Two Approximations to Learning from Examples

L. Diambra[*]
Instituto de Física, Universidade de São Paulo
C.P. 66318, cep 05315-970, São Paulo, Brazil

Received 22 June, 1998

We investigate the learning of a rule from examples of the case of boolean perceptron. Previous studies of this problem have been made using the full quenched theory. We consider here two alternative approaches that can be applied easily. The two-replicas interactions approach considerably improves upon the well-known first-order approach. The mean field approach proved some results that have been obtained previously using the complex full quenched theory. Both approximations have been applied to both continuous weights and discrete weights perceptron.

I Introduction

The replica formulation of the statistical mechanics of disordered systems has become a major research tool in the study of complex systems [1]. In recent years, the learning from examples in feedforward neural networks has been exhaustively studied in the framework of statistical mechanics [2, 3, 4, 5]. In this field, the replica trick (RT) has become a very useful tool for the investigation of learning and generalization processes in neural networks. The popularity of the replica method is based mainly on the elegance and simplicity of the formulation of the so-called quenched theory.

In the context of learning in single-layer feedforward neural networks, the full quenched theory has been applied successfully within the framework of the replica symmetry (RS) and one-step replica symmetry breaking[4, 6, 7]. However, it is possible to extract some information in a simpler way, borrowing techniques from other fields of physics and applying them to the replicated Hamiltonian. Recent studies using high-temperature limit approximation and annealed approximation have been developed and applied to feedforward neural networks. These approaches are able to predict the right behavior in some cases, but cannot introduce the disorder effects produced by the randomness of the examples. These effects become essential with decreasing temperature.

In the present effort we wish to study the generalization process in the perceptron with boolean output, the so-called boolean perceptron (BP), using two different approximations: two-replicas approximation (TRA) obtained by recourse of a pertubative expansion, and mean field approximation (MFA). Our results can be extended to a perceptron with linear output.

We consider learning by a single-layered perceptron [8] within a statistical mechanics environment [9, 10, 11]. Our neural network (NN) has N input units S_i connected to a single output unit σ, whose state is given by $\sigma(\mathbf{S},\mathbf{W}) = g\left(N^{-1}\mathbf{W}\cdot\mathbf{S}\right)$, where $g(x)$ is the transfer function. For each set $\mathbf{W}$ of weights, the NN maps $\mathbf{S}$ onto σ. Learning is said to take place whenever the W_i are chosen so that σ closely approaches the desired, correct map $\sigma_0(\mathbf{S}) = g\left(N^{-1}\mathbf{W}_0\cdot\mathbf{S}\right)$. Within the supervised learning scheme [12] one reaches this goal by recourse to a cost function that is constructed on the basis of P examples $\{\mathbf{S}^l, \sigma_0(\mathbf{S}^l)\}$ with $l = 1, \ldots, P$. Here we assume that the inputs $\mathbf{S}^l$ are randomly selected according to probabilities $D(\mathbf{S})$, (we have considered here a Gaussian distribution) from the input spaces.

The learning process has been regarded as a stochastic dynamics, associated to the minimization of an energy function E_t, where the NN weights evolve according to a Langevin-like relaxation prescription that leads to a Gibbsian probability distribution for the weights [13, 1, 14]

$$P(\mathbf{W}) = Z^{-1}\exp[-\beta E_t(\mathbf{W})], \tag{1}$$

[*] e-mail: diambra@uspif1.if.usp.br

with $\beta = 1/T$ and T a "temperature" characterizing the noise level in the learning process. The normalization factor Z is the partition function. The training energy E_t is defined by

$$E_t\left(\mathbf{W}\right) = \sum_{l=1}^{P} \epsilon\left(\mathbf{W}, \mathbf{S}^l\right), \qquad (2)$$

where $\epsilon\left(\mathbf{W}, \mathbf{S}\right)$ is the mistake function, a measure of the deviation between actual and correct outputs. Here we focus our attention upon perceptrons with binary output, for which $\epsilon\left(\mathbf{W}, \mathbf{S}\right) = \theta\left(-N^{-1/2}\left(\mathbf{W}\cdot\mathbf{S}\right)\left(\mathbf{W}_0\cdot\mathbf{S}\right)\right)$, θ stands for the Heaviside function.

The remainder of the paper is organized as follows: Section II is devoted to a brief recapitulation of basic concepts concerning the replica formulation in the BP. The two approximations that interests us here (TRA and MFA), are derived in Section III. The thermodynamics of the BP, with Ising weights and continuous weights, are analyzed in Section IV. Finally, some conclusions are drawn in Section V.

II The Replica Method

The energy of the systems depends upon the particular training examples selected. Therefore, the associated "macroscopic" observables are evaluated by a double averaging procedure involving two spaces: thermal average over the weight space with probability distribution $P\left(\mathbf{W}\right)$, to be denoted by $\langle ...\rangle_T$ and a so-called "quenched average" over all possible *inputs,* to be represented by $\ll \ ... \ \gg \equiv \int \prod_l d\mu(\mathbf{S}^l)$, where $d\mu(\mathbf{S}^l)$ is some measure. Here we are using the standard Gaussian measure: $d\mu(\mathbf{S}) = D\left(\mathbf{S}\right) d\mathbf{S}$.

The NN free energy F is given in terms of the latter type of average by

$$F\left(T, P\right) = -T \ll \ln Z \gg, \qquad (3)$$

where

$$Z = \int d\mathbf{W} \exp[-\beta E_t\left(\mathbf{W}\right)]. \qquad (4)$$

The NN performance over the space of examples is characterized by the average generalization error ϵ_g, while the performance related to the training set is given by the average training error ϵ_t, i.e.

$$\begin{aligned} \epsilon_g\left(T, P\right) &= \ll \langle \varepsilon\left(\mathbf{W}\right)\rangle_T \gg, & (5) \\ \epsilon_t\left(T, P\right) &= P^{-1} \ll \langle E_t\left(\mathbf{W}\right)\rangle_T \gg . & (6) \end{aligned}$$

where $\varepsilon\left(\mathbf{W}\right) = \int d\mu(\mathbf{S})\epsilon\left(\mathbf{W}, \mathbf{S}\right)$ is the generalization function. Graphs of either $\epsilon_g\left(T, P\right)$ and $\epsilon_t\left(T, P\right)$ versus $\alpha = P/N$ are called *learning curves.*

The RT is the usual tool employed to evaluate the average over the examples [13, 14], and it originated within the context of spin glasses [15, 16]. The RT is recommended whenever it is feasible to evaluate averages of Z, but not the ones for $\ln Z$. The RT exploits the identity

$$\ll \ln Z \gg = \lim_{n\to 0} n^{-1} \ln \ll Z^n \gg, \qquad (7)$$

where Z^n can be regarded as the partition function of n identical non-interacting systems, copies of the original one. They are identified by the label $\gamma = 1, ..., n$. In performing the averaging process over the examples, coupling arises between the distinct copies.

From (3), (4) and (7) the free energy F becomes

$$F = -\beta^{-1} \lim_{n\to 0} \frac{1}{n} \ln \int \prod_{\gamma=1}^{n} d\mathbf{W}_\gamma \exp\left[-N\alpha H\left(\mathbf{W}_\gamma\right)\right], \qquad (8)$$

where the replicated Hamiltonian H is an intensive quantity that does not depend upon the number of examples N, and it is given by

$$H\left(\mathbf{W}_\gamma\right) = -\ln \int D\left(S\right) dS \exp\left[-\beta \sum_{\gamma=1}^{n} \epsilon\left(\mathbf{W}_\gamma, \mathbf{S}\right)\right]. \qquad (9)$$

The evaluation of H in the boolean perceptron is standard by now [4], so we present the results only

$$H\left(R_\gamma, Q_{\gamma\delta}\right) \ = \ -\ln \int \prod_{\gamma}^{n} \frac{d\widehat{x}_\gamma dx_\gamma}{2\pi} \int \frac{d\widehat{y}dy}{2\pi} \exp\left[-\beta \sum_{\gamma=1}^{n} \theta\left(-x_\gamma y\right) + i \sum_{\gamma}^{n} \widehat{x}_\gamma x_\gamma + i\widehat{y}y\right]$$

$$\times \exp\left(-\frac{1}{2}\sum_{\gamma,\delta}^{n}\widehat{x}_{\gamma}\widehat{x}_{\delta}Q_{\gamma\delta}-\widehat{y}\sum_{\gamma}^{n}\widehat{x}_{\gamma}R_{\gamma}-\frac{1}{2}\widehat{y}^{2}\right) \tag{10}$$

This replicated Hamiltonian depends on the weights through the order parameters R_γ and $Q_{\gamma\delta}$ given by

$$R_{\gamma}=N^{-1}\mathbf{W}_{\gamma}\cdot\mathbf{W}_{0},\qquad Q_{\gamma\delta}=N^{-1}\mathbf{W}_{\gamma}\cdot\mathbf{W}_{\delta}. \tag{11}$$

Since $H\left(R_\gamma, Q_{\gamma\delta}\right)$ depends on the weights only through parameters R_γ and $Q_{\gamma\delta}$ above defined, the replicated partition function can be written as an integral over these order parameter introducing auxiliary parameters

$$\ll Z^{n} \gg=\int\prod_{\gamma}^{n}\frac{dR_{\gamma}d\widehat{R}_{\gamma}}{2\pi i}\int\prod_{\gamma<\delta}^{n}\frac{dQ_{\gamma\delta}d\widehat{Q}_{\gamma\delta}}{2\pi i}\exp\left[N\left(S-\alpha H\right)\right],$$

where

$$\begin{aligned} S\left(R_{\gamma},Q_{\gamma\delta},\widehat{R}_{\gamma},\widehat{Q}_{\gamma\delta}\right) &= -\sum_{\gamma}^{n}\widehat{R}_{\gamma}R_{\gamma}-\sum_{\gamma<\delta}^{n}\widehat{Q}_{\gamma\delta}Q_{\gamma\delta}+ \\ & N^{-1}\ln\int\prod_{\gamma}^{n}d\mu\left(\mathbf{W}_{\gamma}\right)\exp\left[N\left(\sum_{\gamma}^{n}\widehat{R}_{\gamma}R_{\gamma}-\sum_{\gamma<\delta}^{n}\widehat{Q}_{\gamma\delta}Q_{\gamma\delta}\right)\right], \end{aligned} \tag{12}$$

is the logarithm of the density of replicated networks with the overlaps R_γ and $Q_{\gamma\delta}$.

In the thermodynamical limit $N \to \infty$ the integral (8) receives an overwhelming contribution from the minimum of the variables R_γ and $Q_{\gamma\delta}$.

Here some physical reasoning is needed in order to simplify things. Since the replicas have no a priori physical meaning, it is reasonable to assume that all replicas have the same overlap with the teacher NN and that, further, the overlaps between two of them are symmetric under permutation of the replica indices. This assumption constitutes the RS ansatz. Therefore, we have

$$\begin{aligned} Q_{\gamma\delta} &= \delta_{\gamma\delta}+\left(1-\delta_{\gamma\delta}\right)q, \\ R_{\gamma} &= R. \end{aligned} \tag{13}$$

Using the substitutions for R_γ and $Q_{\gamma\delta}$ (in a similar way for $\widehat{R}_\gamma$ and $\widehat{Q}_{\gamma\delta}$) in (10) and performing the integrals over $\widehat{x}_\gamma$ and $\widehat{y}_\gamma$, and then evaluating the limit $n \to 0$, the training error for the boolean perceptron takes the form

$$\epsilon_{t}\left(R,q\right)=-2\beta\int_{0}^{\infty}Dy\int_{-\infty}^{\infty}Dt\ln\left[e^{-\beta}+\left(1-e^{-\beta}\right)\frac{1}{2}\left[1-\operatorname{erf}\left(\frac{t\sqrt{q-R^{2}}-yR}{\sqrt{2\left(1-q\right)}}\right)\right]\right], \tag{14}$$

where $Dy = 1/\sqrt{2\pi}e^{-y^2/2}dy$ and $\operatorname{erf}(x)$ is the standard error function. The thermodynamical study of the problem would involve a considerable effort, so it should be useful to study a simpler approach. Borrowing ideas from other fields of physics we consider two approximations. Of course, we pay the customary price: the approximation is valid just for some appropiate range of β.

III Two approximation for BP

A. Two-replica approximation

It is our goal here to introduce a perturbative treatment that enables one to incorporate the disorder effects produced by the randomness in the examples. We shall consider an expansion of H given by (9) in powers of β and then consider the Hamiltonian that incorporates the two-replica interactions [17]

$$H\left[\mathbf{W}_\gamma\right] = \beta H_1 + \frac{1}{2}\beta^2 H_2 + O\left(\beta^3\right), \tag{15}$$

with

$$H_1 = \int D\left(\mathbf{S}\right) d\mathbf{S} \epsilon\left(\mathbf{W},\mathbf{S}\right), \tag{16}$$

H_1 represents the "non-random" part of the training energy, and coincides with the generalization function $\varepsilon\left(\mathbf{W}\right)$, which depends only upon the overlap R, for a BP is given by $1/\pi \cos^{-1}\left(R\right)$ [4, 5]. On the other hand, H_2 represent two-replica coupling arising from the randomness of the training examples. When T diminishes, this coupling becomes more and more important so that one needs to consider H_2 contributions. One has

$$H_2 = e\left(\mathbf{W}_\gamma\right) e\left(\mathbf{W}_\delta\right) - \int D\left(\mathbf{S}\right) d\mathbf{S} \epsilon\left(\mathbf{W}_\gamma,\mathbf{S}\right) \epsilon\left(\mathbf{W}_\delta,\mathbf{S}\right). \tag{17}$$

Of course, higher order terms in β are associated with three-replica coupling, four-replica ones and so on. Replicas can be regarded as *particles* with N degrees of freedom. The first term in (15) describes the coupling of the *particles* with an external field, while the second one represents *two-particle* interactions via an effective potential depending upon the Hamming distance between the replicas.

The H_2 contributions lead to consideration of the integral of correlation

$$C\left(R_\gamma, Q_{\gamma\delta}\right) = \int D\left(\mathbf{S}\right) d\mathbf{S} \epsilon\left(\mathbf{W}_\gamma,\mathbf{S}\right) \epsilon\left(\mathbf{W}_\delta,\mathbf{S}\right). \tag{18}$$

In second order, the replicated Hamiltonian in TRA for the BP reads (see details in the Appendix)

$$H_{TR} = \frac{\beta}{\pi}\sum_{\gamma}^{n} \cos^{-1}\left(R_\gamma\right) - \frac{\beta^2}{2}\sum_{\gamma\delta}^{n} C\left(R_\gamma, Q_{\gamma\delta}\right), \tag{19}$$

where second-order terms in the total number of replicas n have been eliminated. The relevant parameter here is $Q_{\gamma\delta}$, which does not appear at high temperature limits. The temperature T is associated with a coupling constant. It is reasonable to expect our expansion to yield an adequate treatment for $T > 1$. Using the RS approximation for R_γ, $Q_{\gamma\delta}$, $\widehat{R}_\gamma$ and $\widehat{Q}_{\gamma\delta}$, passing to the limit $n \to 0$, we are in a position to write

$$f = \alpha\epsilon_t - Ts, \tag{20}$$

where

$$\epsilon_t = \frac{1}{\pi}\cos^{-1}\left(R\right) - \frac{\beta}{4\pi}\left(\frac{\pi}{2} - \tan^{-1}\left(\frac{q}{\sqrt{1-q^2}}\right)\right), \tag{21}$$

$$s = \frac{1}{2}\left(q-1\right)\widehat{q} - R\widehat{R} + \int D\mathbf{z}\ln\int d\mu\left(\mathbf{W}\right)\exp\left[\mathbf{W}\cdot\left(\sqrt{\widehat{q}}\mathbf{z} + \mathbf{W}_0\widehat{R}\right)\right]. \tag{22}$$

B. Mean field approximation

In some cases, the coupling between replicas produces only minor changes in the learning curves and the phase diagrams. In other cases, such terms can lead to the appearance of qualitalively different phases at low temperatures. These phases are conveniently described by the properties of the matrix $Q_{\gamma\delta}$ which measures the overlap of the weights of two copies of the systems. Since the replicated Hamiltonian is invariant under permutation of the replica indices, one naively would expect that $Q_{\gamma\delta} = q$ for all $\gamma \neq \delta$, where q is given by

$$q = N^{-1} \ll \langle \mathbf{W} \rangle_T \cdot \langle \mathbf{W} \rangle_T \gg .$$

This characterizes the typical overlap of the solutions to the constraints posed by the examples. As α increases, more and more correlations are to be found between the different solutions, and q approaches unity. For $\alpha = \alpha_{cr}$, we have $q = 1$ and the concomitant degeneration is broken. This parameter is known as the Edward-Anderson parameter in spin glass theory, and reflects the degeneracy of the ground states. On the other hand, the expected value of the overlap with the teacher is given by $R = N^{-1} \ll \langle \mathbf{W} \rangle_T \gg \cdot \mathbf{W}_0$. Keeping this correspondence in mind, we are interested in considering an approximation like mean field theory. We substitute $q = R^2$ in (14), and the training error in MFA becomes

$$\epsilon_t = -2\beta \int_{-\infty}^{0} Dy \ln \left[e^{-\beta} + \left(1 - e^{-\beta}\right) \frac{1}{2} \left[1 - \mathrm{erf}\left(\frac{y\, R}{\sqrt{2\left(1 - R^2\right)}} \right) \right] \right] . \tag{23}$$

We can see that this approximation takes into account the degeneration of the ground states because it preserves the structure of the matrix $Q_{\gamma\delta}$ within the framework of the replica symmetry.

IV Analysis of the Results

A. Ising-weights perceptron

Now, evaluation of the expression (22) using the adequate constraint over weights space becomes mandatory. First we consider a BP with Ising weights. In this case, the adequate *a priori* measure of the weights is $d\mu(\mathbf{W}) = \prod_i dW_i \left[\delta(W_i - 1) + \delta(W_i + 1) \right]$ and the expression (22) becomes

$$s = -\frac{1}{2}(1 - q)\,\widehat{q} - R\widehat{R} + \int Dz \ln 2\cosh\left[\left(\sqrt{\widehat{q}} z + \widehat{R} \right) \right] . \tag{24}$$

The free energy function in TRA is given by (20) with ϵ_t and s given by (21) and (24), respectively. Extremalizing the free energy with respect to the parameters $R, \widehat{R}, q$ and $\widehat{q}$ and eliminating $\widehat{R}$ and $\widehat{q}$, we obtain the pertinent *saddle point* equations

$$\begin{aligned} R &= \int Dz \tanh\left(\sqrt{\frac{\beta^2 \alpha}{2\pi} \frac{1}{\sqrt{1 - q^2}}} z + \frac{\alpha\beta}{\pi} \frac{1}{\sqrt{1 - R^2}} \right) \\ q &= \int Dz \tanh^2\left(\sqrt{\frac{\beta^2 \alpha}{2\pi} \frac{1}{\sqrt{1 - q^2}}} z + \frac{\alpha\beta}{\pi} \frac{1}{\sqrt{1 - R^2}} \right) . \end{aligned} \tag{25}$$

In the limit $\beta \to 0$, we recover the high temperature results, with the mean field ansatz $q = R^2$, as a bonus. This relationship cannot be obtained in the first-order treatment, as it does not involve the parameter q. This result indicates that the mean field relationship is exact in the high temperature limit.

On the other hand, since the MFA Hamiltonian depends only on the overlap R, the free energy function is given (20) where ϵ_t is now given by (23), and s is the logarithm of the density of the perceptrons with overlap

R, given by

$$S\left(R\right)=-R\widehat{R}+\ln 2\cosh\widehat{R} \tag{26}$$

The corresponding saddle point equation is obtained extremalizing with respect to the variables R and $\widehat{R}$. Eliminating $\widehat{R}$ we obtain the thermodynamical equilibrium state

$$R=\tanh\left[\frac{2\alpha\left(e^{\beta}-1\right)}{\pi\left(e^{\beta}+1\right)\sqrt{1-R^{2}}}\right] \tag{27}$$

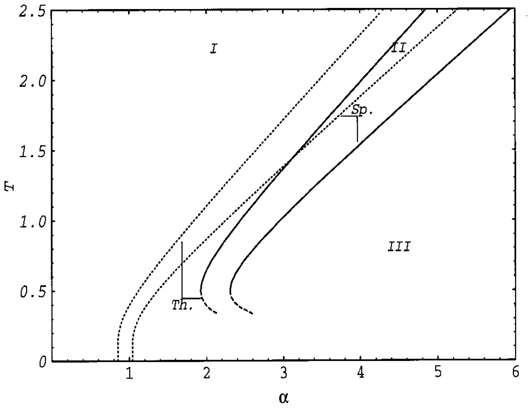

Figure 1. Phase diagram obtained with TRA and with MFA. The full lines correspond to TRA and the dashed lines to MFA. Thermodynamical (Th.) and spinodal (Sp.) curves has been indicated.

Both equations (25) and (27) describe the first-order transition from a state with poor generalization to a state perfect generalization with $R=1$. Fig. 1 depicts the phase diagrams for both approximations. At any fixed T, to left of thermodynamic transition line ($\alpha<\alpha_{th}$) there are two solutions, one with $R=1$, and one with $R<1$. The state of poor generalization ($R<1$) is the equilibrium state, while the state of perfect generalization ($R=1$) is metastable. In the region between the thermodynamic transition line and the spinodal line (first-order transition), the situation reverses, with $R=1$ becoming the equilibrium state, and $R<1$ the metastable state. To right of spinodal line (when $\alpha>\alpha_{sp}$), there is only one solution with $R=1$, there is no metastable state in this phase. Anomalies in the phase diagram arise at low temperatures ($T=0.5$), which is an effect of the approximation in the TRA. The phases of poor generalization, metastable, and perfect generalization are in indicated in the Fig.1 with I, II, and III, respectively.

The training errors in TRA are given by (21), while in MFA by (23). On the other hand, the generalization error is given by $\epsilon_g=1/\pi\cos^{-1}\left(R\right)$. Fig. 2 displays the learning curves for both approximations.

Some features of our approach deserve particular mention. In Fig. 2, the spinodal transition at $T=1$ takes place at $\alpha_{sp}=2.25$ in MFA. While in TRA the spinodal transition takes place at $\alpha_{sp}=2.95$, which agrees with the more elaborate complete quenched theory (CQT) [4]. Our results considerably improve upon the first-order approximation, for which $\alpha_{sp}=2.08$. In addition, unlike high temperature limit approximation, ϵ_t and ϵ_g are different in both TRA and MFA.

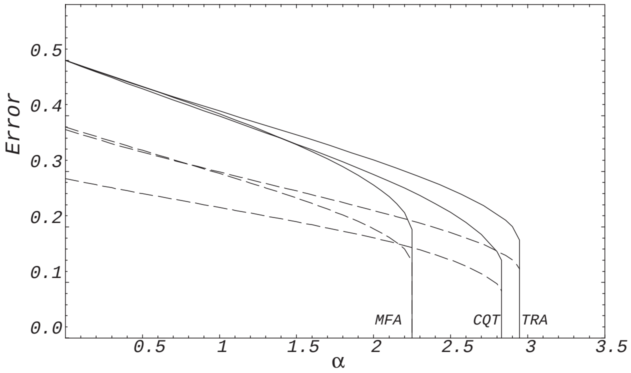

Figure 2. Learning curves for the Ising perceptron computed with TRA, MFA, and complete quenched theory (CQT) at $T=1$. The full lines correspond to the generalization errors, the dashed lines correspond to the training errors.

B. Spherical-weights perceptron

Now, we derive the equilibrium properties for the BP with spherical weights. In this case, the evaluation is somewhat more complicated. We write the *a priori* distribution as

$$d\mu\left(\mathbf{W}\right)=\prod_{i=1}^{N}\frac{dW_i}{\sqrt{2\pi e}}\int_{-i\infty}^{i\infty}\frac{d\lambda}{2\pi i}\exp\left[\lambda\left(\mathbf{W}\cdot\mathbf{W}-N\right)\right], \tag{28}$$

and the entropy (22) is now given by

$$s=-\frac{1}{2}+\frac{1}{2}\lambda+\frac{1}{2}q\widehat{q}-R\widehat{R}-\frac{1}{2}\ln\left(\widehat{q}+\lambda\right)+\frac{1}{2}\frac{\widehat{R}^{2}+\widehat{q}}{\widehat{q}+\lambda}. \tag{29}$$

The additional parameter is the Lagrange multiplier associated with the spherical constraint. Following the basic ideas presented in the previous sections, we derive the equilibrium states by extremalization of the free energy (where the entropy is given now by (29)) with respect to the parameters $R, \widehat{R}, q, \widehat{q}$ and λ. Eliminating $\widehat{R}, \widehat{q}$, and λ we obtain the pertinent *saddle point* equations for the TRA

$$\begin{aligned} R &= \frac{\alpha\beta\left(1-q\right)}{\pi\sqrt{1-R^2}}, \qquad (30)\\ q &= \frac{\alpha\beta^2}{\pi}\left[\frac{\left(1-q\right)^2}{2\sqrt{1-q^2}}+\frac{\alpha\left(1-q\right)^2}{\pi\left(1-R^2\right)}\right]. \end{aligned}$$

In the limit $\beta \to 0$ the parameter q is zero. This result indicates that the coupling between replicas in the perceptrons with spherical weights is weaker than in the case with Ising weights.

Similarly, the free energy function in the MFA can be written as (20) where ϵ_t is given by (23) and the entropy can be computed as the fraction of the weight space with an overlap R, which is simply the volume of the $(N-2)$-dimensional sphere with radius $\sqrt{1-R^2}$. In the thermodynamical limit we have $s = 1/2\ln\left(1-R^2\right)$, and the thermodynamical equilibrium is given by the concomitant *saddle point* equation

$$R = \frac{1}{\sqrt{1+\left(\frac{\pi\left(e^\beta+1\right)}{2\alpha\left(e^\beta-1\right)}\right)^2}}. \qquad (31)$$

Unlike the Ising-weights perceptron, the equations (30) and (31) do not manifest transition to perfect learning at any T and α values. The learning curves fall with a $1/\alpha$ tail for all T, in agreement with the correct power law. The asymptotic behavior of the generalization error in MFA is

$$\epsilon_g\left(\alpha,\beta\right) = \frac{\left(e^\beta+1\right)}{2\left(e^\beta-1\right)}\alpha^{-1}+O\left(\alpha^{-2}\right).$$

Note that at $T = 0$ the prefactor is 0.5, slightly inferior to the correct prefactor 0.625, while the annealed approximation predicts 1.

The training error and the generalization error are displayed in Fig. 3 at $T = 1$ for both approximations.

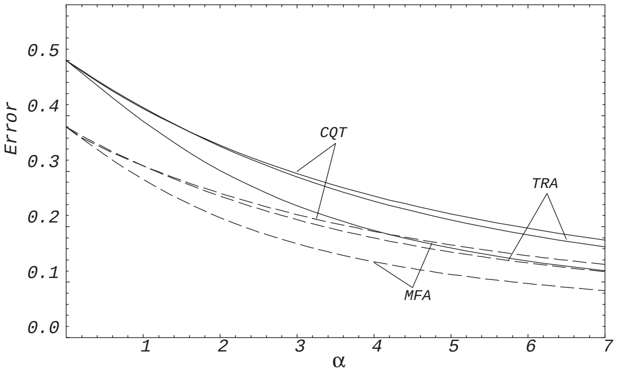

Figure 3. Learning curves for the Ising perceptron computed with TRA, MFA, and complete quenched theory (CQT) at $T = 1$. The full lines correspond to the generalization errors, the dashed lines correspond to the training errors.

V Conclusions

We have presented two approximations for the boolean perceptron. These approaches introduce, in differents way, the disorder effects produced by the random examples. They have been able to reproduce the behavior of the system within the range of β where RS is right. On the other hand, TRA allows us to establish that the coupling between replicas in perceptron with continuous weights is weaker than the one with discrete weights.

In any case, we hope to have convinced the reader that these techniques are satisfactory tools for investigating, at not too low temperatures, the thermodynamics of the learning process.

VI Acknowledgments

The author wishes to thank Coraci P. Malta for helpful conversations. The work of L. D. was supported by Conselho Nacional de Desenvolvimeto Científico e Tecnológico do Brazil, CNPq (Brazil).

VII Appendix

We undertake here the calculation of the correlations (18). We recast the first of them in the form

$$C\left(R_\gamma, Q_{\gamma\delta}\right)=\int d\mathbf{S}D(\mathbf{S})\theta\left(-N^{-1}\left(\mathbf{W}_0\cdot\mathbf{S}\right)\mathbf{W}_\gamma\cdot\mathbf{S}\right)\theta\left(-N^{-1}\left(\mathbf{W}_0\cdot\mathbf{S}\right)\mathbf{W}_\delta\cdot\mathbf{S}\right), \tag{32}$$

i.e.,

$$C=\frac{1}{4}\int D(\mathbf{S})d\mathbf{S}\int d\mathbf{r}\delta\left(x-N^{-1/2}\mathbf{W}_\gamma\cdot\mathbf{S}\right)\delta\left(y-N^{-1/2}\mathbf{W}_\delta\cdot\mathbf{S}\right)\delta\left(z-N^{-1/2}\mathbf{W}_0\cdot\mathbf{S}\right) \tag{33}$$

$$\times\theta\left(-xz\right)\theta\left(-yz\right). \tag{34}$$

By recourse to the representation $\delta\left(x\right)=\frac{1}{2\pi}\int dx'\exp\left(ixx'\right)$ of the delta function and remembering that $D(\mathbf{S})=\prod_i^N\left(2\pi\right)^{-1/2}\exp\left(-\frac{S_i^2}{2}\right)$, the integration process over $d\mathbf{S}$ leads to the (intermediate) result

$$\exp\left(-\left(\frac{1}{2}\mathbf{r}'\cdot\mathbf{r}'+x'zR_\gamma+y'z'R_\delta+x'y'Q_{\gamma\delta}\right)\right). \tag{35}$$

Performing the integrations over the variables $\mathbf{r}$ and $\mathbf{r}'$, we obtain

$$C=\frac{1}{2\pi}\left[\left(\frac{\pi}{2}+\tan^{-1}\left(\frac{Q_{\gamma\delta}}{\sqrt{1-Q_{\gamma\delta}^2}}\right)\right)-\left(\frac{\pi}{2}-\sin^{-1}\left(1-\left(R_\gamma\right)^2-\left(R_\delta\right)^2\right)\right)\right]. \tag{36}$$

Assuming the RS $R_\gamma=R$ and $Q_{\gamma\delta}=\begin{cases}1 & \sigma=\rho\\ q & \sigma\neq\rho\end{cases}$, we find for the n diagonal terms, on one hand,

$$\frac{n}{2}-\frac{n}{2\pi}\left(\frac{\pi}{2}-\sin^{-1}\left(1-2R^2\right)\right),$$

and, for the n^2-n terms, on the other hand (terms of second order in n neglected),

$$\frac{n}{2\pi}\left[\left(\frac{\pi}{2}+\tan^{-1}\left(\frac{q}{\sqrt{1-q^2}}\right)\right)-\left(\frac{\pi}{2}-\sin^{-1}\left(1-2R^2\right)\right)\right].$$

Therefore, within the RS ansatz the correlations are given by

$$\sum_{\gamma\delta}^{n}C\left(R_\gamma,Q_{\gamma\delta}\right)1_{\gamma\delta}=\frac{n}{2\pi}\left(\frac{\pi}{2}-\arctan\left(\frac{q}{\sqrt{1-q^2}}\right)\right). \tag{37}$$

These contributions come from the randomness in the examples.

References

[1] M. Mezard, G. Parisi, and M. A. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987).

[2] E. Gardner, J. Phys A. **21**, 257 (1988).

[3] E. Gardner, and B. Derrida, J. Phys A. **21**, 271 (1988).

[4] H.S. Seung, H. Sompolinsky, and N. Tishby, Phys. Rev. A **45**, 6056 (1992); H. Sompolinsky, N. Tishby, and H.S. Seung, Phys. Rev. Lett **65**, 1683 (1990).

[5] T. Watkin, A. Rau, and M. Biehl, Rev. Mod. Phys. **65**, 499 (1993).

[6] W. Krauth and M. Mezard, J. Phys. (Paris) **50**, 3057 (1989).

[7] J.F. Fontanari and R. Meir, J. Phys. A **26**, 1077 (1993).

[8] D.E. Rumelhart, and J.L. McClelland, *Parallel Distributed Processing*, (MIT, Cambridge, MA., 1986).

[9] N. Tishby, E. Levin, and S. Solla, in *Proceedings of the Internatinal Joint Conference on Neural Networks* (IEEE,New York,1989), Vol 2, pp. 403-409.

[10] E. Levin, N. Tishby, and S. Solla, Proc. IEEE **78**, 1568 (1990).

[11] J.A. Hertz, in *Statitical Mechanics of Neural Networks: Proceedings of the Eleventh Sitges Conference*, edited by L.Garrido (Springer, Berlin,1990).

[12] J. Schrager, T. Hogg, and B.A. Hubermann, Science **242**, 414 (1988).

[13] S.F. Edwards and P.W. Anderson, J. Phys. F **5**, 965 (1980).

[14] G. Parisi, J. Phys. A **13**,1101 (1980).

[15] D. Sherrington and S. Kirkpatrick, Phys. Rev. Lett. **35**, 1792 (1975).

[16] S. Kirkpatrick and D. Sherrington, Phys. Rev. B **17**, 4384 (1978).

[17] L. Diambra and A. Plastino, Phys. Rev. E **53**, 3970 (1996).