Learning from correlated examples in a perceptron

J. Phys. A: Math. Gen. 26 (1993) 3669–3679. Printed in the UK

Learning from correlated examples in a perceptron

Wojciech Tarkowski[†] and Maciej Lewenstein[‡][§]

[†] Centrum Fizyki Teoretycznej, Polskiej Akademii Nauk, Al. Lotników 32/46, 02–668 Warsaw, Poland

[‡] Centre d'Etudes de Saclay, Service des Photons, Atomes et Molécules, Bâtiment 22, 91191 Gif sur Yvette Cedex, France

Received 27 November 1992, in final form 13 April 1993

Abstract. Using the replica method we investigate learning from correlated data in a perceptron. Both types of association—'spatial'—within each example, and 'semantic'—among the different ones—are considered for both Boolean and continuous weights and output functions. 'Spatial' correlations of data may significantly improve the learning properties, whereas the 'semantic' ones do not practically influence them.

1. Introduction

The progress in the theory of neural networks has become possible due to an extensive use of statistical mechanics methods. One of the most significant and interesting problems in this area concerns the theory of learning. A new and powerful line of research was initiated by Gardner's famous paper [1], and then developed in various directions (see, for instance, [2–7]). One of these extensions, which has recently attracted much attention, relates to the so-called learning from examples [8–10]. This problem treats, roughly speaking, the dependence between two networks—the learning one (called a student) and the target one (a teacher)—as a function of learning properties (the amount of presented data, the temperature, etc).

In this paper we investigate learning from associated sets of data, which seems to be a more realistic situation than the previously studied case of learning from random examples. For instance, 'semantic' correlations arise when we consider learning of categories, subcategories of patterns, etc. It has been recently demonstrated that attractors observed in neurophysiological experiments are 'semantically' associated [11, 12]. Here the amount of correlation depends on the temporal interval between the learning times of the corresponding patterns. On the other hand, examples of 'spatial' associations are known from the theory of optical information processing. Visual data usually consist of locally correlated blocks (see [13]).

2. Formulation of the problem

We consider a single-layer perceptron with N input nodes and the output rule g (see [8]). The teacher (target) network is then defined by the updating rule

[§] Permanent address: Centrum Fizyki Teoretycznej Polskiej Akademii Nauk, Al. Lotników 32/46, 02–668 Warsaw, Poland.

$$\sigma_0^\mu = g\left(\sum_j \frac{J_j^0}{\sqrt{N}} S_j^\mu\right) \tag{1}$$

while the student (learning) one by

$$\sigma^\mu = g\left(\sum_j \frac{J_j}{\sqrt{N}} S_j^\mu\right) \tag{2}$$

where $j = 1, \ldots, N$ enumerates the sites, and $\mu = 1, \ldots, \alpha N$ the patterns.

The generalization cost (error) function, which effectively measures the realization of a learning rule (written as the teacher) by the student has the following form

$$\begin{aligned} \varepsilon(\{J_j\}) &= \int \mathrm{d}\mu(\{S_j^\mu\}) \tfrac{1}{2}[\sigma^\mu(\{J_j\}, \{S_j^\mu\}) - \sigma_0^\mu(\{J_j^0\}, \{S_j^\mu\})]^2 \\ &\equiv \int \mathrm{d}\mu(\{S_j^\mu\}) \varepsilon(\{J_j\}, \{S_j^\mu\}) \end{aligned} \tag{3}$$

with a certain measure $\mathrm{d}\mu(\{S_j^\mu\})$ in the example space. For a *realizable rule*, there then exists a weight vector $\{J_j^*\}$, such that

$$\varepsilon(\{J_j^*\}, \{S_j^\mu\}) = 0 \tag{4}$$

for the whole set $\{S_j^\mu\}$.

Function (3) seems to be the most appropriate for a network with a linear output, because of its proportionality to the difference between the local fields $h_\mu^{(0)} = \sum_j J_j^{(0)} S_j^\mu / \sqrt{N}$ of the student and teacher networks. It is the one traditionally used in the problem of learning a rule (see [14, 8, 10]). Other examples of the cost functions, however, have also been formulated and investigated [3, 15, 10, 9].

One then defines the quantities of interest—the generalization and the training errors

$$\varepsilon_{\mathrm{g}}(T, P) \equiv \langle\!\langle \langle \varepsilon(\{J_j\}) \rangle_T \rangle\!\rangle_S \tag{5}$$

$$\varepsilon_{\mathrm{t}}(T, P) \equiv P^{-1} \langle\!\langle \langle E(\{J_j\}) \rangle_T \rangle\!\rangle_S \tag{6}$$

where $\langle\!\langle \cdot \rangle\!\rangle_S$ denotes the quenched average over the distribution of the $\{S_j^\mu\}$, and $\langle \cdot \rangle_T$ means the thermal average with respect to the Gibbs probability distribution

$$\mathcal{P}(\{J_j\}) = \frac{\mathrm{e}^{-\beta E(\{J_j\})}}{\mathcal{Z}} \tag{7}$$

with $P = \alpha N$, $\beta = 1/T$. The training energy $E(\{J_j\})$ is defined by

$$\begin{aligned} E(\{J_j\}) &= \sum_\mu \varepsilon(\{J_j^*\}, \{S_j^\mu\}) \\ &\equiv \sum_\mu \tfrac{1}{2}[\sigma^\mu(\{J_j\}, \{S_j^\mu\}) - \sigma_0^\mu(\{J_j^0\}, \{S_j^\mu\})]^2 \end{aligned} \tag{8}$$

the partition function $\mathcal{Z}$ has the form

$$\mathcal{Z} = \int \mathrm{d}\mu(\{J_j\}) \mathrm{e}^{-\beta E(\{J_j\})} \tag{9}$$

while $d\mu(\{J_j\})$ is a measure in the weight space. The graphs of ε_g and ε_t are called the learning curves.

At the end of this section we write down, as usual, the general form of the replica-symmetric free energy density $\mathcal{F}$, which must be minimized in order to find the values of the proper order parameters

$$-\beta\mathcal{F} = \frac{1}{N}\langle\langle \ln \mathcal{Z}\rangle\rangle = G_1 + G_2 \tag{10}$$

where the functions G_1, G_2 will be specified for each case considered below.

3. 'Spatial' associations

We introduce 'spatial' correlations of the examples (see [16–18]) in the form

$$\langle\langle S_j^\mu \rangle\rangle = 0 \tag{11}$$

$$\langle\langle S_j^\mu S_{j'}^{\mu'} \rangle\rangle = \delta_{\mu\mu'} C_{jj'}. \tag{12}$$

Then we define the measure $d\mu(\{S_j^\mu\})$

$$d\mu(\{S_j^\mu\}) = \frac{(\prod_{\mu,j} dS_j^\mu)}{[(2\pi)^N \det \widehat{C}]^{\alpha N/2}} \exp\left\{-\frac{1}{2}\sum_{\mu,j,j'} S_j^\mu (C^{-1})_{jj'} S_{j'}^\mu\right\}. \tag{13}$$

Note that the inputs $\{S_j^\mu\}$ are continuous in our model. One, however, should stress that, by the validity of the central limit theorem, there is no difference between the continuous inputs introduced above and the binary ones in the thermodynamic limit $N \to \infty$.

Here we divide our analysis into the two subcases depending on the form of the output function g.

3.1. *Linear output function* $g(x) = x$

3.1.1. Continuous weights. We will consider the case of continuous couplings $\{J_j\}$ satisfying the spherical normalization constraint

$$d\mu(\{J_j\}) = \frac{(\prod_j dJ_j)\delta(\sum_j J_j^2 - N)}{\int(\prod_j dJ_j)\delta(\sum_j J_j^2 - N)}. \tag{14}$$

To this end one introduces the order parameters

$$q^{\alpha\beta} = \frac{1}{N}\sum_{j,j'} C_{jj'} J_j^\alpha J_{j'}^\beta \tag{15}$$

$$Q^\alpha = \frac{1}{N}\sum_{j,j'} C_{jj'} J_j^\alpha J_{j'}^\alpha \tag{16}$$

$$R^\alpha = \frac{1}{N}\sum_{j,j'} C_{jj'} J_j^\alpha J_{j'}^0 \tag{17}$$

$$m = \frac{1}{N}\sum_{j,j'} C_{jj'} J_j^0 J_{j'}^0 \tag{18}$$

and their conjugated counterparts $f^{\alpha\beta}$, F^{α}, and r^{α}. We also choose a torus topology for the network, and then the translational (rotational) invariance of the positively defined, symmetric association matrix $\widehat{C}$ (with the elements $C_{jj'}$)

$$C_{jj'} = C(|j - j'|). \tag{19}$$

In order to calculate the free energy density $\mathcal{F}$ (the function G_1) the Fourier transform can be thus simply performed (see [17] for more details). One then may write

$$q^{\alpha\beta} = (1/N)\sum_k C_k \widetilde{J}_k^{\alpha} \widetilde{J}_k^{\beta} \equiv \langle\!\langle C \widetilde{J}_C^{\alpha} \widetilde{J}_C^{\beta} \rangle\!\rangle_C \tag{20}$$

$$Q^{\alpha} = (1/N)\sum_k C_k (\widetilde{J}_k^{\alpha})^2 \equiv \langle\!\langle C (\widetilde{J}_C^{\alpha})^2 \rangle\!\rangle_C \tag{21}$$

$$R^{\alpha} = (1/N)\sum_k C_k \widetilde{J}_k^{\alpha} \widetilde{J}_k^{0} \equiv \langle\!\langle C \widetilde{J}_C^{\alpha} \widetilde{J}^{0} \rangle\!\rangle_C \tag{22}$$

$$m = \frac{1}{N}\sum_k C_k (\widetilde{J}_k^{0})^2 \equiv \langle\!\langle C (\widetilde{J}^{0})^2 \rangle\!\rangle_C \tag{23}$$

with $\{C_k\}$ being a set of eigenvalues of the correlation tensor $C(|j-j'|)$, and $\{\widetilde{J}_k\}$ the Fourier components of $\{J_j\}$. We also assume the self-averaging of the sums over $k = 1, \ldots, N$ *in the thermodynamic limit* $N \to \infty$, *so* $\langle\!\langle \cdot \rangle\!\rangle_C$ *denotes the average with respect to the* eigenvalue spectrum of the matrix $C_{jj'}$ (see [19, 17]). Finally we make a replica-symmetric ansatz $q^{\alpha\beta} = q$, $Q^{\alpha} = Q$, $r^{\alpha} = r$, $R^{\alpha} = R$, $f^{\alpha\beta} = f$, $F^{\alpha} = F$, and $E^{\alpha} = E$, where E^{α} is introduced to assure the normalization constraint (14).

In our model we require that the teacher weights $\{J_j^0\}$ do not depend on the learning process, and then, thanks to the spherical normalization of $\{J_j^0\}$ (14), and to the fact that $\langle\!\langle C \rangle\!\rangle_C = 1$, one may put $m = 1$.

After performing several Gaussian integrals, the functions G_1 and G_2 are obtained

$$\begin{aligned} G_1 = {} & \tfrac{1}{2}E + \tfrac{1}{2}qf - rR - \tfrac{1}{2}QF - \tfrac{1}{2}\langle\!\langle \ln(E + Cf - CF) \rangle\!\rangle_C \\ & + \frac{1}{2}\left\langle\!\!\left\langle \frac{Cf}{E + Cf - CF} \right\rangle\!\!\right\rangle_C + \frac{1}{2}\left\langle\!\!\left\langle \frac{C^2 r^2}{E + Cf - CF} \right\rangle\!\!\right\rangle_C \end{aligned} \tag{24}$$

$$G_2 = -\frac{\alpha}{2}\ln[1 + \beta(Q - q)] - \frac{\alpha}{2}\frac{\beta(q - 2R + 1)}{1 + \beta(Q - q)}. \tag{25}$$

The free energy density $\mathcal{F}$ must now be minimized over all the order parameters and their conjugated counterparts.

To evaluate the learning curves one has to have the appropriate form of the generalization and training errors for the case investigated (see [8])

$$\begin{aligned} \varepsilon_{\mathrm{g}} &= \tfrac{1}{2}(Q + 1) - R \\ \varepsilon_{\mathrm{t}} &= -\frac{1}{\alpha}\frac{\partial G_2}{\partial \beta}. \end{aligned} \tag{26}$$

The values of q, Q and R in the above expressions are to be calculated from the extremized sum of equations (24) and (25).

Both errors may now be calculated numerically, but an alternative way for reaching this task is to use the perturbative approach. This consists in evaluating the proper quantities in the form of series with respect to a small parameter (the details will be described in [20]). One may estimate in this manner the asymptotic behaviour of the generalization and training errors. This is exactly the approximation that we adopt in the present work. We are not generally interested in the behaviour of learning curves far from the perfect learning limit.

For the noiseless case ($T = 0$) near the point $\alpha = 1$ (when $q \to Q$ [17]), evaluating the saddle-point equations in powers of $(1-\alpha)$, one obtains

$$\varepsilon_{\mathrm{g}} = \frac{1-\alpha}{\langle\!\langle 1/C\rangle\!\rangle_C} + \left(\frac{\langle\!\langle 1/C^2\rangle\!\rangle_C}{\{\langle\!\langle 1/C\rangle\!\rangle_C\}^2} - 1\right)\frac{(1-\alpha)^2}{\langle\!\langle 1/C\rangle\!\rangle_C} + \mathcal{O}[(1-\alpha)^3] \tag{28}$$

for α close to 1, and $\varepsilon_{\mathrm{g}} = 0$ for $\alpha \geqslant 1$, while $\varepsilon_{\mathrm{t}} = 0$ for all values of α. From the Hölder inequality we then get

$$\left\{\left\langle\!\!\left\langle\frac{1}{C}\right\rangle\!\!\right\rangle_C\right\}^{1/2} = \{\langle\!\langle C\rangle\!\rangle_C\}^{1/2}\left\{\left\langle\!\!\left\langle\frac{1}{C}\right\rangle\!\!\right\rangle_C\right\}^{1/2} \geqslant \langle\!\langle 1\rangle\!\rangle_C = 1 \tag{29}$$

so the generalization error (28) is smaller than in the case of learning from random examples (where $\varepsilon_{\mathrm{g}} = 1-\alpha$). A transition to perfect learning still occurs, however, at $\alpha = 1$. The numerical solution of the saddle-point equations (obtained by differentiating equations (24) and (25)) indicates that expression (28) also holds for small values of α with quite good accuracy. For $\alpha = 0$, however, one should expect $\varepsilon_{\mathrm{g}} = 1$ independently from the value of T, the kind of association, and from the features of the output functions and weights.

For a finite temperature the perturbative approach (expansion in $1/\alpha$) gives

$$\varepsilon_{\mathrm{g}} = \frac{T}{2\alpha} + \left(4 - T\left\langle\!\!\left\langle\frac{1}{C}\right\rangle\!\!\right\rangle_C\right)\frac{T}{8\alpha^2} + \mathcal{O}(\alpha^{-3}) \tag{30}$$

$$\varepsilon_{\mathrm{t}} = \frac{T}{2\alpha} - \left\langle\!\!\left\langle\frac{1}{C}\right\rangle\!\!\right\rangle_C \frac{T^2}{8\alpha^2} + \mathcal{O}(\alpha^{-3}). \tag{31}$$

The first terms on the left-hand side of equations (30) and (31) agree with the theory of the so-called smooth networks (see [8]). In our case the values of both errors, however, are smaller. For non-zero temperature the transition to perfect learning does not occur (obviously $\varepsilon_{\mathrm{g}} \to 0$, when $\alpha \to \infty$).

3.2. Boolean output function $g(x) = \operatorname{sign} x$

All general expressions and statements from the previous subsection hold for the case of the binary output function the only difference being in the form of the function G_2, which now reads as

$$G_2 = 2\alpha\int_0^{\infty} \mathrm{D}y \int_{-\infty}^{\infty} \mathrm{D}t \ln[\mathrm{e}^{-\beta} + (1-\mathrm{e}^{-\beta})H(p)] \tag{32}$$

where

$$\mathrm{D}(\cdot) = \frac{\mathrm{d}(\cdot)}{\sqrt{2\pi}}\,\mathrm{e}^{-(\cdot)^2/2} \tag{33}$$

$$H(p) = \int_p^{\infty} \mathrm{D}x \tag{34}$$

and

$$p = \frac{t\sqrt{q - R^2} - yR}{\sqrt{Q - q}}. \tag{35}$$

For the binary output function the generalization error has the following form

$$\varepsilon_{\mathrm{g}} = \frac{2}{\pi} \cos^{-1}\left(\frac{R}{\sqrt{Q}}\right) \tag{36}$$

with the saddle-point values of R, Q.

Using, as in the previous case, the perturbative approach, one gets in the noiseless $\beta \to \infty$ limit

$$\begin{aligned}\varepsilon_{\mathrm{g}} \approx 2\sqrt{2}\Bigg[1 - \left(1 - \frac{\sqrt{2}}{\int \mathrm{D}t[H(t/\sqrt{2})]^{-1}}\right)\left\langle\!\!\left\langle \frac{1}{C} \right\rangle\!\!\right\rangle_C \\ + \left(1 - \frac{\sqrt{2}}{\int \mathrm{D}t[H(t/\sqrt{2})]^{-1}}\right)^2 \left(\left\langle\!\!\left\langle \frac{1}{C^2} \right\rangle\!\!\right\rangle_C - \left\{\left\langle\!\!\left\langle \frac{1}{C} \right\rangle\!\!\right\rangle_C\right\}^2\right)\Bigg]\frac{1}{\alpha} + O(\alpha^{-2}).\end{aligned} \tag{37}$$

The exact formula for the generalization error may be obtained using the Bürmann–Lagrange series expansion (see [21]), and does not practically differ from expression (37) in the asymptotic limit $\alpha \to \infty$. It is worth stressing that in the case of a given eigenvalue distribution of the correlation tensor $\widehat{C}$ (for instance, uniform or semicircular), an optimal width for it usually exists, for which the generalization error takes its minimal value.

The presence of 'spatial' associations may, therefore, significantly improve the generalization properties of a perceptron. Note that the best result for learning from random (uncorrelated) examples—using the Bayes algorithm reads as $\varepsilon_{\mathrm{g}} \approx 0.88/\alpha$ (see [22, 10]).

For a finite temperature the calculation is somewhat sophisticated, but it is easy to see that the learning curves decrease with a $1/\alpha$ tail for all T.

3.2.1. Boolean weights. We have not been able to estimate either of the considered errors for the case of 'spatially' associated data learned in a perceptron with binary couplings, because of difficulties in evaluating the explicit form of the function G_1.

4. 'Semantic' correlations

We introduce 'semantic' associations (see [16, 17]) by putting

$$\langle\!\langle S_j^\mu S_{j'}^{\mu'} \rangle\!\rangle = C_{\mu\mu'}\,\delta_{jj'} \tag{38}$$

for unbiased examples

$$\langle\!\langle S_j^\mu \rangle\!\rangle = 0. \tag{39}$$

One then defines the order parameters $q^{\alpha\beta}$ and R^{α}

$$q^{\alpha\beta} = \frac{1}{N}\sum_j J_j^{\alpha} J_j^{\beta} \tag{40}$$

$$R^{\alpha} = \frac{1}{N}\sum_j J_j^{\alpha} J_j^{0} \tag{41}$$

and $f^{\alpha\beta}$, r^{α} as their conjugated counterparts.

The measure $\mathrm{d}\mu(\{S_j^{\mu}\})$ is now taken to be

$$\mathrm{d}\mu(\{S_j^{\mu}\}) = \frac{(\prod_{\mu,j} \mathrm{d}S_j^{\mu})}{[(2\pi)^{\alpha N} \det \widehat{C}]^{N/2}} \exp\left\{-\frac{1}{2}\sum_{\mu,\mu',j} S_j^{\mu}(C^{-1})_{\mu\mu'} S_j^{\mu'}\right\}. \tag{42}$$

As usual, one assumes the translational invariance of the positively defined, symmetric correlation matrix $C_{\mu\mu'}$

$$C_{\mu\mu'} = C(|\mu - \mu'|) \tag{43}$$

and then introduces the Fourier transform in order to evaluate the function G_2.

Here, as in the previous section, we divide the analysis into subcases.

4.1. Linear output function $g(x) = x$

4.1.1. Continuous couplings. We first investigate the case of continuous weights $\{J_j\}$ with the measure defined in expression (14). The functions G_1 and G_2 (which have to be extremized with respect to the order parameters and their conjugated counterparts) take, after making a replica-symmetric ansatz for variables $q^{\alpha\beta}$, R^{α}, $f^{\alpha\beta}$ and r^{α}, the following form

$$G_1 = \frac{1}{2}E + \frac{1}{2}qf - rR - \frac{1}{2}\ln(E+f) + \frac{1}{2}\frac{r^2+f}{E+f} \tag{44}$$

$$G_2 = -\frac{\alpha}{2}\langle\!\langle \ln[1+\beta C(1-q)]\rangle\!\rangle_C - \frac{\alpha}{2}\left\langle\!\!\left\langle \frac{\beta C(q-2R+1)}{1+\beta C(1-q)} \right\rangle\!\!\right\rangle_C \tag{45}$$

where the symbol $\langle\!\langle\cdot\rangle\!\rangle_C$ now denotes the average over the eigenvalues of the tensor $C(|\mu-\mu'|)$. The generalization error is given by $\varepsilon_{\mathrm{g}} = 1-R$, with R taking the saddle-point value. Differentiating equations (44) and (45) over all the variables shows easily that in the noiseless limit the generalization error reads

$$\varepsilon_{\mathrm{g}} = 1 - \alpha \tag{46}$$

for $0 \leqslant \alpha < 1$, while $\varepsilon_{\mathrm{g}} = 0$ for $\alpha \geqslant 1$, and does not differ from the case of learning from random data. The training error $\varepsilon_{\mathrm{t}} = 0$ for all α.

The non-zero temperature expansion for asymptotic values of α (i.e. when $q, R \to 1$) gives

$$\varepsilon_{\mathrm{g}} = \frac{T}{2\alpha} + (4\langle\!\langle C^2\rangle\!\rangle_C - T)\frac{T}{8\alpha^2} + \mathcal{O}(\alpha^{-3}) \tag{47}$$

$$\varepsilon_{\mathrm{t}} = \frac{T}{2\alpha} - \frac{T^2}{8\alpha^2} + \mathcal{O}(\alpha^{-3}). \tag{48}$$

This result practically holds, as numerical simulations indicate, for an arbitrary range of T and not-too-small α. The presence of associations weakly increases the value of ε_{g}.

4.1.2. Binary couplings. We now consider the case of binary values of the weights $\{J_j\}$. The measure $\mathrm{d}\mu(\{J_j\})$ is now taken to be

$$\mathrm{d}\mu(\{J_j\}) = \frac{\prod_j [\mathrm{d}J_j\, \delta(J_j^2 - 1)]}{\int \prod_j [\mathrm{d}J_j\, \delta(J_j^2 - 1)]} \tag{49}$$

the function G_I takes the form

$$G_I = -\tfrac{1}{2}(1-q)f - rR + \int \mathrm{D}t \ln\cosh(t\sqrt{f} + r) \tag{50}$$

while the quantity G_2 is given by expression (45).

In the noiseless limit $T = 0$, as simple linear analysis shows, even one presented example guarantees perfect learning (i.e. $\varepsilon_g = 0$ for $\alpha \cong 0$).

For a finite temperature the perturbative approach gives the asymptotic values of ε_g and ε_t

$$\varepsilon_g = 2\mathrm{e}^{-2\alpha/T} + 2\left(\frac{8\alpha}{T^2}\langle\!\langle C^2\rangle\!\rangle_C - 1\right)\mathrm{e}^{-4\alpha/T} + \mathcal{O}(\mathrm{e}^{-6\alpha/T}) \tag{51}$$

$$\varepsilon_t = 2\mathrm{e}^{-2\alpha/T} + 2\left(\frac{8\alpha}{T^2}\langle\!\langle C^2\rangle\!\rangle_C - \frac{4}{T}\langle\!\langle C^2\rangle\!\rangle_C - 1\right)\mathrm{e}^{-4\alpha/T} + \mathcal{O}(\mathrm{e}^{-6\alpha/T}) \tag{52}$$

which, however, depend weakly (through the $\langle\!\langle C^2\rangle\!\rangle_C$) on the eigenvalue distribution of the correlation matrix $\widehat{C}$.

4.2. *Boolean output function $g(x) = \mathrm{sign}\, x$*

4.2.1. Continuous weights. At the beginning we examined the case of continuous couplings with the measure given by expression (14).

The function G_I is still of the form given by equation (44), whereas G_2 generally reads as

$$G_2 = \alpha \ln \int \mathrm{d}\mu(\{S_j^\mu\}) \exp\left\{-\beta \sum_{\alpha,\mu}\left[\Theta\Big(\sum_j \frac{J_j^\alpha}{\sqrt{N}} S_j^\mu\Big) - \Theta\Big(\sum_j \frac{J_j^0}{\sqrt{N}} S_j^\mu\Big)\right]^2\right\}. \tag{53}$$

It may be easily checked (see [20]) that in the case of asymptotic behaviour $q \to 1$ the generalization and training errors differ from their counterparts obtained for random data in terms of the order of $\mathcal{O}(\alpha^{-2})$. After a simple calculation of the average over $\{S_j^\mu\}$, and shifting the variables (see [8, 20]), one may introduce the new ones ($\bar{x}_\mu^\alpha = x_\mu^\alpha/(1-q)$) and, in this way, change the integral limits, that are, thanks to the presence of the Θ-function in expression (53), of the form:

$$\ln C(\beta) \int_0^\infty \prod_\mu \mathrm{d}y_\mu \exp(-y\widehat{C}^{-1}y) \int_{-Ry_\mu}^\infty \prod_{\alpha,\mu} \mathrm{d}x_\mu^\alpha \exp\{-(\text{'quadratic form' of } x_\mu^\alpha)\}.$$

The proper rescaling of the variables x_μ^α, y_μ implies that at least the terms $\mathcal{O}(\alpha^{-1})$ do not depend on the association tensor $\widehat{C}$. Obviously, the higher-order perturbative corrections to the known result: $\varepsilon_g = 1.25/\alpha + \mathcal{O}(\alpha^{-2})$ for $T = 0$, and learning with $1/\alpha$ tail for $T \neq 0$ (see [8])—are not easy to obtain by a straightforward calculation, but the numerical solutions of the saddle-point equations indicate that they are quite small.

One should add that both errors behave similarly in the case of other cost functions, namely for Gardner–Derrida, perceptron and relaxation functions [3, 15, 9, 20].

4.2.2. Binary couplings. A similar situation to that above also holds for binary weights, but only in the replica-symmetric (RS) regime. In such a case the transition to perfect learning at an arbitrary temperature (for $T = 0$, $\alpha_c = 1.245$) exists. The higher-order perturbative corrections cannot be evaluated easily, but we expect, from the outcomes of the previous stages of this paper, a slightly greater value for α for the transition. However, breaking the RS in the way proposed in [4] implies, roughly speaking, that we must not take the limit $q \to 1$, because the value of q is now determined by the proper saddle-point equations and the entropy function vanishing. In this case the calculation of ε_g and ε_t is by no means easy, but one may hope that the asymptotic values of both errors do not differ too much from those obtained for training from random examples, where the transition to perfect learning (at the noiseless limit) still occurs at $\alpha = 1.245$.

5. Learning unrealizable rules

In this case we deal with a situation which is quite similar to that described above, i.e. semantic correlations do not significantly influence the learning process (see [20, 8]), whereas the spatial ones change some of its properties, but their quantitative account is easy only in the case of the so-called unrealizable threshold ([8]), where the teacher transfer function is changed to be $g(x) = x + \phi$. For such a situation the function G_2 contains the additional term

$$G_2^{\mathrm{add}} = -\frac{\alpha}{2}\frac{\beta\phi^2}{1+\beta(Q-q)} \tag{54}$$

and the generalization error in the noiseless $T = 0$ limit is given by

$$\varepsilon_g = \frac{\phi^2}{2} + \frac{1-\alpha}{\langle\!\langle 1/C\rangle\!\rangle_C} + \left(\frac{\langle\!\langle 1/C^2\rangle\!\rangle_C}{\{\langle\!\langle 1/C\rangle\!\rangle_C\}^2} - 1\right)\frac{(1-\alpha)^2}{\langle\!\langle 1/C\rangle\!\rangle_C} + \mathcal{O}[(1-\alpha)^3] \tag{55}$$

with $\alpha < 1$, and $\varepsilon_g = \frac{1}{2}\phi^2$, while $\alpha \geqslant 1$. The training error $\varepsilon_t = 0$ for approximately (expansion in powers of ϕ)

$$0 \leqslant \alpha \leqslant 1 - \sqrt{\left\langle\!\!\left\langle\frac{1}{C}\right\rangle\!\!\right\rangle_C}\,|\phi| + \frac{\langle\!\langle 1/C^2\rangle\!\rangle_C}{\langle\!\langle 1/C\rangle\!\rangle_C}\frac{\phi^2}{2} + \mathcal{O}(\phi^3) \equiv \alpha_c. \tag{56}$$

On the other hand, if $\alpha > \alpha_c$, $q \to 1$ as $T \to 0$; in such a case the quantity $\beta(1-q)$ remains finite and ε_t grows from 0 to the value

$$\varepsilon_t = \frac{\phi^2}{2} - \frac{\phi^2}{2}\frac{1}{\alpha} + \phi^4\left\langle\!\!\left\langle\frac{1}{C}\right\rangle\!\!\right\rangle_C\frac{1}{8\alpha^2} + \mathcal{O}(\alpha^{-3}) \tag{57}$$

in the asymptotic limit $\alpha \to \infty$.

For a non-zero temperature we obtain

$$\varepsilon_g = \frac{\phi^2}{2} + \frac{T+\phi^2}{2\alpha} + \left(4\phi^2 - 4T\phi^2\left\langle\!\!\left\langle\frac{1}{C}\right\rangle\!\!\right\rangle_C + 4T - 3\phi^4\left\langle\!\!\left\langle\frac{1}{C}\right\rangle\!\!\right\rangle_C - T^2\left\langle\!\!\left\langle\frac{1}{C}\right\rangle\!\!\right\rangle_C\right)\frac{1}{8\alpha^2} + \mathcal{O}(\alpha^{-3}) \tag{58}$$

$$\varepsilon_t = \frac{\phi^2}{2} + \frac{T-\phi^2}{2\alpha} + (\phi^4 - T^2)\left\langle\!\!\left\langle\frac{1}{C}\right\rangle\!\!\right\rangle_C\frac{1}{8\alpha^2} + \mathcal{O}(\alpha^{-3}). \tag{59}$$

In the case of semantic associations one must add to the function G_2 the term

$$G_2^{\mathrm{add}} = -\frac{\alpha}{2}\frac{\beta\phi^2}{1+\beta C_0(1-q)} \tag{60}$$

with

$$C_0 = \sum_{(\mu-\mu')} C(|\mu-\mu'|). \tag{61}$$

The $T = 0$ result then has the following form

$$\varepsilon_{\mathrm{g}} = \tfrac{1}{2}\phi^2 + 1 - \alpha \tag{62}$$

for $\alpha < 1$, and $\varepsilon_{\mathrm{g}} = \frac{1}{2}\phi^2$, when $\alpha \geqslant 1$, while the training error $\varepsilon_{\mathrm{t}} = 0$ for

$$0 \leqslant \alpha \leqslant \frac{1}{2}2 + \frac{\phi^2}{C_0} - \frac{|\phi|}{\sqrt{C_0}}\sqrt{4+\frac{\phi^2}{C_0}} \equiv \alpha_{\mathrm{c}} \tag{63}$$

and, if $\alpha > \alpha_{\mathrm{c}}$, ε_{t} grows from 0 to its asymptotic value

$$\varepsilon_{\mathrm{t}} = \frac{\phi^2}{2} - \frac{C_0\phi^2}{2}\frac{1}{\alpha} + (4C_0^2\phi^2 - 4C_0\phi^2\langle\langle C^2\rangle\rangle_C + C_0^2\phi^4)\frac{1}{8\alpha^2} + \mathcal{O}(\alpha^{-3}) \tag{64}$$

in the $\alpha \to \infty$ limit. One should note that α_{c} in this model (unlike the case of realizable rules) has a similar meaning to that in the problem of storage capacity (see [8]).

For finite temperature $T > 0$ we get

$$\begin{aligned}\varepsilon_{\mathrm{g}} = \frac{\phi^2}{2} + \frac{T + C_0\phi^2}{2\alpha} &+ (12C_0\phi^2\langle\langle C^2\rangle\rangle_C - 3C_0^2\phi^4 - 8C_0^2\phi^2 \\ &+ 4T\langle\langle C^2\rangle\rangle_C - 4TC_0\phi^2 - T^2)\frac{1}{8\alpha^2} + \mathcal{O}(\alpha^{-3})\end{aligned} \tag{65}$$

$$\varepsilon_{\mathrm{t}} = \frac{\phi^2}{2} + \frac{T - C_0\phi^2}{2\alpha} + (4C_0^2\phi^2 - 4C_0\phi^2\langle\langle C^2\rangle\rangle_C + C_0^2\phi^4 - T^2)\frac{1}{8\alpha^2} + \mathcal{O}(\alpha^{-3}). \tag{66}$$

6. Conclusions

Summarizing, we have shown that while the presence of spatial correlations in learned examples may strongly influence the learning and generalization abilities of a perceptron, the semantic ones do not significantly change them. This is in contrast with the results concerning the storage of associated patterns in a perceptron with Boolean output and continuous weights. For that case the spatial critical curves are quite similar to the Gardner's original one [1, 16, 17, 18], while the semantic curves may diverge in the limit of large correlation lengths (see [16, 17]).

It is worth adding that the difference mentioned above (which occurs between both types of data association) in learning and generalization abilities of a perceptron is caused by the existence of two types of connection matrix. In the case of semantically correlated data the weights $\{J_j\}$ are distributed similarly to the one of remembering random examples. On the other hand, for training from spatially associated patterns the couplings tensor has a certain structure (cf [18]). It is, generally speaking, easier for a perceptron with such a connection matrix, which was formed in the learning process, to recognize in an errorless manner the new presented data which are constructed (correlated) in the same way, as previously learned ones.

Acknowledgments

This paper has been financed by KBN Grant No. 2–2417–92–03. ML thanks the Centre d'Etudes de Saclay for hospitality and the French Ministry of Research and Technology for support.

References

[1] Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 257–70

[2] Amit D J 1989 *Modeling Brain Function: The World of Attractor Neural Networks* (Cambridge: Cambridge University Press)

[3] Gardner E and Derrida B 1988 *J. Phys. A: Math. Gen.* **21** 271–84

[4] Krauth W and Mézard M 1989 *J. Physique* **50** 3057–66

[5] Tarkowski W, Komarnicki M and Lewenstein M 1991 *J. Phys. A: Math. Gen.* **24** 4197–217

[6] Barkai E, Hansel D and Sompolinsky H 1992 *Phys. Rev.* A **45** 4146–61

[7] Griniasty M and Grossman T 1992 *Phys. Rev.* A **45** 8924–37

[8] Seung H S, Sompolinsky H and Tishby N 1992 *Phys. Rev.* A **45** 6056–91

[9] Meir R and Fontanari J F 1992 *Phys. Rev.* A **45** 8874–84

[10] Watkin T L H, Rau H and Biehl M 1993 *Rev. Mod. Phys.* in press

[11] Miyashita Y 1988 *Nature* **335** 817–19

[12] Griniasty M, Tsodyks M V and Amit D J 1992 *Preprint* Universitá di Roma 856

[13] Fassnacht C and Zippelius A 1991 *Network* **2** 63–84

[14] Rumelhart D E and McClelland J L (eds) 1986 *Parallel Distributed Processing: Explorations in Microstructure of Cognition* vol I and II (Cambridge, MA: MIT)

[15] Griniasty M and Gutfreund H 1991 *J. Phys. A: Math. Gen.* **24** 715–34

[16] Lewenstein M and Tarkowski W 1993 *Phys. Rev.* A **46** 2139–42

[17] Tarkowski W and Lewenstein M 1993 *J. Phys. A: Math. Gen.* in press

[18] Monasson R 1992 *J. Phys. A: Math. Gen.* **25** 3701–20

[19] Tarkowski W and Lewenstein M 1992 *J. Phys. A: Math. Gen.* **25** 6251–64

[20] Tarkowski W and Lewenstein M 1993 Generalization error in learning from correlated examples in a perceptron (in preparation)

[21] Whittaker E T and Watson G N 1935 *A Course of Modern Analysis* (Cambridge: Cambridge University Press)

[22] Opper M and Haussler D 1991 *Phys. Rev. Lett.* **66** 2677–80