent. A good example of the superiority of the brain is the processing of visual information : a baby which is one year old is much better and faster at recognizing objects, faces, and so on than even the most advanced artificial intelligence system running on the fastest computer.
Our brain (1.5-2kg) consists of 10^{11} neurons. In Fig.(1.1) we depicted a so called pyramidal neuron which play a role in memory functioning. One observes that tree like networks of nerve fiber called dendrites are connected to the cell body, where the cell nucleus is located. Extending from each cell body is a single long fiber called the axon, which eventually branches into strands and substrands. The axon of a typical neuron makes a few thousand synapses with other neurons.

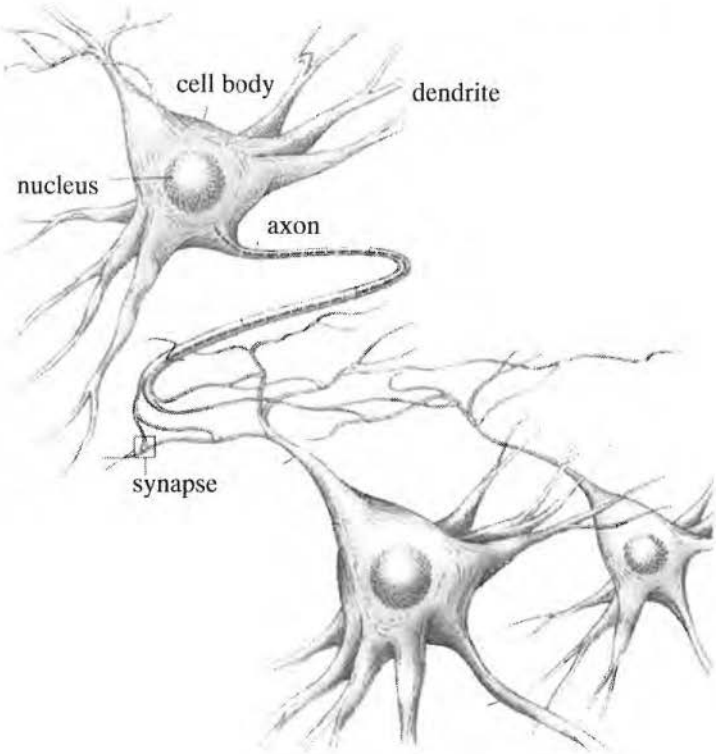

Fig.(1.1): A pyramidal neuron.

The transmission of signals from one nerve cell to another is a complex chemical process in which specific transmitter substances are released. This causes spikes of electrical activity in the axon and its thousands of branches. At the end of each branch, a structure called a synapse converts the activity of the axon into effects that raise or lower the electrical potential inside the body of the receiving cell.

When the potential of a neuron reaches a certain threshold, it sends a pulse of fixed strength down its axon and as a result to all connected nerve cells. Learning occurs by changing the effectiveness of the synapses so that the influence of one neuron on another changes.

Because our knowledge of neurons is incomplete and our computing power is limited, brain models are necessarily idealisations of real networks of neurons. Artificial neural networks are typically composed of interconnected units, which serve as model neurons. The synapse is modeled by a modifiable weight, which is associated with each connection. The electrical output of a neuron is represented by a single number which represents its activity. Each unit converts the incoming activities into a single outgoing signal.

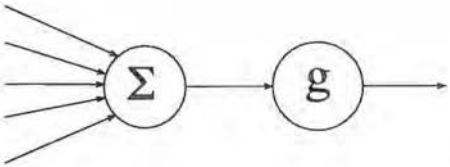

Fig.(1.2): Artificial neuron or unit which first sums all the incoming activities and then calculates the output using the transfer function g.

This conversion is performed in two stages. First all incoming activities are multiplied by the weight or strength of the corresponding synapse and then these weighted inputs are added together to get a quantity called the total input. Second, the unit uses a transfer function g to transform the total input into the outgoing activity. The function g can have various forms, but the most common have a linear, threshold or sigmoidal shape. For treshold units, the output is set at one of two levels, depending on whether the total output is greater than or less than some threshold value.

For sigmoid units, the output varies continuously but not linearly as the input changes. Sigmoid units bear a greater resemblance to real neurons than threshold units, but both must be considered rough approximations.

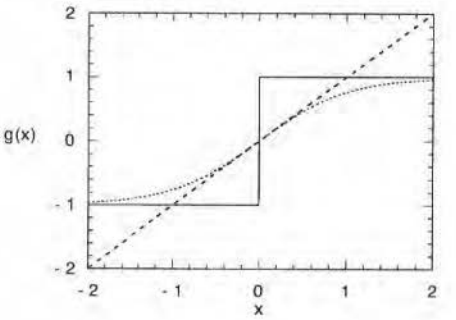

Fig.(1.3): Some of the transferfunctions used in artificial neural networks : $g(x) = \text{sign}(x)$ (full curve), $g(x) = \tanh(x)$ (dotted curve) and $g(x) = x$ (dashed curve).

The simplest neural network architecture one can consider, but the building block for more complicated nets is the perceptron. It consists of an input layer of N units and one output unit with feed-forward connections between them. The input-output relation can be written down in the following way :

$$S_0 = g\left(\sum_{i=1}^{N} J_i\, S_i\right), \tag{1.1}$$

where we have considered a neuron which recieves signals from N other neurons. The activity of the i-th neuron is characterized by S_i and is multiplied with the strength J_i of the corresponding synapse. The connections J_i can have continuous or binary values. For the transfer function one can consider one of the functions plotted in Fig. (1.3).

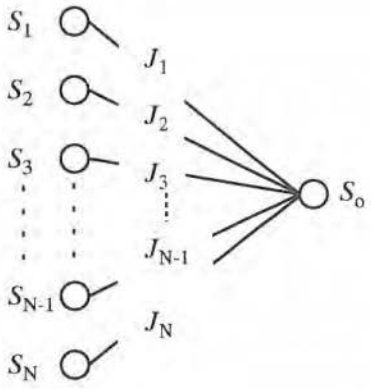

Fig.(1.4): A perceptron with N input units S_i connected with the output unit S_0 through the synapses with weights J_i.

1.2 The teacher-student scenario

To gain more insight into the mechanism of learning one can investigate a very simple scenario, namely that of a student perceptron learning from examples generated by a teacher perceptron. The perceptrons we will consider are characterised by an N-dimensional continuous weight-vector $\mathbf{J}$ which defines a map from the N-dimensional input space to a binary output according to:

$$\xi_0 = \text{sign}\left(\frac{\mathbf{J}.\boldsymbol{\xi}}{\sqrt{N}}\right). \tag{1.2}$$

The vectors $\mathbf{J}$ and $\boldsymbol{\xi}$ are normalised as $\mathbf{J}^2 = \boldsymbol{\xi}^2 = N$.
A teacher perceptron, characterized by the weight vector $\mathbf{T}$ returns the classification

$$\xi_0^\mu = \text{sign}\left(\frac{\mathbf{T}.\boldsymbol{\xi}^\mu}{\sqrt{N}}\right) \quad \forall \mu = 1, \ldots, P, \tag{1.3}$$

on a set of P randomly selected training patterns $\boldsymbol{\xi}^{\mu}$. On the basis of this information one likes to select a student perceptron $\mathbf{J}$ such that it reproduces as closely as possible the classification of the teacher.
The performance of a student perceptron is usually quantified by the generalization error $\varepsilon(\mathbf{J})$ defined as the probability that $\mathbf{J}$ and $\mathbf{T}$ disagree on a randomly chosen question $\mathbf{S}$. One finds for the generalisation error [3] :

$$\varepsilon(\mathbf{J}) = \int d\mu(\mathbf{S})\ \theta\left(-\left(\frac{\mathbf{J}.\mathbf{S}}{\sqrt{N}}\right).\left(\frac{\mathbf{T}.\mathbf{S}}{\sqrt{N}}\right)\right) \tag{1.4}$$

$$= \frac{1}{\pi}\arccos(R), \tag{1.5}$$

where R is the overlap between student $\mathbf{J}$ and teacher $\mathbf{T}$:

$$R = \frac{\mathbf{T}.\mathbf{J}}{N}. \tag{1.6}$$

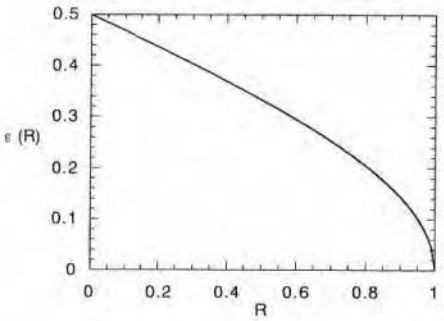

Fig.(1.5): The generalisation error ε as a function of the overlap R between student and teacher.

The patterns we consider are uniformly distributed on the N-dimensional sphere. As a result the integral in (1.4) can be written as :

$$\int d\mu(\mathbf{S})\ldots \sim \int_{-\infty}^{+\infty} d\mathbf{S}\ \delta\left(\mathbf{S}.\mathbf{S} - N\right)\ldots . \tag{1.7}$$

One has to determine the normalisation constant in order that :

$$\int_{-\infty}^{+\infty} d\mu(\mathbf{S}) = 1. \tag{1.8}$$

The expression (1.5) for the generalisation error can easily be derived geometrically. The vectors $\mathbf{J}$ and $\mathbf{T}$ both lie on the surface of an N-dimensional sphere. In Fig.(1.6) we have drawn only the $\mathbf{J} - \mathbf{T}$ plane. One immediatly sees that student and teacher will disagree on a new question $\mathbf{S}$ if its projection into the $\mathbf{J} - \mathbf{T}$ plane lies in the shaded region. Considering random questions one finds:

$$\varepsilon(\mathbf{J}) = \frac{2\theta}{2\pi} \tag{1.9}$$

$$= \frac{1}{\pi} \arccos\left(\frac{\mathbf{T}.\mathbf{J}}{N}\right) \tag{1.10}$$

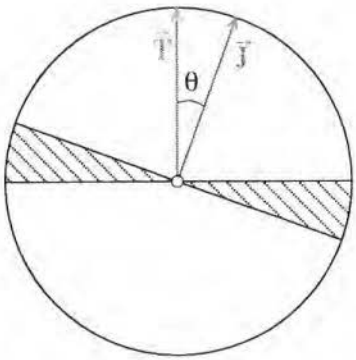

Fig.(1.6): Student and teacher will disagree if the projection of the new pattern $\mathbf{S}$ into the $\mathbf{J} - \mathbf{T}$ plane lies in the shaded region.

The question which now arises is how we will use the information provided by the teacher to construct a student perceptron $\mathbf{J}$ with an as low as possible generalisation error. Several learning algorithms already have been proposed. In the next section we give a short overview of these learning rules and the corresponding generalisation error.

1.3 Learning rules.

1.3.1 The Hebb rule.

Following a hypothesis made by the psychologist Hebb about the way in which synaptic strengths in the brain change in response to experience one can propose a simple learning mechanism. In the case where a student perceptron learns from examples generated by a teacher perceptron one defines the Hebb vector as :

$$\mathbf{J}_{Hebb} = \frac{1}{\gamma N} \sum_{\mu=1}^{P} \boldsymbol{\xi}^{\mu} \xi_0^{\mu}, \tag{1.11}$$

with γ the normalisation factor.
Considering the limit $N \to +\infty$ and $P \to +\infty$ with $\alpha = P/N$ finite, Vallet [22] derived the generalisation error for the Hebb rule and found:

$$\varepsilon_{Hebb} = \frac{1}{\pi} \arccos\left(\frac{1}{\sqrt{1+\frac{\pi}{2\alpha}}}\right). \tag{1.12}$$

For a good learning rule, the generalisation error must rapidly go to zero as the number of examples increases. Therefore it is interesting to study the behaviour of $\varepsilon(\alpha)$ when α becomes large. The large α behaviour of ε_{Hebb} is :

$$\alpha \to +\infty \;\; : \;\; \varepsilon_{Hebb} \sim \frac{0.40}{\sqrt{\alpha}}. \tag{1.13}$$

The Hebb algorithm is simple but has the disadvantage of a finite training error, i.e. not all example patterns $\boldsymbol{\xi}^{\mu}$ are classified correctly. The

training error, i.e. the fraction of misclassified patterns has been calculated in [22] and is given by :

$$e_t = \frac{1}{2} - \int_0^\infty \mathcal{D}t \ \mathrm{erf}\left(t\sqrt{\frac{\alpha}{\pi}} + \frac{1}{\sqrt{2\alpha}}\right). \tag{1.14}$$

The error function erf is defined as :

$$\mathrm{erf}(x) = 2\int_0^x \frac{dt}{\sqrt{\pi}} \exp\left(-t^2\right). \tag{1.15}$$

1.3.2 The Adaline rule.

The problem of learning can also be treated as an optimization process. This approach leads to the definition of a certain cost function $E(\mathbf{J})$ which has a global minimum at the vector $\mathbf{J}$ which has the desired properties. One can, for example, demand that the student vector $\mathbf{J}$ obeys the constraints :

$$\frac{\mathbf{J}.\boldsymbol{\xi}^\mu}{\sqrt{N}}\xi_0^\mu = \kappa \qquad \forall \mu = 1, \ldots, P \tag{1.16}$$

with $\kappa > 0$. In analogy with the Adaline learning algorithm used in the storage problem [10] one can define the cost function :

$$E(\mathbf{J}) = \frac{1}{2}\sum_{\mu=1}^{P}\left(\kappa - \frac{\mathbf{J}.\boldsymbol{\xi}^\mu}{\sqrt{N}}\xi_0^\mu\right)^2. \tag{1.17}$$

This cost function defines the following gradient descent dynamics :

$$\Delta J_i \sim \sum_{\mu=1}^{P}\left(\kappa - \frac{\mathbf{J}.\boldsymbol{\xi}^\mu}{\sqrt{N}}\xi_0^\mu\right)\xi_i^\mu\,\xi_0^\mu. \tag{1.18}$$

A modification of the cost function (1.17) in which κ is set to 1 and where the normalisation condition for the student vector $\mathbf{J}$ is dropped, has been originally introduced by Widrow and Hoff for the storage problem . This learning rule, which is usually called the pseudo-inverse

rule has been studied for the generalisation problem by Opper et al. in [23]. They obtain for the generalisation error:

$$\varepsilon_{PI}(\alpha) = \frac{1}{\pi}\arccos\left(\sqrt{\frac{2\alpha(1-\alpha)}{\pi-2\alpha}}\right) \qquad (\alpha<1), \tag{1.19}$$

$$\varepsilon_{PI}(\alpha) = \frac{1}{\pi}\arccos\left(\sqrt{\frac{2(\alpha-1)}{\pi+2\alpha-4}}\right) \qquad (\alpha>1). \tag{1.20}$$

For large α this leads to a generalisation error :

$$\alpha \to +\infty \;:\; \varepsilon_{PI} \sim \frac{0.24}{\sqrt{\alpha}}. \tag{1.21}$$

1.3.3 The Gibbs rule.

The previous learning rules in general do not lead to the exact classification of all the training patterns $\boldsymbol{\xi}^{\mu}$. Since one is particulary interested in learning rules which lead to student vectors with zero training error, one can study the so called version space. The version space is defined as the set of all properly normalized N-dimensional vectors which reproduce the classification of the training patterns by the teachers perfectly.
In [3], Seung et al. show that the overlap R of a random vector from the version space with the teacher is given by the following transcendental equation :

$$\frac{R}{1-R} = \frac{\alpha}{\pi\sqrt{1-R^2}} \int_{-\infty}^{+\infty} \mathcal{D}t \frac{e^{-\frac{1}{2}Rt^2}}{H\left(\sqrt{R}t\right)}. \tag{1.22}$$

Plugging the resulting R(α) into (1.5) gives the generalisation error of the "Gibbs perceptron" (which is not a unique vector but a sample from the version space).
The large α behaviour of ε_{Gibbs} is [3] :

$$\alpha \to +\infty \;:\; \varepsilon_{Gibbs} \sim \frac{0.625}{\alpha}. \tag{1.23}$$

Note that the generalisation error for the Hebb and Adaline rule exhibit a convergence to zero proportional to $1/\sqrt{\alpha}$ where a typical member of the version space leads to the much faster $1/\alpha$ decay .

1.3.4 The Adatron rule.

The stabilities λ^μ are, in analogy with the storage problem, defined as:

$$\lambda^\mu = \frac{\mathbf{J}.\boldsymbol{\xi}^\mu}{\sqrt{N}} \text{ sign} \left(\frac{\mathbf{T}.\boldsymbol{\xi}^\mu}{\sqrt{N}} \right). \tag{1.24}$$

All vectors from the version space reproduce the classification of the example patterns correctly. As a result they obey the constraints :

$$\lambda^\mu > 0 \qquad \forall \mu = 1, \ldots, P. \tag{1.25}$$

In [12] Anlauf and Biehl propose the Adatron algorithm which makes it possible to construct the maximally stable perceptron. The **J**-vector which characterizes this perceptron satisfies contraints which are stronger than (1.25):

$$\lambda^\mu \geq \kappa, \quad \forall \mu = 1, \ldots, P, \tag{1.26}$$

for the largest possible value of κ.

Geometrically $\mathbf{J}_{MS}$ corresponds to the vector which is the most distant from the borders of the version space. The thermodynamic properties of the maximally stable perceptron are described in [23]. For large α one finds [28] that the generalisation error again goes to zero proportional to $1/\alpha$ but with a smaller coefficient than the one for the Gibbs rule:

$$\alpha \to +\infty \qquad : \qquad \varepsilon_{MS} \sim \frac{0.5005}{\alpha}. \tag{1.27}$$

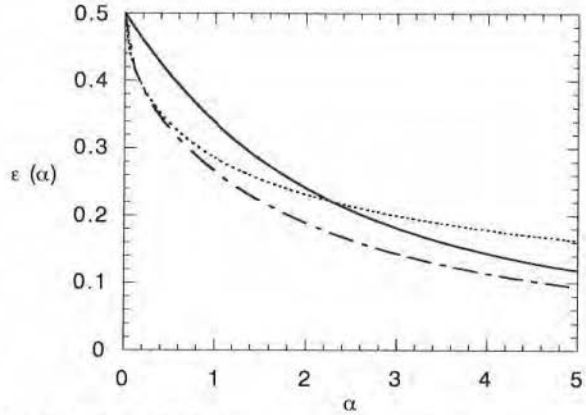

Fig.(1.7): The generalisation error ε as a function of α for the Hebb (dotted line), Gibbs (full line) and the maximally stable perceptron (dashed line).

1.3.5 The Bayes rule.

A very different type of learning rule is the Bayes prescription. Following this rule one classifies a new pattern following the majority vote of all **J**-vectors of the version space. Opper and Haussler [13] calculated the generalisation error of the Bayes classification algorithm and found the following result:

$$\varepsilon_{Bayes}(\alpha) = \frac{1}{\pi} \arccos\left(\sqrt{R_{Gibbs}(\alpha)}\right), \tag{1.28}$$

with R_{Gibbs} the typical overlap of a member of the version space with the teacher $\mathbf{T}$ (solution of the equation (1.22)). For large α one finds for the generalisation error [13] :

$$\alpha \to +\infty : \varepsilon_{Bayes} \sim \frac{0.442}{\alpha} \tag{1.29}$$

A priori it seems unlikely that the Bayes rule can be represented by a perceptron which is the same independent of which new question is being asked.
Surprisingly, Watkin [24] showed that there is indeed such a member, namely the perceptron characterized by the center of mass of all the vectors of the version space. By exploiting results from [26], P.Reimann showed in [32] that the Bayes vector $\mathbf{J}_{Bayes}$ makes the smallest angle with the unknown teacher vector $\mathbf{T}$ among all the student vectors that can be inferred from the given set of examples. Thus one can conclude that the Bayes rule is the optimal learning strategy.

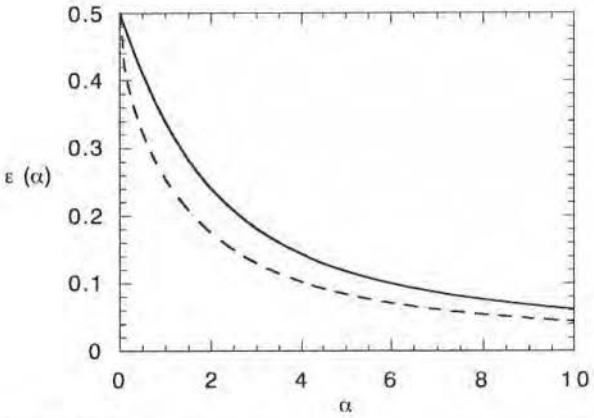

Fig.(1.8): Generalisation error of a typical member of the version space (Gibbs) and the one corresponding with the Bayes prescription (dashed line).

We are aware of 2 algorithms that try to construct the Bayes vector. Watkin [24] proposed the sampler method which is based on the idea of generating random vectors in the version space using the adatron algorithm. The "Billiard-method" [27] of Ruján uses the theory of billiards to generate a long trajectory of a vector in the version space which leads to an estimate of the Bayes vector. Both algorithms are difficult

to implement.

If one considers cost functions with a unique and non-degenerate minimum one can construct a student vector by using a simple gradient descent algorithm. In **Chapter 2** of this thesis, we will present a streamlined method to calculate the generalisation error which is valid whenever the student perceptron can be identified as the unique minimum of a specific cost function. To illustrate our method we will, using simple cost functions, rederive the results for the Hebb, Adaline and Adatron learning rule in a transparant and short way.
We furtermore use our method to study a new class of cost functions that penalises students which are close to the border of the version space. As a result we single out a cost function which makes it possible to construct a student perceptron that leads to a generalisation error extremely close to the one of the Bayes classifier. By calculating the overlap of this vector with the center of mass of the version space we show that the minimum of our optimal cost function is situated very close to the "Bayes-vector". This result is import because we now have the disposal of a practical algorithm to construct a student vector with almost optimal generalisation properties.
In **Chapter 3** we investigate if the cost function which gave results very close to optimal in the case of uniformly distributed patterns also leads to large overlaps in the case of a structured input space. We show that the previously mentioned class of potentials lead to an overlap R which is at least larger than the overlap obtained by using the Adatron learning rule.
In **Chapter 4** we address the problem of learning from a teacher with binary synapses. In view of the binary nature of the components of the teacher, one might expect that a lower error can be achieved by working with the clipped version of the student vector. It turns out that this is not always the case. In this thesis we show that the overlap for a vector with components $f(J_i)$, where f can be any odd function of its argument, is a simple function of the original overlap R.
It turns out that clipping leads to a larger overlap only if the original overlap R is larger than a certain treshold. We furthermore show that the optimal choice of f is a hyperbolic tangent. The corresponding generalisation error can go to zero exponentially fast in α^2, for α large.

Chapter 2

A unified approach : Gradient descent learning.

2.1 General theory

In this section a streamlined procedure will be developed to calculate the generalisation error of a student vector $\mathbf{J}$ obtained by minimizing a cost function $E(\mathbf{J})$:

$$E(\mathbf{J}) = \sum_{\mu=1}^{P} V(\lambda^{\mu}) \tag{2.1}$$

where λ^{μ} is given by (1.26) and will be called the stability of the μ-th pattern. Because of its similarity with mechanics the function $V(\lambda)$ will be called potential.
We restrict ourselves to cost functions $E(\mathbf{J})$ with a unique minimum. In order to calculate the overlap R between student and teacher vector and the corresponding generalisation error $\varepsilon(\alpha)$, the formalism of statistical mechanics turns out to be a powerful tool. Using the cost function $E(\mathbf{J})$ we define the partition function

$$Z = \int d\mu(\mathbf{J})\, e^{-\beta E(\mathbf{J})}, \tag{2.2}$$

with

$$\int d\mu(\mathbf{J}) \sim \int_{-\infty}^{+\infty} d\mathbf{J} \;\; \delta(\mathbf{J}^2 - N). \tag{2.3}$$

Through its dependence on the randomly chosen training patterns, Z is a random variable. In the thermodynamic limit where $N \to +\infty$ and $P \to +\infty$ with $\alpha = P/N$ fixed, one expects the corresponding free energy f to be self-averaging. In this way the free energy per neuron can be calculated as:

$$-\beta f = \frac{1}{N} \langle \ln Z \rangle_{\boldsymbol{\xi}}, \tag{2.4}$$

with $\langle . \rangle_{\xi}$ the average over the pattern distribution. To calculate the quenched average in (2.4) one commonly uses the replica method. The free energy can be written as

$$-\beta f = \lim_{n \to 0} \frac{1}{nN} \ln \langle Z^n \rangle_{\boldsymbol{\xi}}. \tag{2.5}$$

Under assumption of replica symmetry, one finds (for a detailed calculation see Appendix A) :

$$-f = \underset{q,R}{\text{extr}} \left[\frac{q - R^2}{2\beta(1-q)} + \frac{1}{2\beta} \ln(1-q) \right.$$
$$\left. - \frac{\alpha}{\beta} \int_{-\infty}^{+\infty} \mathcal{D}t_1 \int_{-\infty}^{+\infty} \mathcal{D}t_2 \ln \int_{-\infty}^{+\infty} \frac{d\lambda}{\sqrt{2\pi(1-q)}} \exp\left(g(t_1, t_2, q, R, \lambda)\right) \right], \tag{2.6}$$

with

$$g(t_1, t_2, q, R, \lambda) = -\beta V(\lambda \;\text{sign}(t_2)) - \frac{\left(\lambda - R t_2 - \sqrt{q - R^2} t_1\right)^2}{2(1-q)}. \tag{2.7}$$

The meaning of the order parameters is as usual: q is the overlap between two typical $\mathbf{J}$-vectors and R is the overlap between a typical $\mathbf{J}$-vector and the teacher vector $\mathbf{T}$. The word typical refers to the $\mathbf{J}$

vectors which give the exponentially dominant contribution of the free energy.

In order to find the ground state energy, we let $\beta \to +\infty$. Since we take only into account cost functions with a unique non-degenerate minimum, the overlap between two typical **J**-vectors should tend to 1 ($q \to 1$). Introducing the orderparameter $x = \beta(1-q)$ the free energy reduces to:

$$-f^{T=0} = \underset{x,R}{\text{extr}} \Big[\frac{1-R^2}{2x} - 2\alpha \int_{-\infty}^{+\infty} \mathcal{D}t_1 \int_{0}^{+\infty} \mathcal{D}t_2 \min_{\lambda} \Big[V(\lambda) + \frac{(\lambda - t)^2}{2x} \Big] \Big], \tag{2.8}$$

with

$$t = Rt_2 + \sqrt{1-R^2}t_1. \tag{2.9}$$

Let us call $\lambda_0(t,x)$ the value of λ which minimizes the expression

$$V(\lambda) + \frac{(\lambda - t)^2}{2x}. \tag{2.10}$$

This function will play a crucial role in the following. Introducing it in (2.8) easily leads to the extremum equations for x and R. One gets (for details see Appendix A) after an orthogonal transformation of the variables t_1 and t_2 that the saddle point equations can be written as :

$$R = \sqrt{\frac{2}{\pi}}\alpha \int_{-\infty}^{+\infty} \mathcal{D}t \, \lambda_0(\sqrt{1-R^2}t, x), \tag{2.11}$$

$$1 - R^2 = 2\alpha \int_{-\infty}^{+\infty} \mathcal{D}t \, H\left(-\frac{Rt}{\sqrt{1-R^2}} \right) (\lambda_0(t,x) - t)^2. \tag{2.12}$$

Solving these equations immediately leads to the generalisation error of the vector which minimizes $E(\mathbf{J})$ by plugging $R(\alpha)$ into (1.5).
Note that the function $\lambda_0(t,x)$ is identical to the one obtained in the

analogous treatment of the capacity problem [10]. Using $\lambda_0(t,x)$ some interesting quantities can be easily calculated. The ground state energy is given by :

$$e = 2\alpha \int_{-\infty}^{+\infty} \mathcal{D}t_1 \int_{0}^{+\infty} \mathcal{D}t_2 \, V(\lambda_0(t,x)) \tag{2.13}$$

$$= \int_{-\infty}^{+\infty} d\lambda P(\lambda) V(\lambda) \tag{2.14}$$

where the probability density $P(\lambda)$ of the aligned field is given by

$$P(\lambda) = \int_{-\infty}^{+\infty} \mathcal{D}t_1 \int_{0}^{+\infty} \mathcal{D}t_2 \;\; \delta(\lambda - \lambda_0(t,x)). \tag{2.15}$$

The training error ε_t is defined as the fraction of misclassified patterns. Such a pattern corresponds to an overlap $\lambda < 0$. Therefore

$$\varepsilon_t = \int_{-\infty}^{0} d\lambda \; P(\lambda). \tag{2.16}$$

All previous results are given using the RS-ansatz. In Appendix A a local stability analysis has been performed. The resulting Almeida-Thouless condition [16] takes on the the following simple form in terms of λ_0 :

$$2\alpha \int_{-\infty}^{+\infty} \mathcal{D}t \; H\left(-\frac{Rt}{\sqrt{1-R^2}}\right) (\lambda_0{}'(t,x) - 1)^2 < 1. \tag{2.17}$$

Note that a sufficient condition for RS breaking is the presence of a discontinuity of the alignment $\lambda_0(t)$ as a function of t [20].

2.2 Recovering previous results

First we will identify the cost functions which correspond to the algorithms discribed in Section 1.3, i.e. the Hebb, Adaline and Adatron learning rule and recover the well known results. In a later section, we will present a new class of cost functions that allow us to approximate closely the (optimal) Bayes results. Remark that since the Gibbs perceptron is not a unique vector but a sample from the version space our formalism does not apply to this case.

2.2.1 The Hebb potential

If we choose :

$$V(\lambda) = -\lambda, \tag{2.18}$$

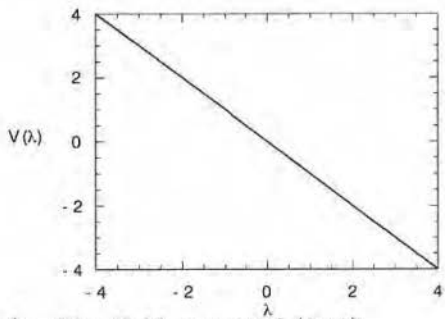

Fig.(2.1): The Hebb potential (2.18).

the cost function $E(\mathbf{J})$ becomes

$$E(\mathbf{J}) = -\sum_{\mu=1}^{P} \frac{\mathbf{J}.\boldsymbol{\xi}^{\mu}}{\sqrt{N}}\, \xi_0^{\mu}. \tag{2.19}$$

The minimum of the cost function $E(\mathbf{J})$ on the surface $\mathbf{J}^2 = N$ is known explicitly, namely

$$\mathbf{J} \sim \sum_{\mu=1}^{P} \boldsymbol{\xi}^{\mu}\, \xi_0^{\mu}. \tag{2.20}$$

With the proper normalisation constant, (2.20) is equal to the Hebb-vector. Applying the method developed in previous section, we first have to determine the function $\lambda_0(t,x)$. This can be done by plugging (2.18) into the expression (2.10) .
This leads to

$$\frac{d}{d\lambda}\left(-\lambda + \frac{(\lambda - t)^2}{2x}\right) = 0, \tag{2.21}$$

and as a result:

$$\lambda_0(t,x) = t + x. \tag{2.22}$$

Inserting this into the saddle point equations (2.11) and (2.12) gives:

$$1 - R^2 = 2\alpha\, x^2 \int_{-\infty}^{+\infty} \mathcal{D}t\; H\left(-\frac{Rt}{\sqrt{1-R^2}}\right), \tag{2.23}$$

$$R = \sqrt{\frac{2}{\pi}}\,\alpha\, x. \tag{2.24}$$

One easily finds that :

$$R = \sqrt{\frac{2\alpha}{\pi + 2\alpha}}, \tag{2.25}$$

$$x = \sqrt{\frac{\pi}{2}}\,\frac{R}{\alpha}. \tag{2.26}$$

From (2.25) and (2.26) the asymptotic behaviour can be derived. One finds:

$$\alpha \to 0 \;:\; \varepsilon_{Hebb} = \frac{1}{2} - \frac{\sqrt{2\alpha}}{\pi^{\frac{3}{2}}} + O(\alpha), \tag{2.27}$$

$$\alpha \to +\infty \;:\; \varepsilon_{Hebb} \sim \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{\alpha}} \sim \frac{0.40}{\sqrt{\alpha}}, \tag{2.28}$$

which corresponds with the results obtained in [22] by Vallet.

2.2.2 The Adaline potential

The Adaline rule corresponds to gradient descent with the potential function :

$$V(\lambda) = \frac{1}{2}(\lambda - \kappa)^2, \tag{2.29}$$

where κ is a positive constant.

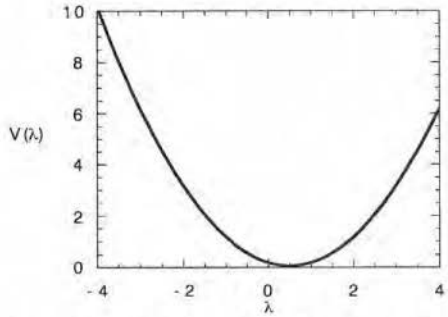

Fig.(2.2): The adaline potential (2.29) with $\kappa = 0.5$.

The function $\lambda_0(t, x)$ which minimizes (2.10) is :

$$\lambda_0(t, x) = \frac{t + \kappa x}{1 + x}. \tag{2.30}$$

This leads to the saddle point equations :

$$1 - R^2 = \alpha \left(\frac{x}{1+x}\right)^2 (\kappa^2 + 1 - 2\sqrt{2/\pi}\kappa R), \tag{2.31}$$

$$R = \sqrt{\frac{2}{\pi}} \alpha \frac{x}{1+x} \kappa. \tag{2.32}$$

For a fixed value of κ, there exists a lower bound for α, given by :

$$\alpha_c = \frac{\pi}{4\kappa^2} \left(1 + \kappa^2 - \sqrt{(1+\kappa^2)^2 - 8\kappa^2/\pi}\right). \tag{2.33}$$

For $\alpha < \alpha_c$ the equations (2.31) and (2.32) have no acceptable solution (i.e. $x \geq 0$ and $0 \leq R \leq 1$). The expression (2.33) for α_c can be obtained by putting $x = +\infty$ in the equations (2.31) and (2.32). Elimination of R leads to $\alpha_c(\kappa)$. The breakdown of the presented formalism for $\alpha < \alpha_c$ is due to the fact that the minimum of $E(\mathbf{J})$ is degenerate in this case.
From (2.31) and (2.32) we obtain that the large α behaviour of the generalisation error is :

$$\varepsilon(\kappa) \sim \frac{1}{\kappa\sqrt{\pi}} \left(\frac{1}{2}(1+\kappa^2) - \frac{2\kappa}{\sqrt{2\pi}} \right)^{\frac{1}{2}} \frac{1}{\sqrt{\alpha}}. \tag{2.34}$$

As an extension one can also study the case where one optimizes the parameter κ for each value of α. By eliminating x from (2.31) and (2.32) one obtains the equation for $R(\alpha, \kappa)$:

$$\frac{2}{\pi}\alpha\kappa^2(1-R^2) = R^2(\kappa^2 + 1 - 2\sqrt{\frac{2}{\pi}}\kappa R). \tag{2.35}$$

Deriving (2.35) with respect to κ and putting $R'(\kappa) = 0$ yields :

$$\frac{2}{\pi}\alpha\kappa^2(1-R^2) = R^2(\kappa^2 - \sqrt{\frac{2}{\pi}}\kappa R). \tag{2.36}$$

Eliminating R between (2.35) and (2.36) leads to the equation for $\kappa_{opt}(\alpha)$:

$$\frac{2}{\pi}\alpha\kappa^2(\kappa^2 - \frac{\pi}{2}) = \frac{\pi}{2}(\kappa^2 - 1). \tag{2.37}$$

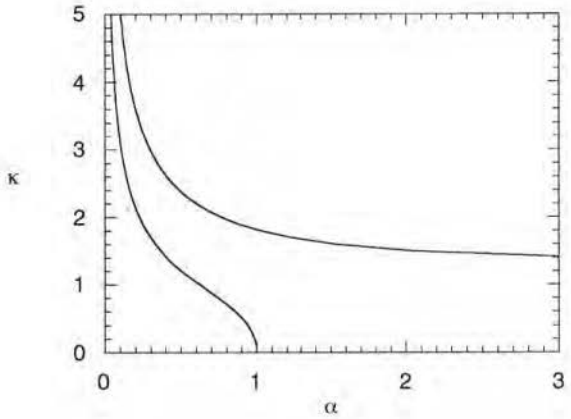

Fig.(2.3): $\kappa_{min}(\alpha)$ and $\kappa_{opt}(\alpha)$ as defined in the text. For $\alpha \to +\infty$, κ_{opt} tends to $\sqrt{\pi/2}$.

The function $\kappa_{opt}(\alpha)$ is shown in Fig. (2.3). The value of κ_{opt} decreases rapidly as α increases and tends to its asymptotic value $\sqrt{\pi/2}$ when $\alpha \to +\infty$. On the same figure we have also plotted $\kappa_{min}(\alpha)$ for $\alpha < 1$ which is the inverse function of $\alpha_c(\kappa)$ defined by (2.33). Fig.(2.4) shows the generalisation error $\varepsilon(\alpha)$ for the Adaline rule with optimal choice of κ and compares it to the result for the Hebb rule. The asymptotic behaviour yields :

$$\varepsilon(\alpha) \sim \frac{0.24}{\sqrt{\alpha}} \tag{2.38}$$

which is considerably better than the Hebb rule (2.28).

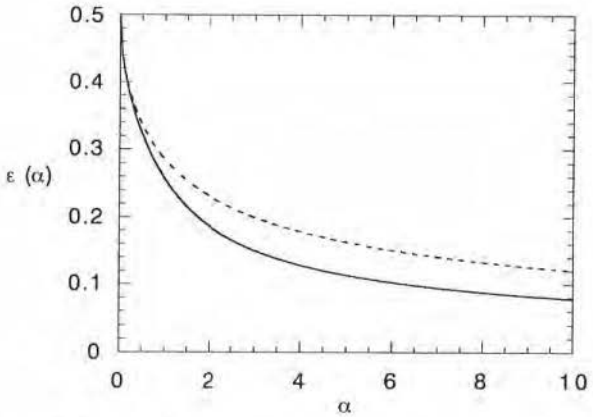

Fig.(2.4): The generalisation error $\varepsilon(\alpha)$ obtained by using the Adaline learning rule with κ_{opt} and as a comparison the Hebb generalisation error.

2.2.3 The Adatron potential

Following the criterion of "maximal stability", one looks for the **J**-vector such that $\lambda^\mu > \kappa$ ($\forall \mu = 1, \ldots, P$), for the largest possible value of κ.
This can be realized by considering the potential:

$$V(\lambda) = \begin{cases} +\infty & \lambda < \kappa \\ 0 & \lambda \geq \kappa \end{cases} \tag{2.39}$$

and determine the largest possible value of κ for which there exists a solution $\boldsymbol{J}$ with cost E equal to zero.
From (2.10), one finds for the function $\lambda_0(t, x)$:

$$\lambda_0(t, x) = t + (\kappa - t)\theta(\kappa - t). \tag{2.40}$$

By inserting this result into the saddle point equations Eq.(2.11) and Eq.(2.12), one finds that the variable x disappears altogether, and one obtains following two equations :

$$\frac{1-R^2}{2\alpha} = \int_{-\infty}^{\kappa} \mathcal{D}t\, H\left(-\frac{Rt}{\sqrt{1-R^2}}\right)(\kappa - t)^2 \tag{2.41}$$

$$\sqrt{\frac{\pi}{2}}\frac{R}{\alpha} = \int_{-\infty}^{\frac{\kappa}{\sqrt{1-R^2}}} \mathcal{D}t\, (\kappa - t\sqrt{1-R^2}), \tag{2.42}$$

which are identical to those given in [23]. The equation for κ arises from the fact that there is a unique value of this parameter, namely precisely the one corresponding to "optimal stability", for which our formalism applies (i.e there is a non-degenerate ground state with zero value of the cost). These equations can now be used to calculate the maximum value of κ and the corresponding value of R.

An alternative and more transparant method uses the following potential :

$$V(\lambda) = \begin{cases} +\infty & \lambda < \kappa \\ \lambda & \lambda \geq \kappa \end{cases} \tag{2.43}$$

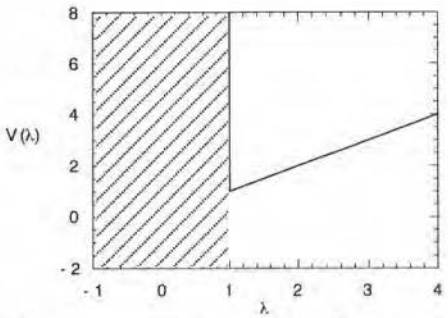

Fig.(2.5): The potential (2.43) with $\kappa = 1.0$.

This leads to :

$$\lambda_0(t,x) = t - x + (\kappa - t + x)\theta(\kappa - t + x). \tag{2.44}$$

Inserting $\lambda_0(t,x)$ into (2.11) and (2.12) leads to the following equations determining R and x for given value of κ and α:

$$\frac{1-R^2}{2\alpha} = \int_{-\infty}^{\kappa+x} \mathcal{D}t\, H\left(-\frac{Rt}{\sqrt{1-R^2}}\right)(\kappa - t)^2 + \int_{\kappa+x}^{+\infty} \mathcal{D}t\, H\left(-\frac{Rt}{\sqrt{1-R^2}}\right) x^2 \tag{2.45}$$

$$\sqrt{\frac{\pi}{2}}\frac{R}{\alpha} = \int_{-\infty}^{\frac{\kappa+x}{\sqrt{1-R^2}}} \mathcal{D}t\, \kappa + \int_{\frac{\kappa+x}{\sqrt{1-R^2}}}^{+\infty} \mathcal{D}t\, (t\sqrt{1-R^2} - x). \tag{2.46}$$

As in the case of the Adaline learning rule we have that for a fixed value of κ these two equations have a "physical" solution only if $0 \leq R \leq 1$ and $0 \leq x \leq +\infty$. This is only the case when α is lying in the interval $[\alpha_{min}(\kappa), \alpha_{max}(\kappa)]$. For $\alpha < \alpha_{min}(\kappa)$ our formalism is not valid since there are a lot of $\mathbf{J}$ vectors which have $E(\mathbf{J}) = 0$. In that case $q < 1$ and as a result $x = +\infty$ for $\beta \to +\infty$. For $\alpha > \alpha_{min}$ one obtains $x < +\infty$. At $\alpha = \alpha_{max}$ one has $x = 0$ and for all $\alpha > \alpha_{max}$ one obtains $E(\mathbf{J}) = +\infty$ for all $\mathbf{J}$ vectors. The values for $\alpha_{min}(\kappa)$ and $\alpha_{max}(\kappa)$ can be calculated by putting $x = +\infty$ and $x = 0$ in the equations (2.45) and (2.46).

If we put $x = +\infty$ the equations (2.45) and (2.46) become :

$$\frac{1-R^2}{2\alpha} = \int_{-\infty}^{+\infty} \mathcal{D}t\, H\left(-\frac{Rt}{\sqrt{1-R^2}}\right)(\kappa - t)^2, \tag{2.47}$$

$$\sqrt{\frac{\pi}{2}}\frac{R}{\alpha} = \int_{-\infty}^{+\infty} \mathcal{D}t\, \kappa = \kappa. \tag{2.48}$$

The integral in (2.47) can be calculated :

$$\frac{1-R^2}{2\alpha} = \frac{1}{2}(\kappa^2+1) - 2\kappa \int_{-\infty}^{+\infty} \mathcal{D}t\, t\, H\left(-\frac{Rt}{\sqrt{1-R^2}}\right) \quad (2.49)$$

$$= \frac{1}{2}(\kappa^2+1) - \frac{2\kappa R}{\sqrt{2\pi}}. \quad (2.50)$$

The equations for α_{min} and R can thus be written as :

$$\frac{1-R^2}{\alpha} = \kappa^2 + 1 - 2\sqrt{\frac{2}{\pi}}\kappa R \quad (2.51)$$

$$\frac{R}{\alpha} = \sqrt{\frac{2}{\pi}}\kappa \quad (2.52)$$

These equations are equal to the saddle point equations (2.31) and (2.32) for the Adaline learning rule with $x = +\infty$. As a result $\alpha_{min} = \alpha_c$ with α_c defined in (2.33). This is not a surprise since at the value of α corresponding with $x = +\infty$ there only exist one $\mathbf{J}$ with zero cost. All stabilities λ_μ are equal to κ and the vector which satisfies these constraints is independent of the shape of the potential $V(\lambda)$(it only has to be minimal at $\lambda = \kappa$).

Let us now put $x = 0$. The equations (2.45) and (2.46) become :

$$\frac{1-R^2}{2\alpha} = \int_{-\infty}^{\kappa} \mathcal{D}t\, H\left(-\frac{Rt}{\sqrt{1-R^2}}\right) (\kappa - t)^2, \quad (2.53)$$

$$\sqrt{\frac{\pi}{2}}\frac{R}{\alpha} = \int_{-\infty}^{\frac{\kappa}{\sqrt{1-R^2}}} \mathcal{D}t\, (\kappa - t\sqrt{1-R^2}). \quad (2.54)$$

For a given value of κ these two equations determine α_{max} and the corresponding value of R.

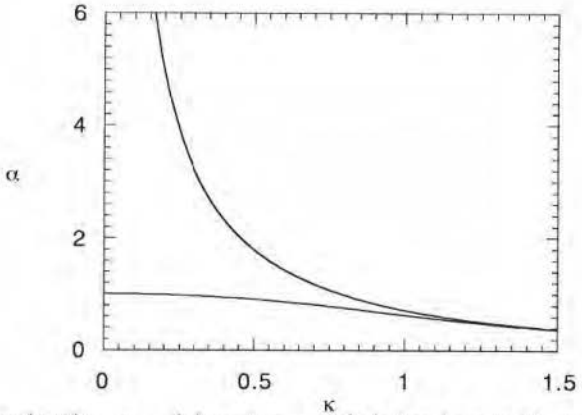

Fig.(2.6): $\alpha_{min}(\kappa)$ and $\alpha_{max}(\kappa)$ defined as in the text.

Remark that the equations (2.53) and (2.54) are identical to the equations (2.41) and (2.42) for the maximally stable perceptron. This is not a surprise since $\alpha_{max}(\kappa)$ which has been plotted in Fig.(2.6) can be interpreted differently. For each value of α, this curve gives the largest value of κ for which a solution with finite value of $E(\mathbf{J})$ exists. But this is clearly the maximal stable perceptron. Similarly, the curve $\alpha_{min}(\kappa)$ is the same as $\kappa_{min}(\alpha)$ for the Adaline rule.

The asymptotic behaviour of ε for the maximal stable student can be determined from (2.53) and (2.54). For $\alpha \to +\infty$, we have that $\kappa \to 0$ and $R \to 1$. First we change the integration variable in equation (2.41) via the coordinate transformation :

$$u = \frac{t}{\sqrt{1-R^2}}. \tag{2.55}$$

This leads to :

$$\frac{1}{2\alpha\sqrt{1-R^2}} = \int_{-\infty}^{\frac{\kappa}{\sqrt{1-R^2}}} \frac{du}{\sqrt{2\pi}}\, e^{-\frac{1}{2}(1-R^2)u^2} H(-Ru)\ (\frac{\kappa}{\sqrt{1-R^2}} - u)^2$$

(2.56)

The equation (2.42) can be rewritten as :

$$\sqrt{\frac{\pi}{2}}\frac{R}{\alpha\sqrt{1-R^2}} = \int_{-\infty}^{\frac{\kappa}{\sqrt{1-R^2}}} \mathcal{D}t\,(\frac{\kappa}{\sqrt{1-R^2}} - t). \tag{2.57}$$

For $\alpha \to +\infty$ we have :$\kappa \to 0$ and $R \to 1$. As a result one can make the ansatz that :

$$A = \alpha\sqrt{1-R^2}, \tag{2.58}$$

$$B = \frac{\kappa}{\sqrt{1-R^2}}, \tag{2.59}$$

stay finite and obtain the equations:

$$\frac{1}{2A} = \int_{-\infty}^{B} \frac{du}{\sqrt{2\pi}} H(-u)\,(B-u)^2 \tag{2.60}$$

$$\sqrt{\frac{\pi}{2}}\frac{1}{A} = \int_{-\infty}^{B} \mathcal{D}t\,(B-t). \tag{2.61}$$

Numerically solving the equations for A and B leads to the following asymptotic form for the generalisation error of the maximum stable perceptron [23] :

$$\varepsilon_g = \frac{\sqrt{1-R^2}}{\pi} = \frac{A}{\pi}\frac{1}{\alpha} = \frac{0.5005}{\alpha}, \tag{2.62}$$

which is equal to the result found by Opper in [28].

2.3 A class of cost functions which favour large overlaps within the version space.

2.3.1 The Hebb rule within the version space.

Since we know that the center of mass of the version space has optimal generalisation properties one might expect that potential functions which "push" the student vector away from the boundaries of the version space will result in a low generalisation error. One of the easiest choices we can make is taking $+\infty$ for $\lambda < 0$ and the Hebb potential inside the version space, i.e.

$$V(\lambda) = \begin{cases} +\infty & \lambda < 0 \\ -\lambda & \lambda > 0. \end{cases} \tag{2.63}$$

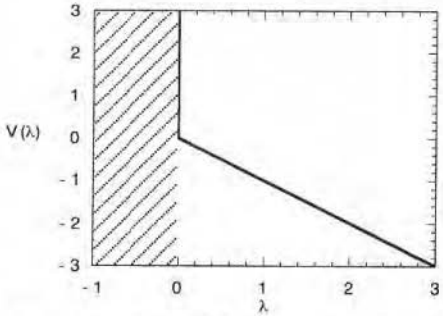

Fig.(2.7): The potential function (2.63).

The corresponding solution of (2.10) is :

$$\lambda_0(t,x) = (t+x)\,\theta(t+x) \tag{2.64}$$

Inserting this in the saddle point equations (2.11) and (2.12) gives :

$$R = \sqrt{\frac{2}{\pi}}\alpha \int_{-\infty}^{+\infty} \mathcal{D}t\,(\sqrt{1-R^2}t+x)\,\theta(\sqrt{1-R^2}t+x) \tag{2.65}$$

$$1 - R^2 = 2\alpha \int_{-\infty}^{+\infty} \mathcal{D}t \; H\left(-\frac{Rt}{\sqrt{1-R^2}}\right) ((t+x)\,\theta(t+x) - t)^2 \tag{2.66}$$

Numerically solving these saddle point equations result in a generalisation error which is plotted together with ε_{Hebb} in Fig.(2.8).

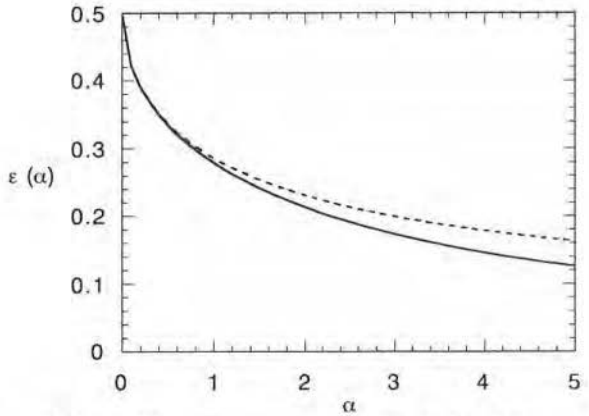

Fig.(2.8): The generalisation error of the Hebb rule (dashed curve) and the one resulting from the equations (2.65) and (2.66).

The large α behaviour of ε can be calculated by rewriting the equations (2.65) and (2.66). Let us begin by considering the first equation. Equation (2.66) can be written as :

$$1 - R^2 = 2\alpha \left[\frac{1}{2}x^2 + \int_{-\infty}^{-x} \mathcal{D}t \; H\left(-\frac{Rt}{\sqrt{1-R^2}}\right)(t^2 - x^2)\right]. \tag{2.67}$$

The transformation :

$$u = \frac{t}{\sqrt{1-R^2}}, \tag{2.68}$$

leads to :

$$\frac{1}{2\alpha\sqrt{1-R^2}} = \frac{x^2}{(1-R^2)^{3/2}} + \int_{-\infty}^{-\frac{x}{\sqrt{1-R^2}}} du\, e^{-\frac{1}{2}(1-R^2)u^2} H(-Ru)(u^2 - \frac{x^2}{1-R^2}). \quad (2.69)$$

The second saddle point equation (2.65) becomes :

$$R = \sqrt{\frac{2}{\pi}}\alpha \int_{-\frac{x}{\sqrt{1-R^2}}}^{+\infty} \mathcal{D}t(x + \sqrt{1-R^2}t). \quad (2.70)$$

If $\alpha \to +\infty$ then $R \to 1$ and $x \to 0$. From (2.69) it is clear that if $\alpha\sqrt{1-R^2}$ is to stays finite in order to obtain a generalisation error proportional to $1/\alpha$, it is necessary that $x^2/(1-R^2)^{3/2}$ remains finite. But then $x/\sqrt{1-R^2}$ must go to zero. Consider now the second saddle point equation (2.65) and write it as :

$$R = \sqrt{\frac{2}{\pi}}\alpha\sqrt{1-R^2} \int_{-\frac{x}{\sqrt{1-R^2}}}^{+\infty} \mathcal{D}t\, (\frac{x}{\sqrt{1-R^2}} + t). \quad (2.71)$$

Since $x/\sqrt{1-R^2} \to 0$ when $\alpha \to +\infty$, this equation becomes :

$$1 = \sqrt{\frac{2}{\pi}}\alpha\sqrt{1-R^2} \int_{0}^{+\infty} \mathcal{D}t\, t, \quad (2.72)$$

from which we get:

$$\alpha\sqrt{1-R^2} = \pi. \quad (2.73)$$

As a result we obtain for the asymptotic form of the generalisation error corresponding with the Hebb rule inside the version space:

$$\varepsilon(\alpha) = \frac{\sqrt{1-R^2}}{\pi} = \frac{1}{\alpha}. \quad (2.74)$$

2.3.2 Bounds for the generalization error associated to a monotonous potential.

It is possible to derive a bound for the generalization error corresponding with cost functions based on potentials, defined within the version space, that are monotonous decreasing functions of λ. Consider the case of a monotonous decreasing potential:

$$\begin{aligned} V(\lambda) &= +\infty \quad \lambda < 0 \\ V'(\lambda) &< 0 \quad \lambda > 0. \end{aligned} \tag{2.75}$$

The function $\lambda_0(t, x)$ that minimizes (2.10) is a solution of the following equation:

$$\lambda - t = -xV'(\lambda). \tag{2.76}$$

provided $\lambda \geq 0$ and it is zero otherwise. Consequently one finds that :

$$\lambda_0(t, x) \geq 0 \quad t \leq 0 \tag{2.77}$$

$$\lambda_0(t, x) \geq t \quad t \geq 0. \tag{2.78}$$

Using this in the saddle point equation (2.11) yields :

$$\sqrt{\frac{\pi}{2}}\frac{R}{\alpha} = \int_{-\infty}^{\infty} Dt \;\; \lambda_0(t\sqrt{1-R^2}, x) \tag{2.79}$$

$$\geq \int_{0}^{\infty} Dt \, t\sqrt{1-R^2} = \frac{\sqrt{1-R^2}}{\sqrt{2\pi}}. \tag{2.80}$$

Hence

$$R^2 \geq \frac{\alpha^2}{\alpha^2 + \pi^2}, \tag{2.81}$$

and thus for every monotonous decreasing potential in the version space we have :

$$\varepsilon(\alpha) \leq \varepsilon^*(\alpha) \quad \forall \alpha > 0, \tag{2.82}$$

with

$$\varepsilon^*(\alpha) = \frac{1}{\pi} \arccos \sqrt{\frac{\alpha^2}{\pi^2 + \alpha^2}}, \tag{2.83}$$

which for large α leads to :

$$\varepsilon^*(\alpha) \sim \frac{1}{\alpha} \tag{2.84}$$

2.3.3 A class of repulsive potentials within the version space.

A generalisation of the potential (2.63) is given by the following general class of monotonic potential functions:

$$V_s^+(\lambda) = \begin{cases} +\infty & \lambda < 0 \\ -\frac{\lambda^s}{s} & \lambda > 0 \end{cases} \tag{2.85}$$

with s real and $\neq 0$. For $s = 0$, we define $V_s^+(\lambda)$ as:

$$V_{s=0}^+(\lambda) = \begin{cases} +\infty & \lambda < 0 \\ -\ln \lambda & \lambda > 0 \end{cases} \tag{2.86}$$

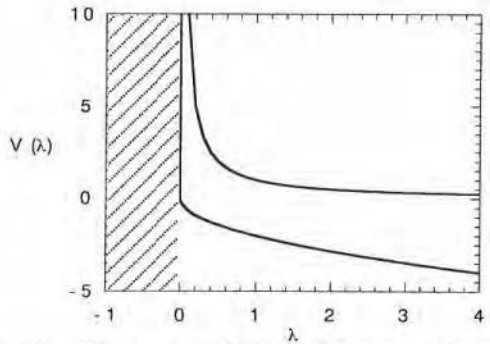

Fig.(2.9): The potential (2.85) for $s = -1.0$ and $s = 0.25$.

We start by pointing out that the parameter s may not be larger than 2 in order to avoid the divergence of the integral determining the value of the free energy, cf. Eq.(2.6). Furthermore, it is shown in Appendix B that the cost functions associated with the above potential is convex $\forall s \leq 1$, hence the minimum is unique and can be found by gradient descent. At $s = 1$, the curvature of the potential switches sign with the result that, for the values $1 < s \leq 2$, one of the conditions in our proof is not met. In fact we will find that the local stability of the replica symmetric solution is violated.

We now use our general method for determining the minimum of the cost function for different values of the parameter s. Since the potential $V_s^+(\lambda)$ has infinite value for $\lambda < 0$ the function λ_0 is defined as :

$$\min_{\lambda \geq 0} \left[V_s^+(\lambda) + \frac{(\lambda - t)^2}{2x} \right]. \tag{2.87}$$

This means that $\lambda_0(t)$ satisfies

$$-\lambda^{s-1} + \frac{\lambda - t}{x} = 0. \tag{2.88}$$

if, for a given value of t, this leads to a solution $\lambda_0 > 0$. If this is not the case we have $\lambda_0 = 0$.
So, the function that minimizes (2.10) for the potential $V_s^+(\lambda)$ is given by :

$$\lambda_0 - t = x\lambda_0^{s-1} \quad (\lambda_0 \geq 0). \tag{2.89}$$

It is not possible to solve $\lambda_0(t, x)$ for general s but one can trivially solve (2.89) for the inverse function:

$$t(\lambda_0, x) = \lambda_0 - x\lambda_0^{s-1} \quad \lambda_0 \geq 0. \tag{2.90}$$

For $s < 1$, the function $t(\lambda_0, x)$ is a monotonously increasing function on the interval $\lambda_0 > 0$ ranging from $-\infty$ at $\lambda_0 = 0$ to $+\infty$ as $\lambda_0 \to +\infty$.

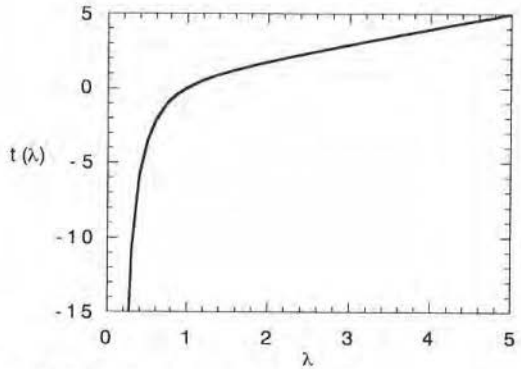

Fig.(2.10): The function $t(\lambda)$ for $s = -1$ and $x = 1.0$.

In this case, it is convenient to use λ_0 as new integration variable in the saddle point equations Eq.(2.11) and Eq.(2.12) and to calculate the integrals numerically.

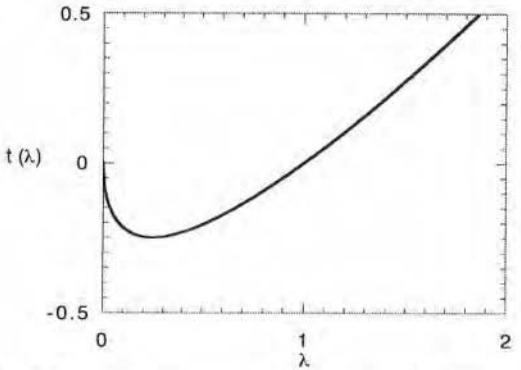

Fig.(2.11): The function $t(\lambda)$ for $s = 1.5$ and $x = 1.0$.

For $1 < s < 2$, however, $t(\lambda_0)$ decreases from the value 0 at $\lambda_0 = 0$ to a minimum $t_m < 0$ at a certain value λ_c and then increases steadily for larger values of λ. In this case, since λ_0 has to be nonnegative, the function $\lambda_0(t, x) \equiv 0$ for $t < t_m$. At $t = t_m$, it makes a finite jump to the value λ_c and follows further the increasing branch of the inverse function $t(\lambda_0, x)$.

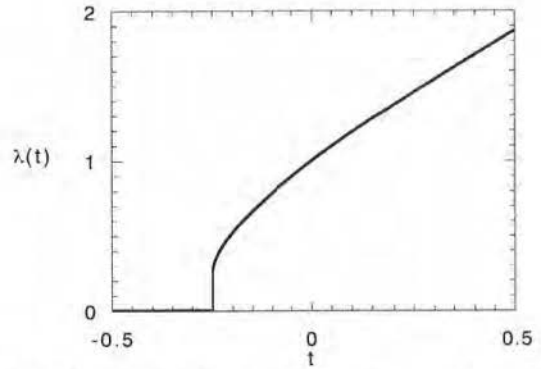

Fig.(2.12): The function $\lambda_0(t)$ for $s = 1.5$ and $x = 1.0$. The discontinuity occurs at $t_m = -0.25$.

The integrals must now be split in a part $-\infty < t < t_m$ where $\lambda_0(t, x) \equiv 0$ and in a part $t_m < t < +\infty$ where λ_0 can again be used as new integration variable. In this way, we never need the explicit solution of Eq.(2.89) for $\lambda_0(t, x)$.

With regard to the stability of the replica symmetric solution it is immediatly clear from the existence of a discontinuity of $\lambda_0(x, t)$ at $t = t_m$ that for $2 > s > 1$ the AT-condition (2.17) is not fulfilled and we have RSB (Replica Symmetry Breaking).

2.3.4 Numerical results.

By using the above discribed technique, one can solve the saddle point equations numerically and find the values for $R(\alpha)$ and $x(\alpha)$. In Fig. (2.13) the generalisation error is plotted for several values of s.

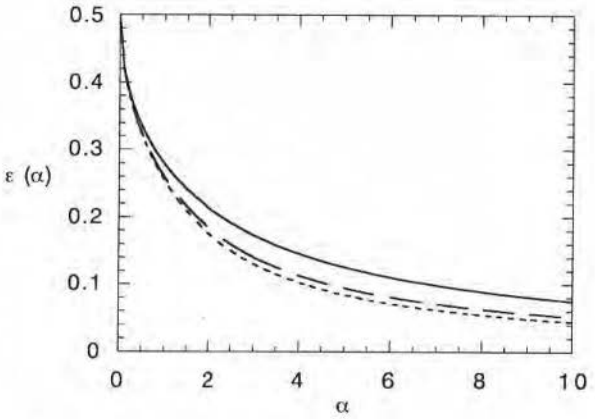

Fig.(2.13): Generalisation error for $s = 1$ (full line), $s = 0.25$ (long dashes) and $s = -1$ (short dashes).

For $s < 1$, one finds by numerical evaluation that the RS solution satisfies the stability condition (2.17). This was to be expected since for this case we showed that the minimum of the cost function is unique and non-degenerate.

2.3.5 Asymptotic behaviour.

To get a more precise idea of how the generalization error depends on the parameter s, we derive the exact asymptotic results for small α and large α values.

For α small, $R \to 0$ and $x \to \infty$. From Eqs.(2.11) and (2.12) one gets:

$$R = \sqrt{\frac{2\alpha}{\pi}} \tag{2.91}$$

hence

$$\varepsilon(\alpha) \sim \frac{1}{2} - \frac{\sqrt{2\alpha}}{\pi^{\frac{3}{2}}} \tag{2.92}$$

which is identical to the small α behaviour for the Hebb rule.

The calculation of the asymptotic behaviour is similar to the derivation for the Hebb rule inside the version space (section 2.1) but more involved and therefore is presented in Appendix C. We briefly discuss the results. For $\alpha \to +\infty$ the order parameters R and x will respectively tend to 1 and 0. Introducing this limit into the saddle point equations results equations for the variables A and B which are defined as :

$$A = \frac{\alpha\sqrt{1-R^2}}{\pi}, \tag{2.93}$$

$$B = \frac{x^{1/(2-s)}}{\sqrt{1-R^2}}. \tag{2.94}$$

The coefficient A is the interesting one because it is directly related to the asymptotic behaviour of $\varepsilon(\alpha)$:

$$\varepsilon(\alpha) \sim \frac{A}{\alpha} \tag{2.95}$$

In solving the equations for A and B, one must distinguish two cases. For $s \geq 1/2$ the equations yield $A = 1$ and $B = 0$. The asymptotic behaviour thus saturates the upper bound (2.84). For $s < 1/2$, B is different from 0 and one should determine A and B numerically . The value of A is represented in Fig. (2.14) as a function of s.

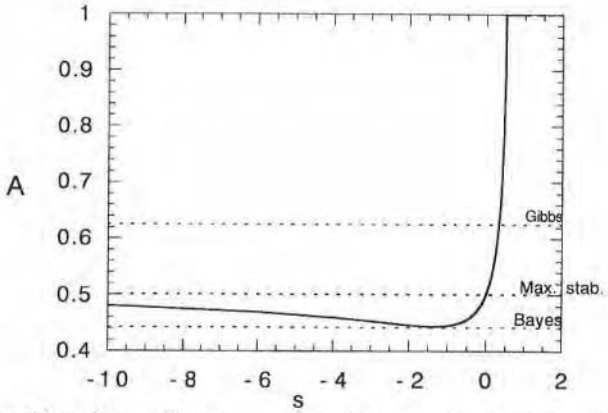

Fig.(2.14): The proportionality constant A describing the assymptotic decay of the generalisation in function of the value of s. On the figure the proportionality constants of the Gibbs, maximal stability and Bayes rule are indicated.

As one moves from large to small s values, one observes that A first takes on a constant plateau value equal to 1 for $\frac{1}{2} \leq s \leq 2$, then decreases from the value 1 for $s = 1/2$ to a minimum value of $A \approx 0.443$ for $s \approx -1.35$ after which it again slightly increases and asymptotically approaches to the value $A \approx 0.50$ which is presumably identical to the asymptotic value for the perceptron with maximal stability. The value for A attains its minimum 0.443 which is extremely close to A_{Bayes} which is 0.442.
In Fig. (2.15) we plot the generalisation curve for the optimal s value -1.35 together with the Bayes generalisation error. The deviation is smaller than 1 percent for all values of α.
It is important to recall that since the cost function has a unique non-degenerate minimum one can easily construct the **J**-vector which minimizes the optimal cost function by applying the gradient descent technique.

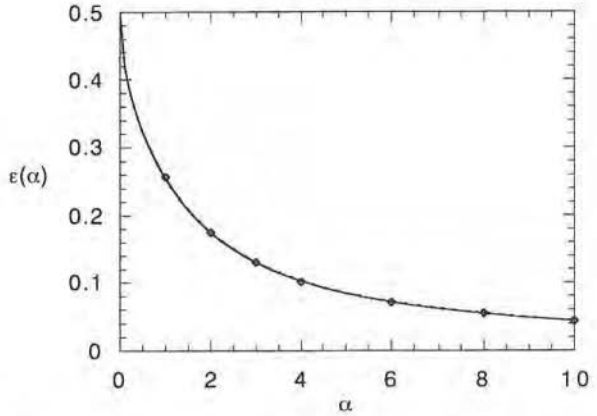

Fig.(2.15): Theoretical results for the Bayes rule (full curve) and for the repulsive potential $V_s^+(\lambda)$ with $s = -1.35$ (dotted line) together with simulation results for a system with 50 neurons.

To check the theoretical results some simulations have been carried out. Since an inverse power law potential $s = -1$ is numerically less time consuming and stil gives results within 1 percent of the Bayes rule we used this potential for the simulations. The results for a system of 50 input neurons are plotted in Fig. (2.15) and show excellent agreement with the theoretical curve.

2.4 A class of potentials which favour small overlaps within the version space

The worst student from the version space is the one with the smallest overlap R with the teacher. Engel and Van den Broeck [25] have calculated the large α behaviour of the generalisation error for this worst

student. A one step Replica Symmetry Breaking (RSB) calculation predicts :

$$\varepsilon(\alpha) \sim \frac{3}{2}\,\frac{1}{\alpha}. \tag{2.96}$$

The question which arises is if we can define a cost function $E(\mathbf{J})$ which after minimisation produces "bad students". Since we know that the version space is convex and that the best possible student that can be constructed on the basis of the training set is the center of mass, we infer that the students with larger generalization error can typically be found close to the boundaries of the version space. To verify this intuition, we consider the following class of functions that favour small values of λ:

$$V_s^-(\lambda) = \begin{cases} +\infty & \lambda < 0 \\ +\frac{\lambda^s}{s} & \lambda > 0. \end{cases} \tag{2.97}$$

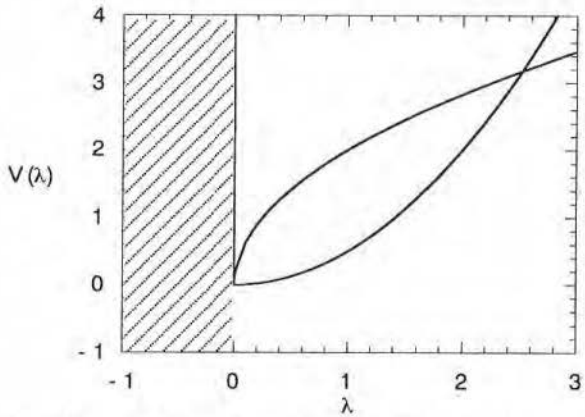

Fig.(2.16): Attractive potential functions for $s = 0.5$ and $s = 2.0$.

The parameter s now may take on positive values only. Negative values must be excluded as they would destroy the convergence of the integral

over λ in (2.6).
The lowest energy $E(\mathbf{J})$ will be 0 for all values of $\alpha < 1$. Indeed, for $\alpha < 1$, many different students will satisfy the stabilty conditions $\lambda^\mu = 0$ $(\mu = 1, ..., p)$ so that Eqs. (2.11) and (2.12), which are based on the assumption of a unique non-degenerate minimum, do not describe this case. We therefore limit ourselves here to $\alpha > 1$.

The function $V_s^-(\lambda)$ is a monotonously increasing function of λ. This makes it possible to calculate a lower bound for the generalisation error. The class of potentials defined by (2.97) satisfies the conditions :

$$\begin{aligned} V(\lambda) &= +\infty \quad \lambda < 0 \\ V'(\lambda) &> 0 \quad \lambda > 0. \end{aligned} \tag{2.98}$$

From (2.10) we find consequently :

$$\lambda_0(t, x) = 0 \quad t \leq 0 \tag{2.99}$$

$$\lambda_0(t, x) \leq t \quad t \geq 0. \tag{2.100}$$

Using this in the saddle point equation Eq.(2.11) yields :

$$\sqrt{\frac{\pi}{2}} \frac{R}{\alpha} = \int_{-\infty}^{\infty} Dt \ \lambda_0(t\sqrt{1-R^2}, x) \tag{2.101}$$

$$\leq \int_0^{\infty} Dt \ t\sqrt{1-R^2} = \frac{\sqrt{1-R^2}}{\sqrt{2\pi}}. \tag{2.102}$$

Hence

$$R^2 \leq \frac{\alpha^2}{\alpha^2 + \pi^2}, \tag{2.103}$$

and thus for every monotonic increasing potential in the version space we have :

$$\varepsilon(\alpha) \geq \varepsilon^*(\alpha) \quad \forall \alpha > 1, \tag{2.104}$$

with

$$\varepsilon^*(\alpha) = \frac{1}{\pi} \arccos \sqrt{\frac{\alpha^2}{\pi^2 + \alpha^2}} \tag{2.105}$$

This means that for large α the generalisation error behaves as :

$$\varepsilon(\alpha) \geq \varepsilon^*(\alpha) \sim \frac{1}{\alpha}. \tag{2.106}$$

The equation for $\lambda_0(t,x)$ now reads

$$\lambda_0 - t = -x\lambda_0^{s-1} \ (\lambda_0 \geq 0). \tag{2.107}$$

From its definition it follows that $x \geq 0$. Since λ_0 must be nonnegative, we immidiatly see that $\lambda_0(t,x) = 0$ for $t \leq 0$.
It is easy to solve Eq.(2.107) for the inverse function

$$t(\lambda_0, x) = \lambda_0 + x\lambda_0^{s-1} \ (\lambda_0 \geq 0). \tag{2.108}$$

For $s > 1$, $t(\lambda_0, x)$ is zero at $\lambda_0 = 0$ and increases steadily with increasing λ_0. This defines the inverse function uniquely.

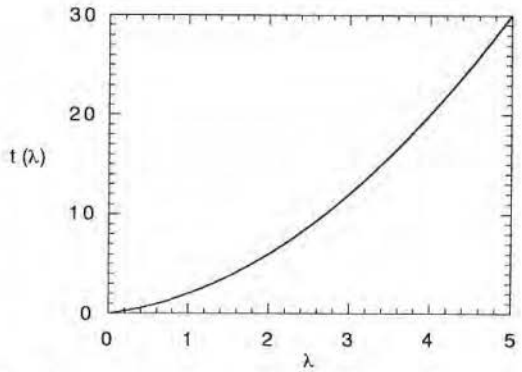

Fig.(2.17): The function $t(\lambda)$ for $s = 3.0$ and $x = 1.0$.

For $0 < s < 1$ on the other hand, $t(\lambda_0, x)$ decreases from $+\infty$ at $\lambda_0 = 0$ to a minimum $t_m > 0$ at a certain value λ_c and then increases steadily for larger values of λ.

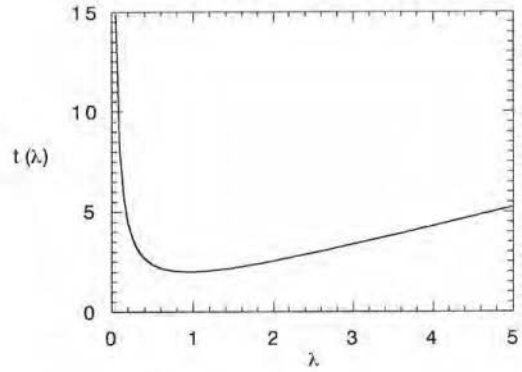

Fig.(2.18): The function $t(\lambda)$ for $s = 0.1$ and $x = 1.0$.

In this case $\lambda_0(t, x) \equiv 0$ for $t < t_m$. At $t = t_m$, it makes a finite jump to the inverse function of $t(\lambda_0, \alpha)$. Using these observations, the integrals in the saddle point equations can again be calculated by splitting the integration interval in a part $-\infty < t < t_m$ where $\lambda_0 = 0$ and a part $t_m < t < +\infty$ where λ_0 can be used as new integration variable. The equations are then easily solved numerically for any value of α and s.

The most interesting point is the behaviour of $\varepsilon(\alpha)$ for $\alpha \to \infty$. Here we proceed as in Appendix C . For $s \geq \frac{1}{2}$, we obtain

$$\varepsilon(\alpha) \sim \frac{1}{\alpha} \tag{2.109}$$

so that we saturate the lower bound ε^* (2.106). For smaller values of s however, the asymptotic behavior is like A/α with the proportionality factor A larger than 1 and reaching a maximum of approximately 1.28 in the limit $s \to 0$, cf. Fig.(2.19) .

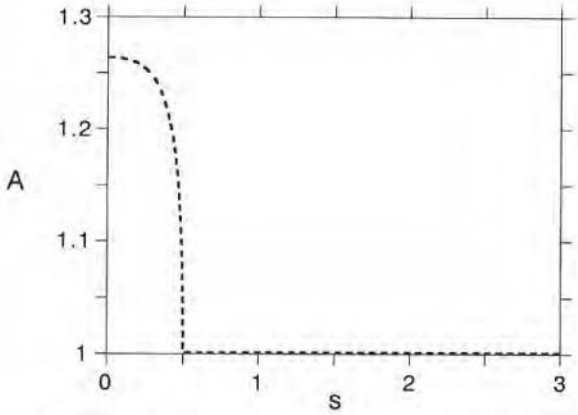

Fig.(2.19): Proportionality constant $A(s)$ for the class of attractive potentials (2.97). The broken line indicates that the replica symmetric solution is unstable (for finite α).

This result should be compared with the result $1.5/\alpha$ for the worst student of the version space, cf. Eq.(2.96).
The results of the present section have been obtained assuming Replica Symmetry. It is however clear that replica symmetry must be broken for $0 < s < 1$, where the function $\lambda_0(t)$ has a discontinuity in function of t. Furthermore, a numerical evaluation of the integral appearing in (2.17) leads to the conclusion that replica symmetry is broken for all values of $s > 0$, except for the limiting values $\alpha \to 1$ and $\alpha \to \infty$, where the replica symmetric solution is marginally stable. A similar behaviour of replica symmetry breaking between two limiting values of α was also observed for the worst case scenario [25] . In this case, it was proven that replica symmetry and one-step replica symmetry breaking lead to an identical asymptotic behavior of the generalization error for $\alpha \to \infty$, which is therefore believed to be exact. For the same reason, we expect that the asymptotic results derived above may also be exact.

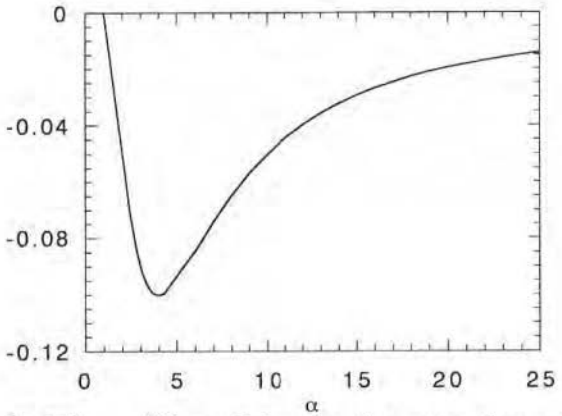

Fig.(2.20): AT condition for the attractive potential (2.97) with $s = 2$. Negativity of the function means that the Replica Symmetry is broken.

2.5 Overlaps between student vectors

In the previous sections several cost functions $E(\mathbf{J})$ have been proposed which after minimization lead to student vectors $\mathbf{J}_E$. Since the Bayes vector has been defined as the center of mass of the version space , it is useful to calculate the overlap S of the vector $\mathbf{J}_E$ with the vector $\mathbf{J}_{CM}$. The center of mass of the version space $\mathbf{J}_{CM}$ is defined as :

$$\mathbf{J}_{CM} = \frac{\sqrt{N}\int_V dm(\mathbf{J})\,\mathbf{J}}{\sqrt{\int_V dm(\mathbf{J})\int_V dm(\mathbf{J}')\,\mathbf{J}.\mathbf{J}'}} \tag{2.110}$$

with

$$\int_V dm(\mathbf{J}) \sim \int_{-\infty}^{+\infty} d\mathbf{J}\;\delta(\mathbf{J}^2 - N)\prod_{\mu=1}^{P}\theta\left(\frac{\mathbf{J}.\boldsymbol{\xi}^\mu}{\sqrt{N}}\,\mathrm{sign}\left(\frac{\mathbf{T}.\boldsymbol{\xi}^\mu}{\sqrt{N}}\right)\right) \tag{2.111}$$

The overlap S can be written as :

$$S = \left\langle \frac{\mathbf{J}_{CM}.\mathbf{J}_E}{N} \right\rangle_{\xi} \tag{2.112}$$

$$= \frac{1}{\sqrt{q}} \left\langle \frac{\mathbf{J}_{av}.\mathbf{J}_E}{N} \right\rangle_{\xi} \tag{2.113}$$

with

$$\mathbf{J}_{av} = \int_V dm(\mathbf{J})\ \mathbf{J} \tag{2.114}$$

and q the typical overlap of two members of the version space and $< . >_{\xi}$ the quenched average over the pattern set. (2.113) follows from the fact that $\mathbf{J}_{av}$ is not a properly normalised vector ($\mathbf{J}_{av}^2 \neq N$). Since we expect our problem to be self averaging the norm of this vector should tend to q.

The overlap S can be calculated using the replica method [29]. The replicated partition function Z^n can be defined as :

$$Z^n = \int_V \prod_{\sigma=1}^{n} d\mathbf{J}_{\sigma}^{*} \int \prod_{\sigma'=1}^{n} d\mu(\mathbf{J}_{\sigma'}) \exp\left(-\beta \sum_{\sigma'=1}^{n} E(\mathbf{J}_{\sigma'})\right) \tag{2.115}$$

where we have introduced two sets of replicas :

$$\mathbf{J}_1, \mathbf{J}_2, \mathbf{J}_3, \ldots, \mathbf{J}_n, \tag{2.116}$$

and

$$\mathbf{J}_1^{*}, \mathbf{J}_2^{*}, \mathbf{J}_3^{*}, \ldots, \mathbf{J}_n^{*}. \tag{2.117}$$

During the replica calculation (for details see Appendix D), the following order parameters have to be introduced (together with their conjugates) :

$$q_{ab} = \frac{\mathbf{J}_a.\mathbf{J}_b}{N} \quad \forall a < b = 1, \ldots, n \tag{2.118}$$

$$r_a = \frac{\mathbf{J}_a.\mathbf{T}}{N} \quad \forall a = 1, \ldots, n \tag{2.119}$$

$$\tilde{q}_{ab} = \frac{\mathbf{J}_a^*.\mathbf{J}_b^*}{N} \quad \forall a < b = 1, \ldots, n \tag{2.120}$$

$$\tilde{r}_a = \frac{\mathbf{J}_a^*.\mathbf{T}}{N} \quad \forall a = 1, \ldots, n \tag{2.121}$$

The order parameters which express the overlap between the two sets of replicas :

$$\tilde{s}_{ab} = \frac{\mathbf{J}_a.\mathbf{J}_b^*}{N} \quad \forall a, b = 1, \ldots, n \tag{2.122}$$

After introduction of the Replica Symmetry ansatz and taking the limit $n \to 0$, the free energy f associated to this problem can be written as :

$$-\beta f = -\beta f_1 - \beta f_2 \tag{2.123}$$

with

f_1 : the free energy of the Gibbs problem [3].

f_2 : the free energy of the generalisation problem with an arbitrary cost function $E(\mathbf{J})$ (section 2.1).

One also finds that (2.123) is independent of the overlap parameter $\tilde{S}$ (and its conjugate $\hat{\tilde{S}}$). To determine $\tilde{S}$ one should derive the saddle point equations before taking the limit $n \to 0$. After eliminating $\hat{\tilde{S}}$ one gets :

$$\tilde{S} - R\tilde{R} = 2\alpha\sqrt{\frac{1-\tilde{q}}{2\pi}} \int_0^{+\infty} \mathcal{D}y \int_{-\infty}^{+\infty} \mathcal{D}t \int_{-\infty}^{+\infty} \mathcal{D}t'$$
$$. \left(\lambda_0(x, R, y, t) - yR + \sqrt{1-R^2}t\right) \frac{\exp\left(-\frac{1}{2}g^2(\tilde{q}, \tilde{R}, y, t, t')\right)}{H\left(g(\tilde{q}, \tilde{R}, y, t, t')\right)} \tag{2.124}$$

where we have used the expression :

$$g(\tilde{q}, \tilde{R}, y, t, t') = \frac{-y\tilde{R} + \sqrt{\tilde{q} - \tilde{R}^2}(at + \sqrt{1-a^2}t')}{\sqrt{1-\tilde{q}}}, \tag{2.125}$$

with

$$a = \frac{\tilde{S} - R\tilde{R}}{\sqrt{1-R^2}\sqrt{\tilde{q} - \tilde{R}^2}}. \tag{2.126}$$

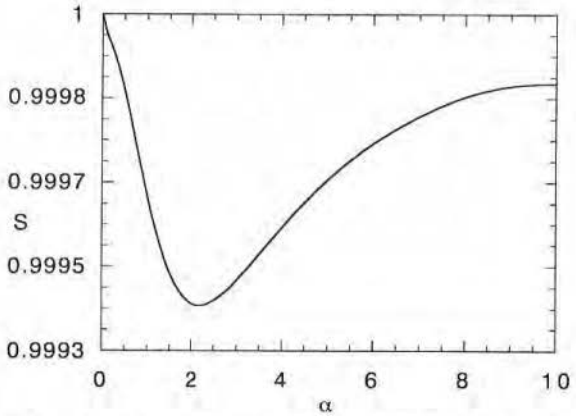

Fig.(2.21): Overlap S between center of mass of the version space and the student constructed using the optimal repulsive potential, (2.85) with $s = -1.35$.

Note that we let $\beta \to +\infty$ and as in section 2.1 introduce $x = \beta(1-q)$. The function $\lambda_0(x, R, y, t)$ minimizes the expression :

$$V(\lambda) + \frac{(\lambda - yR + \sqrt{1-R^2}t)^2}{2x}. \tag{2.127}$$

Remark that the order parameter $\tilde{S}$ gives the overlap of $\mathbf{J}_E$ with $\mathbf{J}_{av}$. To obtain the overlap with the center of mass one still has to divide by

$\sqrt{q}$:

$$S = \frac{\tilde{S}}{\sqrt{q}}. \tag{2.128}$$

The overlap between the center of mass of the version space and the vector obtained by minimizing the cost function (2.85) with optimal s ($s = -1.35$) has been calculated and plotted in Fig. (2.21). One observes that the overlap S is very close to 1 which leads to the conclusion that the cost function (2.85) with optimal s ($s = -1.35$) attains it minimum very close to the center of mass of the version space.

In Fig. (2.22) some other results are plotted. One observes that even a repulsive potential V_s^+ with $s = 0.25$ leads to an overlap S with the center of mass which is larger than 0.98 for all α values.

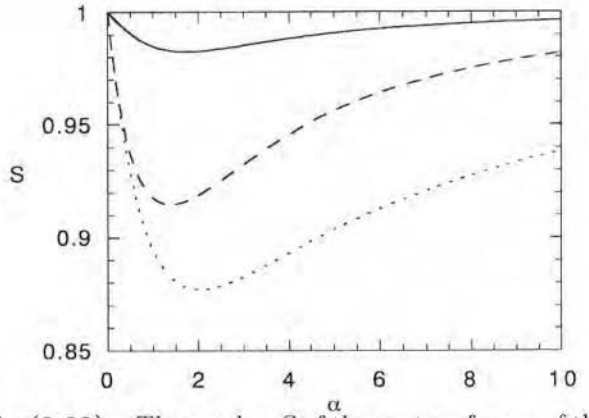

Fig.(2.22): The overlap S of the center of mass of the version space with the Hebb vector (dotted line) and with the vector obtained by using the Hebb rule inside the version space (dashed line). The result for the repulsive potential V_s^+ with $s = 0.25$ (full line) is also displayed.

Following the same procedure one can calculate the overlap of two vectors $\mathbf{J}_{E_1}$ and $\mathbf{J}_{E_2}$ which respectively minimize the cost functions E_1 an

E_2.

$$S = \left\langle \frac{\mathbf{J}_{E_1}.\mathbf{J}_{E_2}}{N} \right\rangle_{\boldsymbol{\xi}}. \tag{2.129}$$

Using an analogue derivation as the one explained in Appendix D leads to :

$$\begin{aligned} S - R\tilde{R} = 2\alpha\sqrt{\frac{1-\tilde{q}}{2\pi}} \int_0^{+\infty} \mathcal{D}y \int_{-\infty}^{+\infty} \mathcal{D}t \int_{-\infty}^{+\infty} \mathcal{D}t' \\ .\left(\lambda_1(x,R,y,t) - yR + \sqrt{1-R^2}t\right) \\ .\left(\lambda_2(\tilde{x},\tilde{R},y,t,t') - y\tilde{R} + \sqrt{1-\tilde{R}^2}(at + \sqrt{1-a^2}t')\right) \end{aligned} \tag{2.130}$$

with

$$a = \frac{S - R\tilde{R}}{\sqrt{1-R^2}\sqrt{\tilde{q}-\tilde{R}^2}} \tag{2.131}$$

The function $\lambda_1(x,R,y,t)$ minimizes the expression :

$$V_1(\lambda) + \frac{(\lambda - yR + \sqrt{1-R^2}t)^2}{2x}, \tag{2.132}$$

and $\lambda_2(\tilde{x},\tilde{R},y,t,t')$ minimizes:

$$V_2(\lambda) + \frac{\left(\lambda - y\tilde{R} + \sqrt{1-\tilde{R}^2}(at + \sqrt{1-a^2}t')\right)^2}{2\tilde{x}}. \tag{2.133}$$

As an illustration we plotted in Fig. (2.23) the overlap between the Hebb-vector and the student defined by (2.63) (Hebb rule inside the version space) and also the overlap between the latter and the student generated by minimizing the cost function (2.85) with $s = 0.25$.

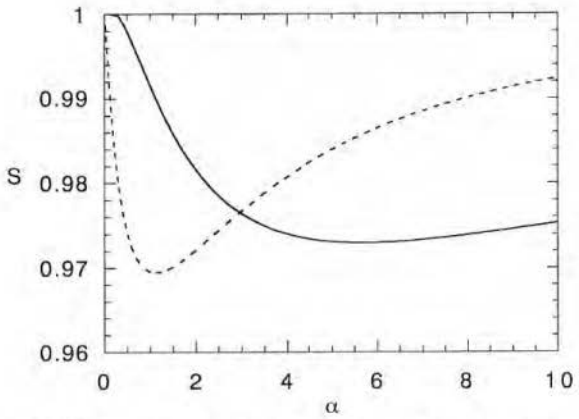

Fig.(2.23): The overlap S between the student defined by (2.63) (Hebb rule inside the version space) and the Hebb vector (full line) and secondly the overlap with the student generated by minimizing the cost function (2.85) with $s = 0.25$ (dashed line).

Chapter 3

Learning from non-uniformly distributed examples.

3.1 The gaussian model.

In Chapter 2 we considered an unstructured distribution of the example patterns, i.e. the training set is drawn from a uniform distribution. In many practical situations however, the distribution of the example patterns will be non-uniform.
Let us assume that the set of patterns $\boldsymbol{\xi}^\mu$, $\mu = 1, \ldots, P$ is generated by P independent samplings from a non-uniform distribution $P^*(\boldsymbol{\xi}|\mathbf{C})$ where $\mathbf{C}$ represents a symmetry breaking orientation. In this chapter we will restrict ourselves to the case where the teacher vector $\mathbf{T}$ coincides with the structure generating vector $\mathbf{C}$. Under assumption of cilindrical symmetry around the $\mathbf{T}$-axis, one can write the probability distribution of the patterns as :

$$P^*(\boldsymbol{\xi}|\mathbf{T}) \sim \delta\left(\boldsymbol{\xi}^2 - N\right) \exp\left(-V^*\left(\frac{\boldsymbol{\xi}.\mathbf{T}}{\sqrt{N}}\right)\right). \tag{3.1}$$

The corresponding distribution of the overlap h between pattern and teacher:

$$h = \frac{\boldsymbol{\xi}.\mathbf{T}}{\sqrt{N}}, \tag{3.2}$$

is given by :

$$P^*(h) \sim \int_{-\infty}^{+\infty} d\boldsymbol{\xi}\ \delta(h - \frac{\boldsymbol{\xi}.\mathbf{T}}{\sqrt{N}})\ P^*(\boldsymbol{\xi}|\mathbf{T}) \tag{3.3}$$

$$\sim \exp\left(-\frac{h^2}{2} - V^*(h)\right). \tag{3.4}$$

In particular for a uniform pattern distribution one has $V^*(h) = 0$. For the perceptron the classification by the teacher of a pattern $\boldsymbol{\xi}^\mu$ as ξ_0^μ automatically implies that the pattern $-\boldsymbol{\xi}^\mu$ is classified as $-\xi_0^\mu$. One can thus consider the set of P alligned patterns

$$\boldsymbol{\xi}^\mu \operatorname{sign}(\frac{\boldsymbol{\xi}^\mu.\mathbf{T}}{\sqrt{N}}) \qquad \forall\, \mu = 1, \ldots, P \tag{3.5}$$

all of which are classified as $+1$.
The distribution $P^{**}(u)$ with

$$u = \frac{|\mathbf{T}.\boldsymbol{\xi}^\mu|}{\sqrt{N}} \tag{3.6}$$

of the overlap of these aligned patterns with the teacher vector is given by :

$$P^{**}(u) = [P^*(u) + P^*(-u)]\,\theta(u). \tag{3.7}$$

If the potential $V^*(h)$ is even one gets :

$$P^{**}(u) = 2\,\theta(u)\,P^*(u). \tag{3.8}$$

As in [32, 33] we will concentrate on the so called gaussian model where the patterns are distributed according to the potential :

$$V^*(h) = a\frac{h^2}{2} \tag{3.9}$$

with a a real parameter which satisfies $-1 < a < +\infty$. The potential (3.9) leads to the following distribution for the overlap u :

$$P^{**}(u) = 2\,\theta(u)\,\sqrt{\frac{1+a}{2\pi}}\,\exp\left(-\frac{(1+a)\,u^2}{2}\right). \tag{3.10}$$

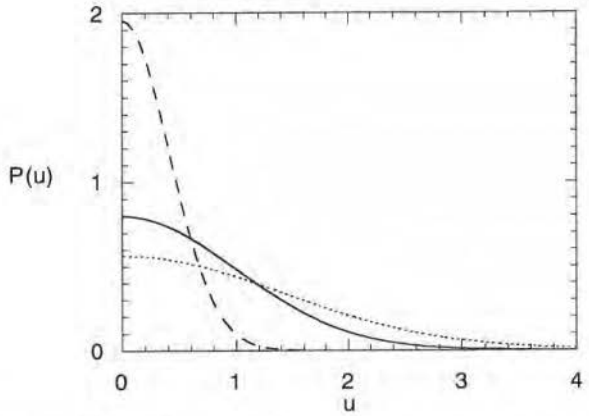

Fig.(3.1): The probability distribution (3.10) for $a = 0$ (full curve), $a = -0.5$ (pointed curve) and $a = 5.0$ (dashed curve).

For $a = 0$ one obtains uniformly distributed patterns. For $a > 0$ the patterns are more densely distributed around the plane orthogonal to the teacher $\mathbf{T}$ (equator). In the $a < 0$ case this zone is less densely populated. Following [34] one can characterize the $a > 0$ case as the problem in which the student has to learn from "difficult" patterns. For $a < 0$ the teacher presents "easy" examples to the student.

3.2 Gibbs and Bayes learning.

P.Reimann and C.Van den Broeck [32] studied Gibbs and Bayes supervised learning from non-uniformly distributed examples by mapping the supervised problem to an unsupervised one. The task in unsupervised learning [35, 36, 37] is to discover a certain distribution from the available examples. This can only be done by using some a-priori knowledge about the structure which has to be inferred. If the form of the a-priori

probability distribution of the patterns (3.1) is known exactly one can obtain the so called a posteriori probability distribution [32]:

$$P(\mathbf{J}|\{\boldsymbol{\xi}^{\mu}\}) \sim \exp\left(-\sum_{\mu=1}^{P} V^{*}\left(\frac{\mathbf{J}.\boldsymbol{\xi}^{\mu}}{\sqrt{N}}\right)\right)\delta(\mathbf{J}.\mathbf{J}-N), \tag{3.11}$$

which is the probability that a particular hypothesis (student) vector $\mathbf{J}$ coincides with the unknown teacher vector $\mathbf{T}$ given the set of example patterns $\{\boldsymbol{\xi}^{\mu}\}$.
The vector which characterizes the Gibbs perceptron is defined as a random sample from the distribution (3.11). The Bayes-vector [36] is defined as the (weighted) center of mass of the distribution (3.11) i.e.

$$\mathbf{J}_B \sim \int d\mathbf{J}\ \mathbf{J}\ P(\mathbf{J}|\{\boldsymbol{\xi}^{\mu}\}) \tag{3.12}$$

with the proper normalisation constant ($\mathbf{J}_B^2 = N$).
For the gaussian model, a replica calculation leads to the following saddle point equation [32] which determines the overlap R_{Gibbs} between the Gibbs perceptron and the teacher :

$$1+a(1-R) = \alpha\left(\frac{a^2(1-R)}{1+a} + \frac{1}{\pi R}\int_{-\infty}^{+\infty} \mathcal{D}z \frac{\exp\left(-\frac{Rz^2}{(1-R)(1+a)}\right)}{H\left(-\sqrt{\frac{R}{(1-R)(1+a)}}z\right)}\right) \tag{3.13}$$

Solving this equation numerically leads to R_{Gibbs} and R_{Bayes} since the relation:

$$R_{Bayes} = \sqrt{R_{Gibbs}} \tag{3.14}$$

stays valid [32].
The small and large α behaviour of R_{Bayes} are given by :

$$\alpha \to 0 \quad : \quad R_{Bayes} = \sqrt{\frac{2\alpha}{\pi(1+a)}} + O(\alpha^2) \tag{3.15}$$

$$\alpha \to +\infty \quad : \quad R_{Bayes} = 1 - \frac{\pi^2}{2}\frac{(0.442)^2}{(1+a)\alpha^2}. \tag{3.16}$$

For $a = 0$, one recovers the well known results for uniformly distributed patterns. As in [34] one can see from (3.15) and (3.16) that an ideal student learns best on the basis of easy examples for small α but in order to obtain optimal generalisation properties for large α one should present example patterns characterized by a large positive a (difficult examples).

3.3 Calculation of the overlap with the teacher using gradient descent learning.

The results from the previous section for optimal and Gibbs learning have been obtained by explicitly taking into account the knowledge about the pattern distribution $P^*(\boldsymbol{\xi}|\mathbf{T})$. In practice one does not know if the example patterns generated by the teacher are correlated . Therefore it is interesting to calculate the performance of some learning rules which do not use the additional information about the distribution from which the training patterns are drawn.
We will apply our general method of Chapter 2 to calculate the overlap R between the teacher $\mathbf{T}$ and the student which minimizes the cost function :

$$E(\mathbf{J}) = \sum_{\mu=1}^{P} V(\lambda^{\mu}) \tag{3.17}$$

with λ^{μ} the stabilty of the μ-th pattern.
A replica calculation very similar to the one performed for the case of uniformly distributed patterns (only the averaging over the patterns is different) leads to the following expression for the free energy :

$$-f = \underset{q,R}{\text{extr}} \Big(\frac{q - R^2}{2\beta(1-q)} + \frac{1}{2\beta} \ln(1-q) - \frac{\alpha}{\beta} \int_{-\infty}^{+\infty} \mathcal{D}^{**}t_2 \int_{-\infty}^{+\infty} \mathcal{D}t_1 \ln \int_{-\infty}^{+\infty} \frac{d\lambda}{\sqrt{2\pi(1-q)}} \exp\left(g(t_1, t_2, q, R, \lambda)\right) \Big)$$

(3.18)

with

$$g(t_1, t_2, q, R, \lambda) = -\beta V(\lambda \ \mathrm{sign}(t_2)) - \frac{\left(\lambda - Rt_2 - \sqrt{q - R^2}t_1\right)^2}{2(1-q)} \quad (3.19)$$

and

$$\mathcal{D}^{**}t_2 = P^{**}(t_2)\, dt_2. \quad (3.20)$$

Remark that, apart from the integration with respect to t_2,equation (3.18) is equal to the expression (2.6).
If one only considers cost functions $E(\mathbf{J})$ with a unique non-degenerate minimum one should let the temperature T go to zero in order to find the ground state energy. Introducing the order parameter $x = \beta(1-q)$ into expression (3.18) leads to :

$$-f^{T=0} = \underset{x,R}{\mathrm{extr}} \left(\frac{1-R^2}{2x} - \alpha \int_{-\infty}^{+\infty} \mathcal{D}^{**}t_2 \int_{-\infty}^{+\infty} \mathcal{D}t_1 \ \min_{\lambda} \left[V(\lambda) + \frac{(\lambda - v)^2}{2x} \right] \right) \quad (3.21)$$

with

$$v = Rt_2 + \sqrt{1 - R^2}t_1. \quad (3.22)$$

The saddle point equations yield :

$$R = \alpha \int_{-\infty}^{+\infty} \mathcal{D}^{**}t_2 \int_{-\infty}^{+\infty} \mathcal{D}t_1 \ (\lambda_0 - v) \left(t_2 - \frac{Rt_1}{\sqrt{1-R^2}} \right) \quad (3.23)$$

$$1 - R^2 = \alpha \int_{-\infty}^{+\infty} \mathcal{D}^{**}t_2 \int_{-\infty}^{+\infty} \mathcal{D}t_1 \ (\lambda_0 - v)^2 \quad (3.24)$$

where the function $\lambda_0(v, x)$ minimizes the expression

$$V(\lambda) + \frac{(\lambda - v)^2}{2x}. \quad (3.25)$$

The Almeida-Thouless stability condition [16] for the Replica Symmetric saddle point takes on the following form :

$$\alpha \int_{-\infty}^{+\infty} \mathcal{D}^{**}t_2 \int_{-\infty}^{+\infty} \mathcal{D}t_1 \left(\frac{\partial \lambda_0}{\partial v} - 1 \right)^2 < 1. \tag{3.26}$$

3.4 Hebb and maximal stability learning.

3.4.1 Hebb learning.

The Hebb potential is :

$$V(\lambda) = -\lambda. \tag{3.27}$$

The corresponding function λ_0 has been determined in section 2.2. Inserting λ_0 in the saddle point equations and using the probability distribution $P^{**}(t)$ (3.10) for the gaussian model leads to :

$$R = \sqrt{\frac{2}{\pi}} \frac{\alpha x}{\sqrt{1+a}}, \tag{3.28}$$

$$1 - R^2 = 2\alpha x^2. \tag{3.29}$$

Elimination of x results in :

$$R_{Hebb}(\alpha) = \sqrt{\frac{2\alpha}{\pi(1+a)+2\alpha}}. \tag{3.30}$$

For large α the overlap R goes to 1 as :

$$\alpha \to +\infty \quad : \quad R_{Hebb} \sim 1 - \frac{\pi(1+a)}{4\alpha}. \tag{3.31}$$

3.4.2 Maximal stability learning.

Maximal stability learning (Adatron algorithm) can be described by considering the potential :

$$V(\lambda) = \begin{cases} +\infty & \lambda < \kappa \\ 0 & \lambda > \kappa \end{cases} \tag{3.32}$$

This leads to the function $\lambda_0(v,x)$:

$$\lambda_0(v,x) = v + (\kappa - v)\,\theta(\kappa - v). \tag{3.33}$$

Inserting this function into the saddle point equations (3.23) and (3.24) gives :

$$R = \alpha \int_{-\infty}^{+\infty} \mathcal{D}^{**} t_2 \int_{-\infty}^{+\infty} \mathcal{D}t_1 (t_2 - \frac{Rt_1}{\sqrt{1-R^2}})(\kappa - v)\,\theta(\kappa - v) \tag{3.34}$$

$$1 - R^2 = \alpha \int_{-\infty}^{+\infty} \mathcal{D}^{**} t_2 \int_{-\infty}^{+\infty} \mathcal{D}t_1 (\kappa - v)^2\,\theta(\kappa - v). \tag{3.35}$$

Numerically solving these equations leads to $\kappa(\alpha)$ and $R(\alpha)$ corresponding with the maximal stability criterion.

3.4.3 Numerical results.

In Fig.(3.2) we present numerical results for the gaussian model with $a = -0.5$. The overlap obtained by using the Hebb and Adatron rule are compared with the results for Gibbs and Bayes learning. One observes that for $\alpha < 5.0$ the Hebb rule is superior to the Adatron rule and produces overlaps which are close to optimal. This can be understood by noting that for $a < 0$ the probability that a pattern has a large overlap with the teacher is higher than in the case of the uniform distributed patterns ($a = 0$). Since the Hebb rule is nothing more than the (normalised) sum of the aligned patterns $\boldsymbol{\xi}^\mu \xi_0^\mu$ this algorithm should lead to large overlaps with the teacher. For $\alpha \approx 5$ one sees that the curves for R_{Hebb} and R_{MS} tend to each other and cross for larger α.

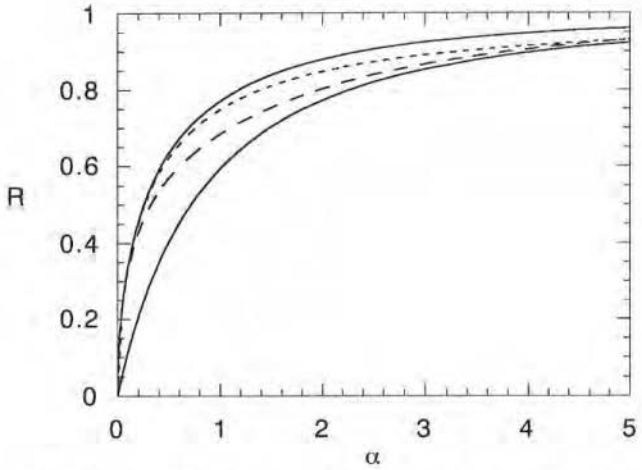

Fig.(3.2): The overlap between student and teacher for the Bayes (upper full curve) and Gibbs (lower full curve) rule for the gaussian model with $a = -0.5$ together with the overlap obtained by using the Hebb (short dashes) and Adatron (long dashes) rule.

In Fig.(3.3) the results for the gaussian model with $a = 5$ are presented.Remind that for $a > 0$ the patterns are lying close to the plane orthogonal to the teacher. One observes that the Hebb rule in this case does not lead to large overlaps. The Adatron algorithm gives larger overlaps than the Hebb rule but for $0.5 < \alpha < 3$ the difference with the optimal overlap is rather large. Therefore it is interesting to investigate if the class of repulsive potentials defined in Chapter 2 produce overlaps which are closer to the optimal ones.

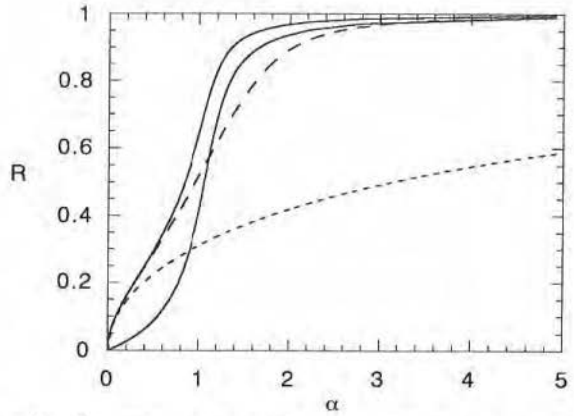

Fig.(3.3): The overlap between student and teacher for the Bayes (upper full curve) and Gibbs (lower full curve) rule for the gaussian model with $a = 5$ together with the overlap obtained by using the Hebb (short dashes) and Adatron (long dashes) rule.

3.5 Gradient descent learning with a repulsive potential.

The class of repulsive potentials defined as :

$$V_s^+(\lambda) = \begin{cases} +\infty & \lambda < 0 \\ -\frac{\lambda^s}{s} & \lambda > 0, \end{cases} \tag{3.36}$$

lead for uniformly distributed example patterns and optimal choice of the parameter s to a student perceptron with almost optimal generalisation properties. Using the general method explained in the previous section, we can calculate the overlap of the student vector obtained by minimizing the cost functions corresponding with the potentials (3.36)

and the teacher for the gaussian model .

First we rewrite the saddle point equations (3.23) and (3.24) by transforming the integration variables t_1 and t_2. If one carries out the transformation $t_2 \rightarrow \sqrt{1+a}\ t_2$ and then introduces the new integration variables t and t' by applying the orthogonal transformation :

$$t = \frac{1}{b}\left(\frac{R}{\sqrt{1+a}}t_2 + \sqrt{1-R^2}t_1\right) \tag{3.37}$$

$$t' = \frac{1}{b}\left(\sqrt{1-R^2}t_2 - \frac{R}{\sqrt{1+a}}t_1\right) \tag{3.38}$$

with

$$b^2 = \frac{1+a(1-R^2)}{1+a}, \tag{3.39}$$

one gets for the saddle point equations :

$$R = 2\alpha \int_{-\infty}^{+\infty} \mathcal{D}t \int_{-\frac{Rt}{\sqrt{(1+a)(1-R^2)}}}^{+\infty} \mathcal{D}t'\ (\lambda_0(bt,x) - bt) \cdot\frac{1}{b}\left(-\frac{aR}{1+a}t + \frac{t'}{\sqrt{(1+a)(1-R^2)}}\right) \tag{3.40}$$

$$1 - R^2 = 2\alpha \int_{-\infty}^{+\infty} \mathcal{D}t\ H\left(-\frac{Rt}{\sqrt{(1+a)(1-R^2)}}\right)(\lambda_0(bt,x) - bt)^2 \tag{3.41}$$

The function $\lambda_0(bt,x)$ now minimizes the expression :

$$V(\lambda) + \frac{(\lambda - bt)^2}{2x}. \tag{3.42}$$

Extremizing (3.42) using (3.36) defines $\lambda_0(bt,x)$:

$$\lambda_0 - bt = x\lambda_0^{s-1} \qquad (\lambda_0 > 0). \tag{3.43}$$

The saddle point equations (3.40) and (3.41) can be solved numerically by applying the same techniques as the ones proposed for the case of the uniformly distributed example patterns. The results for $s = -1.35$ are presented in the next figures .

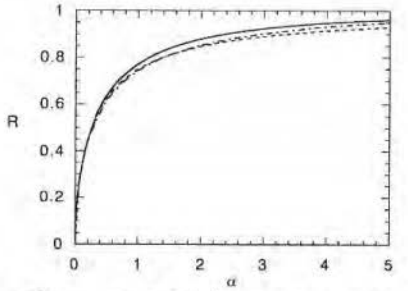

Fig.(3.4): The overlap $R(\alpha)$ between student and teacher for the Bayes rule (full curve) and Hebb rule (short dashes) together with the result obtained by minimizing the repulsive potential with $s = -1.35$ for the gaussian model with $a = -0.5$.

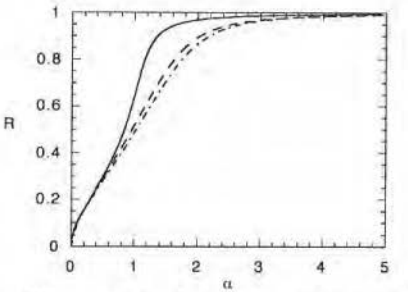

Fig.(3.5): The overlap between student and teacher for the Bayes rule (full curve) and the Adatron rule (long dashes) together with the overlap obtained by minimizing the repulsive potential with $s = -1.35$ for the gaussian model with $a = 5$.

One sees from Fig.(3.4) that for $a = -0.5$ the repulsive potential leads to overlaps which are almost equal to those of the Hebb rule for small α. For bigger α the overlap R becomes very close to the optimal one. From Fig.(3.5) we see that for $a = 5$ the minimization of the repulsive potential with $s = -1.35$ gives overlaps R which, for $\alpha < 5$, are smaller than the ones obtained by using the Adatron algorithm.

Closer examination of the numerical results for different s values lets us conclude that for a different from zero, the optimal value for s is α dependent. The optimal value of s has been calculated numerically for $a = 5$ and is depicted in Fig.(3.6). One gets that for small α one should use a large negative value for s. For large values of α one observes that s_{opt} tends to a value close to -1.5 .

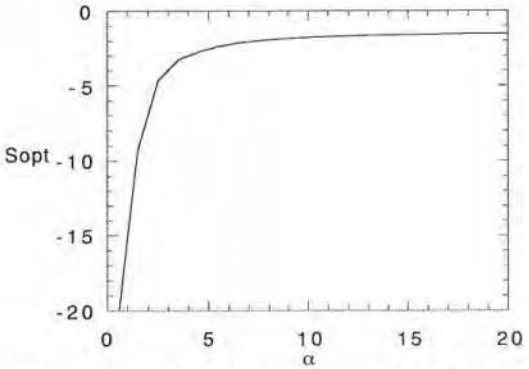

Fig.(3.6): The optimal value for s as a function of α for the gaussian model with $a = 5$.

The overlap R obtained by using s_{opt} , which is plotted in Fig (3.7) together with the $s = -1.35$ curve, is slightly higher than the one of the Adatron algorithm but still is not satisfying for intermediate values of α.

Remark that although for small α values the optimal value s_{opt} differs the most from -1.35 we do not obtain a overlap which is significantly better. Only in the region $0.5 < \alpha < 3$ we obtain a higher overlap by using lower s values.

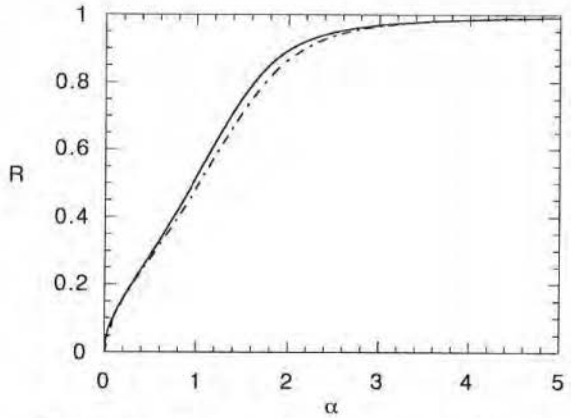

Fig.(3.7): The overlap R obtained by using the optimal value $s_{opt}(\alpha)$ (full curve) together with the result obtained by using $s = -1.35$.

To determine s_{opt} for large α we proceed by using the same methods as in the case for uniform distributed patterns. As explained in Appendix C for uniform distributed examples, we can rewrite the saddle point equations (3.40) and (3.41) as:

$$\frac{1}{2A'} = \frac{B'^2}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \frac{dz}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{B'^2 z^2}{1+a}\right) h(z) \tag{3.44}$$

$$\frac{1}{2A'} = \frac{B'^3}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} dz H\left(-\frac{B'z}{\sqrt{1+a}}\right) \left(h(z) - \frac{z}{\sqrt{1+a}}\right)^2 \tag{3.45}$$

with

$$A' = \frac{\alpha\sqrt{1-R^2}}{\pi} \tag{3.46}$$

$$B' = \frac{x^{\frac{1}{2-s}}}{\sqrt{1-R^2}} \tag{3.47}$$

Taking the $R \to 1$ limit leads to the equations (C.15) and (C.16) of Appendix C with

$$A = \frac{A'}{\sqrt{1+a}}. \tag{3.48}$$

By numerically solving the equations (C.15) and (C.16) we have already calculated the $A(s)$ curve. the function $A(s)$ attains its minimal value 0.443 at $s = -1.35$. As a result we can conclude that for large values of α the repulsive potential V_s^+ with $s = -1.35$ still leads to overlaps between student and teacher which are close to optimal, i.e.

$$\alpha \to +\infty \quad : \quad R = 1 - \frac{\pi^2}{2}\frac{(0.443)^2}{(1+a)\alpha^2}. \tag{3.49}$$

Chapter 4

The Ising teacher problem.

4.1 Introduction

One of the main reasons to describe neural networks with Ising or more generally with discrete valued synapses is the possibility of implementing these systems in hardware. Therefore, the analysis and properties of these networks and the search for learning methods capable of working under these limitations are important research subjects. Clearly the discretization of the connections modifies the learning problem completely because the topology of the phase space differs. Indeed, all Ising vectors are the corners of an N-dimensional hypercube. We will again address the problem of a student perceptron $\mathbf{J}$ which learns from a set of randomly chosen training examples $\boldsymbol{\xi}^\mu$, $\mu = 1, \ldots, P$ whose classification ξ_0^μ is provided by the teacher perceptron $\mathbf{T}$. Now, however, the teacher $\mathbf{T}$ has Ising synapses, i.e.

$$T_i = \pm 1 \quad \forall i = 1, \ldots, N. \tag{4.1}$$

The generalisation error for the Ising Gibbs perceptron, which is obtained by random sampling from the set of Ising vectors belonging to the version space, has been studied in [41] for the zero temperature case and in [3] for the case with temperature different from zero.
For zero temperature one observes a first order phase transition to perfect generalisation ($\varepsilon = 0$) at $\alpha_c = 1.24$. This means that beyond α_c, only one Ising vector lies in the version space which obviously must be

the teacher **T**. The Ising Bayes vector has been defined by Opper in [42] as the center of mass of all Ising Gibbs vectors and in general does not has binary components . As in the case with continuous valued perceptron vectors the generalisation error ε is equal to:

$$\varepsilon_{Bayes} = \frac{1}{\pi} \arccos\left(\sqrt{R_{Gibbs}}\right). \tag{4.2}$$

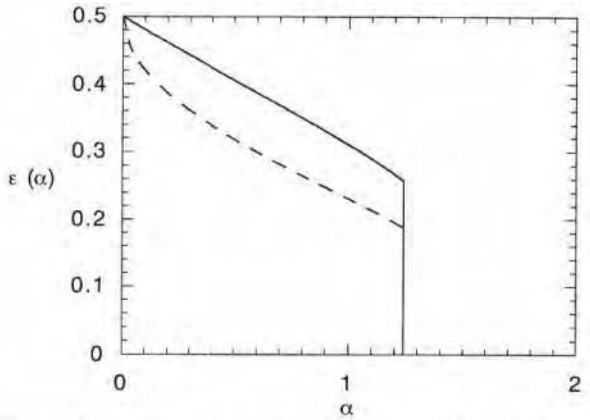

Fig.(4.1): The generalisation error ε for the Ising Gibbs (full curve) and Ising Bayes (dashed curve) perceptron.

Even though it is known that zero generalisation error can be achieved for $\alpha > \alpha_c$ there are no practical algorithms that come near to this result.

As in Chapter 2 we can construct a student vector with continuous components using a certain cost function. But, one might expect that, if one uses the extra information about the the teacher, a lower generalisation error can be achieved.

One of the learning strategies wich can be followed is to construct a student vector with continuous couplings using gradient descent learning and clip or transform its components .

A few simple clipping scenarios have already been studied. Van den Broeck and Bouten [38] have derived the generalisation error of the clipped Hebb perceptron. The synapses of the student are given by :

$$\tilde{J}_i = \text{sign}\left(\frac{1}{\sqrt{P}}\sum_{\mu=1}^{P}\xi_0^\mu\xi_i^\mu\right). \tag{4.3}$$

They obtained for the overlap between student and teacher :

$$\tilde{R} = \frac{\tilde{\mathbf{J}}.\mathbf{T}}{N} \tag{4.4}$$

$$= \text{erf}\left(\sqrt{\frac{\alpha}{\pi}}\right). \tag{4.5}$$

From (4.5) it is clear that no transition to perfect learning is achieved but for large values of α the generalisation error approaches zero exponentially fast.

$$\alpha \to +\infty \;:\; \varepsilon \sim \exp\left(-\frac{\alpha}{2\pi}\right). \tag{4.6}$$

Bollé and Shim [39] generalised the result (4.5) by considering a nonlinear modulation of the Hebb learning rule .

In the next sections we will show that the overlap $\tilde{R}$ of the vector obtained by clipping or transforming the continuous student vector $\mathbf{J}$ can be found as a simple function of the overlap R of $\mathbf{J}$ with the Ising teacher $\mathbf{T}$.

4.2 Clipping.

In [38] and [39] only the Hebb rule has been considered to construct the vector $\mathbf{J}$. Using a gradient descent algorithm one can construct the student vector which minimizes a certain cost function $E(\mathbf{J})$. As in Chapter 2 we will restrict ourselves to cost functions $E(\mathbf{J})$ with a unique non-degenerate minimum where the information about the patterns enters in an additive way :

$$E(\mathbf{J}) = \sum_{\mu=1}^{P} V(\lambda^{\mu}). \tag{4.7}$$

In view of the binary nature of the components of the teacher $\mathbf{T}$ one could expect that a higher overlap can be achieved by clipping the continuous components of the student vector. Following this idea the quantity of interest is $\tilde{R}$:

$$\tilde{R} = \left\langle \frac{1}{N} \sum_{i=1}^{N} \mathrm{sign}(J_i)\, T_i \right\rangle_{\boldsymbol{\xi}}. \tag{4.8}$$

As usual $< . >_{\boldsymbol{\xi}}$ stands for the quenched average over the set of example patterns. With the assumption that all order parameters are self-averaging in the thermodynamical limit a replica calculation has been carried out to determine $\tilde{R}$ as a function of α. In Appendix E, where we have also considered the case of a generic and easy teacher [38, 39], this calculation has been elaborated in details.

At the Replica Symmetric saddle point, one obtains:

$$\tilde{R} = \int_{-\infty}^{+\infty} \mathcal{D}z \left(H\left(-\frac{R - \sqrt{q - R^2}\, z}{\sqrt{1-q}} \right) - H\left(\frac{R + \sqrt{q - R^2}\, z}{\sqrt{1-q}} \right) \right), \tag{4.9}$$

where q is the typical overlap of 2 replicated continuous $\mathbf{J}$ vectors and R is the overlap of a $\mathbf{J}$ vector and the teacher $\mathbf{T}$. The values for this parameters are solely determined by extremizing the free energy (2.6) of the "continuous" problem.

To find the ground state of the function $E(\mathbf{J})$ one considers the $\beta \to +\infty$ limit. Taking into account the fact that the cost function $E(\mathbf{J})$ has a non-degenerate minimum, q should tend to 1. Introducing this limit in the equation (4.9) leads to :

$$\tilde{R} = H\left(-\frac{R}{\sqrt{1-R^2}}\right) - H\left(\frac{R}{\sqrt{1-R^2}}\right) \tag{4.10}$$

$$= \text{erf}\left(\frac{R}{\sqrt{2(1-R^2)}}\right). \tag{4.11}$$

We see that $\tilde{R}(\alpha)$ is determined by a simple expression in terms of $R(\alpha)$, independent of the cost function which was used to construct the continuous student vector $\mathbf{J}$. The function (4.11) is plotted in Fig.(4.2) together with the $\tilde{R} = R$ line.

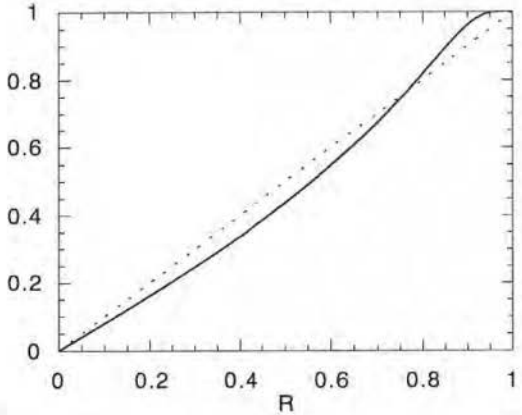

Fig.(4.2): The overlap $\tilde{R}$ of the clipped student $\tilde{\mathbf{J}}$ with the Ising teacher (full line) together with the original overlap R (dotted line).

One observes that the overlap $\tilde{R}$ is a monotonous increasing function of R. The overlap between student and teacher will increase through

clipping only if the prior overlap R is larger than 0.78. For smaller values of R clipping leads to a small decrease of the overlap.

The result (4.11) can also be obtained using a simple geometric approach. Suppose one applies a general odd function f (clipping is a special case) to the components of a student vector $\mathbf{J}$ which has been obtained by an arbitrary learning rule. The components of the properly normalized, transformed vector $\tilde{\mathbf{J}}$ are given by :

$$\tilde{J}_i = \frac{\sqrt{N} f(J_i)}{\sqrt{\sum_{j=1}^{N} f(J_i)^2}}. \tag{4.12}$$

Using the facts that f is an odd function and $\mathbf{T}$ is an Ising teacher, one can write the overlap $\tilde{R}$ between $\tilde{\mathbf{J}}$ and $\mathbf{T}$ as follows :

$$\tilde{R} = \left\langle \frac{1}{N} \sum_{i=1}^{N} \tilde{J}_i \, T_i \right\rangle_{\boldsymbol{\xi}} \tag{4.13}$$

$$= \left\langle \frac{\sum_{i=1}^{N} f(J_i T_i)}{\sqrt{N \sum_{j=1}^{N} f^2(J_i T_i)}} \right\rangle_{\boldsymbol{\xi}} \tag{4.14}$$

$$= \left\langle \frac{1/N \sum_{i=1}^{N} f(J_i T_i)}{\sqrt{1/N \sum_{j=1}^{N} f^2(J_i T_i)}} \right\rangle_{\boldsymbol{\xi}} . \tag{4.15}$$

If we assume $\tilde{R}$ to be self-averaging one can write, for $N \to +\infty$, using the law of large numbers :

$$\frac{1}{N} \sum_{i=1}^{N} f(t_i) = \langle f(t) \rangle \tag{4.16}$$

$$= \int_{-\infty}^{+\infty} dt \; P(t) \, f(t), \tag{4.17}$$

with $P(t)$ the probability density of $t = J_i\, T_i$.

This reduces Eq. (4.15) to :

$$\tilde{R} = \frac{\langle f(t) \rangle}{\sqrt{\langle f^2(t) \rangle}}. \tag{4.18}$$

If the overlap between the **J**-vector and the teacher vector **T** has a given value R, but all locations of the **J**-vector are otherwise equally probable, one finds the following result for $P(t)$:

$$P(t) = \frac{\int_{-\infty}^{+\infty} d\mathbf{J}\; \delta(J_1 T_1 - t)\; \delta(\mathbf{J}^2 - N)\; \delta(\mathbf{J}.\mathbf{T} - NR)}{\int_{-\infty}^{+\infty} d\mathbf{J}\; \delta(\mathbf{J}^2 - N)\; \delta(\mathbf{J}.\mathbf{T} - NR)} \tag{4.19}$$

$$= \frac{\exp\left(-\frac{1}{2}\frac{(t-R)^2}{1-R^2}\right)}{\sqrt{2\pi(1-R^2)}}. \tag{4.20}$$

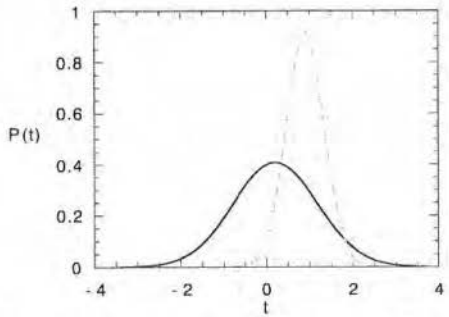

Fig.(4.3): The probability distribution $P(t)$ for $R = 0.2$ (full line) and $R = 0.9$ (dashed line).

The Equations (4.18) and (4.20) are the basic results of this section. They give the new overlap $\tilde{R}$ of the transformed vector $\tilde{\mathbf{J}}$ in terms of the old overlap R for any odd function f.
Returning to the plain clipping scenario we have:

$$f(t) = \text{sign}(t). \tag{4.21}$$

The overlap of the clipped student $\tilde{\mathbf{J}}$ with the teacher $\mathbf{T}$ follows from (4.18) and (4.20) :

$$\tilde{R} = \int_{-\infty}^{+\infty} dt\ P(t)\ \text{sign}(t) \tag{4.22}$$

$$= 2\ H\left(\frac{R}{\sqrt{1-R^2}}\right) - 1 \tag{4.23}$$

$$= \text{erf}\left(\frac{R}{\sqrt{2(1-R^2)}}\right), \tag{4.24}$$

which is equal to the result (4.11) obtained by performing the much longer replica calculation.
To obtain the generalisation error as a function of α, the learning algorithm used to construct the vector $\mathbf{J}$ has to be specified and the corresponding function $R(\alpha)$ has to be substituted in Eq. (4.24).
For example, the Hebb-student has :

$$R_{Hebb}(\alpha) = \sqrt{\frac{2\alpha}{2\alpha+\pi}}, \tag{4.25}$$

inserting this in (4.24) leads to

$$\tilde{R} = \text{erf}\left(\sqrt{\frac{\alpha}{\pi}}\right), \tag{4.26}$$

which is exactly the result derived in [38].

Since it is advantageous to clip starting from the continuous perceptron $\mathbf{J}$ with the largest possible overlap R, the largest possible $\tilde{R}$ can be obtained by clipping the (continuous) Bayes vector. The Bayes vector has been identified as the center of mass of the version space and can be constructed by applying a gradient descent algorithm on the cost function $E(\mathbf{J})$ defined by the potential :

$$V_s^+(\lambda) = \begin{cases} +\infty & \lambda < 0 \\ -\frac{\lambda^s}{s} & \lambda > 0 \end{cases} \tag{4.27}$$

with $s = -1.35$ [40]. In Fig. (4.4) the generalisation error of the Ising student obtained by clipping the optimal continuous student is plotted together with the Ising Gibbs result.

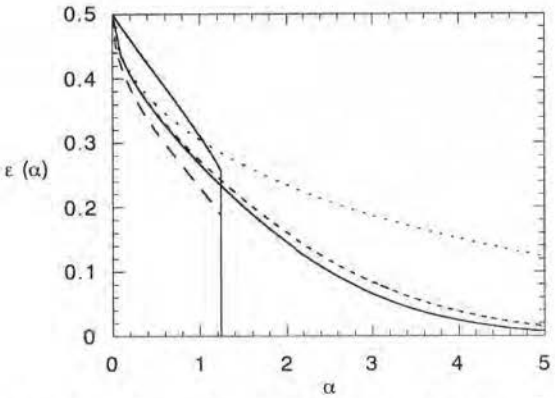

Fig.(4.4): Generalisation error for the Ising Gibbs and Ising Bayes perceptron together with the results for clipped Hebb (dotted line) and the error obtained by clipping the vector which minimizes the cost function defined through the potential (4.27) with $s = 0.25$ (dashed line) and $s = -1.35$ (full line).

One observes that for $\alpha < 1.24$ the generalisation error for the clipped optimal continuous student is smaller than the error for the the Ising Gibbs perceptron but at α_c no transition to perfect learning is achieved. This means that, since for $\alpha > 1.24$ there only remains one Ising vector in the version space (the teacher **T** itself) , clipping produces a vector outside the version space.

The theoretical results have been checked by carrying out some simulations using the inverse power potential ((4.27) with $s = -1.0$) to construct **J**. The results for a system with 100 input units are presented in Fig.(4.5) and show nice agreement with the theoretical results.

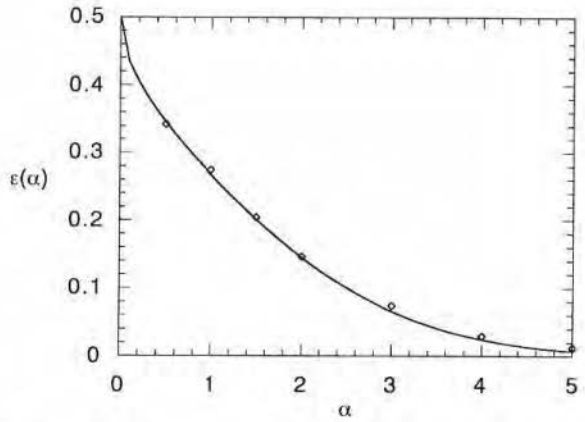

Fig.(4.5): Theoretical generalisation error of the student vector obtained by clipping the vector which minimizes the inverse power cost function (full line) together with simulation results.

From (4.11) one easily obtains the small α behaviour of $\tilde{R}$:

$$R \to 0 \quad : \quad \tilde{R} = \frac{2}{\sqrt{\pi}} \frac{R}{\sqrt{2(1-R^2)}}, \tag{4.28}$$

which leads to

$$\varepsilon = \frac{1}{2} - \frac{\sqrt{2}}{\pi^{3/2}} \frac{R}{1-R^2}. \tag{4.29}$$

For large values of α one gets :

$$R \to 1 \quad : \quad \tilde{R} = 1 - \frac{1}{\sqrt{\pi}} \frac{\exp\left(-\frac{1}{2}\frac{R^2}{1-R^2}\right)}{(R/\sqrt{2(1-R^2)})}, \tag{4.30}$$

which leads to a generalisation error

$$\varepsilon \sim \exp\left(-\frac{1}{2}\frac{R^2}{1-R^2}\right). \tag{4.31}$$

Since we know that for the class of potentials described in the previous chapter the generalisation error behaves as A/α, we conclude that the generalisation error of the clipped student $\tilde{\mathbf{J}}$ will behave as :

$$\varepsilon \sim \exp\left(-C\alpha^2\right), \tag{4.32}$$

with C a constant depending on the chosen potential.

4.3 Partial clipping.

Since plain clipping is a bad strategy for small R values one can consider partially clipped student vectors. The idea is to clip only the components J_i which are, in absolute value, larger than a certain treshold value κ. Following the suggestion of Bollé and Shim [39] one can choose the piece-wise linear function

$$f(t) = \begin{cases} \text{sign}(t) & |t| > \kappa \\ t/\kappa & |t| < \kappa \end{cases} \tag{4.33}$$

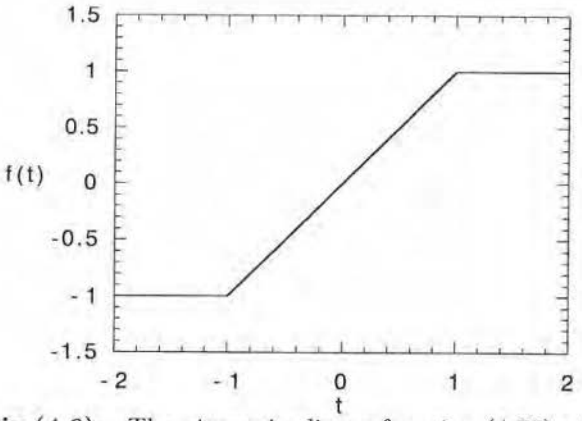

Fig.(4.6): The piece-wise linear function (4.33) with $\kappa = 1.0$.

Using the expressions (4.18) and (4.20) one finds :

$$\tilde{R}(\kappa) = \frac{H\left(\frac{\kappa - R}{\sqrt{1-R^2}}\right) - H\left(\frac{\kappa + R}{\sqrt{1-R^2}}\right) + \int_{-\kappa}^{+\kappa} dt\ P(t)\ (t/\kappa)}{\sqrt{H\left(\frac{\kappa - R}{\sqrt{1-R^2}}\right) + H\left(\frac{\kappa + R}{\sqrt{1-R^2}}\right) + \int_{-\kappa}^{+\kappa} dt\ P(t)\ (t/\kappa)^2}}. \tag{4.34}$$

For each value of R, one can determine the value of the treshold κ which maximizes the overlap $\tilde{R}$. It is not a surprise that the optimal value κ^* is small for large R and increases as R becomes small . The fraction of clipped components is given by :

$$\int_{\kappa^*}^{+\infty} dx\ \ (P(x) + P(-x)) \tag{4.35}$$

and is plotted in Fig.(4.7).

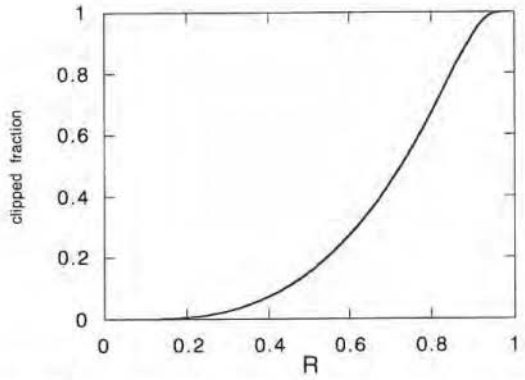

Fig.(4.7): The fraction of clipped synapses corresponding with the optimal value κ^*.

4.4 Optimal transformation

One can now try to determine the function f which maximizes $\tilde{R}$. This function f^* can be found as the solution of the variational problem :

$$\delta\, \tilde{R} = 0. \tag{4.36}$$

Varying (4.18) with respect to f leads to :

$$\int_{-\infty}^{+\infty} dt\, \delta f(t)\, P(t)\, \left[< f^2 > - < f >\, f(t)\right] = 0. \tag{4.37}$$

Assuming $\delta f(t)$ to be odd, one obtains :

$$f^*(t) = \frac{P(t) - P(-t)}{P(t) + P(-t)} \tag{4.38}$$

$$= \tanh\left(\frac{R\, t}{1 - R^2}\right). \tag{4.39}$$

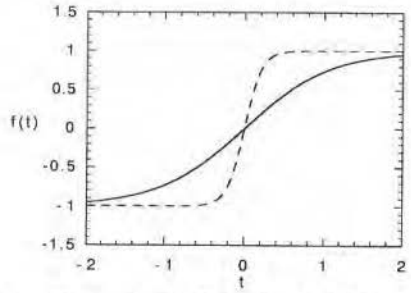

Fig.(4.8): The optimal function $f^*(t)$ for $R = 0.6$ (full curve) and $R = 0.9$ (dashed curve).

The function f^* is in principle only determined up to a multiplicative constant but the value of this constant is of no importance since it drops out of the final expressions. We see that $f^*(x)$ converges to sign(x) in the limit $R \to 1$.
The optimal transformation (4.39) has also been obtained by Bollé and Shim [39] for the modulated Hebb rule but our derivation shows that it remains valid for any learning rule.
The optimal value $\tilde{R}^*$ can be found by combining (4.39) and (4.18). This leads to :

$$< f^*(t) >= \int_{-\infty}^{+\infty} dt \frac{\exp\left(-\frac{1}{2}\frac{(t-R)^2}{1-R^2}\right)}{\sqrt{2\pi(1-R^2)}} \tanh\left(\frac{R\,t}{1-R^2}\right) \tag{4.40}$$

Using the transformation :

$$u = \frac{t-R}{\sqrt{1-R^2}}, \tag{4.41}$$

one can rewrite the equation (4.40) as :

$$< f^*(t) >= \int_{-\infty}^{+\infty} \mathcal{D}u \tanh\left(\frac{R}{1-R^2}\left(u + \frac{R}{1-R^2}\right)\right). \tag{4.42}$$

In an analogue way one obtains :

$$< (f^*(t))^2 >= \int_{-\infty}^{+\infty} \mathcal{D}u \tanh^2\left(\frac{R}{1-R^2}\left(u+\frac{R}{1-R^2}\right)\right). \qquad (4.43)$$

Now one can use the identity :

$$\int_{-\infty}^{+\infty} \mathcal{D}t \ \tanh[\kappa(t+\kappa)] = \int_{-\infty}^{+\infty} \mathcal{D}t \ \tanh^2[\kappa(t+\kappa)], \qquad (4.44)$$

which is valid for every value of κ. It can be proven in the following way : substraction of left and right part of (4.44) leads to:

$$\int_{-\infty}^{+\infty} \mathcal{D}t \ \tanh[\kappa(t+\kappa)] \left(\frac{\cosh[\kappa(t+\kappa)] - \sinh[\kappa(t+\kappa)]}{\cosh[\kappa(t+\kappa)]}\right) =$$
$$\int_{-\infty}^{+\infty} \mathcal{D}t \ \tanh[\kappa(t+\kappa)] \left(\frac{2\exp(-[\kappa(t+\kappa)])}{\cosh[\kappa(t+\kappa)]}\right). \qquad (4.45)$$

Changing the integration variable from t to $u = t + \kappa$ leads to :

$$\exp(-\kappa^2/2) \int_{-\infty}^{+\infty} \mathcal{D}u \ \frac{\sinh(\kappa u)}{\cosh^2(\kappa u)} \qquad (4.46)$$

which obviously is equal to zero. Using this result we get :

$$< f^* >=< f^{*2} > \qquad (4.47)$$

which leads to

$$\tilde{R}^* = \sqrt{< f^* >} \qquad (4.48)$$

$$= \sqrt{\int_{-\infty}^{+\infty} dt \ P(t) \ \tanh\left(\frac{R}{1-R^2}t\right)}. \qquad (4.49)$$

This result is presented in Fig. (4.9) together with the result for partially clipping with optimal threshold $\kappa = \kappa^*$. Note that the latter

result is nearly optimal as conjectured in [39].

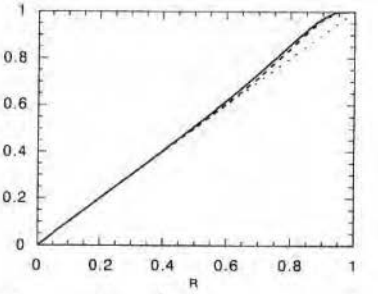

Fig.(4.9): The overlap $\tilde{R}^*$ of the optmally transformed vector (full line) together with the result for partially clipping with optimal threshold $\kappa = \kappa^*$ (dashed line). The dotted line is $\tilde{R} = R$.

To check the theoretical result for the optimal transformation simulations for a system of 100 input neurons have been carried out and they lead to excellent agreement with the theoretical results.

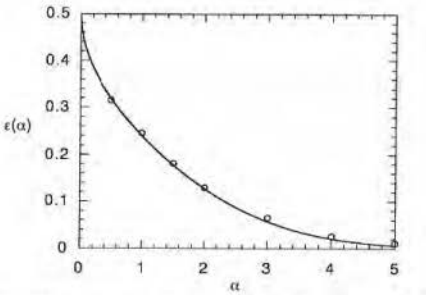

Fig.(4.10): Theoretical generalisation error of the student vector obtained by optimally transforming the vector which minimizes the inverse power cost function (full line) together with simulation results for $N = 100$.

Of particular interest is the behaviour of the overlap $\tilde{R}$ in the limit $R \to 1$. From Eq. (4.49) one obtains

$$R \to 1 \quad : \quad \tilde{R} = 1 - \frac{1}{\sqrt{2\pi}} \frac{\exp\left(-\frac{1}{2}\frac{R^2}{1-R^2}\right)}{(R/\sqrt{(1-R^2)}} \tag{4.50}$$

which is equal to the asymptotic behaviour of plain clipping.
So, also in this case we obtain that for large α the generalisation error goes as

$$\varepsilon \sim \exp\left(-C\alpha^2\right). \tag{4.51}$$

In Fig. (4.11), we present the generalisation error obtained by applying f^* to the Bayes perceptron. Although we are not able to reproduce the first order phase transition to perfect generalisation one observes that below the transition point $\alpha = 1.24$ the curve follows very closely that of the Ising Bayes perceptron which is the best possible result (but for which no practical algorithm exists).

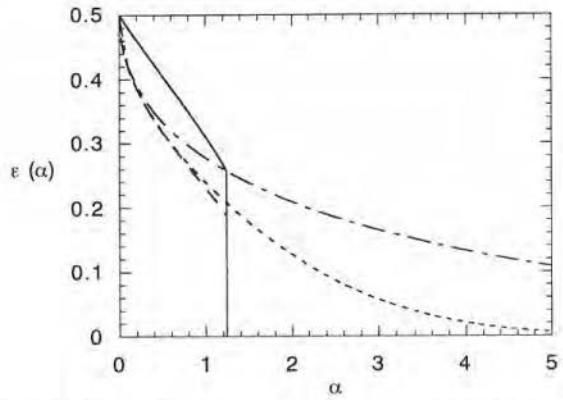

Fig.(4.11): The generalisation error for the optimally transformed continuous Bayes vector (dashed curve) and optimally transformed Hebb vector (long dashes).

Appendix A

Replica calculation for a general cost function.

Using the replica method, the free energy per neuron f is defined as :

$$-\beta f = \frac{1}{N}\langle \ln Z \rangle_{\boldsymbol{\xi}}, \tag{A.1}$$

$$= \lim_{n\to 0} \frac{1}{nN} \ln \langle Z^n \rangle_{\boldsymbol{\xi}}. \tag{A.2}$$

The partition function Z is given by :

$$Z = \int d\mu(\mathbf{J}) e^{-\beta E(\mathbf{J})}. \tag{A.3}$$

The notation $< . >_{\boldsymbol{\xi}}$ is used for the quenched average over the example patterns. The $\mathbf{J}$ -vectors all fulfil the normalisation constraint $\mathbf{J}^2 = N$. As a result one can write the integration in the partition function as :

$$\int d\mu(\mathbf{J}) \ldots \sim \int_{-\infty}^{+\infty} d\mathbf{J}\ \delta(\mathbf{J}^2 - N) \ldots \tag{A.4}$$

$$= \int_{-\infty}^{+\infty} \prod_{i=1}^{N} \mathcal{D}J_i \ldots \tag{A.5}$$

with

$$\mathcal{D}J_i = \frac{dJ_i}{\sqrt{2\pi}} \exp(-\frac{1}{2} J_i^2). \tag{A.6}$$

The cost functions $E(\mathbf{J})$ which we will consider are of the form :

$$E(\mathbf{J}) = \sum_{\mu=1}^{P} V(\lambda^{\mu}) \tag{A.7}$$

with λ^{μ} the "stability" of the μ-th pattern

$$\lambda^{\mu} = \frac{\mathbf{J}.\boldsymbol{\xi}^{\mu}}{\sqrt{N}} \ \text{sign}\left(\frac{\mathbf{T}.\boldsymbol{\xi}^{\mu}}{\sqrt{N}}\right) \tag{A.8}$$

The replicated partition function $< Z^n >$ can be written as :

$$\langle Z^n \rangle_{\boldsymbol{\xi}} = \int \prod_{\sigma=1}^{n} d\mu(\mathbf{J}_{\sigma}) \ \exp(-\alpha N Gr(\mathbf{J}_{\sigma})) \tag{A.9}$$

with the function $Gr(\mathbf{J}_{\sigma})$ defined as :

$$Gr(\mathbf{J}_{\sigma}) = -\ln \int d\mu(\boldsymbol{\xi}) \exp\left(-\beta \sum_{\sigma=1}^{n} V\left(\frac{\mathbf{J}_{\sigma}.\boldsymbol{\xi}}{\sqrt{\mathbf{N}}} \text{sign}\left(\frac{\mathbf{T}.\boldsymbol{\xi}}{\sqrt{N}}\right)\right)\right). \tag{A.10}$$

The measure $d\mu(\boldsymbol{\xi})$ for the patterns is equal to $d\mu(\boldsymbol{J})$ defined in (A.4). Introduction of the variables x_{σ} and y via δ-functions yields :

$$\begin{aligned} Gr = -\ln \int_{-\infty}^{+\infty} \prod_{\sigma=1}^{n} dx_{\sigma} \int_{-\infty}^{+\infty} dy \int_{-\infty}^{+\infty} d\mu(\boldsymbol{\xi}) \prod_{\sigma=1}^{n} \delta(x_{\sigma} - \frac{\mathbf{J}_{\sigma}.\boldsymbol{\xi}}{\sqrt{N}}) \\ .\delta(y - \frac{\mathbf{T}.\boldsymbol{\xi}}{\sqrt{N}}) \exp\left(-\beta \sum_{\sigma=1}^{n} V\left(x_{\sigma} \ \text{sign}(y)\right)\right). \end{aligned} \tag{A.11}$$

Using the exponential expression for the δ-functions and carrying out the integral with respect to $\boldsymbol{\xi}$ gives :

$$\begin{aligned} Gr = -\ln \int_{-\infty}^{+\infty} \prod_{\sigma=1}^{n} \frac{dx_{\sigma} d\hat{x}_{\sigma}}{2\pi} \int_{-\infty}^{+\infty} \frac{dy \ d\hat{y}}{2\pi} \\ .\exp\left(-\beta \sum_{\sigma=1}^{n} V\left(x_{\sigma} \ \text{sign}(y)\right) - \frac{1}{2}\hat{y}^2 - \frac{1}{2}\sum_{a=1}^{n} \hat{x}_a^2\right) \\ .\exp\left(i \sum_{\sigma=1}^{n} x_{\sigma}\hat{x}_{\sigma} + iy\hat{y} - \sum_{a<b=2}^{n} \hat{x}_a \hat{x}_b q_{ab} - \hat{y} \sum_{\sigma=1}^{n} \hat{x}_{\sigma} R_{\sigma}\right). \end{aligned} \tag{A.12}$$

In (A.12) we have introduced the usual overlap orderparameters :

$$q_{ab} = \frac{\mathbf{J}_a.\mathbf{J}_b}{N} \quad \forall a < b = 2, \ldots, n \tag{A.13}$$

$$R_a = \frac{\mathbf{J}_a.\mathbf{T}}{N} \quad \forall \alpha = 1, \ldots, n. \tag{A.14}$$

The replicated partition function $< Z^n >$ can be rewritten as :

$$\langle Z^n \rangle_{\boldsymbol{\xi}} = \int_{-\infty}^{+\infty} \prod_{a<b=2}^{n} \frac{dq_{ab} d\hat{q}_{ab}}{2\pi} \int_{-\infty}^{+\infty} \prod_{a=1}^{n} \frac{dR_a d\hat{R}_a}{2\pi} \int_{-\infty}^{+\infty} \prod_{a=1}^{n} dE_a$$
$$. \exp N \left(\sum_{a=1}^{n} E_a - \sum_{a=1}^{n} R_a \hat{R}_a - \sum_{a<b=2}^{n} q_{ab}\hat{q}_{ab} + G_0 - \alpha G_r \right), \tag{A.15}$$

with the function G_0 defined as :

$$G_0 = \ln \int_{-\infty}^{+\infty} \prod_{\sigma=1}^{n} dJ_\sigma \exp \left(-\sum_{a=1}^{n} E_a + \sum_{a=1}^{n} \hat{R}_a J_a T + \sum_{a<b=2}^{n} \hat{q}_{ab} J_a J_b \right). \tag{A.16}$$

The function G_r is defined in (A.12).

For $N \to +\infty$ (thermodynamic limit), the integral with respect to the orderparameters $q_{\alpha\beta}$ and R_α (and their conjugates) is dominated by the saddle point . The free energy is obtained by analytically continuing the saddle point to $n = 0$:

$$-\beta f = \lim_{n\to 0} \frac{1}{nN} \ln \langle Z^n \rangle_\xi \tag{A.17}$$

$$= \lim_{n\to 0} \frac{1}{n} \underset{(E_a, R_a, q_{ab}, \hat{R}_a, \hat{q}_{ab})}{\text{extr}} \Big(-\sum_{a=1}^{n} E_a - \sum_{a=1}^{n} R_a \hat{R}_a$$
$$- \sum_{a<b=2}^{n} q_{ab}\hat{q}_{ab} + G_0 - \alpha G_r \Big). \tag{A.18}$$

At this point of the calculation one has to introduce an ansatz for the order parameters. According to the Replica Symmetry (RS) ansatz one treats all replicas equivalently and the saddle point takes the form

$$q_{ab} = q, \tag{A.19}$$

$$R_a = R, \tag{A.20}$$

$$\hat{q}_{ab} = \hat{q}, \tag{A.21}$$

$$\hat{R}_a = \hat{R}. \tag{A.22}$$

Introducing this ansatz into (A.18) and letting $n \to 0$ leads to :

$$\begin{aligned} -\beta f = \underset{(E,q,\hat{q},R,\hat{R})}{\text{extr}} \Big(& E + \frac{1}{2} q\hat{q} - R\hat{R} + \frac{1}{2}\frac{\hat{R}^2 + \hat{q}}{2E + \hat{q}} - \frac{1}{2}\ln(2E + \hat{q}) \\ & + 2\alpha \int_0^{+\infty} \mathcal{D}t_2 \int_{-\infty}^{+\infty} \mathcal{D}t_1 \ln \int_{-\infty}^{+\infty} \frac{d\lambda}{\sqrt{2\pi(1-q)}} \\ & \cdot \exp\left(-\beta V(\lambda) + \frac{1}{2}\frac{(\lambda - t_2 R - \sqrt{q - R^2}\, t_1)^2}{1-q} \right) \Big). \end{aligned} \tag{A.23}$$

The adjoint variables E,$\hat{q}$ and $\hat{R}$ can be eliminated using their saddle point equations:

$$2E + \hat{q} = \frac{1}{1-q}, \tag{A.24}$$

$$\hat{R}^2 + \hat{q} = \frac{q}{(1-q)^2}, \tag{A.25}$$

$$\hat{R} = \frac{R}{1-q}. \tag{A.26}$$

Finally we obtain the free energy :

$$f = \underset{(q,R)}{\text{extr}} \Big(-\frac{1}{2}\frac{1 - R^2}{\beta(1-q)} - \frac{1}{2\beta}\ln(1-q)$$

$$-\frac{2\alpha}{\beta}\int_0^{+\infty}\mathcal{D}t_2\int_{-\infty}^{+\infty}\mathcal{D}t_1\ln\int_{-\infty}^{+\infty}\frac{d\lambda}{\sqrt{2\pi(1-q)}}$$
$$.\exp\left(-\beta V(\lambda)+\frac{1}{2}\frac{(\lambda-t_2R-\sqrt{q-R^2}t_1)^2}{1-q}\right)\Bigg). \quad \text{(A.27)}$$

In order to find the ground state energy, we must take the limit $\beta \to +\infty$. Since we concentrate on cost functions $E(\mathbf{J})$ with a non-degenerate minimum q should tend to 1. Introducing the variable $x = \beta(1-q)$ reduces the free energy to the following expression :

$$f^{T=0} = -\operatorname*{extr}_{(x,R)}\Bigg(\frac{1-R^2}{2x}$$
$$-2\alpha\int_0^{+\infty}\mathcal{D}t_2\int_{-\infty}^{+\infty}\mathcal{D}t_1\min_{\lambda}\left[V(\lambda)+\frac{(\lambda-t)^2}{2x}\right]\Bigg), \quad \text{(A.28)}$$

with

$$t = Rt_2+\sqrt{1-R^2}t_1. \quad \text{(A.29)}$$

The integrand of the last integral can also be written as

$$V(\lambda_0(x,R,t_1,t_2))+\frac{(\lambda_0(x,R,t_1,t_2)-t)^2}{2x}, \quad \text{(A.30)}$$

with $\lambda_0(x,R,t_1,t_2)$ the function which satisfies :

$$\frac{d}{d\lambda}\left(V(\lambda)+\frac{(\lambda-Rt_2+\sqrt{1-R^2}t_1)^2}{2x}\right)=0. \quad \text{(A.31)}$$

At last we can extremize the free energy with respect to x and R. This leads to the saddle point equations :

$$R = 2\alpha\int_0^{+\infty}\mathcal{D}t_2\int_{-\infty}^{+\infty}\mathcal{D}t_1\lambda_0(x,R,t_1,t_2)\ (t_2-\frac{Rt_1}{\sqrt{1-R^2}}),$$

(A.32)

$$1 - R^2 = 2\alpha \int_0^{+\infty} \mathcal{D}t_2 \int_{-\infty}^{+\infty} \mathcal{D}t_1 (\lambda_0(x, R, t_1, t_2) - t)^2. \quad \text{(A.33)}$$

A more useful form of these equations is obtained by applying following orthogonal transformation on the integration variables t_1 and t_2:

$$t = Rt_2 + \sqrt{1 - R^2}t_1, \quad \text{(A.34)}$$

$$t' = Rt_1 - \sqrt{1 - R^2}t_2. \quad \text{(A.35)}$$

The saddle point equations finally become :

$$R = \sqrt{\frac{2}{\pi}}\alpha \int_{-\infty}^{+\infty} \mathcal{D}t \ \lambda_0(x, \sqrt{1 - R^2}t) \quad \text{(A.36)}$$

$$1 - R^2 = 2\alpha \int_{-\infty}^{+\infty} \mathcal{D}t \ H\left(-\frac{Rt}{\sqrt{1 - R^2}}\right)(\lambda_0(x, t) - t)^2 \quad \text{(A.37)}$$

The equations (A.36) and (A.37) have been obtained using the RS-ansatz. To check the local stability of the RS solution one should analyse the eigenvalues of the Hessian matrix of the free energy at the RS saddle point. Almeida and Thouless [16] calculated the eigenvectors representing orthogonal fluctuations to the RS-solution. A derivation very analog to the one carried out by Gardner and Derrida in [46] leads to the following condition:

$$(P' - 2Q' + S') \,.\, (P - 2Q + S) < 1. \quad \text{(A.38)}$$

where P',Q',S',P,Q and S are defined in [46]. The first factor is, as in the storage case [46], equal to:

$$(P' - 2Q' + S') = (1 - q)^2. \quad \text{(A.39)}$$

Calculation of the second factor leads to :

$$(P - 2Q + S) =$$

$$2\alpha \int_{-\infty}^{+\infty} \mathcal{D}t\, H\left(-\frac{Rt}{\sqrt{1-R^2}}\right)\left[<< x^2 >> - << x >>^2\right]^2 \quad \text{(A.40)}$$

The average $<< . >>$ is defined as :

$$<< f(x) >> = \frac{\int_{-\infty}^{+\infty} \frac{dxd\lambda}{2\pi} f(x) \exp\left(-\beta V(\lambda) + ix(\lambda - t) - \frac{1}{2}(1-q)x^2\right)}{\int_{-\infty}^{+\infty} \frac{dxd\lambda}{2\pi} \exp\left(-\beta V(\lambda) + ix(\lambda - t) - \frac{1}{2}(1-q)x^2\right)}. \quad \text{(A.41)}$$

We now calculate $<< x >>$ and $<< x^2 >>$. Performing the gaussian integral with respect to x leads to the following expression for $<< x >>$:

$$<< x >> = \frac{i \int_{-\infty}^{+\infty} \frac{d\lambda}{\sqrt{2\pi(1-q)}} e^{\left(-\beta V(\lambda) - \frac{1}{2}\frac{(\lambda-t)^2}{1-q}\right)} \frac{\lambda - t}{1-q}}{\int_{-\infty}^{+\infty} \frac{d\lambda}{\sqrt{2\pi(1-q)}} e^{\left(-\beta V(\lambda) - \frac{1}{2}\frac{(\lambda-t)^2}{1-q}\right)}}. \quad \text{(A.42)}$$

For $\beta \to +\infty$ the integral with respect to λ can be calculated using steepest descent. This yields :

$$<< x >> = i\, \frac{\lambda_0(t,x) - t}{1-q}, \quad \text{(A.43)}$$

with $\lambda_0(t,x)$ defined as above.
Following the suggestion made by Bouten in [20] one remarks that :

$$\left[<< x^2 >> - << x >>^2\right] = i\, \frac{d << x >>}{dt}. \quad \text{(A.44)}$$

Deriving (A.43) with respect to t very easily leads to:

$$\left[<< x^2 >> - << x >>^2\right] = \frac{1}{1-q}\left(1 - \lambda_0'(t,x)\right). \quad \text{(A.45)}$$

Finaly we obtain:

$$(P - 2Q + S) = \frac{2\alpha}{(1-q)^2} \int_{-\infty}^{+\infty} \mathcal{D}t\, H\left(-\frac{Rt}{\sqrt{1-R^2}}\right)\left[\lambda_0' - 1\right]^2 \quad \text{(A.46)}$$

Introducing (A.39) and (A.46) in (A.38) leads to the so called AT-condition for the RS-saddle point:

$$2\alpha \int_{-\infty}^{+\infty} \mathcal{D}t \; H\left(-\frac{Rt}{\sqrt{1-R^2}}\right) [\lambda_0' - 1]^2 < 1. \tag{A.47}$$

Appendix B

Convexity of the cost function associated with the potential $V_s^+(\lambda)$.

In this Appendix we will show that the cost functions associated with the repulsive potentials $V_s^+(\lambda)$ defined in Chapter 2 are convex on the sphere $\mathbf{J}^2 = N$ for $s \leq 1$, hence the minimum is unique and can be found by gradient descent. Indeed, For $s \leq 1$ and $\lambda \geq 0$, the potential $V_s^+(\lambda)$ obeys the following two inequalities:

$$V(a\lambda_1 + (1-a)\lambda_2) \leq aV(\lambda_1) + (1-a)V(\lambda_2), \tag{B.1}$$

which is valid $\forall a$, $0 \leq a \leq 1$ and

$$V(\rho\lambda) \leq V(\lambda) \qquad ,\forall \rho \geq 1. \tag{B.2}$$

Consider now any two vectors $\mathbf{J}_1$ and $\mathbf{J}_2$ with λ_1 and $\lambda_2 \geq 0$ and $\mathbf{J}_1^2 = \mathbf{J}_2^2 = N$ and a vector $\mathbf{J}' = a\mathbf{J}_1 + (1-a)\mathbf{J}_2$, $0 \leq a \leq 1$, lying on the line that connects them.

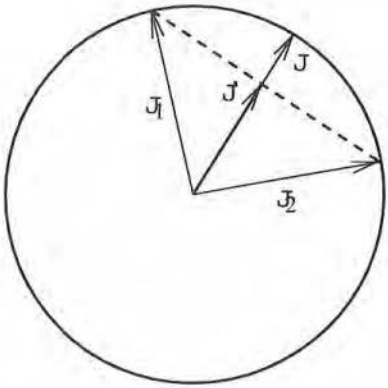

Fig.(2.1): The vectors $\mathbf{J}_1$,$\mathbf{J}_2$,$\mathbf{J}'$ and $\mathbf{J}$ as discribed in text.

By applying inequality (B.1) to every term of the sum defining $E(\mathbf{J}')$, one finds that

$$E(\mathbf{J}') \leq aE(\mathbf{J}_1) + (1-a)E(\mathbf{J}_2) \tag{B.3}$$

This proves the convexity of $E(\mathbf{J})$ within the sphere $\mathbf{J}^2 = N$.
We now consider the vector $\mathbf{J}$ that is parallel and in the same direction of $\mathbf{J}'$, but with the "proper" normalization $\mathbf{J}^2 = N$. Clearly $\mathbf{J} = \rho\mathbf{J}'$ with $\rho \geq 1$, since $\mathbf{J}'$ lies inside the hypersphere ($\mathbf{J}'^2 \leq N$) . Equation (B.2) immediately yields :

$$E(\mathbf{J}) \leq E(\mathbf{J}') \tag{B.4}$$

Combining (B.3) and (B.4) leads to the result :

$$E(\mathbf{J}) \leq aE(\mathbf{J}_1) + (1-a)E(\mathbf{J}_2). \tag{B.5}$$

This proves the convexity of $E(\mathbf{J})$ on the surface of the sphere $J^2 = N$.
At $s = 1$ the curvature of the potential $V(\lambda)$ changes sign with the result that for $1 < s \leq 2$ the condition (B.1) is not met.

Appendix C

Large α behaviour for the generalisation error corresponding with the repulsive potential $V_S^+(\lambda)$.

We consider the class of potential functions :

$$V_s^+(\lambda) = \begin{cases} +\infty & \lambda < 0 \\ -\frac{\lambda^s}{s} & \lambda > 0 \end{cases} \tag{C.1}$$

with s real and $\neq 0$. For $s = 0$, we define $V_s^+(\lambda)$ as:

$$V_{s=0}^+(\lambda) = \begin{cases} +\infty & \lambda < 0 \\ -\ln \lambda & \lambda > 0 \end{cases} \tag{C.2}$$

The function $\lambda_0(t, x)$ should fulfil the condition :

$$\lambda - t - x\lambda^{s-1} = 0 \qquad (\lambda > 0) \tag{C.3}$$

The solution $\lambda_0(t, x)$ of Eq.(C.3) exhibits the following scaling behaviour:

$$\lambda_0(t, x) = x^{\frac{1}{2-s}} \lambda_0 \left(\frac{t}{x^{\frac{1}{2-s}}}, 1 \right). \tag{C.4}$$

This scaling relation can be found by rewriting (C.3) as

$$\lambda[1 - x\lambda^{s-2}] = t, \tag{C.5}$$

or

$$1 - \left(x^{\frac{1}{s-2}}\lambda\right)^{s-2} = \frac{x^{\frac{1}{2-s}}\, t}{x^{\frac{1}{2-s}}\, \lambda} \tag{C.6}$$

If we now define the following variables :

$$h = x^{\frac{1}{2-s}}\lambda, \tag{C.7}$$

$$z = x^{\frac{1}{2-s}}t, \tag{C.8}$$

we obtain the equation :

$$h - h^{s-1} = z. \tag{C.9}$$

Thus, we have :

$$\lambda_0(t,x) = x^{\frac{1}{2-s}}h(\frac{t}{x^{\frac{1}{2-s}}}). \tag{C.10}$$

Using the scaling relation (C.4) one can rewrite the first saddle point equation as :

$$\frac{1}{2\alpha\sqrt{1-R^2}} = \left(\frac{x^{\frac{1}{2-s}}}{\sqrt{1-R^2}}\right)^2 \int_{-\infty}^{+\infty} \frac{du}{\sqrt{2\pi}} e^{-1/2(1-R^2)u^2} H(-Ru)$$
$$\cdot \left[h\left(u\frac{\sqrt{1-R^2}}{x^{\frac{1}{2-s}}}, 1\right) - u\frac{\sqrt{1-R^2}}{x^{\frac{1}{2-s}}}\right]^2, \tag{C.11}$$

where we have introduced the new integration variable $u = t/\sqrt{1-R^2}$. The second saddle point equation becomes :

$$\sqrt{\frac{\pi}{2}}\frac{R}{\alpha\sqrt{1-R^2}} = \frac{x^{\frac{1}{2-s}}}{\sqrt{1-R^2}} \int_{-\infty}^{\infty} Dt \;\; h\left[t\frac{\sqrt{1-R^2}}{x^{\frac{1}{2-s}}}, 1\right]. \tag{C.12}$$

If $\alpha \to +\infty$, R tends to 1 and x to 0. We will make the ansatz that

$$A = \frac{\alpha\sqrt{1-R^2}}{\pi} \tag{C.13}$$

$$B = \frac{x^{1/(2-s)}}{\sqrt{1-R^2}} \tag{C.14}$$

stay finite. As a result one can for, $\alpha \to +\infty$, rewrite the equation (C.11) as :

$$\frac{1}{2\pi A} = B^3 \int_{-\infty}^{+\infty} \frac{dz}{\sqrt{2\pi}} H(-Bz)[h(z)-z]^2 \tag{C.15}$$

where we have introduced the integration variable $z = u/B$.
The second saddle point equation yields :

$$\frac{1}{\sqrt{2\pi}A} = B \int_{-\infty}^{+\infty} \mathcal{D}t \; h(\frac{t}{B}) \tag{C.16}$$

Using (C.9) we can rewrite the integral in (C.15) as :

$$\int_{-\infty}^{+\infty} \frac{dz}{\sqrt{2\pi}} H(-Bz)[h(z)]^{2(s-1)} \tag{C.17}$$

This integral diverges for $2 > s \geq 1/2$ and is convergent for $s < 1/2$. This can be proven in following way.
The divergence of the integrand occurs at $z \to +\infty$. Indeed, if $z \to +\infty$, than one observes by rewriting equation (C.9) as:

$$1 - \frac{z}{h} = h^{s-2}, \tag{C.18}$$

that for $s < 2$, h should tend to $+\infty$ in order to satisfy the above equation. In the limit $z \to +\infty$ equation (C.18) leads to :

$$\lim_{z \to +\infty} \frac{z}{h} = 1. \tag{C.19}$$

This implies that:

$$\lim_{z \to +\infty} \left(\frac{z}{h}\right)^{2(s-1)} = 1. \tag{C.20}$$

As a result the integral (C.17) will converge if the integral

$$\int_{-\infty}^{+\infty} \frac{dz}{\sqrt{2\pi}} H(-Bz) z^{2(s-1)} \tag{C.21}$$

converges.This is the case for $2(s-1) < -1$.

(1) $s \geq 1/2$

Since the integral in (C.15) diverges, B should be equal to zero if we want to obtain a finite value for A. From the scaling relation we obtain :

$$B\, h(\frac{t}{B}) = \lambda(t, B^{2-s}). \tag{C.22}$$

The second saddle point equation yields:

$$\frac{1}{\sqrt{2\pi}A} = \int_{-\infty}^{+\infty} \mathcal{D}t \;\; \lambda_0(t, B^{2-s}) \tag{C.23}$$

Putting $B = 0$ one obtains

$$\frac{1}{\sqrt{2\pi}A} = \int_{-\infty}^{+\infty} \mathcal{D}t \;\; \lambda_0(t, 0). \tag{C.24}$$

But, $\lambda_0(t,0) = t$ for $t > 0$ and $\lambda_0(t,0) = 0$ for $t < 0$. As a result one obtains that :

$$\frac{1}{\sqrt{2\pi}A} = \int_{0}^{+\infty} \mathcal{D}t\, t = \frac{1}{\sqrt{2\pi}} \tag{C.25}$$

Finally one gets the following result :

$$A = \frac{\alpha\sqrt{1-R^2}}{\pi} = 1 \tag{C.26}$$

(ii) $s < \frac{1}{2}$

Numerical evaluation of (C.15) and (C.16) leads to the asymptotic behaviour of the generalisation error .

Appendix D

Replica calculation of the overlap S.

The center of mass of the version space $\mathbf{J}_{CM}$ is defined as :

$$\mathbf{J}_{CM} = \frac{\sqrt{N}\int_V dm(\mathbf{J})\,\mathbf{J}}{\sqrt{\int_V dm(\mathbf{J})\int_V dm(\mathbf{J}')\,\mathbf{J}.\mathbf{J}'}} \tag{D.1}$$

with

$$\int_V dm(\mathbf{J}) = \int_{-\infty}^{+\infty} d\mathbf{J}\;\delta(\mathbf{J}^2 - N)\prod_{\mu=1}^{P}\theta\left(\frac{\mathbf{J}.\boldsymbol{\xi}^\mu}{\sqrt{N}}\,\mathrm{sign}\left(\frac{\mathbf{T}.\boldsymbol{\xi}^\mu}{\sqrt{N}}\right)\right) \tag{D.2}$$

The overlap S can be written as :

$$S = \left\langle \frac{\mathbf{J}_{CM}.\mathbf{J}_E}{N} \right\rangle_{\boldsymbol{\xi}} \tag{D.3}$$

$$= \frac{1}{\sqrt{q}}\left\langle \frac{\mathbf{J}_{av}.\mathbf{J}_E}{N} \right\rangle_{\boldsymbol{\xi}}, \tag{D.4}$$

with

$$\mathbf{J}_{av} = \int_V dm(\mathbf{J})\;\mathbf{J}. \tag{D.5}$$

q is the typical overlap of two members of the version space and $< . >_{\boldsymbol{\xi}}$ stands for the quenched average over the pattern set. The overlap S can be calculated by studying the following partition function using the replica method:

$$Z = \int_V dm(\mathbf{J}^*) \int d\mu(\mathbf{J}) \exp\left(-\beta E(\mathbf{J})\right). \tag{D.6}$$

where $\int d\mu(\mathbf{J})$ is defined in Appendix A.
The replicated partition function Z^n can be written as :

$$Z^n = \int_V \prod_{\sigma=1}^{n} dm(\mathbf{J}_\sigma^*) \int \prod_{\sigma'=1}^{n} d\mu(\mathbf{J}_{\sigma'}) \exp\left(-\beta \sum_{\sigma'=1}^{n} E(\mathbf{J}_{\sigma'})\right). \tag{D.7}$$

One observes that we have introduced two sets of replicas :

$$\mathbf{J}_1, \mathbf{J}_2, \mathbf{J}_3, \ldots, \mathbf{J}_n, \tag{D.8}$$

and

$$\mathbf{J}_1^*, \mathbf{J}_2^*, \mathbf{J}_3^*, \ldots, \mathbf{J}_n^*. \tag{D.9}$$

The patterns $\boldsymbol{\xi}^\mu$ are gaussian distributed :

$$\langle f(\boldsymbol{\xi}) \rangle_{\boldsymbol{\xi}} = \int_{-\infty}^{+\infty} \prod_{i=1}^{N} \frac{d\xi_i}{\sqrt{2\pi}} \exp(-\frac{1}{2}\xi_i^2)\ f(\boldsymbol{\xi}). \tag{D.10}$$

Using the same methods as in Appendix A leads to following expression for $< Z^n >_\xi$:

$$\int \prod_{\sigma=1}^{n} d\mathbf{J}_\sigma^* \delta(\mathbf{J}_\sigma^*.\mathbf{J}_\sigma^* - N) \int \prod_{\sigma'=1}^{n} d\mathbf{J}_{\sigma'} \delta(\mathbf{J}_{\sigma'}.\mathbf{J}_{\sigma'} - N) \exp\left(-\alpha N G r\right), \tag{D.11}$$

with

$$Gr = -ln \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \prod_{a=1}^{n} \frac{d\lambda_a\, dx_a}{2\pi} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \prod_{a=1}^{n} \frac{d\tilde{\lambda}_a\, d\tilde{x}_a}{2\pi}$$

$$
\begin{aligned}
&\cdot \int_{-\infty}^{+\infty}\int_{0}^{+\infty}\prod_{a=1}^{n}\frac{du_a\,dv_a}{2\pi}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\frac{dy\,d\hat{y}}{2\pi} \\
&\cdot \exp\left(i\sum_{a=1}^{n}x_a\lambda_a+\sum_{a=1}^{n}\tilde{x}_a\tilde{\lambda}_a+iy\hat{y}\right) \\
&\cdot \exp\left(i\sum_{a=1}^{n}v_a(u_a-sign(y))\tilde{\lambda}_a)-\beta\sum_{a=1}^{n}V(\lambda_a sign(y))\right) \\
&\cdot \int_{-\infty}^{+\infty}\prod_{i=1}^{N}\mathcal{D}\xi_i\ \exp\left(-i\frac{\xi_i}{\sqrt{N}}(\sum_{a=1}^{n}x_aJ_{ai}+\sum_{a=1}^{n}\tilde{x}_aJ^*_{ai}+\hat{y}T_i)\right) \qquad \text{(D.12)}
\end{aligned}
$$

Carrying out the average over the patterns results in the introduction of the following order parameters :

$$q_{ab} = \frac{\mathbf{J}_a.\mathbf{J}_b}{N} \qquad \forall a<b=2,\ldots,n, \qquad \text{(D.13)}$$

$$r_a = \frac{\mathbf{J}_a.\mathbf{T}}{N} \qquad \forall a=1,\ldots,n, \qquad \text{(D.14)}$$

$$\tilde{q}_{ab} = \frac{\mathbf{J}^*_a.\mathbf{J}^*_b}{N} \qquad \forall a<b=2,\ldots,n, \qquad \text{(D.15)}$$

$$\tilde{r}_a = \frac{\mathbf{J}^*_a.\mathbf{T}}{N} \qquad \forall a=1,\ldots,n. \qquad \text{(D.16)}$$

The order parameters which express the overlap between the two sets of replicas are :

$$\tilde{s}_{ab}=\frac{\mathbf{J}_a.\mathbf{J}^*_b}{N} \qquad \forall a,b=1,\ldots,n. \qquad \text{(D.17)}$$

The conjugate variables are called respectively F_{ab},$\hat{r}_a$, $\tilde{F}_{ab}$, $\hat{\tilde{r}}_a$ and $\hat{\tilde{s}}_{ab}$. After introduction of the Replica Symmetry ansatz one has to use following transformation in order to be able to use the Hubbard-Stratonovich formula to linearise the terms in x_a and $\tilde{x}_a$:

$$
\begin{aligned}
&(q-R^2)\,X^2+2(\tilde{S}-R\tilde{R})X\tilde{X}+(\tilde{q}-\tilde{R}^2)\tilde{X}^2)= \\
&\frac{1}{2}(1+a)\left(\sqrt{q-R^2}X+\sqrt{\tilde{q}-\tilde{R}^2}\tilde{X}\right)^2 \\
&+\frac{1}{2}(1-a)\left(\sqrt{q-R^2}X-\sqrt{\tilde{q}-\tilde{R}^2}\tilde{X}\right)^2, \qquad \text{(D.18)}
\end{aligned}
$$

with $X = \sum_{a=1}^{n} x_a$ and $\tilde{X} = \sum_{a=1}^{n} \tilde{x}_a$.
The variable a is defined as :

$$a = \frac{\tilde{S} - R\tilde{R}}{\sqrt{q - R^2}\sqrt{\tilde{q} - \tilde{R}^2}}. \tag{D.19}$$

Taking the limit $n \to 0$ one obtains for the free energy f per neuron :

$$\begin{aligned}
-\beta f = \mathrm{extr} \Big(& E + \tilde{E} + \frac{1}{2}(qF + \tilde{q}\tilde{R}) \\
& -(\hat{R}R + \hat{\tilde{R}}\tilde{R}) - \frac{1}{2}\ln(2E + F) - \frac{1}{2}\ln(2\tilde{E} + \tilde{F}) \\
& + \frac{1}{2}\frac{\hat{R}^2 + F}{2E + F} + \frac{1}{2}\frac{\hat{\tilde{R}}^2 + \tilde{F}}{2\tilde{E} + \tilde{F}} \\
& + 2\alpha \int_0^{+\infty} \mathcal{D}y \int_{-\infty}^{+\infty} \mathcal{D}t \ln \int_{-\infty}^{+\infty} \frac{d\lambda}{\sqrt{2\pi(1-q)}} \\
& \exp\left(-\beta V(\lambda) - \frac{1}{2}\frac{(\lambda - yR + \sqrt{q - R^2}t)^2}{1 - q} \right) \\
& + 2\alpha \int_0^{+\infty} \mathcal{D}y \int_{-\infty}^{+\infty} \mathcal{D}t \ln \int_0^{+\infty} \frac{d\tilde{\lambda}}{\sqrt{2\pi(1-\tilde{q})}} \\
& \exp\left(-\frac{1}{2}\frac{(\tilde{\lambda} - y\tilde{R} + \sqrt{\tilde{q} - \tilde{R}^2}t)^2}{1 - \tilde{q}} \right) \Big)
\end{aligned} \tag{D.20}$$

Comparing (D.20) with the free energy of Appendix A before taking the $\beta \to +\infty$ limit and with the free energy obtained in [3] by Seung et al. leads to:

$$-\beta f = -\beta f_1 - \beta f_2 \tag{D.21}$$

with f_1 the free energy of the Gibbs problem [3] and f_2 the free energy of the generalisation problem with an arbitrary cost function $E(\mathbf{J})$ (Appendix A).

The free energy (D.21) is independent of the overlap parameter S (and

its conjugate $\hat{S}$). This is not a surprise since the two sets of replicas have there own independend potential, thus their thermodynamic behaviour should not depend on the overlap between two vectors from different replica sets . To determine $\tilde{S}$ one should extrimize the free energy with respect to $\tilde{s}_{ab}$ and $\hat{\tilde{s}}_{ab}$ before introducing the Replica Symmetry ansatz. After taking the $n \to 0$ limit one obtains the following equations:

$$\tilde{S}(2E+F)(2\tilde{E}+\tilde{F}) = \hat{R}\hat{\tilde{R}} + \hat{\tilde{S}} \tag{D.22}$$

and

$$\hat{\tilde{S}}(1-q)(1-\hat{q}) = 2\alpha\sqrt{\frac{1-\tilde{q}}{2\pi}} \int_0^{+\infty} \mathcal{D}y \int_{-\infty}^{+\infty} \mathcal{D}t \int_{-\infty}^{+\infty} \mathcal{D}t' \left(\lambda_0(x,R,y,t) - yR + \sqrt{1-R^2}t\right) \frac{\exp\left(-\frac{1}{2}g^2(\tilde{q},\tilde{R},y,t,t')\right)}{H\left(g(\tilde{q},\tilde{R},y,t,t')\right)} \tag{D.23}$$

where we define the function g as :

$$g(\tilde{q},\tilde{R},y,t,t') = \frac{-y\tilde{R} + \sqrt{\tilde{q}-\tilde{R}^2}(at + \sqrt{1-a^2}t')}{\sqrt{1-\tilde{q}}}. \tag{D.24}$$

After eliminating $\hat{\tilde{S}}$ one gets :

$$\tilde{S} - R\tilde{R} = 2\alpha\sqrt{\frac{1-\tilde{q}}{2\pi}} \int_0^{+\infty} \mathcal{D}y \int_{-\infty}^{+\infty} \mathcal{D}t \int_{-\infty}^{+\infty} \mathcal{D}t' \cdot \left(\lambda_0(x,R,y,t) - yR + \sqrt{1-R^2}t\right) \frac{\exp\left(-\frac{1}{2}g^2(\tilde{q},\tilde{R},y,t,t')\right)}{H\left(g(\tilde{q},\tilde{R},y,t,t')\right)} \tag{D.25}$$

with

$$a = \frac{\tilde{S} - R\tilde{R}}{\sqrt{1-R^2}\sqrt{\tilde{q}-\tilde{R}^2}} \tag{D.26}$$

Note that we let $\beta \to +\infty$ and introduced $x = \beta(1-q)$. The function $\lambda_0(x, R, y, t)$ minimizes the expression :

$$V(\lambda) + \frac{(\lambda - yR + \sqrt{1-R^2}t)^2}{2x}. \tag{D.27}$$

Numerically solving equation (D.25) leads to $\tilde{S}$ and to S, the overlap of the center off mass of the version space with the vector $\mathbf{J}$ minimizing the cost function $E\mathbf{J}$).

Appendix E

Clipping : the replica calculation.

In this appendix the overlap of a clipped perceptron with an Ising teacher will be calculated using the replica method.

$$\tilde{R} = \left\langle \frac{1}{N} \sum_{i=1}^{N} \text{sign}(J_i)\, T_i \right\rangle_{\boldsymbol{\xi}} . \tag{E.1}$$

The teacher perceptron $\mathbf{T}$ has binary synapses and the student perceptron $\mathbf{J}$ has been constructed by minimizing the cost function $E(\mathbf{J})$. The cost function $E(\mathbf{J})$ is defined as :

$$E(\mathbf{J}) = \sum_{\mu=1}^{P} V(\lambda^{\mu}), \tag{E.2}$$

and the corresponding partition function is :

$$Z = \int d\mu(\mathbf{J})\, e^{-\beta E(\mathbf{J})}. \tag{E.3}$$

Using the replica approach one can write the free energy as :

$$-\beta f = \lim_{n \to 0} \frac{1}{nN} \ln \langle Z^n \rangle_{\boldsymbol{\xi}} . \tag{E.4}$$

with Z^n the replicated partition function. The free energy defined by (E.4) is equal to the one defined in Appendix A. In order to be able to

study the properties of the clipped perceptron one can introduce the following identities:

$$1 = \int_{-\infty}^{+\infty} du\ \delta\left(u - \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathrm{sign}(J_i)\ \xi_i\right) \tag{E.5}$$

$$= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\frac{du\ dv}{2\pi}\exp\left(iv\ \left(u - \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathrm{sign}(J_i)\ \xi_i\right)\right). \tag{E.6}$$

This calculation is very similar to the one presented in Appendix A. Therefore, we will only discus the points where it is significantly differs from the one presented in Appendix A.
The integral in which one averages over the gaussian distributed example patterns yields :

$$\prod_{j=1}^{N}\int_{-\infty}^{+\infty}\mathcal{D}\xi_j \exp\left(-\frac{i}{\sqrt{N}}\xi_j\left(\sum_{a=1}^{n}x_a\ J_i^a + \sum_{a=1}^{n}v_a\ \mathrm{sign}(J_i^a) + \hat{y}\ T_i\right)\right) \tag{E.7}$$

Calculation of this simple gaussian integral leads, besides the usual q_{ab} and r_a, to the introduction of the following order parameters :

$$Q_{ab} = \sum_{i=1}^{N}\frac{\mathrm{sign}(J_i^a).\,\mathrm{sign}(J_i^b)}{N}, \tag{E.8}$$

$$\tilde{r}_a = \sum_{i=1}^{N}\frac{\mathrm{sign}(J_i^a).\,\mathrm{sign}(T_i)}{N}. \tag{E.9}$$

One also has to introduce parameters which express the overlap between the replicated student vectors $\mathbf{J}^a$ and the clipped vectors $\mathrm{sign}(\mathbf{J}^a)$.

$$P_c = \frac{1}{N}\sum_{i=1}^{N}|J_i^a| \tag{E.10}$$

$$p_{ab} = \frac{1}{N}\sum_{i=1}^{N}J_i^a\ \ \mathrm{sign}(J_i^b) \tag{E.11}$$

Introduction of the Replica Symmetry ansatz and taking the limit $n \to 0$ leads to the saddle point equations :

$$\hat{\tilde{R}} = \hat{Q} = \hat{P} = \hat{p} = 0 \tag{E.12}$$

The saddle point equations determining $\tilde{R}$, Q , P and p are :

$$\tilde{R} = \int_{-\infty}^{+\infty} \mathcal{D}z \ \langle \mathrm{sign}(J) \rangle \,, \tag{E.13}$$

$$Q = \int_{-\infty}^{+\infty} \mathcal{D}z \ (\langle \mathrm{sign}(J) \rangle)^2 \,, \tag{E.14}$$

$$P = \int_{-\infty}^{+\infty} \mathcal{D}z \ \langle J \rangle \ \langle \mathrm{sign}(J) \rangle \,, \tag{E.15}$$

$$p = \int_{-\infty}^{+\infty} \mathcal{D}z \ \langle |J| \rangle \,, \tag{E.16}$$

where we have used $< f(J) >$ for :

$$\langle f(J) \rangle = \frac{\int_{-\infty}^{+\infty} dJ \ f(J) \ \exp\left(-\frac{1}{2}(2E+F)J^2 + (\hat{R} + \sqrt{F}z)J\right)}{\int_{-\infty}^{+\infty} dJ \ \exp\left(-\frac{1}{2}(2E+F)J^2 + (\hat{R} + \sqrt{F}z)J\right)}, \tag{E.17}$$

with E,F and $\hat{R}$ as defined in Appendix A. The equations (E.13) and (E.14) can be easily calculated :

$$\tilde{R} = \int_{-\infty}^{+\infty} \mathcal{D}z \left(H\left(-\frac{R + \sqrt{q - R^2}z}{\sqrt{1-q}} \right) - H\left(\frac{R + \sqrt{q - R^2}z}{\sqrt{1-q}} \right) \right) \tag{E.18}$$

$$Q = \int_{-\infty}^{+\infty} \mathcal{D}z \left(H\left(-\frac{R + \sqrt{q - R^2}z}{\sqrt{1-q}} \right) - H\left(\frac{R + \sqrt{q - R^2}z}{\sqrt{1-q}} \right) \right)^2 \tag{E.19}$$

As explained in Appendix A we take the $\beta \to +\infty$ limit in order to obtain the ground state energy. If the minimum of the cost function is

unique and non-degenerate q should tend to 1. To carry out this limit one should split the integration interval of different parts:

$$\begin{aligned}\tilde{R} = & \int_{-\infty}^{-\frac{R}{\sqrt{q-R^2}}} \mathcal{D}z\, H\left(-\frac{R+\sqrt{q-R^2}z}{\sqrt{1-q}}\right) \\ & + \int_{-\frac{R}{\sqrt{q-R^2}}}^{+\infty} \mathcal{D}z\, H\left(-\frac{R+\sqrt{q-R^2}z}{\sqrt{1-q}}\right) - \int_{-\infty}^{-\frac{R}{\sqrt{q-R^2}}} \mathcal{D}z\, H\left(\frac{R+\sqrt{q-R^2}z}{\sqrt{1-q}}\right) \\ & - \int_{-\frac{R}{\sqrt{q-R^2}}}^{+\infty} \mathcal{D}z\, H\left(\frac{R+\sqrt{q-R^2}z}{\sqrt{1-q}}\right) \end{aligned} \tag{E.20}$$

Taking the limit $q \to 1$ and using that $H(+\infty) = 0$ and $H(-\infty) = 1$ one gets:

$$\tilde{R} = \int_{-\frac{R}{\sqrt{q-R^2}}}^{+\infty} \mathcal{D}z - \int_{-\infty}^{-\frac{R}{\sqrt{q-R^2}}} \mathcal{D}z. \tag{E.21}$$

Finally, we obtain:

$$\tilde{R} = H\left(-\frac{R}{\sqrt{1-R^2}}\right) - H\left(\frac{R}{\sqrt{1-R^2}}\right) \tag{E.22}$$

$$= \operatorname{erf}\left(\frac{R}{\sqrt{2(1-R^2)}}\right) \tag{E.23}$$

In an analog way we obtain for $q \to 1$ that Q will tend to 1. For the orderparameters p and P we obtain :

$$P = p = R\, \operatorname{erf}\left(\frac{R}{\sqrt{2(1-R^2)}}\right) + \frac{\sqrt{2(1-R^2)}}{\sqrt{\pi}} \exp\left(-\frac{1}{2}\frac{R^2}{1-R^2}\right) \tag{E.24}$$

The result (E.23) has been obtained for a teacher **T** with Ising synapses. In [38] and [39] one also considers different types of teachers. It is easy to generalise the expression (E.23) for a general teacher. One obtains :

$$\tilde{R} = \int_{-\infty}^{+\infty} \mathcal{D}z \int_{-\infty}^{+\infty} d\mu(T)\ T\ \langle \mathrm{sign}(J) \rangle\,, \tag{E.25}$$

with

$$\langle f(J) \rangle = \frac{\int_{-\infty}^{+\infty} dJ\ f(J)\ \exp\left(-\frac{1}{2}(2E+F)J^2 + (\hat{R}T + \sqrt{F}z)J\right)}{\int_{-\infty}^{+\infty} dJ\ \exp\left(-\frac{1}{2}(2E+F)J^2 + (\hat{R}T + \sqrt{F}z)J\right)}. \tag{E.26}$$

This leads to :

$$\tilde{R} = \int_{-\infty}^{+\infty} \mathcal{D}z \int_{-\infty}^{+\infty} d\mu(T)\ T\ \left(H\left(-\frac{RT + \sqrt{q-R^2}z}{\sqrt{1-q}}\right) - H\left(\frac{RT + \sqrt{q-R^2}z}{\sqrt{1-q}}\right)\right). \tag{E.27}$$

Letting q go to 1 gives :

$$\tilde{R} = \int_{-\infty}^{+\infty} d\mu(T)\ T\ \mathrm{erf}\left(\frac{R\,T}{\sqrt{2(1-R^2)}}\right) \tag{E.28}$$

The components of a generic teacher are distributed as:

$$d\mu(T) = dT \frac{\exp\left(-\frac{1}{2}T^2\right)}{\sqrt{2\pi}}. \tag{E.29}$$

Plugging (E.29) in (E.28), leads to:

$$\tilde{R} = \sqrt{\frac{2}{\pi}} R \tag{E.30}$$

By using $R_{Hebb}(\alpha)$ one recovers the result from [3].
For an easy (or diluted) teacher one has :

$$d\mu(T) = dT \left[(1-p_0)\ \delta(T) + \frac{p_0}{2}\left(\delta(T + \frac{1}{\sqrt{p_0}}) + \delta(T - \frac{1}{\sqrt{p_0}})\right)\right] \tag{E.31}$$

The corresponding $\tilde{R}$ yields:

$$\tilde{R} = \sqrt{p_0}\ \mathrm{erf}\left(\frac{1}{\sqrt{p_0}}\frac{R}{\sqrt{2(1-R^2)}}\right). \tag{E.32}$$

Nederlandstalige samenvatting.

Artificiële neurale netwerken

Alhoewel de meeste van de taken die we elke dag opnieuw vervullen eenvoudig lijken, is het zeer moeilijk om robotten te ontwerpen die deze taken even vlug en precies kunnen uitvoeren. Daarom lijkt het interessant om de menselijke hersenen te beschrijven of zelfs na te bouwen. De zogenaamde artificiële neurale netwerken zijn een poging daartoe. Deze netwerken bestaan uit onderling verbonden eenvoudige elementen die we neuronen zullen noemen omdat ze gemodelleerd zijn naar de typische werking van de zenuwcellen in de hersenen.

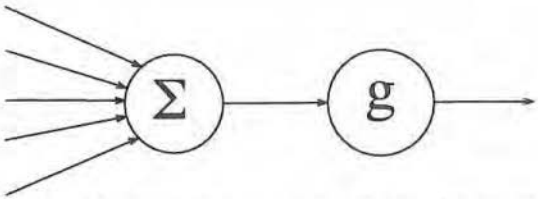

Fig.(N.1): De aktiviteit van een artificiëel neuron wordt berekend door al de inkomende aktiviteiten op te tellen en dan gebruik te maken van de transferfunktie g.

De neuronen, die gekenmerkt worden door een bepaalde aktiviteit, zijn onderling verbonden door middel van konnekties met een welbepaalde sterkte. De aktiviteit van een welbepaald neuron wordt berekend door de aktiviteiten van de verbonden neuronen met de korresponderende sterkte van de konnektie te vermenigvuldigen en dan op te tellen. Met het behulp van de tranferfunctie g wordt deze som dan omgezet in de aktiviteit van het neuron.
Het eenvoudigste neurale netwerk is het perceptron. Dit bestaat uit een laag van N neuronen die alle verbonden zijn met het uitgangsneuron. Deze architectuur is voorgesteld in de volgende tekening.

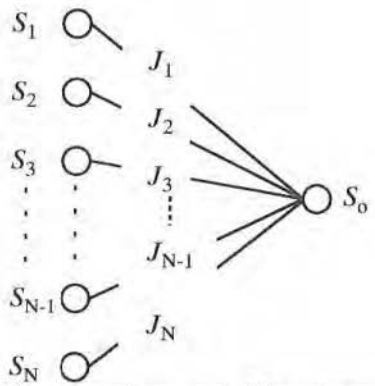

Fig.(N.2): Een perceptron met N input neuronen S_i verbonden met de output S_0 via de konnekties J_i.

Het verband tussen de aktivteit van het uitgangsneuron en de aktiviteiten van de N ingangsneuronen wordt gegeven door de volgende uitdrukking:

$$S_0 = g\left(\sum_{i=1}^{N} J_i\, S_i\right). \tag{N.1}$$

De aktiviteit van het i-de neuron is gekarakteriseerd door S_i en de sterkte van de korresponderende konnektie is gegeven door J_i. In deze

thesis wordt de drempeltransferfunctie $g(x) = \text{sign}(x)$ beschouwd.
Om meer inzicht te krijgen in het leermechanisme kunnen we een zeer eenvoudig scenario onderzoeken, namelijk dat van een leerling perceptron dat informatie krijgt van een leraar perceptron. Konkreet betekent dit dat we op basis van de klassificatie van P willekeurig geselecteerde voorbeelden $\boldsymbol{\xi}^\mu$ (gegenereerd door de leraar), een leerling willen konstrueren die niet alleen alle voorbeelden op de juiste manier klasseert maar ook op een willekeurige ander voorbeeld zo frequent mogelijk hetzelfde antwoord als de leraar geeft.

Formeler kan het leerling-leraar scenario als volgt worden gedefinieerd:

- De leraar is gekarakteriseerd door de op N genormeerde vector $\mathbf{T}$.
- De leraar $\mathbf{T}$ genereert de klassificatie van P willekeurig geselecteerde voorbeelden $\boldsymbol{\xi}^\mu$:

$$\xi_0^\mu = \text{sign}\left(\frac{\mathbf{T}.\boldsymbol{\xi}^\mu}{\sqrt{N}}\right) \quad \forall \mu = 1, \ldots, P. \tag{N.2}$$

- Op basis van de bovenstaande informatie willen we een leerlingsperceptron $\mathbf{J}$ (met $\mathbf{J}^2 = N$) konstrueren dat de klassifikatie van patronen $\boldsymbol{\xi}$ door de leraar zo goed mogelijk reproduceert.

De performantie van de leerling $\mathbf{J}$ wordt gekarakteriseerd door de zogenaamde generalisatiefout ε, gedefinieerd als de kans dat leerling en leraar het oneens zijn over de klassifikatie van een willekeurig gekozen vector $\mathbf{S}$. De generalisatiefout is gegeven door de volgende eenvoudige uitdrukking :

$$\varepsilon(\mathbf{J}) = \frac{1}{\pi} \arccos(R), \tag{N.3}$$

waarbij R de overlap is tussen de vectoren $\mathbf{J}$ en $\mathbf{T}$:

$$R = \frac{\mathbf{J}.\mathbf{T}}{N}. \tag{N.4}$$

Deze uitdrukking kan zeer eenvoudig worden afgeleid. De vectoren $\mathbf{J}$ en $\mathbf{T}$ liggen op het oppervlak van een N dimensionale sfeer. Als we nu enkel het $\mathbf{J} - \mathbf{T}$ vlak tekenen zien we onmiddellijk dat leerling en leraar het oneens zullen zijn over de klassifikatie van een nieuw patroon als de projectie van dat patroon in het gearceerde deel van de tekening ligt. Dit heeft tot gevolg dat:

$$\varepsilon(\mathbf{J}) = \frac{2\theta}{2\pi} \tag{N.5}$$

$$= \frac{1}{\pi} \arccos\left(\frac{\mathbf{T}.\mathbf{J}}{N}\right). \tag{N.6}$$

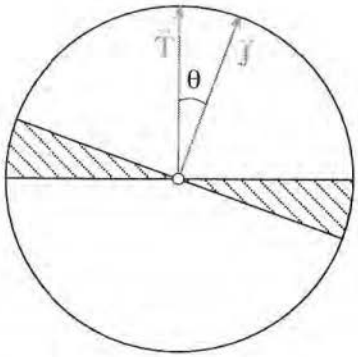

Fig.(N.3): Leraar en leerling zullen het niet eens zijn over de klassifikatie van een nieuw patroon $\mathbf{S}$, als de projektie in het $\mathbf{J} - \mathbf{T}$ vlak in het gearceerde gebied ligt.

Theoretische beschrijving

Een van de meest gebruikte methodes om een leerling te selecteren is het minimaliseren van een passende kost funktie $E(\mathbf{J})$. Als het minimum van de kost funktie uniek is kunnen we het construeren door gebruik te maken van de eenvoudige "gradient descent" techniek.
In deze thesis beperken we ons tot kost functies van de gedaante :

$$E(\mathbf{J}) = \sum_{\mu=1}^{P} V(\lambda^\mu), \tag{N.7}$$

met λ^μ gedefinieerd als :

$$\lambda^\mu = \frac{\mathbf{J}.\boldsymbol{\xi}^\mu}{\sqrt{N}}.\,\mathrm{sign}\left(\frac{\mathbf{T}.\boldsymbol{\xi}^\mu}{\sqrt{N}}\right). \tag{N.8}$$

Merk op dat de classificatie van het μ-de leervoorbeeld door de leerling $\mathbf{J}$ korrekt is als $\lambda^\mu > 0$.
Om nu de generalisatiefout ε te berekenen voor een willekeurige kostfunctie met een uniek minimum kunnen we gebruik maken van de technieken ontwikkelt in de statistische mechanika ter beschrijving van spinglazen. Met de kostfunctie $E(\mathbf{J})$ kunnen we volgende partitiefunctie definiëren:

$$Z = \int_{-\infty}^{+\infty} d\mathbf{J}\ \ \delta(\mathbf{J}^2 - N)\, e^{-\beta E(\mathbf{J})}. \tag{N.9}$$

In de thermodynamische limiet, waarbij $N \to +\infty$ en $P \to +\infty$ met $\alpha = P/N$ vast, kan de vrije energie per neuron geschreven worden als :

$$-\beta f = \frac{1}{N}\langle \ln Z \rangle_\xi \tag{N.10}$$

waarbij $\langle . \rangle_\xi$ staat voor het gemiddelde over de patronen. Gebruik makend van de replica methode vinden we voor de vrije energie:

$$-f^{T=0} = \underset{x,R}{\mathrm{extr}} \left(\frac{1-R^2}{2x} \right.$$
$$\left. -2\alpha \int_{-\infty}^{+\infty} \mathcal{D}t_1 \int_{0}^{+\infty} \mathcal{D}t_2 \ \min_{\lambda} \left[V(\lambda) + \frac{(\lambda - t)^2}{2x} \right] \right) \tag{N.11}$$

met

$$t = Rt_2 + \sqrt{1 - R^2}t_1. \tag{N.12}$$

$R(\alpha)$ kan bepaald worden door de vrije energie af te leiden naar R en x. Dit levert volgende vergelijkingen op:

$$R = \sqrt{\frac{2}{\pi}\alpha} \int_{-\infty}^{+\infty} \mathcal{D}t \; \lambda_0(\sqrt{1 - R^2}t, x) \tag{N.13}$$

$$1 - R^2 = 2\alpha \int_{-\infty}^{+\infty} \mathcal{D}t \; H\left(-\frac{Rt}{\sqrt{1 - R^2}}\right) (\lambda_0(t, x) - t)^2 \tag{N.14}$$

De functie $\lambda_0(t, x)$ is gedefiniëerd door de volgende uitdrukking:

$$\frac{d}{d\lambda}\left(V(\lambda) + \frac{(\lambda - t)^2}{2x}\right) = 0. \tag{N.15}$$

Uit de vergelijkingen (N.13) en (N.14) kan de generalisatiefout voor een willekeurige kostfunctie $E(\mathbf{J})$ berekend worden voor elke waarde van α door $R(\alpha)$ te substitueren in de uitdrukking (N.3).

Resultaten voor een klasse van repulsieve potentialen.

De "version space" is gedefinieerd als de verzameling op N genormeerde vektoren $\mathbf{J}$ die de leervoorbeelden $\boldsymbol{\xi}^\mu$ korrekt klassificeren. Volgens de Bayes leerregel moet de klassifikatie van een nieuw patroon bepaald worden door de meerderheid van de perceptronvektoren die deel uitmaken van de version space. Dus, als er 10 vektoren in de version space liggen en 6 daarvan klassificeren dat nieuwe patroon als $+1$ dan zegt de Bayes regel dat we dat patroon als $+1$ moeten klassificeren.
Watkin toonde aan dat de Bayes leerregel kan gerepresenteerd worden door een perceptron waarbij de konnekties gekarakteriseerd zijn door de massamiddelpuntsvector van de "version space". P.Reimann toonde

aan dat, als we enkel gebruik maken van de informatie vervat in de klassificatie van de P patronen door de leraar, deze vector de kleinst mogelijke hoek maakt met het leraarsperceptron $\mathbf{T}$ en bijgevolg gekenmerkt wordt door de laagst mogelijke generalisatiefout ε .
Geinspireerd door dit resultaat kan men nu op zoek gaan naar een kost functie $E(\mathbf{J})$ die het mogelijk maakt om een leerlingvector te konstrueren die zeer dicht bij de Bayes vektor ligt. We beschouwen daartoe een klasse van kost funkties gedefinieerd door de potentiaal:

$$V_s^+(\lambda) = \begin{cases} +\infty & \lambda < 0 \\ -\frac{\lambda^s}{s} & \lambda > 0 \end{cases} \tag{N.16}$$

met de parameter s reëel en kleiner dan -2. Voor $s = 0$ definiëren we $V_s^+(\lambda)$ als volgt:

$$V_{s=0}^+(\lambda) = \begin{cases} +\infty & \lambda < 0 \\ -\ln\lambda & \lambda > 0 \end{cases} \tag{N.17}$$

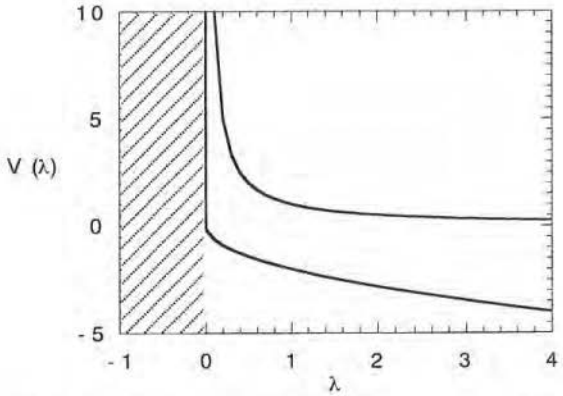

Fig.(N.4): De potentiaal (N.16) voor $s = -1.0$ and $s = 0.25$.

De potentiaal $V_s^+(\lambda)$ is gelijk aan $+\infty$ voor negatieve waarden van λ. Bijgevolg hebben enkel vectoren die gelegen zijn in de version space een eindige waarde van E. Voor $\lambda > 0$ nemen we een monotoon dalende funktie. Aangezien de randen van de version space bepaald worden door de vlakken $\lambda^\mu = 0$ betekent dit dat we leerlingperceptronen die te dicht bij de rand van de version space liggen gaan bestraffen. Men kan aantonen (Appendix B) dat de kostfuncties gedefinieerd door bovenstaande repulsieve potentialen een uniek minimum hebben voor $s \leq 1$.

Als we nu gebruik maken van het formalisme uit de vorige sectie kunnen we de generalisatie fout $\varepsilon(\alpha)$ gaan berekenen voor elke waarde van s. Analyse van de numerieke resultaten leidt tot het besluit dat er een optimale waarde voor s bestaat. Inderdaad, de generalisatie fout ε die we bekomen met de parameter $s_{opt} = -1.35$ is voor alle waarde van α, kleiner dan die voor de andere s waarden.

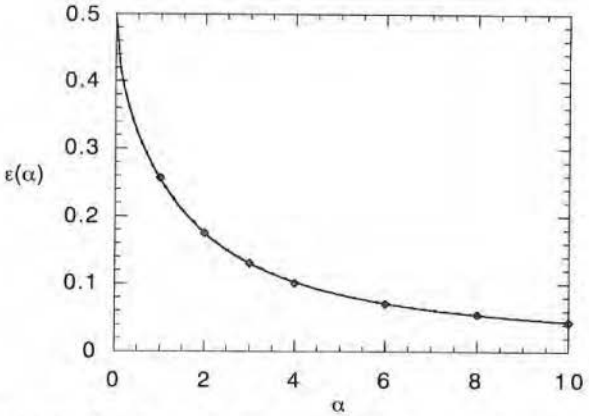

Fig.(N.5): Analytische resultaten voor de Bayes leerregel (volle lijn) en voor de potentiaal $V_s^+(\lambda)$ met $s = -1.35$ (streeplijn) samen met simulatieresultaten voor een systeem van 50 neuronen.

Als we nu de generalisatie fout bekomen door het gebruik van deze optimale repulsieve potentiaal vergelijken met de generalisatiefout van de Bayes leerregel zien we dat op de schaal van de tekening de resultaten bijna niet te onderscheiden zijn (ze verschillen voor elke waarde van α minder dan 1%). Op de tekening zijn ook de resultaten van een computersimulatie te zien voor een systeem van 50 neuronen. We zien dat de simulatieresultaten zeer goed overeenkomen met de numeriek berekende generalisatiefout.

Het belangrijke aan dit resultaat is dat we nu een praktisch algoritme hebben om een leerling perceptron te konstrueren die bijna even goed generaliseert als de Bayes leerregel.

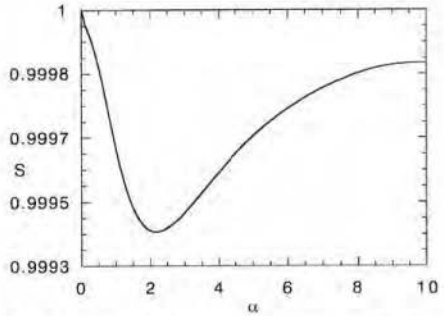

Fig.(N.6): Overlap S tussen de vektor $\mathbf{J}$ die de optimale repulsieve potentiaal minimaliseert en het massamiddelpunt van de version space .

We kunnen de overlap S tussen een vektor $\mathbf{J}_E$ die een zekere kostfunctie E minimaliseert en het massamiddelpunt van de version space gaan berekenen. Het massamiddelpunt van de version space kan op volgende manier worden gedefinieerd:

$$\mathbf{J}_{CM} = \frac{\sqrt{N}\int_V dm(\mathbf{J})\mathbf{J}}{\sqrt{\int_V dm(\mathbf{J})\int_V dm(\mathbf{J}')\mathbf{J}.\mathbf{J}'}}, \qquad \text{(N.18)}$$

met

$$\int_V dm(\mathbf{J}) = \int_{-\infty}^{+\infty} d\mathbf{J}\ \delta(\mathbf{J}^2 - N) \prod_{\mu=1}^{P} \theta\left(\frac{\mathbf{J}.\xi^\mu}{\sqrt{N}} sign\left(\frac{\mathbf{T}.\xi^\mu}{\sqrt{N}}\right)\right). \quad \text{(N.19)}$$

De overlap S wordt dan:

$$S = \left\langle \frac{\mathbf{J}_{CM}.\mathbf{J}_E}{N} \right\rangle \quad \text{(N.20)}$$

Voor de repulsieve potentiaal met $s_{opt} = -1.35$ is het resultaat voorgesteld in de voorgaande figuur. We zien dat voor elke waarde van α deze vektor extreem dicht bij het massamiddelpunt $\mathbf{J}_{CM}$ ligt.

Leraars met binaire koppelingen

Tot nu toe hebben we enkel leraars beschouwd met kontinue koppelingen T_i. We willen nu echter het leerling-leraar scenario bestuderen voor het geval van een Ising leraar. Dit wil zeggen dat de komponenten van de leraar binair zijn, i.e.

$$T_i = \pm 1 \quad \forall i = 1, \ldots, N. \quad \text{(N.21)}$$

Om een student te konstrueren kan natuurlijk terug de kostfunctie gebruikt worden die voor een leraar met kontinue koppelingen bijna optimale resultaten geeft. Rekening houdend met de extra informatie dat de koppelingen van de leraar binair zijn kan men zich echter de vraag stellen of er niet een nog lagere generalisatie fout kan bereikt worden. Een van de strategieën die men kan volgen is een arbitraire leerregel gebruiken om een leerling $\mathbf{J}$ met kontinue koppelingen te konstrueren en de komponenten van die vector te vervangen door ± 1 naargelang hun teken. In het Engels wordt deze strategie "clippen" genoemd. Hier zullen we deze techniek "knippen" noemen.

De komponenten van een algemeen getransformeerde vektor $\tilde{\mathbf{J}}$ zijn gedefinieerd als volgt:

$$\tilde{J}_i = \frac{\sqrt{N} f(J_i)}{\sqrt{\sum_{j=1}^{N} f(J_i)^2}}. \quad \text{(N.22)}$$

We veronderstellen dat de funktie f oneven is. Het knippen van de vektor $\mathbf{J}$ is een speciaal geval van bovenstaande transformatie waarbij $f(x) = \text{sign}(x)$.
We zullen nu aantonen dat de overlap $\tilde{R}$ tussen de getransformeerde leerling $\tilde{\mathbf{J}}$ en de leraar $\mathbf{T}$ een eenvoudige funktie is van de overlap R tussen de vector $\mathbf{J}$ en $\mathbf{T}$. De overlap $\tilde{R}$ wordt gegeven door :

$$\tilde{R} = \left\langle \frac{1}{N} \sum_{i=1}^{N} \tilde{J}_i \, T_i \right\rangle_{\xi} \tag{N.23}$$

$$= \left\langle \frac{\sum_{i=1}^{N} f(J_i T_i)}{\sqrt{N \sum_{j=1}^{N} f^2(J_i T_i)}} \right\rangle_{\xi} \tag{N.24}$$

$$= \left\langle \frac{1/N \sum_{i=1}^{N} f(J_i T_i)}{\sqrt{1/N \sum_{j=1}^{N} f^2(J_i T_i)}} \right\rangle_{\xi} \tag{N.25}$$

Gebruik makend van de wet van de grote getallen kunnen de sommen geschreven worden als integralen:

$$\frac{1}{N} \sum_{i=1}^{N} f(t_i) = \langle f(t) \rangle \tag{N.26}$$

$$= \int_{-\infty}^{+\infty} dt \, P(t) \, f(t) \tag{N.27}$$

met $P(t)$ de kansverdeling van de variabele $t_i = J_i \, T_i$.
$P(t)$ is gegeven door:

$$P(t) = \frac{\exp\left(-\frac{1}{2} \frac{(t-R)^2}{1-R^2}\right)}{\sqrt{2\pi(1-R^2)}} \tag{N.28}$$

Dit reduceert de uitdrukking voor $\tilde{R}$ tot:

$$\tilde{R} = \frac{\langle f(t) \rangle}{\sqrt{\langle f^2(t) \rangle}}. \tag{N.29}$$

Als we nu de functie :

$$f(t) = \text{sign}(t) \tag{N.30}$$

substitueren in de uitdrukking voor $\tilde{R}$ geeft dit als resultaat:

$$\tilde{R} = \text{erf}\left(\frac{R}{\sqrt{2(1-R^2)}}\right), \tag{N.31}$$

met $\text{erf}(t)$ de error funktie.

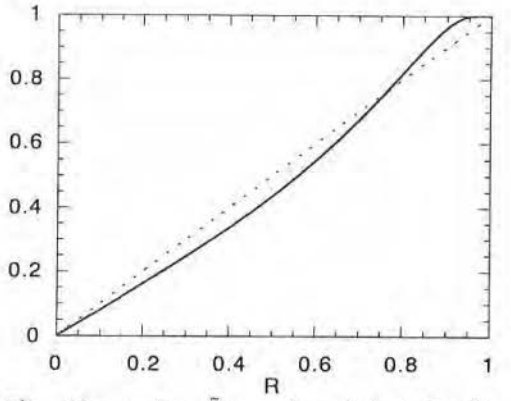

Fig.(N.7): De overlap $\tilde{R}$ van de geknipte leerling $\tilde{\mathbf{J}}$ met de Ising leraar (volle lijn) samen met de originele overlap R (streeplijn).

Uit de figuur is duidelijk dat knippen de overlap enkel vergroot als de oorspronkelijke overlap R groter is dan 0.78. We kunnen ons dan ook afvragen of we geen betere keuze kunnen maken voor de funktie f.
We kunnen bijvoorbeeld enkel de grootste komponenten van de vektor $\mathbf{J}$ gaan knippen.
Dit wordt bereikt door volgende functie f te gebruiken:

$$f(x) = \begin{cases} \text{sign}(x) & |x| > \kappa \\ x/\kappa & |x| < \kappa \end{cases} \tag{N.32}$$

met κ een positieve parameter.
Dit leidt tot volgende uitdrukking voor $\tilde{R}$:

$$\tilde{R} = \frac{H\left(\frac{\kappa-R}{\sqrt{1-R^2}}\right) - H\left(\frac{\kappa+R}{\sqrt{1-R^2}}\right) + \int_{-\kappa}^{+\kappa} dt\, P(t)\,(t/\kappa)}{\sqrt{H\left(\frac{\kappa-R}{\sqrt{1-R^2}}\right) + H\left(\frac{\kappa+R}{\sqrt{1-R^2}}\right) + \int_{-\kappa}^{+\kappa} dt\, P(t)\,(t/\kappa)^2}} \qquad \text{(N.33)}$$

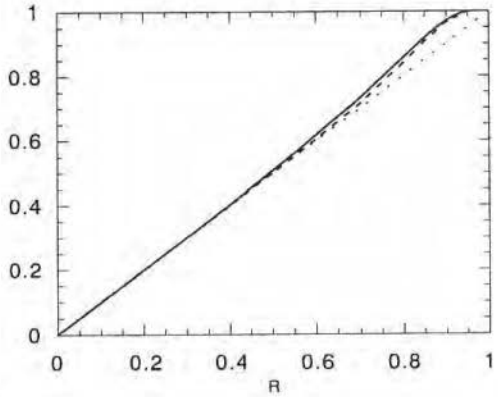

Fig.(N.8): De overlap $\tilde{R}^*$ van de optimaal getransformeerde vektor samen met het resultaat voor het gedeeltelijk geknipte vektor met $\kappa = \kappa^*$ (streeplijn) samen met de $\tilde{R} = R$ rechte.

De optimale transformatie f^* kan bekomen worden door $\tilde{R}$ te varieren naar f:

$$\delta\, \tilde{R} = 0. \qquad \text{(N.34)}$$

Dit levert volgend resultaat op voor de optimale functie $f^*(t)$:

$$f^*(t) = \tanh\left(\frac{R}{1-R^2}t\right) \qquad \text{(N.35)}$$

De bijbehorende waarde voor $\tilde{R}$ is :

$$\tilde{R}^* = \sqrt{\int_{-\infty}^{+\infty} dt\ P(t)\ \tanh\left(\frac{R}{1-R^2}t\right)} \tag{N.36}$$

Dit resultaat is voorgesteld in de vorige figuur samen met het resultaat voor gedeeltelijk knippen (met optimale waarde voor κ).

Bibliography

[1] G.Gyorgyi and N.Tishby, Neural Networks and Spin Glasses, edited by W.K.Theumann and R.Koberle World Scientific, Singapore 1990, p3-36

[2] J.Hertz, A.Krogh and R.G.Palmer, Introduction to the Theory of Neural Computing, Addison-Wesley, Reading (Massachusetts) 1991

[3] H.S.Seung, H.Sompolinsky and N.Tishby , Phys. Rev. A 45, 6056 (1992)

[4] T.L.H.Watkin, A.Rau and M.Biehl, Rev. Mod. Phys. 65, 499 (1993)

[5] M.Opper and W.Kinzel, Statistical Mechanics of Generalisation, to appear in Physics of Neural Networks III, eds. E.Domany, J.L.Van Hemmen and K.Schulten, Springer (Berlin) 1994

[6] C.Van den Broeck, Act. Phys. Pol. B 25, 903 (1994)

[7] A. Engel, Uniform Convergence Bounds for Learning from Examples, Int. J. Mod. Phys., to appear (1994).

[8] T. Grossman, R. Meir and E. Domany, in : "Neural Information Processing Systems 1", p73, D.S. Touretzky Ed., Morgan Kaufmann, San Mateo, CA (1989).

[9] P.Rujan, J. Phys. I France 3, 277 (1993)

[10] M.Griniasty and H.Gutfreund, J. Phys. A: Math. Gen. 24, 715 (1991)

[11] K.Y.M.Wong and D.Sherrington, J. Phys. A: Math. Gen. 23, 4659 (1990)

[12] J.Anlauf and M.Biehl,Europhys. Lett. 10, 587 (1989)

[13] M.Opper and D.Haussler , Phys. Rev. Lett. 66 , 2677 (1991)

[14] A. Blumer, A. Ehrenfeucht, D. Haussler and M. K. Warmuth, J. A.C.M. 36, 929 (1989)

[15] M. Anthony and N. Biggs, Computational Learning Theory, Cambridge University Press, Cambridge (1992)

[16] J.R.L.de Almeida and D.J.Thouless, J. Phys. A: Math. Gen. 11, 271 (1978)

[17] M.Griniasty, Phys. Rev. E 47 , 4496 (1993)

[18] P.Majer, A.Engel and A.Zippelius, J. Phys. A: Math. Gen. 26, 7405 (1993)

[19] K.Y.M.Wong and D.Sherrington, Phys. Rev. E 47, 4465 (1993)

[20] M.Bouten, J. Phys. A: Math. Gen. 27, 6021 (1994)

[21] R.Meir and J.F.Fontanari, Phys. Rev. A 45, 8874 (1992)

[22] F.Vallet and J.C.Cailton, Phys. Rev. A 41, 3059 (1990)

[23] M.Opper, W.Kinzel, J.Kleinz and R.Nehl, J. Phys. A: Math. Gen. 23, L581 (1990)

[24] T.L.H.Watkin, Europhys. Lett. 21, 871 (1993)

[25] A.Engel and C.Van den Broeck, Phys. Rev. Lett. 71, 1772 (1993); Physica A 200, 636 (1993)

[26] T.L.H.Watkin and J.P.Nadal,J. Phys. A: Math. Gen. 27, 1899 (1994)

[27] P.Ruján, Playing Billiard in Version Space, preprint (1995)

[28] M.Opper and W.Kinzel, Statistical Mechanics of Generalisation, Models of Neural Networks, Eds. E.Domany, J.L.van Hemmen, K.Schulten, Springer Verlag (1995)

[29] K.Y.M.Wong, A.Rau and D.Sherrington, Europhys. Lett. 19, p.559 (1992)

[30] C.Marangi, M. Biehl and S.Solla, Supervised learning from clustered input examples, preprint

[31] T.L.H.Watkin, Europhys.Lett. 21 (8),p.871

[32] P.Reimann and C.Van den Broeck, Learning from examples from a non-uniform distribution, to be published Phys.Rev. E

[33] P.Reimann and C.Van den Broeck, A gaussian model for unsupervised learning, preprint

[34] I.Derenyi,T.Geszti and G.Gyorgyi, Phys.Rev E 50, p.3192 (1994).

[35] M.Biehl and A.Mietzner, Europhys. Lett. 24 (1993) p.421.

[36] T.L.H.Watkin and J.P.Nadal, J.Phys.A.:Math.Gen. 27 (1994) p. 1899.

[37] E.Lootens and C.Van den Broeck, Europhys. Lett. 30 (1995) p.381

[38] : C. Van den Broeck and M. Bouten, Europhys. Lett. 22, p. 223 (1993).

[39] : D. Bollé and G.M. Shim, preprint,Nonlinear Hebbian training of the perceptron.

[40] : M. Bouten, J. Schietse and C. Van den Broeck, Phys. Rev. E 52, p.1958 (1995)

[41] : G. Györgyi, Phys. Rev. A 41 , p. 7097 (1990).

[42] : M. Opper and D. Haussler, Phys. Rev. Lett. 66 , p. 2677 (1991).

[43] : J. M. Parrondo and C. Van den Broeck, Europhys. Lett. 22, p. 319 (1993).

[44] : F.Vallet, Europhys. Lett. 8, (1989) p.747

[45] : T.L.H. Watkin, Europhys. Lett. 21, p. 871 (1993)

[46] E.Gardner and B.Derrida,J.Phys.A 21, p.271 (1988)

Publikatielijst

- K.Heyde, J.Schietse and C.De Coster, Proton 4p-4h intruder excitations in heavy even-even nuclei, Phys. Rev. C 44, p.2216 (1991)
- K.Heyde, C.De Coster and J.Schietse, Empirical proton-neutron interactions near closed shells: A simple shell-model approach, Phys. Rev. C 49, p.995 (1994)
- J.Iwanski and J.Schietse, Diluted neural networks with binary couplings: A replica symmetry breaking calculation of the storage capacity, Proceedings European Symposium on Artificial Neural Networks 1994, Editor M.Verleysen, D facto Publications, p.55-60
- J.Schietse, Neural Networks: The storage of patterns, Physicalia Magazine 16, p. 47-56 (1994)
- J.Schietse, M.Bouten and C.Van den Broeck, Optimal learning by gradient descent, Proceedings Belgian Neural Network Contact Group 1994, Editors E.de Bodt and M.Verleysen, p.89-96
- J.Iwanski, J.Schietse and M.Bouten, Replica symmetry breaking in a diluted network with binary couplings, Phys. Rev. E 52, p.888 (1995)
- M.Bouten, J.Schietse and C.Van den Broeck, Gradient descent learning in perceptrons : A review of its possibilities, Phys. Rev. E 52, p.1958 (1995)
- J.Schietse, M.Bouten and C.Van den Broeck, Training Binary Perceptrons by Clipping, Europhys. Lett. 32, p.279-284 (1995)